

Bayesian Computation: MCMC and All That

SCMA V Short-Course, Pennsylvania State University
Alan Heavens, Tom Loredo, and David A van Dyk
11 and 12 June 2011

11 June Morning Session 10.00 - 13.00

| | | |
|---------------|-------------------------------------------------------|---------|
| 10.00 - 11.15 | The Basics: Bayes Theorem, Priors, and Posteriors. | Heavens |
| 11.15 - 11.30 | Coffee Break | |
| 11.30 - 13.00 | Low Dimensional Computing Part I | Loredo |

11 June Afternoon Session 14.15 - 17.30

| | | |
|---------------|-----------------------------------|---------|
| 14.15 - 15.15 | Low Dimensional Computing Part II | Loredo |
| 15.15 - 16.00 | MCMC Part I | van Dyk |
| 16.00 - 16.15 | Coffee Break | |
| 16.15 - 17.00 | MCMC Part II | van Dyk |
| 17.00 - 17.30 | R Tutorial | van Dyk |

12 June Morning Session 10.00 - 13.15

| | | |
|---------------|-------------------------------------|--------------------|
| 10.00 - 10.45 | Data Augmentation and PyBLoCXS Demo | van Dyk |
| 10.45 - 11.45 | MCMC Lab | van Dyk |
| 11.45 - 12.00 | Coffee Break | |
| 12.00 - 13.15 | Output Analysis* | Heavens and Loredo |

12 June Afternoon Session 14.30 - 17.30

| | | |
|---------------|------------------------------------|------------------------------|
| 14.30 - 15.25 | MCMC and Output Analysis Lab | Heavens |
| 15.25 - 16.05 | Hamiltonian Monte Carlo | Heavens |
| 16.05 - 16.20 | Coffee Break | |
| 16.20 - 17.00 | Overview of Other Advanced Methods | Loredo |
| 17.00 - 17.30 | Final Q&A | Heavens, Loredo, and van Dyk |

* Heavens (45 min) will cover basic convergence issues and methods, such as G&R's multiple chains.
Loredo (30 min) will discuss Resampling

The Bayesics



Alan Heavens

University of Edinburgh, UK

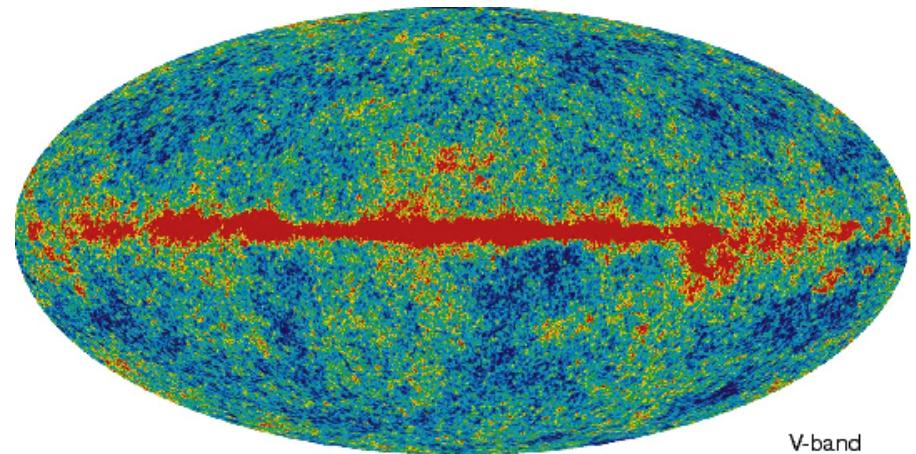
Lectures given at SCMA V, Penn State

June 2011

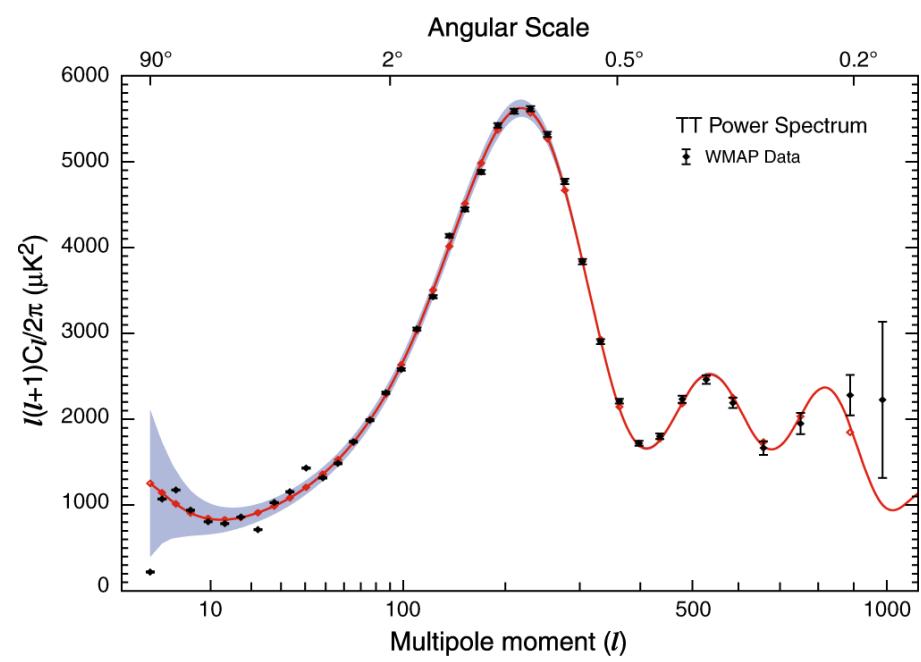
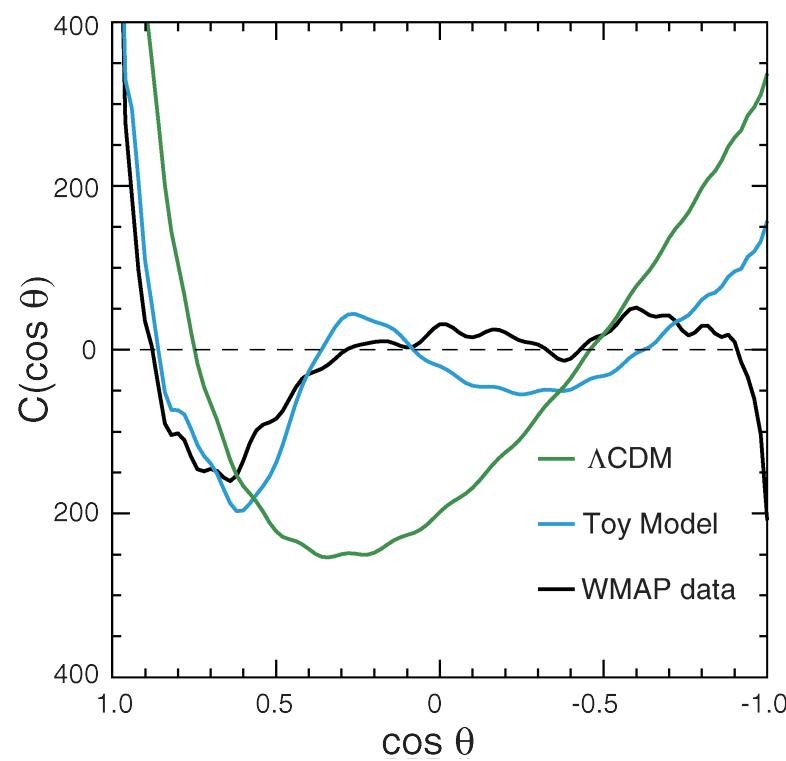
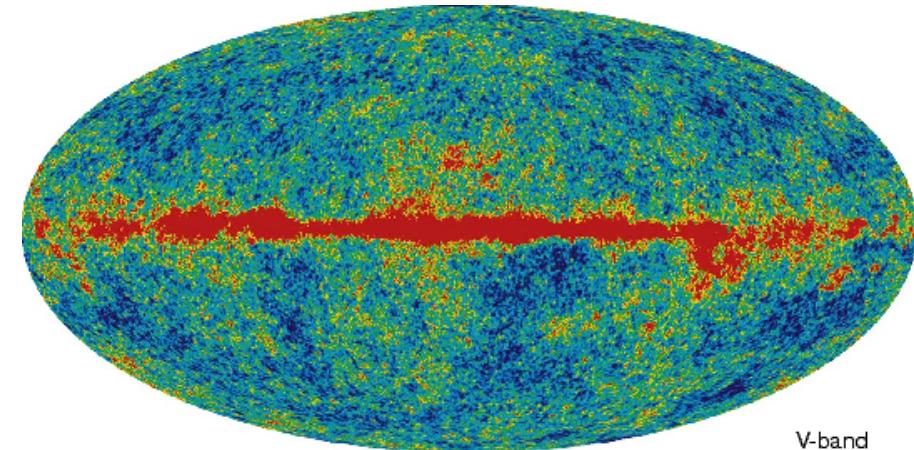


Outline

- Types of problem
- Bayes' theorem
- Parameter Estimation
 - Marginalisation
 - Errors
- Error prediction and experimental design:
Fisher Matrices
- Model Selection



LCDM fits the WMAP data well.



Inverse problems

- Most cosmological problems are *inverse problems*, where you have a set of data, and you want to infer something.
- Examples
 - Hypothesis testing
 - Parameter estimation
 - Model selection

Examples

- Hypothesis testing
 - Is the CMB radiation consistent with (initially) gaussian fluctuations?
- Parameter estimation
 - In the Big Bang model, what is the value of the matter density parameter?
- Model selection
 - Do cosmological data favour the Big Bang theory or the Steady State theory?
 - Is the gravity law General Relativity or higher-dimensional?

What is probability?

- Frequentist view: p describes the relative *frequency of outcomes* in infinitely long trials
- Bayesian view: p expresses our *degree of belief*
- Bayesian view is closer to what we seem to want from experiments: e.g. given the WMAP data, what is the probability that the density parameter of the Universe is between 0.9 and 1.1?
- Cosmology is in good shape for inference because we have decent model(s) with parameters – well-posed problem

Bayes' Theorem

- Rules of probability:
- $p(x) + p(\text{not } x) = 1$ sum rule
- $p(x,y) = p(x|y)p(y)$ product rule
- $p(x) = \sum_k p(x,y_k)$ marginalisation
- Sum -> integral continuum limit ($p=\text{pdf}$)
- $p(x,y) = p(y,x)$ gives $p(y|x) = p(x|y)p(y) / p(x)$
(Bayes' theorem)

$p(x|y)$ is not the same as $p(y|x)$

- $x = \text{female}, y = \text{pregnant}$
- $p(y|x) = 0.03$
- $p(x|y) = 1$

Consistency

- This probability system is consistent e.g. 2 sets of experimental data, x, x'
- We can either combine the data sets, and compute the posterior given any prior information I , $p(\theta | (xx'), I)$
- Or, we can update the prior I with the data x , to get a new prior I' . Then we can calculate $p(\theta | x', I')$
- These give the same answer (exercise)

An exercise in using Bayes' theorem

You choose
this one



?

Do you change your choice?

This is the Monty Hall problem



Bayes' Theorem and Inference

- If we accept p as a degree of belief, then what we often want to determine is*

$$p(\theta|x)$$

θ : model parameter(s), x : the data

To compute it, use Bayes' theorem

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

* This is RULE 1: start by writing down what you want to know

Posteriors, likelihoods, priors and evidence

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

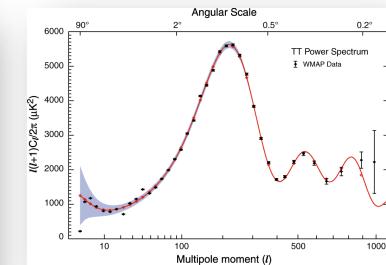
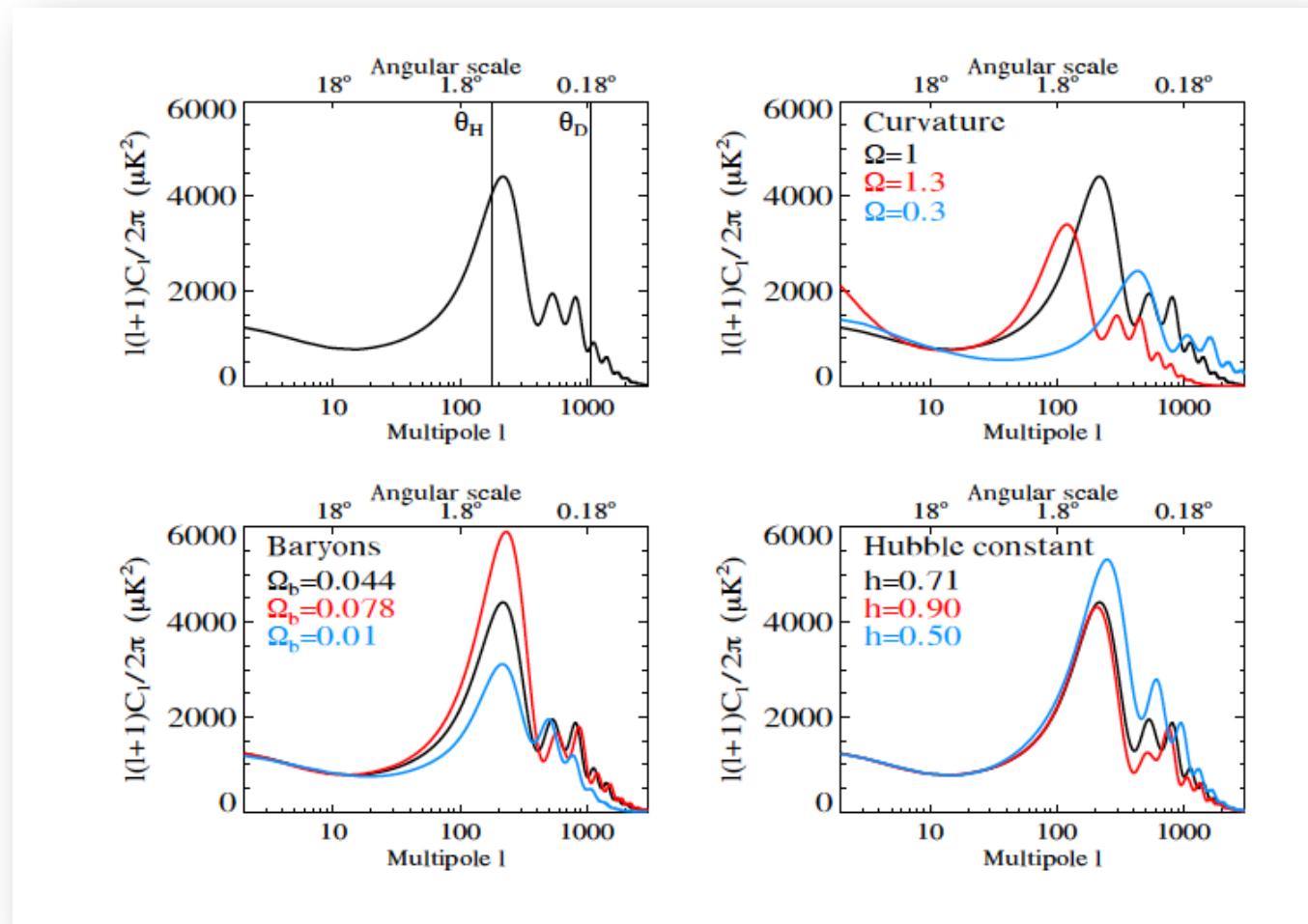
The diagram illustrates the components of Bayes' theorem. At the top is the formula $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$. Below the formula, four labels are positioned: "Posterior" on the far left, "Likelihood L" in the middle-left, "Evidence" in the center, and "Prior" on the far right. Blue arrows point from each label to its corresponding term in the formula: the "Posterior" arrow points to $p(\theta|x)$, the "Likelihood L" arrow points to $p(x|\theta)$, the "Evidence" arrow points to $p(x)$, and the "Prior" arrow points to $p(\theta)$.

Note that we interpret these in the context of a model M , so all probabilities are really conditional on M (and indeed on any prior info I). E.g. $p(\theta) = p(\theta|M)$

The Evidence looks rather odd – what is the probability of the data? For parameter estimation, we can ignore it – it simply normalises the posterior.

Noting that $p(x) = p(x|M)$ makes its role clearer. In model selection (from M and M'), $p(x|M) \neq p(x|M')$

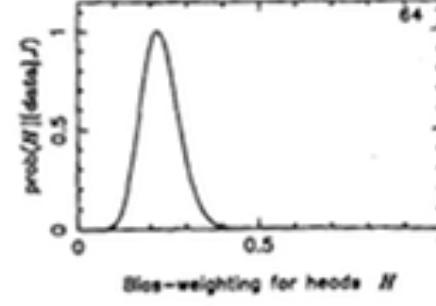
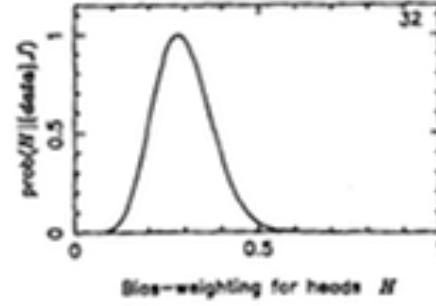
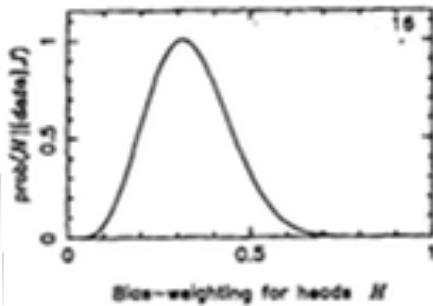
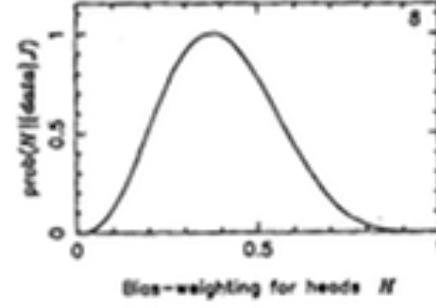
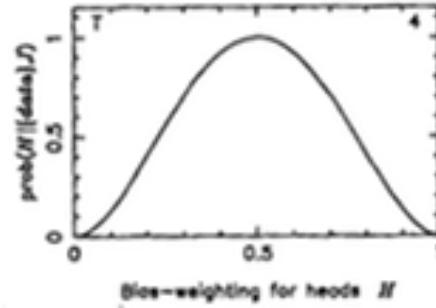
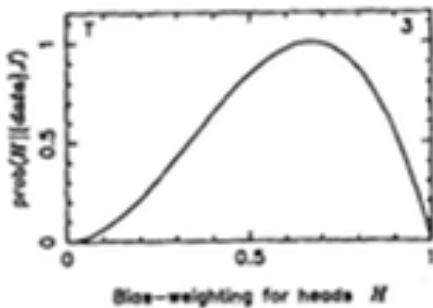
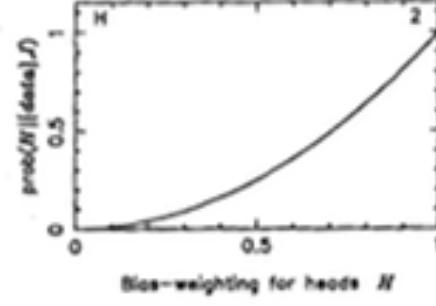
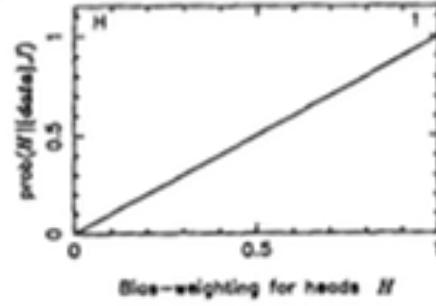
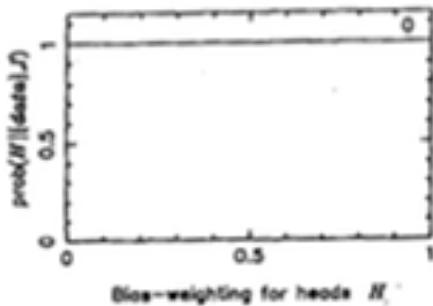
Forward modelling $p(x|\theta)$



