STSCI 4780
Bayesian computation, cont'd: Computation for model comparison

Tom Loredo, CCAPS & SDS, Cornell University

Agenda

1. **Marginal likelihood computation**
   - Cubature
   - Randomized cubature — Nested sampling
   - Posterior density estimation
   - Posterior expectations
   - Importance sampling
2. **Bayes factors via trans-dimensional MCMC**
   - Reversible-jump MCMC
   - Birth-death MCMC
3. **Guidance**

Classes of model uncertainty problems

Single-model inference

Context = choice of single model (specific i)

Parameter estimation: What can we say about θ_i or $f(\theta_i)$?

Prediction: What can we say about future data D'?

Multi-model inference

Context = $M_1 \vee M_2 \vee \cdots$

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

Model averaging:

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

Model checking

Premise = $M_1 \vee$ "all" alternatives

Is M_1 adequate? (predictive tests, calibration, robustness)

Model Comparison

Problem statement

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

Posterior probability for a model

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

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
   - Randomized cubature — Nested sampling
   - Posterior density estimation
   - Posterior expectations
   - Importance sampling
2. **Bayes factors via trans-dimensional MCMC**
   - Reversible-jump MCMC
   - Birth-death MCMC
3. **Guidance**

Agenda

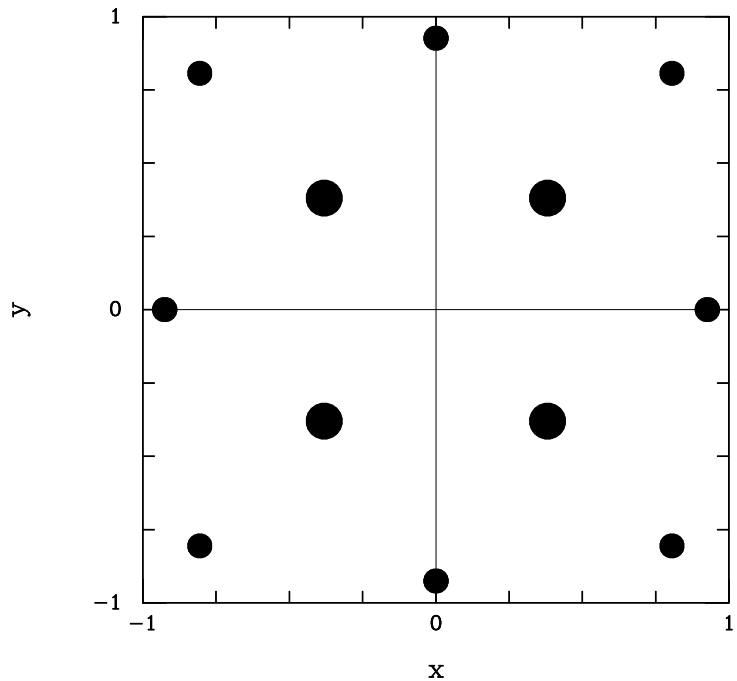
1. **Marginal likelihood computation**
   - Cubature
   - Randomized cubature — Nested sampling
   - Posterior density estimation
   - Posterior expectations
   - Importance sampling
2. **Bayes factors via trans-dimensional MCMC**
   - Reversible-jump MCMC
   - Birth-death MCMC
3. **Guidance**

Agenda

Adaptive Cubature

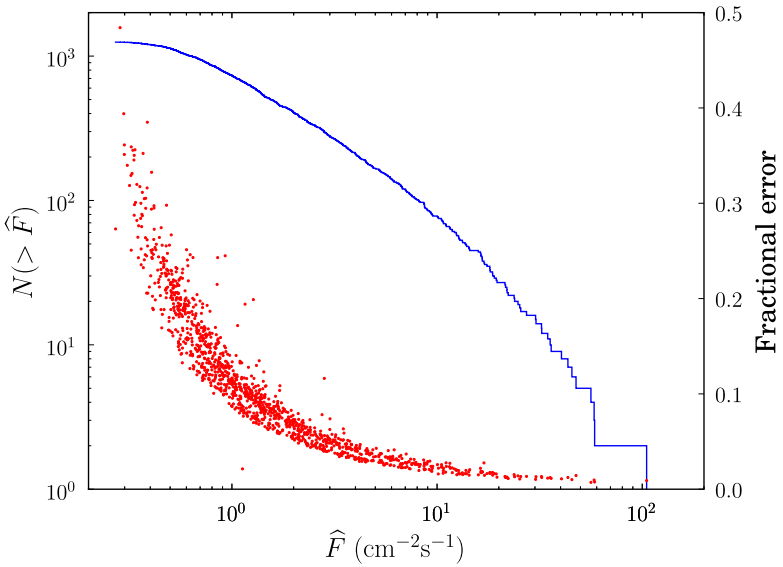
Subregion adaptive cubature: Use a pair of monomial rules (for error estim'n); recursively subdivide regions w/ large error (`ADAPT`, `CUHRE`, `BAYESPACK`, `CUBA`). Concentrates points where most of the probability lies.

