Hierarchical Bayesian modeling

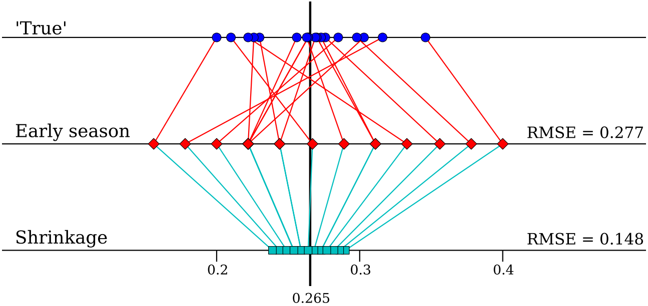
Tom Loredo
Cornell Center for Astrophysics and Planetary Science
http://www.astro.cornell.edu/staff/loredo/bayes/

SAMSI ASTRO — 19 Oct 2016

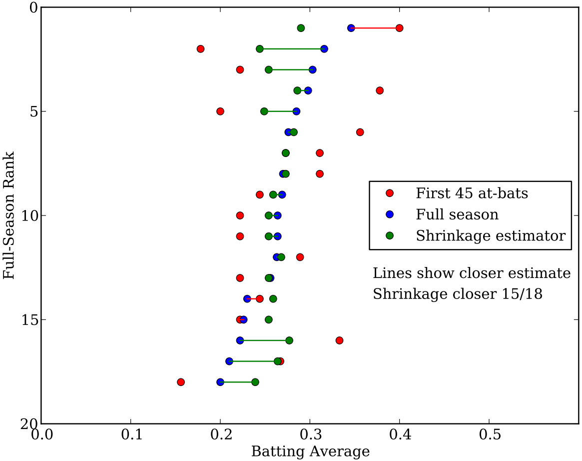
1970 baseball averages

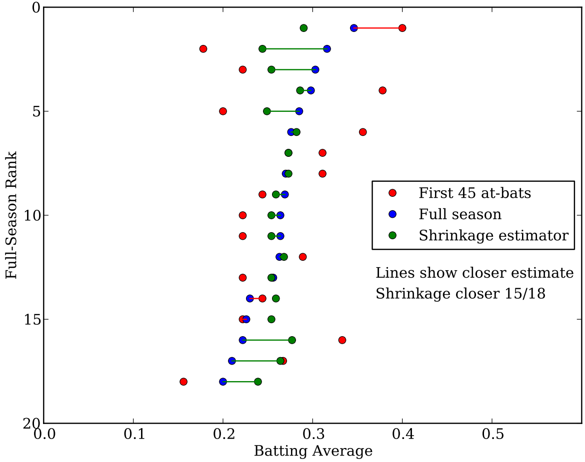
Efron & Morris looked at batting averages of baseball players who had $N = 45$ at-bats in May 1970 — 'large' N & includes Roberto Clemente (outlier!)

Red $= n/N$ maximum likelihood estimates of true averages
Blue $=$ Remainder of season, $N_{\mathrm{rmdr}} \approx 9N$



Cyan $=$ James-Stein estimator: nonlinear, correlated, biased
But *better*!


First 45 at-bats
Full season
Shrinkage estimator
Lines show closer estimate
Shrinkage closer 15/18
Full-Season Rank
Batting Average
0
5
10
15
20
0.0
0.1
0.2
0.3
0.4
0.5



Theorem (independent Gaussian setting): In dimension $\gtrsim 3$, shrinkage estimators always beat independent MLEs in terms of expected RMS error

"The single most striking result of post-World War II statistical theory"
— Brad Efron

All 18 players are *humans playing baseball*—they are members of a population, not arbitrary, unrelated binomial random number generators!

In the absence of data about player i, we may use the performance of the other players to guide a guess about that player's performance—they provide *indirect evidence* (Efron) about player i

But information that is relevant in the absence of data for i remains relevant when we additionally obtain that data; shrinkage estimators account for this

There is "mustering and *borrowing of strength*" (Tukey) across the population

Hierarchical Bayesian modeling is the most flexible framework for generalizing this lesson; *empirical Bayes* is an approximate version with a straightforward frequentist interpretation

Agenda

1. **Basic Bayes recap**
2. **Key idea in a nutshell**
3. **Going deeper**
   - Joint distributions and DAGs
   - Conditional dependence/indepence
   - Example: Binomial prediction
   - Beta-binomial model
   - Point estimation and shrinkage
   - Gamma-Poisson model & Stan
   - Algorithms

Bayesian inference in one slide

Probability as generalized logic

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

Bayes's theorem

$$p(\text{Hypothesis} \mid \text{Data}) \propto p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis})$$

Data *change* the support for a hypothesis $\propto$ ability of hypothesis to *predict* the data