With noise properties we can predict the *Sampling Distribution* (the probability for a general set of data; the *Likelihood* is the probability for the specific data we have)

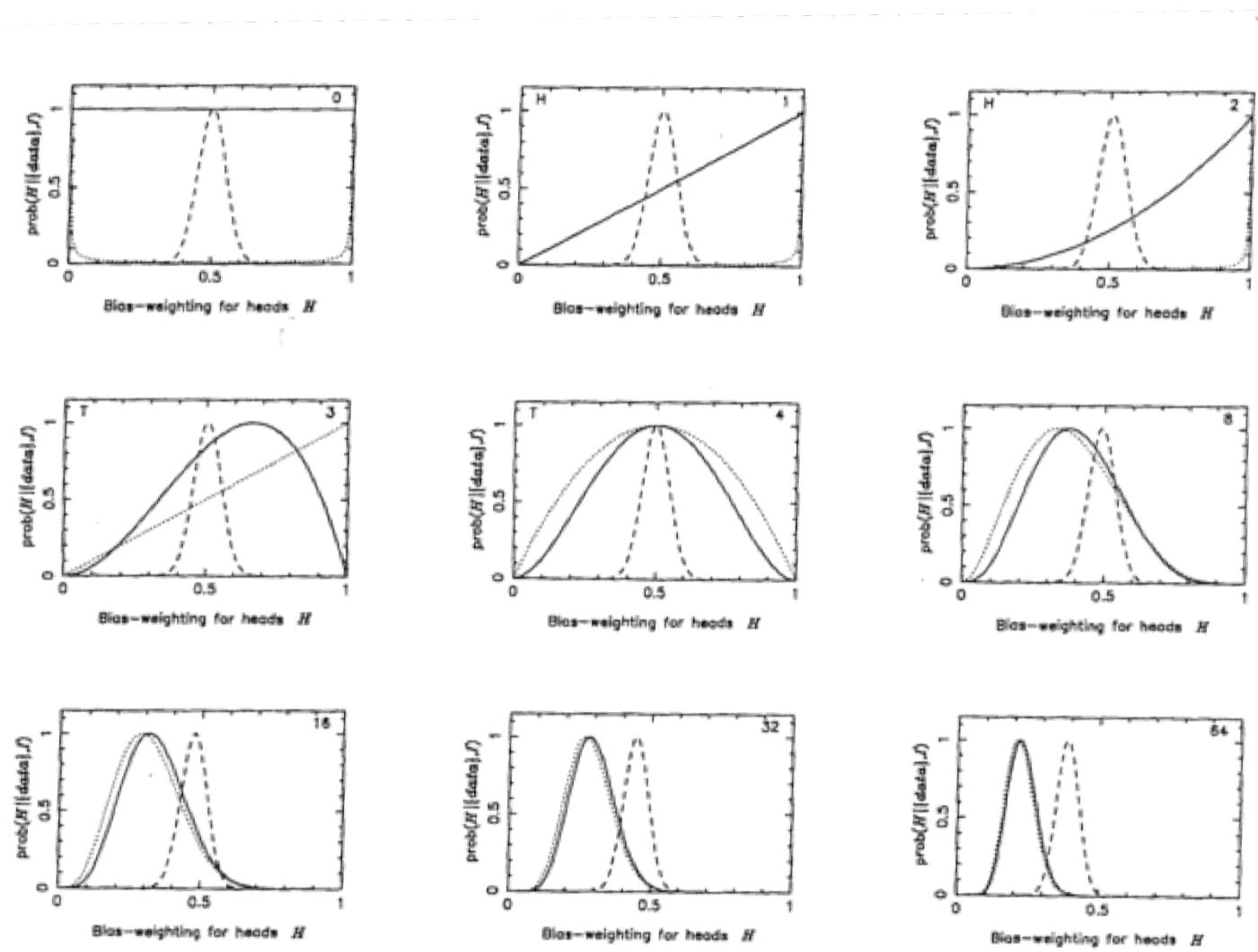
State your priors

- In easy cases, the effect of the prior is simple
- As experiment gathers more data, the likelihood tends to get narrower, and the influence of the prior diminishes
- **Rule of thumb:** if changing your prior to another reasonable one changes the answers significantly, you need more data
- **Reasonable priors?** Uninformative* – constant prior; scale parameters in $[0, \infty)$; uniform in log of parameter (Jeffreys' prior*)
- **Beware:** in more complicated, multidimensional cases, your prior may have subtle effects...
- * Actually, it's better not to use these terms – other people use them to mean different themes – just say what your prior is!



Sivia & Skilling. IS THE COIN FAIR?

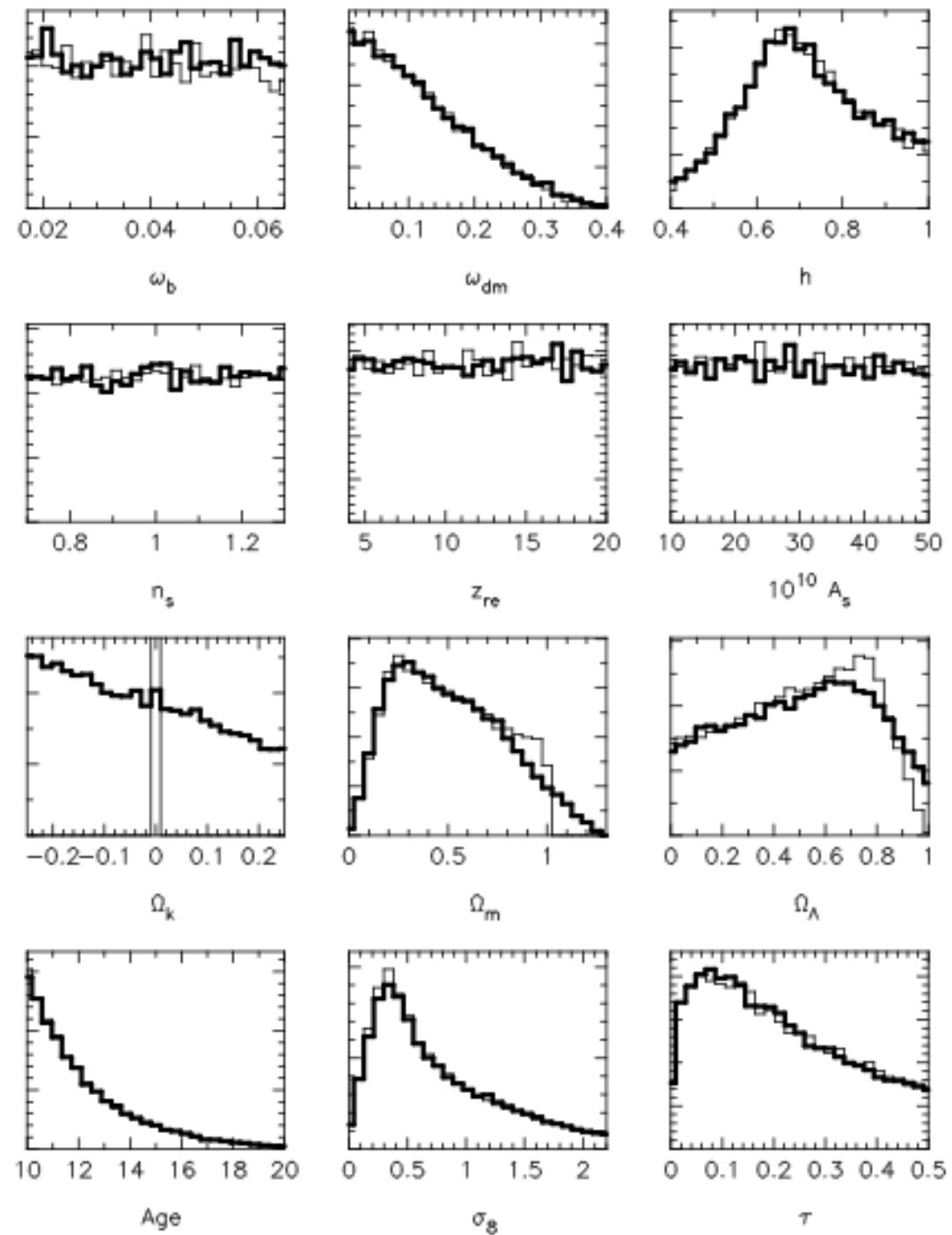
The effect of priors



Sivia & Skilling

- VSA CMB experiment

(Slosar et al 2003)

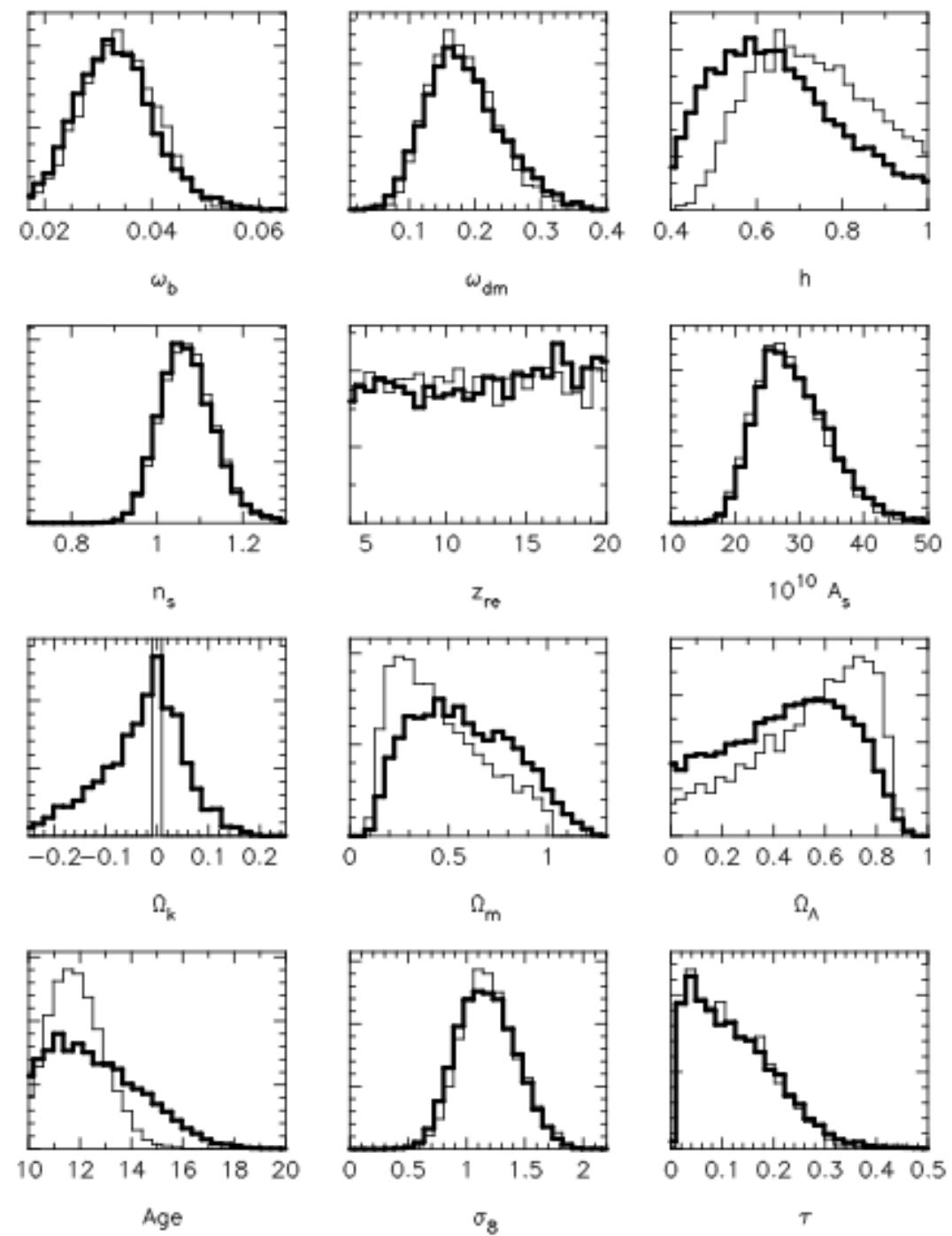


Priors: $\Lambda \geq 0$
 $10 \leq \text{age} \leq 20 \text{ Gyr}$

$$H \approx 0.7 \pm 0.1$$

There are no data in
these plots – it is all
coming from the
prior!

VSA posterior



Estimating the parameter(s)

- Commonly the mode is used (the peak of the posterior)
- If the priors are uniform, then this is the maximum likelihood estimator of frequentist statistics, but in general it differs
- The *posterior mean* may also be quoted

$$\bar{\theta} = \int \theta p(\theta|x) d\theta$$

Errors

If we assume uniform priors, then the posterior is proportional to the likelihood.

If further, we assume that the likelihood is single-moded (one peak at θ_0), we can make a Taylor expansion of $\ln L$:

$$\ln L(x; \theta) = \ln L(x; \theta_0) + \frac{1}{2}(\theta_\alpha - \theta_{0\alpha}) \frac{\partial^2 \ln L}{\partial \theta_\alpha \partial \theta_\beta} (\theta_\beta - \theta_{0\beta}) + \dots$$

$$L(x; \theta) = L_0 \exp \left[-\frac{1}{2}(\theta_\alpha - \theta_{0\alpha}) H_{\alpha\beta} (\theta_\beta - \theta_{0\beta}) + \dots \right]$$

where the Hessian matrix is defined by these equations. Comparing this with a gaussian, the *conditional error* (keeping all other parameters fixed) is

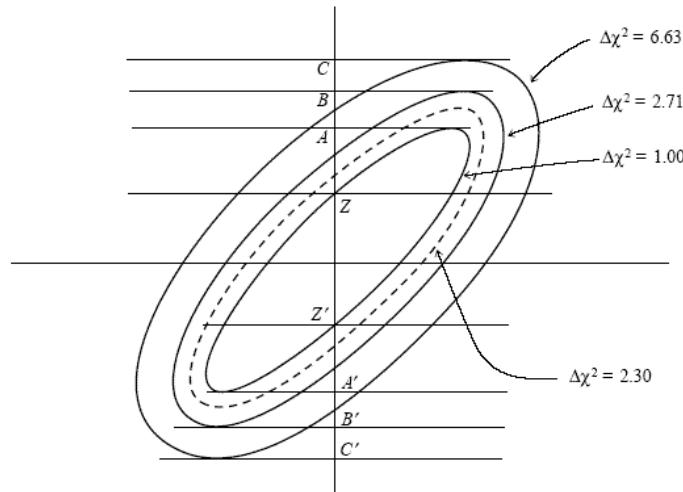
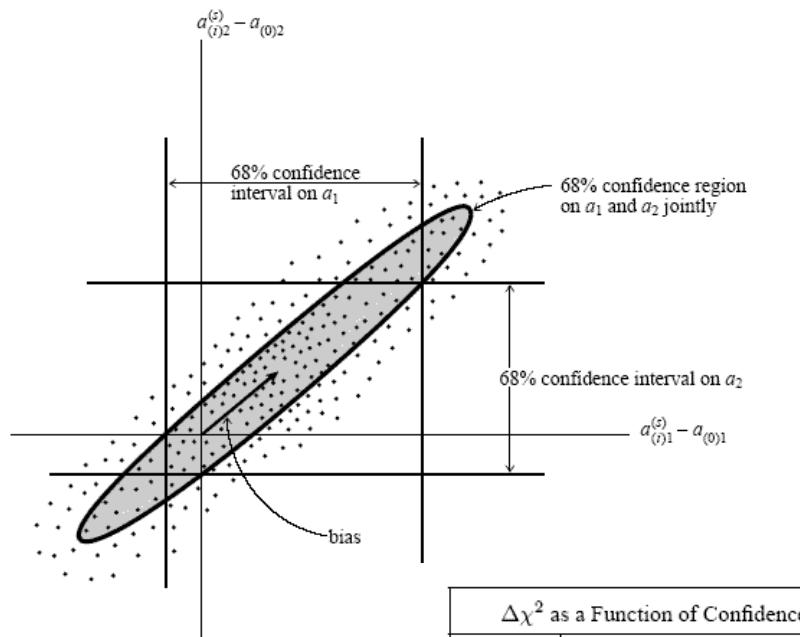
$$\sigma_\alpha = \frac{1}{\sqrt{H_{\alpha\alpha}}}$$

Marginalising over all other parameters gives the *marginal error*

$$\sigma_\alpha = \sqrt{(H^{-1})_{\alpha\alpha}}$$

How do I get error bars in several dimensions?

- Read Numerical Recipes Chapter 15.6



$$L \propto e^{-\frac{1}{2}\chi^2}$$

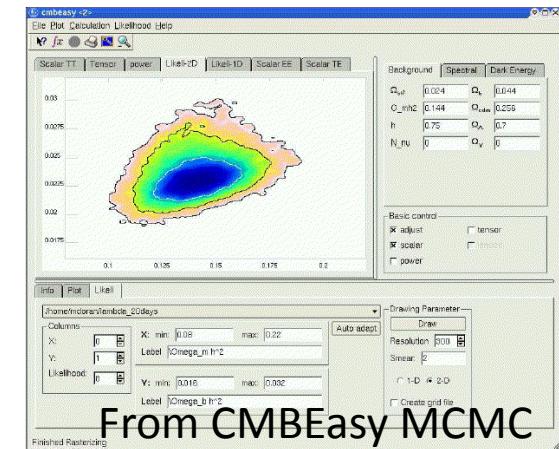
| p | ν | | | | | |
|--------|-------|------|------|------|------|------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 68.3% | 1.00 | 2.30 | 3.53 | 4.72 | 5.89 | 7.04 |
| 90% | 2.71 | 4.61 | 6.25 | 7.78 | 9.24 | 10.6 |
| 95.4% | 4.00 | 6.17 | 8.02 | 9.70 | 11.3 | 12.8 |
| 99% | 6.63 | 9.21 | 11.3 | 13.3 | 15.1 | 16.8 |
| 99.73% | 9.00 | 11.8 | 14.2 | 16.3 | 18.2 | 20.1 |
| 99.99% | 15.1 | 18.4 | 21.1 | 23.5 | 25.7 | 27.8 |

Beware! Assumes gaussian distribution

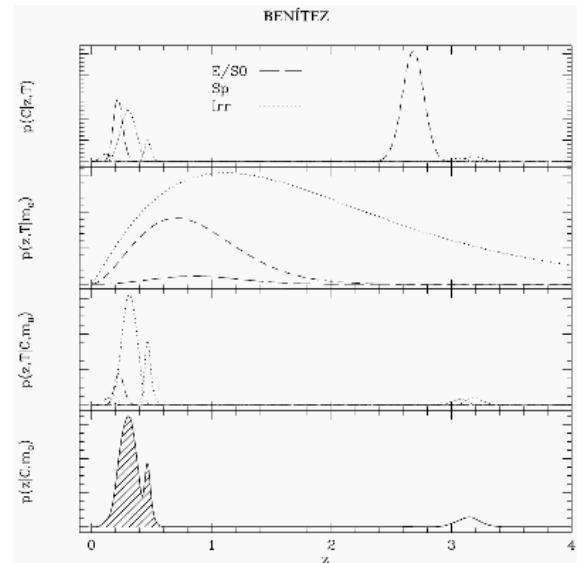
Say what your errors are – e.g.
 1σ , 2 parameter

Multimodal posteriors etc

- Peak may not be gaussian
- If the posterior does not have a single maximum, then characterising it by a mode and an error is probably inadequate. May have to present the full posterior.
- Note that the mean posterior may not be useful in this case – it could be very unlikely, if it is a valley between 2 peaks.