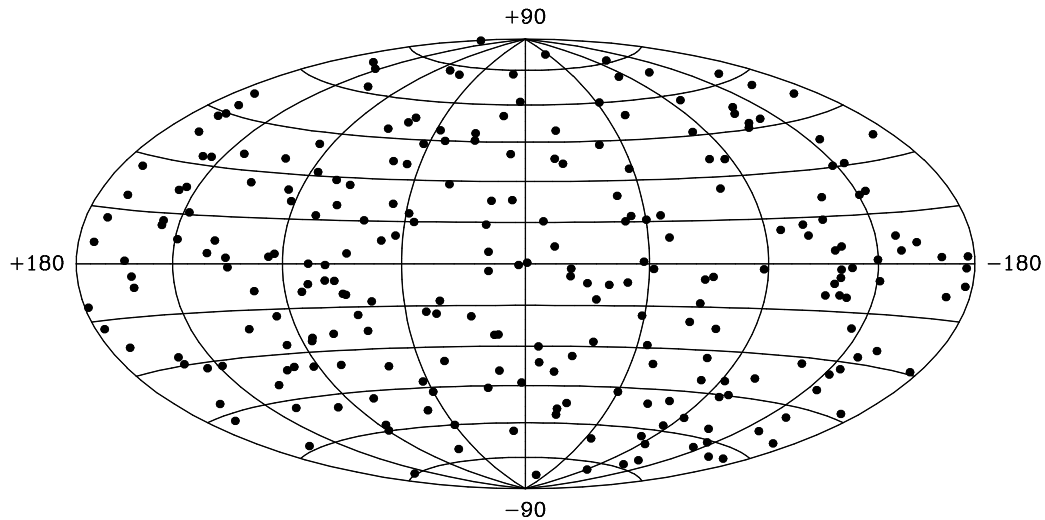
A 7th order rule in 2-d



Hundreds of Parameters. . . Without MCMC

Peak fluxes and directions of GRBs from 4B catalog

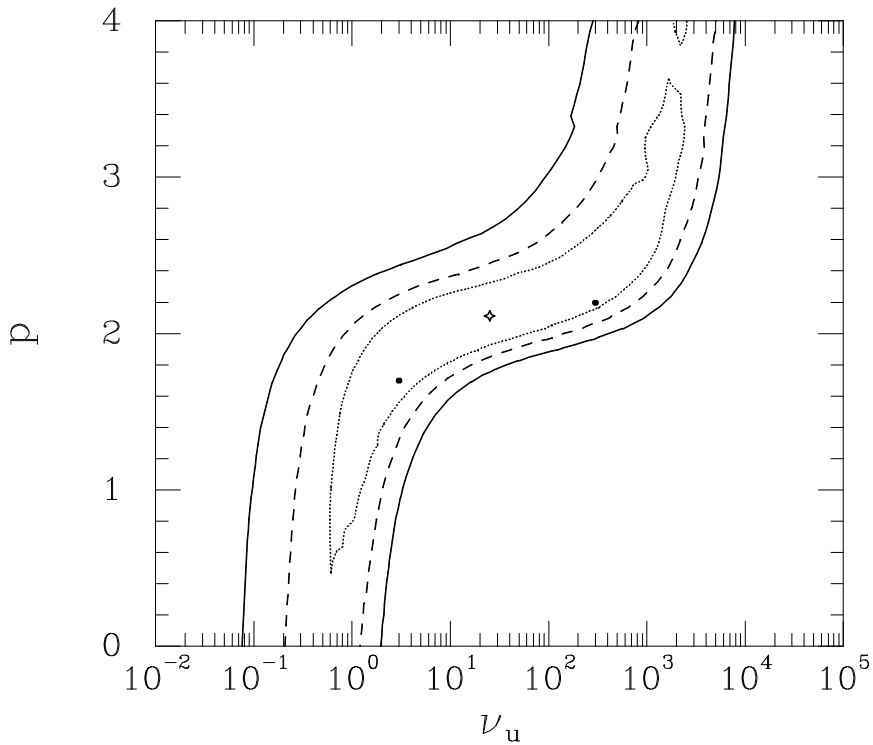




- How luminous and distant are burst sources?
- Indications for local (anisotropic) population?
- Classes, coincidences, many other stat questions. . .

Computation

- Model data using a 2-level hierarchical Bayesian model:
  - Few-parameter GRB luminosity functions (top level)
  - Latent flux & direction parameters for 279 or 463 GRBs—*conditionally independent*
- *Member marginalization:* 1-D or 3-D cubatures for latents; adaptive cubature for top-level



Inevitability of Monte Carlo

The most mature non-MCMC tools cannot handle models with $\gtrsim$5 to 10 dependent parameters

Adaptive importance sampling, sequential Monte Carlo, and nested sampling may push this to dozens of parameters, but presently are complex and not fully understood (and often have an MCMC component)

MCMC is currently the "only game in town" for large problems, and can work with $\sim 10^6$ parameters (e.g., images)

Algorithms are deceptively simple; it is not trivial to *get it right*

Agenda

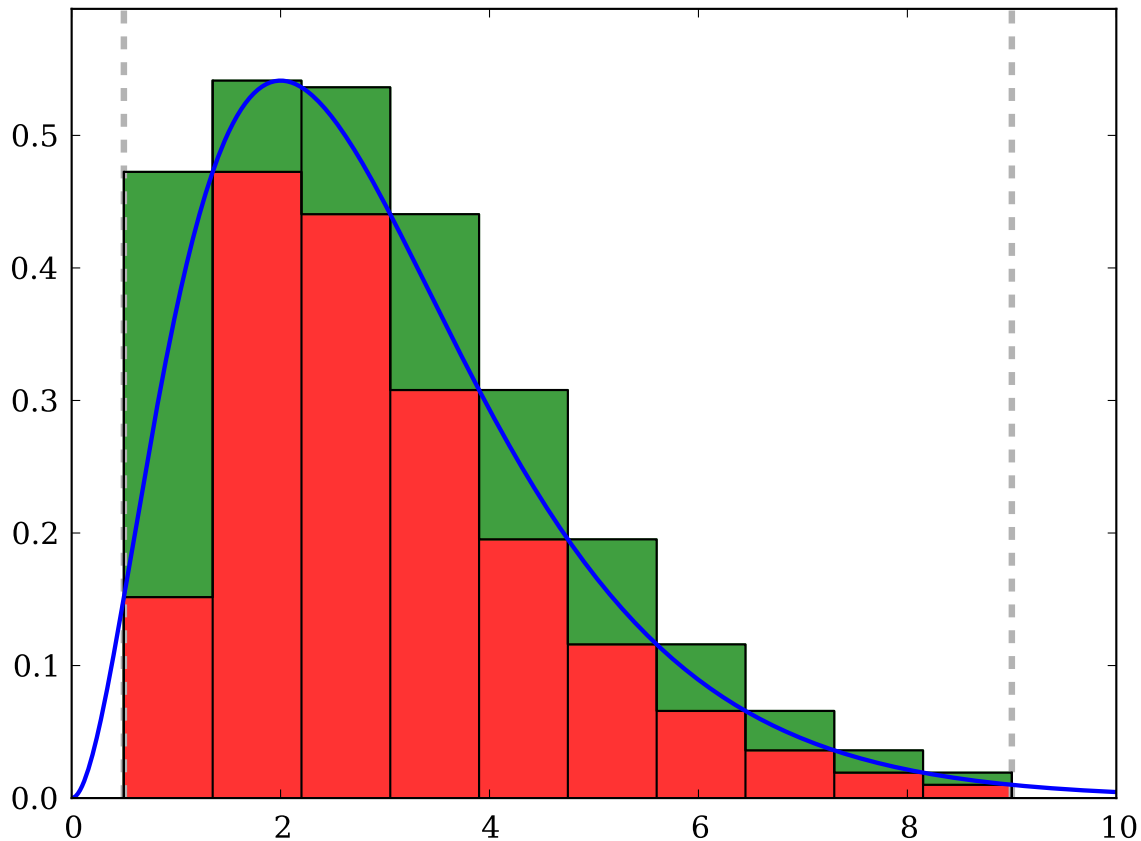
1. **Marginal likelihood computation**
   - Cubature
   - Randomized cubature — Nested sampling
   - Posterior density estimation
   - Posterior expectations
   - Importance sampling
2. **Bayes factors via trans-dimensional MCMC**
   - Reversible-jump MCMC
   - Birth-death MCMC
3. **Guidance**

