

STSCI 4780:

Propagating uncertainty

Tom Lored, CCAPS & SDS, Cornell University

2020-02-18

Recap: Continuous parameter estimation

- Binary data:
 - Bernoulli, binomial, negative binomial dist'ns
 - Beta posterior and prior dist'ns
- Categorical data:
 - Categorical and multinomial dist'ns
 - Dirichlet posterior and prior dist'ns
- Counts in intervals:
 - Poisson point process and count distribution
 - Gamma distribution posterior
- Scalar measurements with additive Gaussian noise:
 - Gaussian distribution; sufficiency
 - Normal posterior; normal-normal conjugacy; stable estim'n
 - Student's t

Inference with parametric models

Models M_i ($i = 1$ to N), each with a *fixed* set of parameters θ_i .

Each model specifies a *sampling dist'n* (conditional predictive dist'n for hypothetical/possible data, D):

$$p(D|\theta_i, M_i)$$

The θ_i dependence when we fix attention on the *observed* data is the *likelihood function*:

$$\mathcal{L}_i(\theta_i) \equiv p(D_{\text{obs}}|\theta_i, M_i)$$

We may be uncertain about i (model uncertainty) or θ_i (parameter uncertainty)

Henceforth we return to considering only the actually observed data, so we drop the cumbersome subscript: $D = D_{\text{obs}}$.

Classes of problems

Single-model inference

Context = choice of single model (specific i)

Parameter estimation: What can we say about θ_i or $f(\theta_i)$?

Prediction: What can we say about future data D' ?

Multi-model inference

Context = $M_1 \vee M_2 \vee \dots$

Model comparison/choice: What can we say about i ?

Model averaging:

- *Systematic error*: $\theta_i = \{\phi, \eta_i\}$; ϕ is common to all
What can we say about ϕ w/o committing to one model?
- *Prediction*: What can we say about future D' , accounting for model uncertainty?

Model checking

Premise = $M_1 \vee$ “all” alternatives

Is M_1 adequate? (predictive tests, calibration, robustness)

Parameter estimation recap

Problem statement

\mathcal{C} = Model M with parameters θ (+ any add'l info)

H_i = statements about θ ; e.g. " $\theta \in [2.5, 3.5]$," or " $\theta > 0$ "

Probability for any such statement can be found using a *probability density function* (PDF) for θ :

$$\begin{aligned} P(\theta \in [\theta, \theta + d\theta] | \dots) &= f(\theta) d\theta \\ &= p(\theta | \dots) d\theta \end{aligned}$$

Posterior probability density

$$p(\theta | D, M) = \frac{p(\theta | M) \mathcal{L}(\theta)}{\int d\theta p(\theta | M) \mathcal{L}(\theta)}$$

Propagating uncertainty

Often the parameters that most directly or simply allow us to model the data are not the quantities we are ultimately interested in

- I model binary outcome data in terms of the success probability, α . What have I learned about the failure probability, $\beta \equiv 1 - \alpha$? Or about the odds favoring success, $o \equiv \frac{\alpha}{1-\alpha}$?
→ *Change of variables*
- To model the data, I need extra (uncertain) parameters beyond those of interest to me—a background level, a noise amplitude, a calibration factor. What do I know about the parameters of interest? → *Marginalization over nuisance parameters*
- I model available data, D , using a parametric model. What can I say about future data, D' ? → *Prediction*
- I have *two or more* rival parametric models for the available data. How strongly does the evidence favor one model over competitors?
→ *Model comparison*

Change of variables: Binomial inference

Recall the binomial inference problem, using success count data, n , and a flat/uniform prior:

$$\pi(\alpha) = 1; \quad \mathcal{L}(\alpha) = \frac{N!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n}$$

$$\rightarrow p(\alpha|n) = \frac{(N+1)!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n}$$

What does this tell us about $\beta \equiv P(\text{failure}) = 1 - \alpha$?

