In J.R. Busemeyer, J.T. Townsend, Z. Wang, & A. Eidels, Eds, Oxford Handbook of Computational and Mathematical Psychology. Oxford University Press, 2015.

CHAPTER 9

Structure and Flexibility in Bayesian Models of Cognition

Joseph L. Austerweil, Samuel J. Gershman, Joshua B. Tenenbaum, and Thomas L. Griffiths

Abstract

Probability theory forms a natural framework for explaining the impressive success of people at solving many difficult inductive problems, such as learning words and categories, inferring the relevant features of objects, and identifying functional relationships. Probabilistic models of cognition use Bayes's rule to identify probable structures or representations that could have generated a set of observations, whether the observations are sensory input or the output of other psychological processes. In this chapter we address an important question that arises within this framework: How do people infer representations that are complex enough to faithfully encode the world but not so complex that they "overfit" noise in the data? We discuss nonparametric Bayesian models as a potential answer to this question. To do so, first we present the mathematical background necessary to understand nonparametric Bayesian models. We then delve into nonparametric Bayesian models for three types of hidden structure: clusters, features, and functions. Finally, we conclude with a summary and discussion of open questions for future research.

Key Words: inductive inference, Bayesian modeling, Nonparametrics, Bias-variance tradeoff, Categorization, Feature representations, Function learning, Clustering

Introduction

Probabilistic models of cognition explore the mathematical principles behind human learning and reasoning. Many of the most impressive tasks that people perform—learning words and categories, identifying causal relationships, and inferring the relevant features of objects—can be framed as problems of inductive inference. Probability theory provides a natural mathematical framework for inductive inference, generalizing logic to incorporate uncertainty in a way that can be derived from various assumptions about rational behavior (e.g., Jaynes, 2003).

Recent work has used probabilistic models to explain many aspects of human cognition, from memory to language acquisition (for a representative sample, see Chater and Oaksford, 2008). There are existing tutorials on some of the key mathematical ideas behind this approach (Griffiths and Yuille, 2006; Griffiths, Kemp, & Tenenbaum, 2008a) and its central theoretical commitments (Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum, Kemp, Griffiths, & Goodman, 2010a; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010a). In this chapter, we focus on a recent development in the probabilistic approach that has received less attention—the capacity to support both structure and flexibility.

One of the striking properties of human cognition is the ability to form structured representations in a flexible way: we organize our environment into meaningful clusters of objects, identify discrete features that those objects possess, and learn relationships between those features, without

apparent hard constraints on the complexity of these representations. This kind of learning poses a challenge for models of cognition: how can we define models that exhibit the same capacity for structured, flexible learning? And how do we identify the right level of flexibility, so that we only postulate the appropriate level of complexity? A number of recent models have explored answers to these questions based on ideas from nonparametric Bayesian statistics (Sanborn, Griffiths, & Navarro, 2010; Austerweil & Griffiths 2013; see Gershman and Blei 2012 for a review), and we review the key ideas behind this approach in detail.

Probabilistic models of cognition tend to focus on a level of analysis that is more abstract than that of many of the other modeling approaches discussed in the chapters of this handbook. Rather than trying to identify the cognitive or neural mechanisms that underlie behavior, the goal is to identify the abstract principles that characterize how people solve inductive problems. This kind of approach has its roots in what Marr (1982) termed the *computational level*, focusing on the goals of an information processing system and the logic by which those goals are best achieved, and was implemented in Shepard's (1987) search for universal laws of cognition and Anderson's (1990) method of *rational analysis*. But it is just the starting point for gaining a more complete understanding of human cognition—one that tells us not just why people do the things they do, but how they do them. Although we will not discuss it in this chapter, research on probabilistic models of cognition is beginning to consider how we can take this step, bridging different levels of analysis. We refer the interested reader to one of the more prominent strategies for building such a bridge, namely rational process models (Sanborn et al., 2010).

The plan of this chapter is as follows. First, we introduce the core mathematical ideas that are used in probabilistic models of cognition—the basics of Bayesian inference—and a formal framework for characterizing the challenges posed by flexibility. We then turn to a detailed presentation of the ideas behind nonparametric Bayesian inference, looking at how this approach can be used for learning three different kinds of representations—clusters, features, and functions.

Mathematical Background

In this section, we present the necessary mathematical background for understanding nonparametric Bayesian models of cognition. First, we describe the basic logic behind using Bayes' rule for inductive inference. Then, we explore two of the main types of hypothesis spaces for possible structures used in statistical models: parametric and nonparametric models.[1] Finally, we discuss what it means for a nonparametric model to be "Bayesian" and propose nonparametric Bayesian models as methods combining the benefits of both parametric and nonparametric models. This sets up the remainder of the article, where we compare the solution given by nonparametric Bayesian methods to how people (implicitly) solve this dilemma when learning associations, categories, features, and functions.

Basic Bayes

After observing some evidence from the environment, how should an agent update her beliefs in the various structures that could have produced the evidence? Given a set of candidate structures and the ability to describe the degree of belief in each structure, Bayes's rule prescribes how an agent should update her beliefs across many normative standards (Oaksford and Chater, 2007; Robert, 1994). Bayes's rule simply states that an agent's belief in a structure or hypothesis h after observing data d from the environment, the *posterior* $P(h|d)$, should be proportional to the product of two terms: her prior belief in the structure, the *prior* $P(h)$, and how likely the observed data d would be had it been produced by the candidate structure, called the *likelihood* $P(d|h)$. This is given by

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')},$$

where $\mathcal{H}$ is the space of possible hypotheses or latent structures. Note that the summation in the denominator is the normalization constant, which ensures that the posterior probability is still a valid probability distribution (sums to one). In addition to specifying how to calculate the posterior probability of each hypothesis, a Bayesian model prescribes how an agent should update her belief in observing new data d_{new} from the environment given the previous observations d

$$\begin{aligned} P(d_{\text{new}}|d) &= \sum_h P(d_{\text{new}}|h)P(h|d) \\ &= \sum_h P(d_{\text{new}}|h)\frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}. \end{aligned}$$

The fundamental assumptions of Bayesian models (i.e., what makes them "Bayesian") are (a) agents express their expectations over structures as probabilities and (b) they update their expectations according

to the laws of probability. These are not uncontroversial assumptions in psychology (e.g., Bowers and Davis 2012; Jones and Love 2011; Kahneman et al. 1982; McClelland et al. 2010, but also look at the replies Chater et al. 2011; Griffiths, Chater, Norris, & Pouget 2012; Griffiths, Chater, Kemp, Perfors, & Tenenbaum 2010b). However, they are extremely useful because they provide methodological tools for exploring the consequences of adopting different assumptions about the kind of structure that appears in the environment.

Parametric and Nonparametric

One of the first steps in formulating a computational model is a specification of the possible structures that could have generated the observations from the environment, the hypothesis space $\mathcal{H}$. As discussed in the Basic Bayes subsection, each hypothesis h in a hypothesis space $\mathcal{H}$ is defined as a probability distribution over the possible observations. So, specifying the hypothesis space amounts to defining the set of possible distributions over events that the agent could observe. To make the model Bayesian, a prior distribution over those hypotheses also needs to be specified.

From a statistical point of view, a Bayesian model with a particular hypothesis space (and prior over those hypotheses) is a solution to the problem of *density estimation*, which is the problem of estimating the probability distribution over possible observations from the environment.[2] In fact, it is the optimal solution given that the hypothesis space faithfully represents how the environment produces observations, and the environment randomly selects a hypothesis to produce observations with probability proportional to the prior distribution.

In general, a probability distribution is a function over the space of observations, which can be continuous, and thus is specified by an infinite number of parameters. So, density estimation involves identifying a function specified by an infinite number of parameters, as theoretically, it must specify the probability of each point in a continuous space. From this perspective, a function is analogous to a hypothesis and the space of possible functions constructed by varying the values of the parameters defines a hypothesis space. Different types of statistical models make different assumptions about the possible functions that define a density, and statistical inference amounts to estimating the parameters that define each function.

The statistical literature offers a useful classification of different types of probability density functions, based on the distinction between *parametric* and *nonparametric* models (Bickel and Doksum, 2007). Parametric models take the set of possible densities to be those that can be identified with a fixed number of parameters. An example of a parametric model is one that assumes the density follows a Gaussian distribution with a known variance, but unknown mean. This model estimates the mean of the Gaussian distribution based on observations and its estimate of the probability of new observations is their probability under a Gaussian distribution with the estimated mean. One property of parametric models is that they assume there exists a fixed set of possible structures (i.e., parameterizations) that does not change regardless of the amount of data observed. For the earlier example, no matter how much data the model is given that is inconsistent with a Gaussian distribution (e.g., a bimodal distribution), its density estimate would still be a Gaussian distribution because it is the only function available to the model.

In contrast, nonparametric models make much weaker assumptions about the family of possible structures. For this to be possible, the number of parameters of a nonparametric model increases with the number of data points observed. An example of a nonparametric statistical model is a Gaussian kernel model, which places a Gaussian distribution at each observation and its density estimate is the average over the Gaussian distributions associated with each observation. In essence, the parameters of this statistical model are the observations, and so the parameters of the model grow with the number of data points. Although *nonparametric* suggests that nonparametric models do not have any parameters, this is not the case. Rather, the number of parameters in a nonparametric model is not fixed with respect to the amount of data.