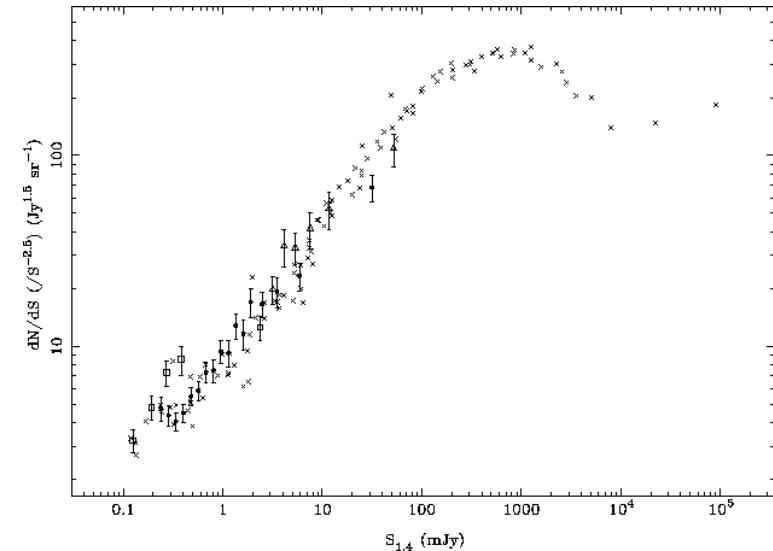
From CMBEasy MCMC



From BPZ

An example

- A radio source is observed with a telescope which can detect sources with fluxes above S_0 . The radio source has a flux $S_1 = 2S_0$.
- What is the slope of the number counts? (Assume $N(S) \propto S^{-\alpha}$)
- Possible answers:
 - Pretty steep ($\alpha > 1.5$)
 - Pretty shallow ($\alpha < 1.5$)
 - We can't tell from one point,
- Stupid
- Sorry – I dozed off



Fisher Matrices

- Useful for forecasting errors, and experimental design
- The likelihood depends on the data collected.
Can we estimate the errors before we do the experiment?
- With some assumptions, yes, using the Fisher matrix

$$F_{\alpha\beta} \equiv - \left\langle \frac{\partial^2 \ln L}{\partial \theta_\alpha \partial \theta_\beta} \right\rangle$$

Gaussian errors

- If the data have gaussian errors (which may be correlated) then we can compute the Fisher matrix easily:

$$F_{\alpha\beta} = \frac{1}{2} \text{Tr}[C^{-1} C_{,\alpha} C^{-1} C_{,\beta} + C^{-1} M_{\alpha\beta}],$$

e.g. Tegmark, Taylor, Heavens 1997

Forecast
marginal error
on parameter α

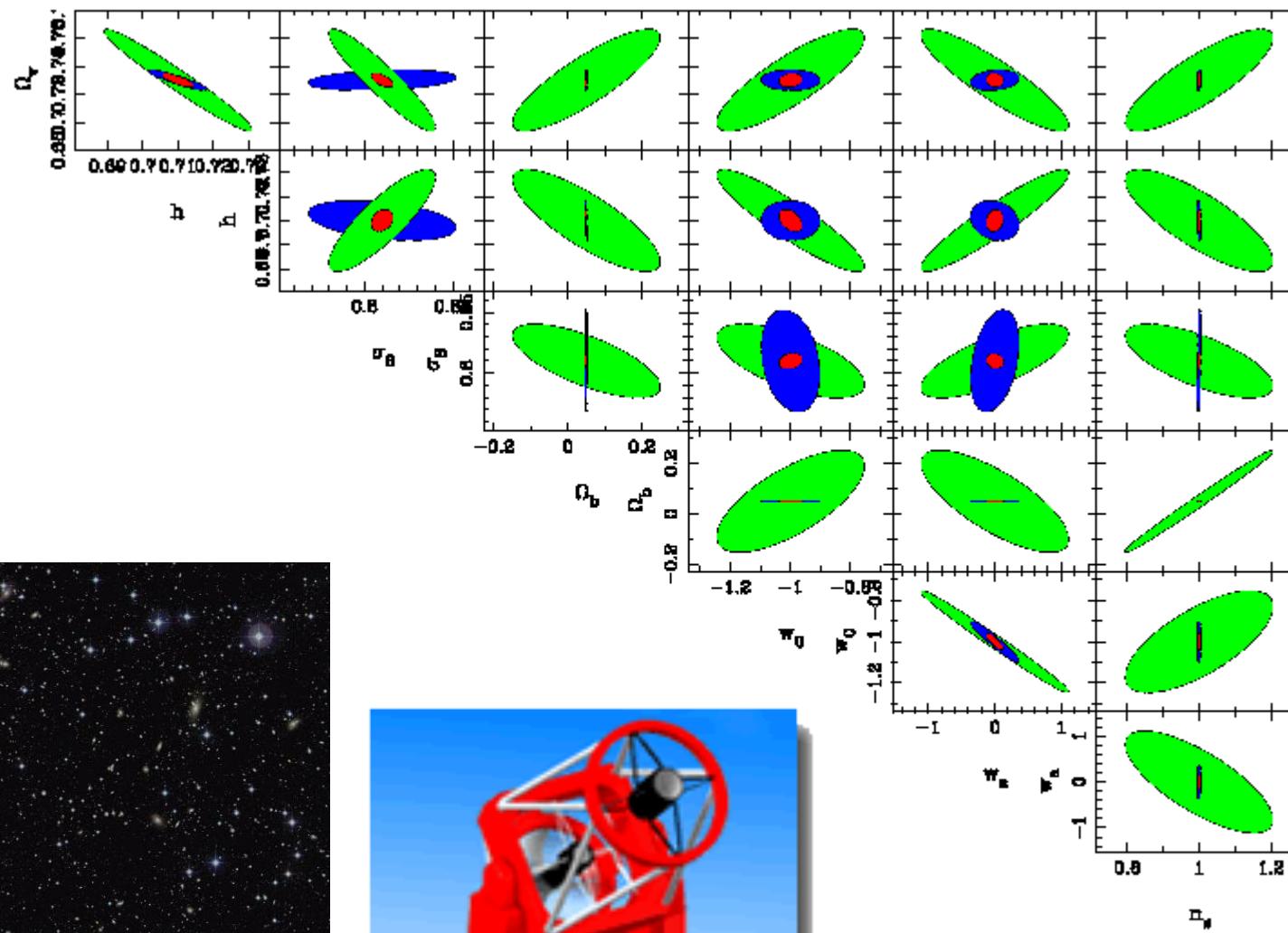
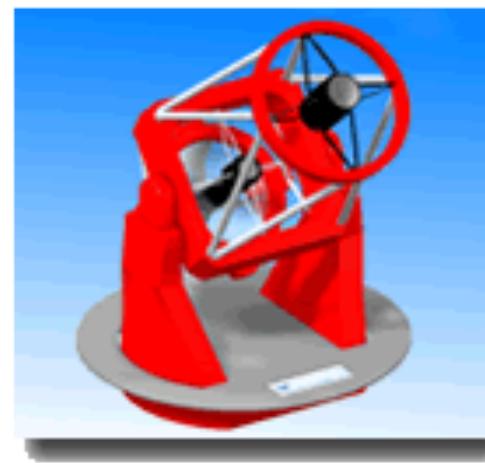
$$\sigma_\alpha = \sqrt{(F^{-1})_{\alpha\alpha}}$$

$$M_{\alpha\beta} = \mu_{,\alpha} \mu_{,\beta}^T + \mu_{,\alpha}^T \mu_{,\beta}$$
$$\mu_\alpha = \langle x_\alpha \rangle \qquad \qquad C_{\alpha\beta} = \langle (x - \mu)_\alpha (x - \mu)_\beta \rangle$$

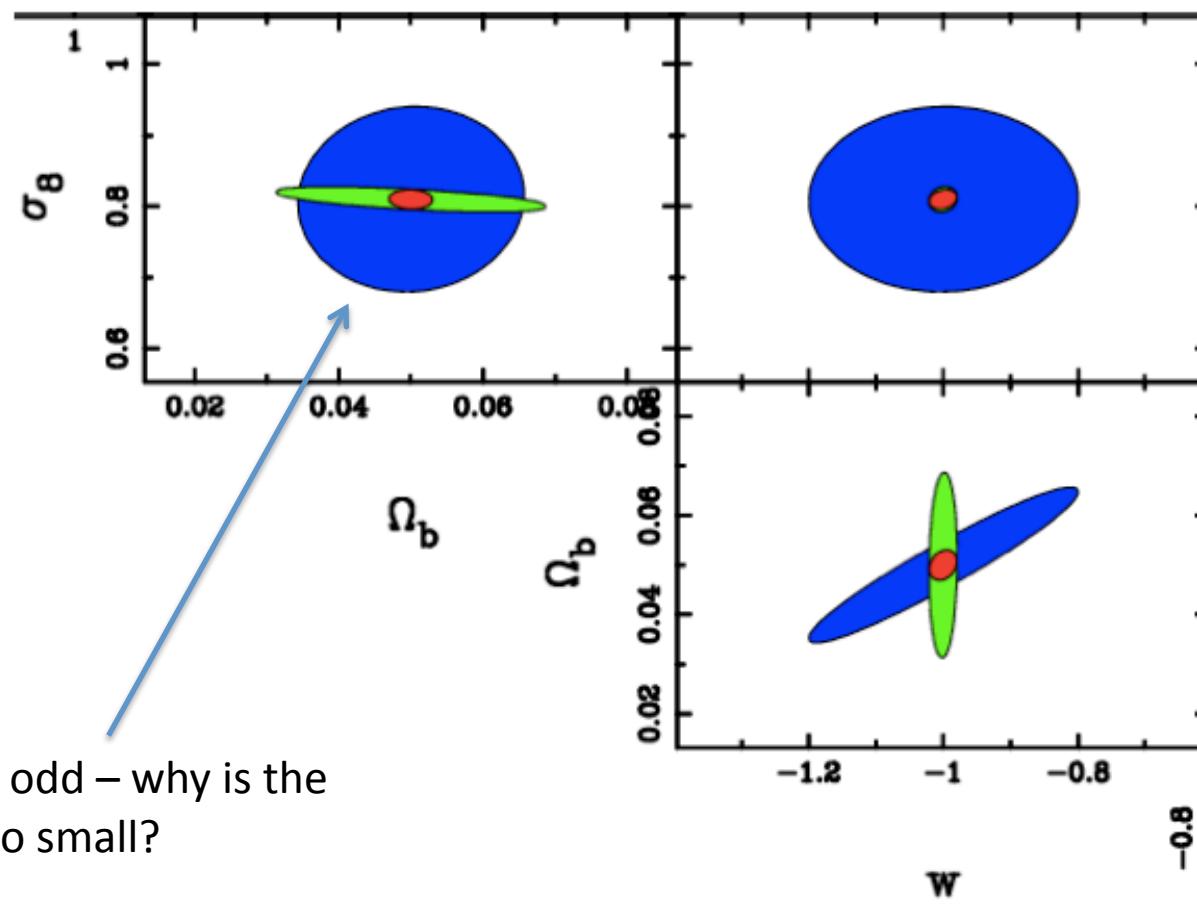
Exercise (not easy): prove this

Smaller problems along the way...

1. Prove that $(C^{-1})_{,\alpha} = -C^{-1}C_{,\alpha}C^{-1}$
2. Prove that $(\ln C)_{,\alpha} = C^{-1}C_{,\alpha}$.
3. Prove that $\ln \det C = \text{Tr } \ln C$.



Combining datasets



Open source Fisher matrices – icosmo.org

[Initiative](#) [Tools](#) [Resources](#) [Help](#) [Contact Us](#) [FAQs](#)

INITIATIVE FOR COSMOLOGY

Welcome!

This site is designed to make cosmology calculations easy and pain-free. Here, you will find a host of tools and resources for performing calculations, ranging from distance calculations to cosmological error predictions for future surveys.

The site also contains a set of tutorials and links that are useful whether you are a newbie to cosmology or a seasoned professional. These resources have been made available in an easy-to-access format and will be continually updated and expanded.

COSMOLOGY TOOLS:
 You can perform a calculation either by using your web browser or by [downloading the source code](#). To get started you can either go to [tools](#), and you will be guided through each step. Alternatively, you can use the QuickStart Calculator to the right.

COSMOLOGY RESOURCES:
 Here you will find general cosmology support materials, such as tutorials and links to external sites. To find the material you need go to [resources](#) or use the QuickStart Tutorial to the right. If you wish to create your own interactive web pages you can use the templates available [here](#). A discussion forum for the tools and resources is provided at [Cosmocoffee](#).

NEWS:
 21/05/2009 - **w(z) eigenfunctions**. Module for [astro-ph/0905.3383](#) to be included in iCosmo v1.2.
 20/05/2009 - **Hardware-Software balance**. Code for [astro-ph/0905.3176](#) can be downloaded here [iCosmo PublicAstroCodes](#).
 11/02/2009 - **Redshift Distortion & ISW**. Module for [astro-ph/0902.1759](#) to be included in iCosmo v1.2.
 21/01/2009 - **Cloud Cosmology**. Article available [here](#). Template web pages available [here](#).

QuickStart Calculator

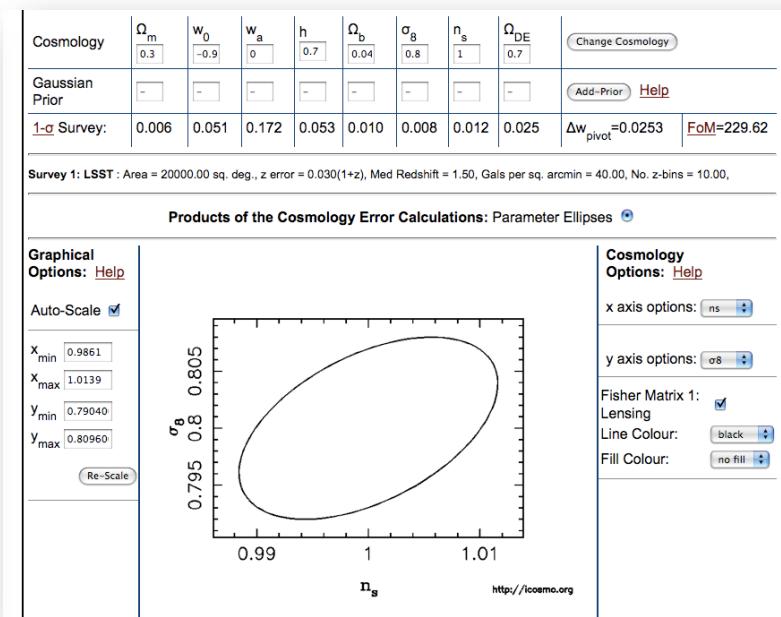
| | | | |
|------------|-------|---------------|-------|
| Ω_m | 0.3 | Ω_{DE} | 0.7 |
| Ω_b | 0.045 | w_0 | -0.95 |
| h | 0.7 | w_a | 0.0 |
| σ_8 | 0.8 | n_s | 1.0 |

[QuickStart Cosmology](#)

QuickStart Tutorial

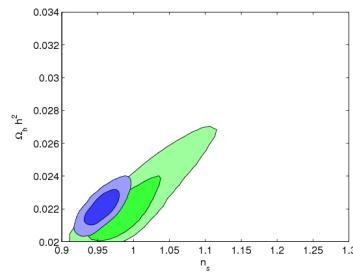
- Gravitational Lensing
- Galaxy Correlations
- CMB

[QuickStart Tutorial](#)



Computing posteriors

- For 2 parameters, a grid is usually possible
 - Marginalise by numerically integrating along each axis of the grid
- For >>2 parameters it is not feasible to have a grid (e.g. 10 points in each parameter direction, 12 parameters = 10^{12} likelihood evaluations)
- MCMC etc





Model Selection



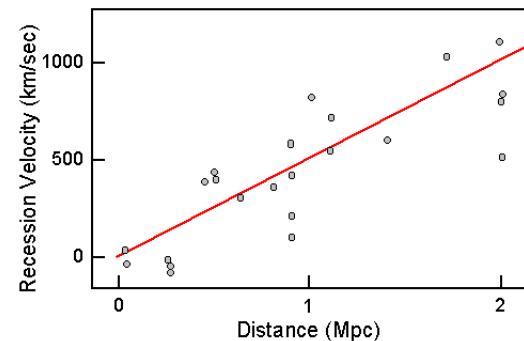
- Model selection: in a sense a higher-level question than parameter estimation
- Is the theoretical framework OK, or do we need to consider something else?
- We can compare widely different models, or want to decide whether we need to introduce an additional parameter into our model (e.g. curvature)
- In the latter case, using likelihood alone is dangerous: the new model will always be at least as good a fit, and virtually always better, so naïve maximum likelihood won't work.

Mr A and Mr B

- Mr A has a theory that $v = 0$ for all galaxies.
- Mr B has a theory that $v = Hr$ for all galaxies, where H is a free parameter.
- Who should we believe?



Hubble's Data (1929)



Bayesian approach

- Let models be M, M'
- Apply Rule 1: Write down what you want to know. Here it is $p(M|x)$ - the probability of the model, given the data.

