Marginalization vs. Profiling

Marginal distribution for signal s, eliminating backgrond b:

$$p(s|D, M) \quad \propto \quad p(s|M)\mathcal{L}_m(s)$$

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

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

For insight: Suppose for a fixed s, we can accurately estimate b with max likelihood $\hat{b}_s$, with small uncertainty δb_s.

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

best b given s

b uncertainty given s

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

$$\mathcal{L}_m(s) \;\approx\; p(\hat{b}_s|s)\,\mathcal{L}(s,\hat{b}_s)\,\delta b_s$$

best b given s

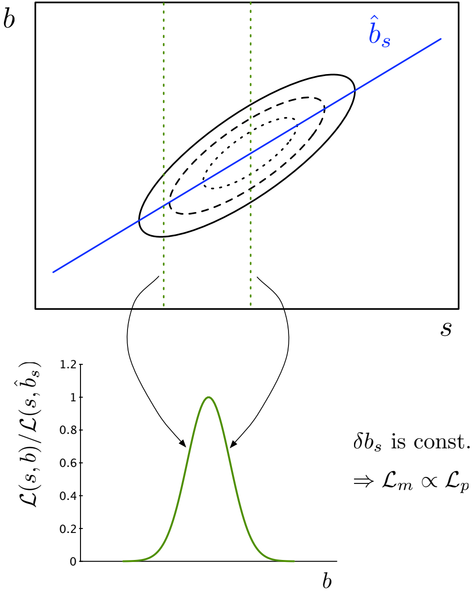
b uncertainty given s

Methods for handling nuisance parameters aim to account for *nuisance parameter uncertainty*.

Profiling takes into account *variation of the best-fit value of b with s*. This will typically be the most important effect of b uncertainty. It accounts for correlation between s and b that is ignored if one just fixes $b = \hat{b}$.

Marginalization implicitly does this, and *additionally accounts for the uncertainty in $\hat{b}_s$*. When δb_s varies with s, one typically finds the marginal is wider than the profile; the profile ignores important uncertainty.

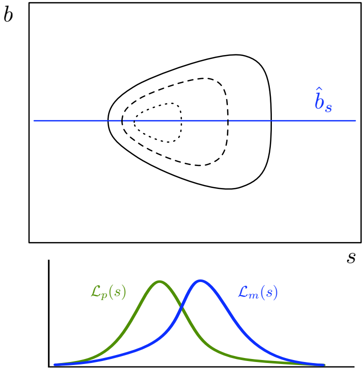
Bivariate normals: $\mathcal{L}_m \propto \mathcal{L}_p$

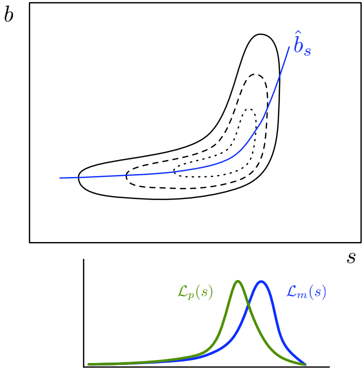


δb_s is const. vs. s

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

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





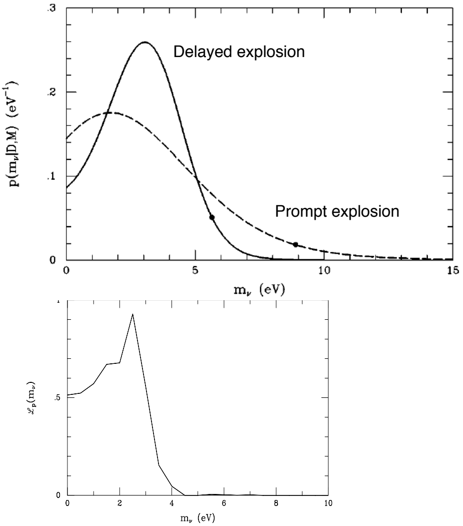
General result: For a linear (in params) model sampled with Gaussian noise, and flat priors, $\mathcal{L}_m \propto \mathcal{L}_p$.

In asymptotically normal regime, $\mathcal{L}_m \propto \mathcal{L}_p$. Otherwise, they will likely *differ*.

In *"measurement error problems"* the difference can have dramatic consequences.

Astrophysics Example: SN 1987A m_ν Limits

Marginal PDF and profile likelihood for $m_{\bar{\nu}_e}$ based on SN 1987A neutrino energies and arrival times; two SN ν emission models.



Discrete Example: Basu's Problem*

Urn contains 1000 colored red ("1") and green ("-1") balls:

- 980 have color θ, uniquely numbered from $\mathcal{S} = \{1, 2, \ldots, 980\}$
- 20 have color $-\theta$, all with the same (unknown) number $\phi \in \mathcal{S}$

What is the color of the majority, θ?