One domain within cognitive science where the distinction between parametric and nonparametric models has been useful is category learning (Ashby and Alfonso-Reese, 1995). The computational problem underlying category learning is identifying a probability distribution associated with each category label. Prototype models (Posner and Keele, 1968; Reed, 1972), approach this problem parametrically, by estimating the mean of a Gaussian distribution for each category. Alternatively, exemplar models (Medin and Schaffer, 1978; Nosofsky, 1986) are nonparametric, using each observation as a parameter; each category's density estimate for

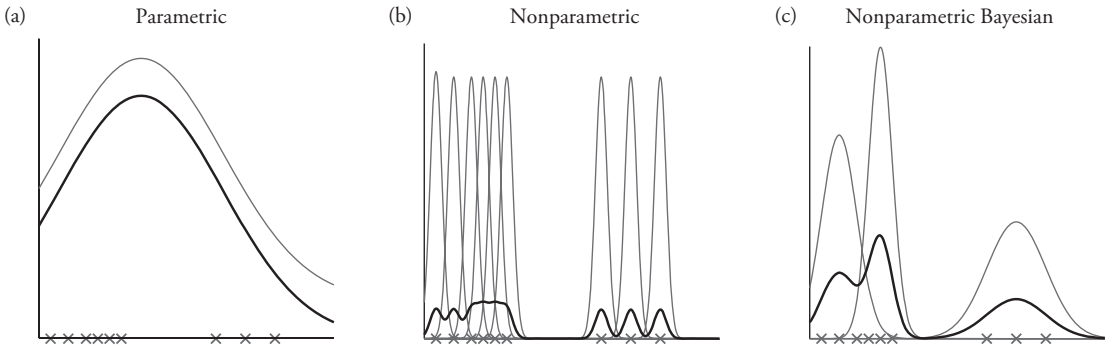

Fig. 9.1 Density estimators from the same observations (displayed in blue) for three types of statistical models: (a) parametric, (b) nonparametric, and (c) nonparametric Bayesian models. The parametric model estimates the mean of a Gaussian distribution from the observations, which results in the Gaussian density (displayed in black). The nonparametric model averages over the Gaussian distributions centered at each data point (in red) to yield its density estimate (displayed in black). The nonparametric Bayesian model puts the observations into three groups, each of which gets its own Gaussian distribution with mean centered at the center of the observations (displayed in red). The density estimate is formed by averaging over the Gaussian distributions associated with each displayed in black.

an new observation is a function of the sum of the distances between the new observation to previous observations. See Figure 9.1(a) and 9.1(b) for examples of parametric and nonparametric density estimation, respectively.

One limitation of parametric models is that the structure inferred by these models will not be the true structure producing the observations if the true structure is not in the model's hypothesis space. On the other hand, many nonparametric models are guaranteed to infer the true structure given enough (i.e., infinite) observations, which is a property known as *consistency* in the statistics literature (Bickel and Doksum, 2007). Thus, as the number of observations increase, nonparametric models have lower error than parametric models when the true structure is not in the hypothesis space of the parametric model. However, nonparametric models typically need more observations to arrive at the true hypothesis when the true hypothesis is in the parametric model's hypothesis space. Box 1 describes the bias-variance trade-off, which expounds this intuition and provides a formal framework for understanding the benefits and problems with each approach.

Putting Them Together: Nonparametric Bayesian Models

Although people are clearly biased toward certain structures by their prior expectations (like parametric models), the bias is a soft constraint, meaning that people seem to be able to infer seemingly arbitrarily complex models given enough evidence (like nonparametric models). In the remainder of the article, we propose nonparametric Bayesian models, which are nonparametric models with prior biases toward certain types of structures, as a computational explanation for how people infer structure but maintain flexibility.

Box 1 The bias-variance trade-off

Given that nonparametric models are guaranteed to infer the true structure, why would anyone use a parametric model? Although it is true that nonparametric models will converge to the true structure, they are only guaranteed to do so in the limiting case of an infinite number of observations. However, people do not get an infinite number of observations. Thus, the more appropriate question for cognitive scientists is how do people infer structures from a small number of observations, and which type of model is more appropriate for understanding human performance? There are many structures consistent with the limited and noisy evidence typically observed by people. Furthermore, when observations are noisy, it becomes difficult for an agent to distinguish between noise and systematic variation due to the underlying structure, a problem known as "overfitting."

When there are many structures available to the agent, as is the case for nonparametric models, this becomes a serious issue. So, although nonparametric models have the upside of guaranteed convergence to the appropriate structure, they have the downside of being

Box 1 continued

prone to overfitting. In this box, we discuss the bias-variance trade-off, which provides a useful framework for understanding the trade-off between ultimate convergence to the true structure and overfitting.

The *bias-variance trade-off* is a mathematical result, which demonstrates that the amount of error an agent is expected to make when learning from observations can be decomposed into the sum of two components (German, Bienenstock, & Doursat, 1992; Griffiths et al., 2010): *bias*, which measures how close the expected estimated structure is to the true structure, and *variance*, which measures how sensitive the expected estimated structure is to noise (how much it is expected to vary across different possible observations). Intuitively, increasing the number of possible structures by using a larger parametric or a fully nonparametric model reduces the bias of the model because a larger hypothesis space increases the likelihood that the true structure is available to the model. However, it also increases the variance of the model because it will be harder to choose among them given noisy observations from the environment. On the other hand, decreasing the possible number of structures available by using a small parametric model increases the bias of the model, because unless the true structure is one of the structures available to the parametric model, it will not be able to infer the true structure. Furthermore, using a small parametric model reduces the variance, because there are fewer structures available to the model that are likely to be consistent with the noisy observations. The bias-variance trade-off presents a trade-off: reducing the bias of a model by using a nonparametric model with fewer prior constraints comes at the cost of less efficient inference and increased susceptibility to overfitting, resulting in larger variance.

How do people resolve the bias-variance trade-off? In some domains, they are clearly biased because some structures are much easier to learn than others (e.g., linear functions in function learning; Brehmer 1971, 1974). So in some respects people act like parametric models, in that they use strong constraints to infer structures. However, given enough training, experience, and the right kind of information, people can infer extremely complex structures (e.g., McKinley and Nosofsky 1995). Thus, in other respects, people act like nonparametric models. How to reconcile these two views remains an outstanding question for theories of human learning. Hierarchical Bayesian models offer one possible answer, where agents maintain multiple hypothesis spaces and infer the appropriate hypothesis space to perform Bayesian inference over, using the distribution of stimuli in the domain (Kemp, Perfors, & Tenenbaum, 2007) and the concepts agents learn over the stimuli (Austerweil and Griffiths, 2010a). In principle, a hierarchical Bayesian model could be formulated that includes both parametric and nonparametric hypothesis spaces, thereby inferring which is appropriate for a given domain. Formulating such a model is an interesting challenge for future research.

Nonparametric Bayesian models are Bayesian because they put prior probabilities over the set of possible structures, which typically include arbitrarily complex structures. They posit probability distributions over structures that can, in principle, be infinitely complex, but they are biased towards "simpler" structures (those representable using a smaller number of units), which reduces the variance that plagues classical nonparametric models. The probability of data under a structure, which can be very large for complex structures that encode each observation explicitly (e.g., each observation in its own category), is traded off against a prior bias toward simpler structures, which allow observations to share parameters. This bias toward simpler structures is a soft constraint, allowing models to adopt more complex structures as new data arrive (this is what makes these models "nonparametric"). Thus, nonparametric Bayesian models combine the benefits of parametric and nonparametric models: a small variance (by using Bayesian priors) and a small bias (by adapting their structure nonparametrically). See Figure 9.1(c) for an example of nonparametric Bayesian density estimation.

Nonparametric Bayesian models can be classified according to the type of hidden structure they posit. For the previously discussed category learning example, the hidden structure is a probability distribution over the space of observations. Thus, the prior is a probability distribution over probability distributions. A common choice for this prior is the

Dirichlet process (Ferguson, 1973), which induces a set of discrete clusters, where each observation belongs to a single cluster and each cluster is assigned to a randomly drawn value. Combining the Dirichlet process with a model of how observed features are generated by clusters, we obtain a Dirichlet-process mixture model (Antoniak, 1974). As we discuss in the following section, elaborations of the Dirichlet process mixture model have been applied to many psychological domains as varied as category learning (Anderson, 1991; Sanborn, Canini, & Navarro, 2008b), word segmentation (Griffiths, & Johnson, 2009), and associative learning (Gershman, Blei, & Niv, 2010; Gershman, and Niv, 2012).

Although many of the applications of nonparametric Bayesian models in cognitive science have focused on the Dirichlet process mixture model, other nonparametric Bayesian models, such as the *Beta process* (Hjort, 1990; Thibaux and Jordan, 2007) and *Gaussian process* (Rasmussen and Williams, 2006), are more appropriate when people infer probability distributions over observations that are encoded using multiple discrete units or continuous units. For example, feature learning is best described by a hidden structure with multiple discrete units. The Beta process (Griffiths and Ghahramani, 2005, 2011; Hjort, 1990; Thibaux and Jordan, 2007) is one appropriate nonparametric Bayesian model for this example, and as we discuss in the section Inferring Features: What Is a Perceptual Unit?, elaborations of the Beta process have been applied to model feature learning (Austerweil and Griffiths, 2011, 2013), multimodal learning (Yildirm and Jacobs, 2012), and choice preferences (Görür, Jäkel; Miller, Griffiths). Finally, Gaussian processes are appropriate when each observation is encoded using one or more continuous units; we discuss their application to function learning in the section Learning Functions: How Are Continuous Quantities Related? (Griffiths, Lucas, Williams, & Kalish, 2009).

Inferring Clusters: How Are Observations Organized into Groups?

One of the basic inductive problems faced by people is organizing observations into groups, sometimes referred to as *clustering*. This problem arises in many domains, including category learning (Clapper and Bower, 1994; Kaplan and Murphy, 1999; Pothos and Chater, 2002), motion perception (Braddick, 1993), causal inference (Kemp, Tenenbaum, Niyogi, & Griffiths, 2010), word segmentation (Werker and Yeung, 2005), semantic representation (Griffiths, Steyvers, & Tenenbaum, 2007), and associative learning (Gershman et al., 2010). Clustering is challenging because in real world situations the number of clusters is often unknown. For example, a child learning language does not know *a priori* how many words there are in the language. How should a learner discover new clusters?

