Computation for Bayesian model comparison

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

Bayesian Computation @ CASt — 11–13 June 2014
Classes of Problems
[Recall CosPop intro Bayes lecture]

**Single-model inference**
Premise = choice of single model, $M$

*Parameter estimation*: What can we say about $\theta$ or $f(\theta)$?
*Prediction*: What can we say about future data $D'$?

**Multi-model inference**
Premise = $\{M_i\}$

*Model comparison/choice*: What can we say about $i$?

*Model averaging*:
- *Systematic error*: $\theta_i = \{\phi_i, \eta_i\}$; $\phi$ is common to all
  What can we say about $\phi$ w/o committing to one model?
- *Prediction*: What can we say about future $D'$, accounting
  for model uncertainty?

**Model checking**
Premise = $M_1 \lor$ “all” alternatives
Is $M_1$ adequate? (predictive tests, calibration, robustness)
Model Comparison

Problem statement

\[ I = (M_1 \lor M_2 \lor \ldots) \] — Specify a set of models.
\[ H_i = M_i \] — Hypothesis chooses a model.

Posterior probability for a model

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

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

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

Varied terminology: Prior predictive = Average likelihood = Global likelihood = Marginal likelihood = (Weight of) Evidence for model
Computation for model comparison

1 Marginal likelihood computation
   Cubature
   Posterior density estimation
   Posterior expectations
   Randomized cubature
   Importance sampling

2 Bayes factors via trans-dimensional MCMC
   Reversible-jump MCMC
   Birth-death MCMC

3 Guidance
Computation for model comparison

1. Marginal likelihood computation
   - Cubature
   - Posterior density estimation
   - Posterior expectations
   - Randomized cubature
   - Importance sampling

2. Bayes factors via trans-dimensional MCMC
   - Reversible-jump MCMC
   - Birth-death MCMC

3. Guidance
Computation for model comparison

1. **Marginal likelihood computation**
   - Cubature
   - Posterior density estimation
   - Posterior expectations
   - Randomized cubature
   - Importance sampling

2. **Bayes factors via trans-dimensional MCMC**
   - Reversible-jump MCMC
   - Birth-death MCMC

3. **Guidance**
Adaptive Cubature

[Recall CosPop lecture on low-D computing]

- **Subregion adaptive cubature**: Use a pair of monomial rules (for error estim’n); recursively subdivide regions w/ large error (ADAPT, CUHRE, 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).

<table>
<thead>
<tr>
<th># nodes used by ADAPT’s 7th order rule</th>
<th>$2^d + 2d^2 + 2d + 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimen</td>
<td>2  3  4  5  6  7  8  9  10</td>
</tr>
<tr>
<td># nodes</td>
<td>17 33 57 93 149 241 401 693 1245</td>
</tr>
</tbody>
</table>
Analysis of Galaxy Polarizations

TJL, Flanagan, Wasserman (1997)
Computation for model comparison

1 Marginal likelihood computation
   Cubature
   Posterior density estimation
   Posterior expectations
   Randomized cubature
   Importance sampling

2 Bayes factors via trans-dimensional MCMC
   Reversible-jump MCMC
   Birth-death MCMC

3 Guidance
Basic marginal likelihood identity

We seek to directly compute the marginal likelihood for a single model considered in isolation:

\[ Z = \int d\theta \, \pi(\theta) \mathcal{L}(\theta) = \int d\theta \, q(\theta) \]

A simple but \textit{bad} idea is based on “candidate’s formula,” aka “basic marginal likelihood identity” (BMI):

\[ p(\theta|D, M) = \frac{\pi(\theta) \mathcal{L}(\theta)}{Z} \]

\[ \rightarrow Z = \frac{\pi(\theta) \mathcal{L}(\theta)}{p(\theta|D, M)}, \quad \text{for any } \theta \text{ in support} \]
Implementation:

\[ Z = \frac{\pi(\theta)L(\theta)}{p(\theta|D, M)} \]

1. Get posterior samples
2. Use samples + a density estimator to estimate \( p(\theta|D, M) \) at some \( \theta = \theta^* \) (probably near the mode is good)
3. Evaluate the formula: \( \hat{Z} = \frac{\pi(\theta^*)L(\theta^*)}{\hat{p}(\theta^*|D, M)} \)

Fails in more than very few dimensions because of the curse of dimensionality for nonparametric density estimation (next slide)

(But see Hsiao, Huang & Chang 2004 for an attempt to fix it)

It has two ideas that appear in other methods (useful and bad!):

- Using a posterior density estimator
- Using an identity from Bayes’s theorem
Curse of dimensionality for KDE

Estimate a normal density at the origin to 10% using Gaussian-kernel KDE with optimal smoothing.

