

Computational Cognitive Science

Lecture 5: Parameters and Probabilities 2

Frank Keller

5 October 2026

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

Slide credits: Chris Lucas, Benjamin Peters

Reading: Chapter 6 of [Farrell and Lewandowsky \(2018\)](#).

Last time, we discussed MLE and MAP estimates:

- MLE: Choose the θ that makes $\mathbf{y}$ most probable, ignoring $p(\theta)$.
- MAP: Choose the θ that is most probable given $\mathbf{y}$.

MAP with non-uniform priors can improve estimates and reduce overfitting.

In general, parameters are continuous, so the MAP maximizes the *density* — the probability that the parameters take those values is still infinitesimal.

- Estimating parameters in a simple discrete-choice experiment
- Compare MLE, MAP, and Bayesian methods
- Brief introduction to conjugate priors

Left-right bias

In 1977, Nisbett and Wilson reported a study where people has been asked to choose between four identical pairs of stockings: A, B, C, D from left to right.¹



This is similar to the coin example from C6 of F&L, but it involves more than two outcomes and is about human decisions.

¹Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231.

Left-right bias

Suppose we observe 8 choices.

Choice	A	B	C	D
Number	0	2	2	4

We want to know how biased people are in general and predict the judgments of the remaining 44 participants.

We can capture this with a multinomial distribution:

$$P(\mathbf{y}|\boldsymbol{\theta}) = \frac{(\sum_i y_i)!}{\prod_i y_i!} \prod_i \theta_i^{y_i}$$

where y_i is the count of choices in the i^{th} category and $\sum_i \theta_i = 1$.

What are our options for estimating $\boldsymbol{\theta}$?

1. MLE
2. MAP
3. Bayesian approaches

1. MLE: $\operatorname{argmax}_{\theta} L(\theta|\mathbf{y})$

The multinomial's parameters are choice probabilities, and one can show that the MLE parameters are just the proportions:

$$\theta_i = \frac{y_i}{\sum_k y_k}$$

Choice	A	B	C	D
Obs. (n=8)	0	2	2	4
θ_{MLE}	0	.25	.25	.5

This maximizes the probability of the data in retrospect, but it's not ideal for predictions.

For example, someone will probably, eventually, choose option A.

2. MAP estimate: $\operatorname{argmax}_{\theta} L(\theta|\mathbf{y})p(\theta)$

If we know or believe something about choices in this setting, we should probably use it.

- We might expect that any bias won't be extreme; a few people will probably choose every option
- Like a coin where we expect to be close to fair

How do we express this?

In choosing priors, we ideally want a distribution that:

- has support for all remotely possible values, i.e., assigns non-zero probability to them
- is easy to interpret and communicate
- allows efficient computation of a posterior distribution

Dirichlet distribution

There are many options, but here we might use a **Dirichlet distribution** with **hyperparameters** α :

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_i \theta_i^{\alpha_i-1}$$

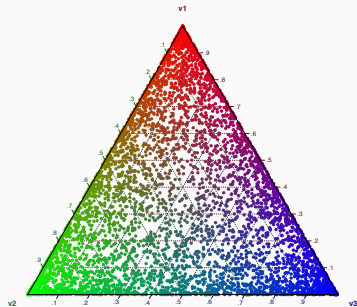
- Can capture intuitions about differences in proportions and in concentration
 - α : “Concentration parameters”
 - $\alpha_i > 0$, one per θ
 - “virtual observations”
 - Can translate beliefs about $P(c_0 < \theta_i < c_1)$ into hyperparameters
- Familiar to many cognitive scientists

Dirichlet distribution

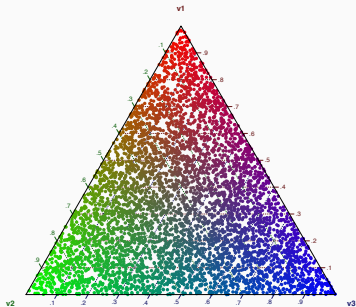
$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_i \theta_i^{\alpha_i-1}$$

The beta distribution (F&L C6) is 2-group Dirichlet distribution.

