Learning Overhypotheses

Charles Kemp, Amy Perfors & Joshua B. Tenenbaum
{ckemp, perfors, jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

Inductive learning is impossible without overhypotheses, or constraints on the hypotheses considered by the learner. Some of these overhypotheses must be innate, but we suggest that hierarchical Bayesian models help explain how the rest can be acquired. The hierarchical approach also addresses a common question about Bayesian models of cognition: where do the priors come from? To illustrate our claims, we consider two specific kinds of overhypotheses — overhypotheses about feature variability (e.g. the shape bias in word learning) and overhypotheses about the grouping of categories into ontological kinds like objects and substances.

Compared to machine-learning algorithms, humans are remarkable for doing so much with so little. A single labelled example is enough for children to learn the meanings of some words, and children develop grammatical constructions that are rarely found in the sentences they hear [2]. These inductive leaps appear even more impressive when we consider the many interpretations of the data that are logically possible but apparently never entertained by human learners [5].

Several authors have proposed that the apparent ease of human learning depends on constraints that guide induction. This view has been applied to many cognitive problems: the M-constraint and the shape bias help explain concept acquisition, universal grammar guides the acquisition of linguistic knowledge [2], and the development of folk biology is guided by the idea that living kinds can be organized hierarchically. Constraints like these may be called framework theories or schemata, but we will borrow a term of Goodman's and refer to them as overhypotheses.[1]

Some overhypotheses must be innate, but others are probably learned. For at least two reasons the acquisition of overhypotheses has received less attention than it deserves. First, the authors who have argued most convincingly for the importance of overhypotheses often suggest that these overhypotheses are innate [2]. Second, the study of overhypothesis acquisition raises some formidable methodological challenges. Designing adult experiments to address the problem is difficult, since adults bring a lifetime of learning experience to any experiment and have already distilled overhypotheses that help them deal with most novel tasks. Infant experiments are challenging for different reasons, but worth pursuing because they can address the acquisition of some of the most fundamental overhypotheses. For instance, Smith and colleagues have explored the development of the shape bias [9, 3], and Piaget and colleagues have explored how abstract kinds of knowledge (such as the concrete operations) can be acquired and used to support many different learning tasks.

This paper argues that hierarchical Bayesian modelling [4] is a formal approach that should help to explain the acquisition of overhypotheses in many different domains. Hierarchical Bayesian models (HBMs) include representations at many levels of abstraction, and show how knowledge can be acquired at levels that are quite remote from the data given by experience. To illustrate our general claim, we describe one of the simplest possible HBMs and use it to suggest how overhypotheses about feature-variability are acquired and used to support categorization. One such overhypothesis is the shape bias, the expectation that shape is a feature that is homogeneous within object categories. We also present an extension of the basic model that groups categories into ontological kinds (e.g. objects and substances) and discovers the features and the patterns of feature variability that are characteristic of each kind.

The problem of overhypothesis acquisition is closely related to a problem raised by Bayesian models of cognition. These models usually rely on a prior distribution chosen by the modeller, and a natural response is to wonder where the prior comes from. HBMs address this question: in the framework we adopt, learning an overhypothesis amounts to learning a prior distribution. The two models we consider provide concrete examples of how priors can be learned from data.

Overhypotheses and HBMs

Goodman introduces the idea of overhypotheses using bags of colored marbles [5]. Suppose that S is a stack containing many bags of marbles. We empty several bags and discover that some bags contain black marbles, others contain white marbles, but that each bag is uniform in color. We now choose a new bag — bag n — and draw a single black marble from the bag. This observation may lead us to endorse the following hypothesis:

[1] Other authors distinguish between theories, schemata, scripts, and overhypotheses. There are important differences between these varieties of abstract knowledge, but it is useful to have a single term (for us, overhypothesis) that includes them all.

