Introduction to Bayesian Inference
Lecture 4:
Multilevel Models for Measurement Error, Basic Bayesian Computation

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

INPE — 16 September 2009

Agenda

1. Bayesian measurement error modeling
   - The Neyman-Scott problem
   - Astrophysical measurement error problems
   - Multilevel models for measurement error
2. Basic Bayesian calculation
   - Large N: Laplace approximations
   - Low d: Cubature, adaptive cubature
   - Available software
   - Closing reflections

Agenda

1. Bayesian measurement error modeling
   - The Neyman-Scott problem
   - Astrophysical measurement error problems
   - Multilevel models for measurement error
2. Basic Bayesian calculation
   - Large N: Laplace approximations
   - Low d: Cubature, adaptive cubature
   - Available software
   - Closing reflections

Measurement Error & Marginalization

Calibrating a noise level

Need to measure several sources with signal amplitudes μ_i, with an "uncalibrated" instrument that adds Gaussian noise with *unknown* but constant σ.

Ideally, either:

- Measure calibration sources of known amplitudes; the scatter of the measurements from the known values allows easy inference of σ.
- Measure one source many times; from many samples we can easily learn both μ_i and σ.

Neyman-Scott problem (1948): Calibrate as-you-go

- No calibration sources are available.
- We have to measure N sources with finite resources, so only a few measurements of each source are available.

The multiple measurements of a single source yield a noisy estimate of σ.

$\longrightarrow$ Pool all the data to learn σ.

Pairs of measurements

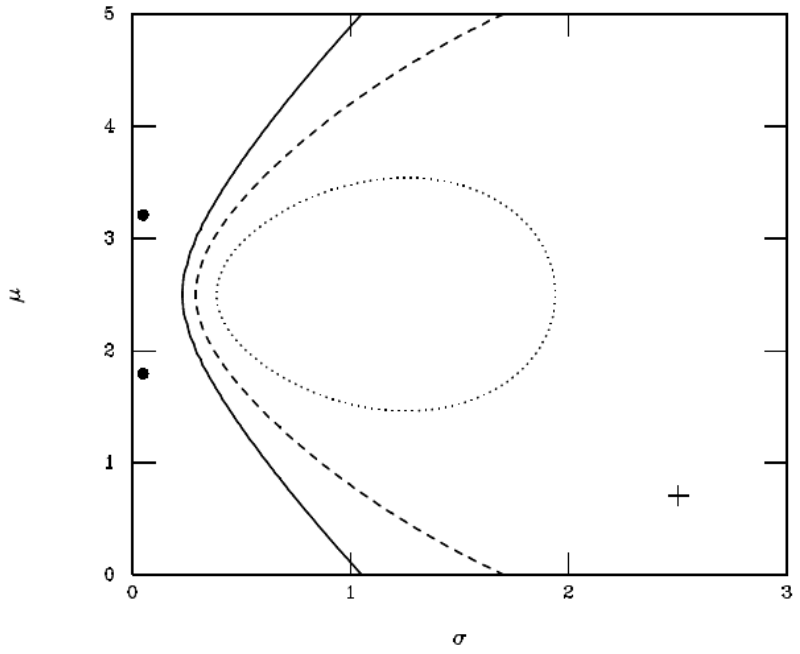
Make 2 measurements (x_i, y_i) for each of the N quantities μ_i.

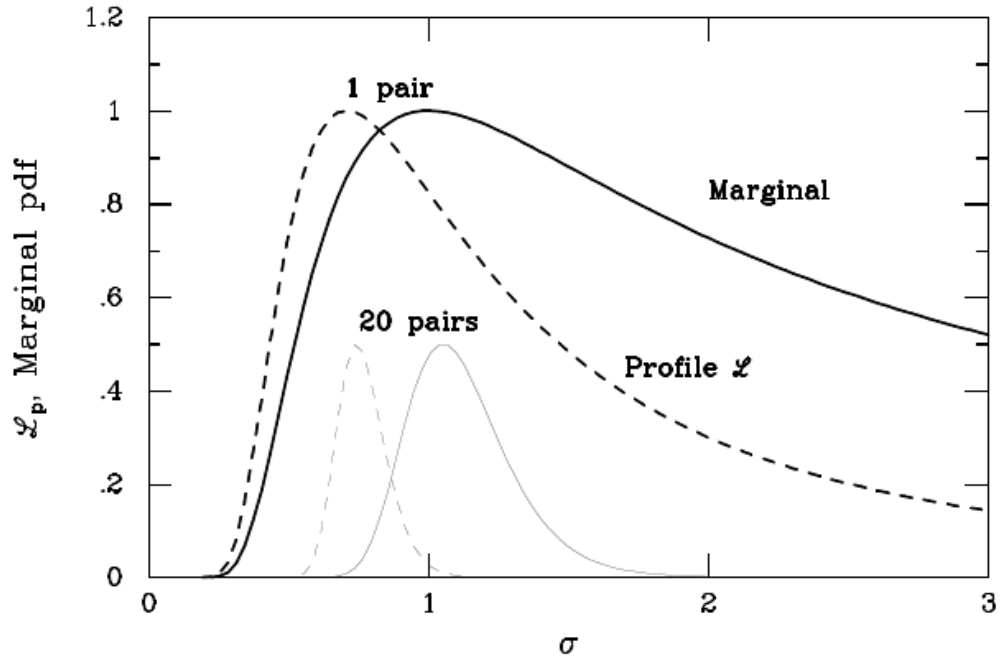
Likelihood:

$$\mathcal{L}(\{\mu_i\}, \sigma) = \prod_i \frac{\exp\left[-\frac{(x_i-\mu_i)^2}{2\sigma^2}\right]}{\sigma\sqrt{2\pi}} \times \frac{\exp\left[-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right]}{\sigma\sqrt{2\pi}}$$

Profile likelihood $\mathcal{L}_p(\sigma) = \max_{\{\mu_i\}} \mathcal{L}(\{\mu_i\}, \sigma)$

Joint & Marginal Results for $\sigma = 1$





The marginal $p(\sigma|D)$ and $\mathcal{L}_p(\sigma)$ differ dramatically!
Profile likelihood estimate converges to $\sigma/\sqrt{2}$.