Law of total probability

$$p(\text{Hypothes}\underline{\textbf{es}} \mid \text{Data}) = \sum p(\text{Hypothes}\underline{\textbf{is}} \mid \text{Data})$$

The support for a *compound/composite* hypothesis must account for all the ways it could be true

Bayes's theorem

$\mathcal{C}$ = context, initial set of premises

Consider $P(H_i, D_{\rm obs}|\mathcal{C})$ using the product rule:

$$\begin{aligned} P(H_i, D_{\rm obs}|\mathcal{C}) &= P(H_i|\mathcal{C})\, P(D_{\rm obs}|H_i, \mathcal{C}) \\ &= P(D_{\rm obs}|\mathcal{C})\, P(H_i|D_{\rm obs}, \mathcal{C}) \end{aligned}$$

Solve for the *posterior probability* (expands the premises!):

$$P(H_i|D_{\rm obs}, \mathcal{C}) = P(H_i|\mathcal{C})\, \frac{P(D_{\rm obs}|H_i, \mathcal{C})}{P(D_{\rm obs}|\mathcal{C})}$$

Theorem holds for any propositions, but for hypotheses & data the factors have names:

posterior $\propto$ *prior* $\times$ *likelihood*

norm. const. $P(D_{\rm obs}|\mathcal{C})$ = *prior predictive*

Law of Total Probability (LTP)

Consider exclusive, exhaustive $\{B_i\}$ ($\mathcal{C}$ asserts one of them must be true),

$$\begin{aligned}\sum_i P(A, B_i|\mathcal{C}) &= \sum_i P(B_i|A,\mathcal{C})P(A|\mathcal{C}) = P(A|\mathcal{C}) \\ &= \sum_i P(B_i|\mathcal{C})P(A|B_i,\mathcal{C})\end{aligned}$$

If we do not see how to get $P(A|\mathcal{P})$ directly, we can find a set $\{B_i\}$ and use it as a "basis"—*extend the conversation*:

$$P(A|\mathcal{C}) = \sum_i P(B_i|\mathcal{C})P(A|B_i,\mathcal{C})$$

If our problem already has B_i in it, we can use LTP to get $P(A|\mathcal{C})$ from the joint probabilities—*marginalization*:

$$P(A|\mathcal{C}) = \sum_i P(A, B_i|\mathcal{C})$$

Example: Take $A = D_{\rm obs}$, $B_i = H_i$; then

$$\begin{aligned} P(D_{\rm obs}|\mathcal{C}) &= \sum_i P(D_{\rm obs}, H_i|\mathcal{C}) \\ &= \sum_i P(H_i|\mathcal{C}) P(D_{\rm obs}|H_i, \mathcal{C}) \end{aligned}$$

prior predictive for $D_{\rm obs}$ = Average likelihood for H_i
(a.k.a. *marginal likelihood*)

Parameter Estimation

Problem statement

$\mathcal{C}$ = Model M with parameters θ (+ any add'l info)

H_i = statements about θ; e.g. "$\theta \in [2.5, 3.5]$," or "$\theta > 0$"

Probability for any such statement can be found using a *probability density function* (PDF) for θ:

$$\begin{aligned} P(\theta \in [\theta, \theta + d\theta] | \cdots) &= f(\theta) d\theta \\ &= p(\theta | \cdots) d\theta \end{aligned}$$

Posterior probability density

$$p(\theta | D, M) = \frac{p(\theta | M) \, \mathcal{L}(\theta)}{\int d\theta \; p(\theta | M) \, \mathcal{L}(\theta)}$$

Summaries of posterior

- "Best fit" values:
  - *Mode*, $\hat{\theta}$, maximizes $p(\theta|D, M)$
  - *Posterior mean*, $\langle\theta\rangle = \int d\theta\, \theta\, p(\theta|D, M)$
- Uncertainties:
  - *Credible region* Δ of probability C:
    $C = P(\theta \in \Delta|D, M) = \int_\Delta d\theta\, p(\theta|D, M)$
    Highest Posterior Density (HPD) region has $p(\theta|D, M)$ higher inside than outside
  - Posterior standard deviation, variance, covariances
- Marginal distributions
  - Interesting parameters ϕ, nuisance parameters η
  - *Marginal dist'n* for ϕ: $p(\phi|D, M) = \int d\eta\, p(\phi, \eta|D, M)$

Many Roles for Marginalization

Eliminate nuisance parameters

$$p(\phi|D, M) = \int d\eta \; p(\phi, \eta|D, M)$$

Propagate uncertainty

Model has parameters θ; what can we infer about $F = f(\theta)$?