In this section, we show how clustering can be formalized as Bayesian inference, focusing in particular on how the nonparametric concepts introduced in the previous section can be brought to bear on the problem of discovering new clusters. We then describe an application of the same ideas to associative learning.

A Rational Model of Categorization

Categorization can be formalized as an inductive problem: given the features of a stimulus (denoted by x), infer the category label c. Using Bayes's rule, the posterior over category labels is given by:

$$P(c|x) = \frac{P(x|c)P(c)}{\sum_{c'} P(x|c')P(c')} = \frac{P(x,c)}{\sum_{c'} P(x,c')}.$$

From this point of view, category learning is fundamentally a problem of *density estimation* (Ashby and Alfonso-Reese, 1995) because people are estimating a probability distribution over the possible observations from each category. Probabilistic models differ in the assumptions they make about the joint distribution $P(x,c)$. Anderson (1991) proposed that people model this joint distribution as a *mixture model*:

$$P(x,c) = \sum_z P(x,c|z)P(z),$$

where $z \in \{1,\ldots,K\}$ denotes the cluster assigned to x (z is the traditional notation for a cluster, and is analogous to the hypothesis h in the previous section). From a generative perspective, observations are generated by a mixture model from the environment according to the following process: to sample observation n, first sample its cluster z_n from $P(z)$, and then the observation x_n and its category c_n from the joint distribution specified by the cluster, $P(x,c|z_n)$. Each distribution specified by a cluster might be simple (e.g., a Gaussian), but their mixture can approximate arbitrarily complicated distributions. Because each observation only belongs to one cluster, the assignments $\mathbf{z}_n = \{z_1,\ldots,z_n\}$ encode a *partition* of the items into K distinct clusters, where a partition is a grouping of items into mutually

exclusive clusters. When the value of K is specified, this generative process defines a simple probabilistic model of categorization, but what should be the value of K?

To address the question of how to select the value of K, Anderson assumed that K is not known *a priori*, but rather learned from experience, such that K can be increased as new data are observed. As the number of clusters grows with observations and each cluster has associated parameters defining its probability distribution over observations, this rational model of categorization is nonparametric. Anderson proposed a prior on partitions that sequentially assign observations to clusters according to:

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

where m_k is the number of items in $\mathbf{z}_{n-1}$ assigned to cluster k, n is the total number of items observed so far, and $\alpha \geq 0$ is a parameter that governs the total number of clusters.[3] As pointed out by Neal (2000), the process proposed by Anderson is equivalent to a distribution on partitions known as the *Chinese restaurant process* (CRP; Aldous, 1985; Blackwell and MacQueen, 1973). Its name comes from the following metaphor (illustrated in Figure 9.2): Imagine a Chinese restaurant with an unbounded number of tables (clusters), where each table can seat an unbounded number of customers (observations). The first customer enters and sits at the first table. Subsequent customers sit at an occupied table with a probability proportional to how many people are already sitting there (m_k), and at a new table with probability proportional to α. Once all the customers are seated, the assignment of customers to tables defines a partition of observations into clusters.

The CRP arises in a natural way from a nonparametric mixture modeling framework (see Gershman and Blei, 2012, for more details). To see this, consider a finite mixture model where the cluster assignments are drawn from:

$$\theta \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$

$$z_i \sim \text{Multinomial}(\theta), \quad \text{for } i = 1, \ldots, n,$$

where Dirichlet($\cdot$) denotes the K-dimensional Dirichlet distribution, where θ roughly corresponds to the relative weight of each block in the partition. As the number of clusters increases to infinity ($K \to \infty$), this distribution on $\mathbf{z}_n$ is equivalent to the CRP. Another view of this model is given by a seemingly unrelated process, the Dirichlet process (Ferguson, 1973), which is a probability distribution over discrete probability distributions. It directly generates the partition and the parameters associated with each block in the partition. Marginalizing over all the possible ways of getting the same partition from a Dirichlet process defines a related distribution, the Pólya urn (Blackwell and MacQueen, 1973), which is equivalent to the CRP when the parameter associated with each block is ignored. For this reason, a mixture model with a CRP prior on partitions is known as a Dirichlet process mixture model (Antoniak, 1974).

One of the original motivations for developing the rational model of categorization was to reconcile two important observations about human category learning. First, in some cases, the confidence with which people assign a new stimulus to a category is inversely proportional to its distance from the average of the previous stimuli in that category (Reed, 1972). This, in conjunction with other data on central tendency effects (e.g., Posner and Keele, 1968), has been interpreted as people abstracting a "prototype" from the observed stimuli. On the other hand, in some cases, people are sensitive to specific stimuli (Medin and Schaffer, 1978), a finding that has inspired exemplar models that memorize the entire stimulus set (e.g., Nosofsky, 1986). Anderson (1991) pointed out that his rational model of categorization can capture both of these findings, depending on the inferred partition structure: when all items are assigned to the same cluster, the model is equivalent to forming a single

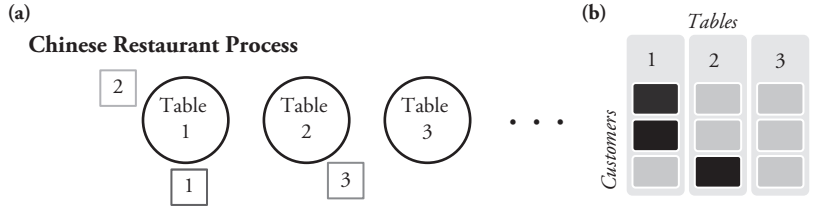

Fig. 9.2 (a) The culinary representation of the Chinese restaurant process and (b) the cluster assignments implied by it.

prototype, whereas when all items are assigned to unique clusters, the model is equivalent to an exemplar model. However, in most circumstances, the rational model of categorization will partition the items in a manner somewhere between these two extremes. This is a desirable property that other recent models of categorization have adopted (Love, Medin, & Gureckis, 2004). Finally the CRP serves as a useful component in many other cognitive models (see Figure 7 and Box 2).

Associative Learning

As its name suggests, associative learning has traditionally been viewed as the process by which associations are learned between two or more stimuli. The paradigmatic example is Pavlovian conditioning, in which a cue and an outcome (e.g., a tone and a shock) are paired together repeatedly. When done with rats, this procedure leads to the rat freezing in anticipation of the shock whenever it hears the tone.

Association learning can be interpreted in terms of a probabilistic causal model, which calculates the probability that one variable, called the cue (e.g., a tone), causes the other variable, called the outcome (e.g., a shock). Here y encodes the presence or absence of the outcome, and x similarly encodes the presence or absence of the cue. The model assumes that y is a noisy linear function of x: $y \sim \mathcal{N}(wx, \sigma^2)$. The parameter w encodes the associative strength between x and y and σ^2 parameterizes the variability of their relationship. This model can be generalized to the case in which multiple cues are paired with an outcome by assuming that the associations are additive: $y \sim \mathcal{N}(\sum_i w_i x_i, \sigma^2)$, where i ranges over the cues. The linear-Gaussian model also has an interesting connection to classical learning theories such as the Rescorla-Wagner model (Rescorla and Wagner, 1972), which can be interpreted as assuming a Gaussian prior on w and carrying out Bayesian inference on w (Dayan, Kakade, Montague 2000; Kruschke, 2008).

Despite the successes of the Rescorla-Wagner model and its probabilistic variants, they incorrectly predict that there should only be learning when the prediction error is nonzero (i.e., when $y - \sum_i w_i x_i \neq 0$), but people and animals can still learn in some cases. For example, in *sensory preconditioning* (Brogden, 1939), two cues (A and B) are presented together without an outcome; when A is subsequently paired with an outcome, cue B acquires associative strength despite never being paired with the outcome. Because A and B presumably start out with zero associative strength and there are no prediction errors during the preconditioning phase, the Rescorla-Wagner model predicts that there should be no learning. The model fails to explain how preconditioning enables B to subsequently acquire associative strength from A-outcome training.

Box 2 Composing Richer Nonparametric Models

Although this chapter has focused on several of the most basic nonparametric Bayesian models and their applications to basic psychological processes, these components can also be composed to build richer accounts of more sophisticated forms of human learning and reasoning (bottom half of Figure 7). These composites greatly extend the scope of phenomena that can be captured in probabilistic models of cognition. In this box, we discuss a number of these composite models.

In many domains, categories are not simply a flat partition of entities into mutually exclusive classes. Often they have a natural hierarchical organization, as in a taxonomy that divides life forms into animals and plants; animals into mammals, birds, fish, and other forms; mammals into canidae, felines, primates, rodents, ...; canidae into dogs, wolves, foxes, coyotes, ...; and dogs into many different breeds. Category hierarchies can be learned via nonparametric models based on nested versions of the CRP, in which each category at one level gives rise to a CRP of subcategories at the level below (Griffiths et al., 2008b; Blei, Griffiths, & Jordan, 2010).

Category learning and feature learning are typically studied as distinct problems (as discussed in the sections Inferring Clusters: How Are Observations Organized into Groups? and Inferring Features: What Is a Perceptual Unit?), but in many real-world situations people can jointly discover how best to organize objects into classes and which features best support these categories. Nonparametric models that combine the CRP and IBP can capture these joint inferences (Austerweil and Griffiths, 2013). The model of Salakhutdinov et al. (2012) extends this idea to hierarchies, using the nested CRP combined with a hierarchical Dirichlet process topic model to jointly learn hierarchically structured object categories and

Box 2 continued

hierarchies of part-like object features that support these categories.