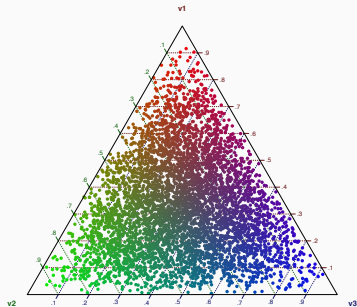
$$\alpha = [.5 \ .5 \ .5]$$



$$\alpha = [1 \ 1 \ 1]$$

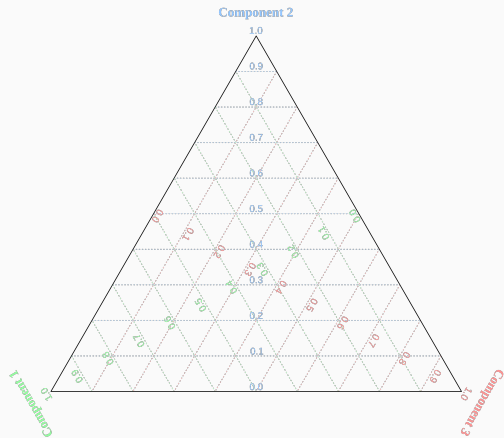
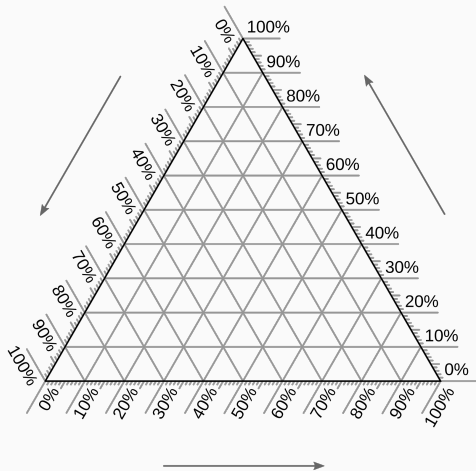


$$\alpha = [2 \ 2 \ 2]$$



Also, it is a **conjugate prior** for the multinomial distribution.

Ternary plots



From [wikipedia page](#) on ternary plots.

Conjugate priors

When the posterior probability function and the prior have the same form, they're *conjugate*.

If we can find a reasonable conjugate prior for our likelihood function, life is easier:

- Simplifies computation
- Makes interpretation of the posterior easier

Conjugate priors

Some commonly-used likelihood/conjugate prior pairs:

Likelihood	Conjugate prior
Bernoulli	beta
binomial	beta
categorical	Dirichlet
multinomial	Dirichlet
normal	normal

Dirichlet-multinomial

Our prior:

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_i \theta_i^{\alpha_i-1} \propto \prod_i \theta_i^{\alpha_i-1}$$

Our likelihood:

$$P(\mathbf{y}|\boldsymbol{\theta}) = \frac{(\sum_i y_i)!}{\prod_i y_i!} \prod_i \theta_i^{y_i} \propto \prod_i \theta_i^{y_i}$$

Our posterior:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto \prod_i \theta_i^{\alpha_i-1} \prod_i \theta_i^{y_i} = \prod_i \theta_i^{\alpha_i-1+y_i}$$

This is an un-normalized Dirichlet distribution; divide by $B(\boldsymbol{\alpha} + \mathbf{y})$ and we have a Dirichlet. See text for more detail in a 2-choice setting.

Our prior is $\text{Dir}(\alpha_1, \dots, \alpha_K)$ and our posterior is $\text{Dir}(\alpha_1 + y_1, \dots, \alpha_K + y_K)$.

- We can think of α as pseudo-observations.
- If we believe a bias for each group is equally likely, $\alpha_i = \alpha$.
- As α increases, the Dirichlet distribution increasingly favors an equal distribution over choices.
- For $\alpha = 1$, all valid parameter combinations are equally likely.
- $\alpha = 1/K$ is a Jeffreys prior; see the text.

Left-right bias

If we think extreme biases are unlikely, we can use $\alpha = 2.0$.²

The mode of a Dirichlet distribution is $\theta_i = \frac{\alpha_i - 1}{\sum_k \alpha_k - K}$.

Choice	A	B	C	D
Obs. (n=8)	0	2	2	4
θ_{MLE}	0	.25	.25	.5
θ_{MAP}	.08	.25	.25	.42

²Under this choice, each parameter's marginal probability of being between .1 and .4 is about 70 percent.

Left-right bias

If we think extreme biases are unlikely, we can use $\alpha = 2.0$. The mode of a Dirichlet distribution is $\theta_i = \frac{\alpha_i - 1}{\sum_k \alpha_k - K}$.

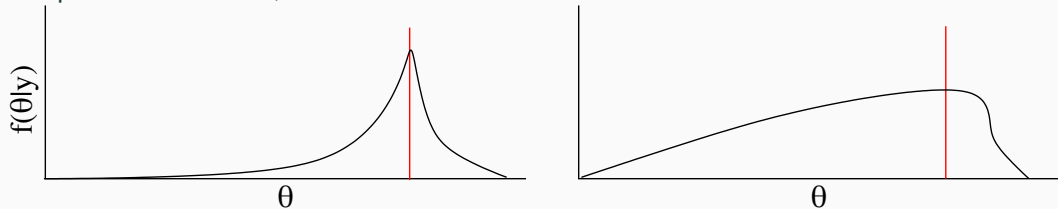
Choice	A	B	C	D
Obs. (n=8)	0	2	2	4
θ_{MLE}	0	.25	.25	.5
θ_{MAP}	.08	.25	.25	.42
Data (n=52)	.12	.17	.31	.40

3. Bayesian approaches

The MAP estimate doesn't account for uncertainty or the shape of $p(\theta|\mathbf{y})$.

