Using Priors to Formalize Theory: Optimal Attention and the Generalized Context Model

Wolf Vanpaemel
Faculty of Psychology and Educational Sciences
University of Leuven

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

Abstract
Formal models in psychology are used to make theoretical ideas precise, and allow them to be evaluated quantitatively against data. We focus on one important—but under-used, and incorrectly maligned—method for building theoretical assumptions into formal models, offered by the Bayesian statistical approach. This method involves capturing theoretical assumptions about the psychological variables in models by placing informative prior distributions on the model parameters representing those variables. Formalizing psychological theory within an informative prior distribution yields a more complete model, and allows standard Bayesian model selection methods to be applied without concerns about the sensitivity of results to the prior. We demonstrate this approach of casting basic theoretical assumptions in an informative prior, by considering a case study that involves the Generalized Context Model (GCM) of category learning. We use prior distributions to formalize and test existing theorizing about the optimal allocation of attention. We argue that the general approach of using psychological theory to guide the specification of informative prior distributions for model parameters is widely applicable, and should be routinely used in psychological modeling.
Introduction

The Bayesian approach is becoming increasingly important and popular in implementing and evaluating psychological models, including models of psychophysical functions (e.g., Kuss, Jäkel, & Wichmann, 2005), stimulus representations (e.g., Lee, 2008), category learning (e.g., Lee & Vanpaemel, 2008; Vanpaemel & Storms, 2010), signal detection (e.g., Rouder & Lu, 2005), response times (e.g., Rouder, Lu, Speckman, Sun, & Jiang, 2005) and decision-making (e.g., Wetzels, Grasman, & Wagenmakers, 2010). It is widely recognized in statistics (e.g., Jaynes, 2003; Gelman, Carlin, Stern, & Rubin, 2004), and increasingly in psychology (e.g., Kruschke, 2010, 2011; Lee & Wagenmakers, 2005), that the Bayesian approach offers a complete and coherent framework for making inferences using models and data. Bayesian parameter estimation allows for model parameters to be estimated in a way that naturally represents uncertainty, and is applicable even when there are few data. Bayesian model selection allows for models to be compared in a way that automatically implements an Ockham’s razor for balancing goodness-of-fit with complexity, including when models are non-nested (e.g., Lee, 2008; Myung & Pitt, 1997; Shiffrin, Lee, Kim, & Wagenmakers, 2008).

Priors as Problems, Priors as Opportunities

Even advocates of the Bayesian approach, however, often view the benefits as coming at a cost, in the form of having to specify prior distributions. In Bayesian parameter estimation, placing priors on parameters is relatively uncontroversial. This is because, as long as the data available are sufficiently informative, the choice of prior has little impact on inference. In contrast, Bayesian model selection is generally sensitive to the exact choice of the prior. Because priors are deemed to bring an unwanted level of arbitrariness to the conclusions, placing prior distributions on parameters is controversial. For example, in their seminal paper on Bayesian model selection in psychology, Myung and Pitt (1997, p. 91) say that “[t]he Bayesian method [for model selection], however, has its drawbacks. One is that parameter priors are required to compute the marginal likelihoods [and hence the Bayes Factor].” In practice, these objections have led researchers often to uniform, flat, or otherwise weakly informative priors, in an attempt to limit the information injected into the model selection inference (e.g., Kemp, Perfors, & Tenenbaum, 2004; Lee, 2004; Pitt, Myung, & Zhang, 2002).
We do not agree that prior distributions on parameters are an unwanted aspect of the Bayesian framework. Within psychological models, the variables that give rise to the behavior of interest are typically unknown, and are therefore represented by parameters. Generally, knowledge about these parameters is gained by estimating their values from observed data. However, it is almost never the case that these values are completely unknown. Knowledge about the parameters is often available before data are observed, based on previous experience or on theory. Expressing knowledge or assumptions about the values parameters are likely to take, and the relationship between them is difficult in orthodox approaches to inference, but is both possible and necessary within the Bayesian approach (Jaynes, 2003; Lindley, 2004).

Therefore, we believe it is wrong to malign priors as a necessary evil. Instead, the need for priors should be welcomed by psychological modelers as yet another advantage of the Bayesian approach, since it provides an additional opportunity to formalize psychological theory in models. We present a concrete example of how a prior can be used to capture important existing psychological theory. Our case study considers the Generalized Context Model of category learning (GCM: Nosofsky, 1984, 1986). We show how basic existing theoretical assumptions—that people tend to allocate their attention over psychological dimensions so as to optimize their classification performance—can be used to derive an informative prior distribution over parameter values. We illustrate how this informative prior affords the possibility of using Bayesian model selection methods to evaluate empirically the theoretical assumption represented by the prior.

