Introduction to Bayesian multilevel models (hierarchical Bayes/graphical models)

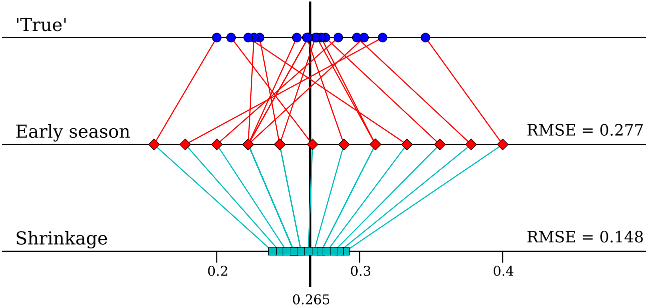
Tom Loredo
Dept. of Astronomy, Cornell University
http://www.astro.cornell.edu/staff/loredo/

IAC Winter School, 3–4 Nov 2014

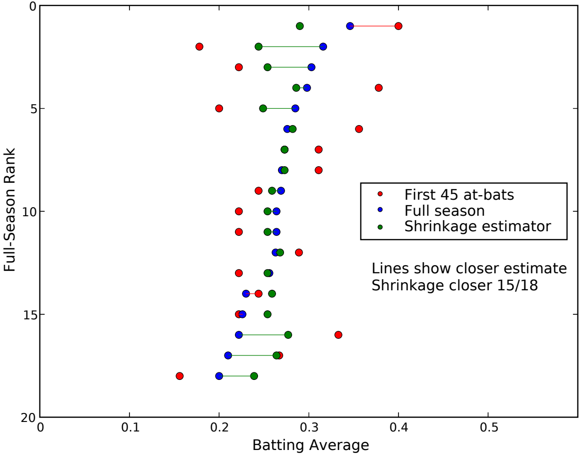
1970 baseball averages

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

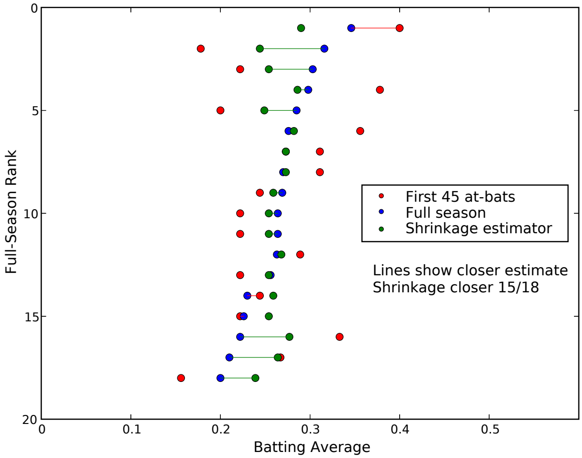
Red $= n/N$ maximum likelihood estimates of true averages
Blue $=$ Remainder of season, $N_{\mathrm{rmdr}} \approx 9N$



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
0.1
0.2
0.3
0.4
0.5



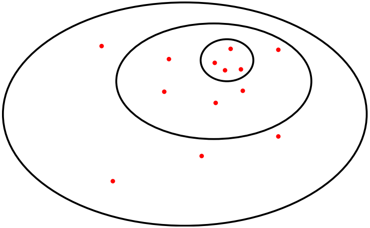
Theorem (independent Gaussian setting): In dimension $\gtrsim 3$, shrinkage estimators always beat independent MLEs in terms of expected RMS error

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

Accounting For Measurement Error

Introduce latent/hidden/incidental parameters

Suppose $f(x|\theta)$ is a distribution for an observable, x.



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

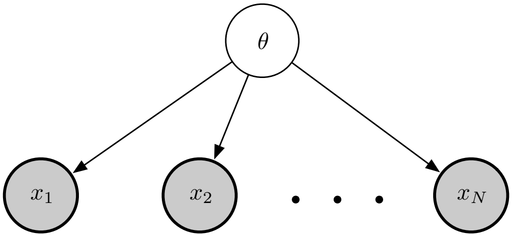
$$\mathcal{L}(\theta) \equiv p(\{x_i\}|\theta) = \prod_i f(x_i|\theta)$$

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

(A *binomial point process*)