$$\begin{aligned} p(F|D, M) &= \int d\theta \; p(F, \theta|D, M) = \int d\theta \; p(\theta|D, M)\, p(F|\theta, M) \\ &= \int d\theta \; p(\theta|D, M)\, \delta[F - f(\theta)] \qquad [\textit{single-valued case}] \end{aligned}$$

Prediction

Given a model with parameters θ and present data D, predict future data D' (e.g., for *experimental design*):

$$p(D'|D, M) = \int d\theta \; p(D', \theta|D, M) = \int d\theta \; p(\theta|D, M)\, p(D'|\theta, M)$$

Model comparison

Marginal likelihood for model M_i:

$$Z_i \equiv p(D|M_i) = \int d\theta_i \; p(\theta_i|M) \, \mathcal{L}_i(\theta_i)$$

Bayes factor $B_{ij} \equiv Z_i/Z_j$
Can write $Z_i = \mathcal{L}_i(\hat{\theta}_i) \cdot \Omega_i$ with Ockham factor
$\Omega_i \approx \delta\theta/\Delta\theta =$ (posterior volume)/(prior volume)

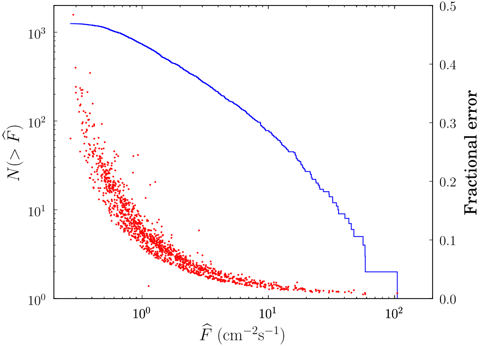
Hierarchical modeling, aka...

- Graphical models — Hierarchical and other structures
- Multilevel models — In regression, linear model settings
- Bayesian networks (Bayes nets) — In AI/ML settings

Agenda

Motivation: Measurement error in surveys

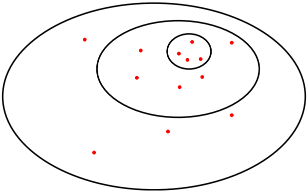
BATSE GRB peak flux estimates



- *Selection effects* (truncation, censoring) — *obvious* (usually)
  Typically treated by "correcting" data
  Most sophisticated: product-limit estimators
- *"Scatter" effects* (measurement error, etc.) — *insidious*
  Typically ignored (average out??? — *No!*)

Accounting For Measurement Error

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

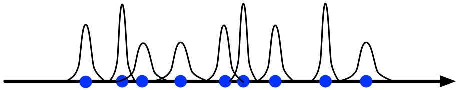


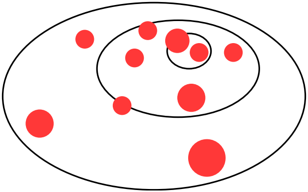
From N precisely measured samples, $\{x_i\}$, we can infer θ from

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

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

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





$\{x_i\}$ are now *uncertain (latent/hidden/incidental) parameters*

We should somehow incorporate $\ell_i(x_i) = p(D_i|x_i)$

The joint PDF for *everything* is

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

The conditional (posterior) PDF for the unknowns is

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

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

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

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

Key point: *Maximizing over x_i (i.e., just using best-fit $\hat{x}_i$) and integrating over x_i can give very different results!*

(See Loredo (2004) for tutorial examples)

To estimate x_1:

$$\begin{aligned} p(x_1|\{x_2,\ldots\}) &= \int d\theta\; p(\theta)\, f(x_1|\theta)\, \ell_1(x_1) \times \prod_{i=2}^{N} \int dx_i\; f(x_i|\theta)\, \ell_i(x_i) \\ &= \ell_1(x_1) \int d\theta\; p(\theta)\, f(x_1|\theta) \mathcal{L}_{m,\check{1}}(\theta) \\ &\approx \ell_1(x_1) f(x_1|\hat{\theta}_{\check{1}}) \end{aligned}$$

with $\hat{\theta}_{\check{1}}$ determined by the remaining data

$f(x_1|\hat{\theta}_{\check{1}})$ behaves like a "prior" that shifts the x_1 estimate away from the peak of $\ell_1(x_1)$; each member's prior depends on all of the rest of the data $\rightarrow$ shrinkage

[*For astronomers*: This generalizes the corrections derived by Eddington, Malmquist and Lutz-Kelker (sans selection effects)]

Agenda

1. Basic Bayes recap
2. Key idea in a nutshell
3. **Going deeper**
   - Joint distributions and DAGs
   - Conditional dependence/indepence
   - Example: Binomial prediction
   - Beta-binomial model
   - Point estimation and shrinkage
   - Gamma-Poisson model & Stan
   - Algorithms

