Introduction to Bayesian Methods

Machine Learning in Cognition Society

Julie J. Lee

Cortex Lab, UCL

15 May 2019

Outline

Outline

1. Essence of Bayesian methods

2. Model fitting

Outline

1. Essence of Bayesian methods
2. Model fitting
3. Model comparison

Outline

Outline

1. Essence of Bayesian methods
2. Model fitting
3. Model comparison
4. Extensions
5. Further reading

To begin

To begin

Would I be successful at Bayesian modeling?

To begin

Would I be successful at Bayesian modeling?

Anybody can learn Bayesian modeling! The math might seem hard at first but after 10 to 50 hours of practice, depending on your background, it is more of the same.

With thanks to Wei Ji Ma

Structure of the talk

Structure of the talk

- Assume basically no prior knowledge

Structure of the talk

- Assume basically no prior knowledge
- Throw a lot of concepts just so they will be familiar next time

Structure of the talk

- Assume basically no prior knowledge
- Throw a lot of concepts just so they will be familiar next time
- Loose with terminology, will be imprecise

Structure of the talk

- Assume basically no prior knowledge
- Throw a lot of concepts just so they will be familiar next time
- Loose with terminology, will be imprecise
- May re-describe things you know, but tailored to this context

Clearing the confusion

Clearing the confusion

What does it mean to "be Bayesian"?

Clearing the confusion

What does it mean to "be Bayesian"?

Bayesian models as a strong *claim* about behavior/the brain

Clearing the confusion

What does it mean to "be Bayesian"?

Bayesian models as a strong *claim* about behavior/the brain

brain: *e.g. "Bayesian Coding Hypothesis", probabilistic population codes*

behavior: *e.g. Bayesian observer models, "Bayesian Brain Hypothesis", human behavior/perception as "Bayes optimal" or performing "Bayesian inference"*

Clearing the confusion

What does it mean to "be Bayesian"?

Bayesian models as a strong *claim* about behavior/the brain

brain: *e.g. "Bayesian Coding Hypothesis", probabilistic population codes*

behavior: *e.g. Bayesian observer models, "Bayesian Brain Hypothesis", human behavior/perception as "Bayes optimal" or performing "Bayesian inference"*

Not going to discuss today!

Clearing the confusion

What does it mean to "be Bayesian"?

Bayesian models as a strong *claim* about behavior/the brain

brain: *e.g. "Bayesian Coding Hypothesis", probabilistic population codes*

behavior: *e.g. Bayesian observer models, "Bayesian Brain Hypothesis", human behavior/perception as "Bayes optimal" or performing "Bayesian inference"*

Not going to discuss today!

Bayesian or Bayes-ish statistics as a method of *analysis*

Clearing the confusion

What does it mean to "be Bayesian"?

Bayesian models as a strong *claim* about behavior/the brain

brain: *e.g. "Bayesian Coding Hypothesis", probabilistic population codes*

behavior: *e.g. Bayesian observer models, "Bayesian Brain Hypothesis", human behavior/perception as "Bayes optimal" or performing "Bayesian inference"*

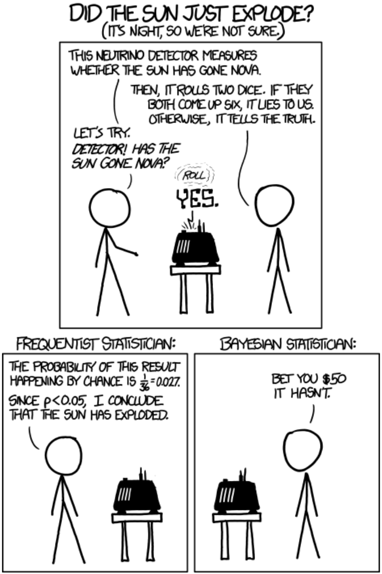
Not going to discuss today!

Bayesian or Bayes-ish statistics as a method of *analysis*

e.g. Bayesian model comparison, Bayesian decoding of neural activity (sensory stimulus, spatial position), Bayesian hierarchical modelling

What does it mean to be Bayesian?

What does it mean to be Bayesian?



What does it mean to be Bayesian?

(a) Representation of beliefs as *probability distributions*, reasoning with *distributions* vs. point estimates

What does it mean to be Bayesian?

