Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Mathematical Psychology 47 (2003) 90–100
Journal of Mathematical Psychology
http://www.elsevier.com/locate/jmp

Tutorial

Tutorial on maximum likelihood estimation

In Jae Myung*

Department of Psychology, Ohio State University, 1885 Neil Avenue Mall, Columbus, OH 43210-1222, USA

Received 30 November 2001; revised 16 October 2002

Abstract

In this paper, I provide a tutorial exposition on maximum likelihood estimation (MLE). The intended audience of this tutorial are researchers who practice mathematical modeling of cognition but are unfamiliar with the estimation method. Unlike least-squares estimation which is primarily a descriptive tool, MLE is a preferred method of parameter estimation in statistics and is an indispensable tool for many statistical modeling techniques, in particular in non-linear modeling with non-normal data. The purpose of this paper is to provide a good conceptual explanation of the method with illustrative examples so the reader can have a grasp of some of the basic principles.

1. Introduction

In psychological science, we seek to uncover general laws and principles that govern the behavior under investigation. As these laws and principles are not directly observable, they are formulated in terms of hypotheses. In mathematical modeling, such hypotheses about the structure and inner working of the behavioral process of interest are stated in terms of parametric families of probability distributions called models. The goal of modeling is to deduce the form of the underlying process by testing the viability of such models.

Once a model is specified with its parameters, and data have been collected, one is in a position to evaluate its goodness of fit, that is, how well it fits the observed data. Goodness of fit is assessed by finding parameter values of a model that best fits the data—a procedure called *parameter estimation*.

There are two general methods of parameter estimation. They are least-squares estimation (LSE) and maximum likelihood estimation (MLE). The former has been a popular choice of model fitting in psychology (e.g., Rubin, Hinton, & Wenzel, 1999; Lamberts, 2000 but see Usher & McClelland, 2001) and is tied to many familiar statistical concepts such as linear regression, sum of squares error, proportion variance accounted for (i.e. r^2), and root mean squared deviation. LSE, which unlike MLE requires no or minimal distributional assumptions, is useful for obtaining a descriptive measure for the purpose of summarizing observed data, but it has no basis for testing hypotheses or constructing confidence intervals.

On the other hand, MLE is not as widely recognized among modelers in psychology, but it is a standard approach to parameter estimation and inference in statistics. MLE has many optimal properties in estimation: sufficiency (complete information about the parameter of interest contained in its MLE estimator); consistency (true parameter value that generated the data recovered asymptotically, i.e. for data of sufficiently large samples); efficiency (lowest-possible variance of parameter estimates achieved asymptotically); and parameterization invariance (same MLE solution obtained independent of the parametrization used). In contrast, no such things can be said about LSE. As such, most statisticians would not view LSE as a general method for parameter estimation, but rather as an approach that is primarily used with linear regression models. Further, many of the inference methods in statistics are developed based on MLE. For example, MLE is a prerequisite for the chi-square test, the G-square test, Bayesian methods, inference with missing data, modeling of random effects, and many model selection criteria such as the Akaike information criterion (Akaike, 1973) and the Bayesian information criteria (Schwarz, 1978).

*Fax: +614-292-5601.
E-mail address: myung.1@osu.edu.

0022-2496/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0022-2496(02)00028-7

In this tutorial paper, I introduce the maximum likelihood estimation method for mathematical modeling. The paper is written for researchers who are primarily involved in empirical work and publish in experimental journals (e.g. *Journal of Experimental Psychology*) but do modeling. The paper is intended to serve as a stepping stone for the modeler to move beyond the current practice of using LSE to more informed modeling analyses, thereby expanding his or her repertoire of statistical instruments, especially in non-linear modeling. The purpose of the paper is to provide a good conceptual understanding of the method with concrete examples. For in-depth, technically more rigorous treatment of the topic, the reader is directed to other sources (e.g., Bickel & Doksum, 1977, Chap. 3; Casella & Berger, 2002, Chap. 7; DeGroot & Schervish, 2002, Chap. 6; Spanos, 1999, Chap. 13).

2. Model specification

2.1. Probability density function

From a statistical standpoint, the data vector $y = (y_1, \ldots, y_m)$ is a random sample from an unknown population. The goal of data analysis is to identify the population that is most likely to have generated the sample. In statistics, each population is identified by a corresponding probability distribution. Associated with each probability distribution is a unique value of the model's parameter. As the parameter changes in value, different probability distributions are generated. Formally, a model is defined as the family of probability distributions indexed by the model's parameters.

