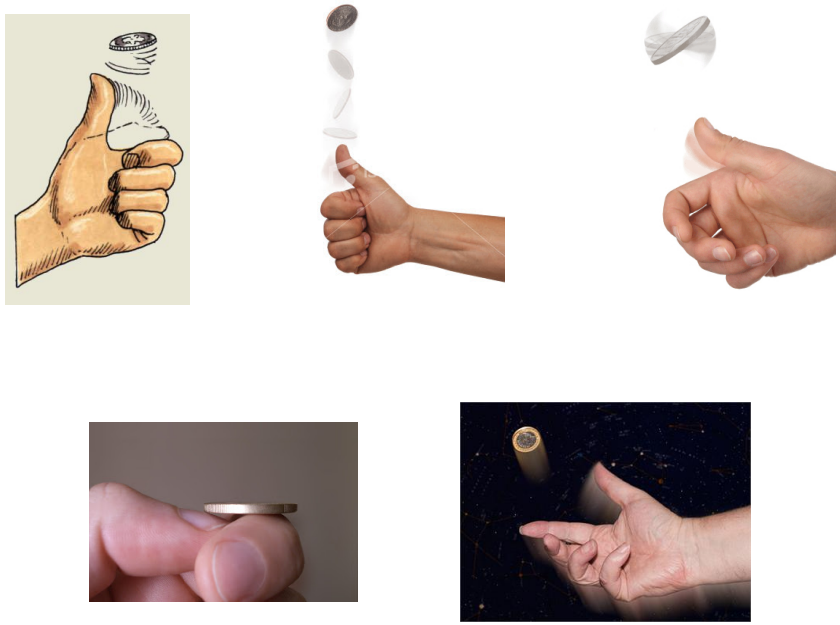
STSCI 4780/5780
Hierarchical/graphical models for measurement error: Density estimation

Tom Loredo, CCAPS & SDS, Cornell University

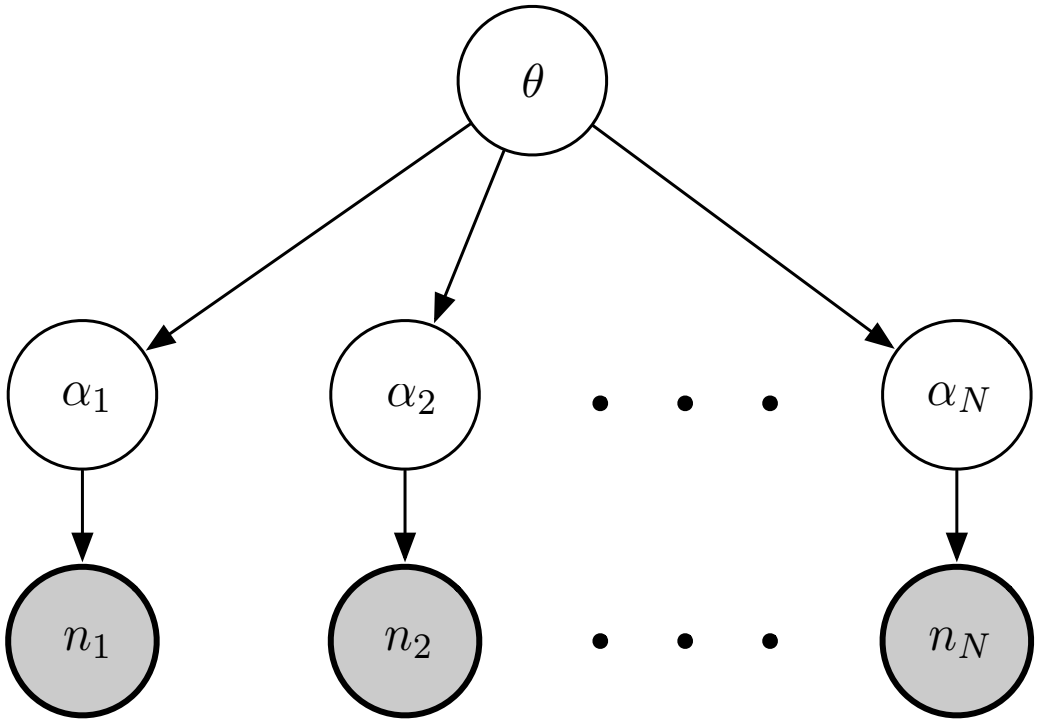
Recap: A population of coins/flippers



Each flipper+coin flips different number of times

- What can we learn about the underlying *population* of coins—the distribution of αs? (E.g., properties of the mint producing the coins, or skills/biases of the flippers)
- How does population membership affect inference for an *individual* coin's α?

Recap: Beta-binomial MLM



$$p(\theta, \{\alpha_i\}, \{n_i\}) = p(\theta) \prod_i p(\alpha_i|\theta)\ p(n_i|\alpha_i)$$

$$= \pi(\theta) \prod_i f(\alpha_i; \theta)\ \ell_i(\alpha_i)$$

Population model

Member or item likelihood

Terminology: θ are *hyperparameters*, $\pi(\theta)$ is the *hyperprior*

DAG describes the joint distribution over hyperparameters and member parameters (open nodes) and data (shaded nodes):

$$p(\theta, \{\alpha_i\}, \{n_i\}) = \pi(\theta) \prod_i p(\alpha_i|\theta)\ p(n_i|\alpha_i)$$

Describes "possible worlds"—choices of:

- Coin flip populations via θ (mean, scale of beta dist'n for pop'n)
- Individual coin parameters, $\alpha_i \sim p(\alpha_i|\theta)$ (IID draws from the beta)
- Number of heads for each coin, $n_i \sim p(n_i|\alpha_i)$ (binomial dist'ns)

The unknowns (open nodes) are $\theta, \{\alpha_i\}$; Bayes's theorem gives the posterior for them by conditioning on the knowns—it describes the subset of possible worlds that match the data, $\{n_i\}$:

$$\begin{aligned} p(\theta, \{\alpha_i\}|\{n_i\}) &= \frac{p(\theta, \{\alpha_i\}, \{n_i\})}{p(\{n_i\})} \\ &\propto \pi(\theta) \prod_i f(\alpha_i; \theta)\ \ell_i(\alpha_i) \end{aligned}$$

If we're primarily interested in the *member properties*, we can marginalize over θ,

$$p(\{\alpha_i\}|\{n_i\}) \propto \int \mathrm{d}\theta\, \pi(\theta) \prod_i f(\alpha_i;\theta)\, \ell_i(\alpha_i)$$

Each member/item has a marginal posterior, found my additionally marginalizing over *all but one* of the αs:

$$p(\alpha_j|\{n_i\}) \propto \int \mathrm{d}\theta\, \pi(\theta)\, f(\alpha_j;\theta)\, \ell_j(\alpha_j) \times \prod_{i\neq j} \int \mathrm{d}\alpha_i\, f(\alpha_i;\theta)\, \ell_i(\alpha_i)$$

$$= \ell_j(\alpha_j) \int \mathrm{d}\theta\, \pi(\theta)\, f(\alpha_j;\theta)\, \mathcal{M}_{\bar{j}}(\theta)$$

where the marginal likelihood function $\mathcal{M}_{\bar{j}}(\theta)$ captures what all of the other α_i measurements say about θ.

If we knew θ, then $f(\alpha_j;\theta)$ would serve as the prior for α_j.

But we *don't* know θ; the integral "mixes" the plausible $f(\alpha_j;\theta)$ distributions to construct a prior for α_j that accounts for θ uncertainty.

If we're primarily interested in the *population properties*, we can marginalize over *all* of the αs:

$$p(\theta|\{n_i\}) \propto \pi(\theta) \times \prod_i \int \mathrm{d}\alpha_i \; f(\alpha_i;\theta) \; \ell_i(\alpha_i)$$
$$= \pi(\theta) \; \mathcal{M}(\theta)$$

where the marginal likelihood function $\mathcal{M}(\theta)$ captures what all of the α_i measurements say about θ.

Note that

$$\mathcal{M}_{\bar{j}}(\theta) = \frac{\mathcal{M}(\theta)}{\int \mathrm{d}\alpha_j \; f(\alpha_j;\theta) \; \ell_j(\alpha_j)}$$

Empirical vs. hierarchical Bayes member estimates

From a fully Bayesian point of view, an individual member estimate should come from

$$p(\alpha_j|\{n_i\}) \propto \int \mathrm{d}\theta\, \pi(\theta)\, f(\alpha_j;\theta)\, \ell_j(\alpha_j) \times \prod_{i\neq j} \int \mathrm{d}\alpha_i\, f(\alpha_i;\theta)\, \ell_i(\alpha_i)$$

$$= \ell_j(\alpha_j) \int \mathrm{d}\theta\, \pi(\theta)\, f(\alpha_j;\theta)\, \mathcal{M}_{\bar{j}}(\theta)$$

E.g., we might report the marginal posterior means or modes as point estimates.

When the data are very informative about θ (e.g., a large pop'n), we expect $\mathcal{M}_{\bar{j}}(\theta) \approx \mathcal{M}(\theta)$ (any one member's data isn't contributing much to learning θ).

Additionally, we expect $\mathcal{M}(\theta)$ to concentrate around some value $\hat{\theta}$ (e.g., the maximum marginal likelihood estimate, MMLE).

Empirical Bayes (EB) estimation finds member estimates by using $f(\alpha_j; \hat{\theta})$ as the prior for α_j, with $\hat{\theta}$ found in a clever way (MMLE, method of moments...).

