Non-Exchangeable Hierarchical Bayes Models for Synthesizing Disparate Information

Robert Wolpert
Duke Department of Statistical Science
rlw@duke.edu

BIRS 18w5095
Statistical Challenges in the Search for Dark Matter
2018 March 01

Outline

Three Running Examples

1. **Toy Example:** Number of "successes" in fixed number of "trials". Could be coin tosses, QC testing, clinical trials, whatever. Model as binomial:

$$Y \sim \mathrm{Bi}(n, \theta)$$

2. **GRBs:** Observe counts of gamma-ray photons, binned by arrival times and energy range. Model continuous-time process as inhomogeneous Poisson process

$$Y_t \sim \mathrm{Po}\Big(f_\theta(t)\, dt\Big)$$

   with some semi-parametric structure on $\{f_\theta\}$.

3. **Cosmic:** Make a variety of observations using a range of instruments, in the hope of learning something deep.

Outline

One data source

- Observe random variable (or vector) Y;
- Believe (or model) $Y \sim f(y)$ for some uncertain pdf $f(\cdot)$;
- Model uncertainty through parametric model

$$f \in \{f_\theta(y) : \ \theta \in \Theta\}$$

for some uncertain parameter θ from a set Θ.

Or, better for us, write in conditional form as

$$f \in \{f(y \mid \theta) : \ \theta \in \Theta\}.$$

Outline

Inference

Frequentist:

$$\hat{\theta}(y) := \operatorname*{argmax}_{\theta \in \Theta} f(y \mid \theta) \qquad = \theta \pm se(y)$$

$$se(y) := \left\{ \mathsf{E}_\theta |\hat{\theta}(Y) - \theta|^2 \right\}^{1/2} \approx \left\{ I(\theta) \right\}^{-1/2}$$

Inference

Frequentist:

$$\hat{\theta}(y) := \operatorname*{argmax}_{\theta\in\Theta} f(y \mid \theta) \qquad = \theta \pm se(y)$$

$$se(y) := \left\{ \mathsf{E}_\theta |\hat{\theta}(Y) - \theta|^2 \right\}^{1/2} \approx \left\{ I(\theta) \right\}^{-1/2}$$

Bayesian:

$$\bar{\theta}(y) := \frac{\int_\Theta \theta \; f(y \mid \theta)\, \pi(d\theta)}{\int_\Theta \; f(y \mid \theta)\, \pi(d\theta)} = \theta \pm sd(y)$$

$$sd(y) := \left\{ \mathsf{E}_Y |\bar{\theta}(Y) - \theta|^2 \right\}^{1/2} \approx \left\{ I(\theta) \right\}^{-1/2}$$

Toy Example

$$Y = \# \text{ of successes in } n \text{ indep trials with same prob}$$
$$\sim \text{Bi}(n, \theta), \qquad \theta \in \Theta = [0, 1]$$

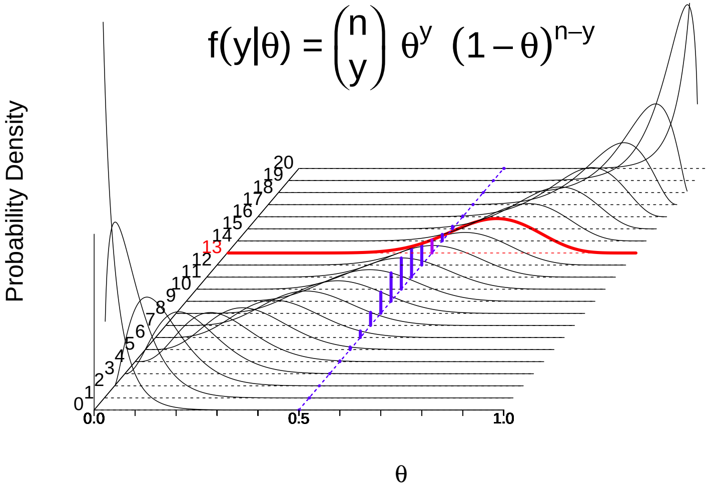
$$f(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Let's take $n = 20$ and observe $y = 13$. Then $f(y \mid \theta)$ can be viewed as either:

Probability Mass Function:	A function of y,	for fixed θ; or
Likelihood Function:	A function of θ,	for fixed y.