Further, we highlight that, besides providing the possibility of using Bayesian model selection methods to evaluate theory represented by the prior, incorporating theoretically meaningful informative priors has several additional important benefits. One is that having an additional mechanism for expressing theoretical assumptions in a model provides a way of constructing richer, stronger and more complete models. Another is that, when models have priors to which the theorist is committed, sensitivity to priors is no longer a problematic aspect of Bayesian model selection.

Theory and Standard Implementation of the GCM

Nosofsky’s (1984, 1986) GCM is one of the most successful and influential models in cognitive psychology. Like any good model, it is built on a set of clear theoretical assumptions. We first describe those included in the standard implementation, and
then turn to an assumption that has failed to be formalized within the model: the attention-optimization hypothesis.

*The Standard Implementation of the GCM*

**Dimensional Stimulus Representation.** The GCM assumes that stimuli can be represented by their values along underlying stimulus dimensions, as points in a multidimensional psychological space. Formally, in the model, the $i$th stimulus is represented in a $D$-dimensional psychological space by the point $x_i = (x_{i1}, \ldots, x_{iD})$.

**Selective Attention.** A key contribution of the GCM to modeling category learning is the assumption that the structure of the psychological space is systematically modified by selective attention. Formally, the $d$th dimension has an attention weight, $w_d$ with $0 \leq w_d \leq 1$, and $\sum_d w_d = 1$. These attention weights act to “stretch” attended dimensions, and “shrink” unattended ones. The psychological distance between the $i$th and $j$th stimuli is given by the weighted city-block ($r = 1$) or Euclidean ($r = 2$) distance between the points, $d_{ij} = [\sum_d w_d |x_{id} - x_{jd}|^r]^{1/r}$.

**Similarity-based Classification.** The GCM assumes classification decisions are based on similarity comparisons with the stored exemplars, with similarity determined as a nonlinearly decreasing function of distance in the psychological space. Formally, the similarity between the $i$th and $j$th stimuli is modeled as an exponential ($\alpha = 1$) or Gaussian ($\alpha = 2$) decay function, $s_{ij} = \exp\left(-cd_{ij}^\alpha\right)$, where $c$ is a generalization parameter.

**Exemplar Representation.** The GCM assumes that categories are represented by individual exemplars. This means that, in determining the overall similarity of a stimulus to a category, every exemplar in that category is considered. Formally, the overall similarity of the $i$th stimulus to Category A is $s_{iA} = \sum_{j \in A} s_{ij}$.

**Choice Rule.** The GCM assumes classification decisions are based on the Luce choice rule, as applied to the overall similarities. Formally, the probability that the $i$th stimulus will be classified as belonging to Category A, rather than to Category B, is often modeled as $p_{iA} = \frac{s_{iA}}{s_{iA} + s_{iB}}$. Extended versions of this choice rule, making additional assumptions about response bias, and the determinism of responding, have also been considered.
PRIORS AND THEORY

Taken together, these five theoretical assumptions lead to the formal implementation of the GCM as it is usually used to estimate parameters or tested against category learning data. Typically, the points representing the stimuli (i.e., $x_i$) are found by multidimensional scaling from separately collected similarity data (Shepard, 1962, 1980). The metric parameter $r$ is set according to knowledge of the separability ($r = 1$) or integrality ($r = 2$) of the stimulus domain and $\alpha$ according to knowledge of the discriminability of the stimuli. With these in place, the generalization parameter $c$ and attention weights $w_d$ remain as free parameters.

Since most applications of the GCM are outside the Bayesian framework, priors for these parameters are never explicitly formulated. In applications, generally, it is implicitly assumed that all parameter values are equally likely. Under a Bayesian interpretation, this corresponds to a uniform (for $w_d$) or flat (for $c$) prior, so we will call this standard implementation of the GCM, assuming all parameter values to be equally likely, $\text{GCM}_{\text{unif}}$.

