

Computational Cognitive Science

Lecture 4: Maximum Likelihood Estimation

Frank Keller

1 October 2026

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

Slide credits: Chris Lucas, Benjamin Peters

Overview

Parameters and Probabilities

What is Likelihood?

Defining a Probability Distribution

Example 1: Reaction times with a Wald distribution

Example 2: GCM plus a data model

Finding the Maximum Likelihood

Properties and Limits of MLE

Reading: Chapter 4 of Farrell and Lewandowsky (2018).

Sharon Goldwater's probability notes.

Parameters and Probabilities

From parameters to probabilistic models

Last time we focused on finding parameters that minimize some loss function, like root mean squared deviation (RMSD).

Today: Probabilistic approaches to parameter estimation. Useful to:

1. Understand which parameter values are better supported by the data.
2. Exploit prior knowledge about what values are reasonable.
3. Build and evaluate models that anticipate specific response distributions.
4. Capture and predict patterns that are difficult to express using standard loss functions.

From parameters to probabilistic models

A cognitive model is probabilistic if it generates a probability distribution over observations $\mathbf{y}$ conditional on its parameters θ , i.e., it generates $P(\mathbf{y}|\theta)$.

We can use the negative log **likelihood of our data** as a loss function:

- Often less ad-hoc to specify a probability distribution than a loss directly.
- Supports more nuanced predictions, e.g., judgments will be extreme but not in a particular direction.
- Offers tools to model individual differences.

We can make inferences about $P(\theta|\mathbf{y})$ if we have a **prior** $P(\theta)$.

Notation

- $\theta = [\theta_1 \ \theta_2 \ \dots]$: parameters.
- $\mathbf{y} = [y_1 \ \dots \ y_K]$: all K observations.
- $P(\mathbf{y}|\theta)$: probability mass function for discrete data.
- $f(\mathbf{y}|\theta)$: probability density function for continuous data.
- $L(\theta|\mathbf{y})$: likelihood function; the same numerical function, but now treated as a function of θ for fixed data $\mathbf{y}$.

Important: $L(\theta|\mathbf{y})$ is not $P(\theta|\mathbf{y})$. Likelihood is not a posterior probability.

What is Likelihood?

Probability function vs. likelihood function

The same expression can be used in two ways.

Prediction

$$f(\mathbf{y}|\boldsymbol{\theta})$$

- Fix the model and parameter values.
- Vary possible data values $\mathbf{y}$.
- Ask: which observations should we expect?

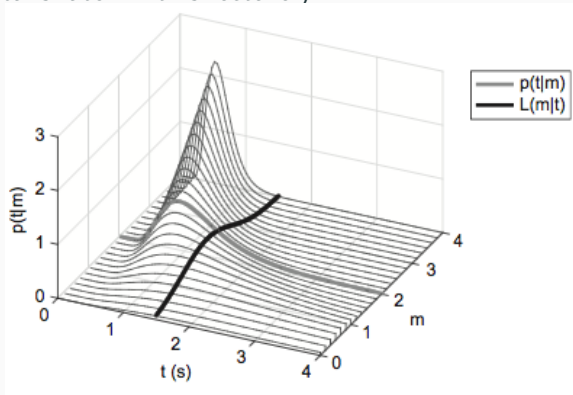
Estimation

$$L(\boldsymbol{\theta}|\mathbf{y})$$

- Fix the observed data.
- Vary possible parameter values $\boldsymbol{\theta}$.
- Ask: which parameter values are best supported?

Example: Reaction times

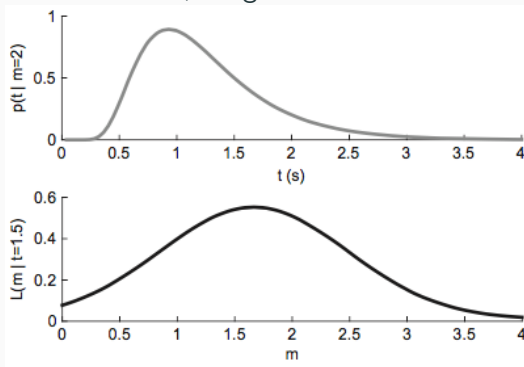
For a single reaction time t and one free parameter m , the Wald function gives a density $f(t|m)$. (Details later in this lecture.)



