Determining Informative Priors for Cognitive Models

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

Wolf Vanpaemel
Faculty of Psychology and Educational Sciences
University of Leuven

Abstract

The development of cognitive models involves the creative scientific formalization of assumptions, based on theory, observation, and other relevant information. In the Bayesian approach to implementing, testing, and using cognitive models, assumptions can influence both the likelihood function of the model, often corresponding to assumptions about psychological processes, *and* the prior distribution over model parameters, often corresponding to assumptions about the psychological variables that influence those processes. The specification of the prior is unique to the Bayesian context, and often causes consternation. Sometimes the concerns stem from philosophical objections, but more often practical difficulties with how priors should be determined are the stumbling block. We survey several sources of information that can help specify priors for cognitive models, discuss some of the methods by which this information can be formalized in a prior, and identify a number of benefits of including informative priors in cognitive modeling. Our discussion is based on three illustrative cognitive models, involving memory, categorization, and decision making.

Keywords: Bayesian statistics, cognitive modeling, informative prior distributions

Introduction

One way to think of cognitive modeling is as a natural extension of data analysis. Both involve developing, using, and testing formal models as accounts of brain and behavioral data. The key difference is the interpretation of the model parameters. Data analysis typically relies on a standard set of statistical models, especially Generalized Linear Models (GLMs) that form the foundations of regression and the analysis of variance. In these models, parameters have generic interpretations, like locations and scales. Cognitive models, in contrast, aim to afford more substantive interpretations. It is natural to interpret the parameters in cognitive process models as variables like memory capacities, attention weights, or learning rates.

For both data-analytic and cognitive models, the likelihood is the function that gives the probability of observed data for a given set of parameter values. For data-analytic models, these likelihoods typically follow from GLMs, as in regression and analysis of variance modeling. Cognitive models often use likelihoods designed to formalize assumptions about psychological processes, such as the encoding of a stimulus in memory, or the termination of search in decision making. Even when a cognitive model uses likelihood functions consistent with GLMs—for example, modeling choice probabilities as weighted linear combinations of stimulus attributes—it is natural to interpret the likelihood as corresponding to cognitive processes, because of the psychological interpretability of its parameters.

The more elaborate interpretation means that cognitive models aim to formalize and use richer information and assumptions than data-analytic models do. In the standard frequentist approach, assumptions can only be used to specify the likelihood, and, less commonly, the bounds of the parameter space. The Bayesian approach offers the additional possibility of expressing assumptions in the prior distribution over the parameters. These prior distributions are representations of the relative probability that a parameter—or more generally, sets of parameters—have specific values, and thus formalize what is known and unknown about the psychological variables the parameters represent.

Conceived in this way, priors are clearly an advantage of the Bayesian approach. They provide a way of formalizing available information and making theoretical assumptions, enabling the evaluation of the assumptions by empirical evidence, and applying what is learned to make more complete model-based inferences and predictions. Priors are often, however, maligned by those resistant to Bayesian methods

(e.g., Edwards, 1991; Trafimow, 2005), or lamented by those otherwise championing Bayesian methods (e.g., I. J. Myung & Pitt, 1997). These concerns are expressed in a number of ways, but almost always come down to indecision about how priors can be determined, and a discomfort with the fact that however the decision is made, the choice of priors will affect inference.

It would be non-sensical if modeling assumptions like priors did not affect inference, and, for this reason alone, it is easy to be dismissive of reservations about priors. A more constructive way to address the concern is to point out that developing likelihoods is just as challenging as developing priors, and inference is also sensitive to choices about likelihoods. Proposing models is a creative scientific act that, in a Bayesian approach, extends to include both priors and likelihoods. The sort of attitudes and practices modelers have in developing, justifying, and testing likelihoods should naturally carry over to priors. Leamer (1983, p.37) insightfully highlights that both the likelihood and the prior are assumptions, and that a perceived difference in their subjectivity may simply reflect the frequency of their use:

> "The difference between a fact and an opinion for purposes of decision making and inference is that when I use opinions, I get uncomfortable. I am not too uncomfortable with the opinion that error terms are normally distributed because most econometricians make use of that assumption. This observation has deluded me into thinking that the opinion that error terms are normal may be a fact, when I know deep inside that normal distributions are actually used only for convenience. In contrast, I am quite uncomfortable using a prior distribution, mostly I suspect because hardly anyone uses them. If convenient prior distributions were used as often as convenient sampling distributions, I suspect that I could be as easily deluded into thinking that prior distributions are facts as I have been into thinking that sampling distributions are facts."

Against this background, the goal of this paper is to discuss how informative priors can be developed for cognitive models so that they are reasonable and useful.[1] We identify several *sources* of information that can help specify priors for cognitive models, and then discuss some of the *methods* by which this information can be

[1] We restrict ourselves to informative priors for cognitive models. For a discussion of priors in data-analytic models, see, for example, Dienes (2011).

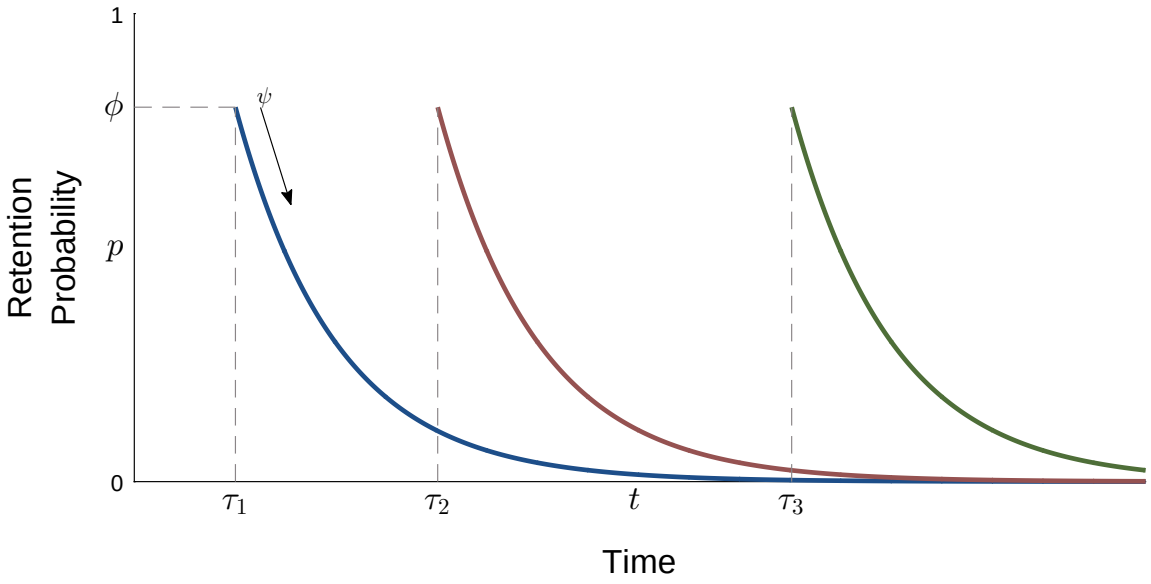

Figure 1. An exponential decay model of memory retention. The x-axis corresponds to time t, and the y-axis corresponds to the probability p that an item will be recalled at a specified time. Retention curves for three items are shown. Each curve starts at the time the item was last rehearsed, corresponding to the parameters τ_1, τ_2, and τ_3. The initial probability of recall at this time of last rehearsal is given by the parameter ϕ. The rate of decrease in the probability of recall as time progresses depends on a decay parameter ψ.

incorporated into formal priors within a model. Finally, we identify a number of *benefits* arising from including informative priors in cognitive models. We mostly rely on published examples of the use of priors in cognitive modeling, but also point to under-used sources and methods that we believe provide important future directions for the field.

Three illustrative cognitive models

To help make some general and abstract ideas clear, we draw repeatedly upon three illustrative cognitive models, involving memory, categorization, and decision making. In this section, we describe these models in some detail.

Exponential decay model of memory retention

A simple and standard model of memory retention assumes that the probability of recalling an item decays exponentially with time (Rubin & Wenzel, 1996). One way to formalize this model is to assume that the probability of recalling the ith

item at time t_i if it was last studied at time τ_i, is $p_i = \phi \exp\{-\psi(t_i - \tau_i)\}$. Figure 1 illustrates this model, showing the study times for three items, and the retention curves assumed by the model.

