Discovering Inductive Biases in Categorization through Iterated Learning

Kevin R. Canini (kevin@cs.berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
University of California, Berkeley, CA 94720 USA

Wolf Vanpaemel (wolf.vanpaemel@psy.kuleuven.be)
Department of Psychology, University of Leuven, Belgium

Michael L. Kalish (kalish@louisiana.edu)
Institute of Cognitive Science, University of Louisiana at Lafayette, Lafayette, LA 70504 USA

Abstract

Progress in studying human categorization has typically involved comparing generalization judgments made by people to those made by models for a variety of training conditions. In this paper, we explore an alternative method for understanding human category learning—iterated learning—which can directly expose the inductive biases of human learners and categorization models. Using a variety of stimulus sets, we compare the results of iterated learning experiments with human learners to results from two prominent classes of computational models: prototype models and exemplar models. Our results indicate that human learning is not perfectly captured by either type of model, lending support to the theory that people use intermediate representations between these two extremes.

Keywords: categorization; inductive bias; iterated learning; Bayesian methods

Introduction

The ability to learn new categories from examples is a basic component of human cognition, and one of the earliest to be studied by psychologists (Hull, 1920). This long history of investigation has resulted in a number of computational models of category learning, including approaches based on inferring decision rules (Ashby & Gott, 1988), extracting prototypes (Reed, 1972), memorizing exemplars (Medin & Schaffer, 1978; Nosofsky, 1986), and combinations of these methods (Nosofsky, Palmeri, & McKinley, 1994; Vanpaemel & Storms, 2008). This proliferation of models has been complemented by an empirical literature comparing the ability of different models to account for human behavior. In a typical experiment, participants are taught the category membership of a set of training stimuli and then asked to generalize to a set of test stimuli. Computational models are evaluated on their ability to predict the resulting patterns of generalization.

Competing models of category learning are commonly presented in terms of their different assumptions about people's mental representations of categories and the processes that translate these representations into behavior. However, we can also think about these models more abstractly: as methods of learning categories that have different *inductive biases*. In machine learning, the inductive bias of a learner is defined to be those factors other than the observed data that lead the learner to favor one hypothesis over another (Mitchell, 1997). Different models of category learning favor different kinds of hypotheses about the structure of categories. For example, a prototype model favors hypotheses in which categories are coherent groups of stimuli, while an exemplar model is more flexible, and can represent categories that consist of multiple clusters of stimuli spread out across a stimulus space (Nosofsky, 1998). Evaluating these models thus becomes a problem of determining the nature of human inductive biases.

In this paper, we use a novel approach to evaluate different models of category learning. Rather than studying the generalizations people make with different training stimuli, we use an experimental method designed to provide direct access to people's and models' inductive biases. In this experimental method, *iterated learning*, each participant is trained with stimuli that are selected from the responses of the previous participant. This results in a sequence of category structures each produced by learning from the previous structure. Mathematical analysis of this process shows that as the sequence gets longer, the structures that emerge will be consistent with the inductive biases of the learners (Griffiths & Kalish, 2007). Intuitively, iterated learning magnifies the small effects that inductive biases have on people's generalizations, until those biases are all that is reflected in the data. We use iterated learning to expose the inductive biases of human learners and compare them to those of categorization models. Our work demonstrates that iterated learning complements traditional categorization experiments and provides a new dataset against which computational models can be compared.

Models of category learning

A wide range of formal approaches have been used to model human categorization. In this paper, we organize our analysis around two of the most prominent models—prototype and exemplar models—illustrating how our approach can be used to evaluate categorization models by empirically exploring human inductive biases. In future work, we hope to extend this analysis to incorporate a more extensive range of models.

Prototype models

Prototype models of categorization represent each category with a single point—the *prototype*—which captures the central tendency of that category (Reed, 1972). The similarity of a novel stimulus x to a category j is given by $\eta_j(x) = \exp\{-d(x,\mu_j)\}$, where μ_j is the prototype of category j, and $d(\cdot,\cdot)$ is some distance metric between stimuli. The distance metric can be chosen to be more sensitive to certain dimensions, reflecting the fact that category members may have more or less variance along each dimension. Given a collec-

tion of observed category members, the probability of classifying a novel object x under category j is calculated as

$$P(j|x) = \frac{\beta_j \eta_j(x)^\gamma}{\sum_{j'} \beta_{j'} \eta_{j'}(x)^\gamma}, \quad (1)$$

where β_j is a response bias towards category j, and γ is a response scaling parameter.

