

# Computational Cognitive Science

## Lecture 4: Maximum Likelihood Estimation

Chris Lucas  
(Slides adapted from Frank Keller's)

School of Informatics  
University of Edinburgh  
clucas2@inf.ed.ac.uk

29 September, 2017

## Introduction

Likelihood

Examples

Maximum Likelihood Estimate

## Defining a Likelihood Function

The SIMPLE Model of Serial Recall

The Data Model

## Readings:

- ▶ Chapter 4 of L&F
- ▶ Sharon Goldwater's notes on basic probability theory ([link](#))

# Probabilistic models

A cognitive model  $M$  is probabilistic if it generates a probability distribution *conditional* on  $M$  and its *parameters*  $\theta$ .

These have some appealing features; today we'll focus on connections to parameter estimation.

# Likelihood


Suppose we have  $K$  observations  $y_1, y_2, \dots, y_K$ . The  $k^{\text{th}}$  individual observation is  $y_k$  and the full sequence of observations  $\mathbf{y}$ .

$P(y_k|\theta, M)$  is the *probability mass function*<sup>1</sup> for an observation  $y_k$  given  $\theta$  and  $M$  ( $M$  is often constant, and thus omitted).

If all of our observations are independent, their joint probability is

$$P(\mathbf{y}|\theta) = \prod^k P(y_k|\theta)$$

---

<sup>1</sup>For discrete observations, e.g., whole-numbers of milliseconds. 

# Likelihood

It's common to treat  $\theta$  as fixed and  $P(\mathbf{y}|\theta)$  as a function of  $\mathbf{y}$ .

Let's instead treat  $\mathbf{y}$  as fixed, and treat  $\theta$  as the varying argument. This is called the *likelihood function*.

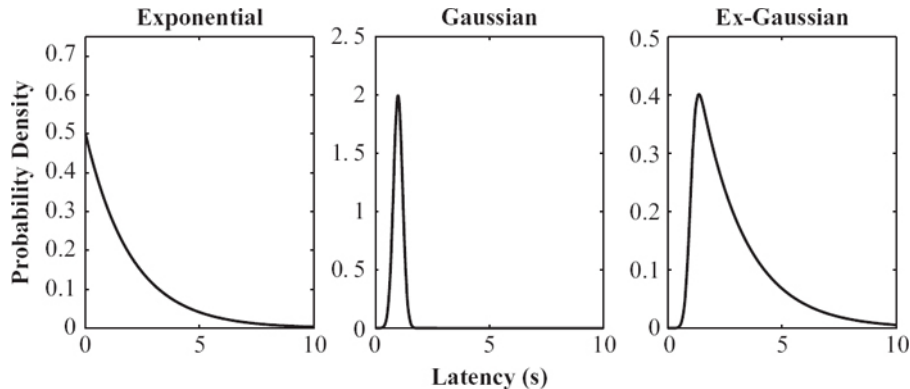
L&F use the notation  $L(\theta|\mathbf{y})$ .

If we want to turn this into a discrepancy function, a common choice is the negative log-likelihood:  $-\log(L(\theta|\mathbf{y}))$ .



## Example: Reaction Times

The exponential Gaussian function captures latencies (reaction times) from a choice experiment.

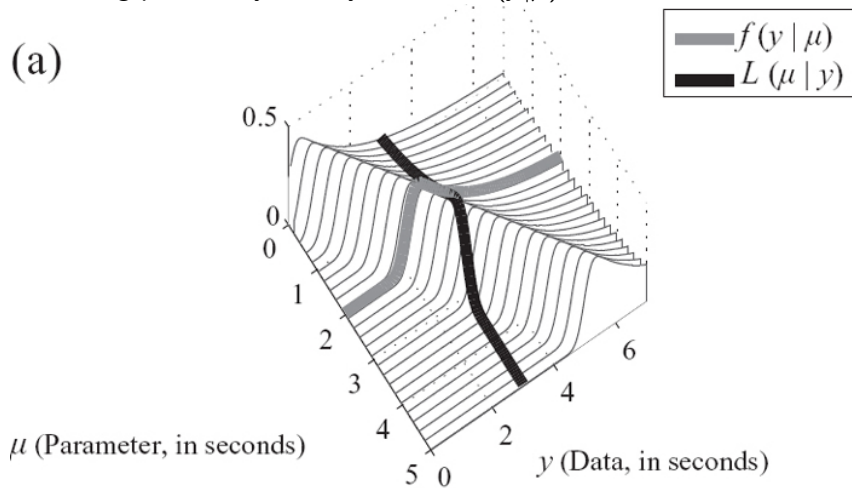


The ex-Gaussian has the following parameters:  $\mu$ ,  $\sigma$ , and  $\tau$ . Let's only consider  $\mu$  for now.

