Constructing informative model priors using hierarchical methods

Wolf Vanpaemel

Department of Psychology, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium

ARTICLE INFO

Article history:

Received 6 February 2010

Received in revised form

8 July 2010

Available online 8 October 2010

Keywords:

Bayes

Hierarchical

Estimation

Priors

Informative

Subjective

Abstraction

Category learning

ABSTRACT

Despite their negative reputation, informative priors are very useful in inference. Priors that express psychologically meaningful intuitions damp out random fluctuations in the data due to sampling variability, without sacrificing flexibility. This article focuses on how an intuitively satisfying informative prior distribution can be constructed. In particular, it demonstrates how the hierarchical introduction of a parameterized generative account of the set of models under consideration naturally imposes a non-uniform prior distribution over the models, encoding existing intuitions about the models. The hierarchical approach for constructing informative model priors is made concrete using a worked example, the Varying Abstraction Model (VAM), a family of categorization models including and expanding the exemplar and prototype models. It is shown how psychological intuitions about the relative plausibilities of the models in the VAM can be formally captured in an informative prior distribution over these models, by specifying a theoretically informed process for generating the models in the VAM. The smoothing effect of the informative prior in estimation is demonstrated by considering ten previously published data sets from the category learning literature.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Rouder, Lu, Speckman, Sun, and Jiang (2005, p. 198) tell the following (true) anecdote from a baseball game:

In the late summer of 2000, the struggling Kansas City Royals were hosting the Boston Red Sox. Pitching for Boston was Pedro Martínez, who was having a truly phenomenal year. Many in the crowd came to see Martínez and his dominant pitching. Contrary to expectation, in the first inning, Kansas City scored five runs, and Boston none. At the end of the first inning, one of our colleagues, who is a loyal Royals fan and an APA editor, predicted a final score of 45–0.

There is a lot of logic to the prediction of 45–0, because it follows from multiplying the score after one inning by the total number of innings in a baseball game. However, it turned out to be far away from the truth. Rouder et al. (2005, p. 198) recount:

After the first inning, Martínez pitched well, allowing only one additional run. Kansas City lost the game by a score of 7–6; Martínez was the winning pitcher.

It seems that it should be possible to come up with a better prediction than the logical one that turned out to be so dramatically wrong. Fortunately, there is an easy remedy: adding knowledge

about baseball to the prediction. Even the slightest knowledge of baseball, such as knowing that there has never been a professional baseball game with a score as extreme as 45–0, would have dramatically improved the prediction of the final score. More detailed knowledge of the game, such as a (rough) distribution of final baseball scores, would have resulted in an even better prediction. A baseball expert knowing that Boston was superior to Kansas City and had the best pitcher might even have come close to nailing the final score. In sum, the prediction could have been vastly improved by combining the observation of the score after one inning with some knowledge about the game.

1.1. The ugly duckling called prior

This example highlights the major theme of this article: using knowledge about what is plausible can *smooth* inference, in the sense that it can prevent one from reaching extreme or unreasonable conclusions. Including such knowledge is greatly facilitated by the adoption of the Bayesian framework, which provides a coherent and principled way to combine information from observed data with information from additional knowledge that does not depend on the data. The Bayesian framework achieves this by augmenting the likelihood function with a distribution for parameters and models, the prior, which provides an opportunity to incorporate additional knowledge and intuitions. A prior could, for example, encode that final scores in professional baseball games rarely exceed 10.

The ability to take relevant prior knowledge into account in inference is one of the major distinctions between the Bayesian

E-mail address: wolf.vanpaemel@psy.kuleuven.be.

URL: <http://ppw.kuleuven.be/concat>.

framework and the standard framework (Lindley & Phillips, 1976). Most standard methods tend to ignore the prior knowledge that is available in a problem. For example, maximum likelihood estimation implicitly and by default assumes a uniform prior (see footnotes 4 and 8), which does not encode any additional information about the plausibility of parameter values or models. Priors that, in contrast, are intended to express this sort of knowledge are called *informative*.

Informative priors are an often maligned aspect of Bayesian inference. As a result, most current psychological modeling explicitly or implicitly relies on uniform priors, thereby closing its eyes to additional knowledge, theory, assumptions or intuitions that might exist. This unfortunate state of affairs has at least two grounds, both conceptual and practical (Goldstein, 2006). A first, conceptual objection to including prior knowledge in the form of an informative prior is that this practice is considered to make inference inherently and irremediably subjective. It is often argued that informative priors corrupt scientific thinking, which seizes the high ground of objectivity, into mere wishful thinking (e.g., Efron, 1986). From this perspective, informative priors are inappropriate for reaching scientific conclusions and have no place in scientific analyses (see, e.g., Lindley, 2004; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008, for discussions).

A second, practical reason for the current underuse of informative priors in psychological modeling relates to a lack of methods for formalizing prior knowledge or intuitions. Even when theorists have clear prior intuitions about the parameters or models they are dealing with, and even if they are convinced that the inference will benefit from including these intuitions, they are still confronted with the challenging problem of quantifying these intuitions in order to make quantitative inferences.

Contrary to the widespread view that the influence of the prior on inference should be minimized, I regard informative priors to be highly useful in inference, chiefly for their capacity to pull the data away from inappropriate and implausible inferences, without losing useful flexibility. Adding prior knowledge or intuitions, in the form of a non-uniform, informative prior should be pursued much more often than is the case now. The present article mainly addresses the practical objection to informative priors, by illustrating a generally applicable mechanism for formally capturing existing intuitions about models in an informative prior distribution over the models.

1.2. Outline

This article provides a detailed demonstration of how a theoretically motivated, informative prior distribution over models in a model family can be obtained by hierarchically extending the model family with a parameterized account of how the models are generated. The demonstration of how hierarchical methods can be used to construct an informative prior over models in a model family focuses on the Varying Abstraction Model (VAM: Vanpaemel & Storms, 2008) as a case study. The VAM is a model family that encompasses a range of categorization models, including the familiar exemplar and prototype models, as well as various other models that can be considered intermediate between these two extremes.

After providing an introduction to the VAM family, it is shown how existing intuitions about the relative plausibilities of the models in the VAM can be formally translated into an informative prior distribution over these models. A key role is played by a process that, informed by psychological intuitions about the models in the VAM, details how these models are generated. Using ten previously published data sets from the category learning literature, it is illustrated how, just like in the baseball example, adding prior knowledge in the mix can smooth inference. Finally, the discussion touches upon the conceptual reason behind the

underuse of informative priors and reflects on the relation between informative priors and subjectivity.

Before we get started, I provide a second example of how prior knowledge can smooth inference, focusing on a simple coin tossing problem. Apart from an additional motivation of why including prior intuitions in inference is desirable, this example also provides the necessary Bayesian background for the more involved VAM example. The coin tossing example thus serves as a mini-tutorial for readers without a working knowledge of Bayesian methods.

2. Using an informative prior to smooth estimation

The Bayesian approach to estimation is most easily introduced using a simple coin tossing example (see also Griffiths, Kemp, & Tenenbaum, 2008; Lee & Wagenmakers, 2005; Lindley & Phillips, 1976). The probability of a sequence of n coin tosses containing h heads and $t = n - h$ tails being generated by a coin which produces heads with probability π is

$$h \sim \text{Binomial}(n, \pi), \quad (1)$$

with

$$\pi \in [0, 1]. \quad (2)$$

Based on an observed sequence of heads and tails, we would like to estimate the value of π .