Lebesgue integration and nested sampling

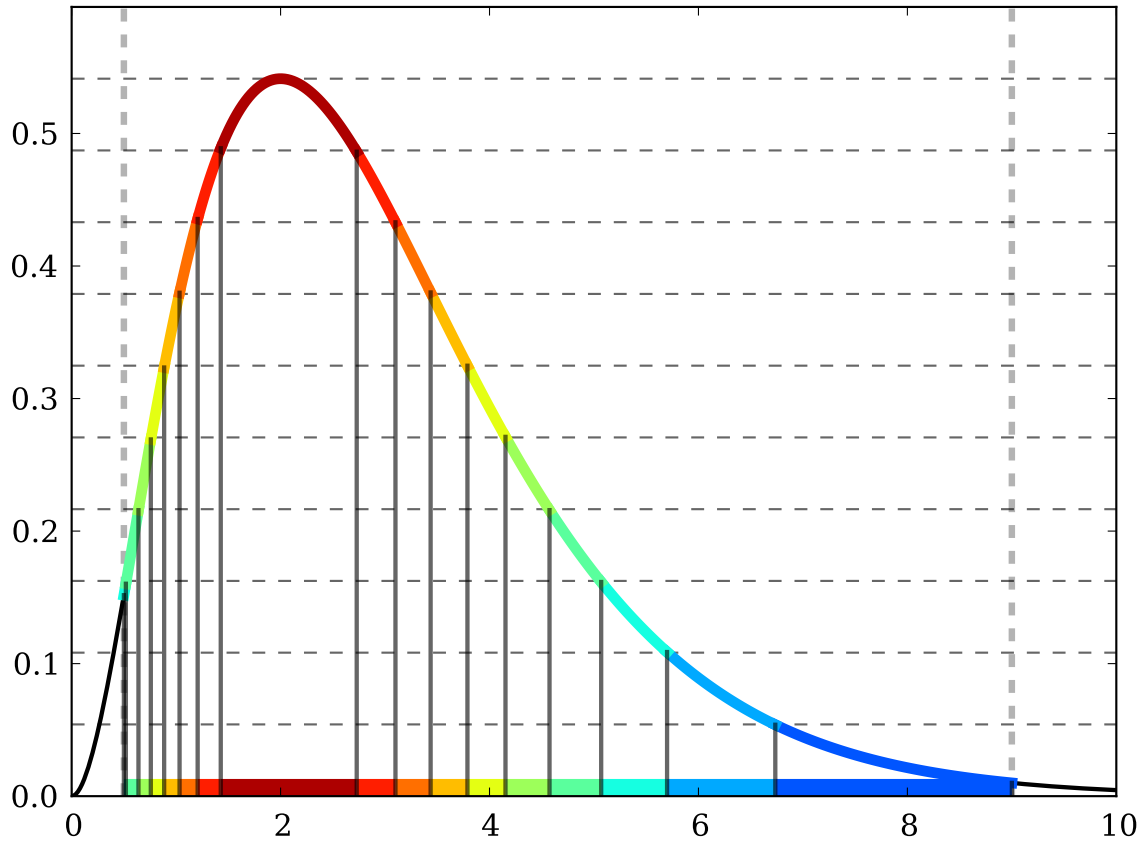
Adaptive quadrature approximates a *Riemann integral* in d-D

But there are other ways to define an integral!

Riemann integral: Partition abscissa



Lebesgue integral: Partition *ordinate*



$$Z_L \approx \sum_i f_i \mu_i(\{x : f \approx f_i\});$$

$$\mu_i(\{x : f \approx f_i\}) = \text{“measure” of } x \text{ in } f_i \text{ bin}$$

Two dimensions

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

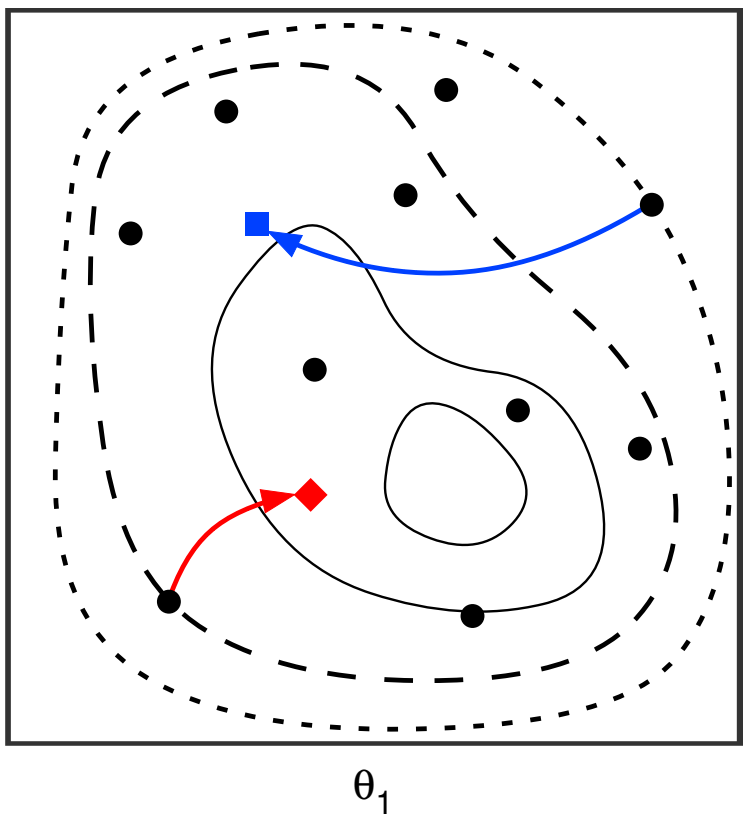
$$Z_L \approx \sum_i f_i \mu_i(\{x, y : f \approx f_i\})$$

