

Computational Cognitive Science

Lecture 6: Bayesian Parameter Estimation 2

Frank Keller

8 October 2026

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

Slide credits: Chris Lucas, Benjamin Peters

Estimating the Posterior

From Coins to Choices

Conjugacy

Dirichlet Distribution

Choosing Priors and Computing Posteriors

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

Estimating the Posterior

MAP is a posterior summary

Last time, we discussed MLE and MAP estimates:

- MLE: Choose the parameters θ that make the data $\mathbf{y}$ most probable, ignoring the prior distribution $p(\theta)$.
- MAP: Choose the θ that is most probable given $\mathbf{y}$ and $p(\theta)$.

MAP can improve accuracy and reduce overfitting.

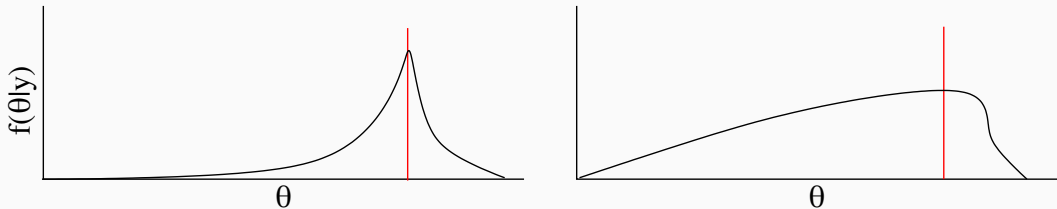
But MAP is still only a **point estimate**: it discards the rest of the posterior distribution.

Point estimate vs posterior distribution

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

If our prior is uniform, MAP = MLE.

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



- The mode of the posterior is the same for both.
- But the uncertainty and posterior mean can differ.

Plan for today's lecture

So far, we have only looked at the **Bernoulli distribution and the Beta prior**.

Today, we will look at more complicated distributions, in particular the **multinomial distribution and the Dirichlet prior**.

So far, we have assumed there's an **analytic solution** for the posterior distribution.

This is often not the case; we will briefly discuss this, but for more detail see Chapters 7 and 8 of **Farrell and Lewandowsky (2018)**.

From Coins to Choices

Left-right bias

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



This is similar to the coin example, but it involves more than two outcomes and is about human decisions.

Multinomial likelihood

Suppose we observe 8 choices.

Choice	A	B	C	D
Number	0	2	2	4

We model the counts with a multinomial distribution:

$$P(\mathbf{y}|\boldsymbol{\theta}) = \frac{(\sum_i y_i)!}{\prod_i y_i!} \prod_i \theta_i^{y_i} \quad \sum_i \theta_i = 1$$

What can we infer about the choice probabilities $\boldsymbol{\theta}$?

MLE for the choice model

The multinomial's parameters are choice probabilities. The MLEs are just the observed 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	.50

This maximizes the probability of the data in retrospect, but it is not ideal for prediction: someone will probably choose option A eventually.

Prior for more than two outcomes

For two outcomes, the Beta distribution is a prior over a single probability θ .

For several outcomes, we need a prior over a vector:

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_K) \quad \sum_i \theta_i = 1$$

The natural extension is the **Dirichlet distribution**:

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

Where $B(\boldsymbol{\alpha})$ is again a normalizing constant.

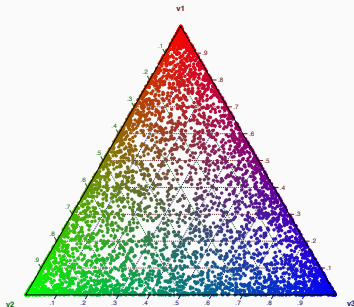
Dirichlet distribution

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

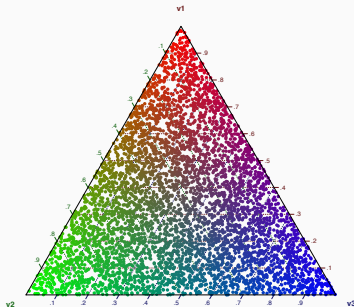
- The Beta distribution is a two-outcome Dirichlet distribution.
- The α_i are concentration parameters.
- They can be interpreted as pseudo-observations.
- Larger values of α_i encode stronger prior information.

