Three Case Studies in the Bayesian Analysis of Cognitive Models

Michael D. Lee
Department of Cognitive Sciences
University of California, Irvine

Abstract

Bayesian statistical inference offers a principled and comprehensive approach for relating psychological models to data. This paper presents Bayesian analyses of three influential psychological models: Multidimensional Scaling models of stimulus representation, the Generalized Context Model of category learning, and a Signal Detection Theory model of decision-making. In each case, the model is recast as a probabilistic graphical model, and evaluated in relation to a previously considered data set. In each case, it is shown that Bayesian inference is able to provide answers to important theoretical and empirical questions easily and coherently. The generality and potential of the Bayesian approach to understanding models and data in cognitive psychology is discussed.

Introduction

Psychology as an empirical science progresses through the development of formal models incorporating theoretical ideas, designed to explain and predict observations of psychological phenomena. This means that progress in psychology relies upon the quality and completeness of the methods it uses to relate models and data. There is little point developing theories and models on the one hand, and collecting data in the laboratory or the field on the other, if the two cannot be brought into contact in useful ways.

In most empirical sciences, Bayesian methods have been or are rapidly being adopted as the most complete and coherent available way to relate models and data. Psychology has long been aware of problems with traditional frequentist and null-hypothesis significance testing approaches to parameter estimation and model selection, and recognition of the Bayesian alternative has followed from a number of recent papers and special volumes addressing the general issues (e.g., Lee & Wagenmakers, 2005; Myung & Pitt, 1997; Myung, Forster, & Browne, 2000; Pitt, Myung, & Zhang, 2002). Beyond the illustrative applications provided in these general treatments, however, there are few worked examples of Bayesian methods being applied to models at the forefront of modern psychological theorizing. Perhaps one reason is that there has been too great a focus on model selection defined in a narrow sense—particularly through the evaluation of Bayes Factors—rather

than a full Bayesian analysis. The perception that all Bayesian methods have to offer for the evaluation of psychological models is a number that quantifies how much more likely one model is than another is dangerously limiting.

In this article, three previous cognitive modeling studies are revisited, in an attempt to demonstrate the generality and usefulness of the Bayesian approach. The three applications involve the Multidimensional Scaling representation of stimulus similarity (Shepard, 1962, 1980), the Generalized Context Model account of category learning (Nosofsky, 1984, 1986), and a Signal Detection Theory account of inductive and deductive reasoning (Heit & Rotello, 2005). These applications are chosen to span a range of cognitive phenomena, involve well-known and influential theories, and put a focus the ability of Bayesian methods to provide useful answers to important theoretical and empirical questions.

Metric Multidimensional Scaling

Theoretical Background

MDS representations of stimuli use a low-dimensional metric space in which points correspond to stimuli, and the distance between points models the (dis)similarity between stimuli (Shepard, 1957, 1962, 1987, 1994). Non-metric varieties of MDS algorithms for inferring these representations from pair-wise similarity data (e.g., Kruskal, 1964) make only weak assumptions about the form of the relationship between distance in the MDS space and stimulus similarity. However, Shepard's (1987) 'Universal Law of Generalization' provides a compelling case for similarity decaying exponentially with distance, at least for relatively low-level perceptual stimulus domains. We make this assumption[1], and so consider the form of metric MDS that uses an exponential decay function to relate distances to similarities.

A classic issue in all MDS modeling has involved the interpretation of different metric assumptions for the representational space. Typically, consideration is restricted to the Minkowskian family of distance metrics. For points $\mathbf{p}_i = (p_{i1}, \ldots, p_{iD})$ and $\mathbf{p}_j = (p_{j1}, \ldots, p_{jD})$ in a D-dimensional space, the Minkowski r-metric distance is given by

$$d_{ij} = \left[\sum_{x=1}^{K} |p_{ix} - p_{jx}|^r \right]^{1/r} . \tag{1}$$

The $r = 1$ (City-Block) and $r = 2$ (Euclidean) cases are usually associated with, respectively, so-called 'separable' and 'integral' stimulus domains (Garner, 1974; Shepard, 1991). The basic idea is that many stimulus domains, like different shapes of different sizes, have component dimensions that can be attended to separately. These are termed separable, and are well modeled by the distance metric that treats each dimension independently in accruing distance. Other stimulus domains, like color, however, have component dimensions that are 'fused', and not easily distinguished, and so the comparison of stimuli involves all of the dimensions simultaneously. These are termed integral, and are well modeled by the familiar Euclidean distance metric. In addition, metrics with $r < 1$ have been given a psychological

[1] We are aware of the argument that it is not clear the exponential decay relationship applies as well to direct judgment of (dis)similarity as it does to the conditional probabilities obtained from identification confusion or generalization experiments. In particular, non-metric analyses of direct judgment data often show a relationship that is more nearly linear.

justification (e.g., Gati & Tversky, 1982; Shepard, 1987, 1991) in terms of modeling stimuli with component dimensions that 'compete' for attention[2].

Despite the theoretical elegance of this framework for relating the Minkowskian metric family to core psychological properties of stimulus domains, there have been few attempts to infer r from similarity data using MDS modeling. Shepard (1991) presents a focused attack on the problem that gives a good account of the capabilities and pitfalls using standard methods. The basic approach is a brute-force one of applying standard non-metric MDS algorithms assuming a large number of different r values, and comparing the solutions on the basis of a measure of goodness-of-fit.

Besides the set of computational problems that are noted by Shepard (1991), which are severe enough to preclude even considering the theoretically interesting possibilities with $r < 1$, this approach suffers from failing to account for the functional form effects of model complexity inherent in varying the metric parameter. Since the value of r dictates how the coordinate location parameters interact, different values of r will certainly change the functional form of parametric interaction, and hence the complexity of the metric space representational model being considered. One of the great attractions of Bayesian inference is that, through its basis in a coherent and axiomatized probabilistic framework for inference, model complexity issues such as these are automatically handled in a principled way.

Graphical Model for MDS

All of the Bayesian models in this paper rely on posterior sampling from graphical models (see Griffiths, Kemp, & Tenenbaum, in press; Jordan, 2004, for psychological and statistical introductions, respectively). In these models, nodes represent variables of interest, and the graph structure is used to indicate dependencies between the variables, with children depending on their parents. We use the conventions of representing continuous variables with circular nodes and discrete variables with square nodes, and unobserved variables without shading and observed variables with shading. For unobserved variables, we distinguish between stochastic variables with single borders and determinstic variables with double borders. We also use plate notation, enclosing with square boundaries subsets of the graph that have independent replications in the model.

Figure 1 presents a graphical model interpretation of metric multidimensional scaling. At the top is the coordinate representation of the points corresponding to stimuli. The p_{ix} node corresponds to the single coordinate value of the ith stimulus on the xth dimension, and the surrounding plates repeat these coordinates over the $i = 1, \ldots, N$ stimuli and $x = 1, \ldots, D$ dimensions. The node is shown as a single-bordered, without shading, and circular because each coordinate dimension is, respectively, stochastic, unknown, and continuous. Under the Bayesian approach, a prior distribution for these coordinate location parameters must be specified. We make the obvious prior assumption that all of the coordinates have equal prior probability of being anywhere in a sufficiently large (hyper)-cube with bounds $(-\delta, +\delta)$,

$$p_{ix} \quad \sim \quad \text{Uniform}\big(-\delta, \delta\big); \quad \delta > 0, \tag{2}$$

[2]These interpretations of $0 < r \leq 2$ also map naturally onto Shepard's (1987) framework, corresponding to the full possible range of -1 to +1 correlations for the 'consequential regions' that underpin his theoretical results. In contrast, other than for the supremum metric, it is difficult to give psychological meaning to metrics with $r > 2$, and so we restrict our attention to $0 < r \leq 2$.

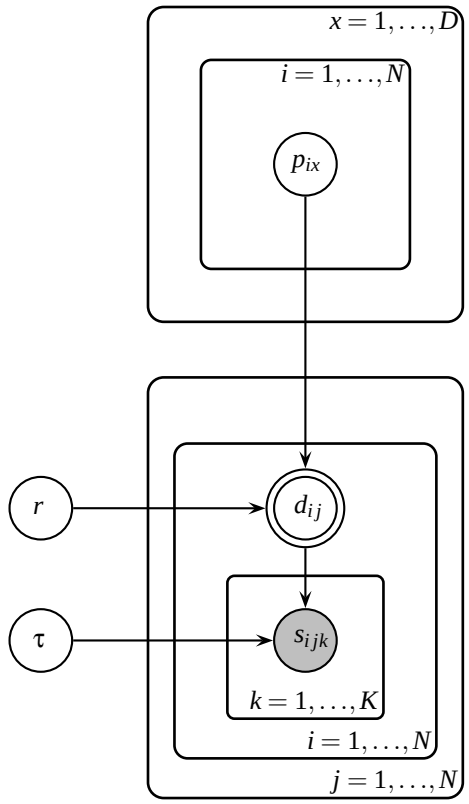

Figure 1. Graphical model for metric multidimensional scaling.

where "sufficiently large" means large enough that increasing δ does not alter the posterior distribution over the coordinate point parameters.

The metric parameter r that is the focus of this application is also a stochastic, unobserved, and continuous node. Following the earlier discussion, a prior distribution is used that is uniform over the theoretically-interpretable interval between zero and two:

$$r \sim \text{Uniform}\big(0, 2\big). \tag{3}$$

Given the value of r, and the coordinate locations p_{ix}, the pairwise distances d_{ij} are automatically given by Equation 1. In the graphical model in Figure 1, this means the d_{ij} node is double-bordered, to indicate it is deterministic, and has as parents the r and p_{ix} nodes. The d_{ij} node is encompassed by two plates, $i = 1, \ldots, N$ and $j = 1, \ldots, N$ to express the repetition over all pairs of N stimuli.