The ϕ parameter has the psychological interpretation of the initial probability of recall, that is, $\phi = p_i$ when $t_i = \tau_i$, while the ψ parameter controls the rate at which recall probabilities change over time. The parameter space is restricted to $\psi > 0$, so that the model formalizes the assumption of decay (e.g., Wickens, 1998). The usual assumption is that the τ_i time intervals are known from the experimental design, based on explicit study presentations, or that all $\tau_i = 0$ corresponding to the end of the study period. We consider a richer model in which the τ_i rehearsal times are treated as parameters, representing the last unobserved mental rehearsal of the item. This extension is made possible by the flexibility of Bayesian methods, and raises interesting questions about determining appropriate priors for the τ_i latent rehearsal parameters.

Generalized Context Model of categorization

The Generalized Context Model (GCM: Nosofsky, 1986) is a seminal model of categorization. It assumes that categorization behavior is based on comparing the attention-weighted similarity of a presented stimulus to known exemplars of the possible alternative categories. In the version of the GCM that we consider, stimuli are represented as points in a multidimensional space. In particular, the ith stimulus is represented by the coordinate location $\boldsymbol{x}_i$, so that the attention-weighted distance between the ith and jth stimuli is $d_{ij} = \sum_k \omega_k |x_{ik} - x_{jk}|$, where ω_k is the attention given to the kth dimension. Accordingly, a dimension receiving more attention will be more influential in determining distances than the one receiving less attention. The similarity between these stimuli is then $s_{ij} = \exp(-\lambda d_{ij})$, and the similarity of the ith stimulus to category A is the sum of the similarities to all the stimuli in the category: $s_{iA} = \sum_{j \in A} s_{ij}$. Finally, the probability of a category response placing the ith stimulus in category A is $p_{iA} = \beta_A s_{iA}^{\gamma} / \sum_C \beta_C s_{iC}^{\gamma}$, where the index C is across all possible categories, β_C is a response bias to category C, and γ controls the extent to which responding is deterministic or probabilistic.

The model is illustrated in Figure 2, for a task in which the stimuli can be represented with two dimensions. The stimulus space is shown on the left, with four exemplars from two categories, denoted by green circles and blue squares, being

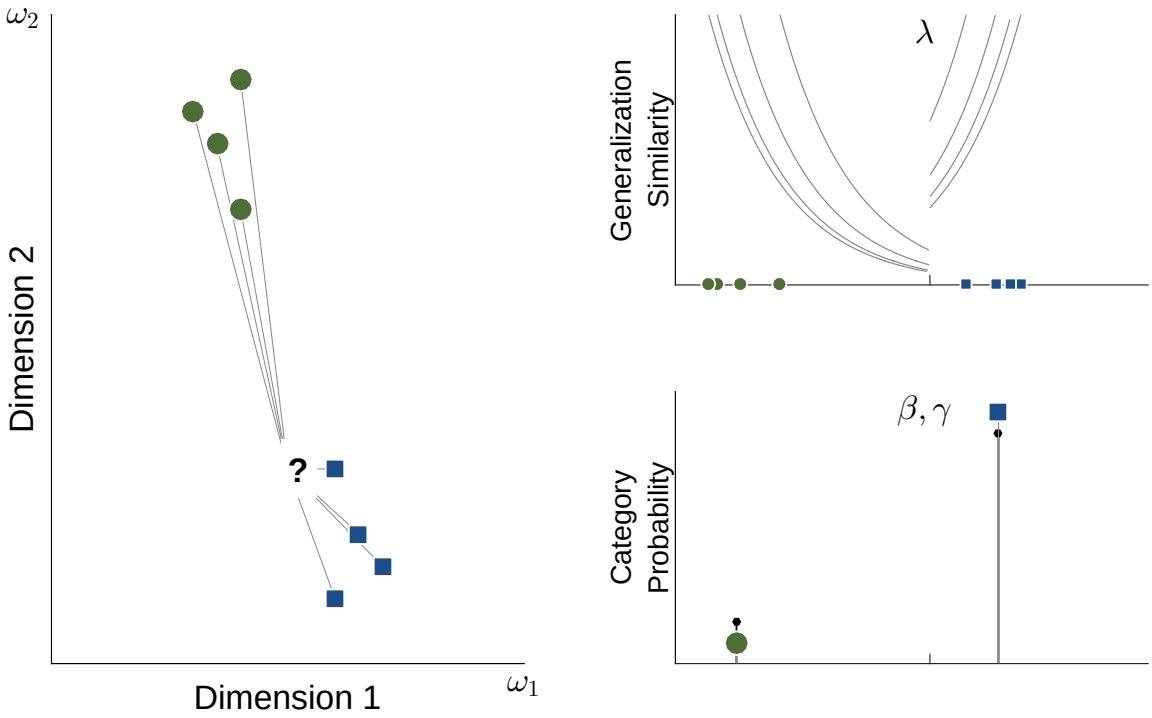

Figure 2. The Generalized Context Model of categorization. The left panel shows the attention-weighted dimensional representation of eight stimuli in a two-dimensional space. Four stimuli in one category are represented by green circles, and four stimuli in an alternative category are represented by blue squares. More attention is given to the second stimulus dimension than the first stimulus dimension, which "stretches" the space to emphasize differences between the stimuli on the second dimension. This attention process is controlled by parameters ω_1 and ω_2 that measure the attention given to each dimension. The upper-right panel shows the generalization gradients from the to-be-categorized stimulus, marked by "?", to the known stimuli. These gradients produce measures of similarity between the known and unknown stimuli, based on their distance in the space, and the steepness of the generalization gradient, controlled by a parameter λ. The bottom-right panel shows the final categorization probabilities of the unknown stimulus as either belonging to the green circle or blue square category. These final probabilities depend on the total similarities of the known stimuli associated with each category, shown by the small black cicles. These total similarities are then subject to a process controlled by a parameter γ that controls to the extent of deterministic versus probabilistic responding. A greater gain leads to more deterministic responding. Finally, these gain-modified simularities are subjected to a possible category-response bias, controlled by a parameter β. In the example, the γ and β parameters are set so there is some level of deterministic responding, but no bias, leading to the final category-response probabilities shown by the green circle and blue square in the bottom-right panel.

used to categorize an unknown stimulus, denoted by a "?". The ω_1 and ω_2 values correspond to the selective attention given to the first and second stimulus dimensions, respectively. The attention-weighted distances between stimuli are calculated in the "stretched" version of the space that applies the attention weights to the dimensions.

The generalization gradients for these attention-weighted distances from the presented stimulus to each exemplar are shown in the generalization panel. The λ parameter controls the generalization gradient, affecting how similar stimuli need to be for what is known about one to affect inferences about the other. In particular, smaller values of λ lead to shallower generalization gradients, so that more dissimilar stimuli influence categorization decisions, while larger values of λ lead to sharper generalization gradients, so that only the most similar stimuli affect decisions about the unknown stimulus.

The total similarity across all of the known stimuli to each of the categories is then calculated. These total similarities are then subjected to two more psychological processes. First, a process that affects the balance between deterministic versus probabilistic responding is applied, controlled by a parameter γ. Different values of γ correspond to different assumptions about how people treat the evidence provided by the total category similarity. When $\gamma = 1$, the similarities are preserved as originally measured, consistent with probability matching behavior. As γ increases above one, the larger similarities are magnified in importance, until eventually only the category with maximal similarity is chosen. As γ decreases below one, the category similarities are progressively made more equal, until eventually, when $\gamma = 0$, they are all identical, and each response becomes equally likely. In this way, the value of γ controls how deterministically categorization behavior follows the total similarities. Finally, the GCM allows for general tendency to prefer some categorization responses over others, controlled by a bias parameter β_k for the kth category. The overall categorization probabilities determined in this way are shown in the bottom-right panel of Figure 2.

Wiener diffusion model of decision making

Sequential sampling models of decision making assume that evidence is gathered from a stimulus over time until enough has been gathered to make a choice (Luce, 1986). The Weiner diffusion (Ratcliff & McKoon, 2008) model is a simple, but widely used, sequential sampling model for two-choice decisions. It assumes evidence takes the form of samples from a Gaussian distribution with mean ν. Total evidence starts

at θ and is summed until it reaches a lower bound of zero or an upper bound of α. The decision made corresponds to the boundary reached, and the response time is proportional to the number of samples, with the inclusion of an additive offset δ.

The decision model is shown Figure 3. The stimulus provides evidence favoring decision A, because the mean ν of the Gaussian characterizing the evidence is greater than zero. The decision and response times are shown by the histograms at each boundary. The shape of the histogram represents the response time distribution for each type of decision, and the area under each histogram represents the probability of each decision. It is clear that decision A is more likely, and both response time distributions have a characteristic non-monotonic shape with a long-tailed positive skew.

