Maximum Likelihood Estimation

In Jae Myung
Department of Psychology
Ohio State University
1885 Neil Avenue Mall
Columbus, Ohio 43210-1222
Email: myung.1@osu.edu

11-21-2001
Submitted for Publication

Abstract

In this paper I provide a tutorial exposition on the maximum likelihood estimation (MLE). Unlike least-squares estimation which is primarily a descriptive tool, MLE is by far the most popular method of parameter estimation 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.

Introduction

In the psychological sciences, we seek to uncover general laws and principles that govern the behavior under investigation. As these 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 mathematical functions 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 the model's goodness of fit, that is, how well the model fits the observed pattern of 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 generally accepted methods of parameter estimation. They are least-squares estimation (LSE) and maximum likelihood estimation (MLE). The former is well known to us as many of the familiar statistical concepts such as linear regression, the sum of squares error, the proportion variance accounted for (i.e., r^2), and the root mean squared deviation are tied to the method. On the other hand, MLE is not widely recognized among modelers in psychology, though it is, by far, the most commonly used method of parameter estimation in the statistics community. LSE might be useful for obtaining a descriptive measure for the purpose of summarizing observed data, but MLE is more suitable for statical inference such as model comparison (i.e., which model should we choose). LSE has no basis for constructing confidence intervals or testing hypotheses whereas both are naturally built into MLE. For example, MLE is a prerequisite for the chi-square test, the G-square test, Bayesian modeling, and many model selection criteria such as the Akaike Information Criterion (Akaike, 1973) and the Bayesian Information Criteria (Schwarz, 1978). In a

sense, LSE can be thought of as a special application of MLE.

In this tutorial paper, I introduce the maximum likelihood estimation method of mathematical 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. As such, the paper was written for the reader who has completed a year of statistics courses. For in-depth, technically more rigorous treatment of the topic, the reader is directed to other sources (e.g., Bickel & Doksum, 1977; Casella & Berger, 2002; Spanos, 1999).

Model Specification

Probability Distribution Function

From a statistical standpoint, the data vector $y = (y_1,...,y_m)$ as the outcome of an experiment 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.

We denote the *probability distribution function* (PDF) by $f(y|w)$ that specifies the probability of observing data y given the parameter w. The parameter vector $w = (w_1,...,w_k)$ 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 $y = (y_1,...,y_m)$ can be expressed as a multiplication of PDFs for individual observations,

$$f(y|w) = f(y_1|w)f(y_2|w)\cdots f(y_m|w) \tag{1}$$

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 independent binary trials (e.g., coin tossing experiment) 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 then given by

$$f(y|w = 0.2) = \frac{10!}{y!(10-y)!}(0.2)^y(0.8)^{10-y} \qquad (y = 0,1,...,10) \tag{2}$$

which is known as the binomial probability distribution. The shape of this PDF is shown in the top panel of Figure 1. If the parameter value is changed to say w = 0.7, a new PDF is obtained as

$$f(y|w = 0.7) = \frac{10!}{y!(10-y)!}(0.7)^y(0.3)^{10-y} \qquad (y = 0,1,...,10) \tag{3}$$

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

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

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

Likelihood Function

Given a set of parameter values, the corresponding PDF describes tells us how some data are more probable than other data. For example, in the previous example, the PDF for w = 0.2, data y = 2 is more likely to occur than data 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, we are interested in finding the one PDF, among all probability distributions 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)

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

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

$$L(w; y = 7) = f(y = 7|w) = \frac{10!}{7!3!} w^7 (1 - w)^3, \quad (0 \le w \le 1) \tag{6}$$

The shape of this likelihood function is shown in Figure 2. For simplicity, we often use L(w) for L(w;y) with the understanding that observed data y is implicit. This notation will be adopted from this point on.

There exist an important difference between the PDF f(y|w) and the likelihood function L(w;y). As illustrated in Figure 1 and Figure 2, the two functions are defined on different axes, and therefore are not directly comparable to each other. Specifically, the PDF in Figure 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, Figure 1 tells us the probability of a particular data value for a fixed parameter. Figure 2 tells us the likelihood ("un-normalized probability") of a particular parameter value for a fixed data set. Note that the likelihood function in this figure is a curve because there is only one parameter. If the model has two parameters instead, the likelihood function would be a surface sitting above parameter space. In general, for a model with k parameters, the likelihood function L(w) takes the shape of a k-dim geometrical "surface" sitting above a k-dim hyperplane spanned by the parameter vector $w = (w_1,...w_k)$.

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 (Figure 1), we are interested in finding the parameter value that corresponds to the desired PDF.

