Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Mathematical Psychology 48 (2004) 310–321

Journal of Mathematical Psychology

www.elsevier.com/locate/jmp

A Bayesian analysis of retention functions

Michael D. Lee*

Department of Psychology, University of Adelaide, Adelaide, SA 5005, Australia

Received 21 January 2004; received in revised form 30 June 2004
Available online 22 October 2004

Abstract

A long standing goal of quantitative research in cognitive psychology has been to provide a lawful description of the retention of information over time. While a number of theoretical alternatives for a retention function have been developed, their empirical evaluation has almost exclusively relied on their ability to fit experimental data. This has meant that the issue of model complexity, which considers the number of parameter in a model and the functional form of parameter interaction, has generally not been considered in a rigorous way. This paper develops a Bayesian method for comparing retention models that naturally considers the competing demands of goodness-of-fit and complexity. We first implement the Bayesian method using numerical techniques, highlighting the basic properties of the method and showing, in particular, how assumptions about the precision of the data affect the inferences that are drawn. We then develop an analytic Bayesian method, based on the Laplacian approximation, that offers some theoretical insights into the inherent complexities of different retention functions, and has the practical advantage of being computationally efficient. We demonstrate both methods by evaluating linear, hyperbolic, exponential, logarithmic and power retention functions against the collection of data sets considered by Rubin and Wenzel (1996).

1. Introduction

From the time of Ebbinghaus (1885/1964), a goal of memory research has been to provide a lawful description of the retention of information over time. A successful descriptive model of retention would provide not only a basic constraint for theorizing about memory, but also a predictive capability with significant potential for application.

The search for a model of memory retention has proceeded on both theoretical and empirical fronts. A range of different theoretical mechanisms for describing memory retention have been developed, based on a variety of conceptual perspectives (e.g., Anderson & Schooler, 1991; Estes, 1997; Laming, 1992). Empirically, there have been many attempts to evaluate candidate theoretical models by testing their ability to capture data that have been collected using a variety of experimental methodologies. For the most part (e.g., Anderson & Schooler, 1991; Rubin & Wenzel, 1996; Rubin, Hinton, & Wenzel, 1999; Wixted & Ebbesen,1991,1997), the empirical evaluation of competing models has relied on goodness-of-fit measures, such as the percentage of variance explained.

Unfortunately, as Roberts and Pashler (2000) have recently argued, the practice of distinguishing between competing psychological models solely on the basis of their ability to fit data faces a number of serious problems. One of these problems, which has been a recent focus in mathematical psychology (e.g., Myung, Balasubramanian, & Pitt, 2000a; Myung, Forster, & Browne, 2000b; Myung & Pitt, 1997; Pitt, Myung, & Zhang, 2002), is that solely measuring goodness-of-fit fails to account for differences in the complexity of competing models. The complexity of a model is basically a measure of its ability to fit any data set well, regardless of whether those data could ever arise from

*Fax: +61-8-8303-3770.
E-mail address: michael.lee@adelaide.edu.au.
URL: http://www.psychology.adelaide.edu.au/members/staff/michaellee.

0022-2496/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2004.06.002

the psychological process being modeled. Overly complicated models achieve high levels of fit through being flexible enough to capture the random variation, or noise, present in a particular set of data, and so do not necessarily provide a better account of the regularities in the data. This means, in turn, that complicated models will generalize poorly to new data observed at another time or in another context, and so will have less predictive capability than simple models that fit the current data well.

For these reasons, psychological models should strive to balance the competing demands of fit and complexity, providing the simplest possible accurate account of the available data. In the absence of any attempt to control for model complexity, there is no guarantee that the best fitting retention models do not achieve their accuracy only through being more complicated than their competitors. Many previous studies have been aware of this issue (e.g., Rubin & Wenzel, 1996; Rubin et al., 1999; Wixted & Ebbesen, 1997), although they seem generally to have equated the complexity of a model complexity with its number of free parameters. As Myung and Pitt (1997) make clear, however, model complexity also has a 'functional form' component, determined by the way parameters interact within a model. This means that two models with the same number of parameters will, in general, have different complexity, because they will assume different functional forms. The rigorous evaluation of competing retention models requires a methodology that takes into account the goodness-of-fit and both sources of model complexity.

This paper presents a Bayesian methodology for evaluating retention models that naturally accounts for both fit and complexity, and provides an intuitive framework for interpreting and understanding the results of these comparisons. The developed method is general, in the sense that it could be used to evaluate any collection of retention functions against any set of data. To demonstrate the method, however, we consider the five retention functions—linear, hyperbolic, exponential, logarithmic and power functions—canvassed by Rubin and Wenzel (1996), and evaluate them in terms of the 210 collated data sets they considered. This is not intended to imply that these particular retention functions are the only serious theoretical possibilities, and is certainly not intended to imply that Rubin and Wenzel's (1996) collated data provide a definitive test-bed. Indeed, the same authors have raised questions about whether the data can support a conclusive test of retention functions (Rubin et al., 1999). What Rubin and Wenzel's (1996) study does provide, however, are specific data sets on which to demonstrate the Bayesian approach.

The structure of this paper is as follows: The next section provides an overview of the Bayesian approach to model selection. A case study, using one of the Rubin and Wenzel (1996) data sets, is then presented, showing how the Bayesian approach balances goodness-of-fit and model complexity, and developing two methods for evaluating retention functions. The more efficient of these methods is then applied to all of the Rubin and Wenzel (1996) data sets, showing how evidence from many sources can be combined to evaluate models. Finally, the general discussion considers variants on the Bayesian methodology.

2. Bayesian model evaluation

This section presents a brief overview of some aspects of Bayesian model evaluation and selection, and provides the necessary statistical background for the various methods used in analyzing the retention models. More detailed treatments of Bayesian methods may be found, for example, in Carlin and Louis (2000), Jaynes (2003), Kass and Raftery (1995), Gill (2002), Leonard and Hsu (1999), and Sivia (1996). Readers familiar with Bayesian methods may safely skip this section.

At the heart of Bayesian analysis is Bayes' theorem, which specifies the way in which the prior probability of a model being true, $p(\mathbf{M})$ is altered by the evidence provided by data to become a posterior probability, $p(\mathbf{M} \mid \mathbf{D})$:

$$p(\mathbf{M} \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \mathbf{M})}{p(\mathbf{D})}\, p(\mathbf{M}).$$

Intuitively, the evidence provided by data is measured by the probability that the observed data would have arisen under the assumption that the model were true, $p(\mathbf{D} \mid \mathbf{M})$, normalized by the probability that the data would have arisen under any set of assumptions, $p(\mathbf{D})$.