The ν parameter, usually called the drift rate, corresponds to the informativeness of the stimulus. Larger absolute values of ν correspond to stimuli that provide stronger evidence in favor of one or other of the decisions. Smaller absolute values of ν correspond to less informative stimuli, with $\nu = 0$ representing a stimulus that provides no overall information about which decision to make.

Figure 3 also shows a number of sample paths of evidence accumulation. All of the paths begin at the starting point θ, which is half-way between the boundaries at $\theta = \alpha/2$. Other starting points would favor one or other decision. The starting point parameter θ can theoretically be conceived as a bias in favor of one of the decisions. Such a bias could arise, psychologically, from prior evidence in favor of a decision, or as a way of incorporating utilities for correct and incorrect decisions of each type.

The α parameter, usually called boundary separation, corresponds to the caution used to make a decision, as manipulated, for example, by speed or accuracy instructions. Larger values of α lead to slower and more accurate decisions, while smaller values lead to faster but more error-prone decisions.

Finally, the offset δ corresponds to the component of the response time not accounted for by the sequential sampling process, such as the time taken to encode the stimulus and produce motor movements for a response. It is shown in Figure 3 as an offset at the beginning of the evidence sampling process, but could also be conceived as having two components, with an encoding part at the beginning, and a responding part at the end.

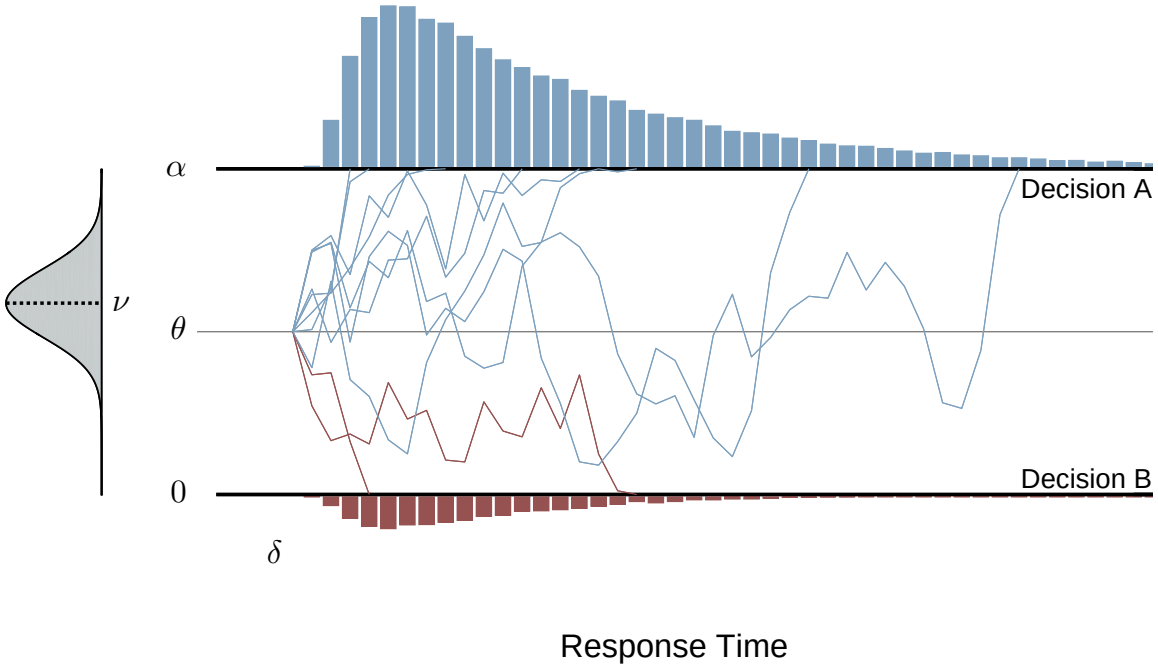

Figure 3. The Wiener diffusion model of decision making. A two-choice decision about a stimulus is made by sampling repeatedly from an evidence distribution for the stimulus, represented by a Gaussian distribution with mean ν. The samples are combined to form an evidence path, and a number of illustrative sample paths are shown. These paths start from an initial evidence value θ, and continue until they reach an upper bound of α or a lower bound of 0. The decision made corresponds to which boundary is reached. The response time corresponds to the number of samples collected, plus a constant δ representing the additional time needed to encode the stimuli and execute the response behavior. The decision and response time behavior is shown by the histograms above and below the decision boundaries. The histogram at each boundary is proportional to the response time distribution for that decision, and the area under each histogram represents the overall probability of that decision.

Sources for specifying priors

In this section, we identify several sources of information that can be used in determining priors, and explain their potential relevance in terms of the parameters of the three illustrative models.

Psychological and other scientific theory

The most important source of information for specifying priors in cognitive models is psychological theory. In cognitive modeling, likelihood functions are largely determined from theoretical assumptions about cognitive processes. The exponential decay memory retention model commits to the way in which information is lost over time, assuming, for example, that the rate of this loss is greatest immediately after information is acquired. The GCM commits to assumptions of exemplar representation, selective attention, and similarity comparisons in categorization. The decision model commits to the sequential sampling of information from a stimulus until a threshold level of evidence is reached. These assumptions are the cornerstones on which the likelihood functions of the models are founded. Analogously, theoretical assumptions about psychological variables should be the cornerstones on which priors are determined (Vanpaemel, 2009, 2010). Ideally, psychological theories should make assumptions about not just psychological processes, but also about the psychological variables that control those processes, leading to theory-informed priors.

One possibility is that theoretical assumptions dictate that some parameter values are impossible, consistent with the non-Bayesian restriction of the parameter space. In the memory retention model, the theoretical assumption that the probability of recall decreases over time constrains the memory retention parameter $\psi > 0$. In the categorization model, the theoretical assumption that generalization gradients decrease as stimuli become less similar, constrains the parameter $\lambda \geq 0$ (Nosofsky, 1986; Shepard, 1987).

Other sorts of theorizing can provide more elaborate information about possible combinations of values for a set of parameters. Theories of attention, for example, often assume it is a capacity-limited resource. In the GCM, this constraint is usually implemented as $\sum_k \omega_k = 1$, so that the values the attention parameters collectively meet a capacity bound. In effect, the theoretical assumption still dictates that some parameter values are impossible, but now the constraint applies jointly to a set of parameters.

As theories become more general and complete they can provide richer information. Theory can provide information beyond which values are possible, and indicate which values are probable. The optimal-attention hypothesis (Nosofsky, 1986) assumes that people distribute their attention near optimally in learning a category structure for a set of stimuli. This assumption implies that values of the ω_k parameters that maximally separate the stimuli in each category from each other are expected. For example, in Figure 2, the stimuli in the two different categories vary more along the second than first dimension. The optimal-attention hypothesis thus assumes that greater attention will be given to the second dimension than to the first dimension, so that $\omega_2 > \omega_1$. To reflect the expectation that the attention ω_2 will be greater, the prior placed on ω_1 should give significant density to values below $\frac{1}{2}$.

We believe that the optimality principle underlying the optimal-attention hypothesis extends to models other than the GCM. The principle that the most likely values of a parameter are those that maximize some aspect of behavioral performance seems generally applicable. Optimality could be a fundamental source for setting priors in cognitive process models, but is currently under-used. Embedding optimality principles within cognitive process models through priors will bring these models in closer contact with the successful rational models of cognition, where optimal behavior is a core theoretical assumption (e.g., Anderson, 1992; Chater, Tenenbaum, & Yuille, 2006; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

A different example of using theory to develop a prior is provided by Rouder, Morey, Speckman, and Pratte (2007), who propose a mass-at-chance model for performance in subliminal priming tasks. Their theoretical expectations are that some people will perform at chance, but others will use a threshold-based detection process to perform above chance. Rouder et al. (2007, see especially their Figure 3) consider different theoretical possibilities about the distribution of detection probabilities for people performing above chance. One possibility is that all detection probabilities are equally likely, so that it is constrained between $\frac{1}{2}$ and 1. Another possibility is that they are only slightly above chance, so that, for example, few people are expected to have a detection probability higher than (say) 70%. A third possibility is that people who are not above chance all have perfect accuracy, so that there are only two possible detection probabilities, $\frac{1}{2}$ and 1. Rouder et al. (2007) consider only the first two options to be reasonable, and express this theoretical assumption by constraining a variance parameter to be smaller than 1. In this way, Rouder et al. (2007)

establish a direct link between an exact range constraint on a variance parameter and substantive theoretical assumptions about the nature of people's performance in the task.