The grey line fixes $m = 2$ and varies t : a probability density. The black line fixes $t = 1.5$ and varies m : a likelihood function.

The same function, two cross-sections

If we plot just the two cross-sections, we get:



- Top: $f(t|m = 2)$, a density over possible reaction times.
- Bottom: $L(m|t = 1.5)$, a likelihood over possible drift values.
- The maximum of the bottom curve is the best-supported value of m for this one observation.

A word of caution about likelihood

A likelihood function is *like* a density over parameters, but it is not a probability distribution over parameters.

- Likelihoods compare parameter values for fixed data.
- They need not integrate or sum to 1.
- Ratios of likelihoods are meaningful: one parameter value can make the observed data more likely than another.
- To obtain $P(\theta|\mathbf{y})$, we need a prior and Bayes' rule.

Defining a Probability Distribution

Why the probability function matters

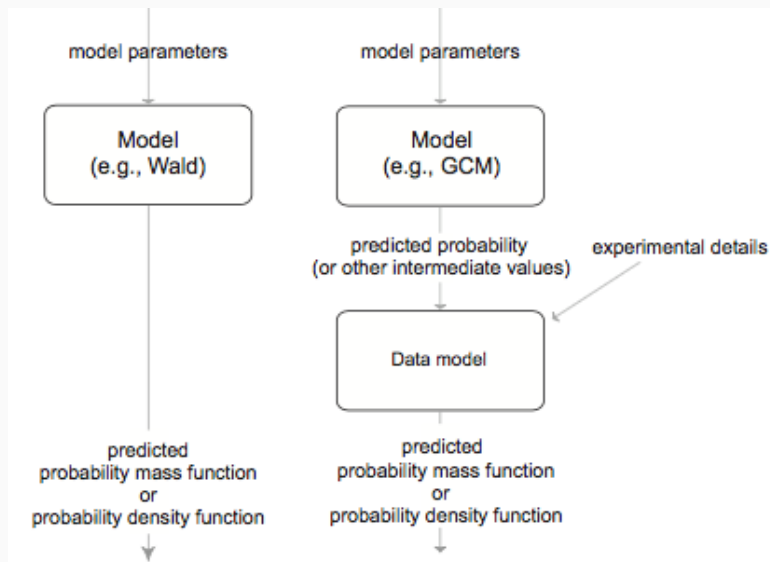
Before we can fit by likelihood, the model must specify a probability function over possible data.

- For discrete data: a probability mass function.
- For continuous data: a probability density function.
- A deterministic model by itself is not enough: it predicts one outcome with certainty and all others with probability zero.

The key modeling question is therefore:

model parameters $\longrightarrow$ probability of the observed data

Two routes to a probability function



Two routes to a probability function

Model directly specifies the distribution

Example: Wald model for reaction times.

- Parameters are fed into the psychological model.
- The model returns a full density over reaction times.
- We can evaluate $f(t|\theta)$ directly.

Model requires a data model

Example: GCM for categorization.

- The psychological model returns intermediate values, e.g., P_A .
- The data model describes sampling variability.
- Together they yield $f(N_A|\theta)$.

Example 1: Reaction times

The **Wald probability function** captures latencies from forced-choice decision experiment (see left/right lines in lecture 2).

Psychological interpretation:

- Evidence accumulates continuously over time.
- A response is made when accumulated evidence crosses a boundary.
- Randomness in the accumulation process generates a distribution of reaction times, not just a single predicted mean.

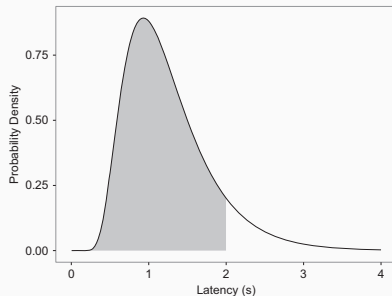
This is why the Wald function can be used directly as a likelihood model for reaction-time data.

