Bayesian Computation—A Survey

(Lecture 6)

Tom Loredo

Dept. of Astronomy, Cornell University

Statistical Integrals

Inference with independent data:

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

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” approach: Report properties of procedure for $\theta = \hat{\theta}$. Asymptotically valid (for large N, expect $\hat{\theta} \to \theta$).

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

Bayesian integrals

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

- $g(\theta) = 1 \to p(D|M)$ (norm. const., model likelihood)
- $g(\theta) =$ 'box' $\to$ credible region
- $g(\theta) = \theta \to$ 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.

"Exact" numerical calculation: For "large" m (> 4 is often enough!) the integrals are often very challenging because of correlations (lack of independence) in parameter space.

Outline

- Asymptotic approximations ($N \gg 1$)
- Methods for low-d models $(m \lesssim 20)$
- Methods for high-d models $(m \sim 10 - -10^6)$

$N =$ # of data

$m =$ # of model parameters

Laplace Approximations

Suppose posterior has a single dominant (interior) mode at $\hat{\theta}$, with m parameters

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

where

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

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

Bayes Factors:

$$\int d\theta\; p(\theta|M)\mathcal{L}(\theta) \approx p(\hat{\theta}|M)\mathcal{L}(\hat{\theta})\; (2\pi)^{m/2}|\hat{\mathbf{I}}|^{-1/2}$$

Marginals:

Profile likelihood $\qquad \mathcal{L}_p(\theta) \equiv \max_{\phi} \mathcal{L}(\theta, \phi)$

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

Expectations:

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

where $\tilde{\theta}$ maximizes $fp\mathcal{L}$

Features

Uses same algorithms as common frequentist calculations (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 $\rightarrow$ Schwarz criterion; Bayesian Information Criterion (BIC)

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

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., $\theta \to \log \theta$ to straighten curved contours)

Asymptotic approximation with no simple diagnostics

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

Low-D $(m \lesssim 10)$: Cubature & Monte Carlo

Quadrature (1-d)/Cubature (2$+$-d) Rules:

$$\int d\theta\; f(\theta) w(\theta) \approx \sum_i w_i\, f(\theta_i) + O(n^{-2}) \text{ or } O(n^{-4})$$

Smoothness $\rightarrow$ fast convergence in 1-D

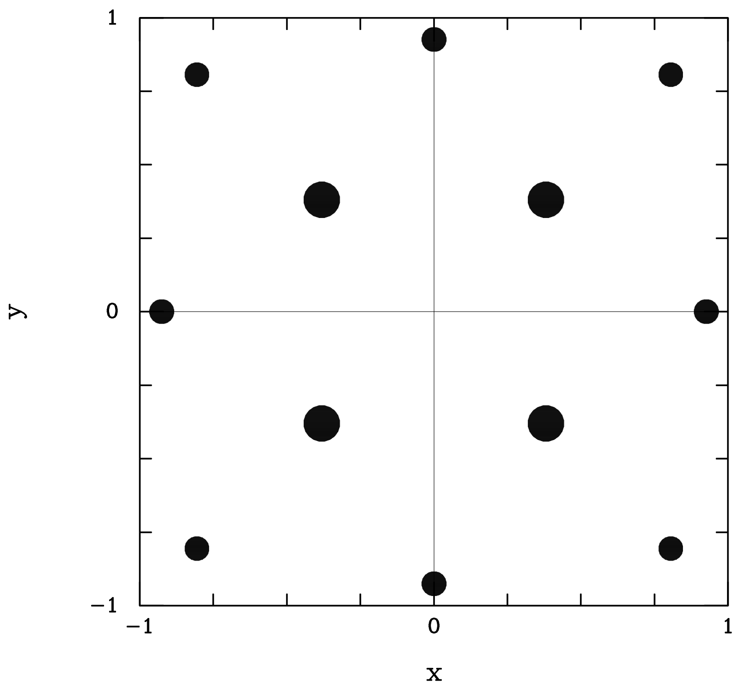
Curse of dimensionality: Cartesian product rules converge slowly, $O(n^{-2/m})$ or $O(n^{-4/m})$ in m-D