If	θ	$<$	0.5	$\Rightarrow$	Y	$\geq$	13	is extremely low;
If	θ	$\approx$	0.65	$\Rightarrow$	Y	$\approx$	13	is rather likely;
If	θ	$>$	0.8	$\Rightarrow$	Y	$\leq$	13	is extremely low.

In pictures:



$f(y|\theta) = \binom{n}{y}\,\theta^y\,(1-\theta)^{n-y}$
Probability Density
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
0.0
0.5
1.0
θ

Toy Example Inference

Again let $Y \sim \text{Bi}(n, \theta)$; define $S := Y$, $F := (n - Y)$.
Perhaps (as before) $y = 13$ and $n = 20$.

Frequentist:

$$\hat{\theta}(y) := \frac{y}{n} \qquad = \frac{13}{20} \qquad = 0.6500$$

$$se(y) := \left\{ \frac{y(n-y)}{n^3} \right\}^{1/2} \qquad = \sqrt{\frac{91}{8000}} \qquad \approx 0.1067$$

Toy Example Inference

Again let $Y \sim \text{Bi}(n, \theta)$; define $S := Y$, $F := (n - Y)$.
Perhaps (as before) $y = 13$ and $n = 20$.

Frequentist:

$$\hat{\theta}(y) := \frac{y}{n} \qquad = \frac{13}{20} \qquad = 0.6500$$

$$se(y) := \left\{ \frac{y(n-y)}{n^3} \right\}^{1/2} \qquad = \sqrt{\frac{91}{8000}} \qquad \approx 0.1067$$

Bayesian, with "Reference" prior $\theta \sim \text{Be}(\frac{1}{2}, \frac{1}{2})$:

$$\bar{\theta}(y) := \frac{y + \frac{1}{2}}{n+1} \qquad = \frac{13.5}{21} \qquad \approx 0.6428$$

$$sd(y) := \left\{ \frac{(y + \frac{1}{2})(n - y + \frac{1}{2})}{(n+1)^2(n+2)} \right\}^{1/2} = \sqrt{\frac{101.25}{9702}} \approx 0.1021$$

Moral

With **one data source**, **moderately large sample size**, and **moderately flat prior**, Frequentist and Bayesian methods both work well and give about the same answers.

GRB Example

Here we model Gamma Ray Burst photon arrivals as a Cox process:

$$Y_t \sim \mathrm{Po}\Big(f_\theta(t)\,dt\Big)$$

for some structured random mean function $f_\theta(t)$. Below we will take

$$f_\theta(t) = B + \sum_{j=1}^{J} A_j k_j(t \mid \theta)$$

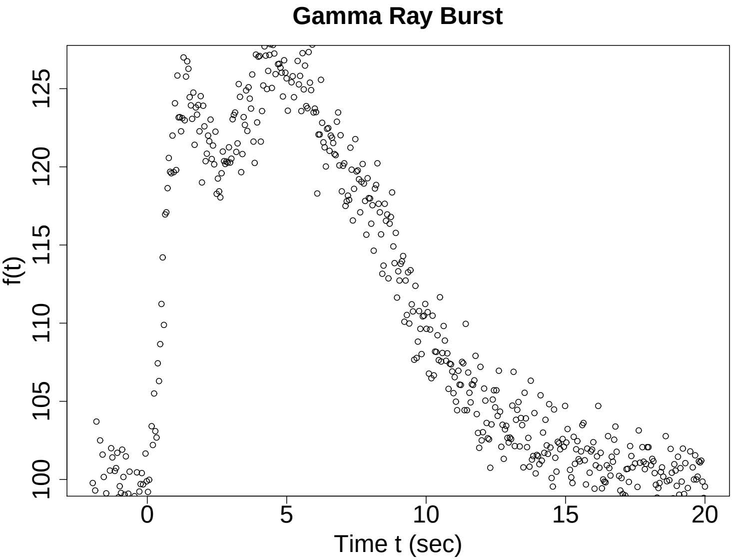
to be a background rate plus a weighted sum of kernels of "Norris" form

$$k_j(t \mid \theta) = \exp\left(2\sqrt{\tau_{1j}\tau_{2j}} - \frac{\tau_{1j}}{(t - t_{0j})} - \frac{(t - t_{0j})}{\tau_{2j}}\right)\mathbf{1}_{t > t_{0j}}$$