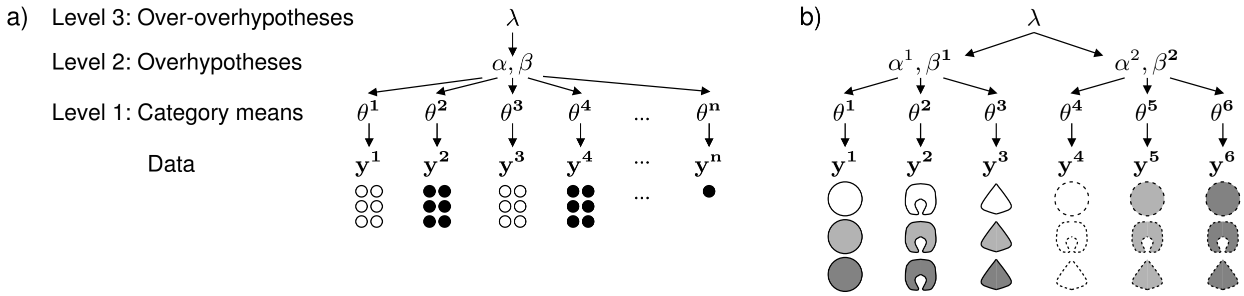

Figure 1: (a) A hierarchical Bayesian model. Each setting of $(\alpha, \boldsymbol{\beta})$ is an overhypothesis: $\boldsymbol{\beta}$ is the expected color distribution for a category, and α represents the variability in color within each category. (b) A model with separate overhypotheses for two ontological kinds meant to correspond loosely to objects and substances. α^1 represents knowledge about feature variability within the first ontological kind (object categories are homogeneous in shape but not in material), and $\boldsymbol{\beta}^{\mathbf{1}}$ captures the characteristic features of the first kind (objects tend to be solid).

H: All marbles in bag n are black.

If asked to justify the hypothesis, we might invoke the following overhypothesis:

O: All bags in S are uniform in color.

Goodman gives a precise definition of 'overhypothesis' but we use the term more generally to refer to any form of abstract knowledge that sets up a hypothesis space at a less abstract level. By this criterion, O is an overhypothesis since it sets up a space of hypotheses about bag n: it could be uniformly black, uniformly white, uniformly green, and so on. To give a very different example, Universal Grammar is an overhypothesis that sets up a space of hypotheses (i.e. a space of possible grammars) for language learning.

Hierarchical Bayesian models [4] capture this notion of overhypothesis by allowing hypothesis spaces at several levels of abstraction. We give an informal introduction to this modelling approach, leaving all technical details for the next section. Suppose that we are given a body of data, and we wish to account for a certain cognitive ability. In Goodman's case, the data are observations of several bags ($\boldsymbol{y^i}$ indicates the observations for bag i) and we are interested in the ability to predict the color of next marble to be drawn from bag n. The first step is to identify a kind of knowledge (level 1 knowledge) that explains the data and that supports the ability of interest. In Goodman's case, level 1 knowledge is knowledge about the color distribution of bags ($\boldsymbol{\theta^i}$ indicates the color distribution for the ith bag).

We then ask how the level 1 knowledge can be acquired, and the answer will make reference to a more abstract body of knowledge (level 2 knowledge). For the marbles scenario, level 2 knowledge is knowledge about the distribution of the $\boldsymbol{\theta}$ variables. As described in the next section, this knowledge can be represented using two parameters, α and $\boldsymbol{\beta}$ (Figure 1a). Roughly speaking, α captures the extent to which individual bags are uniform in colour, and $\boldsymbol{\beta}$ captures the average color distribution across the entire stack of bags. If we now go on to ask how the level 2 knowledge might be acquired, the answer will rely on a body of knowledge at an even higher level, level 3. In Figure 1a, this knowledge is represented by λ, which captures prior knowledge about the values of α and $\boldsymbol{\beta}$. The parameter λ and the pair $(\alpha, \boldsymbol{\beta})$ are both overhypotheses, since each sets up a hypothesis space at the next level down. We will assume that the level 3 knowledge is specified in advance, and show how an overhypothesis can be learned at level 2.