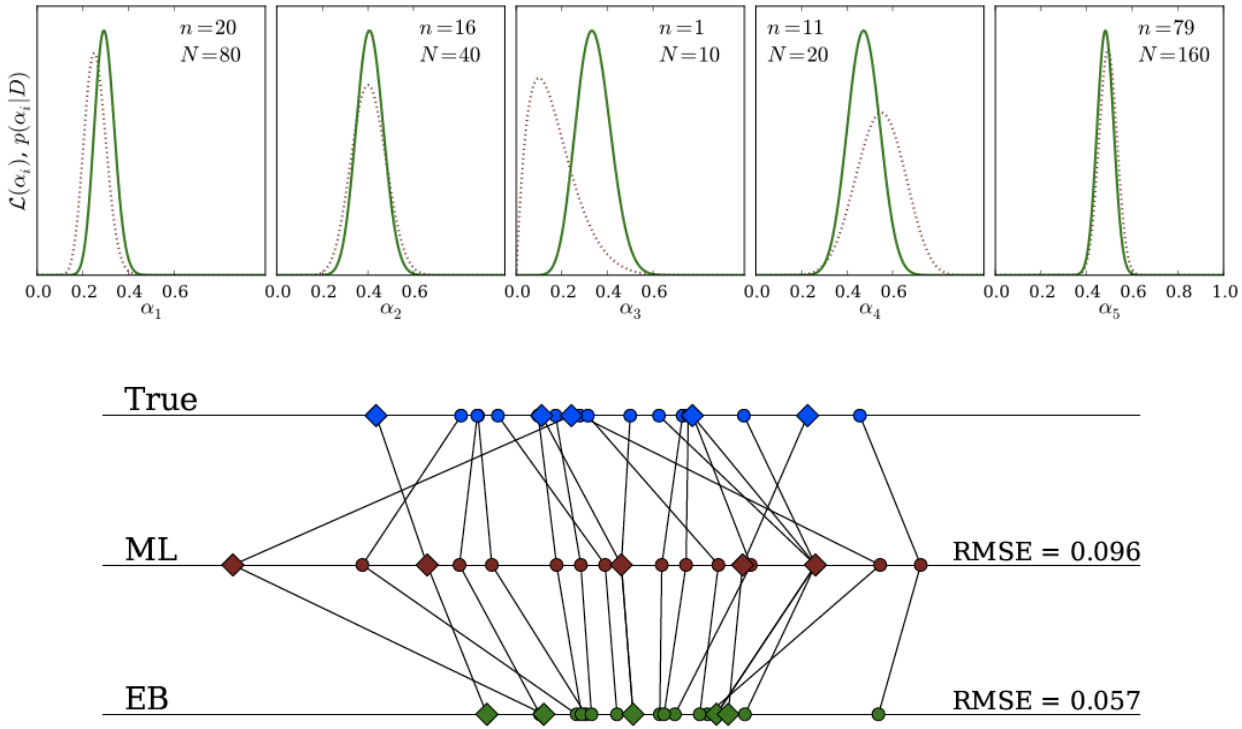
Two different viewpoints:

- An approximation to hierarchical Bayes (HB)
- A frequentist procedure that estimates a prior from the data (in multi-case settings)

HB typically has better frequentist properties than EB because it accounts for hyperparameter uncertainty (e.g., EB intervals tend to undercover—they're too optimistic)

Note the *marginalize-vs-optimize dichotomy* (again)...

Lower level (member/item) estimates



Bayesian outlook

- Marginal posteriors are *narrower* than likelihoods
- Point estimates tend to be closer to true values than MLEs (averaged across the population)
- Joint distribution for $\{\alpha_i\}$ is *dependent*

Frequentist outlook

- Point estimates are biased
- Reduced variance $\rightarrow$ estimates are closer to truth on average (lower MSE in repeated sampling)
- Bias for one member estimate depends on data for all other members

Lingo

- Estimates *shrink* toward prior/population mean
- Estimates "muster and *borrow strength*" across population (Tukey's phrase); increases accuracy and precision of estimates
- Efron* describes shrinkage as a consequence of accounting for *indirect evidence*

*Bradley Efron (2010): "The Future of Indirect Evidence"

Estimating the population distribution

If we knew precise α_i values, we could estimate the population distribution by straightforward inference of θ:

$$p(\theta|\{\alpha_i\}) \propto \pi(\theta) \prod_i f(\alpha_i; \theta)$$

We instead know $\{n_i\}$, giving us imprecise information about the αs; Bayesian analysis tells us to use

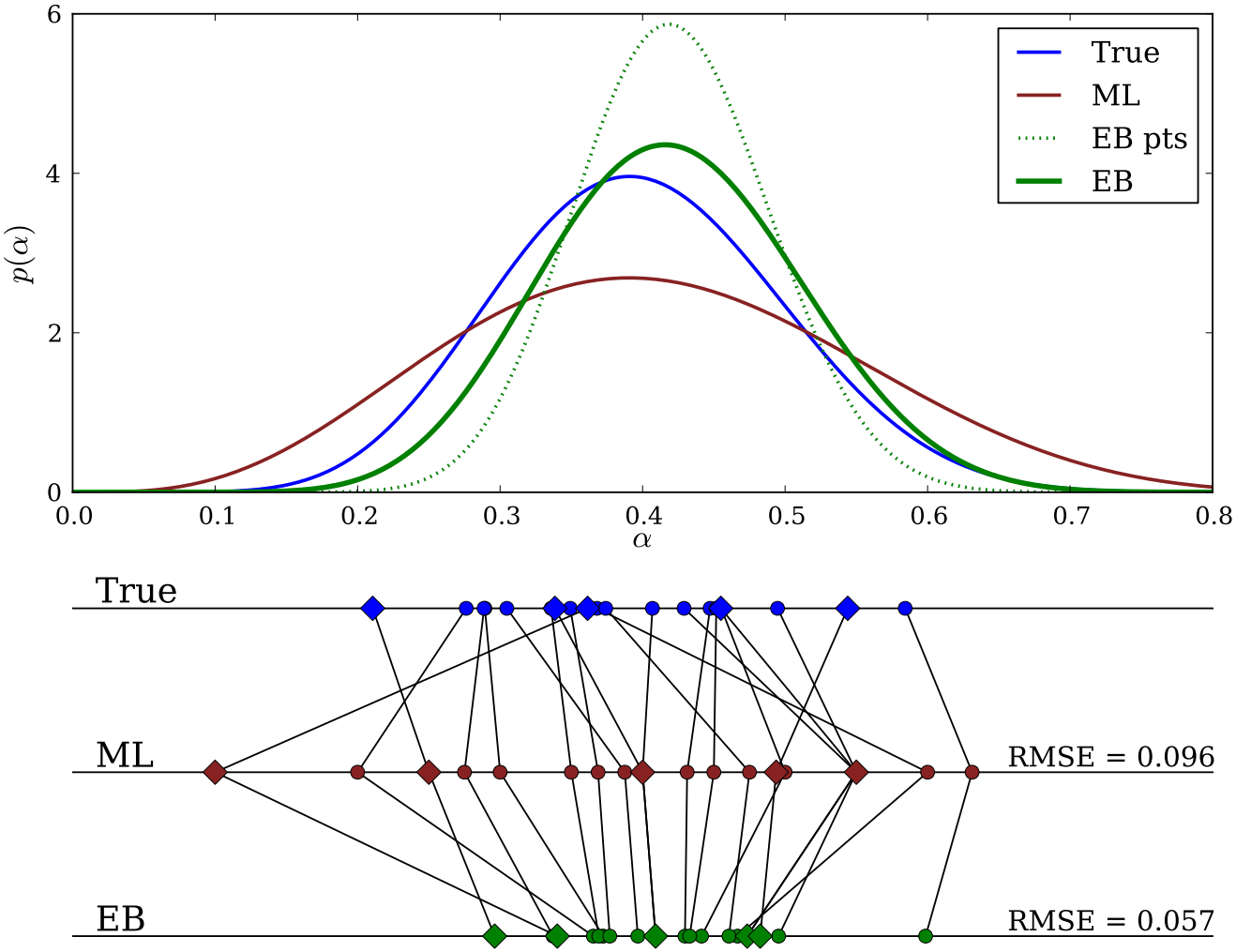
$$p(\theta|\{n_i\}) \propto \pi(\theta)\, \mathcal{M}(\theta)$$