Let $f(y|w)$ denote the *probability density function* (PDF) that specifies the probability of observing data vector y given the parameter w. Throughout this paper we will use a plain letter for a vector (e.g. y) and a letter with a subscript for a vector element (e.g. y_i). The parameter $w = (w_1, \ldots, w_k)$ is a vector defined on a multi-dimensional parameter space. If individual observations, y_i's, are statistically independent of one another, then according to the theory of probability, the PDF for the data $y = (y_1, \ldots, y_m)$ given the parameter vector w can be expressed as a multiplication of PDFs for individual observations,

$$f(y = (y_1, y_2, \ldots, y_n) \mid w) = f_1(y_1 \mid w) f_2(y_2 \mid w) \cdots f_n(y_m \mid w). \quad (1)$$

To illustrate the idea of a PDF, consider the simplest case with one observation and one parameter, that is, $m = k = 1$. Suppose that the data y represents the number of successes in a sequence of 10 Bernoulli trials (e.g. tossing a coin 10 times) and that the probability of a success on any one trial, represented by the parameter w, is 0.2. The PDF in this case is given by

$$f(y \mid n = 10, w = 0.2) = \frac{10!}{y!(10-y)!}(0.2)^y (0.8)^{10-y} \quad (y = 0, 1, \ldots, 10) \quad (2)$$

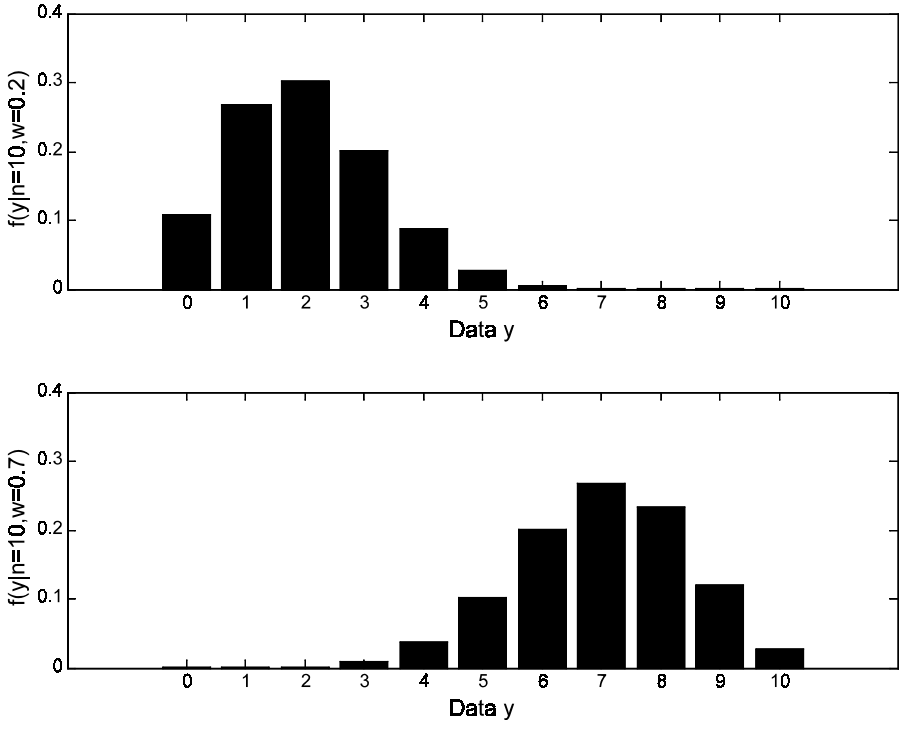

Fig. 1. Binomial probability distributions of sample size $n = 10$ and probability parameter $w = 0.2$ (top) and $w = 0.7$ (bottom).

which is known as the binomial distribution with parameters $n = 10$, $w = 0.2$. Note that the number of trials (n) is considered as a parameter. The shape of this PDF is shown in the top panel of Fig. 1. If the parameter value is changed to say $w = 0.7$, a new PDF is obtained as

$$f(y \mid n = 10, w = 0.7) = \frac{10!}{y!(10-y)!}(0.7)^{y}(0.3)^{10-y} \quad (y = 0, 1, \ldots, 10) \tag{3}$$

whose shape is shown in the bottom panel of Fig. 1. The following is the general expression of the PDF of the binomial distribution for arbitrary values of w and n:

$$f(y|n, w) = \frac{n!}{y!(n-y)!}w^{y}(1-w)^{n-y} \quad (0 \leqslant w \leqslant 1;\ \ y = 0, 1, \ldots, n) \tag{4}$$

which as a function of y specifies the probability of data y for a given value of n and w. The collection of all such PDFs generated by varying the parameter across its range (0–1 in this case for w, $n \geqslant 1$) defines a model.

2.2. Likelihood function

Given a set of parameter values, the corresponding PDF will show that some data are more probable than other data. In the previous example, the PDF with $w = 0.2$, $y = 2$ is more likely to occur than $y = 5$ (0.302 vs. 0.026). In reality, however, we have already observed the data. Accordingly, we are faced with an inverse problem: Given the observed data and a model of interest, find the one PDF, among all the probability densities that the model prescribes, that is most likely to have produced the data. To solve this inverse problem, we define the *likelihood function* by reversing the roles of the data vector y and the parameter vector w in $f(y|w)$, i.e.

$$L(w|y) = f(y|w). \tag{5}$$

Thus $L(w|y)$ represents the likelihood of the parameter w given the observed data y, and as such is a function of w. For the one-parameter binomial example in Eq. (4), the likelihood function for $y = 7$ and $n = 10$ is given by

$$L(w \mid n = 10, y = 7) = f(y = 7 \mid n = 10, w) = \frac{10!}{7!3!}w^{7}(1-w)^{3} \quad (0 \leqslant w \leqslant 1). \tag{6}$$

The shape of this likelihood function is shown in Fig. 2.

