Models, Parameters and Priors in Bayesian Inference

Michael D. Lee (michael.lee@adelaide.edu.au)
Department of Psychology, University of Adelaide
South Australia, 5005, AUSTRALIA

Abstract

Batchelder and Smith (2004) critiqued the use of Bayesian statistical inference for model selection and evaluation, arguing that it can lead to invalid and contradictory conclusions. Their argument centered around a series of concrete examples purporting to show the problems inherent in adopting the Bayesian approach. This paper addresses Batchelder and Smith's critique by re-analyzing their examples using correct Bayesian methods, and demonstrates that their negative conclusions are not warranted. Throughout the re-analysis, the central role of *information* in the Bayesian approach is emphasized. In particular, it is argued a careful consideration of available information is required to understand the relationship between parameters, models, and data, and the setting of prior probability distributions.

Introduction

As an empirical science, cognitive psychology advances primarily through the development of models. Cognitive models provide formal expressions of theories that can be tested and refined against data. Like all scientific models, cognitive models strive to meet the complementary demands of explanation and prediction. They attempt to provide some useful explanation of observed cognitive phenomena, and a capability to generalize to future or different circumstances.

The fundamental role of models in cognitive psychology means that the quality of methods for their evaluation and comparison are a key determinant of scientific progress. It is not suprising, therefore, that considerable effort has been devoted to promoting the use of modern model selection methods—particularly Bayesian and Minimum Description Length (MDL) methods—in cognitive psychology (e.g., Myung et al., 2000a,b; Pitt et al., 2002). There are now worked examples in the literature applying Bayesian and MDL methods for evaluating models of stimulus representation (e.g., Lee, 2001a,b; Navarro and Lee, in press), generalization and concept learning (e.g., Tenenbaum and Griffiths, 2001), judgment (e.g., Griffiths and Tenenbaum, 2004), inference (e.g., Griffiths et al., 2004), decision-making (e.g., Lee and Cummins, 2004), and problem solving (e.g., Lee et al., 2004).

Recently, however, Batchelder and Smith (2004)[1] critiqued the application of Bayesian methods, arguing that they can lead to invalid and contradictory conclusions. Their argument is centered around a specific example purporting to show the problems inherent in adopting the Bayesian approach to model selection. This paper re-analyzes Batchelder and Smith's example using correct Bayesian methods, and demonstrates that their negative conclusions are not warranted.

Throughout our re-analysis, we emphasize the central role of *information* in the Bayesian approach. In particular, we show how a careful consideration of available information is required to understand the relationship between parameters, models, and data, and the setting of prior probability distributions.

Batchelder and Smith's (2004) Example

In this section, we summarize the examples presented by Batchelder and Smith (2004).

The First Two Models

Batchelder and Smith (2004) originally consider two models that predict the probability of four mutually exclusive and exhaustive outcomes, labeled $p_1, \ldots, p_4$. Both models use two parameters, denoted here θ_1 and θ_2 to predict the various probabilities, but combine the parameters in different ways. For the first model, M_a, the predictions are:

$$M_a \doteq \begin{cases} p_1(\alpha_1, \alpha_2) = \alpha_1^2 + 2\alpha_1(1-\alpha_1)(1-\alpha_2) \\ p_2(\alpha_1, \alpha_2) = (1-\alpha_1)^2 \alpha_2 (2-\alpha_2) \\ p_3(\alpha_1, \alpha_2) = 2\alpha_1\alpha_2(1-\alpha_1) \\ p_4(\alpha_1, \alpha_2) = (1-\alpha_1)^2 (1-\alpha_2)^2 , \end{cases}$$

while for the second model, M_b, the predictions are:

$$M_b \doteq \begin{cases} p_1(\beta_1, \beta_2) = \beta_1\beta_2(\beta_1\beta_2 + 2(1-\beta_1)) \\ p_2(\beta_1, \beta_2) = \beta_1(1-\beta_2)(1-2\beta_1(1+\beta_2)) \\ p_3(\beta_1, \beta_2) = 2\beta_1^2\beta_2(1-\beta_2) \\ p_4(\beta_1, \beta_2) = (1-\beta_1)^2 . \end{cases}$$

[1] Graciously, we have resisted the temptation to say "henceforth 'BS'" at this point.