The similarity data considered here provide similarity ratings for each pair of stimuli as generated independently by K participants. The observed similarity between the ith and jth stimuli given by the kth participant is denoted s_{ijk}, and so is enclosed by an additional plate representing the $k = 1, \ldots, K$ participants. These similarities are assumed to be generated as the exponential decay of the distance between these points, but subject to noise, and so are stochastic, observed, and continuous. The noise process is assumed to be

a zero-mean Gaussian with common variance across all participants and stimulus pairs. The precision (i.e., the reciprocal of the variance) is represented by the stochastic, unobserved, and continuous parameter τ, so that

$$s_{ijk} \sim \text{Gaussian}\left(\exp\left(-d_{ij}\right), \tau\right), \tag{4}$$

with the standard (see Spiegelhalter, Thomas, Best, & Gilks, 1996) near non-informative prior distribution for the precision

$$\tau \sim \text{Gamma}\left(\varepsilon, \varepsilon\right), \tag{5}$$

where $\varepsilon = .001$ is set near zero[3].

Posterior Inference in Graphical Models

The graphical model in Figure 1 defines a precise and complete probabilistic relationship between the MDS parameters—the coordinate locations of stimulus points, and the metric parameter—and the observed similarity data. Bayesian inference uses this relationship to update what is known about the parameters, converting prior distributions to posterior distributions on the basis of the evidence provided by data.

The graphical model is a generative one, specifying how stimulus points and a metric combine to produce similarity data. Once similarity data are observed, inference is the conceptually easy process of reversing the generative process, and working out what stimulus points and metric are likely to have produced the data. The posterior probability distribution represents this information, specifying the relative probability of each possible combination of stimulus points and metric being the ones that generated the data.

Although conceptually straightforward, for most interesting cognitive models it will not be possible to find the posterior distribution analytically, and it is also unlikely that standard approximations will be very useful. Modern Bayesian inference for complicated models proceeds computationally, by drawing samples from the posterior distribution. We implement our graphical models using WinBUGS (Spiegelhalter, Thomas, & Best, 2004), which uses a range of Markov Chain Monte Carlo computational methods, including adaptive rejection sampling, splice sampling, and Metropolis-Hastings (see, for example Chen, Shao, & Ibrahim, 2000; Gilks, Richardson, & Spiegelhalter, 1996; Mackay, 2003) to perform posterior sampling.

For the MDS model in Figure 1, each posterior sample lists values for the unobserved variables

$$\left(r, \tau, p_{11}, \ldots, p_{ND}, d_{11}, \ldots, d_{ND}\right). \tag{6}$$

The basic principle of posterior sampling is that, over a large number of samples, the relative frequency of a particular combination of parameter values appearing corresponds to the relative probability of those values in the posterior distribution. This correspondence allows the information that is conceptually in the exact joint posterior distribution to be accessed approximately by simple computations across the posterior samples. For example,

[3]Gelman (2006) points out that this prior distribution, although very widely used, is problematic for small variances. For all of the applications reported here, the inferred variances are large enough to avoid this difficulty.

a histogram of the sampled values of a variable approximates its marginal posterior distribution, and the arithmetic average over these values approximates its expected posterior value. Considering the sampled values of one variable, for only those samples where another variable takes a specific value, corresponds to considering a conditional probability. Considering the combination of values taken by two or more variables corresponds to considering their joint distribution, and so on.

Inference from MDS Data

Our MDS applications consider three sets of individual-participant similarity data. Initial investigations with averaged data, of the type considered by Shepard (1991) showed clearly that the repeated-measures nature of individual participant data was important to make sound inferences about the metric structure of the representational space. This is consistent with results showing that averaging similarity data with individual differences can systematically affect the metric structure of MDS spaces (see Ashby, Maddox, & Lee, 1994; Lee & Pope, 2003). Only three data sets could be found for which raw individual participant data were available, and for which reasonable predictions about the separability or integrality of the stimulus domain could be made.

The first of these related to rectangles of different height with interior line segments in different positions, using eight of the possibilities in a 4×4 factorial design, as reported by Kruschke (1993). The second related to circles of different sizes with radial lines at different angles, following (essentially) a 3×3 factorial design, as reported by Treat, MacKay, and Nosofsky (1999). The third related to ten spectral colors, as reported by Helm (1959). Previous results would strongly suggest the first two of these domains are separable, while the colors are integral. Also on the basis of previous analyses (e.g., Lee, 2001; Shepard, 1962), and the explicit two-factor combinatorial designs for two of these stimulus domains, a two-dimensional representational space was assumed for all three stimulus domains.

For each data set, we calculated 5,000 such samples after a 1,000 sample 'burn-in' period (i.e., a period of sampling that is not recorded, but allows the Markov-chain to converge to sampling from the posterior distribution). We used multiple chains to check convergence, and observed a small proportion of these chains showing a degenerate behavior, with r becoming trapped near zero. While this behavior needs an explanation in the future, these chains were removed from the present analysis. Post-processing of the posterior samples for the coordinate location parameters was also required, to accommodate natural translation, reflection, and axis permutation invariances inherent in the MDS model. We achieved this by translating to center at the origin, reflecting where necessary so that both coordinate values for the first stimulus were positive, and permuting the axes where necessary so that the first coordinate value was larger than the second.

Figure 2 shows 50 randomly selected posterior samples for the stimulus points, displaying the representations that have been inferred from each data set. In Panel A, the eight stimuli are appropriately located within the 4×4 factorial structure. In Panel B, the nine stimuli follow the exhausted 3×3 factorial structure. In Panel C, the stimuli follow the standard 'color circle' representation. In each case, showing samples from the distribution also gives a natural visual representation of the uncertainty in the coordinate locations.

Figure 3 shows the posterior distribution over the metric parameter r for each of the three data sets. It is clear that the color stimulus domain, which is expected to be integral,

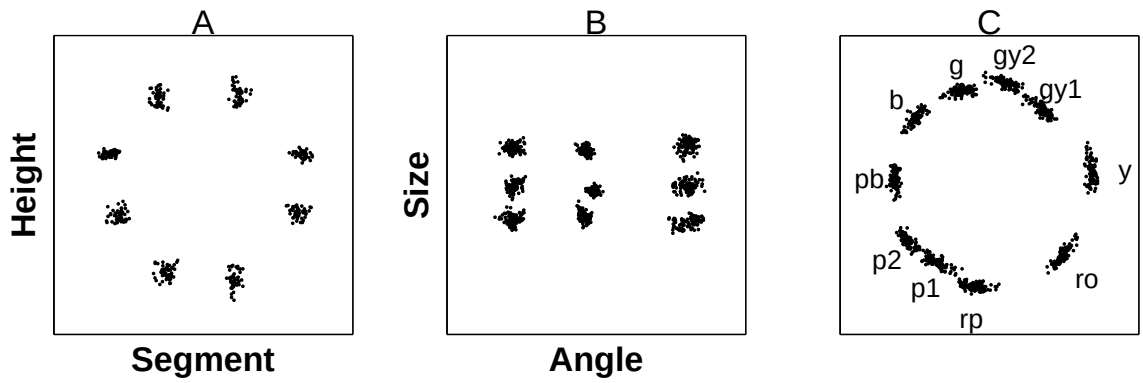

Figure 2. Multidimensional scaling representations for three stimulus domains—relating to (A) rectangles with interior lines, (B) circles with radial lines, and (C) spectral colors—showing samples from the posterior distributions for the representational points.

is distributed between about 1.6 and 2.0. It is not clear whether the mode is slightly below 2.0 because, consistent with previous theorizing, full integrality is not achieved, or as a consequence of the theoretically-driven restriction[4] that r not exceed 2.0. The posterior distribution of r for the radial lines and circles domain is centered about the value 1.0 associated with separability, as would be expected. The rectangle with interior lines stimulus domain has a posterior that lies a little below 1.0. One plausible interpretation of this, again consistent with previous theorizing (e.g., Gati & Tversky, 1982; Shepard, 1991), is that the rectangles and lines compete for attention, and constitute a 'highly separable' stimulus domain.

Summary

This application considered a Bayesian formulation of metric MDS modeling for similarity-based representation. The formulation was not intended to be definitive: It neglected issues of dimensionality determination, and made plausible, but contestable, assumptions about the form of the generalization gradient, and the distribution of empirical similarity. The current model also assumed there are no individual differences in the stimulus representations for different participants. All of these issues await fuller exploration within a Bayesian graphical model framework.

What the application does show, however, is that even with this simple formulation, a Bayesian approach automatically provides useful information not available in previous analyses. It provides a full posterior distribution over the location of the points representing stimuli, without the need to make distributional assumptions about these posteriors, as with many probabilistic MDS methods (e.g., Ramsay, 1982). Under the sampling approach to Bayesian inference, posterior distributions are not constrained to follow any particular distribution, but are free to take the form that follows from the specification of the model

[4]In an alternative analysis, r was given a prior that allowed values greater than 2.0. The posterior for r for the color data under these assumptions retained a mode below 2.0, but had some significant density extending to 2.0 and beyond.