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

Agenda

1. **Marginal likelihood computation**
   - Cubature
   - Randomized cubature — Nested sampling
   - Posterior density estimation
   - Posterior expectations
   - 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 *bad* idea is based on "candidate's formula," aka "basic marginal likelihood identity" (BMI):

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

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

Implementation:

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

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: $\hat{Z} = \frac{\pi(\theta^*)\mathcal{L}(\theta^*)}{\hat{p}(\theta^*|D,M)}$

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

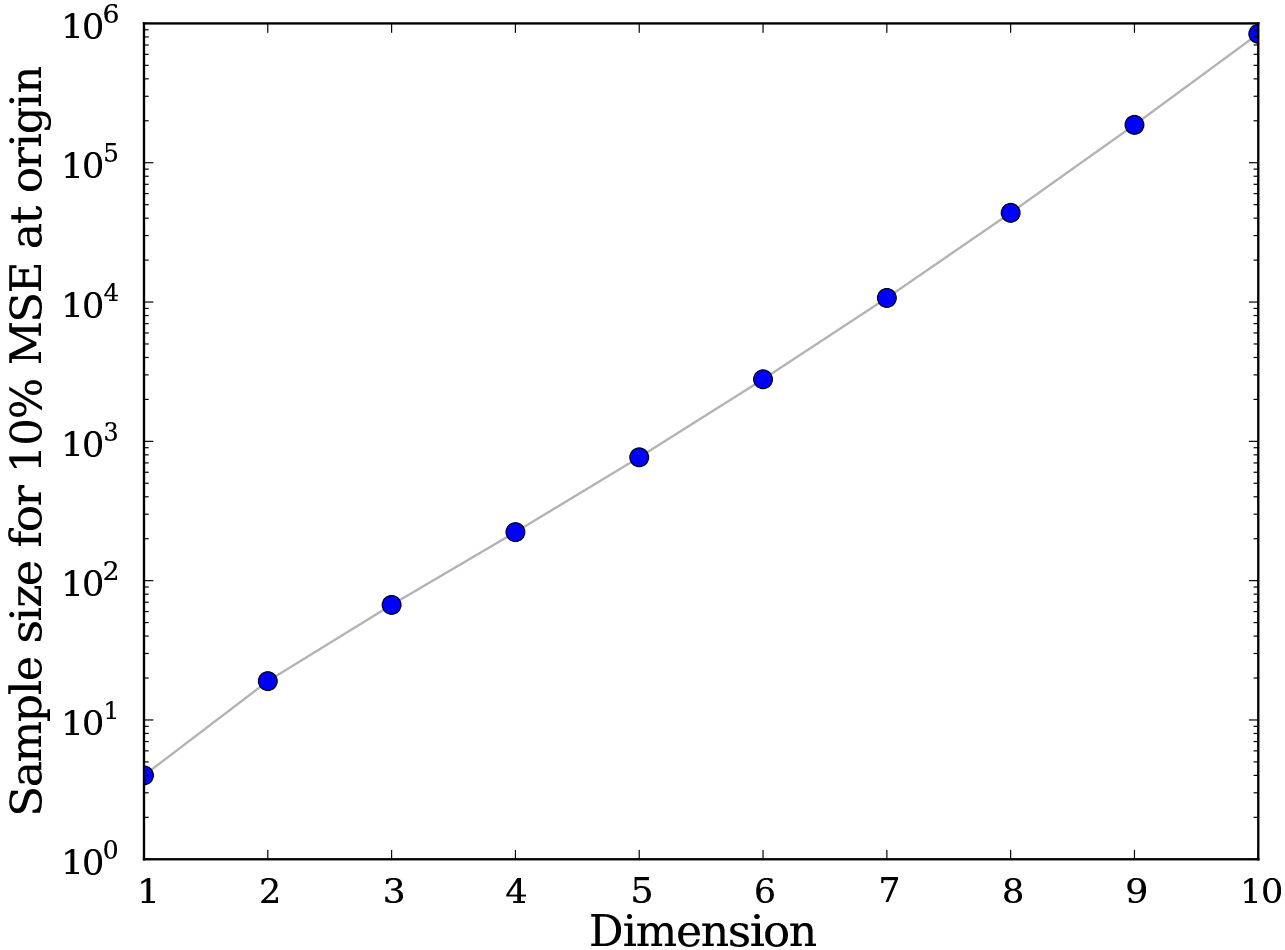
(But see Hsiao, Huang & Chang 2004 for an attempt to fix it)

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.



Silverman (1986)

Chib's method

For marginal likelihoods from *Gibbs sampler* output:

E.g., consider a *missing data* or *latent variable* problem with interesting parameters θ, and missing/latent parameters ψ:

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

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)

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

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

Savage-Dickey density ratio

For model comparison with *nested models:*

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

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

Let $\phi_0 =$ value of ϕ assumed by M_1:

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

Assume priors are independent:

$$\begin{aligned} p(\theta|M_1) &= f(\theta) \\ p(\theta, \phi|M_2) &= f(\theta)\, g(\phi) \end{aligned}$$

(may be relaxed)

Compare models via marginal likelihoods:

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

$$\mathcal{L}(M_2) = \int d\theta d\phi \; f(\theta) \, g(\phi) \, \mathcal{L}_2(\theta, \phi)$$

Due to nesting, integrals appear similar! Recall marginal for ϕ:

$$p(\phi|D, M_2) = \frac{1}{\mathcal{L}(M_2)} \int d\theta \; f(\theta) g(\phi) \mathcal{L}_2(\theta, \phi)$$