The *CrossCat* model (Shafto et al., 2011) allows us to move beyond another limitation of typical models of categorization (parametric or nonparametric): the assumption that there is only a single best way to categorize a set of entities. Many natural domains can be represented in multiple ways: animals may be thought of in terms of their taxonomic groupings or their ecological niches, foods may be thought of in terms of their nutritional content or social role; products may be thought of in terms of function or brand. CrossCat discovers multiple systems of categories of entities, each of which accounts for a distinct subset of the entities' observed attributes, by nesting CRPs over entities inside CRPs over attributes.

Traditional approaches to categorization treat each entity individually, but richer semantic structure can be found by learning categories in terms of how groups of entites relate to each other. The *Infinite Relational Model* (IRM; Kemp et al. 2006, 2010) is a nonparametric model that discovers categories of objects that not only share similar attributes, but also participate in similar relations. For instance, a data set for consumer choice could be characterized in terms of which consumers bought which products, which features are present in which products, which demographic properties characterize which users, and so on. IRM could then discover how to categorize products and consumers (and perhaps also features and demographic properties), and simultaneously uncover regularities in how these categories relate (for) example, that consumers in class X tend to buy products in class Y).

Nonparametric models defined over graph structures, such as the graph-based GP models of Kemp and Tenenbaum (2008, 2009), can capture how people reason about a wider range of dependencies between the properties of entities and the relations between entities, allowing that objects in different domains can be related in qualitatively different ways. For instance, the properties of cities might be best explained by their relative positions in a two-dimensional map, the voting patterns of politicians by their orientation along a one-dimensional liberal-conservative axis, and the properties of animals by their relation in a taxonomic tree. We could also distinguish animals' anatomical and physiological properties, which are best explained by the taxonomy, from behavioral and ecological properties that might be better explained by their relation in a directed graph such as a food web. Perhaps most intriguingly, nonparametric Bayesian models can be combined with symbolic grammars to account for how learners could explore the broad landscape of different model structures that might describe a given domain and arrive at the best model (Kemp and Tenenbaum, 2008; Grosse et al., 2012). A grammar is used to generate a space of qualitatively different model families, ranging from simple to complex, each of which defines a predictive model for the observed data based on a GP, CRP, IBP or other nonparametric process. These frameworks have been used to build more human-like machine learning and discovery systems, but they remain to be tested as psychological accounts of how humans learn domain structures.

Sensory preconditioning and other related findings have prompted consideration of alternative probabilistic models for associative learning. Courville, Daw, & Touretzky, (2006) proposed that people and animals posit *latent causes* to explain their observations in Pavlovian conditioning experiments. According to this idea, a single latent cause generates both the cues and outcomes. Latent cause models are powerful because they can explain why learning occurs in the absence of prediction errors. For example, during sensory preconditioning, the latent cause captures the covariation between the two cues; subsequent conditioning of A increases the probability that B will also be accompanied by the outcome.

An analogous question about how to pick the number of clusters in a mixture model arises in associative learning: How many latent causes should there be in a model of associative learning? To address this question, Gershman and colleagues (Gershman et al., 2010; Gershman and Niv, 2012) used the CRP as a prior on latent causes. This allows the model to infer new latent causes when the sensory statistics change, but otherwise it prefers a small number of latent causes. Unlike previous models that define the number of latent causes

a priori, Gershman et al. (2010) showed that this model could explain why extinction does not tend to erase the original association: Extinction training provides evidence that a new latent cause is active. For example, when conditioning and extinction occur in different contexts, the model infers a different latent cause for each context; upon returning to the conditioning context, the model predicts a renewal of the conditioned response, consistent with empirical findings (see Bouton, 2004). By addressing the question of how agents infer the number of latent causes, the model offered new insight into a perplexing phenomenon.

Inferring Features: What Is a Perceptual Unit?

The types of latent structures that people use to represent a set of stimuli can be far richer than clustering the stimuli into groups. For example, consider the following set of animals: domestic cats, dogs, goldfish, sharks, lions, and wolves. Although they can be represented as clusters (e.g., PETS and WILD ANIMALS or FELINES, CANINES, and SEA ANIMALS), another way to represent the animals is using features, or multiple discrete units per animal (e.g., a cat might be represented with the following features: HAS TAIL, HAS WHISKERS, HAS FUR, IS CUTE, etc.). Feature representations can be used to solve problems arising in many domains, including choice behavior (Tversky, 1972), similarity (Nosofsky, 1986; Tversky, 1977), and object recognition (Palmer, 1999; Selfridge and Neisser, 1960). Analogous to clustering, the appropriate feature representation or even the number of features for a domain is not known *a priori*. In fact, a common criticism of feature-based similarity is that there is an infinite number of potential features that can be used to represent any stimulus and that human judgments are mostly determined by the features selected to be used in a given context (Goldmeier, 1972; Goodman, 1972; Murphy and Medin, 1985). How do people infer the appropriate features to represent a stimulus in a given context?

In this section, we describe how the problem of inferring feature representations can be cast as a problem of Bayesian inference, where the hypothesis space is the space of possible feature representations. Because there is an infinite number of feature representations, the model used to solve this problem will be a nonparametric Bayesian model. Then, we illustrate two psychological applications.

A Rational Model of Feature Inference

Analogous to other Bayesian models of cognition, defining a rational model of feature inference amounts to applying Bayes's rule to a specification of the hypothesis space, how hypotheses produce observations (the likelihood), and the prior probability of each hypothesis (the prior). Following previous work by Austerweil and Griffiths (2011), we first define these three components for a Bayesian model with a finite feature repository and then define a nonparametric Bayesian model by allowing an infinite repository of features. This allows the model to infer a feature representation without presupposing the number of features ahead of time.

The computational problem of feature representation inference is as follows: Given the D-dimensional raw primitives for a set of N stimuli $\mathbf{X}$ (each object is a D-dimensional row vector $\mathbf{x}_n$), infer a feature representation that encodes the stimuli and adheres to some prior expectations. We decompose a feature representation into two components: an $N \times K$ *feature ownership matrix* $\mathbf{Z}$, which is a binary matrix encoding which of the K features each stimulus has (i.e., $z_{nk} = 1$ if stimulus n has feature k, and $z_{nk} = 0$ otherwise), and a $K \times D$ *feature image matrix* $\mathbf{Y}$, which encodes the consequence of a stimulus having each feature. In this case, the hypothesis space is the Cartesian product of possible feature ownership matrices and possible feature image matrices. As we discuss later in further detail, the precise format of feature image matrix $\mathbf{Y}$ depends on the format of the observed raw primitives. For example, if the stimuli are the images of objects and the primitives are D binary pixels that encode whether light was detected in each part of the retina, then a stimulus and a feature image, $\mathbf{x}$ and $\mathbf{y}$ respectively, are both D-dimensional binary vectors. So if $\mathbf{x}$ is the image of a mug, $\mathbf{y}$ could be the image of its handle.

Applying Bayes's rule and assuming that the feature ownership and image matrices are independent *a priori*, inferring a feature representation amounts to optimizing the product of three terms

$$P(\mathbf{Z}, \mathbf{Y}|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{Y}, \mathbf{Z})P(\mathbf{Y})P(\mathbf{Z}),$$

where $P(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$, the likelihood, encodes how well each object $\mathbf{x}_n$ is reconstructed by combining together the feature images $\mathbf{Y}$ of the features the object has, which is given by $\mathbf{z}_n$, $P(\mathbf{Y})$ encodes prior expectations about feature images (e.g., Gestalt laws), and $P(\mathbf{Z})$ encodes prior expectations about feature ownership matrices.[4] As the likelihood and feature image prior are more straightforward to

define and specific to the format of the observed primitives, we first derive a sensible prior distribution over all possible feature ownership matrices before returning our attention to them.

Before delving into the case of an infinite number of potential features, we derive a prior distribution on feature ownership matrices that has a finite and known number of features K. As the elements of a feature ownership matrix are binary, we can define a probability distribution over the matrix by flipping a weighted coin with bias π_k for each element z_{nk}. We do not observe π_k and so, we assume a Beta distribution as its prior. This corresponds to the following generative process

$$\pi_k \overset{iid}{\sim} \text{Beta}(\alpha/K, 1), \quad \text{for } k = 1, \ldots, K$$

$$z_{nk}|\pi_k \overset{iid}{\sim} \text{Bernoulli}(\pi_k), \quad \text{for } n = 1, \ldots, N.$$

Due to the conjugacy of Bernoulli likelihoods and Beta priors, it is relatively simple to integrate out $\pi_1, \ldots, \pi_K$ to arrive at the following probability distribution, $P(\mathbf{Z}|\alpha)$, over finite feature ownership representations. See Bernardo and Smith (1994) and Griffiths and Ghahramani (2011) for details.

Analogous to the method discussed earlier for constructing the CRP as the infinite limit of a finite model, taking the limit of $P(\mathbf{Z}|\alpha)$ as $K \to \infty$ yields a valid probability distribution over feature ownership matrices with an infinite number of potential features.[5] Note that as $K \to \infty$, the prior on feature weights gets concentrated at zero (because $\alpha/K \to 0$). This results in an infinite number of columns that simply contain zeroes, and thus, these features will have no consequence for the set of stimuli we observed (as they are not assigned to any stimuli). Because both the number of columns $K \to \infty$ and the probability of an object taking a feature (probability that $z_{nk} = 1$) $\pi_k \to 0$ at corresponding rates, there is a finite, but random, number of columns have at least one nonzero element (the features that have been taken by at least one stimulus). This limiting distribution is called the *Indian buffet process* (IBP; Griffiths and Ghahramani 2005, 2011), and it is given by the following equation

$$P(Z|\alpha) = \frac{\alpha^{K_+}}{\prod_{h=1}^{2^n-1} K_h} \exp\left\{-\alpha \sum_{n=1}^{N} n^{-1}\right\} \times \prod_{k=1}^{K_+} \frac{(N - m_k)(m_k - 1)}{N},$$

where K_+ is the number of columns with at least one nonzero entry (the number of features taken by at least one object), and K_h is the number of features with history h, where a history can be thought of as the column of the feature interpreted as a binary number. The term containing the history penalizes features that have equivalent patterns of ownership and it is a method for indexing features with equivalent ownership patterns.