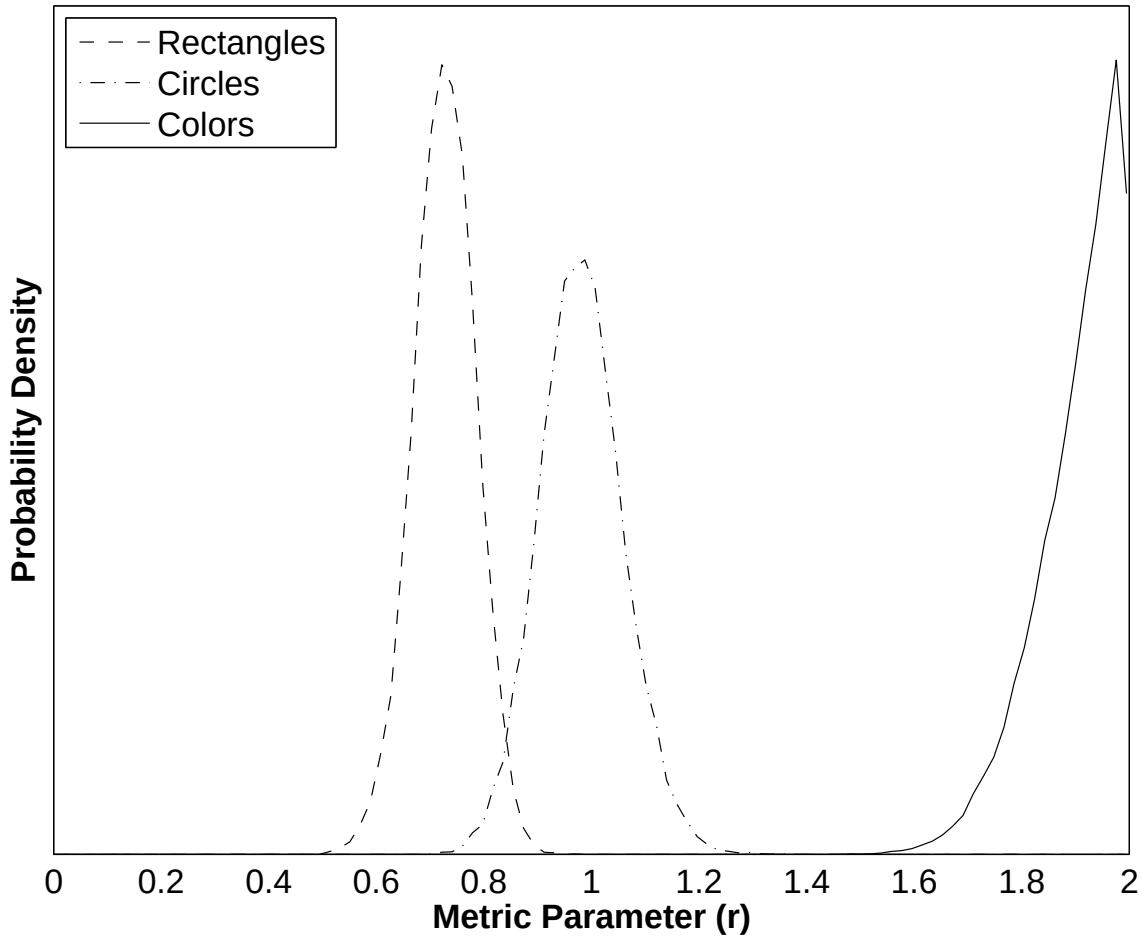

Figure 3. Posterior distributions over the metric parameter r for three stimulus domains.

and the information provided by data. The current application also provides a full posterior distribution over the parameter indexing the metric structure of the space. This posterior is sensitive to differences in the functional form complexity of parameter interaction, unlike the approach based on goodness-of-fit originally considered by Shepard (1991).

Category Learning

Theoretical Background

The Generalized Context Model (GCM) is a highly influential model of exemplar-based category learning (Nosofsky, 1984, 1986). The model assumes stimuli are represented as exemplars in memory according to a previously-derived MDS representation, which is subject to a selective attention process that weights the dimensions of the representation. The similarity between stimuli is modeled as an exponentially decaying function of distance in this transformed space, using a generalization gradient parameter. Category decisions are

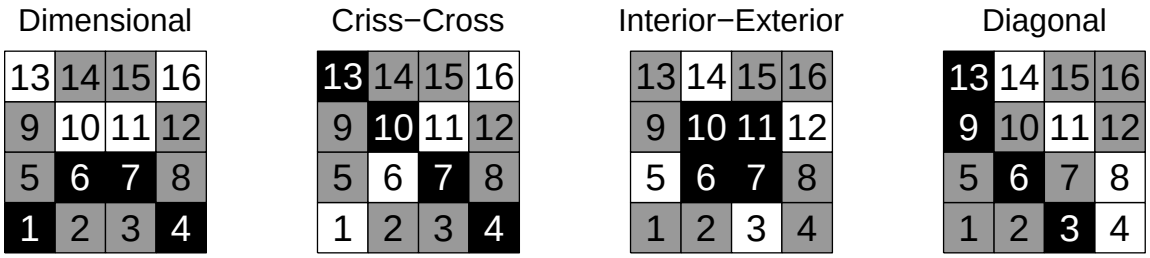

Figure 4. The four category structures used in the Nosofsky (1986) study. Based on Nosofsky (1986, Figure 3).

made probabilistically according to ratio of similarity between the presented stimulus and those in the different categories, using bias parameters that weight the different category responses.

Nosofsky (1986) presented a thorough and impressive study of the performance of the GCM on individual participant data in related identification and two-category learning tasks. Of particular interest are the categorization tasks, which involved four different two-category structures, termed Dimensional, Criss-Cross, Interior-Exterior, and Diagonal. These category structure are shown in Figure 4. The stimuli are arranged in a 4x4 grid, corresponding to their MDS representation. For each structure, the eight stimuli assigned to the two categories are shown as four black and four white squares in the grid, and the unassigned stimuli are shown as grey squares.

Of the many modeling issues Nosofsky (1986) addressed using these category structures, we focus on two. The first is an estimation issue, and relates to the values of the attention, generalization and bias parameters. In several places, theoretical questions are directly addressed by knowledge of the values that these parameters take, and Nosofsky (1986) reports standard tests of significance to decide, for example, if attention is equally distributed over the two components of the stimuli. The posterior distribution over these parameters obtained automatically by Bayesian analysis contains the relevant information for addressing these sorts of inferences.

The second issue is a model selection issue, and relates to the augmented version of the GCM proposed by Nosofsky (1986). In this augmented version, not only the stimuli presented in the category learning task are used in assessing similarity, but also the other stimuli, shown in grey in Figure 4, from the domain encountered in the earlier identification task. These additional stimuli are assumed to have a latent assignment to one of the categories. What inferences can be made about these assignment is not readily amenable to standard statistical testing, and so Nosofsky (1986) considered every possible pattern of latent assignment to draw conclusions. Whether the improved fit of this augmented GCM over the original version warrants the additional model complexity is also a difficult question to answer using standard methods. Nosofsky (1986, p. 48) acknowledged this difficulty, and argued for the appropriateness of the augmented model for just the Criss-Cross and Interior-Exterior category structures on the basis of unspecified "computer simulations".

Graphical Model

Figure 5 presents a graphical model interpretation of the augmented GCM, as applied to the two-dimensional stimulus domain in Nosofsky (1986). The x_i and θ nodes relate only to model comparison applications, and will be explained in that section. At the top of Figure 5 are the observed MDS coordinate locations for the $i = 1, \ldots, N$ stimuli in $x = 1, 2$ dimensions. The attention weight parameter gives the relative emphasis given to the first stimulus dimension over the second. This weight is given a uniform prior distribution over the interval between zero and one,

$$w \sim \text{Uniform}\big(0, 1\big). \tag{7}$$

The version of the GCM used in Nosofsky (1986) models similarity as an exponentially decaying function of the *squared* distance between the representative points. These squared distances are represented by the d_{ij}^2 node, which is deterministically defined in terms of the attention weight and coordinates,

$$d_{ij}^2 = w\left(p_{i1} - p_{j1}\right)^2 + (1 - w)\left(p_{i2} - p_{j2}\right)^2. \tag{8}$$

Given these squared distances, the generalization gradient parameter c determines the similarities between each pair of stimuli,

$$s_{ij} = \exp\Big(-\left(c d_{ij}\right)^2\Big). \tag{9}$$

The c parameter functions as an inverse scale (i.e., $1/c$ scales the distances), implying c^2 functions as a precision, and is given the standard near non-informative prior

$$c^2 \sim \text{Gamma}\big(\varepsilon, \varepsilon\big), \tag{10}$$

where $\varepsilon = .001$ is set near zero. Both the distances and similarities are repeated across all pairs of stimuli, and so are enclosed in two plates.

