Bayesian multilevel modeling of cosmic populations

Truths, subtle truths, and hierarchical Bayes

Tom Loredo
Dept. of Astronomy, Cornell University

With David Chernoff, Martin Hendry, David Ruppert,
Kunlaya Soiaporn, Ira Wasserman

Funded by NSF AAG & NASA ADAP programs

ICIC — 21 Aug 2012


ISAAC D'ISRAELI.
AUTHOR
LIVED HERE.
BORN 1766
DIED 1848

Isaac D'Israeli

1766–1848

Man of letters

Isaac D'Israeli
1766–1848
Man of letters

Wikimedia Commons

Benjamin Disraeli
1804–1881
Prime Minister, 1874–1880

Fogg Museum, Harvard


LCC
BENJAMIN
DISRAELI
Earl of
Beaconsfield
1804-1881
Died Here

Lies, damn lies, and. . .

"There are three kinds of lies: lies, damned lies, and statistics."

Lies, damn lies, and. . .

Mark Twain (1906)

Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark *attributed to* Disraeli would often apply with justice and force:

"There are three kinds of lies: lies, damned lies, and statistics."

Lies, damn lies, and...

Mark Twain (1906)

Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark *attributed to* Disraeli would often apply with justice and force:

"There are three kinds of lies: lies, damned lies, and statistics."

Lie? *That Disraeli ever said or wrote it!*

Lies, damn lies, and. . .

Mark Twain (1906)

Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark *attributed to* Disraeli would often apply with justice and force:

"There are three kinds of lies: lies, damned lies, and statistics."

Lie? *That Disraeli ever said or wrote it!*

Damned lie? *That the statement was originally about statistics!*

Liars, damned liars, and. . .

Sir Robert Giffen (1892)

> An old jest runs to the effect that there are three kinds of comparison among liars. There are liars, there are outrageous liars, and there are *scientific experts*.
>
> This has lately been adapted to throw dirt upon statistics. There are three degrees of comparisons, it is said, in lying. There are lies, there are outrageous lies, and there are statistics.

Liars, damned liars, and...

Sir Robert Giffen (1892)

An old jest runs to the effect that there are three kinds of comparison among liars. There are liars, there are outrageous liars, and there are *scientific experts*.

This has lately been adapted to throw dirt upon statistics. There are three degrees of comparisons, it is said, in lying. There are lies, there are outrageous lies, and there are statistics.

Statisticians can afford to laugh at and profit by jokes at their expense. There is so much knowledge which is unobtainable except by statistics...

"On international statistical comparisons," *Economic Journal* (1892)

See `http://www.york.ac.uk/depts/maths/histstat/lies.htm`

Truths, subtle truths, and hierarchical Bayes

❶ **Multilevel modeling: Key ideas**

❷ **Example applications in astronomy**

❸ **Future directions**

Agenda

1. **Multilevel modeling: Key ideas**
   - Dependence—conditional vs. marginal
   - Graphical models
   - Shrinkage; borrowing strength
   - Cautions: Priors, model checking

2. **Example applications in astronomy**

3. **Future directions**

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

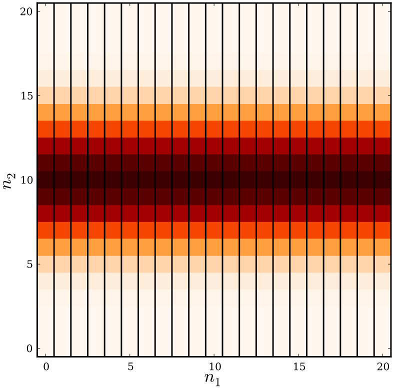
Predicting binomial counts — known α

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* $\alpha = 0.5$

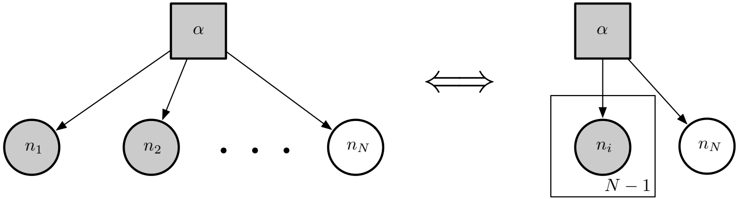
$$p(n_2|n_1, \alpha) = p(n_2|\alpha) \qquad || \, N$$

n_1 and n_2 are *conditionally independent*



Model structure as a graph

- Nodes/vertices = known quantities (squares), uncertain quantities (circles, gray = becomes known)
- 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