For parameterized models, measuring the probability of observed data arising under the assumption that the model is true involves considering the match between the data and the model in all of its parameterizations. This means that $p(\mathbf{D} \mid \mathbf{M})$ becomes a marginal probability, obtained by integrating the probability of the data for each of the possible parameter combinations, $p(\mathbf{D} \mid \theta, \mathbf{M})$, as weighted by the prior probability of each of these combinations $p(\theta \mid \mathbf{M})$:

$$p(\mathbf{D} \mid \mathbf{M}) = \int p(\mathbf{D} \mid \theta, \mathbf{M}) p(\theta \mid \mathbf{M})\, d\theta. \tag{1}$$

A particularly useful form of Bayes' theorem considers the posterior odds for two models. This is simply the ratio of their prior probabilities, multiplied by the ratio of the marginal probabilities:

$$\frac{p(\mathbf{M}_i \mid \mathbf{D})}{p(\mathbf{M}_j \mid \mathbf{D})} = \frac{p(\mathbf{D} \mid \mathbf{M}_i)}{p(\mathbf{D} \mid \mathbf{M}_j)} \frac{p(\mathbf{M}_i)}{p(\mathbf{M}_j)},$$

where the ratio of the posterior to the prior odds is usually called the 'Bayes factor'. This form is useful because it directly compares two models in a way that is naturally interpretable. As Kass and Raftery (1995, p. 777) argue, probabilities lie on a scale defined by betting, and so the ratio of posterior probabilities is a meaningful number. A posterior probability ratio of ten, for example, means that the first model is ten times more likely than the second model. In the same way, the Bayes factor is easily interpreted as quantifying the evidence that the observed data provides for one model over the other. Kass and Raftery (1995, p. 777) give a number of alternative interpretative frameworks for these probability ratios, which essentially amount to suggested standards of scientific evidence.

Another useful form of Bayes' theorem involves model averaging, where the probability of the data being observed $p(\mathbf{D})$ is partitioned into the probability of it being observed under each of an exhaustive set of alternative models, so that

$$p(\mathbf{M}_i \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \mathbf{M}_i)}{\sum_j p(\mathbf{D} \mid \mathbf{M}_j)}.$$

As Carlin and Louis (2000, p. 50) note, there are often too many plausible alternative models that need to be considered for model averaging to be a practical method of calculating the posterior probability of a model. In these cases, it may be useful to consider a small set of competing models that are of particular interest, and define a measure

$$H_i = \frac{p(\mathbf{D} \mid \mathbf{M}_i)}{\sum_j p(\mathbf{D} \mid \mathbf{M}_j)}, \tag{2}$$

which effectively measures the posterior probability of model $\mathbf{M}_i$ *relative* to the other competing models.

Finally, in Bayesian analyses, it is important to consider the concept of precision, which measures the level of noise in empirical data that obscures the regularities coming from an underlying cognitive process. When observed data are precise (i.e., relatively noise free), the introduction of additional complexity into a model to achieve a greater level of descriptive accuracy may well be warranted. When observed data are imprecise (i.e., relatively noisy), however, the same increase in complexity will not be warranted, because the extra complexity will tend to be used to fit the noise. As argued by Lee (2001, p. 155), this means that a quantitative estimate of data precision is needed to determine the appropriate balance between goodness-of-fit and model complexity. If such an estimate is not available, it is necessary either to make explicit assumptions about data precision, or undertake broader analyses that consider the entire plausible range of data precision possibilities. Note that this conceptualization of precision makes it fundamentally a property of the data, in the sense that it parameterizes the way data are summarized before any particular model has been considered.

3. A case study

This section develops the methodology for Bayesian analysis of retention functions by focusing on one particular data set from Rubin and Wenzel (1996).

3.1. The retention functions

We consider the five retention models given particular attention by Rubin and Wenzel (1996), which assume linear, hyperbolic, exponential, logarithmic and power functions. For a particular time interval t_i, the linear model predicts a retention value, $\hat{y}_i = -mt_i + b$, where m and b are non-negative parameter values. Different predictions are made by the other models, using the functional relationships $\hat{y}_i = 1/(mt_i + b)$ for the hyperbolic model, $\hat{y}_i = b\exp(-mt_i)$ for the exponential model, $\hat{y}_i = b - m \ln t_i$ for the logarithmic model, and $\hat{y}_i = bt_i^{-m}$ for the power model.

3.2. The retention data

The particular data set chosen was that presented by Squire (1989, Fig. 1A), measuring the average accuracy with which television shows were recalled after time periods ranging up to 15 years. The data may be represented as two vectors, $\mathbf{t} = (t_1, t_2, \ldots, t_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, respectively denoting the time (measured in years), and the recall (measured in terms of proportion correct). Each (t_i, y_i) pair gives the y_i

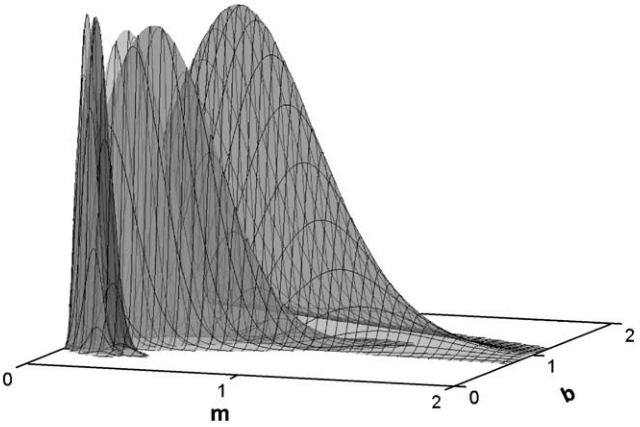

Fig. 1. Surface plot showing the level of fit for the five retention functions on the Squire (1989) data, across the parameter space ranging from 0 to 2 for both the m and b parameters. From left to right, the functions are the logarithmic, power, linear, exponential and hyperbolic functions.

proportion of shows correctly recalled after t_i years. The fit of a memory retention function to these data is a measure of the agreement between the recall it predicts, $\hat{y}_i$, and the observed recall, y_i over all of the n data points.