The probability of responding to the ith stimulus is determined by the similarities between the stimuli, the response bias, and the assignment of the stimuli to the categories. The response bias b is stochastic, unobserved and continuous, and is given a uniform prior distribution over the interval between zero and one,

$$b \sim \text{Uniform}\big(0, 1\big). \tag{11}$$

The assignment of stimuli to the two categories is determined by two indicator variables. The indicator variables a_j represent the known assignment of the jth presented stimulus, with $j = 1, \ldots, A$ ranging over the $N/2$ such stimuli in each category structure. The indicator variables z_j represent the latent assignment of the jth unassigned stimulus, ranging over the $N/2$ such stimuli in each category structure. The latent variables are stochastic, and are given a Bernoulli prior distribution with rate $1/2$, making the stimuli equally likely a priori to be assigned to each category,

$$z_j \sim \text{Bernoulli}\big(1/2\big). \tag{12}$$

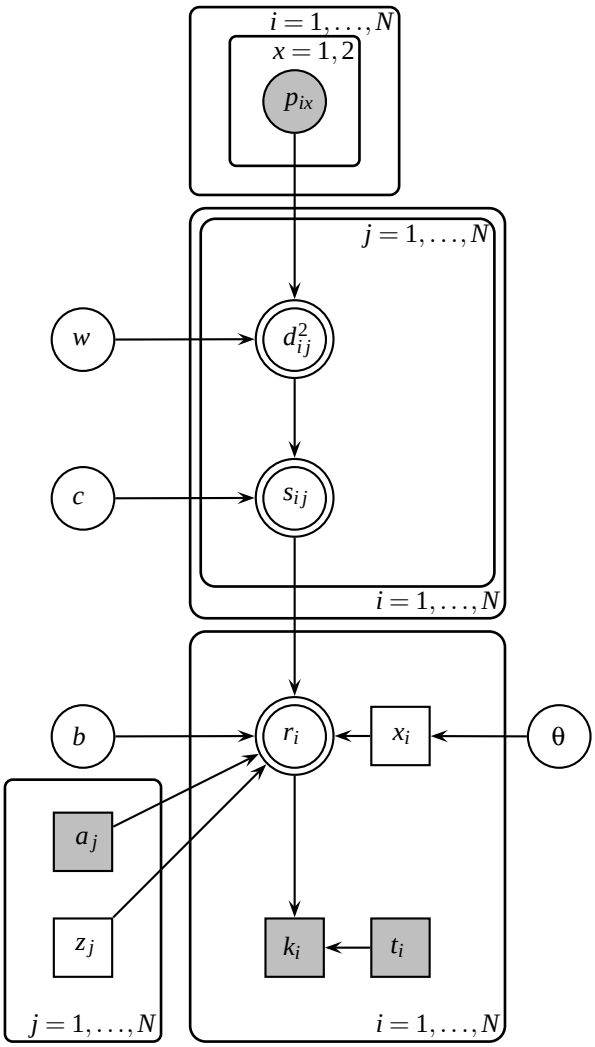

Figure 5. Graphical model for the augmented GCM.

From the similarities, bias and assignments, the response probability for the ith stimulus being chosen as a member of the first category ("Category A") is

$$r_i = \frac{b\left[\sum_{a\in A} s_{ia} + \sum_{z\in A} s_{iz}\right]}{b\left[\sum_{a\in A} s_{ia} + \sum_{z\in A} s_{iz}\right] + (1-b)\left[\sum_{a\in B} s_{ia} + \sum_{z\in B} s_{iz}\right]}. \tag{13}$$

The GCM uses these response probabilities to account for the observed data, which are the counts, k_i of the number of times the ith stimulus was chosen in Category A out of the t_i trials it was presented. Accordingly, the counts k_i follow a Binomial distribution

$$k_i \sim \text{Binomial}\big(r_i, t_i\big). \tag{14}$$

Inference from Data

For each of the four category learning data sets for Subject 1 from Nosofsky (1986), 100 chains were run collecting 1,000 posterior samples drawn after a burn-in of 1,000 samples. Each of the chains used a different random initial pattern of assignment. For two of the category structures—the Dimensional and Interior-Exterior ones—a single pattern of latent assignment was observed to dominate the posterior. These patterns are shown, together with the original category structures, in Figure 6, and match exactly those reported by Nosofsky (1986, Table 5).

Given the consistency in latent assignments, it is straightforward to interpret the posterior distributions for the attention, generalization and bias parameter for each category structure, as shown in Figure 7. These distributions are entirely consistent with the maximum-likelihood estimates reported by Nosofsky (1986, Table 5). The posterior distributions in Figure 7, however, carry useful additional information, since they provide a complete characterization of the uncertainty in knowledge of each parameter. It is clear, for example, that attention in the Interior-Exterior condition has significant density at the theoretically important value of 0.5. It is also clear that the Dimensional and Interior-Exterior bias and generalization parameters are very likely to be different, since their posterior distributions do not significantly overlap.

Finally, it is worth noting that the posterior of the attention parameter for the Dimensional condition shows how Bayesian methods naturally handle the theoretical restriction of their range. Frequentist confidence intervals based on asymptotic assumptions are unlikely to be suitable for inference in cases like these, and more difficult and ad hoc methods such as bootstrapping would probably be required. In contrast, the Bayesian approach handles the constraint automatically and naturally.

Across the 100 chains, the posterior samples for the Criss-Cross and Diagonal category structure revealed, respectively, two and four different patterns of latent stimulus

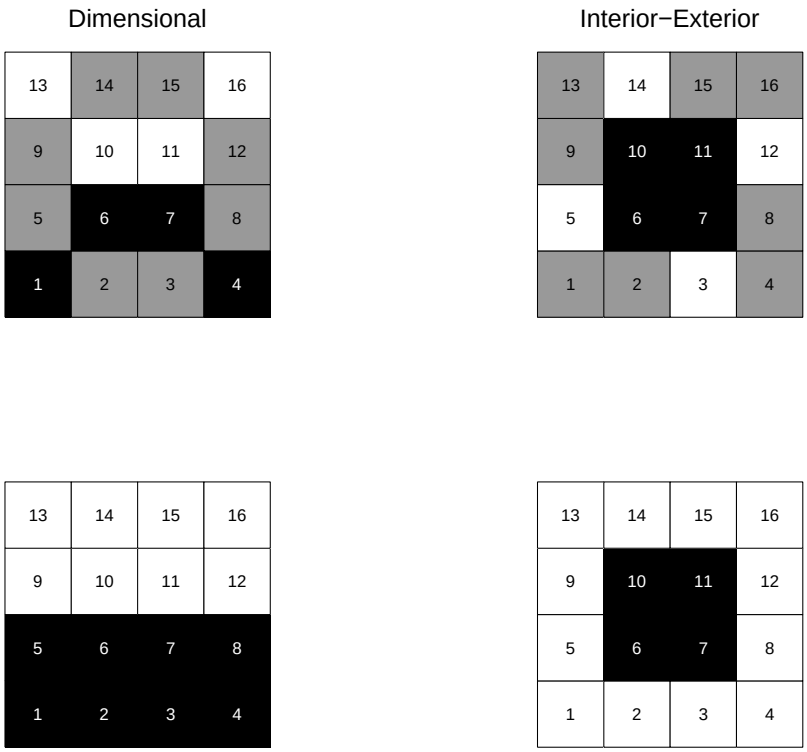

Figure 6. The augmented GCM latent assignments for the Dimensional and Interior-Exterior structures.

assignment[5]. These latent assignments are shown in Figure 8.

Only the latent assignments CI, DII and DIV were identified by Nosofsky (1986). Considering the additional latent assignments for the Diagonal structure is particularly satisfying, because the total four patterns exhaust the possibilities for stimuli 7 and 10. In this way, the two newly found assignments complement and complete those already established. Of course, it is possible by lowing the threshold of goodness-of-fit used to find latent assignments, (Nosofsky, 1986) could also have found these new assignments. But it is important to understand that, unlike the Bayesian results reported here, such an analysis would not be sensitive to differences in the complexity of different assignments. Formally, Nosofsky (1986) considered patterns of latent assignment based on their *maximum* likelihood, $p\left(z \mid w^*, , c^*, b^*, D\right)$, whereas the posterior samples in Figure 8 come from the *marginal* likelihood $p\left(z \mid D\right)$, where D is the category learning data. Only the marginal density accounts for model complexity, because it considers how likely a category representation is averaged

[5]In this case, it would be have been highly desirable to have observed mixing within the chains, rather than relying on random initial assignments for a large number of chains. In other words, what posterior sampling should ideally be able to produce is a single chain that includes all of the patterns of latent assignment in proportion to their posterior density. What we have done is combine many different chains to approximate this output. Nevertheless, the patterns of latent assignments we observe are intuitively sensible, and we believe they can be used to make inferences. But it is obviously necessary to treat measures like the posterior densities of each latent assignment that could be derived from this analysis with caution.

