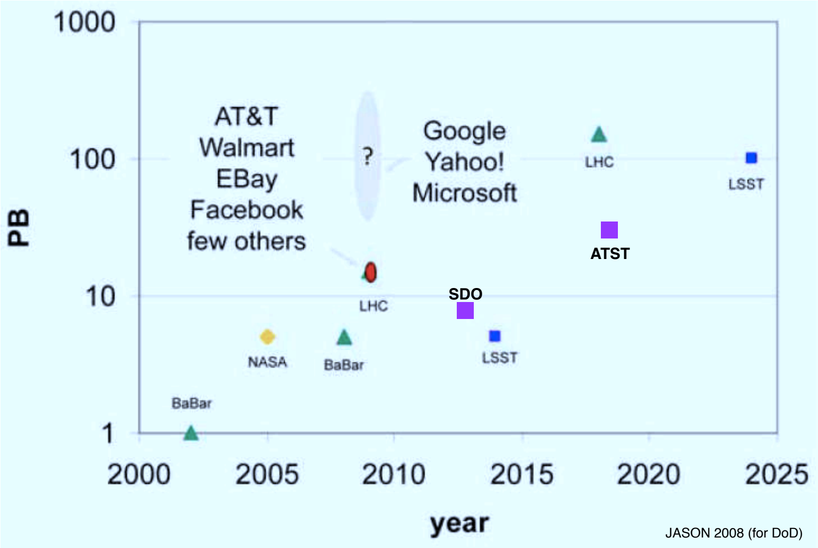
Big Data:
Handle With Care!

Tom Loredo
Dept. of Astronomy, Cornell University

17 Feb 2012 — SolarStat @ CfA

“Big data” problems are ubiquitous

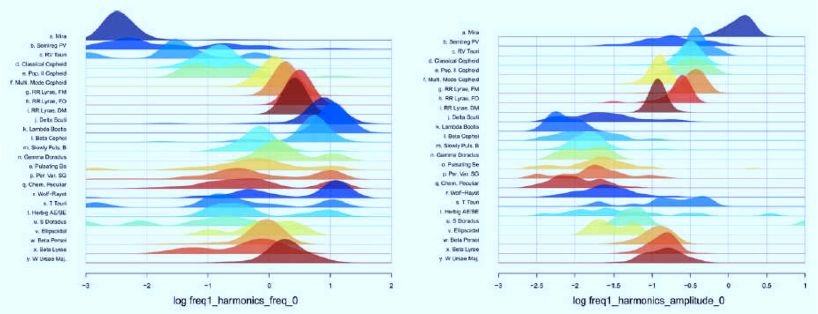


JASON 2008 (for DoD)

Solar/LSST points of contact

- LSST will produce high-dimensional data products:
  - Multicolor images
  - *functional data* for objects (stars, galaxies, minor planets)
- Real-time/streaming analysis crucial
- Co-located processing resources for user community
- Changing perspectives:
  - Populations of light curves
  - Feature-based data processing

UC Berkeley group (Richards et al. 2011)



Used to compare various classifiers (random forest most accurate but too slow!)

Spurring growth of emerging subdisciplines: Astrostatistics, astroinformatics

Asymptopia is tantalizing

Asymptopia
by Peter Sprangers

Come with me and we shall go
a place that only *n* has known

a kingdom distant and sublime
whose ruler is the greatest prime

a land where infinite sums can rest
and undergrads shall take no test

a place where every child you see
writes poems about the C.L.T.

where cdf's converge to one
and every day is filled with sun.

where we can jump time's famous hurdle
and watch Achilles beat the turtle

and every stat plucked from a tree
is, without proof, U.M.V.U.E.

where joy o'erflows the cornucopia
in this, the land of Asymptopia.

Asymptopia is tantalizing

Asymptopia
by Peter Sprangers

Come with me and we shall go
a place that only *n* has known

a kingdom distant and sublime
whose ruler is the greatest prime

a land where infinite sums can rest
and undergrads shall take no test

a place where every child you see
writes poems about the C.L.T.

where cdf's converge to one
and every day is filled with sun.

where we can jump time's famous hurdle
and watch Achilles beat the turtle

and every stat plucked from a tree
is, without proof, U.M.V.U.E.

where joy o'erflows the cornucopia
in this, the land of Asymptopia.

Punishment of Tantalus

Agenda

❶ **Big is never big enough**

❷ **Measurement error in number-size distributions**

❸ **Bayesian computation for streaming data**

Outline

1. Big is never big enough

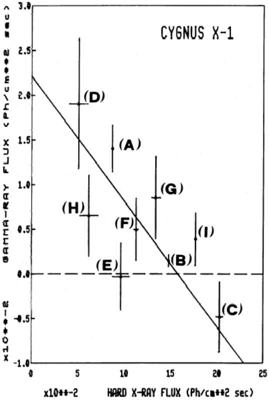
2. Measurement error in number-size distributions

3. Bayesian computation for streaming data

Is N ever large?

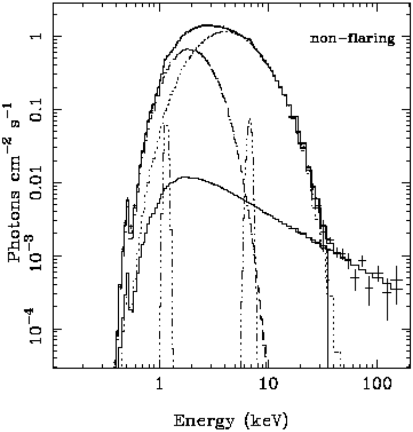
Cyg X-1 γ- vs. X-ray

Bassani$^+$ 1989



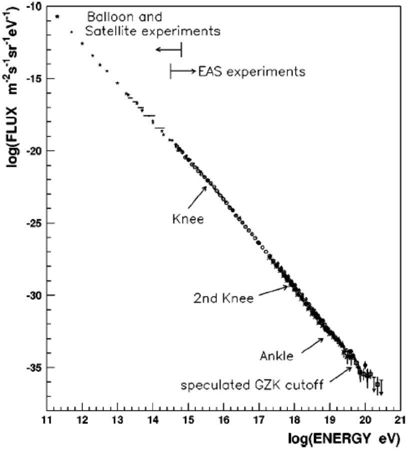
BeppoSAX spectrum, GX 349+2

Di Salvo et al. 2001

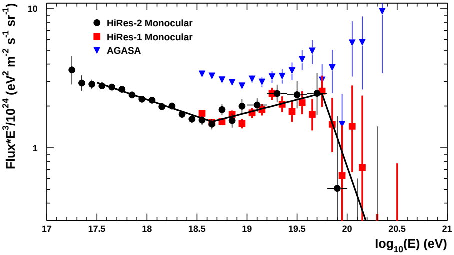


Spectrum of Ultrahigh-Energy Cosmic Rays

Nagano & Watson 2000



HiRes Team 2007



N is never large (enough)

Sample sizes are never large. If N is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once N is 'large enough,' you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc etc). N is never enough because if it were 'enough' you'd already be on to the next problem for which you need more data.

— Andrew Gelman (blog entry, 31 July 2005)

N is never large (enough)

Sample sizes are never large. If *N* is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once *N* is 'large enough,' you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc etc). *N* is never enough because if it were 'enough' you'd already be on to the next problem for which you need more data.

Similarly, you never have quite enough money. But that's another story.

— Andrew Gelman (blog entry, 31 July 2005)

Outline

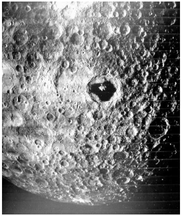
1. Big is never big enough

2. **Measurement error in number-size distributions**

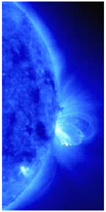
3. Bayesian computation for streaming data

We survey everything!

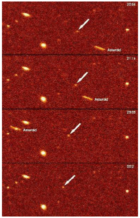
Lunar Craters



Solar Flares

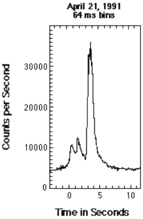


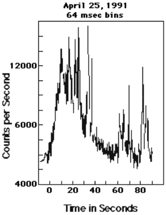
TNOs



Stars & Galaxies

GRBs

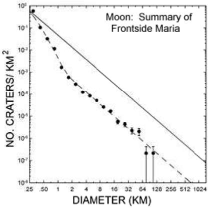




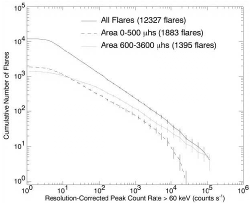
Number-size distributions

aka size-frequency distributions, number counts, log N–log S...

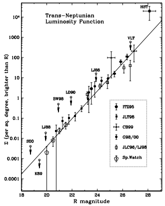
Lunar Craters



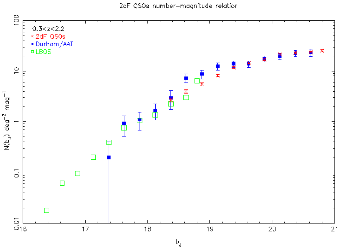
Solar Flares



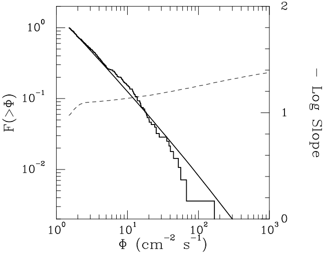
TNOs



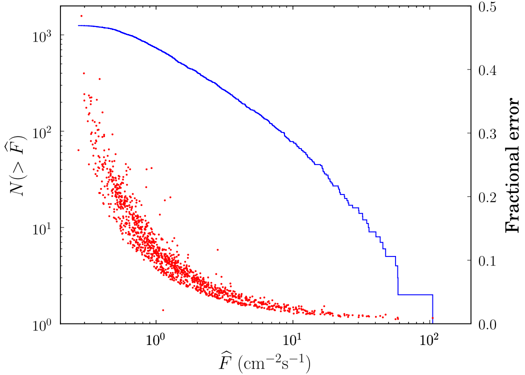
Quasars



GRBs



Selection effects and measurement error

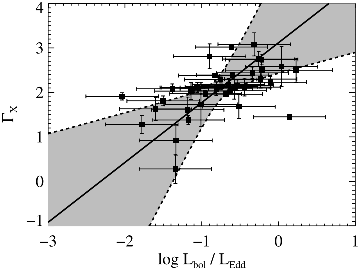


- Selection effects (truncation, censoring) — *obvious* (usually)
  Typically treated by "correcting" data
  Most sophisticated: product-limit estimators

- "Scatter" effects (measurement error, etc.) — *insidious*
  Typically ignored (average out?)

Measurement error for line & curve fitting

QSO hardness vs. luminosity (Kelly 2007, 2011)

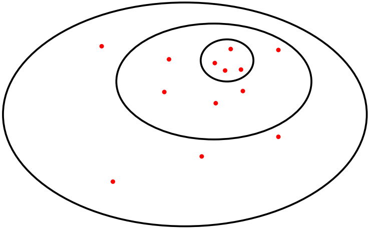


“Regression with easurement error” in statistics refers to the case of errors in both x and y

Accounting For Measurement Error

Introduce latent/hidden/incidental parameters

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

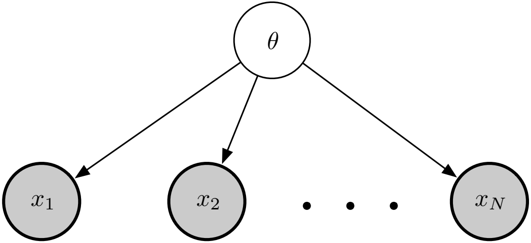


From N precisely measured samples, $\{x_i\}$, we can infer θ from

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

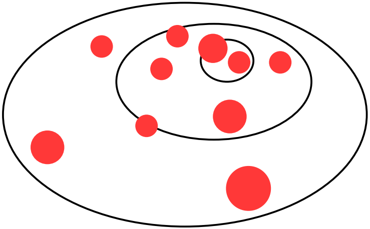
Bayesian graphical model

- Nodes/vertices = uncertain quantities
- Edges specify conditional dependence
- Absence of an edge denotes conditional *in*dependence



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

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



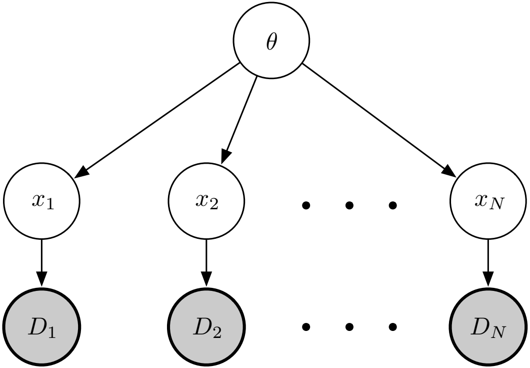
We should somehow incorporate $\ell_i(x_i) = p(D_i|x_i)$

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

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!*

Graphical representation



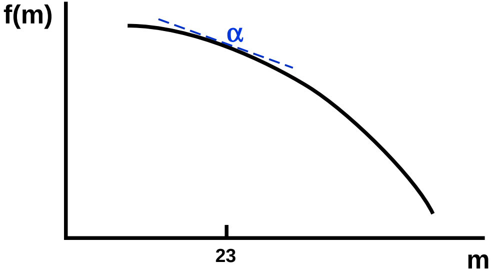
$$p(\theta, \{x_i\}, \{D_i\}) \quad = \quad p(\theta) \prod_i f(x_i|\theta) p(D_i|x_i) = \prod_i f(x_i|\theta) \ell_i(x_i)$$

A two-level *multi-level model* (MLM) or *hierarchical model*

Example—Distribution of Source Fluxes

Measure $m = -2.5\log(\text{flux})$ from sources following a "rolling power law" distribution (inspired by trans-Neptunian objects)

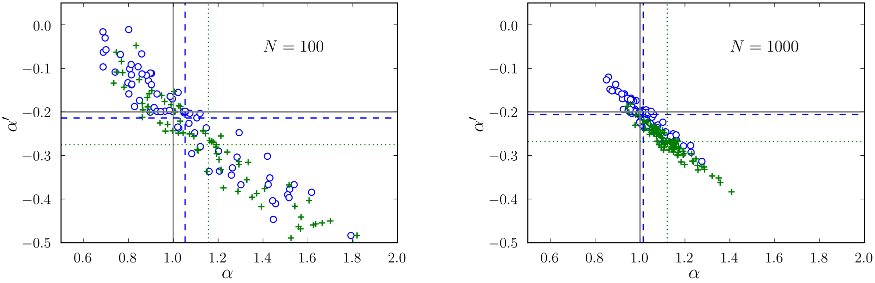
$$f(m) \propto 10^{[\alpha(m-23)+\alpha'(m-23)^2]}$$



Simulate 100 surveys of populations drawn from the same dist'n.
Simulate data for photon-counting instrument, fixed count threshold.
Measurements have uncertainties 1% (bright) to ≈ 30% (dim).

Analyze simulated data with maximum ("profile") likelihood and Bayes.

Parameter estimates from Bayes (circles) and maximum likelihood (crosses):

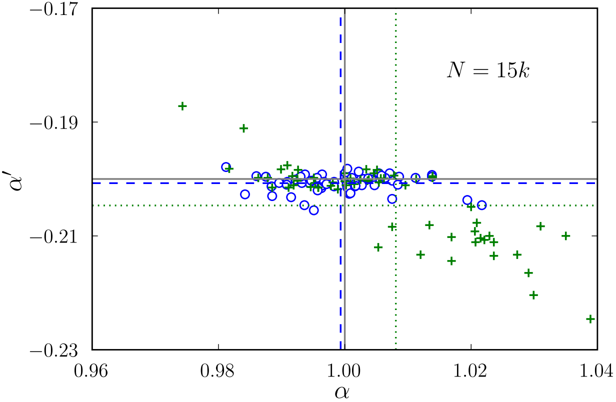


Uncertainties don't average out!

Applications: GRBs (TL & Wasserman 1993+); TNOs (Gladman$^+$ 1998+, Petit$^+$ 2006+)

Smaller measurement error only postpones the inevitable:

Similar toy survey, with parameters to mimic SDSS QSO surveys (few % errors at dim end):



What's going on?

- Implicitly or explicitly, each datum brings with it a new parameter
- Middle level distribution $f(x_i|\theta)$ acts as prior on lower level latent parameters x_i (observables)
- Separate "copies" of the prior → data never overwhelm effect of prior
- Marginalization accounts for volume in $\{x_i\}$ parameter space
- "Mustering and borrowing of strength" (Tukey):
  - Pool information from individuals about upper (pop'n) level θ
  - Pooled information feeds back; x_i estimate is affected by all other $\{x_i\}$ through what they say about θ
  - Shrinkage: esulting x_i estimates are biased but *better* (lower MSE)

Outline

1. Big is never big enough

2. Measurement error in number-size distributions

3. **Bayesian computation for streaming data**

Posterior sampling

Monte Carlo integration wrt posterior distribution:

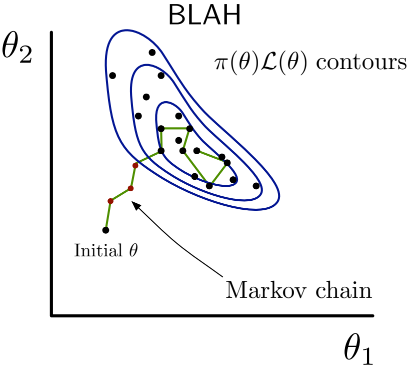
$$\int d\theta\, g(\theta) p(\theta|D) \approx \frac{1}{n} \sum_{\theta_i \sim p(\theta|D)} g(\theta_i) + O(n^{-1/2})$$

When $p(\theta)$ is a posterior distribution, drawing samples from it is called *posterior sampling*:

- *One set of samples* can be used for many different calculations (so long as they don't depend on low-probability events)
- This is the most promising and general approach for Bayesian computation in *high dimensions*—though with a twist (MCMC!)

Challenge: How to build a RNG that samples from a posterior?

Markov chain Monte Carlo (MCMC)



$$\begin{aligned} p(\theta|D) &= \frac{q(\theta)}{Z} \\ q(\theta) &\equiv p(\theta)p(D|\theta) \end{aligned}$$

Metropolis-Hastings algorithms:

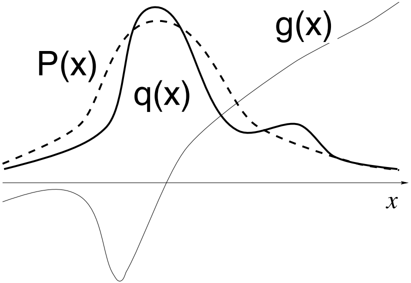
- Propose candidate θ' from *proposal dist'n*
- Accept new/repeat old depending on $q(\theta')/q(\theta)$ (and proposal ratio)

Requires evaluating $q(\theta)$ for every candidate

Importance sampling

$$\int d\theta\; g(\theta) q(\theta) = \int d\theta\; g(\theta) \frac{q(\theta)}{P(\theta)} P(\theta) \approx \frac{1}{n} \sum_{\theta_i \sim P(\theta)} g(\theta_i) \frac{q(\theta_i)}{P(\theta_i)}$$

Choose P to make variance small. (Not easy!)

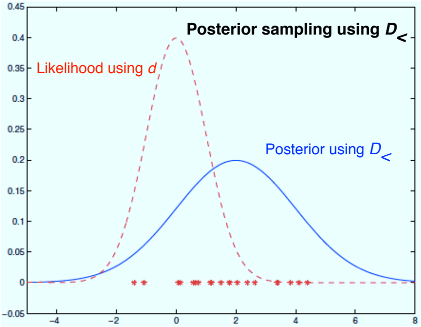


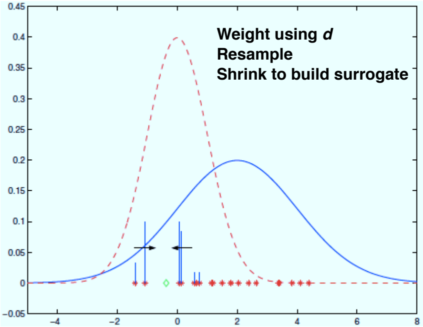
Sequential Monte Carlo for streaming data

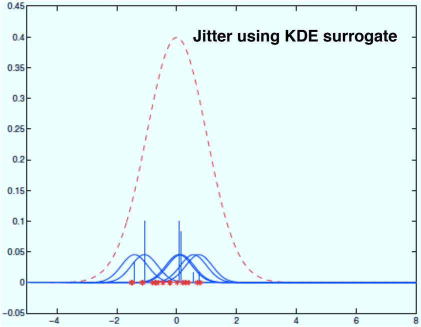
Balakrishnan and Madigan (2006)

- Get posterior samples $\{\theta_i\}$ from initial chunk $D_<$
- Revise samples by bringing in $D_>$ in small chunks d:
  - Weight by new data likelihood $p(d|\theta_i)$
  - Resample via weights (like bootstrap) → degeneracy
  - Jitter via MCMC using modified KDE surrogate for current posterior

Applied to Bayesian logistic regression (7 predictors) and 4×4 Markov transition regression (5 – 20 samples), $N \sim 10^6$


Posterior sampling using $D_<$
Likelihood using d
Posterior using $D_<$
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−4
−2
0
2
4
6
8


Weight using d
Resample
Shrink to build surrogate
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−4
−2
0
2
4
6
8


Jitter using KDE surrogate
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−4
−2
0
2
4
6
8

Main Points

- Solar, stellar & extragalactic astronomy share "big data" problems → Cross fertilization opportunities?
- Statistics may not become simpler with "large *N*"
- Relative importance of measurement error *grows* with sample size; it doesn't "average out"
- MLMs handle such complications by *marginalizing over latent parameters*
- Streaming Bayesian computation via sequential Monte Carlo — a research front

Note: 2012/2013 SAMSI Program on Statistical and Computational Methodology for Massive Datasets (www.samsi.info)