Exemplar models

Exemplar models (Medin & Schaffer, 1978; Nosofsky, 1986) represent a category with all of its observed members. Rather than calculating a single prototype for each category, exemplar models sum over all previously observed examples, the *exemplars*. The similarity of a novel stimulus to category j is given by $\eta_j(x) = \sum_{y \in j} \exp\{-d(x,y)\}$, where y is an exemplar belonging to category j, and $d(\cdot,\cdot)$ is again some suitable distance metric between stimuli. Given a collection of categories and observations, the probability of classifying a novel object x under category j is the same as in the prototype models, given by Equation 1.

Interpolating between prototypes and exemplars

Prototype and exemplar models can be viewed as opposite ends of a spectrum of models which vary in the complexity of their representations. Prototype models use the simplest representation: a single point for each category, while exemplar models use the most complex representation: all observed category members. Recently, models have been developed which interpolate between these extremes by grouping the observed stimuli into clusters and representing each cluster using a single point. These models adopt a flexible representation where clusters are added as warranted by the data. Examples include SUSTAIN (Love, Medin, & Gureckis, 2004), the varying abstraction model (Vanpaemel & Storms, 2008), the rational model of categorization (Anderson, 1991), and the hierarchical Dirichlet process (Griffiths, Canini, Sanborn, & Navarro, 2007; Teh, Jordan, Beal, & Blei, 2006). Because these models can behave like prototype models, exemplar models, or anything in-between, they can potentially explain experimental results that suggest that people use flexible representations to learn categories.

Iterated learning

The categorization models introduced in the previous section correspond to learners with different capabilities and preferences for category representations. In categorization research, comparisons of different models typically proceed by presenting each model (as well as human participants) with a set of training data and comparing the generalization predictions made by the learners. While this method allows us to quantitatively measure the degree to which each model explains the human data, it does not directly expose the underlying inductive biases of the learners. Iterated learning is an experimental method designed to give a pure estimate of inductive biases (Griffiths & Kalish, 2007).

The central concept of the iterated learning framework is that the training data given to a learner (either a human participant or a learning model) are not directly specified by the experimental design; rather, they are sampled from a previous learner's generalization responses. The learners are arranged into a chain, where the responses from the first learner are used as training data for the second learner, and so on. Because each learner's responses depend only on the previous learner's, the chain is formally a Markov process, and therefore the responses will converge to a stationary distribution.

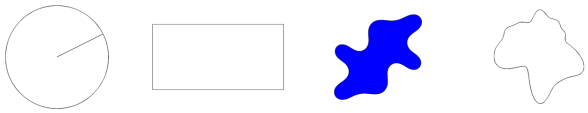


Figure 1: Stimuli used in the experiment. (a) Shepard circles, (b) rectangles, (c) Cortese blobs, and (d) Shepard blobs.

Griffiths and Kalish (2007) provided an analysis of iterated learning under the assumption that learners use Bayesian inference, sampling hypotheses from the posterior distribution given by Bayes' rule: $P(h|d) \propto P(h)P(d|h)$. In this case, the observed responses in the iterated learning chain will converge to the prior distribution $P(h)$, therefore allowing us to directly expose the inductive biases of the learners in the form of the prior over hypotheses.