For all of the retention functions considered by Rubin and Wenzel (1996), the predictions made depend not only upon the independent variable time, but also upon two parameters, m and b. If it is assumed that each of the observed recall proportions y_i comes from a Gaussian distribution with common variance σ^2, for which we have an estimate s^2, then a probabilistic measure of fit for a particular parameterization of a retention function is given by:[1]

$$p(\mathbf{y} \mid m, b, \mathbf{t}) = \prod_{i=1}^{n} \frac{1}{(2\pi s^2)^{\frac{1}{2}}} \exp\left(-\frac{(y_i - \hat{y}_i)^2}{2s^2}\right) = \frac{1}{(2\pi s^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2s^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right). \quad (3)$$

The value of s measures the precision of the data, taking smaller values for more precisely observed recall proportions. Lee (2001, p. 155) suggests that, for averaged data, the mean of the sample standard deviation for each of the data points furnishes an appropriate estimate. Squire (1989, p. 244) reports that the standard deviations for the recall proportions were similar, symmetric around the mean, and ranged from 24.2% to 28.3%. In light of this information, a reasonable estimate for normalized data might be $s \approx 0.25$, although it is still important to undertake an analysis across a broader range of s values, to test the sensitivity of any conclusions to the exact assumption made regarding data precision.

With the probabilistic measure of fit given by Eq. (3) in place, it is possible to measure the marginal probability of Eq. (1) required for a Bayesian analysis. This involves measuring the fit across the entire plausible set of parameter values m and b. For all of the retention models, this parameter space is simply the quadrant of the two-dimensional plane corresponding to positive m and b values.

3.3. Numerical approximation

The most straightforward approach to evaluating the integral that defines the marginal probability uses a simple numerical method. By using a large number of parameter values $\{\theta_1, \theta_2, \ldots, \theta_N\}$ according to their prior probability, the marginal probability may be estimated as

$$p(\mathbf{D} \mid \mathbf{M}) \approx \frac{1}{N}\sum_{i=1}^{N} p(\mathbf{D} \mid \theta_i, \mathbf{M}).$$

In terms of the retention data at hand, it is computationally feasible to impose a grid on the two-dimensional parameter space, and sample at each point on the grid. For sets of m and b parameter values given by $\{m_1, m_2, \ldots, m_N\}$ and $\{b_1, b_2, \ldots, b_N\}$, which define the extent and resolution of the grid, the numerical approximation becomes

$$p(\mathbf{D} \mid \mathbf{M}) \approx Z\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} p(\mathbf{y} \mid m_i, b_j, \mathbf{t}),$$

where Z is a constant relating to the spacing of the grid from which samples are drawn, and so is the same for all models. This approximation was applied to the Squire (1989) data, using a grid that extended from 0 to 2 for both m and b parameters, and increasing in steps of 0.005 for both parameters, which proved a small enough step size that further refinement did not change the final Bayes factors significantly. A range of precision estimates were considered, increasing from $s = 0.05$ to $s = 0.5$ in steps of 0.05.

The results of these calculations are depicted graphically for $s = 0.25$ in Figs. 1 and 2. The surfaces in Fig. 1 indicate by their heights the level of fit achieved by each of the five functions at each point in the two-dimensional parameter space. In particular, the highest peaks for each function correspond to the maximum goodness-of-fit at the best parameter values. It can be seen that the logarithmic and power function are the best fitting, followed by the hyperbolic, then the exponential, then the linear function. It is also clear,

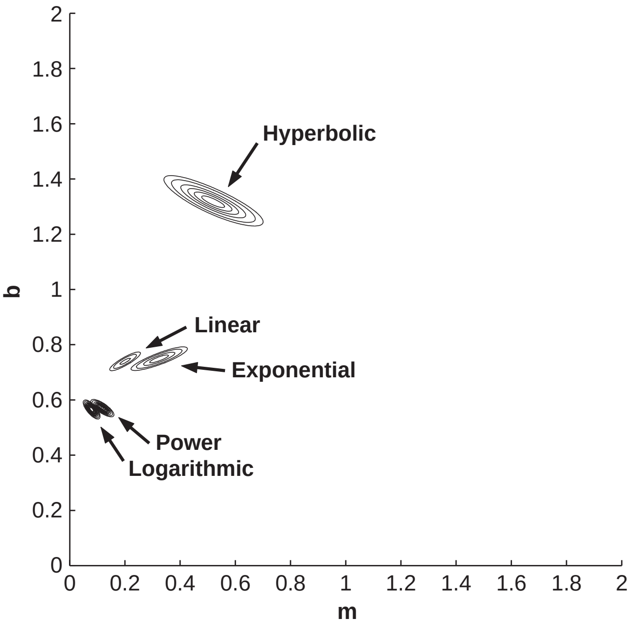

Fig. 2. A contour plot showing the level of fit for the five retention functions on the Squire (1989) data, across the parameter space ranging from 0 to 2 for both the m and b parameters.

[1] Later we provide a critical discussion of the Gaussian assumption.

however, that the good fits of the logarithmic and power functions occur at a very narrow set of parameter values, while the hyperbolic function fits reasonably well at a large number of parameterizations. These differences in complexity are even clearer in Fig. 2, which shows contour plots of the same information. It can be seen that the hyperbolic function fits across a broader range of parameter values than the exponential function, which in turn has a broader range than the linear, power and logarithmic functions.

Figs. 1 and 2 provide intuitive graphical support for two of the basic assumptions made in the Bayesian analysis. First, it is clear that the range of the parameter space is sufficient to cover all of the appreciable density being integrated, and so extending the grid would not change the results. Secondly, it is clear that prior assumptions about the values of the m and b parameters would affect the conclusions only if they gave appreciably different density to those regions of the parameter space where there is reasonable fit. Any vague prior, like the uniform prior used, that does not modulate the peaks in Fig. 2, will result in the same conclusions. This is because the data quickly dominate the prior through the likelihood function.

Fig. 3 presents the basic pattern of results evident in Figs. 1 and 2 in a different way, by showing the data and those parameterizations of each function that exceeded a pre-determined threshold level of fit, where the threshold has been chosen to make the display visually informative. The hyperbolic function clearly has the most parameterizations that meet the threshold, followed by the exponential function, with the linear, power and logarithmic fitting less often. Both the power and logarithmic functions, however, have parameterizations that achieve better fit to the data than the hyperbolic or exponential functions. The linear function, meanwhile, does not exhibit close levels of fit for any parameterization, nor do many parameterizations reach the threshold.