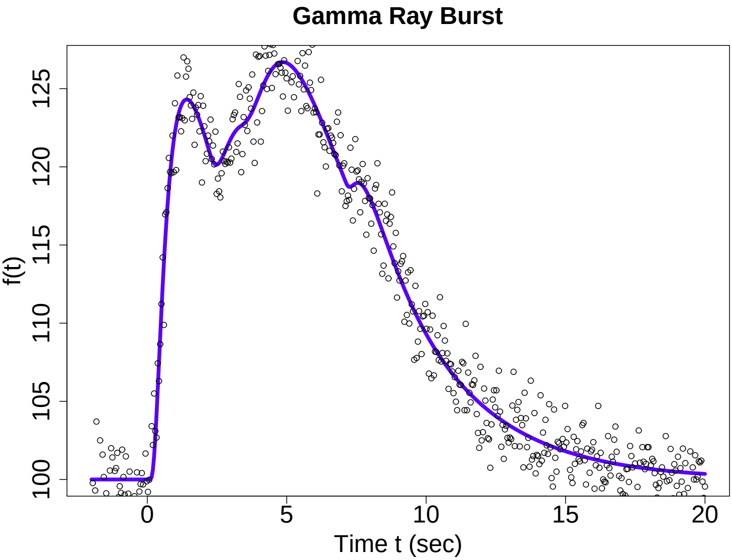
with uncertain polydimensional parameter

$$\theta = \big(B, J, \vec{A}, \vec{t_0}, \vec{\tau}_1, \vec{\tau}_2\big)$$

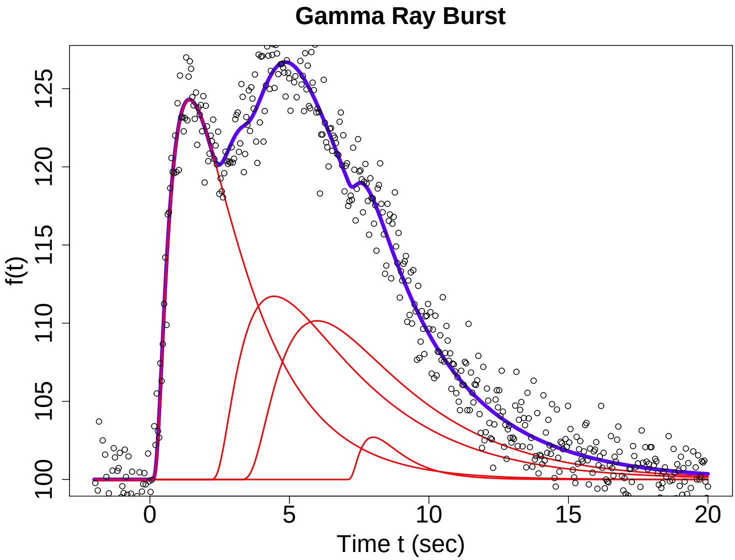
GRB Data (BATSE Poisson bin counts):



GRB Smoothed estimate of Poisson mean:



GRB Resolution of burst into pulses:

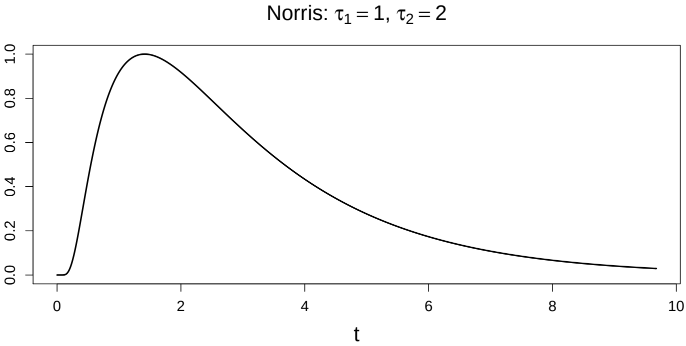


FRED: Norris Kernels

$$f(t \mid \theta) = B + \sum_{j=1}^{J} A_j \exp\left(2\sqrt{\tau_{1j}\tau_{2j}} - \frac{\tau_{1j}}{(t - t_{0j})} - \frac{(t - t_{0j})}{\tau_{2j}}\right)\mathbf{1}_{t > t_{0j}}$$