There exist an important difference between the PDF $f(y|w)$ and the likelihood function $L(w|y)$. As illustrated in Figs. 1 and 2, the two functions are defined on different axes, and therefore are not directly comparable to each other. Specifically, the PDF in Fig. 1 is a function of the data given a particular set of parameter values, defined on the *data scale*. On the other hand, the likelihood function is a function of the parameter given a particular set of observed data, defined on the *parameter scale*. In short, Fig. 1 tells us the probability of a particular data value for a fixed parameter, whereas Fig. 2 tells us the likelihood ("unnormalized probability") of a particular parameter value for a fixed data set. Note that the likelihood function in this figure is a curve

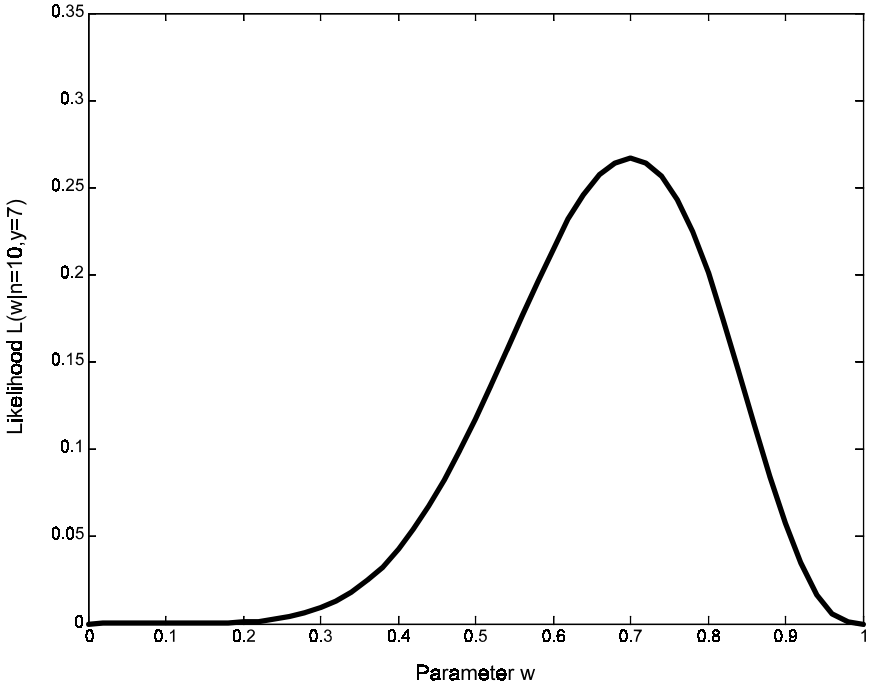

Fig. 2. The likelihood function given observed data $y = 7$ and sample size $n = 10$ for the one-parameter model described in the text.

because there is only one parameter beside n, which is assumed to be known. If the model has two parameters, the likelihood function will be a surface sitting above the parameter space. In general, for a model with k parameters, the likelihood function $L(w|y)$ takes the shape of a k-dim geometrical "surface" sitting above a k-dim hyperplane spanned by the parameter vector $w = (w_1, \ldots, w_k)$.

3. Maximum likelihood estimation

Once data have been collected and the likelihood function of a model given the data is determined, one is in a position to make statistical inferences about the population, that is, the probability distribution that underlies the data. Given that different parameter values index different probability distributions (Fig. 1), we are interested in finding the parameter value that corresponds to the desired probability distribution.

The principle of *maximum likelihood estimation* (MLE), originally developed by R.A. Fisher in the 1920s, states that the desired probability distribution is the one that makes the observed data "most likely," which means that one must seek the value of the parameter vector that maximizes the likelihood function $L(w|y)$. The resulting parameter vector, which is sought by searching the multi-dimensional parameter space, is called the *MLE estimate*, and is denoted by $w_{\mathrm{MLE}} = (w_{1,\mathrm{MLE}}, \ldots, w_{k,\mathrm{MLE}})$. For example, in Fig. 2, the MLE estimate is $w_{\mathrm{MLE}} = 0.7$ for which the maximized likelihood value is $L(w_{\mathrm{MLE}} = 0.7|n = 10, y = 7) = 0.267$. The probability distribution corresponding to this MLE estimate is shown in the bottom panel of Fig. 1. According to the MLE principle, this is the population that is most likely to have generated the observed data of $y = 7$. To summarize, maximum likelihood estimation is a method to seek the probability distribution that makes the observed data most likely.

3.1. Likelihood equation

MLE estimates need not exist nor be unique. In this section, we show how to compute MLE estimates when they exist and are unique. For computational convenience, the MLE estimate is obtained by maximizing the log-likelihood function, $\ln L(w|y)$. This is because the two functions, $\ln L(w|y)$ and $L(w|y)$, are monotonically related to each other so the same MLE estimate is obtained by maximizing either one. Assuming that the log-likelihood function, $\ln L(w|y)$, is differentiable, if w_{MLE} exists, it must satisfy the following partial differential equation known as the *likelihood equation*:

$$\frac{\partial \ln L(w|y)}{\partial w_i} = 0 \tag{7}$$

at $w_i = w_{i,\mathrm{MLE}}$ for all $i = 1, \ldots, k$. This is because the definition of maximum or minimum of a continuous differentiable function implies that its first derivatives vanish at such points.

