Computational Cognitive Science

Lecture 4: Parameters and probabilities

Chris Lucas

School of Informatics

University of Edinburgh

September 27, 2024

Readings

- Chapter 4 of F&L
- Chapter 6 of F&L
- Sharon Goldwater's probability notes

Parameters

Last time we focused on finding estimates for parameters that minimize some loss function, like RMSD.

Today: Probabilistic approaches to parameter estimation. Useful to:

1. Understand what parameter values are probable in light of 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.

Notation

- $\boldsymbol{\theta} = [\theta_1 \ \theta_2 \ ...]$: parameters.
- $\mathbf{y} = [y_1 \ ... \ y_K]$: all K observations.

For probability distributions, we'll omit what's clear from context, e.g., $P(y_k)$ rather than $P(Y_k = y_k | M = m)$.

- $P(\mathbf{y}|\boldsymbol{\theta})$: *Probability mass function* for $\mathbf{y}$ conditional on $\boldsymbol{\theta}$.
- $f(\mathbf{y}|\boldsymbol{\theta})$: *Probability density function* for $\mathbf{y}$ conditional on $\boldsymbol{\theta}$. Sometimes $p(\mathbf{y}|\boldsymbol{\theta})$ (notice the lowercase)
- $L(\boldsymbol{\theta}|\mathbf{y})$: Likelihood function, treating either of the above as a unary function of $\boldsymbol{\theta}$. **Not** $P(\boldsymbol{\theta}|\mathbf{y})$ or $P^{-1}(\mathbf{y}|\boldsymbol{\theta})$.

(Also know: *cumulative density function* and *cumulative mass function*).

Probabilistic models

A cognitive model is probabilistic if it generates a probability distribution over $\mathbf{y}$ conditional on its parameters $\boldsymbol{\theta}$.

- 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 $P(\boldsymbol{\theta}|\mathbf{y})$ if we have a *prior* $P(\boldsymbol{\theta})$.

Likelihood

Again, $P(y_k|\boldsymbol{\theta})$ is the *probability mass function* for an observation y_k given $\boldsymbol{\theta}$ (m, the model, is constant here, and thus omitted for simplicity).

If our observations are conditionally independent, their joint conditional probability is

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

Likelihood

If we're happy with our parameters and are making predictions, we might treat $\boldsymbol{\theta}$ as fixed and $P(\mathbf{y}|\boldsymbol{\theta})$ as a function of $\mathbf{y}$.

If we're trying to fit or assess our model given data, we treat $\mathbf{y}$ as fixed, and treat $\boldsymbol{\theta}$ as the varying argument to a *likelihood function*.

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

Negative log likelihood

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

- Products of probabilities become sums: $\log(\prod_k x_k) = \sum_k \log(x_k)$
- More manageable and comparable numbers; avoids underflow. Minimum at zero if using mass functions.

Example: Independent Gaussians

Suppose a model predicts judgment k will have a mean of $\hat{y}_k$ and a variance of σ^2. Judgment k has a probability density of

$$\mathcal{N}(y_k; \hat{y}_k, \sigma^2) = \frac{1}{Z} e^{-\frac{(y_k - \hat{y}_k)^2}{2\sigma^2}}$$

The log likelihood is $-\log Z - \frac{(y_k - \hat{y}_k)^2}{2\sigma^2}$ where Z (i.e., $\sqrt{2\pi\sigma^2}$) doesn't depend on y or $\hat{y}$.

If σ^2 is fixed and the $\hat{y}_k$ values are the parameters:

$$-\log(L(\boldsymbol{\theta}|\mathbf{y})) = K \log Z + \frac{1}{2\sigma^2} \sum_{k=1}^{K} (y_k - \hat{y}_k)^2$$

This is just a constant plus a scaled sum squared error. Minimizing it is equivalent to minimizing RMSD.

Example: Independent Gaussians

Note that if we allow σ^2 to vary, this loss function will reward models that are well-calibrated with respect to uncertainty (i.e., making larger errors when variance is higher).

Example: Reaction Times

The Wald probability function captures latencies (reaction times) from a choice experiment.

It describes the time it takes a continuous random walk to drift past a threshold.

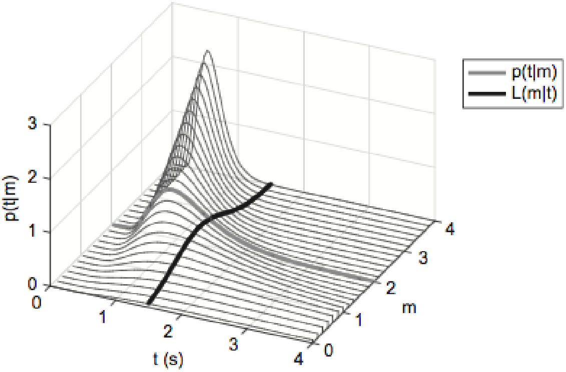
The Wald function has the following parameters:

- m: drift
- a: boundary position
- T: added non-decision time

Let's only consider m for now.