Shifted Wald density

For response time t , boundary a , drift m , and non-decision time T :

$$f(t|a, m, T) = \frac{a}{\sqrt{2\pi(t-T)^3}} \exp\left(-\frac{[a - m(t-T)]^2}{2(t-T)}\right), \quad t > T$$

- m : drift rate, i.e., average rate of evidence accumulation.
- a : boundary position, i.e., how much evidence is required.
- T : non-decision time, e.g., encoding and motor components.



Likelihood for one reaction time

Suppose we observe one reaction time $t = 1.5$ seconds.

If a and T are fixed and we estimate only m , then:

$$L(m|t = 1.5) = f(1.5|a, m, T)$$

The maximum-likelihood estimate is the value of m that maximizes this curve:

$$\hat{m}_{\text{MLE}} = \underset{m}{\operatorname{argmax}} L(m|t = 1.5)$$

With one data point the estimate is fragile; with multiple reaction times we multiply the densities, or add their log-likelihoods.

Likelihood for multiple reaction times

For reaction times $\mathbf{t} = [t_1, \dots, t_K]$:

$$L(a, m, T|\mathbf{t}) = \prod_{k=1}^K f(t_k|a, m, T)$$

In practice we work with the log likelihood:

$$\log L(a, m, T|\mathbf{t}) = \sum_{k=1}^K \log f(t_k|a, m, T)$$

This gives a surface over the parameter space. With two free parameters, we can visualize it; with more parameters, we need numerical optimization.

Time for a short quiz on WooClap!

Go to wooclap.com

Enter the event code in the top banner

Event code

EBLNMAS



Example 2: GCM is not yet a likelihood model

The **Generalized Context Model** (GCM) of categorization predicts response probabilities from similarities to stored exemplars (see lecture 1).

For a two-category task, GCM gives an intermediate prediction:

$$P_A = P(\text{respond } A | \text{stimulus}, c, \mathbf{w})$$

- c : sensitivity/scaling parameter for similarity.
- $\mathbf{w}$: attention weights over stimulus dimensions.

But the data may be **counts**: N_A means A responses out of N trials. So we need a data model to map P_A to probabilities over possible counts N_A .

Binomial data model for GCM

If each trial is an A/B response and the GCM predicts P_A , then the number of A responses is modeled as binomial:

$$f(N_A|P_A, N) = \binom{N}{N_A} P_A^{N_A} (1 - P_A)^{N - N_A}$$

The full likelihood for the GCM parameters is therefore:

$$L(c, \mathbf{w}|N_A) = f(N_A|P_A(c, \mathbf{w}), N)$$

The psychological model predicts P_A ; the data model predicts the sampling variability in observed counts.

Worked GCM + data model example

For one stimulus in a categorization experiment:

- Suppose $N = 160$ responses are observed.
- Suppose $N_A = 155$ are A responses.
- For a proposed parameter setting $(c, \mathbf{w})$, GCM predicts $P_A(c, \mathbf{w})$.
- The binomial data model evaluates the probability of observing exactly 155 A responses:

$$L(c, \mathbf{w} | N_A = 155) = \binom{160}{155} P_A(c, \mathbf{w})^{155} [1 - P_A(c, \mathbf{w})]^5$$

Different values of $(c, \mathbf{w})$ imply different P_A , and hence different likelihoods for the same observed count.

Across many stimuli

If we have counts for many stimuli $i = 1, \dots, I$:

$$L(c, \mathbf{w} | \mathbf{N}_A) = \prod_{i=1}^I \binom{N_i}{N_{Ai}} P_{Ai}(c, \mathbf{w})^{N_{Ai}} [1 - P_{Ai}(c, \mathbf{w})]^{N_i - N_{Ai}}$$

For numerical fitting, use the log likelihood:

$$\log L(c, \mathbf{w} | \mathbf{N}_A) = \sum_{i=1}^I \log f(N_{Ai} | P_{Ai}, N_i)$$

This is the objective optimized when fitting GCM parameters by maximum likelihood.

Finding the Maximum Likelihood

Maximum likelihood estimation

Maximum likelihood estimation finds the parameter values that make the observed data most likely:

$$\theta_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} L(\theta|\mathbf{y})$$

Equivalently, because the log is monotonic:

$$\theta_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} \log L(\theta|\mathbf{y})$$

Most optimizers minimize rather than maximize, so we often minimize the **negative log likelihood** or the deviance.