$$\theta = \left(B, J, \vec{A}, \vec{t_0}, \vec{\tau}_1, \vec{\tau}_2\right)$$

A and B are *amplitudes* (in s^{-1}); trigger t_0 and time constants τ_1, τ_2 are *times* (in s). Note Fast Rise Exponential Decay, or FRED shape.



GRB Example Inference I: Frequentist

Frequentist:

- Try fitting one pulse to light curve:

$$f(t \mid \theta) = B + A_1 \exp\left(2\sqrt{\tau_{11}\tau_{21}} - \frac{\tau_{11}}{(t - t_{01})} - \frac{(t - t_{01})}{\tau_{21}}\right)\mathbf{1}_{t > t_{01}}$$

GRB Example Inference I: Frequentist

Frequentist:

- Try fitting one pulse to light curve:

$$f(t \mid \theta) = B + A_1 \exp\left(2\sqrt{\tau_{11}\tau_{21}} - \frac{\tau_{11}}{(t - t_{01})} - \frac{(t - t_{01})}{\tau_{21}}\right)\mathbf{1}_{t>t_{01}}$$

- Like it? Quit and report $\theta = \big(B, J = 1, A_1, t_0, \tau_1, \tau_2\big)$.

GRB Example Inference I: Frequentist

Frequentist:

- Try fitting one pulse to light curve:

$$f(t \mid \theta) = B + A_1 \exp\left(2\sqrt{\tau_{11}\tau_{21}} - \frac{\tau_{11}}{(t - t_{01})} - \frac{(t - t_{01})}{\tau_{21}}\right)\mathbf{1}_{t > t_{01}}$$

- Like it? Quit and report $\theta = \big(B, J = 1, A_1, t_0, \tau_1, \tau_2\big)$.
- No? Try two:

$$f(t \mid \theta) = B + \sum_{j=1}^{2} A_j \exp\left(2\sqrt{\tau_{1j}\tau_{2j}} - \frac{\tau_{1j}}{(t - t_{0j})} - \frac{(t - t_{0j})}{\tau_{2j}}\right)\mathbf{1}_{t > t_{0j}}$$

GRB Example Inference I: Frequentist

Frequentist:

- Try fitting one pulse to light curve:

$$f(t \mid \theta) = B + A_1 \exp\left(2\sqrt{\tau_{11}\tau_{21}} - \frac{\tau_{11}}{(t - t_{01})} - \frac{(t - t_{01})}{\tau_{21}}\right)\mathbf{1}_{t > t_{01}}$$

- Like it? Quit and report $\theta = \big(B, J = 1, A_1, t_0, \tau_1, \tau_2\big)$.
- No? Try two:

$$f(t \mid \theta) = B + \sum_{j=1}^{2} A_j \exp\left(2\sqrt{\tau_{1j}\tau_{2j}} - \frac{\tau_{1j}}{(t - t_{0j})} - \frac{(t - t_{0j})}{\tau_{2j}}\right)\mathbf{1}_{t > t_{0j}}$$

- Like it? Quit and report $\theta = \big(B, J = 2, \vec{A}, \vec{t_0}, \vec{\tau}_1, \vec{\tau}_2\big)$
- No? Try three... or four... until you do, and report θ.