Sample size for 10% MSE at origin vs. Dimension

Silverman (1986)
Chib’s method
For marginal likelihoods from *Gibbs sampler* output:

E.g., consider a *missing data* or *latent variable* problem (MLM!) with interesting parameters $\theta$, and missing/latent parameters $\psi$:

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

Implement via Gibbs or Metropolis-with-Gibbs, alternating:

- $\theta_i \sim p(\theta|\psi, D)$
- $\psi_i \sim p(\psi|\theta, D)$

Suppose the full conditional $p(\theta|\psi, D)$ is known, *including its normalization constant* (analytically or numerically)

$$\Rightarrow p(\theta^*|D) \approx \frac{1}{N} \sum_{i=1}^{N} p(\theta^*|\psi_i, D)$$

Use this *bespoke* finite mixture density estimator in the BMI
Applications in astronomy

- **Exoplanet RV data**: Tuomi (2011) analysis of RV data from Gliese 581, arguing for 4 planets
  Uses Metropolis-Hastings generalization due to Chib & Jeliazkov (2001)—the proposal dist’n and acceptance probabilities appear in the mixture

- **Ultra-high energy cosmic ray directions**: Soiaporn$^+$ (2012) implement Bayesian cross-matching of UHECR and local AGN directions in an MLM accounting for selection and measurement error
  Metropolis-within-Gibbs algorithm enables use of Chib estimate (using numerical normalization of a full conditional)
Savage-Dickey density ratio

For model comparison with *nested models*:

$M_1$: Parameters $\theta$, likelihood $\mathcal{L}_1(\theta)$

$M_2$: Parameters $(\theta, \phi)$, likelihood $\mathcal{L}_2(\theta, \phi)$

Let $\phi_0 = \text{value of } \phi \text{ assumed by } M_1$:

$$\mathcal{L}_1(\theta) = \mathcal{L}_2(\theta, \phi_0)$$

Assume priors are independent:

$$p(\theta|M_1) = f(\theta)$$

$$p(\theta, \phi|M_2) = f(\theta)g(\phi)$$

(may be relaxed).
Compare models via marginal likelihoods:

\[ \mathcal{L}(M_1) = \int d\theta \, f(\theta) \mathcal{L}_2(\theta, \phi_0) \]

\[ \mathcal{L}(M_2) = \int d\theta d\phi \, f(\theta) f(\phi) \mathcal{L}_2(\theta, \phi) \]

Due to nesting, integrals appear similar! Note:

\[ p(\phi\mid D, M_2) = \frac{1}{\mathcal{L}(M_2)} \int d\theta \, f(\theta) g(\phi) \mathcal{L}_2(\theta, \phi) \]

Now calculate \( \mathcal{L}(M_1) \), using \( \mathcal{L}_2(\theta, \phi_0) \):

\[ \mathcal{L}(M_1) = \int d\theta \, f(\theta) g(\phi_0) \mathcal{L}_2(\theta, \phi_0) \times \frac{1}{g(\phi_0)} \]

\[ = \frac{p(\phi_0\mid D, M_2)}{p(\phi_0\mid M_2)} \mathcal{L}(M_2) \]

\[ \rightarrow B_{21} = \frac{p(\phi_0\mid M_2)}{p(\phi_0\mid D, M_2)} \sim \frac{1}{p(\phi_0\mid D, M_2)} \]

Can approximate this via MCMC with only \( M_2 \), as long as \( \phi_0 \) isn’t too far in tail and \( \phi \) is low-dimensional (1 or 2!)
Trotta (2007) – Comparison of nested cosmological models

Table 2. Summary of model comparison results from WMAP data combined with small-scale CMB measurements, SDDS, HST and SNIa data. WMAP3+ext refers to WMAP 3-year data release, WMAP1 + ext to WMAP first-year data. The most spectacular improvement from WMAP1 to WMAP3 is the moderate evidence against a scale-invariant spectral index. Errors in the Bayes factor are obtained by computing the variance of the SDDR estimate from five subchains (see Appendix C for details). The ‘estimate’ column gives the value obtained by employing the Gaussian approximation to the likelihood, equation (A9) for a Gaussian prior or equation (A10) for a flat prior.

