Bayesian Inference: A Practical Primer

Tom Loredo

Department of Astronomy, Cornell University

loredo@spacenet.tn.cornell.edu

http://www.astro.cornell.edu/staff/loredo/bayes/

Outline

- Parametric Bayesian inference
  - Probability theory
  - Parameter estimation
  - Model uncertainty
- What's different about it?
- Bayesian calculation
  - Asymptotics: Laplace approximations
  - Quadrature
  - Posterior sampling and MCMC

Bayesian Statistical Inference: Quantifying Uncertainty

Inference:

- Reasoning from one proposition to another
- *Deductive Inference:* Strong syllogisms, logic; quantify with Boolean algebra
- *Plausible Inference:* Weak syllogisms; quantify with probability

  Propositions of interest to us are descriptions of data (D), and hypotheses about the data, H_i

Statistical:

- *Statistic:* Summary of what data say about a particular question/issue
- Statistic $= f(D)$ (value, set, etc.); implicitly also f(question)
- Statistic is chosen & interpreted via probability theory
- Statistical inference = Plausible inference using probability theory

Bayesian (vs. Frequentist):

What are valid arguments for probabilities $P(A|\cdots)$?

- Bayesian: *Any* propositions are valid (in principle)
- Frequentist: Only propositions about *random events* (data)

How should we use probability theory to do statistics?

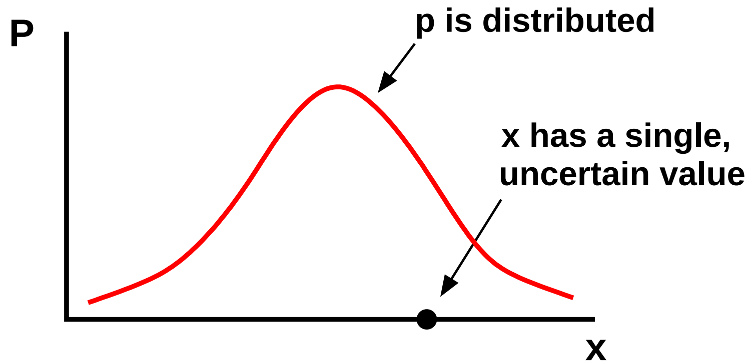
- Bayesian: Calculate $P(H_i|D,\cdots)$ vs. H_i with $D = D_{obs}$
- Frequentist: Create methods for choosing among H_i with good *long run behavior* determined by examining $P(D|H_i)$ for all possible hypothetical D; apply method to D_{obs}

What is distributed in $p(x)$?

Bayesian: Probability describes uncertainty

Bernoulli, Laplace, Bayes, Gauss...

$p(x)$ describes how probability (plausibility) is distributed among the possible choices for x in the case at hand. Analog: a mass density, $\rho(x)$



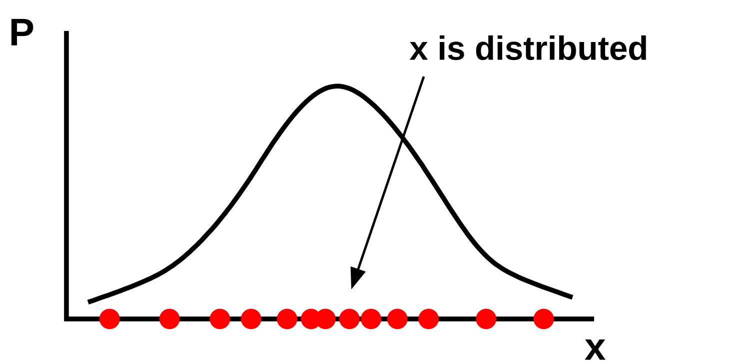
Relationships between probability and frequency were demonstrated mathematically (large number theorems, Bayes's theorem).

Frequentist: Probability describes "randomness"

Venn, Boole, Fisher, Neymann, Pearson...

x is a *random variable* if it takes different values throughout an infinite (imaginary?) ensemble of "identical" sytems/experiments.

$p(x)$ describes how x is distributed throughout the infinite ensemble.



Probability $\equiv$ frequency.

Interpreting Abstract Probabilities

Symmetry/Invariance/Counting

- Resolve possibilities into equally plausible "microstates" using symmetries
- Count microstates in each possibility

Frequency from probability

Bernoulli's laws of large numbers: In repeated trials, given P(success), predict

$$\frac{N_{\text{success}}}{N_{\text{total}}} \to P \quad \text{as} \quad N \to \infty$$

Probability from frequency

Bayes's "An Essay Towards Solving a Problem in the Doctrine of Chances" $\to$ Bayes's theorem

Probability $\neq$ Frequency!

Bayesian Probability: A Thermal Analogy

Intuitive notion	*Quantification*	*Calibration*
Hot, cold	Temperature, T	Cold as ice = 273K Boiling hot = 373K
uncertainty	Probability, P	Certainty = 0, 1 $p = 1/36$: plausible as "snake's eyes" $p = 1/1024$: plausible as 10 heads

The Bayesian Recipe

Assess hypotheses by calculating their probabilities $p(H_i|\ldots)$ conditional on known and/or presumed information using the rules of probability theory.