The principle of *maximum likelihood estimation* (MLE), originally developed by R. A. Fisher

in the 1920s, states that the desired probability distribution be the one that makes the observed data most likely, which is obtained by seeking the value of the parameter vector that maximizes the likelihood function L(w). The resulting parameter, which is sought by searching the multi-dimensional parameter space, is called the *MLE estimate*, denoted by $w_{MLE} = (w_{1,MLE}, ..., w_{k,MLE})$. For example, in Figure 2, the MLE estimate is $w_{MLE} = 0.7$ for which the maximized likelihood value is $L(w_{MLE} = 0.7) = 0.267$, as indicated by the dotted lines. The probability distribution corresponding to this MLE estimate is shown in the bottom panel of Figure 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 by which the probability distribution that makes the observed data most likely is sought.

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, in practice, the MLE estimate is obtained by maximizing the log-likelihood function, ln L(w), instead of L(w). This is because the two functions, ln L(w) and L(w), are monotonically related to each other so the same MLE estimate is obtained by maximizing either one. Assuming the log-likelihood function, ln L(w), is differentiable, if w_{MLE} exists, it must satisfy the following partial differential equation known as the *likelihood equation*:

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

at $w_i = w_{i,MLE}$ for all i=1,...,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 $L(w_{MLE})$ 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 $w_i = w_{i, MLE}$ for i=1,...,k.[1],

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

To illustrate the MLE procedure, let us again consider the previous one-parameter binomial example. First, by taking the logarithm of the likelihood function L(w) in Eq. (6), we obtain the log-likelihood as

---

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

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

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

$$\frac{d \ln L(w)}{dw} = \frac{7}{w} - \frac{3}{1 - w} = \frac{7 - 10w}{w(1 - w)} \tag{10}$$

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

$$\frac{d^2 \ln L(w)}{dw^2} = -\frac{7}{w^2} - \frac{3}{(1 - w)^2} = -47.62 < 0 \tag{11}$$

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 nonlinear. In such situations, the MLE estimate must be sought numerically using nonlinear optimization algorithms. The basic idea of nonlinear optimization is to quickly find optimal parameters that maximize the log-likelihood. This is done by 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 Figure 3, continues until the parameters are judged to have converged on the optimal set of parameters on an appropriately predefined criterion (i.e., point B in Figure 3). 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.

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 Figure 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 remedies 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 could conclude with some confidence that a global maximum may have been found.[2]

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 i-th observation. Note that SSE(w) is a function of the parameter vector $w = (w_1,...,w_k)$.

As in MLE, finding the parameter values that minimize SSE generally requires use of a nonlinear 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. 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.

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, 1996; 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 following two:

$$\begin{aligned} &\textbf{Power model}: && p(w,t) = w_1 t^{-w_2} && (w_1, w_2 > 0) \\ &\textbf{Exponential model}: && p(w,t) = w_1 \exp(-w_2 t) && (w_1, w_2 > 0) \end{aligned} \tag{13}$$

where $w = (w_1, w_2)$ is the parameter vector, t is the time, and p(w,t) is the model's prediction of the probability of correct recall at time t. Suppose that data $y = (y_1,...,y_m)$ consists of m observations in

[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 realistic option in practice.

which y_i ($0 \le y_i \le 1$) represents an observed proportion of correct recall at time t_i (i = 1,..., 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), $y_i = x_i/n$ ($0 \le y_i \le 1$) 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

$$\begin{aligned}
&\textbf{Power}: f(x_i|w) = \frac{n!}{(n-x_i)!\,x_i!}\left(w_1 t_i^{-w_2}\right)^{x_i}\left(1-w_1 t_i^{-w_2}\right)^{n-x_i} \\
&\textbf{Exponential}: f(x_i|w) = \frac{n!}{(n-x_i)!\,x_i!}\left(w_1 \exp(-w_2 t_i)\right)^{x_i}\left(1-w_1 \exp(-w_2 t_i)\right)^{n-x_i}
\end{aligned} \tag{14}$$

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

Regarding the PDFs in the above equation, there are two points to be made. First, note that what is being modeled is the probability parameter of a binomial probability distribution (i.e., w in Eq. (4)). Therefore, the desired PDF for each model, as shown in Eq. (13), is obtained simply by substituting the model equation, p(w,t), in Eq. (13) for the probability parameter. 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, $f(y_i|w)$, is obtained simply by replacing x_i in $f(x_i|w)$ with $n \cdot y_i$.