2.1. The maximum a posteriori (MAP) estimate

Treating π as a random variable, the initial (i.e., before the data are collected) assumptions or knowledge about which values of π are likely and unlikely is represented in a distribution, $p(\pi)$. To reflect that this distribution is independent of observed data and thus can be expressed before data are observed, it is referred to as the prior. After observing data d , such as the number of heads in a coin tossing sequence, the prior distribution $p(\pi)$ is updated to the posterior distribution $p(\pi | d)$, representing the knowledge about π based upon the combination of the initial assumptions and the observed data. Updating takes place according to Bayes' rule:

$$p(\pi | d) = \frac{P(d | \pi)p(\pi)}{P(d)}, \quad (3)$$

where $P(d | \pi)$ is the likelihood, indicating the probability of the data given π , and $P(d)$ a normalizing constant given by

$$P(d) = \int_{\Omega_\pi} P(d | \pi)p(\pi)d\pi, \quad (4)$$

with Ω_π the range of π .

One commonly used method to obtain a point estimate¹ from the posterior distribution is maximum a posteriori (MAP) estimation: choosing the value of π that maximizes the posterior probability,

$$\pi_{\text{MAP}} = \arg \max p(\pi | d) = \arg \max \frac{P(d | \pi)p(\pi)}{P(d)}. \quad (5)$$

Note that, since $P(d)$ does not depend upon π , this term does not affect π_{MAP} and can be omitted in the computation of π_{MAP} ; i.e., $\pi_{\text{MAP}} = \arg \max P(d | \pi)p(\pi)$. Accordingly, computing π_{MAP} does not require the integral of (4) to be evaluated, but only requires an optimization algorithm.

¹ Because reducing the full distribution $p(\pi | d)$ to a single number like π_{MAP} discards useful information, such as the uncertainty about the estimate, it is often preferable to maintain the full posterior distribution.

Fig. 1. Prior, likelihood and posterior for $h = n = 2$, assuming that $\alpha = \beta = 1$ (panel A), $\alpha = \beta = 2$ (panel B) and $\alpha = \beta = 10$ (panel C).

2.2. The Beta binomial

As a concrete example, let us assume the Beta distribution² as the prior distribution over π :

$$\pi \sim \text{Beta}(\alpha, \beta). \quad (7)$$

The Beta distribution is often used as a prior since it can express widely different expectations. Using the prior from (7) with the likelihood from (1), the posterior distribution over π produced by a sequence with h heads and t tails becomes³

$$\pi | h \sim \text{Beta}(h + \alpha, t + \beta), \quad (8)$$

which implies that the maximum a posteriori estimate of π equals

$$\pi_{\text{MAP}(\alpha, \beta)} = \frac{h + \alpha - 1}{h + t + \alpha + \beta - 2}. \quad (9)$$

The MAP estimate reflects the influence of, and provides a compromise between, both the observed data (in terms of h and t) and the prior knowledge (in terms of α and β).

2.2.1. Uniform prior

A prior that is often used, both explicitly and implicitly, is the uniform distribution over π (i.e., $\alpha = \beta = 1$), expressing the belief that all values of π are a priori equally likely to have generated the data. The mode of the posterior distribution, $\pi_{\text{MAP}(1,1)}$, is then given by⁴

$$\pi_{\text{MAP}(1,1)} = \frac{h}{h + t}. \quad (10)$$

This estimate is not always free from absurdity. For example, based on a sequence of heads only (i.e., $t = 0$), one would infer that the

coin has two heads, and will never produce tails ($\pi_{\text{MAP}(1,1)} = 1$; see Panel A of Fig. 1). After observing a very long all-heads sequence, concluding that the coin has two heads might be reasonable, but $\pi_{\text{MAP}(1,1)} = 1$ for a sequence of any length. Earlier experience with coins suggests that a normal, randomly chosen coin tends to be (nearly) fair, so after seeing a coin producing, say, two heads in a row, it would be reasonable to conclude that the coin is approximately fair rather than two-headed.

Inferring, based on a short all-heads sequence, that the coin is two-headed seems highly implausible, and is probably false, just like estimating a final score of 45–0 after a first inning score of 5–0 was unreasonable and false. The same remedy that improved the implausible baseball estimate can also be applied to adjust the implausible coin toss estimate: adding prior knowledge in the mix. In particular, taking into account the prior knowledge that coins tend to be nearly fair will vastly improve the estimate.

2.2.2. Informative prior

The inference using the uniform prior can lead to conclusions that are not in line with existing knowledge about coins. The reason is that nowhere in the inference does this knowledge play a part—it is simply ignored by the uniform prior. Knowledge or intuitions about what is plausible can be expressed by means of an informative prior, which, unlike the uniform prior, tends to be narrow and sharply peaked. In particular, the expectation that coins are nearly fair can be expressed by centering the Beta distribution of (7) on $\frac{1}{2}$, which can be achieved by setting $\alpha = \beta > 1$. The strength of the expectation is governed by the size of $\alpha = \beta$: the larger $\alpha = \beta$, the stronger the expectation that the coin is fair.

Even a prior with a very weak expectation of the coin being fair will improve the estimate. If $\alpha = \beta = 2$, seeing a sequence of two heads in a row leads to an estimate of $\pi_{\text{MAP}(2,2)} = \frac{3}{4} = 0.75$ (see Panel B of Fig. 1). This estimate is, like $\pi_{\text{MAP}(1,1)}$, tilted towards heads, but is a far less extreme one. The implausibility of the estimate of π can further be reduced by using a prior that encodes a stronger belief about the value of π . A Beta prior with $\alpha = \beta = 10$ more strongly favors values of π close to $\frac{1}{2}$. Using this prior, seeing a sequence of two heads in a row results in an estimate of $\pi_{\text{MAP}(10,10)} = \frac{11}{20} = 0.55$ (see Panel C of Fig. 1). This estimate again clearly reflects the influence both of the prior expectations (towards $\frac{1}{2}$) and of the observed data (towards 1).

The above example highlights the role of informative priors in estimation. When estimation relies upon a uniform prior, which does not favor any values of π over others, the prior knowledge that coins tend to be fair is ignored. Ignoring this knowledge implies that, based on a sequence of two heads in a row, one would reach the highly implausible conclusion that the coin has two heads. When estimation relies upon an informative prior,

² The Beta distribution is parameterized by two positive parameters, α and β , and, for $x \in (0, 1)$, has a probability density function given by

$$p(x | \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad (6)$$

where the normalizing constant $B(\alpha, \beta)$ is known as the Beta function. $\text{Beta}(\alpha, \beta)$ has a mode at $\frac{\alpha-1}{\alpha+\beta-2}$ (if $\alpha > 1$ and $\beta > 1$), and a variance of $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, indicating that the distribution gradually becomes more centered on the mode as $\alpha + \beta$ becomes large. The Beta distribution contains the uniform distribution on $(0, 1)$ as a special case, when $\alpha = \beta = 1$.

³ Note that the prior and the posterior distributions are of the same type, both being Beta distributions. When the likelihood does not change the form of the distribution going from the prior to the posterior, the prior is said to be conjugate to the likelihood (see, e.g., Diaconis & Ylvisaker, 1979).

⁴ This is, of course, the standard maximum likelihood estimate. More generally, from (5) it follows that, when $p(\pi)$ does not depend on π , $\pi_{\text{MAP}} = \arg \max P(d | \pi)$, which is the standard definition of the maximum likelihood estimate.

taking into account the knowledge about what is plausible, the implausibility of the estimate is drastically reduced. Thus, relying on an informative prior, expressing the knowledge that coins tend to be approximately fair, guards the inference against implausible estimates.