Graphical representation

- Nodes/vertices = uncertain quantities (gray $\rightarrow$ known)
- Edges specify conditional dependence
- Absence of an edge denotes *conditional independence*

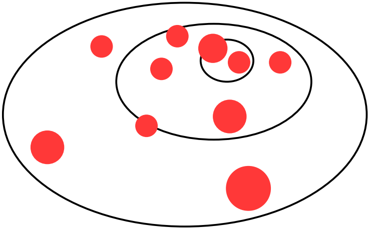


Graph specifies the form of the *joint distribution*:

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

Posterior from BT: $p(\theta|\{x_i\}) = p(\theta, \{x_i\})/p(\{x_i\})$

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 $\{x_i\}$.

Key point: *Maximizing over x_i and integrating over x_i can give very different results!*

To estimate x_1:

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

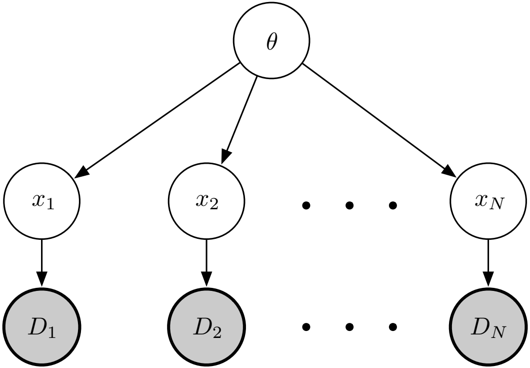
with $\hat{\theta}$ determined by the remaining data (EB)

$f(x_1|\hat{\theta})$ behaves like a prior that shifts the x_1 estimate away from the peak of $\ell_1(x_i)$

This generalizes the corrections derived by Eddington, Malmquist and Lutz-Kelker (sans selection effects)

(Landy & Szalay (1992) proposed adaptive Malmquist corrections that can be viewed as an approximation to this.)

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")

Multilevel models

1. Conditional and marginal dependence/independence

2. Populations and multilevel modeling

3. MLMs for cosmic populations

Multilevel models

1. **Conditional and marginal dependence/independence**

2. **Populations and multilevel modeling**

3. **MLMs for cosmic populations**

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

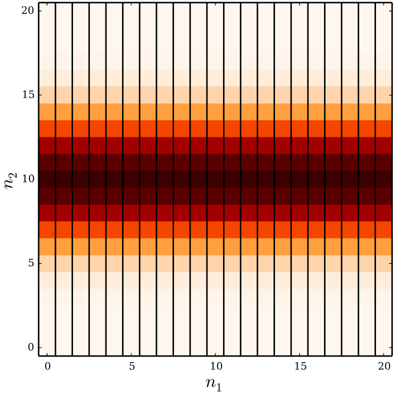
Predicting binomial counts — known α

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

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

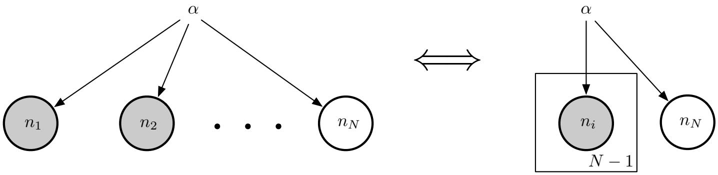
$$p(n_2|n_1, \alpha) = p(n_2|\alpha) \qquad || \, N$$

n_1 and n_2 are *conditionally independent*



Model structure as a graph

- Circular nodes/vertices = a priori uncertain quantities (gray = becomes known as data)
- Edges specify conditional dependence
- Absence of an edge indicates conditional *in*dependence



$$p(\{n_i\}|\alpha) = \prod_i p(n_i|\alpha)$$

Knowing α lets you predict each n_i, independently

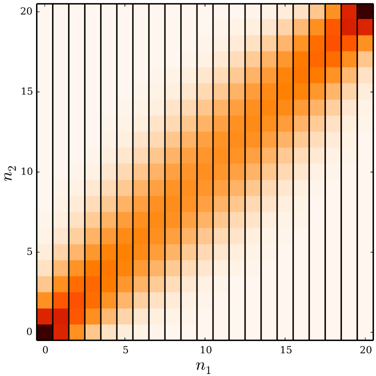
Predicting binomial counts — uncertain α