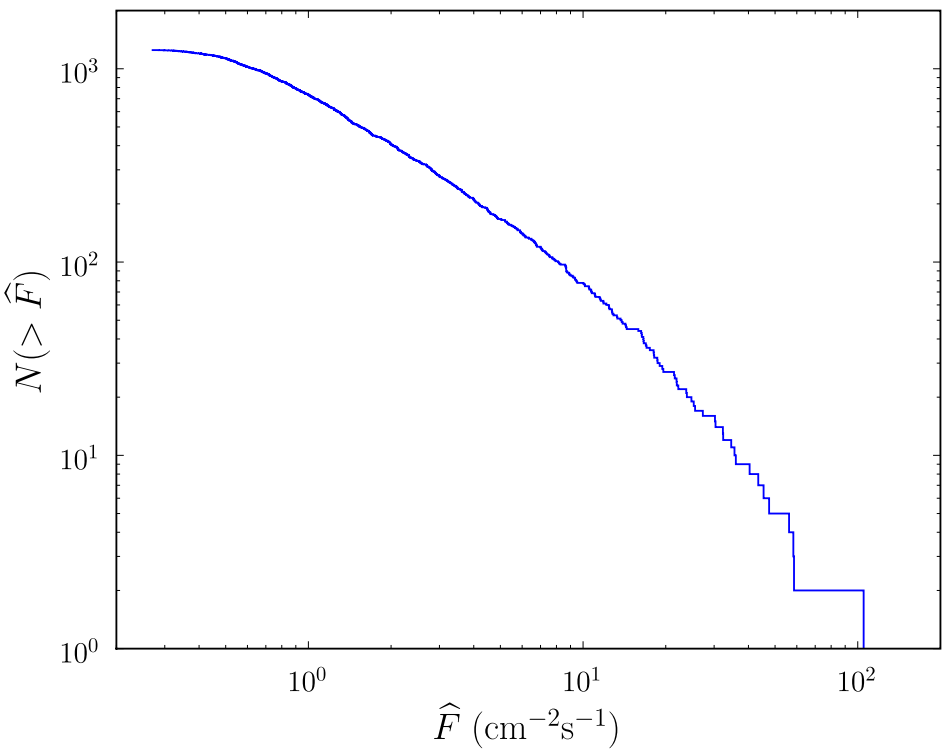
The total # of parameters grows with the # of data.
$\Rightarrow$ *Volumes along μ_i do not vanish as $N \to \infty$.*

Empirical Number Counts Distributions

Star counts, galaxy counts, GRBs, TNOs . . .

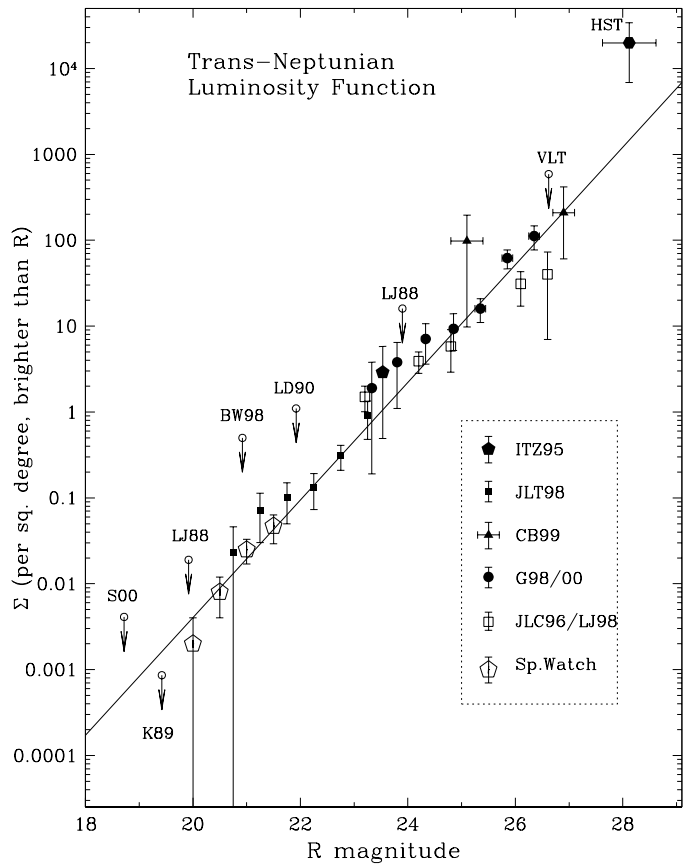
BATSE 4B Catalog (≈ 1200 GRBs)

$F \propto L/d^2$ [$\times$ cosmo, extinct'n]