More Bayes:

$$p(M|\mathbf{x}) = \frac{p(\mathbf{x}|M)p(M)}{p(\mathbf{x})}$$

$$\frac{p(M'|\mathbf{x})}{p(M|\mathbf{x})} = \frac{p(M')}{p(M)} \frac{\int d\boldsymbol{\theta}' p(\mathbf{x}|\boldsymbol{\theta}', M')p(\boldsymbol{\theta}'|M')}{\int d\boldsymbol{\theta} p(\mathbf{x}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)}$$

Define the Bayes factor as the ratio of evidences:

$$B \equiv \frac{\int d\boldsymbol{\theta}' p(\mathbf{x}|\boldsymbol{\theta}', M')p(\boldsymbol{\theta}'|M')}{\int d\boldsymbol{\theta} p(\mathbf{x}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)}$$

$$B \equiv \frac{\int d\boldsymbol{\theta}' p(\mathbf{x}|\boldsymbol{\theta}', M') p(\boldsymbol{\theta}'|M')}{\int d\boldsymbol{\theta} p(\mathbf{x}|\boldsymbol{\theta}, M) p(\boldsymbol{\theta}|M)}$$

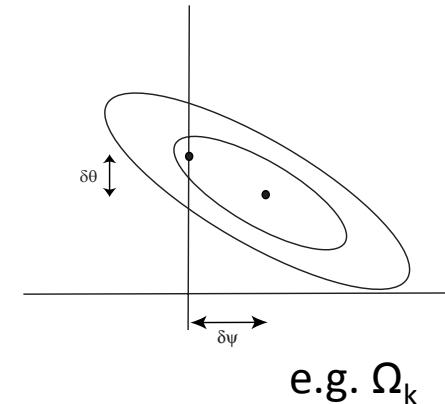
Let us assume flat priors:

$$p(\boldsymbol{\theta}|M) = (\Delta\boldsymbol{\theta}_1 \dots \Delta\boldsymbol{\theta}_n)^{-1}$$

So priors do not entirely cancel:

$$\frac{\Delta\boldsymbol{\theta}_1 \dots \Delta\boldsymbol{\theta}_n}{\Delta\boldsymbol{\theta}'_1 \dots \Delta\boldsymbol{\theta}'_{n'}} = \Delta\boldsymbol{\theta}_{n'+1} \dots \Delta\boldsymbol{\theta}_{n'+p}$$

Let us consider
NESTED MODELS



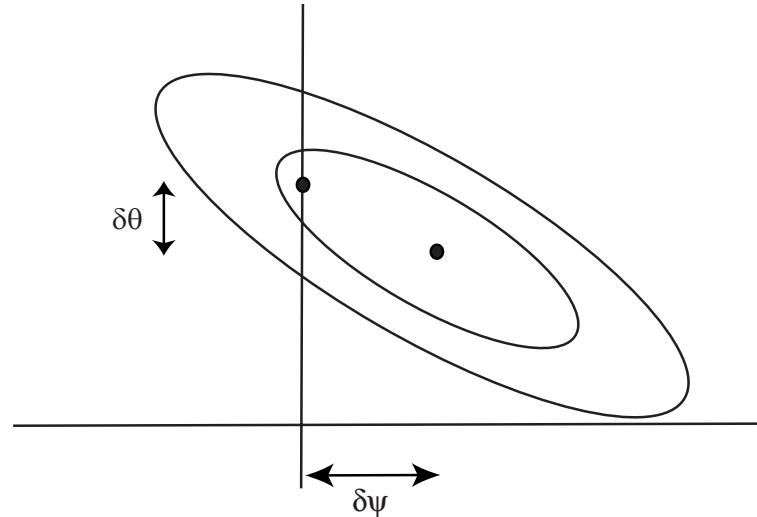
e.g. Ω_k

Laplace approximation (analogue of Fisher matrix approach, but for model selection)

$$\langle p(\mathbf{x}|\boldsymbol{\theta}, M) \rangle = L_0 \exp \left[-\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)_\alpha F_{\alpha\beta} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)_\beta \right]$$

$$\langle B \rangle = (2\pi)^{-p/2} \frac{\sqrt{\det F}}{\sqrt{\det F'}} \frac{L'_0}{L_0} \Delta\boldsymbol{\theta}_{n'+1} \dots \Delta\boldsymbol{\theta}_{n'+p}$$

$$L'_0 = L_0 \exp \left(-\frac{1}{2} \delta \boldsymbol{\theta}_\alpha F_{\alpha\beta} \delta \boldsymbol{\theta}_\beta \right)$$



$$\delta \boldsymbol{\theta}'_\alpha = -(F'^{-1})_{\alpha\beta} G_{\beta\zeta} \delta \psi_\zeta$$

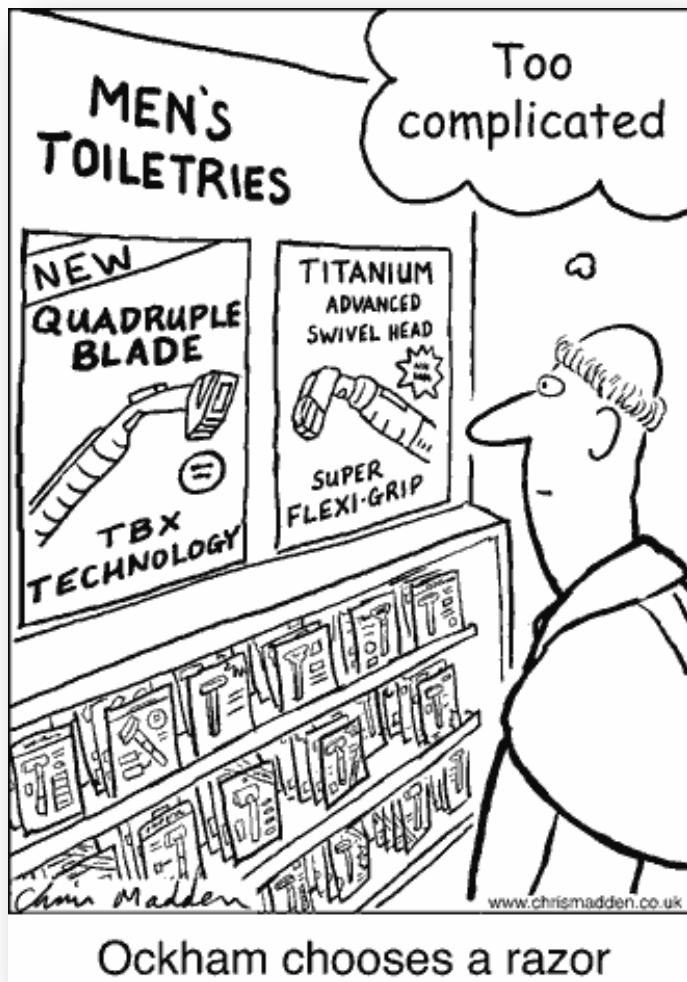
G is a subset of the Fisher matrix

$$\langle B \rangle = (2\pi)^{-p/2} \frac{\sqrt{\det F}}{\sqrt{\det F'}} \exp \left(-\frac{1}{2} \delta \boldsymbol{\theta}_\alpha F_{\alpha\beta} \delta \boldsymbol{\theta}_\beta \right) \prod_{q=1}^p \Delta \boldsymbol{\theta}_{n'+q}$$

Occam's razor... tends to favour simpler model.



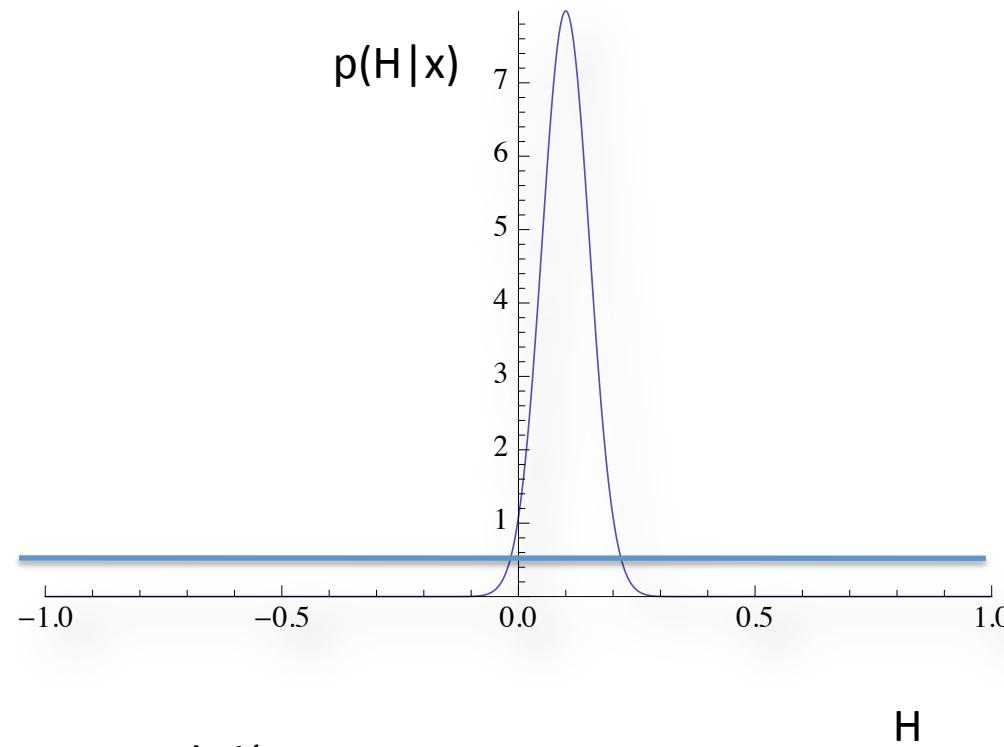
Occam's razor



- "entities should not be multiplied unnecessarily."
- "The simplest explanation for a phenomenon is most likely the correct explanation."
- "Make everything as simple as possible, but not simpler."
- Einstein

Which model is more likely?

Mr A and Mr B

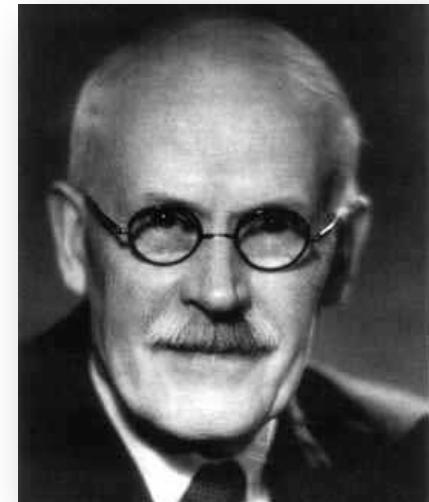


Prior of extra parameter is $\frac{1}{2}$

$$\frac{p(\text{Model A})}{p(\text{Model B})} = \frac{1.1}{0.5} = 2.2$$

Jeffreys' criteria

- Evidence:
- $1 < \ln B < 2.5$ 'substantial'
- $2.5 < \ln B < 5$ 'strong'
- $\ln B > 5$ 'decisive'
- These descriptions seem too aggressive:
 - $\ln B=1$ corresponds to a posterior probability for the less-favoured model which is 0.37 of the favoured model



Neutrino hierarchy

$\delta\alpha = -0.62$

| | σ | δ |
|----------------|-----------------------|-----------------------|
| $\Omega_b h^2$ | 10^{-4} | -10^{-4} |
| $\Omega_c h^2$ | 0.00064 | 0.00080 |
| H_0 | 0.53 | -0.62 |
| τ | 0.0028 | -0.0022 |
| n_s | 0.0019 | -0.0014 |
| A_s | $1.42 \cdot 10^{-11}$ | $-8.4 \cdot 10^{-12}$ |
| $\sum m_\nu$ | 0.027eV | 0.038eV |
| $dn_s/d \ln k$ | 0.0023 | 0.0041 |



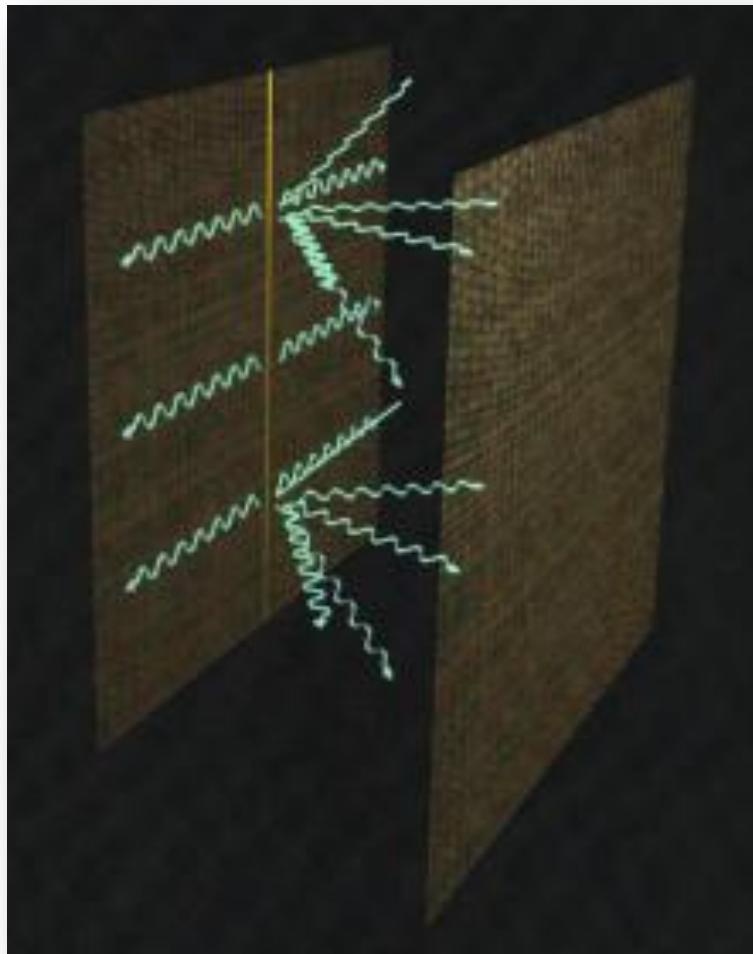
$\delta\alpha = -0.90$

| | σ | δ |
|----------------|-----------------------|------------------------|
| $\Omega_b h^2$ | 10^{-4} | $-1.6 \cdot 10^{-4}$ |
| $\Omega_c h^2$ | 0.00064 | 0.0011 |
| H_0 | 0.53 | -0.91 |
| τ | 0.0028 | -0.0031 |
| n_s | 0.0019 | -0.0020 |
| A_s | $1.42 \cdot 10^{-11}$ | $-1.23 \cdot 10^{-11}$ |
| $\sum m_\nu$ | 0.027eV | 0.055eV |
| $dn_s/d \ln k$ | 0.0023 | 0.0060 |

Normal, or inverted



Extra-dimensional gravity?



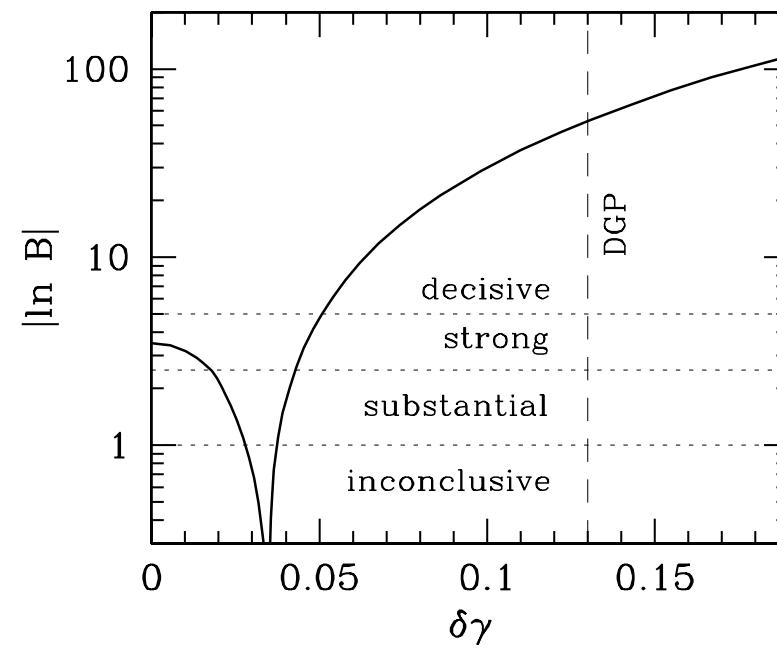
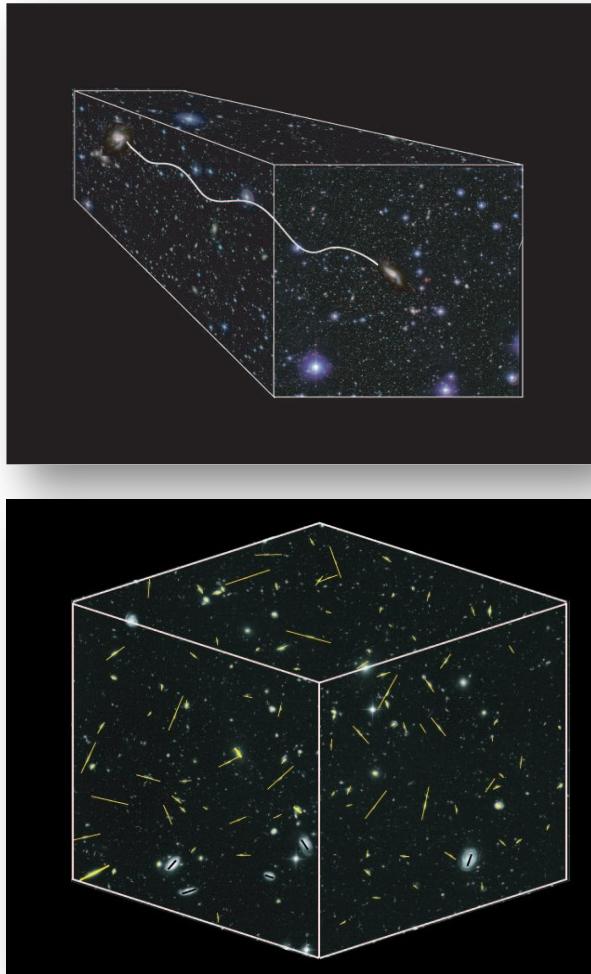
Evidence for beyond-Einstein gravity

- How would we tell? Different growth rate

$$\frac{\delta_m}{a} \equiv g(a) = \exp \left\{ \int_0^a \frac{da'}{a'} [\Omega_m(a')^\gamma - 1] \right\}$$

- $\gamma = 0.55$ (GR) 0.68 (Flat DGP model)
- Do the data demand an additional parameter, γ ?

Expected Evidence: braneworld gravity?