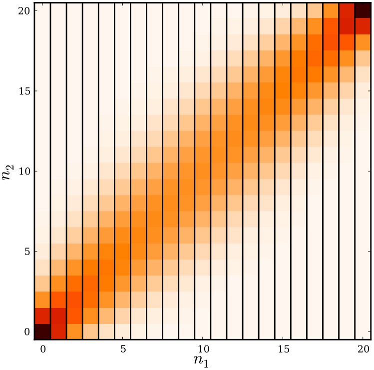
Predicting binomial counts — unknown α

Consider the same setting, but with α *unknown*

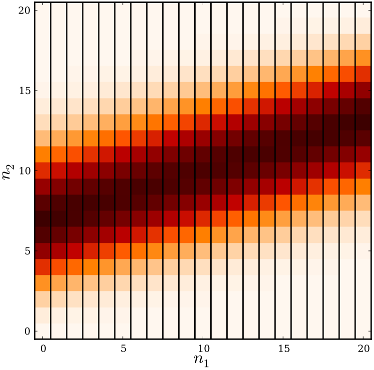
Outcomes are *physically* independent, but n_1 tells us about $\alpha \rightarrow$ outcomes are *marginally dependent*:

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

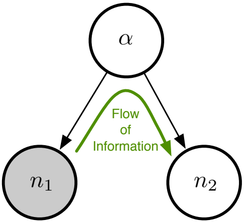
Flat prior on α



Prior: $\alpha = 0.5 \pm 0.1$



Graphical model — "Probability for everything"



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

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

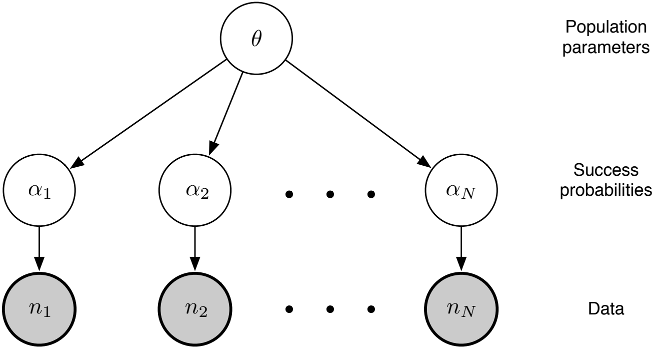
Knowledge of α affects predictions for $n_2 \rightarrow$ dependence on n_1

A population of coins/flippers





Each flipper+coin flips different number of times

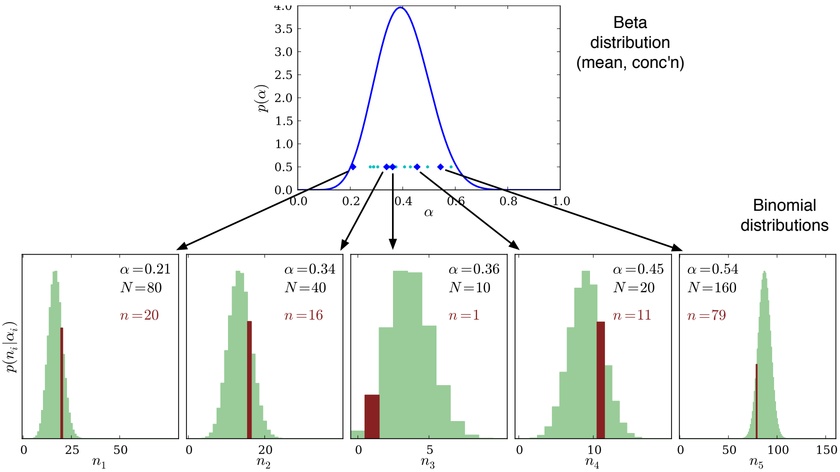


$$
\begin{aligned}
p(\theta, \{\alpha_i\}, \{n_i\}) &= \pi(\theta) \prod_i p(\alpha_i|\theta)\; p(n_i|\alpha_i) \\
&= \pi(\theta) \prod_i p(\alpha_i|\theta)\; \ell_i(\alpha_i)
\end{aligned}
$$