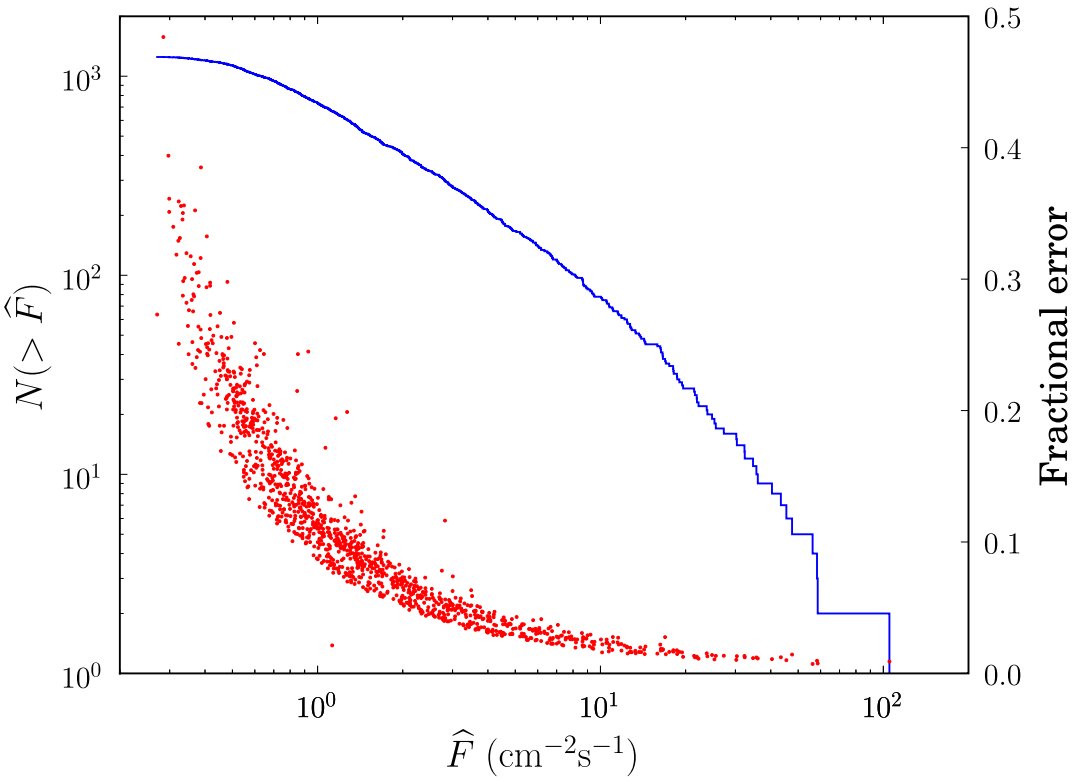


13 TNO Surveys (c. 2001)

$F \propto \nu D^2/(d_\odot^2 d_\oplus^2)$



Selection Effects and Measurement Error

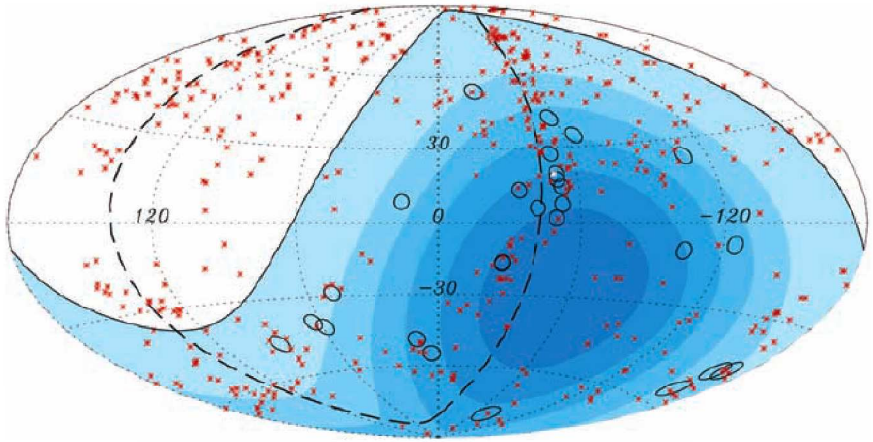


- Selection effects (truncation, censoring) — *obvious* (usually)
  Typically treated by "correcting" data
  Most sophisticated: product-limit estimators

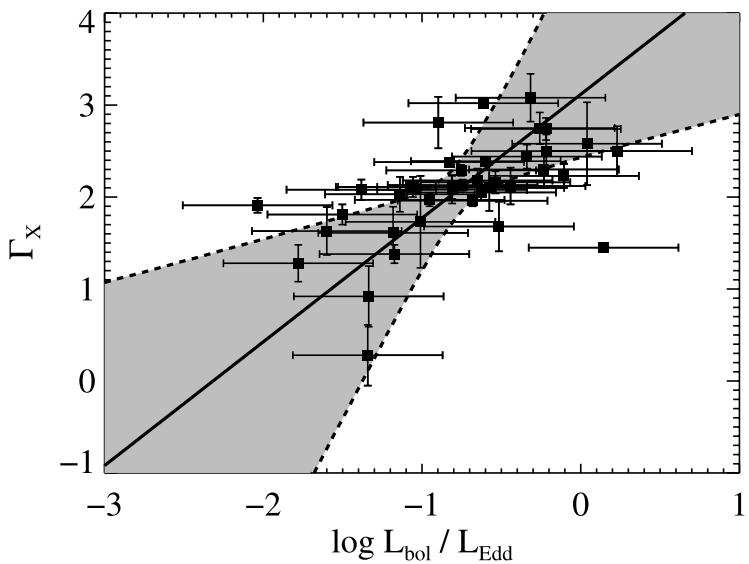
- "Scatter" effects (measurement error, etc.) — *insidious*
  Typically ignored (average out?)

Many Guises of Measurement Error

Auger data above GZK cutoff (Nov 2007)

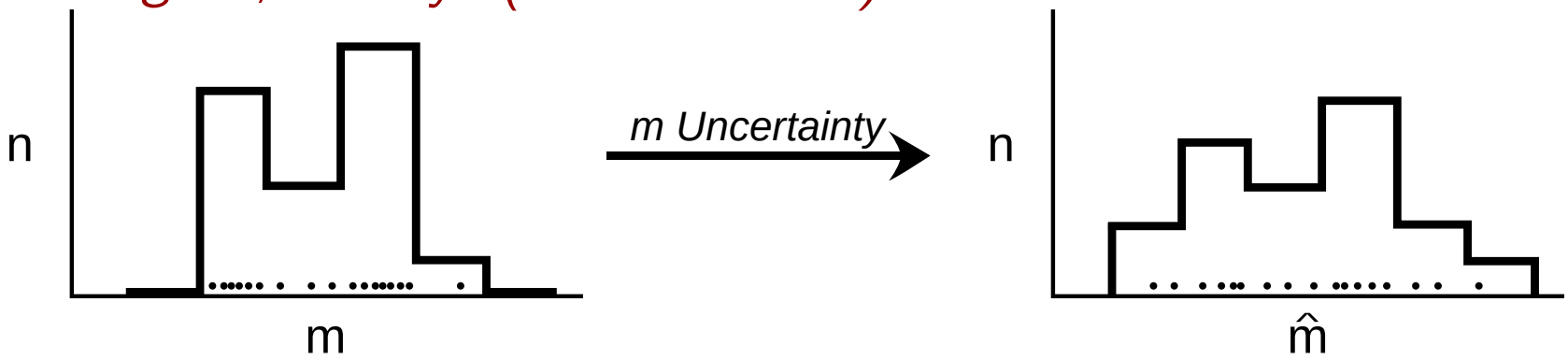


QSO hardness vs. luminosity (Kelly 2007)



History

Eddington, Jeffreys (1920s – 1940)



Malmquist, Lutz-Kelker

- Joint accounting for truncation and (intrinsic) scatter in 2-D data (flux + distance indicator, parallax)
- Assume homogeneous spatial distribution

Many rediscoveries of "scatter biases"

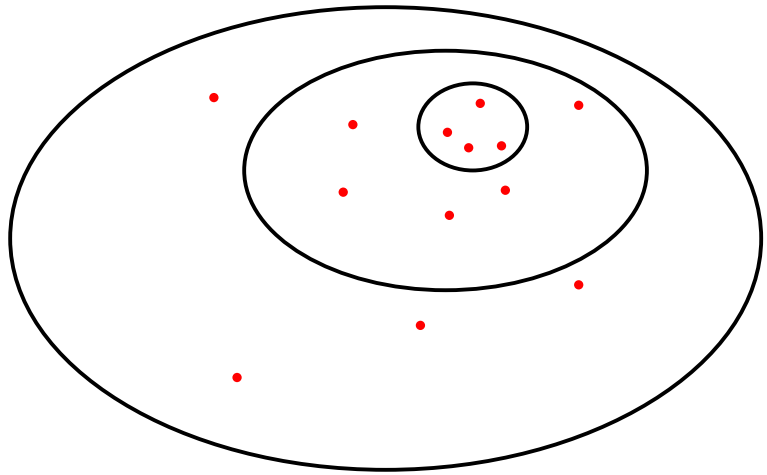
- Radio sources (1970s)
- Galaxies (Eddington, Malmquist; 1990s)
- Linear regression (1990s)
- GRBs (1990s)
- X-ray sources (1990s; 2000s)
- TNOs/KBOs (c. 2000)
- Galaxy redshift dist'ns (2007+)
- ...