Example: Reaction Times

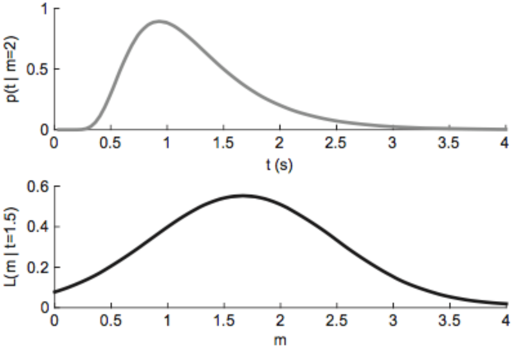
For a single data point t and the parameter m, we get the following probability density function $f(t|m)$:



The gray line marks $f(t|m = 2)$, the black one $L(m|t = 1.5)$.

Example: Reaction Times

If we just plot $f(t|m = 2)$ and $L(m|t = 1.5)$, we get:



We can optimize $L(m|t)$ using the optimizer of our choice (e.g., Nelder-Mead).

Maximum Likelihood Estimation

We've been talking about finding parameter values that maximize the likelihood of the data:

$$\boldsymbol{\theta}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y})$$

These *maximum-likelihood estimates* (MLEs) are frequently used in cognitive models.

These are (usually) different from *maximum a posteriori estimates* (MAPs):

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathbf{y})$$

MAP estimates and other alternatives

- If we have any a priori information about parameters, MAP estimates can be preferable
- If we don't, MAP estimate is equal to MLE.

Posterior probability:

$$P(\boldsymbol{\theta}|\mathbf{y}) = \frac{P(\mathbf{y}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}'} P(\mathbf{y}|\boldsymbol{\theta}')P(\boldsymbol{\theta}')} \propto P(\mathbf{y}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

$$\arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathbf{y}) = \arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

If $P(\boldsymbol{\theta}) \propto 1$, then

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

MAP estimates and other alternatives

If we really want to know about $\boldsymbol{\theta}$, a point estimate is often not enough

- Doesn't tell us how likely it is that a parameter is greater than zero (or another parameter)
- Often not the most useful point estimate.

Consider coin flips:

- Flip a coin twice; get two heads.
- What's the MLE for the coin's bias (i.e., $P(H = 1)$ for the next flip)?
- Same issue tends to apply to MAP estimates.

One alternative: Expected value.

Defining a Likelihood Function

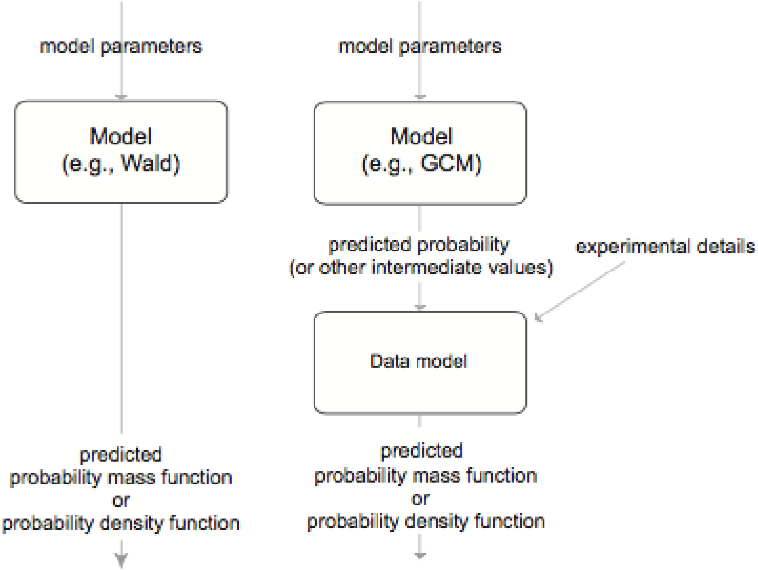
In the last two examples, we were given probability density functions.

We often don't have that luxury; sometimes a model will:

- Give a best or mostly-likely option without associated probabilities.
- Specify a process that generates judgments.
- Give probabilities or utilities that an agent might assign to options.

In these cases, we need a *data model* to assign probabilities to data.

The Data Model



The Data Model

Even if our model produces probabilities, we may need to do some work. For example:

- The GCM gives response probabilities for different categories, but our data may be numbers of people choosing each category. Here, a multinomial distribution may be appropriate.

The Data Model

Even if our model produces probabilities, we may need to do some work. For example:

- A model might give probabilities that an agent should assign, subjectively, to events. Will discrete judgment probabilities match those probabilities?
  - *Probability matching* appears to be common, but why should this happen?
  - *Maximizing* is arguably more rational in many contexts
  - *Soft maximization* or "Softmax" includes matching, max, random as special cases: $P(r) \propto P(b)^{\gamma}$

The Data Model

Even if our model produces probabilities for judgments, we may need to do some work. For example:

- A model might assign probabilities of zero to some outcomes – should we plan to bin the model if one such outcome occurs?
  - E.g., adding a category "C" to options under the GCM, but no category-C exemplars.

Summary

- Maximum-likelihood estimates: $\arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y})$
  - Often preferable to least-squares or other alternatives
  - Probabilistic but only a “halfway house” to fully Bayesian methods
- Alternerative: MAP estimate
- Sometimes we need (or want) a data model to predict a probability distribution from the output of our model, even if the model gives probabilities.