The Attention Optimization Hypothesis

The idea that people selectively allocate their attention called for the introduction of the free attention parameters $w_d$. However, the existence of an attention weight is not the limit of theorizing, and some progress has been made in understanding the allocation of attention. In particular, the first GCM paper suggested that “it is reasonable to hypothesize that, with learning, subjects will distribute attention among the component dimensions in a way that tends to optimize performance” (Nosofsky, 1984, p. 109), and the following seminal paper said “As a working hypothesis, it is assumed that subjects will distribute attention among the component dimensions so as to optimize performance in a given categorization paradigm. That is, it is assumed that the $w_k$ [i.e., $w_d$] parameters will tend toward those values that maximize the average percentage of correct categorizations.” (Nosofsky, 1986, p. 41). In a more recent summary paper, Nosofsky (1998a) observes that “The theme of optimal performance has always played a central role in theorizing involving Nosofsky’s (1986) generalized context model.”

Despite being one of the original key assumptions of the theory, the assumption that people learn to allocate their attention over psychological dimensions so as to optimize their classification performance is not incorporated formally in the standard implementation of the GCM. By assuming all values of the attention weights to be
equally likely, the attention-optimization hypothesis is ignored. This is surprising, since the idea that people allocate their attention optimally seems to be a clear theoretical position that should—as with assumptions about representation, selective attention, similarity, selective attention and choice behavior—be able to be formalized within a model.

We think the reason the attention-optimization hypothesis was never formally implemented within the GCM is that, unlike the other assumptions, it is an assumption about the distribution of parameters. The exemplar assumption dictates how categories are represented, and the selective attention, similarity and choice assumptions dictate how information is processed and behavior is produced in a category learning task. The attention optimization hypothesis, in contrast, does not affect how processing proceeds, but rather provides information about the values of the attention parameters that partly control the processing. This sort of assumption is usually not easy to express in the likelihood function. Rather, it is exactly the sort of information that can be expressed in the prior distribution over the possible values of the model parameters, and so requires the adoption of the Bayesian framework.

An Implementation of the GCM with Optimal Attention

In this section, we develop an implementation of the GCM’s existing theoretical assumptions that extends the standard implementation by formally including the attention optimization hypothesis. This version of the GCM, called GCM\textsubscript{opt}, is identical to GCM\textsubscript{unif}, except for one key assumption—whether or not people optimally allocate attention. While GCM\textsubscript{unif} does not include as part of its formal description any mechanism that corresponds to optimal attention, by assuming a uniform prior over the attention parameters, GCM\textsubscript{opt} explicitly embodies the assumption of optimal attention, by assuming an informative prior over the attention parameters. Because of the informative prior it assumes, expressing additional theory, GCM\textsubscript{opt} is a richer and more complete model than GCM\textsubscript{unif}.

Our application also highlights a particularly useful by-product of using informative priors. If priors are used for expressing theories about parameters, these theories can be tested using Bayesian model selection methods. Since the prior over the attention weights is the only difference between the models, the empirical selection between GCM\textsubscript{unif} and GCM\textsubscript{opt} amounts to a direct quantitative evaluation of
Data

Our demonstration of completing the GCM to include an informative prior, and using this completed version to test the theory captured in the prior relies on four previously published data sets, reported by Nosofsky (1989), which are briefly discussed first.

Stimuli. Nosofsky’s (1989) experiment involved a set of 16 semicircles, with an embedded radial line drawn from the center of the semicircle to the rim, as shown in Nosofsky’s (1989) Figure 2. The stimuli varied in the size of the semicircle and in the angle of orientation of the line. A two-dimensional representation of the stimulus space was used, based on similarity data derived from a separate identification experiment. Following Nosofsky (1989), we used the Gaussian similarity and Euclidean distance version.

Category Structures. Four different category structures were defined using the 16 stimuli, as shown in Figure 1. For all of the structures, three or four stimuli were assigned to Category A, three or four to Category B, and the remaining eight or nine stimuli were left unassigned. In two of the category structures, called ‘angle’ and ‘size’, attending to a single dimension is sufficient to learn the categories. The other structures, ‘criss-cross’ and ‘diagonal’, require attention to both dimensions.