In terms of the balance between goodness-of-fit and complexity, these results indicate that the power and logarithmic models are capable of achieving better fit to the data, but are more complicated than the hyperbolic and exponential functions, because their fit is less robust across parametric variation. This raises the possibility that the ability of the power and logarithmic models to fit the data may have less to do with the functions they use being the appropriate ones to capture the retention data, and more to do with the functions being sufficiently flexible to accommodate any data (including data that would never be observed in a memory recall experiment) by adjusting their parameter values. The hyperbolic and exponential models, in contrast, robustly predict retention curves that resemble the observed data, even though their best fits are not as impressive.

Resolving the trade-off between the better descriptive accuracy of the power and logarithmic models, and the lesser complexity of the hyperbolic and exponential

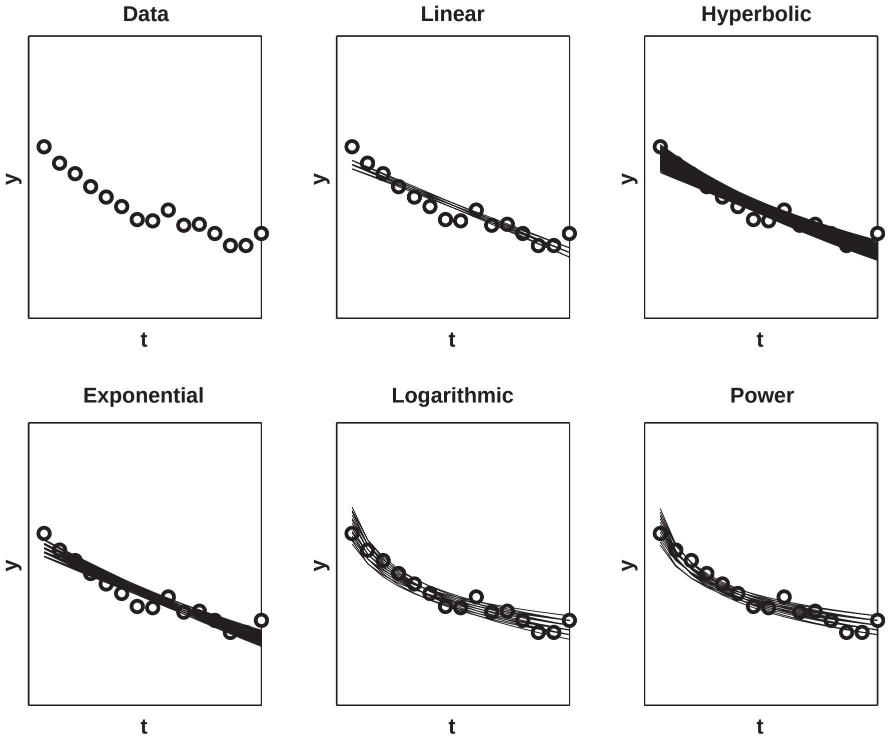

Fig. 3. The Squire (1989) data, and the parameterizations of each of the five retention functions that exceed a threshold level of fit to the data.

models, can only be made through recourse to the precision of the observed data. If the data were measured arbitrarily precisely (corresponding to a precision value s approaching zero), the issue of robustness across different parameterizations would not arise. The best model would simply be the model that provided the most accurate description of the data, as measured by its best level of fit. As the data become less precise, however, the robustness or complexity of the models becomes relatively more important, and should be weighted progressively more highly than fit. In the limit, as data become infinitely noisy, they are not capable of providing evidence for or against any model, and neither fit nor complexity help distinguish between competing alternatives.

This pattern is consistent with the numerical analysis. The marginal probability of the data having arisen under the assumption that a particular retention model is true is given by the area under the distributions shown in Figs. 1 and 2. For precise data, corresponding to small values of s, these distributions will fall away from their peak at the best fitting parameter values quickly. This means that the height of the peak, corresponding to the maximal level of fit, will largely determine the area under the distribution. However, as the data become less precise, and s becomes larger, the distribution will fall away less quickly, and the robustness of the fit across parameter values will play a progressively greater role in determining the area. Finally, as s tends towards infinity, every function will fit the data equally poorly at every parameter value, the distributions will become completely flat, and will each encompass the same area.

For the Squire (1989) data, because a precision estimate of $s \approx 0.25$ is available, it is possible to strike an appropriate balance between fit and complexity. This is naturally done through calculating the Bayes' factors for each pair of models. It turns out that, using numerical analysis, the hyperbolic model has the greatest posterior probability, being approximately 3.4, 6.7, 13.0 and 16.4 times more likely than the exponential, linear, power and logarithmic models, respectively. Obviously, the Bayes' factors between any other pair of models can be derived from these ratios. The important result is that, at the estimated level of data precision, the hyperbolic model constitutes the best balance between fit and inherent complexity, and is most strongly supported by Squire's (1989) data.

Despite being able to provide these sort of insights, numerical analysis suffers from two shortcomings. First, it is computationally intensive when a broad range of parameter values must be sampled, or when the resolution of the imposed grid must be increased to improve accuracy. It is for this reason that a wide range of computational mechanisms for improving the accuracy of numerical analysis have been developed (e.g., Gilks, Richardson, & Spiegelhalter, 1996; Smith & Roberts, 1993). The second shortcoming, however, is more fundamental. Numerical analysis does not offer the possibility of gaining any analytic insight into the inherent complexities of competing models, since it only provides a method for dealing with a particular set of data.

3.4. The Laplacian approximation

Both of the shortcomings of numerical analysis are potentially overcome by using what is known as the Laplacian approximation to the marginal probability (see, for example, Carlin & Louis, 2000, pp. 122–129; Kass & Raftery, 1995, pp. 777–778; Leonard & Hsu, 1999, pp. 191–194). This approximation treats each of the distributions shown in Fig. 1 as Gaussian distributions, with a mean given by the best fitting parameter values, and a co-variance matrix given by a matrix of second derivatives. By making this assumption, the area under the distribution is easily calculated, and the form of the co-variance matrix provides information relating to the complexity of a model.

For a general model with a vector of P parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_P)$, with all of the possible parameterizations being a priori equally likely, the Laplacian approximation is given by

$$p(\mathbf{D} \mid \mathbf{M}) \approx \frac{(2\pi)^{P/2} p(\mathbf{D} \mid \theta^*, \mathbf{M})}{(\det \mathbf{I}(\theta^*))^{\frac{1}{2}}}, \qquad (4)$$

where θ^* are the best fitting parameter values. The matrix $\mathbf{I}(\theta^*)$ is the Hessian matrix of second derivatives of the log probability, defined by

$$I_{ij}(\theta) = \frac{\partial^2 \ln p(\mathbf{D} \mid \theta, \mathbf{M})}{\partial\theta_i \partial\theta_j},$$

evaluated at the best fitting parameter values.