Joint and conditional distributions

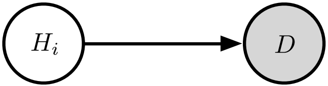
Bayesian inference is largely about the interplay between *joint* and *conditional* distributions for related quantities

Ex: Bayes's theorem relating hypotheses and data ($||\mathcal{C}$):

$$P(H_i|D) = \frac{P(H_i)P(D|H_i)}{P(D)} = \frac{P(H_i, D)}{P(D)} = \frac{\text{joint for everything}}{\text{marginal for knowns}}$$

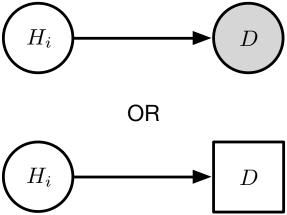
The usual form identifies an available factorization of the joint

Express this via a *directed acyclic graph* (DAG):



Joint distribution structure as a graph

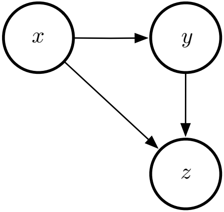
- Graph = *nodes/vertices* connected by *edges/links*
- Circular/square nodes/vertices = a priori uncertain quantities (gray/square = becomes known as data)
- Directed edges specify conditional dependence
- Absence of an edge indicates conditional *in*dependence
  $\rightarrow$ *the most important edges are the missing ones*



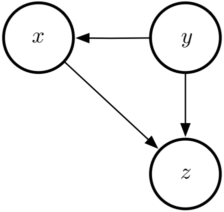
$$P(H_i, D) = P(H_i) \times P(D|H_i)$$

$p(x, y, z)$

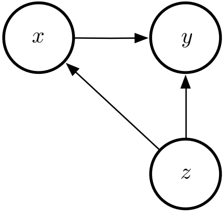
$p(x)p(y|x)p(z|x, y)$



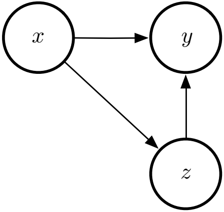
$p(y)p(x|y)p(z|y, x)$



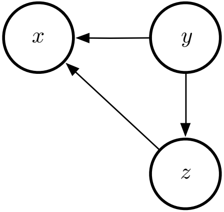
$p(z)p(x|z)p(y|z, x)$



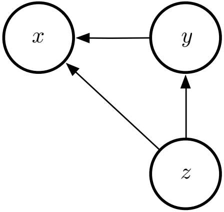
$p(x)p(z|x)p(y|x, z)$



$p(y)p(z|y)p(x|y, z)$

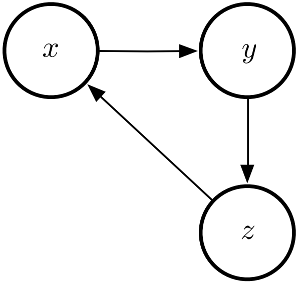


$p(z)p(y|z)p(x|z, y)$



Cycles not allowed

$$p(x|z) \times p(y|x) \times p(z|y)?$$

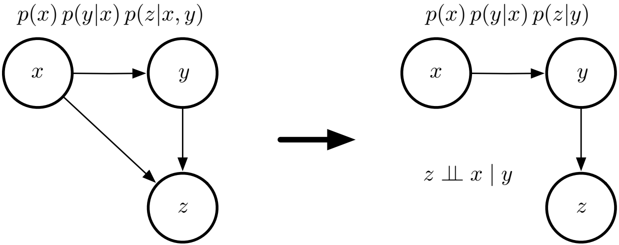


Conditional independence

Suppose for the problem at hand z is independent of of x when y is known:

$$p(z|x, y) = p(z|y)$$

"z is *conditionally independent* of x, given y": $z \perp\!\!\!\perp x \mid y$

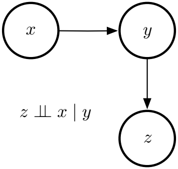


Absence of an edge indicates conditional *in*dependence
Missing edges indicate simplification in structure
$\rightarrow$ *the most important edges are the missing ones*

DAGs with missing edges

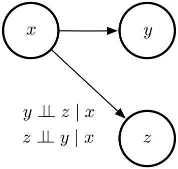
Conditional independence

$p(x)\,p(y|x)\,p(z|y)$



"Causal chain"

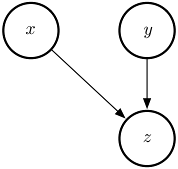
$p(x)\,p(y|x)\,p(z|x)$



"Common cause"

Conditional dependence

$p(x)\,p(y)\,p(z|x,y)$



"Common effects"

Conditional vs. complete independence

"z is *conditionally* independent of x, given y"

$$\neq$$

"z is independent of x"