In some modeling situations, the likelihood can carry little theoretical content, and the theoretically most-relevant information is about the parameters. One example is provided by Lee (2015a), in a Bayesian implementation of a model originally developed by Hilbig and Moshagen (2014), for inferring which of a number of decision strategies people used in a cue-based decision-making task. The likelihood function is made up of simple binomial distributions, corresponding to how often an alternative is chosen for the trials within each decision type. Because different strategies predict different choice patterns, all of the important theoretical content is reflected in constraints on the choice parameters within the binomial distributions. For example, the new strategy introduced by Hilbig and Moshagen (2014) assumes an ordering for the probability of choice of different types of questions, and this information is represented by order constraints on the parameters corresponding to these probabilities in a joint prior. A similar earlier example in which the prior is theoretically more important than the likelihood is provided by J. I. Myung, Karabatsos, and Iverson (2005), who formalize several decision-making axioms, such as the monotonicity of joint receipt axiom and the stochastic transitivity axiom, using order constraints on the parameters representing probabilities.

Finally, we note that sciences other than psychology can and should provide relevant theoretical information. Physics, for example, provides the strong constraint—unless the controversial assumption of the existence of extra-sensory perception is made—that an item in a memory task cannot be rehearsed before it has been presented. This means, in the memory model, that each τ_i rehearsal parameter is constrained not to come before the actual time the item was presented. A good example of the potential relevance of multiple other scientific fields to determine priors is provided by the offset parameter δ in the decision model. Neurobiological and chemical processes, such as the time taken for physical stimulus information to transmit through the brain regions responsible for low-level visual processing, should constrain the component of this parameter that corresponds to the time needed to encode stimuli. Physiological theories specifying, for example, distributions of the speeds of sequences of motor movements, should constrain the component of the parameter that corresponds to the time taken to produce an overt response. Thus, a theoret-

ically meaningful prior for δ in the decision model could potentially be determined almost entirely by theories from scientific fields outside cognitive psychology.

Logic and invariances

The meaning of parameters can have logical implications for their prior distribution. Logic can dictate, for example, that some values of a parameter are impossible Taagepera (2007). Probabilities are logically constrained to be between 0 and 1, and variances and other scale parameters are constrained to be positive. In the memory, categorization, and decision models, the probability parameters ϕ, β, and θ are all logically constrained to be between 0 and 1.

The nature of a modeling problem can also provide logical constraints. The decision model has no meaning unless the starting point θ is between 0 and the boundary α, and has the same substantive interpretation under the transformation $(\alpha, \theta) \rightarrow (-\alpha, -\theta)$ that "flips" the boundary and starting point below zero. This invariance leads to the constraints $\alpha,\ \theta > 0$ and $0 < \theta < \alpha$ to make the model meaningful.

In general, superficial changes to a modeling problem that leave the basic problem unchanged should not affect inference, and priors must be consistent with this. In our memory and decision-making models, for example, inferences should not depend on whether time are measured in seconds of milliseconds, and the way priors over (ϕ, ψ, τ) and $(\alpha, \theta, \nu, \delta)$ are determined should lead to the same results regardless of the unit of measurement. This constraint is known as transformation invariance (Jaynes, 2003, Ch. 12; see also Lee & Wagenmakers, 2005). We think it is an important principle for determining priors, but it is difficult to find examples in cognitive modeling.

Previous data and modeling

Cognitive psychology has a long history as an empirical science, and has accumulated a wealth of behavioral data. Empirical regularities for basic cognitive phenomena are often well established. These regularities provide an accessible and substantial source of information for constructing priors. For example, response time distributions typically have a positive skew (e.g., Luce, 1986) and people often probability match in categorization, which means their probability of choosing each alternative is given by the relative evidence for that alternative (Shanks, Tunney, & McCarthy,

2002). This last observation is a good example of how empirical regularities can help determine a prior, and is applicable to the γ parameter in the categorization model. Different values of this parameter correspond to different assumptions about how people convert evidence for response alternatives into a single choice response. When $\gamma = 1$, decisions are made by probability matching. As γ increases above one, decision making become progressively more deterministic in choosing the alternative with the most evidence. As γ decreases below one, the evidence plays a lesser role in guiding the choice until, when $\gamma = 0$, choices are made at random. Thus, previous empirical findings that provide evidence as to whether people respond deterministically, probability match, and so on, can naturally provide useful information for determining a prior distribution over the γ parameter (e.g., Lee, Abramyan, & Shankle, 2015).

Cognitive psychology is also a model-based science, and so there are many reported applications of models to data. These efforts provide estimates or inferences about parameters that can inform the development of priors. For each of the memory, categorization, and decision models, there are many published relevant applications to data, including inferred parameter values (e.g., Nosofsky, 1991; Rubin & Wenzel, 1996; Ratcliff & Smith, 2004). The approach of relying on previous parameter inferences to determine priors for related models is becoming more frequent in cognitive modeling. Some recent examples include Gu et al. (2016) in psychophysics, Gershman (2016) for reinforcement learning models, Vincent (2015) in the context of temporal discounting, Wiehler, Bromberg, and Peters (2015) for different clinical sub-populations in the context of gambling, and Donkin, Tran, and Le Pelley (2015) in the context of a visual working memory model. In an interesting application of the latter model, Kary, Taylor, and Donkin (2015) defined vague priors for key parameters, and used the data from the first half of their participants to derive the posterior distributions. These posteriors were subsequently used as a basis for informative priors in the analysis of the data from the remaining half of the participants (see also Kruschke & Vanpaemel, 2015).

Elicitation

There is a reasonably well-developed literature on methods designed to elicit priors from people (e.g., Albert et al., 2012; Garthwaite, Kadane, & O'Hagan, 2005; Kadane & Wolfson, 1998; O'Hagan et al., 2006). These methods are used quite extensively in modeling in some empirical sciences, but do not seem to be used rou-

tinely in cognitive modeling. Elicitation methods are designed to collect judgments from people—often with a focus on experts—that allow inferences about a probability distribution over unknown quantities. The most common general approach involves asking for estimates of properties of the required distribution. These methods can be as simple as asking for a minimum and maximum possible value, or the bounds on (say) an 80% credible interval for an observed quantity. More intensive methods involve a larger number of percentiles, and may ask directly about latent parameters of interest, or about predicted observable quantities implied by values of those parameters. In these more complicated cases, a statistical model is needed to relate people's judgements to the desired probability distributions. Often people's estimates are accompanied with judgments of confidence, and this information is also incorporated into the inference about the distribution.

Another approach to elicitation used in applied settings require a series of judgments between discrete options, from which a probability distribution representing uncertainty can be derived (e.g., Welsh, Begg, Bratvold, & Lee, 2004). Along these lines one potentially useful recent development is the elicitation procedure known as iterated learning (Kalish, Griffiths, & Lewandowsky, 2007; Lewandowsky, Griffiths, & Kalish, 2009) This clever procedure requires a sequence of people to do a task, such as learning a category structure, or the functional relationship between variables. Each person's task depends on the answers provided by the previous person, in a way that successively amplifies the assumed common prior information, or inductive bias, people bring to the task. Applying this procedure to categorization, Canini, Griffiths, Vanpaemel, and Kalish (2014) found that learners have a strong bias for a linear category boundary on a single dimension, provided that such a dimension can be identified. Translating this observation to the ω_k parameters in the GCM implies that, in absence of any other information about category structures, these parameters are expected to be close to 0 or 1. It is a worthwhile topic for future research to find ways of formally translating this sort of information in a prior for a cognitive model.

Methods for specifying priors

The sources of information identified in the previous section are a pre-cursor to the complete formalization of a prior distribution. Knowing, for example, that some values of a parameter are theoretically impossible does not determine what distribution should be placed on the possible values. In the memory retention model, the

non-Bayesian parameter space constraint that $\psi > 0$ is not specific enough to enable the model to make detailed quantitative predictions. The Bayesian requirement of formalizing the prior distribution over ψ does lead to a model that makes predictions, consistent with the basic goals of modeling in the empirical sciences (Feynman, 1994, Chapter 7). In this section, we identify some methods for taking relevant information, and using it to construct a formal prior distribution.