Within cognitive science, linguists have provided the most familiar example of this style of model building. Language comprehension can be explained using structural descriptions of sentences (level 1 knowledge). Structural descriptions, in turn, can be explained with reference to a grammar (level 2 knowledge), and the acquisition of this grammar can be explained with reference to Universal Grammar (level 3 knowledge). There are few settings where cognitive modellers have gone beyond three levels, but there is no principled reason to stop at level 3. Ideally, we should continue adding levels until the knowledge at the highest level is simple enough or general enough that it can be plausibly assumed to be innate.

As the grammar-learning example suggests, it has long been known that hierarchical models are capable in principle of explaining the acquisition of overhypotheses. The value of hierarchical *Bayesian* models in particular is that they explain how overhypotheses can be acquired by rational statistical inference. Given observations at the lowest level of a HBM, statistical inference can be used to compute posterior distributions over entities at the higher levels. In the model of Figure 1a, for instance, acquiring an overhypothesis is a matter of acquiring knowledge at level 2. The posterior distribution $p(\alpha, \boldsymbol{\beta}|\boldsymbol{y})$ represents a normative belief about level 2 knowledge — the belief, given the data y, that most bags are close to uniform in color.

We have argued that HBMs go beyond previous hierarchical models proposed by cognitive scientists, but they also represent an advance over the standard Bayesian models used in cognitive science. A standard Bayesian model has two levels of knowledge: the elements in its hypothesis space represent level 1 knowledge, and the prior (generally fixed) represents knowledge at level 2. A common objection to Bayesian modelling is that priors can be chosen to approximate almost any pattern

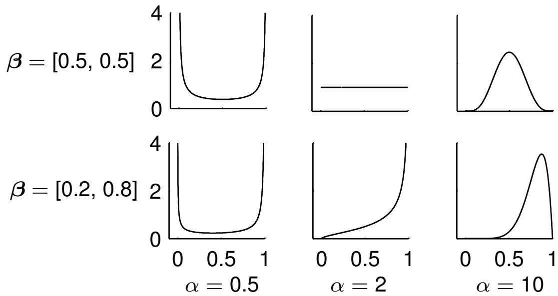

Figure 2: The 2-dimensional Dirichlet distribution serves as a prior on $\boldsymbol{\theta}$, the color distribution of a bag of marbles. Let θ_2 be the proportion of black marbles within the bag. Shown here are distributions on θ_2 when the parameters of the Dirichlet distribution (α and $\boldsymbol{\beta}$) are systematically varied. When α is small, most bags are near-uniform in color (θ_2 is close to 0 or close to 1). When α is large, θ_2 is expected to be close to β_2.

of data, and that the success of a Bayesian model depends critically on the modeller's ability to choose the right prior. HBMs disarm this objection by showing that knowledge at level 2 need not be specified in advance, but can be learned from raw data. Of course, the prior at the highest level must still be specified in advance, but the ultimate goal is to design models where this prior is simple enough to be unobjectionable.

Computational Theory

We now describe one formal instantiation of the model in Figure 1a. There may be other ways to formalize overhypotheses about feature variability, but ours is perhaps the simplest account of how these overhypotheses can be acquired and used to guide learning at lower levels. Suppose we are working with a set of k colors. Initially we set $k = 2$ and use white and black as the colors. Let $\boldsymbol{\theta^i}$ indicate the true color distribution for the ith bag in the stack: if 60% of the marbles in bag 7 are black, then $\boldsymbol{\theta^7} = [0.4, 0.6]$. Let y^i indicate a set of observations of the marbles in bag i. If we have drawn 5 marbles from bag 7 and all but one are black, then $\boldsymbol{y^7} = [1, 4]$.