Compare:

- Using MLEs $\hat{\alpha}_i = n_i/N_i$ (frequencies) as if they were precise
- Using shrunken EB member estimates as if they were precise
- The MMLE estimate, $\hat{\theta} = \arg\max \mathcal{M}(\theta)$

Plot below shows *typical* behavior—for most datasets, MMLE is closest to truth

- Using MLEs $\hat{\alpha}_i = n_i/N_i$ (frequencies) as if they were precise (red)
- Using shrunken EB member estimates as if they were precise (green dots)
- The MMLE estimate, $\hat{\theta} = \arg\max \mathcal{M}(\theta)$ (green)



Beware of point estimates!

"Shrunken" member estimates provide improved & reliable estimate for population member properties

But they are *under-dispersed* in comparison to the true values $\rightarrow$ not optimal for estimating *population* properties*

No point estimates of member properties are optimal for all tasks!

Don't think of shrinkage estimators as "correcting the data," with the corrected data then usable in place of the true values for *any* data analysis purpose

Adjusted point estimates account for uncertainties in a particular way, for a particular task (e.g., optimizing pop'n-averaged error of member params), but otherwise don't communicate uncertainties

Accurate "downstream" inference ideally should fully use the member likelihood functions, not point estimates

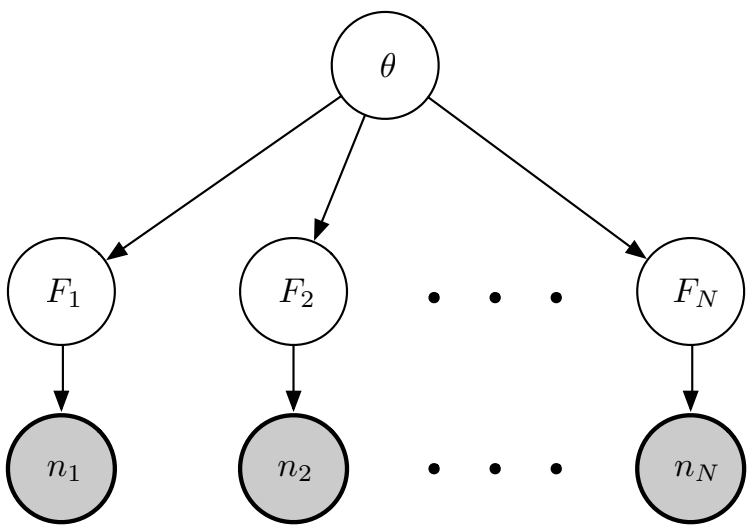
*Louis (1984) finds point estimates optimal for three different tasks. Eddington noted this problem in 1940!

Another conjugate MLM: Gamma-Poisson

Goal: Learn a distribution of event rates from event counts
a.k.a.: Estimating a *number-size distribution*

Examples: learn infection rates from area-specific disease counts; learn a star or galaxy brightness dist'n from photon counts

