Uncovering Mental Representations with Markov Chain Monte Carlo

Adam N. Sanborn
Indiana University, Bloomington

Thomas L. Griffiths
University of California, Berkeley

Richard M. Shiffrin
Indiana University, Bloomington

Abstract

A key challenge for cognitive psychology is the investigation of mental representations, such as object categories, subjective probabilities, choice utilities, and memory traces. In many cases, these representations can be expressed as a non-negative function defined over a set of objects. We present a behavioral method for estimating these functions. Our approach uses people as components of a Markov chain Monte Carlo (MCMC) algorithm, a sophisticated sampling method originally developed in statistical physics. Experiments 1 and 2 verified the MCMC method by training participants on various category structures and then recovering those structures. Experiment 3 demonstrated that the MCMC method can be used estimate the structures of the real-world animal shape categories of giraffes, horses, dogs, and cats. Experiment 4 combined the MCMC method with multidimensional scaling to demonstrate how different accounts of the structure of categories, such as prototype and exemplar models, can be tested, producing samples from the categories of apples, oranges, and grapes.

Determining how people mentally represent different concepts is one of the central goals of cognitive psychology. Identifying the content of mental representations allows us to explore the correspondence between those representations and the world, and to understand their role in cognition. Psychologists have developed a variety of sophisticated techniques for estimating certain kinds of mental representations, such as the spatial or featural repre-

---

The authors would like to thank Jason Gold, Rich Ivry, Michael Jones, Woojae Kim, Krystal Klein, Tania Lombrozo, Chris Lucas, Angela Nelson, Rob Nosofsky and Jing Xu for helpful comments. ANS was supported by an Graduate Research Fellowship from the National Science Foundation and TLG was supported by grant number FA9550-07-1-0351 from the Air Force Office of Scientific Research while completing this research. Experiments 1, 2, and 3 were conducted while ANS and TLG were at Brown University. Preliminary results from Experiments 2 and 3 were presented at the 2007 Neural Information Processing Systems conference.

sentations underlying similarity judgments, from behavior (e.g., Shepard, 1962; Shepard & Arabie, 1979; Torgerson, 1958). However, the standard method used for investigating the vast majority of mental representations, such as object categories, subjective probabilities, choice utilities, and memory traces, is asking participants to make judgments on a set of stimuli that the researcher selects before the experiment begins. In this paper, we describe a novel experimental procedure that can be used to efficiently identify mental representations that can be expressed as a non-negative function over a set of objects, a broad class that includes all of the examples mentioned above.

The basic idea behind our approach is to design a procedure that produces samples of stimuli that are not chosen before the experiment begins, but are adaptively selected to concentrate in regions where the function in question has large values. More formally, for any non-negative function $f(x)$ over a space of objects $\mathcal{X}$, we can define a corresponding probability distribution $p(x) \propto f(x)$, being the distribution obtained when $f(x)$ is normalized over $\mathcal{X}$. Our approach provides a way to draw samples from this distribution. For example, if we are investigating a category, which is associated with a probability distribution or similarity function over objects, this procedure will produce samples of objects that have high probability or similarity for that category. We can also use the samples to compute any statistic of interest concerning a particular function. A large number of samples will provide a good approximation to the means, variances, covariances, and higher moments of the distribution $p(x)$. This makes it possible to test claims about mental representations made by different cognitive theories. For instance, we can compare the samples that are produced with the claims about subjective probability distributions that are made by probabilistic models of cognition (e.g., Anderson, 1990; Ashby & Alfonso-Reese, 1995; Oaksford & Chater, 1998).

The procedure that we develop for sampling from these mental representations is based on a method for drawing samples from complex probability distributions known as Markov chain Monte Carlo (MCMC) (an introduction is provided by Neal, 1993). In this method, a Markov chain is constructed in such a way that it is guaranteed to converge to a particular distribution, allowing the states of the Markov chain to be used in the same way as samples from that distribution. Transitions between states are made using a series of local decisions as to whether to accept a proposed change to a given object, requiring only limited knowledge of the underlying distribution. As a consequence, MCMC can be applied in contexts where other Monte Carlo methods fail. The first MCMC algorithms were used for solving challenging probabilistic problems in statistical physics (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), but they have become commonplace in a variety of disciplines, including statistics (Gilks, Richardson, & Spiegelhalter, 1996) and biology (Huelsenbeck, Ronquist, Nielsen, & Bollback, 2001). These methods are also beginning to be used in modeling of cognitive psychology (Farrell & Ludwig, in press; Griffiths, Steyvers, & Tenenbaum, 2007; Morey, Rouder, & Speckman, 2008; Navarro, Griffiths, Steyvers, & Lee, 2006)

In standard uses of MCMC, the distribution $p(x)$ from which we want to generate samples is known – in the sense that it is possible to write down at least a function proportional to $p(x)$ – but difficult to sample from. Our application to mental representations is more challenging, as we are attempting to sample from distributions that are unknown to the researcher, but implicit in the behavior of our participants. To address this problem, we

designed a task that will allow people to act as elements of an MCMC algorithm, letting us sample from the distribution $p(x)$ associated with the underlying representation $f(x)$. This task is a simple two-alternative forced choice. Much research has been devoted to relating the magnitude of psychological responses to choice probabilities, resulting in mathematical models of these tasks (e.g., Luce, 1963). We point out an equivalence between a model of human choice behavior and one of the elements of an MCMC algorithm, and then use this equivalence to develop a method for obtaining information about different kinds of mental representations. This allows us to harness the power of a sophisticated sampling algorithm to efficiently explore these representations, while never presenting our participants with anything more exotic than a choice between two alternatives.

While our approach can be applied to any representation that can be expressed as a non-negative function over objects, our focus in this paper will be on the case of object categories. Grouping objects into categories is a basic mental operation that has been widely studied and modeled. Most successful models of categorization assign a category to a new object according to the similarity of that object to the category representation. The similarity of all possible new objects to the category representation can be thought of as a non-negative function over these objects, $f(x)$. This function, which we are interested in determining, is a consequence of not only category membership, but also of the decisional processes that extend partial category membership to objects that are not part of the category representation.

Our method is certainly not the first developed to investigate categorization. Researchers have invented many methods to investigate the categorization process: some determine the psychological space in which categories lie, others determine how participants choose between different categories, and still others determine what objects are members of a category. All of these methods have been applied in simple training experiments, but they do not always scale well to the large stimulus spaces needed to represent natural stimuli. Consequently, the majority of experiments that investigate human categorization models are training studies. Most of these training experiments use simple stimuli with only a few dimensions, such as binary dimensional shapes (Shepard, Hovland, & Jenkins, 1961), lengths of a line (Huttenlocher, Hedges, & Vevea, 2000), or Munsell colors (Nosofsky, 1987; Roberson, Davies, & Davidoff, 2000), though more complicated stimuli have also been used (e.g., Posner & Keele, 1968). Research on natural perceptual categories has focused on exploring the relationships of category labels (Storms, Boeck, & Ruts, 2000; Rosch & Mervis, 1975; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), with a few studies examining members of a perceptual category using simple parameterized stimuli (e.g., Labov, 1973).

It is difficult to investigate natural perceptual categories with realistic stimuli because natural stimuli have a great deal of variability. The number of objects that can be distinguished by our senses is enormous. Our MCMC method makes it possible to explore the structure of natural categories in relatively large stimulus spaces. We illustrate the potential of this approach by estimating the probability distributions associated with a set of natural categories, and asking whether these distributions are consistent with different models of categorization. An enduring question in the categorization literature is whether people represent categories in terms of exemplars or prototypes (J. D. Smith & Minda, 1998; Nosofsky & Zaki, 2002; Minda & Smith, 2002; Storms et al., 2000; Nosofsky, 1988; Nosofsky & Johansen, 2000; Zaki, Nosofsky, Stanton, & Cohen, 2003; E. E. Smith, Patalano, & Jonides,

1998; Heit & Barsalou, 1996). In an exemplar-based representation, a category is simply the set of its members, and all of these members are used when evaluating whether a new object belongs to the category (e.g., Medin & Schaffer, 1978; Nosofsky, 1986). In contrast, a prototype model assumes that people have an abstract representation of a category in terms of a typical or central member, and use only this abstract representation when evaluating new objects (e.g., Posner & Keele, 1968; Reed, 1972). These representations are associated with different probability distributions over objects, with exemplar models corresponding to a nonparametric density estimation scheme that admits distributions of arbitrary structure and prototype models corresponding to a parametric density estimation scheme that favors unimodal distributions (Ashby & Alfonso-Reese, 1995). The distributions for different natural categories produced by MCMC could thus be used to discriminate between these two different accounts of categorization.

The plan of the paper is as follows. The next section describes MCMC in general and the Metropolis method and Hastings acceptance rules in particular. We then describe the experimental task we use to connect human judgments to MCMC, and outline how this task can be used to study the mental representation of categories. Experiment 1 tests three variants of this task using simple one-dimensional stimuli and trained distributions, while Experiment 2 uses one of these variants to show that our method can be used to recover different category structures from human judgments. Experiment 3 applies the method to four natural categories, providing estimates of the distributions over animal shapes that people associate with giraffes, horses, cats, and dogs. Experiment 4 combines the MCMC method with multidimensional scaling in order to explore the structure of natural categories in a perceptual space. This experiment illustrates how the suitability of a simple prototype model for describing natural categories could be tested by combining these two methodologies. Finally, in the General Discussion we discuss how MCMC relates to other experimental methods, and describe some potential problems that can be encountered in using MCMC, together with possible solutions.

Markov chain Monte Carlo

Generating samples from a probability distribution is a common problem that arises when working with probabilistic models in a wide range of disciplines. This problem is rendered more difficult by the fact that there are relatively few probability distributions for which direct methods for generating samples exist. For this reason, a variety of sophisticated Monte Carlo methods have been developed for generating samples from arbitrary probability distributions. Markov chain Monte Carlo (MCMC) is one of these methods, being particularly useful for correctly generating samples from distributions where the probability of a particular outcome is multiplied by an unknown constant – a common situation in statistical physics (Newman & Barkema, 1999) and Bayesian statistics (Gilks et al., 1996).

A Markov chain is a sequence of random variables where the value taken by each variable (known as the state of the Markov chain) depends only on the value taken by the previous variable (Norris, 1997). We can generate a sequence of states from a Markov chain by iteratively drawing each variable from its distribution conditioned on the previous variable. This distribution expresses the *transition probabilities* associated with the Markov chain, which are the probabilities of moving from one state to the next. An *ergodic* Markov chain, having a non-zero probability of reaching any state from any other state in a finite and

aperiodic number of iterations, eventually converges to a *stationary distribution* over states. That is, after many iterations the probability that a chain is in a particular state converges to a fixed value, the stationary probability, regardless of the initial value of the chain. The stationary distribution is determined by the transition probabilities of the Markov chain.

Markov chain Monte Carlo algorithms are schemes for constructing Markov chains that have a particular distribution (the "target distribution") as their stationary distribution. If a sequence of states is generated from such a Markov chain, the states towards the end of the sequence will arise with probabilities that correspond to the target distribution. These states of the Markov chain can be used in the same way as samples from the target distribution, although the correlations among the states that result from the dynamics of the Markov chain will reduce the effective sample size. The original scheme for constructing these types of Markov chains is the Metropolis method (Metropolis et al., 1953), which was subsequently generalized by Hastings (1970).

The Metropolis method

The Metropolis method specifies a Markov chain by splitting the transition probabilities into two parts: a *proposal distribution* and an *acceptance function*. A proposed state is drawn from the proposal distribution (which can depend on the current state) and the acceptance function determines the probability with which that proposal is accepted as the next state. If the proposal is not accepted, then the current state is taken as the next state. This procedure defines a Markov chain, with the probability of the next state depending only on the current state, and it can be iterated until it converges to the stationary distribution. Through careful choice of the proposal distribution and acceptance function, we can ensure that this stationary distribution is the target distribution.

A sufficient, but not necessary, condition for specifying a proposal distribution and acceptance function that produce the appropriate stationary distribution is *detailed balance*, with

$$p(x)q(x^*|x)a(x^*;x) = p(x^*)q(x|x^*)a(x;x^*) \tag{1}$$

where $p(x)$ is the stationary distribution of the Markov chain (our target distribution), $q(x^*|x)$ is the probability of proposing a new state x^* given the current state x, and $a(x^*;x)$ is the probability of accepting that proposal. We can motivate detailed balance by first imagining that we are running a Markov chain in which we draw a new proposed state from a uniform distribution over all possible proposals and then accept the proposal with some probability. The transition probability of the Markov chain is equal to the probability of accepting this proposal. We would like for the states of the Markov chain to act as samples from a target distribution. A different way to think of this constraint is to require the ratio of the number of visits of the chain to state x to the number of visits of the chain to state x^* to be equal to the ratio of the probabilities of these two states under the target distribution. A sufficient, but not necessary, condition for satisfying this requirement is to enforce this constraint for every transition probability. Formally we can express this equality between ratios as

$$\frac{p(x)}{p(x^*)} = \frac{a(x;x^*)}{a(x^*;x)} \tag{2}$$

Now, to generalize, instead of assuming that each proposal is drawn uniformly from the space of parameters, we assume that each proposal is drawn from a distribution that depends on

the current state, $q(x^*|x)$. In order to correct for a non-uniform proposal we need to multiply each side of Equation 2 by the proposal distribution. The result is equal to the detailed balance equation.

The Metropolis method specifies an acceptance function which satisfies detailed balance for any symmetric proposal distribution, with $q(x^*|x) = q(x|x^*)$ for all x and x^*. The Metropolis acceptance function is

$$a(x^*; x) = \min\Big(\frac{p(x^*)}{p(x)}, 1\Big). \tag{3}$$

meaning that proposals with higher probability than the current state are always accepted. To apply this method in practice, a researcher must be able to determine the relative probabilities of any two states. However, it is worth noting that the researcher need not know $p(x)$ exactly. Since the algorithm only uses the ratio $p(x^*)/p(x)$, it is sufficient to know the distribution up to a multiplicative constant. That is, the Metropolis method can still be used even if we only know a function $f(x) \propto p(x)$.

Hastings acceptance functions

Hastings (1970) gave a more general form for acceptance functions that satisfy detailed balance, and extended the Metropolis method to admit asymmetric proposal distributions. The key observation is that Equation 1 is satisfied by any acceptance rule of the form

$$a(x^*; x) = \frac{s(x^*, x)}{1 + \frac{p(x)}{p(x^*)}\frac{q(x^*|x)}{q(x|x^*)}} \tag{4}$$

where $s(x^*, x)$ is a symmetric in x and x^* and $0 \leq a(x^*; x) \leq 1$ for all x and x^*. Taking

$$s(x^*, x) = \begin{cases} 1 + \frac{p(x)}{p(x^*)}\frac{q(x^*|x)}{q(x|x^*)} \text{ , if } \frac{p(x^*)}{p(x)}\frac{q(x|x^*)}{q(x^*|x)} \geq 1 \\ \\ 1 + \frac{p(x^*)}{p(x)}\frac{q(x|x^*)}{q(x^*|x)} \text{ , if } \frac{p(x^*)}{p(x)}\frac{q(x|x^*)}{q(x^*|x)} \leq 1 \end{cases} \tag{5}$$

yields the Metropolis acceptance function for symmetric proposal distributions, and generalizes it to

$$a(x^*; x) = \min\Big(\frac{p(x^*)}{p(x)}\frac{q(x|x^*)}{q(x^*|x)}\text{ , } 1\Big) \tag{6}$$

for asymmetric proposal distributions. With a symmetric proposal distribution, taking $s(x^*, x) = 1$ gives the Barker acceptance function

$$a(x^*; x) \quad = \quad \frac{p(x^*)}{p(x^*) + p(x)} \tag{7}$$

where the acceptance probability is proportional to the probability of the proposed and the current state under the target distribution (Barker, 1965). While the Metropolis acceptance function has been shown to result in lower asymptotic variance (Peskun, 1973) and faster convergence to the stationary distribution (Billera & Diaconis, 2001) than the Barker acceptance function, it has also been argued that neither rule is clearly dominant across all situations (Neal, 1993).

Summary

Markov chain Monte Carlo provides a simple way to generate samples from probability distributions for which other sampling schemes are infeasible. The basic procedure is to start a Markov chain at some initial state, chosen arbitrarily, and then apply one of the methods for generating transitions outlined above, proposing a change to the state of the Markov chain and then deciding whether or not to accept this change based on the probabilities of the different states under the target distribution. After allowing enough iterations for the Markov chain to converge to its stationary distribution (known as the "burn-in"), the states of the Markov chain can be used to answer questions about the target distribution in the same way as a set of samples from that distribution.

An acceptance function from human behavior

The way that transitions between states occur in the Metropolis-Hastings algorithm already has the feel of a psychological task: two objects are presented, one being the current state and one the proposal, and a choice is made between them. We could thus imagine running an MCMC algorithm in which a computer generates a proposed variation on the current state and people choose between the current state and the proposal. In this section, we consider how to lead people to choose between two objects in a way that would correspond to a valid acceptance function.

From a task to an acceptance function

Assume that we want to gather information about a mental representation characterized by a non-negative function over objects, $f(x)$. In order for people's choices to act as an element in an MCMC algorithm that produces samples from a distribution proportional to $f(x)$, we need to design a task that leads them to choose between two objects x^* and x in a way that corresponds to one of the valid acceptance functions introduced in the previous section. In particular, it is sufficient to construct a task such that the probability with which people choose x^* is

$$a(x^*; x) = \frac{f(x^*)}{f(x^*) + f(x)} \tag{8}$$

this being the Barker acceptance function (Equation 7) for the distribution $p(x) \propto f(x)$.

Equation 8 has a long history as a model of human choice probabilities, where it is known as the Luce choice rule or the ratio rule (Luce, 1963). This rule has been shown to provide a good fit to human data when participants choose between two stimuli based on a particular property (Bradley, 1954; Clarke, 1957; Hopkins, 1954), though the rule does not seem to scale with additional alternatives (Morgan, 1974; Rouder, 2004; Wills, Reimers, Stewart, Suret, & McLaren, 2000). When $f(x)$ is a utility function or probability distribution, the behavior described in Equation 8 is known as probability matching and is often found empirically (Vulkan, 2000). The ratio rule has also been used to convert psychological response magnitudes into response probabilities in many models of cognition such as the Similarity Choice Model (Luce, 1963; Shepard, 1957), the Generalized Context Model (Nosofsky, 1986), the Probabilistic Prototype model (Ashby, 1992; Nosofsky, 1987), the Fuzzy Logical Model of Perception (Oden & Massaro, 1978), and the TRACE model (McClelland & Elman, 1986).

There are additional simple theoretical justification for using Equation 8 to model people's response probabilities. One justification is to use the logistic function to model a soft threshold on the log odds of two alternatives (cf. Anderson, 1990; Anderson & Milson, 1989). The log odds in favor of x^* over x under the distribution $p(x)$ will be $\Lambda = \log \frac{p(x^*)}{p(x)}$. Assuming that people receive greater reward for a correct response than an incorrect response, the optimal solution to the problem of choosing between the two objects is to deterministically select x^* whenever $\Lambda > 0$. However, we could imagine that people place their thresholds at slightly different locations or otherwise make choices non-deterministically, such that their choices can be modeled as a logistic function on Λ with the probability of choosing x^* being $1/(1 + \exp\{-\Lambda\})$. An elementary calculation shows that this yields the acceptance rule given in Equation 8.

Another theoretical justification is to assume that log probabilities are noisy and the decision is deterministic. The decision is made by sampling from each of the noisy log probabilities and choosing the higher sample. If the additive noise follows a Type I Extreme Value (e.g., Gumbel) distribution, then it is known that the resulting choice probabilities are equivalent to the ratio rule (McFadden, 1974; Yellott, 1977).

The correspondence between the Barker acceptance function and the ratio rule suggests that we should be able to use people as elements in MCMC algorithms in any context where we can lead them to choose in a way that is well-modeled by the ratio rule. We can then explore the representation characterized by $f(x)$ through a series of choices between two alternatives, where one of the alternatives that is presented is that selected on the previous trial and the other is a variation drawn from an arbitrary proposal distribution. The result will be an MCMC algorithm that produces samples from $p(x) \propto f(x)$, providing an efficient way to explore those parts of the space where $f(x)$ takes large values.