(a) Representation of beliefs as *probability distributions*, reasoning with *distributions* vs. point estimates (this is ultimately the distinction viz. frequentist statistics)

What does it mean to be Bayesian?

(a) Representation of beliefs as *probability distributions*, reasoning with *distributions* vs. point estimates (this is ultimately the distinction viz. frequentist statistics)

(b) Quantification of uncertainty (width of distributions)

What does it mean to be Bayesian?

(a) Representation of beliefs as *probability distributions*, reasoning with *distributions* vs. point estimates (this is ultimately the distinction viz. frequentist statistics)

(b) Quantification of uncertainty (width of distributions)

(c) "Updating" beliefs in light of new evidence

What does it mean to be Bayesian?

(a) Representation of beliefs as *probability distributions*, reasoning with *distributions* vs. point estimates (this is ultimately the distinction viz. frequentist statistics)

(b) Quantification of uncertainty (width of distributions)

(c) “Updating” beliefs in light of new evidence

(d) “Baking in” prior knowledge/experience

What does it mean to be Bayesian?

(a) Representation of beliefs as *probability distributions*, reasoning with *distributions* vs. point estimates (this is ultimately the distinction viz. frequentist statistics)
(b) Quantification of uncertainty (width of distributions)
(c) "Updating" beliefs in light of new evidence
(d) "Baking in" prior knowledge/experience

Although use of priors tends to be the typical association/complaint with Bayesian statistics, this is often **not** so much what it is about!

Probability primer

Probability primer

A random variable X takes on a specific value x_i according to a probability distribution. This is denoted $P(X = x_i)$.

Probability primer

A random variable X takes on a specific value x_i according to a probability distribution. This is denoted $P(X = x_i)$.

Probability distributions always sum to 1.

Probability primer

Probability distributions can be discrete.

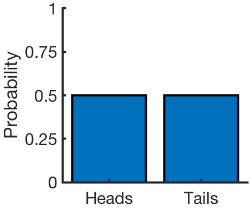
Probability primer

Probability distributions can be discrete.
e.g. the outcome of a coin flip can be *heads* or *tails*.

Probability primer

Probability distributions can be discrete.
e.g. the outcome of a coin flip can be *heads* or *tails*.
For a fair coin,

$$P(\mathit{flip} = \mathit{heads}) = P(\mathit{flip} = \mathit{tails}) = 0.5$$



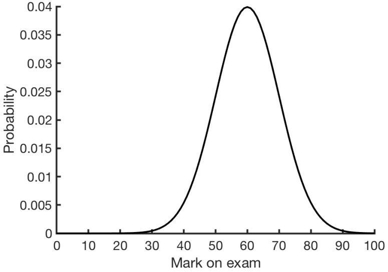
Probability primer

Probability distributions can be continuous.

Probability primer

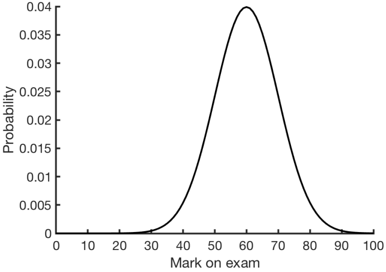
Probability distributions can be continuous.

e.g. a student's mark on an exam can be any number between 0 and 100.



Probability primer

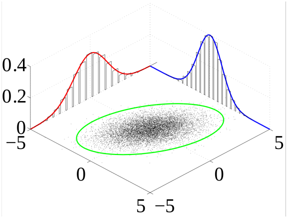
Probability distributions can be continuous.
e.g. a student's mark on an exam can be any number between 0 and 100.



Chance of getting a First, $P(70 \leq \mathit{mark} \leq 100) \approx 0.16$

Probability primer

The *joint probability* of two outcomes is denoted $P(X = x_i, Y = y_i)$.



e.g. in a deck of cards, what is the probability that $P(\mathit{suit} = \mathit{hearts}, \mathit{number} = \mathit{even})$

Probability primer

Marginalization
Sum/integrate over *all possible values* of the other variable.

Probability primer

Marginalization
Sum/integrate over *all possible values* of the other variable.

$$P(X) = \int P(X, Y) dy \text{ (continuous)} \quad P(X) = \sum_{Y} P(X, Y) \text{ (discrete)}$$

Probability primer

Marginalization

Sum/integrate over *all possible values* of the other variable.

$$P(X) = \int P(X, Y)dy \text{ (continuous)} \quad P(X) = \sum_{Y} P(X, Y) \text{ (discrete)}$$

Conditionalization

Probability of Y "given" that X has occurred/has been observed.