2.3. Balancing expectations and data

Inferring an extreme final score based on the one-inning score in the baseball game, and inferring that a coin that has produced two heads in a row will always produce heads is problematic for the very same reason: the inference is based on a few observations only (one inning, two tosses). Much more than large data sets, small data sets are riddled with a high degree of sampling variability. In an estimate that does not take intuitions about what is plausible into account, this sampling variability is heavily reflected and can lead to wildly implausible conclusions.

An obvious strategy that might improve the estimate is reducing the sampling variability by collecting more data. For example, after the eighth inning, with a score of, say, 6–4 for Kansas, one might still incorrectly predict Kansas to be the final winner, but not with an outrageous score of 45–0. There is a sense in which, in the above example, the informative prior exactly did this: adding more data. To see this, note that, according to (8), the posterior is influenced by both the observed data (h and t) and the prior (α and β). Crucially, from the point of view of the posterior distribution, there is no difference between starting with $\alpha = 5$ and observing $h = 2$ or starting with $\alpha = 2$ and observing $h = 5$. Indeed, α (and β) are pooled with the observed data h (and t) to produce the posterior. In effect, while h and t are the real, observed numbers of heads and tails, α and β act as virtual, imaginary observations of heads and tails. For the posterior distribution, it is as if both the real and virtual tosses had been observed in the same data set. Thus, the informative estimate proceeds *as if* more data have been observed, by pooling the real, observed tosses with the virtual, prior tosses. The equivalence between an informative prior and virtual data underscores that Bayesian inference can be seen as an information processing device, where information can derive from different sources. Information is not only provided by the observed data but also by the specification of the model, which includes the choice of the prior.

As increasing numbers of data are observed, the influence of an individual data point (say, the outcome of the second toss) on the final estimate decreases. This observation holds not only for real data, but also for virtual data. Consequently, in its capacity of corresponding to virtual data, the influence of the prior also decreases with an increasing number of observed data. For example, seeing a sequence of 1000 heads in a row results in an estimate $\pi_{\text{MAP}(2,2)} = \frac{1001}{1002} \approx 0.999$ or $\pi_{\text{MAP}(10,10)} = \frac{1009}{1018} \approx 0.991$, which is nearly identical to the estimate $\pi_{\text{MAP}(1,1)} = 1$. For a data set this large, the exact prior hardly matters. Although a fair coin was expected, the observed data were so indicative about the biasedness of the coin that one is forced to give up this expectation. The data have overwhelmed the prior.

This example highlights that adding prior knowledge, in the form of an informative prior distribution, is not intended to totally remove possibilities from consideration (such as an unfair coin) or to override the evidence provided by the data. Rather, Bayesian estimation assuming an informative prior strikes a balance of having expectations (coins tend to be fair) without losing flexibility by simply assuming the expectations to be true (coins are necessarily fair). Given overwhelming data, such as 1000 tosses, the prior intuitions will be overturned where inappropriate. Given sparse data, such as two tosses, the prior intuitions will help guide more intuitive inferences. In sum, the danger of reaching implausible conclusions, and hence the benefit of adding knowledge about what is plausible to the estimation, is especially high when data sets are small.

2.4. Conclusion

Both the baseball and the coin examples highlighted that, without adding prior knowledge, unreasonable and implausible conclusions are hard to avoid, especially with small data sets. The Bayesian framework offers a complete and coherent approach for balancing empirical data and prior expectations. Taking relevant prior knowledge into account, expressed in an informative prior, can correct extreme and implausible estimates, and adjust them toward more reasonable and plausible ones. The intuitions expressed in the informative prior act to smooth the estimations based on the observed data, damping out random fluctuations. Smoothing is the power of the informative prior in estimation.

The remainder of this article gives a worked example of an informative prior improving inference in a category learning model family. The same basic property of hierarchical models is used a number of times in this special issue, especially by Morey (2011), who shows the inferential advantages of including additional relevant theory in a hierarchical extension of a basic memory model.

3. The Varying Abstraction Model (VAM)

A fundamental debate in the category learning literature concerns the role of abstraction in representing categories, often taking the form of a dichotomous juxtaposition (e.g., Murphy, 2002; Smith & Medin, 1981). On the one hand, exemplar theorists (Brooks, 1978; Medin & Schaffer, 1978; Nosofsky, 1986) speculate that category learning does not involve abstraction—people learn a category by learning the category members themselves (i.e., the exemplars). According to prototype theorists (Minda & Smith, 2001; Reed, 1972), on the other hand, people learn a category by extracting abstract, category-level information from the category members. Most often, this summary information (i.e., the prototype) is defined as the average of all exemplars.

Vanpaemel and Storms (2008) argued that focusing on the exemplar and prototype representations only when investigating abstraction in category learning provides a limited window on abstraction, and ignores a wealth of other representations that are psychologically viable. The exemplar and prototype representations correspond to extreme levels on the continuum of abstraction (minimal and maximal, respectively), but there seems to be no principled reason to exclude the possibility that people rely on partial abstraction (see Love, Medin, & Gureckis, 2004, for a similar observation). Vanpaemel and Storms (2008) then proposed a set of models that intermediate between the prototype and exemplar models, formalizing the possibility that abstraction is partial. The resulting family of models, including the exemplar, prototype and all the intermediate models, is referred to as the Varying Abstraction Model (VAM), to reflect that the model family spans a range of models assuming different levels of representational abstraction.

3.1. The models in the VAM family

All models in the VAM family can be considered variations to the Generalized Context Model (GCM: Nosofsky, 1984, 1986), a very successful exemplar model of category learning. The VAM adopts all but one of the assumptions of the GCM. In particular, while the GCM assumes a single representation (the exemplar representation, consisting of all category members), the VAM relaxes this assumption and considers all possible representations that can be obtained by merging subsets of category members.

Fig. 2. The 15 possible representations for a category with four category members considered by the VAM, split up in four levels of abstraction. The two extreme levels of abstraction correspond to the exemplar representation (top) and the prototype representation (bottom). The representations shown in the second and third rows correspond to partial abstraction.

3.1.1. Representations

Following the GCM, it is assumed that stimuli can be represented by their values along underlying stimulus dimensions, as points in a multidimensional psychological space. Such a multidimensional representation is typically derived from identification confusion data or from similarity ratings using multidimensional scaling (MDS). In this space, similar stimuli lie close together, whereas dissimilar stimuli lie far apart. A visualization of a category with four members, each represented by a two-dimensional point in the psychological space, is provided by the top panel of Fig. 2. Together, this collection of the four members makes of the exemplar representation.

Apart from the exemplar representation, the VAM also considers the prototype representation, in which all four category members are merged into a single item, as shown in the bottom panel of Fig. 2. This prototype is denoted in black, and the original category members are shown in white, joined by lines to their merged item. Additionally, the VAM defines a host of new representations, that are intermediate between the two extremes. The full set of additional representations that are possible are shown in the two intermediate rows. The second row shows all the representations that can be created when two of the category members are merged, resulting in a representation with three items, shown in black. The third row shows all representations that result from two merges, leaving a representation consisting of two items, again shown in black. Each of these 15 representations is considered by the VAM.