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

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

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, thereby not affecting 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, ..., t_6) = (1, 3, 6, 9, 12, 18) (i.e., m = 6), and the proportion recall on each retention interval was calculated based on 100 independent trials (i.e., n = 100) to yield the observed data (y_1,

..., 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,..,6. In Figure 4, the proportion recall data are shown as squares.

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

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 low shows the maximized log-likelihood value for MLE or the minimized sum of squares error value for LSE. Shown in the parenthesis is the proportion variance accounted for (i.e., r^2). The second and third rows show MLE or LSE parameter estimates. The above results were obtained using Matlab code described 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 (smaller SSE) than for the latter. The same conclusion can be drawn even in terms of r^2. Also in the table 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 not normally distributed but binomially 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.

MLE Interpretation

What does it mean when a model fits the data better than its competitor model? It is important not to jump to the conclusion that the former model does 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 only a necessary condition but not a sufficient condition for such a conclusion. A superior fit 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. The central question is then how one should decide among a set of competing models given data. 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 predict future data, not based on its goodness of fit, which is no more than a descriptive statistic. 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, in press).

Concluding Remarks

The goal of the present article is to provide a tutorial exposition of the maximum likelihood estimation. Although it is less known to modelers in the behavioral sciences, 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 tried to provide a simple, intuitive explanation of the method so the reader can have a grasp of some of the basic principles. Ultimately we hope the reader will apply the method in his or her mathematical modeling efforts so a plethora of widely available, MLE-based analyses can be performed on data thereby extracting as much information and insight as possible into the underlying mental process under investigation.

References

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

Bickel, P. J. & Doksum, K. A. (1977). *Mathematical statistics*, chapter 3. Oakland, CA: Holden-day, inc.

Casella, G. & Berger, R. L. (2002). *Statistical inference* (2nd edition), chapter 7. Pacific Grove, CA: Duxberry.

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

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

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

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

Pitt, M. A., Myung, I. J., & Zhang, S. (in press). Toward a method of selecting among computational models of cognition. *Psychological Review*.

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

Spanos, A. (1999). *Probability theory and statistical inference*, chapter 13. Cambridge, UK: Cambridge University Press.

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

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.

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);
      % 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
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
[w1,lik1,exit1]=fmincon('power_mle',ini_w,[],[],[],[],low_w,upr_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

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 < ones(6,1)).*p+(p > ones(6,1))*1;% ensure that p lies between 0 and 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 < ones(6,1)).*p+(p > ones(6,1))*1;% ensure that p lies between 0 and 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
```

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,'Diagnostics','off','MaxIter',200);
      % 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,exit1]=lsqnonlin('power_lse',init_w,low_w,up_w,opts,y);
      % optimization for power model
      % w1: LSE estimates
      % sse1: minimized SSE value
      % exit1: optimization has converged if exit1 >0 or not otherwise
[w2,sse2,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
```

Acknowledgments

This work was supported by research grant R01 MH57472 from the National Institute of Mental Health. The author thanks Mark Pitt for valuable comments on earlier versions of this paper.
Correspondence concerning this article should be addressed to In Jae Myung, Department of Psychology, Ohio State University, 1885 Neil Avenue, Columbus, OH, 43210-1222; myung.1@osu.edu.