GRB Example Inference II: Bayes

Bayesian:

- Choose joint prior on $\theta = \{J, \vec{A}, \vec{t_0}, \vec{\tau}_1, \vec{\tau}_2\}$.
  For J and the amplitudes $\{A_j\}$ with $A_j \geq \varepsilon$ for some threshold $\varepsilon > 0$, we[1] use Lévy processes built on Gamma processes
  (so J has Poisson dist'n, and $\{(\tau_{1j}, \tau_{2j})\}_{1 \leq j \leq J}$ are iid, given J.)
- Use Reversible Jump (varying J) Metropolis/Hastings MCMC to sample $\{\theta^{(t)}\}_{t \in \mathbb{N}}$ from posterior distribution.
- Report marginal distributions (or means & variances) of any feature of interest— like $\{J\}$ or total number of photons or max amplitude or duration at half-max-height or ...

[1] Joint work with Tom Loredo, Jon Hakkila, and Duke Stats PhD Mary Beth Broadbent

Moral

With **complex models**, Bayesian methods offer richer inferential products than Frequentist ones do, with better reflection of **uncertainty**.
But either approach can deliver point estimates.

Cosmic Example

Cosmological Parameters for ΛCDM Universe
(one parameter choice)

Baryon Density	Ω_b	0.0486
DM Density	Ω_c	0.2589
Age of Universe	t_0	$13.799 \cdot 10^9$ yr
Scalar Spectral Density	n_s	0.997
Curvature fluct. amplitude	Δ_R^2	$2.441 \cdot 10^{-9}$
Reionization optical depth	τ	0.066

Data bearing on these include CMB anisotropy measurements, brightness/redshift relation for SNe, baryon acoustic oscillation feature of large-scale galaxy clustering, WGL, etc. For any of these, find LH:

$$\theta := (\Omega_b, \Omega_c, t_0, n_s, \Delta_R^2, \tau)$$
$$\mathcal{L}(\theta) \propto f(Y \mid \theta)$$

Outline

$J > 1$ Similar Experiments

If we have $J \geq 1$ independent experiments, each generating evidence

$$Y_j \sim f_j(y \mid \theta)$$

about the same uncertain quantity θ for its own likelihood function $f_j(y \mid \theta)$ (these can differ for different js, but all depend on the *same* uncertain $\theta \in \Theta$), then the *vector* $\mathbf{Y} = (Y_1, \ldots, Y_J)$ has pdf $f_Y(\mathbf{x} \mid \theta) = \prod_{j \leq J} f_j(y_j \mid \theta)$, and so the total evidence is embodied by the product likelihood function:

$$\mathcal{L}(\theta \mid \mathbf{Y}) \propto \prod_{j \leq J} f_j(y_j \mid \theta)$$

Freq: $\hat{\theta}(\mathbf{Y}) := \operatorname*{argmax}_{\theta \in \Theta} \mathcal{L}(\theta \mid \mathbf{Y})$ Bayes: $\bar{\theta}(\mathbf{Y}) := \dfrac{\int_\Theta \theta \; \mathcal{L}(\theta \mid \mathbf{Y})\, \pi(d\theta)}{\int_\Theta \mathcal{L}(\theta \mid \mathbf{Y})\, \pi(d\theta)}$

Toy Example Again

If we have $Y_j \sim \text{Bi}(n_j, \theta)$ successes in n_j independent trials, all with the *same* probability $\theta \in \Theta = [0, 1]$ of success, and if we express our prior ignorance about θ by the flat prior $\pi(\theta) \equiv 1$, then the aggregate evidence about θ is given by

$$
\begin{aligned}
\mathcal{L}(\theta \mid \mathbf{Y}) &\propto \prod_{j \le J} f_j(y_j \mid \theta)\, \pi(\theta) \\
&\propto \prod_{j \le J} \binom{n_j}{y_j} (\theta)^{y_j} (1-\theta)^{n_j - y_j} \\
&\propto (\theta)^{\sum y_j} (1-\theta)^{\sum (n_j - y_j)} \\
&\sim \text{Be}(1 + y_+, \ \ 1 + n_+ - y_+),
\end{aligned}
$$