Probability primer

Marginalization

Sum/integrate over *all possible values* of the other variable.

$$P(X) = \int P(X, Y)dy \text{ (continuous)} \quad P(X) = \sum_{Y} P(X, Y) \text{ (discrete)}$$

Conditionalization

Probability of Y "given" that X has occurred/has been observed.

$$P(Y|X) = \frac{P(X, Y)}{P(X)}$$

Joint $P(X, Y)$ divided by marginal $P(X)$

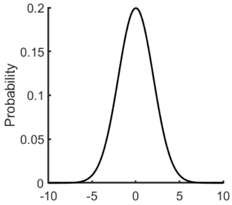
Probability primer

The Gaussian/normal distribution:

Probability primer

The Gaussian/normal distribution:

$$P(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

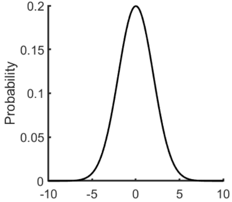


Mean (μ) and variance (σ^2) in this case is $\mu = 0$, $\sigma^2 = 4$

Probability primer

The Gaussian/normal distribution:

$$P(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$



Mean (μ) and variance (σ^2) in this case is $\mu = 0$, $\sigma^2 = 4$

FYI: Product of Gaussians is another Gaussian (add the exponents, rearrange)

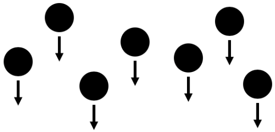
Likelihoods as *hypotheses*

With thanks to Wei Ji Ma

Likelihoods as *hypotheses*

With thanks to Wei Ji Ma

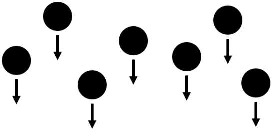
e.g. "Law of common fate"



Likelihoods as *hypotheses*

With thanks to Wei Ji Ma

e.g. "Law of common fate"

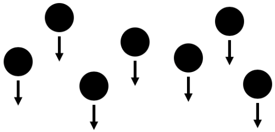


If you *observe* all seven dots moving downwards, and
$P(move = down) = P(move = up) = 0.5$,

Likelihoods as *hypotheses*

With thanks to Wei Ji Ma

e.g. "Law of common fate"



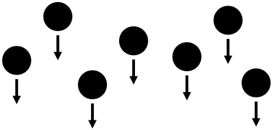
If you *observe* all seven dots moving downwards, and
$P(move = down) = P(move = up) = 0.5$,

What is the *likelihood* that (a) they have a common cause vs. (b) they are operating separately and *happen* to all move in the same direction?

Likelihoods as *hypotheses*

With thanks to Wei Ji Ma

e.g. "Law of common fate"



If you *observe* all seven dots moving downwards, and

$$P(move = down) = P(move = up) = 0.5,$$

What is the *likelihood* that (a) they have a common cause vs. (b) they are operating separately and *happen* to all move in the same direction?

Answer: (a) P(common cause) $= 0.5$ (b) P(separate) $= 0.5^7 \approx 0.008$

Bayes' rule:

Bayes' rule: Likelihood $\times$ prior $\propto$ Posterior

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Bayes' rule: Likelihood $\times$ prior $\propto$ Posterior

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

$P(Y)$ is just the "normalizer" and ensures the result sums to 1.

Bayes' rule: Likelihood $\times$ prior $\propto$ Posterior

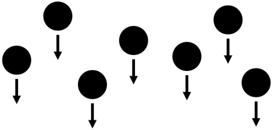
$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

$P(Y)$ is just the "normalizer" and ensures the result sums to 1.

$$P(X|Y) \propto P(Y|X)P(X)$$

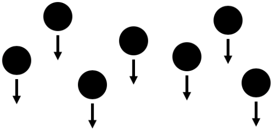
Returning to the law of common fate

With thanks to Wei Ji Ma



Returning to the law of common fate

With thanks to Wei Ji Ma

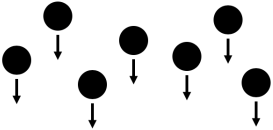


Say the *prior probability* of the common cause (a) is 1/4 (25%) and the dots being separate (b) is 3/4 (75%).