A simple multilevel model

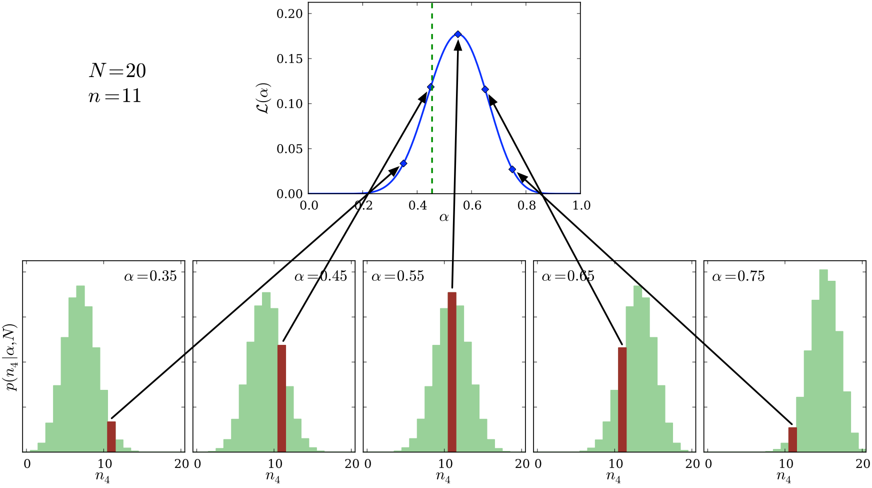
Goal: Learn a population-level "prior" by pooling data