The likelihood equation represents a necessary condition for the existence of an MLE estimate. An additional condition must also be satisfied to ensure that $\ln L(w|y)$ is a maximum and not a minimum, since the first derivative cannot reveal this. To be a maximum, the shape of the log-likelihood function should be convex (it must represent a peak, not a valley) in the neighborhood of w_{MLE}. This can be checked by calculating the second derivatives of the log-likelihoods and showing whether they are all negative at $w_i = w_{i,\mathrm{MLE}}$ for $i = 1, \ldots, k$,[1]

$$\frac{\partial^2 \ln L(w|y)}{\partial w_i^2} < 0. \tag{8}$$

To illustrate the MLE procedure, let us again consider the previous one-parameter binomial example given a fixed value of n. First, by taking the logarithm of the likelihood function $L(w|n = 10, y = 7)$ in Eq. (6), we obtain the log-likelihood as

$$\ln L(w \mid n = 10, y = 7) = \ln \frac{10!}{7!3!} + 7 \ln w + 3 \ln(1 - w) \tag{9}$$

Next, the first derivative of the log-likelihood is calculated as

$$\frac{d \ln L(w \mid n = 10, y = 7)}{dw} = \frac{7}{w} - \frac{3}{1 - w} = \frac{7 - 10w}{w(1 - w)}. \tag{10}$$

By requiring this equation to be zero, the desired MLE estimate is obtained as $w_{\mathrm{MLE}} = 0.7$. To make sure that the solution represents a maximum, not a minimum, the second derivative of the log-likelihood is calculated and evaluated at $w = w_{\mathrm{MLE}}$,

$$\begin{aligned} \frac{d^2 \ln L(w \mid n = 10, y = 7)}{dw^2} &= -\frac{7}{w^2} - \frac{3}{(1 - w)^2} \\ &= -47.62 < 0 \end{aligned} \tag{11}$$

which is negative, as desired.

In practice, however, it is usually not possible to obtain an analytic form solution for the MLE estimate, especially when the model involves many parameters and its PDF is highly non-linear. In such situations, the MLE estimate must be sought numerically using non-linear optimization algorithms. The basic idea of non-linear optimization is to quickly find optimal parameters that maximize the log-likelihood. This is done by

[1] Consider the Hessian matrix $H(w)$ defined as $H_{ij}(w) = \frac{\partial^2 \ln L(w)}{\partial w_i \partial w_j}$ $(i, j = 1, \ldots, k)$. Then a more accurate test of the convexity condition requires that the determinant of $H(w)$ be *negative definite*, that is, $z' H(w = w_{\mathrm{MLE}}) z < 0$ for any kx1 real-numbered vector z, where z' denotes the transpose of z.

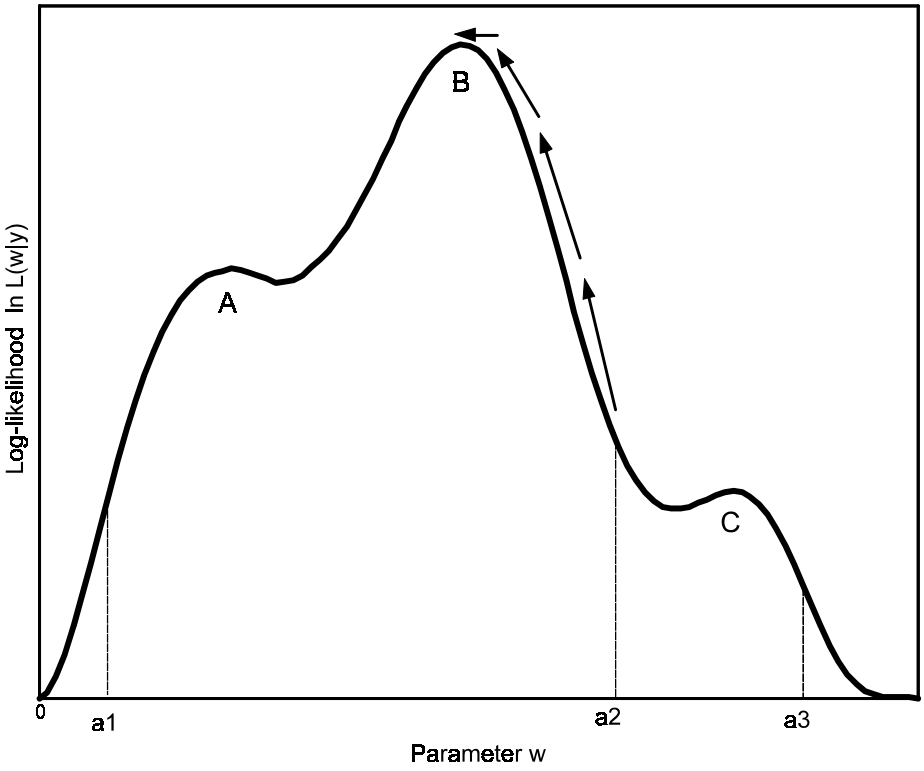

Fig. 3. A schematic plot of the log-likelihood function for a fictitious one-parameter model. Point B is the global maximum whereas points A and C are two local maxima. The series of arrows depicts an iterative optimization process.

searching much smaller sub-sets of the multi-dimensional parameter space rather than exhaustively searching the whole parameter space, which becomes intractable as the number of parameters increases. The "intelligent" search proceeds by trial and error over the course of a series of iterative steps. Specifically, on each iteration, by taking into account the results from the previous iteration, a new set of parameter values is obtained by adding small changes to the previous parameters in such a way that the new parameters are likely to lead to improved performance. Different optimization algorithms differ in how this updating routine is conducted. The iterative process, as shown by a series of arrows in Fig. 3, continues until the parameters are judged to have converged (i.e., point B in Fig. 3) on the optimal set of parameters on an appropriately predefined criterion. Examples of the stopping criterion include the maximum number of iterations allowed or the minimum amount of change in parameter values between two successive iterations.