What is the *posterior probability* of (a) and (b)?

Returning to the law of common fate

With thanks to Wei Ji Ma



Say the *prior probability* of the common cause (a) is 1/4 (25%) and the dots being separate (b) is 3/4 (75%).

What is the *posterior probability* of (a) and (b)?

Answer: since the *likelihood* of (b) was very small, (a) still wins.

$$(a) 0.5 \times 0.25 = 0.125 \qquad (b) \approx 0.008 \times 0.75 = 0.006$$

Returning to the law of common fate

With thanks to Wei Ji Ma

As will often be the case, here, **most of the action is in the likelihood.**

This waves away objections about the prior (and therefore Bayesian methods) being "too subjective".

Bayes' rule: Likelihood $\times$ prior $\propto$ Posterior

Example for a Gaussian

Bayes' rule: Likelihood $\times$ prior $\propto$ Posterior

Example for a Gaussian

Prior $p(s)$ – how probable are various stimuli in the world

$$p(s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(s-\mu)^2}{2\sigma_s^2}}$$

Bayes' rule: Likelihood $\times$ prior $\propto$ Posterior

Example for a Gaussian

Prior $p(s)$ – how probable are various stimuli in the world

$$p(s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(s-\mu)^2}{2\sigma_s^2}}$$

Likelihood $p(x_{obs}|s)$ – probability of an *observation* under potential stimuli

$$L(s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s-x_{\text{obs}})^2}{2\sigma^2}}$$

Bayes' rule: Likelihood $\times$ prior $\propto$ Posterior

Example for a Gaussian

Prior $p(s)$ – how probable are various stimuli in the world

$$p(s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(s-\mu)^2}{2\sigma_s^2}}$$

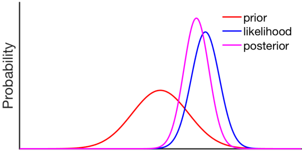
Likelihood $p(x_{obs}|s)$ – probability of an *observation* under potential stimuli

$$L(s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s-x_{\text{obs}})^2}{2\sigma^2}}$$

Recall: product of Gaussians (here, product is the posterior) is another Gaussian.

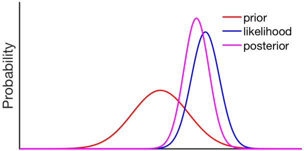
Bayes' rule: Likelihood $\times$ prior $\propto$ Posterior

Example for a Gaussian



Bayes' rule: Likelihood $\times$ prior $\propto$ Posterior

Example for a Gaussian

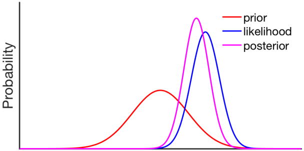


After rearranging, likelihood $\times$ prior is another Gaussian with:

$$\mu_{combined} = \frac{\frac{\mu}{\sigma_s^2} + \frac{x_{obs}}{\sigma^2}}{\frac{1}{\sigma_s^2} + \frac{1}{\sigma^2}} \qquad \sigma^2_{combined} = \frac{1}{\frac{1}{\sigma_s^2} + \frac{1}{\sigma^2}}$$

Bayes' rule: Likelihood $\times$ prior $\propto$ Posterior

Example for a Gaussian



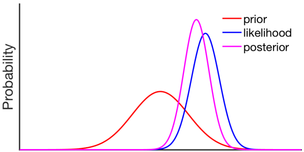
After rearranging, likelihood $\times$ prior is another Gaussian with:

$$\mu_{combined} = \frac{\frac{\mu}{\sigma_s^2} + \frac{x_{obs}}{\sigma^2}}{\frac{1}{\sigma_s^2} + \frac{1}{\sigma^2}} \qquad \sigma^2_{combined} = \frac{1}{\frac{1}{\sigma_s^2} + \frac{1}{\sigma^2}}$$

Posterior is a **weighted combination** of the prior and the likelihood.