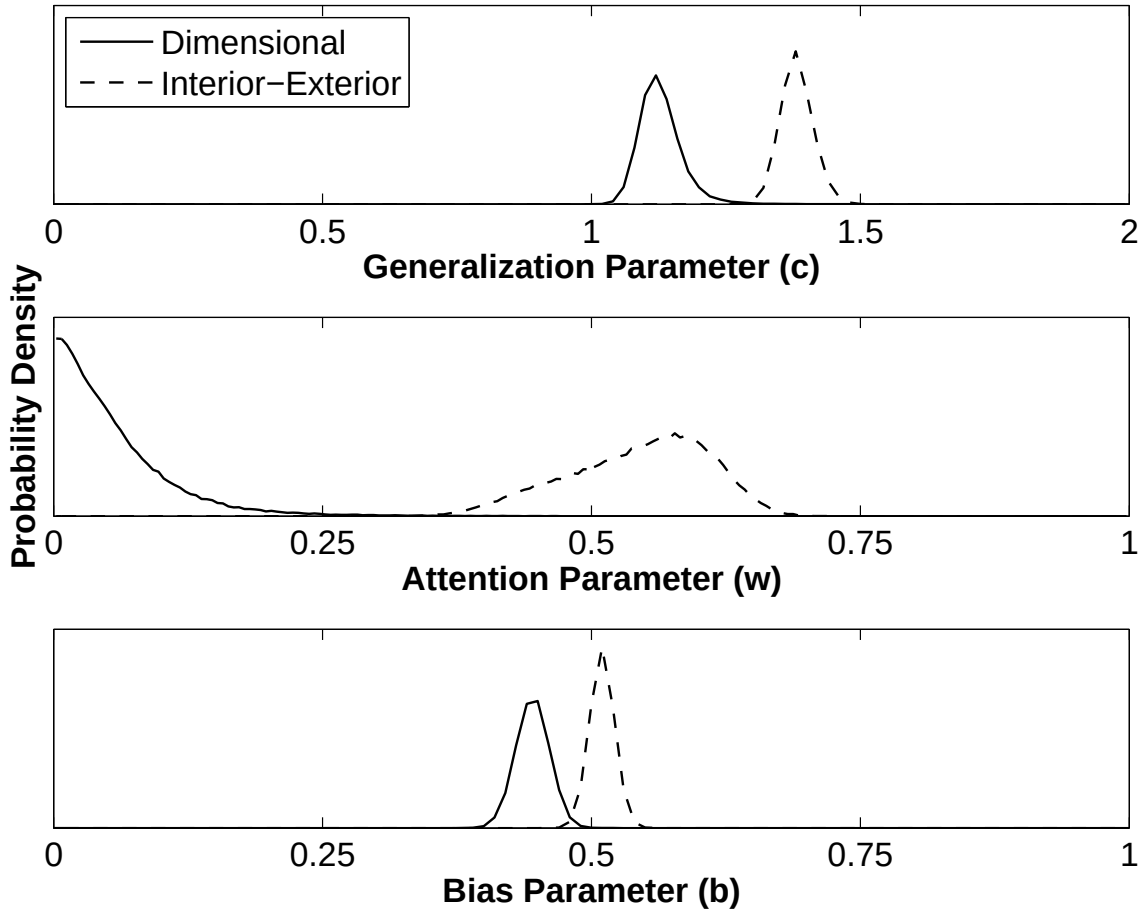

Figure 7. Posterior distributions for the augmented GCM parameters for three category learning structures.

across all of the different possible values for attention, generalization and bias.

Model Comparison

To address the model comparison issue of whether the additional complexity involved in allowing latent assignments in the augmented GCM is justified by the data, the x_i and θ nodes in Figure 5 are used. The x_i nodes are latent indicator variables for each of the i stimuli being categorized. These indicators determines whether or not the associated response probability r_i uses the latent stimulus assignments z_j as per the augmented GCM, or simply relies on the fixed assignments from the category learning task given by a_j.

Formally, this extension can be expressed by updating Equation 13 to

$$r_i = \begin{cases} \dfrac{b \sum_{a \in A} s_{ia}}{b \sum_{a \in A} s_{ia} + (1-b) \sum_{a \in B} s_{ia}} & \text{if } x_i \text{ is } 0 \\ \dfrac{b \left[\sum_{a \in A} s_{ia} + \sum_{z \in A} s_{iz}\right]}{b \left[\sum_{a \in A} s_{ia} + \sum_{z \in A} s_{iz}\right] + (1-b) \left[\sum_{a \in B} s_{ia} + \sum_{z \in B} s_{iz}\right]} & \text{if } x_i \text{ is } 1. \end{cases} \tag{15}$$

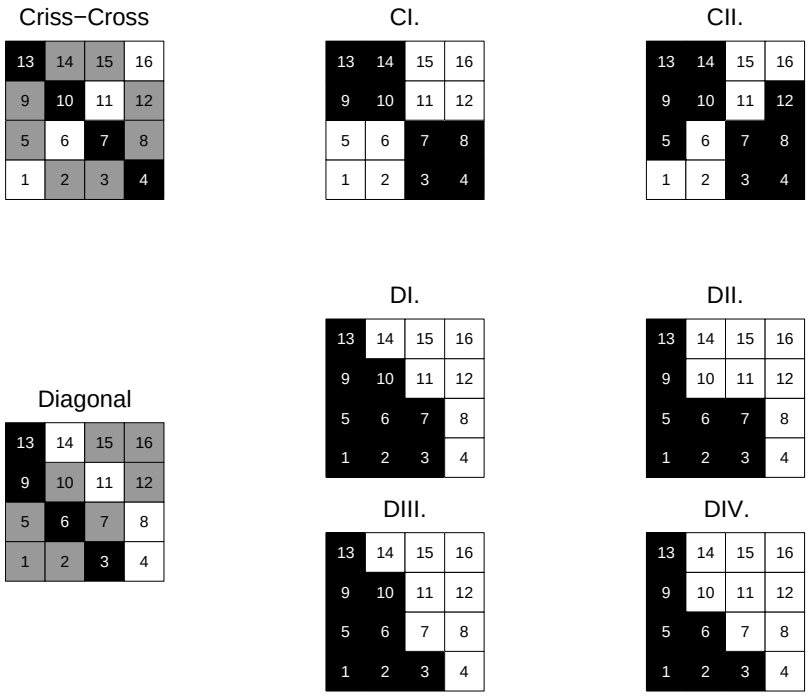

Figure 8. The augmented GCM latent assignments for the Criss-Cross and Diagonal category structures.

All of the indicators x_i are assumed to support the standard and augmented GCM accounts according to a fixed underlying rate of use, θ. This rate of use is given a uniform prior distribution

$$\theta \quad \sim \quad \text{Uniform}\,\big(0,1\big), \tag{16}$$

and its posterior provides a measure of the relative usefulness of the standard and augmented GCM accounts.

The posterior rate of use provides a measure of the relative importance of the two models in accounting for the way the participant categorized all of the stimuli. The nature of the measure is best understood by noting its relationship to the standard Bayes Factor (see Kass & Raftery, 1995). If θ were given a prior that only allowed the possibility that *every* stimulus was categorized by the standard GCM, or *every* stimulus was categorized by the augmented GCM, the posterior distribution would naturally allow the estimation of the Bayes Factor. That is, the Bayes Factor is a form of mixture estimation, when the only possible mixing rates are zero and one, because exactly one of the models is true. The assumption of a uniform prior employed here corresponds to allowing the possibility that neither model is exactly and exclusively true, but both might be useful, and the issue of relative merit is the issue of what mixture of standard to augmented GCM can be inferred from the data.

This is the information provided by the posteriors for rate of use shown in Figure 9 for the four category structures. It is clear that the augmented GCM is rarely used for the

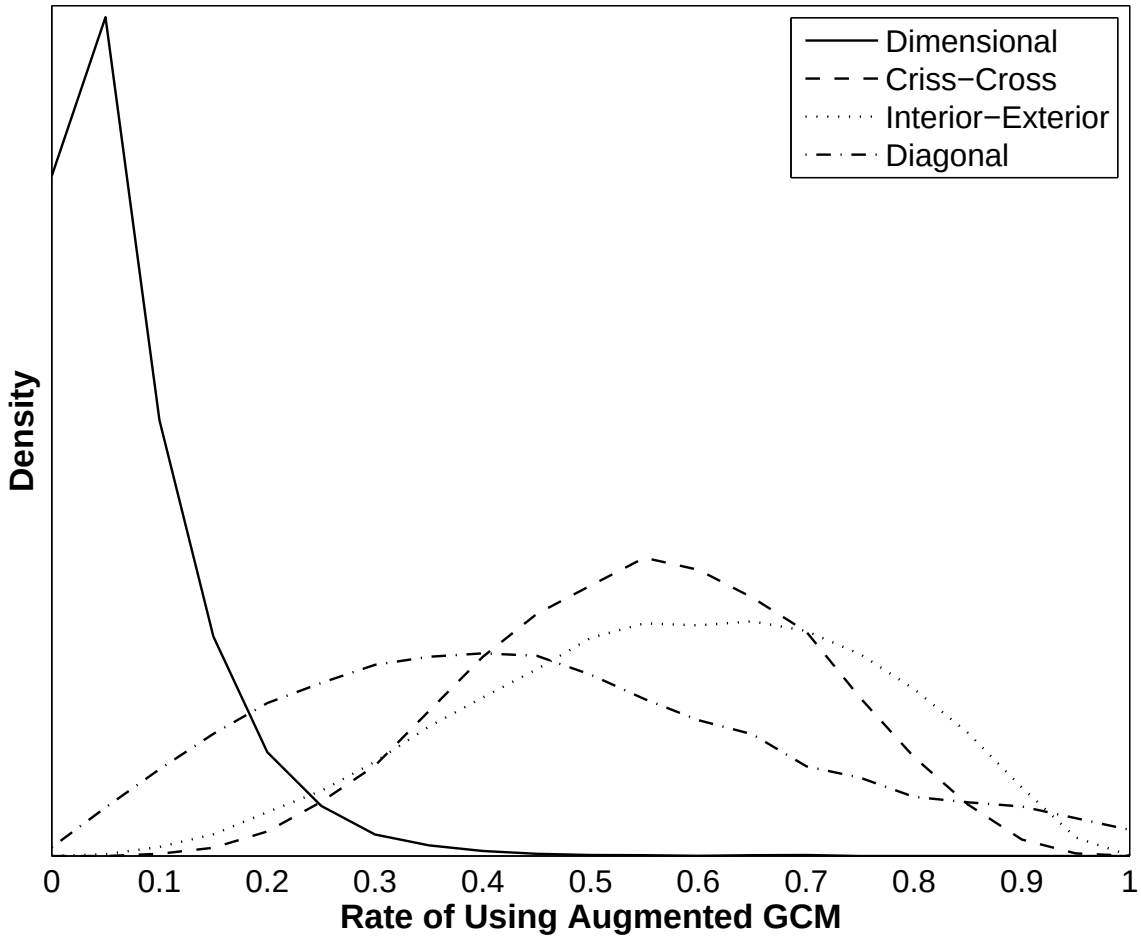

Figure 9. Posterior distribution of the rate at which stimuli are assigned to the augmented GCM rather than the standard GCM, for each of the four category structures.