## Example: Reaction Times

For a single data point  $y$  and the parameter  $\mu$ , we get the following probability density function  $f(y|\mu)$ :

(a)



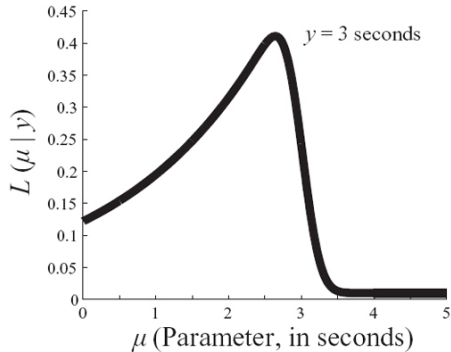
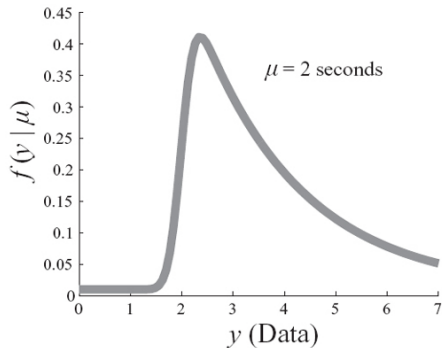
The gray line marks  $f(y|\mu = 2)$ , the black one  $L(\mu|y = 3)$ .



## Example: Reaction Times

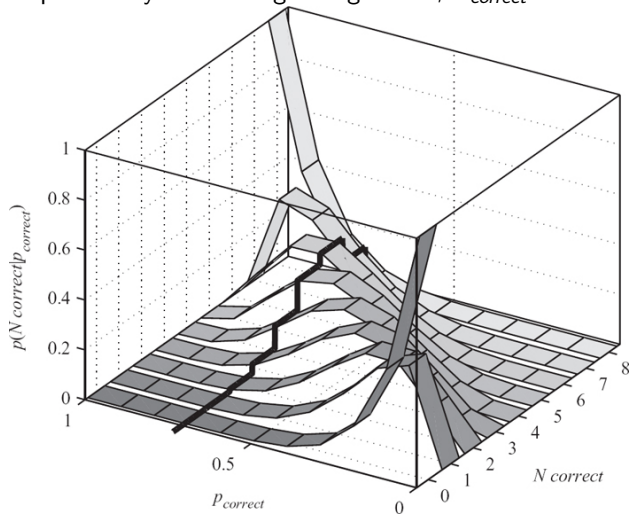
Typically, we have collected some data, and want to estimate the parameters of our model. The likelihood function  $L(\mu|y)$  tells us how probable a parameter value is given this data.

If we just plot  $f(y|\mu = 2)$  and  $L(\mu|y = 3)$ , we get:



## Example: Recall Scores

Distribution over recalled items ( $N_{correct}$ ) in a memory experiment.  
Parameter: probability of recalling a single item,  $P_{correct}$ .



Black line:  $f(N_{correct} | P_{correct} = 0.7)$ ; ribbons:  $L(P_{correct} | N_{correct})$ .

# Maximum Likelihood Estimate

Idea behind maximum likelihood estimation: determine parameter values such that they maximize the likelihood of the data.

The *maximum likelihood estimate* (MLE)  $\hat{\theta}$  of a parameter  $\theta$  is:

$$\hat{\theta} = \arg \max_{\theta} L(\theta|y)$$

This is not the same as maximizing the probability of the parameter given the data:

$$\hat{\theta} = \arg \max_{\theta} P(\theta|y) = \arg \max_{\theta} \frac{P(y|\theta)P(\theta)}{P(y)}$$

This is the *maximum a posteriori estimate* to which we return in when we discuss Bayesian estimation.

# Maximum Likelihood Estimate

Idea behind maximum likelihood estimation: determine parameter values such that they maximize the likelihood of the data.

The *maximum likelihood estimate* (MLE)  $\hat{\theta}$  of a parameter  $\theta$  is:

$$\hat{\theta} = \arg \max_{\theta} L(\theta|y) = \arg \max_{\theta} P(y|\theta)$$

This is not the same as maximizing the probability of the parameter given the data:

$$\hat{\theta} = \arg \max_{\theta} P(\theta|y) = \arg \max_{\theta} \frac{P(y|\theta)P(\theta)}{P(y)}$$

This is the *maximum a posteriori estimate* to which we return in when we discuss Bayesian estimation.