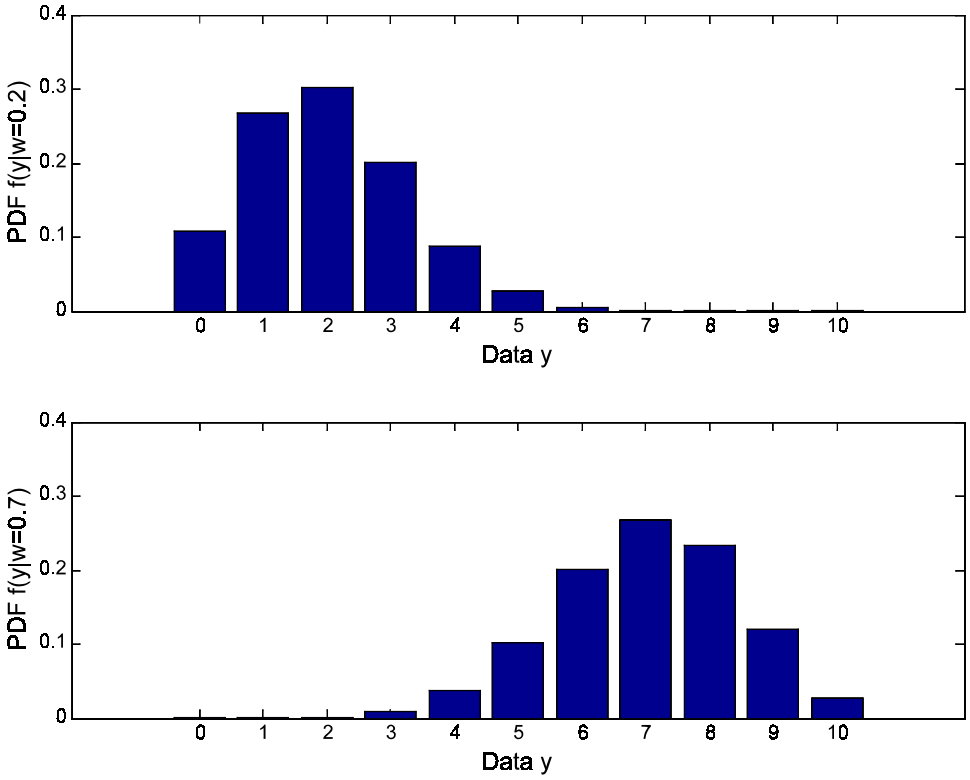

Figure 1. Binomial probability distributions of sample size n = 10 and probability parameter w = 0.2 (top) and w = 0.7 (bottom)

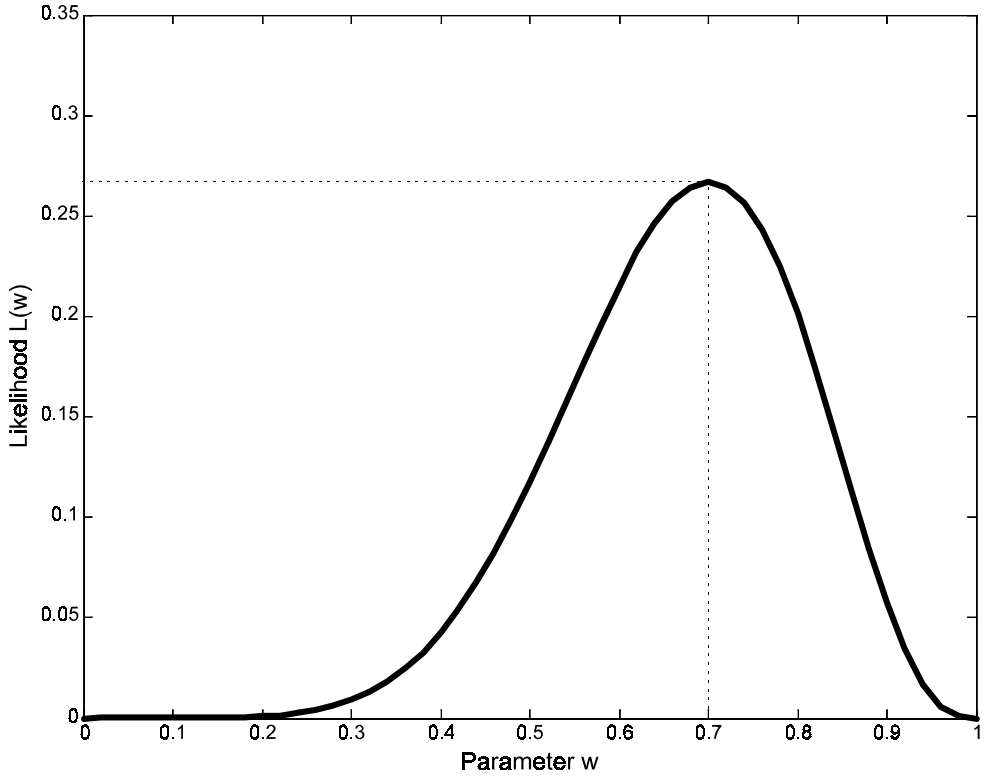

Figure 2. The likelihood function given observed data y = 7 and sample size n = 10 for the one-parameter model described in the text