Dimensional category structure, but is used significantly often for the other three structures, particularly in the case of the Criss-Cross and Interior-Exterior structures. In general, exactly what rate of use is required before a model is declared necessary, or superior to a competitor, is a question of the standards of scientific evidence needed, and must be made by researchers in each specific context. For the current application, we would conclude from Figure 9 that the augmented GCM is a useful and justified theoretical extension for all but the Dimensional category structure.

Summary

This application demonstrated the ability of Bayesian methods to improve both parameter estimation and model comparison for the GCM account of category learning. The parameter posterior distributions provide a complete representation of what is known and unknown about the psychological variables—selective attention, stimulus generalization and

response bias—used by the model to explain the observed category learning behavior. Under the posterior sampling approach to Bayesian inference, these distributions are again not constrained to follow any particular distribution, but are free to take the form that follows from the specification of the model and the information provided by data.

A different sort of parameter estimation is demonstrated by the patterns of latent assignment in Figure 6. The augmented version of the GCM involves an additional set of membership parameters, which indicate the assignment of untrained stimuli to the two categories. In contrast to the difficulties encountered by Nosofsky (1986) using standard methods, Bayesian inference applies exactly the same principles to estimating these discrete parameters as it does for the continuous attention, generalization and bias parameters. For two of the category structures, the Bayesian analysis found the same augmented assignments originally found by Nosofsky (1986), but for the Criss-Cross anad Diagonal structures it found additional and intuitively satisfying patterns of associating untrained stimuli with the categories, and made these inferences with sensitivity to model complexity.

Finally, this application shows how Bayesian inference can provide answers for a difficult model selection problem that was not addressed in any formal way by Nosofsky (1986). Using a mixture modeling approach to compare the standard and augmented GCM accounts, strong evidence was found for the additional complexity of the augmented account for three of the four category structures.

Signal Detection Model of Reasoning

Theoretical Background

Heit and Rotello (2005) present a clever model-based evaluation of the conjecture that inductive and deductive reasoning both involve the same single psychological dimension of 'argument strength' (Rips, 2001). Heit and Rotello (2005) use Signal Detection Theory (SDT: see Green & Swets, 1966; MacMillan & Creelman, 2004, for detailed treatments) to model this conjecture. The basic idea is to assume that the strength of an argument is unidimensional, but that different decision criteria control inductive and deductive reasoning. In particular, a relatively lesser criterion of argument strength is assumed to decide between 'weak' and 'strong' arguments for induction, while a relatively greater criterion decides between 'invalid' and 'valid' arguments for deduction. Under this conception, deduction is simply a more stringent form of induction. Accordingly, empirical evidence for or against the SDT model has strong implications for the many-threaded contemporary debate over the existence of different kinds of reasoning systems or processes (e.g., Chater & Oaksford, 2000; Heit, 2000; Parsons & Osherson, 2001; Sloman, 1998).

As empirical evidence to evaluate the SDT model, Heit and Rotello (2005) tested the inductive and deductive judgments of 80 participants on eight arguments. They used a between-subjects design, so that 40 subjects were asked induction questions about the arguments (i.e., whether or not the conclusion was "plausible"), while the other 40 participants were asked deduction questions (i.e., whether or not the conclusion was "necessarily true"). For each participant, there were four 'signal' questions, where the conclusions were plausible or necessarily true, and four 'noise' questions, where the conclusions were not plausible or necessarily true. Accordingly, the decisions made by participants have a natural characterization in terms of hit and false-alarm rates, which can then be converted to

standard measures of discriminability (or, synonymously, sensitivity) and bias using SDT.

One of key analyses of Heit and Rotello (2005) used standard significance testing to reject the null hypothesis that there was no difference between discriminability for induction and deduction conditions. Their analysis involved calculating the mean discriminabilities for each participant, using edge-corrections where perfect performance was observed. These sets of discriminabilities gave means of 0.93 for the induction condition and 1.68 for the deduction condition. By calculating via the t statistic, and so assuming associated Gaussian sampling distributions, and observing the p-value was less than 0.01, Heit and Rotello (2005) rejected the null hypothesis of equal means. According to Heit and Rotello (2005), this finding of different discriminabilities provided evidence against the criterion-shifting uni-dimensional account offered by SDT.

While the statistical inference methods used by Heit and Rotello (2005) are widely used and accepted, they explicitly or implicitly make a number of problematic assumptions that can be dealt with effectively using the Bayesian approach. First, the uncertainty about the discriminability of each individual is ignored, since it is represented by a single point estimate. Intuitively, making decisions corresponding, for example, to three hits and one false alarm is consistent, to varying degrees, with a range of possible hit and false-alarm rates, and hence, to varying degrees, with a range of discriminabilities. The Bayesian approach naturally represents this uncertainty by making prior assumptions about hit and false-alarm rates, and then using the evidence provided by the decisions to calculate posterior distributions. These posterior distributions are naturally mapped into posterior distributions for discriminability and bias according to SDT, which avoids the need for ad-hoc edge corrections.

In addition, and perhaps more importantly, the statistical analyses undertaken by Heit and Rotello (2005) implicitly assume there are no individual differences across participants within each condition. The mean discriminabilities for each group they calculate are based on the statistical assumption there is exactly one underlying point that generates the behavior of every participant in that group. That is, all of the individual participant data are used to estimate a single discriminability, with a standard error representing only the uncertainty about this single point. However, it seems psychologically implausible that there are not some individual differences in higher-order cognitive abilities like reasoning. Ideally, what ought to be estimated is a *distribution* of individual participant discriminabilities, with the parameters of this distribution becoming more certain as additional data become available. Bayesian methods naturally achieve this extension to accommodate individual differences using hierarchical models.

Graphical Model

Figure 10 shows a graphical model for a hierarchical version of signal detection theory that allows for individual differences in discriminability and bias across participants, and is very similar to that developed by Rouder and Lu (2005). The plate represents repetitions over participants. Within the plate, the graphical model shows the relationships for the ith participant between their discriminability d_i, bias c_i, hit rate h_i, and false-alarm rate f_i, and their observed counts of hit k_i^h and false-alarm k_i^f decisions.

The discriminability of each participant is assumed to be a unique value drawn from an over-arching Gaussian distribution with mean m_d and precision τ_d. Similarly, the bias

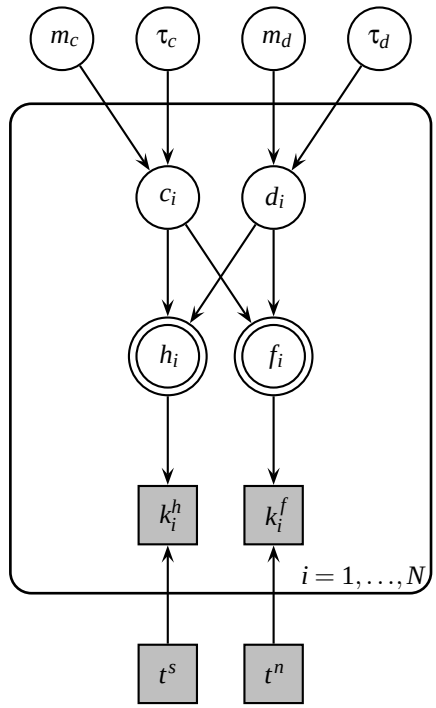

Figure 10. Graphical model for signal detection theory analysis allowing for Gaussian variation in discriminability and bias across participants.

of each participant is drawn from a Gaussian with mean m_c and precision τ_c. This means that

$$\begin{aligned} d_i &\sim \text{Gaussian}\left(m_d\, \tau_d\right) \\ c_i &\sim \text{Gaussian}\left(m_c, \tau_c\right). \end{aligned} \tag{17}$$

These over-arching Gaussians represent the individual differences in discriminability and bias over participants. Their mean and precision parameters are given standard near non-informative priors

$$\begin{aligned} m_d &\sim \text{Gaussian}\left(0, \varepsilon\right) \\ m_c &\sim \text{Gaussian}\left(0, \varepsilon\right) \\ \tau_d &\sim \text{Gamma}\left(\varepsilon, \varepsilon\right) \\ \tau_c &\sim \text{Gamma}\left(\varepsilon, \varepsilon\right), \end{aligned} \tag{18}$$

where $\varepsilon = .001$ is set near zero.

The discriminability and bias variables for each participant can be re-parameterized according to equal-variance SDT into hit and false-alarm rates, according to

$$\begin{aligned} h_i &= \Phi\big(\frac{1}{2} d_i - c_i\big) \\ f_i &= \Phi\big(-\frac{1}{2} d_i - c_i\big), \end{aligned} \tag{19}$$

where $\Phi(\cdot)$ is the standard cumulative Gaussian function. Finally, the counts of hit and false-alarm decisions follow a Binomial distribution with respect to the hit and false-alarm rates, and the number of 'signal' t^s (i.e., valid or strong) and 'noise' t^n (i.e., invalid or weak) arguments presented, so that

$$\begin{aligned} k_i^h &\sim \text{Binomial}\big(h_i, t^s\big) \\ k_i^f &\sim \text{Binomial}\big(f_i, t^n\big). \end{aligned} \tag{20}$$

Inference

The graphical model for SDT with individual differences was applied to both the induction and deduction condition data of Heit and Rotello (2005), drawing 100,000 samples after a burn-in period of 1,000 samples. One useful analysis of the full joint posterior distribution, concentrating on the group-level means of discriminability and bias for each condition, is shown in Figure 11. The main panel shows 500 random samples from the joint posterior of the means m_d and m_c, shown as circles for the induction condition, and crosses for the deduction condition. The side panels show the marginal distribution for each of these means.