3.1.2. Models

Adding the category learning processes and mechanisms of the GCM to a representation for each category results in a categorization model. The formal details of the processing assumptions of the GCM (and the VAM) are not relevant for the present demonstration of how intuitions about models can be formally captured in an informative prior distribution, and are given in the Appendix. If no category members are merged (i.e., all subsets are singletons), the GCM results. If all members of a category are merged into a single

item (i.e., there is one subset per category), the MDS-based prototype model (MPM) results. Intermediate models arise by merging some subsets of the category members, and can take the form of multiple prototype models (see also Nosofsky, 1987).

In sum, the VAM is a family of categorization models, each of which has the same free parameters, taken from the GCM. All models of the family are identical to each other, except for the category representation they assume. The VAM includes the familiar exemplar and prototype models, as well as a wide variety of models that are intermediate between these extremes. The number of models in the VAM, denoted by K , is countable and depends on the number of members each category has. For example, in an experimental design with two categories containing four members each, the VAM contains $15 \times 15 = 225$ different parameterized models. The models in the VAM family will be indexed by M_i , where $i \in [1, 2, \dots, K]$.

3.2. The posterior distribution over M_i

It is useful to keep in mind that, although equipped with more nuts and bolts, the VAM is conceptually no different than the simple coin model presented in the previous section. In the VAM, every different value of the model index M_i corresponds to a different model. In the coin model, every different value of the parameter π can be regarded as corresponding to a different model. Hence, also the coin model could be regarded as a model family, containing (uncountably) infinitely many models. The VAM family's model index M_i and the coin model family's parameter π play the exact same role.⁵

⁵ Unlike the coin model family, which contains an (uncountably) infinite number of models, the VAM contains a finite number of models only (e.g., 225 for an experimental design with two categories with four members each). This difference relates to the fact that π is a continuously varying parameter, whereas the index M_i is a discretely varying parameter. The fact that, unlike π , M_i is discrete is reflected in the subscript i .

Just like the coin model family can be used to infer from observed data which value of π is the most probable to have generated the data (or which values of π , if the full posterior distribution is maintained), the VAM can be used for estimation—inferring from observed data which value(s) of M_i (i.e., which model(s) of the VAM family) is (are) the most likely to have generated the data. Using the VAM to estimate M_i from data d observed in a category learning task is conceptually identical to using the coin model to estimate π from the observed number of heads in a sequence of coin tosses. Estimation proceeds by calculating the posterior distribution, which can be obtained using Bayes' rule (see (3)). Rather than over π , the distribution of interest is the posterior distribution over M_i :

$$P(M_i | d) = \frac{P(d | M_i)P(M_i)}{P(d)}. \quad (11)$$

As with the coin model, the posterior distribution requires three ingredients. The first is the normalizing constant $P(d)$, which is given by

$$P(d) = \sum_{i=1}^K P(d | M_i)P(M_i). \quad (12)$$

The normalizing constant is computed by summation rather than integration as in (4), because, unlike π , M_i is discrete.

The second ingredient to calculate the posterior is the likelihood $P(d | M_i)$. In the coin tossing example, the likelihood is straightforwardly defined by (1). In the VAM, the likelihood is somewhat more complicated. In the coin model family, the model that results at a given value of π is a parameter-free point hypothesis, and it therefore makes unambiguous (albeit probabilistic) predictions. In contrast, in the VAM family, the model that results at a given value of M_i contains free parameters (in particular, the ones adopted from the GCM). For a parameterized model as M_i , $P(d | M_i)$ is known as the marginal likelihood, and it can be obtained by integrating out, or marginalizing over, the parameters (see, e.g., Lee, 2004; Myung & Pitt, 1997):

$$P(d | M_i) = \int_{\Omega_\tau} P(d | \tau, M_i)p(\tau | M_i)d\tau, \quad (13)$$

where Ω_τ indicates the prior range of the parameter (vector) τ , $p(\tau | M_i)$ indicates the prior distribution over τ , and $P(d | \tau, M_i)$ indicates the likelihood. The range, prior, and likelihood are discussed in the Appendix.

The third part in the calculation of $P(M_i | d)$ is the prior distribution $P(M_i)$. One particularly easy, and to many appealing, way to compute $P(M_i)$ is to assume equal prior probabilities for all models indexed by M_i (i.e., $P(M_i)$ is the uniform distribution). Most current modeling in psychology, explicitly or most often implicitly, relies on this uniform assumption. However, as illustrated in the coin tossing example, taking prior knowledge into account in estimation, in the form of an informative prior distribution, can adjust implausible estimates and improve inference. In the context of the VAM as well, adding prior information through an informative prior distribution over M_i has the desirable effect of smoothing the inferences made by the VAM, by avoiding implausible estimates.

While in the coin tossing example the translation of the expectations about π is fairly straightforward, by setting α and β to specific values in (7), this is much less so for M_i . It is at this point where hierarchical methods come in useful: as a way to formally translate theory or intuitions into a non-uniform, informative prior distribution over models. The next section illustrates how the hierarchical introduction of a theoretically meaningful parameterized generative process, motivated by existing psychological intuitions about the relative plausibilities of the models, naturally gives rise to an informative prior distribution $P(M_i)$.

4. Constructing an informative prior

Using a hierarchical extension to formally translate theory or intuitions about models in a model family into a prior model distribution involves three steps (e.g., Lee, 2006). The first step involves deciding which intuitions and expectations about the models under consideration seem reasonable. The second step is, inspired by these intuitions, to define a process that generates the models. This generative process then automatically imposes a theoretically informed and psychologically meaningful, non-uniform prior distribution over the models. These three steps are illustrated for the VAM, taking the lead from Lee and Vanpaemel (2008).

4.1. Priors intuitions about the models in the VAM family

The first step of the hierarchical extension of the VAM is articulating which models are intuitively more likely than others. To this end, it is useful to note that the models in the VAM family differ in two aspects. First, they differ in their level of abstraction, roughly reflected by *how many* category members are merged. Second, at a given level of abstraction, the extreme ones aside, different models are possible, differing in *which* members exactly are merged. In the present demonstration, two intuitions will be formalized, one about each aspect.

As far as the number of merges is concerned, there exists ample evidence for prototype and exemplar models, suggesting they are viable accounts of category learning. Therefore, these models are given a high prior mass, inversely proportional to the number of abstraction levels implied by the experimental design. As far as the type of merges is concerned, it seems reasonable to expect that merging is driven by similarity. Taking the category *fruit* as an example, one could sensibly expect that similar members, such as two different lemons or a lemon and an orange, are merged rather than dissimilar members, such as a lemon and a banana. In the context of the VAM, the model at the very right end of the second row in Fig. 2 seems intuitively more likely than its immediate (left) neighbor. This bias towards similarity is a core assumption of other models that are similar to the VAM, like SUSTAIN (Love et al., 2004) and the Rational Model of Categorization (Anderson, 1991; Griffiths, Canini, Sanborn, & Navarro, 2007).

4.2. Generating the models in the VAM family

The second step of the hierarchical extension of the model family is to capture these prior intuitions in a process for generating the models. In this application, the generative process starts with the exemplar model (at the top of Fig. 2) and, because it is inspired by two intuitions, it has two parts and two hierarchical parameters. How many merges are made – how deep the process continues in Fig. 2 – is controlled by the first part of the process. When an additional merge is undertaken, different merges are possible, as indicated by the six (or seven) possibilities in the second (or third) row of Fig. 2. Which members are merged – how far to the left or the right the process continues in Fig. 2 – is controlled by the second part of the process. Consistent with the intuition that merging is driven by similarity, the probability of being merged should depend on the similarity of the black items (i.e., the original category members or their merged replacement) in a current model.