Analogous to the CRP, the probability distribution given by this limiting process is equivalent to the probability distribution on binary matrices implied by a simple sequential culinary metaphor. In this culinary metaphor, "customers," corresponding to the stimuli or rows of the matrix, enter an Indian buffet and take dishes, corresponding to the features or columns of the matrix, according to a series of probabilistic decisions based on how the previous customers took dishes. When the first customer enters the restaurant, she takes a number of new dishes sampled from a Poisson distribution with parameter α. As customers sequentially enter the restaurant, each customer n takes a previously sampled dish k with probability m_k/n and then samples a number of new dishes sampled from a Poisson distribution with parameter α/n.

Figure 9.3(a) illustrates an example of a potential state of the IBP after three customers have entered the restaurant. The first customer entered the restaurant and sampled two new dishes from the Poisson probability distribution with parameter α. Next, the second customer entered the restaurant and took each of the old dishes with probability 1/2 and sampled one new dish from the Poisson probability distribution with parameter $\alpha/2$. Then, the third customer entered the restaurant and took the first dish with probability 2/3, did not take the second dish with probability 1/3, and did not take the third dish with probability 2/3. The equivalent feature ownership matrix represented by this culinary metaphor is shown in Figure 9.3(b).

As previously encountered features are sampled with probability proportional to the number of times they were previously taken and the probability of new features decays as more customers enter the restaurant (it is Poisson distributed with parameter given by α/N where N is the number of customers), the IBP favors feature representations that are sparse and have a small number of features. Thus, it encodes a natural prior expectation toward feature representations with a few features, and can be interpreted as a simplicity bias.

Now that we have derived the feature ownership prior $P(\mathbf{Z})$, we turn to defining the feature image

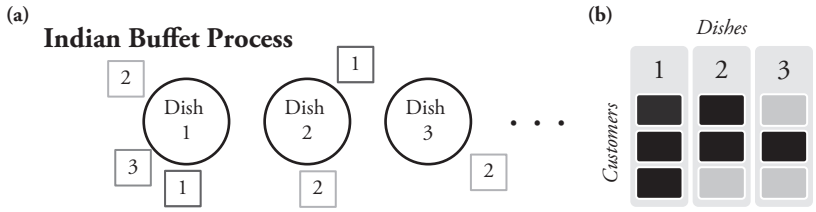

Fig. 9.3 (a) The culinary representation of the Indian buffet process and (b) the feature ownership matrix implied by it.

prior $P(\mathbf{Y})$ and the likelihood $P(\mathbf{X}|\mathbf{Y},\mathbf{Z})$, which will finish a specification of the nonparametric Bayesian model. Remember that the feature image matrix contains the consequences of a stimulus having a feature in terms of the raw primitives. Thus, in the nonparametric case, the feature image matrix can be thought of containing the D-dimensional consequence of each of the K_+ used features. The infinite unused features can be ignored because a feature only affects the representation of a stimulus if it is used by that stimulus. In most applications to date, the feature image prior is mostly "knowledge-less." For example, the standard prior used for stimuli that are binary images is an independent Bernoulli prior, where each pixel is on with probability ϕ independent of the other pixels in the image. One exception is one of the simulations by Austerweil and Griffiths (2011), who demonstrated that using a simple proximity bias (an Ising model that favors adjacent pixels to share values, Geman and Geman, 1984) as the feature image prior results in more psychologically plausible features: The feature images without using the proximity bias were not contiguous and were speckled, whereas the feature images using the proximity bias were contiguous. For grayscale and color images (Austerweil and Griffiths, 2011; Hu, Zhai, Williamson, & Boyd-Graber, 2012), the standard "knowledge-less" prior generates each pixel from a Gaussian distribution independent of the other pixels in the image.

Analogous to the feature image priors, the choice of the likelihood depends on the format of the raw dimensional primitives. In typical applications, the likelihood assumes that the reconstructed stimuli is given by the product of the feature ownership and image matrices, $\mathbf{ZY}$ and penalizes the deviation between the reconstructed and observed stimuli (Austerweil and Griffiths, 2011). For binary images, the noisy-OR likelihood (Pearl, 1988; Wood, Griffiths, & Ghahramani, 2006) is used, which amounts to thinking of each feature as a "hidden cause" and has support as a psychological explanation for how people reason about observed effects being produced by multiple hidden causes (Cheng, 1997; Griffiths and Ghahramani, 2005). For grayscale images, the linear-Gaussian likelihood is typically used (Griffiths and Ghahramani, 2005; Austerweil and Griffiths, 2011), which is optimal under the assumption that the reconstructed stimuli is given by $\mathbf{ZY}$ and that the metric of success is the sum squared error between the reconstructed and observed stimuli. Recent work in machine learning has started to explore more complex likelihoods, such as explicitly accounting for depth and occlusion (Hu, Zhai, Williamson, & Boyd-Graber 2012). Formalizing more psychologically valid feature image priors and likelihoods is a mostly unexplored area of research that demands attention.

After specifying the feature ownership and image prior and the likelihood, a feature representation can be inferred for a given set of observations using standard machine learning inference techniques, such as Gibbs sampling (Geman and Geman, 1984) or particle filtering (Gordon, Salmond, & Smith, 1993). We refer the reader to Austerweil and Griffiths (2013), who discuss Gibbs sampling and particle filtering for feature inference models and analyze their psychological plausibility.

What features should people use to represent the image in Figure 9.4(a)? When the image is in the context of the images in Figure 9.4(b), Austerweil and Griffiths (2011) found that people and the IBP model infer a single feature to represent it, namely the object itself, which is shown in Figure 9.4(d). Alternatively, when the image is in the context of the images in Figure 9.4(c), people and the IBP model infer a set of features to represent it, which are three of the six parts used to create the images, which are shown in Figure 9.4(e).[6] Importantly, Austerweil and Griffiths (2011) demonstrated that two of the most popular machine-learning techniques for inferring features from a set of images, principal component analysis (Hyvarinen,

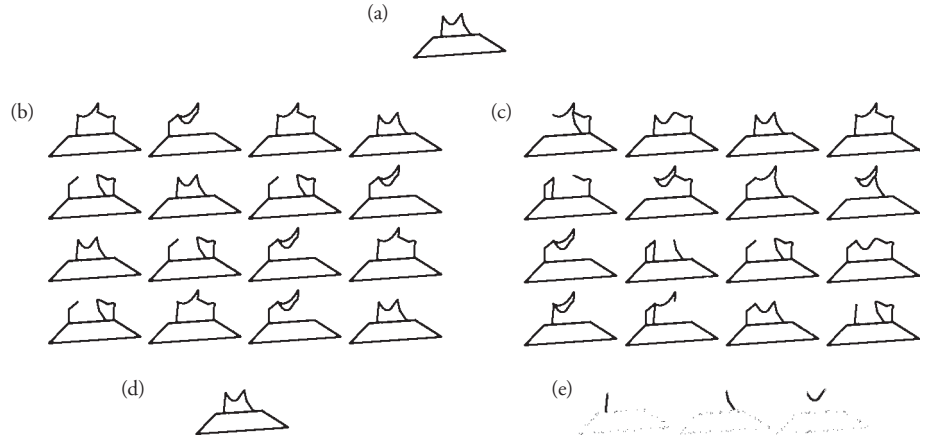

Fig. 9.4 Effects of context on the features inferred by people to represent an image (from Experiment 1 of Austerweil & Griffiths, 2011). What features should be used to represent the image shown in (a)? When (a) is in the context of the images shown in (b), participants and the nonparametric Bayesian model represent it with a single feature, the image itself, which is shown in (d). Conversely, when (a) is in the context of the images shown in (c), it is represented as a conjunction of three features, shown in (e). Participants and the nonparametric Bayesian model generalized differently due to using different representations.

Karhunen, & Oja 2001) and independent component analysis (Hyvarinen et al., 2001), did not explain their experimental results. This suggests that the simplicity bias given by the IBP is similar to the prior expectations people use to infer feature representations, although future work is necessary to more precisely test and formally characterize the biases people use to infer feature representations. Furthermore, this contextual effect was replicated in grayscale 3-D rendered images and in a conceptual domain, suggesting that it may be a domain-general capability.

Unlike the IBP and most other computational models of feature learning, people tend to use features that are invariant over transformations (e.g., translations or dilations) because properties of features observed from the environment do not occur identically across appearances (e.g., the retinal image after eye or head movements). By modifying the culinary metaphor such that each time a customer takes a dish she draws a random spice from a spice rack, the transformed IBP (Austerweil and Griffiths, 2010b) can infer features invariant over a set of transformations. Austerweil and Griffiths (2013) explain a number of previous experimental results and outline novel contextual effects predicted by the extended model that they subsequently confirm through behavioral experiments. Other extensions to this same framework have been used to explain multimodal representation learning (Yildirm and Jacobs, 2012), and in one extension, the IBP and CRP are used together to infer features diagnostic for categorization (Austerweil and Griffiths, 2013).