Figure 11 shows that the two conditions have different patterns of mean discriminability and bias. In particular, the induction condition seems to have worse mean discriminability than the deduction condition. It is also clear that there is a large negative bias for the induction condition, indicating a tendency to over-respond 'strong', whereas the deduction condition shows little if any bias towards over-responding 'valid'.

Summary

Our conclusion from the Bayesian analyses is that, in complete agreement with Heit and Rotello (2005), it is important to allow discriminability in the induction condition to be different from that in the deduction condition. The contribution of the Bayesian analysis is that this conclusion has been reached, unlike the Heit and Rotello (2005) analysis, allowing for the possibility of individual differences in discriminability and bias across participants, and accommodating the clear limitations in how accurately hit and false-alarm rates can be estimated from only four observations per participant. In this way, the application demonstrates the ability of Bayesian methods to implement more realistic theoretical and methodological assumptions in drawing inferences from data.

Discussion

This paper aimed to demonstrate that Bayesian methods can be applied generally and usefully to aid the understanding and evaluation of psychological models. The three applications tried to span a range of cognitive models, and demonstrate a range of Bayesian analyses for addressing interesting theoretical questions informed by the available empirical data. In each case, the idea was to learn something useful through the Bayesian approach that would be difficult to achieve with the ways of relating models to data traditionally used in psychological modeling.

We concede it is probably the case that the collection of ad hoc methods dominating current practice could be enlarged further with specific methods to achieve the outcomes

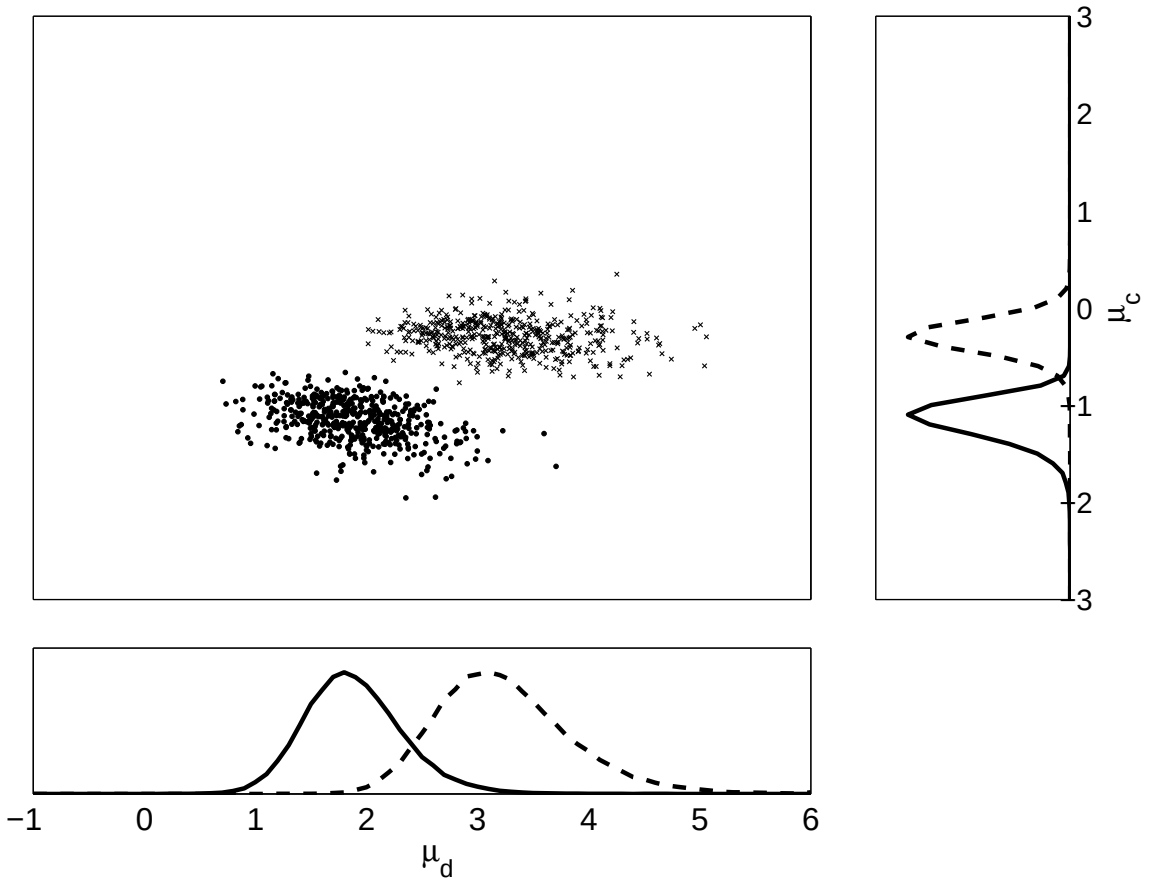

Figure 11. The main panel shows samples from the joint distribution of mean discriminability and mean bias, using circles for the induction condition, and crosses for the deduction condition. The side panels show the corresponding marginal distributions, using solid lines for the induction condition, and broken lines for the deduction condition.

reported here (after all, that is what "ad hoc" means). But the conceptual insight and technical skills needed to develop new methods stands in stark contrast to the conceptual simplicity and ease of implementation and analysis for the Bayesian graphical modeling approach.

To the extent our applications succeeded in encouraging the use of Bayesian methods, a number of obvious questions arise. One asks to what extent Bayesian methods can be applied to diverse types of cognitive models and cognitive modeling approaches. Another asks about the scalability of computational forms of Bayesian inference to large-scale models and data sets. A final question asks to what extent Bayesian inference is being used "just for data analysis", rather than as a model of human cognition. We attempt some tentative answers.

Generality of Bayesian Methods

One thing that is clear from the three applications presented is that Bayesian analysis can be possible for models not originally developed in Bayesian terms. But a natural question is how generally psychological models can be accommodated in the structured probabilistic framework needed for graphical model interpretation. Deterministic (or 'qualitative') models that do not have an error theory—algorithmic models of decision-making like Take-the-Best (Gigerenzer & Goldstein, 1996) constitute one prominent example—clearly need some additional assumptions before being amenable to probabilistic treatment. One promising source for providing principled error theories to such models are 'entropification' methods arising from the Minimum Description Length coding approach to model evaluation (Grünwald, 1998, 1999), which have already been applied successfully to a number of psychological models (e.g., Lee & Cummins, 2004; Lee, 2006).

Another class of psychological models that present a challenge are those that do not have a fixed set of parameters. Examples include the original version of the ALCOVE model (Kruschke, 1992) or the SUSTAIN model (Love, Medin, & Gureckis, 2004) of category learning, which specify processes for introducing additional representation nodes within a task, and so their parameter sets change as a function of the data. The mechanics of the applications presented here, with their reliance on WinBUGS and its associated standard Markov Chain Monte Carlo methods, do not extend automatically to these types of models. Instead, a more general Bayesian model-theoretic approach is required, using Bayesian non-parametric (also known as 'infinite dimensional') methods (Escobar & West, 1995; Ferguson, 1973; Ghosh & Ramamoorthi, 2003; Neal, 2000). Navarro, Griffiths, Steyvers, and Lee (2006) provide a general introduction to many Bayesian non-parametric ideas, and a specific application to modeling individual differences in psychology.

Scalability of Bayesian Methods

Bayesian model theoretics relies upon the full joint-posterior distribution over the model parameters as the basis for understanding and evaluating models against data. This is powerful, because the posterior represents everything that is known and unknown about the psychologically interesting variables represented by the parameters. Analytic power, however, comes with a computational burden, and it is reasonable to ask how well the approach scales to large models or data sets. There seem to be grounds for optimism on this front. The Topics Model (Griffiths & Steyvers, 2002, 2004) of language processing, for example, has been applied to a text corpus with about two million words, and hence has successfully made inferences from data about the joint distribution of about two million latent variables. Pioneering hierarchical and generative Bayesian models in vision have also succeeded at impressive scales (e.g., Yuille & Kersten, 2006). The application of Bayesian model theoretics in other fields, such as biology and machine-learning, give successful examples at very large scales (e.g., Ridgeway & Madigan, 2003).

More generally, there is a natural tension in psychological modeling between building models and addressing data that have the scale and complexity needed to account for what goes on in the real world, and maintaining the ability to evaluate those models and data in rigorous ways. The advantage of the Bayesian approach to model theoretics is that it guarantees a principled relationship between models and data. The potential of the Bayesian

approach is that it will be able to accommodate progressively larger and more sophisticated models and data.

Bayesian Modeling Versus Data Analysis

It is possible to draw a distinction between two ways that Bayesian ideas can be applied to the modeling of human cognition. One is to assume that the mind solves the inference problems it faces in a Bayesian way. That is, a theoretical assumption is made that the mind does Bayesian inference. Good recent examples of this approach include models of concept and category learning (e.g., Anderson, 1991; Sanborn, Griffiths, & Navarro, 2006; Tenenbaum & Griffiths, 2001), and models of inductive inference and decision-making (e.g., Griffiths & Tenenbaum, 2005). These are impressive models, and have significantly increased our understanding of the basic abilities of human cognition they address.