# Maximum Likelihood Estimate

Idea behind maximum likelihood estimation: determine parameter values such that they maximize the likelihood of the data.

The *maximum likelihood estimate* (MLE)  $\hat{\theta}$  of a parameter  $\theta$  is:

$$\hat{\theta} = \arg \max_{\theta} L(\theta|y) = \arg \max_{\theta} P(y|\theta)$$

This is not the same as maximizing the probability of the parameter given the data:

$$\hat{\theta} = \arg \max_{\theta} P(\theta|y) = \arg \max_{\theta} \frac{P(y|\theta)P(\theta)}{P(y)}$$

This is the *maximum a posteriori estimate* to which we return in when we discuss Bayesian estimation.

## Defining a Likelihood Function

We first need to specify a function  $f(y|\theta)$  that maps data values (outcomes of an experiment) onto probabilities.

Sometimes, the probability density function can be specified directly, e.g., the ex-Gaussian function for reaction times:

$$f(y_k|\mu, \sigma, \tau) = \frac{1}{\tau} \exp\left(\frac{\mu - y_k}{\tau} + \frac{\sigma^2}{2\tau^2}\right) \Phi\left(\frac{y_k - \mu}{\sigma} - \frac{\sigma}{\tau}\right)$$

where  $\Phi$  is the Gaussian cumulative distribution.

We assume that there are two components that generate reaction times: *time to make a decision* (exponential) and *time for encoding and motor movement* (Gaussian).

However, sometimes we additionally need a *data model* to relate data values to model probabilities.

# The SIMPLE Model of Serial Recall

The SIMPLE model (Brown, Neath, & Chater, 2007) predicts serial recall, i.e., the accuracy of recalling an item at a certain position in a list.

Assumptions:

- ▶ distinctiveness (i.e., the extent to which an item differs from other items) determines accuracy of recall;
- ▶ the distinctiveness criterion is time: during recall, we select the target items among items that occurred at around the same time;
- ▶ time is logarithmic, i.e., items that occurred longer ago are harder to distinguish.

# The SIMPLE Model of Serial Recall

Let  $T_r$  be the time of retrieval, and  $T_i$  and  $T_j$  the times associated with items  $i$  and  $j$ . The psychological distance in time is then given by  $M_i = \log(T_r - T_i)$  and  $M_j = \log(T_r - T_j)$ .

The similarity of two items is then given by:

$$\eta_{ij} = \exp(-c|M_i - M_j|)$$

where  $c$  is a scaling parameter. Then the probability of recalling item  $j$  given a probe  $i$  is:

$$P(j|i) = \frac{\eta_{ij}}{\sum_k \eta_{ik}}$$

where  $k$  ranges over all candidates. Note this is like the GCM, but using temporal similarity instead of feature similarity.



## The SIMPLE Model of Serial Recall

If we set  $j = i$ , we get the probability of correctly recalling an item:

$$P_{correct}(i) = \frac{1}{\sum_k \eta_{ik}}$$

```
function pcor = SIMPLEserial(c, presTime, recTime, J)

pcor = zeros(1,J);
Ti = cumsum(repmat(presTime,1,J));
Tr = Ti(end) + cumsum(repmat(recTime,1,J));

for i=1:J % i indexes output + probe position
    M = log(Tr(i)-Ti);
    eta = exp(-c*abs(M(i)-M));
    pcor(i) = 1./sum(eta);
end
```

# The Data Model

SIMPLE predicts a probability of correct recall for each serial position, but it doesn't predict a probability *distribution*.

In an experiment, the observed proportion correct varies from trial to trial because of sampling variability.

We can use the *binomial distribution* to model this:

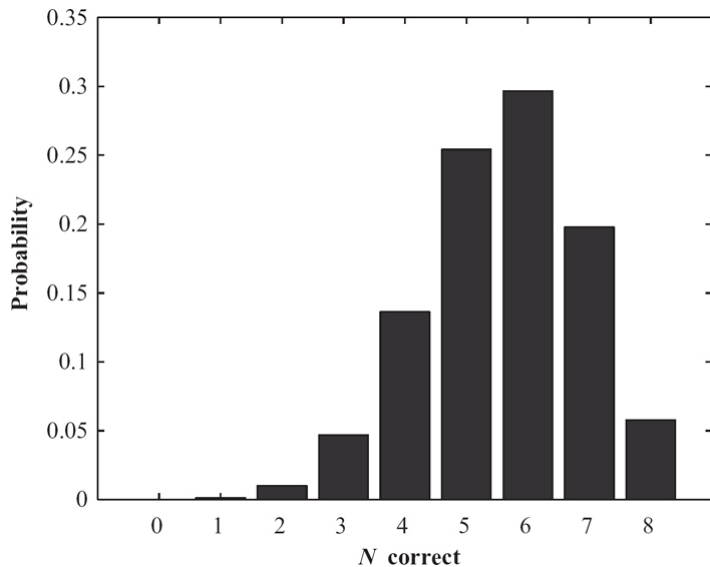
$$b(x|\theta, n) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