Figure 1. The four category structures used by Nosofsky (1989). Each numbered location corresponds to a stimulus, with the stimuli assigned to the two categories shown by squares (Category A) and circles (Category B) for each structure. Based on Nosofsky (1989, Figure 1).
**Procedure.** Each category structure was learned by a separate group of participants under a standard training-test procedure. Nosofsky (1989) restricted modeling analyses to those participants who achieved a sufficiently high accuracy in classifying the assigned stimuli during the final 125 trials of the training phase, leaving 41, 37, 43 and 37 participants for the angle, criss-cross, diagonal, and size structures respectively. The data reported in Nosofsky’s (1989) Table 4 are the counts of Category A responses made during the test phase, aggregated over participants.

**Formalizing the Attention Optimization Hypothesis**

The stimuli are two-dimensional, so there is a single free attention weight parameter \( w \). One naive way to include the attention optimization hypothesis in a formal implementation of the GCM is to fix this attention parameter to the single optimal value. However, this does not correspond well to the theoretical assumption, which is described in terms of tendencies and inclinations. It is more natural to express the tendency to optimality in terms of a *distribution* over the possible values of the attention parameter. We view this distribution as uncertainty on the part of the modeler, representing the relative likelihood they assign to different possible attention values a participant in an experiment might be using. Under this view, the uncertainty corresponds exactly to the Bayesian prior.

It seems reasonable for the prior distribution to be centered on the optimal value. Following Nosofsky (1984, 1986), this value is found by maximizing the average proportion correct \( APC(c, w) = \sum_{i \in A} p_{iA}(c, w) + \sum_{i \in B} p_{iB}(c, w) \). One complication with finding the optimal value is that \( APC \) simultaneously depends on both \( c \) and \( w \), implying that there is no unique attention weight maximizing \( APC \). Previous authors have pragmatically solved this issue by fixing \( c \) to a single value, either estimated from the empirical data, or specified independently of the data (e.g., Lamberts, 1995; Nosofsky, 1984, 1986; Rehder & Hoffman, 2005). Our remedy starts from the observation that, as \( c \) increases, self-similarity starts to dominate the total similarity, and consequently the GCM starts to behave as a rote memorization model rather than as the exemplar-based generalization model it is intended to be. Based on this consideration, of all \((c, w)\) pairs that maximize \( APC \), we choose the value \( w \) that implies perfect performance with the smallest possible value of \( c \). As different modelers can have different degrees of uncertainty about the possible attention values used by a subject, we considered two different priors for \( w \) for each learning task. Both priors
are centered on the same optimal value, but differ in the level of uncertainty they represent. Formally, Beta distributions were derived with variances of 0.01 (more certain) and 0.05 (less certain), as shown for each category structure in Figure 2.

The attention optimization hypothesis posits that, for category structures in which only a single dimension is relevant for performing the classification, observers will be inclined to attend selectively to this dimension, which translates to the attention weights having extreme values. As Nosofsky (1998b, p. 330) makes clear, the attention optimization hypothesis does not predict that for category structures in which both dimensions are relevant, both dimensions will be attributed exactly equal attention \((w = \frac{1}{2})\). What the hypothesis does predict is that, for such structures, extreme values of the attention weight, as with one dimensional classification problems, are very unlikely. These expectations accord with the priors as is shown in Figure 2: For the size and angle structures, attention is centered close to 0 or 1, respectively; for the criss-cross and diagonal structures, less extreme values are favored. These theory-informed distributions represent the attention values GCM\(_{\text{opt}}\) believes correspond to people’s attention.\(^1\)

**Models and Model Evaluation Method**

Both GCM\(_{\text{unif}}\) and GCM\(_{\text{opt}}\) assume the data-generating process given earlier, and assume joint independence in the prior distribution, so that \(p(w, c) = p(w) p(c)\). In particular, GCM\(_{\text{unif}}\) and GCM\(_{\text{opt}}\) use a uniform prior for \(c\), so that \(c \sim \text{Uniform}(0, 1)\). The only difference between GCM\(_{\text{unif}}\) and GCM\(_{\text{opt}}\) is the prior for \(w\). GCM\(_{\text{unif}}\) is noncommittal about allocation of attention, so that \(w \sim \text{Uniform}(0, 1)\); GCM\(_{\text{opt}}\), in contrast, assumes that people tend to optimally allocate attention, as expressed in the informative priors for \(w\) shown in Figure 2.