Dirichlet distribution: examples

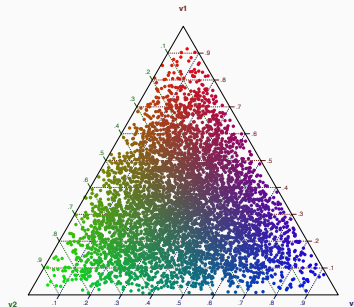
$$\alpha = [.5 \ .5 \ .5]$$



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



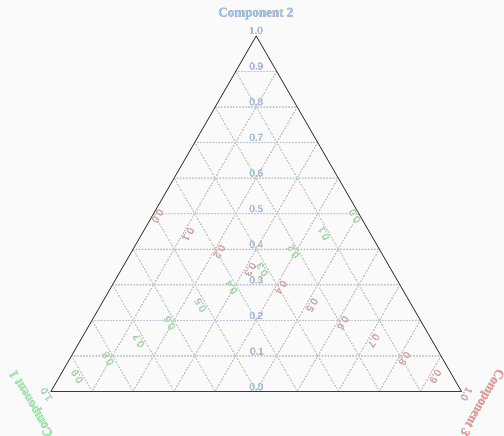
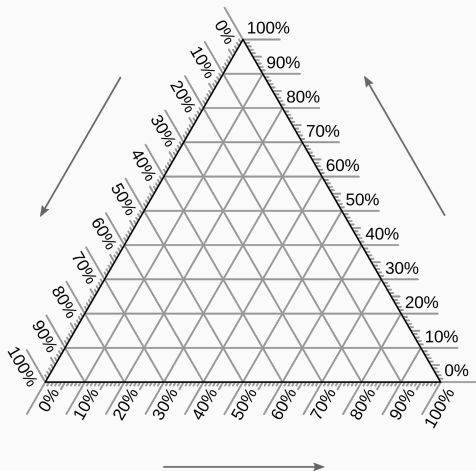
$$\alpha = [2 \ 2 \ 2]$$



Different concentration parameters express different assumptions about how evenly probability mass is distributed over categories.

Ternary plots

For three outcomes, possible values of θ can be visualized in a triangle.



Conjugacy

Recall from the last lecture: For Bernoulli data, we worked out the posterior and found:

$$P(\theta|h, n) = \text{Beta}(\theta|\alpha + h, \beta + n - h)$$

When the prior and posterior belong to the same distribution family, the prior is **conjugate** to the likelihood.

Beta prior + Bernoulli data $\Rightarrow$ Beta posterior

This can be generalized to other distributions.

Conjugate prior pairs

Here are some more common likelihood/conjugate-prior pairs:

Likelihood	Conjugate prior
Bernoulli	Beta
Binomial	Beta
Categorical	Dirichlet
Multinomial	Dirichlet
Normal	Normal

Time for a short quiz on WooClap!

Go to wooclapp.com

Enter the event code in the top banner

Event code

TGUPBYK



Dirichlet Distribution

Dirichlet–multinomial posterior

Now back to the Dirichlet prior and the multinomial likelihood.

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:

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

Dirichlet–multinomial update

The posterior is again Dirichlet:

$$\text{Dir}(\alpha_1, \dots, \alpha_K) \Rightarrow \text{Dir}(\alpha_1 + y_1, \dots, \alpha_K + y_K)$$

- This is the same logic as the Beta–Bernoulli update.
- The prior contributes pseudo-observations.
- The data add real observations.

MAP for the left-right example

If we think extreme biases are unlikely, we can use a symmetric Dirichlet prior with $\alpha_i = 2.0$.¹

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	.50
θ_{MAP}	.08	.25	.25	.42

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

Posterior predictive distribution

For a Dirichlet–multinomial model:

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

Choice	A	B	C	D
Obs. (n=8)	0	2	2	4
θ_{MLE}	0	.25	.25	.50
θ_{MAP}	.08	.25	.25	.42
$P(x_{N+1} = i)$	.12	.25	.25	.38
Data (n=52)	.12	.17	.31	.40

What else can we ask?

Once we have the posterior distribution, we can answer questions directly:

- How likely is it that people choose option D more than 25 percent of the time?

$$P(\theta_4 > .25|\mathbf{y})$$

- How likely is it that choices are close to uniform?

$$P(\theta_i = .25 \pm \epsilon \text{ for all } i|\mathbf{y})$$

- What is the probability that $\theta_1 + \theta_2 > \theta_3 + \theta_4$?

Choosing Priors and Computing Posteriors

Non-informative priors are not trivial

The textbook asks: what should the prior be if we know nothing?

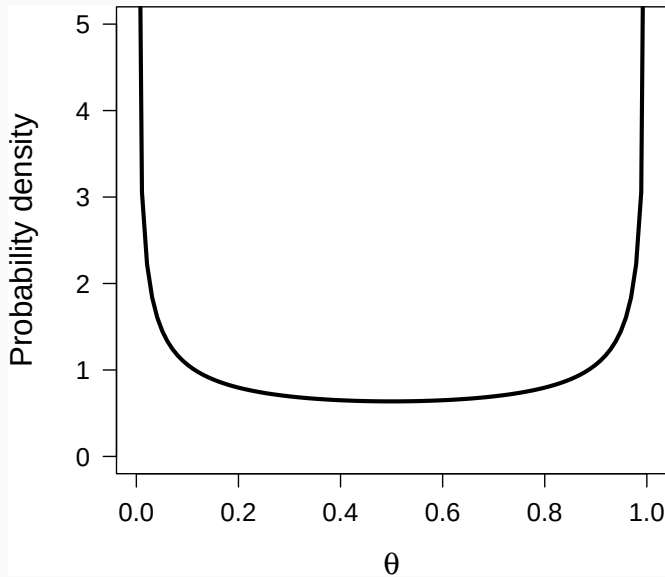
- A uniform prior over θ seems natural: $\text{Beta}(1, 1)$.
- But “uniform” depends on how we parameterize the problem.
- This is the **transformation problem**.