Now calculate $\mathcal{L}(M_1)$, using this marginal with $\phi = \phi_0$:

$$\begin{aligned} \mathcal{L}(M_1) &= \int d\theta \; f(\theta) g(\phi_0) \mathcal{L}_2(\theta, \phi_0) \times \frac{1}{g(\phi_0)} \\ &= \frac{p(\phi_0|D, M_2)}{p(\phi_0|M_2)} \mathcal{L}(M_2) \\ \rightarrow B_{21} &= \frac{p(\phi_0|M_2)}{p(\phi_0|D, M_2)} \quad \sim \quad \frac{\text{small for broad } \phi \text{ prior}}{\text{small if } \phi_0 \text{ far from } \hat{\phi}} \end{aligned}$$

Can approximate this via MCMC with *only* M_2, as long as ϕ_0 isn't too far in tail and ϕ is low-dimensional (1 or 2!)

Agenda

1. **Marginal likelihood computation**
   - Cubature
   - Randomized cubature — Nested sampling
   - Posterior density estimation
   - Posterior expectations
   - 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:

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

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

$$\int \mathrm{d}\theta \, \frac{\pi(\theta)}{Z} = \frac{1}{Z} = \int d\theta \, \frac{p(\theta|D,M)}{\mathcal{L}(\theta)}$$

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

$$\widehat{Z}_{HM} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\mathcal{L}(\theta_i)}$$

Appealingly simple, but. . .

"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."

$$\widehat{Z}_{HM} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\mathcal{L}(\theta_i)}$$

Qualitative explanation

Recall that $Z \approx \mathcal{L}(\hat{\theta}) \frac{\delta\theta}{\Delta\theta}$ (posterior width / prior width)

$\widehat{Z}_{HM}$ 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): $\widehat{Z}_{HM}$ 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:

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

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

$$\frac{1}{Z} \int_{\Omega} d\theta \, \pi(\theta) \quad = \quad \int_{\Omega} d\theta \, \frac{p(\theta|D)}{\mathcal{L}(\theta)}$$

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

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

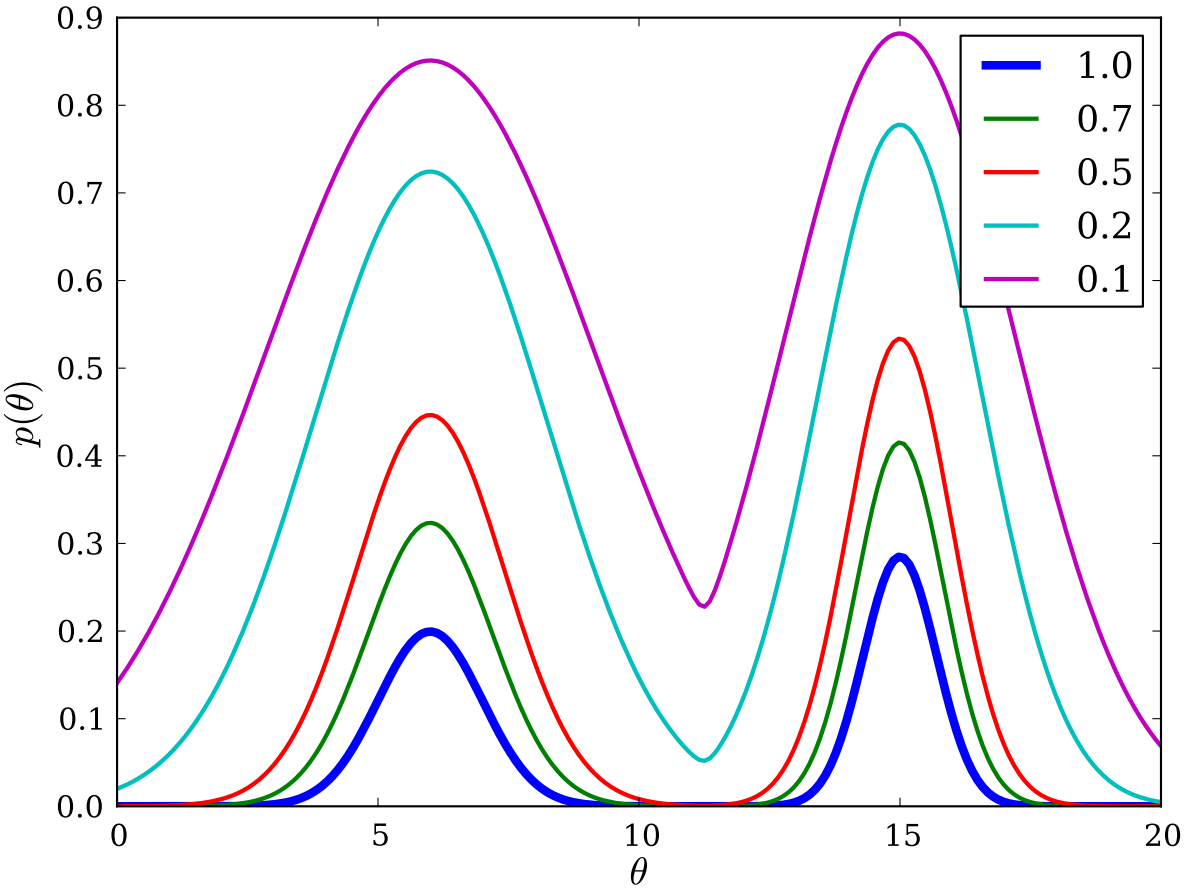
Thermodynamic integration

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

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

for a sequence of increasing "tempers" ("inverse temperatures") β_i with $\beta_0 = 0$, $\beta_N = 1$

Tempered likelihood functions



Estimating the ratios

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

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

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

$$\begin{aligned} \text{where } R(\beta) &\equiv \frac{Z'(\beta)}{Z(\beta)} = \frac{1}{Z(\beta)} \frac{d}{d\beta} \int d\theta \, \pi(\theta) \exp\left[\beta \log \mathcal{L}(\theta)\right] \\ &= \frac{1}{Z(\beta)} \int d\theta \, \log[\mathcal{L}(\theta)] \pi(\theta) \mathcal{L}^{\beta}(\theta) \end{aligned}$$

