Bayesian Computation: Overview and Methods for Low-Dimensional Models

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

IAC Winter School, 3–4 Nov 2014

Computation overview, low-dimensional models

1. **Bayesian integrals**
2. **Large N: Laplace approximations**
3. **Cubature**
4. **Monte Carlo integration**
   - Posterior sampling
   - Importance sampling

Computation overview, low-dimensional models

1. **Bayesian integrals**
2. **Large N: Laplace approximations**
3. **Cubature**
4. **Monte Carlo integration**
   - Posterior sampling
   - Importance sampling

Notation

$$
\begin{aligned}
p(\theta|D, M) &= \frac{p(\theta|M)p(D|\theta, M)}{p(D|M)} \\
&= \frac{\pi(\theta)\mathcal{L}(\theta)}{Z} = \frac{q(\theta)}{Z}
\end{aligned}
$$

- M = model specification
- D specifies observed data
- θ = model parameters
- $\pi(\theta)$ = prior pdf for θ
- $\mathcal{L}(\theta)$ = likelihood for θ (likelihood function)
- $q(\theta) = \pi(\theta)\mathcal{L}(\theta) =$ "quasiposterior"
- $Z = p(D|M) =$ (marginal) likelihood for the model

Marginal likelihood:

$$Z = \int d\theta \; \pi(\theta)\mathcal{L}(\theta) = \int d\theta \; q(\theta)$$

Use "Skilling conditional" for common conditioning info:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)} \qquad || \; M$$

Suppress such conditions when clear from context

Parameter space integrals

For model with m parameters, we need to evaluate integrals like:

$$\int d^m\theta\, g(\theta)\, \pi(\theta)\, \mathcal{L}(\theta) \;=\; \int d^m\theta\, g(\theta)\, \overbrace{q(\theta)}^{\pi(\theta)\,\mathcal{L}(\theta)}$$

- $g(\theta) = 1 \rightarrow p(D|M)$ (norm. const., model likelihood)
- $g(\theta) = \theta \rightarrow$ posterior mean for θ
- $g(\theta) =$ 'box' $\rightarrow$ probability $\theta \in$ credible region
- $g(\theta) = 1$, integrate over subspace $\rightarrow$ marginal posterior
- $g(\theta) = \delta[\psi - \psi(\theta)] \rightarrow$ propagate uncertainty to $\psi(\theta)$

The Bayesian computation challenge

Asymptotic approximations

- Most probability is usually in regions near the mode
- Taylor expansion of $\log p \rightarrow$ leading order is quadratic
- Integrand may be well-approximated by a multivariate (correlated) normal: the *Laplace approximation*

Requires ingredients familiar from frequentist calculations

Bayesian calculation is *not significantly harder* than frequentist calculation in this limit.

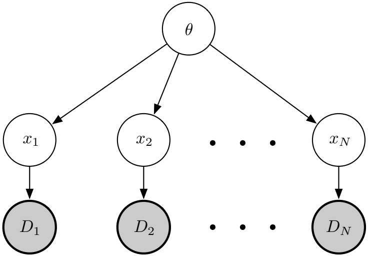
Inference with independent data

Analytical: For exponential family models & conjugate priors, integrals are often tractable and simpler than frequentist counterparts (e.g., normal credible regions, Student's t)

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

Inference with conditionally independent parameters

In multilevel (hierarchical) models—e.g., for "measurement error" and latent variable problems—a layer of variables may be independent given higher level variables → numerically tractable marginals



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

$$
\text{so} \quad \mathcal{L}_m(\theta) = \prod_i \int dx_i \; \ell_i(x_i) f(x_i|\theta)
$$

Bayesian Computation Menu

Large sample size, N: Laplace approximation

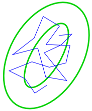
- Approximate posterior as multivariate normal $\rightarrow$ det(covar) factors
- Uses ingredients available in χ^2/ML fitting software (MLE, Hessian)
- Often accurate to $O(1/N)$ (better than $O(1/\sqrt{N})$)

Modest-dimensional models ($m \lesssim 10$ to 20)