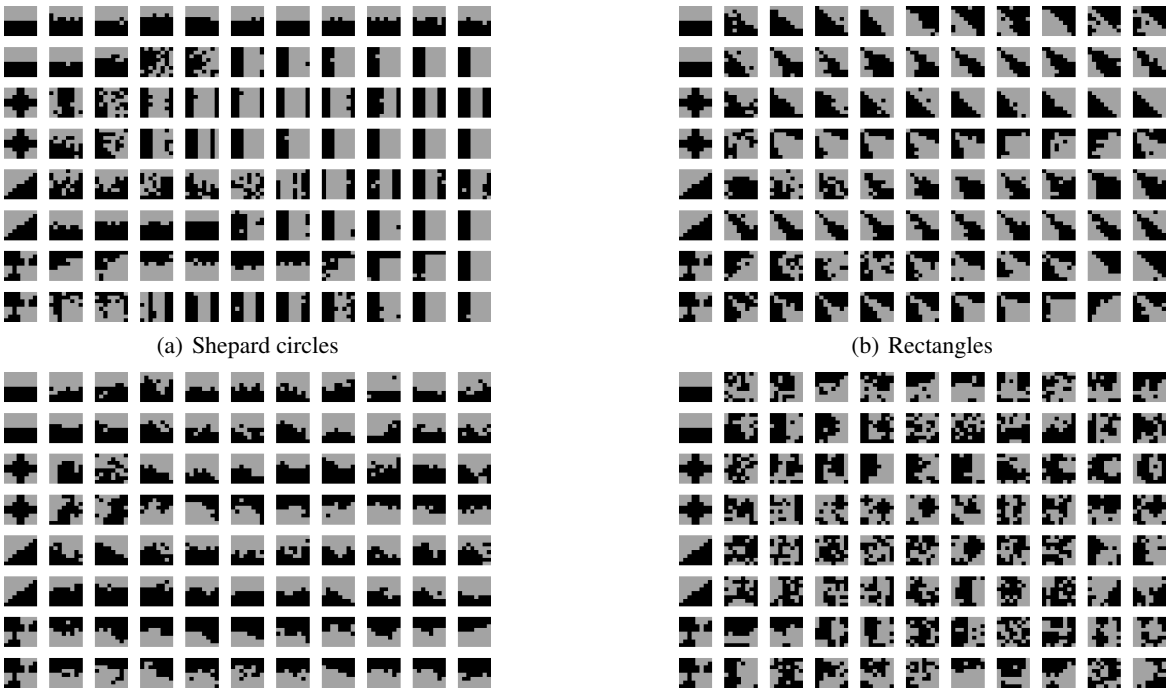
Exploring human inductive biases

Using the iterated learning methodology, we performed a categorization experiment to explore the inductive biases of both people and models and to create a new dataset which can be used as a resource by other researchers.

Method

Participants The experimental participants included 640 workers from Amazon Mechanical Turk, who received a payment of $0.50, and 160 students at the University of California, Berkeley, who received course credit, for a total of 800. The experiment had 16 conditions, resulting from the combination of four stimulus sets and four initial category structures. The eight conditions with the Cortese blobs and Shepard circles were each replicated six times, and the eight conditions with the other two stimulus sets were each replicated four times. Each replication of each condition consisted of an iterated learning chain of 10 generations. Each participant was randomly assigned to a chain in their pool (either Mechanical Turk or Berkeley students), occupying the next available generation in the chain.

Stimuli The experiment involved four different sets of stimuli, each of which varied on two dimensions (see Figure 1). Two of the stimulus sets had separable dimensions, meaning the dimensions on which they varied are easily differentiated. These were rectangles that varied in their width and height, and "Shepard circles" (Shepard, 1964): circles of a varying diameter with a radius drawn at a varying angle. The other two stimulus sets had integral dimensions: their dimensions are not readily apparent, leaving no preferred coordinate sys-



(a) Shepard circles (b) Rectangles

(c) Cortese blobs (d) Shepard blobs

Figure 2: Samples of human data from the experiment. Each row is an iterated learning chain; two replications of each condition are shown. Black vs. gray pixels indicate category membership, and each image is the generalization responses of a single learner. Each learner learned from examples from the categories shown to the immediate left. In (a), the *y*-axis is the circle's diameter and the *x*-axis is the angle of its radius. In (b), the *y*-axis is the rectangle's width and the *x*-axis is its height.

tems for these stimuli in psychological space. These were both amoeba-like shapes, one from Cortese and Dyre (1996) which we call "Cortese blobs", and the other from Shepard and Cermak (1973) which we call "Shepard blobs". The construction of both of these stimulus sets involves varying the amplitudes and phase-shifts of components of periodic, trigonometric functions, which are then converted to closed loops. For each stimulus set, we constructed an equal-spaced, 8-by-8 grid of stimuli and used these 64 to train and test the human learners.