Allowing a wider range of behavior

Probability matching, as expressed in the ratio rule, provide a good description of human choice behavior in many situations, but motivated participants can produce behavior that is more deterministic than probability matching (Vulkan, 2000; Ashby & Gott, 1988). Several models of choice, particularly in the context of categorization, have been extended in order to account for this behavior (Ashby & Maddox, 1993) by using an exponentiated version of Equation 8 to map category probabilities onto response probabilities,

$$a(x^*; x) = \frac{f(x^*)^\gamma}{f(x^*)^\gamma + f(x)^\gamma} \tag{9}$$

where γ raises each term on the right side of Equation 8 to a constant. The parameter γ makes it possible to interpolate between probability matching and purely deterministic responding: when $\gamma = 1$ participants probability match, and when $\gamma = \infty$ they choose deterministically.

Equation 9 can be derived from an extension to the "soft threshold" argument given above. If responses follow a logistic function on the log posterior odds with gain γ, the probability of choosing x^* is $1/(1+\exp\{\gamma\Lambda\})$ where Λ is defined above. The case discussed above corresponds to $\gamma = 1$, and as γ increases, the threshold moves closer to a step function at $\Lambda = 0$. Carrying out the same calculation as for the case where $\gamma = 1$ in the general case yields the acceptance rule given in Equation 9.

The acceptance function in Equation 9 contains an unknown parameter, so we will not gain as much information about $f(x)$ as when we make the stronger assumption that participants probability match (i.e. fixing $\gamma = 1$). Substituting Equation 9 into Equation 1 and assuming a symmetric proposal distribution we can show that the exponentiated choice rule is an acceptance function for a Markov chain with stationary distribution

$$p(x) \propto f(x)^{\gamma} \tag{10}$$

Thus, using the weaker assumptions of Equation 9 as a model of human behavior, we can estimate $f(x)$ up to a constant exponent. The relative probabilities of different objects will remain ordered, but cannot be directly evaluated. Consequently, the distribution defined in Equation 10 has the same peaks as the function produced by directly normalizing $f(x)$, and the ordering of the variances for the dimensions is unchanged, but the exact values of the variances will depend on the parameter γ.

Categories

The results in the previous section provide a general method for exploring mental representations that can be expressed in terms of a non-negative function over a set of objects. In this section, we consider how this approach can be applied to the case of categories, providing a concrete example of a context in which this method can be used and an illustration of how tasks can be constructed to meet the criteria for defining an MCMC algorithm. We begin with a brief review of the methods that have been used to study mental representations of category structure. In the following discussion it is important to keep in mind that performance in any task used to study categories involves both the mental representations with which the category members, or category abstractions, are held in memory, and the processes that operate on those representations to produce the observed data. In this article our aim is to map the category structure that results from both the representations and processes acting together. Much research in the field has of course been aimed at specifying the representations, the processes, or both, as described briefly in the review that follows.

Methods for studying categories

Cognitive psychologists have extensively studied the ways in which people group objects into categories, investigating issues such as whether people represent categories in terms of exemplars or prototypes (J. D. Smith & Minda, 1998; Nosofsky & Zaki, 2002; Minda & Smith, 2002; Storms et al., 2000; Nosofsky, 1988; Nosofsky & Johansen, 2000; Zaki et al., 2003; E. E. Smith et al., 1998; Heit & Barsalou, 1996). Many different types of methodologies have been developed in an effort to distinguish the effects of prototypes and exemplars, and to test other theories of categorization. Two of the most commonly used methods are multidimensional scaling (MDS; Shepard, 1962; Torgerson, 1958) and additive clustering (Shepard & Arabie, 1979). These methods each take a set of similarity judgments and find the representation that best respects these judgments. MDS uses similarity judgments to constrain the placement of objects within a similarity space – objects judged most similar are placed close together and those judged dissimilar are place far apart. Additive clustering works in a similar manner, except that the resulting representation is

based on discrete features that are either shared or not shared between objects. These two methodologies are used in order to establish the psychological space for the stimuli in an experiment, assisting in the evaluation of categorization models. MDS has been applied to many natural categories, including Morse code (Rothkopf, 1957) and colors (Nosofsky, 1987). Traditional MDS does not scale well to a large number of objects because all pairwise judgments must be collected, but alternative formulations have been developed to deal with large numbers of stimuli (Hadsell, Chopra, & LeCun, 2006; de Silva & Tenenbaum, 2003; Goldstone, 1994).

Discriminative tests are a second avenue for comparing theories about categorization. In these tests, participants are shown a single object and asked to pick the category to which it belongs. The choices that are made reflect the structures of all the categories under consideration and can be analyzed to determine the decision boundaries between categories. This is perhaps the most commonly used method for distinguishing between many different types of models (e.g., Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Ashby & Gott, 1988), and it has been combined with MDS to provide a more diagnostic test (e.g, Nosofsky, 1986, 1987). In addition, discriminative tests can be applied to naturalistic stimuli with large numbers of parameters. In psychophysics, the response classification technique is a discriminative test that has been used to infer the decision boundary that guides perceptual decisions between two stimuli (Ahumada & Lovell, 1971). This methodology has been applied to visualize the decision boundary between images with as many as 10,000 pixels (Gold, Murray, Bennett, & Sekuler, 2000).

A third method for exploring category structure is to evaluate the strength of category membership for a set of the objects. In categorization studies, this goal is generally accomplished by asking participants for typicality ratings of stimuli (e.g., Storms et al., 2000; Bourne, 1982). This procedure was originally used to justify the existence of a category prototype (Rosch & Mervis, 1975; Rosch, Simpson, & Miller, 1976), but typicality ratings have also been used to support exemplar models (Nosofsky, 1988). Typicality ratings are difficult to collect for a large set of stimuli, because a judgment must be made for every object. In situations in which there are a large number of objects and few belong to a category, this method will be very inefficient. The key claim of this paper is the idea that the MCMC method can provide the same insight into categorization as given by typicality judgments, but in a form that scales well to the large numbers of stimuli needed to explore natural categories.

Representing categories

Computational models of categorization quantify different hypotheses about how categories are represented. These models were originally developed as accounts of the cognitive processes involved in categorization, and specify the structure of categories in terms of a similarity function over objects. Subsequent work identified these models as being equivalent to schemes for estimating a probability distribution (a problem known as *density estimation*), and showed that the underlying representations could be expressed as probability distributions over objects (Ashby & Alfonso-Reese, 1995). We will outline the basic ideas behind two classic models – exemplar and prototype models – from both of these perspectives. While we will use the probabilitistic perspective throughout the rest of the paper, the equivalence between these two perspectives mean that our analyses carry over

equally well to the representation of categories as a similarity function over objects.

Exemplar and prototype models share the basic assumption that people assign stimuli to categories based on similarity. Given a set of $N-1$ stimuli $x_1, \ldots, x_{N-1}$ with category labels $c_1, \ldots, c_{N-1}$, these models express the probability that stimulus N is assigned to category c as

$$p(c_N = c|x_n) = \frac{\eta_{N,c}\beta_c}{\sum_{c'} \eta_{N,c'}\beta_{c'}} \tag{11}$$

where $\eta_{N,c}$ is the similarity of the stimulus x_N to category c and β_c is the response bias for category c. The key difference between the models is in how $\eta_{N,c}$, the similarity of a stimulus to a category, is computed.

In an exemplar model (e.g., Medin & Schaffer, 1978; Nosofsky, 1986), a category is represented by all of the stored instances of that category. The similarity of stimulus N to category c is calculated by summing the similarity of the stimulus to all stored instances of the category. That is,

$$\eta_{N,c} = \sum_{i|c_i=c} \eta_{N,i} \tag{12}$$

where $\eta_{N,i}$ is a symmetric measure of the similarity between the two stimuli x_N and x_i. The similarity measure is typically defined as a decaying exponential function of the distance between the two stimuli, following Shepard (1987). In a prototype model (e.g., Reed, 1972), a category c is represented by a single prototypical instance. In this formulation, the similarity of a stimulus N to category c is defined to be

$$\eta_{N,c} = \eta_{N,p_c} \tag{13}$$

where p_c is the prototypical instance of the category and η_{N,p_c} is a measure of the similarity between stimulus N and the prototype p_c, as used in the exemplar model. One common way of defining the prototype is as the centroid of all instances of the category in some psychological space, i.e.,

$$p_c = \frac{1}{N_c} \sum_{i|c_i=c} x_i \tag{14}$$

where N_j is the number of instances of the category (i.e. the number of stimuli for which $c_i = c$). There are obviously a variety of possibilities within these extremes, with similarity being computed to a subset of the exemplars within a category, and these have been explored in more recent models of categorization (e.g., Love, Medin, & Gureckis, 2004; Vanpaemel, Storms, & Ons, 2005).

Taking a probabilistic perspective on the problem of categorization, a learner should believe stimulus N belongs to category c with probability

$$p(c_N = c|x_N) = \frac{p(x_N|c_N = c)p(c_N = c)}{\sum_{c'} p(x_N|c_N = c')p(c_N = c')} \tag{15}$$

with the posterior probability of category c being proportional to the product of the probability of an object with features x_N being produced from that category and the prior probability of choosing that category. The distribution $p(x|c)$, which we use as shorthand for $p(x_N|c_N = c)$, reflects the learner's knowledge of the structure of the category, taking

into account the features and labels of the previous $N-1$ objects. Ashby and Alfonso-Reese (1995) observed a connection between this Bayesian solution to the problem of categorization and the way that choice probabilities are computed in exemplar and prototype models (i.e. Equation 11). Specifically, $\eta_{N,c}$ can be identified with $p(x_N|c_N = c)$, while β_c corresponds to the prior probability of category c, $p(c_N = c)$. The difference between exemplar and prototype models thus comes down to different ways of estimating $p(x|c)$.

The definition of $\eta_{N,c}$ used in an exemplar model (Equation 12) corresponds to estimating $p(x|c)$ as the sum of a set of functions (known as "kernels") centered on the x_i already labeled as belonging to category c, with

$$p(x_N|c_N = c) \propto \sum_{i|c_i=c} k(x_N, x_i) \tag{16}$$

where $k(x, x_i)$ is a probability distribution centered on x_i.[1] This is a method that is widely used for approximating distributions in statistics, being a simple form of nonparametric density estimation (meaning that it can be used to identify distributions without assuming that they come from an underlying parametric family) called *kernel density estimation* (e.g., Silverman, 1986). The definition of $\eta_{N,c}$ used in a prototype model (Equation 13) corresponds to estimating $p(x|c)$ by assuming that the distribution associated with each category comes from an underlying parametric family, and then finding the parameters that best characterize the instances labeled as belonging to that category. The prototype corresponds to these parameters, with the centroid being an appropriate estimate for distributions whose parameters characterize their mean. Again, this is a common method for estimating a probability distribution, known as parametric density estimation, in which the distribution is assumed to be of a known form but with unknown parameters (e.g., Rice, 1995). As with the similarity-based perspective, there are also probabilistic models that interpolate between exemplars and prototypes (e.g., Rosseel, 2002).

Tasks resulting in valid MCMC acceptance rules

Whether they are represented as similarity functions (being non-negative functions over objects), or probability distributions, it should be possible to use our MCMC method to investigate the stucture of categories. The critical step is finding a question that we can ask people that leads them to respond in the way identified by Equation 8 when presented with two alternatives. That is, we need a question that leads people to choose alternatives with probability proportional to their similarity to the category or probability under that category. Fortunately, it is possible to argue that several kinds of questions should have this property. We consider three candidates.

Which stimulus belongs to the category?. The simplest question we can ask people is to identify which of the two stimuli belongs to the category, on the assumption that this is true for only one of the stimuli. We give a Bayesian analysis of this question in Appendix A, showing that stimuli should be selected in proportion to their probability under the category distribution $p(x|c)$ under some minimal assumptions. It is also possible to argue for this claim via an analogy to the standard categorization task, in which Equation 8 is

[1] The constant of proportionality is determined by $\int k(x, x_i)dx$, being $\frac{1}{N_c}$ if $\int k(x, x_i)\ dx = 1$ for all i, and is absorbed into β_c to produce direct equivalence to Equation 12.

widely used to describe choice probabilities. In the standard categorization task, people are shown a single stimulus and asked to select one of two categories. In our task, people are shown two stimuli and asked to identify which belongs to a single category. Just as people make choices in proportion to the similarity between the stimulus and the category in categorization, we expect them to do likewise here.

Which stimulus is a better example of the category?. The question of which stimulus belongs to the catgory is straightforward for participants to interpret when it is clear that one stimulus belongs to the category and the other does not. However, if both stimuli have very high or very low values under the probability distribution, then the question can be confusing. An alternative is to ask people which stimulus is a better example of the category – asking for a typicality rating rather than a categorization judgment. This question is about the representativeness of an example for a category. Tenenbaum and Griffiths (2001) provide evidence for a representativeness measure that is the ratio of the posterior probabilities of the alternatives. In Appendix A, we show that under the same assumptions as those for the previous question, answers to this question should also follow the Barker acceptance function. As above, this question can also be justified by appealing to similarity-based models: Nosofsky (1988) found that a ratio rule combined with an exemplar model predicted response probabilities when participants were asked to choose the better example of a category from a pair of stimuli.

Which stimulus is more likely to have come from the category?. An alternative means of relaxing the assumption that only one of the two stimuli came from the category is to ask people for a judgment of relative probability rather than a forced choice. This question directly measures $p(x|c)$ or the corresponding similarity function. Assuming that participants apply the ratio rule to these quantities, their responses should yield the Barker acceptance function for the appropriate stationary distribution.

Summary

Based on the results in this section, we can define a simple method for drawing samples from the probability distribution (or normalized similarity function) associated with a category using MCMC. Start with a parameterized set of objects and start with an arbitrary member of that set. On each trial, a proposed object is drawn from a symmetric distribution around the original object. A person chooses between the current object and the proposed new object, responding to a question about which object is more consistent with the category. Assuming that people's choice behavior follows the ratio rule given in Equation 8, the stationary distribution of the Markov chain is the probability distribution $p(x|c)$ associated with the category, and samples from the chain provide information about the mental representation of that category. As outlined above, this procedure can also provide information about the relative probabilities of different objects when people's behavior is more deterministic than probability matching. To explore the value of this method for investigating category structures we conducted a series of four experiments. The first two experiments examine whether we can recover artificial category structures produced through training, while the other two experiments apply it to estimation of the distributions associated with natural categories.

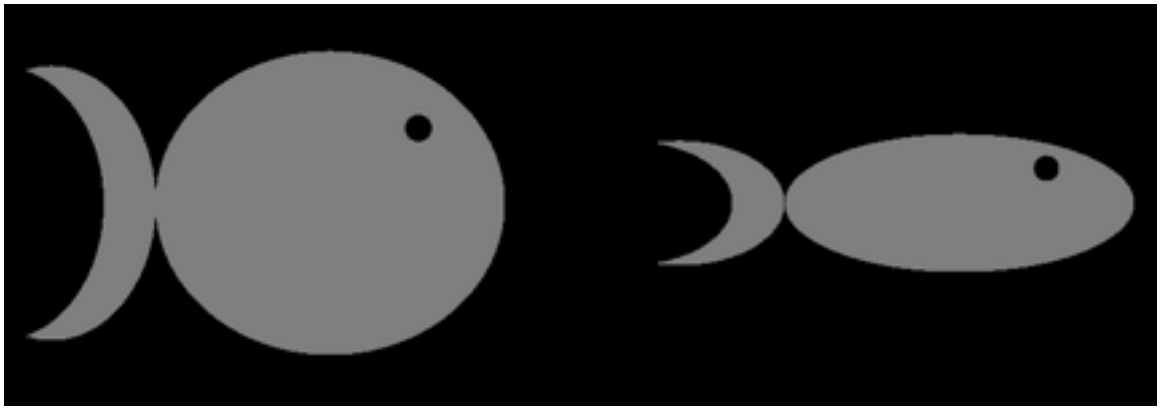

Figure 1. Examples of the largest and smallest fish stimuli presented to participants during training in Experiments 1 and 2. The relative size of the fish stimuli are shown here; true display sizes are given in the text.

Experiment 1

To test whether our assumptions about human decision making were accurate enough to use MCMC to sample from people's mental representations, we trained people on a category structure and attempted to recover the resulting distribution. A simple one-dimensional categorization task was used, with the height of schematic fish (see Figure 1) being the dimension along which category structures were defined. Participants were trained on two categories of fish height – a uniform distribution and a Gaussian distribution – being told that they were learning to judge whether a fish came from the ocean (the uniform distribution) or a fish farm (the Gaussian distribution). Once participants were trained, we collected MCMC samples for the Gaussian distributions by asking participants to judge which of two fish came from the fish farm using one of the three questions introduced above. The three questions conditions were used to find empirical support for the theoretical conclusion that all three questions would allow samples to be drawn from a probability distribution. Assuming participants accurately represent the training structure and use a ratio decision rule, the samples drawn will reflect the structure on which participants were trained for each question condition.

Method

Participants. Forty participants were recruited from the Brown University community. The data from four participants were discarded due to computer error in the presentation. Two additional participants were discarded for reaching the upper or lower bounds of the parameter range. These participants were replaced to give a total of 12 participants using each of the three MCMC questions. Each participant was paid $4 for a 35 minute session.

Stimuli. The experiment was presented on a Apple iMac G5 controlled by a script running in Matlab using PsychToolbox extensions (Brainard, 1997; Pelli, 1997). Participants were seated approximately 44 cm away from the display. The stimuli were a modified

version of the fish stimuli used in Huttenlocher et al. (2000). The fish were constructed from three ovals, two gray and one black, and a gray circle on a black background. Fish were all 9.1 cm long with heights drawn from the Gaussian and uniform distributions in training. Because the display did not allow a continuous range of fish heights to be displayed, the heights varied in steps of 0.13 cm. Examples of the smallest and largest fish are shown in Figure 1. During the the MCMC trials, fish were allowed to range in height from 0.03 cm to 8.35 cm.

Procedure.

Each participant was trained to discriminate between two categories of fish: ocean fish and fish farm fish. Participants were instructed, "Fish from the ocean have to fend for themselves and as a result they have an equal probability of being any size. In contrast, fish from the fish farm are all fed the same amount of food, so their sizes are similar and only determined by genetics." These instructions were meant to suggest that the ocean fish were drawn from a uniform distribution and the fish farm fish were drawn from a Gaussian distribution, and that the variance of the ocean fish was greater than the variance of the fish farm fish.

Participants saw two types of trials. In a training trial, either the uniform or Gaussian distribution was selected with equal probability, and a single sample (with replacement) was drawn from the selected distribution. The uniform distribution had a lower bound of 2.63 cm and an upper bound of 5.75 cm, while the Gaussian distribution had a mean of 4.19 cm and a standard deviation of 0.39 cm. The sampled fish was shown to the participant, who chose which distribution produced the fish. Feedback was then provided on the accuracy of this choice. In an MCMC trial, two fish were presented on the screen. Participants chose which of the two fish came from the Gaussian distribution. Neither fish had been sampled from the Gaussian distribution. Instead, one fish was the state of a Markov chain and the other fish was the proposal. The state and proposal were unlabeled and they were randomly assigned to either the left or right side of the screen. Three MCMC chains were interleaved during the MCMC trials. The start states of the chains were chosen to be 2.63 cm, 4.20 cm, and 5.75 cm. Relative to the training distributions, the start states were over-dispersed, facilitating assessment of convergence. The proposal was chosen from a symmetric discretized pseudo-Gaussian distribution with a mean equal to the current state and standard deviation equal to the training Gaussian standard deviation. The probability of proposing the current state was set to zero.

The question asked on the MCMC trials was varied between-participants. During the test phase, participants were asked, "Which fish came from the fish farm?", "Which fish is a better example of a fish from the fish farm?", or "Which fish was more likely to have come from the fish farm?". Twelve participants were assigned to each question condition. All participants were trained on the same Gaussian distribution, with mean $\mu = 4.19$ cm and standard deviation $\sigma = 3.9$ mm.

The experiment was broken up into blocks of training and MCMC trials. The experiment began with 120 training trials, followed by alternating blocks of 60 MCMC trials and 60 training trials. Training and MCMC trials were interleaved to keep participants from forgetting the training distributions. After a total of 240 MCMC trials, there was one final block of 60 test trials. These were identical to the training trials, except participants

were not given feedback. The final block provided an opportunity to test the discriminative predictions of the MCMC method.

Results and Discussion

Participants were excluded if the state of any chain reached the edge of the parameter range or if their chains did not converge to the stationary distribution. We used a heuristic for determining convergence to the stationary distribution: every chain had to cross another chain.[2] For the participants that were retained, the first 20 trials were removed from each chain. Figure 2 provides justification for removing the first 20 trials, as each example participant (one from each question condition) required about this many trials for the chains to converge. The remaining 60 trials from each chain were pooled and used in further analyses.