Qualitative



Quantitative

Population parameters

$$\theta = (\alpha, s) \text{ or } (\mu, \sigma)$$

$$\pi(\theta) = \mathrm{Flat}(\mu, \sigma)$$

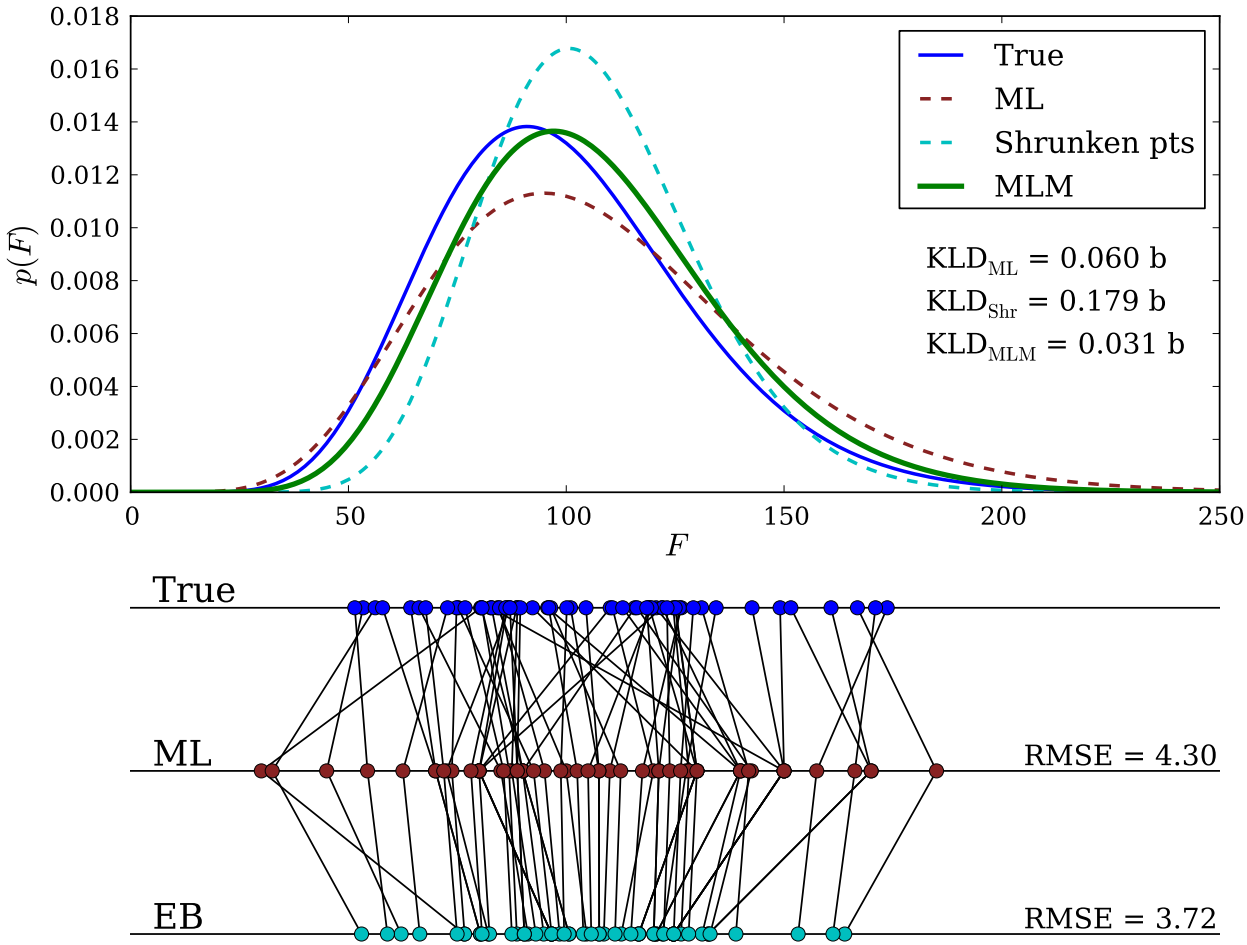
Source properties

$$p(F_i|\theta) = \mathrm{Gamma}(F_i|\theta)$$

Observed data

$$p(n_i|F_i) = \mathrm{Pois}(n_i|\epsilon_i F_i)$$

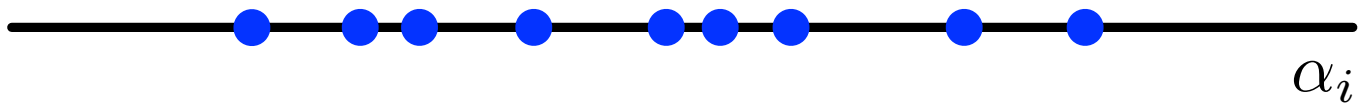
Gamma-Poisson population and member estimates



Simulations: $N = 60$ sources from gamma with $\langle F \rangle = 100$ and $\sigma_F = 30$; exposures spanning dynamic range of $\times 16$

Measurement error perspective

If the data provided *precise* $\{\alpha_i\}$ values (coin measurements, flip physics), we could easily model them as points drawn from a (beta) population PDF with params θ:



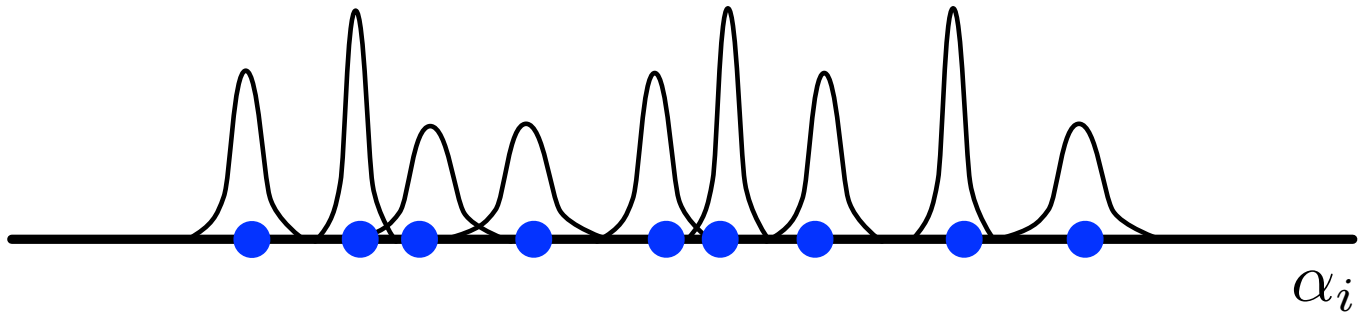
$$D = \{\alpha_i\}$$

$$p(D|\theta) = \prod_i p(\alpha_i|\theta)$$

$$= \prod_i \mathrm{Beta}(\alpha_i|\theta)$$

(A *binomial point process*)