Choice Behavior

According to feature-based choice models, people choose between different options depending on their preference for the features (called an aspect) that each option has. One influential feature-based choice model is the elimination by aspects model (Tversky, 1972). Under this model, the preference of a feature is assumed to be independent of other features and defined by a single, positive number, called a *weight*. Choices are made by repeatedly selecting a random feature weighted by their preference and removing all options that do not contain the feature until there is only one option remaining. For example, when a person is choosing which television to purchase, she is confronted with a large array of possibilities that vary on many features, such as IS LCD, IS PLASMA, IS ENERGY EFFICIENT, IS HIGH DEFINITION, and so on, each with their own associated weight. Because the same person may make different choices even when they are confronted with the same set of options, the probability of choosing option i over option j, p_{ij}, is proportional to the weights of the features that option i has, but option j does not have, relative to the features that option j has, but option i does not have. Formally, this is given by

$$p_{ij} = \frac{\sum_k w_k z_{ik}(1 - z_{jk})}{\sum_k w_k z_{ik}(1 - z_{jk}) + \sum_k w_k (1 - z_{ik}) z_{jk}},$$

where $z_{ik} = 1$ denotes that option i has feature k and w_k is the preference weight given to feature k.

Although modeling human choice behavior using the elimination by aspects model is straightforward when the features of each option and their preference weights are known; it is not straightforward how to infer what features a person uses to represent a set of options and their weights given only the person's choice behavior. To address this issue, Görür et al. (2006) proposed the IBP as a prior distribution over possible feature representations and a Gamma distribution as a prior over the feature weights, which is a flexible distribution that only assigns probability to positive numbers. They demonstrated that human choice for which celebrities participants from the early 1970s one would prefer to chat with (Rumelhart and Greeno, 1971) is just as well described by the elimination by aspects model given a set of features defined by a modeler or when the features are inferred with the IBP as the prior over possible feature representations. As participants are familiar with the celebrities and the celebrities are related to each other according to a hierarchy (i.e., politicians, actors, and athletes), Miller et al. (2008) extended the IBP such that it infers features in a manner that respects the given hierarchy (i.e., options are more likely to have the same features to the degree that they are close to each other in the hierarchy). They demonstrated that the extended IBP explains human choice judgments better and uses fewer features to represent the options. Although the IBP-based extensions helps the elimination by aspects model overcome some of its issues, the extended models are unable to account for the full complexity of human choice behavior (e.g., attraction effects; Huber, Payne, & Puto, 1982). Regardless, exploring choice models that include a feature-inference process is a promising direction for future research, because such models can potentially be incorporated with more psychologically valid choice models (e.g., sequential sampling models of preferential choice; Roe, Busemeyer, & Townsend 2001).

Learning Functions: How Are Continuous Quantities Related?

So far, we have focused on cases in which the latent structure to be inferred is discrete—either a category or a set of features. However, latent structures can also be continuous. One of the most prominent examples is function learning, in which a relationship is learned 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 solution to this problem that can learn complex functions in a manner that favors simple solutions.

Viewed abstractly, the computational problem behind function learning is to learn a function $y = f(x)$ 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_n is assumed to be the true value obscured by some kind of additive noise (i.e., $t_n = y_n + \epsilon_n$, where ϵ_n is some type of 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 Williams (1998).

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 $\mathbf{x}$ and $\mathbf{y}$. We can do this by applying Bayes's rule, with

$$p(f|\mathbf{x}_N, \mathbf{t}_N) = \frac{p(\mathbf{t}_N|f, \mathbf{x}_N)p(f)}{\int_{\mathcal{F}} p(\mathbf{t}_N|f, \mathbf{x}_N)p(f)\, df}, \quad (1)$$

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 likelihood of observing the values of $\mathbf{t}_N$ if f were the true function, and $p(f|\mathbf{x}_N, \mathbf{t}_N)$ is the posterior distribution over functions given the observations $\mathbf{x}_N$ and $\mathbf{y}_N$. In many cases, the likelihood is defined by assuming that the values of t_n are independent given f and x_n, and each follows a Gaussian distribution with mean $y_n = f(x_n)$ and variance σ^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.

Performing the calculations outlined in the previous paragraph for a general hypothesis space $\mathcal{F}$ is challenging, but it becomes straightforward if we limit the hypothesis space to certain specific classes of functions. If we take $\mathcal{F}$ to be all linear functions of the form $y = b_0 + xb_1$, then we need to define a prior $p(f)$ over all linear functions. As 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 Eq. 1, then, results in a multivariate Gaussian posterior distribution on $\mathbf{b}$ (see Rasmussen and Williams, 2006, for details) with

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

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

where $\mathbf{X}_N = [\mathbf{1}_N \ \mathbf{x}_N]$ (i.e., a matrix with a vector of ones horizontally concatenated with $\mathbf{x}_{N+1}$). Because y_{N+1} is simply a linear function of $\mathbf{b}$, the predictive distribution is Gaussian, 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^2 to the variance.

Basis Functions and Similarity Kernels

Although 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 classes of functions. Many classes 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), $x, x^2, \ldots, x^k$. Letting $\phi^{(1)}, \ldots, \phi^{(k)}$ denote a set of basis 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 $f(x) = b_0 + \phi^{(1)}(x)b_1 + \cdots + \phi^{(k)}(x)b_k$ 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 $\boldsymbol{\phi}(\mathbf{x}_N) = [\boldsymbol{\phi}(x_1) \ldots \boldsymbol{\phi}(x_N)]^T$.

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, by simply defining a joint distribution on $\mathbf{y}_{N+1}$ given $\mathbf{x}_{N+1}$ and conditioning on $\mathbf{y}_N$. For example, we might take $\mathbf{y}_{N+1}$ to be jointly Gaussian, with covariance matrix

$$\mathbf{K}_{N+1} = \begin{pmatrix} \mathbf{K}_N & \mathbf{k}_{N,N+1} \\ \mathbf{k}_{N,N+1}^T & k_{N+1} \end{pmatrix}$$

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 $k_{N+1} - \mathbf{k}_{N,N+1}^T \mathbf{K}_N^{-1} \mathbf{k}_{N,N+1}$. 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 y_{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 function whose argument is a pair of stimuli 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. One common kernel is the radial basis function, with

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

indicating that values of y for which values of x are close are likely to be highly correlated. See Schölkopf and Smola (2002) for further details regarding kernels. 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. Some examples are shown in Figure 9.5.

Just as we can express a covariance matrix in terms of its eigenvectors and eigenvalues, we can express a given kernel $K(x_i, x_j)$ in terms of its eigenfunctions $\boldsymbol{\phi}$ and eigenvalues $\boldsymbol{\lambda}$, with

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

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 kernel, there exists a corresponding prior on functions that yields the same predictions. This result is a consequence of Mercer's theorem (Mercer, 1909). Thus, Bayesian linear regression and prediction with Gaussian processes are just two views of the same class of solutions to regression problems.

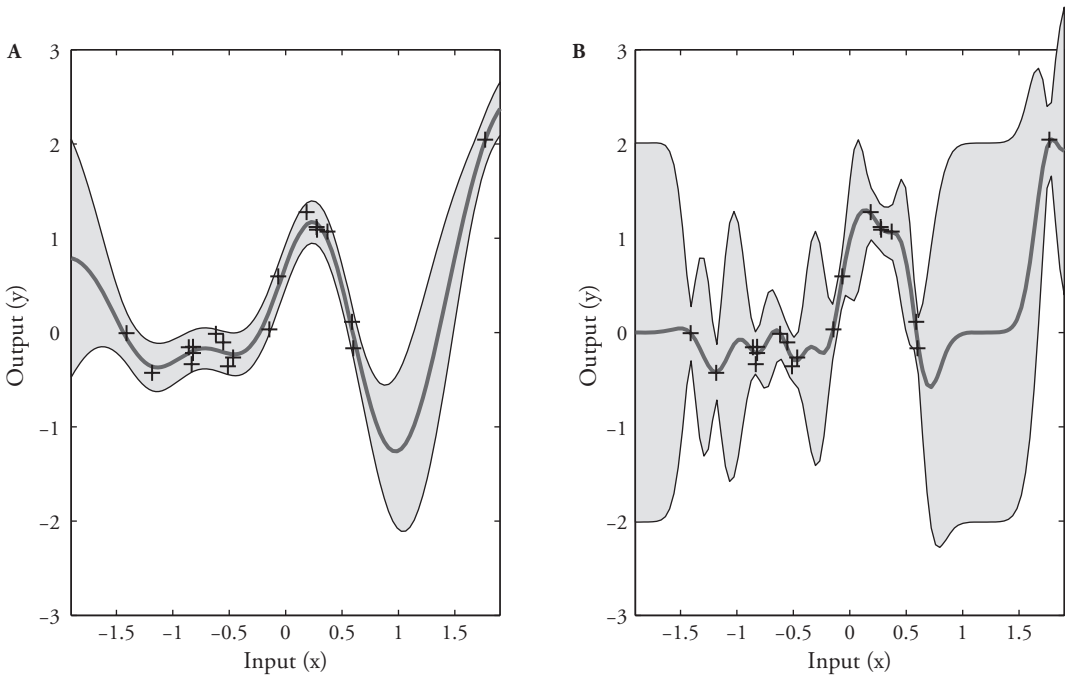

Fig. 9.5 Modeling functions with Gaussian processes. The data points (crosses) are the same in both panels. (A) Inferred posterior using a radial basis covariance function (Eq. 2) with $\theta_2 = 1/4$. (B) Same as panel A, but with $\theta_1 = 1$ and $\theta_2 = 1/8$. Notice that as θ_2 gets smaller, the posterior mean more closely fits the observed data points, and the posterior variance is larger in regions far from the data.

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 and Meyer, 1991) suggests that people are learning an explicit function from a given class, such as the polynomials of degree D. 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) 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 it faces challenges in explaining how people generalize beyond their experience. More recently, hybrids of these two approaches have been proposed (e.g., Kalish, Lewandowsky, & Kruschke, 2004; McDaniel and Busemeyer, 2005). For example, the population of linear experts (POLE; Kalish et al. 2004) uses associative learning to learn a set of linear functions and their expertise over regions of dimensional space.