The acceptance rates of the proposals should be reasonably low in MCMC applications. A high acceptance rate generally results from proposal distributions that are narrow compared to the category distribution and thus do not explore the space well. Very low acceptance rates occur if the proposals are too extreme and the low acceptance rate also causes the algorithm to explore the space more slowly. A rate of 20% to 40% is often considered a good value (Roberts, Gelman, & Gilks, 1997). In this experiment, participants accepted between 33% and 35% of proposals in the three conditions. A more extensive exploration of the effects of acceptance rates is presented in the General Discussion.

The chains for one example participant per condition are shown in Figure 2. The right side of the figure compares the training distribution to two estimates of the stationary distribution formed from the samples: a nonparametric kernel density estimate, and a parametric estimate made by assuming that the stationary distribution was Gaussian. The nonparametric kernel density estimator used a Gaussian kernel with width optimal for producing Gaussian distributions (Bowman & Azzalini, 1997), although we should note that this kernel does not necessarily produce Gaussian distributions. The parametric density estimator assumed that the samples come from a Gaussian distribution and estimates the parameters of the Gaussian distribution that generated the samples. Both density estimation methods show that the samples do a good job of reproducing the Gaussian training distributions. A more quantitative measure of how well the participants reproduced the training distribution is to evaluate the means and standard deviations of the samples relative to the training mean and standard deviation. The averages over participants are shown in Figure 3. Over all participants, the mean of the samples is not significantly different from the training mean and the standard deviation of the samples is not significantly different from the standard deviation of the training distribution for any of the questions. The t-tests for the mean and standard deviation for each question are shown in Table 1.

The final testing block provided an opportunity to test the predictions of samples collected by the MCMC method for a separate set of discriminative judgments. The fish farm distribution was assumed to be Gaussian with the parameters determined by the mean and standard deviation of the MCMC samples. The uniform distribution on which participants were trained was assumed for the ocean fish category. The category with the

[2]Many heuristics have been proposed for assessing convergence (Brooks & Roberts, 1998; Mengersen, Robert, & Guihenneuc-Jouyaux, 1999; Cowles & Carlin, 1996). The heuristic we used is simple to apply in a one-dimensional state space.

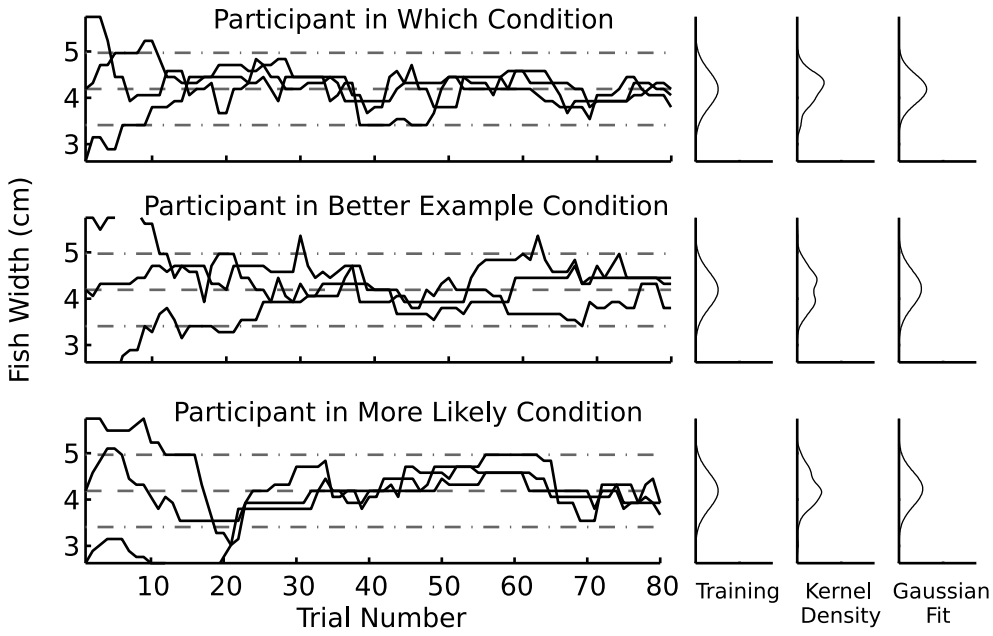

Figure 2. Markov chains produced by participants in Experiment 1. The three rows are participants from the three question conditions. The panels in the first column show the behavior of the three Markov chains per participant. The black lines represent the states of the Markov chains, the dashed line is the mean of the Gaussian training distribution, and the dot-dashed lines are two standard deviations from the mean. The second column shows the densities of the training distributions. These training densities can be compared to the MCMC samples, which are described by their kernel density estimates and Gaussian fits in the last two columns.

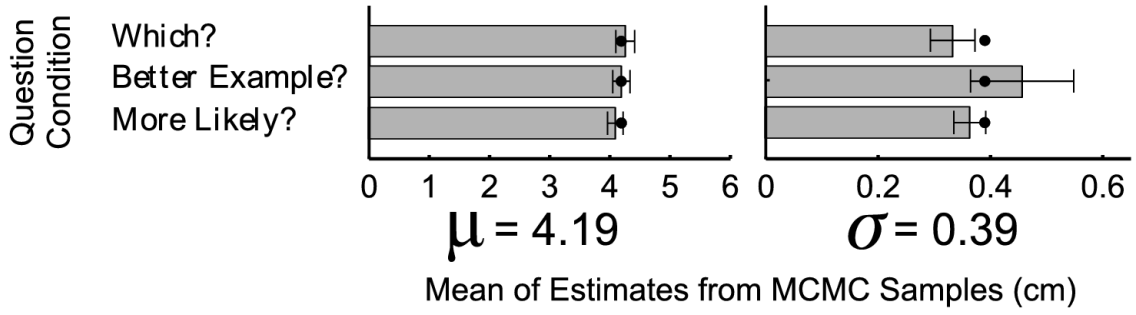

Figure 3. Results of Experiment 1. The bar plots show the mean of μ and σ across the MCMC samples produced by participants in all three question conditions. Error bars are one standard error. The black dot indicates the actual value of μ and σ for each condition, which corresponds closely with the MCMC samples.

Table 1: Statistical Significance Tests for Question Conditions in Experiment 1

Question	Statistic	df	t	p
Which?	Mean	11	0.41	0.69
	Std Dev	11	-1.46	0.17
Better Example?	Mean	11	0.00	1.00
	Std Dev	11	0.72	0.48
More Likely?	Mean	9	-0.95	0.37
	Std Dev	9	-0.22	0.83

higher probability for each fish size was used as the prediction, consistent with assuming that participants make optimal decisions. The discriminative responses matched the prediction on 69% of trials, which was significantly greater than the chance value of 50%, $z = 17.7, p < 0.001$.

Experiment 2

The results of Experiment 1 indicate that the MCMC method can accurately produce samples from a trained structure for a variety of questions. In the second experiment, the design was extended to test whether participants trained on different structures would show clear differences in the samples that they produced. Specifically, participants were trained on four distributions differing in their mean and standard deviation to observe the effect on the MCMC samples.

Method

Participants. Fifty participants were recruited from the Brown University community. Data from one participant was discarded for not finishing the experiment, data from another was discarded because the chains reached a boundary, and the data of eight others were discarded because their chains did not cross. After discarding participants, there were ten participants in the low mean and low standard deviation condition, ten participants in the low mean and high standard deviation condition, nine participants in the high mean and low standard deviation condition, and eleven participants in the high mean and high standard deviation condition.

Stimuli. Stimuli were the same as Experiment 1.

Procedure. The procedure for this experiment is identical to that of Experiment 1, except for three changes. The first change is that participants were all asked the same question, "Which fish came from the fish farm?" The second difference is that the mean and the standard deviation of the Gaussian were varied in four between-participant conditions. Two levels of the mean, $\mu = 3.66$ cm and $\mu = 4.72$ cm, and two levels of the standard deviation, $\sigma = 3.1$ mm and $\sigma = 1.3$ mm, were crossed to produce the four between-participant conditions. The uniform distribution was the same across training distributions and was bounded at 2.63 cm and 5.76 cm. The final change was that the standard deviation of the proposal distribution was set to 2.2 mm in this experiment.

Results and Discussion

As in the previous experiment, it took approximately 20 trials for the chains to converge, so only the remaining 60 trials per chain were analyzed. The acceptance rates in the four conditions ranged from 38% to 45%. The distributions on the right hand side of Figure 4 show the training distribution and the nonparametric and parametric estimates of the stationary distribution produced from the MCMC samples. The distributions estimated for the participants shown in this figure match well with the training distributions. The mean, μ, and standard deviation, σ, was computed from the MCMC samples produced by each participant. The average of these estimates for each condition is shown in Figure 5. As predicted, μ was higher for participants trained on Gaussians with higher means, and σ was higher for participants trained on Gaussians with higher standard deviations. These differences were statistically significant, with a one-tailed Student's t-test for independent samples giving $t(38) = 7.36, p < 0.001$ and $t(38) = 2.01, p < 0.05$ for μ and σ respectively. The figure also shows that the means of the MCMC samples corresponded well with the actual means of the training distributions. The standard deviations of the samples tended to be higher than the training distributions, which could be a consequence of either perceptual noise (increasing the effective variation in stimuli associated with a category) or choices being made in a way consistent with the exponentiated choice rule with $\gamma < 1$. Perceptual noise seems to be implicated because both standard deviations in this experiment (which were overestimated) were less than the standard deviation in the previous experiment (which was correctly estimated). The prediction of the MCMC samples for discriminative trials was determined by the same method used in Experiment 1. The discriminative responses matched the prediction on 75% of trials, significantly greater than the chance value of 50%, $z = 24.5, p < 0.001$.

Experiment 3

The results of Experiments 1 and 2 suggest that the MCMC method can accurately reproduce a trained structure. Experiment 3 used this method to gather samples from the natural (i.e., untrained) categories of giraffes, horses, cats, and dogs, representing these animals using nine-parameter stick figures (Olman & Kersten, 2004). These stimuli are less realistic than full photographs of animals, but have a representation that is tuned to the category, so that perturbations of the parameters will have a better chance of producing a similar animal than than pixel perturbations. The complexity of these stimuli is still fairly high, hopefully allowing the capture of a large number of interesting aspects of natural categories. The samples produced by the MCMC method were used to examine the mean animals in each category, what variables are important in defining membership of a category, and how the distributions of the different animals are related.

The use of natural categories also provided the opportunity to compare the results of the MCMC method with other approaches to studying category representations. These stick figure stimuli were developed by Olman and Kersten (2004) to illustrate that classification images, one of the discriminative approaches introduced earlier in the paper, could be applied to parameter spaces as well as to pixel spaces. Discriminative methods like classification images can be used to determine the decision bound between categories, but the results of discrimination judgments should not be used to estimate the parameters charac-

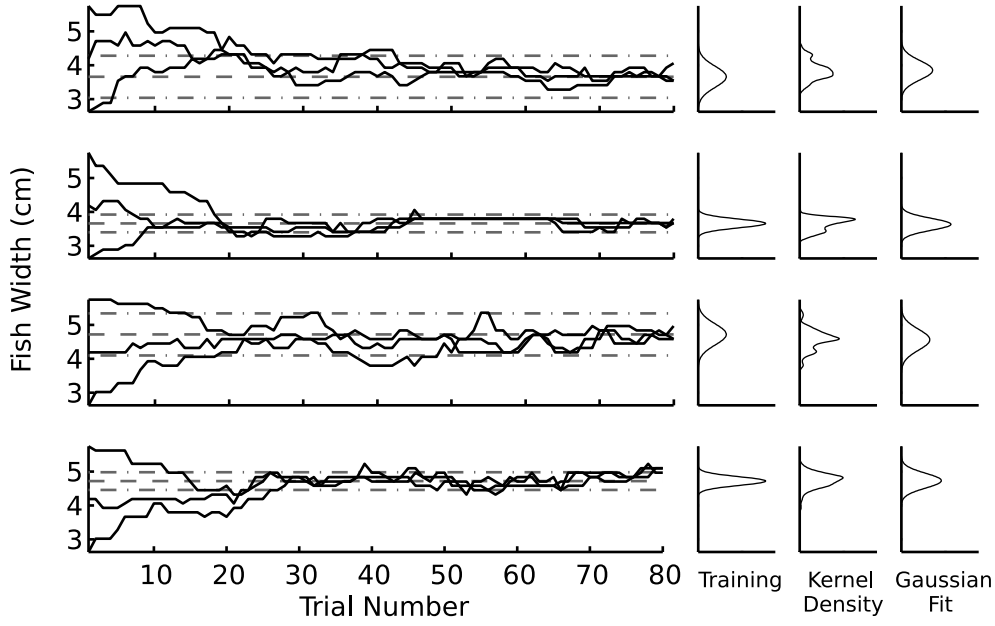

Figure 4. Markov chains produced by participants in Experiment 2. The four rows are participants from each of the four conditions. The panels in the first column show the behavior of the three Markov chains per participant. The black lines represent the states of the Markov chains, the dashed line is the mean of the Gaussian training distribution, and the dot-dashed lines are two standard deviations from the mean. The second column shows the densities of the training distributions. These training densities can be compared to the MCMC samples, which are described by their kernel density estimates and Gaussian fits in the last two columns.

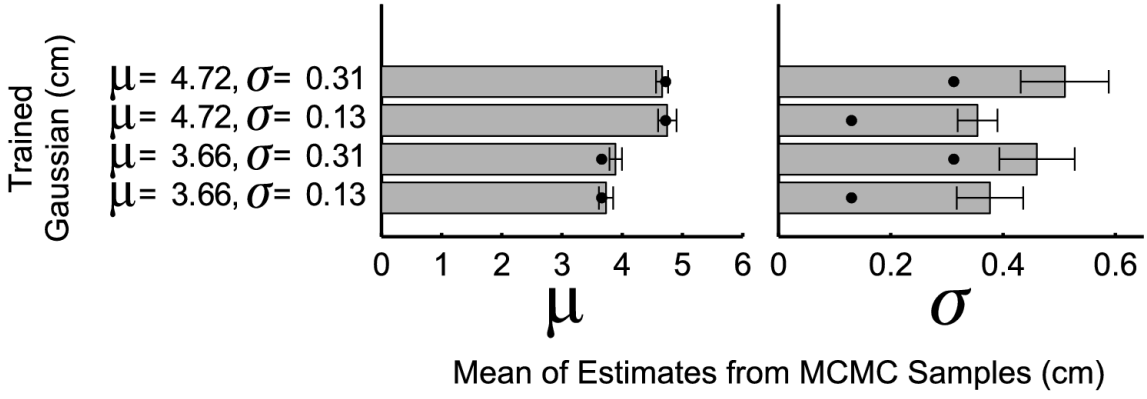

Figure 5. Results of Experiment 2. The bar plots show the mean of μ and σ across the MCMC samples produced by participants in all four training conditions. Error bars are one standard error. The black dot indicates the actual value of μ and σ for each condition, which corresponds with the MCMC samples.

terizing the mean of a category. The experiment empirically highlights why discriminative judgments should not be used for this purpose, and serves as a complement to the theoretical argument given in the discussion. In addition, we analyzed the results to determine whether it would have been feasible to use typicality judgments on a random subset of examples to identify the underlying category structure. Collecting typicality judgments is efficient when most stimuli belong to a category, but can be inefficient when a category distribution is non-zero in only a small portion of a large parameter space. The MCMC method excels in the latter situation, as it uses a participant's previous responses to choose new trials.

Method

Participants. Eight participants were recruited from the Brown University community. Each participant was tested in five sessions, each of which could be split over multiple days. A session lasted for a total of one to one-and-a-half hours for a total time of five to seven-and-a-half hours. Participants were compensated with $10 per hour of testing.

Stimuli. The experiment was presented on a Apple iMac G5 controlled by a script running in Matlab using PsychToolbox extensions (Brainard, 1997; Pelli, 1997). Participants were seated approximately 44 cm away from the display. The nine-parameter stick figure stimuli that we used were developed by Olman and Kersten (2004). The parameter names and ranges are shown in Figure 6. From the side view, the stick figures had a minimum possible width of 1.8 cm and height of 2.5 mm. The maximum possible width and height were 10.9 cm and 11.4 cm respectively. Each stick figure was plotted on the screen by projecting the three dimensional description of the lines onto a two dimensional plane from a particular viewpoint. To prevent viewpoint biases, the stimuli were rotated on the screen by incrementally changing the azimuth of the viewpoint, while keeping the elevation constant at zero. The starting azimuth of the pair of stick figures on a trial was drawn from a uniform distribution over all possible values and the period of rotation was fixed at six seconds. The perceived direction of rotation is ambiguous in this experiment.

Procedure. The first four sessions of the experiment were used to sample from a participant's category distribution for four animals: giraffes, horses, cats, and dogs. Each session was devoted to a single animal and the order in which the animals were tested was counterbalanced across participants. On each trial, a pair of stick figures was presented, with one stick figure being the current state of the Markov chain and the other the proposed state of the chain. Participants were asked on each trial to pick "Which animal is the _____", where the blank was filled by the animal tested during that session. Participants made their responses via keypress. The proposed states were drawn from a multivariate discretized pseudo-Gaussian with a mean equal to the current state and a diagonal covariance matrix. Informal pilot testing showed good results were obtained by setting the standard deviation of each variable to 7% of the range of the variable. Proposals that went beyond the range of the parameters were automatically rejected and were not shown to the participant.

To assess convergence, three independent Markov chains were run during each session. The starting states of the Markov chains were the same across sessions. These starting states were at the vertices of an equilateral triangle embedded in the nine-dimensional parameter space. The coordinates of each vertex were either 20% of the range away from the minimum

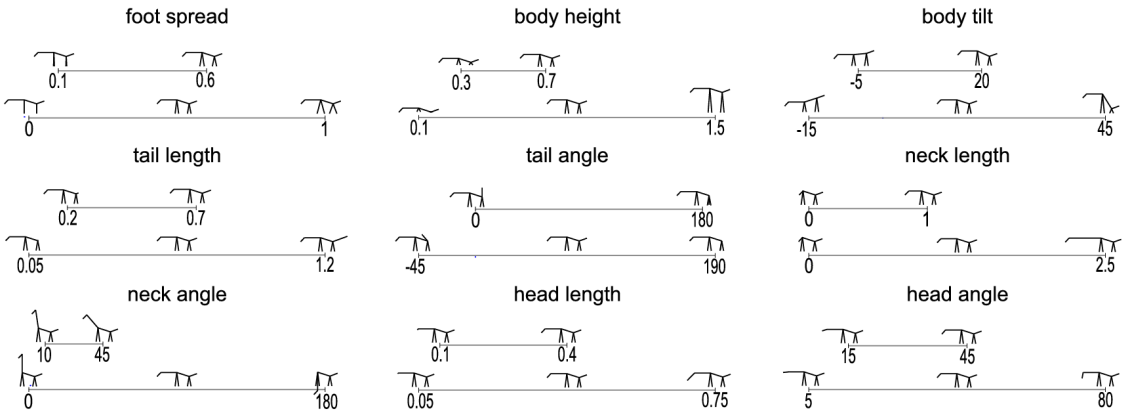

Figure 6. Range of stick figures shown in Experiment 3 by parameter. The top line for each parameter is the range in the discrimination session, while the bottom line is the range in the MCMC sessions. Lengths are relative to the fixed body length of one.

value or 20% of the range away from the maximum value for each parameter. Participants responded to 999 trials in each of the first four sessions, with 333 trials presented from each chain. The automatically rejected out-of-range trials did not count against the number of trials in a chain, so more samples were collected than were presented. The number of collected trials ranged from 1252 to 2179 trials with an average of 1553 trials.

The fifth and final session was a 1000 trial discrimination task, reproducing the procedure that Olman and Kersten (2004) used to estimate the structure of visual categories of animals. On each trial, participants saw a single stick figure and made a forced response of giraffe, horse, dog, or cat. Stimuli were drawn from a uniform distribution over the restricted parameter range shown in Figure 6. This range was used in Olman and Kersten (2004), and provided a higher proportion of animal-like stick figures at the expense of not testing portions of the animal distributions. Unlike in the MCMC sessions, the stimuli presented were independent across trials.