Color data only

Draw a ball; observe only its color, x.

Sampling distribution: Knowing ϕ does not help you predict the color $\rightarrow$ the sampling dist'n does not depend on ϕ:

$$p(x|\theta, \phi) = \begin{cases} 0.98 & \text{for } x = \theta \\ 0.02 & \text{for } x = -\theta \end{cases}$$

Maximum likelihood guess is $\theta = x$.
This will be correct with long-run frequency 0.98.

* D. Basu (1975) "Statistical information and likelihood," *Sankhya*, A37, 1–71

Color & number data

Draw a ball; observe its color, x, and number, n.

Sampling distribution:

$$
\begin{aligned}
p(x, n|\theta, \phi) &= p(x|\theta, \phi)p(n|x, \theta, \phi) \\
&= \begin{cases} 0.98 \times \frac{1}{980} = 0.001 & \text{for } \theta = x, \text{ any } \phi \\ 0.02 \times 1 = 0.02 & \text{for } \theta = -x,\ n = \phi \\ 0.02 \times 0 = 0 & \text{for } \theta = -x,\ n \neq \phi \end{cases}
\end{aligned}
$$

Profile likelihood: Plug in $\hat{\phi}(\theta)$:

$$
\begin{aligned}
\mathcal{L}_p(\theta) &\equiv p(x, n|\theta, \hat{\phi}(\theta)) \\
&= \begin{cases} 0.001 & \text{if } \theta = x \\ 0.02 & \text{if } \theta = -x \end{cases}
\end{aligned}
$$

Maximum profile likelihood guess is $\theta = -x$.
This will be correct with long-run frequency 0.02.

Marginal likelihood: Use flat prior over $\mathcal{S}$ for ϕ:

$$\mathcal{L}_m(\theta) \equiv \sum_{\phi=1}^{980} \frac{1}{980} p(x, n|\theta, \phi) = \begin{cases} 0.98/980 & \text{for } \theta = x \\ 0.02/980 & \text{for } \theta = -x \end{cases}$$

Maximum marginal likelihood guess is $\theta = x$.

Example: Draw a red ticket numbered 42.

The one hypothesis with $(\theta = \text{Green}, \phi = 42)$ has larger likelihood and posterior probability than any hypothesis with $\theta = \text{Red}$.

But there are *so many* hypotheses with $\theta = \text{Red}$ that it is more plausible (probable!) that one of them is true, than that $\theta = \text{Green}$.

We must somehow account for the size of the plausible ϕ space.

Continuous Example: The Neyman-Scott problem

Calibrating a noise level

Need to measure several sources with signal amplitudes μ_i, with an "uncalibrated" instrument that adds Gaussian noise with *unknown* but constant σ.

Ideally, either:

- Measure calibration sources of known amplitudes; the scatter of the measurements from the known values allows easy inference of σ.
- Measure one source many times; from many samples we can easily learn both μ_i and σ.

Neyman-Scott problem (1948): Calibrate as-you-go

- No calibration sources are available.
- We have to measure N sources with finite resources, so only a few measurements of each source are available.

The multiple measurements of a single source yield a noisy estimate of σ.

$\rightarrow$ Pool all the data to more precisely estimate σ.

Pairs of measurements

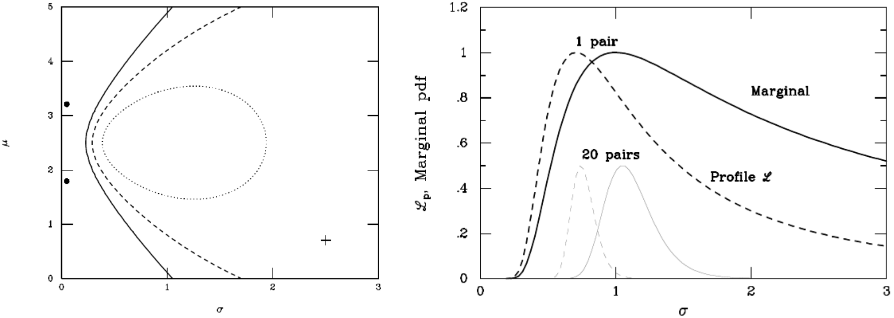
Make 2 measurements (x_i, y_i) for each of the N quantities μ_i.