Here the finite number of flips provide *noisy measurements of each* α_i, described by the member likelihood functions $\ell_i(\alpha_i)$;



$$D = \{n_i\}$$

$$\begin{aligned} p(D|\theta) &= \prod_i \int d\alpha_i \; p(D, \{\alpha_i\}|\theta) \\ &= \prod_i \int d\alpha_i \; p(\alpha_i|\theta)\, p(n_i|\alpha_i) \\ &= \prod_i \int d\alpha_i \; \mathrm{Beta}(\alpha_i|\theta)\; \mathrm{Binom}(n_i|\alpha_i) \end{aligned}$$

This is a prototype for *measurement error problems*

Agenda

- (Joint) Density estimation with measurement error (density deconvolution)
- Regression with measurement error (next lecture)

Agenda

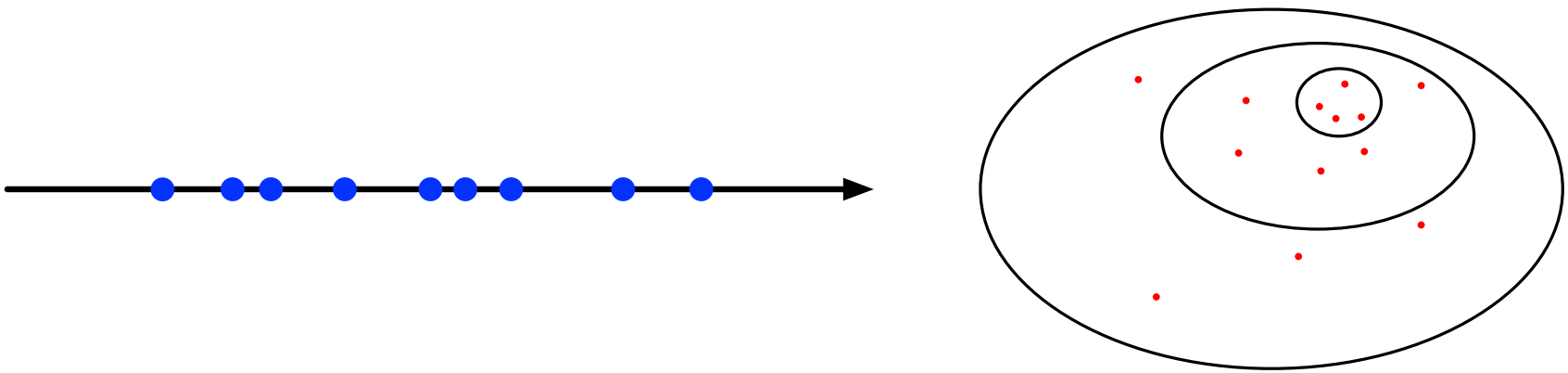
1. **Density estimation with measurement error**

Accounting For Measurement Error

Introduce latent/hidden/incidental parameters

Suppose $f(x|\theta)$ is a distribution for an observable, x (scalar or vector, $\vec{x} = (x, y, \ldots)$)

Scalar x **Vector** $\vec{x}$



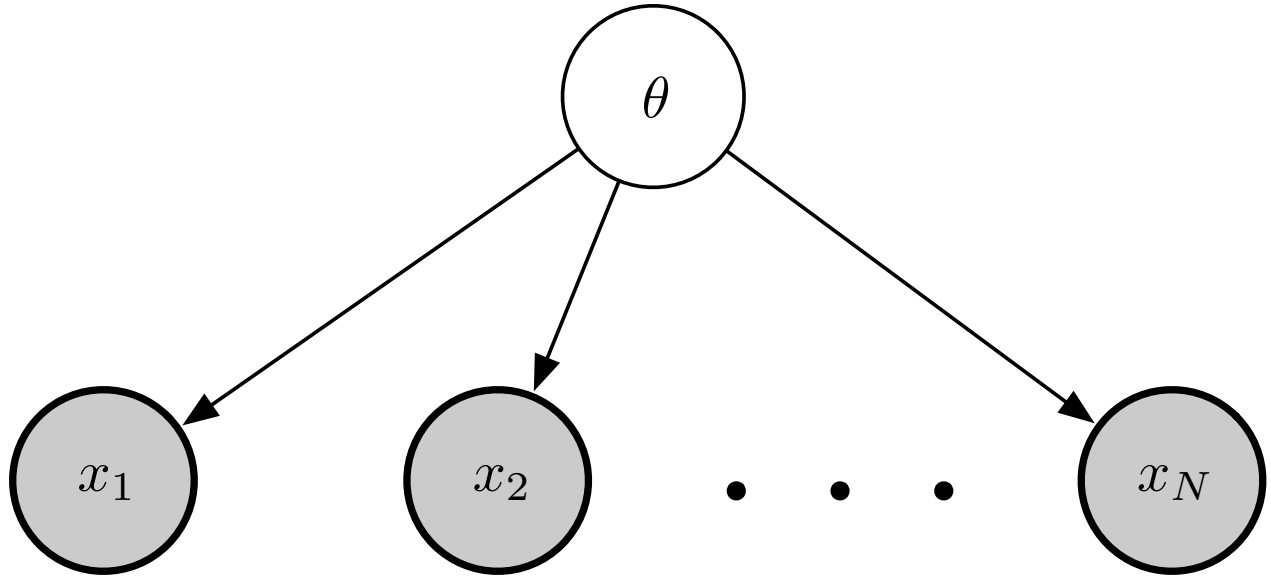
From N *precisely measured* samples, $\{x_i\}$, we can infer θ using