(See Loredo 2007, SCMA IV proceedings, for review)

Accounting For Measurement Error

Introduce latent/hidden/incidental parameters

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

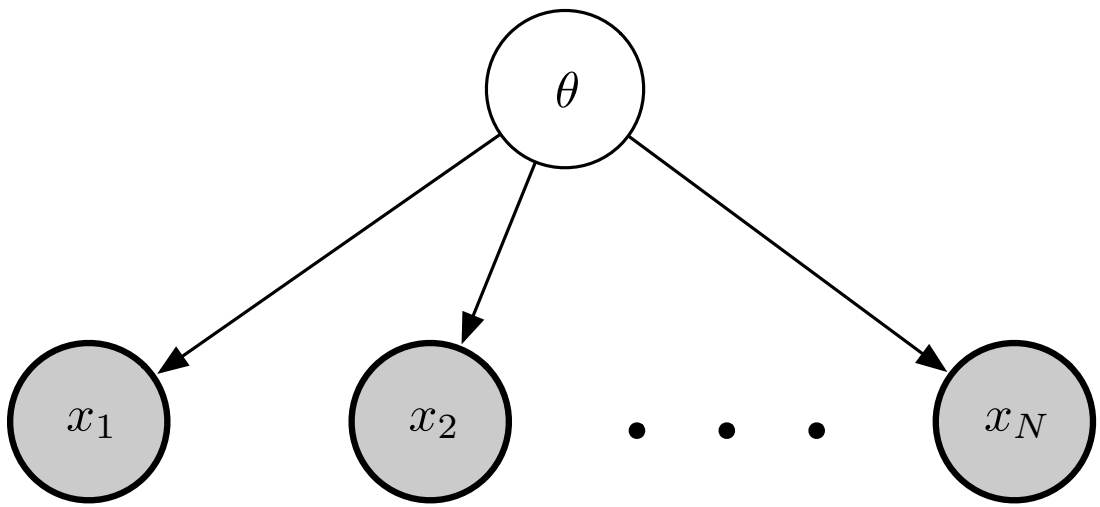


From N precisely measured samples, $\{x_i\}$, we can infer θ from

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

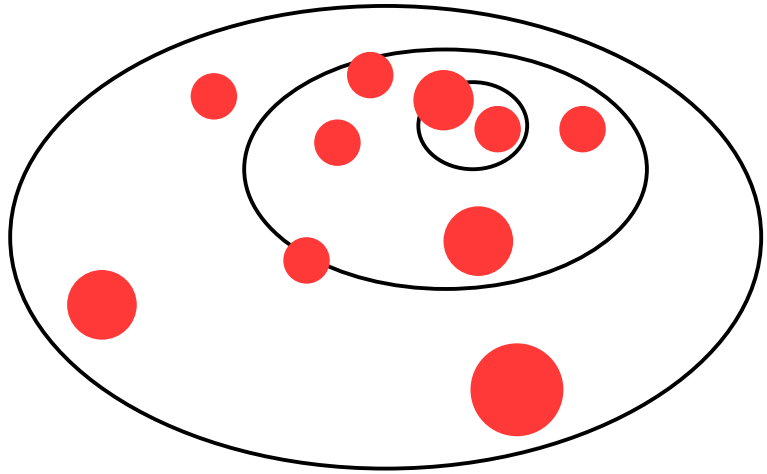
Graphical representation

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



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

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



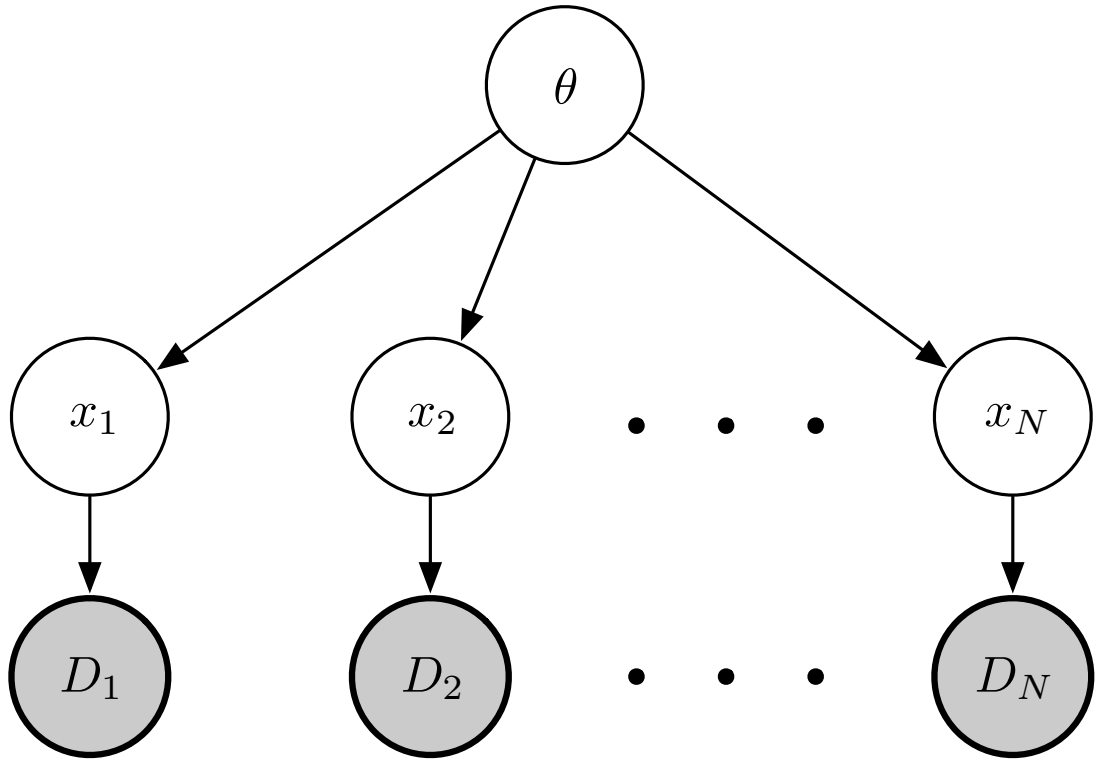
We should somehow incorporate $\ell_i(x_i) = p(D_i|x_i)$

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

Marginalize (sum probabilities) over $\{x_i\}$ to summarize for θ.
Marginalize over θ to summarize results for $\{x_i\}$.