Constraint satisfaction

If available information, whether by theoretical assumption, out of logical necessity, or from some other source, constrains parameter values, these constraints can be used as bounds. To determine the prior distribution within these bounds, the maximum-entropy principle provides a powerful and general approach (Jaynes, 2003, Ch. 11; Robert, 2007). Conceptually, the idea of maximum entropy is to specify a prior distribution that satisfies the constraints, but is otherwise as uninformative as possible. In other words, the idea is for the prior to capture the available information, but no more. Common applications of this approach in Bayesian cognitive modeling include setting uniform priors between 0 and 1 on probabilities, and setting a form of inverse-gamma prior on variances (see Gelman, 2006, for discussion).

The maximum-entropy principle is also often used in enforcing order constraints between parameters (e.g., Hoijtink, Klugkist, & Boelen, 2008; Lee, 2015a). A good example involves the τ_k rehearsal parameters in the memory model, if they are subject to the constraint indicating that an item cannot be rehearsed before it has been presented. Figure 4 shows the resultant joint prior on (τ_1, τ_2, τ_3) if the three study items are presented at times t_1, t_2, and t_3. Only rehearsal parameter combinations that are in the shaded cube have prior density. The uniformity of the prior in this region follows from the maximum-entropy principle. The justification for the prior is that it satisfies the known constraints about when the items could be rehearsed, but otherwise carries as little information as possible.

More general applications of the maximum-entropy principle are rare in the cognitive modeling literature. Vanpaemel and Lee (2012) present an example that is conceptually close, relating to setting the prior on the attention-weight parameter ω in the categorization model. The prior is assumed to be a beta distribution, and the optimal-attention hypothesis is used to set the mean of the prior to the value that best separates the stimuli from the different categories. The optimal-attention hypothesis,

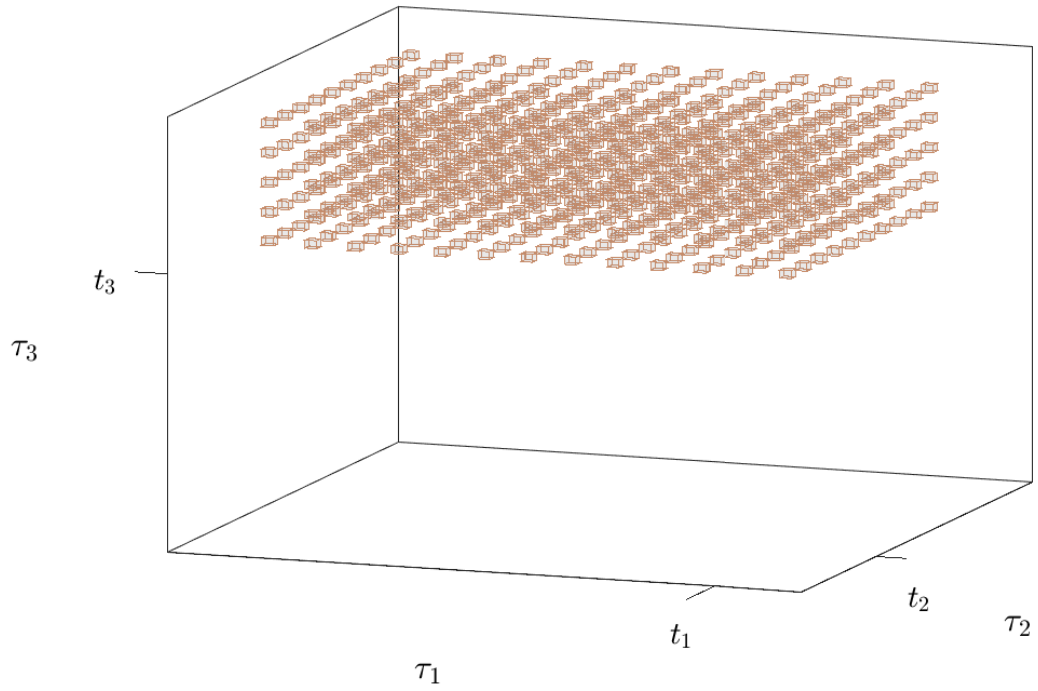

Figure 4. A prior specified by constraint satisfaction for the memory model. The three axes correspond to the last rehearsal times of three studied items, represented by the model parameters τ_1, τ_2, and τ_3. The case considered involves these items having been first presented at known times t_1, t_2, and t_3. The shaded region corresponds to the set of all possible rehearsal times (τ_1, τ_2, τ_3) that satisfy the logical constraint that an item can only be rehearsed after it is presented, so that $\tau_1 \geq t_1$, $\tau_2 \geq t_2$, and $\tau_2 \geq t_2$. The uniform distribution of prior probability within this constraint satisfaction region is justified by the maximum-entropy principle.

however, is not precise enough to determine an exact shape for the prior, but the precision of the beta distribution could have been determined in a more principled way by maximum-entropy methods. This would have improved on the heuristic approach actually used by Vanpaemel and Lee (2012) to set the precision. We think maximum-entropy methods are under-used, and that they are an approach cognitive modeling should adopt and develop, especially given the availability of general statistical results that relate known constraints to maximum-entropy distributions (e.g., Lisman & Van Zuylen, 1972).

Prior prediction

One of the clear benefits of priors is that it becomes possible to calculate prior predictive distributions, which are predictions about the relative probability of all possible data sets, based solely on modeling assumptions. If information is available about possible or plausible data patterns, most likely based on previously established empirical regularities, then one approach is to develop priors that lead to prior predictive distributions consistent with this information. A very similar approach is Parameter Space Partitioning (PSP: Pitt, Kim, Navarro, & Myung, 2006), which divides the entire parameter space into mutually exclusive regions that correspond to different qualitative data patterns a model can generate. Priors can then be determined by favoring those regions of the parameter space that generate data patterns consistent with expected patterns, and down-weighting or excluding regions corresponding to less plausible or implausible data patterns.

A closely-related approach involves considering the priors over psychologically meaningful components of a model that are implied by priors over their parameters. If information is available about the plausible form of these parts of models, most likely based on theory, it makes sense to define parameter priors that produce reasonable prior distributions for them. Figure 5 shows an example of this second approach using the decision model. Each combination of the starting point θ and offset δ parameters, which lie in the two-dimensional parameter space on the left, corresponds to a single joint decision probability and response time distributions for the two alternative choices shown on the right. Two different joint prior distributions over the parameters are considered. The first prior distribution, shown by red circles in parameter space, has a truncated Gaussian prior for θ with a mean of 0.5 and a standard deviation of 0.1 in the valid range $0 < \theta < 1$, and a truncated Gaussian prior for δ with a mean of 0.2 and a standard deviation of 0.05 in the valid ranges $\delta > 0$. The second prior, shown by the blue crosses, simply uses uniform priors on reasonable ranges for the parameters: $0 < \theta < 1$, and $0 < \delta < 0.4$,

The consequences of these different assumptions are clear from the corresponding distributions shown in the model space, which shows response time distributions generated by the decision models corresponding to both priors, for the same assumptions about boundary separation and the distribution of drift rates. The predictions of the decision model with the first prior distribution, shown by solid red lines, cover the sorts of possibilities that might be expected, in terms of their qualitative posi-

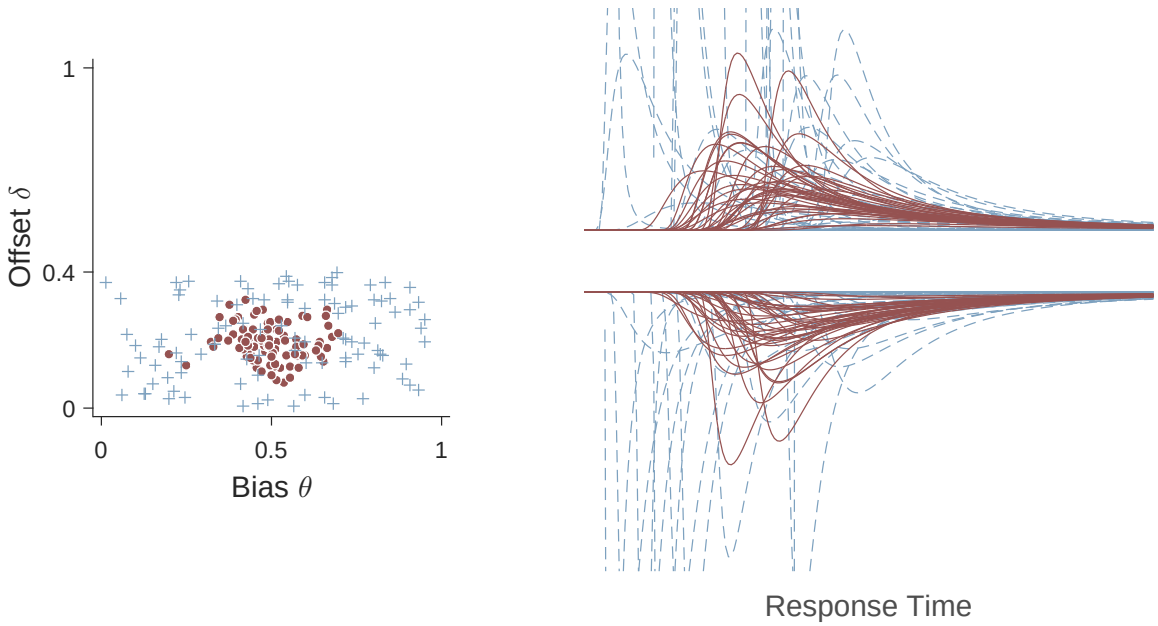

Figure 5. Developing prior distributions using prior prediction, for the decision model. The left panel shows the joint parameter space for the starting point θ and offset δ parameters. The right panel shows the joint decision and response-time distributions generated by the model. Two specific prior distributions are considered, represented by red circles and blue crosses in the parameter space, with corresponding solid red and broken blue lines in the model space. The prior represented by the red circles make stronger assumptions about both bias and offset, and predicts a more reasonable set of response time distribution than the prior represented by the blue crosses.