Toy Example Again

If we have $Y_j \sim \text{Bi}(n_j, \theta)$ successes in n_j independent trials, all with the *same* probability $\theta \in \Theta = [0, 1]$ of success, and if we express our prior ignorance about θ by the flat prior $\pi(\theta) \equiv 1$, then the aggregate evidence about θ is given by

$$\begin{aligned}
\mathcal{L}(\theta \mid \mathbf{Y}) &\propto \prod_{j \le J} f_j(y_j \mid \theta)\, \pi(\theta) \\
&\propto \prod_{j \le J} \binom{n_j}{y_j} (\theta)^{y_j} (1-\theta)^{n_j - y_j} \\
&\propto (\theta)^{\sum y_j} (1-\theta)^{\sum (n_j - y_j)} \\
&\sim \text{Be}(1 + y_+, \;\; 1 + n_+ - y_+),
\end{aligned}$$

Exactly the same as a single experiment with same total numbers

$$y_+ := \sum_{j \le J} y_j \qquad \text{and} \qquad (n_+ - y_+) = \sum_{j \le J} (n_j - y_j)$$

of successes & failures in $n_+ := \sum_{j \le J} n_j$ trials. Pretty unrealistic.

Uh oh...

But what if the experiments aren't "similar"?

In the Toy example, if the binomials are from clinical trials in different countries or even different medical centers, the success probabilities θ_j may vary.

Uh oh...

But what if the experiments aren't "similar"?

In the Toy example, if the binomials are from clinical trials in different countries or even different medical centers, the success probabilities θ_j may vary.

In the GRB example, we cannot expect multiple GRBs $Y_j(t)$ to all have the *same* parameter $\theta = \big(B, J, \vec{A}, \vec{t_0}, \vec{\tau}_1, \vec{\tau}_2\big)$— same number of pulses, same amplitudes, same shapes, etc.

Uh oh...

But what if the experiments aren't "similar"?

In the Toy example, if the binomials are from clinical trials in different countries or even different medical centers, the success probabilities θ_j may vary.

In the GRB example, we cannot expect multiple GRBs $Y_j(t)$ to all have the *same* parameter $\theta = \big(B, J, \vec{A}, \vec{t_0}, \vec{\tau}_1, \vec{\tau}_2\big)$— same number of pulses, same amplitudes, same shapes, etc.

In the Cosmic example, we will want to synthesize evidence from different telescopes, different nights, different spectral bands, perhaps GWs...

Uh oh...

But what if the experiments aren't "similar"?

In the Toy example, if the binomials are from clinical trials in different countries or even different medical centers, the success probabilities θ_j may vary.

In the GRB example, we cannot expect multiple GRBs $Y_j(t)$ to all have the *same* parameter $\theta = \big(B, J, \vec{A}, \vec{t_0}, \vec{\tau}_1, \vec{\tau}_2\big)$— same number of pulses, same amplitudes, same shapes, etc.

In the Cosmic example, we will want to synthesize evidence from different telescopes, different nights, different spectral bands, perhaps GWs...

We must be prepared for different observations, in complex examples; and different data sources, even in simple examples; to be **dis**similar.

Outline

$J > 1$ *Dis*similar Experiments

If we have $J \geq 1$ independent experiments, each generating evidence

$$Y_j \sim f_j(y \mid \theta_j)$$

for likelihood functions $\{f_j(y \mid \theta_j)\}$ that depend on the *different* uncertain parameters $\{\theta_j \in \Theta_j\}$, then the *vector* $\mathbf{Y} = (Y_1, \ldots, Y_J)$ has pdf

$$f(\vec{y} \mid \theta_1, \cdots, \theta_J) = \prod_{j \leq J} f_j(y_j \mid \theta_j);$$

how can we use this to derive a **synthesis** of evidence about whatever we care about?

$J > 1$ *Dis*similar Experiments (cont'd) I: Frequentist

A Frequentist solution:

$J > 1$ *Dis*similar Experiments (cont'd) I: Frequentist

A Frequentist solution:

I don't know a Frequentist solution to this problem.

