# STSCI 4780/5780: Propagating uncertainty, 1

Tom Loredo, CCAPS & SDS, Cornell University

# Recap: Univariate parameter estimation

- Binary data:
  - Bernoulli, binomial, negative binomial dist'ns
  - Beta 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
  - Marginalizing $\sigma$ & Student's $t$

# Inference with parametric models

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

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

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

The $\theta_i$ 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 $\theta_i$ (parameter uncertainty)

*Henceforth we almost always* **consider only the actually observed data**, *so we typically drop the cumbersome subscript:* $D = D_{\rm obs}$.

*When needed, we'll sometimes use $D_{hyp}$ to refer to hypothetical data.*

# Classes of problems

*Single-model inference*

Context = choice of single model (specific $i$)

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

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

*Multi-model inference ("M-closed")*

Context = $M_1 \vee M_2 \vee \cdots$

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

*Model averaging*:

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

*Model checking ("M-open")*

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 $\theta$ (+ any add'l info)

$H_i$ = statements about $\theta$; 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 $\theta$:

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

*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.

- To model the data, I need extra (uncertain) parameters beyond those of interest to me—a background level, a noise scale, 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, *accounting for parameter uncertainty*? $\rightarrow$ *Model comparison*
- I model binary outcome data in terms of the success probability, $\alpha$. 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*

The LTP will play a key role in addressing these problems.

# 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: Gaussian noise with unknown* $\sigma$

In Lec07 we had parameters $(\mu, \sigma)$, but we were only interested in $\mu$

*Example: Signal + background*

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 $s$?

*Whiteboard work — What do the full likelihoods look like?*

# 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: $P(D|H_i)$ can be directly evaluated when $i$ is specified

- Discrete hypothesis spaces (binary classification; Monte Hall): $P(D|H_i)$ was a column of numbers (or a table when we computed the sampling dist'n over $D_{\rm hyp}$)
- Continuous hypothesis spaces (multinomial, Poisson, Gaussian): Specifying a parameter, $\theta$, determined $p(D|\theta)$ as an explicit function of $\theta$ (a kind of infinite column of numbers), or a joint function of $\theta$ and $D_{\rm hyp}$

## *Composite/compound hypotheses*

Specifying a *composite* hypothesis narrows down the choice of the sampling distribution or likelihood function, 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),

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

What is $p(H|D, \mathcal{C})$? We could try using BT directly:

$$p(H|D, \mathcal{C}) = \frac{p(H|\mathcal{C})\, p(D|H, \mathcal{C})}{p(D|\mathcal{C})}$$

But what is $p(D|H, \mathcal{C})$?

## *LTP and composite hypotheses*

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

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

since $p(H|\theta,\ldots) = 1$ when $\theta$ is in the interval, and 0 otherwise.

We can similarly handle conditioning on an interval hypothesis:

$$\begin{aligned} p(D|H,\mathcal{C}) &= \int \mathrm{d}\theta\, p(D,\theta|H,\mathcal{C}) \\ &= \int \mathrm{d}\theta\, p(\theta|H,\mathcal{C})\, p(D|\theta,H,\mathcal{C}) \end{aligned}$$

Use BT for $p(\theta|H,\mathcal{C})$:

$$p(\theta|H,\mathcal{C}) = \frac{p(\theta|\mathcal{C})\, p(H|\theta,\mathcal{C})}{p(H|\mathcal{C})}$$

Since $p(H|\theta,\ldots) = 1$ when $\theta$ is in the interval and 0 otherwise, this is just the prior, renormalized over the interval, i.e., vanishing outside the interval, and scaled up inside it:

$$p(\theta|H,\mathcal{C}) = \frac{p(\theta|\mathcal{C})}{\int_{\theta_l}^{\theta_u} \mathrm{d}\theta\ p(\theta|\mathcal{C})} = \frac{p(\theta|\mathcal{C})}{F}, \quad \text{with } F < 1$$

Inside the interval, $p(D|\theta,H,\mathcal{C}) = p(D|\theta,\mathcal{C})$ (likelihood!), so

$$p(D|H,\mathcal{C}) = \int_{\theta_l}^{\theta_u} \mathrm{d}\theta\ \frac{p(\theta|\mathcal{C})}{F}\, p(D|\theta,\mathcal{C})$$

# Marginal posterior distribution

Specifying the value of one parameter in a multiparameter problem is a composite hypothesis: Specifying just $s$ in a problem requiring $(s, b)$ 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*:

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

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 $\rightarrow$ for a fixed $s$, we can accurately estimate $b$ with max likelihood $\hat{b}_s$, with small uncertainty $\delta b_s$.