- Adaptive cubature
- Monte Carlo integration (importance & stratified sampling, adaptive importance sampling, quasirandom MC)

High-dimensional models ($m \gtrsim 5$)

- Posterior sampling — create RNG that samples posterior
- Markov Chain Monte Carlo (MCMC) is the most general framework



Computation overview, low-dimensional models

1. Bayesian integrals
2. Large N: Laplace approximations
3. Cubature
4. Monte Carlo integration
   - Posterior sampling
   - Importance sampling

Laplace Approximations

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

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

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

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

Marginal likelihoods

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

Marginal posterior densities

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

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

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

Posterior expectations

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

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 → approximation is often $O(1/N)$ or better

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

*Some optimizers provide approximate Hessians, e.g., Levenberg-Marquardt for modeling data with additive Gaussian noise. For more general cases, see Kass (1987) "Computing observed information by finite differences" (beware typos): central 2nd differencing + Richardson extrapolation.

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 banana-shaped contours)

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

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

Relationship to BIC

Laplace approximation for marginal likelihood:

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

$$\sim \pi(\hat{\theta})\mathcal{L}(\hat{\theta})\, (2\pi)^{m/2} \prod_{k=1}^{m} \sigma_{\theta_k}$$

We expect asymptotically $\sigma_{\theta_k} \propto 1/\sqrt{N}$

Bayesian Information Criterion (BIC; aka Schwarz criterion):

$$-\frac{1}{2}\text{BIC} = \ln \mathcal{L}(\hat{\theta}) - \frac{m}{2} \ln N$$

This is a *very* crude approximation to $\ln Z$; it captures the asymptotic N dependence, but omits factors $O(1)$. Can justify in some i.i.d. settings using "unit info prior."

BIC $\sim$ Bayesian counterpart to adjusting χ^2 for d.o.f., but partly accounts for parameter space volume ($\rightarrow$ consistent model choice, unlike fixed-α hyp. tests)

Can be useful for identifying cases where an an accurate but hard Z calculation is useful (esp. for nested models, where some missing factors cancel)

Computation overview, low-dimensional models

1. Bayesian integrals

2. Large N: Laplace approximations

3. **Cubature**

4. Monte Carlo integration
   - Posterior sampling
   - Importance sampling

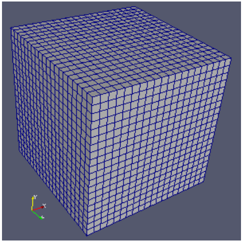
Modest-D: Quadrature & Cubature

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

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

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



Wikipedia

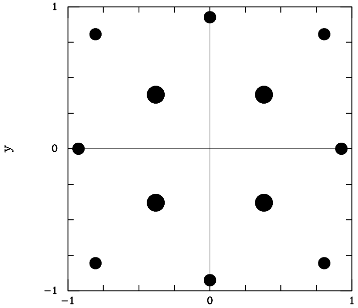
Monomial Cubature Rules

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

$$f(x,y,z) = \mathrm{MVN}(x,y,z)\sum_{ijk} a_{ijk}x^i y^j z^k \qquad \text{for } i+j+k \leq 7$$

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`, `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.

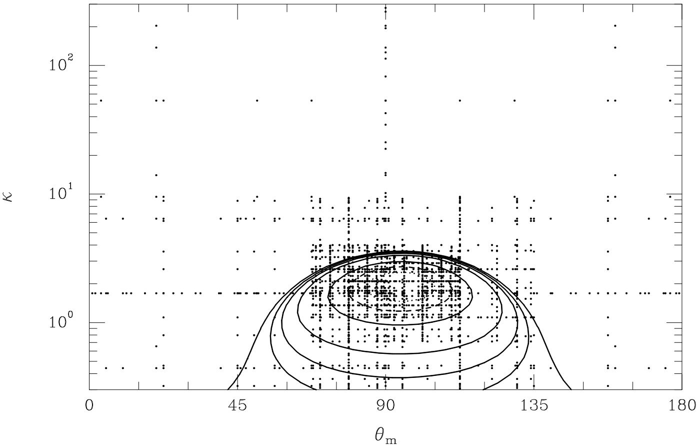
These provide diagnostics (error estimates or measures of reparameterization quality).

