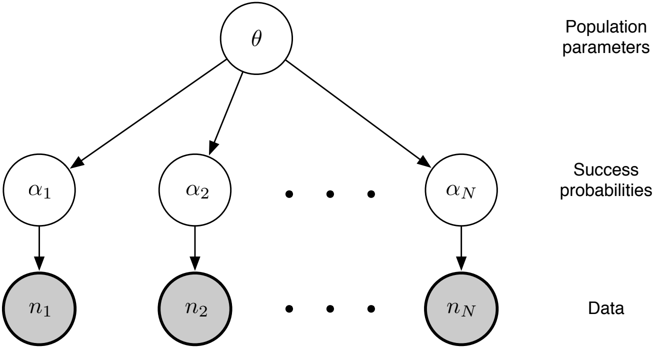
STSCI 4780
Hierarchical/graphical models for measurement error

Tom Loredo, CCAPS & SDS, Cornell University

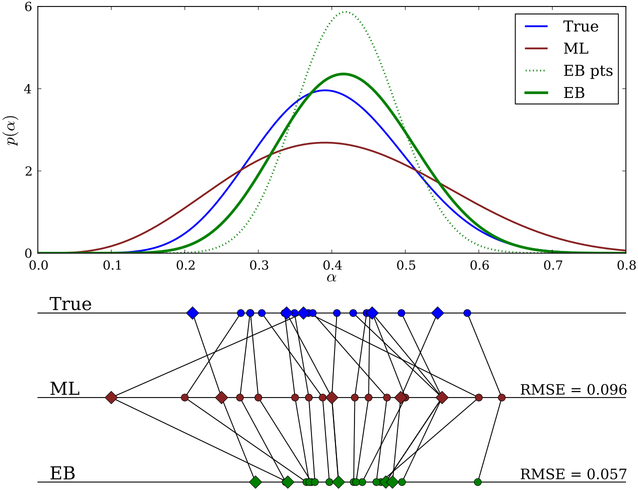
Recap: Beta-binomial MLM



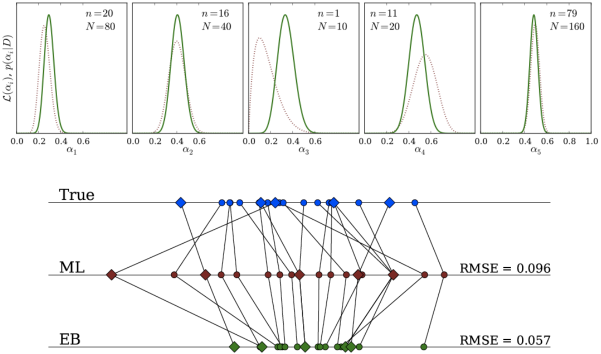
$$
\begin{aligned}
p(\theta, \{\alpha_i\}, \{n_i\}) &= \pi(\theta) \prod_i p(\alpha_i|\theta)\; p(n_i|\alpha_i) \\
&= \pi(\theta) \prod_i p(\alpha_i|\theta)\; \ell_i(\alpha_i)
\end{aligned}
$$

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

Population and member estimates



Lower level 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"

Competing data analysis goals

"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 good for all tasks!

We should view population data tables/catalogs as providing *descriptions of member likelihood functions*, not "estimates with errors"

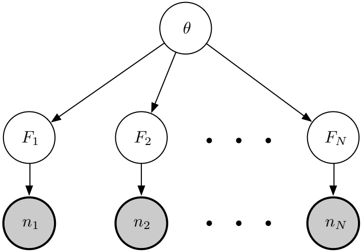
*Louis (1984); Eddington noted this in 1940!

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

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

Population parameters

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

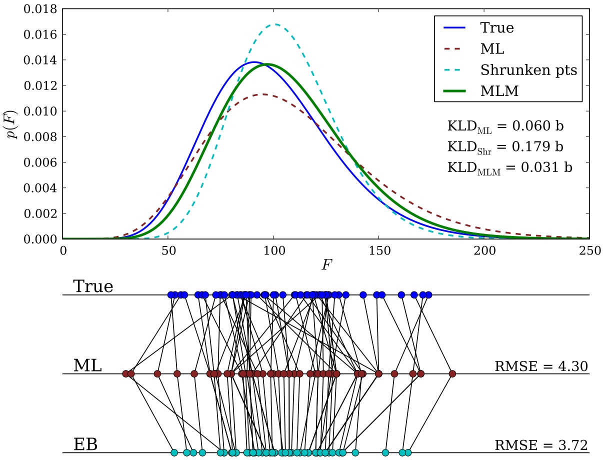
Source properties

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

Observed data

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

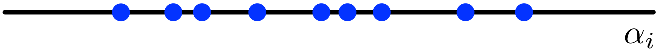
Gamma-Poisson population and member estimates



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:

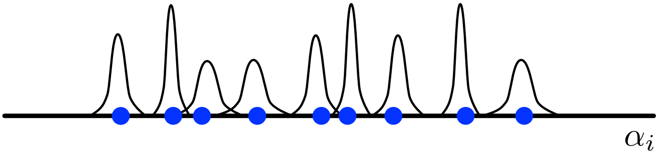


$$D = \{\alpha_i\}$$

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

(A *binomial point process*)

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



$$D = \{n_i\}$$

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

This is a prototype for *measurement error problems*

Agenda

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

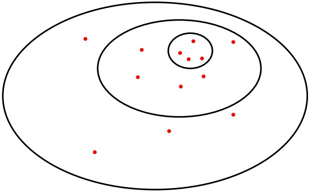
Agenda

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, $\vec{x} = (x, y, \ldots)$)



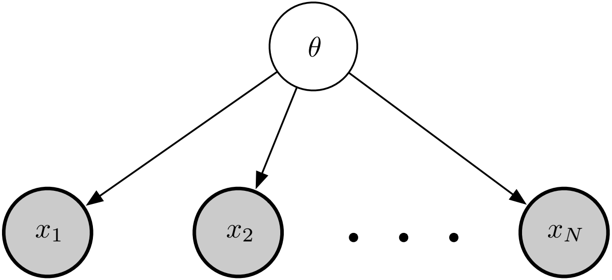
From N precisely measured samples, $\{x_i\}$, we can infer θ from

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

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

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

Graphical representation



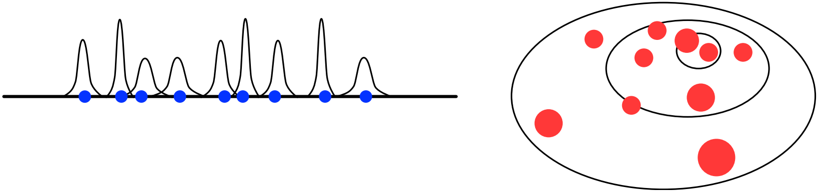
Joint distribution:

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

Posterior from BT:

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

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



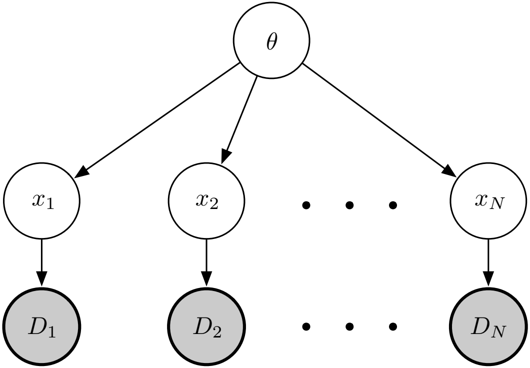
$\{x_i\}$ are now *uncertain (latent) parameters*
We should somehow incorporate $\ell_i(x_i) = p(D_i|x_i)$:

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

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:

$$p(x_j|\{D_i\}) \propto \int \mathrm{d}\theta\; p(\theta)\, f(x_1|\theta)\, \ell_1(x_1) \times \prod_{i \neq j} \int \mathrm{d}x_i\; f(x_i|\theta)\, \ell_i(x_i)$$

Key point: *Maximizing over x_i (treating best-fit values as if they were precise) and integrating over x_i can give very different results!*

To estimate x_1:

$$
\begin{aligned}
p(x_1|\{D_2,\ldots\}) &\propto \int \mathrm{d}\theta\; p(\theta)\, f(x_1|\theta)\, \ell_1(x_1) \times \prod_{i=2}^{N} \int \mathrm{d}x_i\; f(x_i|\theta)\, \ell_i(x_i) \\
&= \ell_1(x_1) \int \mathrm{d}\theta\; \left[p(\theta)\, \mathcal{L}_{\mathrm{marg},\check{1}}(\theta)\right]\; f(x_1|\theta) \\
&\qquad \text{with } \mathcal{L}_{\mathrm{marg},\check{1}}(\theta) \equiv \prod_{i=2}^{N} \int \mathrm{d}x_i\; f(x_i|\theta)\, \ell_i(x_i) \\
&\approx \ell_1(x_1)\, f(x_1|\hat{\theta}_{\check{1}}), \quad \text{(using a plug-in approx'n for } \theta\text{)}
\end{aligned}
$$

with $\hat{\theta}_{\check{1}} = \arg\max \mathcal{L}_{\mathrm{marg},\check{1}}(\theta)$ determined by the remaining data

$f(x_1|\hat{\theta})$ behaves like a prior that shifts the x_1 estimate away from the peak of $\ell_1(x_i)$; learning it from the data can lead to *shrinkage*

Algorithms

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

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

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 $N_{\rm mem}$ such factors specifying its θ dependences $\Rightarrow$ *even for 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*

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 $\{x_i\} \to$ explore the reduced-dimension marginal for θ via MCMC $\to \{\theta_i\} \sim p(\theta|D)$
- 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

Metropolis-within-Gibbs 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*

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 parameters (to estimate the item/member params)
- Computational methods: Monolithic (Stan), member marginalization, Metropolis-within-Gibbs