(Complete) independence between z and x ("$z \perp\!\!\!\perp x$") would imply:

$$p(z|x) = p(z) \quad \text{(i.e., not a function of } x\text{)}$$

Conditional independence *given y* ("$z \perp\!\!\!\perp x \mid y$") is weaker:

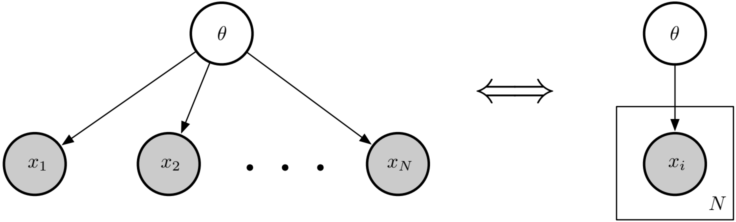
$$\begin{aligned} p(z|x) &= \int dy\ p(z, y|x) \\ &= \int dy\ p(y|x)p(z|x, y) \\ &= \int dy\ p(y|x)p(z|y) \quad \text{since } z \perp\!\!\!\perp x \mid y \end{aligned}$$

Although x drops out of the last factor, x dependence remains in $p(y|x)$

x *does* provide information about z, but it only does so through the information it provides about y (which directly influences z)

Bayes's theorem with IID samples

For model with parameters θ predicting data $D = \{x_i\}$ that are IID given θ:



$$p(\theta, D) = p(\theta)p(\{x_i\}|\theta) \; = \; p(\theta)\prod_{i=1}^{N} p(x_i|\theta)$$

To find the posterior for the unknowns (θ), divide the joint by the marginal for the knowns ($\{x_i\}$):

$$p(\theta|\{x_i\}) = \frac{p(\theta)\prod_{i=1}^{N} p(x_i|\theta)}{p(\{x_i\})} \quad \text{with} \quad p(\{x_i\}) = \int d\theta \, p(\theta)\prod_{i=1}^{N} p(x_i|\theta)$$

Binomial counts

   . . . n_1 heads in N flips

   . . . n_2 heads in N flips

Suppose we know n_1 and want to predict n_2

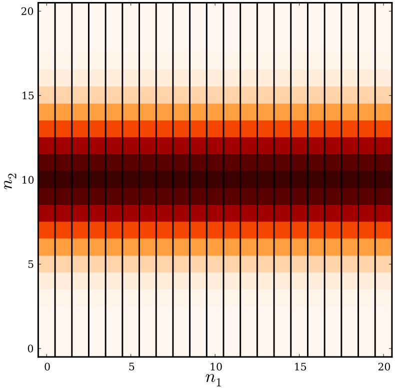
Predicting binomial counts — known α

Success probability $\alpha \to p(n|\alpha) = \frac{N!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n} \qquad || \, N$

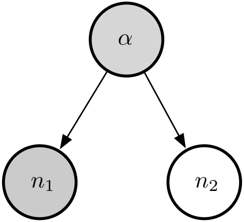
Consider two successive runs of $N = 20$ trials, *known* $\alpha = 0.5$

$$p(n_2|n_1, \alpha) = p(n_2|\alpha) \qquad || \, N$$

n_1 and n_2 are *conditionally independent*



DAG for binomial prediction — known α



$$p(\alpha, n_1, n_2) = p(\alpha)p(n_1|\alpha)p(n_2|\alpha)$$

$$\begin{aligned} p(n_2|\alpha, n_1) &= \frac{p(\alpha, n_1, n_2)}{p(\alpha, n_1)} \\ &= \frac{p(\alpha)p(n_1|\alpha)p(n_2|\alpha)}{p(\alpha)p(n_1|\alpha)\sum_{n_2} p(n_2|\alpha)} \\ &= p(n_2|\alpha) \end{aligned}$$

Knowing α lets you predict each n_i, independently

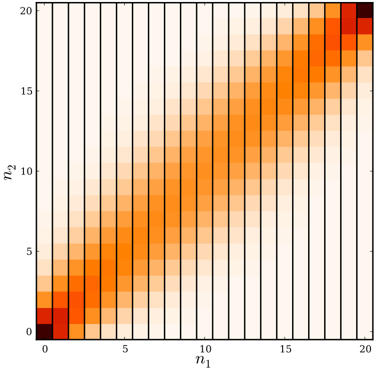
Predicting binomial counts — uncertain α