Bayesian linear regression resembles explicit rule learning, estimating the parameters of a function, whereas 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 that prototype and exemplar models correspond to different types of solutions of the same statistical problem, 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. (2009) compared a Gaussian process model with a mixture of kernels (linear, quadratic, and radial basis) to human performance. Figure 9.6 shows mean human predictions when trained on a linear, exponential, and quadratic function (from DeLosh et al., 1997), together with

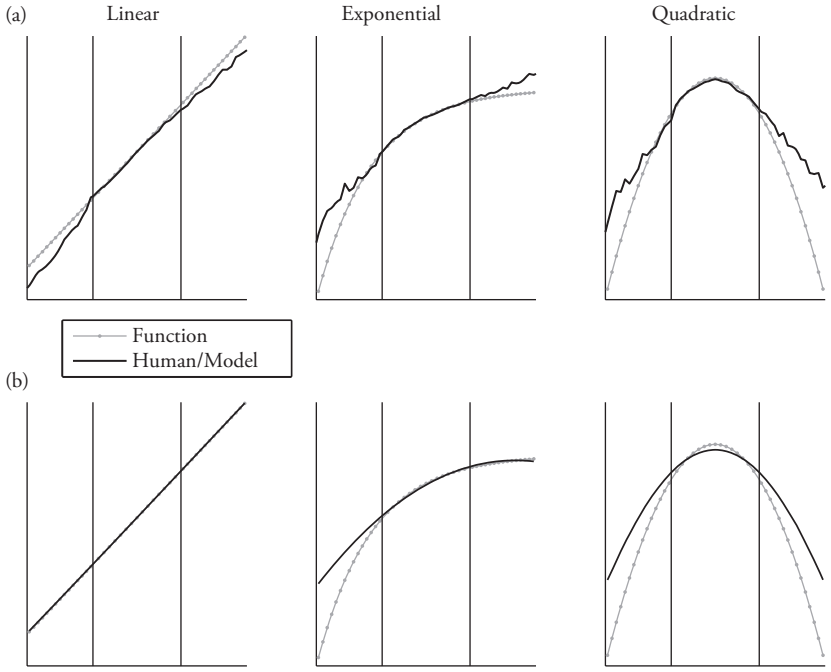

Fig. 9.6 Extrapolation performance in function learning. Mean predictions on linear, exponential, and quadratic functions for (a) human participants (from DeLosh et al. 1997) and (b) a Gaussian process model with linear, quadratic, and nonlinear kernels. Training data were presented in the region between the vertical lines, and extrapolation performance was evaluated outside this region. Figure reproduced from Griffiths et al. (2009).

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 toward a linear slope in the extrapolation of the exponential and quadratic functions, with extreme values of the quadratic and exponential function being overestimated.

Conclusions

Probabilistic models form a promising framework for explaining the impressive success that people have in solving different inductive problems. As part of performing these feats, the mind constructs structured representations that flexibly adapt to the current set of stimuli and context. In this chapter, we reviewed how these problems can be described from a statistical viewpoint. To define models that infer representations that are both flexible and structured, we described three main classes of nonparametric Bayesian models and how the format of the observed stimuli determines which of the three classes of models should be used.

Our presentation of these three classes of models is only the beginning of an ever-growing literature using nonparametric Bayesian processes in cognitive science. Each class of model can be thought of as providing primitive units, which can be composed in various ways to form richer models. For example, the IBP can be interpreted as providing primitive units that can be composed together using logical operators to define a categorization model, which learns its own features along with a propositional rule to define a category. Figure 9.7 presents some of these applications and Box 2 provides a detailed discussion of them.

Concluding remarks

1. Although it is possible to define probabilistic models that can infer any desired structure, due to the bias-variance trade-off, prior expectations over a rich class of structures are needed to capture the structures inferred by people when given a limited number of observations.

2. Nonparametric Bayesian models are probabilistic models over arbitrarily complex structures that are biased toward simpler structures.

Representational primitive	Key Psychological Application	Nonparametric Bayesian Process	Key References
Partitions	Category assignment	Chinese restaurant process (CRP)	Aldous (1985)
Probability distributions over competing discrete units	Category learning/Density estimation	Dirichlet process (DP)	Ferguson (1973)
Probability distributions over independent discrete units	Feature assignment	Indian buffet process (IBP)/Beta process (BP)[1]	Griffiths and Ghahramani (2005); Thibaux and Jordan (2007)
Distributions over continuous units	Function learning	Gaussian process (GP)	Rasmussen and Williams (2006)
Composites	Hierarchical category learning	Nested CRP	Griffiths et al. (2008b); Blei et al. (2010)
	Jointly learning categories and features	CRP + IBP, nested CRP + hierarchical DP topic model	Austerweil and Griffiths (2013); Salakhutdinov et al. (2012)
	Cross-cutting category learning	CRP over entities embedded inside CRP over attributes	Shafto et al. (2011)
	Relational category learning	Product of CRPs over multiple entity types	Kemp et al. (2006, 2010)
	Property induction	GP over latent graph structure	Kemp and Tenenbaum (2009)
	Domain structure learning	GP/CRP/IBP + grammar over model forms	Kemp and Tenenbaum (2008); Grosse et al. (2012)

Fig. 9.7 Different assumptions about the type of structure generating the observations from the environment results in different types of nonparametric Bayesian models. The most basic nonparametric models define distributions over core representational primitives, while more advanced models can be constructed by composing these primitives with each other and with other probabilistic modeling and knowledge representation elements (see Box 2). Typically, researchers in cognitive science do not distinguish between the CRP and DP, or the IBP and BP. However, they are all distinct mathematical objects, where the CRP and IBP are distributions over the assignment of stimuli to units and the DP and BP are distributions over the assigment of stimuli to units and the parameters associated with those units. The probability distribution given by only considering the number of stimuli assigned to each unit by a DP and BP yields a distribution over assignments equivalent to the CRP and IBP, respectively.

Thus, they form a middle ground between the two extremes of models that infer overly simple structures (parametric models) and models that infer overly complex structures (nonparametric models).

3. Using different nonparametric models result in different assumptions about the format of the hidden structure. When each stimulus is assigned to a single latent unit, multiple latent units, or continuous units, the Dirichlet process, Beta process, and Gaussian process are appropriate, respectively. These processes are compositional in that they can be combined with each other and other models to infer complex latent structures, such as relations and hierarchies.

Some Future Questions

1. How similar are the inductive biases defined by nonparametric Bayesian models to those people use when inferring structured representations?

2. What are the limits on the complexity of representations that people can learn? Are nonparametric Bayesian models too powerful?

3. How do nonparametric Bayesian models compare to other computational frameworks that adapt their structure with experience, such as neural networks?

Notes

1. Note that we define parametric, nonparametric, and other statistical terms from the Bayesian perspective. We refer the reader interested in the definition of these terms from the frequentist perspective and a comparison of frequentist and Bayesian perspectives to Young and Smith (2010).

2. Our definition of "density estimation" includes estimating any probability distribution over a discrete or continuous space, which is slightly broader than its standard use in statistics, estimating any probability distribution over a continuous space.

3. Our formulation departs from Anderson's by adopting the notation typically used in the statistics literature. However, the two formulations are equivalent.

4. We have written the posterior probability as proportional to the product of three terms because the normalizing constant (the denominator) for this example is intractable to compute when there is an infinite repository of features.

5. Technically, to ensure that the infinite limit of $P(\mathbf{Z}|\alpha)$ is valid requires defining all feature ownership matrices that differ only in the order of the columns to be equivalent. This is due to identifiability issues and is analogous to the arbitrariness of the cluster (or table) labels in the CRP.

6. Austerweil and Griffiths (2011) tested whether people represent the objects with the parts as features by seeing if they were willing to generalize a property of the set of objects (being found in a cave on Mars) to a novel combination of three of the six parts used to create the images. See Austerweil and Griffiths (2011) and Austerweil and Griffiths (2013) for a discussion of this methodology and the theoretical implications of these results.

7. Although we collapsed the distinction between IBP and BP, they are distinct nonparametric Bayesian processes. See the caption of Figure 2 and the glossary for more details.

Glossary

Beta process: a stochastic process that assigns a real number between 0 and 1 to a countable set of units, which makes it a natural prior for latent feature models (interpreting the number as a probability)

bias: the error between the true structure and the average structure inferred based on observations from the environment

bias-variance trade-off: to reduce the error in generalizing a structure to new observations, an agent has to reduce both its bias and variance

Chinese restaurant process: a culinary metaphor that defines a probability distribution over partitions, which yields an equivalent distribution on partitions as the one implied by a Dirichlet process when only the number of stimuli assigned to each block is considered

computational level: interpreting the behavior of a system as the solution to an abstract computational problem posed by the environment

consistency: given enough observations, the statistical model infers the true structure producing the observations

Dirichlet process: a stochastic process that assigns a set of non-negative real numbers that sum to 1 to a countable set of units, which makes it a natural prior for latent class models (interpreting the assigned number as the probability of that unit)

exchangeability: a sequence of random variables is exchangeable if and only if their joint probability is invariant to reordering (does not change)

Gaussian process: a stochastic process that defines a joint distribution on a set of variables that is Gaussian, which makes it a natural prior for function learning models

importance sampling: approximating a distribution by sampling from a surrogate distribution and then reweighting the samples to compensate for the fact that they came from the surrogate rather than the desired distribution

Indian buffet process: a culinary metaphor for describing the probability distribution over the possible assignments of observations to multiple discrete units, which yields an equivalent distribution on discrete units as the one implied by a Beta process when only the number of stimuli assigned to each unit is considered

inductive inference: a method for solving a problem that has more than one logically possible solution

likelihood: the probability of some observed data given a particular structure or hypothesis is true

Markov chain Monte Carlo: approximating a distribution by setting up a Markov chain whose limiting distribution is that distribution

Monte Carlo: using random number sampling to solve numerical problems

nonparametric: a model whose possible densities belongs to a family that includes arbitrary distributions

parametric: a model that assumes possible densities belongs to a family that is parameterized by a fixed number of variables

particle filtering: a sequentially adapting importance sampler where the surrogate distribution is based on the approximated posterior at the previous time step

partition: division of a set into nonoverlapping subsets

posterior probability: an agent's belief in a structure or hypothesis after some observations

prior probability: an agent's belief in a structure or hypothesis before any observations

rational analysis: interpreting the behavior of a system as the ideal solution to an abstract computational problem posed by the environment usually with respect to some assumptions about the system's environment

rational process models: a process model that is a statistical approximation to the ideal solution given by probability theory

variance: the degree that the inferred structure changes across different possible observations from the environment

References

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

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

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

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

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

Austerweil, J. L. & Griffiths, T. L. (2010a). Learning hypothesis spaces and dimensions through concept learning. In S. Ohlsson & R. Camtrabone, (Ed.). *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 73–78. Austin, TX: Cognitive Science Society.