<table>
<thead>
<tr>
<th>Data</th>
<th>$\ln B_{01}$ from SDDR (numerical)</th>
<th>$\ln B_{01}$ from SDDR (estimate)</th>
<th>Odds in favour of simpler model</th>
<th>Probability of simpler model</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 to 17</td>
<td>0.05</td>
<td></td>
<td></td>
<td>Moderate evidence for non-scale invariance</td>
</tr>
<tr>
<td>WMAP3+ext</td>
<td>$-2.86 \pm 0.28$</td>
<td>$-3.00$</td>
<td></td>
<td></td>
<td>Inconclusive result</td>
</tr>
<tr>
<td>WMAP1+ext</td>
<td>$0.68 \pm 0.04$</td>
<td>$0.71$</td>
<td>$2$ to $1$</td>
<td>$0.66$</td>
<td>Moderate evidence for a flat universe</td>
</tr>
<tr>
<td>WMAP3+ext</td>
<td>$3.37 \pm 0.05$</td>
<td>$3.25$</td>
<td>$29$ to $1$</td>
<td>$0.97$</td>
<td>Moderate evidence for a flat universe</td>
</tr>
<tr>
<td>WMAP1+ext</td>
<td>$2.70 \pm 0.09$</td>
<td>$2.68$</td>
<td>$15$ to $1$</td>
<td>$0.94$</td>
<td>Moderate evidence for a flat universe</td>
</tr>
<tr>
<td>WMAP3+ext</td>
<td>$7.62 \pm 0.02$</td>
<td>$7.63$</td>
<td>Adiabaticity: $f_{\text{iso}} = 0$ versus $-100 \leq f_{\text{iso}} \leq 100$ (flat)</td>
<td>$0.9995$</td>
<td>Strong evidence for adiabatic conditions</td>
</tr>
<tr>
<td>WMAP1+ext</td>
<td>$7.50 \pm 0.03$</td>
<td>$7.53$</td>
<td>$1800$ to $1$</td>
<td>$0.9994$</td>
<td>Strong evidence for adiabatic conditions</td>
</tr>
</tbody>
</table>
Computation for model comparison

1 Marginal likelihood computation
   Cubature
   Posterior density estimation
   Posterior expectations
   Randomized cubature
   Importance sampling

2 Bayes factors via trans-dimensional MCMC
   Reversible-jump MCMC
   Birth-death MCMC

3 Guidance
Harmonic mean of the likelihood

Take the reciprocal of the BMI:

\[
\frac{1}{Z} = \frac{p(\theta|D, M)}{\pi(\theta)L(\theta)}
\]

If we integrate over \( \theta \), the RHS will look like a posterior expectation. To control LHS, multiply by a density—e.g., the prior:

\[
\frac{1}{Z} = \int d\theta \frac{p(\theta|D, M)}{L(\theta)}
\]

Estimate by Monte Carlo via posterior samples \( \{\theta_i\} \):

\[
\hat{Z}_{HM} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{L(\theta_i)}
\]

Appealingly simple, but...
The Harmonic Mean of the Likelihood: Worst Monte Carlo Method Ever

2008-08-17 at 12:09 am | 36 comments

Many Bayesian statisticians decide which of several models is most appropriate for a given dataset by computing the marginal likelihood of each model (also called the integrated likelihood or the evidence). The marginal likelihood is the probability that the model gives to the observed data, averaging over values of its parameters with respect to their prior distribution. If \( x \) is the entire dataset and \( t \) is the entire set of parameters, then the marginal likelihood is

\[
P(x) = \int P(x|t) P(t) \, dt
\]