These models can be evaluated against data, D, that detail how n observations distribute across the four outcomes. The number of times the ith outcome is observed is counted by k_i, with $\sum_{i=1}^{4} k_i = n$, and so $D = (k_1, \ldots, k_4)$. Accordingly, the likelihood function relating the model and its parameters to the observed data are the multinomials

$$p\left(D \mid \alpha_1, \alpha_2, M_a\right) = \binom{n}{k_1 \ldots k_4} \prod_{i=1}^{4} \left[p_i\left(\alpha_1, \alpha_2\right)\right]^{k_i}. \tag{1}$$

and

$$p\left(D \mid \beta_1, \beta_2, M_b\right) = \binom{n}{k_1 \ldots k_4} \prod_{i=1}^{4} \left[p_i\left(\beta_1, \beta_2\right)\right]^{k_i}. \tag{2}$$

Batchelder and Smith (2004) consider the specific data set for $n = 16$ observations with $D = (8, 3, 4, 1)$. By assuming uniform priors in both models,

$$\pi_a\left(\alpha_1, \alpha_2\right) \propto 1,$$

and

$$\pi_b\left(\beta_1, \beta_2\right) \propto 1,$$

they calculate the Bayes Factor to be

$$\begin{aligned} & \frac{p\left(D \mid M_a\right)}{p\left(D \mid M_b\right)} \\ = \; & \frac{\int_0^1 \int_0^1 p\left(D \mid \alpha_1, \alpha_2, M_a\right) \pi_a\left(\alpha_1, \alpha_2\right) d\alpha_1 d\alpha_2}{\int_0^1 \int_0^1 p\left(D \mid \beta_1, \beta_2, M_b\right) \pi_b\left(\beta_1, \beta_2\right) d\beta_1 d\beta_2} \\ = \; & 1.4. \end{aligned}$$

This means that the data provide evidence in favor M_a over M_b, given the assumptions made in the analysis.

The 'True' Model

At this point, Batchelder and Smith (2004) reveal that both M_a and M_b are both one-to-one and onto reparameterizations of a 'true' model, M_t, which makes the predictions

$$M_t \doteq \begin{cases} p_1\left(\theta_1, \theta_2\right) = \theta_1^2 + 2\theta_1(1 - \theta_1 - \theta_2) \\ p_2\left(\theta_1, \theta_2\right) = \theta_2^2 + 2\theta_1(1 - \theta_1 - \theta_2) \\ p_3\left(\theta_1, \theta_2\right) = 2\theta_1\theta_2 \\ p_4\left(\theta_1, \theta_2\right) = \left(1 - \theta_1 - \theta_2\right)^2 \end{cases}.$$

In addition, Batchelder and Smith (2004) give a substantive interpretation of M_t as the blood group model (e.g., Uhlenbruck and Prokop, 1969). This model is depicted in Figure 1, showing how the four blood groups are determined by a hierarchical probabilistic process. As indicated on Figure 1, this process is naturally parameterized in terms of two rates, θ_1 and θ_2.

Batchelder and Smith (2004) conclude that there is a problem with the Bayesian analysis, because the Bayes Factor favors M_a over M_b, even though both are simple reparameterizations of the same underlying model. Their intuition is that neither model should be favored.

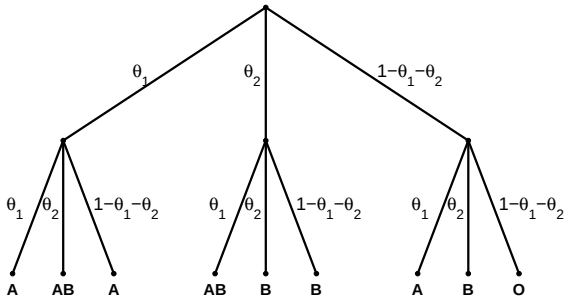

Figure 1: The interpretable parameterization of the blood group model.