3.2. Local maxima

It is worth noting that the optimization algorithm does not necessarily guarantee that a set of parameter values that uniquely maximizes the log-likelihood will be found. Finding optimum parameters is essentially a heuristic process in which the optimization algorithm tries to improve upon an initial set of parameters that is supplied by the user. Initial parameter values are chosen either at random or by guessing. Depending upon the choice of the initial parameter values, the algorithm could prematurely stop and return a sub-optimal set of parameter values. This is called the *local maxima* problem. As an example, in Fig. 3 note that although the starting parameter value at point a2 will lead to the optimal point B called the *global maximum*, the starting parameter value at point a1 will lead to point A, which is a sub-optimal solution. Similarly, the starting parameter value at a3 will lead to another sub-optimal solution at point C.

Unfortunately, there exists no general solution to the local maximum problem. Instead, a variety of techniques have been developed in an attempt to avoid the problem, though there is no guarantee of their effectiveness. For example, one may choose different starting values over multiple runs of the iteration procedure and then examine the results to see whether the same solution is obtained repeatedly. When that happens, one can conclude with some confidence that a global maximum has been found.[2]

[2] A stochastic optimization algorithm known as simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983) can overcome the local maxima problem, at least in theory, though the algorithm may not be a feasible option in practice as it may take an realistically long time to find the solution.

3.3. Relation to least-squares estimation

Recall that in MLE we seek the parameter values that are *most likely* to have produced the data. In LSE, on the other hand, we seek the parameter values that provide the *most accurate* description of the data, measured in terms of how closely the model fits the data under the square-loss function. Formally, in LSE, the *sum of squares error* (SSE) between observations and predictions is minimized:

$$SSE(w) = \sum_{i=1}^{m} (y_i - prd_i(w))^2, \tag{12}$$

where $prd_i(w)$ denotes the model's prediction for the ith observation. Note that $SSE(w)$ is a function of the parameter vector $w = (w_1, \ldots, w_k)$.

As in MLE, finding the parameter values that minimize SSE generally requires use of a non-linear optimization algorithm. Minimization of LSE is also subject to the local minima problem, especially when the model is non-linear with respect to its parameters. The choice between the two methods of estimation can have non-trivial consequences. In general, LSE estimates tend to differ from MLE estimates, especially for data that are not normally distributed such as proportion correct and response time. An implication is that one might possibly arrive at different conclusions about the same data set depending upon which method of estimation is employed in analyzing the data. When this occurs, MLE should be preferred to LSE, unless the probability density function is unknown or difficult to obtain in an easily computable form, for instance, for the diffusion model of recognition memory (Ratcliff, 1978).[3] There is a situation, however, in which the two methods intersect. This is when observations are independent of one another and are normally distributed with a constant variance. In this case, maximization of the log-likelihood is equivalent to minimization of SSE, and therefore, the same parameter values are obtained under either MLE or LSE.

4. Illustrative example

In this section, I present an application example of maximum likelihood estimation. To illustrate the method, I chose forgetting data given the recent surge of interest in this topic (e.g. Rubin & Wenzel, 1996; Wickens, 1998; Wixted & Ebbesen, 1991).

Among a half-dozen retention functions that have been proposed and tested in the past, I provide an example of MLE for the two functions, power and exponential. Let $w = (w_1, w_2)$ be the parameter vector, t time, and $p(w,t)$ the model's prediction of the probability of correct recall at time t. The two models are defined as

$$\begin{aligned} &\text{power model}: \quad p(w,t) = w_1 t^{-w_2} \quad (w_1, w_2 > 0), \\ &\text{exponential model}: \quad p(w,t) = w_1 \exp(-w_2 t) \quad (w_1, w_2 > 0). \end{aligned} \tag{13}$$

Suppose that data $y = (y_1, \ldots, y_m)$ consists of m observations in which $y_i (0 \leqslant y_i \leqslant 1)$ represents an observed proportion of correct recall at time t_i $(i = 1, \ldots, m)$. We are interested in testing the viability of these models. We do this by fitting each to observed data and examining its goodness of fit.