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

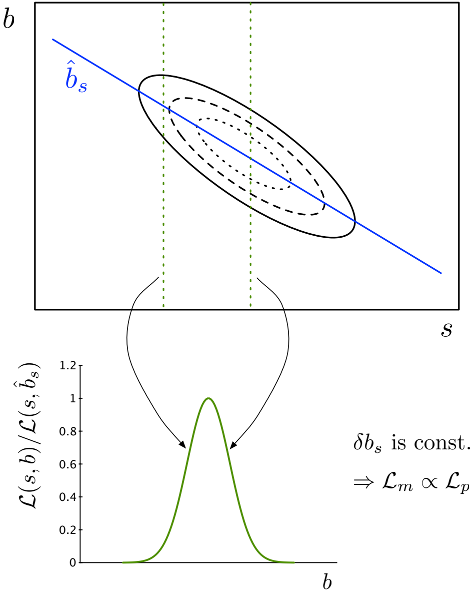
$$\approx p(\hat{b}_s|s)\, \mathcal{L}(s, \underbrace{\boxed{\hat{b}_s}}_{\text{best } b \text{ given } s})\; \underbrace{\boxed{\delta b_s}}_{b \text{ uncertainty given } s}$$

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

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

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

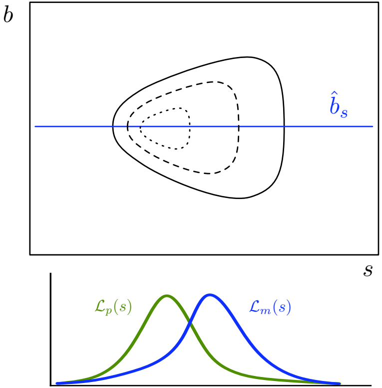
# Bivariate normals: $\mathcal{L}_m \propto \mathcal{L}_p$

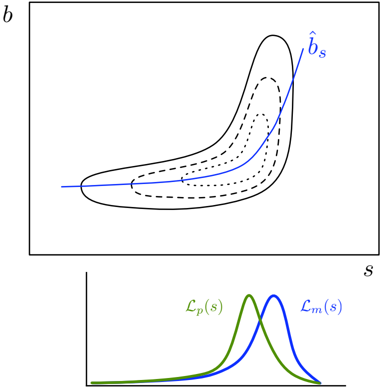



$\delta b_s$ is const. vs. $s$

$\Rightarrow \mathcal{L}_m \propto \mathcal{L}_p$

## 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

# Prediction

Context: Model $M$ with parameters $\theta$
Data: Available data $D$; *future data $D'$*
What does $D$ tell us about $D'$ in the context of the model?

Calculate the *posterior predictive dist'n*:

$$\begin{aligned}
p(D'|D,M) &= \int d\theta\; p(\theta, D'|D,M) \\
&= \int d\theta\; p(\theta|D,M)\, p(D'|\theta,M) \\
&= \int d\theta\; (\text{posterior for } \theta) \times (\text{sampling dist'n for } D')
\end{aligned}$$

Typically the last factor is easy to compute (e.g., binomial, Poisson, or normal dist'n with parameters *given*)

This is propagation of uncertainty (from $\theta$ to $D'$), with a probabilistic rather than deterministic relationship—i.e., $p(D'|\theta, M)$ is not a $\delta$-function

*Supplementary material*

# Predicting a future Bernoulli outcome

FFSSSSFSSSFS ($n = 8$ successes in $N = 12$ trials)

*Bernoulli process likelihood function*

$$p(\mathcal{S}|\alpha, M) = \alpha^n (1-\alpha)^{N-n}$$

*Binomial likelihood function*

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

*Flat prior posterior PDF (beta dist'n)*

$$\begin{aligned} p(\alpha|n, M) &= \frac{(N+1)!}{n!(N-n)!}\alpha^n (1-\alpha)^{N-n} \\ &= \mathrm{Beta}(\alpha|a = n+1, b = N-n+1) \end{aligned}$$

## *Probability for next outcome*

Next outcome $o = 0$ (F) or $o = 1$ (S)

$$
\begin{aligned}
p(o|n, M) &= \int d\alpha \; p(\alpha, o|n, M) \\
&= \int d\alpha \; p(\alpha|n, M) \, p(o|\alpha, M) \\
&= \int d\alpha \; \frac{(N+1)!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n} \times \alpha^o (1-\alpha)^{1-o} \\
&= \frac{(N+1)!}{n!(N-n)!} \int d\alpha \; \alpha^{n+o} (1-\alpha)^{N-n+1-o} \\
&= \frac{(N+1)!}{n!(N-n)!} \times \frac{(n+o)!(N-n-o+1)!}{(N+2)!}
\end{aligned}
$$

$$
\begin{aligned}
p(o|n,M) &= \frac{(N+1)!}{n!(N-n)!} \times \frac{(n+o)!(N-n-o+1)!}{(N+2)!} \\
&= \begin{cases} \frac{n+1}{N+2} & \text{for } o=1 \\ \frac{N-n+1}{N+2} & \text{for } o=0 \end{cases} \\
&\approx \begin{cases} \frac{n}{N} & \text{for } o=1 \\ \frac{N-n}{N} & \text{for } o=0 \end{cases} \quad \text{for } N,n \gg 1
\end{aligned}
$$

*Laplace's rule of succession*:

$$P(\text{next outcome}|\text{past}) \approx \text{Frequency of outcome in the past}$$

Provides a justification for inductive reasoning in IID settings