nodes used by ADAPT's 7th order rule

$$2^d + 2d^2 + 2d + 1$$

Dimen	2	3	4	5	6	7	8	9	10
# nodes	17	33	57	93	149	241	401	693	1245

Analysis of Galaxy Polarizations



TJL, Flanagan, Wasserman (1997)

Computation overview, low-dimensional models

1. **Bayesian integrals**

2. **Large N: Laplace approximations**

3. **Cubature**

4. **Monte Carlo integration**
   - Posterior sampling
   - Importance sampling

Monte Carlo Integration

$\int g \times p$ is just the *expectation of g*; suggests approximating with a *sample average*:

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

This is like a cubature rule, with *equal weights* and *random nodes*

Ignores smoothness $\rightarrow$ poor performance in 1-D, 2-D

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

Why/when it works

- Independent sampling & law of large numbers $\rightarrow$ asymptotic convergence in probability
- Error term is from CLT; requires finite variance

Practical problems

- $p(\theta)$ must be a density we can draw IID samples from—perhaps the prior, but...
- $O(n^{-1/2})$ multiplier (std. dev'n of g) may be large

$\rightarrow$ *IID* Monte Carlo can be hard if dimension* $\gtrsim$ *5–10*

*IID = independently, identically distributed

Posterior sampling

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

When $p(\theta)$ is a posterior distribution, drawing samples from it is called *posterior sampling*:

- *One set of samples* can be used for many different calculations (so long as they don't depend on low-probability events)
- This is the most promising and general approach for Bayesian computation in *high dimensions*—though with a twist (MCMC!)

Challenge: How to build a RNG that samples from a posterior?

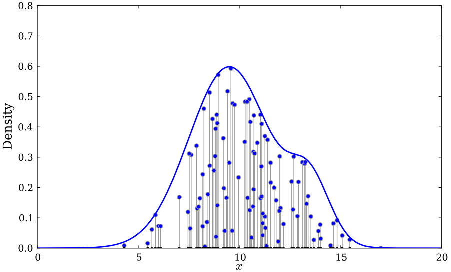
Accept-Reject Algorithm

Goal: Given $q(\theta) \equiv \pi(\theta)\mathcal{L}(\theta)$, build a RNG that draws samples from the probability density function (pdf)

$$f(\theta) = \frac{q(\theta)}{Z} \quad \text{with} \quad Z = \int d\theta\, q(\theta)$$

The probability for a region under the pdf is the *area (volume) under the curve (surface)*.

$\rightarrow$ Sample points uniformly in volume under q; their θ values will be draws from $f(\theta)$.



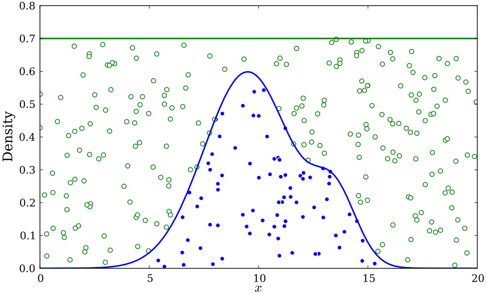
The fraction of samples with θ ("x" in the fig) in a bin of size $\delta\theta$ is the fractional area of the bin.

How can we generate points uniformly under the pdf?

Suppose $q(\theta)$ has compact support: it is nonzero over a finite contiguous region of θ-space of length/area/volume V.

Generate *candidate* points uniformly in a rectangle enclosing $q(\theta)$.

Keep the points that end up under q.



Basic accept-reject algorithm

1. Find an upper bound Q for $q(\theta)$
2. Draw a candidate parameter value θ' from the uniform distribution in V
3. Draw a uniform random number, u
4. If the ordinate $uQ < q(\theta')$, record θ' as a sample
5. Goto 2, repeating as necessary to get the desired number of samples.

Efficiency = ratio of areas (volumes), $Z/(QV)$.

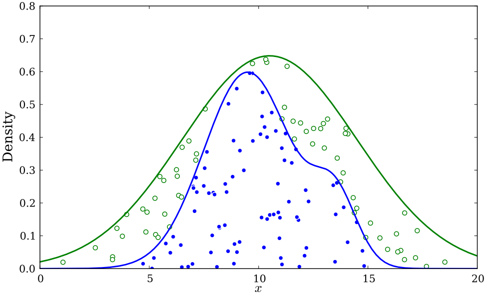
Two issues

- Increasing efficiency
- Handling distributions with infinite support

Envelope Functions

Suppose there is a pdf $h(\theta)$ that we know how to sample from and that roughly resembles $q(\theta)$:

- Multiply h by a constant C so $Ch(\theta) \geq q(\theta)$
- Points with coordinates $\theta' \sim h$ and ordinate $uCh(\theta')$ will be distributed uniformly under $Ch(\theta)$
- Replace the hyperrectangle in the basic algorithm with the region under $Ch(\theta)$



Accept-Reject Algorithm

1. Choose a tractable density $h(\theta)$ and a constant C so Ch bounds q
2. Draw a candidate parameter value $\theta' \sim h$
3. Draw a uniform random number, u
4. If $q(\theta') < Ch(\theta')$, record θ' as a sample
5. Goto 2, repeating as necessary to get the desired number of samples.

Efficiency = ratio of volumes, Z/C.

In problems of realistic complexity, the efficiency is intolerably low for parameter spaces of more than several dimensions.

Take-away idea: *Propose candidates that may be accepted or rejected*

Markov Chain Monte Carlo

Accept/Reject aims to produce *independent* samples—each new θ is chosen irrespective of previous draws.

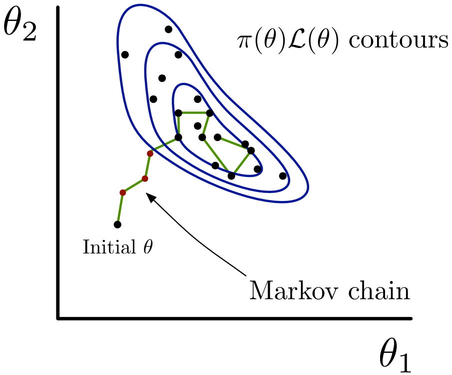
To enable exploration of complex pdfs, let's introduce *dependence*: Choose new θ points in a way that

- Tends to *move toward* regions with higher probability than current
- Tends to *avoid* lower probability regions

The simplest possibility is a *Markov chain*:

$$
\begin{aligned}
&p(\text{next location}|\textit{current}\ \textbf{and}\ \textit{previous locations}) \\
&\qquad = p(\text{next location}|\textit{current location})
\end{aligned}
$$

A Markov chain "has no memory."

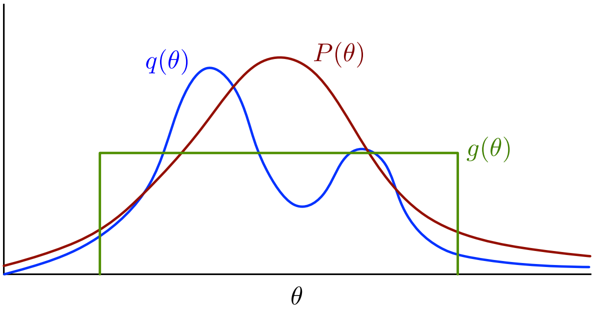


Covered later!

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.