$J > 1$ *Dis*similar Experiments (cont'd) II: Bayes

A Bayesian Solution:

- Identify just what it is that we care about from these experiments— Hubble constant H_0? Mean number λ_J of pulses in GRBs? Something else? Let's call it "ε".
- Identify a vector θ of whatever is **uncertain** and **common to two or more of** $\{\theta_j\}$, in the sense that the collection $\{\theta_j\}$ and ε are conditionally independent *a priori* given θ— so the conditional prior distribution (given θ) can be written as

$$\pi(d\theta_1 \cdots d\theta_J \, d\varepsilon \mid \theta) = \pi(d\varepsilon \mid \theta) \prod_{j \le J} \pi(d\theta_j \mid \theta)$$

This entails some thoughtful modeling and some compromises and approximations, in the hope of finding a **low dimensional** feature θ that "separates" ε and $\{\theta_j\}$, with **simple and tractable** distributions $\pi_j(d\theta_j \mid \theta)$.

$J > 1$ *Dis*similar Experiments (cont'd)

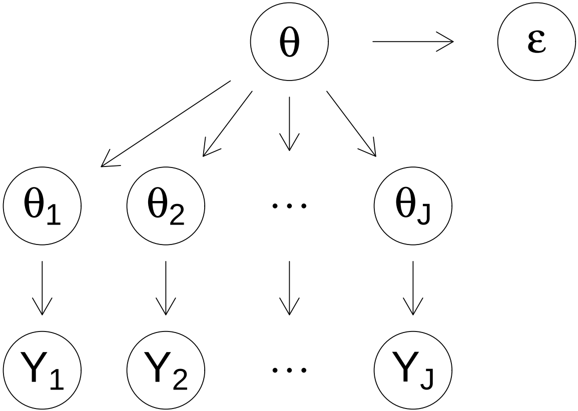
Now the posterior distribution for *everything* can be written

$$\pi(\varepsilon, \theta, \theta_1, \cdots, \theta_J \mid \mathbf{Y}) \propto \Big\{ \prod_{j \le J} f_j(y_j \mid \theta_j) \pi(d\theta_j \mid \theta) \Big\} \pi(d\varepsilon \mid \theta) \pi(d\theta)$$
$$= \Big\{ \prod_{j \le J} f_j(y_j \mid \theta_j) \pi(d\theta_j \mid \theta) \Big\} \pi(d\theta \mid \varepsilon) \pi(d\varepsilon)$$

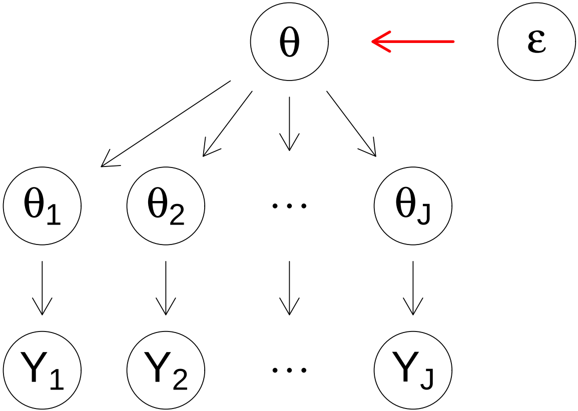
and so the "marginal likelihood for ε" (Berger, Liseo, Wolpert 1999) is available by dividing by $\pi(d\varepsilon)$ and integrating away everything else:

$$\mathcal{L}(\varepsilon) = \int_\Theta \prod_{j \le J} \Big\{ \int_{\Theta_j} f_j(y_j \mid \theta_j) \pi(d\theta_j \mid \theta) \Big\} \pi(d\theta \mid \varepsilon)$$
$$\pi(\varepsilon \mid \mathbf{Y}) \propto \mathcal{L}(\varepsilon)\ \pi(d\varepsilon).$$

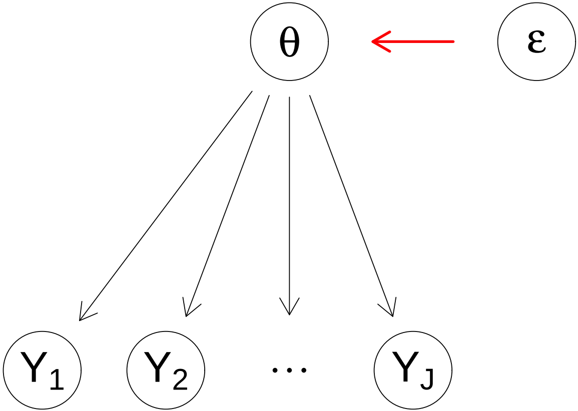
In pictures, as a DAG...