To compare these models, we relied on the standard Bayesian model selection approach provided by Bayes Factors (e.g., Kass & Raftery, 1995). Let \(y = (k_1, \ldots, k_{16})\) denote the observed category learning data, where \(k_i\) is the number of times the \(i\)th stimulus is classified as being in Category A. Formally, the Bayes Factor based on these data is given by the ratio of marginal likelihoods \(BF_{\text{uo}} = \frac{p(y | \text{GCM}_{\text{unif}})}{p(y | \text{GCM}_{\text{opt}})}\). Each marginal likelihood integrates over the prior distribution of parameters\(^2\), so that \(p(y | \text{GCM}) = \int p(y | w, c, \text{GCM}) p(w, c | \text{GCM}) \, d(w, c)\).

\(^1\)Note that GCM\(_{\text{unif}}\) could be viewed as an extremely weak version of GCM\(_{\text{opt}}\).

\(^2\)Note that \(c\) is only held fixed when deriving the optimal value for \(w\). When computing the marginal likelihood, \(c\) is no longer held fixed.
Figure 2. Distributions over the attention parameter $w$, corresponding to expectations about the allocation of attention for each of the four category structures used in the experiment reported by Nosofsky (1989). Panel A shows prior distributions representing the strongest assumption, with variance 0.01; Panel B shows prior distributions representing the weakest assumption, with variance 0.05.

Marginal likelihoods have the well-known and important property of automatically accounting for model complexity by finding the average fit of each model to the data across the entire prior distribution of parameters. A crucial feature of marginal likelihoods for our application is that they—unlike simple approximations like the Akaike Information Criterion or Bayesian Information Criterion—are also sensitive to the prior distributions on parameters. Thus, if people use optimal attention weights consistent with the assumptions of GCM_{opt}, the marginal likelihood will reveal a better average fit. On the other hand, if people make classification decisions consistent with non-optimal attention allocation, the marginal likelihood will prefer GCM_{unif}. Thus, a BF_{unif} > 1 indicates that, based on the evidence provided by the data, GCM_{unif} is preferred.

Results

We evaluated each marginal likelihood using standard numerical grid sampling, using steps of 0.001 for $w$ and 0.01 for $c$. The calculated marginal likelihoods for
Table 1: Marginal likelihoods, expressed as minus log likelihoods, for GCM\textsubscript{unif}, and for the strong (with variance 0.01) and weak (with variance 0.05) versions of GCM\textsubscript{opt}, for each of the four category learning tasks considered by Nosofsky (1989).

<table>
<thead>
<tr>
<th>Category</th>
<th>GCM\textsubscript{opt} (strong version)</th>
<th>GCM\textsubscript{opt} (weak version)</th>
<th>GCM\textsubscript{unif}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angle</td>
<td>63.97</td>
<td>64.19</td>
<td>64.57</td>
</tr>
<tr>
<td>Size</td>
<td>45.49</td>
<td>46.16</td>
<td>46.83</td>
</tr>
<tr>
<td>Criss-cross</td>
<td>68.13</td>
<td>63.64</td>
<td>62.74</td>
</tr>
<tr>
<td>Diagonal</td>
<td>61.68</td>
<td>56.23</td>
<td>55.96</td>
</tr>
</tbody>
</table>

GCM\textsubscript{unif} and the strong and weak version of GCM\textsubscript{opt}, expressed on the standard logarithmic scale, are presented in Table 1. When specifying the prior for \(c\), the upper bound of the uniform distribution was picked rather arbitrarily. When priors are not conceptually grounded, it is crucial that a sensitivity analysis is performed. Reassuringly, the results reported below are almost identical using alternative values of 5, 15, 20, 50 and 100 as upper bounds for \(c\).

The difference between these log marginal likelihoods for the two models is the log of the Bayes Factor that compares them. Thus, for the angle category structures, \(BF_{ou}\) shows that the strong version of GCM\textsubscript{opt} is \(\exp(64.57 - 63.97) \approx 1.8\) times more likely than GCM\textsubscript{unif}, providing mild evidence for optimal attention. When the belief in the attention optimization hypothesis is less strong, GCM\textsubscript{opt} becomes more similar to GCM\textsubscript{unif}, which is indicated by drop in the Bayes Factor to \(\exp(64.57 - 64.19) \approx 1.5\), still supporting GCM\textsubscript{opt} but to a lesser degree. Similarly, for the size category structure, there is evidence for GCM\textsubscript{opt}, with \(BF_{ou} = \exp(46.83 - 45.49) \approx 3.8\), decreasing to \(\exp(46.83 - 46.16) \approx 2.0\) when the belief in optimal attention decreases.