Results and Discussion

In order to determine the number of trials that were required for the Markov chains to converge to the stationary distribution, the samples were projected onto a plane. The plane used was the best possible plane for discriminating the different animal distributions for a particular participant, as calculated through linear discriminant analysis (Duda, Hart, & Stork, 2000). Trials were removed from the beginning of each chain until the distributions looked as if they had converged. Removal of 250 trials appeared to be sufficient for most chains, and this criterion was used for all chains and all participants. The acceptance rate of the proposals across subjects and conditions ranged between 14% and 31%. Averaged across individuals, it ranged from 21% to 26%, while averaged across categories, acceptance rates ranged from 19% to 27%. Figure 7 shows the chains of the giraffe condition of Participant 3 projected onto the plane that best discriminates the categories. The largest factor in the horizontal component was neck length, while the largest factor in the vertical component was head length. The left panel shows the starting points and all of the samples for each

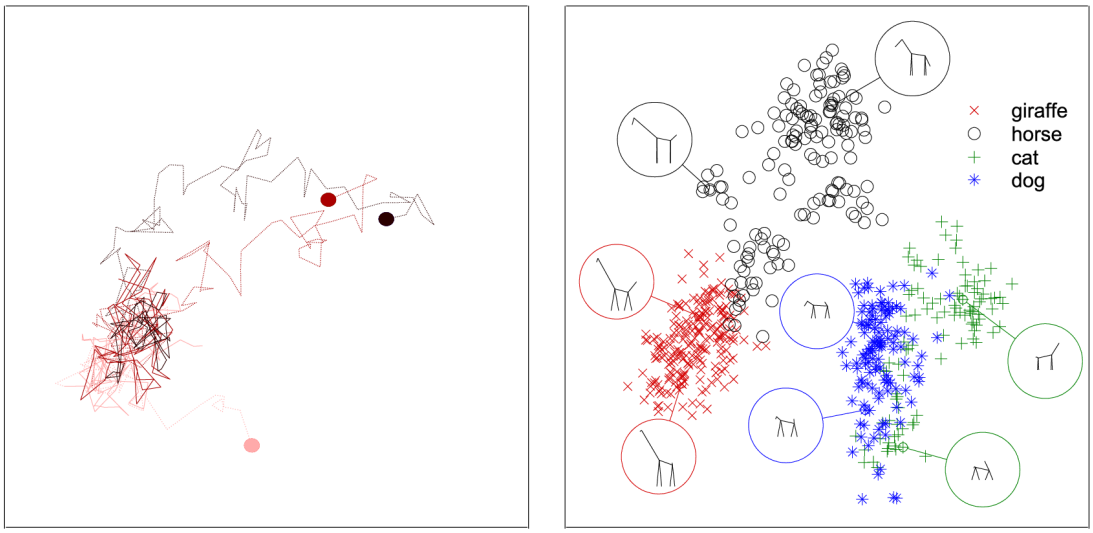

Figure 7. Markov chains and samples from Participant 3 in Experiment 3. The left panel displays the three chains from the giraffe condition. The discs represent starting points, the dotted lines are discarded samples, and the solid lines are retained samples. The right panel shows the samples from all of the animal conditions for Participant 3 projected onto the plane that best discriminates the classes. The quadrupeds in the bubbles are examples of MCMC samples.

chain. The right panel shows the remaining samples after the burn-in was removed. From this figure, we can see that the giraffe and horse distributions are relatively distinct, while the cat and dog distributions show more overlap.

Displaying a nine-dimensional joint distribution is impossible, so we must look at summaries of the distribution. A basic summary of the animal distributions that were sampled is the mean over samples. The means for all combinations of participants and quadrupeds are shown in Figure 8. As support for the MCMC method, these means tend to resemble the quadrupeds that participants were questioned about. While the means provide face validity for the MCMC method, the distributions associated with the different categories contain much more interesting information. Samples from each animal category are shown in Figure 7, which give a sense of variation within a category. Also, the marginal distributions shown in Figure 9 give a more informative picture of the joint distribution than the means. The marginal distributions were estimated by taking a kernel density estimate along each parameter for each of the quadrupeds on the data pooled over participants. The Gaussian kernel used had its width optimized for estimating Gaussian distributions (Bowman & Azzalini, 1997). In Figure 9 we can see that neck length and neck angle distinguish giraffes from the other quadrupeds. Horses generally have a longer head length than the other quadrupeds. Dogs have a somewhat distinct head length and head angle. Cats have the longest tails, shortest head length, and a narrowest foot spread.

The data summaries along one or two dimensions may exaggerate the overlap between categories. It is possible that the overlap differs between participants or the distributions

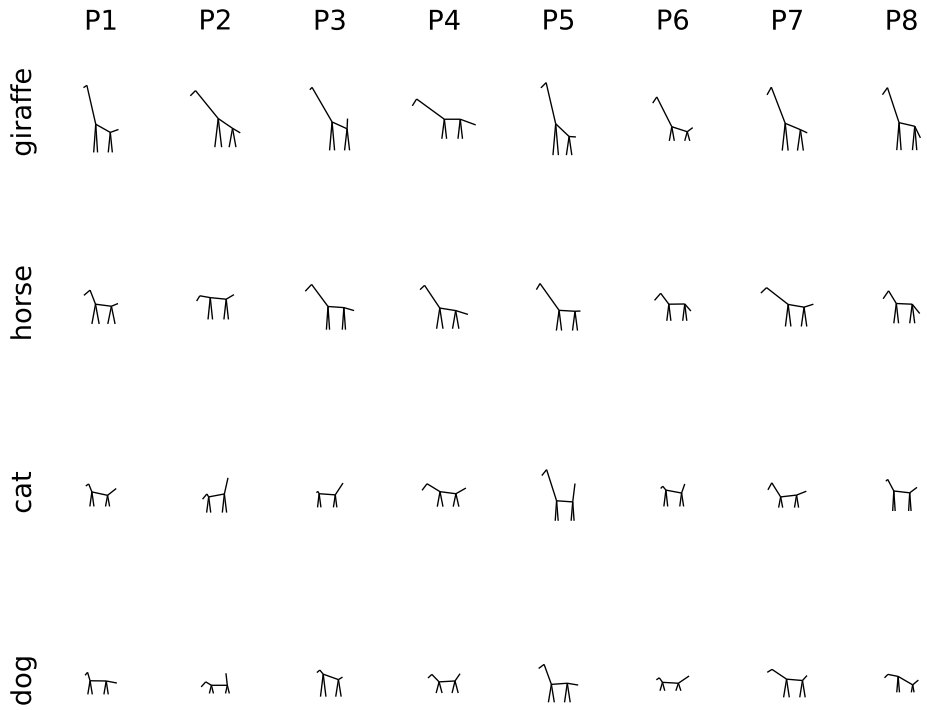

Figure 8. Means across samples for all participants and all quadrupeds in Experiment 3. The rows of the figure correspond to giraffes, horses, cats, and dogs respectively. The columns correspond to the eight participants.

are even more separated in the full parameter space. In order to examine how much the joint distributions overlap, a multivariate Gaussian distribution was fit to each set of animal samples. Using these distributions for each quadruped, We calculated how many samples from each quadruped fell within the 95% confidence region of each animal distribution. Table 2 shows the number of samples that fall within each distribution averaged over participants. The diagonal values are close to the anticipated value of 95%. Between quadrupeds, horses and giraffes share a fair amount of overlap and dogs and cats share a fair amount of overlap. Between these two groups, cats and dogs overlap a bit less with horses than they do with each other. Also, cats and dogs overlap very little with giraffes. There is an interesting asymmetry between the cat and dog distributions. The cat distribution contains 14% of the dog samples, but the dog distribution only contains 3.7% of the cat samples.

The discrimination session allows us to compare the outcome of the MCMC method with that of a more standard experimental task. The mean animals from all of the MCMC trials and from all of the discrimination trials (both pooled over participants) are shown in Figure 10. The discrimination judgments were made on stimuli drawn randomly from the parameter range used by Olman and Kersten (2004). The mean stick figures that

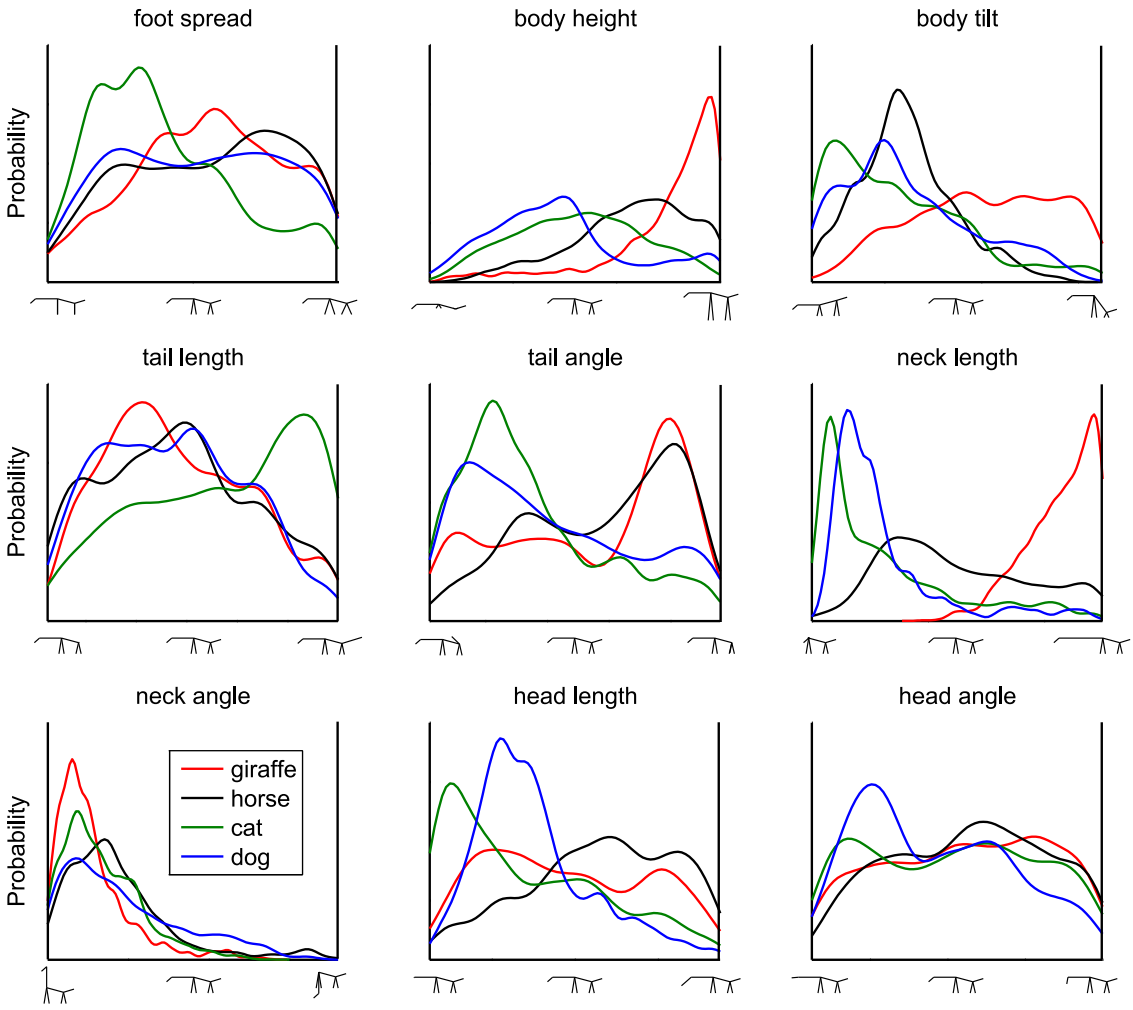

Figure 9. Marginal kernel density estimates for each of the nine stimulus parameters, aggregating over all participants in Experiment 3. Examples of a parameter's effect are shown on the horizontal axis. The middle stick figure is given the central value in the parameter space. The two flanking stick figures show the extremes of that particular parameter with all other parameters held constant.

Table 2: Proportion of Samples that Fall Within Quadruped 95% Confidence Regions (Averaged Over Participants) in Experiment 3

Distribution	Samples			
	Giraffe	Horse	Cat	Dog
Giraffe	0.954	0.048	0.006	0.000
Horse	0.097	0.960	0.046	0.058
Cat	0.000	0.044	0.936	0.143
Dog	0.021	0.044	0.037	0.954

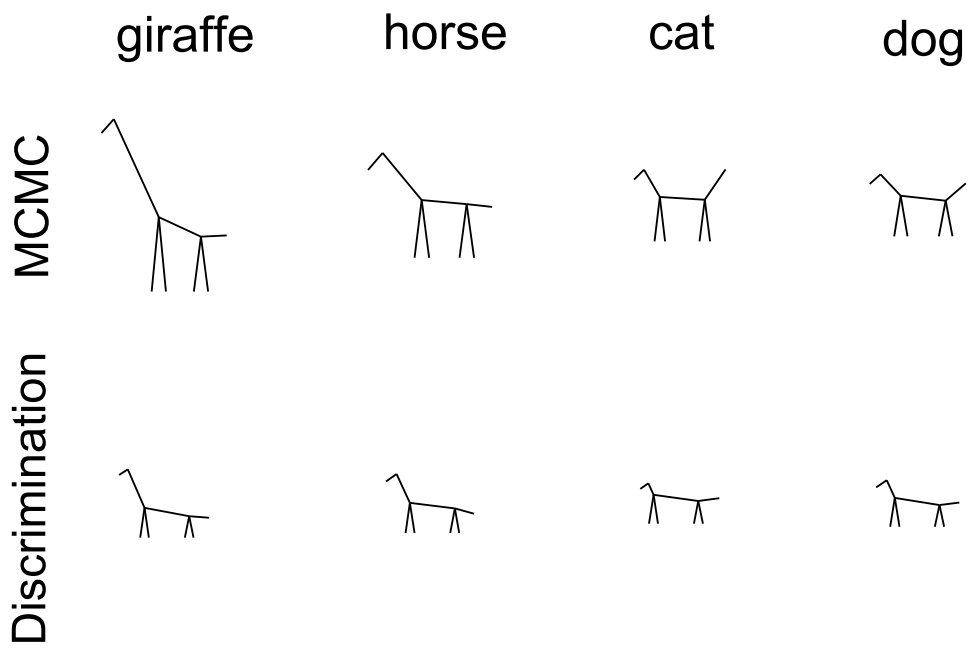

Figure 10. Mean quadrupeds across participants for the MCMC and discrimination sessions in Experiment 3. The rows of the figure correspond to the data collection method and the columns to giraffes, horses, cats, and dogs respectively.

result from these discrimination trials look reasonable, but MCMC pilot results showed that the marginal distributions were often pushed against the edges of the parameter ranges. The parameter ranges were greatly expanded for the MCMC trials as shown in Figure 6, resulting in a space that was approximately 3,800 times the size of the discrimination space. Over all participants and animals, only 0.4% of the MCMC samples fell within the restricted parameter range used in the discrimination session. This indicates that the true category distributions lie outside this restricted range, and that the mean animals from the discrimination session are actually poor examples of the categories – something that would have been hard to detect on the basis of the discrimination judgments alone. This result was borne out by asking people whether they thought the mean MCMC animals or the corresponding mean discrimination animals better represented the target category. A separate group of 80 naive participants preferred most mean MCMC animals over the corresponding mean discrimination animals. Given a forced choice between the MCMC and discrimination mean animals, the MCMC version was chosen by 99% of raters for giraffes, 73% of raters for horses, 25% of raters for cats, and 69% of raters for dogs.

With an added assumption the results from the MCMC method can be used to predict discriminative judgments. The MCMC samples estimate the category densities and it is straightforward to use this information to predict discrimination judgments if optimal decision bounds between categories are assumed. Multivariate Gaussian distributions were

Table 3: Proportion of Parameter Space Occupied by Individual Participant Samples in Experiment 3

Giraffe	Horse	Cat	Dog
0.00004	0.00006	0.00003	0.00002

estimated for each category and for each stimulus shown in the discrimination session. These four probabilities were normalized and then the prediction was the category with the maximum probability from four animal distributions, consistent with an optimal decision. Using this method, the predictions matched human responses on 32% of trials, significantly higher than the chance value of 25%, $z = 14.5, p < 0.001$. (Other assumptions, such as probability matching to make predictions or using Gaussian kernels to construct the distribution produced very similar results.) The discrimination results were confined to a small region of the parameter space in which few MCMC samples were drawn, which may have contributed to the weakness of the prediction.

In a large space, randomly selecting trials for typicality judgments will be inefficient when the category regions are small and spread out. This is exactly the situation encountered in Experiment 3. The projection of the samples from the nine-dimensional parameter space onto a plane, as in Figure 7, exaggerates the size of the relevant region for a category relative to the whole parameter space. The volume of the space occupied by the different categories can be measured by constructing convex hulls around all of the post-burn-in samples from each participant for each animal. Convex hulls connect the most extreme points in a cloud to form a surface that has no concavities and contains all of the points (Barnett, 1976). Dividing the volume of the space taken up by each animal (averaged over participants) by the total volume of the parameter space result in the very small ratios shown in Table 3: the categories take up only a tiny part of the nine-dimensional space of stick-figures. This analysis depends on the bounds of the parameters: selecting a much larger space than necessary would artificially decrease the volume ratios. We tested whether these ratios are artificially decreased by computing the size of the least enclosing hypercube for all of the samples pooled across participants and animal conditions. The ratio of this hypercube to the total parameter space was 0.99, negating the criticism that the parameter space was larger than necessary. The MCMC method thus efficiently honed in on the small portion of the space where each category was located, while typicality judgments on a random subset of examples would have been mostly uninformative.

Experiment 4

In this experiment, we outline how the MCMC method can be used to test models of categorization with natural categories. We do not provide a definitive result, but instead outline how difficulties in applying the MCMC method to testing models of natural categories might be overcome. Theoretically, we are able to test categorization models, such as prototype and exemplar models, because they make different predictions about the distribution over objects associated with categories. As outlined above, prototype models produce unimodal distributions, while exemplar models are more flexible, approximating a distribution with a mixture of kernels centered on each example. The MCMC method produces samples from this distribution, which can be used to distinguish these models.

Categorization models are usually tested with training experiments, but the efficiency of the MCMC method in exploring complex parameter spaces allows for the testing of categorization models using more realistic stimuli and natural categories. Using this method, we can examine whether the distributions associated with natural categories more closely resemble those predicted by prototype or exemplar models. More precisely, we can test whether these distributions have properties that rule out a prototype model, as an exemplar model can emulate a prototype model given an appropriate set of exemplars. For example, an exemplar model with two exemplars that are very near one another would be indistinguishable from a unimodal distribution. As a result, the test is strong in only one direction: a multimodal distribution is evidence against a simple prototypical category representation. In the other direction we would only be able to claim that the prototype model could fit the data, but not provide direct evidence that the exemplar model is unable to do so.

In Experiment 3, there was a suggestion of multiple modes in the marginal distribution of tail angle for several animal shape categories. However, beyond any arguments about the strength of the results, the design of the experiment did not have all of the controls necessary to provide strong evidence against the prototype model. Most importantly, the current experiment demonstrates how to test for category multimodality in a psychological space, instead of testing in an experimenter-defined parameter space. Categorization models are defined over a psychological space (Reed, 1972; Nosofsky, 1987), and it is possible that a transformation from one space to another would change the shape of the function. In order to assess multimodality, the samples must be projected to a psychological space and then the probability distributions can be assessed there.

Multidimensional scaling (MDS) is the standard method used to create a psychological space, and was used in this experiment to project the MCMC samples into a psychological space. Classical MDS is most commonly applied MDS method in psychological experiments, but it was impractical in this situation because a very large number of pairwise similarity judgments would have been needed to construct a scaling solution for the MCMC samples. Instead, an MDS variant known as Dimensionality Reduction by Learning an Invariant Mapping (DrLIM; Hadsell et al., 2006) was used. DrLIM has not been used in psychological experiments to our knowledge, but was chosen because its advantages over classical MDS. First, it utilizes neighborhood judgments instead of requiring all pairwise similarity judgments between examples. Neighborhood judgments can be made more quickly than the full set of pairwise similarity judgments – while the similarity of an item to every other item needs to be implicitly determined in order to find its neighbors, participants need only respond with the target item's neighbors. Second, DrLIM uses these neighborhood judgments to train an explicit function that maps the parameter space onto the similarity space. Classical MDS ignores the parameters of the items and does not produce an explicit mapping function. As a result, DrLIM can use the neighborhood judgments collected for a small set of stimuli in order to produce a solution for all of the stimuli in the entire parameter space.