Consider the same setting, but with α *uncertain*

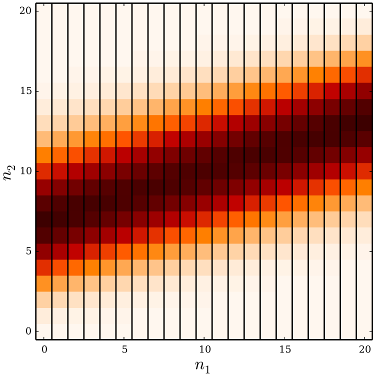
Outcomes are *physically* independent, but n_1 tells us about $\alpha \rightarrow$ outcomes are *marginally dependent*:

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

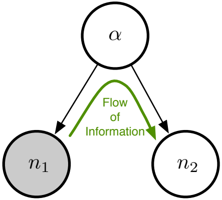
Flat prior on α



Prior: $\alpha = 0.5 \pm 0.1$



Graphical model — "Probability for everything"



$$p(\alpha, n_1, n_2) = \pi(\alpha) \prod_i p(n_i|\alpha) \equiv \pi(\alpha) \prod_i \ell_i(\alpha)$$

member likelihood

From joint to conditionals:

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

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

Observing n_1 lets you learn about α

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

Multilevel models

1. Conditional and marginal dependence/independence

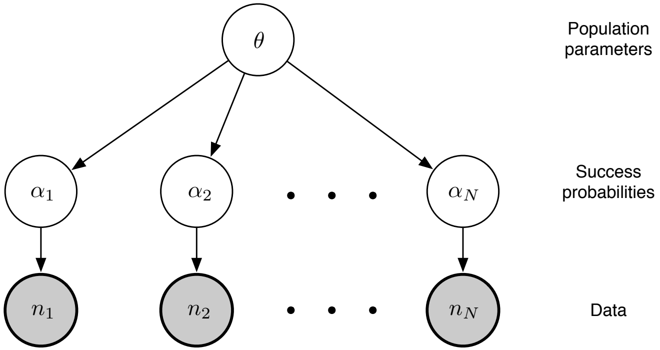
2. **Populations and multilevel modeling**

3. MLMs for cosmic populations

A population of coins/flippers




Each flipper+coin flips different number of times



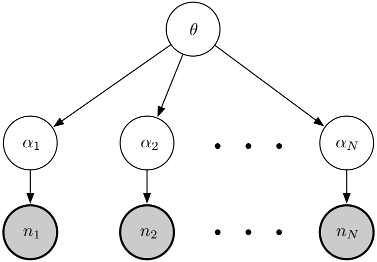
$$p(\theta, \{\alpha_i\}, \{n_i\}) = \pi(\theta) \prod_i p(\alpha_i|\theta)\; p(n_i|\alpha_i)$$
$$= \pi(\theta) \prod_i p(\alpha_i|\theta)\; \ell_i(\alpha_i)$$

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

A simple multilevel model: beta-binomial

Goal: Learn a population-level “prior” by pooling data

Qualitative | Quantitative



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

Population parameters

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

Success probabilities

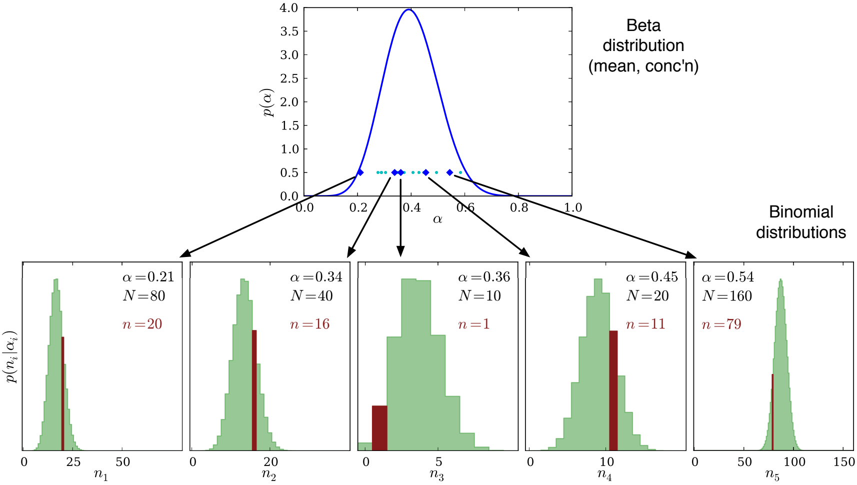
$$p(\alpha_i|\theta) = \mathrm{Beta}(\alpha_i|\theta)$$