“The good news is that the Law of Large Numbers guarantees that this estimator is consistent. . . . The bad news is that the number of points required for this estimator to get close to the right answer will often be greater than the number of atoms in the observable universe. The even worse news is that it’s easy for people to not realize this, and to naively accept estimates that are nowhere close to the correct value of the marginal likelihood.”
\[ \hat{Z}_{\text{HM}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\mathcal{L}(\theta_i)} \]

Qualitative explanation

Recall that \( Z \approx \mathcal{L}(\hat{\theta}) \frac{\delta \theta}{\Delta \theta} \) posterior width
\[ \hat{Z}_{\text{HM}} \] has to give very different answers for \( \Delta \theta = 5 \times \delta \theta \) and \( \Delta \theta = 500 \times \delta \theta \), even though posteriors are very similar and it only knows about \( \mathcal{L}(\theta_i) \rightarrow \) its value must be dominated by rare contributions from the tails

Theoretical explanation

Wolpert & Schmidler (2012): \( \hat{Z}_{\text{HM}} \) converges in distribution to a one-sided stable law with parameters such that the rate of convergence is typically \( N^{-\epsilon} \) with \( \epsilon = 0.1 \) or even 0.01.
“Those who don’t know history . . .”

“Probabilities of exoplanet signals from posterior samplings”
tries to fix HM but creates an even worse (i.e., inconsistent) estimator; see Christian Robert’s blog (3 Jan 2012)

Potential fixes

- **Weighted harmonic mean:** Gelfand & Dey (1994) integrate the reciprocal BMI with a PDF $g(\theta)$ different from the prior:

$$\hat{Z}_{WHM} = \frac{1}{N} \sum_{i=1}^{N} \frac{g(\theta_i)}{\pi(\theta_i) \mathcal{L}(\theta_i)}$$

- **Subdomain approaches:** Weinberg$^+$ (2012, 2013; see BIE) use posterior samples to identify a high-probability subregion for integrals, avoiding variability from tail contributions:

$$\frac{1}{Z} \int_{\Omega} d\theta \, \pi(\theta) = \int_{\Omega} d\theta \, \frac{p(\theta|D)}{\mathcal{L}(\theta)}$$
Thermodynamic integration, bridge & path sampling

Underlying idea
Calculating a single complicated $N$-D integral may be hard, but calculating the ratio of similar $N$-D integrals can be easier

⇒ Calculate $Z$ via a product of tractable ratio calculations along a path of integrands from a simple one to $q(\theta)$

Thermodynamic integration

Let $Z(\beta) \equiv \int d\theta \pi(\theta) \mathcal{L}^\beta(\theta)$

$Z = \frac{Z(1)}{Z(0)} = \frac{Z(1)}{Z(.5)} \times \frac{Z(.5)}{Z(0)} = \prod_{i=0}^{N-1} \frac{Z(\beta_{i+1})}{Z(\beta_i)}$

for a sequence of increasing “tempers” (“inverse temperatures”) $\beta_i$ with $\beta_0 = 0$, $\beta_N = 1$
Tempered likelihood functions

\[ p(\theta) \]

\[ \begin{align*}
\theta &\quad 0.0 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 \\
p(\theta) &\quad 1.0 & 0.7 & 0.5 & 0.2 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 
\end{align*} \]
**Estimating the ratios**

Write $\beta_{i+1} = \beta_i + \delta_i$; consider small $\delta_i$

$$Z = \prod_{i=0}^{N-1} \frac{Z(\beta_{i+1})}{Z(\beta_i)} \quad \rightarrow \quad \ln Z = \sum_i \ln \frac{Z(\beta_i + \delta_i)}{Z(\beta_i)}$$

$$\ln Z \approx \sum_i \ln \left[ 1 + \delta_i \frac{Z'(\beta_i)}{Z(\beta_i)} \right] \approx \sum_i \delta_i R(\beta_i) \quad \rightarrow \quad \int d\beta \, R(\beta)$$

where $R(\beta) \equiv \frac{Z'(\beta)}{Z(\beta)} = \frac{1}{Z(\beta)} \frac{d}{d\beta} \int d\theta \, \pi(\theta) \exp [\beta \log L(\theta)]$

$$= \frac{1}{Z(\beta)} \int d\theta \, \log[L(\theta)] \pi(\theta) L^\beta(\theta)$$