Monomial/lattice 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

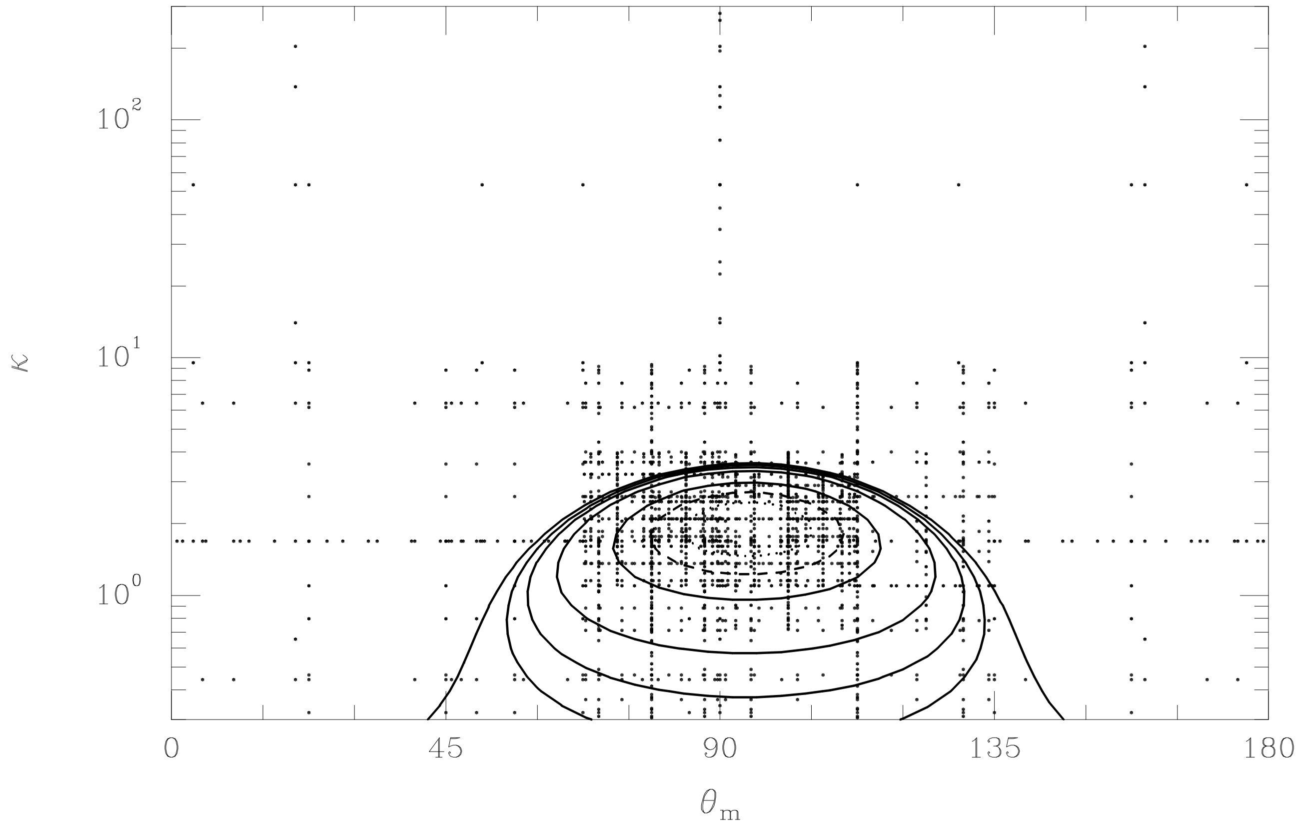


Adaptive Cubature:

- Subregion adaptive cubature: Use a pair of lattice rules (for error estim'n); recursively subdivide regions w/ large error (`ADAPT`, `DCUHRE`, `BAYESPACK` by Genz et al.). Concentrates points where most of the probability lies.
- Adaptive grid adjustment: Naylor-Smith method Iteratively reparameterize $\rightarrow$ update abscissas and weights to make the (unimodal) posterior approach normality

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

Analysis of Galaxy Polarizations



Monte Carlo Integration:

Choose points randomly rather than deterministically:

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

[$\sim O(n^{-1})$ with quasi-MC]

Ignores smoothness $\rightarrow$ poor performance in 1-D

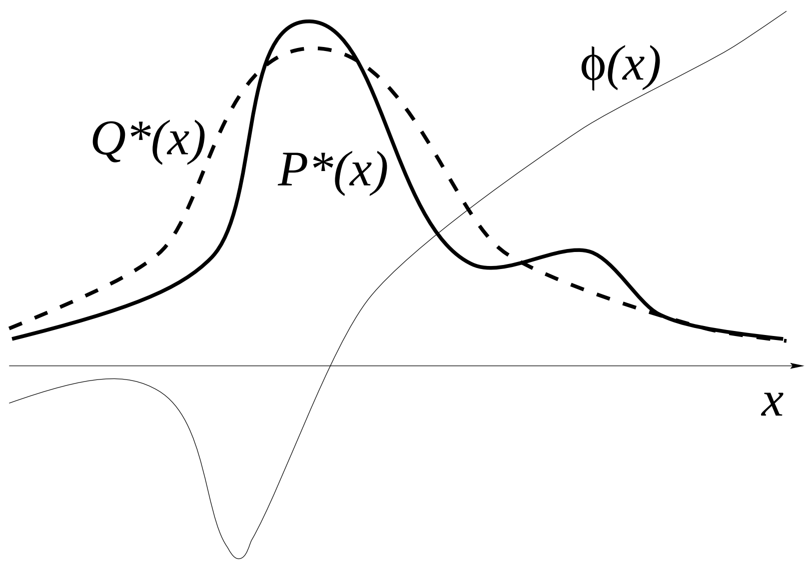
Avoids curse: $O(n^{-1/2})$ regardless of dimension