Data

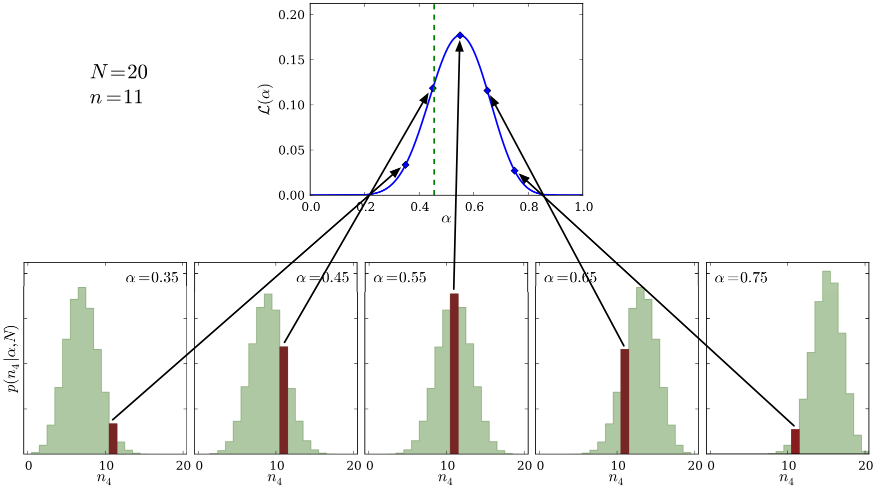
$$p(n_i|\alpha_i) = \binom{N_i}{n_i} \alpha_i^{n_i} (1-\alpha_i)^{N_i - n_i}$$