Probability Theory Axioms ("grammar"):

$$\text{'OR' (sum rule)} \quad P(H_1 + H_2|I) = P(H_1|I) + P(H_2|I) - P(H_1, H_2|I)$$

$$\begin{aligned} \text{'AND' (product rule)} \quad P(H_1, D|I) &= P(H_1|I)\,P(D|H_1, I) \\ &= P(D|I)\,P(H_1|D, I) \end{aligned}$$

Direct Probabilities ("vocabulary"):

- Certainty: If A is certainly true given B, $P(A|B) = 1$
- Falsity: If A is certainly false given B, $P(A|B) = 0$
- Other rules exist for more complicated types of information; for example, invariance arguments, maximum (information) entropy, limit theorems (tying probabilities to frequencies), bold (or desperate!) presumption...

Important Theorems

Normalization:

For *exclusive, exhaustive* H_i

$$\sum_i P(H_i|\cdots) = 1$$

Bayes's Theorem:

$$P(H_i|D,I) = P(H_i|I)\,\frac{P(D|H_i,I)}{P(D|I)}$$

posterior $\propto$ prior $\times$ likelihood

Marginalization:

Note that for exclusive, exhaustive $\{B_i\}$,

$$\begin{aligned}\sum_i P(A,B_i|I) &= \sum_i P(B_i|A,I)P(A|I) = P(A|I)\\ &= \sum_i P(B_i|I)P(A|B_i,I)\end{aligned}$$

$\rightarrow$ We can use $\{B_i\}$ as a "basis" to get $P(A|I)$. This is sometimes called "extending the conversation."

Example: Take $A = D$, $B_i = H_i$; then

$$\begin{aligned}P(D|I) &= \sum_i P(D,H_i|I)\\ &= \sum_i P(H_i|I)P(D|H_i,I)\end{aligned}$$

prior predictive for $D =$ Average likelihood for H_i

Inference With Parametric Models

Parameter Estimation

$I =$ Model M with parameters θ (+ any add'l info)

$H_i =$ statements about θ; e.g. "$\theta \in [2.5, 3.5]$," or "$\theta > 0$"

Probability for any such statement can be found using a *probability density function* (PDF) for θ:

$$\begin{aligned} P(\theta \in [\theta, \theta + d\theta] | \cdots) &= f(\theta) d\theta \\ &= p(\theta | \cdots) d\theta \end{aligned}$$

Posterior probability density:

$$p(\theta | D, M) = \frac{p(\theta | M)\ \mathcal{L}(\theta)}{\int d\theta\ p(\theta | M)\ \mathcal{L}(\theta)}$$

Summaries of posterior:

- "Best fit" values: mode, posterior mean
- Uncertainties: Credible regions
- Marginal distributions:
  - Interesting parameters ψ, nuisance parameters ϕ
  - Marginal dist'n for ψ:

$$p(\psi | D, M) = \int d\phi\, p(\psi, \phi | D, M)$$

Generalizes "propagation of errors"

Model Uncertainty: Model Comparison

$I = (M_1 + M_2 + \ldots)$ — Specify a set of models.

$H_i = M_i$ — Hypothesis chooses a model.

Posterior probability for a model:

$$p(M_i|D,I) = p(M_i|I)\frac{p(D|M_i,I)}{p(D|I)} \\ \propto p(M_i)\mathcal{L}(M_i)$$

But $\mathcal{L}(M_i) = p(D|M_i) = \int d\theta_i\, p(\theta_i|M_i)p(D|\theta_i, M_i)$.

Likelihood for model = Average likelihood for its parameters

$$\mathcal{L}(M_i) = \langle \mathcal{L}(\theta_i) \rangle$$

Posterior odds and Bayes factors:

Discrete nature of hypothesis space makes odds convenient:

$$\begin{aligned} O_{ij} &\equiv \frac{p(M_i|D,I)}{p(M_j|D,I)} \\ &= \frac{p(M_i|I)}{p(M_j|I)} \times \frac{p(D|M_i)}{p(D|M_j)} \\ &= \text{Prior Odds} \times \text{Bayes Factor } B_{ij} \end{aligned}$$

Often take models to be equally probable a priori $\rightarrow O_{ij} = B_{ij}$.

Model Uncertainty: Model Averaging

Models have a common subset of interesting parameters, ψ.

Each has different set of nuisance parameters ϕ_i (or different prior info about them).