The second way Bayesian ideas can improve our understanding of cognition is to use them to relate model to data, improving the ability to make inferences about parameters and models using the incomplete and uncertain information provided by empirical data. The applications in this paper are largely of this type, and are part of a more general enterprise that has addressed diverse areas from similarity modeling and structure learning (e.g., Navarro & Lee, 2004), response time distributions (Lee, Fuss, & Navarro, 2007; Rouder, Lu, Speckman, Sun, & Jiang, 2005), and individual differences (e.g. Lee & Webb, 2005; Navarro et al., 2006; Rouder & Lu, 2005).

While the distinction between 'Bayesian models of cognition' and 'Bayesian analyses of models of cognition' is an intuitively appealing and practically useful one, it can mask a number of important issues. One issue is that the mere act of analyzing a model from the Bayesian perspective almost always requires making additional theoretical assumptions, and so changes the model itself to some degree. Most obviously, this happens in specifying prior distributions for parameters, as in all of the applications presented here. Occasionally, existing theory will suggest a form for these priors, but more often the goal will be to specify priors that affect the posteriors following from data as little as possible. In either case, the introduction of priors makes new theoretical assumptions about the psychologically meaningful variables used by a model. In this sense, the adoption of Bayesian methods can never amount to 'just data analysis'.

In some cases, applying Bayesian methods can have more dramatic theoretical consequences. One example is the Rescorla-Wagner model of classical conditioning which, under a non-Bayesian treatment, does not predict backward blocking effects (Rescorla & Wagner, 1972). Historically, this failure has been remedied by making additional theoretical assumptions and augmenting the basic model (e.g., Van Hamme & Wasserman, 1994). Dayan and Kakade (2001) show, however, that a Bayesian treatment of the basic learning mechanism underlying the Rescorla-Wagner model automatically predicts backward blocking. A similar recent example is provided by Lee et al. (2007), who show that Bayesian posterior predictive distributions for a basic Weiner diffusion model of response time exhibit empirically observed cross-over effects. Previously, the inability of the basic model to show these effects using standard analyses motivated the introduction of several additional parameterized noise processes (e.g., Ratcliff & Rouder, 1998), significantly complicating the original model both theoretically and practically. In both of these cases, simple and intuitive models have been shown able to make key empirical predictions when uncertainty about parameter

values is handled in a coherent Bayesian way. In both cases, this lessens the appeal of the more complicated models that have been developed, and so the Bayesian analysis makes a strong contribution to the development of theory.

A final point is that Bayesian inference, by itself, will often not provide all of the ideas needed to model any significant part of human cognition, and so will often require additional theory to be applied. This means it will be rare to have a purely Bayesian model of some aspect of cognition. As Griffiths et al. (in press) argue:

> "Bayesian inference stipulates how rational learners should update their beliefs in the light of evidence. The principles behind Bayesian inference can be applied whenever we are making inferences from data, whether the hypotheses involved are discrete or continuous, or have one or more unspecified free parameters. However, developing probabilistic models that can capture the richness and complexity of human cognition requires going beyond these basic ideas."

Two good recent examples are models of feature induction and stimulus similarity (Kemp, Bernstein, & Tenenbaum, 2005; Kemp, Perfors, & Tenenbaum, 2004), and of sequential decision-making behavior (Lee, 2006). Both of these are hierarchical Bayesian models, and rely entirely on the Bayesian approach to statistical inference to relate model parameters to data. Both also apply Bayesian methods to model the mind where those ideas are available, using, for example, Bayes rule as an account of how information updates mental representations, and how model averaging combines different mental hypotheses. But both models need to introduce non-Bayesian components to address the full range of phenomena they aim to explain. Kemp et al. (2005) uses generative node-replacement graph grammars and diffusion processes over graphical structure to generate stimulus representations and model their relationships to one another. Lee (2006) relies on a simple finite state account for generating thresholds to guide decision-making. None of these theoretical mechanisms could be regarded as following directly and uniquely from Bayesian principles, although all are compatible. Accordingly, both models use Bayesian inference as theoretical accounts of the mind *and* as a method for analyzing data.

Conclusion

The underlying philosophy of Bayesian inference is to represent uncertainty about quantitites of interest using probability distributions, and then use relevant information to update these representations as it comes to hand, all within the coherent framework for inference offered by probability theory. These capabilities fit naturally with the goals of modeling in empirical sciences like psychology. Psychological variables and processes are given formal expression by parameters and models, and the rationale for collecting experimental data is to refine our understanding of those variables and processes.

In this paper, we have aimed to give worked examples applying Bayesian methods to cognitive models, showing how it provides the tools to make inferences about hard but important research questions. We have emphasized the use of graphical models and posterior sampling as easy and powerful methods to undertake Bayesian analyses. The adoption of Bayesian methods for analysis promises to improve the way cognitive models are related to data, maximizing what we can learn from our ongoing efforts to develop theoretical models and gather empirical information.

Acknowledgments

I thank Geoffrey Iverson, Simon Dennis, Charles Kemp, E.-J. Wagenmakers, Wolf Vanpaemel and Jared Smith, and an anonymous reviewer, and also Evan Heit and Teresa Treat for generously sharing their raw data.

References

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

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*(3), 144–151.

Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese*, *122*, 93–131.

Chen, M. H., Shao, Q. M., & Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Spinger-Verlag.

Dayan, P., & Kakade, S. (2001). Explaining away in weight spac. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (pp. 451–457). Cambridge, MA: MIT Press.

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

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

Garner, W. R. (1974). *The Processing of Information and Structure*. Potomac, MD: Erlbaum.

Gati, I., & Tversky, A. (1982). Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(2), 325–340.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–534.

Ghosh, J. K., & Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. New York, NY: Springer.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov Chain Monte Carlo in Practice*. Boca Raton (FL): Chapman & Hall/CRC.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (in press). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Modeling*. Cambridge University Press.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In W. G. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th annual conference of the cognitive science society* (pp. 381–386). Mahwah, NJ: Erlbaum.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*, 5228–5235.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354–384.

Grünwald, P. D. (1998). *The Minimum Description Length Principle and Reasoning Under Uncertainty.* University of Amsterdam: Institute for Logic, Language and Computation.

Grünwald, P. D. (1999). Viewing all models as 'probabilistic'. In *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT' 99).* Santa Cruz: ACM Press.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, *7*, 569–592.

Heit, E., & Rotello, C. (2005). Are there two kinds of reasoning? In B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 923–928). Mahwah, NJ: Erlbaum.

Helm, C. E. (1959). *A Multidimensional Ratio Scaling Analysis of Color Relations.* (Princeton, NJ: Princeton University and Educational Testing Service)

Jordan, M. I. (2004). Graphical models. *Statistical Science*, *19*, 140–155.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.

Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 720–725). Mahwah, NJ: Erlbaum.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3–36.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology*, *45*(1), 149–166.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 555-580.

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and "rational" models. *Psychonomic Bulletin & Review*, *11*(2), 343–352.

Lee, M. D., Fuss, I. G., & Navarro, D. J. (2007). A Bayesian approach to diffusion models of decision-making and response time. In B. Schölkopf, J. C. Platt, & T. Hofmann (Eds.), *Advances in Neural Information Processing Systems 19.* Cambridge, MA: MIT Press.

Lee, M. D., & Pope, K. J. (2003). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical Psychology*, *47*, 32–46.

Lee, M. D., & Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*(3), 662–668.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605–621.

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

Mackay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambirdge University Press.

MacMillan, N., & Creelman, C. D. (2004). *Detection theory: A user's guide (2nd ed.)*. Hillsdale, NJ: Erlbaum.

Myung, I. J., Forster, M., & Browne, M. W. (2000). A special issue on model selection. *Journal of Mathematical Psychology*, *44*, 1–2.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet 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*(6), 961–974.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet processes. *Journal of Computational and Graphical Statistics*, *9*, 619–629.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114.

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

Parsons, L. M., & Osherson, D. (2001). New evidence for distinct right and left brain systems for deductive and probabilistic reasoning. *Cerebral Cortex*, *11*, 9545–965.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472–491.

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society: Series A*, *145*(3), 285–312.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two–choice decisions. *Psychological Science*, *9*, 347–356.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcemen. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning ii* (pp. 64–99). Appleton-Century-Crofts.

Ridgeway, G., & Madigan, D. (2003). A sequential Monte Carlo methods for Bayesian analysis of massive datasets. *Data Mining and Knowledge Discovery*, *7*(3), 301–319.

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, *12*, 129–134.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *32*, 573–604.

Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*, 195-223.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 726–731). Mahwah, NJ: Erlbaum.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*, 390–398.

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

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In J. R. Pomerantz & G. L. Lockhead (Eds.), *The Perception of Structure: Essays in Honor of Wendell R. Garner* (pp. 53–71). Washington, DC: American Psychological Association.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review*, *1*(1), 2–28.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, *35*, 1–33.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (2004). *WinBUGS Version 1.4 User Manual*. Cambridge, UK: Medical Research Council Biostatistics Unit.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). *BUGS Examples Volume 1, Version 0.5*. Cambridge, UK: MRC Biostatistics Unit.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Treat, T. A., MacKay, D. B., & Nosofsky, R. M. (1999). *Probabilistic Scaling: Basic Research and Clinical Applications.* Paper presented at the meeting of the Society for Mathematical Psychology, Santa Cruz.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonrepresentation of compound stimulus elements. *Learning and Motivation*, *25*, 127–151.

Yuille, A. L., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.