So $R(\beta) = \langle \log \mathcal{L} \rangle_\beta$, the (annealed) posterior expectation of the log-likelihood

Thermodynamic integration algorithm

$$\ln Z = \int d\beta \, \langle \log \mathcal{L} \rangle_\beta$$

- For each of an "annealing schedule" of β_i values:
  - Use MCMC to get $\{\theta_j\}$ with $\theta \sim \pi(\theta)\mathcal{L}^\beta(\theta)$
  - Find $\hat{R}(\beta) = \frac{1}{N} \sum_j \log \mathcal{L}(\theta_j)$
- 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 β range near $\beta = 1 \rightarrow$ need more tempers than for param estimation

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

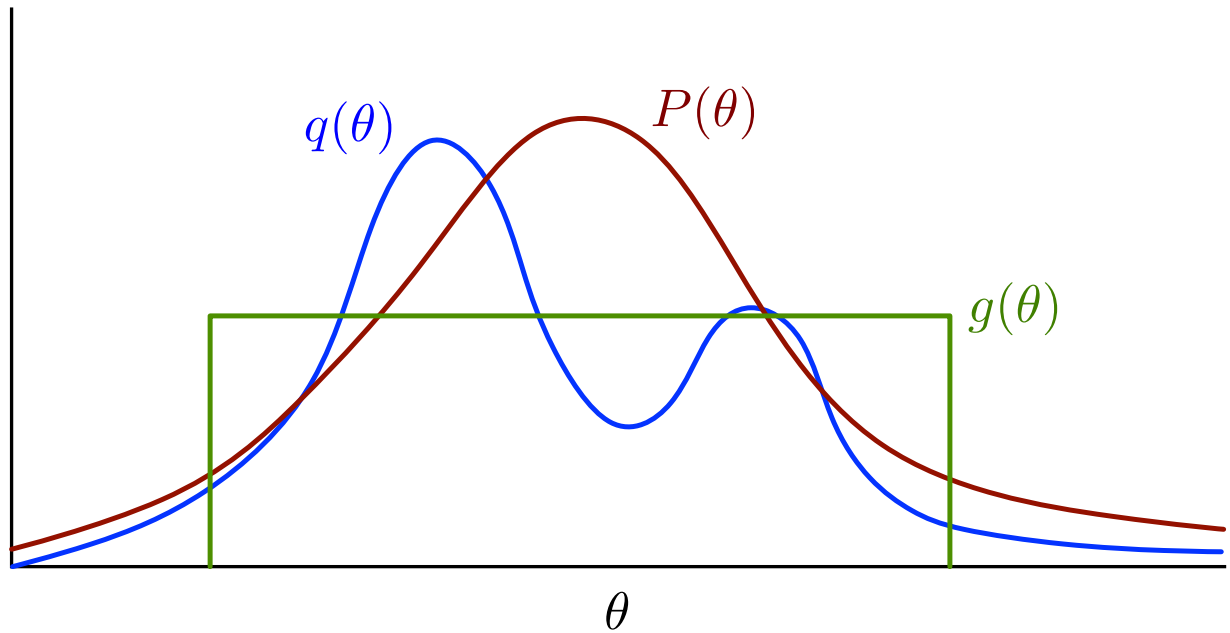
Agenda

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

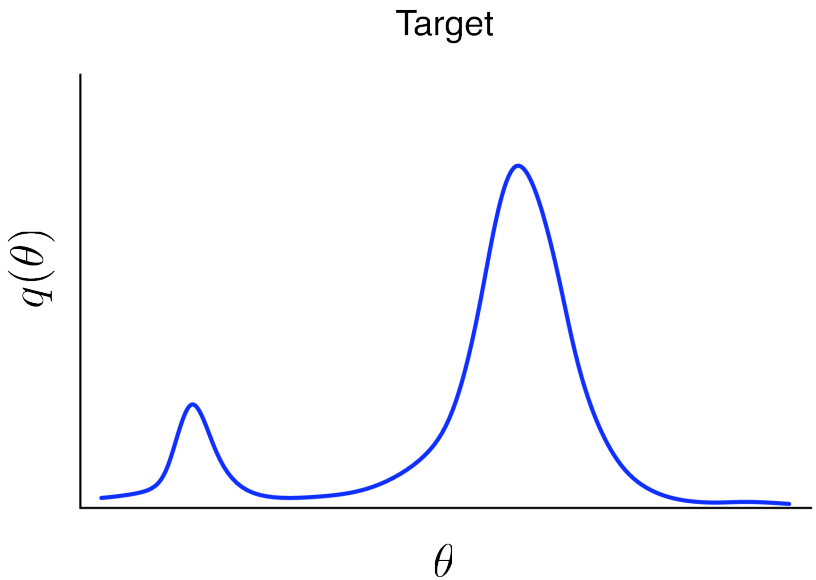
Building a Good Importance Sampler

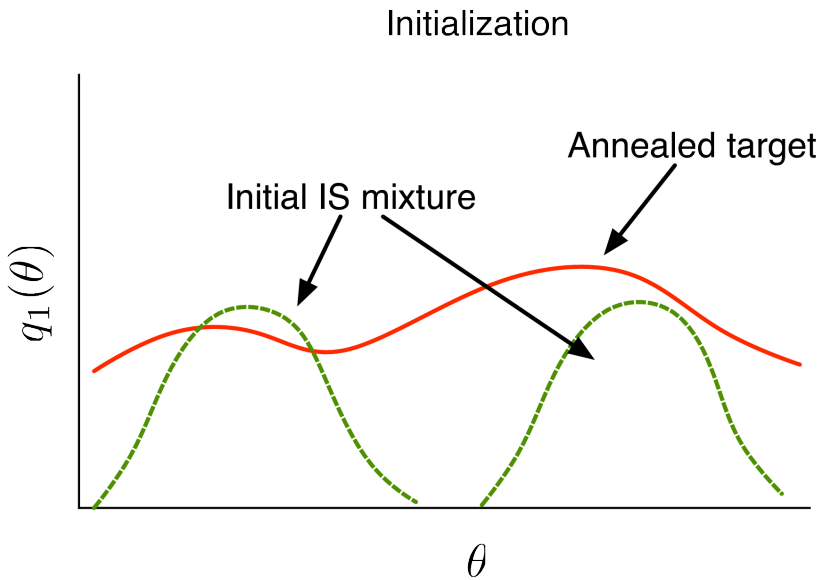
Estimate an annealing target density, π_n, using a mixture of multivariate Student-t distributions, P_n:

$$q_n(\theta) = [q_0(\theta)]^{1-\lambda_n} \times [q(\theta)]^{\lambda_n}, \qquad \lambda_n = 0 \ldots 1$$
$$P_n(\theta) = \sum_j \mathrm{MVT}(\theta; \mu_j^n, S_j^n, \nu)$$

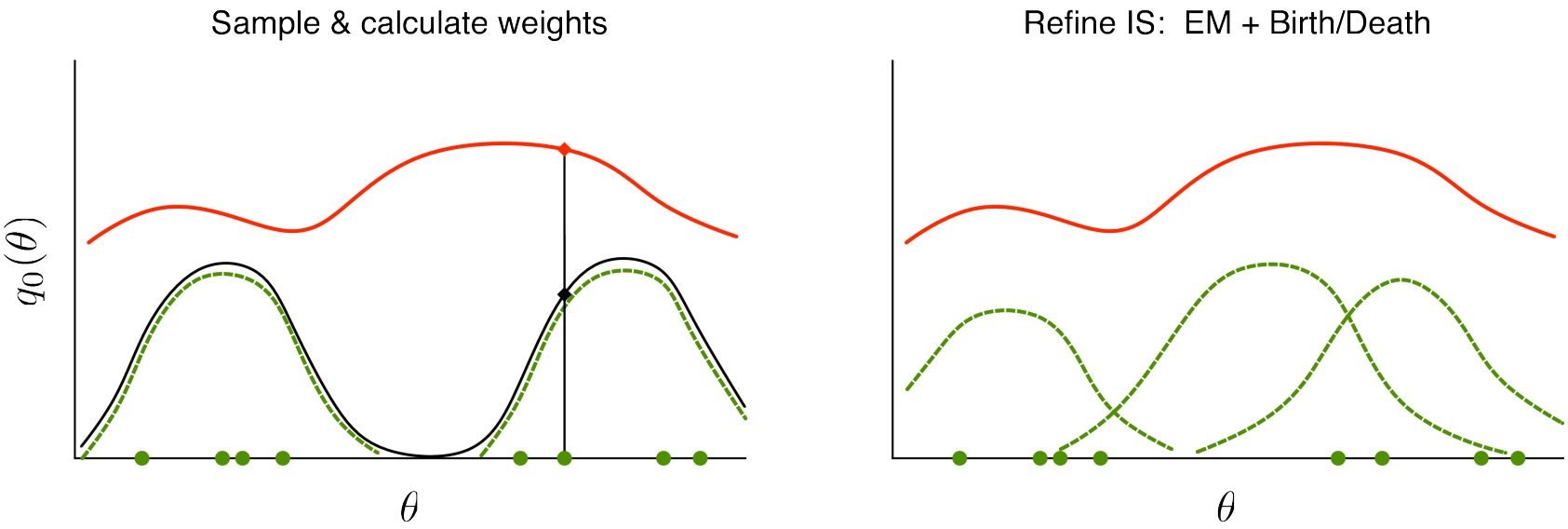
Adapt the mixture to the target using ideas from *sequential Monte Carlo* $\rightarrow$ *Adaptive annealed importance sampling (AAIS)*

Initialization

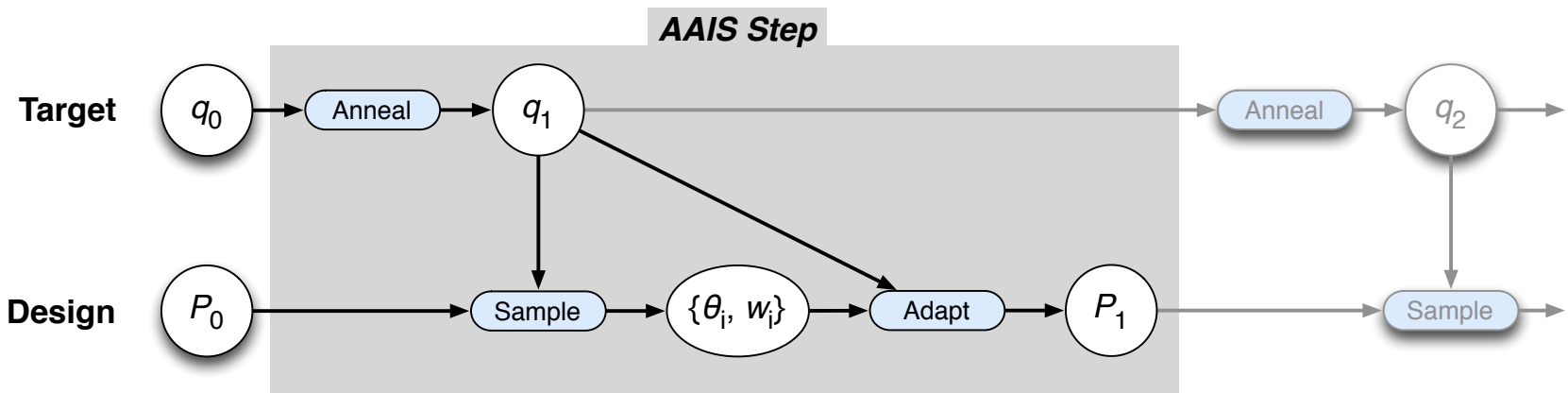




Sample, weight, refine

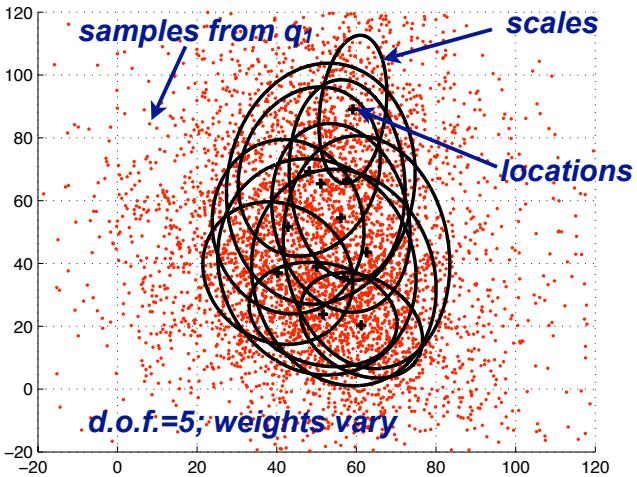


Overall algorithm

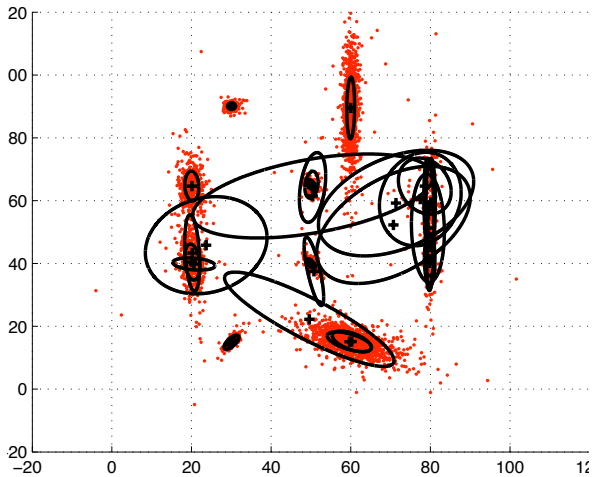


2-D Example:
Many well-separated correlated normals

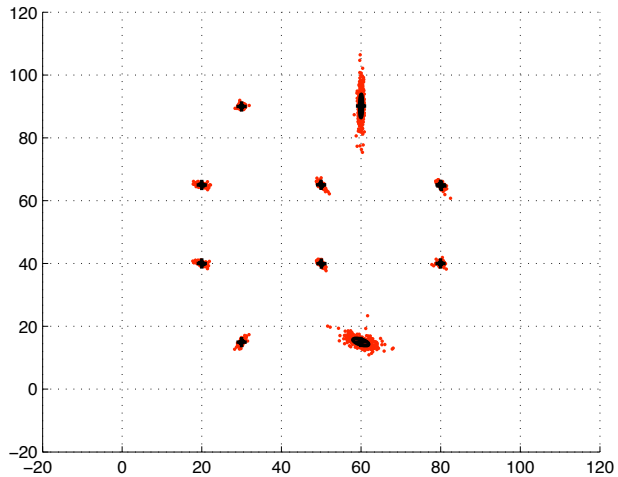
$\lambda_1 = 0.01$



$\lambda_3 = 0.11$



$\lambda_8 = 1$



Agenda

1. Marginal likelihood computation
   - Cubature
   - Randomized cubature — Nested sampling
   - Posterior density estimation
   - Posterior expectations
   - 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, θ_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)

Agenda

1. **Marginal likelihood computation**
   - Cubature
   - Randomized cubature — Nested sampling
   - Posterior density estimation
   - Posterior expectations
   - 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, $M_1 : \theta$ $M_2 : \theta_1, \theta_2$

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

$$\theta_1 = \theta + u; \qquad \theta_2 = \theta - u$$

Adjust the usual MH α by factors accounting for the move probabilities, the dist'n for u, and the Jacobian

$$|\partial(\theta, u)/\partial(\theta_1, \theta_2)|$$

Agenda

1. **Marginal likelihood computation**
   - Cubature
   - Randomized cubature — Nested sampling
   - Posterior density estimation
   - Posterior expectations
   - Importance sampling
2. **Bayes factors via trans-dimensional MCMC**
   - Reversible-jump MCMC
   - Birth-death MCMC
3. **Guidance**

Birth-death MCMC

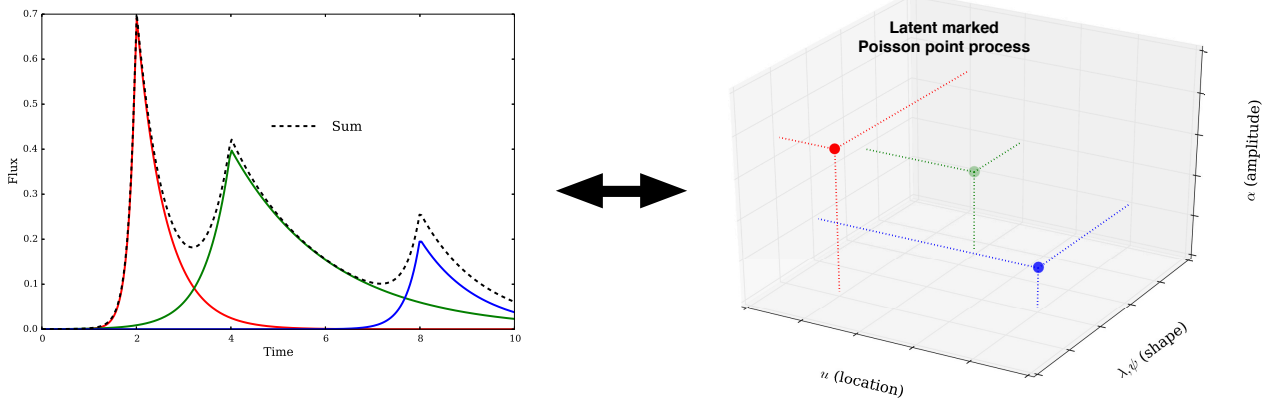
Setting

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

Agenda

1. **Marginal likelihood computation**
   - Cubature
   - Randomized cubature — Nested sampling
   - Posterior density estimation
   - Posterior expectations
   - 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 ($\lesssim$15): Mixture-based importance sampling guided by MCMC output is often easiest to implement with good accuracy [also try cubature]
- Explore robustness to priors!

SAMSI '03 tech report, "Marginal Likelihood Computation: A Comparative Study" (Rui Paulo 2003)
SCMA '06 review paper, "Current Challenges in Bayesian Model Choice" (Clyde$^+$ 2007)