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

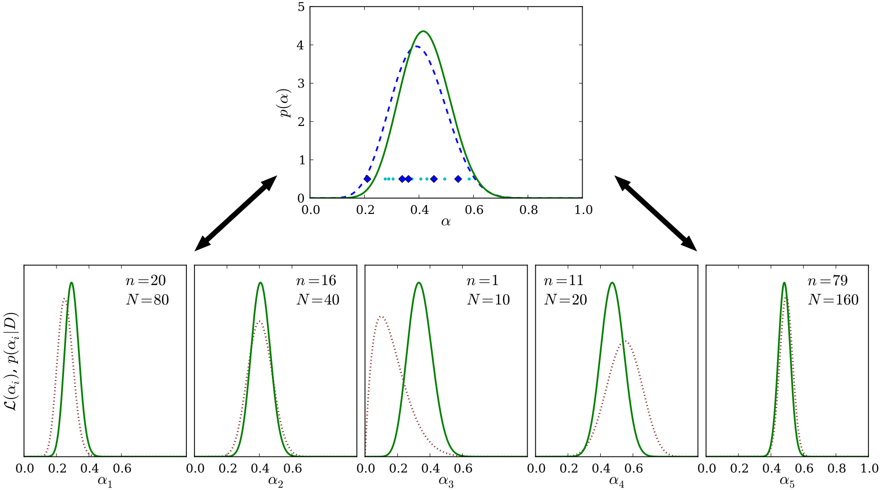
Generating the population & data



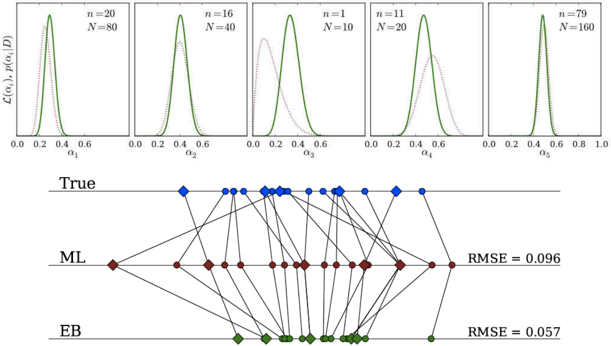
Likelihood function for one member's α



Learning the population distribution



Lower level estimates



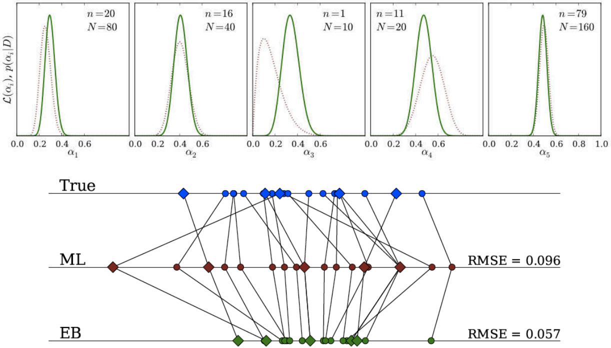
Two approaches

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

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

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

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*

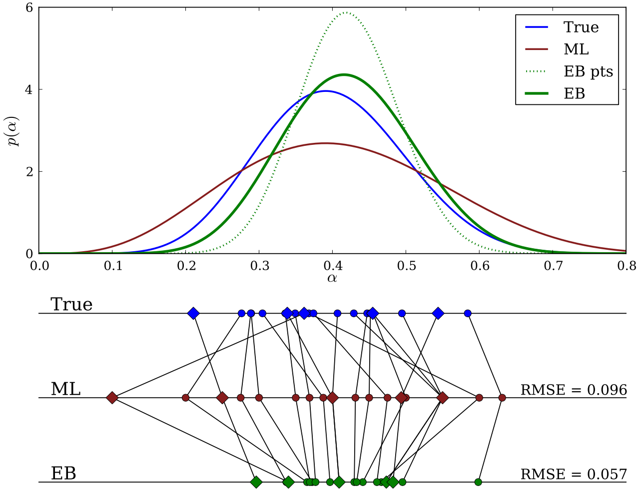
Frequentist outlook

- 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

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 → not optimal for estimating *population* properties*

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

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

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

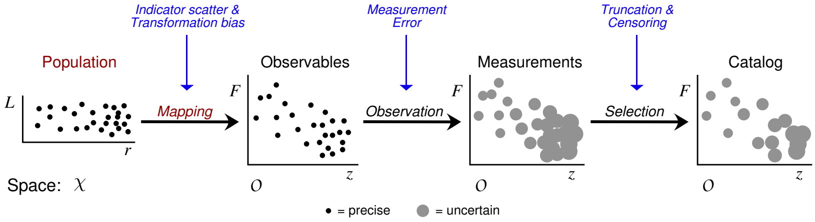
Multilevel models

❶ Conditional and marginal dependence/independence

❷ Populations and multilevel modeling

❸ **MLMs for cosmic populations**

Observing and modeling cosmic populations



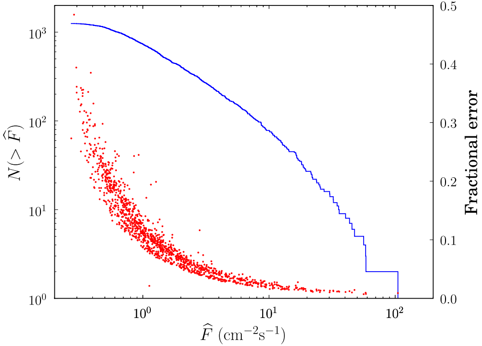
Science goals

- *Density estimation*: Infer the distribution of source characteristics, $p(\chi)$
- *Regression/Cond'l density estimation*: Infer relationships between different characteristics
- *Map regression*: Infer parameters defining the mapping from characteristics to observables

Notably, seeking improved point estimates of source characteristics is seldom a goal in astronomy.