Generating the population & data

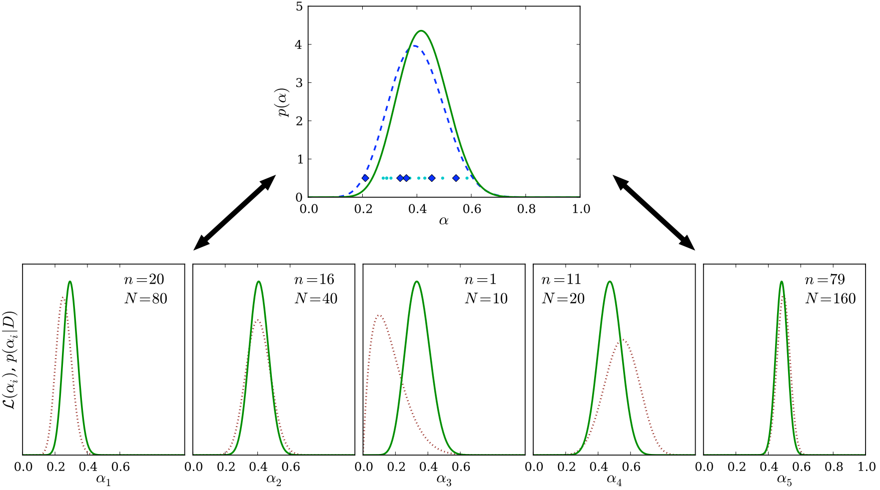


Likelihood function for one member's α

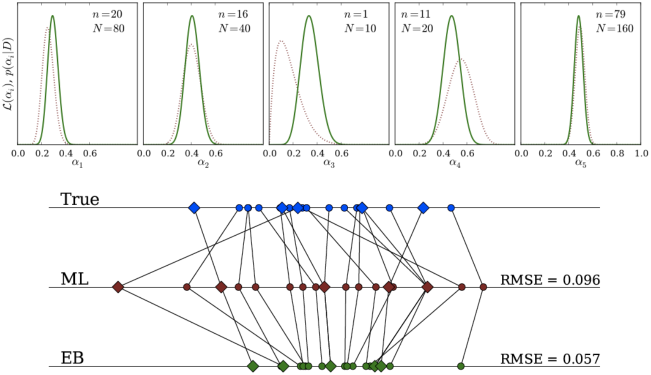
$N = 20$
$n = 11$



Learning the population distribution



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*

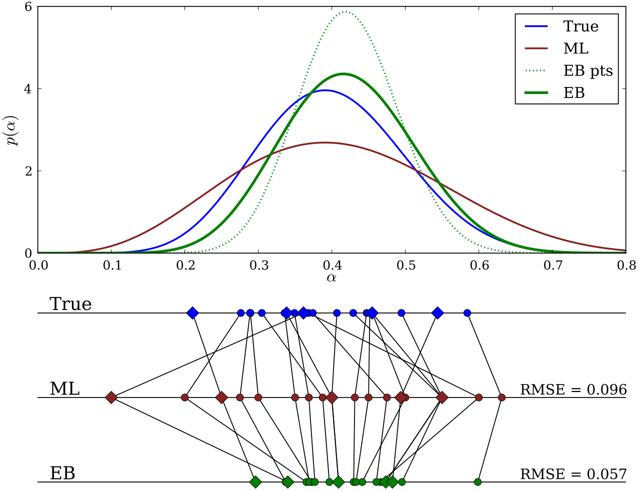
Frequentist outlook

- Point estimates are biased
- Reduced variance → 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 $\rightarrow$ not optimal for estimating *population* properties*

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

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

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

From flips to fluxes

- $\alpha_i \rightarrow$ 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

Cautions

Hyperpriors for population parameters

- Information gain from the data weakens going up the hierarchy
- Weakens dependence of lower level inference on upper levels $\rightarrow$ some robustness
- Improper priors that are okay for single-level inference can be dangerous (e.g., $1/\sigma$ is bad!)

Model checking

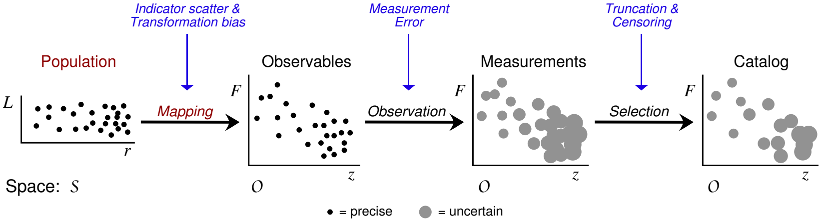
- Sinharay & Stern 2003:
  "[With posterior predictive checks] it is very difficult to detect violations of the assumptions made about the population distribution of the paramters unless the extent of violation is huge or the observed data have small standard errors."

- Bayarri & Castellanos 2007:
  "Both the posterior empirical Bayes and predictive posterior measures are *extremely* conservative, indicating almost perfect agreement of the observed data with the quite obviously wrong null models."
  Advocate *partial posterior predictive p-values*

Agenda

1. Multilevel modeling: Key ideas
2. **Example applications in astronomy**
   - Marked point process framework
   - Number-size dist'ns for GRBs. . .
   - Directional coincidence assessment
3. Future directions

Surveying and "Un-surveying"



$\Leftarrow$ *Inference goes this way!*

Inverse methods

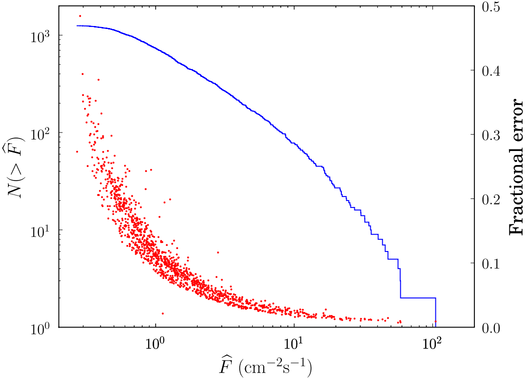
- Try to "correct" or "debias" data via adjustments/weights
- Focus on moments & empirical dist'n function (EDF)

Forward modeling methods

- Try to predict data by applying obs. process to pop'n model
- Focus on likelihood

Selection Effects and Measurement Error

BATSE Gamma-ray burst peak fluxes (EDF)



- Selection effects (truncation, censoring) — *obvious* (usually)
  Typically treated by "correcting" data
  Most sophisticated: product-limit estimators

- "Scatter" effects (measurement error, etc.) — *insidious*
  Classical "bias corrections" in some cases (Eddington. . .)
  Sometimes ignored (average out???)