The following experiment demonstrates how to test the simple prototype model using the categories of apples, grapes, and oranges. These categories are basic-level categories (Rosch, Mervis, et al., 1976; Liu, Golinkoff, & Sak, 2001), like the animal categories (Snodgrass & Vanderwart, 1980), the level at which the prototype model is defined (Rosch,

Mervis, et al., 1976). However, there is reason to expect a stronger result of multimodality with these categories: the stimuli were colored shapes and varieties of apples and grapes have different colors.

Method

Participants. Eight participants were recruited from the Indiana University community. Each participant was tested in two phases, with the second phase allowed to extend over multiple days. Participants completed the experiment in five to seven sessions with a total time ranging from five to seven-and-a-half hours and were compensated $10 per hour.

Stimuli. The experiment was presented in a darkened room on a Apple eMac controlled by a script running in Matlab using PsychToolbox extensions (Brainard, 1997; Pelli, 1997). The display was calibrated using a Spyder2 colorimeter to the sRGB standard with a D65 temperature point. Observers were seated approximately 44 cm away from the display. Stimuli were six-parameter shapes with three parameters determining the shape and three parameters determining the color. The three shape parameters were the radii of the circles, the horizontal distance between circle centers, and the vertical distance between circle centers as shown in Figure 11. The range was $[0, 0.5]$ for the the radii, $[0, 1]$ for the horizontal distance, and $[-0.866, 0.866]$ for the vertical distance. The maximum values for each of these parameters produced a shape with maximum size: 2.5 cm in the MDS session and 7.5 cm in the MCMC sessions. Negative values of the horizontal and vertical parameters resulted in the changes in the relative position of the circle centers (e.g., a negative vertical parameter resulted in an internal structure of two circles above a single circle). The three color parameters corresponded to the three parameters in the CIE Lab standard. The allowable range for L was $[0, 100]$, while the ranges for a and b were $[-100, 100]$. Each stimulus was topped by a green sprig (sRGB value of $[34, 139, 34]$) with a height and width 1/10th and 1/40th that of the maximum fruit size respectively. Each fruit stimulus was placed on a "dinner plate" constructed of three circles: two white and one black. The three circles were overlayed to produce a white plate with a black ring. The plate was 1.3512 times the maximum fruit size with an inner white region radius that was 90% of the total plate radius. These values were chosen so that largest possible stimulus would not touch the black ring on the plate. A schematic ruler was always present on the screen, indicating that participants should behave as if the plates were 10 inches in diameter.

Procedure. There were two phases: an MDS session and an MCMC session that interleaved three basic-level categories (apples, grapes, and oranges). The MDS session consisted of ninety trials of neighborhood judgments. On each trial one of the stimuli was used as the target and participants chose the neighbors from all of the remaining stimuli. Participants were instructed to think of the shapes as fruit and to make their neighbor judgments on that basis. These stimuli were picked so that they would include the most common fruit. Table 4 shows the factors and levels that were crossed to make the ninety stimuli. (For a shape parameter x, $x = sx'$.) A screen capture of this phase of the experiment is shown in Figure 12. The number of potential neighbors exceeded the number of stimuli that could be comfortably displayed on a single screen, so several pages of stimuli were used, with a maximum of 32 stimuli per page. The locations of the potential neighbors were randomized on each trial to reduce the possible bias caused by the ordering

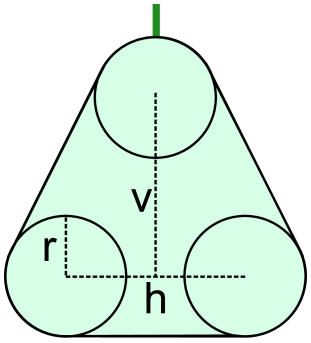

Figure 11. Shape parameters used in constructing fruit stimuli in Experiment 4. The r, h, and v correspond to the radii, horizontal, and vertical parameters. Black lines are for reference only and were not visible as the shape was a single monochromatic color. A green sprig was placed at the top of each stimulus to signify where the vine would attach.

Table 4: Stimulus Properties Crossed to Produce Multidimensional Scaling Stimulus Set in Experiment 4

Color (L, a, b)		Shape (r', h', v')		Size (s)	
Red	$(53.24, 80.09, 67.20)$	Circle	$(0.5, 0, 0)$	Small	(0.33)
Cardinal	$(42.72, 62.89, 28.48)$	Fat Circle	$(0.5, 0.3, 0)$	Large	(0.67)
Forest Green	$(50.59, -49.59, 45.02)$	Very Fat Circle	$(0.5, 1, 0)$		
Yellow	$(97.14, -21.56, 94.48)$	Stawberry-shaped	$(0.5, 1, 0.866)$		
Brown	$(40.44, 27.50, 50.14)$	Pear-shaped	$(0.5, 1, -0.866)$		
Eggplant	$(34.88, 64.05, -30.59)$				
Orange	$(74.93, 23.93, 78.95)$				
Peach	$(91.95, 1.80, 27.19)$				
Tan	$(74.98, 5.02, 24.43)$				

of the potential neighbors. Participants were allowed to navigate back and forth between the pages to choose the nearest neighbors to the target stimulus (by putting those stimuli on the neighbor plates). Once all of the pages had been viewed and five nearest neighbors had been selected, the END button appeared and participants were allowed to end the trial whenever they were satisfied with their choices.

A series of MCMC sessions followed the MDS session. Each MCMC trial displayed two stimuli on plates and participants were asked to respond via keypress to the stimulus that was a better example of a particular category. One stimulus was the current state of a Markov chain and the other was the proposed state. Whichever stimulus was selected became the new state of the chain. For each tested category, three chains were interleaved. Three chains for each of the three basic level categories (apples, grapes, and oranges) were all interleaved randomly with one another. There were 667 trials per chain yielding 2001 trials presented per basic level category. The number of samples was often far greater than the number of presented trials because all proposals outside of the parameter ranges were automatically rejected. The number of samples in a category condition for a participant ranged from 3,572 to 7,416 with an average of 4,650. The proposal distribution and chain start points were the same across all categories and participants. The proposal distribution

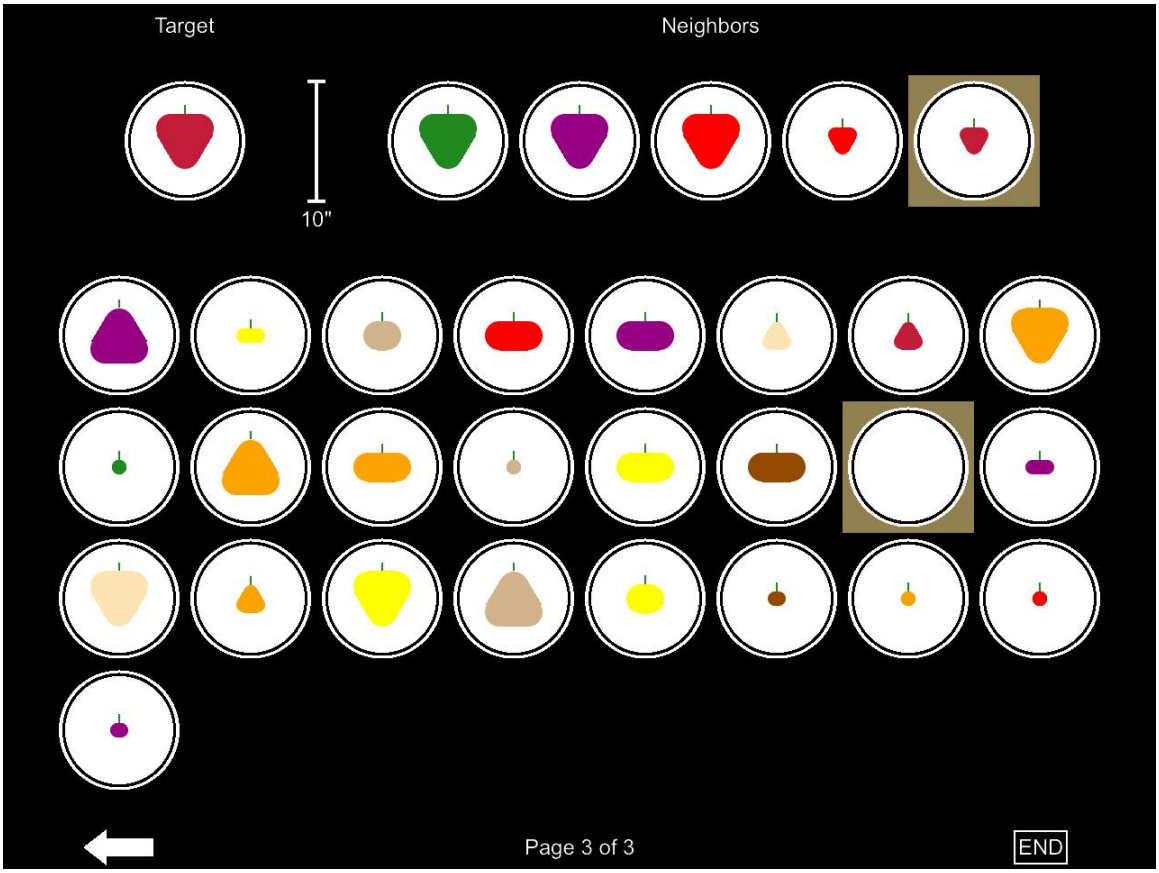

Figure 12. Sample screen from the session collecting nearest neighbor data in Experiment 4. Participants selected the stimuli that they considered the nearest neighbors of each stimulus.

was a mixture with a 0.8 weighting on a diagonal-covariance Gaussian with standard deviations equal to 7% of each parameter's range, a 0.1 weight on a uniform distribution across all shape parameters, and a 0.1 weight on a uniform distribution across all color parameters. (When proposing on a subset of parameters the other parameters remained fixed.) The uniform distributions were used in order to make large jumps between potentially isolated peaks in the subjective probability distributions. The large jumps were proposed separately for shape and color parameters because these dimensions are psychologically separable (Garner & Felfoldy, 1970).

Multidimensional Scaling Analysis

This section provides details on the method used to analyze the neighborhood data and the scaling results. DrLIM is a MDS method that uses neighborhood judgments to find the best function to map stimuli from a parameter space to a similarity space. The best function is selected by finding the parameters for a class of functions that best minimize a particular loss function. This loss function increases as the distance between items that are neighbors increases and as the distance between items that are not neighbors decreases.

Non-neighbors only impact the loss function within a certain margin. The exact loss function is

$$L(W, Y, \vec{X}_1, \vec{X}_2) = (1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}(\max\{0, m - D_W\})^2 \quad (17)$$

where W is the set of parameters, Y is the response (neighbors are 1 and non-neighbors are -1), $\vec{X}_1$ and $\vec{X}_2$ are the two members of the pair, m is the margin outside of which the non-neighbors do not contribute to the loss function, and D_W is the Euclidean distance between the function's outputs given $\vec{X}_1$ and $\vec{X}_2$ as inputs and parameters W.

In order to allow as large a class of nonlinear functions as possible, a feedforward artificial neural network (ANN) with a hidden layer was used. The number of inputs was equal to the number of parameters of the stimuli and the number of outputs determined the dimensionality of the similarity space. (Values on these outputs give the coordinates of the items in the similarity space.) To produce a distance measure between items, a siamese ANN was used, meaning that a single set of parameters governed the behavior of the pair of networks required to produce the distance between a pair of items. Pairs of stimuli were considered neighbors when if either of the stimuli was the target, the other was chosen to be one of its five nearest neighbors.

The network structure is shown in Figure 13. Thirty hidden nodes were used and network weights were updated via stochastic gradient descent. Weights were initially drawn from a uniform distribution with a range of $[-0.7, 0.7]$. Training error was backpropogated through the network with a total gradient equal to the sum of the gradients of the network for each input, multiplied by the gradient of the loss function, and a learning rate parameter. The gradient of the loss function was D_W for neighbors, $-(m-D_W)$ for non-neighbors within the margin, and zero for non-neighbors outside the margin. The margin was set to 0.3. The learning rate parameter was equivalent to dividing the total gradient by the number of training pairs. The network was trained for 80 epochs on each run and 30 runs with new random starting weights were used to avoid local minimums in the loss function. This simulation was repeated for each participant for each possible number of output dimensions (one to six).

The neighborhood judgments of an example participant, Participant 7, are shown in Figure 14. These judgments show a strong influence of color, while shape is also be a factor. Across different colors, some were considered better neighbors than others, in particular peach and tan were close neighbors. These neighborhood judgments and those of the other participants were the input into the DrLIM method described above. The values of the loss function for each of the participants for each possible number of dimensions in the solution are shown in Figure 15. The loss function tended to decrease with more dimensions, and the small departures from monotonicity are due to the difficulty in finding the best weights for each output dimension. An inspection of this figure reveals an elbow point at about three dimensions, but the remainder of the analyses will use the two dimensional solution to simplify the presentation.

The variability across participants in neighborhood judgments is illustrated by plots of the MDS stimuli onto the best two dimensional solutions in Figure 16. The MDS solution for Participant 7 reflects the trends in the neighborhood judgments (Figure 14): a strong clustering by color, but with a contribution of shape as well. This pattern was common across participants, with only Participant 5 showing a very different pattern. Shape is the

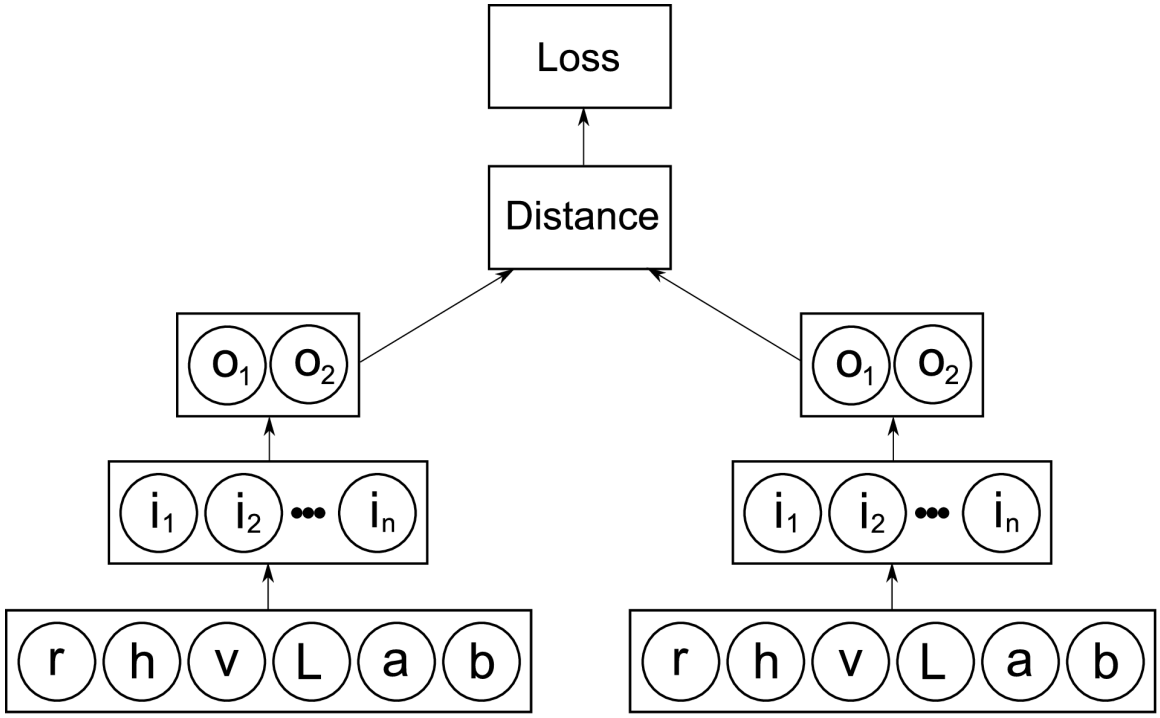

Figure 13. Illustration of the structure of the artificial neural network used to implement the DrLIM method for the multidimensional scaling analysis in Experiment 4. An input layer feeds the parameters forward to a hidden layer and then an output layer for each of the stimuli. The output layer in this illustration is for a two dimensional solution. A pair of stimuli are fed through the network (parameters are the same in the two networks) and the values of the outputs are fed into a distance function, and then into a loss function based on a participant's neighborhood judgments.

strongest cue for Participant 5 and size also appears to be more important than color. Participant 4 shows the same basic color clustering as Participant 7, but with the colors divided into groups. Participant 2 shows this phenomenon as well, but in this case only the green fruit shapes are isolated.

While the scaling solution presented here satisfies the loss function used in DrLIM, there are many possible ways to define what is meant by a good scaling solution. Research in dimensionality reduction has produced many different algorithmic approaches to this problem and these approaches can produce different solutions (Hadsell et al., 2006; Hinton & Salkhutdinov, 2006; Roweis & Saul, 2000; Tenenbaum, de Silva, & Langford, 2000). In order to make strong claims about models it will be important for the neighborhood judgments to lead to consistent representations under a variety of plausible assumptions.

Results and Discussion

The key analysis step in testing whether a prototype model is a good description of the selected categories is to determine the number of clusters in the MCMC samples. A larger and more comprehensive investigation of these category structures could use a variety of methods for determining the number of clusters, including fitting mixture models to the

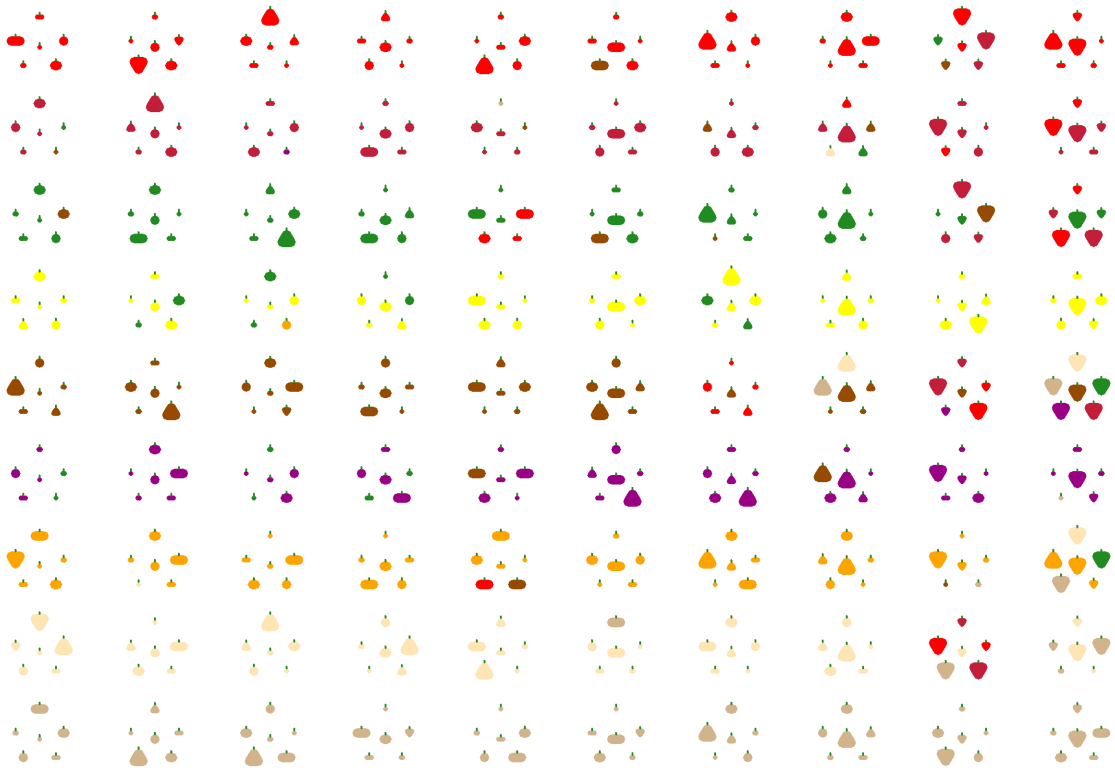

Figure 14. Neighborhood judgments made by Participant 7 in Experiment 4. The center of each cluster was the target and the five surrounding shapes were the neighbors selected by the participant on that trial.

data. In one mixture model approach, a certain number of components (or clusters) is assumed and then the parameters that produce the maximum likelihood of the data are found. The maximum likelihoods from different numbers of clusters are then compared using a model selection criterion, such as the Bayesian Information Criterion (Schwartz, 1978). More thoroughly Bayesian options exist as well, such as inferring a distribution over the number of clusters by using a Dirichlet process mixture model (Antoniak, 1974; Ferguson, 1983; Neal, 1998). Since our goal here is primarily to illustrate how the MCMC approach could be used to address questions about prototype and exemplar models, we will restrict ourselves to a qualitative analysis of the data.