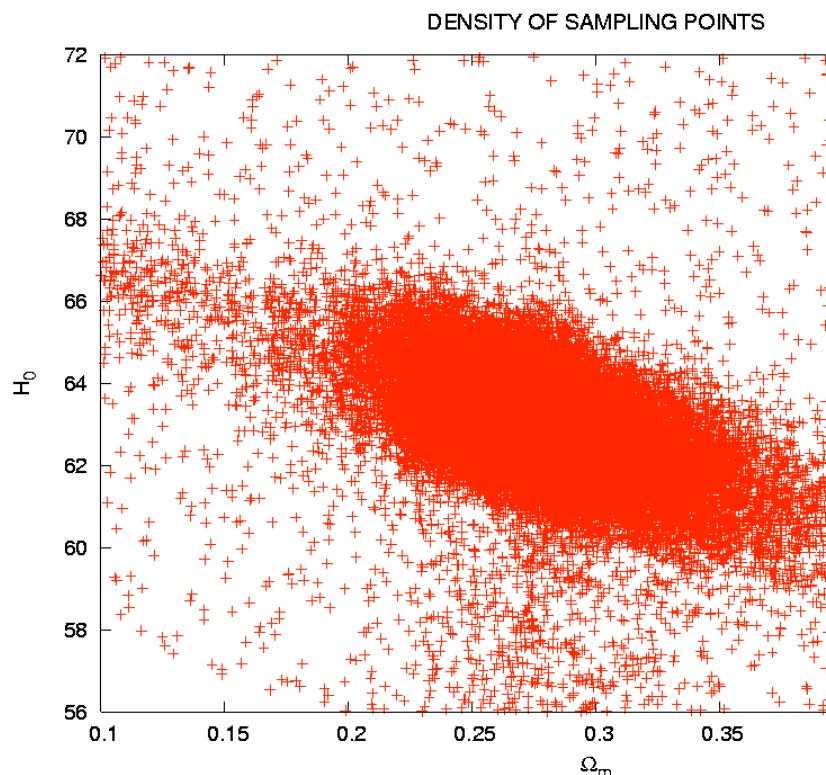


Heavens, Kitching & Verde 2007

VEGAS sampling

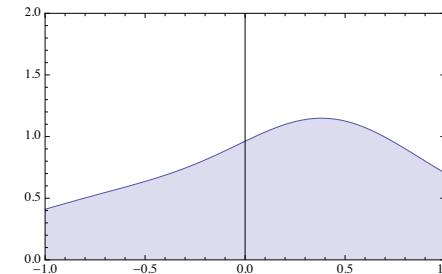
Want a way to sample from a *separable* function of the parameters.

$$P(\theta) = p_1(\theta_1) p_2(\theta_2) \dots p_N(\theta_N)$$

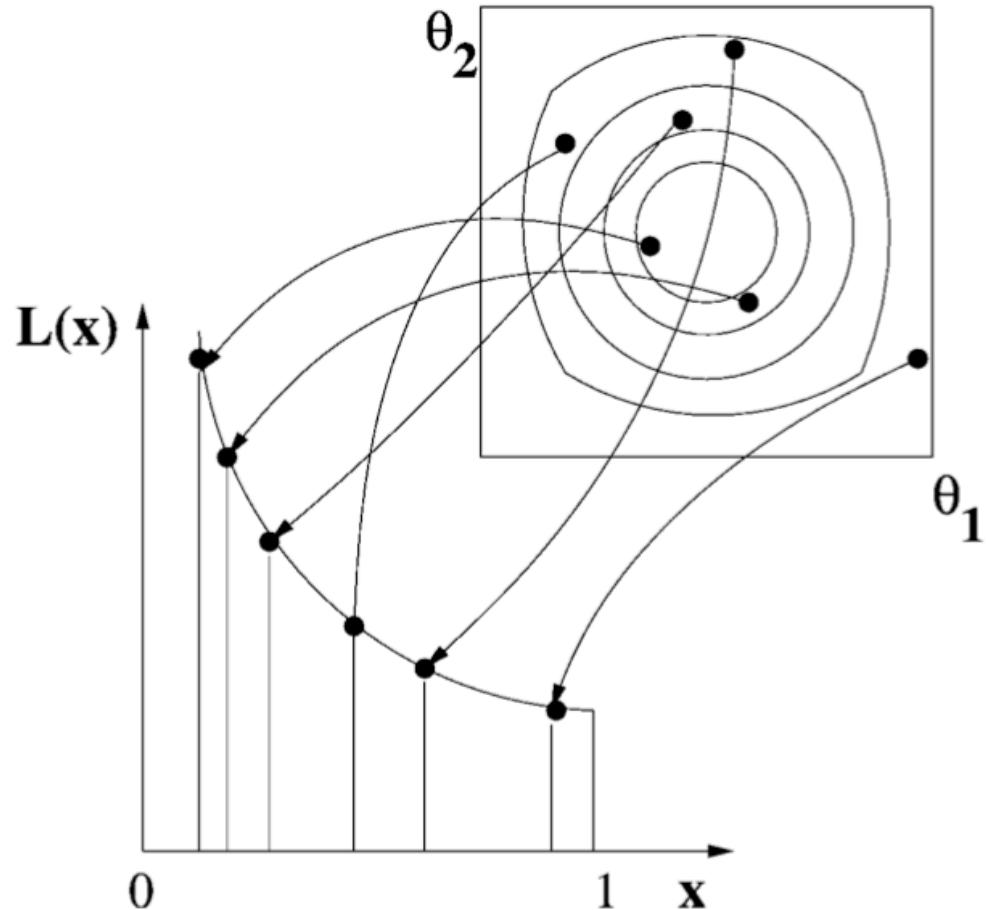


For correlated parameters,
rotate the axes first (do a
preliminary MCMC).

Use rejection method
in 1D, for example



Nested Sampling



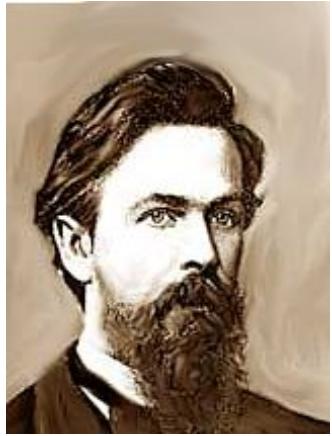
Skilling (2004)
Sample from the prior
volume, replacing the
lowest point with one from
a higher target density.

See: CosmoNEST (add-on
for CosmoMC)

Multimodal? MultiNEST

Numerical Sampling methods

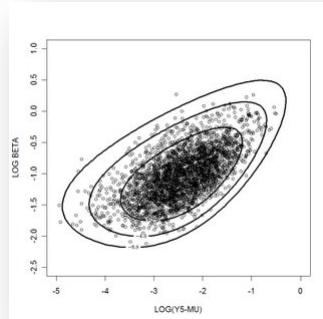
- MCMC (Markov Chain Monte Carlo)
- HMC (Hamiltonian Monte Carlo)



MCMC



Aim of MCMC: generate a set of points in the parameter space whose distribution function is the same as the target density.



MCMC follows a Markov process - i.e. the next sample depends on the present one, but not on previous ones.

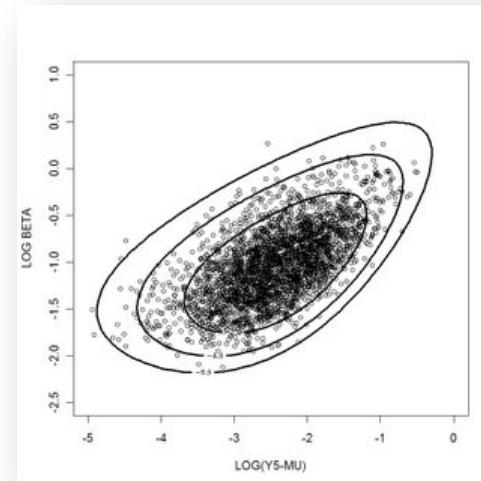
Target density

The *target density* is approximated by a set of delta functions (you may need to normalise):

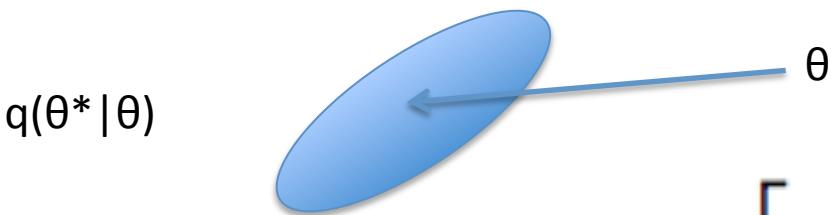
$$p(\theta) \simeq \frac{1}{N} \sum_{i=1}^N \delta(\theta - \theta_i)$$

and we can estimate any function f by

$$f(\theta) \simeq \frac{1}{N} \sum_{i=1}^N f(\theta_i).$$



Metropolis-Hastings algorithm


$$p(\text{acceptance}) = \min \left[1, \frac{p(\theta^*)q(\theta^*|\theta)}{p(\theta)q(\theta|\theta^*)} \right]$$

Metropolis algorithm (special case):

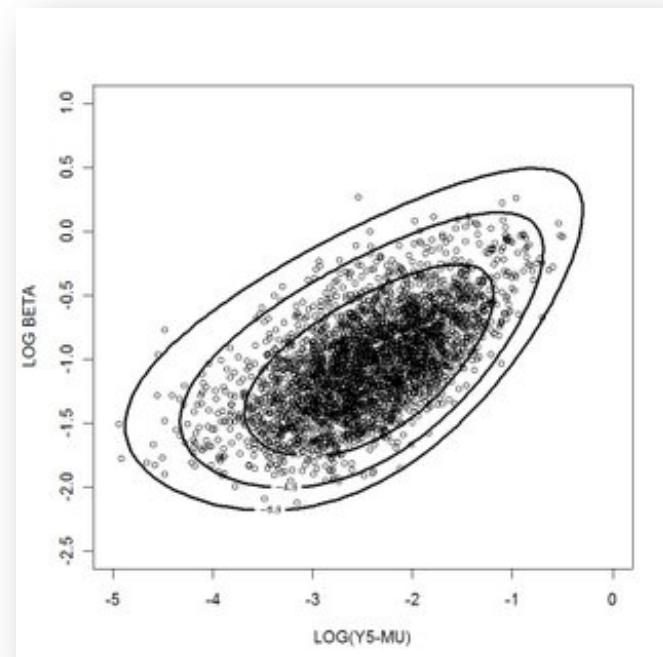
$$\min \left[1, \frac{p(\theta^*)}{p(\theta)} \right]$$

MCMC Algorithm

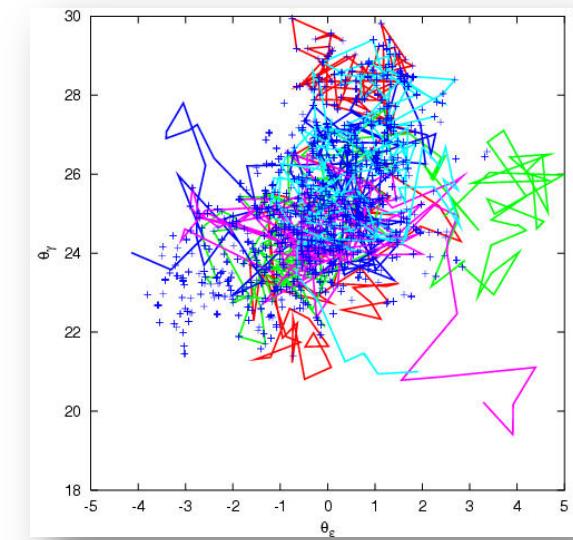
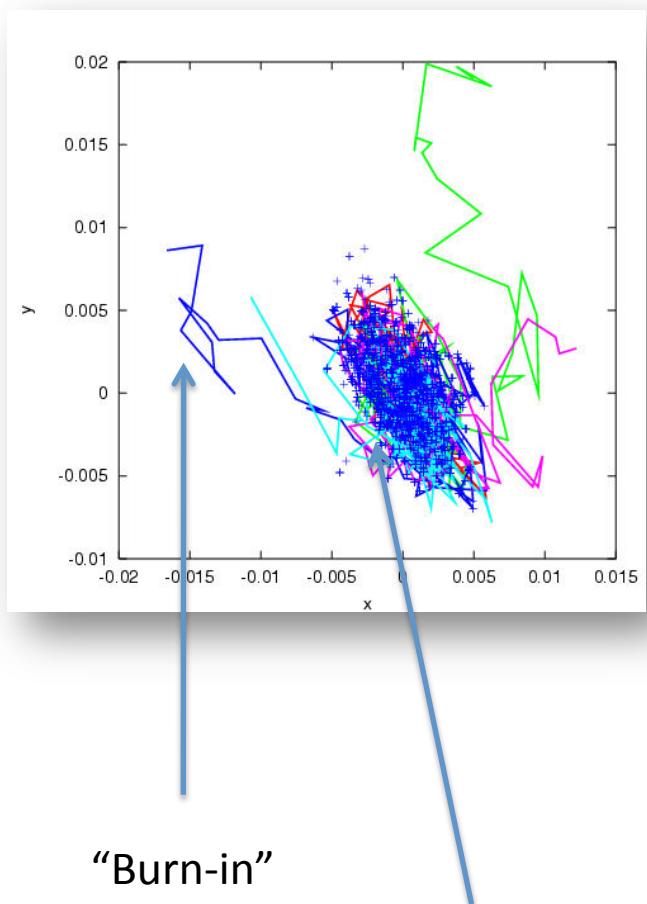
- Choose a random initial starting point in parameter space, and compute the target density.}
- Repeat:
 - Generate a step in parameter space from a proposal distribution, generating a new trial point for the chain.
 - Compute the target density at the new point, and accept it (or not) with the Metropolis-Hastings algorithm.
 - If the point is not accepted, the previous point is repeated in the chain.
- End Repeat:

The proposal distribution

- Too small, and it takes a long time to explore the target
- Too large and almost all trials are rejected
- $q \sim \text{'Fisher size'}$ is good.



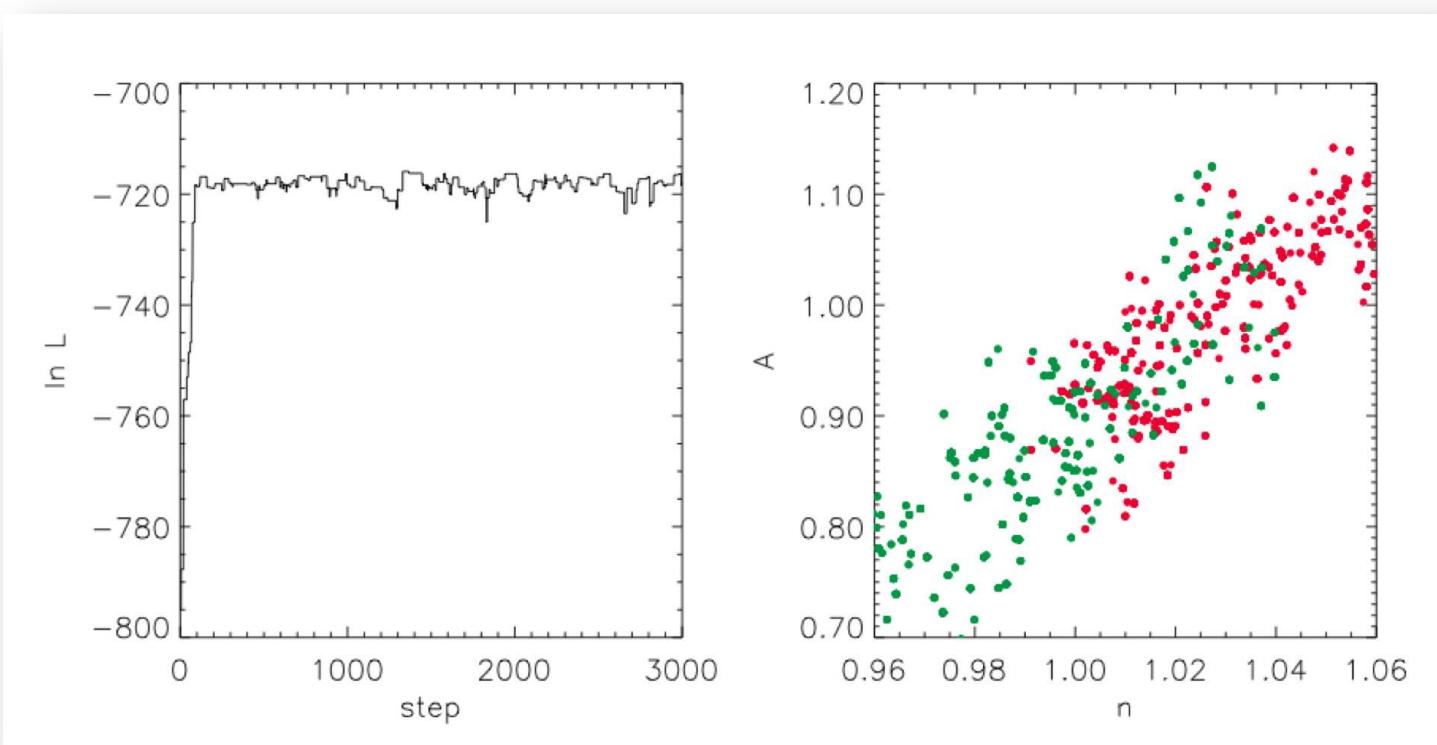
Burn-in and convergence



You **must** use a convergence test.
Gelman-Rubin test is most common (see notes)

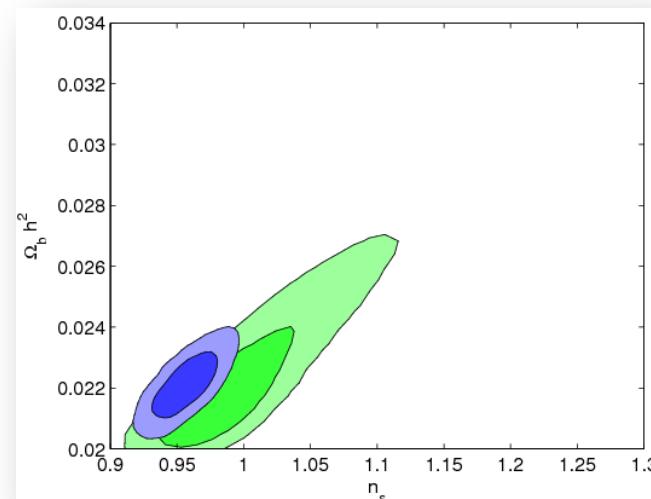
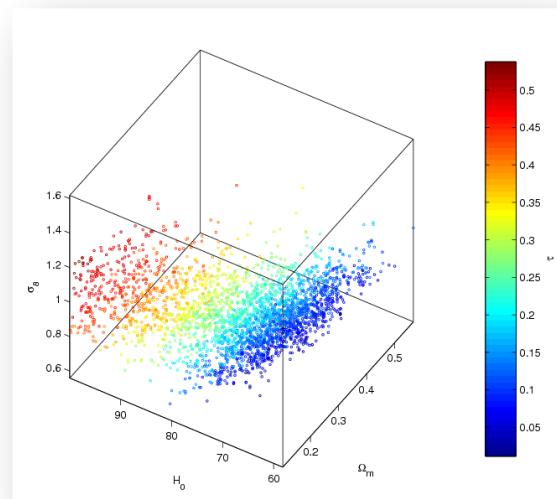
Points are correlated