Procedure Each participant completed a training phase and a test phase. In the training phase, the participant was trained to reproduce the category memberships of a random selection of 32 of the 64 stimuli. In the test phase, each participant classified all 64 items in random order without feedback. In each training trial, the participant classified a single stimulus from the training set with feedback. For first-generation learners, this feedback was based on one of four initial category structures, which are shown in the first columns of Figure 2. Two of the initializations—the first and third distinct ones—are simple linear boundaries compatible with a prototype model. The other two are discretized versions of category structures described by McKinley and Nosofsky (1995). For the remaining generations, feedback was provided according to the test phase responses of the participant in the previous generation. Participants were not made aware that their test responses would be used in later generations and did not have any contact with other learners from different generations. The training phase was organized into blocks containing 32 trials each, with the order of presentation of the stimuli randomized within each block.

If the participant correctly answered at least 22 of the 32 training trials[1] in any training block, they continued to the test phase. Otherwise, they completed another block of the training phase. If after 20 blocks or 25 minutes, a participant had not yet reached the learning threshold, the experiment was ended, and the data collected so far were not included in further analyses. There were 21 participants who reached the maximum number of blocks and 16 who reached the time limit without achieving the criterion. These participants were replaced by others to fill in their positions in the chains.

Results

No significant differences were found between the two participant pools, so their data were combined in all further analyses. Figure 2 shows two representative chains of 10 generations for each of the 16 conditions, with gray vs. black pixels indicating category membership.[2] In each row, the first panel shows the initial category structure, and all other panels show the category assignments made by a learner in the test phase

[1] 22/32 correct responses indicates with $p < 0.05$ that the responses are not purely random, according to an exact Binomial test.

[2] To promote further exploration of the results by other researchers, the full set of results is available online at `http://cocosci.berkeley.edu/iteratedCatData/`.

		Prototype model				Exemplar model			
Stimulus set	Dimensions	Distribution*	Covariance*	γ	ε	r*	c	γ	ε
Shepard circles	Separable	Laplace	independent	1.3731	0.1516	1	0.8245	2.0678	0.1448
Rectangles	Separable	Laplace	independent	0.9034	0.1662	1	0.9257	1.5144	0.1651
Cortese blobs	Integral	Normal	full	0.6516	0.0434	2	0.3717	3.7737	0.0417
Shepard blobs	Integral	Normal	full	1.0195	0.5093	2	0.8171	1.1993	0.3096

Table 1: The model parameters fit to the human data. * shows parameters fixed by the experimenter rather than fit to the data. γ are response scaling parameters, ε are noise mixture parameters, r is the exponent of the distance metric, and c is specificity.

after being trained on the category structure to its left.

Most of the Shepard circle chains converged to fairly simple structures using categorization boundaries aligned with one of the dimensions. For the rectangles, people seem to prefer three main types of category structures: one with a category of items on or near the main diagonal (corresponding to squares and square-like rectangles), one with a boundary between the categories along the main diagonal (corresponding to wide vs. tall rectangles), and one with a category along the top and left borders (corresponding to very narrow or very short rectangles). The Cortese blob chains seem to favor boundaries which are roughly aligned with the horizontal axis, but with some variability in their curvature. The results for the Shepard blobs seem quite noisy. Perhaps people interpreted these stimuli in feature spaces which are rather different from the dimensions we used to plot the results, or perhaps because these stimuli are difficult to interpret, people's inductive biases about them are very weak.

Convergence analysis For all of the stimulus sets, the chains appear to have converged to their stationary distributions. To quantitatively verify this, we performed a clustering analysis of the test phase data. The category structures from each generation of each chain were clustered using the k-means algorithm, with the variation of information (VI) metric (Meila, 2003) used as the distance function between pairs of category structures. The VI metric is a measure of the distance between partitions, so it depends only on how stimuli are classified, and not the locations of those stimuli in the feature space. The VI metric is invariant to relabelings of the categories, so two structures which are identical but switch the category labels would have a VI distance of zero.

Clustering the results from all conditions and generations of human data, we found that using 10 clusters gave a reasonable result. We used a χ^2 test on the histograms of the number of responses in each of the 10 clusters, comparing across pairs of generations in all the chains. We found statistically significant differences ($p < 0.05$) between the initial category structures and all others, as well as between the first generation of learners and each of the last two generations. This analysis suggests that the overall distribution of responses has converged to the stationary distribution by the second generation. To be conservative, we used only the last five generations of human data in our further evaluations.

