STSCI 4780: Propagating uncertainty

Tom Loredo, 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*:

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

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: $D = D_{obs}$.

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 = $M_1 \vee M_2 \vee \cdots$
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:

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

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, $o \equiv \frac{\alpha}{1-\alpha}$?
  $\rightarrow$ *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? $\rightarrow$ *Marginalization over nuisance parameters*

- I model available data, D, using a parametric model. What can I say about future data, D'? $\rightarrow$ *Prediction*

- I have *two or more* rival parametric models for the available data. How strongly does the evidence favor one model over competitors?
  $\rightarrow$ *Model comparison*

Change of variables: Binomial inference

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

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

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

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

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

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

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

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!

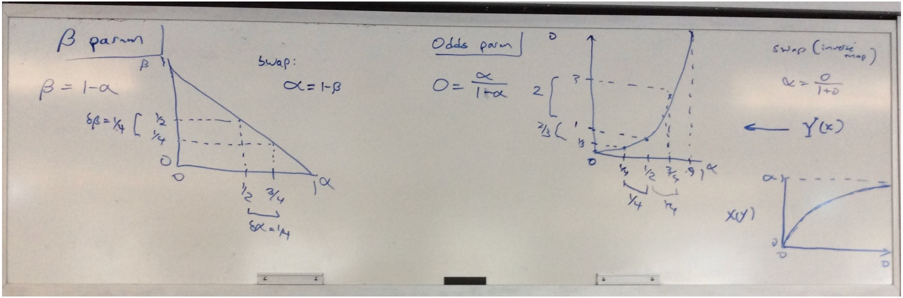
$$\pi(o) = 1 \quad \rightarrow \quad \int_0^\infty do\, \pi(o) = \infty$$

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

Why simple variable replacement fails



Univariate change of variables

Recall the definition of a PDF for x:

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

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:

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

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

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

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

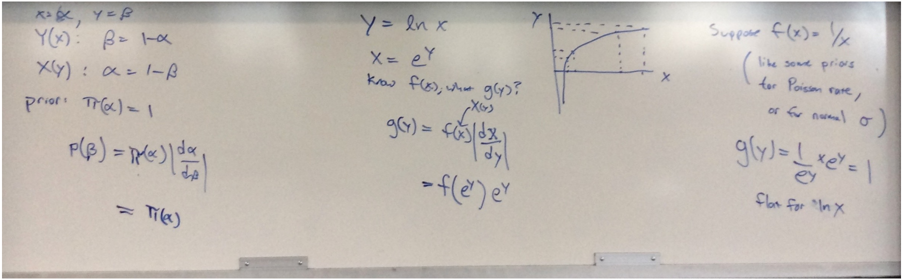
$$\begin{aligned} x + \delta x &= X(y + \delta y) \\ X(y) + \delta x &\approx X(y) + X'(y)\delta y \\ \rightarrow \quad \delta x &= X'(y)\delta y \end{aligned}$$

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: $g(y)\,\mathrm{d}y = f(x)\,\mathrm{d}x \quad \rightarrow \quad g(y) = f(x)|\,\mathrm{d}x/\,\mathrm{d}y|$

Two examples



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 continous 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,

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

Marginal posterior distribution

Specifying the value of one parameter in a multiparameter problem is a composite hypothesis: Specifying just s corresponds to saying one hypothesis in the set $\{(s, b) : b \in [b_l, b_u]\}$ holds

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*:

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