$H_i =$ statements about ψ

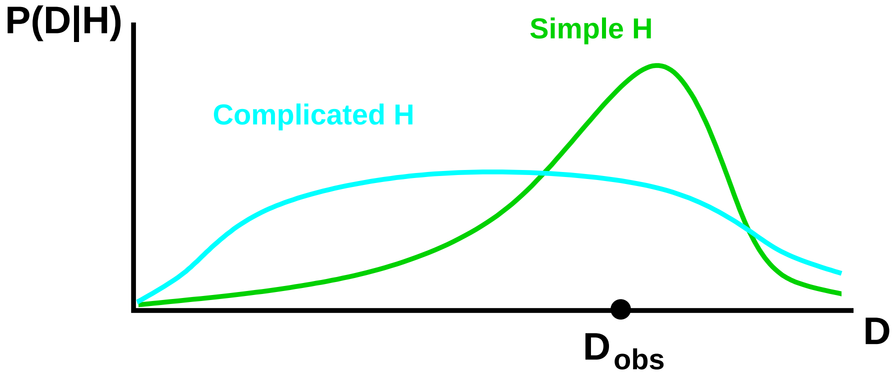
Calculate posterior PDF for ψ:

$$
\begin{aligned}
p(\psi|D, I) &= \sum_i p(\psi|D, M_i) p(M_i|D, I) \\
&\propto \sum_i \mathcal{L}(M_i) \int d\theta_i \, p(\psi, \phi_i|D, M_i)
\end{aligned}
$$

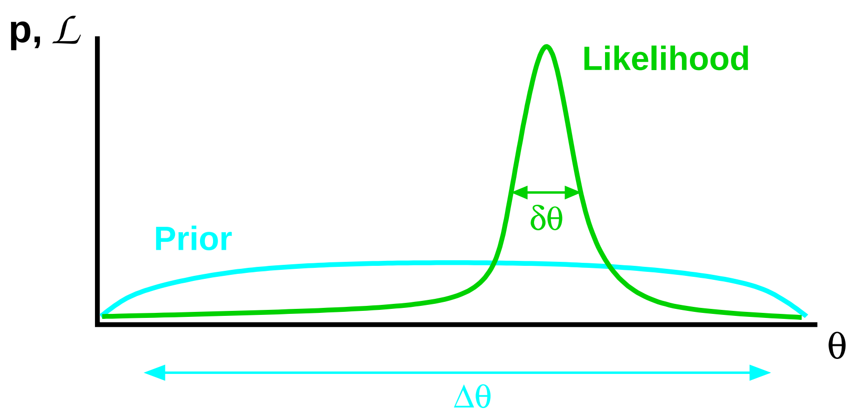
The model choice is itself a (discrete) nuisance parameter here.

An Automatic Occam's Razor

Predictive probabilities prefer simpler models:



The Occam Factor:



$$
\begin{aligned}
p(D|M_i) &= \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i) \\
&\approx p(\hat{\theta}_i|M)\mathcal{L}(\hat{\theta}_i)\delta\theta_i \\
&\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i} \\
&= \text{Maximum Likelihood} \times \text{Occam Factor}
\end{aligned}
$$

Models with more parameters usually make the data more probable *for the best fit*.

The Occam factor penalizes models for "wasted" volume of parameter space.

Comparison of Bayesian & Frequentist Approaches

Bayesian Inference (BI):

- Specify at least two competing hypotheses and priors
- Calculate their probabilities using the rules of probability theory
  - Parameter estimation:

$$p(\theta|D,M) = \frac{p(\theta|M)\mathcal{L}(\theta)}{\int d\theta\, p(\theta|M)\mathcal{L}(\theta)}$$

  - Model Comparison:

$$O \propto \frac{\int d\theta_1\, p(\theta_1|M_1)\,\mathcal{L}(\theta_1)}{\int d\theta_2\, p(\theta_2|M_2)\,\mathcal{L}(\theta_2)}$$

Frequentist Statistics (FS):

- Specify null hypothesis H_0 such that rejecting it implies an interesting effect is present
- Specify statistic $S(D)$ that measures departure of the data from null expectations
- Calculate $p(S|H_0) = \int dD\, p(D|H_0)\delta[S - S(D)]$ (e.g. by Monte Carlo simulation of data)
- Evaluate $S(D_{\text{obs}})$; decide whether to reject H_0 based on, e.g., $\int_{>S_{\text{obs}}} dS\, p(S|H_0)$

Crucial Distinctions

The role of subjectivity:

BI exchanges (implicit) subjectivity in the choice of null & statistic for (explicit) subjectivity in the specification of alternatives.

- Makes assumptions explicit
- Guides specification of further alternatives that generalize the analysis
- Automates identification of statistics:

  BI is a problem-solving approach

  FS is a solution-characterization approach

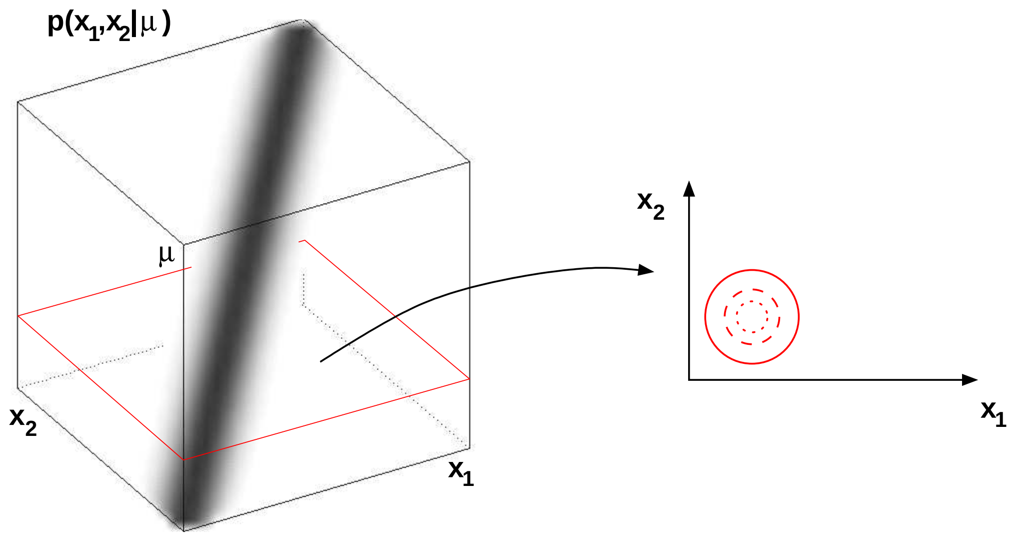
The types of mathematical calculations:

The two approaches require calculation of very different sums/averages.

- BI requires integrals over hypothesis/parameter space
- FS requires integrals over sample/data space

A Frequentist Confidence Region

Infer μ : $\quad x_i = \mu + \epsilon_i; \qquad p(x_i|\mu, M) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$



68% confidence region: $\bar{x} \pm \sigma/\sqrt{N}$

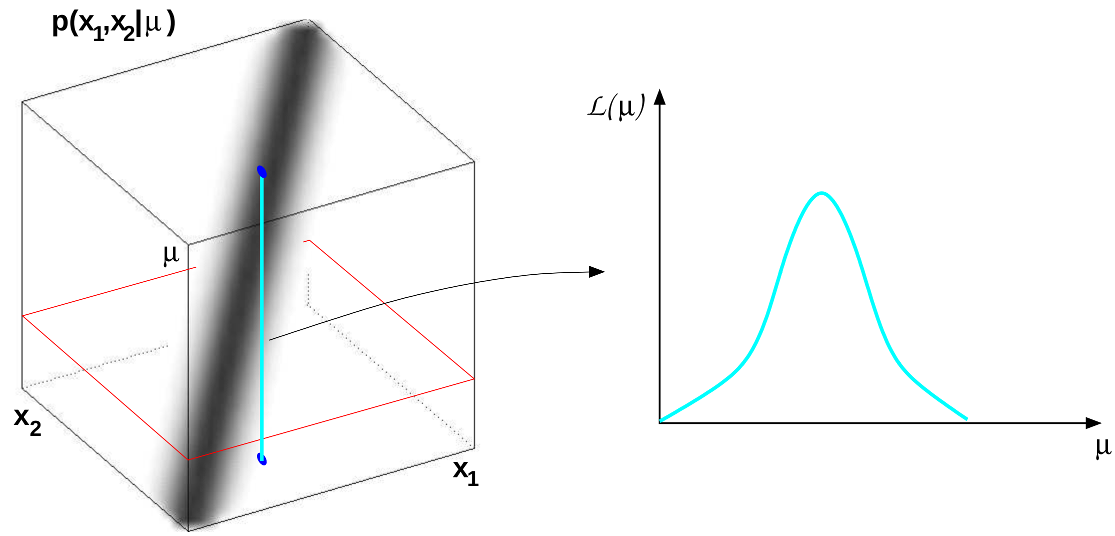
1. Pick a null hypothesis, $\mu = \mu_0$
2. Draw $x_i \sim N(\mu_0, \sigma^2)$ for $i = 1$ to N
3. Find $\bar{x}$; check if $\mu_0 \in \bar{x} \pm \sigma/\sqrt{N}$
4. Repeat $M >> 1$ times; report fraction (≈ 0.683)
5. *Hope result is independent of μ_0!*

A Monte Carlo calculation of the N-dimensional integral:

$$\int dx_1 \frac{e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \cdots \int dx_N \frac{e^{-\frac{(x_N-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \times [\mu_0 \in \bar{x} \pm \sigma/\sqrt{N}] \approx 0.683$$

A Bayesian Credible Region

Infer μ :   Flat prior; $\mathcal{L}(\mu) \propto \exp\left[-\frac{(\bar{x}-\mu)^2}{2(\sigma/\sqrt{N})^2}\right]$



68% credible region: $\bar{x} \pm \sigma/\sqrt{N}$

$$\frac{\int_{\bar{x}-\sigma/\sqrt{N}}^{\bar{x}-\sigma/\sqrt{N}} d\mu \, \exp\left[-\frac{(\bar{x}-\mu)^2}{2(\sigma/\sqrt{N})^2}\right]}{\int_{-\infty}^{\infty} d\mu \, \exp\left[-\frac{(\bar{x}-\mu)^2}{2(\sigma/\sqrt{N})^2}\right]} \approx 0.683$$

Equivalent to a Monte Carlo calculation of a 1-d integral:

1. Draw μ from $N(\bar{x}, \sigma^2/N)$ (i.e., prior $\times\mathcal{L}$)
2. Repeat $M >> 1$ times; histogram
3. Report most probable 68.3% region

This simulation uses hypothetical *hypotheses* rather than hypothetical *data*.