Formally, the probability that an additional merge will take place, and the process will continue, is instantiated by a parameter $0 \leq \theta \leq 1$. This means that, at any stage of the process, there is a $1 - \theta$ probability that the current model will be maintained as the final one. If θ is close to 1, merging is very likely, and the prototype model almost always results. If, instead, θ is close to 0, merging is very unlikely, and the exemplar model is almost always retained.

Further, the probability of choosing to merge the pair of items (i, j) is given by an exponentiated Luce choice rule,

$$p_{ij} = \frac{s_{ij}^\gamma}{\sum_x \sum_{y \geq x} s_{xy}^\gamma}, \quad (14)$$

where s_{ij} , the similarity between the i th and j th item, is modeled as an exponentially decaying function of the Minkowski r -metric distance between their points, $s_{ij} = \exp[-\sum_k (|v_{ik} - v_{jk}|^r)^{1/r}]$, with v_{ik} the coordinate location on the k th dimension for the point that represents the i th item. The parameter $\gamma \geq 0$ controls the level of emphasis given to similarity in determining the pair to be merged. When $\gamma = 0$, similarity is not taken into account, so all pairs of items are equally likely to be merged. Since the pair that is merged is chosen at random, all models within a given level of abstraction are equally probable. As γ increases, the maximally similar pair dominates the others, and will be chosen as the pair to be merged with probability approaching one. For example, when γ is very large (e.g., $\gamma = 10$), only the most similar items are merged. This would imply that, in the second row of Fig. 2, the model to the far right will be the only one that is generated at that level of abstraction. Values of γ between these extremes result in intermediate behavior. In the second row of Fig. 2, the model to the far right will be the most likely one, but also other models at the same level of abstraction will have a non-negligible probability of being generated.

4.3. The prior distribution over the models in the VAM family

The generative process assigns, for each value of θ and γ , a probability to each of the models in the VAM family, $P(M_i | \theta, \gamma)$. In the third and final step of the construction of the informative prior, the prior distribution over M_i is obtained by integrating out the hierarchical parameters θ and γ , weighted by their prior:

$$P(M_i) = \int_{\Omega_\theta} \int_{\Omega_\gamma} p(M_i | \theta, \gamma) p(\theta, \gamma) d\theta d\gamma. \quad (15)$$

Thus the model prior $P(M_i)$ is the average of all of the distributions that result from some combination of θ and γ , weighted by the priors for θ and γ . In the present applications, θ and γ were assumed to be independent: i.e., $p(\theta, \gamma) = p(\theta)p(\gamma)$. For γ , it was assumed that

$$\gamma \sim \text{Gamma}(2, 1), \quad (16)$$

which gives support to all positive values, but has most density around the modal value one, corresponding to the prior expectation that similarity plays an important role in merging. For θ , it was assumed that

$$\theta \sim \text{Beta}(3.2, 1.8). \quad (17)$$

Consistent with prior intuitions, this distribution assures that the prototype and exemplar models both have a prior mass of roughly $\frac{1}{6}$, reflecting the number of levels of abstraction in an experimental design involving two categories with four members each.⁶

In sum, the distribution $P(M_i)$ falls out of the way in which the generative process works, and the priors on the hierarchical parameters that control the process. This means that it directly follows from theoretical assumptions about how the models in the VAM are generated, and is fully specified before any data have been observed, or inferences have been made.

5. Application to data from category learning tasks

In this section, the VAM is used to estimate M_i based on ten previously published data sets. All data sets were taken from three seminal articles presented by Nosofsky and his collaborators (Nosofsky, 1986, 1987; Nosofsky, Clark, & Shin, 1989). The category learning tasks involve learning various two-category structures over a small number of stimuli varying on two dimensions (for example, color chips varying in saturation and brightness). In each task, a subset of the stimuli is assigned to categories A and B, and the remaining stimuli are left unassigned. Most tasks consist of a training phase, in which the category structure is learned, followed by a test phase. During the training phase, only the assigned stimuli are presented, and the participant classifies each presented stimulus into either category A or category B. Following each response, corrective feedback is presented. During the test phase, both the trained and the untrained stimuli are presented. The human data used in the estimation of M_i are the categorization responses to the stimuli presented in the test phase. Some details about these data sets are provided in Table 1.

To illustrate how the hierarchical extension to the VAM and the informative prior over the models it implies make a difference in the estimation, I focus on two versions of the VAM. The first one, VAM_{uni}, assumes a uniform prior over the models, so that each model of the VAM is a priori equally likely (formally, $P(M_i) = \frac{1}{K}$). The second version is VAM_{sim}, which assumes the informative prior obtained by the generative process described in the previous section. According to VAM_{uni}, all levels of abstraction are equally likely, and all sorts of merging are equally likely. VAM_{sim}, in contrast, captures two intuitions: extreme levels of abstraction are more likely than intermediate ones; similarly-based intermediate models are more likely than intermediate models based on the merging of dissimilar category members.

These biases are illustrated in Fig. 3, showing, for data set 6, the 13 models with the highest prior mass. The bottom row of Fig. 3 shows the models, using the graphical conventions adopted earlier (i.e., the items representing the category are shown in black and are connected by lines to the original category members, shown in white). Squares relate to category A and circles relate to category B. The top row of Fig. 3 indicates, by the height of the bar, the prior mass the model is given under VAM_{sim}. The four models with the highest mass, on the left, show the bias towards extreme levels of abstraction, corresponding to the exemplar model, the prototype model, and the two hybrid mixture models where one category has an exemplar representation and the other one has a prototype representation. The remaining nine models show the bias towards similarity-based merging. If merging takes place, it always involves stimuli that lie very close, and are thus similar to each other. For example, the stimulus represented by the square in the left bottom corner, which is clearly separated from the three other stimuli in the same category, is only merged if this merging results in the prototype model.

The aim of the application is to demonstrate the effect of an informative prior in estimation, rather than providing insight into category learning. The substantive insights that follow from the application of the VAM, as well as the implications for the prototype versus exemplar debate, are discussed elsewhere (Lee & Vanpaemel, 2008; Vanpaemel & Storms, 2008).

5.1. Results

The primary interest concerns the inferences that can be drawn from the posterior distribution $P(M_i | d)$.⁷ As noted earlier, the full

⁶ This is the design of seven of the ten data sets considered in the rest of this article. The remaining three data sets involve one category with three members and another category with four members, and required a Beta(2.4, 1.5) prior for θ , assigning a prior mass of $\frac{1}{5}$ to both the exemplar and prototype models.

⁷ Note that the parameter θ can be readily interpreted as a measure of the extent of abstraction, and γ as a measure of the reliance on similarity in forming

Table 1
Details of the ten data sets.