It's tempting to swap in $\alpha = 1 - \beta$:

$$\pi(\beta) = 1; \quad \mathcal{L}(\beta) = \frac{N!}{n!(N-n)!} (1-\beta)^n \beta^{N-n}$$

$$\rightarrow p(\beta|n) = \frac{(N+1)!}{n!(N-n)!} (1-\beta)^n \beta^{N-n}$$

This has worked, *but only by accident!*

What do the data tell us about the *odds*,

$$o \equiv \frac{\alpha}{1 - \alpha}, \quad \text{with } o \in [0, \infty]$$

Try parameter swapping:

$$o - o\alpha = \alpha \quad \rightarrow \quad o = \alpha(1 + o) \quad \rightarrow \quad \alpha = \frac{o}{1 + o}$$

We're already in trouble with the prior!

$$\pi(o) = 1 \quad \rightarrow \quad \int_0^\infty do \pi(o) = \infty$$

The swap-in posterior can be improper (not normalizable):

$$\alpha^n (1 - \alpha)^{N-n} \quad \rightarrow \quad \left(\frac{o}{1 + o} \right)^n \left(\frac{1}{1 + o} \right)^{N-n}$$

For $N = 2$ and $n = 1$, we expect equal probability for $o < 1$ and $o > 1$, but the integral diverges

Whiteboard work...

Univariate change of variables

Recall the definition of a PDF for x :

$$P(x_* \in [x, x + dx] \mid \dots) = f(x) dx \quad \text{for small } dx$$

Let $y = Y(x)$, with a one-to-one function $Y(x)$, so y is a relabeling of the hypotheses labeled by x

There is a PDF for y :

$$P(y_* \in [y, y + dy] \mid \dots) = g(y) dy \quad \text{for small } dy$$

What $g(y)$ assigns probabilities to y intervals consistent with the probabilities $f(x)$ assigns to the corresponding x intervals?

We'll use the inverse map, from y to x : $x = X(y)$

Consistency condition: Require $f(x)$ and $g(y)$ to assign the same (small) probability to *corresponding* intervals δy and δx :

$$g(y)|\delta y| = f(x)|\delta x|$$

We want to relate δx and δy so that

$$[x, x + \delta x] \iff [y, y + \delta y]$$

For the left boundary, set $x = X(y)$. For the right boundary:

$$\begin{aligned}x + \delta x &= X(y + \delta y) \\X(y) + \delta x &\approx X(y) + X'(y)\delta y \\ \rightarrow \delta x &= X'(y)\delta y\end{aligned}$$

The consistency cond'n becomes $g(y)|\delta y| = f[X(y)] \times |X'(y)\delta y|$,
so

$$\boxed{g(y) = f[X(y)] |X'(y)|}$$

Mnemonic: $g(y) dy = f(x) dx \quad \rightarrow \quad g(y) = f(x) |dx/dy|$

Nuisance Parameters and Marginalization

To model most data, we need to introduce parameters besides those of ultimate interest: *nuisance parameters*

That is, the hypotheses of actual interest (about the *interesting* parameters) are *composite* hypotheses—we would have to specify the nuisance parameters in order to predict the data

Example

We have data from measuring a rate $r = s + b$ that is a sum of an interesting signal s and a background b .

We have additional data just about b .

What do the data tell us about s ?

Simple vs. composite hypotheses

Simple hypotheses

For a set of simple hypotheses, specifying the hypothesis completely determines the sampling distribution (conditional predictive distribution) for possible data: $P(D|H_i)$ is a fully determined function of D when i is specified

- Discrete hypothesis spaces (binary classification; Monte Hall): $P(D|H_i)$ was a table of numbers
- Continuous hypothesis spaces (multinomial, Poisson, Gaussian): Specifying a parameter, θ , determined $p(D|\theta)$ as an explicit function of D (a kind of infinite table of numbers)

Composite/compound hypotheses

Specifying a *composite* hypothesis narrows the choice of the sampling distribution, but requires further information for the distribution to be fully determined

Simple example: An interval hypothesis about a continuous parameter (e.g., for a credible region),

$$H : \theta \in [\theta_l, \theta_u]$$

We can resolve a composite hypothesis into simple components, using LTP to compute it's overall probability. E.g., for an interval hypothesis,

$$\begin{aligned} P(H|\dots) &= \int d\theta p(H, \theta|\dots) \\ &= \int d\theta p(\theta|\dots) p(H|\theta, \dots) \\ &= \int_{\theta_l}^{\theta_u} d\theta p(\theta|\dots) \end{aligned}$$

Marginal posterior distribution

To summarize implications for s , accounting for b uncertainty, *marginalize*:

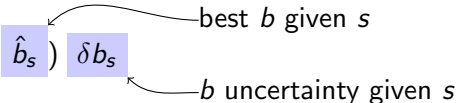
$$\begin{aligned} p(s|D, M) &= \int db \, p(s, b|D, M) \\ &\propto p(s|M) \int db \, p(b|s, M) \mathcal{L}(s, b) \\ &= p(s|M) \mathcal{L}_m(s) \end{aligned}$$

with $\mathcal{L}_m(s)$ the *marginal likelihood function* for s :

$$\mathcal{L}_m(s) \equiv \int db \, p(b|s) \mathcal{L}(s, b)$$

Marginalization vs. Profiling

For insight: Suppose the prior is broad compared to the likelihood
→ for a fixed s , we can accurately estimate b with max likelihood \hat{b}_s , with small uncertainty δb_s .

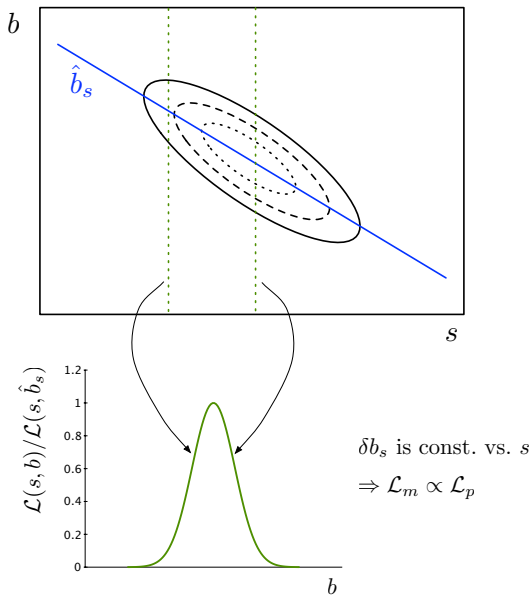
$$\begin{aligned}\mathcal{L}_m(s) &\equiv \int db \, p(b|s) \mathcal{L}(s, b) \\ &\approx p(\hat{b}_s|s) \mathcal{L}(s, \hat{b}_s) \delta b_s\end{aligned}$$


Profile likelihood $\mathcal{L}_p(s) \equiv \mathcal{L}(s, \hat{b}_s)$ gets weighted by a *parameter space volume factor*

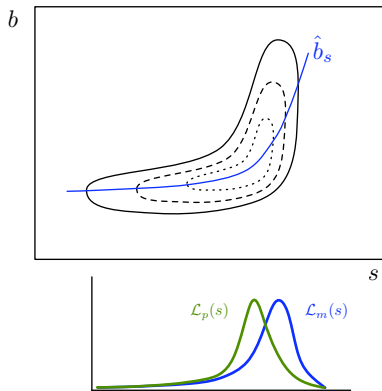
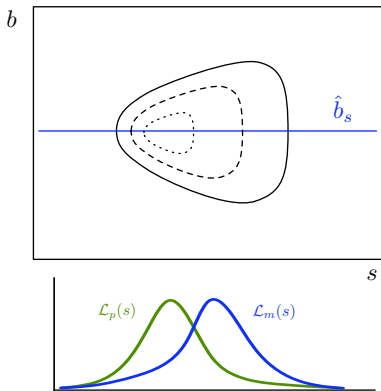
E.g., Gaussians: $\hat{s} = \hat{r} - \hat{b}$, $\sigma_s^2 = \sigma_r^2 + \sigma_b^2$

Background *subtraction* is a special case of background *marginalization*.

Bivariate normals: $\mathcal{L}_m \propto \mathcal{L}_p$



Flared/skewed/bannana-shaped: \mathcal{L}_m and \mathcal{L}_p differ



General result: For a linear (in params) model sampled with Gaussian noise, and flat priors, $\mathcal{L}_m \propto \mathcal{L}_p$

Otherwise, they will likely *differ*

In “*measurement error problems*” the difference can be dramatic

Model comparison

Problem statement

$\mathcal{C} = (M_1 \vee M_2 \vee \dots)$ — Specify a set of models.

$H_i = M_i$ — Hypothesis chooses a model.