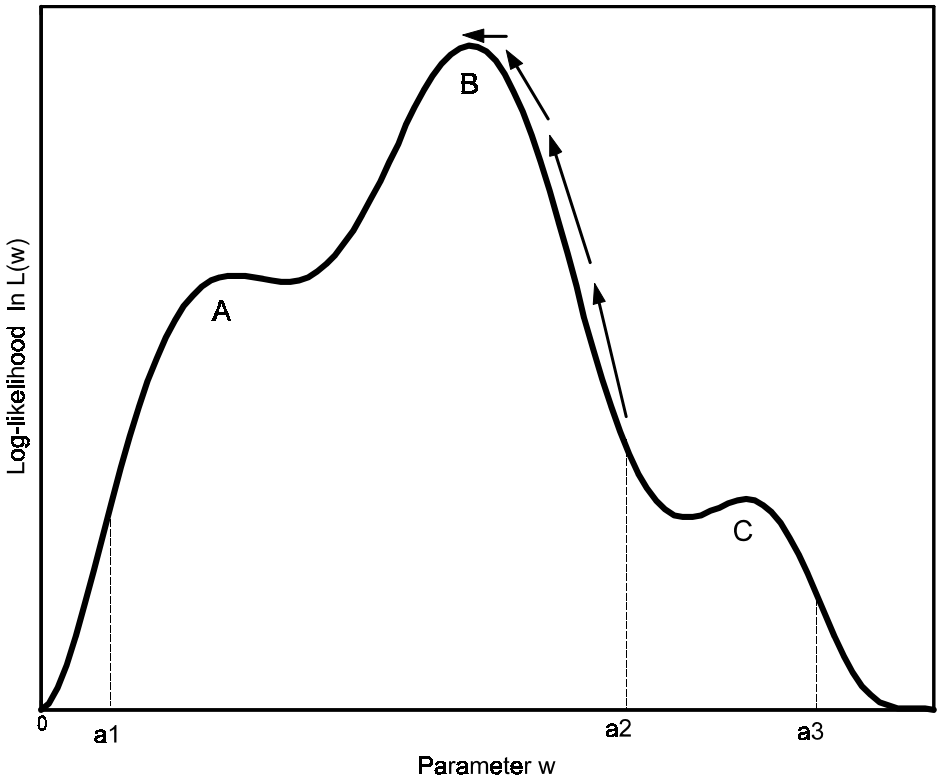

Figure 3. A schematic plot of the log-likelihood function for a fictitious one-parameter model. Point B is the global maximum whereas points A & C are two local maxima. The series of arrows depicts an iterative optimization process.

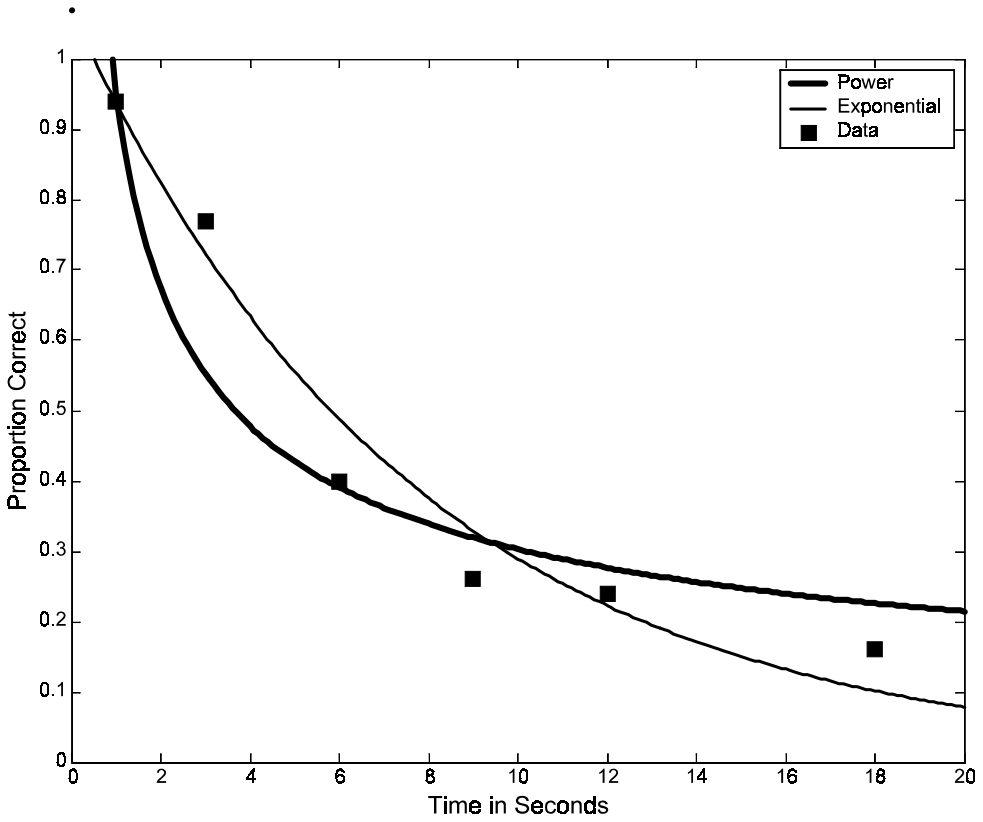

Figure 4. Modeling forgetting data. Solid rectangles represent the data in Murdock (1961). The thick and thin curves are best fits by the power and exponential models, respectively