Following Myung and Pitt (1997), the Laplacian approximation may be interpreted in terms of two components. The term $p(\mathbf{D} \mid \theta^*, \mathbf{M})$ measures the fit of the model to the available data. For models with the same number of parameters P, the denominator $\sqrt{\det \mathbf{I}(\theta^*)}$ is the only way in which model complexity contributes to the posterior probability, and in this sense provides a measure of the functional form complexity of the model.

For the probabilistic measure of the fit of a retention function given by Eq. (3), the log probability is essentially the familiar sum-squared error, scaled by the data precision estimate:

$$-\ln p(\mathbf{y} \mid m, b, \mathbf{t}) = \frac{1}{2s^2} \sum_i (y_i - \hat{y}_i)^2 + \frac{n}{2} \ln 2\pi s^2.$$

The required Hessian matrix of this log probability is the 2 × 2 matrix

$$\mathbf{I}(m,b) = \frac{1}{2s^2}\begin{bmatrix} \partial^2/\partial m^2 \sum_i (y_i - \hat{y}_i)^2 & \partial^2/(\partial m \partial b)\sum_i (y_i - \hat{y}_i)^2 \\ \partial^2/(\partial b \partial m)\sum_i (y_i - \hat{y}_i)^2 & \partial^2/\partial b^2 \sum_i (y_i - \hat{y}_i)^2 \end{bmatrix}.$$

The derivation of the determinant of this matrix for each of the five retention functions is a straightforward algebraic exercise. The results of the derivation may be summarized by setting $\mathbf{I} = \frac{1}{s^2}\mathbf{G}$, and giving the determinant of the matrix $\mathbf{G}$. Under the Laplacian approximation, it is the square root of this determinant that measures the functional form complexity of a model. For the linear function, it turns out that

$$\det \mathbf{G} = n \sum_i t_i^2 - \left(\sum_i t_i \right)^2,$$

which is simply n times the variance of the time data values $\mathbf{t}$. This means, as has previously been noted by MacKay (1992), that the complexity of a linear function fit to data may be minimized by choosing values along the independent variable with maximal variance. A related result is achieved for the logarithmic function, where

$$\det \mathbf{G} = n \sum_i (\ln t_i)^2 - \left(\sum_i \ln t_i \right)^2,$$

which corresponds to n times the variance of $\ln \mathbf{t}$. This is to be expected, given the relationship between the linear and logarithmic models. Intuitively, as noted by Rubin and Wenzel (1996, p. 749), the difference is that the linear model assumes equal intervals of time are important, while the logarithmic model assumes that equal ratios of time are important. Their different complexity measures simply assess the extent to which time intervals have been chosen that are consistent with these assumptions.

The complexity results for the other functions are less interpretable, but nonetheless provide efficient formulae for calculating the complexity component of the Laplacian approximation. For the exponential function, the result is

$$\det \mathbf{G} = \frac{1}{b^2} \sum_i \sum_j \hat{y}_i \hat{y}_j t_j (2\hat{y}_j - y_j)(\hat{y}_i t_j - t_i t(2\hat{y}_i - y_i));$$

for the power function, the result is

$$\det \mathbf{G} = \frac{1}{b^2} \sum_i \sum_j \hat{y}_i \hat{y}_j \ln t_j (2\hat{y}_j - y_j) \times (\hat{y}_i \ln t_j - \ln t_i (2\hat{y}_i - y_i));$$

and for the hyperbolic function, the result is

$$\det \mathbf{G} = \sum_i \sum_{j>i} \hat{y}_i^3 \hat{y}_j^3 (2y_i - 3\hat{y}_i)(2y_j - 3\hat{y}_j)(t_i - t_j)^2.$$

3.5. Accuracy of the Laplacian approximation

The accuracy of the Laplacian approximation may be assessed in a direct way by comparing the posterior probabilities it produces with those obtained by the numerical analysis, which makes no approximating assumptions. This was done for ten of the data sets considered by Rubin and Wenzel (1996), chosen to cover a breadth of sample sizes, experimental methodologies, and differences in the fits of the five retention functions. The particular data sets used were those reported by Bahrick, Bahrick, and Wittlinger (1975, Table 4, free recall condition), Bean (1912, Table 6), Burtt and Dobell (1925, Table 1), Conway, Cohen, and Stanhope (1991, Fig. 4), Finkenbinder (1913, Table 2), Jarrard and Moise (1970, Fig. 1), Luh (1922, Table 6), Squire (1989, Table 1A), Thompson (1982, Fig. 2), and Tsai (1924, Table 2). These studies were conducted between 1912 and 1991, and have sample sizes as low as 5 and as high as 15, which are essentially the extremes available in the Rubin and Wenzel (1996) data. They involved free recall, serial recognition, construction, relearning, typing, maze running, delayed matching, grouping and dating tasks. They used (in no particular order) adults, graduate students, primates and undergraduate students as participants, and they measured retention across time periods ranging from 1–28 s, to 1–15 years.

A regression analysis was undertaken comparing the (logarithm of) the Bayes' factors produced by the numerical analysis with the (logarithm of) the Bayes' factors produced by the Laplacian approximation, for all $5 \times 4/2 = 10$ possible pairs of retentions functions, at each estimated s value ranging from 0.05 to 0.50 in steps of 0.05. This revealed a linear relationship with a slope of 0.984 and an intercept of 0.000 that explained 99.7% of the variance, indicating that the two measures are virtually identical. On these grounds, it seems safe to accept the Laplacian approximation of marginal probabilities as being sufficiently accurate to make the sort of model selection decisions that are part of a Bayesian analysis.

3.6. Other approaches

These results show that more sophisticated approximations to the Bayesian posterior, such as the Stochastic Complexity Criterion (SCC: Rissanen, 1996), and the Geometric Complexity Criterion (GCC: Myung et al., 2000a), are not necessary on the grounds of accuracy. Measures like the SCC and GCC do offer reparameterization invariance, but we believe this is not fundamentally important. Models of retention, like models of most psychological phenomena, seek to use parameters with meaningful interpretations to gain insight and understanding. This means that reparameterization invariance, while clearly a desirable property if all other things are equal, is not crucial in the same way it might be for 'black box' models that are nothing more than indexable probability distributions over data.