When Will Results Differ?

When models are linear in the parameters and have additive Gaussian noise, frequentist results are identical to Bayesian results with flat priors.

This mathematical coincidence will not occur if:

- The choice of statistic is not obvious
  (no sufficient statistics)
- There is no identity between parameter space and sample space integrals (due to nonlinearity or the form of the sampling distribution)
- There is important prior information

In addition, some problems can be quantitatively addressed only from the Bayesian viewpoint; e.g., systematic error.

Benefits of Calculating in Parameter Space

- Provides probabilities *for hypotheses*
  - Straightforward interpretation
  - Identifies weak experiments
  - Crucial for global (hierarchical) analyses (e.g., pop'n studies)
  - Allows analysis of systematic error models
  - Forces analyst to be explicit about assumptions
- Handles nuisance parameters via marginalization
- Automatic Occam's razor
- Model comparison for > 2 alternatives; needn't be nested
- Valid for all sample sizes
- Handles multimodality
- Avoids inconsistency & incoherence
- Automated identification of statistics
- Accounts for prior information (including other data)
- Avoids problems with sample space choice:
  - Dependence of results on "stopping rules"
  - Recognizable subsets
  - Defining number of "independent" trials in searches
- Good frequentist properties:
  - Consistent
  - Calibrated—E.g., if you choose a model only if $B > 100$, you will be right $\approx 99\%$ of the time
  - Coverage as good or better than common methods

Challenges from Calculating in Parameter Space

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:

$$\int dx_1\, p(x_1|\theta) \int dx_2\, p(x_2|\theta) \cdots \int dx_N\, p(x_N|\theta) f(D)$$

Seek integrals with properties independent of θ. Such rigorous frequentist integrals usually cannot be identified.

Approximate results are easy via Monte Carlo (due to independence).

Bayesian integrals:

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

Such integrals are sometimes easy if analytic (especially in low dimensions).

Asymptotic approximations require ingredients familiar from frequentist calculations.

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

Bayesian Integrals: Laplace Approximations

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

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

$$\text{where} \quad \mathbf{I} = \frac{\partial^2 \ln[p(\theta|M)\mathcal{L}(\theta)]}{\partial^2\theta}\Big|_{\hat{\theta}}, \qquad \text{Info matrix}$$

Bayes Factors:

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

Marginals:

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

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

Uses same ingredients as common frequentist calculations

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

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

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

Low-D Models ($m \lesssim 10$): Quadrature & MC Integration

Quadrature/Cubature Rules:

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

Smoothness $\rightarrow$ fast convergence in 1-D

Curse of dimensionality $\rightarrow O(n^{-2/m})$ or $O(n^{-4/m})$ in m-D

Monte Carlo Integration:

$$\int d\theta\, g(\theta) p(\theta) \approx \sum_{\theta_i \sim p(\theta)} g(\theta_i) + O(n^{-1/2}) \qquad \left[\begin{array}{c} \sim O(n^{-1}) \text{ with} \\ \text{quasi-MC} \end{array} \right]$$

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

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

Practical problem: multiplier is large (variance of g)
$\rightarrow$ hard if $m \gtrsim 6$ (need good "importance sampler" p)

Randomized Quadrature:

Quadrature rule + random dithering of abscissas
$\rightarrow$ get benefits of both methods