Understanding Bayesian Inference

The job of statistical inference is to deal with information under conditions of uncertainty. Within the Bayesian approach, probability distributions are used to represent explicitly, at all stages of analysis, what is known and unknown about every variable of interest. Bayesian inference then proceeds by the routine application of principled methods for updating probability distributions, based on three different, but potentially equally important, sources of information (see Lee and Wagenmakers, in press, for an overview).

Information Inherent in the Problem

Just as zero is the natural starting point for counting, the natural initial representation for Bayesian inference is one corresponding to complete ignorance. Jaynes (see 2003, ch. 12) describes principled methods for defining prior distributions corresponding to complete ignorance. These methods rely on establishing transformational invariances inherent in problems that constrain the choice of prior distribution.

Intuitively, the idea is to consider ways in which a problem could be restated, so that it remains fundamentally the same problem, but is expressed in a different formal way. Prior distributions must necessarily be invariant under these transformations, since otherwise different ways of stating the same problem would lead to different inferences being drawn. In general, the requirement of invariance from information inherent in a problem provides strong constraints on the choice of prior distribution, and often determines them uniquely.

Relevant Prior Information

If other relevant prior information is available, Bayesian analysis incorporates it into the prior distributions. This is done using maximum entropy methods. Intuitively, the idea is to update the prior distributions so that they capture the known constraints provided by additional information, but otherwise remain as uncertain as possible.

Information Provided by Data

As relevant data from experimental or other observations become available, Bayesian analysis uses them to update the probability distribution according to Bayes Theorem.

The Correct Bayesian Analysis

In this section, we apply the Bayesian method of analysis just outlined to the problems presented by Batchelder and Smith (2004).

The First Two Models

As the first two models are presented, no information is available other than the way a set of meaningless parameters lead to predictions about a set of four probabilities. The only way to use this information to determine priors for the parameters is to ensure reparameterization invariance. As Batchelder and Smith (2004) correctly observe later in their paper, this can only be achieved using Jeffreys priors.

In general, the Jeffreys' prior for model with a (possibly) multidimensional parameterization γ is given by

$$\pi(\gamma) \propto \sqrt{\det J(\gamma)}$$

where

$$J_{ij}(\gamma) = E_\gamma \left[\frac{-\partial^2 \ln p(D \mid \gamma)}{\partial \gamma_i \partial \gamma_j} \right].$$

For several common classes of likelihood functions, including multinomials and Gaussians, Su et al. (in press) provide a simple alternative method for calculating Jeffreys priors. For the multinomial likelihood in Eq. 2, their result (see also Schervish, 1995, p. 115) is that $J = P^{\mathrm{T}} \Lambda^{-1} P$ where

$$P = \begin{bmatrix} \partial p_1/\partial \gamma_1 & \partial p_2/\partial \gamma_1 & \partial p_3/\partial \gamma_1 & \partial p_4/\partial \gamma_1 \\ \partial p_1/\partial \gamma_2 & \partial p_2/\partial \gamma_2 & \partial p_3/\partial \gamma_2 & \partial p_4/\partial \gamma_2 \end{bmatrix}$$

and

$$\Lambda = \begin{bmatrix} p_1 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 \\ 0 & 0 & p_3 & 0 \\ 0 & 0 & 0 & p_4 \end{bmatrix}.$$

Using this result, it is a straightforward (but not very enlightening) algebraic exercise to calculate Jeffreys priors for M_a and M_b, which we denote $\pi_a^J(\alpha_1, \alpha_2)$ and $\pi_b^J(\beta_1, \beta_2)$ respectively.