Data Set	Reference	Experiment	Stimuli	r	α	Categories	K	Condition	N	n	s
1	Nosofsky (1986)	–	Shepard circles	2	2	Diagonal	225	Participant 1	150 ^m	250 ^a	1
2	Nosofsky (1986)	–	Shepard circles	2	2	Diagonal	225	Participant 2	150 ^m	225 ^a	1
3	Nosofsky (1986)	–	Shepard circles	2	2	Dimensional	225	Participant 1	150 ^m	225 ^a	1
4	Nosofsky (1986)	–	Shepard circles	2	2	Dimensional	225	Participant 2	150 ^m	200 ^a	1
5	Nosofsky (1987)	2	Munsell colors	2	1	Criss-cross	225	–	–	10	24
6	Nosofsky (1987)	2	Munsell colors	2	1	Saturation A	225	–	–	7.5 ^a	24
7	Nosofsky (1987)	2	Munsell colors	2	1	Saturation B	225	–	–	10	40
8	Nosofsky et al. (1989)	1	Shepard circles	1	1	Interior–exterior	75	Free	43 ^a	5	122
9	Nosofsky et al. (1989)	2	Shepard circles	1	1	Interior–exterior	75	Rule 1	43 ^a	5	30
10	Nosofsky et al. (1989)	2	Shepard circles	1	1	Interior–exterior	75	Rule 2	43 ^a	5	28

Notes: r = metric; α = similarity function; K = number of models in the VAM family; N = number of trials per stimulus per participant in the training phase; n = number of trials per stimulus per participant in the test phase; s = number of participants; ^m = minimal value (actual value unreported); ^a = approximate average value.

Fig. 3. Models with the highest prior mass, indicated by the height of the bar, for data set 6.

posterior distribution contains more information than a point estimate, such as the MAP estimate. Nevertheless, since the purpose of this application is illustrative, only the MAP estimate (i.e., the model of the VAM family with the highest posterior mass) is reported. Focusing on the MAP estimate rather than on the full distribution facilitates the comparison between estimation using an informative prior and a uniform prior. The MAP estimates made by VAM_{uni} and by VAM_{sim} for the ten data sets are shown in Fig. 4.⁸

5.1.1. The informative prior accords with the data

The models inferred by VAM_{uni} based on data sets 1, 5, and 8 seem very plausible. All models rely on the merging of one or two pairs of similar category members. As the inferred models comply with the intuitions about the models captured in VAM_{sim} , the inference does not change when these prior intuitions are added to the estimation. The inferences made by VAM_{sim} are identical to the inferences made by VAM_{uni} .

5.1.2. The informative prior overrides the data

The inferences for five data sets illustrate how the intuitions captured in the informative prior of VAM_{sim} result in more plausible estimates than those made by VAM_{uni} , which lacks such an informative prior. For example, the effect of the preference in VAM_{sim} for the exemplar model is evident in the inference based on data set 7. The adjusting influence of the bias towards extreme levels of abstraction and similarity-based models is demonstrated best by considering data sets 2, 3, 4, and 6. The models inferred by VAM_{uni} as the most likely ones to have generated these data sets are counterintuitive in the sense that they involve merging of category members that are very dissimilar to each other (an observation also made by Vanpaemel and Storms (2008) in relation to data set 6). VAM_{sim} adjusts these implausible estimates and pulls them towards the intuitions captured in its prior. Specifically, VAM_{sim} infers the prototype model (data set 2), the exemplar model (data set 3), and highly plausible intermediate models, based on the merging of similar members (data sets 4 and 6).

5.1.3. The data override the informative prior

The above examples highlighted how the informative prior of VAM_{sim} has smoothed the inference. Of course, the intuitions captured in the prior of VAM_{sim} might be wrong ones. As noted earlier, the goal of adding the informative prior is not to try to override the evidence provided by data. Rather, the intuitions captured in the prior should be overturned by the data when inappropriate, just like observing 1000 heads in 1000 tosses overturned the expectation that the coin is fair.

The intuition that merging is driven by similarity has not eliminated all models where merging is not driven by similarity

category representations. The posterior distributions of these parameters, $p(\theta | d)$ and $p(\gamma | d)$, are of interest as well, since they directly convey the conclusions that can be drawn about the use abstraction and the role of similarity. Details and demonstrations of this style of analysis can be found elsewhere (Lee & Vanpaemel, 2008; Vanpaemel & Lee, 2007).

⁸ From (11) it follows that, under VAM_{uni} , the model with the highest posterior mass coincides with the model with the highest marginal likelihood, since both $P(d)$ and $P(M_i)$ do not depend on the model M_i . This is, of course, equivalent to the fact that, in the coin tossing example, the maximum a posteriori estimate assuming a uniform prior over π coincides with the maximum likelihood estimate.

Fig. 4. MAP estimates under VAM_{uni} (left columns) and VAM_{sim} (right columns).

from consideration in VAM_{sim} . The bias towards similarity-based merging is not operationalized by assigning models based on the merging of dissimilar members a prior mass of zero, but rather by assigning such models a lower prior mass than the models in which similar members are merged. Assigning a non-zero mass to implausible models assures that, if unexpectedly the data turn out to provide evidence for an implausible model, the prior can be overwhelmed by the data.

Data sets 9 and 10 present two cases where the data swamp the prior. The inferences made by VAM_{uni} reveal quite counterintuitive models as being the most likely ones, as these models involve the merging of very disparate category B members. However, unlike with the counterintuitive models inferred for data sets 2, 3, 4, and 6, VAM_{sim} infers the same models as VAM_{uni} . The evidence in favor of these counterintuitive models provided by the data has outweighed the bias against these models.⁹

5.2. Conclusion

When the evidence provided by the data in favor of implausible models was weak, the intuitions about which models are plausible overturned the evidence. The informative prior in VAM_{sim} guarded the inference against implausible estimates. When the evidence provided by the data in favor of implausible models was strong, the intuitions about which models are plausible did not overcome the evidence. The informative prior in VAM_{sim} was overwhelmed by the evidence provided by the data. When the evidence provided by the data favored plausible models, the intuitions about which models are plausible had hardly any effect. In sum, the informative prior of VAM_{sim} strikes a balance of having theoretically based expectations (levels of abstraction tend to be extreme; merging tends to be driven by similarity) without losing flexibility by simply assuming the basic tenets of those theories (levels of abstraction are always extreme; merging is always driven by similarity).

6. General discussion

Not all models in the VAM family seem equally plausible, irrespective of their ability to capture behavioral data. In particular,

⁹ It is informative to make sense of why the preference for similarity-based merging has been overcome by data. One plausible reason is that data sets 9 and 10 involve participants being told to use a certain rule to learn the category structure. This instruction contrasts with typical category learning tasks, where no special instructions are given about the strategy to be used.

the VAM contains the exemplar and prototype models, for which much previous evidence exists. Further, it includes models relying on very dissimilar category members being merged, as well as models in which similar category members are merged together. These latter models seem intuitively more likely than the former (e.g., Love et al., 2004). A detailed demonstration was provided of how these prior biases for some models over others can be formally translated in a non-uniform, informative prior distribution over the models in the VAM family. The key to constructing the informative prior was the introduction of a higher-level parameterized process of how the models in the VAM are generated. This generative process, together with sensible priors on the parameters that control the process, automatically gave rise to the desired prior distribution.

Incorporating prior intuitions through the informative prior had the desirable effect of smoothing, i.e., damping out random fluctuations in the data due to sampling variability. The informative prior corrected implausible estimates, adjusting them toward more reasonable ones. Crucially, the informative prior did not sacrifice flexibility: when inappropriate, the informative prior was overturned by the data.

To understand the goal of bringing the VAM in contact with observed data, it is helpful to think back to the conceptual equivalence of the coin model and the VAM. Connecting the coin model (family) with data is not informative about whether the coin model is viable. The goal is estimation (of a parameter in a model, or equivalently of a model in a model family), not testing.¹⁰ Analogously, the applications of the VAM are not intended to test whether the intuitions captured in VAM_{uni} or in VAM_{sim} are supported by the data.¹¹

¹⁰ One approach to test the coin model is the computation of the posterior distribution of the model, rather than of the parameter π .