Inspection of the data suggested that the procedure used in this preliminary experiment did not provide enough samples for the MCMC algorithm to explore the entire stationary distribution. Acceptance rates across subjects and questions ranged from 2.5% to 15%, which indicates that the proposal distribution may have been too wide. Discarding the first half of the samples from each chain results in the left column of Figure 17. Each panel shows the states of the three Markov chains sampling from one of Participant 7's categories, plotted on the best two-dimensional MDS solution. The three chains show a fair

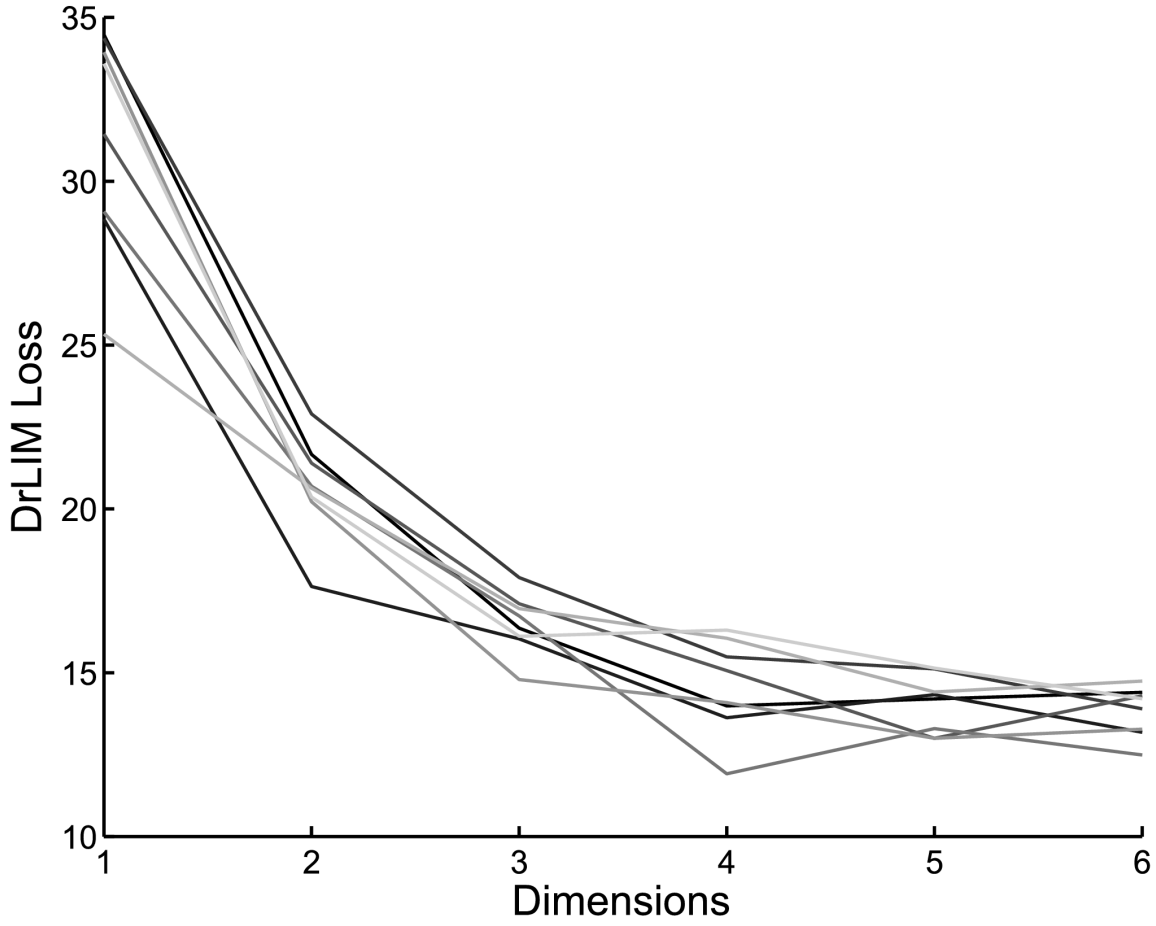

Figure 15. Value of the loss function optimized by the DrLIM algorithm for multidimensional scaling in Experiment 4, as a function of dimension. Each line represents an individual participant.

amount of convergence: each region of the samples seems to involve at least two chains. However, the ideal would be for every chain to have a presence in each cluster of samples. The samples pooled across chains for each category are shown in the right column. The separation between samples shown in Figure 17 suggests that they were drawn from a multimodal distribution, but simulations showed that similar numbers of clusters could have been produced from three non-converged chains sampling from a category distribution with a single mode.

This experiment outlines the procedures necessary for the MCMC method to produce evidence against a simple prototype model of natural categories. An advanced MDS method was adapted to use for human experiments for the purpose of determining an explicit mapping between the stimulus space and the psychological space. Basic-level categories were tested using the MCMC method. While this experiment was primarily intended to illustrate how the MCMC method can be used to explore theoretical questions about the representation of category structures, the preliminary results provide some suggestion of multimodality. Further, more comprehensive, experiments will be required to obtain

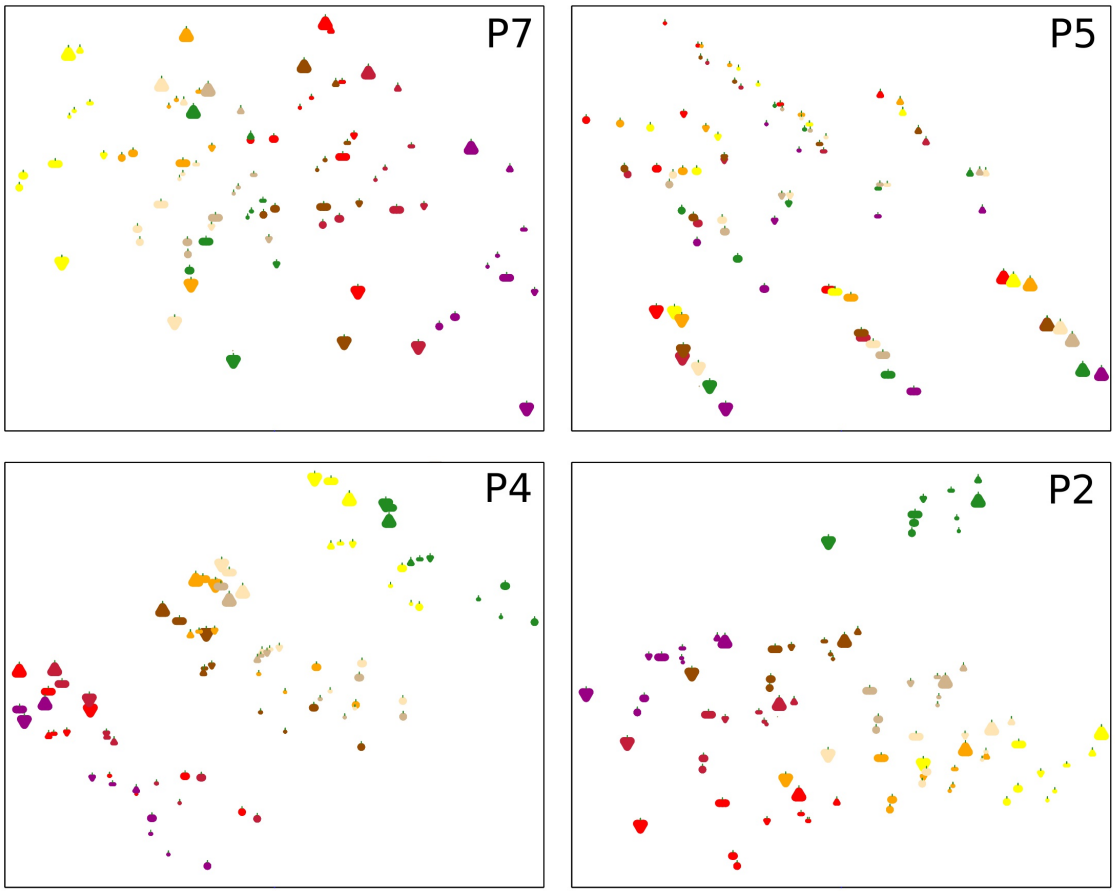

Figure 16. Two dimensional scaling solutions for the stimuli presented during the neighborhood task in Experiment 4. Panels correspond to individual participants. The locations of individual fruit stimuli reflect their locations in the recovered psychological space.

definitive evidence as to whether these categories have structures that are inconsistent with prototype models.

General Discussion

We have developed a Markov chain Monte Carlo method for exploring probability distributions associated with different mental representations. This method allows people to act as a component of an MCMC algorithm, constructing a task for which choice probabilities follow a valid acceptance function. By choosing between the current state and a proposal according to these assumptions, people produce a Markov chain with a stationary distribution that is informative about their mental representations. Our experiments illustrate how this approach can be used to investigate the mental representation of category structures. The results of Experiments 1 and 2 indicate that the MCMC method accurately uncovers differences in mental representations that result from training people on categories with different structures. However, the potential of the method lies in being able to dis-

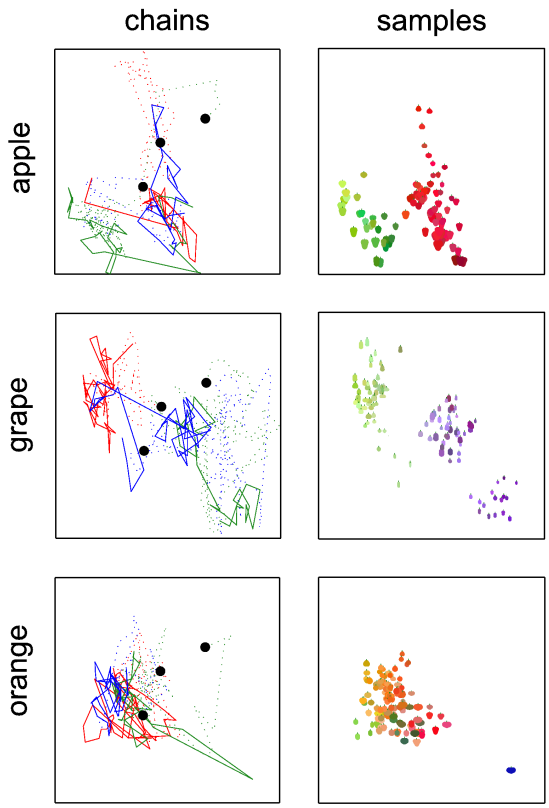

Figure 17. States of the Markov chains and samples for Participant 7 in Experiment 4, plotted on the best two dimensional MDS solution. In the first column, each of the three chains for a fruit condition are displayed in a different color. Dotted lines are discarded samples, solid lines are retained samples, and the black discs are the starting points for each chain. The second column shows the samples that were retained in each fruit condition.

cover the structures that people associate with categories that they have learned through interacting with the world. Experiment 3 collected samples from natural categories of animals over a nine-dimensional stick figure space. These samples illustrated the differences between the representations of these animals along the parameters of the stimuli, giving more insight into the representations than could be gleaned by finding the discrimination boundary alone. Experiment 4 illustrated a test of a central question about representation of natural categories. Multidimensional scaling was combined with MCMC to probe the category structure of apples, grapes, and oranges. The results provided a demonstration of how to test a models of natural categories and open up the possibilities of future empirical investigations with this method.

In the remainder of the paper, we consider the impact of relaxing the assumptions we have made about human decision making, how the MCMC method relates to other methods that have been used for exploring mental representations, and provide some cautions and recommendations for users of this method. The latter draw on both our experience in running these experiments and the theoretical properties of MCMC, and are intended to

make it easier for other researchers to use the MCMC method to explore psychological questions. We close with a brief discussion of the implications of our results, and some possible future directions.

Relaxing the assumptions

To match human decision making to the MCMC algorithm, we made several strong assumptions about human decisions:

1. Participants use a ratio rule to make decisions
2. Category representations are unchanged throughout the experiment
3. Decisions are based only on the current stimulus pair

We have already addressed sensitivity to the first of these assumptions, showing that the algorithm will produce meaningful results for behavior that is more or less deterministic than probability matching. In the remainder of this section, we address the other two assumptions. The implications of relaxing the second assumption are considered in the section on *category drift* and the robustness of the method to violations of the third assumption are considered in the *context effects* section.

Category drift. As the MCMC analysis is based on the assumption that the category distribution remains constant over time, a concern in the experiments was to make sure that participants' categories were not influenced by the test trials. In a pilot version of Experiment 1, training and test trials were not interleaved, instead all of the training trials were presented at the beginning of the experiment. Data from an example participant from the pilot experiment is shown in Figure 18. The Markov chains converge, but once they converge the chains tend to move together. This result, if not simply due to chance, is due to changes in the underlying category structure. If the distribution we are drawing samples from is not constant, then the samples will reflect a combination of all of the states of the dynamic distribution, not the stationary distribution that is assumed. In Experiments 1 and 2, we attempted to reduce category drift through interleaving blocks of training and test trials. This strategy seemed effective as the behavior in Figure 18 was not observed. This caution is necessary for natural categories as well. If the categories participants have formed are fragile, then they could change throughout a MCMC test session through the influence of test trials. And if the stimuli have more than one dimension, it will be much more difficult to identify category drift by examining whether the Markov chains are traveling together.

Context effects. Two types of context effects are considered in this section: context effects arising from between-trial dependencies and context effects that are constant over the course of the experiment. Sequential effects are nearly ubiquitous in cognitive science, with examples including representations of spatial or temporal intervals (Gilden, Thornton, & Mallon, 1995), serial position effects in free recall (e.g., Murdock, 1962; Raaijmakers & Shiffrin, 1981), and priming (Schooler, Shiffrin, & Raaijmakers, 2001; Howes & Osgood, 1954). However, MCMC was developed for the digital computer and assumes that decisions depend only on the current stimuli and are not influenced by any previous trials. In the MCMC task, there are a multiple plausible between-trial dependencies that could arise.

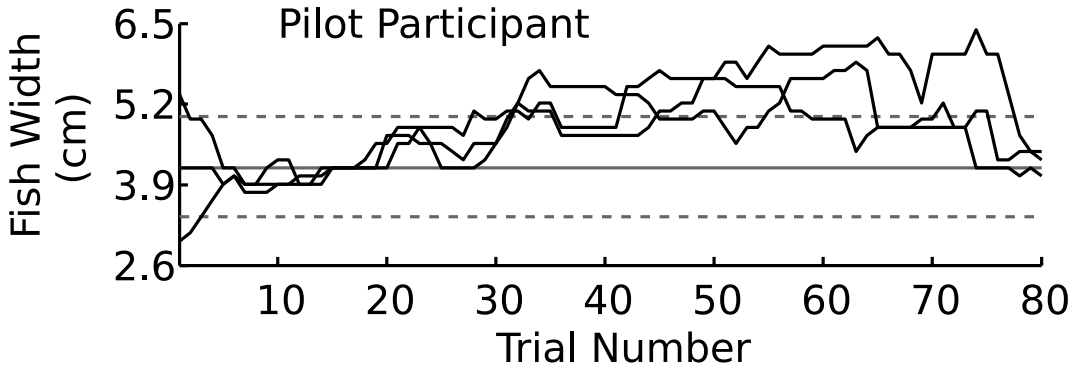

Figure 18. Markov chain states for a pilot participant that illustrates category drift. The chains converge and then move together over time.

Participants could grow attached to (or tired with) a particular stimulus and change the probability of accepting a new stimulus.

We can attempt to model these context effects with a more general decision rule. Instead of Equation 7, we will use a rule that assumes participants can remember which example was presented on a previous trial. This rule assumes that participants have a fixed chance of ignoring the category probabilities of the stimuli and instead are biased to choose either the state of the Markov chain or the proposal of the Markov chain,

$$a(x^*; x) = gb + (1 - g)\frac{p(x^*)}{p(x^*) + p(x)} \tag{18}$$

where g is the probability of making a biased choice, and b is the probability of choosing the proposal when making a biased choice. Interestingly, if participants are biased towards remaining in the same state, the stationary distribution does not change. We can see this in Equation 18 if $b = 0$. In this case the acceptance probability of a new state is scaled downward by $1 - g$ and the acceptance function still satisfies the detailed balance equation (Equation 1). Simulations verified this result, as shown in Figure 19. The left panel shows one million states from an unbiased decision rule that is sampling from a mixture of Gaussians distribution. The middle panel shows one million states from a decision rule that flips a fair coin and either uses the ratio rule or remains in the current state. These simulations illustrate how the stationary distribution does not change when participants are biased towards the current state as in Equation 18.

The outcome is not as faithful to the stationary distribution for participants who can remember the previous choice, and are then biased towards choosing the new stimulus. This bias can be modeled by setting $b = 1$ in Equation 18. We do not have analytical results for this case, but the right panel in Figure 19 shows simulated results from a mixture of two Gaussians. The modes of the distribution seem to be fairly unbiased, while the variance of each of the Gaussian components seems to have increased. As the exact spread of the distribution cannot be determined if participants are using an exponentiated decision rule (Equation 9), it does not seem like a simple bias toward the proposed state will much harm the interpretation of the results. Context effects arising from between-trial dependencies can

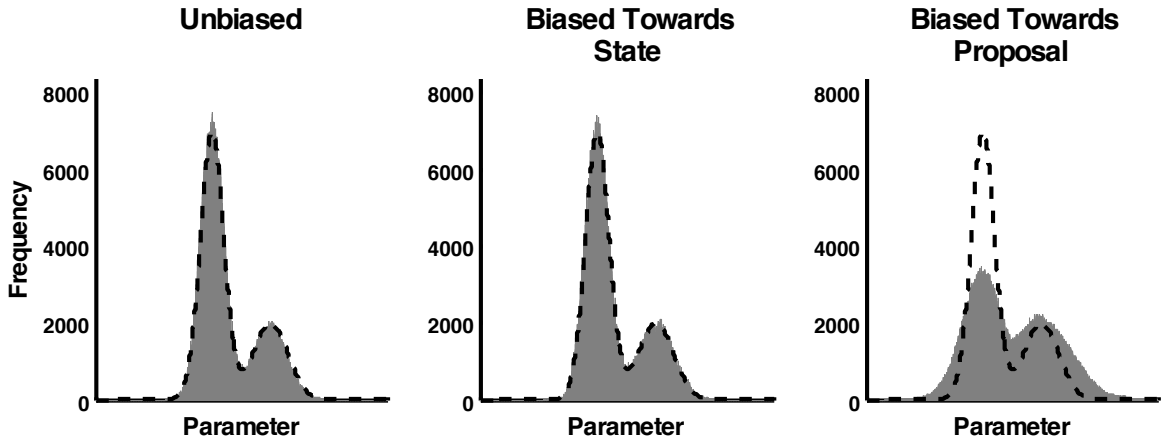

Figure 19. States of a Markov chain for unbiased and biased decision rules. The states are shown in a histogram with the expected results from the true stationary distribution plotted as a dashed curve over the bars. The left panel is the unbiased decision, the middle panel is a decision biased by always choosing the current state on half of the trials ($b = 0$ and $g = 0.5$), and the right panel is a decision by always choosing the proposal on the half the trials ($b = 1$ and $g = 0.5$).

certainly take a more complicated form than Equation 18, but the simulations for a simple form of bias show that the method has some robustness to between-trial dependencies.

These bias effects are undesirable, even if they do not change the asymptotic results. This reason is that the convergence of the Markov chain will be slowed because fewer trials are informative. An experimental method for mitigating these effects is to interleave multiple Markov chains from multiple categories. Multiple chains introduce variety into the stimuli participants judge during any short period of time and chains from different categories increase that variety greatly. This diversity should help counter any fatigue or affinity participants build up toward a stimulus or set of stimuli.

Context effects that remain constant during an experiment can come from a large variety of sources including the internal and external conditions of the experiment (Criss & Shiffrin, 2004). Instructions in particular can have effects that are unforeseen by the researcher (McKoon & Ratcliff, 2001). In the MCMC experiments, the previous categories on which a participant has been tested may influence a participant's conception of other category structures. The label used for the category could have an effect as well (for instance, cats versus felines), though Experiment 1 shows that the results are robust to the form of the question. There is no way to produce a context-free experiment, so the samples gathered using the MCMC method must be considered with respect to the context of the experiment.