Bayes' rule: Likelihood $\times$ prior $\propto$ Posterior

Example for a Gaussian



After rearranging, likelihood $\times$ prior is another Gaussian with:

$$\mu_{combined} = \frac{\frac{\mu}{\sigma_s^2} + \frac{x_{obs}}{\sigma^2}}{\frac{1}{\sigma_s^2} + \frac{1}{\sigma^2}} \qquad \sigma^2_{combined} = \frac{1}{\frac{1}{\sigma_s^2} + \frac{1}{\sigma^2}}$$

Posterior is a **weighted combination** of the prior and the likelihood.

Intuition: Prior "pulls posterior away" from the observations, depending on the relative narrowness (inverse variance, "certainty") of each.

Model fitting: The recipe

Model fitting: The recipe

Given some data (e.g. participant responses),

Model fitting: The recipe

Given some data (e.g. participant responses),

- Hypothesize a model that could have *produced* that data

Model fitting: The recipe

Given some data (e.g. participant responses),

- Hypothesize a model that could have *produced* that data (and with it, candidate parameter settings θ_M)

Model fitting: The recipe

Given some data (e.g. participant responses),

- Hypothesize a model that could have *produced* that data (and with it, candidate parameter settings θ_M)
- Compute “likelihood”, probability of data given model (and parameters) $P(D|M, \theta_M)$.

Model fitting: The recipe

Given some data (e.g. participant responses),

- Hypothesize a model that could have *produced* that data (and with it, candidate parameter settings θ_M)
- Compute "likelihood", probability of data given model (and parameters) $P(D|M, \theta_M)$.
- Hypothesize a prior distribution for the parameters $P(\theta_M|M)$

Model fitting: The recipe

Given some data (e.g. participant responses),

- Hypothesize a model that could have *produced* that data (and with it, candidate parameter settings θ_M)
- Compute "likelihood", probability of data given model (and parameters) $P(D|M, \theta_M)$.
- Hypothesize a prior distribution for the parameters $P(\theta_M|M)$ (it may seem tricky to "come up with" a prior, but the choice is not actually super important)

Model fitting: The recipe

Given some data (e.g. participant responses),

- Hypothesize a model that could have *produced* that data (and with it, candidate parameter settings θ_M)
- Compute "likelihood", probability of data given model (and parameters) $P(D|M, \theta_M)$.
- Hypothesize a prior distribution for the parameters $P(\theta_M|M)$ (it may seem tricky to "come up with" a prior, but the choice is not actually super important)
- Combine the two: $P(\theta_M|M, D) \propto P(D|M, \theta_M)P(\theta_M|M)$, revealing the most probable parameter settings. Optimize.

Model fitting: The recipe

Given some data (e.g. participant responses),

- Hypothesize a model that could have *produced* that data (and with it, candidate parameter settings θ_M)
- Compute "likelihood", probability of data given model (and parameters) $P(D|M, \theta_M)$.
- Hypothesize a prior distribution for the parameters $P(\theta_M|M)$ (it may seem tricky to "come up with" a prior, but the choice is not actually super important)
- Combine the two: $P(\theta_M|M, D) \propto P(D|M, \theta_M)P(\theta_M|M)$, revealing the most probable parameter settings. Optimize.
- Evaluate *model evidence* $P(D|M)$.

Model fitting: The recipe

Given some data (e.g. participant responses),

- Hypothesize a model that could have *produced* that data (and with it, candidate parameter settings θ_M)
- Compute "likelihood", probability of data given model (and parameters) $P(D|M, \theta_M)$.
- Hypothesize a prior distribution for the parameters $P(\theta_M|M)$ (it may seem tricky to "come up with" a prior, but the choice is not actually super important)
- Combine the two: $P(\theta_M|M, D) \propto P(D|M, \theta_M)P(\theta_M|M)$, revealing the most probable parameter settings. Optimize.
- Evaluate *model evidence* $P(D|M)$. If you remember only one thing, remember that this is what Bayesian model comparison is about: evaluating the **probability** of the observed data under a **hypothesized model**.

Model fitting: The recipe

Given some data (e.g. participant responses),

- Hypothesize a model that could have *produced* that data (and with it, candidate parameter settings θ_M)
- Compute "likelihood", probability of data given model (and parameters) $P(D|M, \theta_M)$.
- Hypothesize a prior distribution for the parameters $P(\theta_M|M)$ (it may seem tricky to "come up with" a prior, but the choice is not actually super important)
- Combine the two: $P(\theta_M|M, D) \propto P(D|M, \theta_M)P(\theta_M|M)$, revealing the most probable parameter settings. Optimize.
- Evaluate *model evidence* $P(D|M)$. If you remember only one thing, remember that this is what Bayesian model comparison is about: evaluating the **probability** of the observed data under a **hypothesized model**.

  i.e. How *consistent* is the model with the data?

