STSCI 4780: Composite hypotheses: Model comparison, marginalization, and prediction

Tom Loredo, CCAPS & SDS, Cornell University

2020-02-20

Recap/agenda: Uncertainty propagation and composite hypotheses

- Parametric inference problems: single model, multimodel, model checking
- Univariate uncertainty propagation—change of variables
- Simple vs. composite hypotheses
- Marginalizing over nuisance parameters
- *Today*
  - More on uncertainty propagation, marginalization
  - [Prediction—*more in a few weeks*]
  - Model comparison
  - [Model averaging]

Propagating uncertainty with δ functions

If we can compute $p(\theta| \ldots)$ for some scalar parameter θ, but we are interested in $f = F(\theta)$, we've established that

$$p(f| \ldots) = p(\theta| \ldots)\, |\Theta'(f)|,$$

where $\theta = \Theta(f)$ is the inverse map between f and θ

If θ is a vector, and f is a vector with the same number of components,

$$p(f| \ldots) = p(\theta| \ldots) \left| \frac{\partial \Theta(f)}{\partial f} \right|,$$

where the last factor is the determinant Jacobian of the transformation

Often we are interested in $F(\theta)$ that is lower-dimensional than θ, e.g., a scalar function of multiple parameters:

- The difference between two parameters, $s = r - b$
- The ratio of two parameters, $\rho = \theta/\psi$
- Some complex function: planet mass $M_p = f(K, \tau, e)$

The univariate case can be recast using the LTP and δ functions:

$$\begin{aligned} p(f) &= \int d\theta\, p(f,\theta) \\ &= \int d\theta\, p(\theta)\, p(f|\theta) \\ &= \int d\theta\, p(\theta)\, \delta[f - F(\theta)] \end{aligned}$$

Using the definition of δ as a limit of a narrow "hat" function,

$$\delta[f - F(\theta)] = \delta[\theta - \Theta(f)]\ |\Theta'(f)|$$

$$\begin{aligned} \rightarrow \quad p(f) &= \int d\theta\, p(\theta)\, \delta[\theta - \Theta(f)]\ |\Theta'(f)| \\ &= p[\theta = \Theta(f)]\ |\Theta'(f)| \end{aligned}$$

This reproduces our earlier result, but now in a generalizable setting

Two-parameter case: Parameters (θ, ψ) with $f = F(\theta, \psi)$

$$\begin{aligned} p(f) &= \int d\theta \int d\psi\, p(f, \theta) \\ &= \int d\theta \int d\psi\, p(\theta, \psi)\, p(f|\theta, psi) \\ &= \int d\theta \int d\psi\, p(\theta, psi)\, \delta[f - F(\theta, \psi)] \end{aligned}$$

Two ways this may be computed:

- Pick one of θ or ψ, find the inverse map, and use $\delta[\theta - \Theta(f, \psi)]$ or $\delta[\psi - \Psi(f, \theta)]$
- Monte Carlo: Draw samples $\{(\theta_i, \psi_i)\}$ from $p(\theta, \psi)$; compute $\{f_i\}$ with $f_i = F(\theta_i, \psi_i)$

Nuisance Parameters and Marginalization

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

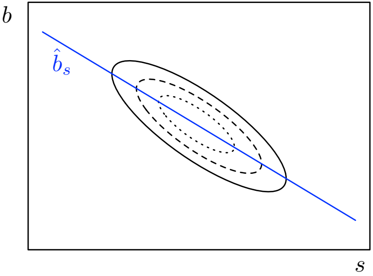
That is, the hypotheses of actual interest (about the *interesting* parameters) are *composite* hypotheses—we would have to specify the nuisance parameters in order to predict the data

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?

Marginal posterior distribution

To summarize implications for s, accounting for b uncertainty, find the *joint* posterior PDF for (s, b), and *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.

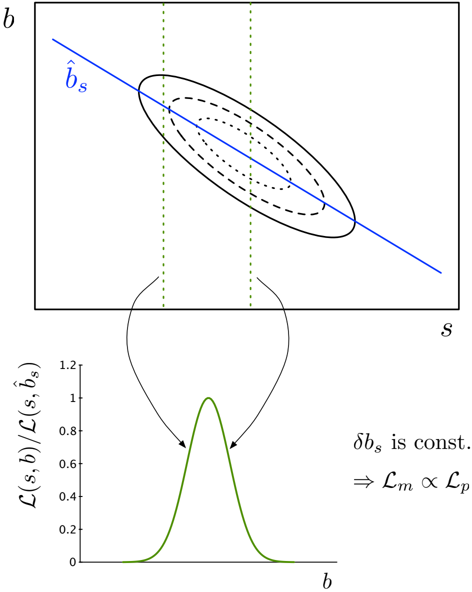
$$
\begin{aligned}
\mathcal{L}_m(s) &\equiv \int db\, p(b|s)\, \mathcal{L}(s,b) \\
&\approx p(\hat{b}_s|s)\, \mathcal{L}(s, \overbrace{\hat{b}_s}^{\text{best } b \text{ given } s})\; \underbrace{\delta b_s}_{b \text{ uncertainty given } s}
\end{aligned}
$$