Consider the same setting, but with α *uncertain*

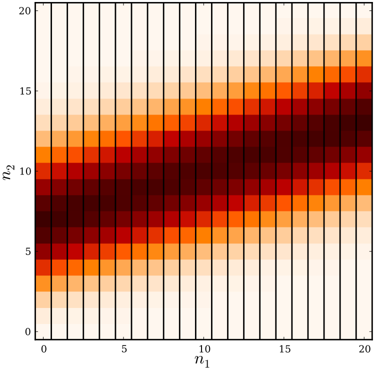
Outcomes are *physically* independent, but n_1 tells us about $\alpha \rightarrow$ outcomes are *marginally dependent* (see Lec 12 for calculation):

$$p(n_2|n_1, N) = \int d\alpha\ p(\alpha, n_2|n_1, N) \ = \ \int d\alpha\ p(\alpha|n_1, N)\, p(n_2|\alpha, N)$$

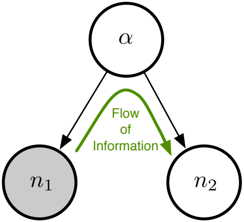
Flat prior on α



Prior: $\alpha = 0.5 \pm 0.1$



DAG for binomial prediction



$$p(\alpha, n_1, n_2) = p(\alpha)p(n_1|\alpha)p(n_2|\alpha)$$

From joint to conditionals:

$$p(\alpha|n_1, n_2) = \frac{p(\alpha, n_1, n_2)}{p(n_1, n_2)} = \frac{p(\alpha)p(n_1|\alpha)p(n_2|\alpha)}{\int d\alpha\; p(\alpha)p(n_1|\alpha)p(n_2|\alpha)}$$

$$p(n_2|n_1) = \frac{\int d\alpha\, p(\alpha, n_1, n_2)}{p(n_1)}$$

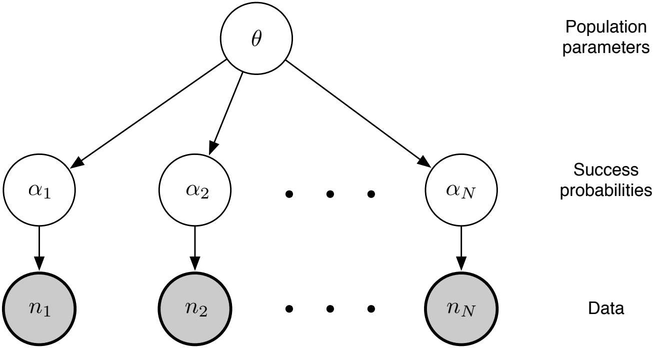
Observing n_1 lets you learn about α

Knowledge of α affects predictions for $n_2 \rightarrow$ dependence on n_1

A population of coins/flippers

Each flipper+coin flips different number of times

- What do we learn about the *population* of coins—the distribution of αs?
- How does population membership effect inference for a single coin's α?



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

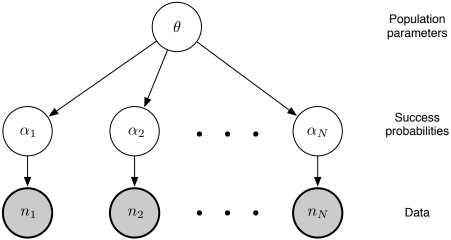
Terminology: θ are *hyperparameters*, $\pi(\theta)$ is the *hyperprior*

A simple multilevel model: beta-binomial

Goals:

- Learn a population-level "prior" by pooling data
- Account for population membership in member inferences

Qualitative



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

Quantitative

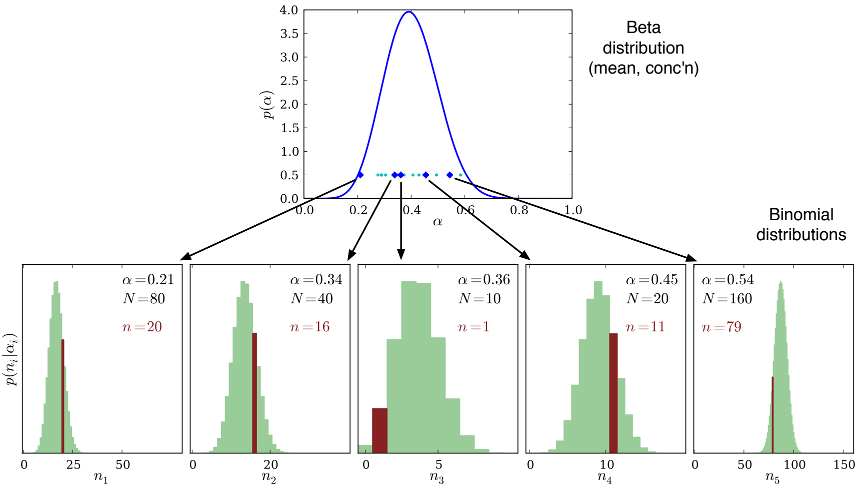
$$\theta = (a, b) \text{ or } (\mu, \sigma)$$