So $R(\beta) = \langle \log L \rangle_\beta$, the (annealed) posterior expectation of the log-likelihood
Thermodynamic integration algorithm

\[
\ln Z = \int d\beta \langle \log L \rangle_\beta
\]

- For each of an “annealing schedule” of \( \beta \) values:
  - Use MCMC to get \( \{\theta_j\} \) with \( \theta \sim \pi(\theta)L^\beta(\theta) \)
  - Find \( \hat{R}(\beta) = \frac{1}{N} \sum_j \log L(\theta_j) \)
- Estimate \( \ln \hat{Z} = \int d\beta \hat{R}(\beta) \) via 1-D quadrature
- Issues:
  - Requires MCMC of multiple tempered posteriors
  - Much of the integral can be in small \( \beta \) range near \( \beta = 1 \) → need more tempers than for param estimation

Generalizations: bridge and path sampling (related to importance sampling), Gelman & Meng (1998)
## Applications in astronomy

**Gregory (2005)**
Parallel tempering and thermo integ’n for exoplanet model comparison

| Model   | Period (days) | Mean $v$ (m s$^{-1}$) | $p(D|\text{Model, } I)$ | Bayes Factor | Probability |
|---------|---------------|------------------------|--------------------------|--------------|-------------|
| $M_0$... | ...           | 0                      | $9.6 \times 10^{-232}$   | $3.4 \times 10^{-188}$ | $2 \times 10^{-149}$ |
| $M_{0a}$ | ...           | 69                     | $2.8 \times 10^{-44}$    | 1.0          | $6 \times 10^{-7}$ |
| $M_1$... | 128           | 0                      | $(7.5 \pm 0.03) \times 10^{-40}$ | ...          | ...         |
| ...     | 190           | 0                      | $(2.2 \pm 0.02) \times 10^{-40}$ | ...          | ...         |
| ...     | 376           | 0                      | $(1.5 \pm 0.2) \times 10^{-38}$ | ...          | ...         |
| ...     | All           | 0                      | $(1.6 \pm 0.2) \times 10^{-38}$ | $5.7 \times 10^5$ | 0.36        |
| $M_{1s}$ | 128           | 7                      | $(1.2 \pm 0.03) \times 10^{-39}$ | ...          | ...         |
| ...     | 190           | 11                     | $(1.1 \pm 0.01) \times 10^{-39}$ | ...          | ...         |
| ...     | 376           | 8                      | $(2.6 \pm 0.3) \times 10^{-38}$ | ...          | ...         |
| ...     | All           | 8                      | $(2.8 \pm 0.3) \times 10^{-38}$ | $9.3 \times 10^5$ | 0.64        |

**Littenberg & Cornish (2009)**
Parallel tempering and thermo integ’n for grav’l wave detection

*Integrand vs. SNR*
Computation for model comparison

1. **Marginal likelihood computation**
   - Cubature
   - Posterior density estimation
   - Posterior expectations
   - Randomized cubature
   - Importance sampling

2. **Bayes factors via trans-dimensional MCMC**
   - Reversible-jump MCMC
   - Birth-death MCMC

3. **Guidance**
Adaptive simplex cubature

Another bad method!

Motivation: Use MCMC sample locations and densities

Suppose you were given \( \{ \theta_i, q_i \} \) and told to estimate \( Z = \int d\theta q(\theta) \) for this 1-d \( q(\theta) \).

Use a quadrature approximation that doesn't require specific abscissas: histogram, trapezoid, etc. These weight by "volume" factors:

\[
Z = \sum_{\text{intervals}} \text{(length)} \times \text{(avg. height)}
\]

In 2-d intervals are triangles (2-simplices); length \( \rightarrow \) area. Make the triangles via Delaunay triangulation.
Higher dimensions: Combine $n$-d Delaunay triangulation and $n$-d simplex trapezoidal rules of Lyness & Genz (1980)

Note that this is valid for points drawn from *any* distribution that covers the support; it’s not obvious that posterior samples are optimal
**Performance**

Explored up to 6-d with a variety of standard test-case normal mixtures, using samples as vertices. **Qhull** used for triangulation.

Triangulation is *expensive* → use a small number of vertices.