¹¹ One approach to test the VAM is the computation of the posterior distribution of the model family, rather than of the models M_i :

$$P(VAM | d) = \frac{P(d | VAM)P(VAM)}{P(d)}, \quad (18)$$

where

$$P(d | VAM) = \sum_{i=1}^K P(d | M_i, VAM)P(M_i | VAM). \quad (19)$$

As before, $P(M_i | VAM)$ and $P(d | M_i, VAM)$ indicate the prior probability and the marginal likelihood of model M_i , respectively.

Further, it is important to realize that the generative process is not intended to represent the literal description of the psychological process people go through when learning a category, neither does it reflect a normative goal of what people should be doing. Rather, it is a technical process, which details how the different models encompassed by the VAM can be formed. For example, the fact that the process starts with the exemplar model and then, potentially, evolves into more abstract models, should not be given the psychological interpretation that, when people are faced with a new category learning task, they rely on an exemplar representation at first, but, as learning matures, start to adopt a more abstract one. In fact, it has been suggested that people do the exact opposite (Smith & Minda, 1998; Vanpaemel & Navarro, 2007). Thus, while the end result of the process – the prior distribution over the models – is intended to capture psychological intuitions, the process itself is just a means to get there.

Finally, the approach to construct an informative prior over models in a model family by introducing a hierarchical generative account of the models is generally applicable, provided that there are intuitions available. The hierarchical extension to a model family provides an avenue for formalizing intuitions, but does not provide a way to come up with intuitions. It cannot resolve the indecision of theorists about which intuitions they hold. Like model building in general, constructing an informative prior distribution over models cannot be automated, but requires effort, care and thought.

6.1. Informative priors and subjectivity

The applications showed how the informative prior of VAM_{sim} smoothed the inference without losing useful flexibility. Despite this important feature, informative priors have not been uncontroversial, to say the least. In particular, it has been argued that using an informative prior does not let the data speak for themselves, and makes inference subjective and “unreliable, impractical and not scientific” (Wasserman, 2000, p. 106). The uniform prior of VAM_{uni} , in contrast, does not a priori favor any models over others, and might therefore seem more objective, and more applicable for scientific analyses.

Unlike the prior of VAM_{sim} , the uniform prior of VAM_{uni} can be seen as non-informative.¹² The connection between informative priors and subjectivity on the one hand, and between non-informative priors and objectivity on the other, is so strong that informative priors are often referred to as “subjective priors” whereas non-informative priors are sometimes honored with the epithet “objective” (e.g., Berger, 2006). From this perspective, the inferences made by VAM_{uni} are objective, unlike those made by VAM_{sim} .

6.1.1. The illusion of objectivity

There is no denying that inference assuming an informative prior is subjective and does not let the data speak for themselves. However, in spite of its name, it is a mistake to assume that inferences based on objective priors are, in fact, objective. Sadly, “calling things objective does not make them so, and is dishonest when they are not so” (O’Hagan, 2006, p. 445).

Consider, for example, the common practice of contrasting the MPM and the GCM (see Nosofsky, 1992; Vanpaemel & Storms, 2010, for overviews). Implicitly, these analyses assume equal prior probabilities for both models, and therefore have never been seen

as problematically subjective. However, this objectivity is only an illusion. In the small model family spanned by the GCM and the MPM, giving both models equal prior mass corresponds to a uniform, non-informative prior. But from the perspective of the larger, encompassing model family provided by the VAM, the very same practice is equivalent to using a version of the VAM with a prior that is anything but uniform, giving weight to only two of the possible models, to the neglect of all other models. Inferences made by a model family with such a highly non-uniform prior would be seen as very subjective. Thus, what seems to be objective and unproblematic in one model family (MPM and GCM) is highly subjective and problematic in another model family (VAM). By the same token, the objectivity of VAM_{uni} is merely illusory. Inferences made by VAM_{uni} are as subjective (or as objective, for that matter) as those made by VAM_{sim} .

The trouble is, of course, that all inference relies on models and data, not just on data. Therefore, inference inevitably requires a decision of which models to consider, and different theorists might use different models for inference based on the same data (Romeijn & van de Schoot, 2008). Choosing which models to consider is not only a positive choice, but as the example with the MPM and the GCM illustrates, also a negative one, in the sense that any model choice inevitably excludes potential models from consideration. Model choice is therefore subject to what Kass and Wasserman (1996) refer to as the “partitioning paradox”. There might exist a model family in which a given model choice corresponds to a uniform model prior, but there is always another model family in which that same choice corresponds to a non-uniform model prior. In sum, it does not seem to make much sense to consider a uniform distribution over models objective, as it is only objective with respect to a subjectively chosen model family. Model choice causes all inference to be subjective at its core.

6.1.2. The reality of subjectivity

The realization that inference not only depends on data, but on models and data, implies that different reasonable model choices and formulations can yield substantially different conclusions. The inevitable subjectivity that permeates science and the scientific uncertainty it brings about might be seen as unfortunate by many, but is hard to deny. Press and Tanur (2001) debunk the myth of scientific objectivity from a historical perspective, by providing an account of how many giants of science were far from being objective.

The pill of subjectivity might be sugared to some extent by realizing that subjectivity should not be confused with, and should not provide an excuse for, sloppiness or arbitrariness. Unlike subjective decisions, arbitrary decisions are made without any theoretical justification, motivation or intuition. Arbitrary decisions can sneak into all aspects of inference, including data collection (e.g., the number of participants, the choice of the experimental design), model choice and formulation (which includes, but is not restricted to, the choice of the prior) and several technical pragmatics (e.g., the starting point of the optimization algorithm, or the number of samples in a Monte Carlo integration).

Being theoretically motivated, an informative prior does not necessarily correspond to an arbitrary whim. Assuming equal prior probabilities when contrasting the GCM and the MPM is subjective, but it should not be seen as arbitrary. Similarly, the informative prior assumed in VAM_{sim} is subjective, but non-arbitrary. In fact, as noted earlier, other theorists have argued for and relied upon intuitions very similar to those captured in the informative prior of VAM_{sim} (Anderson, 1991; Love et al., 2004). The fact that these intuitions are expressed in the prior does not make them more or less subjective.

It is the responsibility of the theorist to keep the number of arbitrary decisions to a minimum, by making the subjective choices

¹² This is not a general claim about the relationship between non-informativeness and uniformity. Non-informative priors should not necessarily be uniform, and uniform priors are not always non-informative (Kass & Wasserman, 1996).

non-arbitrary and well motivated. Further, where possible, the robustness of the conclusions against any arbitrary decision that was impossible to avoid should be checked. Finally, given that subjectivity permeates all aspects of science, it is essential to make sure that all subjectivity is transparent, fully in view, and available for scrutiny. It speaks for the informative Bayesian approach that it makes (part of) the inevitable subjectivity explicit and precise, in contrast to many non-Bayesian approaches relying on implicit and default assumptions.

6.2. Model expansion and extension

In many debates in psychology, progress is sought by developing a limited number of models that adopt different theoretical positions, and evaluating these models on the basis of empirical data. These evaluations then lead to choosing one model to the exclusion of all others. Considering a limited number of possibilities only provides a narrow window on the possible ways in which people can engage in a task. Expanding the number of theoretical positions results in a freshening broadening of the debate, a strategy known as model expansion in statistics (see, e.g., Draper, 1995). Using a large model family rather than a limited set of models facilitates approaching the debate from a model estimation perspective rather than from a model selection perspective.