Most useful in settings resembling Gaussian quadrature

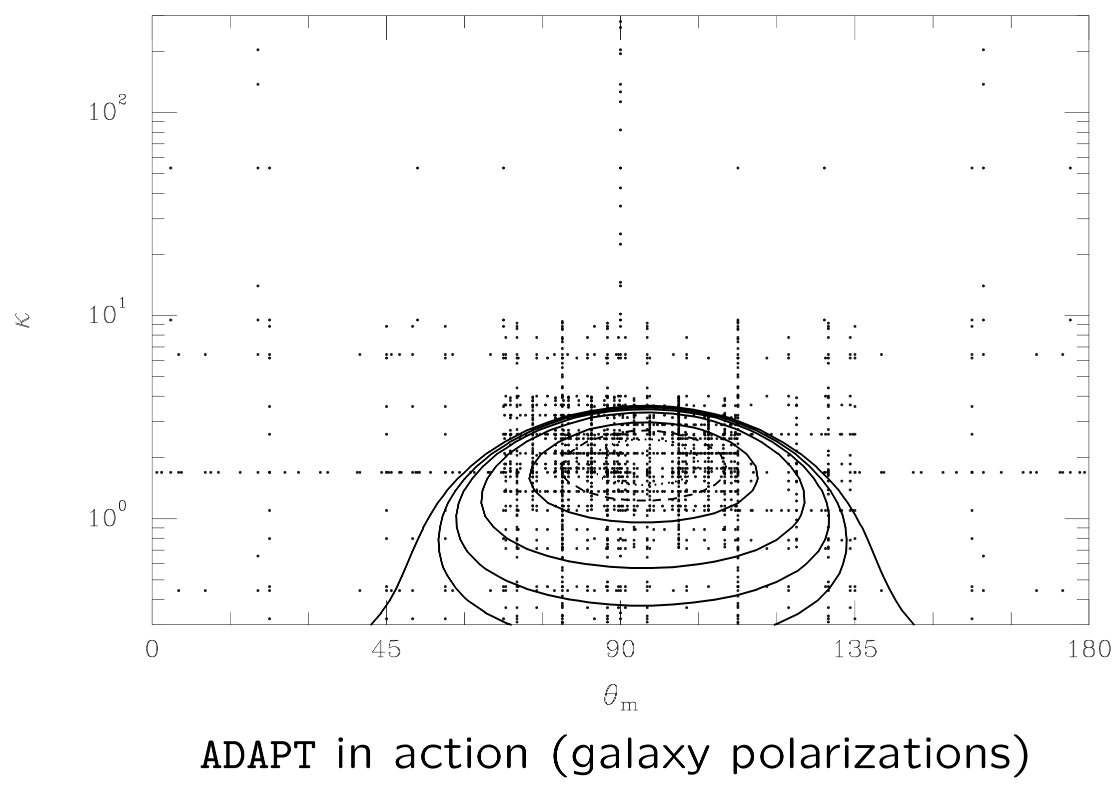
Subregion-Adaptive Quadrature/MC:

Concentrate points where most of the probability lies via recursion

Adaptive quadrature: Use a pair of lattice rules (for error estim'n), subdivide regions w/ large error

- `ADAPT` (Genz & Malik) at GAMS (`gams.nist.gov`)
- `BAYESPACK` (Genz; Genz & Kass)—many methods
  Automatic; regularly used up to $m \approx 20$

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



`ADAPT` in action (galaxy polarizations)

High-D Models ($m \sim 5$–10^6): Posterior Sampling

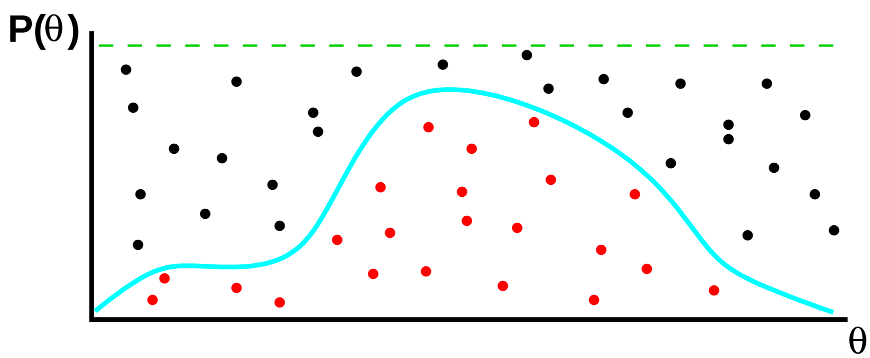
General Approach:

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

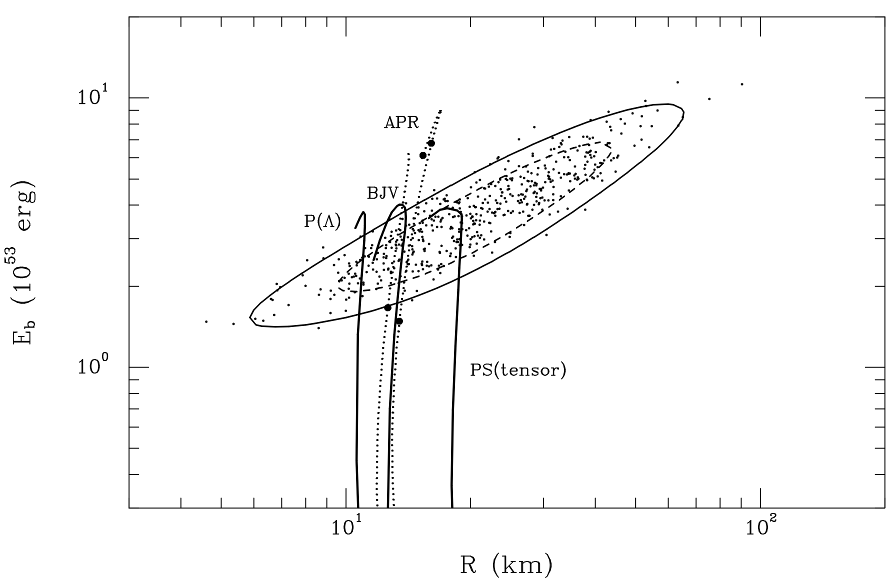
- Integrals, moments easily found via $\sum_i f(\theta_i, \phi_i)$
- $\{\theta_i\}$ are samples from $p(\theta|D, M)$

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

Rejection Method:



Hard to find efficient comparison function if $m \gtrsim 6$



A 2-D marginal of a 6-D posterior

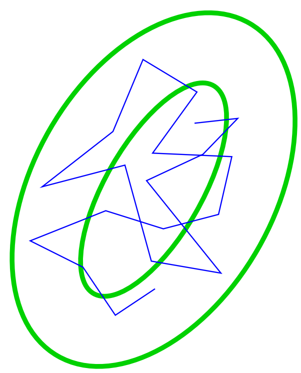
Markov Chain Monte Carlo (MCMC):

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

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

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

Markov chain methods are standard: Metropolis; Metropolis-Hastings; molecular dynamics; hybrid Monte Carlo; simulated annealing



The MCMC Recipe:

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

- Draw a candidate θ_{i+1} from a kernel $T(\theta_{i+1}|\theta_i)$
- Enforce "detailed balance" by accepting with $p = \alpha$

$$\alpha(\theta_{i+1}|\theta_i) = \min\left[1, \frac{T(\theta_i|\theta_{i+1})p(\theta_{i+1})}{T(\theta_{i+1}|\theta_i)p(\theta_i)}\right]$$

Choosing T to minimize "burn-in" and corr'ns is an art! Coupled, parallel chains eliminate this for select problems ("exact sampling").

Summary

What's different about Bayesian Inference:

- Problem-solving vs. solution-characterizing approach
- Calculate in parameter space rather than sample space

Bayesian Benefits:

- Rigorous foundations, consistent & simple interpretation
- Automated identification of statistics
- Numerous benefits from parameter space vs. sample space

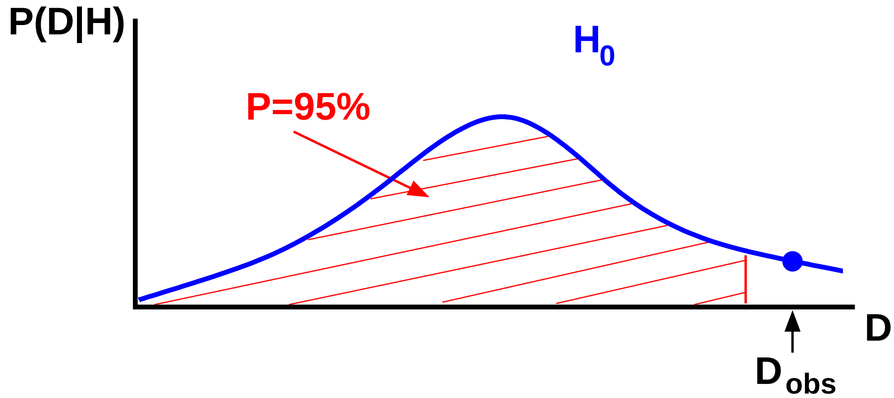
Bayesian Challenges:

- More complicated problem specification (≥ 2 alternatives; priors)
- Computational difficulties with large parameter spaces
  - Laplace approximation for "quick entry"
  - Adaptive & randomized quadrature for lo-D
  - Posterior sampling via MCMC for hi-D

Compare or Reject Hypotheses?

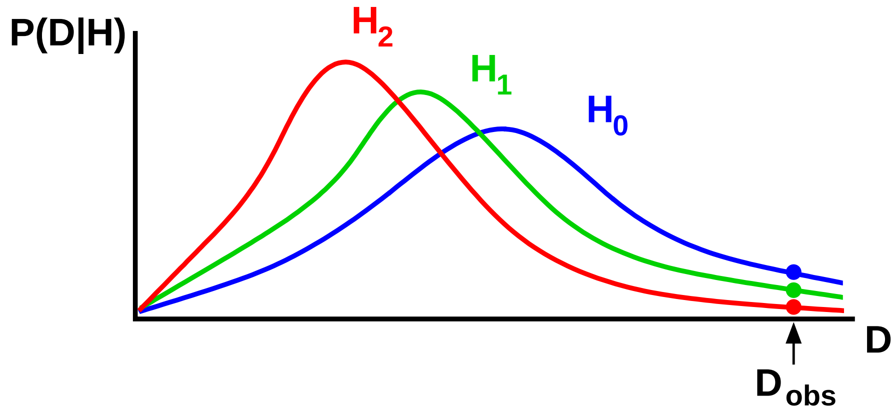
Frequentist Significance Testing (G.O.F. tests):

- Specify simple null hypothesis H_0 such that rejecting it implies an interesting effect is present
- Divide sample space into probable and improbable parts (for H_0)
- If D_{obs} lies in improbable region, reject H_0; otherwise accept it



Bayesian Model Comparison:

- Favor the hypothesis that makes the observed data most probable (up to a prior factor)



If the data are improbable under M_1, the hypothesis *may* be wrong, *or* a rare event may have occured. GOF tests reject the latter possibility at the outset.