Negative log likelihood and deviance

Two common objective functions:

$$\text{negative log likelihood} = -\log L(\boldsymbol{\theta}|\mathbf{y})$$

$$\text{deviance} = -2 \log L(\boldsymbol{\theta}|\mathbf{y})$$

Why log likelihoods?

- Products become sums: $\log \prod_k x_k = \sum_k \log x_k$.
- Numerical values are more manageable and less prone to underflow.
- For many models, the log form is easier to simplify and optimize.

Likelihood, log likelihood, deviance

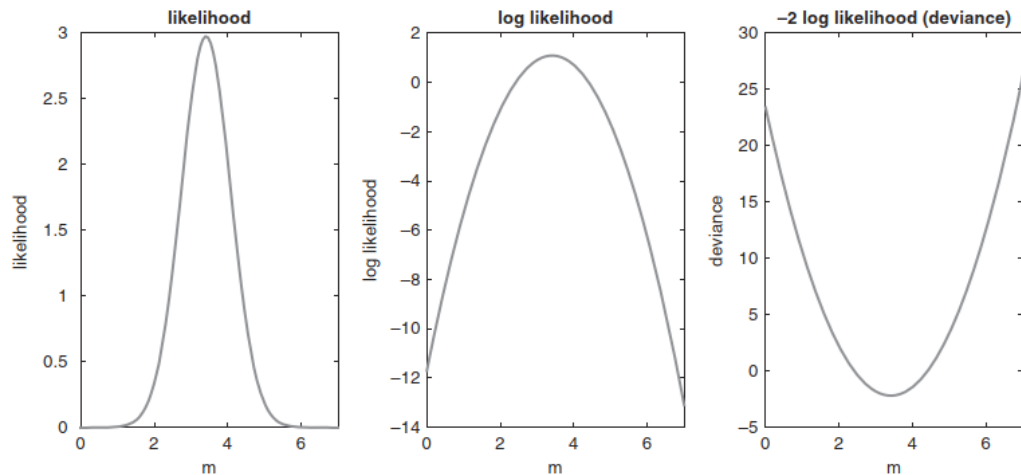


Figure 4.9 A likelihood function (left panel), and the corresponding log-likelihood function (middle) and deviance function ($-2 \log$ likelihood; right panel).

Practical optimization

In realistic models, we usually cannot find the maximum analytically.

Typical workflow:

1. Write a function that takes θ and returns $-\log L(\theta|\mathbf{y})$ or deviance.
2. Impose sensible bounds or transformations on parameters.
3. Run a numerical optimizer, often from multiple starting points.
4. Inspect the solution, the predictions, and possible local optima.

For example, in the GCM case the optimizer searches over c , attention weights, noise, and bias parameters, while the objective sums the binomial deviance across stimuli.

Properties and Limits of MLE

Useful properties of MLE

Under regularity conditions, maximum-likelihood estimators have several important properties.

- **Sufficiency**: the likelihood captures the information in the data relevant to the parameter.
- **Parameterization invariance**: transforming a parameter and then maximizing gives the transformed MLE.
- **Consistency**: estimates become more accurate on average as sample size increases.
- **Efficiency**: asymptotically, MLEs have low variance among consistent estimators.

Limits and cautions

MLE is powerful, but not magic.

- It depends on the likelihood being correctly specified.
- Small-sample estimates can be biased.
- Participant-level MLEs can be overdispersed when used to infer population variability (variance is artificially inflated).
- Point estimates do not express full uncertainty about parameters.

This is why likelihood is a foundation for later topics: **model comparison, Bayesian parameter estimation, and hierarchical models.**

Summary

- Likelihood uses the same probability function as prediction, but fixes the data and varies the parameters.
- A likelihood model requires a probability function over possible data.
- Some cognitive models provide this directly, e.g., the Wald model for reaction times.
- Other models need a data model, e.g., GCM plus a binomial or multinomial model for response counts.
- MLE chooses θ to maximize $L(\theta|\mathbf{y})$, or equivalently to minimize negative log likelihood or deviance.

Farrell, Simon and Stephan Lewandowsky. 2018. *Computational Modeling of Cognition and Behavior*. Cambridge University Press, Cambridge.