In few-d, requires many fewer points than subregion-adaptive cubature (DCUHRE), but underestimates integrals in $>4$-D. There is lots of volume in the outer “shell” so even though density is low, it contributes a lot.

**Modifications**

- Tempered/derivative-weighted resampling (seems to work to 6- or 7-D)
- Non-optimal triangulations
- Weinberg$^+$ (2013) — Subdomain sampling/cubature method
Lebesgue integration and nested sampling
Adaptive simplex quadrature implements a Riemann integral in $d$-D

But there are other ways to define an integral!

Riemann integral: Partition abscissa
Lebesgue integral: Partition *ordinate*

\[ Z_L \approx \sum_{i} f_i \mu_i(x); \quad \mu_i(x) = \text{“measure” of } x \text{ at } f_i \]
Two dimensions

\[ Z_R \approx \sum_i \sum_j f(x_i, y_j) \delta x \delta y \]

\[ Z_L \approx \sum_i f_i \mu_i(x, y) \]

where now the measure is the area in contours about \( f_i \).
Skilling’s Nested Sampling

Nested sampling is a kind of numerical Lebesgue integral, with a random twist:

- $\mu(\theta)$ for contour interval is estimated statistically
- The contour levels $f_i$ are specified randomly, marching up in likelihood

- Achille’s heel: How to sample inside contour(s)
- MultiNest does this approximately; performance uncertain
- See Brewer’s diffusive nested sampling for a more rigorous approach
Crowded field star detection/measurement, diffusive nested sampling

100-star image

Posterior sample catalogs
(1000 star case)

Number-size dist’n

Inference
SExtractor
True

Number (Flux > f)

0.0

1.0

10.0

100.0

Test Case 1
Computation for model comparison

1 Marginal likelihood computation
   Cubature
   Posterior density estimation
   Posterior expectations
   Randomized cubature
   Importance sampling

2 Bayes factors via trans-dimensional MCMC
   Reversible-jump MCMC
   Birth-death MCMC

3 Guidance
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!)

Can be useful for both model comparison (marginal likelihood calculation), and parameter estimation.
Building a Good Importance Sampler

Estimate an annealing target density, $\pi_n$, using a mixture of multivariate Student-t distributions, $P_n$:

$$q_n(\theta) = [q_0(\theta)]^{1-\lambda_n} \times [q(\theta)]^{\lambda_n}, \quad \lambda_n = 0 \ldots 1$$

$$P_n(\theta) = \sum_j \text{MVT}(\theta; \mu_j^n, S_j^n, \nu)$$

Adapt the mixture to the target using ideas from sequential Monte Carlo $\rightarrow$ Adaptive annealed importance sampling (AAIS)

Initialization
Sample, weight, refine

Sample & calculate weights

Refine IS: EM + Birth/Death

Overall algorithm

Target

Design

AAIS Step
2-D Example:
Many well-separated correlated normals

\( \lambda_1 = 0.01 \)
\( \lambda_3 = 0.11 \)
\( \lambda_8 = 1 \)
Observed Data:
HD 73526 (2 planets)

Data and RV Curve for 2-Planet Fit
Periods: 188 d, 377 d (weakly resonant)

Bayes factors:
1 vs 0 planet: $6.5 \times 10^6$
2 vs 1 planet(s): $8.2 \times 10^4$

1-D and 2-D Marginals for Orbital Parameters
(longer-period planet)

Sampling efficiency of final mixture $\text{ESS}/N \approx 65\%$
Computation for model comparison

1. Marginal likelihood computation
   - Cubature
   - Posterior density estimation
   - Posterior expectations
   - Randomized cubature
   - Importance sampling

2. Bayes factors via trans-dimensional MCMC
   - Reversible-jump MCMC
   - Birth-death MCMC

3. Guidance
Trans-dimensional MCMC

Trans-dimensional MCMC performs posterior sampling on the *dimensionally inhomogeneous* space of model index and parameters, \((M_i, \theta_i)\)

The posterior probability for model \(i\) is just the frequency of sampling that model

Frameworks: Reversible-jump MCMC, product-space MCMC, birth-death processes

Particularly suited to large model spaces where most probability will be in a few models; trans-
\(D\) MCMC can often find them

Not well-suited to settings where you need to know the value of a large or small Bayes factor, e.g., for just a few competing models (frequencies may be small or zero)
Computation for model comparison

1 Marginal likelihood computation
   Cubature
   Posterior density estimation
   Posterior expectations
   Randomized cubature
   Importance sampling

2 Bayes factors via trans-dimensional MCMC
   Reversible-jump MCMC
   Birth-death MCMC

3 Guidance
Reversible-jump MCMC

Supplement the usual MH algorithm with a set of moves from one model to another, and a varying number of auxiliary parameters so that the total number of parameters is constant.