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

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

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

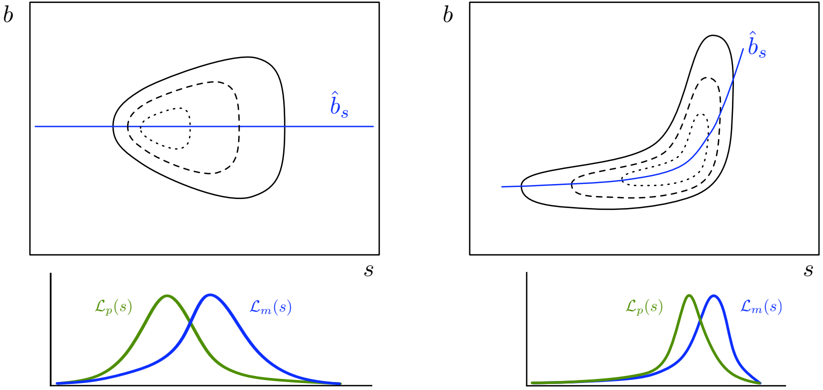
Bivariate normals: $\mathcal{L}_m \propto \mathcal{L}_p$



δb_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 θ
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 θ to D'), with a probabilistic rather than deterministic relationship—i.e., $p(D'|\theta,M)$ is not a δ-function

Model comparison

Problem statement

$\mathcal{C} = (M_1 \vee M_2 \vee \ldots)$ — 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:

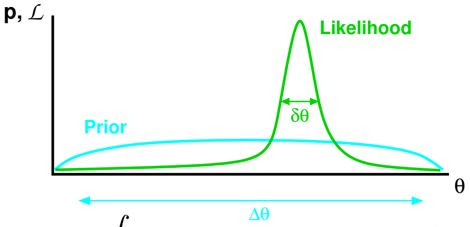
$$
\begin{aligned}
O_{ij} &\equiv \frac{p(M_i|D,\mathcal{C})}{p(M_j|D,\mathcal{C})} \\
&= \frac{p(M_i|\mathcal{C})}{p(M_j|\mathcal{C})} \times \frac{p(D|M_i,\mathcal{C})}{p(D|M_j,\mathcal{C})}
\end{aligned}
$$

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*

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

Example: Equal probabilities for binary outcomes?

M_1: $\alpha = 1/2$ (a simple hypothesis)
M_2: $\alpha \in [0, 1]$ with flat prior
$\mathcal{C}$: $M_1 \vee M_2$; $D =$ FFSSSSFSSSFS — 8 successes in 12 trials

Maximum Likelihood ratio

From Bernoulli trials model:

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

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

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

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

Binary outcomes Bayes factor

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

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

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

Example: Signal detection with Gaussian noise

Data consist of N measurements with additive noise:

$$d_i = \mu + \epsilon_i, \qquad i = 1 \text{ to } N$$

Noise contributions are independent, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, with σ known

Consider two models:

- M_1: $\mu = \mu_1$ (perhaps zero)
- M_2: μ is uncertain with flat prior over $[\mu_l, \mu_u]$ (search range)

Likelihood functions

The form of the sampling dist'n is the same for both models (they just say different things about what μ to use):

$$\begin{aligned}\mathcal{L}(\mu) &= \prod_i p(d_i|\mu, \sigma, \mathcal{C}) \\ &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right)\end{aligned}$$

M_1 marginal likelihood

M_1 just sets $\mu = \mu_1$; equivalently, it uses a δ function prior: $p(\mu|M_1) = \delta(\mu - \mu_1)$:

$$\mathcal{L}(M_1) = C \exp\left(-\frac{N(\mu_1 - \overline{d})^2}{2\sigma^2}\right)$$

M_2 marginal likelihood

M_2 has $p(\mu|M_2) = 1/\Delta$ inside $[\mu_l, \mu_u]$, with $\Delta \equiv \mu_u - \mu_l$, so

$$\begin{aligned}\mathcal{L}(M_2) &\equiv p(D|M_2) = \int \mathrm{d}\mu\, p(\mu|M_2)\, \mathcal{L}(\mu) \\ &= \frac{C}{\Delta}\int_{\mu_l}^{\mu_u} \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \\ &\approx C\, \frac{\sqrt{2\pi}(\sigma/\sqrt{N})}{\Delta} \equiv C\,\Omega\end{aligned}$$

with Ockham factor Ω

Maximum likelihood ratio

For M_2, the likelihood is maximized for $\mu = \overline{d}$ (if it's in the prior range), so the likelihood ratio is

$$R_{12} \equiv \frac{\mathcal{L}(\mu_1)}{\mathcal{L}(\overline{d})} \;=\; \exp\left(-\frac{N(\mu_1 - \overline{d})^2}{2\sigma^2}\right)$$

This is always ≤ 1 (equality only if $\mu_1 = \overline{d}$), disfavoring M_1

Bayes factor

$$\begin{aligned} B_{12} &\equiv \frac{\mathcal{L}(M_1)}{\mathcal{L}(M_2)} \\ &= \frac{\Delta}{\sqrt{2\pi}\sigma/\sqrt{N}}\, R_{12} \end{aligned}$$

Divides the MLR by Ockham factor (typically < 1)

$$\Omega = \frac{\sqrt{2\pi}(\sigma/\sqrt{N})}{\Delta}$$

Model averaging

Problem statement

$I = (M_1 \vee M_2 \vee \ldots)$ — 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! This is *the signature feature* of the Bayesian approach.

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

- Credible regions integrate over parameter space
- Uncertainty propagation integrates over parameter space
- Marginalization weights the profile likelihood by a volume factor for the nuisance parameters
- Prediction integrates over parameter space
- Model (marginal) likelihoods & Bayes factors have Ockham factors resulting from parameter space volume factors

Many/most interesting hypotheses are really *composite*. Many virtues of Bayesian methods can be attributed to 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

$$\begin{aligned} \text{Heat} \quad Q &= \int d\vec{r}\,[\rho(\vec{r})c(\vec{r})]\,T(\vec{r}) \\ \text{Probability} \quad P &\propto \int d\theta\; p(\theta)\mathcal{L}(\theta) \end{aligned}$$

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.