Under Jeffreys priors, the Bayes Factor is now

$$\begin{aligned} & \frac{p(D \mid M_a)}{p(D \mid M_b)} \\ = & \frac{\int_0^1 \int_0^1 p(D \mid \alpha_1, \alpha_2, M_a) \pi_a^J(\alpha_1, \alpha_2) \, d\alpha_1 d\alpha_2}{\int_0^1 \int_0^1 p(D \mid \beta_1, \beta_2, M_b) \pi_b^J(\beta_1, \beta_2) \, d\beta_1 d\beta_2} \\ = & 1. \end{aligned}$$

This means that neither model is favored by the data, as we require.

This result is represented graphically in Figure 2, which shows the prior-weighted likelihood function across the parameter space for both models, using the Jeffreys priors. The integrations in the numerator and denominator of the Bayes Factor correspond to the volumes enclosed by these surfaces. Because these volumes area equal, the ratio is one, and both models are equally likely based on the evidence provided by the data.

Thinking in terms of the information available, the problem with the original analysis of Batchelder and Smith (2004) is that the assumption of uniform priors corresponds to fabricating information that is not, in fact available. Uniform priors do express complete ignorance in some situations—an example of which is described in more detail later—but the models M_a and M_b do not correspond to such a situation. Given that the Bayesian analysis has been misled by the information it was provided, it is not surprising that it reaches an unsatisfactory conclusion. In other words, the mismatch between the (correct) intuition of Batchelder and Smith (2004) and the incorrect Bayes Factor they obtained arises because of a mismatch between the information on which their intuition was based, and the information they formalized in defining prior distributions over their parameters.

The 'True' Model

A more subtle issue arising from the Batchelder and Smith (2004) problem is how the 'true' model, M_t, should be treated. It would be straightforward to find Jeffreys priors for this model, and conduct exactly the same analysis as for the uninterpreted 'black box' models M_a and M_b. But M_t is not a black box model, because Batchelder and Smith (2004) provided *additional information* in revealing that M_t corresponds to an interpretable parameterization of the blood group model. In particular, the parameters θ_1 and θ_2 are now known to correspond to rates, with their parameter space being $(\theta_1, \theta_2) \in [0, 1] \times [0, 1]$.

For a rate parameter, complete ignorance is expressed by the prior distribution known as Haldane's prior, which is shown in Figure 3. A rigorous derivation of this prior using transformational invariance is given by Jaynes (2003, pp. 382–385); a more intuitive justification is provided by Zhu and Lu (2004)). Applying this result to the model M_t gives the prior $\pi_t(\theta_1, \theta_2) = 1/(\theta_1(1-\theta_1)\theta_2(1-\theta_2))$

While we do not know Batchelder and Smith's private thoughts, we guess they would find this prior counter-intuitive. We guess this because Batchelder and Smith (2004) say they find Jeffreys prior for a rate parameter, also shown in Figure 3, to be counter-intuitive, and it has a similar qualitative form to the Haldane prior. Of course, as Jaynes (2003) argues, en-