Other methods for studying mental representations

The results of Experiments 3 and 4 demonstrate that the MCMC method can be used to explore categories that people have built up through real-world experience. However, this is certainly not the only method that has been developed to gain insight into people's concepts and categories. In this section, several existing methods of gathering and analyzing categorization data are reviewed and compared to the MCMC method.

Multidimensional scaling and additive clustering. Multidimensional scaling (MDS; Shepard, 1962; Torgerson, 1958) and additive clustering (AC; Shepard & Arabie, 1979) have been effectively used in many domains to convert judgments of similarity into a psychological structure. Similarity judgments are assumed to reflect the distances between stimuli in a mental space. MDS uses the relationship between similarity and distance to construct a space for the stimuli that best respects people's similarity judgments, while AC constructs a featural representation. As demonstrated in Experiment 4, MDS can be used as a complementary method to the MCMC method. MDS/AC are methods for representing the stimuli, while MCMC determines the probability distribution. As a result, samples gathered by MCMC can be projected from the original parameter space into the psychological space derived by MDS. This transformation gives the MCMC samples additional significance and allows researchers to test categorization models, such as the prototype model, that are defined within a psychological space.

The judgments made during an MDS or AC trial share properties with the judgments made during an MCMC trial. MDS and AC data are collected by one of two methods: directly, by rating the similarity or dissimilarity of pairs of stimuli (e.g., Navarro & Lee, 2004), or indirectly, by counting the confusions between pairs of stimuli (e.g., Rothkopf, 1957; Nosofsky, 1987). The MCMC task also requires judgments between pairs of stimuli – specifically which stimulus is a better example of a particular category. Both of these tasks depend on an instructional context. In Experiment 4, the MDS context was fruit, while the MCMC context was a particular fruit category. The difference between the tasks is that participants are asked about different aspects of the stimulus pair. MDS/AC data are the psychological distances between a pair of stimuli, while MCMC data are the members of each pair that are the better examples. Even if asked in the same context, these questions will produce different information, which are suitable for different applications.

Discriminative methods. A discriminative choice is defined here as a decision between a set of category labels for a single stimulus. These data are used by discriminative methods, such as the estimation of classification images (Ahumada & Lovell, 1971; Gold et al., 2000; Li, Levi, & Klein, 2004; Olman & Kersten, 2004; Victor, 2005; Tjan & Nandy, 2006), to make inferences about category structures. Classification images provide a way to estimate the boundaries between categories in high-dimensional stimulus spaces. In a typical classification image experiment, participants classify stimuli that have had noise added. A common choice for visual stimuli is to draw samples from pixel-wise independent Gaussian distributions with a mean equal to one of the images participants are asked to find. Participants are presented this noisy stimulus and asked to make a forced choice as to the identity of the noiseless image. Based on the combination of the actual stimulus and the response made, a classification image is a hyperplane that separates the responses into different classes (Murray, Bennett, & Sekuler, 2002).

Discriminative methods make different assumptions about human categorization than the MCMC method, and for the purposes of predicting discriminative judgments, can be best understood using the distinction between discriminative and generative approaches to classification that is used in machine learning (Jordan, 1995; Ng & Jordan, 2001). Following this distinction, *discriminative* methods implicitly view human categorization as involving some function that maps stimuli directly to category labels, and attempt to estimate the

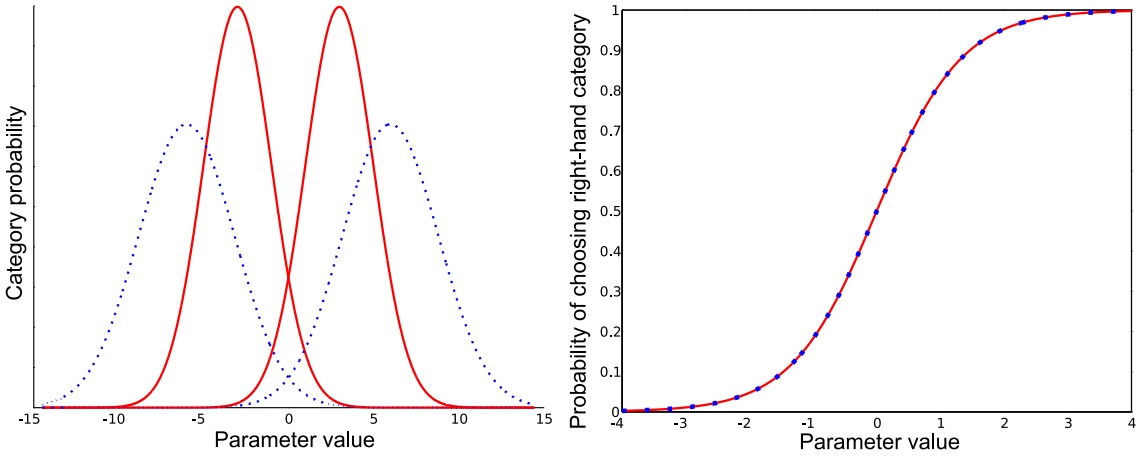

Figure 20. Discriminative choices do not determine category means. The left panel shows two sets of Gaussian distributions with equal within-set variance and an optimal decision boundary at zero. The right panel shows the response probabilities for the right-hand category for each pair of distributions. Despite different means in the left panel, the response probabilities in the right panel are equal.

parameters of that process. This approach provides information about how a judgment is made about two or more alternatives, and it is useful when the question of interest is how people make these discriminations, and what features of the stimuli are relevant. However, other information, such as the mean and variance of the category, is lost when using discriminative methods.

Experiment 3 showed empirically that the mean of a set of discriminative judgments does not provide a good estimate of the natural category mean. In training studies the exemplars are known, and the category means for a particular model can be estimated from the exemplars so this problem does not arise. However, in general, this approach faces a significant obstacle: discriminative trials do not contain the information necessary to determine a category mean unless very restrictive assumptions are made about the structure of a category. Figure 20 shows the probability distributions associated with two sets of categories. While these categories look very different, they lead to the same predictions about discriminative responses. The left panel shows two pairs of Gaussian distributions that have the same optimal decision boundary, but different means and variances. Using the ratio rule of Equation 8, the response probabilities for choosing the distribution of each pair with the higher mean were computed and are displayed in the right panel. Despite the difference in category centers, the response probabilities for the two sets of distributions in a discriminative paradigm are exactly the same. The multivariate analysis in Appendix B extends this result, showing that the means of the categories that can mimic a particular discriminative response function do not have to be proportional to the original means. These results demonstrate that it is not possible to determine category means from discriminative data, unless even stronger assumptions than optimal boundaries, Gaussian distributions, and equal covariance are made.

In contrast, the MCMC method corresponds to a *generative* view of discriminative

judgments, assuming that each category is represented as a probability distribution over objects, and that categorization is the consequence of using Bayes' rule to work back from these distributions to a decision as to how to categorize new objects. The samples gathered approximate the actual distributions themselves, allowing us to learn more about the categories than if a discriminative model had been used. A researcher can perform any test on these distributions, participant to there being enough data to support the test. The advantage of the generative nature of the MCMC method is that additional information is obtained, such as estimates of category means and variability as in Experiment 3.

Discriminative judgments can be predicted using the probability distributions estimated by MCMC, by making assumptions about the decision process that participants use. The discrimination predictions based on MCMC judgments and optimal decision boundaries were compared to actual discrimination judgments in Experiments 1, 2, and 3. The match was above chance, but was far from perfect. In general, if the experimenter is only interested in a very particular and limited set of information about category discriminations, then the lack of flexibility of the discriminative model can be an advantage. Fewer trials will be necessary to produce the same results, assuming the assumptions made by the discriminative model are correct (Jordan, 1995; Ng & Jordan, 2001). Classification images provide an example of the efficiency of discriminative methods in collecting boundary information, having been successfully applied to estimating linear boundaries between images of 10,000 pixels (e.g., Gold et al., 2000). The MCMC method has the advantage of revealing gradient information, but would be impractical for estimating a decision boundary in this situation. Its strength is in providing a much richer picture of the structure of a category than information about the boundaries that determine discrimination judgments.

Typicality ratings. Typicality ratings are a useful tool for determining how well objects represent a category. These ratings are most commonly used to explore natural categories and were instrumental in supporting the view that natural basic-level categories have a prototypical representation (Rosch & Mervis, 1975). Later work showed that typicality ratings in a training paradigm were well-described by an exemplar model (Nosofsky, 1988). The most commonly used categorization tasks are discrimination tasks, but the information gathered by typicality ratings is often more stable across different contexts. As an example, the boundaries of color categories show a lot of variation across different cultures. In constrast, focal colors, or the best examples of colors, do not show much variation across cultures (Lakoff, 1987). Other experiments suggest that the same result occurs for odors (Chrea, Valentin, Sulmont-Ross, Hoang Nguyen, & Abdi, 2005). These results and others (Labov, 1973), show that category boundaries are more sensitive to context effects than category centers. Thus, typicality judgments may be more useful for studying basic cognitive processes than discrimination judgments.

A typicality rating task and the task used in the MCMC method are very similar. The ratings can be typicality judgments of single objects (e.g., Storms et al., 2000) or a pairwise judgment of which object is more typical (e.g., Bourne, 1982). The pairwise judgment is exactly the task used in the MCMC method. However, the MCMC method provides a way to estimate the distributions associated with natural categories that is more efficient than previous methods for gathering typicality ratings. Experiments that collect typicality ratings have relied on the researcher to choose the test objects (e.g., Labov, 1973;

Rosch, Mervis, et al., 1976; Nosofsky, 1988). If the researcher has a lot of knowledge about the structure of the category, then this method will be efficient. However, with natural categories the researcher often has less knowledge, because he or she does not know what examples were used to construct the category. Out of many hypotheses about a natural category structure, few will be true. The resulting experiment will require a large set of stimuli to be tested, and few trials will be informative.

Experiment 3 demonstrated just this kind of situation: the distributions associated with each individual animal shape were only non-zero in a small region of the parameter space required to test all of the animals. Choosing a set of animal shapes for participants to rate would have been inefficient because almost all would fall outside the bounds of the categories. The MCMC method is more effective in this situation because it uses a participant's previous responses to choose informative new test objects. In general, the MCMC algorithm is particularly effective for sampling from distributions that occupy a small region of a parameter space (Neal, 1993). As a result, the MCMC method presented here is well suited for determining the distribution associated with a natural category.

Categorization models. Categorization research has benefited from intensive investigation using a large number of mathematical models. These models have been developed to fit the data from a variety of tasks including binary classifications (e.g., Nosofsky et al., 1994; Ashby & Gott, 1988; Nosofsky, 1987), typicality judgments (e.g., Storms et al., 2000; Nosofsky, 1988), or stimulus reconstructions (e.g., Huttenlocher et al., 2000; Vevea, 2006). In order to produce predictions, the models need to be adapted for each task in which the researcher is interested. The MCMC method is not a model of categorization, but instead is a method for collecting data. To make this discussion more concrete, we will describe how the MCMC method relates to a very well known model of categorization: the Generalized Context Model (GCM; Nosofsky, 1986).

The GCM stores all exemplars that a participant has observed. In order to make a category judgment, the summed similarity of a new stimulus to each of the examples from a category is computed, as shown in Equation 12. These summed similarities for a category are combined in a ratio rule with the summed similarities of other categories to response probabilities, as in Equation 11. In essence, the exemplars are peaks of a similarity function over the stimulus parameter space. The generalization gradient and dimension weights of the GCM determine the spread around these peaks. Each category has a separate similarity function and the strength of an example in a particular category is determined by the value of the similarity function at the parameters of the test stimulus. An illustration of the similarity function created by the GCM is shown in Figure 21.

There is already an explicit model of the MCMC task – participants respond according to the ratio rule based on the similarity functions of the different categories. This is the same assumption as that used in the GCM, except we make this assumption as part of the design of our experiment rather than part of the analysis. If the assumption holds, the results our method produces samples from a probability distribution that is proportional to the summed similarity of the objects to the exemplars. Making this assumption earlier in the process of data collection allows it to be leveraged to collect more informative data via MCMC. However, it should be noted that this method does not draw samples from the stored exemplars in the GCM. If people are behaving according to the GCM, then

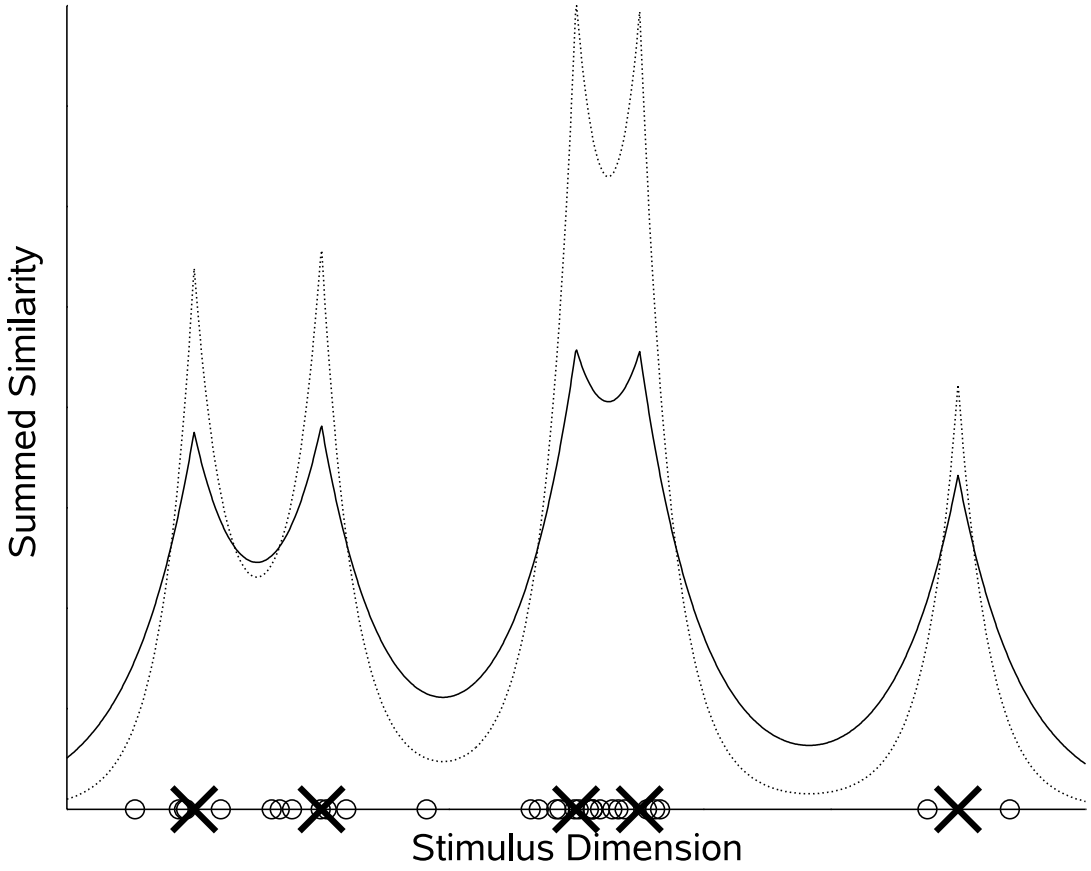

Figure 21. Relationship of samples to exemplars in the Generalized Context Model. The crosses on the horizontal axis are exemplars, the solid curve is the resulting similarity function, the dashed curve is the exponentiated function, and the circles are samples from the exponentiated normalized function.

the samples will be a proportional approximation to the similarity function that the GCM produces for a category, as shown in Figure 21. These samples will reflect the exemplars plus the generalization gradient and the dimension weights, not the exemplars themselves. Also, if participants deviate from probability matching and follow the choice rule in Equation 9, the resulting probability distributions will show an effect of the γ parameter and will not be equivalent to the similarity function used to represent exemplars in the GCM.

Psychophysical staircases. The psychophysical method of adaptive staircases (Cornsweet, 1962) is used to find a participant's threshold in a task in which the intensity of the stimulus is varied. A simple implementation of this method is to show a participant a series of trials, half of which contain a stimulus and half of which are noise alone. The participant responds as to whether the stimulus is present. If the response is correct, then the intensity of the stimulus is lowered to make the task more difficult. If the response is incorrect, the intensity is raised to make it easier. This staircase automatically converges

on the 50% accuracy threshold. The number of consecutive correct or incorrect responses required can be manipulated to change the percent accuracy to which the staircase will converge (Levitt, 1971).

The MCMC method resembles the method of adaptive staircases in that they both adjust trial presentation based on participant responses. Both have a current state in the stimulus space and transition to a new state based on human responses. Both methods use Markov chains – for psychophysical staircases, the next state depends only on the responses to the current state. However, the Markov chains of the two methods use different contingencies for making transitions. The MCMC method transitions to the state that the participant chooses, while a staircase transitions based on the correctness of the participant's response. The stationary distribution of the MCMC method is the category distribution while the stationary distribution of a psychophysical staircase is unknown apart from its target mean. The most salient difference between the two methods is that the MCMC method has no ground truth. Rather than converging to a distribution centered around a point with a certain objective level of accuracy, it converges to a distribution reflecting a participant's subjective degrees of belief about the stimuli.

Cautions and recommendations

While the MCMC method has many desirable theoretical properties and the experiments illustrate that it can produce meaningful results for natural categories, there are several issues that can arise in its implementation. In this section, we describe some of the potential problems and some ways in which they can be solved.

Designing the experiment and analyzing the data

Burn-in. As mentioned previously, MCMC requires a number of trials to converge to the stationary distribution. These pre-convergence trials are not samples from the distribution of interest and must be discarded. Unfortunately, there are few theoretical guarantees about when a Markov chain has converged to its stationary distribution. A large and varied set of heuristics have been developed for assessing convergence, and we applied different heuristics in different experiments. In Experiments 1 and 2, we waited until the chains crossed, a simple heuristic that can be applied in one dimension. Since the chains will definitely cross once they reach their stationary distibution (as each chain should visit every state with probability proportional to its stationary probability), this gives us a simple criterion to check for convergence. In Experiment 3, a variant of an early standard known as the "thick pen technique" (Gelfand, Hills, Racine-Poon, & Smith, 1990) was applied. This heuristic involves looking at a fixed number of chains and removing the early samples from each until they are visually indistinguishable. More objective measures have been developed to determine whether a Markov chain has converged, but are usually more difficult to implement and can be computationally expensive. There is quite a large number of these methods and they vary in their applicability to different sampling problems (see Brooks & Roberts, 1998; Mengersen et al., 1999; Cowles & Carlin, 1996 for reviews).

Dependent samples. An issue with MCMC that affects the power of the method is that there is a dependency between successive samples. The states of a Markov chain are not drawn independently – the overall probability of transitioning to a new state depends

on the current state. There are many ways to deal with sample dependence. A common method is to keep every nth sample and discard the rest, as the dependence between samples decreases as the distance between samples increases. Thus, by dropping out a large enough number of intermediate samples, the remaining set will be pseudo-independent. In the experiments, all of the samples were retained. A sample can be dependent when computing the moments of a distribution, or any other expectation with respect to that distribution. However, if a calculation that involved the number of samples (e.g., t-test, ANOVA) were used, the number of states of the Markov chain cannot be used as the sample size. One alternative is to use the pseudo-independent sample in the calculation. A second alternative is to compute the effective sample size of the Markov chain states and using that as the sample size.

Bounded parameter spaces. The MCMC method uses a symmetric proposal distribution, which requires care when applying to bounded parameter spaces. For states near the boundary, a large number of proposals will be outside the bounds. A common impulse in this situation is to truncate the proposal distribution and normalize. However, this scheme will violate detailed balance (Equation 1) because the proposal probabilities between a state at the edge of the space and a state in the middle of the space will not be the same. Experiments 1 and 2 dealt with this problem by allowing proposals outside of the parameter range and rejecting participants who accepted out-of-bounds proposals. This method is not very appealing, especially for experiments in which each participant requires a long time to run. A more sophisticated approach was adopted for Experiments 3 and 4. Proposals of parameter values outside of the parameter range were allowed, but these proposals were automatically rejected – adding another sample of the Markov chain at the current state without presenting a trial to the participant.

The justification for this approach is straightforward. In committing ourselves to a restricted range of parameter values, we only want to measure the subjective function $f(x)$ over those values. Thus, we treat parameters outside the range as having $f(x)$ equal to zero. As a consequence, we will automatically reject any proposal that goes outside the range, and converge to a stationary distribution that is proportional to $f(x)$ normalized over the set of parameter values allowed in the experiment. In principle, we can actually run the MCMC procedure without bounds on the values of the parameters, since it should converge to a sensible distribution regardless, but we have not tested this in practice.