Likelihood:

$$\mathcal{L}(\{\mu_i\}, \sigma) = \prod_i \frac{\exp\left[-\frac{(x_i-\mu_i)^2}{2\sigma^2}\right]}{\sigma\sqrt{2\pi}} \times \frac{\exp\left[-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right]}{\sigma\sqrt{2\pi}}$$

Profile likelihood $\mathcal{L}_p(\sigma) = \max_{\{\mu_i\}} \mathcal{L}(\{\mu_i\}, \sigma)$
Plugs in $\hat{\mu}_i = \frac{1}{2}(x_i + y_i)$

Joint & Marginal Results for $\sigma = 1$



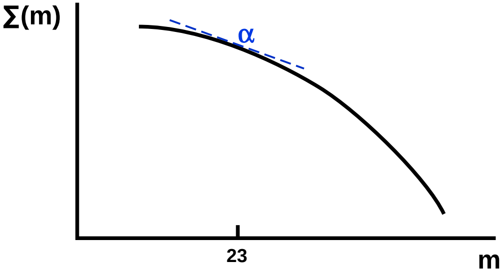
The marginal $p(\sigma|D)$ and $\mathcal{L}_p(\sigma)$ differ dramatically!
Profile likelihood estimate converges to $\sigma/\sqrt{2}$.

The total # of parameters grows with the # of data.
$\Rightarrow$ *Volumes along μ_i do not vanish as $N \to \infty$.*

Astro Example—Distribution of Source Magnitudes

Measure m_i of sources following a "rolling power law" flux dist'n (i.e., a "rolling exponential" magnitude dist'n; inspired by TNOs)

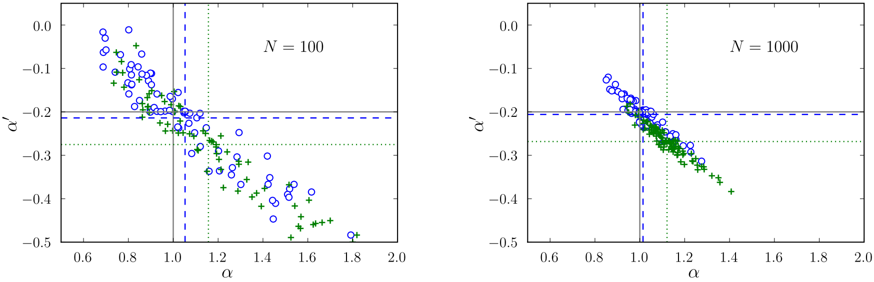
$$\Sigma(m) \propto 10^{[\alpha(m-23)+\alpha'(m-23)^2]}$$



Simulate 100 surveys of populations drawn from the same dist'n.
Simulate data for photon-counting instrument, fixed count threshold.
Measurements have uncertainties 1% (bright) to ≈ 30% (dim).

Analyze simulated data with maximum ("profile") likelihood and Bayes.

Parameter estimates from Bayes (circles) and profile likelihood (crosses):



Uncertainties don't average out!

This failure of profile likelihood has been (re)discovered several times in various astronomical sub-disciplines.

A Generalized Wilks Theorem*

Setting

Test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$

Log likelihood ratio:

$$\lambda(\theta) = \log \mathcal{L}(\hat{\theta}) - \log \mathcal{L}(\theta)$$

Test using maximum log likelihood ratio, $\lambda_0 = \lambda(\theta_0)$.
What is the asymptotic distribution for λ_0?

Conditions (crudely summarize!)

- The MLE converges to the true value, but in a weaker sense than requiring asymptotic normality
- Likelihood contours are "fan-shaped" (i.e., scaled versions of a single shape)
- The size of the contours grows like a power of λ

* Fan, Hung, & Wong (2000) "Geometric understanding of likelihood ratio statistics," *JASA* **95**, 451

A Generalized Wilks Theorem*

Setting

Test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$

Log likelihood ratio:

$$\lambda(\theta) = \log \mathcal{L}(\hat{\theta}) - \log \mathcal{L}(\theta)$$

Test using maximum log likelihood ratio, $\lambda_0 = \lambda(\theta_0)$.
What is the asymptotic distribution for λ_0?

Conditions (crudely summarize!)

- The MLE converges to the true value, but in a weaker sense than requiring asymptotic normality
- Likelihood contours are "fan-shaped" (i.e., scaled versions of a single shape)
- The size of the contours grows like a power of λ

* Fan, Hung, & Wong (2000) "Geometric understanding of likelihood ratio statistics," *JASA* **95**, 451

Result

Theorem: $\lambda_0 \sim \text{Gamma}(rp)$ for p parameters, and $r =$ power for how contour size grows with λ.

Examples given:

- Multivariate exponential, where contours are hypertriangles and MLE is exponentially distributed; $2\lambda \sim \chi^2_{2p}$
- Multivariate uniform
- Nonlinear normal, $N(\theta^3, I_p)$; MLE $\sim$ cube root of a normal; contours are ellipses; here $r = 1/2$
- Different asymptotic behavior in different directions
- Nuisance parameters