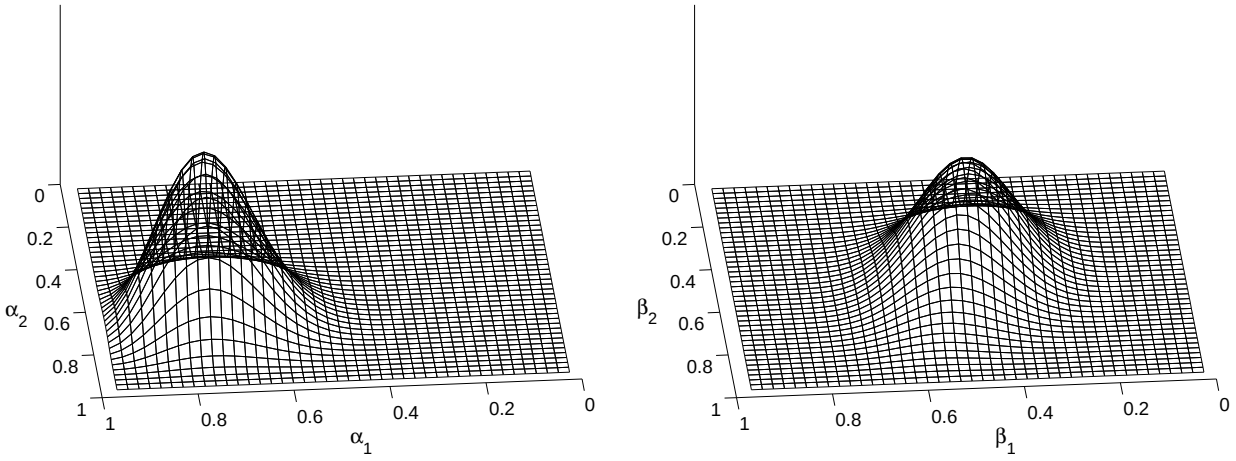

Figure 2: The prior-weighted likelihood across the parameter space for the two models, M_a (left panel) and M_b (right panel), given the data considered by Batchelder and Smith (2004), and using Jeffreys priors. The volume encompassed by both surfaces is equal.

countering a counter-intuitive result in this way ought to be treated as an opportunity to educate and correct our intuition, rather than abandoning a coherent framework for statistical inference.

In this case, the form of the Haldane prior for rates arises because many phenomena always happen (i.e., have a rate of one) or never happen (i.e., have a rate of zero), and relatively few have an actual 'rate' between these extremes. As a concrete example, consider undertaking an experiment to determine the 'rate' at which two chemicals, when mixed, turn green. The first observation will be very informative in this situation: if the mixed chemicals turn green, a rate of one seems likely; if they do not turn green, a rate of zero seems likely. Many phenomena share this characteristic of having 'rates' that are actually either zero or one. Unless we know both outcomes are possible, a true rate between these extremes is less likely. Accordingly, under the assumption of *complete* ignorance, where the "always" and "never" possibilities are not known to be false, the Haldane prior makes rates of one and zero more probable, while still allowing for the possibility of a rate somewhere between.

Besides requiring the Haldane prior, the other important contribution of the additional information provided in the 'true' model is that the parameters now have meaning, and their posterior distribution are useful and interpretable. Applying the laws of probability, the required posterior is just

$$p\left(\theta_1, \theta_2 \mid D\right) \propto p\left(D \mid \theta_1, \theta_2\right) \pi_t\left(\theta_1, \theta_2\right),$$

where the proportionality is handled by normalizing the posterior to sum to one, as required for it to be a probability distribution. Figure 4 summarizes the posterior for the data considered by Batchelder and Smith (2004), showing the maximum posterior density, together with 50%, 90%, 95% and 99% credible regions.

Of course, it may be the case—Batchelder and Smith (2004) never say one way or the other—that it is known the rate parameters in the blood group correspond to processes where both binomial possibilities can occur. If this additional information is available, maximum entropy methods lead to the Haldane prior being updated to a uniform prior (Jaynes, 2003, pp. 385), also shown in Figure 3. As should be the case, the availability of different information will generally lead to different results. For this example, however, the posterior distribution under the uniform prior is visually indistinguishable from Figure 4. This is because the 16 available data provide information that dominate the subtle difference between initially available information that lead to the Haldane and uniform prior. The fact that Bayesian measures are generally relatively insensitive to priors when sufficient data are available is fundamentally a statement about the relative contribution of different sources of information to reducing uncertainty; it is a pity that it is often presented (by both Bayesian and non-Bayesian) as merely a practical apologetic for putative deficiencies in the framework of Bayesian inference.