Application of MLE requires specification of the PDF $f(y|w)$ of the data *under each model*. To do this, first we note that each observed proportion y_i is obtained by dividing the number of correct responses (x_i) by the total number of independent trials (n), $y_i = x_i/n$ $(0 \leqslant y_i \leqslant 1)$ We then note that each x_i is binomially distributed with probability $p(w,t)$ so that the PDFs for the power model and the exponential model are obtained as

$$\begin{aligned} &\text{power}: \quad f(x_i \mid n, w) = \frac{n!}{(n-x_i)!x_i!} (w_1 t_i^{-w_2})^{x_i} (1 - w_1 t_i^{-w_2})^{n-x_i}, \\ &\text{exponential}: \quad f(x_i \mid n, w) = \frac{n!}{(n-x_i)!x_i!} (w_1 \exp(-w_2 t_i))^{x_i} (1 - w_1 \exp(-w_2 t_i))^{n-x_i}, \end{aligned} \tag{14}$$

where $x_i = 0, 1, \ldots, n, \ \ i = 1, \ldots, m$.

There are two points to be made regarding the PDFs in the above equation. First, the probability parameter of a binomial probability distribution (i.e. w in Eq. (4)) is being modeled. Therefore, the PDF for each model in Eq. (14) is obtained by simply replacing the probability parameter w in Eq. (4) with the model equation, $p(w,t)$, in Eq. (13). Second, note that y_i is related to x_i by a fixed scaling constant, $1/n$. As such, any statistical conclusion regarding x_i is applicable directly to y_i, except for the scale transformation. In particular, the PDF for y_i, $f(y_i|n,w)$, is obtained by simply replacing x_i in $f(x_i|n,w)$ with ny_i.

Now, assuming that x_i's are statistically independent of one another, the desired log-likelihood function for the power model is given by

$$\begin{aligned} &\ln L(w = (w_1, w_2) | n, x) \\ &\quad = \ln(f(x_1|n,w) \cdot f(x_2 \mid n, w) \cdots f(x_m \mid n, w)) \\ &\quad = \sum_{i=1}^{m} \ln f(x_i|n,w) \\ &\quad = \sum_{i=1}^{m} (x_i \ln(w_1 t_i^{-w_2}) + (n - x_i) \ln(1 - w_1 t_i^{-w_2}) \\ &\qquad + \ln n! - \ln(n - x_i)! - \ln x_i!). \end{aligned} \tag{15}$$

[3] For this model, the PDF is expressed as an infinite sum of transcendental functions.

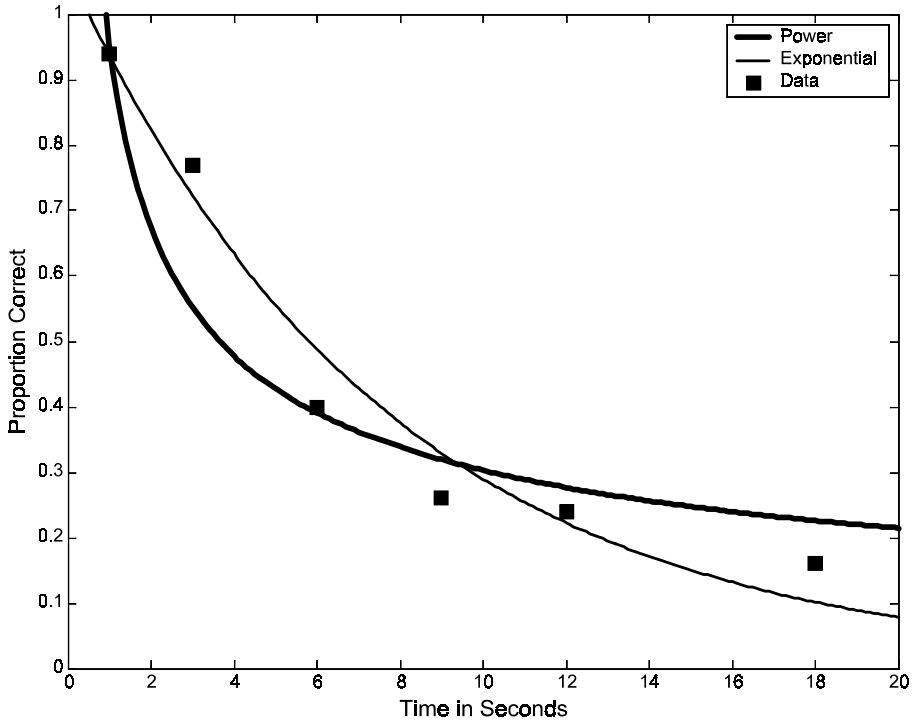

Fig. 4. Modeling forgetting data. Squares represent the data in Murdock (1961). The thick (respectively, thin) curves are best fits by the power (respectively, exponential) models.

Table 1
Summary fits of Murdock (1961) data for the power and exponential models under the maximum likelihood estimation (MLE) method and the least-squares estimation (LSE) method.

	MLE		LSE	
	Power	Exponential	Power	Exponential
Loglik/SSE (r^2)	−313.37 (0.886)	−305.31 (0.963)	0.0540 (0.894)	0.0169 (0.967)
Parameter w_1	0.953	1.070	1.003	1.092
Parameter w_2	0.498	0.131	0.511	0.141

Note: For each model fitted, the first row shows the maximized log-likelihood value for MLE and the minimized sum of squares error value for LSE. Each number in the parenthesis is the proportion of variance accounted for (i.e. r^2) in that case. The second and third rows show MLE and LSE parameter estimates for each of w_1 and w_2. The above results were obtained using Matlab code described in the appendix.