Backgrounds as Nuisance Parameters

Background marginalization with Gaussian noise:

Measure background rate $b = \hat{b} \pm \sigma_b$ with source off.

Measure total rate $r = \hat{r} \pm \sigma_r$ with source on.

Infer signal source strength s, where $r = s + b$.

With flat priors,

$$p(s, b|D, M) \propto \exp\left[-\frac{(b - \hat{b})^2}{2\sigma_b^2}\right] \times \exp\left[-\frac{(s + b - \hat{r})^2}{2\sigma_r^2}\right]$$

Marginalize b to summarize the results for s (complete the square to isolate b dependence; then do a simple Gaussian integral over b):

$$p(s|D, M) \propto \exp\left[-\frac{(s - \hat{s})^2}{2\sigma_s^2}\right] \qquad \begin{aligned} \hat{s} &= \hat{r} - \hat{b} \\ \sigma_s^2 &= \sigma_r^2 + \sigma_b^2 \end{aligned}$$

Background *subtraction* is a special case of background *marginalization*.

Analytical Simplification: The Jaynes-Bretthorst Algorithm

Superposed Nonlinear Models

N samples of a superpos'n of nonlinear functions plus Gaussian errors,

$$\begin{aligned} d_i &= \sum_{\alpha=1}^{M} A_\alpha g_\alpha(x_i;\theta) + \epsilon_i \\ \text{or} \qquad \vec{d} &= \sum_\alpha A_\alpha \vec{g}_\alpha(\theta) + \vec{\epsilon}. \end{aligned}$$

The log-likelihood is a quadratic form in A_α,

$$\mathcal{L}(A,\theta) \propto \frac{1}{\sigma^N} \exp\left[-\frac{Q(A,\theta)}{2\sigma^2}\right]$$

$$\begin{aligned} Q &= \left[\vec{d} - \sum_\alpha A_\alpha \vec{g}_\alpha\right]^2 \\ &= d^2 - 2\sum_\alpha A_\alpha \vec{d}\cdot\vec{g}_\alpha + \sum_{\alpha,\beta} A_\alpha A_\beta \eta_{\alpha\beta} \end{aligned}$$

$$\eta_{\alpha\beta} = \vec{g}_\alpha \cdot \vec{g}_\beta$$

Estimate θ given a prior, $\pi(\theta)$.

Estimate amplitudes.

Compare rival models.

The Algorithm

- Switch to orthonormal set of models, $\vec{h}_\mu(\theta)$ by diagonalizing $\eta_{\alpha\beta}$; new amplitudes $B = \{B_\mu\}$.

$$Q = \sum_\mu \left[B_\mu - \vec{d}\cdot\vec{h}_\mu(\theta)\right]^2 + r^2(\theta, B)$$

$$\text{residual} \qquad \vec{r}(\theta, B) = \vec{d} - \sum_\mu B_\mu \vec{h}_\mu$$

$$p(B, \theta|D, I) \quad \propto \quad \frac{\pi(\theta)J(\theta)}{\sigma^N} \exp\left[-\frac{r^2}{2\sigma^2}\right] \exp\left[\frac{-1}{2\sigma^2}\sum_\mu (B_\mu - \widehat{B}_\mu)^2\right]$$

$$\text{where} \qquad J(\theta) = \prod_\mu \lambda_\mu(\theta)^{-1/2}$$

- Marginalize B's analytically.

$$p(\theta|D, I) \propto \frac{\pi(\theta)J(\theta)}{\sigma^{N-M}} \exp\left[-\frac{r^2(\theta)}{2\sigma^2}\right]$$

$$r^2(\theta) = \begin{array}{l}\text{residual sum of squares}\\ \text{from least squares}\end{array}$$

- If σ unknown, marginalize using $p(\sigma|I) \propto \frac{1}{\sigma}$.

$$p(\theta|D, I) \propto \pi(\theta)J(\theta)\left[r^2(\theta)\right]^{\frac{M-N}{2}}$$

Frequentist Behavior of Bayesian Results

Bayesian inferences have good long-run properties, sometimes better than conventional frequentist counterparts.

Parameter Estimation:

- Credible regions found with flat priors are typically confidence regions to $O(n^{-1/2})$.
- Using standard nonuniform "reference" priors can improve their performance to $O(n^{-1})$.
- For handling nuisance parameters, regions based on marginal likelihoods have superior long-run performance to regions found with conventional frequentist methods like profile likelihood.

Model Comparison:

- Model comparison is asymptotically consistent. Popular frequentist procedures (e.g., χ^2 test, asymptotic likelihood ratio ($\Delta\chi^2$), AIC) are not.
- For separate (not nested) models, the posterior probability for the true model converges to 1 exponentially quickly.
- When selecting between more than 2 models, carrying out multiple frequentist significance tests can give misleading results. Bayes factors continue to function well.