Finally, we note that Batchelder and Smith (2004), like some Bayesian authors (e.g., Gill, 2002, pp. 135-137), express the concern that maximum entropy methods are not reparameterization invariant. This is true, but not relevant. Following the stages of infor-

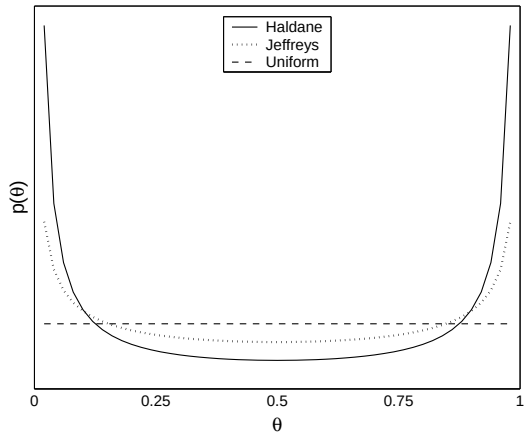

Figure 3: Haldane, Jeffreys and uniform priors for a rate parameter θ.

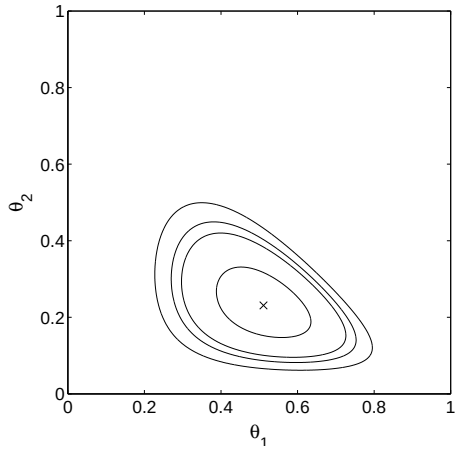

Figure 4: Posterior over the parameter space for the 'true' model M_t, given the data considered by Batchelder and Smith (2004), using Haldane's prior. The maximum is shown by the cross, together with 50%, 90%, 95% and 99% credible regions.

mation analysis done in Bayesian inference, the application of maximum entropy methods (for incorporating prior information about parameters) is done after analysis of the information in the problem itself (which determines complete ignorance priors with respect to a parameterization). The role of maximum entropy is therefore to refine an existing probability distribution over an existing parameterization, and there is no need for reparameterization invariance. Our analysis of the Batchelder and Smith (2004) example provides a concrete demonstration of this. The complete ignorance Haldane prior determined by transformational invariance corresponds to a specific parameterization. When it becomes known that both outcomes are possible, maximum entropy methods only have to deal with the existing parameterization, and update the Haldane to a uniform prior.

Discussion

Considering the Batchelder and Smith (2004) example highlights the importance of understanding what information is available about parameters and models when making Bayesian inferences. At their most basic level, models are just a set of probability distributions across a data space indexed by one or more parameters. Sometimes in cognitive modeling, parameters do nothing other than fulfill this indexing role, and exist only to allow the full range of model predictions to be observed. If this is all that is known, priors need to be reparameterization invariant, and this is achieved by using Jeffreys priors.

Usually, however, cognitive models carry information about the parameters. Often, the level of theoretical committment to parameters, and the meaning attached to them, is as important as the model itself. For example, something like the ALCOVE (Kruschke, 1992) model of category learning makes theoretical claims about cognitive processes (i.e., selective attention, error-driven learning, and generalization are important for categorization), but also seeks to interpret its parameters as corresponding to meaningful psychological variables (i.e., learning rates, levels of generalization, and response strategies). The only reason, for example, it is sensible to study individual differences through fitting models like ALCOVE is that variations in parameters are psychologically meaningful (e.g., Nosofsky et al., 1994; Nosofsky and Johansen, 2000; Treat et al., 2001; Webb and Lee, 2004).

In even more extreme cases, some cognitive models view the parameters as being more meaningful and important than the model itself. For example, many models in psychometrics take very simple forms (e.g., linear models in factor analysis) that are not part of a strong theoretical committment, but serve primarily as vehicles for inferring their parameters from data. It is these parameters that correspond to the psychological variables of interest, such as levels of cognitive abilities, and are the important outcome of the modeling.