For Bernoulli data, Jeffreys' prior is:

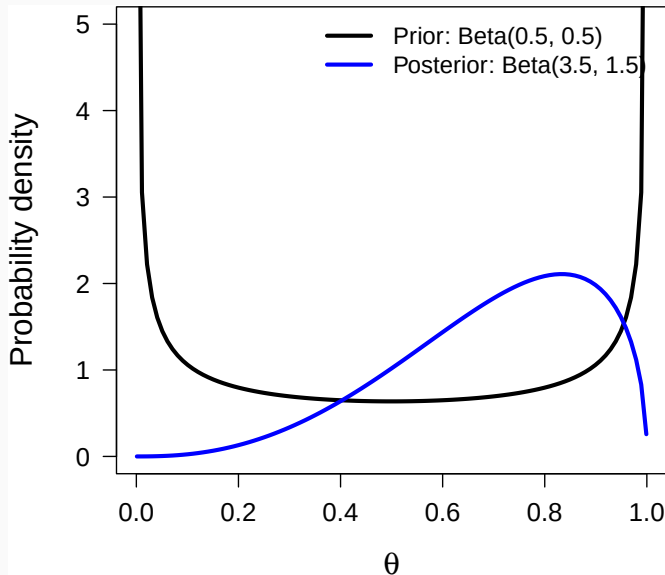
$$\text{Beta}(.5, .5)$$

For a symmetric Dirichlet prior with K categories, the analogous choice is often $\alpha_i = 1/K$.

Jeffreys' prior



Jeffreys' prior



Priors should be chosen deliberately and reported clearly.

- If we have domain knowledge, the prior can encode it.
- If we want the data to dominate, use weakly informative priors.
- If results depend strongly on the prior, that is important information.

A Bayesian analysis makes assumptions explicit: both the likelihood and the prior matter.

When analytic methods are not enough

The biased-coin example is analytically tractable:

$$\text{Beta prior} + \text{Bernoulli data} \Rightarrow \text{Beta posterior}$$

Many cognitive models are not so convenient.

- The posterior may not have a closed form.
- The evidence $P(\mathbf{y})$ may be difficult or impossible to compute analytically.
- We need numerical methods, especially Markov Chain Monte Carlo (MCMC) methods.

Chapter 7 of [Farrell and Lewandowsky \(2018\)](#) discusses MCMC methods.

Why use Bayesian methods?

Some advantages over pure MLE:

- Priors and posterior averaging help avoid overfitting.
- We obtain uncertainty over parameters, not just point estimates.
- We can answer probabilistic questions directly.
- Bayesian models naturally support the modeling of hierarchical structure and individual differences.

Why not always use Bayesian methods?

Practical barriers:

- Conjugate priors are not always appropriate.
- Inference can be computationally expensive.
- Bayesian methods expose more mathematical detail.
- Priors can be philosophically or practically controversial.

But faster computers and better tools have made Bayesian modeling more common.

Examples include [Stan and JAGS](#), languages for specifying Bayesian models. They use MCMC algorithms to obtain samples from posterior distributions (see Chapter 8 of [Farrell and Lewandowsky 2018](#)).

Bayesian models are everywhere

The Bayesian approach to modeling cognition is very influential and has been applied to many different domains.

This is reflected in the Bayesian models we discuss in this course:

- concept learning (lecture 9);
- causal inference (lecture 12);
- active learning (lectures 13 and 14);
- language acquisition (lecture 15).

- Bayesian estimation turns likelihood plus prior into a posterior distribution.
- For multinomial data, the Dirichlet prior gives an analytic Dirichlet posterior.
- In general, it's desirable if likelihood and prior are conjugate.
- Prior choice matters and should be explicit.
- Often the posterior distribution cannot be compute analytically and needs to be sampled from
- Methods like MCMC are often used, and languages such as Stan and JAGS.

- Farrell, Simon and Stephan Lewandowsky. 2018. *Computational Modeling of Cognition and Behavior*. Cambridge University Press, Cambridge.
- Nisbett, Richard E. and Timothy D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84(3):231–259.