Comparing human and model inductive biases

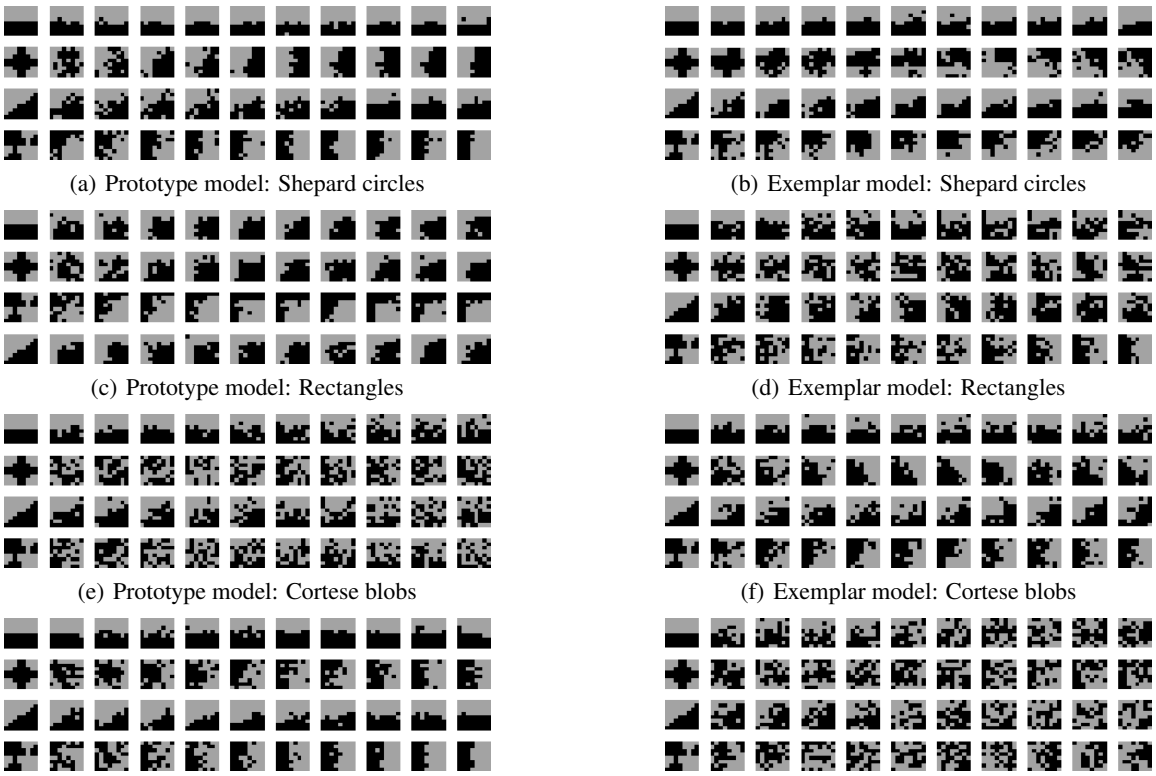
The experimental results described above give a picture of the inductive biases of human learners for various stimulus sets. To evaluate whether human inductive biases are consistent with those of the categorization models, we performed the same iterated learning procedure using the models.

Deriving the inductive biases of the models

We first set the various model parameters, fixing some based on the properties of the stimulus sets and fitting others to the human data. The results of the experiment with human learners using the rectangle stimulus set suggest that people prefer an alternative set of dimensions: the logarithms of the area (width $\times$ height) and aspect ratio (width $\div$ height) of the rectangles. These dimensions roughly correspond to the main diagonals in the plots in Figure 2(b). Indeed, previous work indicates that the logarithms of the area and aspect ratio are more psychologically salient dimensions than width and height (Krantz & Tversky, 1975). Correspondingly, we performed the model fitting and subsequent analyses using this alternative feature space for the rectangle stimulus set.