Assuming  $N_C(i)$  correct recalls out of  $n$  trials, we get:

$$P(N_C(i)|P_{correct}(i), n) = \binom{n}{N_C(i)} P_{correct}^{N_C(i)} (1 - P_{correct})^{n-N_C(i)}$$

# The Data Model

Distribution predicted by SIMPLE with binomial data model for  $P_{correct} = 0.7$ :



# The Data Model

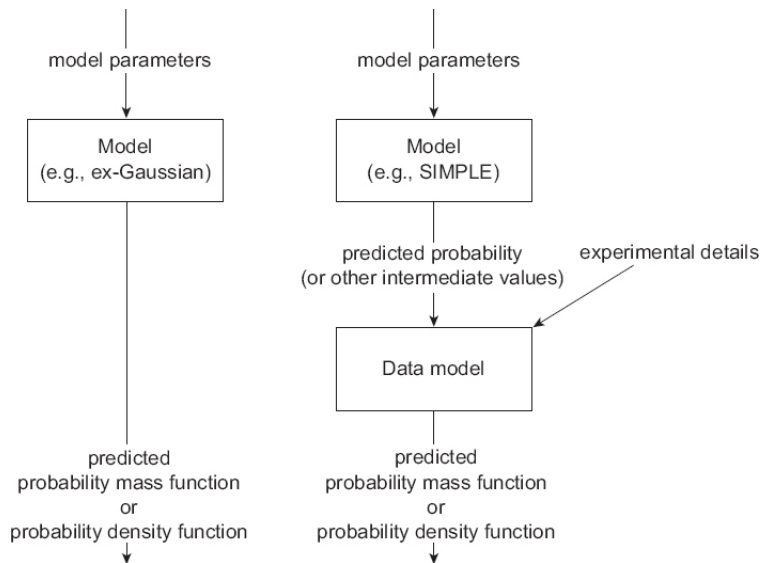
Implementation of SIMPLE with data model:

```
function pmf = SIMPLEserialBinoPMF(c, presTime, recTime, J, Nc, N)
% c is the parameter of SIMPLE
% presTime and recTime are the effective temporal
% separation of items at input and output
% J is the length of the list
% Nc (a vector) is no. of items correctly recalled at each pos.
% N is the number of trials at each position

pmf = zeros(1,J);
Ti = cumsum(repmat(presTime,1,J));
Tr = Ti(end) + cumsum(repmat(recTime,1,J));
for i=1:J % i indexes output + probe position
    M = log(Tr(i)-Ti);
    eta = exp(-c*abs(M(i)-M));
    pcor = 1./sum(eta);
    pmf(i) = binomPMF(Nc(i), N, pcor);
end
```

# The Data Model

If the model doesn't predict a distribution, we use a *data model*:



# The Data Model

We need to be careful to distinguish the following different probabilities:




- ▶ the probability  $P_{correct}(i)$  of correct recall predicted by SIMPLE;
- ▶ the probability of correct recall in the data,  $\frac{N_C}{N}$ ;
- ▶ the probability of each outcome  $N_C$  predicted by SIMPLE after applying the data model.

We can extend the data model to accommodate multiple outcomes (not just correct/incorrect, but type of error), see L&F, ch. 4.3.4.

# Summary

- ▶ The likelihood of a model  $L(\theta|y)$  is the probability of the data given the parameters, where we keep the data fixed;
- ▶ maximum likelihood estimation computes  $\arg \max_{\theta} L(\theta|y)$ , i.e., the parameters that maximize model likelihood;
- ▶ the SIMPLE model of memory recall assumes that temporal determines accuracy of recall;
- ▶ sometimes we need a data model to predict a probability distribution from the output of our model.

# References

-  Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, 3(114), 539–76.
-  Goldwater, S. (2017). *Basic probability theory*. Retrieved from <http://homepages.inf.ed.ac.uk/sgwater/teaching/general/probability.pdf>
-  Lewandowsky, S. & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage.