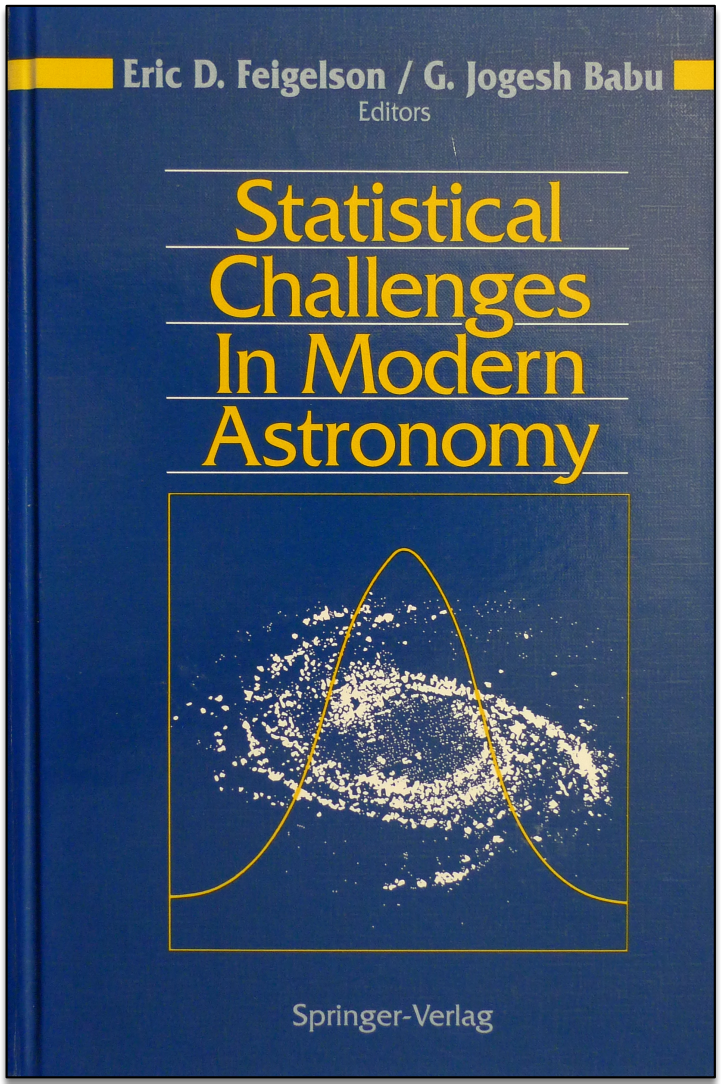
Commentary: Bayesian multilevel modeling at SCMA 6

Tom Loredo
Cornell Center for Astrophysics and Planetary Science

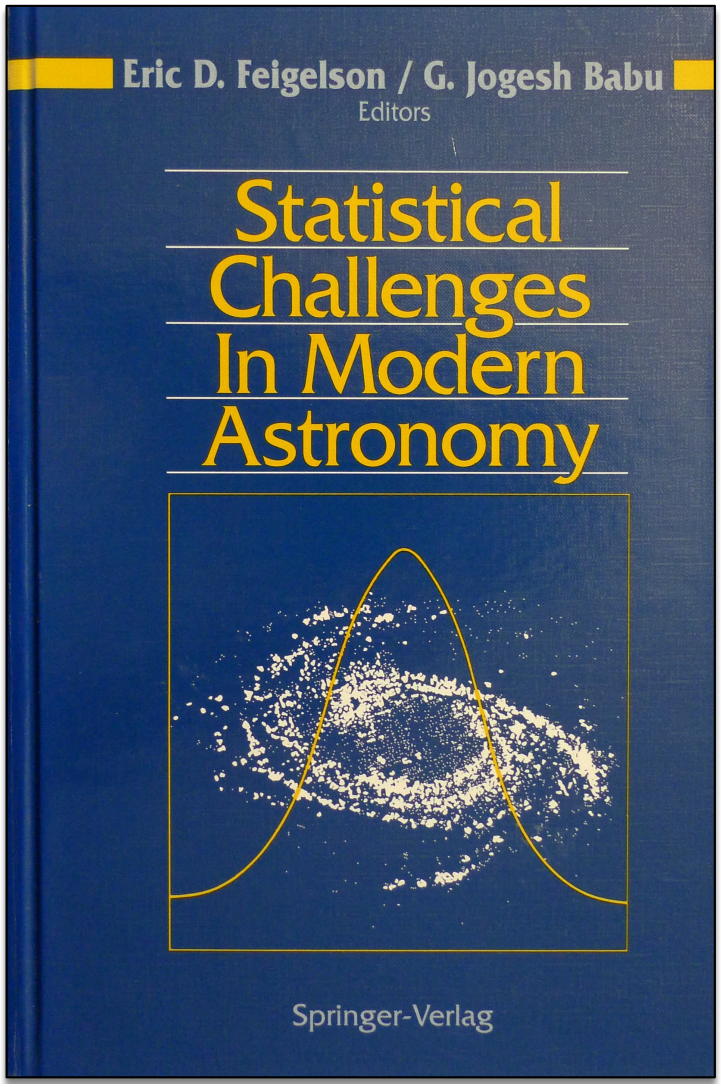
SCMA 6 — 9 June 2016

SCMA demographics



The conference was held on August 11–14, 1991, at the University Park campus of the Pennsylvania State University. Of the 131 participants, approximately 40% were statisticians and 60% astronomers. Twenty percent were graduate students, divided equally between the fields. Participants arrived from 12 countries. Several distinguished statisticians (including 11 Fellows of the IMS, 8 Fellows of the ASA, 2 past presidents of the IMS, and editors of several important journals) and astronomers (including a Fellow of the Royal Society) participated.

SCMA demographics



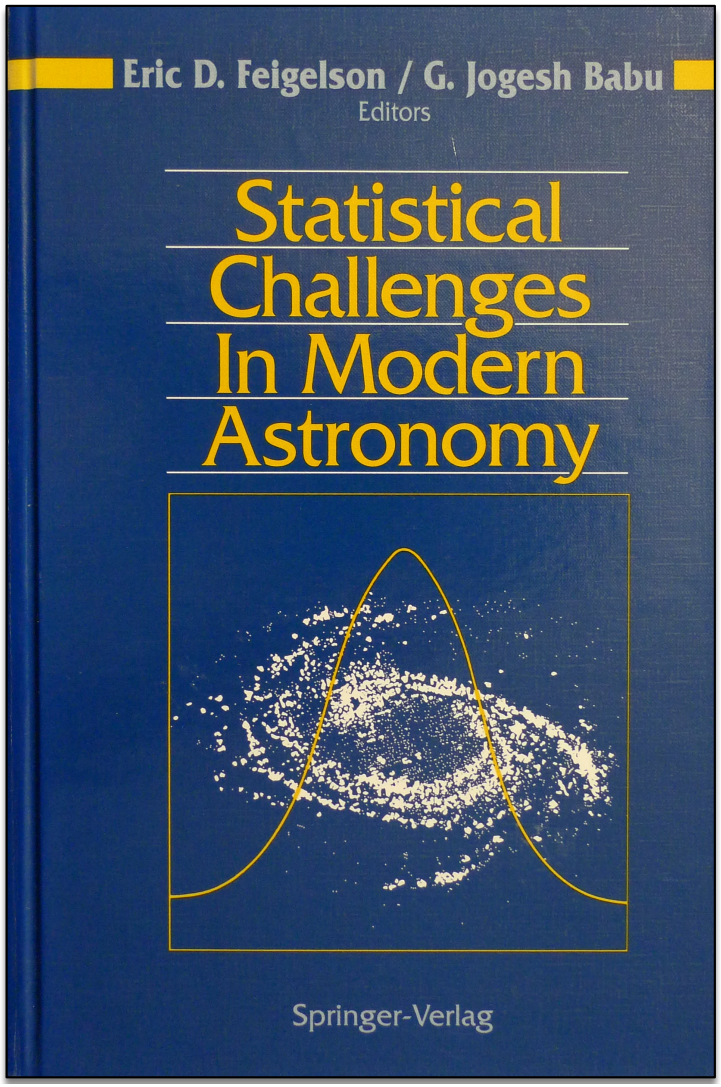
The conference was held on August 11–14, 1991, at the University Park campus of the Pennsylvania State University. Of the 131 participants, approximately 40% were statisticians and 60% astronomers. Twenty percent were graduate students, divided equally between the fields. Participants arrived from 12 countries. Several distinguished statisticians (including 11 Fellows of the IMS, 8 Fellows of the ASA, 2 past presidents of the IMS, and editors of several important journals) and astronomers (including a Fellow of the Royal Society) participated.

Bayesian content

12 Promise of Bayesian Inference for Astrophysics.
Thomas J. Loredo
Discussion by Gutti Jogesh Babu
Discussion by Mike West

14 Bayesian Methods of Deconvolution and Shape Classification, B.D. Ripley
Discussion by Fionn Murtagh
Discussion by Nicholas Weir

SCMA demographics



The conference was held on August 11–14, 1991, at the University Park campus of the Pennsylvania State University. Of the 131 participants, approximately 40% were statisticians and 60% astronomers. Twenty percent were graduate students, divided equally between the fields. Participants arrived from 12 countries. Several distinguished statisticians (including 11 Fellows of the IMS, 8 Fellows of the ASA, 2 past presidents of the IMS, and editors of several important journals) and astronomers (including a Fellow of the Royal Society) participated.

Bayesian content

12 Promise of Bayesian Inference for Astrophysics.
Thomas J. Loredo
Discussion by Gutti Jogesh Babu
Discussion by Mike West

14 Bayesian Methods of Deconvolution and
Shape Classification, B.D. Ripley
Discussion by Fionn Murtagh
Discussion by Nicholas Weir

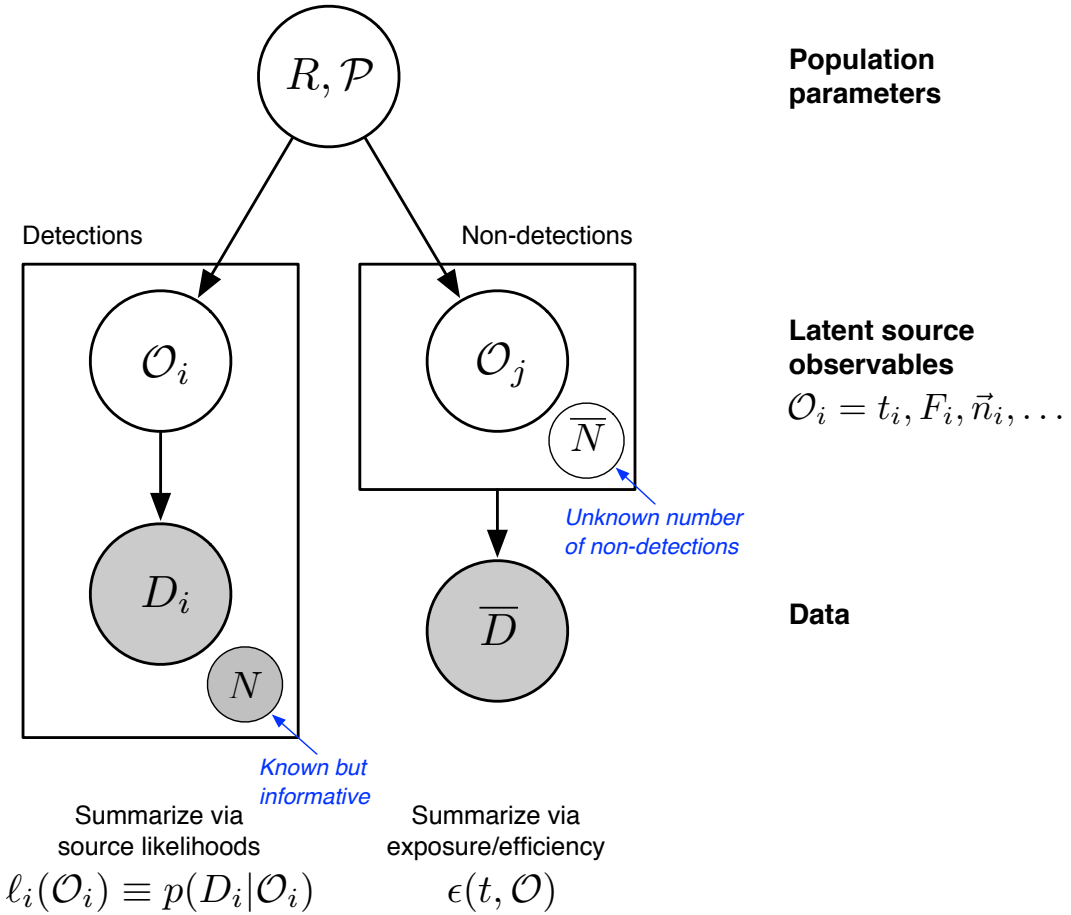
Female speakers

1 Surveys of Galaxy Redshifts, Martha P. Haynes

19 General Discussion: Time Series,
France Córdova (Moderator)

Cosmic demographics

The rise of Bayesian multilevel modeling



1989: SN 1987A ν dynamic spectroscopy: $N \approx 24$

1998: BATSE GRB fluxes, directions, distances: $N \approx 10^3$

Thinned latent point process marginal likelihood function

In LW95 we gave a detailed derivation of the form of the likelihood function for gamma-ray burst (GRB) data such as that provided by BATSE. It can be written as

$$\mathscr{L}(\mathscr{P}) = \exp\left[-T\int d\Phi \int d\boldsymbol{n}\,\bar{\eta}(\Phi, \boldsymbol{n})\,\frac{dR}{d\Phi\, d\boldsymbol{n}}\right] \times \prod_i \int d\Phi \int d\boldsymbol{n}\,\mathscr{L}_i(\Phi, \boldsymbol{n})\,\frac{dR}{d\Phi\, d\boldsymbol{n}}\,. \quad (7)$$

Here T is the duration of the observations, $\bar{\eta}(\Phi, \boldsymbol{n})$ is the time-averaged detection efficiency for bursts of flux Φ from direction $\boldsymbol{n}$, and $\mathscr{L}_i(\Phi, \boldsymbol{n})$ is the probability for seeing the data for burst i, presuming it comes from a burst with peak flux Φ and direction $\boldsymbol{n}$. We call $\mathscr{L}_i(\Phi, \boldsymbol{n})$ the individual burst likelihood function; it is the function one would use to infer the properties of a particular burst. LW95 derive expressions for $\bar{\eta}(\Phi, \boldsymbol{n})$ and $\mathscr{L}_i(\Phi, \boldsymbol{n})$ in terms of raw photon count data in the eight BATSE detectors.

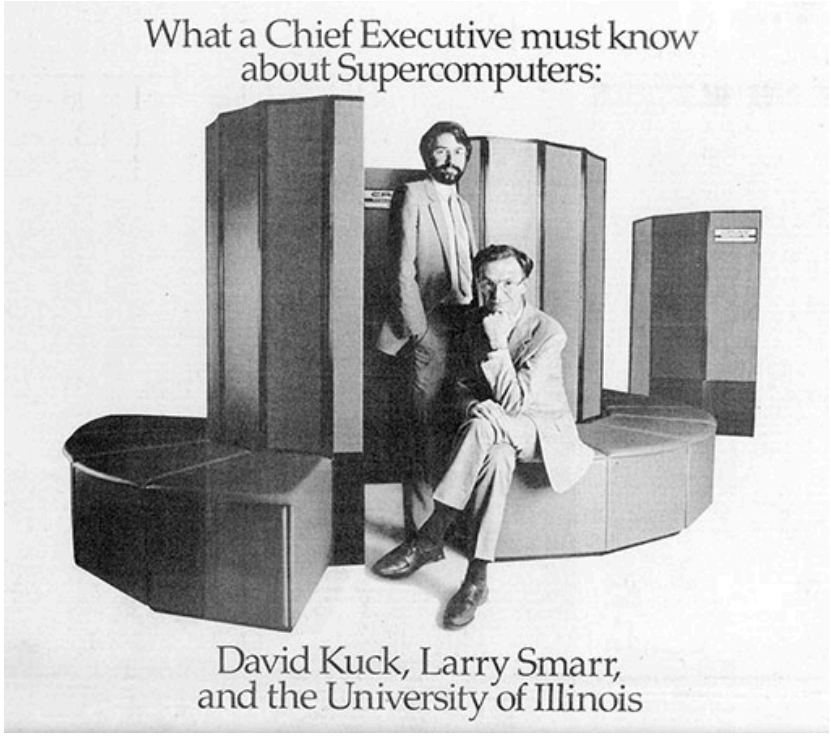
TL & Ira Wasserman 1993, 1995, 1998

Handling 24 latent parameters, ca. 1990

FPS 164

computerhistory.org

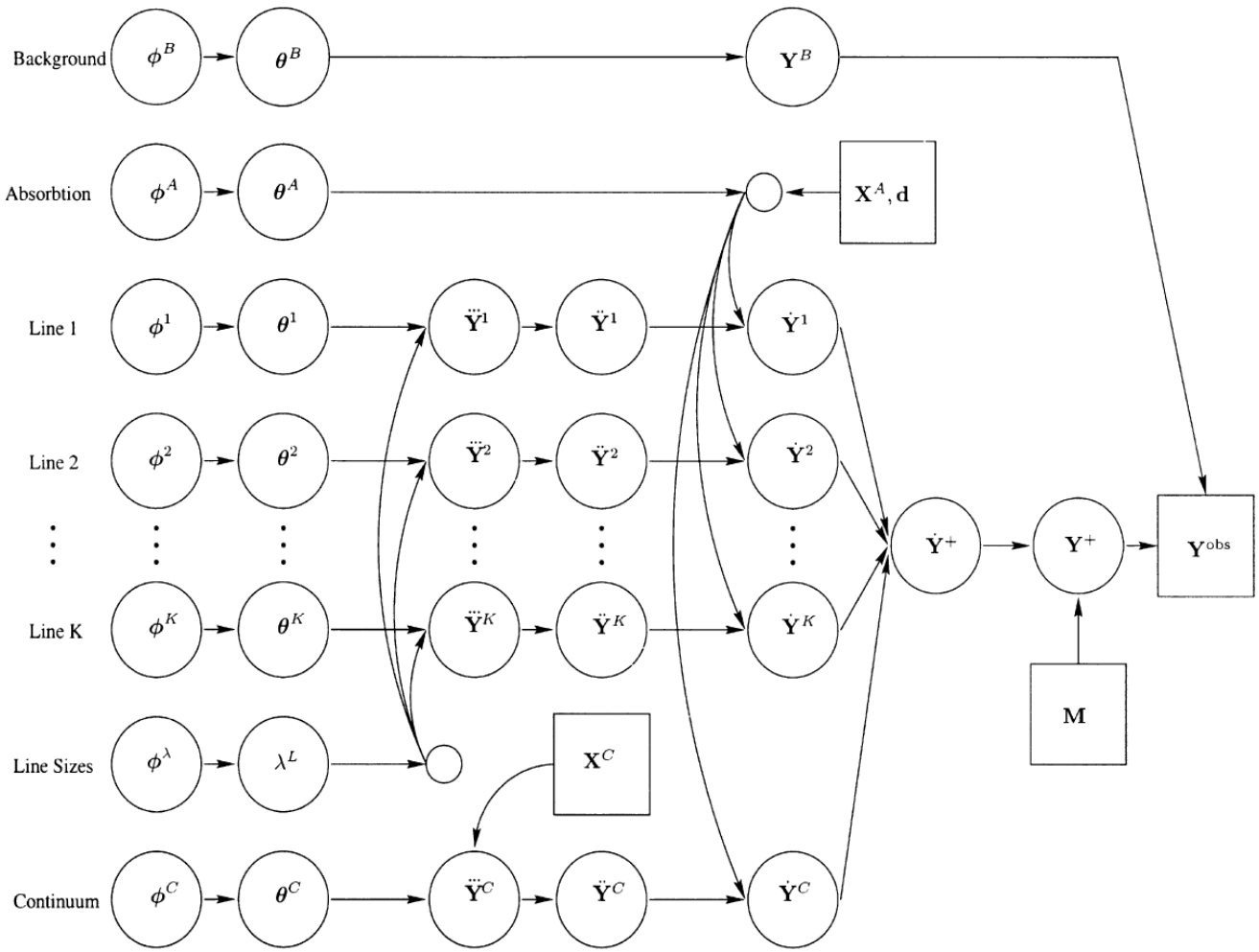
Cray Supercomputer



NCSA

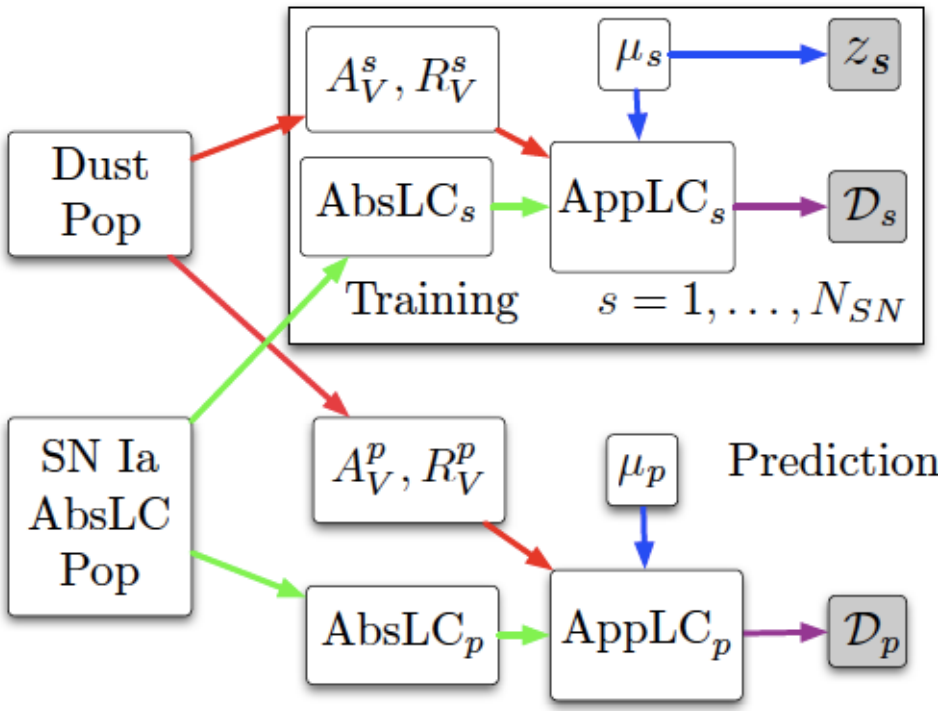
Modern cosmic MLMs

CHASC's x ray spectroscopy model, 2001

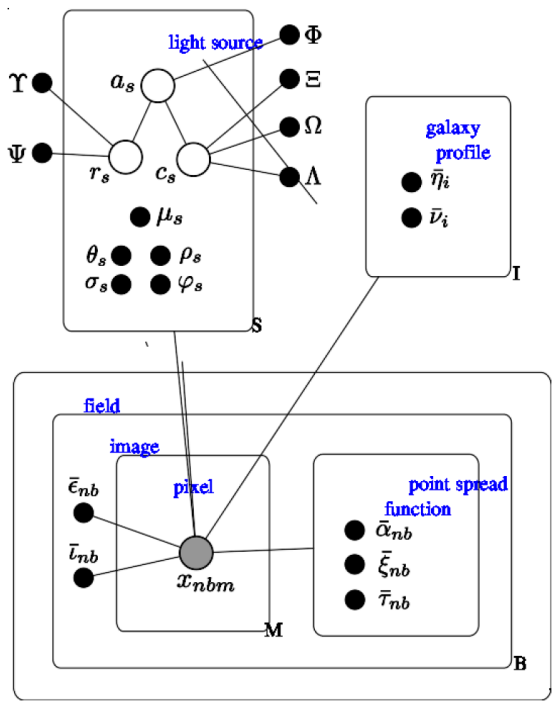


SCMA 5 & 6

BayeSN (Mandel 2011, 2016)



Celeste (McAuliffe 2016)



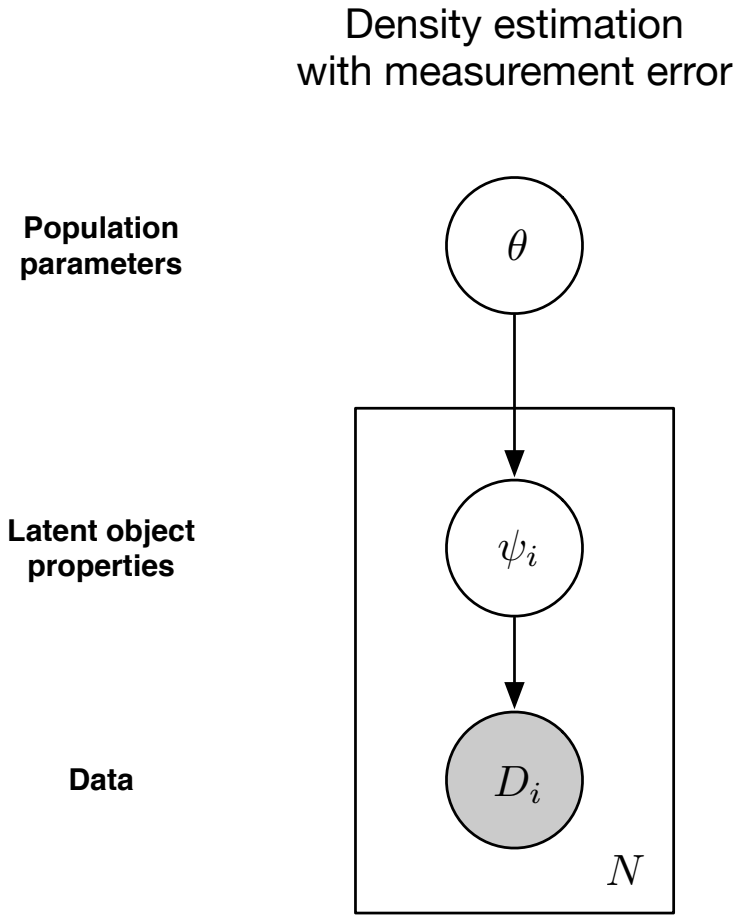
Menu

1. **Catalogs: Report likelihood functions**
2. **Point estimates: Fogetaboutit!**
3. **Population distributions: Multiplying vs. summing**
4. **Low-D MLMs: Cubature vs. importance sampling**
5. **Functional data analysis: Demographics of light curves & spectra**

Menu

1. **Catalogs: Report likelihood functions**
2. Point estimates: Fogetaboutit!
3. Population distributions: Multiplying vs. summing
4. Low-D MLMs: Cubature vs. importance sampling
5. Functional data analysis: Demographics of light curves & spectra

Basic cosmic demography



For $\psi_i = z_i$: Galaxy redshift distribution function

For $\psi_i = F_i$: Number-size dist'n, number counts, log N–log S

Building catalogs

Catalogs enable the community to do science based on an imposing dataset without requiring analysis/modeling of the raw data

Catalog content (data summaries, software, documentation) should:

- Enable sound science with minimized effort
  Avoid doing things that users must undo!
- Discourage misuse/misinterpretation
  Don't make the catalog look like something it isn't!

Catalog builders should do demographic science with their data—they may be the best people to do so—but they should do it *elsewhere*!

From SCMA 5 (and earlier!)

A crazy position?

Since no estimates are useful for all purposes, to make survey information most useful to future investigators, *survey catalogs should not report estimates* of source properties

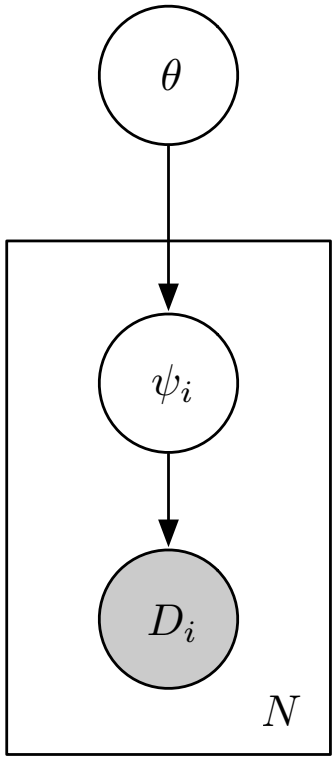
The goal of a survey catalog should be to enable analysts to evaluate likelihoods for diverse models of the survey data $\rightarrow$ catalogs should report *summaries of individual source (marginal) likelihood functions*

This should be done both for detections *and nondetections* (requires more information than "upper limits")

Goal: Enable *chains* of discovery (c.f. Weinberg's remarks on archival value of data)

See arXiv:1208.3036

Likelihoods are what MLMs require (not posteriors!)



A graphical (multilevel/hierarchical) model specifies the joint dist'n for data and parameters (pop'n and latent):

$$\begin{aligned} p(\theta, z, D) &= p(\theta) \prod_{i=1}^{N} p(z_i|\theta)\, p(D_i|z_i) \qquad || \; M \\ &\propto \pi(\theta) \prod_{i} f(z_i;\theta)\, \ell_i(z_i) \end{aligned}$$

with *member likelihood functions*

$$\ell_i(z_i) \;\propto\; p(D_i|z_i)$$

Bayes's theorem gives the posterior for all params:

$$p(\theta, z|D) \;=\; \frac{p(\theta, z, D)}{p(D)} \;\propto\; p(\theta, z, D)$$

$$p(\theta, z|D) \quad \propto \quad \pi(\theta) \prod_i f(z_i;\theta)\, \ell_i(z_i)$$

Marginal posterior for pop'n params:

$$p(\theta|D) \quad \propto \quad \pi(\theta) \prod_i \int dz_i \; f(z_i;\theta)\, \ell_i(z_i)$$

Marginal posterior for member properties ("hierarchical Bayes"):

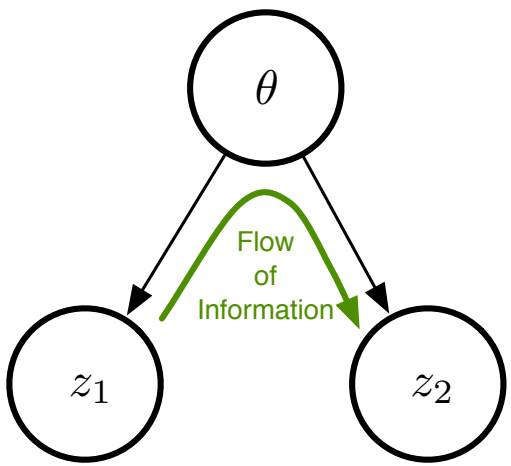
$$\begin{aligned} p(z|D) \quad &= \quad \int d\theta \, p(\theta, z|D) \\ &\propto \quad \left[\prod_i \ell_i(z_i)\right] \times \int d\theta \, \pi(\theta) \prod_i f(z_i;\theta) \\ &\equiv \quad \left[\prod_i \ell_i(z_i)\right] \times F(z_1, \ldots, z_N) \end{aligned}$$

$$p(z|D) \quad \propto \quad \left[\prod_i \ell_i(z_i)\right] \times F(z_1, \ldots, z_N)$$

where $F(z_1, \ldots, z_N)$ is an exchangeable joint prior for the member properties (deFinetti's theorem!)—*correlated*

The z_i are all dependent because the population dist'n is being learned from all of them

The least precisely measured properties are the most dependent—and usually the most numerous



Providing posteriors necessitates de-prioring. . .

Suppose the cataloger provides posteriors based on a pop'n dist'n $h(z;\psi)$, *with a fixed value of* ψ (a priori, or a point estimate)

$$p_i(z_i|\psi, M') \propto h(z_i;\psi)\,\ell_i(z_i)$$

The analyst using M with $f(z;\theta)$ now must "de-prior":

$$p(\theta, z|D) \quad \propto \quad \pi(\theta)\prod_i f(z_i;\theta)\,\frac{p_i(z_i|\psi, M')}{h(z_i;\psi)}$$

Okay, that's doable—but a hassle

. . . unless the catalog makes it impossible!

Suppose the cataloger is "ambitious" and presents a "fully probabilistic catalog" via hierarchical Bayes, e.g., providing posterior samples from

$$p(z|D, M') \propto \left[\prod_i \ell_i(z_i)\right] \times H(z_1, \ldots, z_N)$$

Now it's not easy to recover the member likelihoods!

If N is uncertain, it's harder still—"label switching problem"

There are ways to work around this—but let's try to *avoid* it

Separate cataloging from scientific analysis, and have catalogs report quantities that are as easy as possible to use

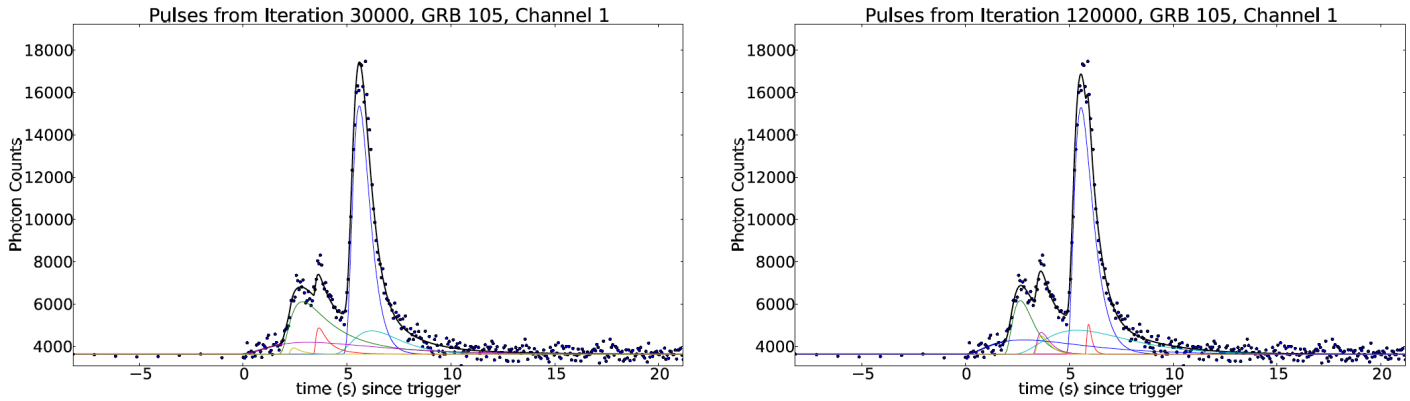
This will sometimes not be easy! E.g., crowded fields. . .

Rethink detection: Rather than a source detection decision boundary, there should be decision boundaries that determine what type of likelihood summary is associated with each potential candidate source location (See Budavári, Szalay, TL forthcoming)

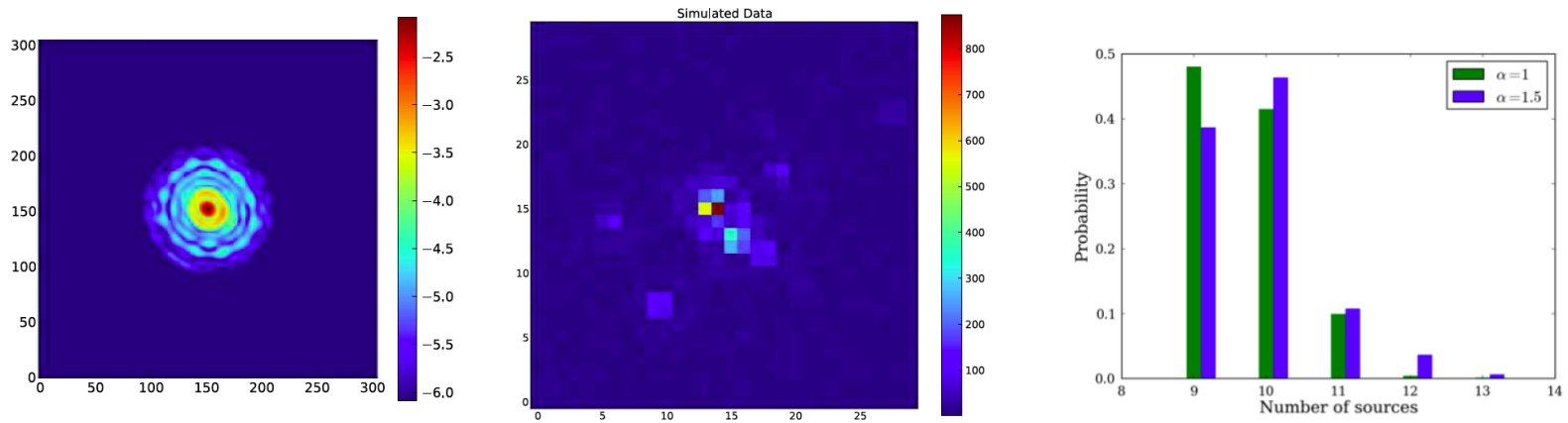
We (Broadbent, Wolpert, TL, Hakkila) like hierarchical, trans-dimensional modeling—for specific science goals. It tends to be inflexible for re-use.

Bayesian droplets (Lévy adaptive regression kernels)

BATSE GRB droplet decomposition



HST image droplet decomposition (simulated)



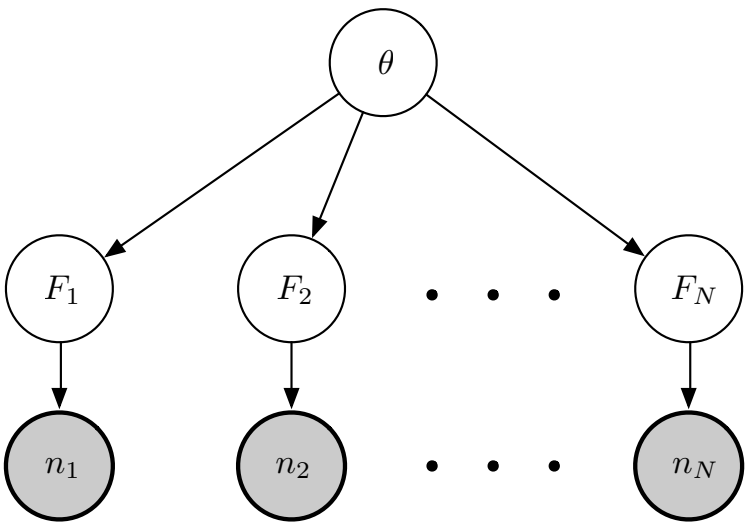
Menu

1. Catalogs: Report likelihood functions
2. **Point estimates: Fogetaboutit!**
3. Population distributions: Multiplying vs. summing
4. Low-D MLMs: Cubature vs. importance sampling
5. Functional data analysis: Demographics of light curves & spectra

A conjugate MLM: Gamma-Poisson

Goal: Learn a flux dist'n from photon counts

Qualitative



Quantitative

Population parameters

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

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

Source properties

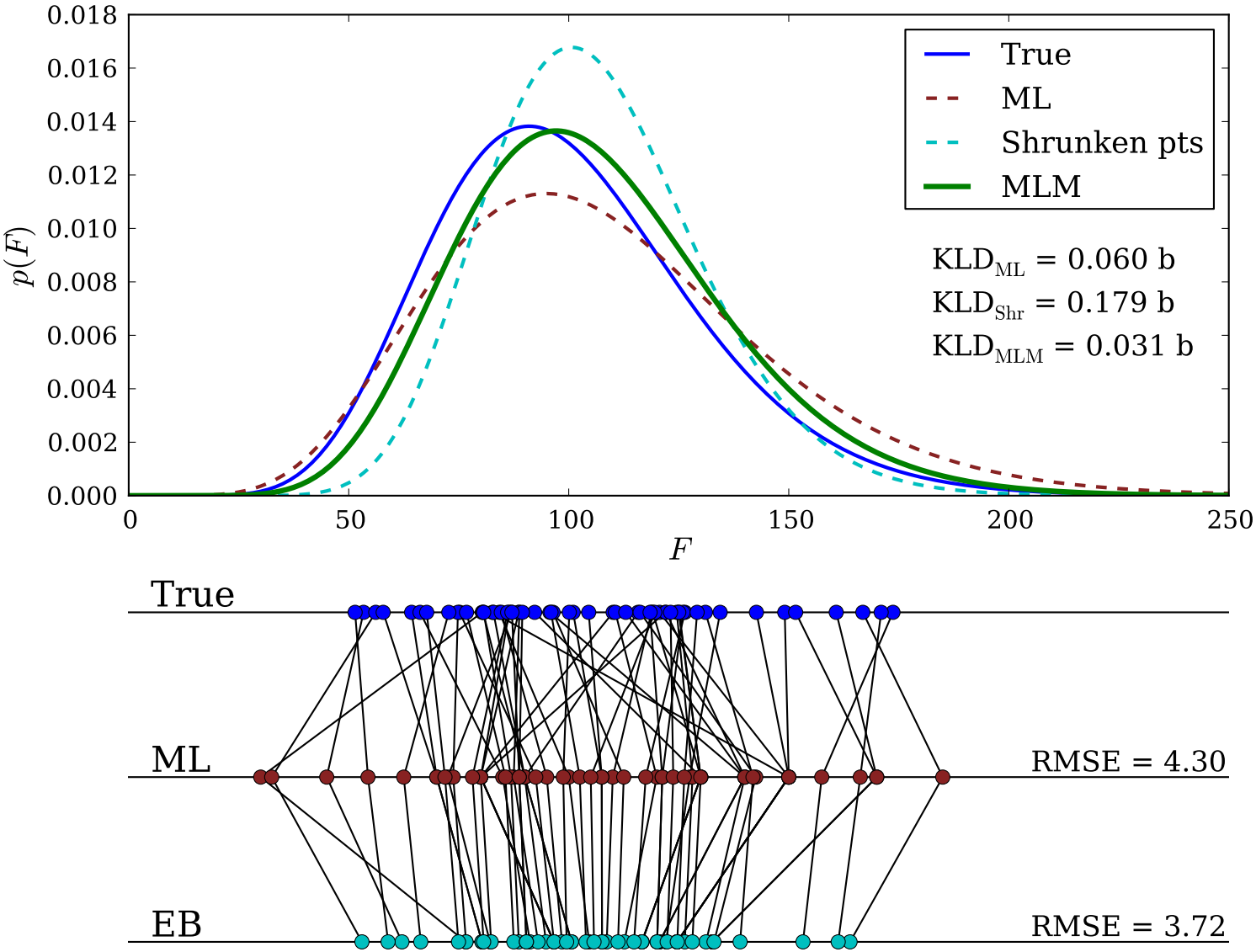
$$p(F_i|\theta) = \mathrm{Gamma}(F_i|\theta)$$

Observed data

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

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

Gamma-Poisson 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 data catalogs as providing
descriptions of member likelihood functions,
not "estimates with errors"

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

Menu

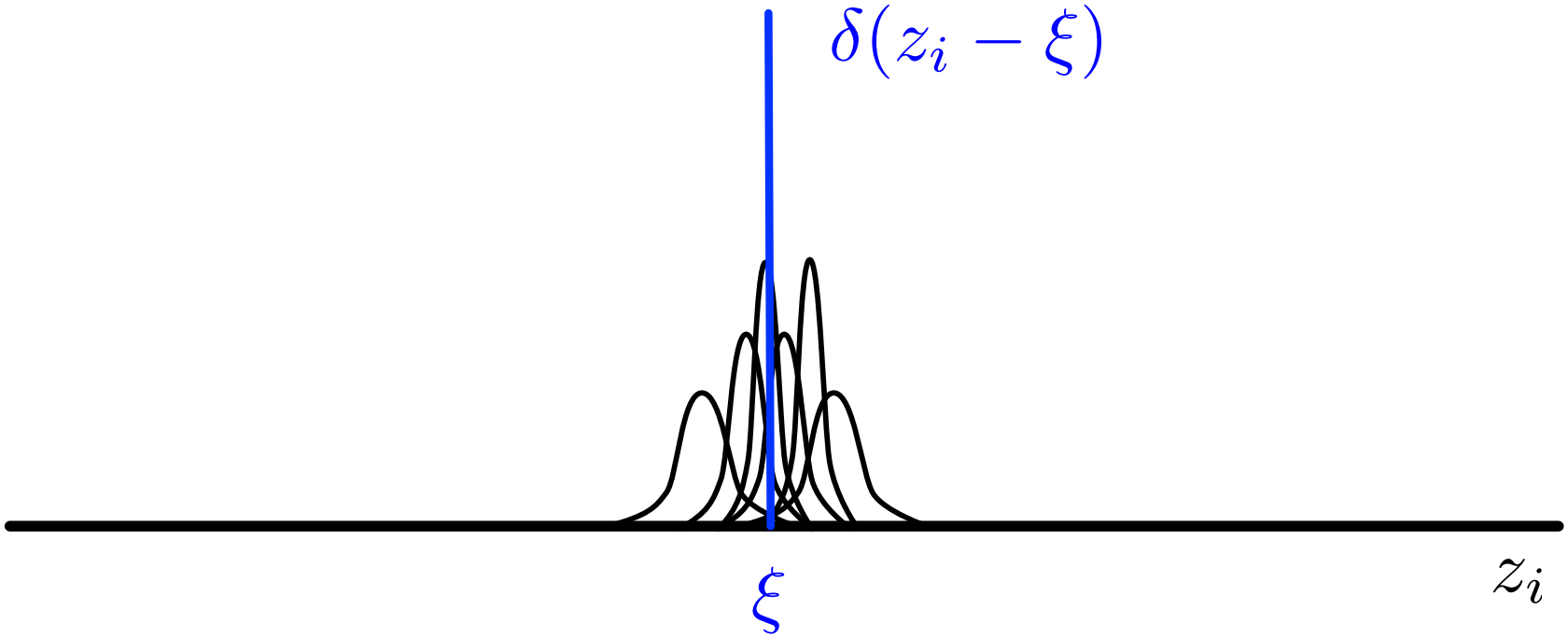
1. Catalogs: Report likelihood functions
2. Point estimates: Fogetaboutit!
3. **Population distributions: Multiplying vs. summing**
4. Low-D MLMs: Cubature vs. importance sampling
5. Functional data analysis: Demographics of light curves & spectra

From likelihoods to distributions

Delta-function model

Suppose all galaxies are at the same (*unknown*) redshift, ξ

How to use the member likelihoods to infer ξ?

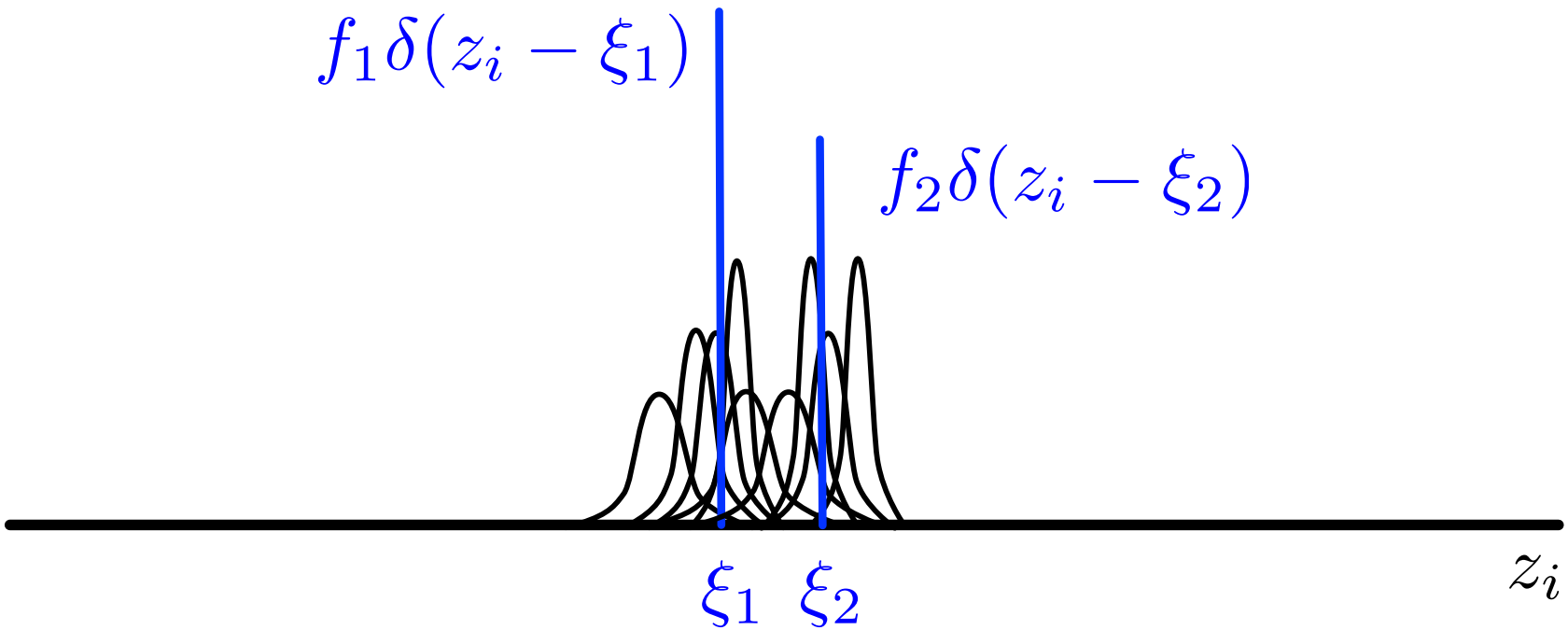


The population parameter is $\theta = \xi$

$$
\begin{aligned}
p(\xi, z|D) &\propto \pi(\xi) \prod_i \delta(z_i - \xi)\, \ell_i(z_i) \\
p(\xi|D) &\propto \pi(\xi) \prod_i \int dz_i\, \delta(z_i - \xi)\, \ell_i(z_i) \\
&= \pi(\xi) \prod_i \ell_i(\xi)
\end{aligned}
$$

Sum of deltas

Now consider a sum of deltas at *fixed* locations ξ_j, but with *unknown amplitudes* f_j to be estimated



The population parameters are $\theta = f \equiv \{f_j\}$ for $j = 1$ to M:

$$\begin{aligned} p(f, z|D) &\propto \pi(f) \prod_i \left[f_1 \delta(z_i - \xi_1) + f_2 \delta(z_i - \xi_2) + \cdots \right] \ell_i(z_i) \\ p(f|D) &\propto \pi(f) \prod_i \left[f_1 \ell_i(\xi_1) + f_2 \ell_i(\xi_2) + \cdots \right] \\ &= \pi(f) \prod_i \left[\sum_j f_j \, \ell_i(\xi_j) \right] \end{aligned}$$

A product-of-sums, with sums corresponding to the $\ell_i(z_i)$ integrals

Product-of-sums:

$$p(f|D) \;=\; \pi(\xi)\prod_i \left[f_1\ell_i(\xi_1) + f_2\ell_i(\xi_2) + \cdots\right]$$

Gather terms, or use the law of total probability, to get a sum-of-products representation, summing over partitions ϖ of the members into sets S_j of N_j objects assigned to component j:

$$p(f|D) \;\propto\; \pi(f)\sum_{\varpi}\left[\frac{N!}{N_1!\cdots N_M!}f_1^{N_1}\cdots f_M^{N_1} \times \left(\prod_{i\in S_1}\ell_i(\xi_1)\right)\cdots\left(\prod_{i\in S_M}\ell_i(\xi_M)\right)\right]$$

A sum over ways to slice the member likelihoods and assign the slices to redshift "bins"

The "sum the uncertainties" intuition is sound, if you are careful over exactly what alternatives you are summing over—assignments of objects to redshift intervals

Menu

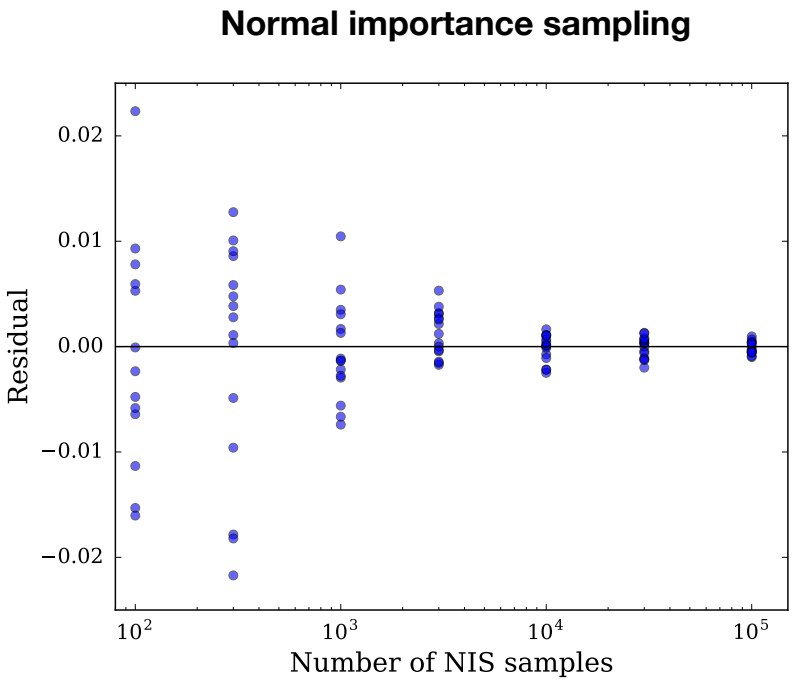
1. Catalogs: Report likelihood functions
2. Point estimates: Fogetaboutit!
3. Population distributions: Multiplying vs. summing
4. **Low-D MLMs: Cubature vs. importance sampling**
5. Functional data analysis: Demographics of light curves & spectra

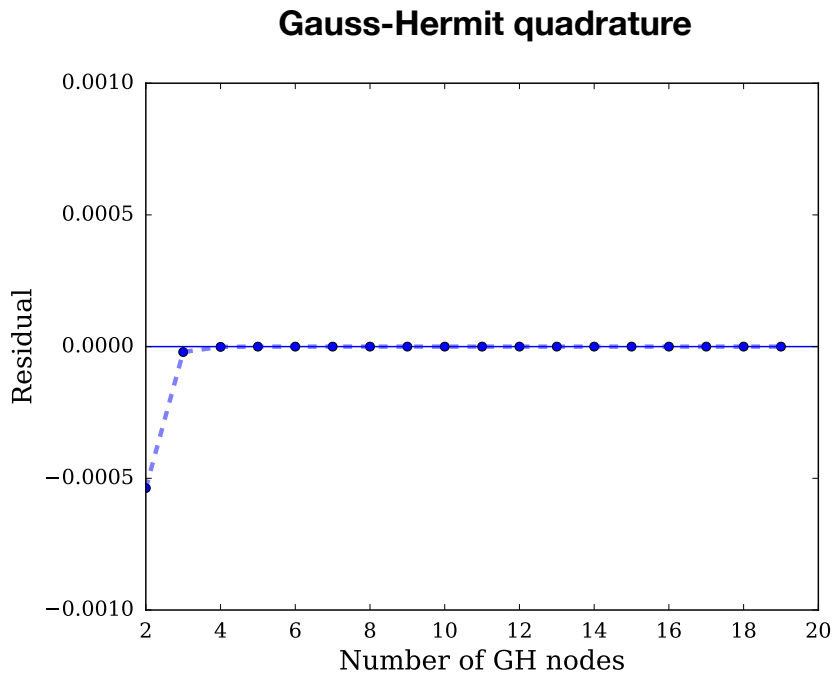
1-D lower-level MLM marginalization

For a Gaussian member likelihood function and power-law $f(z;\theta)$, compute:

$$\int dz_i \; f(z_i;\theta)\,\ell_i(z_i)$$

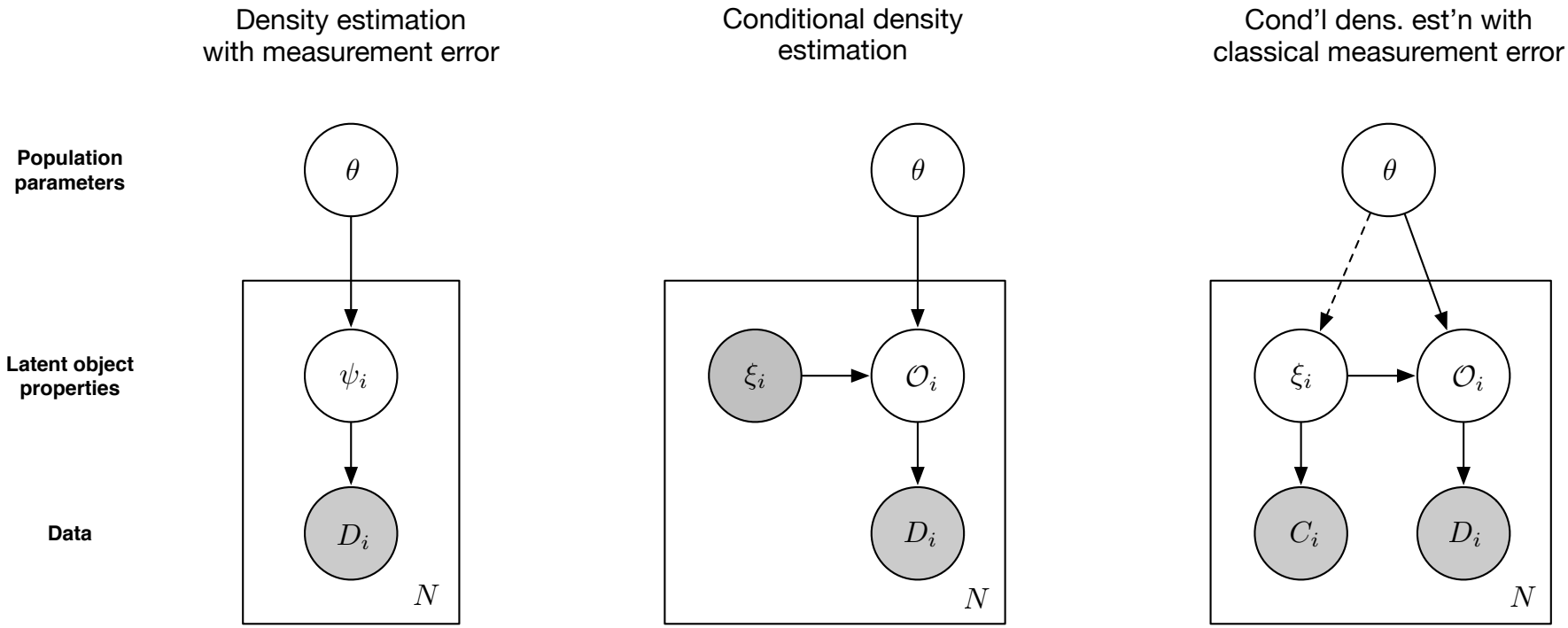
Compare normal-kernel importance sampling (i.e., using posterior samples based on flat interim prior) and Gauss-Hermite quadrature





CUDAHM for big-data single-plate MLMs

Tamás Budavári, Brandon Kelly, TL, János Szalai-Gindl



Conditional independence $\Rightarrow$ plates are "embarassingly parallel":

- Metropolis-with-Gibbs: Sample θ on CPU, member properties on GPU using Robust Adaptive Metropolis (RAM)
- Quadrature and cubature rules for 1-D to 3-D member properties

Menu

1. Catalogs: Report likelihood functions
2. Point estimates: Fogetaboutit!
3. Population distributions: Multiplying vs. summing
4. Low-D MLMs: Cubature vs. importance sampling
5. **Functional data analysis: Demographics of light curves & spectra**

Functional Data Analysis (FDA)

LSST: the greatest ever movie of the Universe and associated astro-statistical challenges

Željko Ivezić (Bill)

Actually the greatest, *weirdest* time-lapse movie. . .

- Each frame is one color (filter)—colors not synchronized
- Frame rate is highly uneven; large gaps
- Patches of the sky observed differently
- Many dim, noisy sources

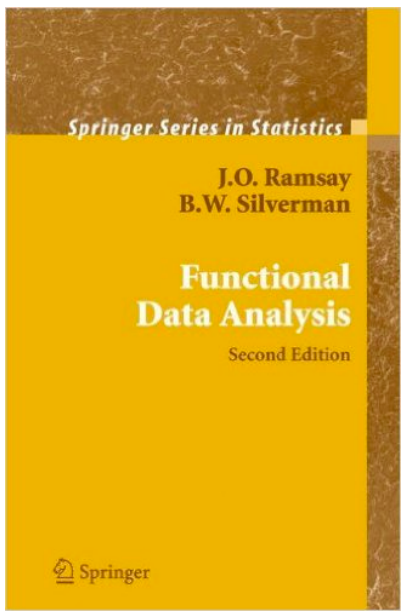
Data are multiband, irregularly sampled,asynchronous, light curves; with heteroskedastic & asymmetric errors; censored

Functional Data Analysis (FDA)

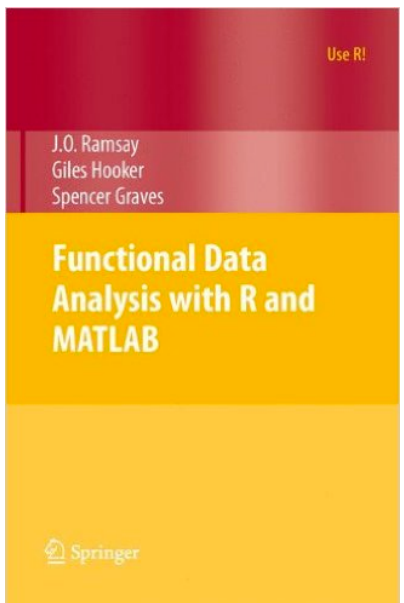
FDA = statistical methods tailored for learning, classifying, predicting with *populations of functions*

Frequentist FDA: Registration, functional PCA, functional mixed effects, function-on-function regression...

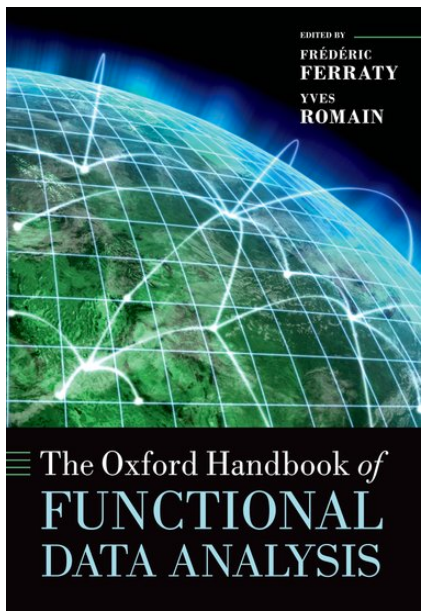
Bayesian FDA: Modeling strategies for multilevel models for functional data (parametric & nonparametric)



1997, 2005



2009



2011