Posterior probability for a model

$$\begin{aligned} p(M_i|D, \mathcal{C}) &= p(M_i|\mathcal{C}) \frac{p(D|M_i, \mathcal{C})}{p(D|\mathcal{C})} \\ &\propto p(M_i|\mathcal{C}) \mathcal{L}(M_i) \end{aligned}$$

$$\mathcal{L}(M_i) \equiv p(D|M_i) = \int d\theta_i p(\theta_i|M_i) p(D|\theta_i, M_i)$$

Likelihood for model = Average likelihood for its parameters

$$\mathcal{L}(M_i) = \langle \mathcal{L}(\theta_i) \rangle$$

Varied terminology: Prior predictive = *Marginal likelihood* =

Average likelihood = Global likelihood = (Weight of) Evidence for model

Odds and Bayes factors

A ratio of probabilities for two propositions using the same premises is called the *odds* favoring one over the other:

$$\begin{aligned} O_{ij} &\equiv \frac{p(M_i|D, \mathcal{C})}{p(M_j|D, \mathcal{C})} \\ &= \frac{p(M_i|\mathcal{C})}{p(M_j|\mathcal{C})} \times \frac{p(D|M_i, \mathcal{C})}{p(D|M_j, \mathcal{C})} \end{aligned}$$

The data-dependent part is called the *Bayes factor*:

$$B_{ij} \equiv \frac{p(D|M_i, \mathcal{C})}{p(D|M_j, \mathcal{C})}$$

It is a *likelihood ratio*; the BF terminology is usually reserved for cases when the likelihoods are marginal/average likelihoods for *composite hypotheses*

An automatic Ockham's razor

"Entities must not be multiplied without necessity"

Consider *nested models*:

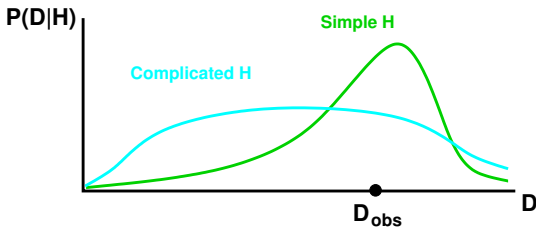
- Simpler model M_1 with parameters θ_1
- "Larger" rival M_2 with parameters $\theta_2 = (\theta_1, \eta)$

$$\Rightarrow \mathcal{L}(\hat{\theta}_2) \geq \mathcal{L}(\hat{\theta}_1)$$

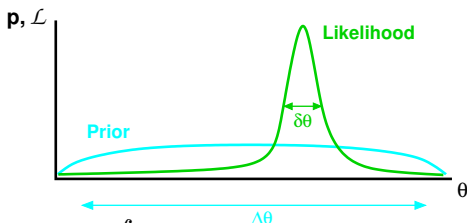
But what about $p(D|M_i) = \int d\theta_i p(\theta_i|M_i) \mathcal{L}(\theta_i)$?

Prior predictive distributions

Normalization implies *there must be data that favor M_1* :



The Ockham Factor



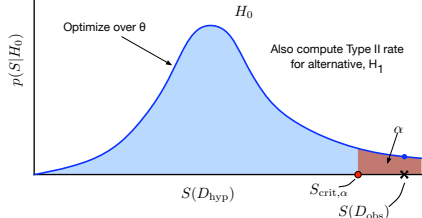
$$\begin{aligned} p(D|M_i) &= \int d\theta_i p(\theta_i|M_i) \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M_i) \mathcal{L}(\hat{\theta}_i) \delta\theta_i \\ &\approx \mathcal{L}(\hat{\theta}_i) \frac{\delta\theta_i}{\Delta\theta_i} \\ &= \text{Maximum Likelihood} \times \text{Ockham Factor} \end{aligned}$$

Models with more parameters often make the data more probable — *for the best fit*

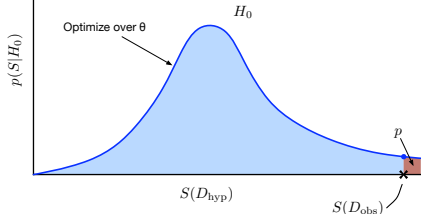
Ockham factor penalizes models for “wasted” *volume of parameter space*

Quantifies intuition that models shouldn't require fine-tuning

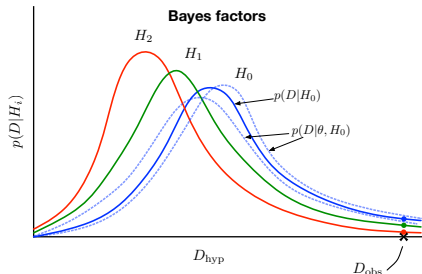
Neyman-Pearson test with Type I error rate α