For stimuli with separable dimensions, it is appropriate to use an ℓ_1 (city-block) distance metric, while for stimuli with integral dimensions, an ℓ_2 (Euclidean) distance metric is appropriate (Shepard, 1964). Therefore, for the separable stimuli, the prototype model's similarity function was chosen to be the product of Laplace distribution functions on each dimension d: $\eta_j(x) = \prod_d \exp\{-|x_d - \mu_{j,d}|/b_{j,d}\}/2b_{j,d}$, where $\mu_{j,d}$ and $b_{j,d}$ are the parameters of the category prototype. This implies the distance function $d(x,(\mu_j,b_j)) = \sum_d(|x_d - \mu_{j,d}|/b_{j,d} - 2b_{j,d})$. The prototype parameter $\mu_{j,d}$ was set as the sample median of the observed values on dimension d, and we set $b_{j,d} = 1/N_j \sum_i |x_{i,d} - \mu_{j,d}|$, the average absolute difference between the category members and the median $\mu_{j,d}$. These correspond to maximum likelihood parameters for the Laplace distribution. For the integral stimuli, the prototype model's similarity function was chosen to be the multivariate normal distribution, using the maximum likelihood estimates for the mean and covariance matrix parameters. For all stimuli, we set the response biases $\beta_j = \frac{1}{2}$.

The distance function of the exemplar model was set to $d(x,y) = c(\sum_d |x_d - y_d|^r)^{1/r}$, with $r = 1$ for the separable stimulus sets and $r = 2$ for the integral sets, corresponding to ℓ_1 and ℓ_2 distance metrics, respectively. The parameter c is the model's *specificity*—analogous to the variance-tuning parameters in the prototype models—and was fit to the human data separately for each stimulus set. The response bias β_j was set to be $1/N_j$, the inverse of the number of category members, to remove the inherent bias of the exemplar model to prefer categories with more observed members, a bias which is not present in the prototype model and would



(a) Prototype model: Shepard circles

(b) Exemplar model: Shepard circles

(c) Prototype model: Rectangles

(d) Exemplar model: Rectangles

(e) Prototype model: Cortese blobs

(f) Exemplar model: Cortese blobs

(g) Prototype model: Shepard blobs

(h) Exemplar model: Shepard blobs

Figure 3: Fitted model simulations. The format is the same as Figure 2. One replication per condition and model is shown. The rectangle stimuli were presented to the models using the alternative dimensions of log-area and log-aspect ratio, but the plots above use height and width as the x and y axes.

otherwise introduce a confounding factor in their comparison.

The response scaling parameter γ was fit to the human data separately for each stimulus set and each model. Additionally, we found that a certain proportion of the human learners appeared to be responding at random during the test phase, so all models were mixed with a noise component, from which responses were assumed to be generated uniformly at random. Participants were probabilistically assigned to the noise component using the expectation-maximization (EM) algorithm, with the prior probability ε of noise component membership being fit to the data. The results of the model fitting procedures are summarized in Table 1.

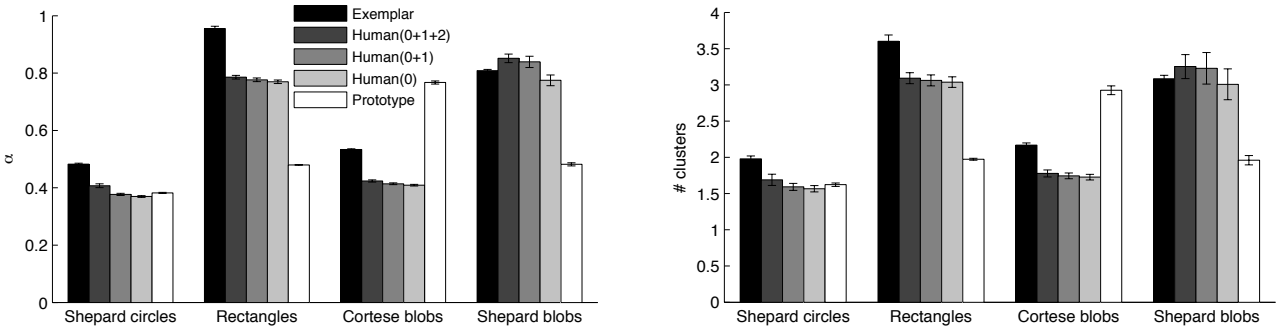
The fitted models were run through the same iterated learning experimental framework as the human learners, with four replications of each condition. As with the human data, only the last five generations of each chain were used for analysis. A sample of the results is shown in Figure 3.