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

Graphical representation



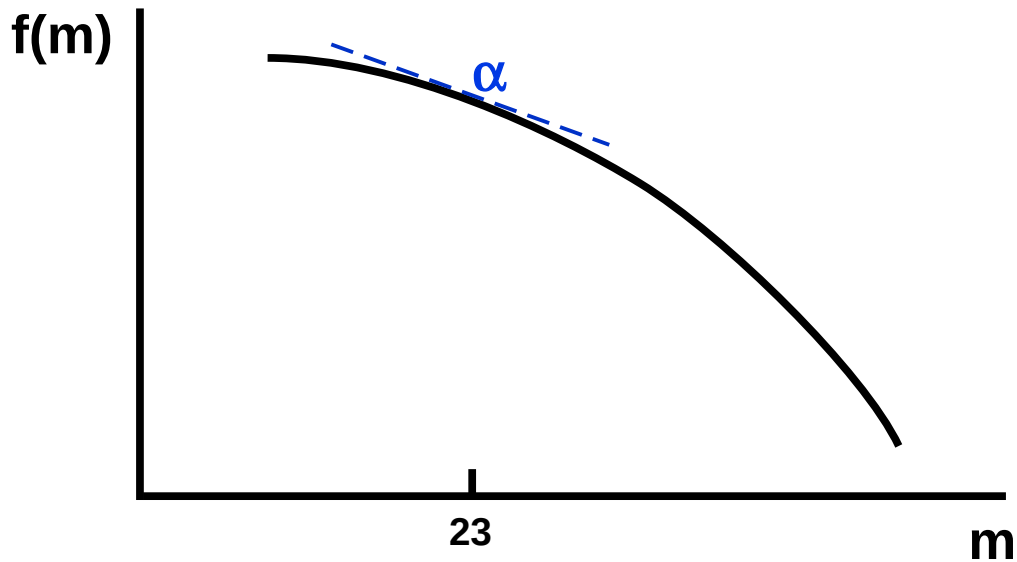
$$\begin{aligned}\mathcal{L}(\theta, \{x_i\}) &\equiv p(\{D_i\}|\theta, \{x_i\}) \\ &= \prod_i p(D_i|x_i) f(x_i|\theta) = \prod_i \ell_i(x_i) f(x_i|\theta)\end{aligned}$$

A two-level *multi-level model* (MLM).

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)

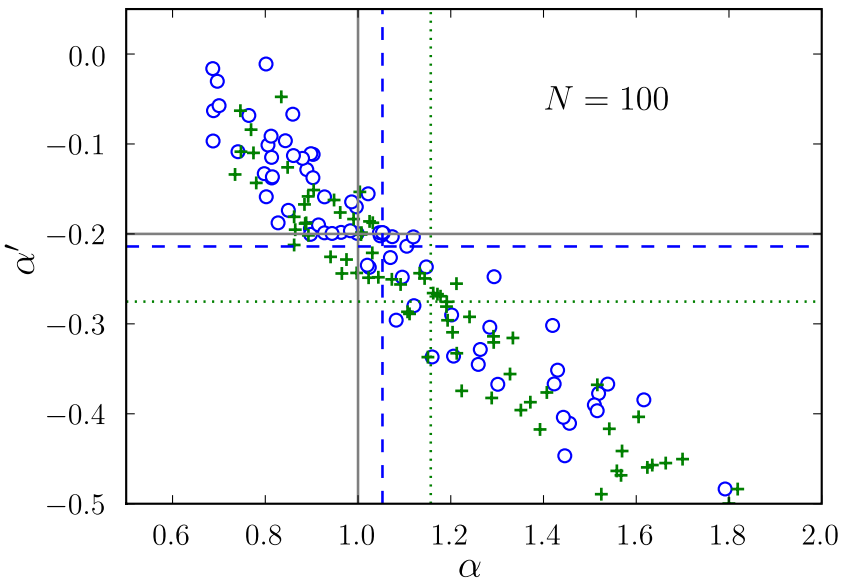
$$f(m) \propto 10^{[\alpha(m-23)+\alpha'(m-23)^2]}$$

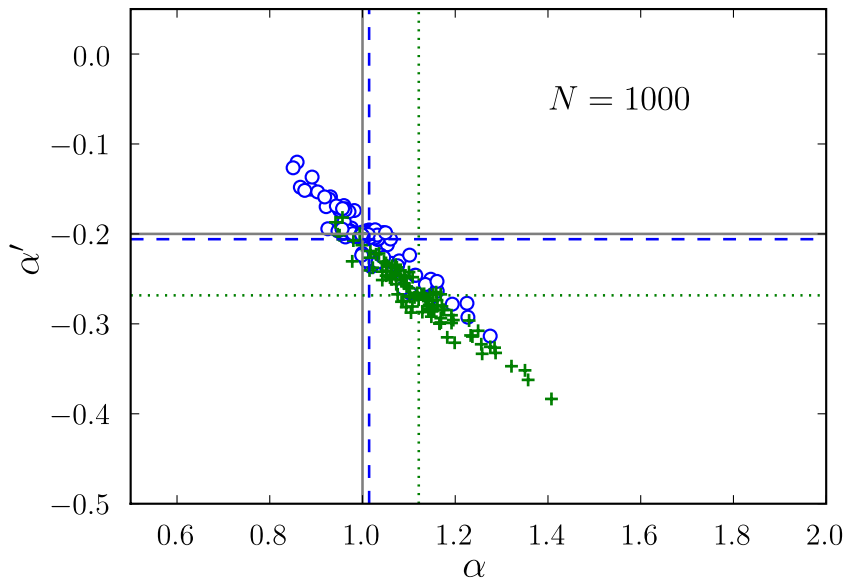


Simulate 100 surveys of populations drawn from the same dist'n.
Simulate data for photon-counting instrument, fixed count threshold.
Measurements have uncertainties 1% (bright) to $\approx$ 30% (dim).

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

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





Uncertainties don't average out!

Bayesian MLMs in Astronomy

- **Directional & spatio-temporal coincidences:**
  - GRB repetition (Luo$^+$ 1996; Graziani$^+$ 1996)
  - GRB host ID (Band 1998; Graziani$^+$ 1999)
  - VO cross-matching (Badavári & Szalay 2008)
- **Magnitude surveys/number counts/"log N–log S":**
  - GRB peak flux dist'n (Loredo & Wasserman 1998);
  - TNO/KBO magnitude distribution (Gladman$^+$ 1998; Petit$^+$ 2008)
  - Malmquist-type biases in cosmology (Loredo & Hendry 2009)
- **Dynamic spectroscopy:** SN 1987A neutrinos, uncertain energy vs. time (Loredo & Lamb 2002)
- **Linear regression:** QSO hardness vs. luminosity (Kelly 2007)

Agenda

1. Bayesian measurement error modeling
   - The Neyman-Scott problem
   - Astrophysical measurement error problems
   - Multilevel models for measurement error

2. Basic Bayesian calculation
   - Large N: Laplace approximations
   - Low d: Cubature, adaptive cubature
   - Available software
   - Closing reflections

Statistical Integrals

Inference with independent data

Consider N data, $D = \{x_i\}$; and model M with m parameters.