For the criss-cross and diagonal structures, in contrast, \(BF_{uo}\) shows that the empirical data indicate that GCM\textsubscript{unif} is, respectively, \(\exp(68.13 - 62.74) > 200\) and \(\exp(61.68 - 55.96) > 300\) times more likely than the strong version of GCM\textsubscript{opt}. If the strength of the belief in attention optimization declines, GCM\textsubscript{unif} is still supported, although to a much lesser degree, with Bayes Factors \(\exp(63.64 - 62.74) \approx 2.5\) and \(\exp(56.23 - 55.96) \approx 1.3\).

Overall, in the two category structures where a single dimension is relevant for performing a classification, there is evidence for attention optimization. In the two category structures where both dimensions are relevant, there is overwhelming evidence for the implementation of the GCM using uniform priors. In these conditions,
participants do not seem to allocate their attention optimally over both of the stimulus dimensions. When learning categories requires attention to more than a single dimension, attention allocation appears not to be governed by optimality mechanisms alone (see Nosofsky, 1998b; Nosofsky & Johansen, 2000, for discussions). These conclusions are consistent with those of Nosofsky (1989), made on the basis of point parameter estimates found by fitting the GCM without explicit priors on parameters.

Discussion

In a discussion on the theoretical status of the free attention parameters in the GCM, Nosofsky (1998b, p. 330) correctly observed that “parameters are ubiquitous in mathematical modeling in psychology. They count against a model when one evaluates its parsimony, and the need to estimate free parameters represents an important source of incompleteness in theorizing.” Often faced with incompleteness, current modeling in psychology has regularly relied on uniform or flat priors. We believe this state of affairs is unfortunate, and damaging to the field.

Theory about psychological variables is often incomplete, calling for the use of parameters in models. When theory is totally mute on the values of parameters, the use of non-informative priors logically follows. However, theory, while usually incomplete, often has at least something to say. Within the coherent Bayesian approach to statistical inference, it is possible to include even incomplete and inexact theorizing about psychological variables within informative prior distributions on the parameters representing these variables. Specifying an informative prior lies between succumbing to the bland default specification of uniform or flat priors, pretending no theory at all is available, and overconfidently fixing parameters, pretending that a theory is complete and exact.

To illustrate that prior distributions over parameters can capture psychological theory, and are therefore integral parts of a psychological model, we focused on the GCM, which contains parameters corresponding to dimensional attention. Current theorizing about attention allocation is not yet developed to the point at which the attention parameter can be fixed to a single value. However, some theory is available, relating to optimal attention allocation, which can be formally represented in an informative prior on the attention parameter.

More generally, we think our demonstration of the ability of prior distributions in the Bayesian setting to carry theoretical information underscores an under-
appreciated merit of the Bayesian approach. Non-Bayesian statistical frameworks do not make it easy to include theorizing about more or less likely combinations of parameter values in models. This is unfortunate, since specifying informative prior distributions over the parameters is an important avenue for expressing theoretical ideas and has several attractive advantages.

One is that expressing psychological theory in a prior allows theory represented by the prior to be evaluated using Bayesian model selection methods. Our case study highlighted how implementing the attention optimization hypothesis in an informative prior provides the ability to test this theory directly using Bayes factors. Our model selection approach contrasts with the usual approach of investigating the attention-optimization hypothesis, which uses a model that explicitly does not try to capture the relevant theory, and then relies on post-hoc assessment of parameter estimates to reach conclusions. The approach taken in this paper—testing a model that explicitly formalizes the assumption—follows one of the rationales for building formal models in psychology, which is to make psychological theories precise and testable.

Another important benefit of using theory to define informed prior distributions over parameters involves the merits of model selection methods, like Bayes Factors, that integrate over the parameter space. While these model selection methods have the attractive feature of providing an automatic Ockham’s razor, a serious reservation about these methods is that the results can be very sensitive to priors, including often near-arbitrary decisions such as how to bound the parameter space (e.g., Liu & Aitkin, 2008). Essentially unprincipled bounding can be found in landmark papers such as Pitt et al. (2002, footnote 9), and the important effect the choice of bounds can have on model selection conclusions is nicely demonstrated, for example, by the critique presented by Liu and Aitkin (2008) showing flaws in the retention function comparison conducted by Lee (2004). The same problem occurred in our case study, because the choice of 10 as the upper bound of $c$ was largely unprincipled, reflecting a serious incompleteness in theorizing about the generalization parameter (see Nosofsky & Johansen, 2000, for some initial suggestions in this regard). A partial solution is to perform a sensitivity analysis, as we did. These sorts of problems are substantially addressed, however, if the prior distributions used in calculating Bayes Factors are not arbitrary, but come directly from theory, as with the attention parameter in our optimal attention implementation of the GCM.