We can also check that rejecting proposals outside the parameter range is valid by separately showing detailed balance holds for in-bounds and out-of-bounds proposals. Within the in-bounds region detailed balance holds because the proposal probability between any two states using a symmetric distribution is equal and a Barker acceptance function is assumed. Between the in-bounds and out-of-bounds regions proposals are allowed, but an always-reject acceptance function has been adopted. Assuming that in-bounds proposals from an out-of-bounds state are always rejected (which will never actually be implemented), detailed balance is satisfied – between any in-bounds and out-of-bounds state both sides of the detailed balance equation are zero.

Selection of a proposal distribution. When implementing the MCMC method, a proposal distribution must be chosen. In the limit, any distribution that is symmetric and allows the Markov chain to visit every state in the space will produce samples from the

same distribution. However for any practical number of trials, some proposal distributions will be more efficient than others. In Experiments 1-3, we used Gaussian distributions that had a variance along each parameter equal to a fixed percentage of that parameter's range. The percentage for each experiment was chosen based on pilot studies, looking for a percentage that would give an average acceptance probability of around 20% - 40%. This value was sought because it is the optimal acceptance probability when sampling from a Gaussian distribution (Roberts et al., 1997) and we anticipated that our subjective probability distributions would be similar to Gaussians, consistent with several standard models of human category learning (Ashby & Alfonso-Reese, 1995; Reed, 1972).

Recent research in statistics has focused on developing automated methods for adjusting a proposal distribution. A variety of related algorithms collectively known as adaptive MCMC were developed for this reason (Gilks, Roberts, & George, 1994; Gelfand & Sahu, 1994; Gilks, Roberts, & Sahu, 1998). These algorithms work by using samples from the chain in order to adjust the proposal distribution to be more efficient. Similar methods could be used to adjust proposal distributions when using MCMC as an experimental procedure, potentially increasing the efficiency of the method by shortening burn-in and ensuring better exploration of the distributions associated with different categories.

Isolated modes. A concern in any attempt to characterize a natural category is that there may be peaks of the probability distribution that are isolated from the rest of the peaks by regions of stimuli that are not members of the category. This concern was especially valid in Experiment 4 in which the mode of red apples might be isolated from the mode of green apples by a region of poor category members. Drawing samples from isolated peaks can be difficult under MCMC. The width of the proposal distribution can be increased to raise the probability of proposing a state in an isolated mode, but has the downside of decreasing the overall probability of accepting proposals. In Experiment 4, we used a mixture distribution that combined a high proportion of local proposals with the possibility of making large jumps in the parameter space. Despite using this proposal distribution, the chains did not mix well in this experiment, which could be due to chains not switching between isolated modes.

One alternative approach is to use algorithms developed in the machine learning literature to specifically deal with this issue. The key to these methods is to sample from the target distribution at different *temperatures*, meaning that samples are drawn from distributions proportional to the target distribution raised to different powers. Low temperatures refer to large exponents and high temperatures refer to small exponents. Metropolis-coupled MCMC (Geyer, 1991) runs several chains with transition kernels at different *temperatures* in parallel. The states of the chains are occasionally exchanged, allowing chains with small proposal widths to make large jumps. Simulated tempering (Marinari & Parisi, 1992) is a similar algorithm, but instead of running several chains in parallel, one chain is run and transition kernels of different temperatures are swapped. These ideas provide help in most applications, and it may be possible to apply them in the human MCMC procedure. In our setting, the "temperature" of a chain corresponds to the value of the parameter γ used in Equation 9. Consequently, MCMC methods that exploit variation in temperature could be implemented by developing a way to adjust the γ parameter of a participant's acceptance function, perhaps through instructions or manipulation of the motivation of the participant

via rewards.

Conclusion

Markov chain Monte Carlo is one of the basic tools in modern statistical computing, providing the basis for numerical simulations conducted in a wide range of disciplines. The results presented in this paper suggest that this class of algorithms can also be of use in psychology, not just for numerical simulation, but also for uncovering mental representations. The experiments in this paper have applied the MCMC method to categorization tasks, but it could also be used in a variety of other domains as well. Many formal models of human cognition (Oaksford & Chater, 1998; Chater, Tenenbaum, & Yuille, 2006; Anderson, 1990; Shiffrin & Steyvers, 1997) assume that there is a range of degrees of belief in different hypothetical outcomes. The MCMC method is designed to estimate these subjective functions. Using this technique, interesting data could be collected in many different areas of study, including investigating the subjective response to a probe of memory or exploring the internal values of different response alternatives in decision making. We are particularly excited about the prospects for using this approach together with probabilistic models of cognition, making it possible to estimate many of the distributions that otherwise have to be assumed in these models.

Beyond the potential of MCMC for uncovering mental representations, we view one of the key contributions of this research to be the idea that we can use people as elements in randomized algorithms, provided we can predict their behavior, and use those algorithms to learn more about the mind. The early success of this approach suggests that other MCMC methods could be used to explore mental representations, either by using the Barker acceptance function or by establishing other connections between statistical inference and human behavior. The general principle of allowing people to act as components of randomized algorithms can teach us a great deal about cognition.

References

Ahumada, A. J., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America*, *49*, 1751-1756.

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

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*, 703-719.

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

Ashby, F. G. (1992). *Multidimensional models of perception and cognition.* Hillsdale, NJ: Erlbaum.

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

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

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372-400.

Barker, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, *18*, 119-133.

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society. Series A (General)*, *139*, 318-355.

Billera, L. J., & Diaconis, P. (2001). A geometric interpretation of the Metropolis-Hastings algorithm. *Statistical Science*, *16*, 335-339.

Bourne, L. (1982). Typicality effects in logically defined categories. *Memory & Cognition*, *10*, 3-9.

Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: The kernel approach with S-plus illustrations.* Oxford: Oxford University Press.

Bradley, R. A. (1954). Incomplete block rank analysis: On the appropriateness of the model of a method of paired comparisons. *Biometrics*, *10*, 375-390.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433-436.

Brooks, S., & Roberts, O. (1998). Assessing convergence of Markov chain Monte Carlo. *Statistics and Computing*, *8*, 319-335.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Special issue on "Probabilistic models of cognition". *Trends in Cognitive Sciences*, *10*(7).

Chrea, C., Valentin, D., Sulmont-Ross, C., Hoang Nguyen, D., & Abdi, H. (2005). Semantic, typicality and odor representation: a cross-cultural study. *Chemical Senses*, *30*, 37-49.

Clarke, F. R. (1957). Constant-ratio rule for confusion matrices in speech communication. *The Journal of the Acoustical Society of America*, *29*, 715-720.

Cornsweet, T. N. (1962). The staircase-method in psychophysics. *The American Journal of Psychology*, *75*, 485-491.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91*, 883-904.

Criss, A. H., & Shiffrin, R. M. (2004). Context noise and item noise jointly determine recognition memory: a comment on Dennis and Humphreys (2001). *Psychological Review*, *111*, 800-807.

de Silva, V., & Tenenbaum, J. B. (2003). Global versus local methods in nonlinear dimensionality reduction. In S. T. S. Becker & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 705–712). Cambridge, MA: MIT Press.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification.* New York: Wiley.

Farrell, S., & Ludwig, C. (in press). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin & Review*.

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

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions of various types of information processing. *Cognitive Psychology*, *1*, 225-241.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, *85*, 972-985.

Gelfand, A. E., & Sahu, S. K. (1994). On Markov chain Monte Carlo acceleration. *Journal of Computational and Graphical Statistics*, *3*, 261-276.

Geyer, C. (1991). Markov chain Monte Carlo maximum likelihood. In E. M. Keramidas (Ed.), *Proceedings of the 23rd symposium on the interface: Computing science and statistics.* Interface Foundation.

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995, March). 1/f Noise in Human Cognition. *Science*, *267*, 1837-1839.

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

Gilks, W. R., Roberts, G. O., & George, E. I. (1994). Adaptive direction sampling. *The Statistician*, *43*, 179-189.

Gilks, W. R., Roberts, G. O., & Sahu, S. K. (1998). Adaptive Markov chain Monte Carlo through regeneration. *Journal of the American Statistical Association*, *93*, 1045-1054.

Gold, J., Murray, R., Bennett, P., & Sekuler, A. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology*, *10*(11), 663-666.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavioral Research Methods, Instruments, & Computers*, *26*, 381-386.

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

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on* (Vol. 2, p. 1735-1742).

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97-109.

Heit, E., & Barsalou, L. W. (1996). The instantiation principle in natural categories. *Memory*, *4*, 413-451.

Hinton, G. E., & Salkhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*, 504-507.

Hopkins, J. W. (1954). Incomplete block rank analysis: Some taste test results. *Biometrics*, *10*, 391-399.

Howes, D., & Osgood, C. E. (1954). On the combination of associative probabilities in linguistic contexts. *The American Journal of Psychology*, *67*, 241-258.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, *294*, 2310-2314.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, *129*, 220-241.

Jordan, M. I. (1995). *Why the logistic function? a tutorial discussion on probablities and neural networks* (Computational Cognitive Science Technical Report No. 9503).

Labov, W. (1973). The boundaries of words and their meanings. In B. Aarts (Ed.), *New ways of analyzing variation in english* (p. 340-373). Oxford University Press.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: University of Chicago Press.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *2*, 467-477.

Li, R. W., Levi, D. M., & Klein, S. A. (2004). Perceptual learning improves efficiency by re-tuning the decision 'template' for position discrimination. *Nature Neuroscience*, *7*(2), 178-183.

Liu, J., Golinkoff, R. M., & Sak, K. (2001). One cow does not an animal make: Young children can extend novel words at the superordinate level. *Child Development*, *72*, 1674-1694.

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

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, volume 1* (p. 103-190). New York and London: John Wiley and Sons, Inc.

Marinari, E., & Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters*, *19*, 451-458.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of econometrics.* Academic Press.

McKoon, G., & Ratcliff, R. (2001). Counter model for word identification: reply to Bowers (1999). *Psychological Review*, *108*, 674-681.

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

Mengersen, K. L., Robert, C. P., & Guihenneuc-Jouyaux, C. (1999). MCMC convergence diagnostics: a 'reviewww'. In J. Berger, B. J., A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 6* (p. 415-440). Oxford: Oxford Sciences.

Metropolis, A. W., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1092.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 275-292.

Morey, R. D., Rouder, J. N., & Speckman, P. L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology*, *52*, 21-36.

Morgan, B. J. T. (1974). On Luce's choice axiom. *Journal of Mathematical Psychology*, *11*, 107-123.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482-488.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: weighted sums. *Journal of Vision*, *2*, 79-104.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirchlet processes. *Journal of Mathematical Psychology*, *50*, 101-122.

Navarro, D. J., & Lee, M. D. (2004). Common and distinctive features in stimulus similarity: a modified version of the contrast model. *Psychonomic Bulletin & Review*, *11*, 961-974.

Neal, R. M. (1993). *Probabilistic inference using Markov Chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). Department of Computer Science, University of Toronto.

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

Newman, M. E. J., & Barkema, G. T. (1999). *Monte carlo methods in statistical physics*. Oxford: Clarendon Press.

Ng, A. Y., & Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Nips* (p. 841-848).

Norris, J. R. (1997). *Markov chains*. Cambridge, UK: Cambridge University Press.

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

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87-108.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classifcation, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 700-708.

Nosofsky, R. M., Gluck, M., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition*, *22*, 352-369.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, *7*, 375-402.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 924-940.

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition*. Oxford University Press.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*, 172-191.

Olman, C., & Kersten, D. (2004). Classification objects, ideal observers, and generative models. *Cognitive Science*, *28*, 227-239.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437-442.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, *60*, 607-612.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353-363.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93-134.

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

Rice, J. A. (1995). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, *129*, 369-398.

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *Annals of Applied Probability*, *7*, 110-120.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573-605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 491-502.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*, 178-210.

Rothkopf, E. Z. (1957). A measure of stimulus similiarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, *53*, 94-101.

Rouder, J. N. (2004). Modeling the effects of choice-set size on the processing of letters and words. *Psychological Review*, *111*, 80-93.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*, 2323-2326.

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review*, *108*.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325-345.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function: I. *Psychometrika*, *27*, 124-140.

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

Shepard, R. N., & Arabie, P. (1979). Additive clutering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*, 87-123.

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

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review*, *4*, 145-166.

Silverman, B. W. (1986). *Density estimation.* London: Chapman and Hall.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition*, *65*, 167-196.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1411-1436.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 174-215.

Storms, G., Boeck, P. D., & Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language*, *42*, 51-73.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319-2323.

Tenenbaum, J. B., & Griffiths, T. L. (2001). The rational basis of representativeness. In *Proceedings of the 23rd annual meeting of the cognitive science society.*

Tjan, B., & Nandy, A. (2006). Classification images with uncertainty. *Journal of Vision*, *6*, 387-413.

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: Wiley.

Vanpaemel, W., Storms, G., & Ons, B. (2005). A varying abstraction model for categorization. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Vevea, J. L. (2006). Recovering stimuli from memory: a statistical method for linking discrimination and reproduction responses. *British Journal of Mathematical and Statistical Psychology*, *59*, 321-346.

Victor, J. (2005). Analyzing receptive fields, classification images and functional images: challenges with opportunities for symmetry. *Nature Neuroscience*, *8*, 1651-1656.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*, 101-118.

Wills, A. J., Reimers, S., Stewart, N., Suret, M., & McLaren, I. P. L. (2000). Tests of the ratio rule in categorization. *The Quarterly Journal of Experimental Psychology*, *53A*, 983-1011.

Yellott, J. I. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, *15*, 109-144.

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1160-1173.

Appendix A
Bayesian analysis of choice tasks

Consider the following task. You are shown two objects, x_1 and x_2, and told that one of those objects comes from a particular category, c. You have to choose which object you think comes from that category. How should you make this decision?

We can analyze this choice task from the perspective of a rational Bayesian learner. The choice between the objects is a choice between two hypotheses: The first hypothesis, h_1, is that x_1 is drawn from the probability distribution associated with a category $p(x|c)$ and x_2 is drawn from $g(x)$, an alternative distribution that governs the probability of a non-category object being presented to the participant. The second hypothesis, h_2, is that x_1 is from the alternative distribution and x_2 is from the category distribution. The posterior probability of the first hypothesis given the data is determined via Bayes' rule,

$$\begin{aligned} p(h_1|x_1,x_2) &= \frac{p(x_1,x_2|h_1)p(h_1)}{p(x_1,x_2|h_1)p(h_1)+p(x_1,x_2|h_2)p(h_2)} \\ &= \frac{p(x_1|c)g(x_2)p(h_1)}{p(x_1|c)g(x_2)p(h_1)+p(x_2|c)g(x_1)p(h_2)} \end{aligned} \tag{19}$$

where we use the category distribution $p(x|c)$ and its alternative $g(x)$ to calculate $p(x_1,x_2|h)$.

We will now make two assumptions. The first assumption is that the prior probabilities of the hypotheses are the same. Since there is no a priori reason to favor one of the

objects over the other, this assumption seems reasonable. The second assumption is that the probabilities of the two stimuli under the alternative distribution are approximately equal, with $g(x_1) \approx g(x_2)$. If people assume that the alternative distribution is uniform, then the probabilities of the two stimuli will be exactly equal. However, the probabilities will still be roughly equal under the weaker assumption that the alternative distribution is fairly smooth and x_1 and x_2 differ by only a small amount relative to the support of that distribution. The difference between x_1 and x_2 is likely to be small if the proposal distribution is narrow. With these assumptions Equation 19 becomes

$$p(h_1|x_1, x_2) \approx \frac{p(x_1|c)}{p(x_1|c) + p(x_2|c)} \tag{20}$$

with the posterior probability of h_1 being set by the probabilities of x_1 and x_2 in that category.

The analysis of the "Which is more typical?" question proceeds similarly. Tenenbaum and Griffiths (2001) defined a Bayesian measure of the representativeness of a stimulus x for a category c to be

$$r(x, c) = \frac{p(x|c)}{\sum_{c' \neq c} p(x|c')p(c')} \tag{21}$$

where $p(c')$ is the probability of an alternative category, normalized to exclude c. This measure indicates the extent to which x provides evidence for c, as opposed to all other categories. Applying the Luce choice rule, we might expect that participants asked to decide whether x_1 or x_2 were more typical of c would choose x_1 with probability

$$p(\text{choose } x_1) = r(x_1, c)/(r(x_1, c) + r(x_2, c)) \tag{22}$$

If x_1 and x_2 are approximately equally likely, averaging over all categories, then $\sum_{c' \neq c} p(x_1|c')p(c')$ will be approximately the same as $\sum_{c' \neq c} p(x_2|c')p(c')$. As above, the fact that x_1 and x_2 should be relatively similar makes this assumption likely to hold. In this case, Equation 22 reduces to the familiar Barker acceptance rule for $p(x|c)$ (Equation 7).

Appendix B
The information available from discriminative data

Assume we have two categories with Gaussian distribution representations. The two categories have equal covariance matrices Σ and assume that discriminative judgments are made via a ratio rule (Equation 8). Without loss of generality, let the midpoint between the means of categories 1 and 2 be zero, so that $\mu_1 = -\mu_2$. The probability of assigning a stimulus x to category 1 is

$$P(c = 1|x) = \frac{N(x; \mu_1, \Sigma)}{N(x; \mu_1, \Sigma) + N(x; \mu_2, \Sigma)} \tag{23}$$

where $N(x; \mu, \Sigma)$ is the probability density function for a Gaussian with mean μ and covariance matrix Σ, and we assume that the two categories have equal prior probability. Dividing through by the numerator we obtain

$$P(c = 1|x) = \frac{1}{1 + \exp\{-\frac{1}{2}[(x - \mu_2)^T \Sigma^{-1}(x - \mu_2) - (x - \mu_1)^T \Sigma^{-1}(x - \mu_1)]\}} \tag{24}$$

being a logistic function. Using the distributive property and $\mu_1 = -\mu_2$, the argument of the exponential in the denominator reduces to $-\mu_1^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_1$, which simplifies to $-2\mu_1^T\Sigma^{-1}x$, being a linear function of x.

From the results in the previous paragraph, we can see that the mean μ_1 can be multiplied by any constant and produce the same discriminative judgments, as long as the covariance Σ is multiplied by the same constant. This result was used to produce the univariate example in Figure 20. In the multivariate case, this result means that discriminative judgments can at least determine the the means of the categories up to a multiplicative constant. However, by diagonalizing the covariance matrix, we can show that the multivariate category means are even less constrained. The covariance matrix is diagonalized by $\Sigma = Q^T DQ$, factoring it into matrices of eigenvectors Q and a diagonal matrix of eigenvalues D. The eigenvector matrices can be used to transform μ_1 and x into a new basis, giving

$$-2\mu_1^T Q^T D^{-1} Qx = -2(\mu_1')^T Dx' \tag{25}$$

where μ_A' and x' are the vectors transformed into the new basis and $Q^T = Q^{-1}$ because Q is an orthogonal matrix.

In the right-hand side of Equation 25, each entry of the transformed mean vector is multiplied only by the variance along that dimension. Now each element of the mean vector can be multiplied by a different coefficient, as long as the corresponding variance is divided by that coefficient. Assume we now have two more categories, categories 3 and 4, with $\mu_3 = -\mu_4$ and a shared covariance matrix Σ'. Let K be a diagonal matrix with each entry corresponding to the ratio of an entry of the transformed mean of category 3, $\mu_3 Q$, divided by an entry of the transformed mean of category 1, $\mu_1 Q$. Then μ_3 can take any value where Σ' is positive semi-definite, with

$$\Sigma' = Q^T DKQ \tag{26}$$

and still produce an equivalent boundary between categories 3 and 4 to that between categories 1 and 2.

Not all values of μ_3 will produce a positive semi-definite Σ', but μ_3 can take many values that are not a multiplicative constant of μ_1. This analysis shows that discriminative judgments between a pair of categories cannot determine the category means without additional information. Even knowing that the categories are multivariate Gaussian with equal covariance, the best a researcher can say is that the means are equidistant from a particular point (or, more generally, a particular hyperplane) in a metric determined by their covariance matrix, as this is the relevant criterion for producing the same discriminative judgment. Though this result was shown assuming a ratio rule, odds ratios of multivariate Gaussians, such as $N(x;\mu_1,\Sigma)/N(x;\mu_2,\Sigma)$, have the same property. Dividing the ratio by the numerator results in Equation 24 without the additive constant in the denominator.