Suppose $\mathcal{L}(\theta) = p(x_1|\theta)\, p(x_2|\theta) \cdots p(x_N|\theta)$.

Frequentist integrals

Find long-run properties of procedures via sample space integrals:

$$\mathcal{I}(\theta) = \int dx_1\, p(x_1|\theta) \int dx_2\, p(x_2|\theta) \cdots \int dx_N\, p(x_N|\theta) f(D, \theta)$$

Rigorous analysis must explore the θ dependence; rarely done in practice.

"Plug-in" approximation: Report properties of procedure for $\theta = \hat{\theta}$. *Asymptotically* accurate (for large N, expect $\hat{\theta} \to \theta$).

"Plug-in" results are easy via Monte Carlo (due to independence).

Bayesian integrals

$$\int d^m\theta \; g(\theta)\, p(\theta|M)\, \mathcal{L}(\theta)$$

- $g(\theta) = 1 \rightarrow p(D|M)$ (norm. const., model likelihood)
- $g(\theta) =$ 'box' $\rightarrow$ credible region
- $g(\theta) = \theta \rightarrow$ posterior mean for θ

Such integrals are sometimes easy if analytic (especially in low dimensions), often easier than frequentist counterparts (e.g., normal credible regions, Student's t).

Asymptotic approximations: Require ingredients familiar from frequentist calculations. Bayesian calculation is *not significantly harder* than frequentist calculation in this limit.

Numerical calculation: For "large" m (> 4 is often enough!) the integrals are often very challenging because of structure (e.g., correlations) in parameter space. This is usually pursued *without making any procedural approximations*.

Bayesian Computation

Large sample size: Laplace approximation

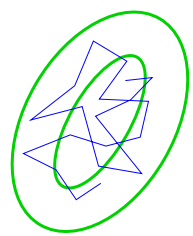
- Approximate posterior as multivariate normal $\rightarrow$ det(covar) factors
- Uses ingredients available in χ^2/ML fitting software (MLE, Hessian)
- Often accurate to $O(1/N)$

Low-dimensional models ($d \lesssim 10$ to 20)

- Adaptive cubature
- Monte Carlo integration (importance & stratified sampling, adaptive importance sampling, quasirandom MC) — *Hedibert's lectures*

Hi-dimensional models ($d \gtrsim 5$)

- Posterior sampling—create RNG that samples posterior
- MCMC is most general framework — *Esther's & Hedibert's lectures*



Laplace Approximations

Suppose posterior has a single dominant (interior) mode at $\hat{\theta}$. For large N,

$$\pi(\theta)\mathcal{L}(\theta) \approx \pi(\hat{\theta})\mathcal{L}(\hat{\theta}) \exp\left[-\frac{1}{2}(\theta-\hat{\theta})\hat{\mathbf{I}}(\theta-\hat{\theta})\right]$$

$$\begin{aligned} \text{where} \quad \hat{\mathbf{I}} &= -\frac{\partial^2 \ln[\pi(\theta)\mathcal{L}(\theta)]}{\partial^2\theta}\bigg|_{\hat{\theta}} \\ &= \text{Negative Hessian of } \ln[\pi(\theta)\mathcal{L}(\theta)] \\ &= \text{“Observed Fisher info. matrix” (for flat prior)} \\ &\approx \text{Inverse of covariance matrix} \end{aligned}$$

E.g., for 1-d Gaussian posterior, $\hat{\mathbf{I}} = 1/\sigma_\theta^2$

Marginal likelihoods

$$\int d\theta\, \pi(\theta)\mathcal{L}(\theta) \approx \pi(\hat{\theta})\mathcal{L}(\hat{\theta})\, (2\pi)^{m/2} \left|\hat{\mathbf{I}}\right|^{-1/2}$$

Marginal posterior densities

Profile likelihood $\mathcal{L}_p(\phi) \equiv \max_\eta \mathcal{L}(\phi, \eta) = \mathcal{L}(\phi, \hat{\eta}(\phi))$

$$\rightarrow p(\phi|D, M) \quad \propto \quad \pi(\phi, \hat{\eta}(\phi))\mathcal{L}_p(\phi)\left|\mathbf{I}_\eta(\phi)\right|^{-1/2}$$

with $\mathbf{I}_\eta(\phi) = \partial_\eta \partial_\eta \ln(\pi\mathcal{L})|_{\hat{\eta}}$

Posterior expectations

$$\int d\theta\, f(\theta)\pi(\theta)\mathcal{L}(\theta) \quad \propto \quad f(\tilde{\theta})\pi(\tilde{\theta})\mathcal{L}(\tilde{\theta})\, (2\pi)^{m/2} \left|\tilde{\mathbf{I}}\right|^{-1/2}$$

where $\tilde{\theta}$ maximizes $f\pi\mathcal{L}$

Tierney & Kadane, "Accurate Approximations for Posterior Moments and Marginal Densities," *JASA* (1986)

Features

Uses output of common algorithms for frequentist methods (optimization, Hessian)

Uses ratios $\rightarrow$ approximation is often $O(1/N)$ or better

Includes volume factors that are missing from common frequentist methods (better inferences!)

Using "unit info prior" in i.i.d. setting $\rightarrow$
Bayesian Information Criterion (BIC; aka Schwarz criterion):

$$\ln B \approx \ln \mathcal{L}(\hat{\theta}) - \ln \mathcal{L}(\tilde{\theta}, \tilde{\phi}) + \frac{1}{2}(m_2 - m_1) \ln N$$

Bayesian counterpart to adjusting χ^2 for d.o.f., but partly accounts for parameter space volume (consistent!)

Drawbacks

Posterior must be smooth and unimodal (or well-separated modes)

Mode must be away from boundaries (can be relaxed)

Result is parameterization-dependent—try to reparameterize to make things look as Gaussian as possible (e.g., $\theta \rightarrow \log \theta$ to straighten curved contours)

Asymptotic approximation with no simple diagnostics (like many frequentist methods)

Empirically, it often does not work well for $m \gtrsim 10$

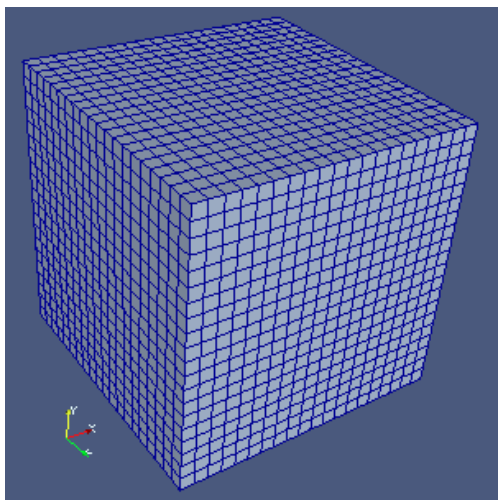
Quadrature Rules

Quadrature rules for 1-D integrals (with weight function $h(\theta)$):

$$\begin{aligned} \int d\theta\, f(\theta) &= \int d\theta\, h(\theta)\, \frac{f(\theta)}{h(\theta)} \\ &\approx \sum_i w_i\, f(\theta_i) + O(n^{-2}) \text{ or } O(n^{-4}) \end{aligned}$$