In pictures, as a DAG...



In pictures, as a DAG...



Summary

In summary: to synthesize evidence from disparate sources about a quantity of interest ε,

1. For each source $j \in \{1, \ldots, J\}$, find the **parameter space** Θ_j and the likelihood function $\Big\{ f_j(y_j \mid \theta_j) : \ \theta_j \in \Theta_j \Big\}$ governing Y_j;
2. Identify a small vector $\theta \in \Theta$ s.t. $\{\theta_j\}$ and ε are **conditionally indep**. given (*i.e.*, share no common uncertain features outside of) θ;
3. Identify **conditional prior** distributions $\pi_j(d\theta_j \mid \theta)$ and $\pi(\varepsilon \mid \theta)$ and a marginal prior $\pi(d\theta)$;
4. For each j, compute the "**adjusted likelihood function**" for θ:

$$\mathcal{L}_j(\theta) := \Big\{ \int_{\Theta_j} f_j(y_j \mid \theta_j) \pi(d\theta_j \mid \theta) \Big\}$$

5. Find the "**marginal likelihood function**" for ε as

$$\mathcal{L}(\varepsilon) := \int_{\Theta} \Big\{ \prod \mathcal{L}_j(\theta) \Big\} \pi(d\theta \mid \varepsilon)$$

6. Your choice— find $\hat{\varepsilon} := \operatorname{argmax}_{\varepsilon} \mathcal{L}(\varepsilon)$, plot and explore $\mathcal{L}(\varepsilon)$, or use it to find posterior probabilities and expectations.

The hardest bit:

2 Identify a small vector $\theta \in \Theta$ s.t. $\{\theta_j\}$ and ε are **conditionally indep**. given (*i.e.*, share no common uncertain features outside of) θ;

- Could always take $\theta = (\theta_1, \dots, \theta_J, \varepsilon)$, but that's too big making the "$d\theta$" integral unmanageable;
- Could hope to take $\theta = (\varepsilon)$, but that's too small making the conditional independence untennable;
- Need a Goldilocks solution.

Something that often works:

- Identify "Ideal" experiment whose low-dim parameter $\theta \in \Theta$ would completely determine $\varepsilon = \varepsilon(\theta)$, and
- Identify **key covariates** z_j in jth experiment and
- Function $\phi_j : \Theta \times \mathcal{Z} \to \Theta_j$ quantifying how un-ideal jth study is, s.t.
- $\theta_j = \phi_j(\theta, z_j)$

Then $\{\theta_j\}$ and ε are **c.i.**. given θ. Yay.

Examples:

- **2nd-hand smoke Ca risk** with evidence from 6 country groups;
- **volcano magma chamber** with seismic, magnetic, gravitational, acoustic, electrical probes;
- **populations of GRBs**;

Morals

1. In simple problems with plentiful data, Frequentist and Bayes methods give similar results;
2. In more complex problems with dicier data, Frequentist and Bayes methods both offer point estimates but Bayes methods offer more meaningful **uncertainty quantification**;
3. In **population-based** problems (like exoplanet searches), exchangeable hierarchical Bayes methods offer a principled way to discover both features of individual systems and of the population characterization; and, finally,

Morals

1. In simple problems with plentiful data, Frequentist and Bayes methods give similar results;
2. In more complex problems with dicier data, Frequentist and Bayes methods both offer point estimates but Bayes methods offer more meaningful **uncertainty quantification**;
3. In **population-based** problems (like exoplanet searches), exchangeable hierarchical Bayes methods offer a principled way to discover both features of individual systems and of the population characterization; and, finally,
4. In hard, complex, multi-source problems (e.g.: EM+GW, maybe DM), NON-exchangeable hierarchical Bayes methods are the best choice I know.

Thanks!

More details (references, this talk in .pdf, related work) are available on request from

`rlw@duke.edu`.

Thanks to Jessi, Tom, Jon, BIRS, SAMSI, NASA, and the NSF!