This quantity is to be maximized with respect to the two parameters, w_1 and w_2. It is worth noting that the last three terms of the final expression in the above equation (i.e., $\ln n! - \ln(n - x_i)! - \ln x_i!$) do not depend upon the parameter vector, thereby do not affecting the MLE results. Accordingly, these terms can be ignored, and their values are often omitted in the calculation of the log-likelihood. Similarly, for the exponential model, its log-likelihood function can be obtained from Eq. (15) by substituting $w_1 \exp(-w_2 t_i)$ for $w_1 t_i^{-w_2}$.

In illustrating MLE, I used a data set from Murdock (1961). In this experiment subjects were presented with a set of words or letters and were asked to recall the items after six different retention intervals, $(t_1, \ldots, t_6) = (1, 3, 6, 9, 12, 18)$ in seconds and thus, $m = 6$. The proportion recall at each retention interval was calculated based on 100 independent trials (i.e. $n = 100$) to yield the observed data $(y_1, \ldots, y_6) = (0.94, 0.77, 0.40, 0.26, 0.24, 0.16)$, from which the number of correct responses, x_i, is obtained as $100y_i$, $i = 1, \ldots, 6$. In Fig. 4, the proportion recall data are shown as squares.

The curves in Fig. 4 are best fits obtained under MLE. Table 1 summarizes the MLE results, including fit measures and parameter estimates, and also include the LSE results, for comparison. Matlab code used for the calculations is included in the appendix.

The results in Table 1 indicate that under either method of estimation, the exponential model fit better than the power model. That is, for the former, the log-likelihood was larger and the SSE smaller than for the latter. The same conclusion can be drawn even in terms of r^2. Also note the appreciable discrepancies in parameter estimate between MLE and LSE. These differences are not unexpected and are due to the fact

that the proportion data are binomially distributed, not normally distributed. Further, the constant variance assumption required for the equivalence between MLE and LSE does not hold for binomial data for which the variance, $\sigma^2 = np(1-p)$, depends upon proportion correct p.

4.1. MLE interpretation

What does it mean when one model fits the data better than does a competitor model? It is important not to jump to the conclusion that the former model does a better job of capturing the underlying process and therefore represents a closer approximation to the true model that generated the data. A good fit is a necessary, but not a sufficient, condition for such a conclusion. A superior fit (i.e., higher value of the maximized log-likelihood) merely puts the model in a list of candidate models for further consideration. This is because a model can achieve a superior fit to its competitors for reasons that have nothing to do with the model's fidelity to the underlying process. For example, it is well established in statistics that a complex model with many parameters fits data better than a simple model with few parameters, even if it is the latter that generated the data. The central question is then how one should decide among a set of competing models. A short answer is that a model should be selected based on its generalizability, which is defined as a model's ability to fit current data but also to predict future data. For a thorough treatment of this and related issues in model selection, the reader is referred elsewhere (e.g. Linhart & Zucchini, 1986; Myung, Forster, & Browne, 2000; Pitt, Myung, & Zhang, 2002).

5. Concluding remarks

This article provides a tutorial exposition of maximum likelihood estimation. MLE is of fundamental importance in the theory of inference and is a basis of many inferential techniques in statistics, unlike LSE, which is primarily a descriptive tool. In this paper, I provide a simple, intuitive explanation of the method so that the reader can have a grasp of some of the basic principles. I hope the reader will apply the method in his or her mathematical modeling efforts so a plethora of widely available MLE-based analyses (e.g. Batchelder & Crowther, 1997; Van Zandt, 2000) can be performed on data, thereby extracting as much information and insight as possible into the underlying mental process under investigation.

Acknowledgments

This work was supported by research Grant R01 MH57472 from the National Institute of Mental Health. The author thanks Mark Pitt, Richard Schweickert, and two anonymous reviewers for valuable comments on earlier versions of this paper.

Appendix

This appendix presents Matlab code that performs MLE and LSE analyses for the example described in the text.