In cases like these, where information is available about model parameters, using Jeffreys priors is suboptimal because it does not use all of the available information. Rather, the more general notion of transformational invariance approach advocated by Jaynes (2003, ch. 12) should be applied, because it incorporates what is known about the parameters. At least the first part of this conclusion should be obvious. One of the defining philosophical features of the Bayesian

approach is that, unlike Orthodox methods based on sampling distributions, it is possible to know something about parameters before any data have been observed, or, indeed, before any model has been proposed. Jeffreys priors are defined with respect to a model, and so cannot possibly always be the appropriate method for determining priors.

When the available information is analysed carefully, however, Bayesian inference provides a complete and general method for making quantitative judgments about models, parameters, and data. In this sense, it provides the ideal framework for dealing with the uncertainty that surrounds the enterprise of cognitive modeling.

Acknowledgments

I thank Dan Navarro and Eric-Jan Wagenmakers.

References

Batchelder, W. H. and Smith, J. B. (2004). What if model selection becomes the metric for scientific acceptance? Paper presented at the 37th Annual Meeting of the Society for Mathematical Psychology, Ann Arbor, MI.

Gill, J. (2002). *Bayesian Methods: A Social and Behavioral Sciences Approach*. Chapman & Hall/CRC, Boca Raton: FL.

Griffiths, T. L., Baraff, E. R., and Tenenbaum, J. B. (2004). Using physical theories to infer hidden causal structure. In Forbus, K., Gentner, D., and Regier, T., editors, *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 500–505, Mahwah, NJ. Erlbaum.

Griffiths, T. L. and Tenenbaum, J. B. (2004). From algorithmic to subjective randomness. *Advances in Neural Information Processing Systems*, 16.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, New York.

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

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

Lee, M. D. (2001b). On the complexity of additive clustering models. *Journal of Mathematical Psychology*, 45(1):131–148.

Lee, M. D. and Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and "rational" models. *Psychonomic Bulletin & Review*, 1(2):343–352.

Lee, M. D., O'Connor, T. A., and Welsh, M. B. (2004). Human decision making on the full-information secretary probllem. In Forbus, K., Gentner, D., and Regier, T., editors, *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 819–824, Mahwah, NJ. Erlbaum.

Lee, M. D. and Wagenmakers, E. J. (in press). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*.

Myung, I. J., Balasubramanian, V., and Pitt, M. A. (2000a). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, 97:11170–11175.

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

Navarro, D. J. and Lee, M. D. (in press). Common and distinctive features in stimulus similarity: A modified version of the contrast model. *Psychonomic Bulletin & Review*.

Nosofsky, R. M. and Johansen, M. K. (2000). Exemplar-based accounts of 'mutliple-system' phenomena in perceptual organization. *Psychonomic Bulletin & Review*, 7(3):375–402.

Nosofsky, R. M., Palmeri, T. J., and McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101:53–79.

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

Schervish, M. J. (1995). *Theory of Statistics*. Springer, New York.

Su, Y., Myung, J. I., and Pitt, M. A. (in press). Minimum description length and cognitive modeling. In Grünwald, P., Myung, J. I., and Pitt, M. A., editors, *Advances in Minimum Description Length: Theory and Applications*. MIT Press, Cambridge, MA.

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

Treat, T. A., McFall, R., Viken, R. J., and Kruschke, J. K. (2001). Using cognitive science emthods to assess the role of social information processing in sexually coercive behavior. *Psychological Assessment*, 13(4):549–565.

Uhlenbruck, G. and Prokop, O. (1969). *Human Blood and Serum Groups*. McLaren and Sons, London.

Webb, M. R. and Lee, M. D. (2004). Modeling individual differences in category learning. In Forbus, K., Gentner, D., and Regier, T., editors, *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 1440–1445. Erlbaum, Mahwah, NJ.

Zhu, M. and Lu, A. Y. (2004). The counter-intuitive non-informative prior for the Bernoulli family. *Journal of Statistics Education [Online]*, 12(2).