tion and shape. The predictions for the second prior distribution, shown by broken blue lines, however, are much less reasonable. Many of the predicted response-time distributions start too soon, and are too peaked. These weaknesses can be traced directly to the vague priors allowing starting points too close to the boundaries, and permitting very fast non-decision times. This analysis suggests that the sorts of assumptions about the starting point and offset made in forming the first prior may be good ones for the decision model. In this way, the relationship between prior distributions and psychologically interpretable components of the model provides a natural way to apply relevant knowledge in developing priors.

Using prior prediction to determine prior distributions in cognitive modeling is a general and relatively easy approach. Theorists often have clear expectations about model components like retention functions, generalization gradients, or the shapes of

response time distributions, as well as about the data patterns that will be observed in specific experiments, which can be examined in prior predictive distributions. While it is currently hard to find cognitive modeling examples of priors being developed by the examination of prior predictions (see Lee, 2015b; Lee & Danileiko, 2014, for exceptions), we expect this state of affairs will change quickly. One reason for this optimism is that prior predictions are slowly starting to appear in the cognitive modelling literature, with goals that are closely related to setting priors. For example, Kary et al. (2015) and Turner, Dennis, and Van Zandt (2013) examine the prior predictions of memory models, as a sanity check before application. In addition, prior predictive distributions have been used for assessing model complexity (Vanpaemel, 2009), for evaluating the falsifiability of a model, and for testing a model against empirical data (Vanpaemel, submitted).

Hierarchical modeling

An especially important method for developing priors in cognitive modeling involves extending the cognitive model itself. The basic idea is to extend the model so that priors on parameters are determined as the outcome of other parts of an extended model. This involves incorporating additional theoretical assumptions in to the model, and is naturally achieved by hierarchical or multi-level model structures (Lee, 2011; Vanpaemel, 2011). None of the illustrative memory, categorization, or decision models, as we presented them, have this property, which is representative of the field as a whole. The parameters in these models represent psychological variables that initiate a data generating process, and so priors must be placed explicitly on these parameters. The key insight of hierarchical modeling is that these psychological variables do not exist in isolation in a complete cognitive system, but can be conceived as the outcomes of other cognitive processes. Including those other processes within a more complete model thus naturally defines a prior for the original parameters.

An example of this approach is provided by Lee and Vanpaemel (2008), in their development of a hierarchical model of categorization. The model extends the GCM by allowing for different sorts of category representations, ranging from an exemplar representation in which every stimulus in each category is represented, to a prototype representation in which each category is represented by a single point. These possibilities are shown in the 7 bottom panels in Figure 6, for a case in which there are two categories with four stimuli each. The representation on the far left is

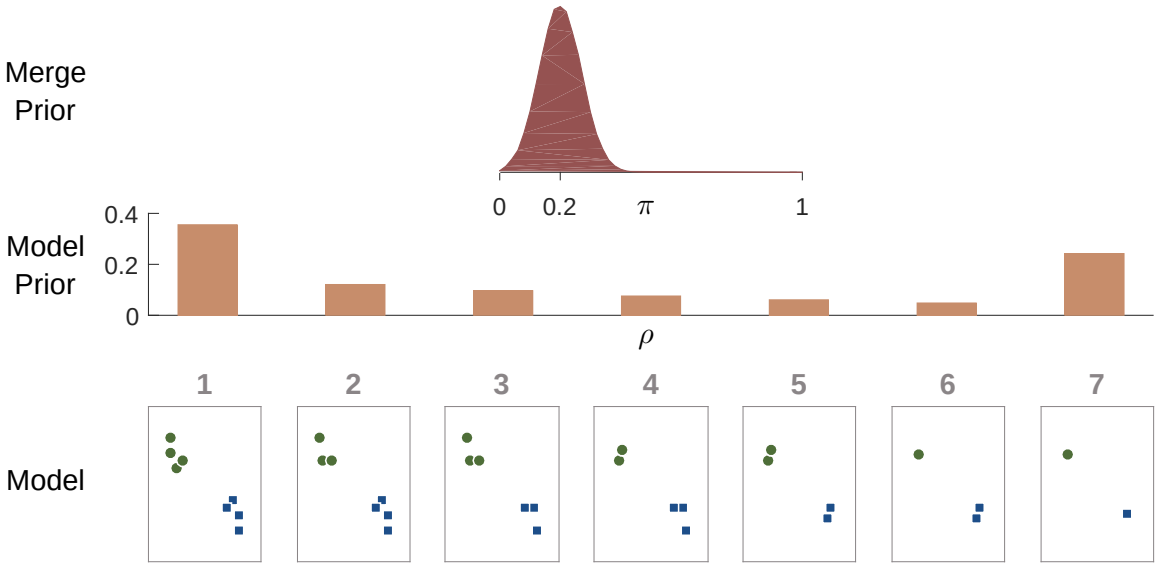

Figure 6. A hierarchical approach to determining a prior distribution for the representation index parameter ρ in an extended model of categorization. The top panel shows an assumed prior distribution over a parameter π that corresponds to the probability of merging a pair of stimuli in an exemplar representation. The bottom panels show a selection of 7 possible representations generated by this merging process, for a categorization problem with four stimuli in each of two categories, distinguished as green circles and blue squares. The full exemplar representation is shown on the left, the prototype representation is shown on the right, and some of the representations with intermediate levels of abstraction are shown between. The bar graph shows the prior probability on the representational index parameter ρ implied by the merging process and the prior distribution on π.

the exemplar representation, as assumed by the GCM, while the representation on the far right is the prototype representation. The intermediate representations show different levels of abstraction, as the detail of exemplar representation gives way to summary representations of the categories. In the non-hierarchical version of this model, called the Varying Abstraction Model (VAM: Vanpaemel & Storms, 2008), the inference about which representation is used is controlled by a discrete parameter ρ, which simply indexes which representation used. In the example in Figure 6, ρ is a number between 1 and 7, and requires a prior distribution that gives the prior probabilities that each of these 7 possibilities is the one used.

The hierarchical extension of the VAM is shown by the remainder of Figure 6. A new cognitive process is included in the model, which generates the different possible representations. This process begins with the full exemplar representation, but can

successively merge pairs of stimuli. At each stage, the probability of a merge is given by a new model parameter π. At each stage in the merging process, two stimuli are merged with probability π, otherwise the merging process stops and the current representation is used. Thus, there is probability $1-\pi$ that the full exemplar representation is used, probability $\pi\,(1-\pi)$ that a representation with a single merge is used, and so on. The key point is that, having formalized this merging process as a model of representational abstraction, a prior over the parameter π automatically corresponds to a prior over the indexing parameter ρ. Figure 6 shows a Gaussian prior[2] over π with a mean near the merge probability 0.2, and the bar graph below shows the implied prior this places on ρ for the 7 different representations. More prior mass is placed on the exemplar and prototype representations, while allowing some prior probability for the intermediate representations. This prior on ρ is non-obvious, and seems unlikely to be have been proposed in the original non-hierarchical VAM. In the hierarchical approach in Figure 6, it arises through psychological theorizing about how different representations might be generated by merging stimuli, and related prior assumptions about the probability of each merge. This sort of information is not uniquely linked to the prior. Other similar models have expressed similar theorizing in their likelihood (e.g., Love, Medin, & Gureckis, 2004).