$$\mathcal{L}(\theta) \equiv p(\{x_i\}|\theta) = \prod_i f(x_i;\theta)$$

$$p(\theta|\{x_i\}) \propto p(\theta)\mathcal{L}(\theta) = p(\theta, \{x_i\})$$

A *binomial point process* model (Poisson if N is random)

Graphical representation



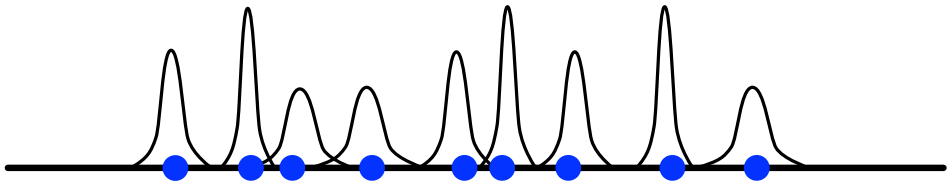
Joint distribution:

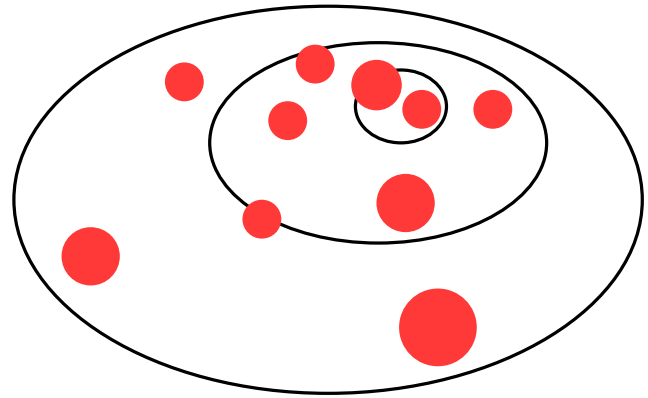
$$p(\theta, \{x_i\}) \;=\; p(\theta)\, p(\{x_i\}|\theta) \;=\; p(\theta) \prod_i f(x_i; \theta)$$

Posterior from BT:

$$p(\theta|\{x_i\}) = \frac{p(\theta, \{x_i\})}{p(\{x_i\})}$$

But what if the x data are *noisy*, $D_i = x_i + \epsilon_i$?





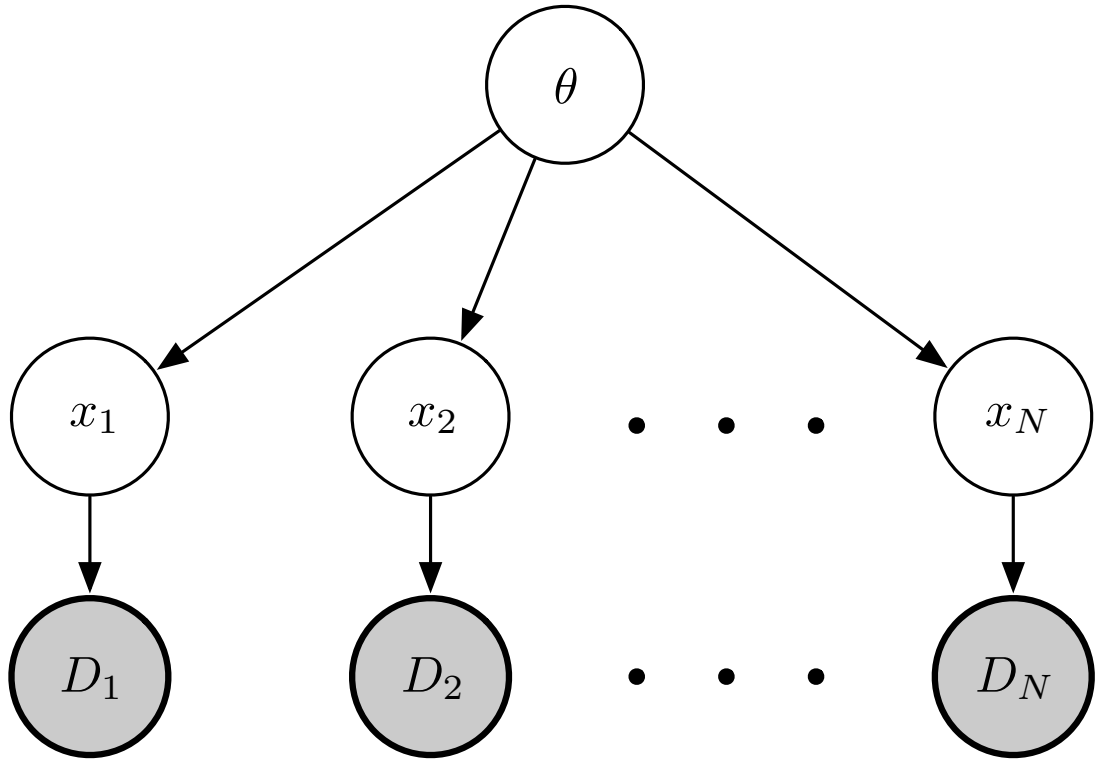
$\{x_i\}$ are now *uncertain (latent) parameters*
We should somehow incorporate *member likelihoods*
$\ell_i(x_i) = p(D_i|x_i)$ quantifying the uncertainties in the measurements:

$$\begin{aligned} p(\theta, \{x_i\}, \{D_i\}) &= p(\theta)\, p(\{x_i\}|\theta)\, p(\{D_i\}|\{x_i\}) \\ &= p(\theta) \prod_i f(x_i;\theta)\, \ell_i(x_i) \end{aligned}$$

Marginalize over $\{x_i\}$ to summarize inferences for θ.

Marginalize over θ to summarize inferences for $\{x_i\}$.

Graphical representation



$$
\begin{aligned}
p(\theta, \{x_i\}, \{D_i\}) &= p(\theta)\; p(\{x_i\}|\theta)\; p(\{D_i\}|\{x_i\}) \\
&= p(\theta) \prod_i f(x_i;\theta)\; p(D_i|x_i) \;=\; p(\theta) \prod_i f(x_i;\theta)\; \ell_i(x_i)
\end{aligned}
$$

(sometimes called a "two-level MLM" or "two-level hierarchical model")

Joint for everything

$$p(\theta, \{x_i\}, \{D_i\}) = p(\theta) \prod_i f(x_i; \theta)\, \ell_i(x_i)$$

Population-level inference

Condition on data, marginalize over latent member params:

$$p(\theta|\{D_i\}) \propto p(\theta) \prod_{i=1}^{N} \int \mathrm{d}x_i\, f(x_i; \theta)\, \ell_i(x_i)$$

Conditional independence $\rightarrow$ the $O(N)$–D integral over $\{x_i\}$ is the *product of N* **independent**, **low-D** *integrals*

Member-level inference

Condition on data, marginalize over population dist'n params:

$$p(x_j|\{D_i\}) \propto \int \mathrm{d}\theta\; p(\theta)\, f(x_j; \theta)\, \ell_j(x_j) \times \prod_{i \neq j} \int \mathrm{d}x_i\; f(x_i; \theta)\, \ell_i(x_i)$$

Algorithms

Consider the posterior PDF for θ and $\{\alpha_i\}$ in the beta-binomial MLM:

$$p(\theta, \{\alpha_i\}|\{n_i\}) \propto \pi(\theta) \prod_{i=1}^{N_{\mathrm{mem}}} \mathrm{Beta}(\alpha_i|\theta)\, \mathrm{Binom}(n_i|\alpha_i)$$

For each member, the $\mathrm{Beta} \times \mathrm{Binom}$ factor is $\propto$ a beta distribution for α_i; but as a function of θ (e.g., (a, b) or (μ, σ)) it is **not simple**

The full posterior has a product of N_{mem} such factors specifying its θ dependences $\Rightarrow$ *even using a conjugate model for the lower levels, the overall model is typically analytically intractable*

Posterior sampling over the joint population/member parameter space is challenging; `Stan` does it **all-at-once** using *Hamiltonian Monte Carlo* (HMC)

Two approaches exploit *conditional independence of member-level parameters*

Metropolis-within-Gibbs (MWG) algorithm

Block the full parameter space:

- Block of m population parameters, θ
- N blocks of (latent) member parameters, x_i

Get posterior samples by iterating back and forth between:

- m-D Metropolis-Hastings sampling of θ from $p(\theta|\{x_i\}, D)$

  This requires a problem-specific proposal distribution

- N *independent* samples of x_i from the conditional $p(x_i|\theta, D_i)$

  This can often exploit conjugate structure

  E.g., Beta-binomial: $\alpha_i \sim \mathrm{Beta}(\alpha_i|\theta)\ \mathrm{Binom}(n_i|\alpha_i)$, which is just a Beta for α_i

MWG explicitly displays the *feedback between population and member inference*

Member marginalization

$$p(\theta|\{D_i\}) \propto p(\theta) \prod_{i=1}^{N} \int \mathrm{d}x_i \, f(x_i;\theta) \, \ell_i(x_i)$$

- Analytically or numerically integrate over $\{x_i\} \rightarrow$ explore the (greatly!) reduced-dimension marginal for θ via MCMC $\rightarrow \{\theta_i\} \sim p(\theta|D)$
- If x_i are of interest, sample them from their conditionals, conditioned on θ_i:
  - Pick a θ from $\{\theta_i\}$
  - Draw $\{x_i\}$ by *independent* sampling from their conditionals (give θ)
  - Iterate

GPUs can accelerate this for application to large datasets

Only useful for low-dimensional latent parameters x_i

Seldom used in *B* literature; frequently used in *F* "random effects" literature

Takeaways

- *Density deconvolution* — Estimating a PDF when the "points" are measured with error
- Hierarchical/multilevel models treat density estimation with measurement error via *latent parameters* — the uncertain true values underlying noisy measurements
- *Hierarchical Bayes* marginalizes over everything; *empirical Bayes* optimizes over the population-level (hyper)parameters (to estimate the item/member params)
- Computational methods: Posterior sampling all-at-once (Stan) or by blocking (Metropolis-within-Gibbs); member marginalization to focus on population parameters