We assume that $\boldsymbol{y^i}$ is drawn from a multinomial distribution with parameter $\boldsymbol{\theta^i}$: in other words, the marbles responsible for the observations in $\boldsymbol{y^i}$ are drawn independently at random from the ith bag, and the color of each depends on the color distribution $\boldsymbol{\theta^i}$ of that bag. The vectors $\boldsymbol{\theta^i}$ are drawn from a Dirichlet distribution parameterized by a scalar α and a vector $\boldsymbol{\beta}$. Here $\boldsymbol{\beta}$ represents the expected distribution of colors across the stack and α captures the notion of feature variability (Figure 2). The larger the value of α, the more likely that the color distribution for any given bag will be close to the vector $\boldsymbol{\beta}$. When α is small, however, each individual bag is likely to be near-uniform in color, and $\boldsymbol{\beta}$ will determine the relative proportions of 'mostly black' and 'mostly white' bags.

Each possible setting of $(\alpha, \boldsymbol{\beta})$ is an overhypothesis. In order to discover values for these variables, we need prior distributions on $\boldsymbol{\beta}$ and α. We use a uniform distribution (Dirichlet($\mathbf{1}$)) on $\boldsymbol{\beta}$ and an exponential distribution with mean λ on α (Figure 3a). For all simulations in this paper we set $\lambda = 1$.

This model is known to statisticians as a Dirichlet-multinomial model [4]. Using statistical notation, it can be written as:

$$\begin{aligned}
\alpha &\sim \text{Exponential}(1) \\
\boldsymbol{\beta} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\theta^i} &\sim \text{Dirichlet}(\alpha, \boldsymbol{\beta}) \\
\boldsymbol{y^i} \mid n^i &\sim \text{Multinomial}(\boldsymbol{\theta^i})
\end{aligned}$$

where n^i is the number of observations for bag i. As written, the model assumes we are working with a single dimension — for Goodman, marble color. Perhaps, however, some marbles are made from metal and others are made from glass. We deal with multiple dimensions by assuming that each dimension is independently generated according to the model, and introducing separate values of α and $\boldsymbol{\beta}$ for each dimension. When working with multiple features, we often use α to refer to the collection of α values along all dimensions, and $\boldsymbol{\beta}$ for the set of all $\boldsymbol{\beta}$ vectors.

To fit the model to data we assume that counts y are observed for one or more bags, and use a Markov chain Monte Carlo (MCMC) scheme to draw a sample from $p(\alpha, \boldsymbol{\beta}|y)$, the posterior distribution on $(\alpha, \boldsymbol{\beta})$. Figures 3b and 3c show posterior distributions on α and $\boldsymbol{\beta}$ for two sets of counts. Predictions about $\boldsymbol{\theta}^{\text{new}}$, the color distribution of a new, sparsely observed bag can be computed by calculating the mean prediction made by all pairs $(\alpha, \boldsymbol{\beta})$ in the MCMC sample. Note that each possible setting of $(\alpha, \boldsymbol{\beta})$ specifies a prior distribution on $\boldsymbol{\theta}^{\text{new}}$. By showing how knowledge about $(\alpha, \boldsymbol{\beta})$ can be acquired, we therefore show how prior distributions can be acquired.

Modelling behavioral data

Since Goodman, psychologists have confirmed that adults [8] and children [7] have overhypotheses about feature variability, and use them to make inductive leaps given very sparse data. Nisbett et al. [8] asked subjects to imagine they were exploring an island in the Southeastern Pacific. As part of the task, subjects were told that they had encountered a single member of the Barratos tribe, and that the tribesman was brown and obese. Based on this single example, subjects concluded that most Barratos were brown, but gave a much lower estimate of the proportion of obese Barratos (Figure 3d). When asked to justify their responses, subjects often said that tribespeople were "homogeneous with respect to color" but "heterogeneous with respect to body weight."