Austerweil, J. L. & Griffiths, T. L. (2010b). Learning invariant features using the transformed indian buffet process. In R. Zemel & J. Shawne-Taylor, (Ed.). *Advances in neural information processing systems* (Vol. 23, pp. 82–90. Cambridge, MA. MIT Press.

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

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

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

Bickel, P. J. & Doksum, K. A. (2007). *Mathematical statistics: basic ideas and selected topics*. Upper Saddle River, NJ: Pearson.

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

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM, 57*, 1–30.

Bouton, M. (2004). Context and behavioral processes in extinction. *Learning & Memory, 11*(5):485–494.

Bowers, J. S. & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*(3):389–414.

Braddick, O. (1993). Segmentation versus integration in visual motion processing. *Trends in neurosciences, 16*(7), 263–268.

Brehmer, B. (1971). Subjects' ability to use functional rules. *Psychonomic Science, 24*, 259–260.

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.

Brogden, W. (1939). Sensory pre-conditioning. *Journal of Experimental Psychology, 25*(4), 323–332.

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

Chater, N., Goodman, N., Griffiths, T. L., Kemp, C., Oaksford, M., & Tenenbaum, J. B. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. *Behavioral and Brain Sciences, 34*(4), 194–196.

Chater, N. & Oaksford, M. (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. New York, NY: Oxford University Press.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367–405.

Clapper, J. & Bower, G. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(2), 443–460.

Courville, A., Daw, N., & Touretzky, D. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences, 10*(7), 294–300.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience, 3*, 1218–1223.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: the sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 968–986.

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

Freedman, D. & Diaconis, P. (1983). On inconsistent Bayes estimates in the discrete case. *Annals of Statistics, 11*(4), 1109–1118.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias-variance dilemma. *Neural Computation, 4*, 1–58.

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., Blei, D., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review, 117*(1), 197–209.

Gershman, S. & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior, 40*(3), 255–268.

Gershman, S. J. & Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology, 56*(1), 1–12.

Ghosal, S. (2010). The Dirichlet process, related priors, and posterior asymptotics. In N. L. Hjort, C. Holmes, P. Müller, & S. G. Walker, (Eds.), *Bayesian nonparametrics* (pp. 35–79). Cambridge UK, Cambridge University Press.

Goldmeier, E. (1972). Similarity in visually perceived forms. *Psychological Issues, 8*, 1–136. Original written in German and published in 1936.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112*, 21–54.

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

Gordon, N., Salmond, J., & Smith, A. (1993). A novel approach to non-linear/non-Gaussian Bayesian state estimation. *IEEE Proceedings on Radar and Signal Processing, 140*, 107–113.

Görür, D., Jäkel, F., & Rasmussen, C. E. (2006). A choice model with infinitely many latent features. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, pages 361–368, New York. ACM Press.

Griffiths, T., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. (2010a). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*, 357–364.

Griffiths, T., Steyvers, M., & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244.

Griffiths, T. L. (2010). Bayesian models as tools for exploring inductive biases. In Banich, M. & Caccamise, D., editors, *Generalization of knowledge: Multidisciplinary perspectives*. Psychology Press, New York.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010b). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*(8), 357–364.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin, 138*(3), 415–422.

Griffiths, T. L. & Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. (Technical Report 2005-001, Gatsby Computational Neuroscience Unit).

Griffiths, T. L. & Ghahramani, Z. (2011). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research, 12*, 1185–1224.

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

Griffiths, T. L., Lucas, C., Williams, J. J., & Kalish, M. L. (2009). Modeling human function learning with Gaussian processes. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams & A. Culotta, (Eds.). *Advances in Neural Information Processing Systems 21*, pp. 556–563. Red Hook, NY: Curran.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008b). Categorization as nonparametric Bayesian

density estimation. In N. Chater, & M. Oaksford, (Eds.). *The probabilistic mind*. Oxford, England: Oxford University Press.

Griffiths, T. L. & Yuille, A. (2006). A primer on probabilistic inference. *Trends in Cognitive Sciences, 10*(7), 1–11.

Grosse, R. B., Salakhutdinov, R., Freeman, W. T., & Tenenbaum, J. B. (2012). Exploiting compositionality to explore a large space of model structures. In N. de Freitas & K. Murphy, (Eds.). *Conference on Uncertainty in Artificial Intelligence*, pp. 306–315. Corvallis, OR: AUAI Press.

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

Hu, Y., Zhai, K., Williamson, S., & Boyd-Graber, J. (2012). Modeling images using transformed Indian buffet processes. *International Conference of Machine Learning*, Edinburgh, UK.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research, 9*(1), 90–98.

Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York, NY: Wiley.

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

Jolliffe, I. T. (1986). *Principal component analysis*. New York, NY: Springer.

Jones, M. & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34*, 169–231.

Kahneman, D., Slovic, P., & Tversky, A., editors (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review, 111*(4), 1072–1099.

Kaplan, A. & Murphy, G. (1999). The acquisition of category structure in unsupervised learning. *Memory & Cognition, 27*(4), 699–712.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science, 10*(3), 307–321.

Kemp, C., Tenenbaum, J., Niyogi, S., & Griffiths, T. (2010). A probabilistic model of theory formation. *Cognition, 114*(2), 165–196.

Kemp, C. & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*(31), 10687–10692.

Kemp, C. & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116*(1), 20–58.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In Y. Gil & R. J. Mooney, (Eds.). *Proceedings of the 21st National Conference on Artificial Intelligence*, pp. 381–388. Menlo Park, CAAAAI Press.

Koh, K. & Meyer, D. E. (1991). Function learning: induction of continuous stimulus–response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(5), 811–836.

Kruschke, J. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior, 36*(3), 210–226.

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

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

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences, 14*(8), 348–356.

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 & Review, 12*(1), 24–42.

McKinley, S. C. & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance, 21*(1), 128–148.

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

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A, 209*, 415–446.

Miller, K. T., Griffiths, T. L., & Jordan, M. I. (2008). The phylogenetic Indian Buffet Process: A non-exchangeable nonparameteric prior for latent features. In D. McAllester & P. Myllymaki, (Eds.). *Proceedings of the Twenty-Fourth Conference on Uncertainity in Artificial Intelligence*, (pp. 403–410). Corvallis, Oregon: AUAI Press.

Murphy, G. L. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics, 9*(2), 249–265.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39–57.

Oaksford, M. & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. New York, NY: Oxford University Press.

Palmer, S. E. (1999). *Vision Science*. Cambridge, MA: MIT Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.

Posner, M. I. & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3p1), 353.

Pothos, E. & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26*(3), 303–343.

Rasmussen, C. E. & Williams, C. K. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

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

Rescorla, R. A. & Wagner, A. R. (1972). A theory of of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. Black, & W. Prokasy, (Eds.), *Classical conditioning II: Current research*

and theory (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Robert, C. P. (1994). *The Bayesian choice: A decision-theoretic motivation*. New York, NY: Springer.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review, 108*(2), 370–392.

Rumelhart, D. & Greeno, J. (1971). Similarity between stimuli: An experimental test of the Luce and Restle choice models. *Journal of Mathematical Psychology, 8*, 370–381.

Salakhutdinov, R., Tenenbaum, J. B., & Torralba, A. (2012). Learning to learn with compound hierarchical-deep models. In *Advances in Neural Information Processing Systems*.

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.

Schölkopf, B. & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge, MA.

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

Selfridge, O. G. & Neisser, U. (1960). Pattern recognition by machine. Scientific American, 203*, 60–68.*

Shafto, P., Kemp, C., Manishka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. Cognition, 120*(1), 1–25.*

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Science*, 10:309–318.*

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331*, 1279–1285.*

Thibaux, R. & Jordan, M. I. (2007). Hierarchical Beta processes and the Indian buffet process. In Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*, pages 564–571.*

Tversky, A. (1972). Elimination by aspects: A theory of choice. Psychological Review, 79*, 281–299.*

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

Werker, J. & Yeung, H. (2005). Infant speech perception bootstraps word learning. Trends in Cognitive Sciences, 9*(11), 519–527.*

Williams, C. K. (1998). Prediction with gaussian processes: From linear regression to linear prediction and beyond. In M. Jordan, (Ed.). Learning in graphical models *(pp. 599–621). Cambridge, MA: MIT Press.*

Wood, F., Griffiths, T. L., & Ghahramani, Z. (2006). A nonparametric Bayesian method for inferring hidden causes. Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence (UAI '06)*, 536–543.*

Yildirm, I. & Jacobs, R. A. (2012). A rational analysis of the acquisition of multisensory representations. Cognitive Science, 36*, 305–332.*

Young, G. A. & Smith, R. L. (2010). Essentials of statistical inference*. Cambridge, UK: Cambridge University Press.*