The important invariances for psychological modeling, we believe, are those inherent in the description of the problem itself. Jaynes (2003, Chapter 12) presents a compelling case that the rational approach to defining 'complete ignorance' priors is by defining transformations that leave a problem unchanged in substance, but affect aspects of its quantitative description. The prior distributions for parameters must necessarily be invariant under these transformations, since they leave the problem unchanged. The invariances, in turn, often provides a strong constraint of the form of priors, or define them uniquely. Retention modeling has one obvious transformational invariance, relating to the measurement scale of the times. The performance of a retention model should give the same results for the same data set whether the times are represented, for example, in second or milliseconds. This implies that the appropriate prior distributions for retention model parameters should lead to the same performance under positive scalar multiplication of the time data. It is a worthwhile topic for future research to find these priors, which will be allied to a specific parameterization of a retention model, and would generally change under reparameterization. In the meantime, the Laplacian approximation used here, like other approximations, makes different assumptions about prior distributions, and so relies on the evidence provided by data dominating prior information to justify its results. As noted earlier, for the data analyzed here, any diffuse prior will lead to essentially the same conclusions.

In considering alternative approaches, it is also worth noting that Minimum Description Length (MDL) methods (e.g., Grünwald, 2000) have the theoretical advantage that they do not make Bayesian assumptions about 'true' models, but focus on 'useful' models that compress data by finding regularities. How this theoretical distinction manifests itself is not entirely clear. We are not aware of any practical examples in psychology where Bayesian inference behaves inappropriately because the statistical process that generates data is not contained in the set of models being considered, even though it seems this must almost always be the case. Finally, it is interesting to observe that the Bayesian Laplacian measure is closely related to the MDL approach used by the SCC.[2] Basically, the Laplacian measure is an approximation to the SCC under the simplifying assumption that the Hessian matrix **I** can be evaluated only at the best fitting parameters θ^*, rather than being integrated across the entire parameter space.

Table 1
The fit and complexity of the retention functions for Squire's (1989) data

Function	Proportion of variance explained	Functional form complexity
Linear	0.884	4.32
Hyperbolic	0.922	2.02
Exponential	0.904	2.98
Logarithmic	0.955	11.3
Power	0.946	11.5

3.7. Application of the Laplacian approximation

Table 1 details the goodness-of-fit and complexity of the five retention models found by applying the Laplacian approximation to the Squire (1989) data. The goodness-of-fit is measured in terms of the proportion of variance explained, while the complexity is measured by $\sqrt{\det \mathbf{G}}$. These results confirm the patterns observed in the numerical analysis. The logarithmic and power models provide the best fit, but are more complicated than the linear model, which is in turn more complicated than the exponential and hyperbolic models.

As before, the competing claims of fit and complexity for the models can only be resolved by understanding the precision of the data. For any data precision estimate s, the values in Table 1 may be used in the Laplacian approximation of Eq. (4) to estimate marginal probabilities, Bayes' factors, or the various other interpretable measures discussed earlier. Perhaps the most useful analysis is in terms of the relative posterior probabilities defined in Eq. (2), which may be re-written as

$$H_i(s) = \frac{p(\mathbf{D} \mid \mathbf{M}_i)}{\sum_j p(\mathbf{D} \mid \mathbf{M}_j)}$$

to indicate explicitly the dependence on data precision.

Fig. 4 shows the relative posterior probabilities for each of the five retention models, for the Squire (1989) data, ranging from $s = 0.01$ to $s = 0.50$. It can be seen that, if the data are assumed to be very precise, the relative probabilities of models reflects their accuracy. In particular, the logarithmic model, which has the best fit, is strongly preferred. As the assumed level of precision decreases, however, the models are ordered in terms of their complexity, with the hyperbolic model having the greatest value. This preference is less marked, however, with imprecise data providing some evidence for all of the models. Overall, the pattern of change in the H

[2] I thank Jay Myung for drawing my attention to this correspondence.

values concurs with the preceding analysis of the data, and provides a simple means of visualizing the fit and complexity behavior of the competing models across different levels of data precision.

4. General evaluation

Having established an efficient method, based on the Laplacian approximation, for evaluating the relative posterior probabilities of the five retention models for any given data set, we now examine all of the data sets collated by Rubin and Wenzel (1996). Following the assertion of these authors that the autobiographical retention data are fundamentally different, five data sets were not considered. For the remaining 205 data sets, the goodness-of-fit and functional form complexity of each of the five retention models were calculated. The results of this analysis are given in Table 2, which is based on the rank orderings of fit and complexity. The entries in Table 2 give the percentage of data sets for which each function has the best fit (1st), the second best fit (2nd), through to the worst fit (5th), and the percentage of data sets for which each function has the lowest complexity (1st), the second lowest complexity (2nd), through to the highest complexity (5th).

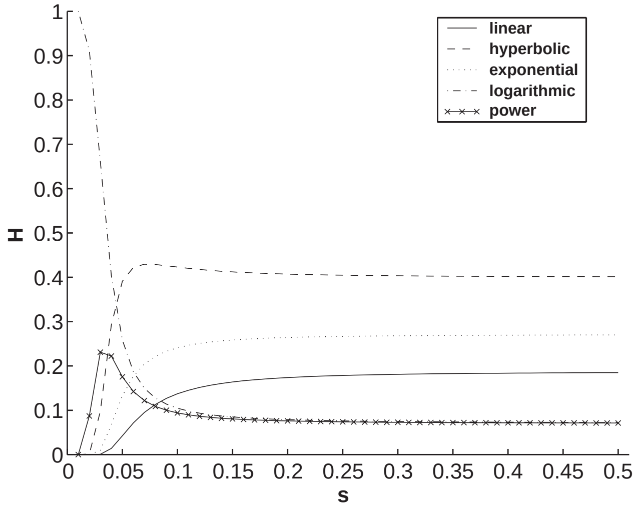

Fig. 4. The relative posterior probabilities for each of the five retention models on the Squire (1989) data, with assumed levels of data precision s ranging from 0.01 to 0.50.

In terms of goodness-of-fit, Table 2 shows that the hyperbolic, logarithmic and power functions were the best fitting functions approximately equally often, while the linear function was most often the worst fitting function. In terms of complexity, Table 2 shows that the hyperbolic function was always the least complicated and the exponential function was almost always the second least complicated, while the logarithmic function was always the most complicated.