Smoothness $\longrightarrow$ fast convergence in 1-D

Curse of dimensionality: Cartesian product rules converge slowly, $O(n^{-2/m})$ or $O(n^{-4/m})$ in m-D



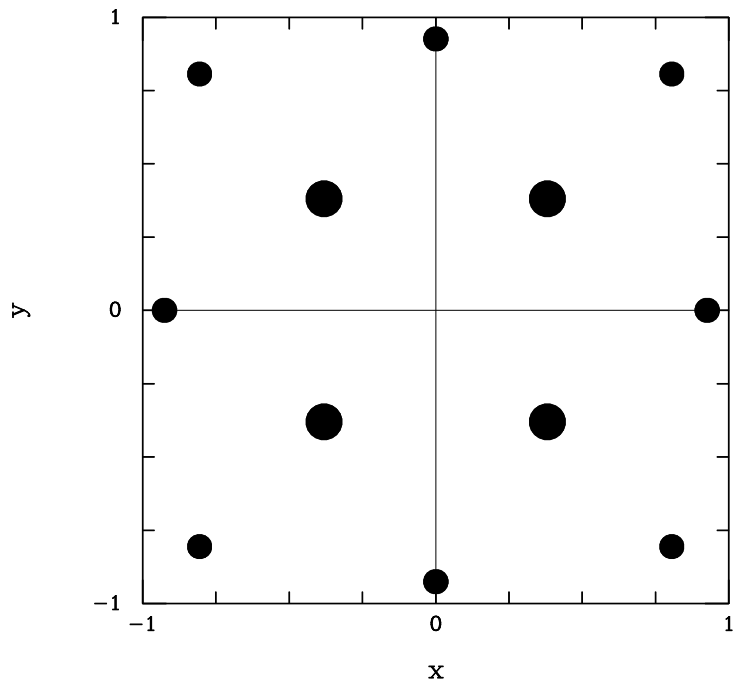
Wikipedia

Monomial Cubature Rules

Seek rules exact for multinomials ($\times$ weight) up to fixed monomial degree with desired lattice symmetry.

Number of points required grows much more slowly with m than for Cartesian rules (but still quickly)

A 7th order rule in 2-d

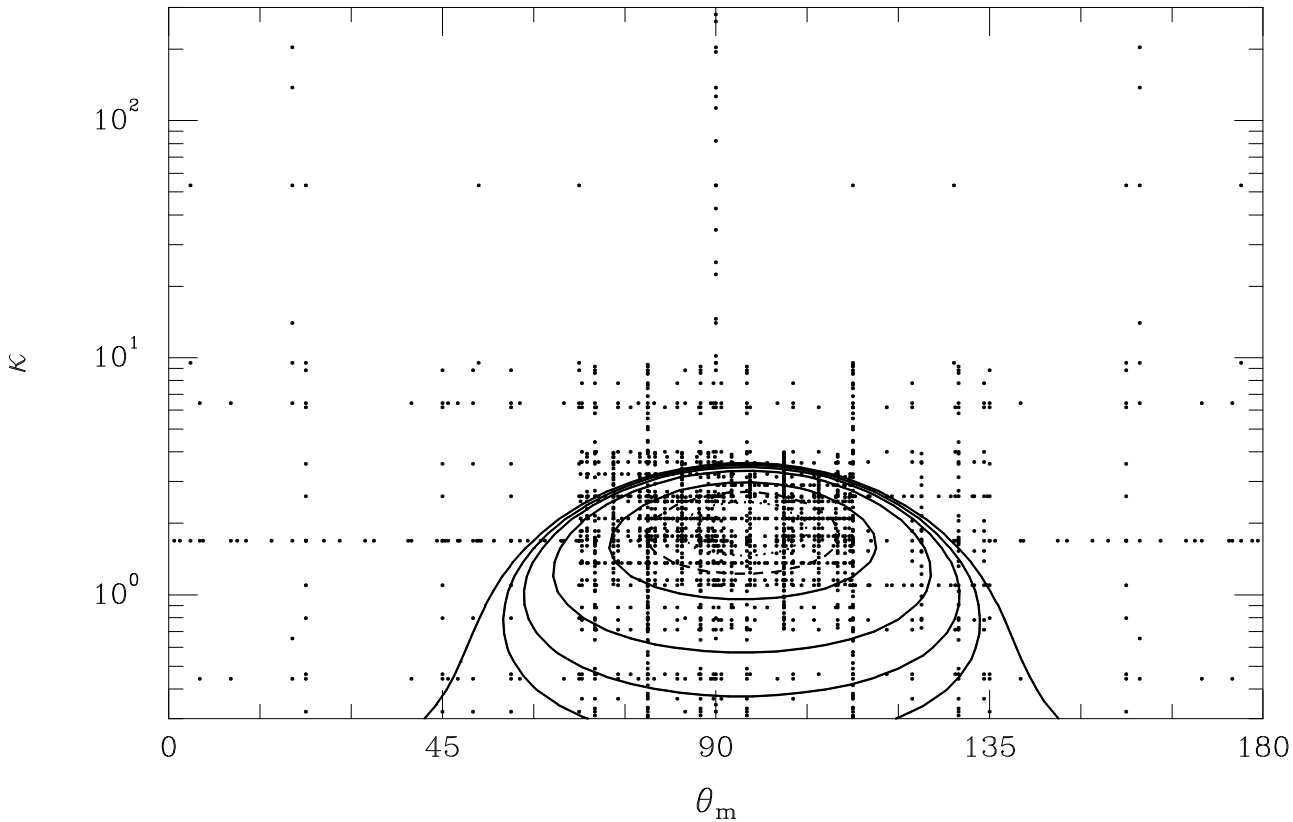


Adaptive Cubature

- Subregion adaptive cubature: Use a pair of monomial rules (for error estim'n); recursively subdivide regions w/ large error (`ADAPT`, `DCUHRE`, `BAYESPACK`, `CUBA`). Concentrates points where most of the probability lies.
- Adaptive grid adjustment: Naylor-Smith method
  Iteratively update abscissas and weights to make the (unimodal) posterior approach the weight function.

These provide diagnostics (error estimates or measures of reparameterization quality).