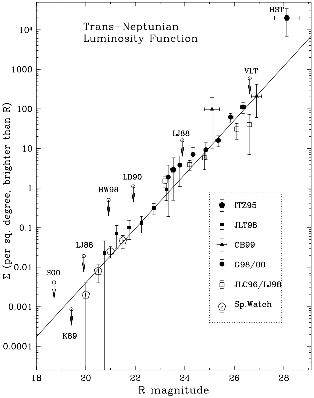
Number counts, luminosity functions

GRB peak fluxes



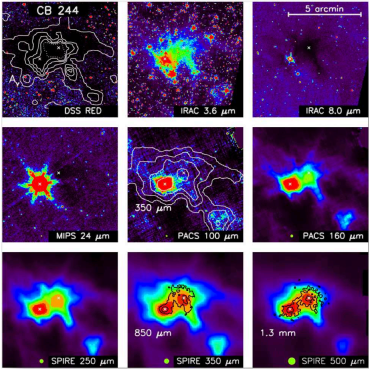
Loredo & Wasserman 1993, 1995, 1998: GRB luminosity/spatial dist'n via hierarchical Bayes

TNO magnitudes



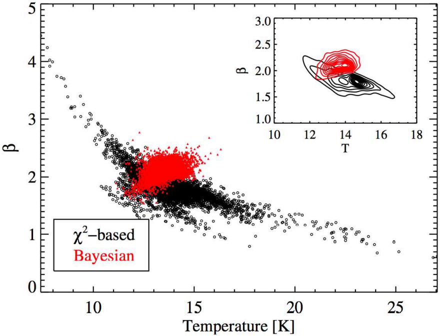
Gladman^{+}1998, 2001, 2008: TNO size distribution via hierarchical Bayes

CB244 molecular cloud



Herschel data from Stutz$^+$ 2010

SED properties vs. position

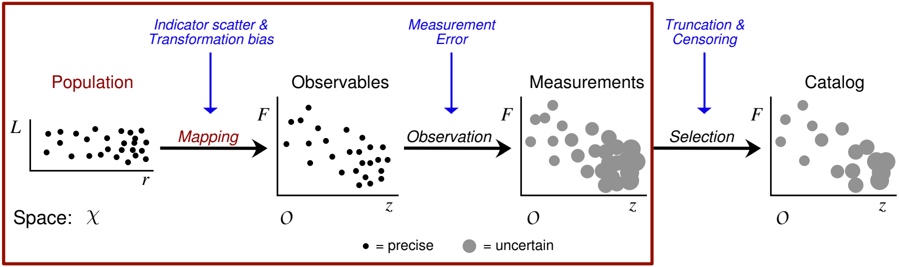


Kelly$^+$2012: Dust parameter correlations via hierarchical Bayes

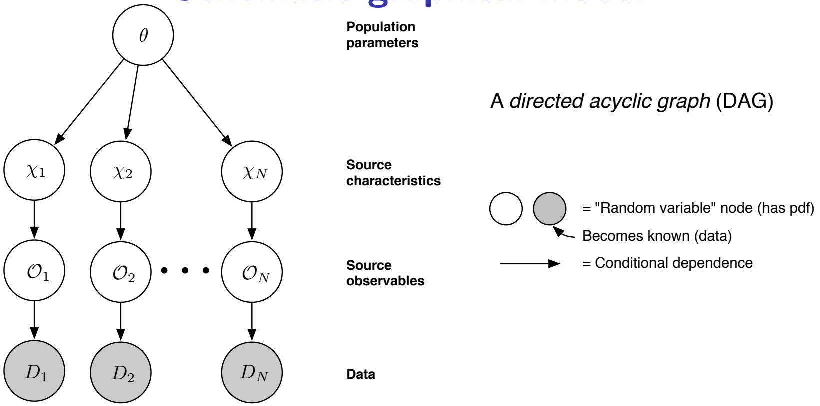
β = power law modification index

Expect $\beta \to 0$ for large grains

Measurement error models for cosmic populations



Schematic graphical model



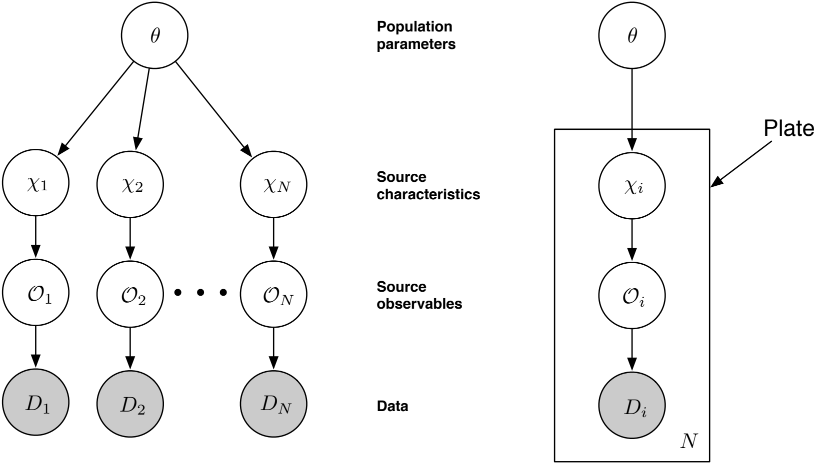
Graph specifies the form of the *joint distribution*:

$$p(\theta, \{\chi_i\}, \{\mathcal{O}_i\}, \{D_i\}) \; = \; p(\theta) \prod_i p(\chi_i|\theta)\; p(\mathcal{O}_i|\chi_i)\; p(D_i|\mathcal{O}_i)$$

Posterior from Bayes's theorem:

$$p(\theta, \{\chi_i\}, \{\mathcal{O}_i\}|\{D_i\}) = p(\theta, \{\chi_i\}, \{\mathcal{O}_i\}, \{D_i\}) \;/\; p(\{D_i\})$$

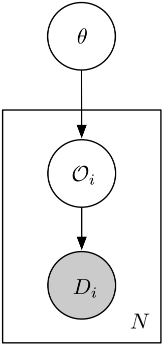
Plates



θ
Population parameters
θ
Plate
χ_1
χ_2
χ_N
Source characteristics
χ_i
$\mathcal{O}_1$
$\mathcal{O}_2$
$\mathcal{O}_N$
Source observables
$\mathcal{O}_i$
D_1
D_2
D_N
Data
D_i
N

"Two-level" effective models

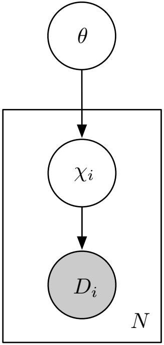
Number counts
$\mathcal{O}$ = flux



Calculate flux dist'n using "fundamental eqn" of stat astro

(Analytically/numerically marginalize over $\chi = (L, r)$)

Dust SEDs
χ = spectrum params



Observables = fluxes in bandpasses
Fluxes are *deterministic* in χ_i

From flips to fluxes

Simplified number counts model

- $\alpha_i \rightarrow$ source flux, F_i
- Upper level $\pi(\alpha) \rightarrow \log N$–$\log S$ dist'n
- $n_i \rightarrow$ counts in CCD pixels

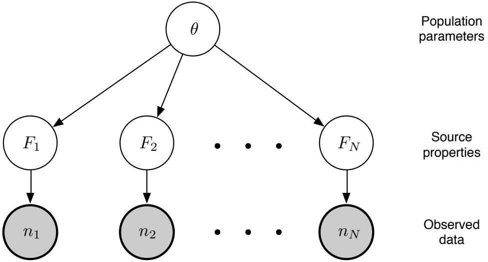
$\Rightarrow$ "Eddington bias" in disguise,
with both member *and* population inference
and uncertainty quantification

Another conjugate MLM: Gamma-Poisson

Goal: Learn a flux dist'n from photon counts

Qualitative

Quantitative



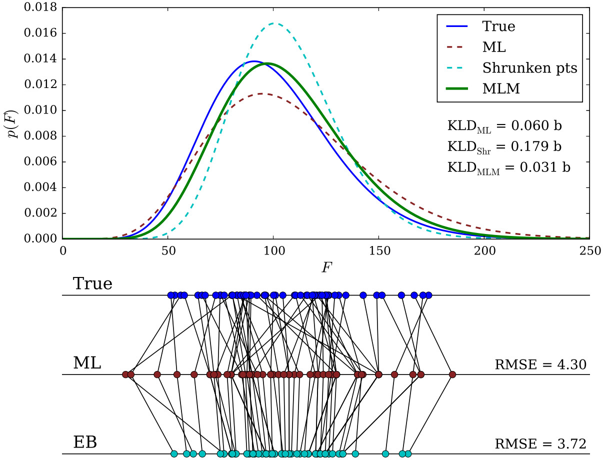
$$\theta = (\alpha, s) \text{ or } (\mu, \sigma)$$