Fig. 5 shows the results of a relative posterior probability analysis across the 205 data sets. The pattern of change in the mean relative posterior probability for each function as data precision changes is shown, together with error bars showing one standard error. These curves are based on a set of precision values ranging from 0.025 to 0.50 in steps 0.025, together with a relative posterior probability for the case of arbitrarily precise data ($s = 0$), calculated by comparing only the goodness-of-fit of the competing models.

Fig. 5 indicates that the hyperbolic, power and logarithmic models provide the best fits for data sets approximately equally often. This is indicated by all three models having approximately the same relative posterior probabilities of about 0.3 at $s = 0$, where effectively only goodness-of-fit is measured. However, as the assumed level of data precision worsens, the inherent complexity of the power and logarithmic functions means that the hyperbolic function comes to have the greatest relative posterior probability. Similarly, the lower complexity of the exponential functional form leads to it also assuming significant relative posterior probability for larger s values. Meanwhile, the linear model, which has both poor fit and moderately high

Table 2
The relative goodness-of-fit and complexity rankings of the five retention functions across all of the Rubin and Wenzel (1996) data sets, except for the five dealing with autobiographical recall

Function	Goodness-of-fit					Complexity				
	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
Linear	6	1	12	9	72	0	0	29	71	0
Hyperbolic	28	14	56	1	1	100	0	0	0	0
Exponential	8	24	11	57	0	0	98	2	0	0
Logarithmic	28	39	11	20	2	0	0	0	0	100
Power	30	22	10	13	25	0	2	69	29	0

For fit, the percentage of data sets for which each function was ranked in each position is shown, going from best fitting (1st) to worst fitting (5th). For complexity, the percentage of data sets for which each function was ranked in each position is shown, going from least complicated (1st) to most complicated (5th).

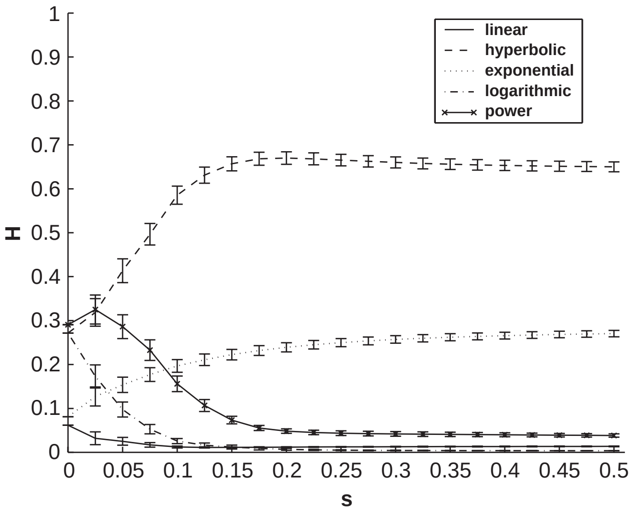

Fig. 5. The mean relative posterior probabilities, as a function of data precision, for each of the five retention models on all of the Rubin and Wenzel (1996) data sets, except for the five dealing with autobiographical recall.

complexity, sustains almost no relative posterior probability at any stage.

It should be acknowledged that the averaging process used to generate the results shown in Fig. 5 is not entirely justified. Averaging would be most appropriate if the underlying curves for all of the data sets were the same, except for measurement error. Visual inspection of the individual curves suggests, however, that there are qualitative differences between some of the curves, and developing an understanding of these differences is an important area for future research. For the moment, however, Fig. 5 provides a convenient graphical summary representation that captures the main message of the Bayesian analysis: That the simplicity of the hyperbolic function makes it a preferable model to the power and logarithmic models for retention data that are not very precise.

Ideally, it would be possible to estimate the precision using the sample standard deviation, and determine the vertical line (or a limited region) in Fig. 5 that should be used to make model selection decisions. Unfortunately, most published retention data sets do not report the necessary statistics, nor do they provide the relevant raw data. As Rubin and Wenzel (1996, p. 756) lament 'In the existing literature, confidence intervals are rarely reported, cannot be calculated from reported statistics in most procedures, and when they can are too large to reject functions'. The reported standard deviations of about 0.25 for the Squire (1989) data accord with this belief in large confidence intervals, although the 25 year time span covered by this data raises the possibility that it is less precise than others. A few of the Rubin and Wenzel (1996, p. 756) datasets are available in the raw form required for precision estimation, and these happen to be studies involved much shorter (less than 20 s) time spans. Estimating the precision for these data sets, therefore, may provide some guidance as to how precise the retention data might plausibly be. To this end, the data reported in Jans and Catania (1980, Table 1) were found to have sample standard deviations of 0.117 and 0.101 for the 'standard' and 'activity' conditions respectively, while the Harnett, McCarthy, and Davison (1984, Appendix B) data were found to have a sample standard deviation of 0.0827.

Without access to all of the necessary data, any determination of data precision must be heavily subjective. However, all of the precision estimates obtained from individual data sets favor the hyperbolic model, and Fig. 5 makes it clear that a precision estimate of $s<0.05$ for the data would required before the power and logarithmic models rivaled the hyperbolic model. Given that Rubin et al. (1999, p. 1161) believe their data are not precise, it seems reasonable to conclude that the Bayesian analysis presented here favors the hyperbolic model.

5. General discussion

The conclusions from the Bayesian evaluation of the Rubin and Wenzel (1996) data make sense. The imprecise data essentially contain only one regularity, in the form of a negatively accelerating downward change in retention over time. This regularity is incompatible with the linear model, but is accommodated by the other four functions. Accordingly, the Bayesian analysis chooses the simplest functional form that has the observed regularity, which is the hyperbolic function. Because the Bayesian approach is sensitive to the precision of data, evaluating the same functions using more precise empirical data might well lead to different conclusions. All that can be concluded from the analysis presented here is that the data were not sufficiently noise free to warrant any more complicated negatively accelerating downward function than the hyperbolic function. In this sense, the Bayesian methodology for evaluating retention functions naturally and directly places the onus on researchers to collect precise measures of the retention phenomena they wish to model.

Despite this strength, there are a number of limitations of the Bayesian approach developed here that should be acknowledged. First, the evidence for any of the retention models provided by the Bayesian analysis is merely relative evidence, showing that a particular function is preferable to its current competitors. The introduction of additional candidate models into the analysis could, of course, significantly alter conclusions. For example, the derivation of Laplacian approximations for any of the 105 functions considered by Rubin and Wenzel (1996) would be straightforward, and could be included in future analysis. Other less analytically tractable models may require the use of numerical methods, but the basic Bayesian approach to model selection will still apply.