- If our prior is uniform, MAP = MLE; we're back to estimating zero-probabilities.

Compare two densities, where θ is the bias of a coin toward heads:



- The mode of the posterior is the same for both.
- Which are we more confident of?
- Which coin do we think is more likely to come up heads?

3. Bayesian approaches

If we have a posterior distribution, we can ask:

- What is the expected value of θ_i ?³

$$E[\theta_i|\mathbf{y}] = \int_{\theta_i} \theta_i f(\theta_i|\mathbf{y}) d\theta_i$$

- What is the probability that the next choice/toss will be in category i ($x_{N+1} = i$)?

$$P(x_{N+1} = i|\mathbf{y}) = \int_{\theta_i} P(x_{N+1} = i|\theta_i) f(\theta_i|\mathbf{y}) d\theta_i$$

~~Because θ_i is $P(x_{N+1} = i|\theta_i)$, these are the same (here).~~

³Notice that we're writing down $f(\theta_i|\mathbf{y})$ directly – we get this distribution for free; another nice property of the Dirichlet distribution.

3. Bayesian approaches

For a Dirichlet-multinomial,

$$P(x_{N+1} = i | \mathbf{y}, \boldsymbol{\alpha}) = \frac{\alpha_i + y_i}{\sum_j (\alpha_j + y_j)}$$

Choice	A	B	C	D
Obs. (n=8)	0	2	2	4
θ_{MLE}	0	.25	.25	.5
θ_{MAP}	.08	.25	.25	.42
$P(x_{N+1} = i)$	.12	.25	.25	.38
Data (n=52)	.12	.17	.31	.40

A slightly modified example for visualization

A *modified* example with six observations in **three** categories.

Choice	A	B	C
Obs. (n=6)	0	2	4
θ_{MLE}	0	.33	.67
θ_{MAP}	.11	.33	.56
$P(x_{N+1} = i)$	.17	.33	.50

A slightly modified example for visualization (ctd)

[missing figure here.]

The colour scheme visualizes the probability density of the prior (left, Dirichlet with $\alpha_i = \alpha = 2$) and posterior (right, $\text{Dir}(2, 4, 6)$).

The likelihood (center) is **not** a probability density (it does not integrate to 1 over all possible values for θ). The likelihood follows a multinomial distribution, which is proportional to $\text{Dir}(1, 3, 5)$ and visualized here.

3. Bayesian approaches

We can also answer other questions, e.g.,

- How likely is it that people are choosing option D more than 25 percent of the time?
 - $P(\theta_4 > .25|\mathbf{y})$
- How likely is it that people are choosing uniformly (null hyp)?
 - For all i , $P(\theta_i = .25 \pm \epsilon|\mathbf{y})$
- What is the standard deviation of θ_i ?
- What is the probability that $\theta_1 + \theta_2 > \theta_3 + \theta_4$?

3. Bayesian approaches

To summarize, some advantages of Bayesian approaches over MLE:

- Sensible priors and averaging both help us avoid overfitting
 - This allows more complex models, including cases where MLEs aren't unique
- Can answer diverse questions, e.g., support *for* null hypotheses
- Naturally lead to hierarchical models
 - Individual differences – next time
- We used a classic conjugate prior
 - Not always so easy; see C7 of F&L

3. Bayesian approaches

Why doesn't everyone use Bayesian methods?

1. Computational complexity

- Conjugate priors aren't always appropriate
- Inference can be computationally expensive

3. Bayesian approaches

Why doesn't everyone use Bayesian methods?

2. Convention, momentum, philosophical differences

- More psychologists use/understand* frequentist methods
- Out-of-the-box hypothesis tests are less work
- Suspicion about priors

3. Bayesian approaches

Why doesn't everyone use Bayesian methods?

3. Technical barriers

- Bayesian methods expose more mathematical detail
- Until recently, few good tools for running non-trivial Bayesian analyses

But:

- Faster computers
- Friendlier/better tools, e.g. Stan
- Wider adoption and better dissemination
 - Bayesian analyses much more common than ~ 10 years ago
 - Materials for wider audiences, e.g., Kruschke's "puppy book"

Summary

- MLE and MAP generate point estimates of the parameters
- Sensible priors can mitigate overfitting until MAP estimation
- Better yet: Bayesian methods – priors and integrating over parameters
 - less prone to overfitting and allows better use of informative priors
 - allows more questions to be answered more directly
- Conjugate prior distributions, where the prior and the posterior have the same form given a particular likelihood function
 - Easily-interpretable posteriors
 - Closed-form expressions for many quantities of interest

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