Unconverged chains

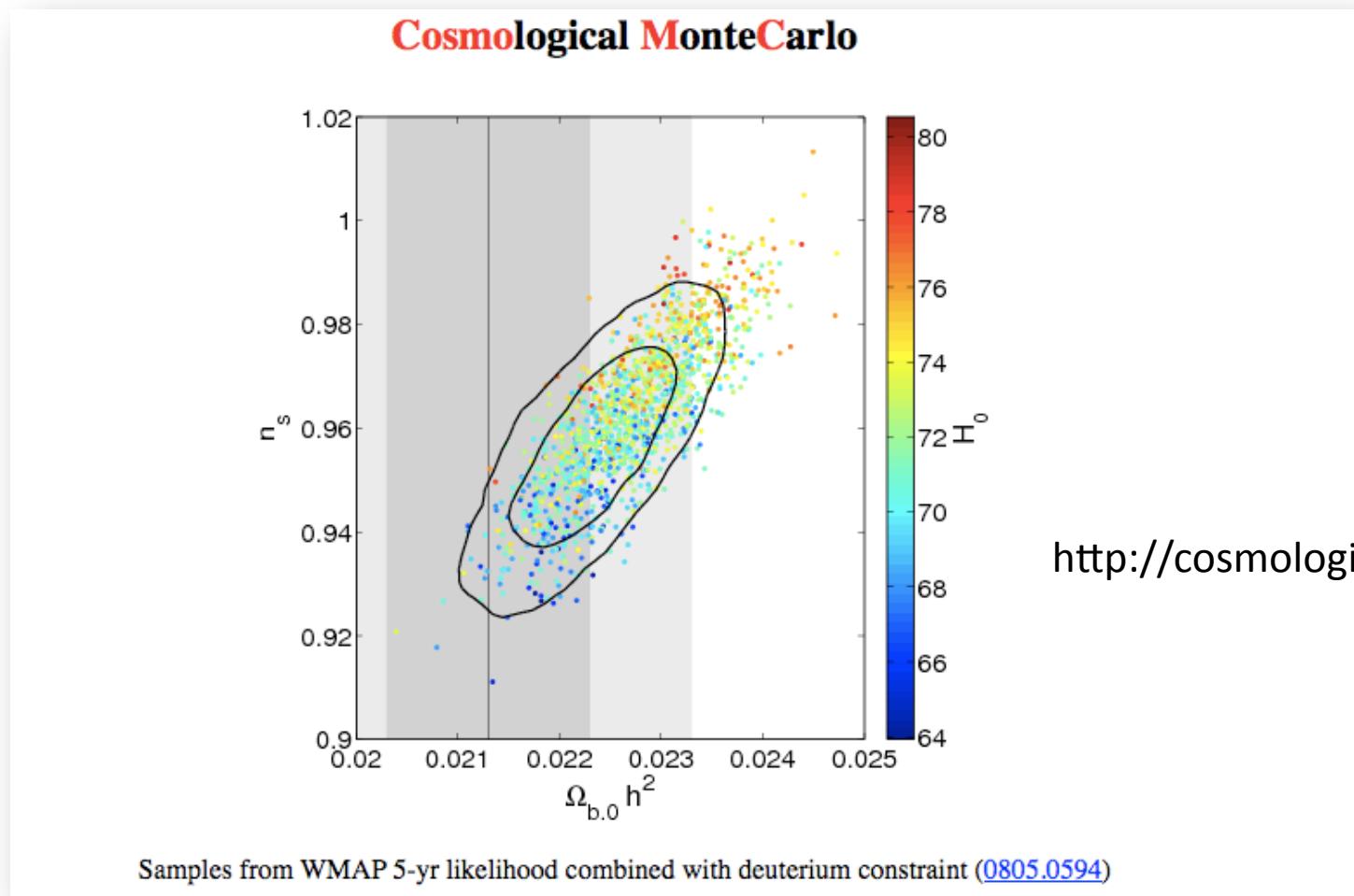


Marginalisation

- Marginalisation is trivial
 - Each point in the chain is labelled by all the parameters
 - To marginalise, just ignore the labels you don't want



CosmoMC



Hamiltonian Monte Carlo

- We would like to increase the acceptance rate to improve efficiency
- HMC works by sampling from a *larger* parameter space:
- M auxiliary variables, one for each parameter in the model.
- Imagine each of the parameters in the problem as a coordinate.
- Target distribution = effective potential
- For each coordinate HMC generates a generalised momentum.
- It then samples from the extended target distribution in $2M$ dimensions.
- It explores this space by treating the problem as a dynamical system, and evolving the phase space coordinates by solving the dynamical equations.
- Finally, it ignores the momenta (marginalising, as in MCMC), and this gives a sample of the original target distribution.

Theory

- Potential $U(\theta) = -\ln p(\theta)$
- For each θ_α , generate a momentum u_α .
- K.E. $K = u^T u / 2$
- Define a Hamiltonian

$$H(\theta, \mathbf{u}) \equiv U(\theta) + K(\mathbf{u})$$

- and define an extended target density

$$p(\theta, \mathbf{u}) = \exp [-H(\theta, \mathbf{u})]$$

Magic of HMC

- Evolve as a dynamical system

$$\begin{aligned}\dot{\theta}_\alpha &= u_\alpha \\ \dot{u}_\alpha &= -\frac{\partial H}{\partial \theta_\alpha}\end{aligned}$$



William Rowan Hamilton

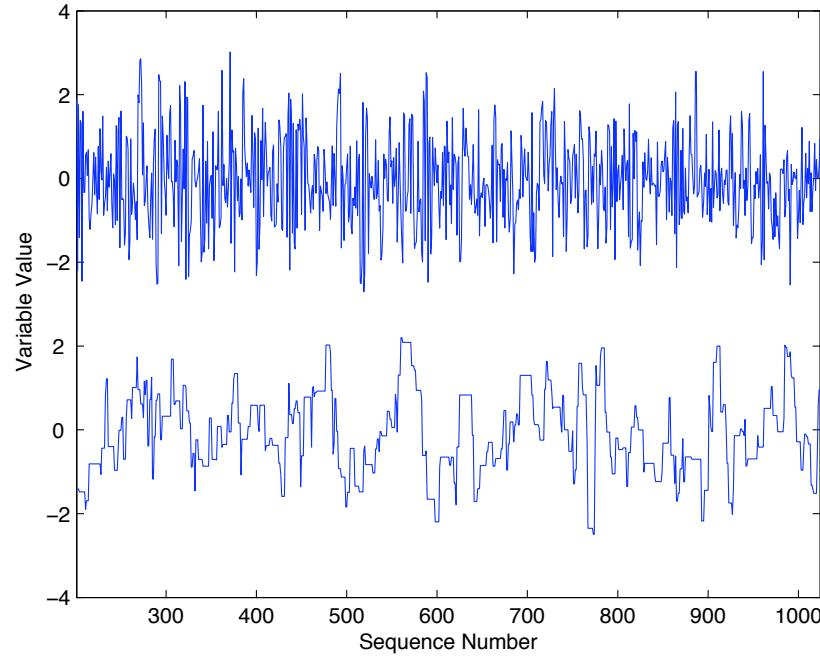
- H remains constant, so extended target density is uniform – all points get accepted!
- Also, you can make big jumps – good mixing, if you generate a new u each time a point is accepted

Complications

- Evolving the system takes time. Take big steps.
- We don't know $U = -\ln p$ (it's what we are looking for)
- We approximate U (from a short MCMC)
- H is therefore not constant
- Use Metropolis-Hastings. Accept new point with probability

$$\min \{1, \exp [-H(\boldsymbol{\theta}^*, \mathbf{u}^*) + H(\boldsymbol{\theta}, \mathbf{u})]\}$$

HMC vs MCMC



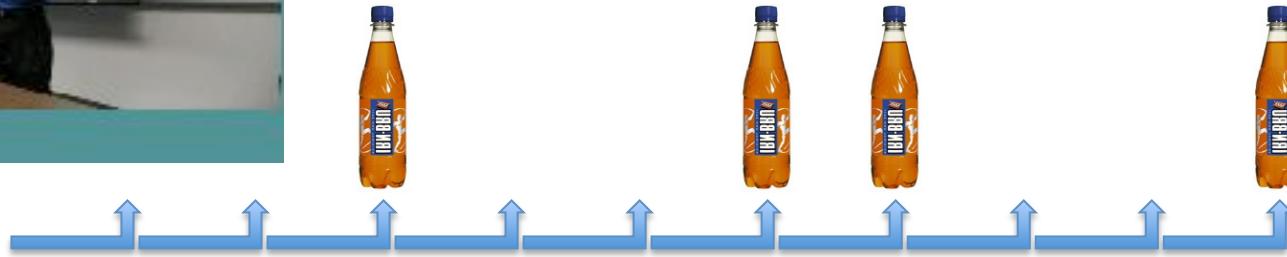
HMC should be $\sim M$ times as fast as MCMC.
Typical speed-ups: factor 4.

Monty Hall solution

- Rule 1: write down what you want
- $a = \text{Irn Bru is behind Door A}$
- $B = \text{Monty Hall opened Door B}$
- It is $p(a|B)$
- Now $p(a|B) = p(B|a)p(a)/p(B)$
- $p(B) = p(B,a) + p(B,b) + p(B,c)$ (marginalisation)
 - $p(B) = p(B|a)p(a) + p(B|b)p(b) + p(B|c)p(c)$
 - $p(B) = \frac{1}{2} \times \frac{1}{3} + 0 + 1 \times \frac{1}{3} = \frac{1}{2}$
- $p(a|B) = \frac{1}{2} \times \frac{1}{3} / \frac{1}{2} = \frac{1}{3}$ BETTER TO CHANGE



Binomial drinking*



Eric thinks about drinking a bottle of Irn Bru once every minute. He decides to drink one with probability $p = 0.1$.

The distribution tells us that the mean time between drinks is $1/p = 10$ minutes

You check at **random** times and note the time of the last drink, and the next drink, and record the **time between drinks**

The expectation value of the time you measure between drinks is $2/p-1 = 19$ minutes

WHY IS IT > 10 minutes?

* Not to be confused with Poisson drinking, which is Drinking Like a Fish

Beware!

- Be careful if your data depend on your model
 - Results on parameter errors do not apply then

Solution: choose as your data some statistics (=numbers determined in a fixed way from the data) which are based on some *fiducial set* of parameters.

e.g. Take LCDM as the fiducial model. Compute the power spectrum with this model.
Data = “Power spectrum measurements assuming LCDM”

As parameters are varied, compute the expected value of the data (which won’t be the power spectrum in the trial cosmologies).

Bayesian Computation: Overview and Methods for Low-Dimensional Models

Tom Loredo

Dept. of Astronomy, Cornell University

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

Bayesian Computation Tutorials — 11-12 June 2011

Overview/Low-Dimensional Models

- ① Bayes recap: Parameter space integrals
- ② Bayesian vs. frequentist computation
- ③ Geometry & probability in high dimensions
- ④ Large N : Laplace approximations
- ⑤ Cubature
- ⑥ Monte Carlo integration
 - Posterior sampling
 - Importance sampling
- ⑦ Bootstrapping vs. posterior sampling

Overview/Low-Dimensional Models

- ① Bayes recap: Parameter space integrals
- ② Bayesian vs. frequentist computation
- ③ Geometry & probability in high dimensions
- ④ Large N : Laplace approximations
- ⑤ Cubature
- ⑥ Monte Carlo integration
 - Posterior sampling
 - Importance sampling
- ⑦ Bootstrapping vs. posterior 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:

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

Suppress such conditions when clear from context

Recap of Key Bayesian Ideas

Probability as generalized logic

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

Bayes's theorem

$$p(\text{Hypothesis} \mid \text{Data}) \propto$$

$$p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis}) \parallel \text{Context}$$

Data change the support for a hypothesis \propto ability of hypothesis to predict the data

Law of total probability

$$p(\text{Hypotheses} \mid \text{Data}) = \sum p(\text{Hypothes}is \mid \text{Data}) \parallel \text{Context}$$

The support for a *compound/composite* hypothesis must account for all the ways it could be true

Roles of the prior

Prior has two roles

- Incorporate any relevant prior information
- Convert likelihood from “intensity” to “measure”
→ account for *size of parameter space*

Physical analogy

$$\text{Heat } Q = \int d\mathbf{r} c_v(\mathbf{r}) T(\mathbf{r})$$

$$\text{Probability } P \propto \int d\theta p(\theta) \mathcal{L}(\theta)$$

Maximum likelihood focuses on the “hottest” parameters.

Bayes focuses on the parameters with the most “heat.”

A high- T region may contain little heat if its c_v is low or if its volume is small.

A high- \mathcal{L} region may contain little probability if its prior is low or if its volume is small.

Nuisance Parameters and Marginalization

To model most data, we need to introduce parameters besides those of ultimate interest: *nuisance parameters*.

Example

We have data from measuring a rate $r = s + b$ that is a sum of an interesting signal s and a background b .

We have additional data just about b .

What do the data tell us about s ?

Marginal posterior distribution

To summarize implications for s , accounting for b uncertainty,
marginalize:

$$\begin{aligned} p(s|D, M) &= \int db \, p(s, b|D, M) \\ &\propto p(s|M) \int db \, p(b|s) \mathcal{L}(s, b) \\ &= p(s|M) \mathcal{L}_m(s) \end{aligned}$$

with $\mathcal{L}_m(s)$ the *marginal likelihood for s*:

$$\mathcal{L}_m(s) \equiv \int db \, p(b|s) \mathcal{L}(s, b)$$

Marginalization vs. Profiling

For insight: Suppose the prior is broad compared to the likelihood
→ for a fixed s , we can accurately estimate b with max likelihood \hat{b}_s , with small uncertainty δb_s .

$$\begin{aligned}\mathcal{L}_m(s) &\equiv \int db p(b|s) \mathcal{L}(s, b) \\ &\approx p(\hat{b}_s|s) \mathcal{L}(s, \hat{b}_s) \delta b_s\end{aligned}$$

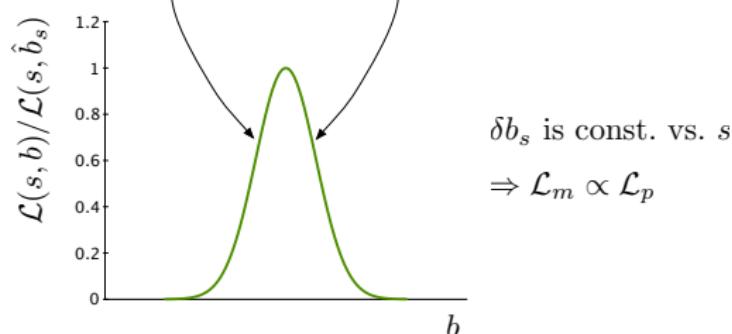
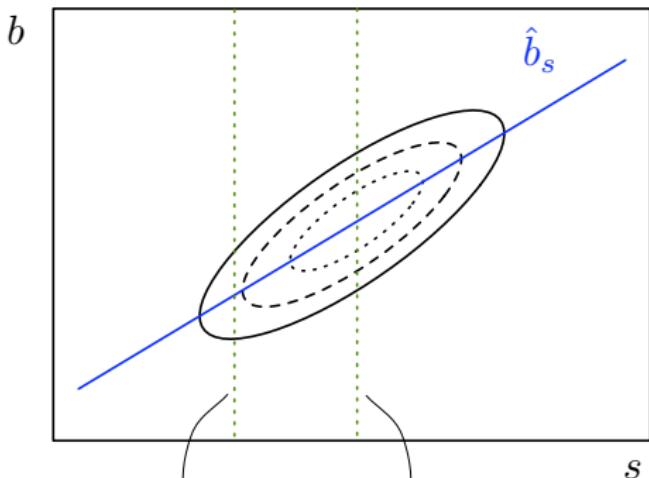
best b given s
b uncertainty given s

Profile likelihood $\mathcal{L}_p(s) \equiv \mathcal{L}(s, \hat{b}_s)$ gets weighted by a *parameter space volume factor*

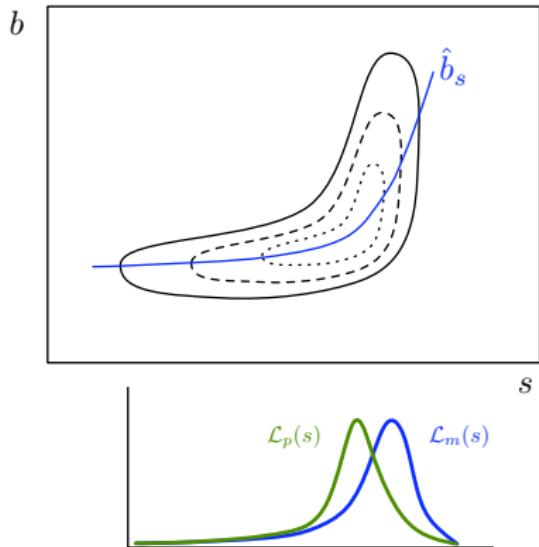
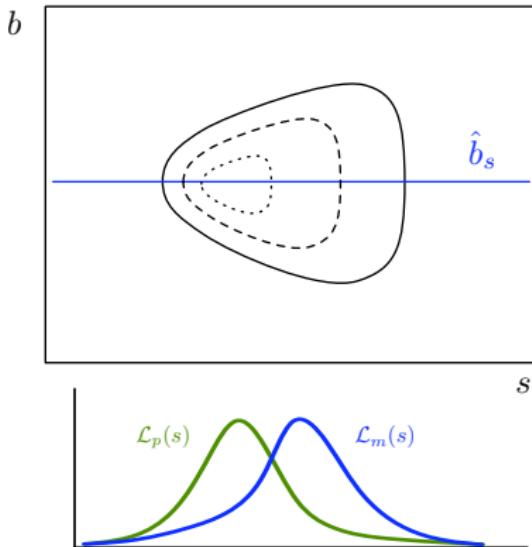
E.g., Gaussians: $\hat{s} = \hat{r} - \hat{b}$, $\sigma_s^2 = \sigma_r^2 + \sigma_b^2$

Background subtraction is a special case of background marginalization.

Bivariate normals: $\mathcal{L}_m \propto \mathcal{L}_p$



Flared/skewed/banana-shaped: \mathcal{L}_m and \mathcal{L}_p differ



General result: For a linear (in params) model sampled with Gaussian noise, and flat priors, $\mathcal{L}_m \propto \mathcal{L}_p$. Otherwise, they will likely differ.

In *measurement error problems* (SCMA lectures!) the difference can be dramatic.

Many Roles for Marginalization

Eliminate nuisance parameters

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

Propagate uncertainty

Model has parameters θ ; what can we infer about $F = f(\theta)$?

$$\begin{aligned} p(F|D, M) &= \int d\theta \ p(F, \theta|D, M) = \int d\theta \ p(\theta|D, M) p(F|\theta, M) \\ &= \int d\theta \ p(\theta|D, M) \delta[F - f(\theta)] \quad [\text{single-valued case}] \end{aligned}$$

Prediction

Given a model with parameters θ and present data D , predict future data D' (e.g., for *experimental design*):

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

Model comparison

Posterior odds for model i vs. model j :

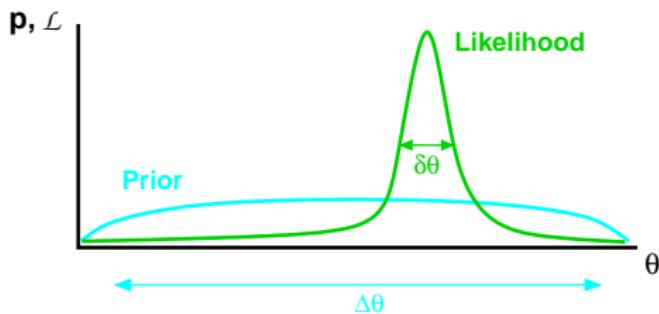
$$\begin{aligned} O_{ij} &\equiv \frac{p(M_i|D, I)}{p(M_j|D, I)} \\ &= \frac{p(M_i|I)}{p(M_j|I)} \times \frac{p(D|M_i, I)}{p(D|M_j, I)} \end{aligned}$$

The data-dependent part is the *Bayes factor*:

$$B_{ij} \equiv \frac{p(D|M_i, I)}{p(D|M_j, I)}$$

It is a *likelihood ratio*; the BF terminology is usually reserved for cases when the likelihoods are *marginal/average likelihoods*:

$$p(D|M_i) = \int d\theta_i p(\theta_i|M_i) p(D|\theta_i, M_i)$$



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

Models with more parameters often make the data more probable — *for the best fit*

Occam factor penalizes models for “wasted” volume of parameter space

Quantifies intuition that models shouldn’t require fine-tuning

Theme: Parameter Space Volume

Bayesian calculations sum/integrate over parameter/hypothesis space!

(Frequentist calculations average over *sample* space & typically *optimize* over parameter space.)