The hierarchical approach to determining priors is broadly applicable, because it is a natural extension of theory- and model-building. It is naturally also applied, for example, in both the memory and decision models. In the memory model, a theory of rehearsal should automatically generate a prior for the τ parameters. For example, one prominent idea is that rehearsal processes are similar to free recall processes themselves (e.g., Rundus, 1971; Tan & Ward, 2008). Making this assumption, it should be possible to make predictions about whether and when presented items will be rehearsed—in the same way it is possible to make predictions about observed recalled behavior itself—and thus generate a prior for the latent rehearsal τ parameters. In the decision model, the boundary separation parameter α could be modeled as coming from control processes that respond to task demands, such as speed or accuracy instructions, as well as the accuracy of previous decisions. There are some cognitive models of these control processes, involving, for example, theories of reinforcement learning (Simen, Cohen, & Holmes, 2006), or self-regulation (Lee, Newell,

[2]Ideally, the sources and methods discussed earlier should be used to set this top-level prior.

& Vandekerckhove, 2015; Vickers, 1979), that could augment the decision model to generate the decision bound, and thus effectively place a prior on its possible values.

Benefits of informative priors

Determining priors for parameters is sometimes regarded by those who use Bayesian methods in cognitive modeling as a cost that must be borne to reap the benefits of complete and coherent inference. This lack of interest in the prior often results in what Gill (2014) terms "Bayesians of convenience", who use priors they label vague, flat, non-committal, weakly informative, default, diffuse, or something else found nearby in a thesaurus. Our view is that priors should capture the available theoretical, logical, and empirical information that is available about the psychological variables they represent. This is the sense in which priors should be informative: they should correspond to the available information. Only when modelers genuinely have no information about their parameters will informative priors be vague. In the usual and desirable situation in which something is known about parameters, assuming a vague prior loses useful information. Weiss (n.d.) expresses the issue nicely, saying:

> "In class I exhort my students that if they're not using a proper, informative prior then they're leaving money on the table.'

The problem is put most emphatically by Gill (personal communication, August 2015)[3]

> "Prior information is all over the place in the social sciences. I really don't want to read a paper by authors who didn't know *anything* about their topic before they started."

Modelers do not strive to make likelihoods vague, but aim to make them consistent with theory, empirical regularities, and other relevant information. Since, in the Bayesian approach, priors and likelihoods combine to form the predictive distribution over data that *is* the model, priors should also aim to be informative. It seems ironic to make the effort of developing a likelihood that is as informative as possible, only to dilute the predictions of the model by choosing a prior of convenience that ignores relevant theory, data, and logic. A worked example from psychophysics, showing

[3]We thank Richard Morey for drawing our attention to this quotation.

how the unthinking assumption of vague priors can undo the theoretical content of a likelihood, is provided by (Lee, in press, see especially Figures 9 and 11).

Informative priors offer significant benefits for cognitive modeling. The additional information they provide can solve basic statistical issues, related to model identifiability. These occur regularly in cognitive models that use latent mixtures, which is sometimes done to model qualitative or discrete individual differences. Latent-mixture models involve a set of model components that mix to produce data, and are notorious for being statistically unidentifiable, in the sense that the likelihood of data is the same under permutation of the mixture components (Marin, Mengersen, & Robert, 2011). The use of priors that gives each component a different meaning—by, for example, asserting that one sub-group of people has a higher value on a parameter than the other sub-group— makes the model interpretable, and makes it easier to analyze (e.g., Bartlema, Lee, Wetzels, & Vanpaemel, 2014).

Informative priors can address modeling problems relating not only to statistical ambiguity, but also those relating to theoretical ambiguity. The starting point parameter θ in the decision model provides a good example. It has sensible psychological interpretations as a bias capturing information about base-rate of correct decisions on previous trials, or as an adjustment capturing utility information about payoffs for different sorts of correct or incorrect decisions. In practice, these different psychological interpretations will typically correspond to different priors on θ and, in this sense, specifying a prior encourages a modeller to disambiguate the model theoretically.

Informative priors often make a model simpler, by constraining and focusing its predictions. The γ parameter in the categorization model provides an example of this. Sometimes the γ parameter is not included in the GCM, on the grounds that its inclusion necessarily increases the complexity of the model (J. D. Smith & Minda, 2002). It turns out, however, that introducing γ with a prior that emphasizes the possibility of near-deterministic responding can result in a simpler model. This is because the range of predictions becomes more constrained as deterministic responding is given high prior probability. This example shows that equating model complexity with counts of parameters can be mis-leading, and that the omission of a parameter does not necessarily represent theoretical neutrality or agnosticism. The omission of the γ parameter corresponds to a strong assumption that people always probability match, which turns out to make the model flexible and imprecise in its predictions.

Thus, in this case, a prior on the γ parameter that captures additional psychological theory by allowing for both probability matching and more deterministic responding reduces the model's complexity.

Constraining predictions in this sort of way has several important scientific benefits. First, it raises what Popper (1959) terms the "empirischer Gehalt" or empirical content of a model, which can be thought of as the amount of information a model conveys (see also Glöckner & Betsch, 2011; Popper, 1959; Vanpaemel & Lee, 2012). Empirical content is directly related to falsifiability and testability. A model that makes sharper predictions is more likely to rule out plausible outcomes, and therefore runs a higher risk of being falsified by empirical observation (Lakatos, 1978; Roberts & Pashler, 2000; Vanpaemel, submitted). In addition, the support a model gains through the confirmation of a prediction increases with the degree of riskiness of the prediction.

Perhaps most importantly, however, priors offer the opportunity to place additional substantive content in a model, making it a better formalization of the theory on which it is based. As noted by Vanpaemel and Lee (2012), the GCM categorization model is a good example of this. Most of the theoretical assumptions on which the GCM is explicitly founded—involving exemplar representation, selective attention, and so on—are formalized in the likelihood of the model. The theoretical assumption that is conspicuously absent is the optimal-attention hypothesis. The difference is that most of the assumptions are about psychological processes, and so are naturally formalized in the likelihood function. The optimal-attention assumption, however, relates to a psychological variable, and so is most naturally formalized in the prior.

A similar story recently played out in the literature dealing with sequential sampling models very much like the decision model in Figure 3. In a critique of these sorts of decision models, Jones and Dzhafarov (2014a) allowed the drift-rate parameter ν to have little variability over trials. P. L. Smith, Ratcliff, and McKoon (2014) argued that this allowance was contrary to guiding theory, pointing out that it implied a deterministic growth process, which conflicts with the diffusion process assumptions on which the model is founded Ratcliff and Smith (2004). Jones and Dzhafarov (2014b) rejoindered that there is nothing in the standard model-fitting approach used by Ratcliff and Smith (2004) and others that precludes inferring parameters corresponding to the reduced deterministic growth model. From a Bayesian perspective, the problem is that theoretically available information about the vari-

ability of the distribution affecting the drift rate was not specified in the traditional non-Bayesian modeling setting used by Ratcliff and Smith (2004). Because the theory makes assumptions about the plausible values of a parameter, rather than a process, it is naturally incorporated in the prior, which requires a Bayesian approach.

Discussion

One consequence of using priors for cognitive modeling is the need to conduct additional sensitivity analyses. As our survey of information sources and methods makes clear, there is no automatic method for determining a prior. A combination of creative theorizing, logical analysis, and knowledge of previous data and models is required. Different conclusions will be reached using the same data for different choices of priors, just as they would if different likelihoods were used. This means, where there is some subjectivity or arbitrariness in the specification of the prior—because the available information and methods do not allow complete determination of a formal prior specification—a sensitivity analysis is appropriate.

There is nothing inherent to the prior that makes it uniquely subject to some degree of arbitrariness. It is often the case that the likelihoods in models are defined with some arbitrariness, and it is good practice to undertake sensitivity analyses for likelihoods. Rubin and Wenzel (1996) consider a large number of theoretically plausible likelihoods for modeling memory retention, including many variants of exponential, logarithmic, and hyperbolic curves. A number of different forms of the GCM have been considered, including especially different response rules for transforming category similarity to choice probabilities (e.g., Nosofsky, 1986, 1992). Ratcliff (2013) reports a sensitivity analysis for some theoretically unconstrained aspects of the likelihood of a diffusion model of decision making. The same approach and logic applies to the part of cognitive modeling that involves choosing priors. Sensitivity analyses highlight whether and where arbitrariness in model specification is important—in the sense that it affects the inferences that address the current research questions—and so guides where clarifying theoretical development and empirical work is needed.