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

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

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

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 $\times 16$

Benefits and requirements of cosmic MLMs

Benefits

- Selection effects quantified by *non-detection data*
  - *vs.* $V/V_{\rm max}$ *and "debiasing" approaches*
- Source uncertainties propagated via *marginalization*
  - *Adaptive generalization of Eddington/Malmquist "corrections"*
  - *No single adjustment addresses source & pop'n estimation*

Requirements

- Data summaries for non-detection intervals
  (exposure, efficiency)
- *Likelihood functions* (*not* posterior dist'ns) for detected source characteristics
  (Perhaps a role for *interim priors*)

Some Bayesian MLMs in astronomy

Surveys (number counts/"log N–log S"/Malmquist):

- GRB peak flux dist'n (Loredo & Wasserman 1998^{+})
- TNO/KBO magnitude distribution (Gladman^{+} 1998; Petit^{+} 2008)
- Malmquist-type biases in cosmology; MLM tutorial (Loredo & Hendry 2009 in *BMIC* book)
- "Extreme deconvolution" for proper motion surveys (Bovy, Hogg, & Roweis 2011)
- Exoplanet populations (2014 Kepler workshop)

Directional & spatio-temporal coincidences:

- GRB repetition (Luo^{+} 1996; Graziani^{+} 1996)
- GRB host ID (Band 1998; Graziani^{+} 1999)
- VO cross-matching (Budavári & Szalay 2008)

Linear regression with measurement error:

- QSO hardness vs. luminosity (Kelly 2007)

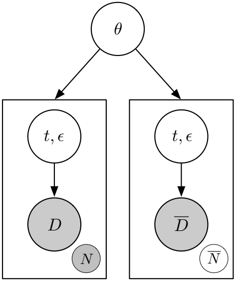
Time series:

- SN 1987A neutrinos, uncertain energy vs. time (Loredo & Lamb 2002)
- Multivariate "Bayesian Blocks" (Dobigeon, Tourneret & Scargle 2007)
- SN Ia multicolor light curve modeling (Mandel$^+$ 2009$^+$)

How far we've come

SN 1987A neutrinos, 1990

Marked Poisson point process
Background,
thinning/truncation,
measurement error

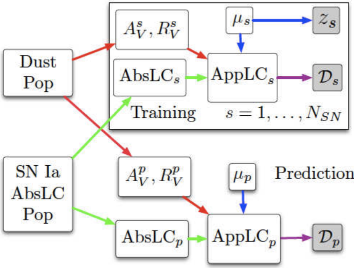


Model checking via
examining conditional
predictive dist'ns

SN Ia light curves

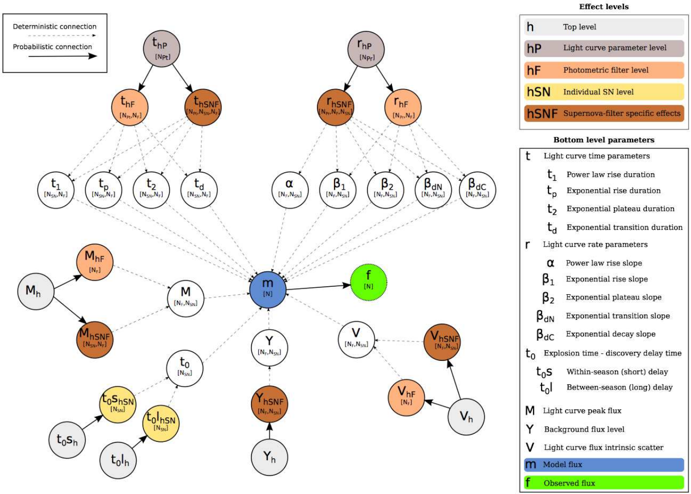
Mandel 2009, 2011

Nonlinear regression,
Gaussian process regression,
measurement error



Model checking via cross validation

SN IIP light curves (Sanders$^+$ 2014)



Recap of Key Ideas

- Conditional & marginal dependence/independence
- Latent parameters for measurement error
- Graphical models, multilevel models, hyperparameters
- Beta-binomial & gamma-Poisson conjugate MLMs
- Shrinkage estimators (member point estimates)
  - Empirical Bayes
  - Hierarchical Bayes
- Member vs. population inference—competing goals