Matlab Code for MLE

```
%   This is the main program that finds MLE estimates. Given a model, it
%   takes sample size (n), time intervals (t) and observed proportion correct
%   (y) as inputs. It returns the parameter values that maximize the log-
% likelihood function
global n t x; % define global variables
opts = optimset ('DerivativeCheck','off','Display','off','TolX',1e-6,'TolFun',1e-6,
'Diagnostics','off','MaxIter',200,LargeScale','off');
   % option settings for optimization algorithm
n = 100 ;% number of independent Bernoulli trials (i.e., sample size)
t = [1 3 6 9 12 18]';% time intervals as a column vector
y = [.94 .77 .40 .26 .24 .16]';% observed proportion correct as a column vector
x = n*y;% number of correct responses

init_w = rand(2,1);% starting parameter values
low_w = zeros(2,1);% parameter lower bounds
up_w = 100*ones(2,1);% parameter upper bounds

while 1,
[w1,lik1,exit1] = fmincon ('power_mle',init_w,[],[],[],[],low_w,up_w,[],opts);
```

```
   % optimization for power model that minimizes minus log-likelihood (note that minimization of
   minus log-likelihood is equivalent to maximization of log-likelihood)
   % w1: MLE parameter estimates
   % lik1: maximized log-likelihood value
   % exit1: optimization has converged if exit1 >0 or not otherwise
[w2, lik2, exit2] = FMINCON('EXPO_MLE', INIT_W,[], [], [], [], LOW_W, UP_W, [], OPTS);
   % optimization for exponential model that minimizes minus log-likelihood
prd1 = w1(1, 1)*t.^(-w1(2, 1));% best fit prediction by power model
r2(1, 1) = 1-sum((prd1-y).^2)/sum((y-mean(y)).^2); % r^2 for power model
prd2 = w2(1, 1)*exp(-w2(2, 1)*t);% best fit prediction by exponential model
r2(2, 1) = 1-sum((prd2-y).^2)/sum((y-mean(y)).^2); %r^2 for exponential model

if sum(r2>0) == 2
        break;
else
     init_w = rand(2, 1);
end;
end;

format long;
disp(num2str([w1 w2 r2],5));% display results
disp(num2str([lik1 lik2 exit1 exit2],5));% display results
end % end of the main program

function loglik = power_mle(w)
% POWER_MLE The log-likelihood function of the power model
   global n t x;
   p = w(1, 1)*t.^(-w(2, 1));% power model prediction given parameter
p = p + (p == zeros(6, 1))*1e-5 - (p == ones(6, 1))*1e-5; % ensure 0<p<1
  loglik = (-1)*(x.*log(p) + (n-x).*log(1-p));
   % minus log-likelihood for individual observations
  loglik = sum(loglik);% overall minus log-likelihood being minimized

function loglik = expo_mle(w)
% EXPO_MLE The log-likelihood function of the exponential model
   global n t x;
   p = w(1, 1)*exp(-w(2, 1)*t);% exponential model prediction
   p = p + (p == zeros(6, 1))*1e-5 - (p == ones(6, 1))*1e-5; % ensure 0<p<1
   loglik = (-1)*(x.*log(p) + (n-x).*log(1p));
        % minus log-likelihood for individual observations
loglik = sum(loglik);% overall minus log-likelihood being minimized
```

Matlab Code for LSE

```
% This is the main program that finds LSE estimates. Given a model, it
% takes sample size (n), time intervals (t) and observed proportion correct
% (y) as inputs. It returns the parameter values that minimize the sum of
% squares error
global t; % define global variable
opts = optimset('DerivativeCheck','off','Display','off','TolX',1e-6,'TolFun',1e-6, 'Diagnostic-
s','off','MaxIter',200,'LargeScale','off');
             % option settings for optimization algorithm
n = 100; % number of independent binomial trials (i.e., sample size)
t = [1 3 6 9 12 18]';% time intervals as a column vector
```

```
y = [.94 .77 .40 .26 .24 .16]';% observed proportion correct as a column vector

init_w = rand(2,1);% starting parameter values
low_w = zeros(2,1);% parameter lower bounds
up_w = 100*ones(2,1);% parameter upper bounds
[w1, sse1, res1, exit1] = lsqnonlin('power_lse', init_w, low_w, up_w, opts, y);
            % optimization for power model
            % w1: LSE estimates
            % sse1: minimized SSE value
            % res1: value of the residual at the solution
            % exit1: optimization has converged if exit1 >0 or not otherwise
[w2, sse2, res2, exit2] = lsqnonlin('expo_lse', init_w, low_w, up_w, opts, y);
            % optimization for exponential model

r2(1,1) = 1-sse1/sum((y-mean(y)).^2);% r^2 for power model
r2(2,1) = 1-sse2/sum((y-mean(y)).^2);% r^2 for exponential model

format long;
disp(num2str([w1 w2 r2],5));% display out results
disp(num2str([sse1 sse2 exit1 exi2],5));% display out results
end % end of the main program

function dev = power_lse(w,y)
% POWER_LSE The deviation between observation and prediction of the power
% model
   global t;
   p = w(1,1)*t.^(-w(2,1));% power model prediction
   dev = p - y;
                % deviation between prediction and observation, the square of which is
                    being minimized

function dev = expo_lse(w,y)
% EXPO_LSE The deviation between observation and prediction of the
% exponential model
     global t;
     p = w(1,1)*exp(-w(2,1)*t);% exponential model prediction
     dev = p - y;
             % deviation between prediction and observation, the square of which is
             being minimized
```

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrox, B.N., & Caski, F. *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Batchelder, W. H., & Crowther, C. S. (1997). Multinomial processing tree models of factorial categorization. *Journal of Mathematical Psychology*, *41*, 45–55.

Bickel, P. J., & Doksum, K. A. (1977). *Mathematical statistics*. Oakland, CA: Holden-day, Inc.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxberry.

DeGroot, M. H., & Schervish, M. J. (2002). *Probability and statistics* (3rd ed.). Boston, MA: Addison-Wesley.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review*, *107*(2), 227–260.

Linhart, H., & Zucchini, W. (1986). *Model selection*. New York, NY: Wiley.

Murdock Jr., B. B. (1961). The retention of individual items. *Journal of Experimental Psychology*, *62*, 618–625.

Myung, I. J., Forster, M., & Browne, M. W. (2000). Special issue on model selection. *Journal of Mathematical Psychology*, *44*, 1–2.

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

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

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Spanos, A. (1999). *Probability theory and statistical inference*. Cambridge, UK: Cambridge University Press.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice; The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*(3), 424–465.

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

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