Evaluation of human and model results

To quantitatively compare the human data to the model results, we fit a Dirichlet process mixture model (DPMM) (Ferguson, 1973) to each set of responses. The DPMM is a model which probabilistically clusters a set of observed data, where the number of clusters is inferred from the data rather than specified as a parameter. One of its hyperparameters, α, indirectly controls the number of inferred clusters by making additional clusters more or less likely. When α is large, more clusters are inferred, and in the limit $\alpha \to \infty$, each datapoint is assigned to its own cluster. When α is small, fewer clusters are inferred, and in the limit $\alpha \to 0$, only a single cluster is inferred. In this way, the DPMM generalizes both the prototype and exemplar models, depending on the choice of α (Sanborn, Griffiths, & Navarro, 2006). By specifying a prior distribution, the value of α can be inferred from the observed data rather than being set at a fixed value.

While the DPMM is a useful model of categorization (Griffiths et al., 2007), it also provides us a way of analyzing the responses of human category learners and categorization models by inspecting the inferred number of clusters in their category structures. Using a Gibbs sampling procedure, we fit a DPMM to each set of responses from the human data and model results, collecting a set of samples from the posterior distribution over the number of clusters and the value of α. The results of this analysis are summarized in Figure 4.

For the Shepard circles, the prototype model seems to provide a better fit to the human data, while for the Shepard blobs, the exemplar model is a better match. For the rectangles, neither model seems to capture the inductive bias of human learners, and the results suggest that a model using intermediate representations might be a better fit. For the Cortese blobs, the prototype model produces results which have more clusters than either the human data or the exemplar results;



(a) Inferred α value

(b) Inferred number of clusters per category

Figure 4: Evaluation results for the human data and model simulations. Error bars have length twice the standard error. The bars labeled "Human(0+1+2)" reflect all the human data, "Human(0+1)" are participants assigned to the noise component of at most one model, and "Human(0)" are those not assigned to the noise component of either model.

this can be explained by the relatively low response scaling parameter that was fit for this stimulus set (see Table 1). A more psychologically plausible feature space might improve the modeling results for the Cortese blobs.

Conclusions and future work

As a whole, our results suggest that the human learners' inductive biases are not always consistent with those of prototype or exemplar models, but can vary depending on the stimuli. This supports the notion that models which use more flexible representations and can interpolate between the behavior of prototypes and exemplars provide a better explanation of the variable nature of human categorization. However, perhaps our most significant contribution is the creation of a new dataset for evaluating categorization models, which we hope will be subjected to further analyses by other researchers.

In future work, we plan to conduct analyses using more psychologically plausible feature spaces for these stimuli, which can be obtained through multidimensional scaling studies. We also plan to extend our analysis to include intermediate models of categorization mentioned earlier, which interpolate between prototypes and exemplars. Work by Griffiths et al. (2007) has shown that in traditional categorization studies, the hierarchical Dirichlet process is capable of explaining human data that neither prototypes nor exemplars adequately model; we hope that these results can also be replicated using the iterated learning experimental method.

Acknowledgements This work was supported by grants IIS-0845410 and BCS-0704034 from the National Science Foundation.

References

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

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *J. Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33-53.

Cortese, J., & Dyre, B. (1996). Perceptual similarity of shapes generated from Fourier descriptors. *J. Experimental Psychology: Humam Perception and Performance*, *22*(1), 133–43.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*(2), 209–230.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the 29th annual conference of the cognitive science society* (p. 323-328).

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science: A Multidisciplinary Journal*, *31*(3), 441–480.

Hull, C. (1920). Quantitative aspects of the evolution of concepts. *Psychological Monographs*, *XXVIII*(1).

Krantz, D. H., & Tversky, A. (1975). Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology*, *12*(1), 4–34.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309-332.

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*(3), 207-238.

Meila, M. (2003). Comparing clusterings by the variation of information. In B. Schlkopf & M. K. Warmuth (Eds.), *Learning theory and kernel machines* (Vol. 2777, p. 173-187). Springer.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw Hill.

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

Nosofsky, R. M. (1998). Optimal performance and exemplar models of classification. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (p. 218-247). Oxford University Press.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53-79.

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

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th annual conference of the cognitive science society.*

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*(1), 54–87.

Shepard, R. N., & Cermak, G. W. (1973). Perceptual-cognitive explorations of a toroidal set of free-form stimuli. *Cognitive Psychology*, *4*(3), 351–377.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566-1581.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*(4), 732-749.