Fisherian p -value



Bayes factors



- NP & Fisher give H_0 a special role
- NP & Fisher optimize over θ , integrate over D_{hyp}
- Bayes considers rival H_1 symmetrically
- Bayes integrates over θ , uses only D_{obs}

Bayes factors can only compare rival models;
they don't measure "goodness-of-fit"

Posterior predictive p -values are a BDA alternative
for measuring "surprisingness" of data for model checking;
they integrate over both data and parameter spaces

See "p-value note" online

Binary Outcomes: Equal Probabilities?

$M_1: \alpha = 1/2$

$M_2: \alpha \in [0, 1]$ with flat prior

$\mathcal{C}: M_1 \vee M_2; \quad D = \text{FFSSSSFSSSFS} \text{ — 8 successes in 12 trials}$

Maximum Likelihood ratio

From Bernoulli trials model:

$$M_1: \quad p(D|M_1) = \frac{1}{2^N} = 2.44 \times 10^{-4}$$

$$M_2: \quad \mathcal{L}(\hat{\alpha}) = \left(\frac{2}{3}\right)^n \left(\frac{1}{3}\right)^{N-n} = 4.82 \times 10^{-4}$$

$$\frac{p(D|M_1)}{p(D|\hat{\alpha}, M_2)} = 0.51$$

Maximum likelihood (slightly) favors M_2 (on the basis of best-fit α)

Binary outcomes Bayes factor

$$p(D|M_1) = \frac{1}{2^N}; \quad \text{and} \quad p(D|M_2) = \frac{n!(N-n)!}{(N+1)!}$$

$$\begin{aligned} \rightarrow B_{12} &\equiv \frac{p(D|M_1)}{p(D|M_2)} = \frac{(N+1)!}{n!(N-n)!2^N} \\ &= 1.57 \end{aligned}$$

Bayes factor (odds) favors M_1 (equiprobable)

Note that for $n = 6$, $B_{12} = 2.93$; for this small amount of data, we can never be very sure results are equiprobable

If $n = 0$, $B_{12} \approx 1/315$; if $n = 2$, $B_{12} \approx 1/4.8$; for extreme data, 12 flips *can* be enough to lead us to strongly suspect outcomes have different probabilities

Model averaging

Problem statement

$I = (M_1 \vee M_2 \vee \dots)$ — Specify a set of models

Models all share a set of “interesting” parameters, ϕ

Each has different set of nuisance parameters η_i (or different prior info about them)

H_i = statements about ϕ

Model averaging

Calculate posterior PDF for ϕ :

$$\begin{aligned} p(\phi|D, \mathcal{C}) &= \sum_i p(M_i|D, \mathcal{C}) p(\phi|D, M_i) \\ &\propto \sum_i \mathcal{L}(M_i) \int d\eta_i p(\phi, \eta_i|D, M_i) \end{aligned}$$

The model choice is a (discrete) nuisance parameter here

Theme: Parameter space volume

Bayesian calculations sum/integrate over parameter/hypothesis space!

(Frequentist calculations average over *sample* space & typically *optimize* over parameter space.)

- Credible regions integrate over parameter space
- Marginalization weights the profile likelihood by a volume factor for the nuisance parameters
- Model likelihoods have Ockham factors resulting from parameter space volume factors
- Prediction and uncertainty propagation (later topics!) require similar integration over parameter space

Many virtues of Bayesian methods can be attributed to this accounting for the “size” of parameter spaces when considering composite hypotheses. This idea does not arise naturally in frequentist statistics (but it can be added “by hand”).

Roles of the prior

Prior has two roles

- Incorporate any relevant prior information
- Convert likelihood from “intensity” to “measure”
→ account for *size of parameter space*

Physical analogy

$$\text{Heat } Q = \int d\vec{r} [\rho(\vec{r})c(\vec{r})] T(\vec{r})$$

$$\text{Probability } P \propto \int d\theta p(\theta)\mathcal{L}(\theta)$$

Maximum likelihood focuses on the “hottest” parameters.

Bayes focuses on the parameters with the most “heat.”

A high- T region may contain little heat if ρc is low or if its volume is small.

A high- \mathcal{L} region may contain little probability if its prior is low or if its volume is small.