The hierarchical extension to a model family demonstrated in this article seems to be most valuable for the areas of modeling in psychology where the few models typically considered can be or have been sensibly expanded into a broader model family. Having expanded the models into a large model family, the hierarchical extension has the ability to generate psychologically meaningful, informative priors over the models and to tie the models together at a higher level in the hierarchy. Jointly, the vertical expansion and the horizontal hierarchical extension can result in a rich and powerful estimation tool that can be used to advance and deepen our understanding.

Acknowledgments

The research reported in this article was funded by Grant FWO G.0513.08. I wish to thank Gert Storms for his support, Jan-Willem Romeijn for his insightful and detailed comments, and Michael Lee for both.

Appendix. Formal details of the GCM and the VAM

In the experimental designs considered in this article, stimuli are represented as points in a two-dimensional space. Let p_{ik} denote the coordinate location on the k th dimension for the point that represents the i th stimulus presented in the test phase. Let v_{ik} denote the coordinate location on the k th dimension for the point that represents the i th category member. The assignment of the category members to the two categories, as defined by the category structure to be learned, is represented by indicator variables, so that $a_i = 1$ if the i th category member belongs to category A, and $a_i = 0$ otherwise.

The response probability for the i th test stimulus being chosen as a member of category A is determined by the similarity between the stimulus and both categories, according to a Luce choice rule¹³

$$r_i = \frac{s_{iA}}{s_{iA} + s_{iB}}. \quad (20)$$

¹³ A more general response rule with two additional parameters, the bias κ and the response scaling λ , has also been proposed: $r_i = \frac{\kappa s_{iA}^\lambda}{\kappa s_{iA}^\lambda + (1-\kappa)s_{iB}^\lambda}$ (Ashby & Maddox, 1993; Navarro, 2007). This more general form is not used in the present applications.

In the GCM, s_{iA} , the similarity between the i th test stimulus and category A, is computed by the sum of the similarities between the stimulus and all members of category A:

$$s_{iA} = \sum_j a_j s_{ij}, \quad (21)$$

and similarly

$$s_{iB} = \sum_j (1 - a_j) s_{ij}, \quad (22)$$

where j indexes all category members. The similarity between the i th test stimulus and the j th category member, s_{ij} , is modeled as an exponentially decaying function of their distance, d_{ij} :

$$s_{ij} = \exp\{-(cd_{ij})^\alpha\}, \quad (23)$$

where c is the sensitivity parameter measuring the rate at which similarity declines with distance, and α determines the shape of the function. The distance between the i th test stimulus and the j th category member is calculated according to the Minkowski r -metric:

$$d_{ij}^r = w(p_{i1} - v_{j1})^r + (1 - w)(p_{i2} - v_{j2})^r, \quad (24)$$

where w is the attention weight parameter measuring the relative emphasis given to the first stimulus dimension over the second.

The only way in which the VAM differs from the GCM is that it does not assume that a category is necessarily represented by the category members, but also considers other representations. These representations are formed by merging subsets (possibly singletons) of category members. This means that the representations considered in the VAM are made up of items that can correspond to the original category members, but also to the merged replacements of the category members (i.e., the black circles in Fig. 2). Formally, then, the similarity between a test stimulus and a category is computed as the sum of the similarities of that stimulus to the items in the VAM category representation, rather than to the category members. This implies that a_i represents the membership of the items rather than of the category members, that in (21) and (22) the index j runs over the items rather than over the category members, and that in (24) v_{ik} denotes the coordinate location on the k th dimension for the point that represents the i th item, rather than the i th category member.

The present applications relied on the Binomial likelihood

$$k_i \sim \text{Binomial}(t_i, r_i), \quad (25)$$

with k_i the number of times the i th test stimulus was chosen in category A and t_i the number of times it was presented, and on the following priors for w and c :

$$w \sim \text{Uniform}(0, 1), \quad (26)$$

and

$$c^2 \sim \text{Gamma}(\varepsilon, \varepsilon), \quad (27)$$

with $\varepsilon = .001$ set near zero, with the range of c being restricted to run between 0 and 10.

References

- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98, 409–429.
- Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37, 372–400.
- Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1, 385–402.
- Brooks, L. R. (1978). Non-analytic concept formation and memory for instances. In E. Rosch, & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Lawrence Erlbaum.
- Diaconis, P., & Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics*, 7, 269–281.

- Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B*, 57, 45–97.
- Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, 40, 1–5.
- Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. *Bayesian Analysis*, 1, 403–420.
- Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In D. S. McNamara, & J. G. Trafton (Eds.), *Proceedings of the 29th annual conference of the cognitive science society* (pp. 323–328). Austin, TX: Cognitive Science Society.
- Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling* (pp. 59–100). Cambridge: Cambridge University Press.
- Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343–1370.
- Lee, M. D. (2004). A Bayesian analysis of retention functions. *Journal of Mathematical Psychology*, 48, 310–321.
- Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, 30, 555–580.
- Lee, M. D., & 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 statistical inference in psychology: comment on Trafimow (2003). *Psychological Review*, 112, 662–668.
- Lindley, D. V. (2004). That wretched prior. *Significance*, 1, 85–87.
- Lindley, D. V., & Phillips, L. D. (1976). Inference for a Bernoulli process (a Bayesian view). *The American Statistician*, 30, 112–119.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review*, 111, 309–332.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207–238.
- Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 775–799.
- Morey, R. (2011). A Bayesian hierarchical model for the measurement of working memory capacity. *Journal of Mathematical Psychology*, 55(1), 8–24.
- Murphy, G. L. (2002). *The big book of concepts*. Boston, MA: MIT Press.
- Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychonomic Bulletin & Review*, 4, 79–95.
- Navarro, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. *Journal of Mathematical Psychology*, 51, 85–98.
- Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 104–114.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57.
- Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 87–108.
- Nosofsky, R. M. (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 honor of William K. Estes: vol. 1* (pp. 149–167). Hillsdale, NJ: Lawrence Erlbaum.
- Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 282–304.
- O'Hagan, A. (2006). Science, subjectivity and software (comment on articles by Berger and by Goldstein). *Bayesian Analysis*, 1, 445–450.
- Press, S. J., & Tanur, J. M. (2001). *The subjectivity of scientists and the Bayesian approach*. New York: Wiley-Interscience.
- Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3, 392–407.
- Romeijn, J.-W., & van de Schoot, R. (2008). A philosopher's view on Bayesian evaluation of informative hypotheses. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 329–357). New York: Springer.
- Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, 12, 195–223.
- Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.
- Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: the early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1411–1436.
- Vanpaemel, W., & Lee, M. D. (2007). A model of building representations for category learning. In D. S. McNamara, & J. G. Trafton (Eds.), *Proceedings of the 29th annual conference of the cognitive science society* (pp. 1605–1610). Austin, TX: Cognitive Science Society.
- Vanpaemel, W., & Navarro, D. J. (2007). Representational shifts during category learning. In D. S. McNamara, & J. G. Trafton (Eds.), *Proceedings of the 29th annual conference of the cognitive science society* (pp. 1599–1604). Austin, TX: Cognitive Science Society.
- Vanpaemel, W., & Storms, G. (2008). In search of abstraction: the varying abstraction model of categorization. *Psychonomic Bulletin & Review*, 15, 732–749.
- Vanpaemel, W., & Storms, G. (2010). Abstraction and model evaluation in category learning. *Behavior Research Methods*, 42, 421–437.
- Wagenmakers, E. J., Lee, M. D., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York: Springer.
- Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44, 92–107.