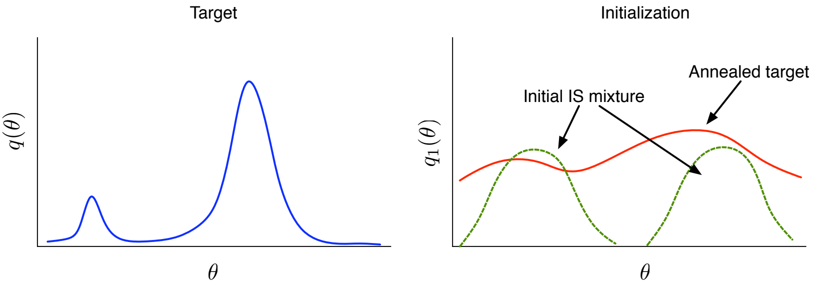
Building a Good Importance Sampler

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

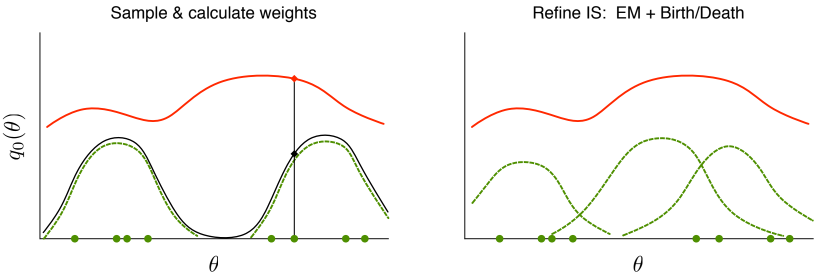
$$q_n(\theta) = [q_0(\theta)]^{1-\lambda_n} \times [q(\theta)]^{\lambda_n}, \qquad \lambda_n = 0 \ldots 1$$
$$P_n(\theta) = \sum_j \mathrm{MVT}(\theta; \mu_j^n, S_j^n, \nu)$$

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

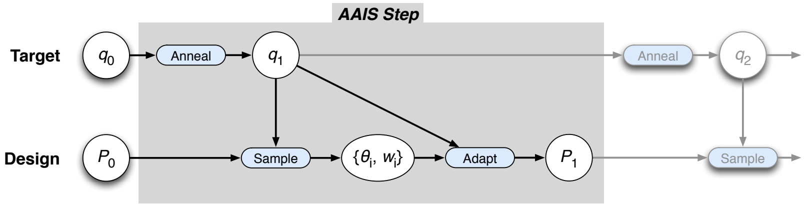
Initialization



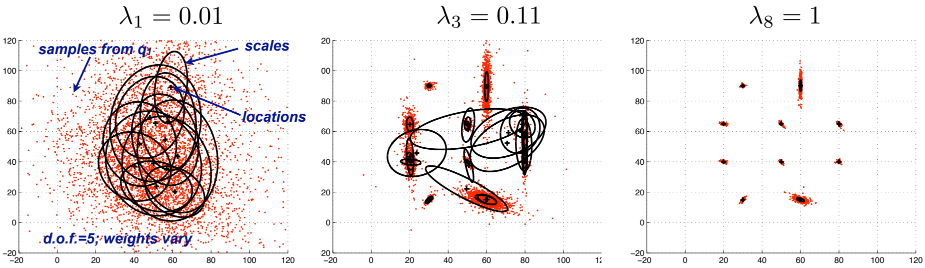
Sample, weight, refine



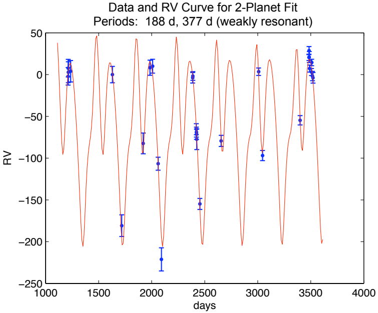
Overall algorithm



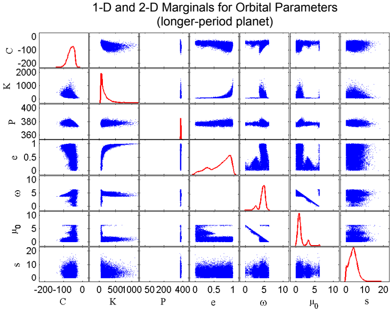
2-D Example:
Many well-separated correlated normals



Observed Data: HD 73526 (2 planets)



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



Sampling efficiency of final mixture ESS/$N \approx 65\%$

See Liu (2014): "Adaptive Annealed Importance Sampling"