Secondly, it is important to note that the best description of retention data that have been averaged across participants does not necessarily constitute the best description of the retention in the individual participants themselves. The basic idea that averaging data can influence the outcomes of psychological model testing was examined by Estes (1956), and has been considered in the context of retention models by a number of authors (e.g., Anderson & Tweney, 1997; Wixted & Ebbesen, 1997). A theoretical response surface analysis, in the context of comparing power and exponential functions, is provided by (Myung, Kim, & Pitt, 2000c; see also Heathcote, Brown, & Mewhort, 2000). Since the vast majority of the data sets considered by Rubin and Wenzel (1996) are averaged across participants, the results of the Bayesian analysis presented here do not necessarily indicates that the retention of individuals is best modeled using the hyperbolic function.

Finally, it is worth considering the adequacy of the basic modeling assumptions made in Eq. (3). As noted above, most of the Rubin and Wenzel (1996) retention data take the form of averages across blocks of binary response measures. The assumption that these averages have Gaussian distributions follows from its approximation to the binomial distribution for moderately large numbers of observations. One weakness of the Gaussian assumption is that it gives some probability density to data values outside the possible interval [0,1], particularly for average values at the extremes of this interval. Another weakness is that the current assumption of equal variance Gaussian distributions probably needs refinement, particularly when retention performance is at floor or ceiling levels. An analysis relying on the binomial distribution would overcome both these problems, but is only possible with access to more detailed data, so that the necessary counts can be formed.

An additional attraction of using the binomial distribution, with access to individual participant data, is that the question of individual differences in retention could be tackled in a principled way (Webb & Lee, 2004). Different parameterizations of a retention function could be applied to different subsets of participants, by combining the counts for each participants in the same subset, and finding the best fitting parameterization for each of these combinations. The overall adequacy of competing models formed in this way could be measured using the same Bayesian methodology presented here. Where genuine individual differences existed, the additional complexity required by using additional functions to model the different groups would be justified. Where between-subject variation constituted noise, in the sense that it did not show meaningful structure in terms of the retention functions being proposed, simpler models proposing fewer individual variations would be preferred. The study of individual differences in retention using this approach, and using the theoretically preferable binomial distribution, is an important avenue for future research.

In the meantime, however, the Bayesian method presented here allows goodness-of-fit and model complexity, including functional form complexity, to be considered when assessing retention models. It provides a computationally straightforward and easily interpreted method for evaluating rival models against data, and so has the potential to contribute to model development in a basic enterprise for cognitive psychology: describing and predicting how information is retained over time.

Acknowledgments

The research reported in this paper was supported by the Australian Defence Science and Technology Organization. The author wishes to thank David Rubin for providing the collated retention data, Helen Braithwaite, Andrew Heathcote, Daniel Navarro, Mark Pitt, Douglas Vickers and John Wixted for helpful comments, and Eric-Jan Wagenmakers, Jeffrey Rouder, Rich Shiffrin and an anonymous reviewer for thoughtful reviews. Particular thanks are given to Jay Myung for his extensive comments, which greatly improved the paper.

References

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408.

Anderson, J. R., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, *25*, 724–730.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: a cross-sectional approach. *Journal of Experimental Psychology: General*, *104*, 54–75.

Bean, C. H. (1912). The curve of forgetting. *Archives of Psychology*, *2*, 1–47.

Burtt, H. E., & Dobell, E. M. (1925). The curve of forgetting for advertising material. *Journal of Applied Psychology*, *9*, 5–21.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). New York: Chapman & Hall.

Conway, M. A., Cohen, G., & Stanhope, N. (1991). On the very long term retention of knowledge acquired through formal education: twelve years of cognitive psychology. *Journal of Experimental Psychology: General*, *120*, 395–409.

Ebbinghaus, H., 1964. *Memory: a contribution to experimental psychology*. (H. A. Rutger & C. E. Bussenius, Trans.). New York: Dover (original work published 1885).

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140.

Estes, W. K. (1997). Processes of memory loss, recovery, and distortion. *Psychological Review*, *104*(1), 148–169.

Finkenbinder, E. D. (1913). The curve of forgetting. *American Journal of Psychology*, *24*, 8–32.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Gill, J. (2002). *Bayesian methods: a social and behavioral sciences approach*. Boca Raton: FL: Chapman & Hall/CRC.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*(1), 133–152.

Harnett, P., McCarthy, D., & Davison, M. (1984). Delayed signal detection, differential reinforcement, and short term memory in the pigeon. *Journal of the Experimental Analysis of Behavior*, *42*, 87–111.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: the case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207.

Jans, J. E., & Catania, C. (1980). Short-term remembering of discriminative stimuli in pigeons. *Journal of the Experimental Analysis of Behavior*, *34*, 177–183.

Jarrard, L. E., & Moise, S. L. (1970). Short term memory in the stump tail macaque: effect of restraint of behavior on performance. *Learning and Motivation*, *1*, 267–275.

Jaynes, E. T. (2003). *Probability theory: the logic of science*. New York: Cambridge University Press.

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

Laming, D. (1992). Analysis of short-term retention: models for Brown–Peterson experiments. *Journal of Experimental Psychology: Learning Memory, and Cognition*, *18*(6), 1342–1365.

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

Leonard, T., & Hsu, J. S. J. (1999). *Bayesian methods: an analysis for statisticians and interdisciplinary researchers*. New York: Cambridge University Press.

Luh, C. W. (1922). The conditions of retention. *Psychological Monographs*, *31* (whole no. 142).

MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*, 590–604.

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

Myung, I. J., Kim, C., & Pitt, M. A. (2000c). Toward an explanation of the power law artifact: insights from response surface analysis. *Memory & Cognition*, *28*(5), 832–840.

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

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

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*(1), 40–47.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Rubin, D. C., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1161–1176.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: a quantitative description of retention. *Psychological Review*, *103*(4), 734–760.

Sivia, D. S. (1996). *Data analysis: a Bayesian tutorial*. Oxford: Clarendon Press.

Smith, A. F. M., & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, *55*, 3–23.

Squire, L. R. (1989). On the course of forgetting in very long term memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*, 241–245.

Thompson, C. P. (1982). Memory for unique personal events: the roommate study. *Memory & Cognition*, *10*, 324–332.

Tsai, C. (1924). A comparative study of retention curves for motor habits. *Comparative Psychology Monographs*, *2*, 1–29.

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

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, *2*(6), 409–415.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: a quantitative analysis of individual subject forgetting functions. *Memory & Cognition*, *25*(5), 731–739.