In fact, when priors express theory, the use of Bayes Factors is encouraged,
because it is sensitive to all theoretical assumptions made by the model, including those formalized in the prior. Any method that is insensitive to the prior only provides a partial evaluation of the theory formalized in the model (e.g., Vanpaemel, 2010). Therefore, Bayes Factors could not only be used to directly evaluate the theory in the prior, as in our example, but also to compare the GCM to, for example, the prototype model (e.g., Reed, 1972) or to models from the general recognition theory framework (e.g., Ashby & Townsend, 1986), provided the priors assumed in these models capture meaningful theory.

Perhaps the most important reason to use informative priors is that they help build theoretically richer and more complete models. There is a sense in which GCM_{opt} is closer to the theoretical foundations of the GCM, as outlined by Nosofsky in the 1980s, than the standard implementation. Generally, a model with an informative prior will be more constrained, stronger and less complex and will make more stringent predictions than its counterpart with a uniform or flat prior (see, e.g., Vanpaemel, 2009, for an example). The benefits of taking a strong position is illustrated by considering the conclusions based on the strong and weak version of GCM_{opt}. As our case study demonstrated, the exact level of the evidence against or in favor of GCM_{opt} depends on how strongly one is willing to believe in the theory. A model making strong assumptions has the advantage that when the assumptions are consistent with behavior, the evidential reward is larger, as illustrated with the angle and size category structures. If a model making bold claims does not accommodate the data well, however, the evidential penalty is larger, as illustrated with the criss-cross and diagonal category structures. Stronger tests of theories and higher scientific payoffs come from more complete models.

The approach we have advocated and demonstrated is applicable in any situation where theory has something to say about which combinations of parameter values in a model are more likely than others. In psychological models, this should very often be the case, because parameters usually correspond to meaningful psychological variables, and are a primary focus of theory.

There are many ways theorists can formalize their ideas within the Bayesian approach. One common possibility is to place order constraints on parameters (e.g., Hoijtink, Klugkist, & Boelen, 2008; Mulder et al., 2009), or to constrain the range of a parameter (e.g., Navarro, Pitt, & Myung, 2004; Vanpaemel, 2010). More advanced methods involve maximum entropy priors, where complicated constraints on
variables can be embedded in distributions (e.g. Jaynes, 2003, Ch. 9), and hierar-
chical Bayesian approaches, where theories about where the basic model parameters
themselves come from are implemented formally within the model (Lee & Vanpaemel,
2008; Vanpaemel, 2011).

Of course, it will often be challenging to translate expectations about psycho-
logical variables into formal statements about prior probability distributions. This is
exactly the same sort of challenge as is faced in building any sort of formal model. The
original formulation of the GCM must have presented problematic challenges when
casting basic theoretical assumptions as formal model mechanisms. Assumptions
about stimulus representation took the form of points in a psychological space, rather
than other representational possibilities; assumptions about similarity-based catego-
rization took the form of families of exponential curves relating distance to similarity,
rather than other possible generalization gradients; and assumptions about choice
took the form of the Luce choice rule, rather than alternative choice functions. In the
same way, the particular approach we used to formalize the optimal attention assump-
tion is surely not the only possibility, but it is a formal and testable implementation
consistent with the basic theoretical motivation.

Most generally, we think the belief—held, in our experience, by many who con-
struct and test psychological models—that priors on parameters should be as vague
and non-informative as possible is misguided. When researchers build models, a pri-
mary goal is to capture regularities in data, finding the principles and processes that
govern people’s behavior. In this sense, model building aims to make informed predic-
tions about data, and the goal is certainly not to be non-informative, and render all
possible data sets and behaviors equally likely. We think the same argument applies
to the psychological variables represented by parameters. A key goal of modeling is
to develop theories that make informative statements about these variables, so that
placing non-informative priors on parameters is the hallmark of theoretical immatu-

Acknowledgments

MDL acknowledges the support of KULeuven/BOF Senior Fellowship
SF/08/015.
References