To apply our model to this task, we replace bags of marbles with tribes. Suppose we have observed 20 members from each of 20 tribes. Half the tribes are brown and the other half are white, but all of the individuals in a given tribe share the same skin color. Given these data, the model learns a posterior distribution on α indicating that α is probably small, which means that skin color

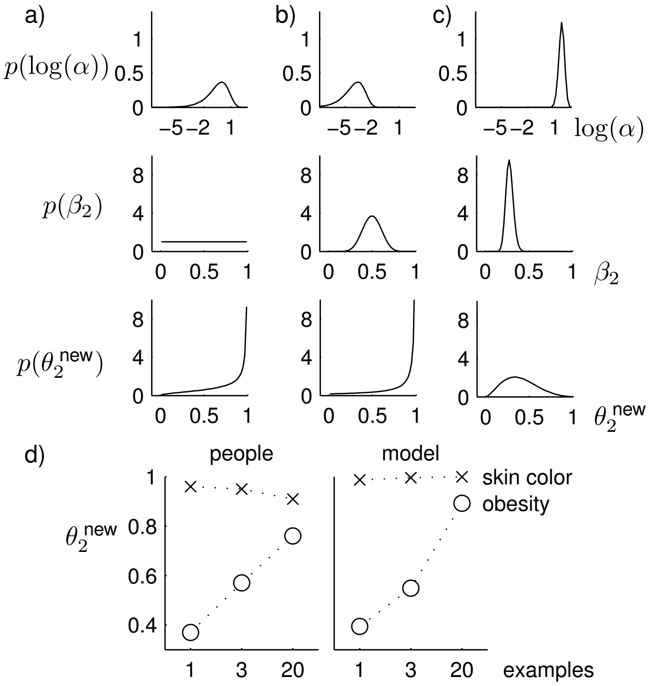

Figure 3: (a-c) Distributions on $\log(\alpha)$ and $\boldsymbol{\beta}$ given three patterns of data: (a) before any data have been observed; (b) after observing 10 all-white bags and 10 all-black bags; (c) after observing 20 mixed bags inspired by the obesity condition of the Barratos task. The final row shows beliefs about the color distribution ($\boldsymbol{\theta}^{\text{new}}$) of a new bag from which a single black marble has been drawn. After 20 bags that are either all white or all black (b), the model realizes that most bags are near-uniform in color (α is small), and that about half of these bags are black (β_2 is around 0.5). These posterior distributions allow the model to predict that nearly all the marbles in the new, sparsely observed bag will be black (θ_2^{new} is close to 1). (d) Generalizations about a new tribe after seeing 1, 3, or 20 obese, brown-skinned individuals from that tribe. Human generalizations are replotted from Nisbett et al. [8].

tends to be homogeneous within tribes (Figure 3b). We can also make predictions about a sparsely observed new tribe: having observed a single, brown-skinned member of a new tribe, the posterior distribution on $\boldsymbol{\theta}^{\text{new}}$ indicates that most members of the tribe are likely to be brown (Figures 3b and 3d).

Suppose now that obesity is a feature that varies within tribes: a quarter of the 20 tribes observed have an obesity rate of 10%, and the remaining quarters have rates of 20%, 30%, and 40%. Obesity is represented in our model as a second binary feature, and the posterior distributions on α and $\boldsymbol{\beta}$ (Figure 3c) indicate that obesity varies within tribes (α is high), and that the base rate of obesity is around 30% (β_2 is around 0.3). Again, we can use these posterior distributions to make predictions about a new tribe, and our model now requires many observations before it concludes that most members of the new tribe are obese (Figure 3b).

a)

	Training	T_1	T_2
Category	1 1 2 2 3 3 4 4	1	5
Shape	1 1 2 2 3 3 4 4	1	5
Texture	1 2 3 4 5 6 7 8	1	9
Color	1 2 3 4 5 6 7 8	1	9
Size	1 2 1 2 1 2 1 2	1	1

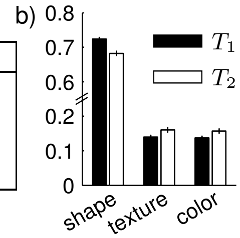

Figure 4: (a) Data based on Smith et al. [9]. Each column represents an object, and there are 10 possible colors, textures, and shapes, and 2 possible sizes. First and second-order generalization were tested using exemplars T_1 and T_2. (b) Generalizations for the data in (a). Three possible matches were provided for each exemplar. The plot shows probabilities (normalized) that each match belongs to the same category as the exemplar. The model makes exact predictions about these probabilities: we computed 30 MCMC estimates of these predictions, and the error bars represent the standard error of the mean.