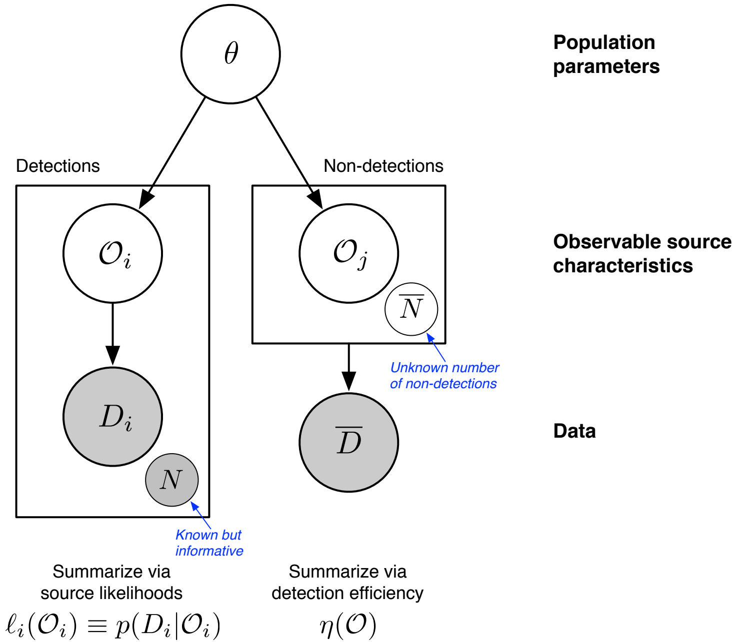
Marked point process framework

Catalog construction

- Systematically search through a *scan space* for sources
  - GRBs, cosmic rays: Scan in *time*
  - Stars/galaxies: Scan in *direction*
- Estimate *observable source characteristics* for candidates
  - GRBs: time, direction, peak flux, hardness, duration...
  - Cosmic rays: time, direction, energy...
  - Stars/galaxies: direction, multiband photometry...
- Collect information about *non-detections*:
  limits for candidates, thresholds, exposure/detection efficiency

Multilevel modeling of catalog data

- Model sources as a *point process in the scan space*, with *source observables as marks*
  - Phenomenological models: Model observables directly (e.g., $\log N$–$\log S$)
  - Physical models: Model in a population space; map to observables
- Measurement error: Data produce *source likelihoods*, $\ell_i(\mathcal{O})$
  - Straightforward to handle candidate sources/upper limits
- Model *detection and nondetection data*, accounting for detection criteria (thinning/truncation, $\eta(\mathcal{O})$)


θ
Population parameters
Detections
Non-detections
$\mathcal{O}_i$
$\mathcal{O}_j$
$\overline{N}$
Observable source characteristics
Unknown number of non-detections
D_i
$\overline{D}$
N
Data
Known but informative
Summarize via source likelihoods
$\ell_i(\mathcal{O}_i) \equiv p(D_i|\mathcal{O}_i)$
Summarize via detection efficiency
$\eta(\mathcal{O})$

Modeling GRB fluxes and directions

Loredo & Wasserman 1993, 1995, 1998

Observables: time, peak flux, direction (ignorable for cosmo models)

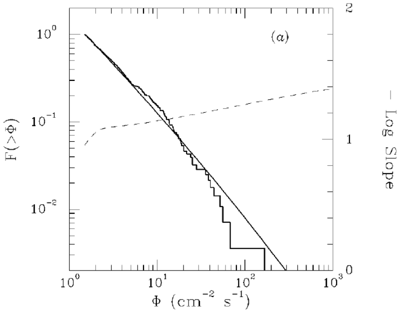
$R(\mathcal{O}_i; \theta) =$ Poisson point process intensity function for $\mathcal{O}$

$$p(\theta, \{\mathcal{O}_i\}|D, \overline{D}) \propto \pi(\theta) \exp\left[-\int d\mathcal{O}\, \eta(\mathcal{O}) R(\mathcal{O}; \theta)\right] \prod_{i=1}^{N} \ell_i(\mathcal{O}_i) R(\mathcal{O}_i; \theta)$$

$\mathcal{O}_i$ integrands are conditionally independent $\Rightarrow$ marginalize with 1-D or $1 \otimes$ 2-D quadrature rules

Modeling GRB fluxes and directions

"EDF" and
broken power law CDF



Phenomenological models (isotropic):

- Power law (PL)
- Broken power laws

Astrophysical models:

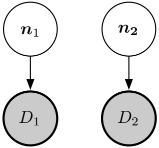
- Cosmological: Std candles, density evolution, power-law luminosity function
- Cosmo + halo models

Compare with Bayes factors vs. PL
(all $\sim$ .2 to 5)