Model fitting: The recipe

Given some data (e.g. participant responses),

- Hypothesize a model that could have *produced* that data (and with it, candidate parameter settings θ_M)
- Compute "likelihood", probability of data given model (and parameters) $P(D|M, \theta_M)$.
- Hypothesize a prior distribution for the parameters $P(\theta_M|M)$ (it may seem tricky to "come up with" a prior, but the choice is not actually super important)
- Combine the two: $P(\theta_M|M, D) \propto P(D|M, \theta_M)P(\theta_M|M)$, revealing the most probable parameter settings. Optimize.
- Evaluate *model evidence* $P(D|M)$. If you remember only one thing, remember that this is what Bayesian model comparison is about: evaluating the **probability** of the observed data under a **hypothesized model**.

  i.e. How *consistent* is the model with the data?
- Compare for different models.

Parameter estimation: *What* to fit

Maximum Likelihood Estimate ("MLE")

Parameter estimation: *What* to fit

Maximum Likelihood Estimate ("MLE")

If you want to estimate θ_M, maximize the likelihood $P(D|\theta_M, M)$.

Parameter estimation: *What* to fit

Maximum Likelihood Estimate ("MLE")

If you want to estimate θ_M, maximize the likelihood $P(D|\theta_M, M)$.

In practice use the *log* likelihood

Parameter estimation: *What* to fit

Maximum Likelihood Estimate ("MLE")

If you want to estimate θ_M, maximize the likelihood $P(D|\theta_M, M)$.

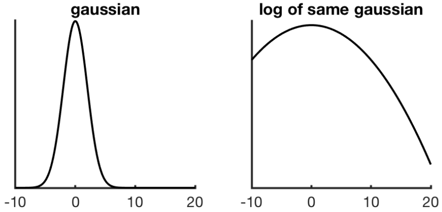
In practice use the *log* likelihood (the log likelihood is concave and sums are more numerically stable).

Parameter estimation: *What* to fit

Maximum Likelihood Estimate ("MLE")

If you want to estimate θ_M, maximize the likelihood $P(D|\theta_M, M)$.

In practice use the *log* likelihood (the log likelihood is concave and sums are more numerically stable).

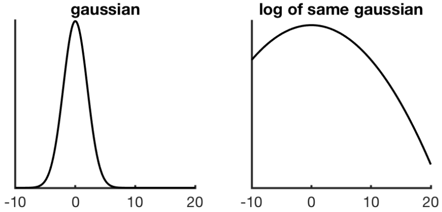


Parameter estimation: *What* to fit

Maximum Likelihood Estimate ("MLE")

If you want to estimate θ_M, maximize the likelihood $P(D|\theta_M, M)$.

In practice use the *log* likelihood (the log likelihood is concave and sums are more numerically stable).



In fact, use the *negative* log likelihood (because standard packages usually *minimize* – "*convex* optimization").

Parameter estimation: *What* to fit

Maximum a posteriori (“MAP”) estimate

Parameter estimation: *What* to fit

Maximum a posteriori ("MAP") estimate
Posterior *mode* (maximum value)

Parameter estimation: *What* to fit

Maximum a posteriori ("MAP") estimate
Posterior *mode* (maximum value)

Why is $P(M|D) \propto P(D|M)P(D)$ rather than $P(M|D) = \dfrac{P(D|M)P(D)}{P(M)}$ okay?

Parameter estimation: *What* to fit

Maximum a posteriori ("MAP") estimate
Posterior *mode* (maximum value)

Why is $P(M|D) \propto P(D|M)P(D)$ rather than $P(M|D) = \dfrac{P(D|M)P(D)}{P(M)}$ okay?

Taking the parameters which yield the maximum does not need normalization.

Parameter estimation: *What* to fit

Maximum a posteriori ("MAP") estimate
Posterior *mode* (maximum value)

Why is $P(M|D) \propto P(D|M)P(D)$ rather than $P(M|D) = \dfrac{P(D|M)P(D)}{P(M)}$ okay?

Taking the parameters which yield the maximum does not need normalization.