Novel feature values

Our model of the Barratos task does not address an important kind of reasoning that overhypotheses support: reasoning about *new* feature values. At least two overhypotheses might be induced from the data in Figure 1a: the first states that all bags are uniform in color, and the second states that every bag is entirely white or entirely black. These two overhypotheses have very different consequences: for example, only the first can handle a case where a single green marble is drawn from a new bag.

Many real-world problems involve inferences about novel features. Children know, for example, that animals of the same species tend to make the same sound. Observing one horse neigh is enough to conclude that most horses neigh, even though a child may never have heard an animal neigh before. Similarly, children show a "shape bias:" they know that shape tends to be homogeneous within object categories. Given a single exemplar of a novel object category, children extend the category label to similarly shaped objects ahead of objects that share the same texture or color as the exemplar.

The model in Figure 1a deals naturally with inferences like these. We illustrate using stimuli inspired by the work of Smith et al. [9]. In their second experiment, Smith et al. [9] trained 17-month olds on four novel categories with two exemplars each. Category labels were provided during training. Within each category, the two exemplars had the same shape but differed in size, texture and color (Figure 4a). After training, the authors tested *first-order* generalization by presenting T_1, an exemplar from one of the training categories, and asking children to choose another exemplar from the same category as T_1. Three possible matches were provided, each of which matched T_1 in exactly one feature (shape, color or size). Children preferred the shape match, showing that they were sensitive to feature distributions within a known category. Smith et al. [9] also tested *second-order* generalization by presenting children with T_2, an exem-

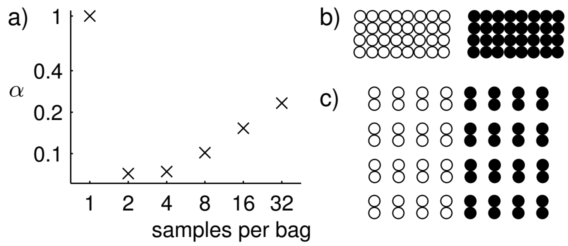

Figure 5: (a) Mean α values after seeing 32 white marbles and 32 black marbles. Low values of α indicate that bags are expected to be homogeneous in color. The model is most confident that bags are homogeneous when given the data in (c). (b) Data set with 32 samples per bag (c) Data set with 2 samples per bag

plar from a novel category. Again, children preferred the shape match, revealing knowledge that shape in general is a reliable indicator of category membership. Note that this result depends critically on the training summarized by Figure 4a: children of this age do not normally reveal a shape bias on tests of second-order generalization.

We supplied our model with counts $\boldsymbol{y^i}$ computed from the feature vectors in Figure 4a. The key modelling step is to allow for more values along each dimension than appear in the training set. We allowed for 10 shapes, 10 colors, 10 textures and 2 sizes. For example, the count vector $\boldsymbol{y^1}$ says that the observed exemplars of category 1 include 2 objects with shape value 1 and no objects with shape value 10. This policy allows the model to handle shapes, colors and textures it has never seen during training, but assumes, of course, that the model is able to recognize a novel shape as a kind of shape, a novel color as a kind of color, and so on.

Figure 4b shows the patterns of generalization predicted by the model. Smith et al. [9] report that shape matches are chosen 88% of the time given exemplar T_1, and 70% of the time given exemplar T_2. The model reproduces this general pattern: shape matches are preferred in both cases, and are preferred more strongly when the exemplar belongs to a familiar category.