Create a consistent set of mappings that use the auxiliary parameters to determine parameters for a proposed model from the parameters of the current model. This must be a bijection.

Add factors to the Metropolis-Hastings acceptance ratio accounting for the model moves and the mappings.

Now just follow the MH recipe!
Reversible jump example

Two models, \( M_1 : \theta \quad M_2 : \theta_1, \theta_2 \)

Two between-model moves (besides within-model moves):

- Go from 2 to 1 with probability \( r_1 \), setting
  \[
  \theta = \frac{1}{2}(\theta_1 + \theta_2)
  \]

- Go from 1 to 2 with probability \( r_2 \), picking a random \( u \) and setting
  \[
  \theta_1 = \theta + u; \quad \theta_2 = \theta - u
  \]

Adjust the usual MH \( \alpha \) by factors accounting for the move probabilities, the dist’n for \( u \), and the Jacobian

\[
|\partial(\theta, u)/\partial(\theta_1, \theta_2)|
\]
FIG. 2 (color online). First $3 \times 10^4$ iteration of the RJMCMC output when comparing a seven- and a five-dimensional LTP model (models X and Y, respectively). Since the models are not competitive when the SNR = 60, the blue line tends asymptotically to zero.

FIG. 4. A RJMCMC run on a set of nested LTP models. There is a clear preference for the five-dimensional model for the given data set. The data were produced with a “perfect” model in which the two respective actuators were identical.

Also:

- Umstaetter & Tinto (2008) — Detection of gravitational waves from coalescing binaries (chirps)
- Stroeer & Veitch (2009) — Extracting WD binary signals from LISA’s colored noise
Computation for model comparison

1 Marginal likelihood computation
   Cubature
   Posterior density estimation
   Posterior expectations
   Randomized cubature
   Importance sampling

2 Bayes factors via trans-dimensional MCMC
   Reversible-jump MCMC
   Birth-death MCMC

3 Guidance
Birth-death MCMC

Setting
Competing models with different numbers of components of the same form but with different parameter values:

- Finite mixture model for density estimation
- Superposition of pulses
- Superposition of (non)linear regression components

Approach
Represent the competing models as realizations of a marked point process

Explore via *birth–death–split–merge* moves; no auxiliary parameters needed
Bayesian droplet decomposition of BATSE GRB trigger #540

Broadbent \(^+\) (in prep.)

Two representative GiG decompositions

\[ \chi^2 = 220.0 \text{ on } 68 \text{ df} \]

\[ \chi^2 = 220.0 \text{ on } 76 \text{ df} \]

\( \# \text{ of pulses} \)
Computation for model comparison

1. Marginal likelihood computation
   - Cubature
   - Posterior density estimation
   - Posterior expectations
   - Randomized cubature
   - Importance sampling

2. Bayes factors via trans-dimensional MCMC
   - Reversible-jump MCMC
   - Birth-death MCMC

3. Guidance
Provisional guidance

From 2003 and 2006 SAMSI programs

- Calculate marginal likelihoods directly when comparing a small set of models; use trans-dimensional MCMC when exploring a large model space (with a small but unknown subset likely to be favored)
- “It is important to try to implement more than one method and test code on examples with known marginals, if nothing else because it is very easy to make mistakes in coding!”
- Methods using posterior samples will likely require much longer runs than are needed for parameter estimation; too-short runs can produce severe errors
- Chib’s method often performs well when it can be easily implemented; complex Gibbs sampling, and the M-H variant, appear less stable
- Low-D (<15): Mixture-based importance sampling guided by MCMC output is often easiest to implement with good accuracy [also try cubature]
- Explore robustness to priors!

SCMA ’06 review paper, “Current Challenges in Bayesian Model Choice” (Clyde† 2007)