- Credible regions integrate over parameter space
- Marginalization weights the profile likelihood by a volume factor for the nuisance parameters
- Prediction averages parameter-dependent predictive distributions over parameter uncertainties
- Model likelihoods have “Ockham factors” resulting from parameter space volume factors

Many virtues of Bayesian methods can be attributed to this accounting for the “size” of parameter space. This idea does not arise naturally in frequentist statistics (can be added “by hand”).

Overview/Low-Dimensional Models

- ① Bayes recap: Parameter space integrals
- ② Bayesian vs. frequentist computation
- ③ Geometry & probability in high dimensions
- ④ Large N : Laplace approximations
- ⑤ Cubature
- ⑥ Monte Carlo integration
 - Posterior sampling
 - Importance sampling
- ⑦ Bootstrapping vs. posterior sampling

Frequentist Computation

Sample space integrals

Integrate the sampling distribution over D to quantify variability of a procedure. Examples:

Bias of an estimator, $\hat{\theta}(D)$:

$$b(\theta) = \int dD p(D|\theta) [\hat{\theta}(D) - \theta]$$

If $b(\theta) = b$, we can easily make an unbiased estimator.

Coverage of an interval, $\Delta(D)$:

$$C(\theta) = \int dD p(D|\theta) \underbrace{[\theta \in \Delta(D)]}_{\text{Indicator}}$$

If $C(\theta) = C$, the interval is a *strict* confidence interval with confidence level $CL = C$. Otherwise, it is a *conservative* confidence interval with confidence level $CL = \min_{\theta} C(\theta)$

A major theoretical focus is finding good procedures with properties independent of the (*unknown!*) parameters

“Plug-in” approximation

Report properties of procedure for $\theta = \hat{\theta}$

Can be *asymptotically* accurate: for large N , expect $\hat{\theta} \rightarrow \theta$
(under regularity conditions)

Inference with independent data

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

Suppose $p(D|\theta) = p(x_1|\theta) p(x_2|\theta) \cdots p(x_N|\theta)$

Analytically: If the statistics depend on sums of the data,
characteristic functions (Fourier transforms) are helpful; can
evaluate N -D integral with 1-D integrals

Numerically: Independence makes computations easy *in the*
plug-in approximation, using Monte Carlo simulation of data

Confidence Interval for a Normal Mean

Suppose we have a sample of $N = 5$ values x_i ,

$$x_i \sim N(\mu, 1)$$

We want to estimate μ , including some *quantification of uncertainty* in the estimate: an interval *with a probability attached*

Frequentist approaches: method of moments, BLUE, least-squares/ χ^2 , maximum likelihood

Focus on likelihood (equivalent to χ^2 here); this is closest to Bayes:

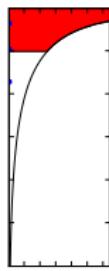
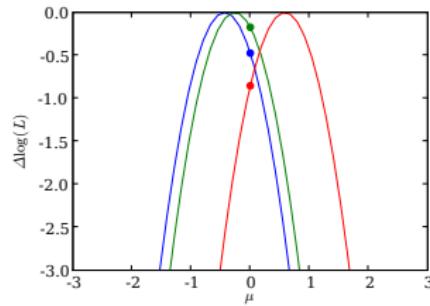
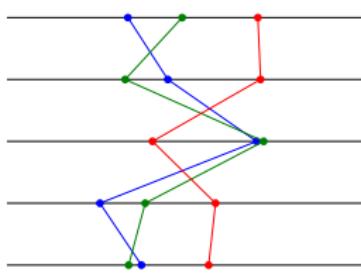
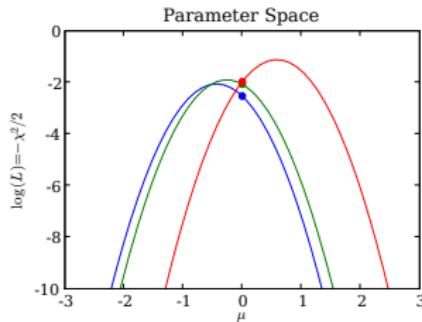
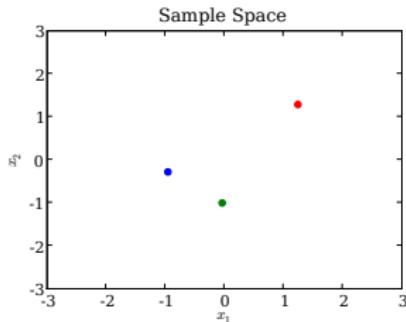
$$\begin{aligned}\mathcal{L}(\mu) &= p(\{x_i\}|\mu) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}; \quad \sigma = 1 \\ &\propto e^{-\chi^2(\mu)/2}\end{aligned}$$

Estimate μ from maximum likelihood (minimum χ^2)

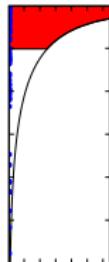
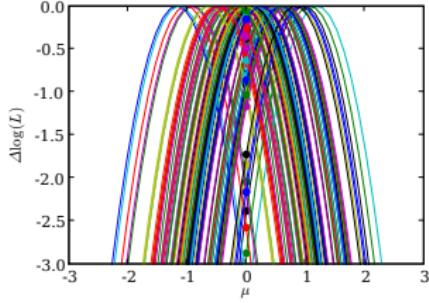
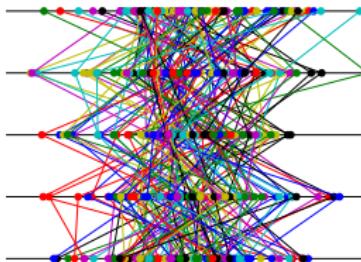
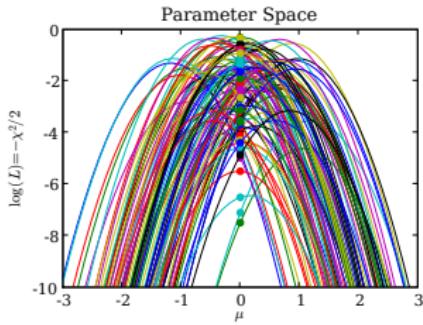
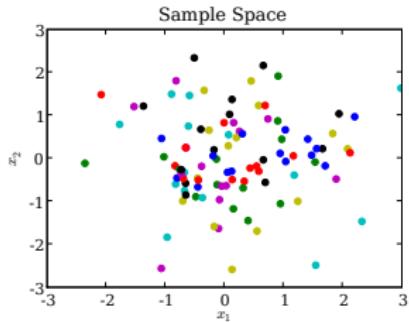
Define an interval and its coverage frequency from the $\mathcal{L}(\mu)$ curve

Construct an Interval Procedure for Known μ

Likelihoods for 3 simulated data sets, $\mu = 0$

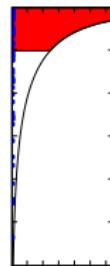
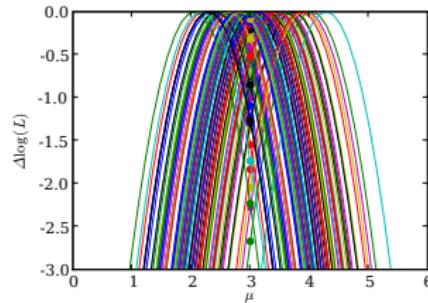
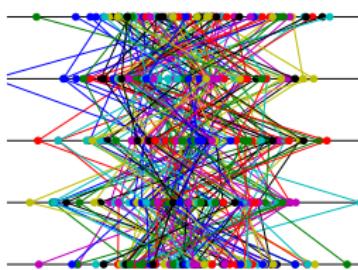
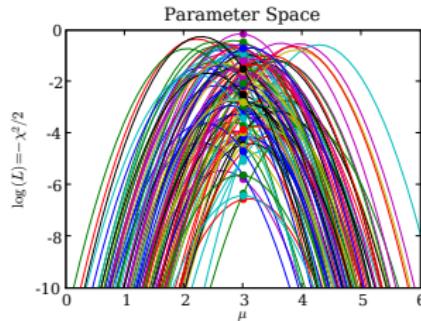
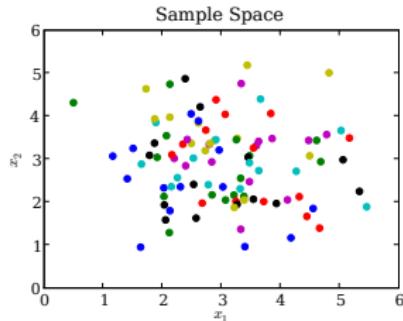


Likelihoods for 100 simulated data sets, $\mu = 0$



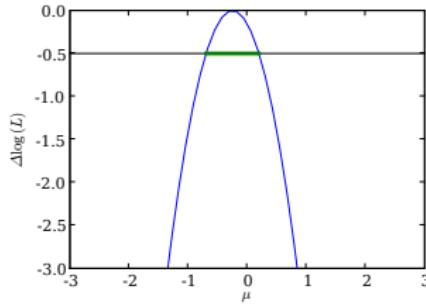
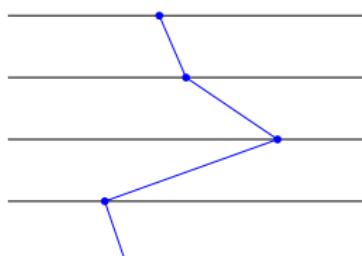
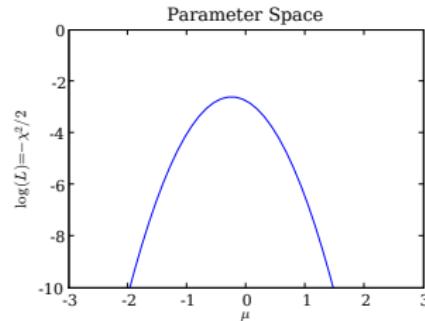
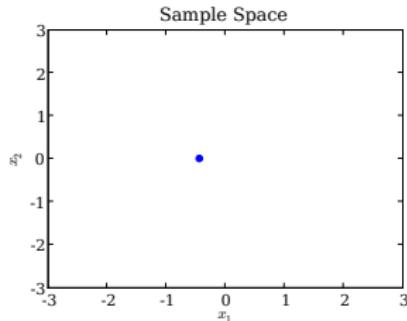
Explore Dependence on μ

Likelihoods for 100 simulated data sets, $\mu = 3$



Luckily the $\Delta \log \mathcal{L}$ distribution is the same!

Apply to Observed Sample



Report the green region, with coverage as calculated for ensemble of hypothetical data (red region, previous slide).

Bayesian Computation

Parameter space integrals

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

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

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

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.

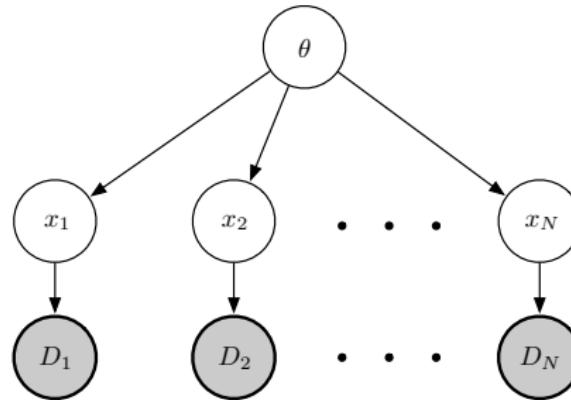
Inference with independent data

Analytically: 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. This is often 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

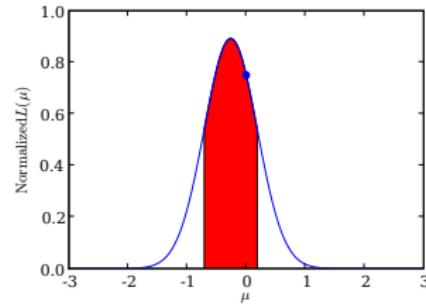
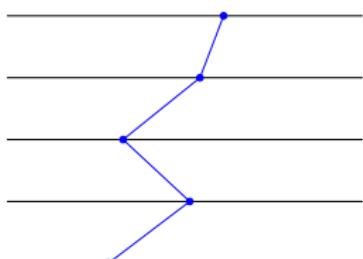
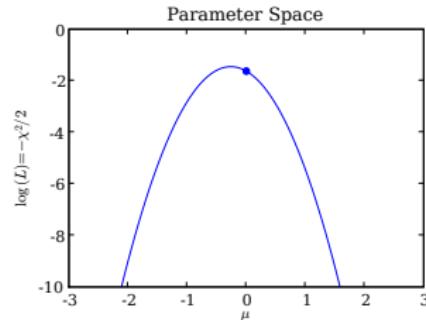
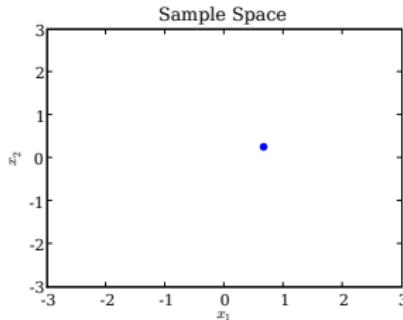


$$\begin{aligned}\mathcal{L}(\theta, \{x_i\}) &\equiv p(\{D_i\} | \theta, \{x_i\}) \\ &= \prod_i p(D_i | x_i) f(x_i | \theta) = \prod_i \ell_i(x_i) f(x_i | \theta)\end{aligned}$$

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

Credible Region for a Normal Mean

Normalize the likelihood for the observed sample; report the region that includes 68.3% of the normalized likelihood.



When They'll Differ

Both approaches report $\mu \in [\bar{x} - \sigma/\sqrt{N}, \bar{x} + \sigma/\sqrt{N}]$, and assign 68.3% to this interval (with different meanings).

This matching is a *coincidence*!

When might results differ? (\mathcal{F} = frequentist, \mathcal{B} = Bayes)

- If \mathcal{F} procedure doesn't use likelihood directly
- If \mathcal{F} procedure properties depend on params (nonlinear models, pivotal quantities)
- If \mathcal{F} properties depend on likelihood shape (conditional inference, ancillary statistics, recognizable subsets)
- If there are extra uninteresting parameters (nuisance parameters, corrected profile likelihood, conditional inference)
- If \mathcal{B} uses important prior information

Also, for a different task—comparison of parametric models—the approaches are qualitatively different (significance tests & info criteria vs. Bayes factors)

Bayesian Computation Menu

Large sample size, N : Laplace approximation

- Approximate posterior as multivariate normal $\rightarrow \det(\text{covar})$ 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



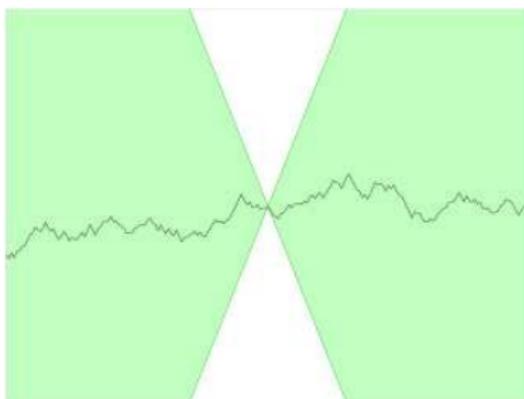
Overview/Low-Dimensional Models

- ① Bayes recap: Parameter space integrals
- ② Bayesian vs. frequentist computation
- ③ Geometry & probability in high dimensions
- ④ Large N : Laplace approximations
- ⑤ Cubature
- ⑥ Monte Carlo integration
 - Posterior sampling
 - Importance sampling
- ⑦ Bootstrapping vs. posterior sampling

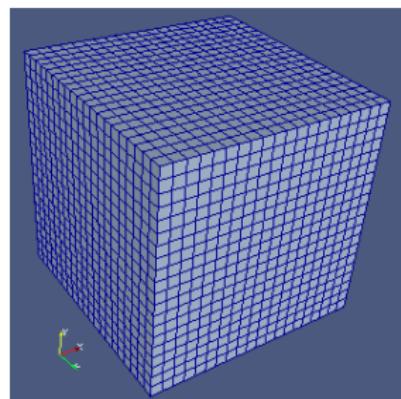
Curse of Dimensionality

Bellman's (1961) phrase concerning exhaustive enumeration on product spaces

Lipschitz-continuous function



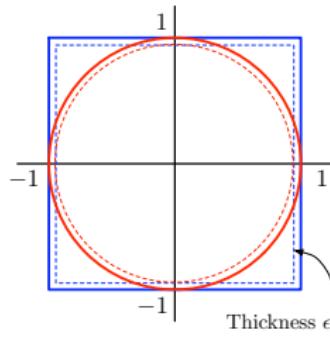
Cartesian grid



Wikipedia

Optimizing, interpolating, or integrating a smooth d -D function to error ϵ requires $O(1/\epsilon^d)$ evaluations.

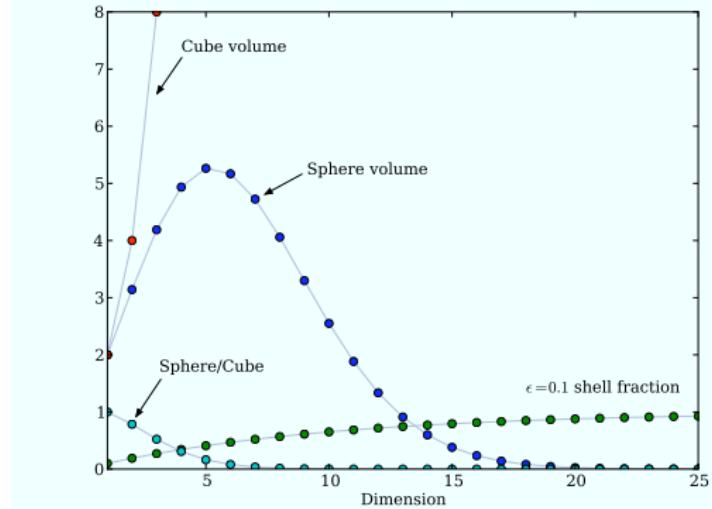
The “Excluded Middle”



$$V_{\square} = 2^d$$

$$V_o = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)}$$

$$\text{Shell fraction} = 1 - (1 - \epsilon)^d$$



- Hi-D unit-radius spheres have volume quickly decreasing with d
- Spherical core of a hypercube has negligible volume
- Volume in a simple d -D region is mostly near the boundary

Uniform Distribution in Hi-D

Consider a large sample of points from $U[0, 1]^d$.