Are we Bayesian yet?

Parameter estimation: *What* to fit

Maximum a posteriori ("MAP") estimate
Posterior *mode* (maximum value)

Why is $P(M|D) \propto P(D|M)P(D)$ rather than $P(M|D) = \dfrac{P(D|M)P(D)}{P(M)}$ okay?

Taking the parameters which yield the maximum does not need normalization.

Are we Bayesian yet?

Not really – MAP is just MLE "regularized" with a prior.

Parameter estimation: *What* to fit

Maximum a posteriori ("MAP") estimate
Posterior *mode* (maximum value)

Why is $P(M|D) \propto P(D|M)P(D)$ rather than $P(M|D) = \dfrac{P(D|M)P(D)}{P(M)}$ okay?

Taking the parameters which yield the maximum does not need normalization.

Are we Bayesian yet?

Not really – MAP is just MLE "regularized" with a prior.

If the prior is uniform (flat), multiplying it with the likelihood literally does nothing, i.e. reduces to MLE anyway.

Parameter estimation: *How* to fit

Parameter estimation: *How* to fit

Very much not specific to Bayesian methods.

Parameter estimation: *How* to fit

Very much not specific to Bayesian methods.

- Grid search – take the minimum after an exhaustive search

Parameter estimation: *How* to fit

Very much not specific to Bayesian methods.

- Grid search – take the minimum after an exhaustive search
  Computationally expensive (too dense) and/or misses intermediate settings (too coarse)

Parameter estimation: *How* to fit

Very much not specific to Bayesian methods.

- Grid search – take the minimum after an exhaustive search
  Computationally expensive (too dense) and/or misses intermediate settings (too coarse)
- Gradient descent

Parameter estimation: *How* to fit

Very much not specific to Bayesian methods.

- Grid search – take the minimum after an exhaustive search
  Computationally expensive (too dense) and/or misses intermediate settings (too coarse)
- Gradient descent
- Other standard optimization packages e.g. `fminunc` for MATLAB

Model evaluation: Using the model evidence

In actuality, the model M has parameters θ, so you need to "integrate out" the parameters to compute the **model evidence** $P(D|M)$.

Model evaluation: Using the model evidence

In actuality, the model M has parameters θ, so you need to "integrate out" the parameters to compute the **model evidence** $P(D|M)$.

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)d\theta$$

Model evaluation: Using the model evidence

In actuality, the model M has parameters θ, so you need to "integrate out" the parameters to compute the **model evidence** $P(D|M)$.

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)d\theta$$

The model evidence is also called the **marginal likelihood** because you *marginalize* (recall: marginalizing is summing/integrating!).

Model evaluation: Using the model evidence

In actuality, the model M has parameters θ, so you need to "integrate out" the parameters to compute the **model evidence** $P(D|M)$.

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)d\theta$$

The model evidence is also called the **marginal likelihood** because you *marginalize* (recall: marginalizing is summing/integrating!).

Here you can see the value of using the *log* model evidence to turn products into sums.

Model evaluation: Using the model evidence

In actuality, the model M has parameters θ, so you need to "integrate out" the parameters to compute the **model evidence** $P(D|M)$.

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)d\theta$$

The model evidence is also called the **marginal likelihood** because you *marginalize* (recall: marginalizing is summing/integrating!).

Here you can see the value of using the *log* model evidence to turn products into sums.

Notice a big problem: Integrating over all parameter settings is difficult if you have many parameters!

Model comparison: metrics

How to determine *which model* best explains the data?

In practice: the more metrics the merrier!

Model comparison: metrics

How to determine *which model* best explains the data?

By eye: Meh...

Model comparison: metrics

How to determine *which model* best explains the data?

Using maximum likelihood:

Model comparison: metrics

How to determine *which model* best explains the data?

Using maximum likelihood:
(log) likelihood ratio for nested models, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC)

$$\mathrm{BIC} = \log(n)k - 2\log(\widehat{L})$$

Model comparison: metrics

How to determine *which model* best explains the data?

Using maximum likelihood:
(log) likelihood ratio for nested models, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC)

$$\mathrm{BIC} = \log(n)k - 2\log(\widehat{L})$$

Penalizes for number of parameters k.

Model comparison: metrics

How to determine *which model* best explains the data?