Practical problem: multiplier (std. dev'n of g) is large and uncertain $\rightarrow$ hard if $m \gtrsim$ 5–10

Importance sampling:

$$\int d\theta\ g(\theta)p(\theta) = \int d\theta\ g(\theta)\frac{p(\theta)}{q(\theta)}q(\theta) \approx \sum_{\theta_i \sim q(\theta)} g(\theta_i)\frac{p(\theta_i)}{q(\theta_i)}$$

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



MacKay 2003

Adaptive Monte Carlo: Build the importance sampler on-the-fly (e.g., `VEGAS`, `miser` in *Numerical Recipes*)

High-D Models: Posterior Sampling

General Approach:

Draw samples of θ, ϕ from $p(\theta, \phi | D, M)$; then:

- Integrals, moments easily found via $\sum_i f(\theta_i, \phi_i)$
- $\{\theta_i\}$ are samples from $p(\theta | D, M)$

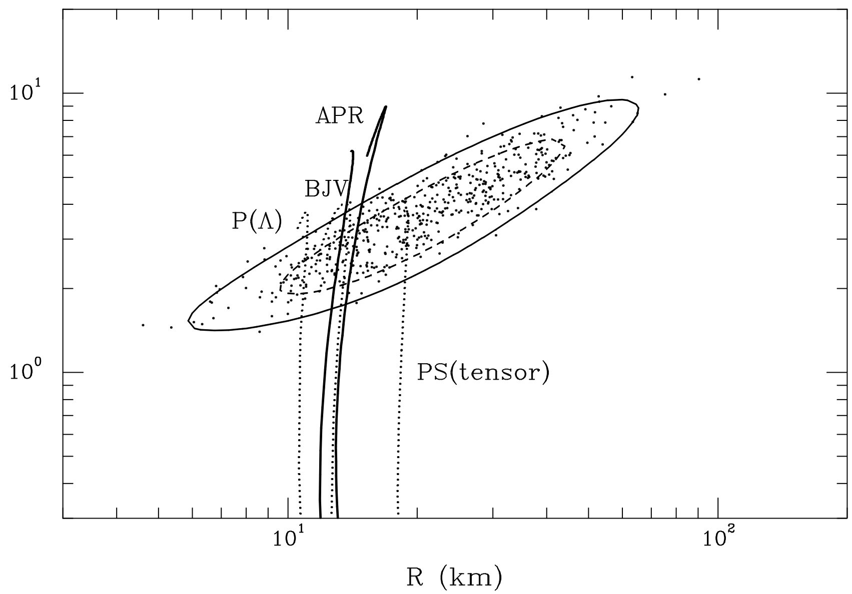
But how can we obtain $\{\theta_i, \phi_i\}$?

A Complicated Marginal Distribution

Nascent neutron star properties inferred from neutrino data from SN 1987A.

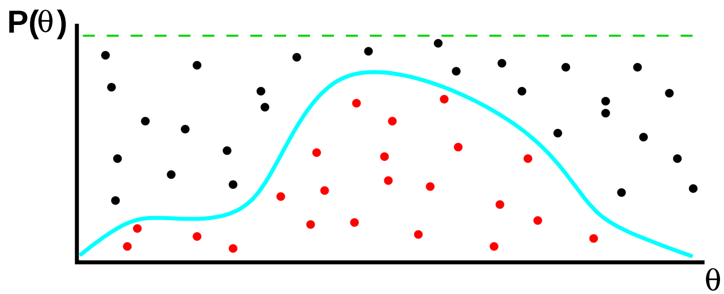
Signal model has 9 parameters; multi-modal.

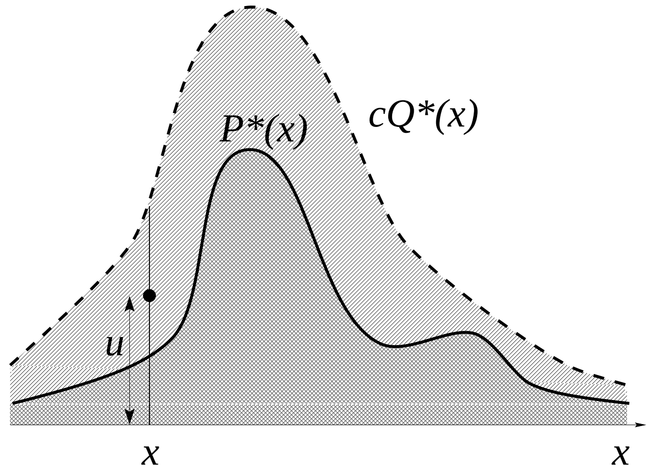
Two interesting parameters are the NS radius and its binding energy—a functional of the signal model.



Rejection Method:

Instead of sampling θ directly, sample the area under the $p(\theta)$ curve.





Adds an auxiliary variable, $y = p(\theta)$, samples unformly over $\{(\theta, y) : 0 < y < p(\theta)\}$, and keeps θ

Hard to find efficient comparison function if $m \gtrsim$5–10.

Markov Chain Monte Carlo (MCMC)

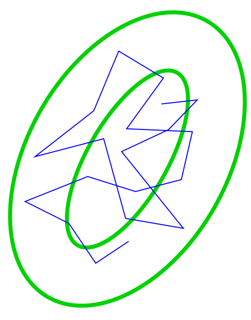
Let $\quad -\Lambda(\theta) = \ln\left[p(\theta|M)\, p(D|\theta, M)\right]$

Then $\quad p(\theta|D, M) = \dfrac{e^{-\Lambda(\theta)}}{Z} \qquad Z \equiv \int d\theta\; e^{-\Lambda(\theta)}$

Bayesian integration looks like problems addressed in computational statmech and Euclidean QFT.

Methods share a common element: make a proposal *that depends on the current state* $\rightarrow$ Markov chains

Goal: An iterative algorithm that wanders around the posterior with time $\propto$ probability.



The Metropolis-Hastings MCMC Recipe:

Create a “time series” of samples θ_i from $p(\theta)$:

- Draw a candidate θ_{i+1} from a proposal $Q(\theta_{i+1};\theta_i)$
- Calculate

$$\alpha = \frac{Q(\theta_i;\theta_{i+1})p(\theta_{i+1})}{Q(\theta_{i+1};\theta_i)p(\theta_i)}$$

- If $\alpha \geq 1$, accept the proposal
- Otherwise, accept it with probability α; otherwise repeat the previous sample

What this gets you:

Let $T(\theta_{i+1};\theta_i)$ be the transition probability.

For a wide variety of choices of Q, one can show:

$p(\theta)$ is the *stationary dist'n*,

$$\int d\theta T(\theta';\theta)p(\theta) = p(\theta')$$

$p(\theta)$ is a *limiting dist'n*: even if $p_0 \neq p$,

$$p_i(\theta) \rightarrow p(\theta)$$

The chain is *ergodic*,

$$\frac{1}{K}\sum_{i=1}^{K} f(\theta_i) \rightarrow \int d\theta\, f(\theta)p(\theta)$$

Only *ratios* of p's and Q's need be known.

What Proposal Distribution?

Almost anything will work—if you wait long enough! But most simple choices will take very, very long.

Development of new methods is one of the hottest research areas; very many to choose from.

Good choices tend to be problem-specific.

Some themes:

- Reparameterize wisely
- Adaptively tune the proposal
- Add extra variables (e.g., hybrid Monte Carlo)
- Run parallel chains, possibly interacting
- Temper/anneal if there is multimodality

Transdimensional MCMCM: Methods that can jump between models of different dimensionality (“reversible jump”)

MCMC Output Diagnostics

How many iterations until the sample distribution is “close” to $p(\theta)$? (“burn-in”)

How many timesteps to use to guarantee mixing/ergodicity?

How correlated are the output samples?

Seek diagnostics both for guiding algorithm tuning, and for alerting failure.

Several approaches:

- Monitor trends in simulation output
- Compare within- and between-chain variation for several chains
- Monitor algorithm characteristics (acceptance rate, transition or posterior probabilities)

Summary of Tools

- Asymptotic (large N) approximations: Laplace approximations
- Low-d models $(m \lesssim 20)$:
  - Quadrature/Cubature (esp. adaptive methods)
  - Monte Carlo integration (imp. sampling, adaptive)
- High-d Models ($m \sim 10$ to 10^6):
  - Posterior Sampling (MCMC)

Outlook

- There are many useful methods, but there is no panacea
- Method choice depends not just on model dimension but on model/posterior structure
- All methods can fail without obvious notice—compare!
- Plenty of room for future developments!
- Several software packages exist/in development implementing multiple methods