Smith et al. [9] also measured real-world generalization by tracking vocabulary growth over an eight week period. They showed that experience with the eight exemplars in Figure 4a led to a significant increase in the number of object names used by children. Our model helps to explain this striking result. Even though the training set includes only four categories, the results in Figure 4b show that it contains enough statistical information to establish or reinforce the shape bias. Similarly, our model explains why providing only two exemplars per category is sufficient. In fact, if the total number of exemplars is fixed, our model predicts that the best way to teach the shape bias is to provide just two exemplars per category. We illustrate by returning to the marbles scenario.

Each point in Figure 5 represents a simulation where 64 observations of marbles are evenly distributed over some number of bags. The marbles drawn from any given bag are uniform in color — black for half of the bags and white for the others. When 32 observations are provided for each of two bags, the model has strong evidence about the color distribution of each bag: one is mostly black and the other is mostly white. The case where two observations are provided for each of 32 bags represents the other extreme. Now the evidence about the composition of any single bag is weak, but taken together, these observations provide strong support for the idea that α is low and most bags are homogeneous.

The U-shaped curve in Figure 5 is a novel prediction of our model that could be tested in developmental experiments. More generally, the curve illustrates how a learner might become relatively certain about an overhypothesis (e.g. the value of α) even though she is uncertain about the individual entities described by the overhypothesis (e.g. the $\boldsymbol{\theta}$ values for categories when only two exemplars per category are observed). This insight appears relevant to many learning problems: to give just one additional example, a hierarchical model of grammar induction may be able to explain how a learner becomes confident about some aspect of a grammar even though most of the individual sentences that support this conclusion are poorly understood.

Discovering ontological kinds

The model in Figure 1a is a simple HBM that acquires something like the shape bias, but to match the capacities of a child it is necessary to apply the shape bias selectively — to object categories, for example, but not to substance categories. Selective application of the shape bias appears to demand knowledge that categories are grouped into ontological kinds and that there are different patterns of feature variability within each kind. Before the age of three, for instance, children appear to know that shape tends to be homogeneous within object categories but heterogeneous within substance categories, that color tends to be homogeneous within substance categories but heterogeneous within object categories, and that both shape and texture tend to be homogeneous within animate categories.

Figure 1b shows how we can give our model the ability to discover ontological kinds. The model assumes that categories *may* be grouped into ontological kinds, and that each kind is associated with a different α and $\boldsymbol{\beta}$. The model, however, is not told which categories belong to the same kind, and is not even told how many different kinds it should look for. Instead, we give it a prior distribution on the partition of categories into kinds. Intuitively, this prior should assign some probability to all possible category partitions, but favor the simpler partitions — those that use a small number of kinds. We satisfy both criteria using a prior generated by a Chinese Restaurant Process [1].

The new model can be written as:

$$
\begin{aligned}
z \mid n_{\text{cat}} &\sim \text{CRP}(\gamma) \\
\alpha^k &\sim \text{Exponential}(\lambda) \\
\boldsymbol{\beta^k} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\theta^i} &\sim \text{Dirichlet}(\alpha^{z_i}\boldsymbol{\beta^{z_i}}) \\
\boldsymbol{y^i} \mid n^i &\sim \text{Multinomial}(\boldsymbol{\theta^i})
\end{aligned}
$$

a)

	Training	S	N
Category	1 1 2 2 3 3 4 4	5	6
Shape	1 1 2 2 3 4 5 6	7	8
Material	1 2 3 4 5 5 6 6	7	8
Size	1 2 1 2 1 2 1 2	1	1
Solidity	1 1 1 1 2 2 2 2	1	2

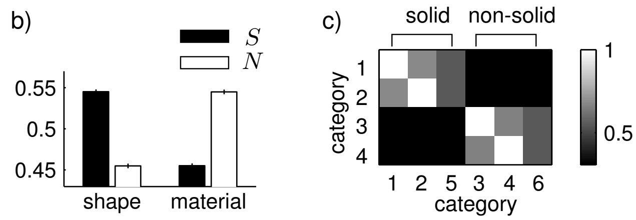

Figure 6: (a) Data used to test the model in Figure 1b. Second-order generalization was tested using solid and non-solid exemplars (S and N). (b) Generalizations for the data in (a). The model chooses the shape match given the solid exemplar and the material match given the non-solid exemplar. (c) Entry (i, j) in the matrix is the posterior probability that categories i and j belong to the same ontological kind. The model groups categories 1, 2 and 5 (solid categories) and 3, 4 and 6 (non-solid categories).

where n_{cat} is the number of categories, γ is the concentration parameter for the CRP, z_i is the kind label for category i and there is a separate α^k and $\boldsymbol{\beta^k}$ for each ontological kind k.