Using full likelihood function:

Model comparison: metrics

How to determine *which model* best explains the data?

Using full likelihood function:

Bayes factors or "log (model) evidence ratio"

Model comparison: metrics

How to determine *which model* best explains the data?

Using full likelihood function:

Bayes factors or “log (model) evidence ratio”

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)P(M_1)}{P(D|M_2)P(M_2)}$$

Model comparison: metrics

How to determine *which model* best explains the data?

Using full likelihood function:

Bayes factors or "log (model) evidence ratio"

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)P(M_1)}{P(D|M_2)P(M_2)}$$

Note: here the normalizing constant $P(D)$ gets cancelled out anyway.

Model comparison: metrics

How to determine *which model* best explains the data?

Using full likelihood function:

Bayes factors or "log (model) evidence ratio"

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)P(M_1)}{P(D|M_2)P(M_2)}$$

Note: here the normalizing constant $P(D)$ gets cancelled out anyway.

Interesting fact: *implicitly* penalizes over-fitting (Bayesian "Occam's Razor")

Getting the full posterior

Analytic/closed-form solution

Getting the full posterior

Analytic/closed-form solution

If it's fair to model your likelihood function as a Gaussian, can pick a "conjugate prior" to make computing the posterior very easy.

Getting the full posterior
Approximations

If your function is not literally a Gaussian, just approximate it as one.

If your function is not literally a Gaussian, just approximate it as one.

Commonly, a *Laplace approximation* is used.
(It is a second-order Taylor series expansion: just means it will have the same first and second derivative, i.e. same slope and curvature)

Extensions

- **Hierarchical model-fitting**
  What if your participants are using different strategies? Group-level distributions over *models*. e.g. use expectation maximization to find.

Extensions

- **Hierarchical model-fitting**
  What if your participants are using different strategies? Group-level distributions over *models*. e.g. use expectation maximization to find.
- **Choosing a prior**: Subjective vs. "non-informative" priors
  Don't use a flat/uniform prior!
  https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Extensions

- **Hierarchical model-fitting**
  What if your participants are using different strategies? Group-level distributions over *models*. e.g. use expectation maximization to find.
- **Choosing a prior**: Subjective vs. "non-informative" priors
  Don't use a flat/uniform prior!
  https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations
- **Sampling-based methods**: Approximating the full posterior by *sampling*
  (look up: Markov Chain Monte Carlo (MCMC), Metropolis-Hastings, Gibbs sampling)

Model-fitting in practice

Model-fitting in practice

It is useful to simulate data with known parameters and try to *recover* those parameters. You may discover you *can't* recover parameters because they “trade-off” in unexpected ways.

For Further Reading

Specific articles

- Trial-by-trial data analysis using computational models (Daw, 2011)
- Bayesian model selection for group studies (Stephan, 2009)

For Further Reading

Specific articles

- Trial-by-trial data analysis using computational models (Daw, 2011)
- Bayesian model selection for group studies (Stephan, 2009)

More resources

- Wei Ji Ma's course notes on Bayesian modeling http://www.cns.nyu.edu/malab/courses.html
- Zoubin Ghahramani's talks and tutorials http://mlg.eng.cam.ac.uk/zoubin
- Michael Betancourt's blog posts https://betanalpha.github.io/writing/
- Quentin Huys' teaching material https://quentinhuys.com/teaching.html
- Hanneke Den Ouden's tutorial on fitting reinforcement learning models https://hannekedenouden.ruhosting.nl/RLtutorial/Instructions.html

For Further Reading

Specific articles

- Trial-by-trial data analysis using computational models (Daw, 2011)
- Bayesian model selection for group studies (Stephan, 2009)

More resources

- Wei Ji Ma's course notes on Bayesian modeling http://www.cns.nyu.edu/malab/courses.html
- Zoubin Ghahramani's talks and tutorials http://mlg.eng.cam.ac.uk/zoubin
- Michael Betancourt's blog posts https://betanalpha.github.io/writing/
- Quentin Huys' teaching material https://quentinhuys.com/teaching.html
- Hanneke Den Ouden's tutorial on fitting reinforcement learning models https://hannekedenouden.ruhosting.nl/RLtutorial/Instructions.html

Textbooks

- Pattern Recognition and Machine Learning (Bishop)
- Information Theory, Inference, and Learning Algorithms (MacKay)