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

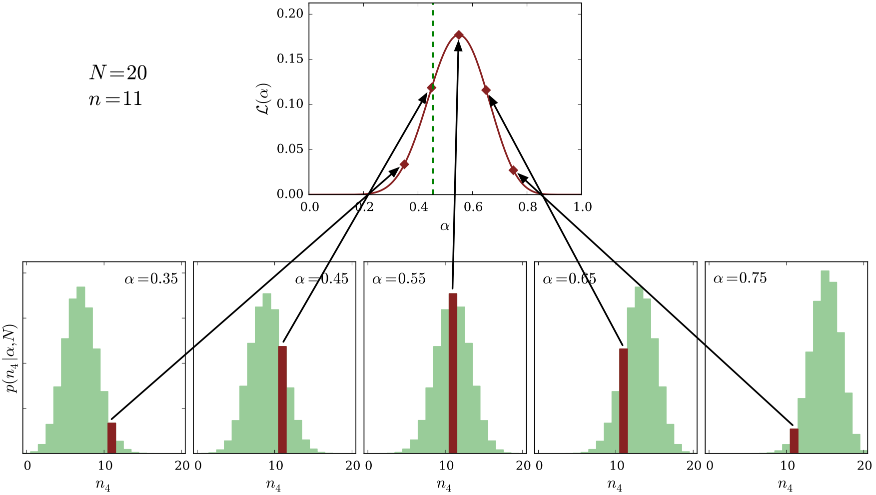
$$p(\alpha_i|\theta) = \mathrm{Beta}(\alpha_i|\theta)$$

$$p(n_i|\alpha_i) = \binom{N_i}{n_i} \alpha_i^{n_i}(1-\alpha_i)^{N_i - n_i}$$

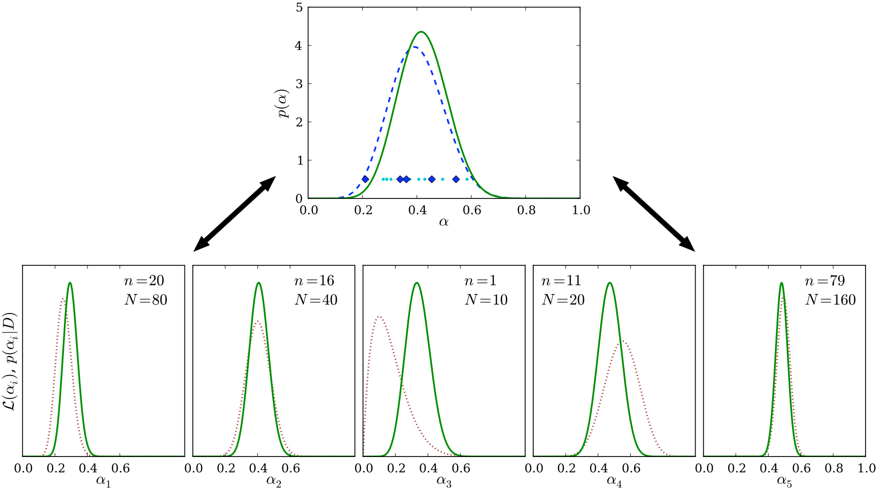
Generating the population & data



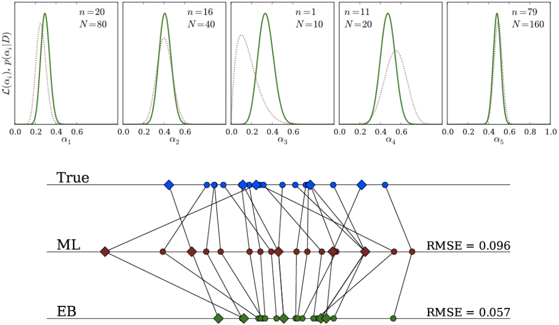
Likelihood function for one member's α



Learning the population distribution



Lower level estimates



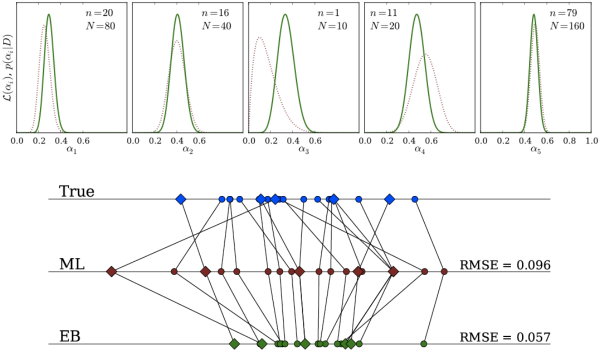
Two approaches

- **Hierarchical Bayes (HB):** Calculate marginals

$$p(\alpha_j|\{n_i\}) \propto \int d\theta\, \pi(\theta) \prod_{i \neq j} \int d\alpha_i\, p(\alpha_i|\theta)\, p(n_i|\alpha_i)$$