Jones and Smith [6] have shown that training young children on a handful of suitably structured categories can promote the acquisition of ontological knowledge. We gave our model a data set of comparable size (Figure 6a). During training, the model saw two exemplars from each of four categories: two object categories and two substance categories. Exemplars of each object category were solid, matched in shape, and differed in material and size. Exemplars of each substance category were non-solid, matched in material, and differed in shape and size. Second-order generalization was tested using exemplars from novel categories — one test exemplar (S) was solid and the other (N) was not. Figure 6b shows that the model chooses a shape match for the solid exemplar and a material match for the non-solid exemplar.

Figure 6c confirms that the model correctly groups the stimuli into two ontological kinds: object categories and substance categories. This discovery is based on the characteristic features of ontological kinds ($\boldsymbol{\beta}$) as well as patterns of feature variability within each kind (α). If the object categories are grouped into kind k, α^k indicates that shape is homogeneous within categories of that kind, and $\boldsymbol{\beta^k}$ indicates that categories of that kind tend to be solid. The $\boldsymbol{\beta}$ parameter, then, is responsible for the inference that the test exemplar S should be grouped with the two object categories, since all three categories contain solid objects.

Discussion

We presented hierarchical Bayesian models that help explain the acquisition of the shape bias, and of overhypotheses about feature variability within ontological kinds. We know of no previous attempts to provide rational computational theories of the acquisition of the shape bias, or of other overhypotheses about feature variability. Colunga and Smith [3] present a connectionist model that acquires knowledge of this sort, but our approach is different in emphasis and explanatory effect. We provided a computational theory but have not attempted to specify the psychological mechanisms by which it might be implemented, and Colunga and Smith [3] describe a process model but do not provide a rational computational theory. A second difference is that our probabilistic approach extends naturally to contexts where structured representations are required: computational linguists, for example, work with probabilistic grammars that generate parse trees. It is less clear how a connectionist approach might deal with representations more complex than lists of features.

Although we have argued that overhypotheses can be acquired by HBMs, we do not claim that overhypotheses can be generated out of thin air. Any HBM will assume that the process by which each level is generated from the level above is known, and that the prior at the topmost level is provided. Any account of induction must rely on *some* initial knowledge: the real question for a learning framework is whether it allows us to build models that require no initial assumptions beyond those we are willing to make. Whether the hierarchical Bayesian approach will meet this challenge is not yet clear, but it deserves to be put to the test.

Acknowledgments Supported by the William Asbjornsen Albert memorial fellowship (CK), a NDSEG fellowship (AP) and the Paul E. Newton chair (JBT).

References

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

[2] Chomsky, N. (1980). *Rules and Representations.* Basil Blackwell, Oxford.

[3] Colunga, E. and Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychological Review*, 112(2).

[4] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis.* Chapman & Hall, New York.

[5] Goodman, N. (1955). *Fact, Fiction, and Forecast.* Harvard University Press, Cambridge.

[6] Jones, S. S. and Smith, L. B. (2002). How children know the relevant properties for generalizing object names. *Developmental Science*, 5(2):219–232.

[7] Macario, J. F., Shipley, E. F., and Billman, D. O. (1990). Induction from a single instance: formation of a novel category. *Journal of Experimental Child Psychology*, 50:179–199.

[8] Nisbett, R. E., Krantz, D. H., Jepson, C., and Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 90(4):339–363.

[9] Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., and Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, 13(1):13–19.