Analysis of Galaxy Polarizations



Tools for Computational Bayes

Astronomer/Physicist Tools

- **BIE** `http://www.astro.umass.edu/~weinberg/proto_bie/`
  Bayesian Inference Engine: General framework for Bayesian inference, tailored to astronomical and earth-science survey data. Built-in database capability to support analysis of terabyte-scale data sets. Inference is by Bayes via MCMC.
- **XSpec**, **CIAO/Sherpa**
  Both environments have some basic Bayesian capability (including basic MCMC in XSpec)
- **CosmoMC** `http://cosmologist.info/cosmomc/`
  Parameter estimation for cosmological models using CMB, etc., via MCMC
- **ExoFit** `http://zuserver2.star.ucl.ac.uk/~lahav/exofit.html`
  Adaptive MCMC for fitting exoplanet RV data
- **CDF Bayesian Limit Software**
  `http://www-cdf.fnal.gov/physics/statistics/statistics_software.html`
  Limits for Poisson counting processes, with background & efficiency uncertainties
- **root/RooStats**
  `http://root.cern.ch/root/html/ROOFIT_ROOSTATS_Index.html`
  Statistical tools for particle physicists; Bayesian support being incorporated
- **CUBA** `http://www.feynarts.de/cuba/`
  Multidimensional integration via adaptive cubature; adaptive importance sampling & stratification; QMC (C/C++, Fortran, and Mathematica)
- **Inference** Forthcoming at `http://inference.astro.cornell.edu/`
  Several self-contained Bayesian modules; Parametric Inference Engine

Python

- **PyMC** `http://trichech.us/pymc`
  A framework for MCMC via Metropolis-Hastings; also implements Kalman filters and Gaussian processes. Targets biometrics, but is general.
- **SimPy** `http://simpy.sourceforge.net/`
  Intro to SimPy http://heather.cs.ucdavis.edu/ matloff/simpy.html SimPy (rhymes with "Blimpie") is a process-oriented public-domain package for discrete-event simulation.
- **RSPython** `http://www.omegahat.org/`
  Bi-directional communication between Python and R
- **MDP** `http://mdp-toolkit.sourceforge.net/`
  Modular toolkit for Data Processing: Current emphasis is on machine learning (PCA, ICA...). Modularity allows combination of algorithms and other data processing elements into "flows."
- **Orange** `http://www.ailab.si/orange/`
  Component-based data mining, with preprocessing, modeling, and exploration components. Python/GUI interfaces to $C++$ implementations. Some Bayesian components.
- **ELEFANT** `http://rubis.rsise.anu.edu.au/elefant`
  Machine learning library and platform providing Python interfaces to efficient, lower-level implementations. Some Bayesian components (Gaussian processes; Bayesian ICA/PCA).

R and S

- **CRAN Bayesian task view**
  `http://cran.r-project.org/web/views/Bayesian.html`
  Overview of many R packages implementing various Bayesian models and methods; pedagogical packages; packages linking R to other Bayesian software (BUGS, JAGS)
- **Omega-hat** `http://www.omegahat.org/`
  RPython, RMatlab, R-Xlisp
- **BOA** `http://www.public-health.uiowa.edu/boa/`
  Bayesian Output Analysis: Convergence diagnostics and statistical and graphical analysis of MCMC output; can read BUGS output files.
- **CODA**
  `http://www.mrc-bsu.cam.ac.uk/bugs/documentation/coda03/cdaman03.html`
  Convergence Diagnosis and Output Analysis: Menu-driven R/S plugins for analyzing BUGS output

Java

- **Omega-hat** `http://www.omegahat.org/`
  Java environment for statistical computing, being developed by XLisp-stat and R developers
- **Hydra** `http://research.warnes.net/projects/mcmc/hydra/`
  HYDRA provides methods for implementing MCMC samplers using Metropolis, Metropolis-Hastings, Gibbs methods. In addition, it provides classes implementing several unique adaptive and multiple chain/parallel MCMC methods.
- **YADAS** `http://www.stat.lanl.gov/yadas/home.html`
  Software system for statistical analysis using MCMC, based on the multi-parameter Metropolis-Hastings algorithm (rather than parameter-at-a-time Gibbs sampling)

C/C++/Fortran

- **BayeSys 3** `http://www.inference.phy.cam.ac.uk/bayesys/`
  Sophisticated suite of MCMC samplers including transdimensional capability, by the author of MemSys
- **fbm** `http://www.cs.utoronto.ca/~radford/fbm.software.html`
  Flexible Bayesian Modeling: MCMC for simple Bayes, Bayesian regression and classification models based on neural networks and Gaussian processes, and Bayesian density estimation and clustering using mixture models and Dirichlet diffusion trees
- **BayesPack, DCUHRE**
  `http://www.sci.wsu.edu/math/faculty/genz/homepage`
  Adaptive quadrature, randomized quadrature, Monte Carlo integration
- **BIE**, **CDF Bayesian limits**, **CUBA** (see above)

Other Statisticians' & Engineers' Tools

- **BUGS/WinBUGS** `http://www.mrc-bsu.cam.ac.uk/bugs/`
  Bayesian Inference Using Gibbs Sampling: Flexible software for the Bayesian analysis of complex statistical models using MCMC
- **OpenBUGS** `http://mathstat.helsinki.fi/openbugs/`
  BUGS on Windows and Linux, and from inside the R
- **JAGS** `http://www-fis.iarc.fr/~martyn/software/jags/`
  "Just Another Gibbs Sampler;" MCMC for Bayesian hierarchical models
- **XLisp-stat** `http://www.stat.uiowa.edu/~luke/xls/xlsinfo/xlsinfo.html`
  Lisp-based data analysis environment, with an emphasis on providing a framework for exploring the use of dynamic graphical methods
- **ReBEL** `http://choosh.csee.ogi.edu/rebel/`
  Library supporting recursive Bayesian estimation in Matlab (Kalman filter, particle filters, sequential Monte Carlo).

Closing Reflections

Philip Dawid (2000)

> What is the principal distinction between Bayesian and classical statistics? It is that Bayesian statistics is fundamentally boring. There is so little to do: just specify the model and the prior, and turn the Bayesian handle. There is no room for clever tricks or an alphabetic cornucopia of definitions and optimality criteria. I have heard people use this 'dullness' as an argument against Bayesianism. One might as well complain that Newton's dynamics, being based on three simple laws of motion and one of gravitation, is a poor substitute for the richness of Ptolemy's epicyclic system.
>
> All my experience teaches me that it is invariably more fruitful, and leads to deeper insights and better data analyses, to explore the consequences of being a 'thoroughly boring Bayesian'.

Dennis Lindley (2000)

> The philosophy places more emphasis on model construction than on formal inference. . . I do agree with Dawid that 'Bayesian statistics is fundamentally boring'. . . My only qualification would be that the theory may be boring but the applications are exciting.