A standard concern in the application of Bayesian methods to cognitive modeling is that model selection measures like Bayes factors are highly sensitive to priors, but parameter inference based on posterior distributions and their summaries are far less so. Part of the reason for the greater sensitivity of the Bayes factor probably stems from the fundamentally different inferential question it solves, and its formalization

in optimizing zero-one loss. But it is also possible some of the perceived relative insensitivity of parameter inference to priors stems from the use of vague priors. It seems likely that informative priors will make inferences more sensitive to their exact specification. As a simple intuitive example, an informative prior that expresses an order constraint will dramatically affect inference about a parameter if the unconstrained inference has significant density around the values where the constraint is placed. In general, the heightened sensitivity of parameter inference to priors that capture all of the available information makes conceptual sense. These priors will generally make stronger theoretical commitments and more precise predictions about data, and Bayesian inferences will automatically represent the compromise between the information in the prior and the data.

In this paper, we have identified sources of information that can be used to develop informative priors for cognitive models, have surveyed a set of methods that can be used for this development, and have highlighted the benefits of capturing the available information in the prior. The sources and methods we have discussed are not routinely use in cognitive modeling, and we certainly do not claim they are complete, nor that they constitute a general capability for all modeling challenges. In addition, the use of informative priors in cognitive modeling is not yet extensive or mature enough to provide a tutorial on best practice in the field. But, as we have argued, priors make models more complete as account of human cognition, increase their empirical content, and make their predictions more precise, testable, falsifiable, and useful.

Acknowledgments

We thank John Kruschke, Mike Kalish, and two anonymous reviewers for very helpful comments on an earlier version of this paper.

References

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., & Rousseau, J. (2012). Combining expert opinions in prior elicitation. *Bayesian Analysis*, *7*, 503–532.

Anderson, J. R. (1992). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471–517.

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150.

Canini, K. R., Griffiths, T. L., Vanpaemel, W., & Kalish, M. L. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. *Psychonomic Bulletin & Review*, *21*, 785–793.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Donkin, C., Tran, S. C., & Le Pelley, M. (2015). Location-based errors in change detection: A challenge for the slots model of visual working memory. *Memory & Cognition*, *43*, 421–431.

Edwards, A. F. W. (1991). Bayesian reasoning in science. *Nature*, *352*, 386–387.

Feynman, R. (1994). *The character of physical law.* Modern Press.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*, 680–701.

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

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6.

Gill, J. (2014). *Bayesian methods: A social and behavioral sciences approach* (Vol. 20). CRC press.

Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making*, *6*, 711–721.

Gu, H., Kim, W., Hou, F., Lesmes, L. A., Pitt, M. A., Lu, Z.-L., & Myung, J. I. (2016). A hierarchical Bayesian approach to adaptive vision testing: A case study with the contrast sensitivity function. *Journal of Vision*, *16*, 15–17.

Hilbig, B. E., & Moshagen, M. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic Bulletin & Review*, *21*, 1431–1443.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses.* New York: Springer.

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

Jones, M., & Dzhafarov, E. N. (2014a). unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review*, *121*, 1–32.

Jones, M., & Dzhafarov, E. N. (2014b). analyzability, ad hoc restrictions, and excessive flexibility of evidence-accumulation models: Reply to two critical commentaries. *Psychological Review*, 689–695.

Kadane, J., & Wolfson, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*, 3–19. Retrieved from `http://dx.doi.org/10.1111/1467-9884.00113` doi: 10.1111/1467-9884.00113

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*, 288–294.

Kary, A., Taylor, R., & Donkin, C. (2015). Using Bayes factors to test the predictions of models: A case study in visual working memory. *Journal of Mathematical Psychology*.

Kruschke, J. K., & Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. *The Oxford Handbook of Computational and Mathematical Psychology*, 279–299.

Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge: Cambridge University Press.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*(1), 31–43.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.

Lee, M. D. (2015a). Bayesian outcome-based strategy classification. *Behavior Research Methods*, 1–13.

Lee, M. D. (2015b). Evidence for and against a simple interpretation of the less-is-more effect. *Judgment and Decision Making*, *10*, 18–33.

Lee, M. D. (in press). Bayesian methods in cognitive modeling. In *The stevens' handbook of experimental psychology and cognitive neuroscience* (Fourth ed.).

Lee, M. D., Abramyan, M., & Shankle, W. R. (2015). New methods, measures, and models for analyzing memory impairment using triadic comparisons. *Behavior Research Methods*, 1–16.

Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, *9*, 259–273.

Lee, M. D., Newell, B. R., & Vandekerckhove, J. (2015). Modeling the adaptation of the termination of search in human decision making. *Decision*, 223–251.

Lee, M. D., & Vanpaemel, W. (2008). Exemplars, prototypes, similarities and rules in category representation: An example of hierarchical Bayesian analysis. *Cognitive Science*, *32*, 1403–1424.

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

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*, 969–998.

Lisman, J., & Van Zuylen, M. (1972). Note on the generation of most probable frequency distributions. *Statistica Neerlandica*, *26*, 19–23.

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

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Marin, J. M., Mengersen, K., & Robert, C. P. (2011). Bayesian modelling and inference on mixtures of distributions. In D. Dey & C. R. Rao (Eds.), *Essential Bayesian models. Handbook of statistics: Bayesian thinking – modeling and computation 25.* Elsevier.

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

Myung, J. I., Karabatsos, G., & Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, *49*, 205–225.

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

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 3–27.

Nosofsky, R. M. (1992). Exemplars, prototypes and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honour of William K. Estes vol. 1.* Hillsdale, NJ: Lawrence Erlbaum.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, R., Garthwaite, P., Jenkinson, D., ... Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities.* Wiley.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*, 57–83.

Popper, K. R. (1959). *The logic of scientific discovery.* Routledge.

Ratcliff, R. (2013). Parameter variability and distributional assumptions in the diffusion model. *Psychological Review*, *120*, 281–292.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*, 333–367.

Robert, C. P. (2007). *The Bayesian choice*. New York, NY: Springer-Verlag.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Pratte, M. S. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin & Review*, *14*, 597–605.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*, 734–760.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of experimental psychology*, *89*, 63–77.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*, 233–250.

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

Simen, P., Cohen, J. D., & Holmes, P. (2006). Rapid decision threshold modulation by reward rate in a neural network. *Neural Networks*, *19*, 1013–1026.

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 800–811.

Smith, P. L., Ratcliff, R., & McKoon, G. (2014). The diffusion model is not a deterministic growth model: Comment on Jones and Dzhafarov (2014). *Psychological Review*, *121*, 679–688.

Taagepera, R. (2007). Predictive versus postdictive models. *European Political Science*, *6*, 114–123.

Tan, L., & Ward, G. (2008). Rehearsal in immediate serial recall. *Psychonomic Bulletin & Review*, *15*, 535–542.

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

Trafimow, D. (2005). The ubiquitous Laplacian assumption: Reply to Lee and Wagenmakers (2005). *Psychological Review*, *112*, 669–674.

Turner, B. M., Dennis, S., & Van Zandt, T. (2013). Likelihood-free Bayesian analysis of memory models. *Psychological Review*, *120*, 667–678.

Vanpaemel, W. (2009). Measuring model complexity with the prior predictive. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 1919–1927). Red Hook, NY: Curran Associates Inc.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology*, *55*, 106–117.

Vanpaemel, W. (submitted). Complexity, data prior and the persuasiveness of a good fit: Comment on Veksler, Myers and Gluck (2015).

Vanpaemel, W., & Lee, M. D. (2012). Using priors the formalize theory: Optimal attention and the Generalized Context Model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.

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

Vickers, D. (1979). *Decision Processes in Visual Perception.* New York, NY: Academic Press.

Vincent, B. (2015). Hierarchical Bayesian estimation and hypothesis testing for delay discounting tasks. *Behavior Research Methods*, 1–13.

Weiss, R. (n.d.). *Kathryn Chaloner 1954–2014.* (https://faculty.biostat.ucla.edu/robweiss/taxonomy/term/107)

Welsh, M., Begg, S., Bratvold, R., & Lee, M. (2004). Problems with the elicitation of uncertainty. In *Spe annual technical conference and exhibition.* Richardson, TX: Society for Petroleum Engineers.

Wickens, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel (1996): A quantitative description of retention. *Psychological Review*, *105*, 379–386.

Wiehler, A., Bromberg, U., & Peters, J. (2015). The role of prospection in steep temporal reward discounting in gambling addiction. *Frontiers in Psychiatry*, *6*.