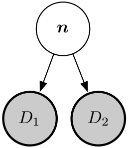
Bayesian Coincidence Assessment

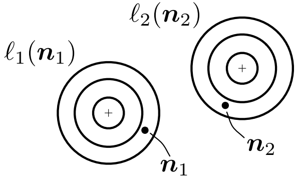
Luo, Loredo & Wasserman 1996; Graziani & Lamb 1996
Budavári & Szalay 2008

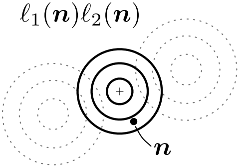
Not associated



Associated







$$
\begin{aligned}
p(d_1, d_2|H_0) &= \int d\boldsymbol{n}_1\; p(\boldsymbol{n}_1|H_0)\, \ell_1(\boldsymbol{n}_1) \\
&\times \int d\boldsymbol{n}_2 \cdots
\end{aligned}
$$

$$
p(d_1, d_2|H_1) = \int d\boldsymbol{n}\; p(\boldsymbol{n}|H_1)\, \ell_1(\boldsymbol{n})\, \ell_2(\boldsymbol{n})
$$

Challenge: Large hypothesis spaces

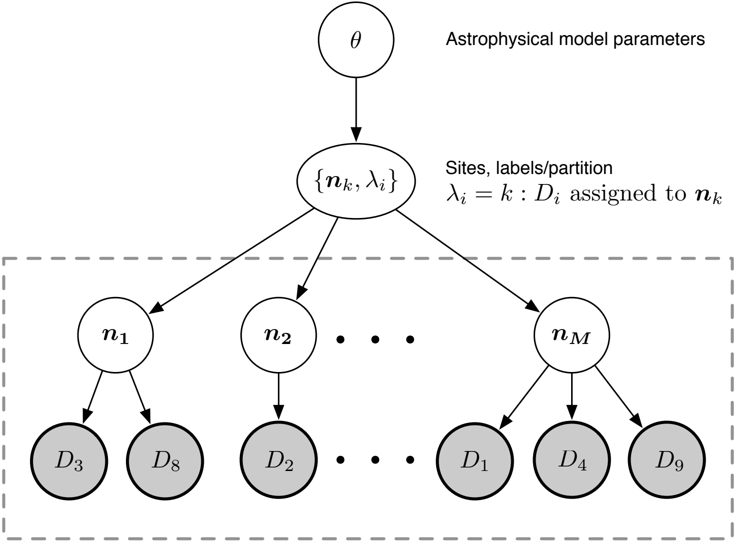
For $N = 2$ events, there was a single coincidence hypothesis, H_1

For $N = 3$ events:

- Three doublets: $1 + 2$, $1 + 3$, or $2 + 3$
- One triplet

The number of alternatives (partitions, ϖ) grows combinatorially!

- *Model building:* Assign sensible priors to partitions
- *Computation:* Find & sum over important partitions

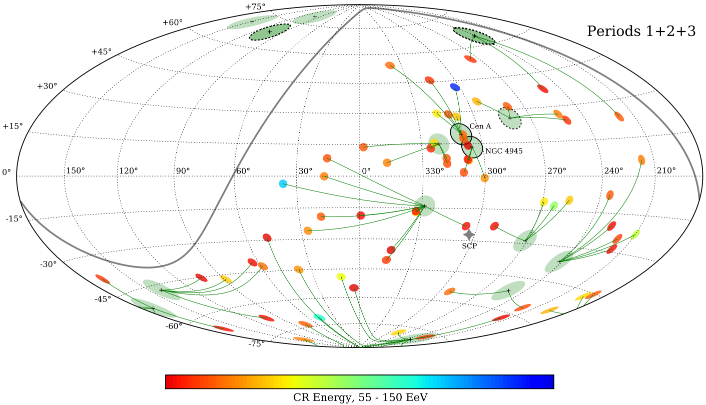


θ
Astrophysical model parameters
$\{\boldsymbol{n}_k, \lambda_i\}$
Sites, labels/partition
$\lambda_i = k : D_i$ assigned to $\boldsymbol{n}_k$
$\boldsymbol{n}_1$
$\boldsymbol{n}_2$
$\boldsymbol{n}_M$
D_3
D_8
D_2
D_1
D_4
D_9