- **Empirical Bayes (EB):** Plug in an optimum $\hat{\theta}$ and estimate $\{\alpha_i\}$
  View as approximation to HB, or a frequentist procedure that estimates a prior from the data

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

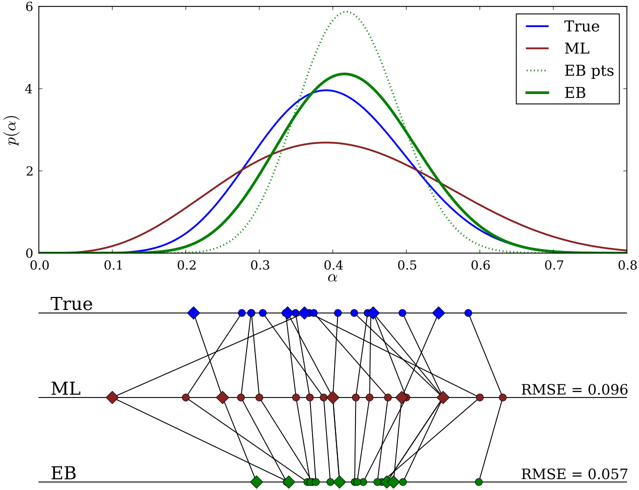
Lingo

- 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"

Beware of point estimates!

Population and member estimates



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*

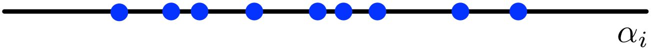
No point estimates of member properties are good for all tasks!

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!

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:

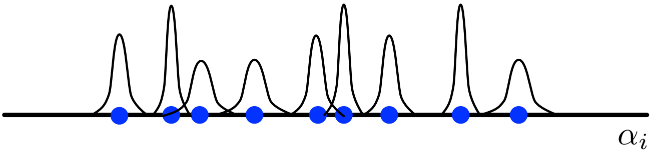


$$D = \{\alpha_i\}$$

$$\begin{aligned} p(D|\theta) &= \prod_i p(\alpha_i|\theta) \\ &= \prod_i \mathrm{Beta}(\alpha_i|\theta) \end{aligned}$$

(A *binomial point process*)

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



$$D = \{n_i\}$$

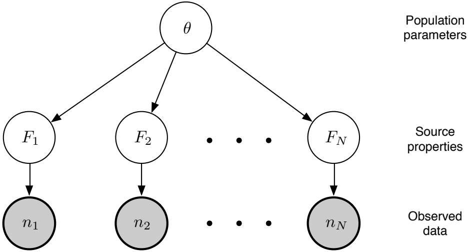
$$\begin{aligned} p(D|\theta) &= \prod_i \int d\alpha_i \; p(D, \{\alpha_i\}|\theta) \\ &= \prod_i \int d\alpha_i \; p(\alpha_i|\theta) \, p(n_i|\theta) \\ &= \prod_i \int d\alpha_i \; \mathrm{Beta}(\alpha_i|\theta) \; \mathrm{Binom}(n_i|\theta) \end{aligned}$$

This is a prototype for *measurement error problems*

Another conjugate MLM: Gamma-Poisson

Goal: Learn a rate dist'n from count data
(E.g., learn a star or galaxy brightness dist'n from photon counts)

Qualitative | Quantitative



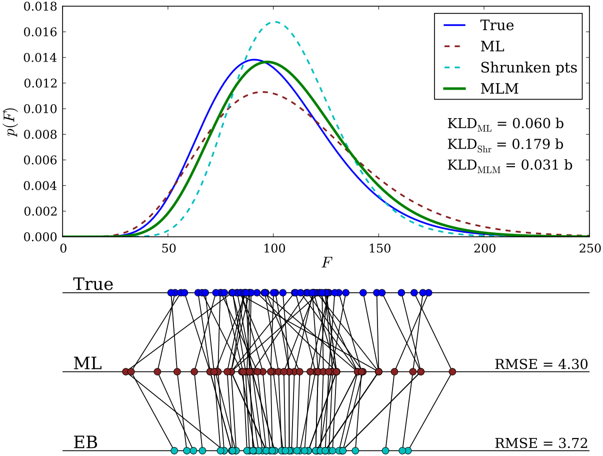
$$\theta = (\alpha, s) \text{ or } (\mu, \sigma)$$

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

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

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

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 $\times 16$

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*

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

Member marginalization

- Analytically or numerically integrate over $\{x_i\}$ $\rightarrow$ explore the 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

Metropolis-within-Gibbs algorithm

Block the full parameter space:

- Block of m population parameters, θ
- N blocks of lower level (latent) 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