Hunting for ultra-high energy cosmic ray sources

69 UHECRs from Pierre Auger Observatory (PAO)

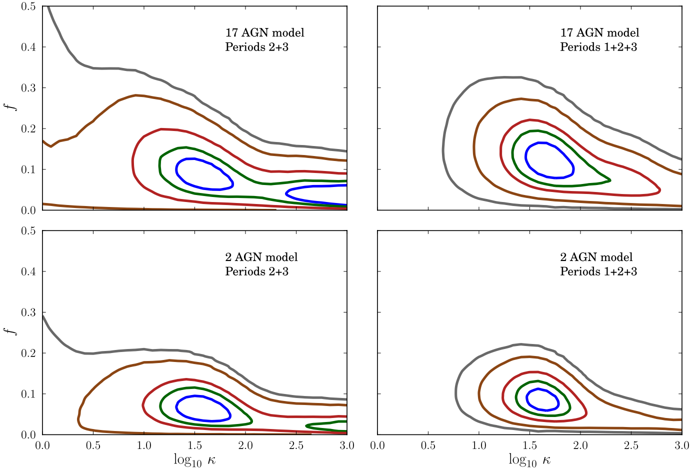
17 AGN from a volume-complete survey to 15 Mpc



Arcs connect each CR to its nearest AGN

Bayesian treatments: Watson$^+$ 2011; Soiaporn$^+$ 2012

Estimation of magnetic deflection (κ), AGN fraction (f)



Also assignment probabilities, change point models, predictive checks...

Simplistic models + significant issues due to "tuning" of published data
$\rightarrow$ results only suggestive

Agenda

1. Multilevel modeling: Key ideas
2. Example applications in astronomy
3. **Future directions**
   - Likelihood function catalogs
   - Adaptive scatter distortion corrections
   - Bayesian FDA for light curve ensembles

Likelihood function catalogs

MLM lessons

- Data are conditionally independent at lowest level
- Data enter both source-level and population-level inference via $\ell_i(\mathcal{O}) \equiv p(D_i|\mathcal{O})$
- No collection of point estimates is optimal for both source-level and population-level inference

Implications for survey reporting

- Report $\ell_i(\mathcal{O})$ to enable optimal inferences
- Naive likelihood summaries are *not* optimal estimates of source properties
- The required summaries are *not pdfs* for source properties; independent pdfs are typically not possible
- Report probabilistic summaries of non-detection data
- For targeted (counterpart) surveys replace "upper limits" with $\ell_i(\mathcal{O})$ summaries for candidate sources

This is in progress for BATSE GRBs, CFHTLS galaxy shapes

Adaptive scatter distortion corrections

Landy & Szalay on "Malmquist bias" for distances (1992)

Data D_i provide estimates $\hat{r}_i$; true distances are r_i

Prior $p(r_i) \propto r^2 n(r)$; likelihood $\mathcal{L}(r_i) =$ lognormal

$$\begin{aligned} p(r_i|D_i) &= \frac{r_i^2 n(r_i)\mathcal{L}(r_i)}{p(D_i)} \\ p(D_i) &= \int dr_i \, r_i^2 n(r_i)\mathcal{L}(r_i) \end{aligned}$$

LS92 set $p(D_i) = p(\hat{r}_i) = \Psi(\hat{r}_i)$, a smoothed fit to $\{\hat{r}_i\}$

$\rightarrow$ *moments of* $p(r_i|\hat{r}_i)$ *can be found from* $\Psi(\hat{r}_i)$

Use these to calculate corrections to $\hat{r}_i$

A *quasi*-empirical Bayes approach

Issues

- "Double counts" the data
- Doesn't account for uncertainty in $n(r)$ from $\hat{r}_i$ uncertainty or finite sample size

$\rightarrow$ *Revisit this as an explicit MLM*

Light curve ensembles

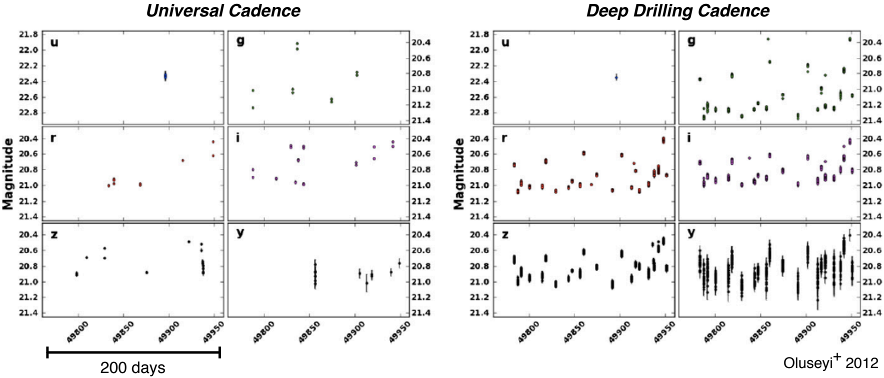
Current (CRTS, PTF, Pan-STARRS. . .) and future (LSST. . .) synoptic surveys $\rightarrow$ *large ensembles of multi-band light curves*

Underlying dynamic spectrum: $F(\lambda, t)$

Fluxes in bands: $F_\alpha(t) = \int d\lambda \, R_\alpha(\lambda) \, F(\lambda, t)$

Data produce sparse, asynchronous, noisy estimates of $\{F_{\alpha_i}(t_i)\}$

Simulated LSST RR Lyr Observations



Oluseyi+ 2012

Functional data analysis

Caricature: "Curves as data points"

Analysis of data probing *ensembles of functions* on a continuum: curves, surfaces, pdfs *over* time, space, wavelength...

Ramsey & Silverman 2005 (2nd ed.)

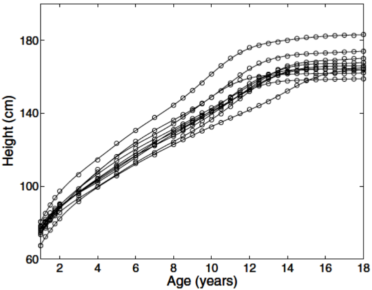


Figure 1.1. The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.

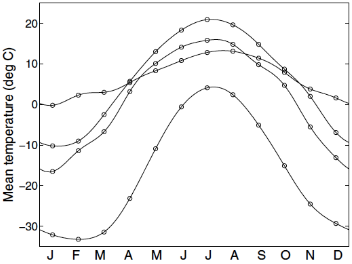


Figure 1.6. Mean monthly temperatures for the Canadian weather stations. In descending order of the temperatures at the start of the year, the stations are Prince Rupert, Montreal, Edmonton, and Resolute.

Emerging generalization: *Object oriented data analysis*—collections of curves, points on a shape manifold, graphs/trees...

FDA themes

- Registration of curves
- Smoothing of individual curves (estimate a function from samples)
- Nonparametric modeling
- Dimension reduction (functional PCA)
- Functional regression (using functions to predict scalars)
- *Until recently:*
    - Many samples
    - Synchronous samples
    - Negligible measurement error

Emerging area: Bayesian FDA

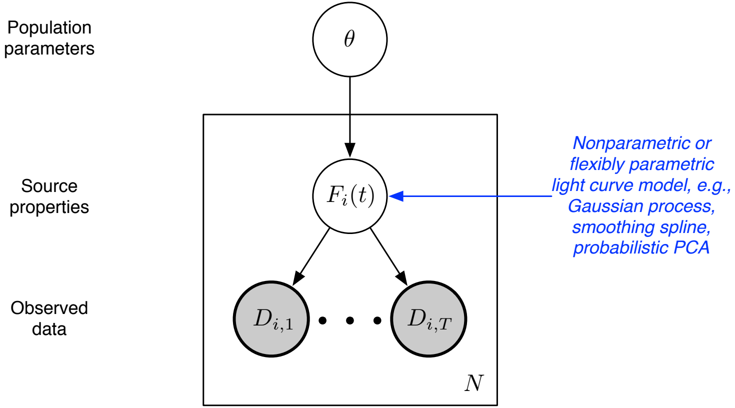
Arising for treatment of sparse, misaligned data with significant measurement errors

Motivation (Morris^{+} 2001):

- Frequentist study of colon cancer growth in rats fed corn or fish oil using parametric and kernel-based regression
- DNA indicators measured sparsely and non-coincident in time/space, with measurement error
- Key insight: Accurate population-level inference requires *under*smoothing of individual functions

This is a *functional counterpart to Eddington bias*

Subsequent work by M. D. Anderson group uses wavelet-based nonparametric regression in a MLM framework


Population parameters
θ
Source properties
$F_i(t)$
Nonparametric or flexibly parametric light curve model, e.g., Gaussian process, smoothing spline, probabilistic PCA
Observed data
$D_{i,1}$
$D_{i,T}$
N

A prototype

Mandel's BayeSN
SN Ia light curve model