$\langle \# \text{ pts in volume } \delta V \rangle \propto \delta V \rightarrow \text{volume effects map over}$

Empty space phenomenon (Scott & Thompson 1983)

- Most cells in a grid will be empty even for large samples
- Most points are near boundaries: Most points appear extreme/surprising in some respect
- Points are all near a $(d - 1)$ -D manifold
- Spherical neighborhoods of a point will be nearly empty

Concentration of Euclidean norm

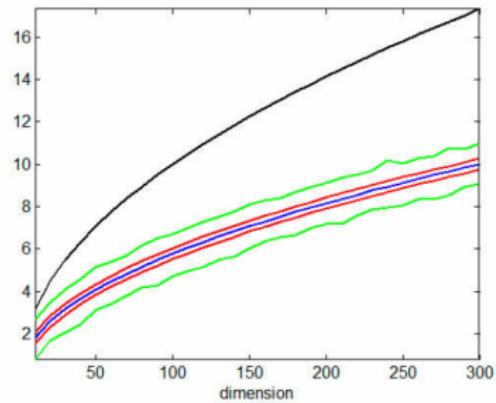
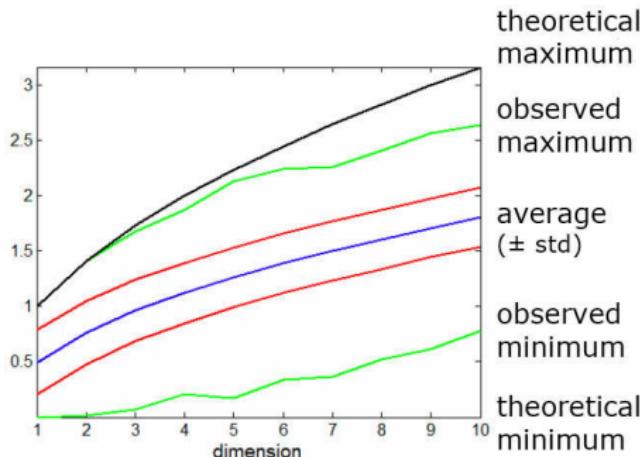
$$r^2 = \sum_{i=1}^d x_i^2; \quad x_i^2 \text{ has mean } 1/3, \text{ variance } 4/45$$

$\approx d \times \text{mean of } d \text{ draws from } N(1/3, 4/45)$

$\sim d \times \text{draw from } N(1/3, 4/(45 \cdot d))$

$\rightarrow r$ concentrates near $\sqrt{d/3}$ with *constant* variance

Average norms of 10^4 draws from $U[0, 1]^d$



Damien Francois (2005)

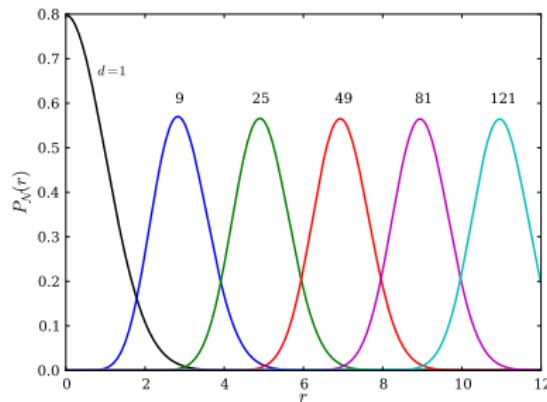
Standard Normal Distribution in Hi-D

Normal distribution has infinite range, with high-density region is localized near origin

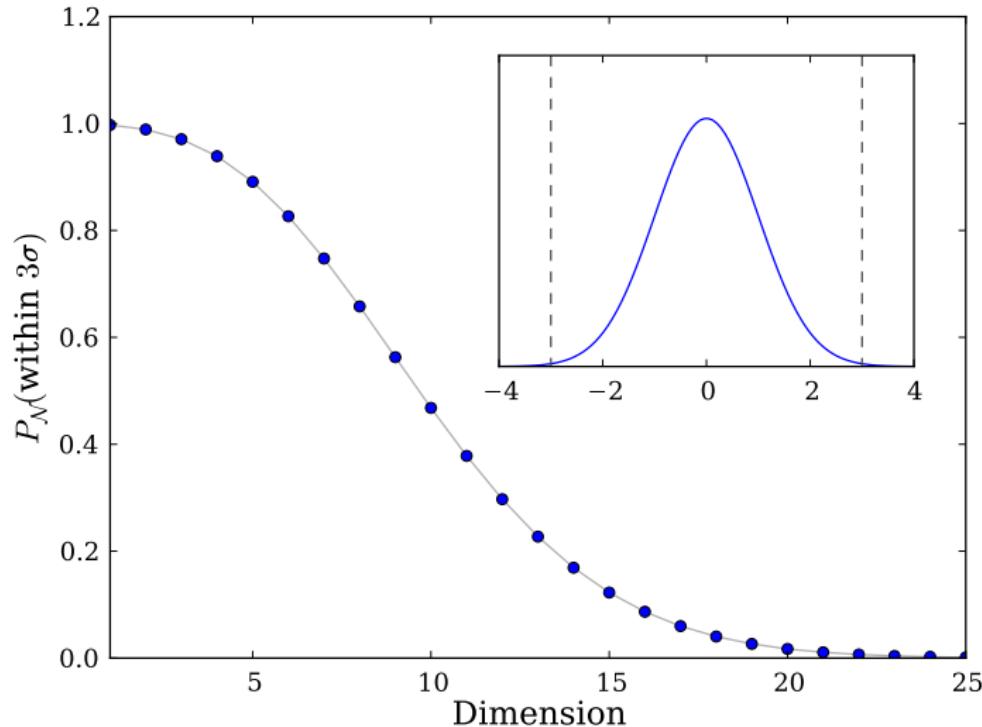
Basic facts:

- Squared radius $\sum_{i=1}^d x_i^2$ is χ_d^2
- $\langle \chi_d^2 \rangle = d$; std dev'n = $\sqrt{2d}$

Most points are in thin shells

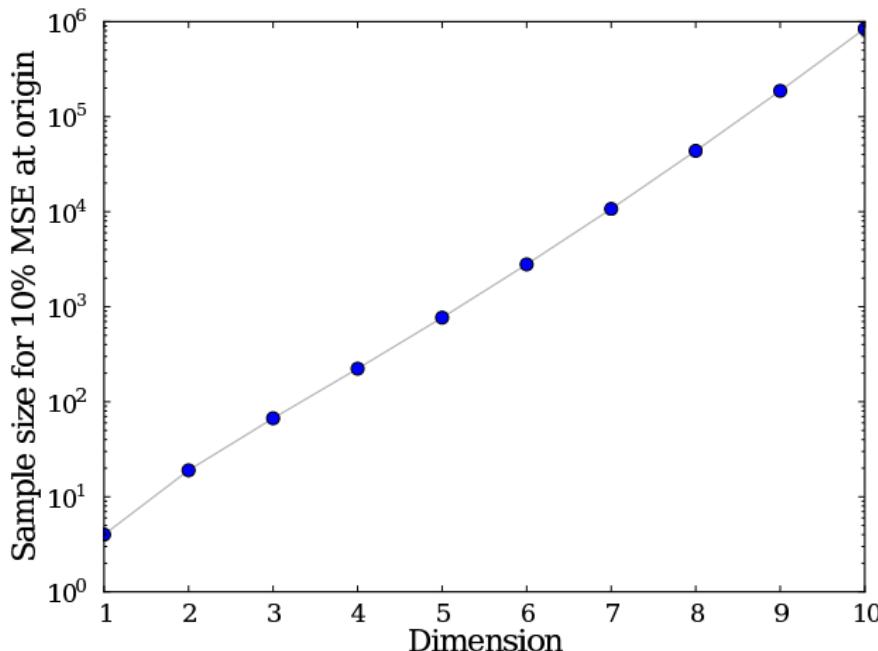


Most points are in the tails



Curse of Dimensionality for KDE

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



Silverman (1986)

Concentration of Measure

Both the uniform and normal settings exhibited Gaussian-like concentration into small volumes.

How generic is this? *Very!*

For random vector with IID components with 8 finite moments:

$$\begin{aligned} E(|\vec{x}|) &= \sqrt{ad - b} + O(1/d) \\ Var(|\vec{x}|) &= b + O(1/\sqrt{d}) \end{aligned}$$

Constants a, b depend on 1st 4 moments

- Norm grows like \sqrt{d} but variance \approx const.
- If you contain a region of the space with a substantial fraction of probability, a small expansion includes nearly all of it
- Smooth functions of d random variables become approximately constant for large d

Overview/Low-Dimensional Models

- ① Bayes recap: Parameter space integrals
- ② Bayesian vs. frequentist computation
- ③ Geometry & probability in high dimensions
- ④ Large N : Laplace approximations
- ⑤ Cubature
- ⑥ Monte Carlo integration
 - Posterior sampling
 - Importance sampling
- ⑦ Bootstrapping vs. posterior sampling

Laplace Approximations

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

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

where $\hat{\mathbf{I}} = -\frac{\partial^2 \ln[\pi(\theta)\mathcal{L}(\theta)]}{\partial^2 \theta}\Bigg|_{\hat{\theta}}$

= Negative Hessian of $\ln[\pi(\theta)\mathcal{L}(\theta)]$

= “Observed Fisher info. matrix” (for flat prior)

\approx Inverse of covariance matrix

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

Marginal likelihoods

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

Marginal posterior densities

$$\begin{aligned} \text{Profile likelihood } \mathcal{L}_p(\phi) &\equiv \max_{\eta} \mathcal{L}(\phi, \eta) = \mathcal{L}(\phi, \hat{\eta}(\phi)) \\ \rightarrow p(\phi | D, M) &\propto \pi(\phi, \hat{\eta}(\phi)) \mathcal{L}_p(\phi) |\mathbf{I}_{\eta}(\phi)|^{-1/2} \end{aligned}$$

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

Posterior expectations

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

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., $\theta \rightarrow \log \theta$ to straighten curved 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:

$$\begin{aligned} Z &\equiv \int d\theta \pi(\theta) \mathcal{L}(\theta) \approx \pi(\hat{\theta}) \mathcal{L}(\hat{\theta}) (2\pi)^{m/2} |\mathbf{I}|^{-1/2} \\ &\sim \pi(\hat{\theta}) \mathcal{L}(\hat{\theta}) (2\pi)^{m/2} \prod_{k=1}^m \sigma_{\theta_k} \end{aligned}$$

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

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 (consistent!)

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

Overview/Low-Dimensional Models

- ① Bayes recap: Parameter space integrals
- ② Bayesian vs. frequentist computation
- ③ Geometry & probability in high dimensions
- ④ Large N : Laplace approximations
- ⑤ Cubature
- ⑥ Monte Carlo integration
 - Posterior sampling
 - Importance sampling
- ⑦ Bootstrapping vs. posterior sampling

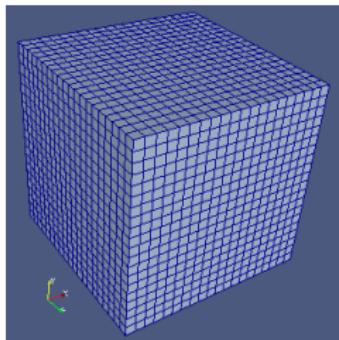
Modest-D: Quadrature & Cubature

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

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

Smoothness \rightarrow fast convergence in 1-D

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



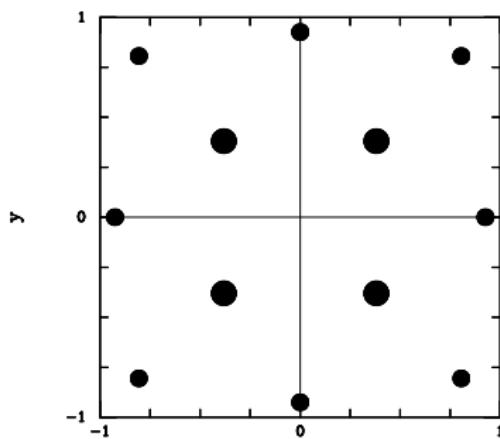
Monomial Cubature Rules

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

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

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

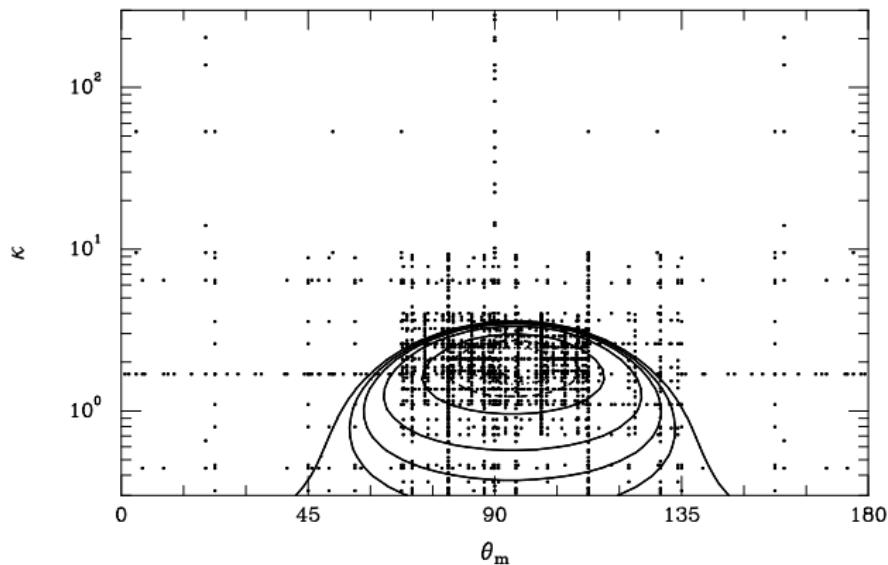
- Subregion adaptive cubature: Use a pair of monomial rules (for error estim'n); recursively subdivide regions w/ large error (ADAPT, DCUHRE, BAYESPACK, CUBA). Concentrates points where most of the probability lies.
- Adaptive grid adjustment: Naylor-Smith method
Iteratively update abscissas and weights to make the (unimodal) posterior approach the weight function.

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

$$\begin{aligned} & \text{\# nodes used by ADAPT's 7th order rule} \\ & 2^d + 2d^2 + 2d + 1 \end{aligned}$$

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

Analysis of Galaxy Polarizations



Overview/Low-Dimensional Models

- ① Bayes recap: Parameter space integrals
- ② Bayesian vs. frequentist computation
- ③ Geometry & probability in high dimensions
- ④ Large N : Laplace approximations
- ⑤ Cubature
- ⑥ Monte Carlo integration
 - Posterior sampling
 - Importance sampling
- ⑦ Bootstrapping vs. posterior sampling

Monte Carlo Integration

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

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

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 → 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

→ *IID* Monte Carlo can be hard if dimension $\gtrsim 5\text{--}10$*

*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?

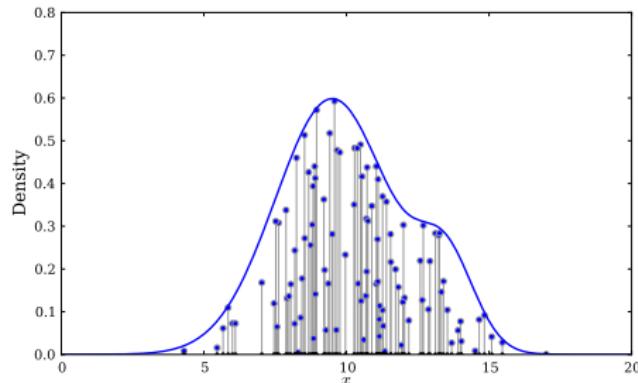
Accept-Reject Algorithm

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)*.

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



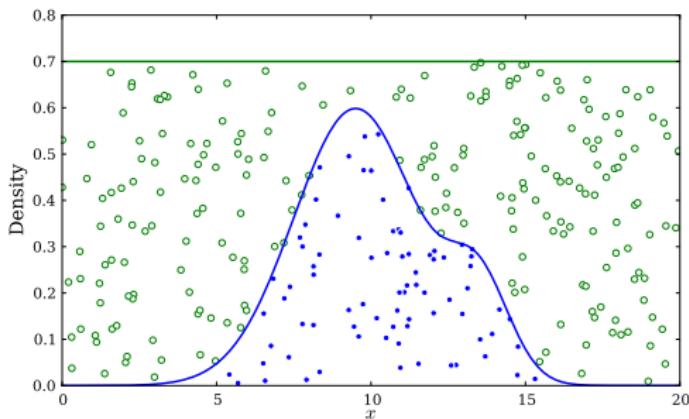
The fraction of samples with θ ("x" in the fig) in a bin of size $\delta\theta$ is the fractional area of the bin.

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), $Z/(QV)$.

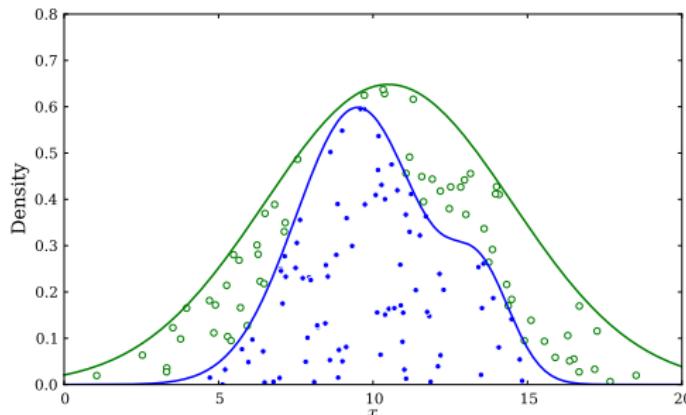
Two issues

- 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.

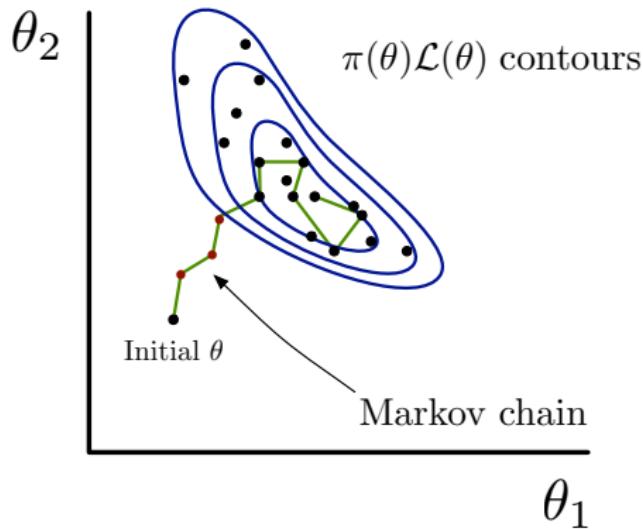
To enable exploration of complex pdfs, let's introduce *dependence*: Choose new θ points in a way that

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

The simplest possibility is a *Markov chain*:

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

A Markov chain “has no memory.”

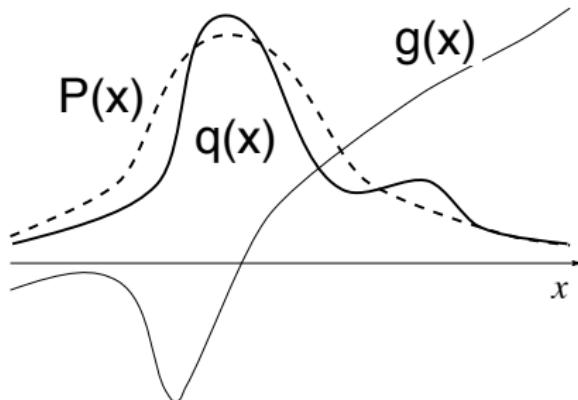


Covered in subsequent tutorials!

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 used for both model comparison (marginal likelihood calculation), and parameter estimation.

Adaptive importance sampling: Build the importance sampler on-the-fly (e.g., VEGAS, miser in *Numerical Recipes*); annealing adaptive importance sampling (Liu⁺ 2011)...

Overview/Low-Dimensional Models

- ① Bayes recap: Parameter space integrals
- ② Bayesian vs. frequentist computation
- ③ Geometry & probability in high dimensions
- ④ Large N : Laplace approximations
- ⑤ Cubature
- ⑥ Monte Carlo integration
 - Posterior sampling
 - Importance sampling
- ⑦ Bootstrapping vs. posterior sampling

Bootstrapping vs. posterior sampling

"Bootstrapping" is a framework that aims to improve simple but approximate frequentist methods:

- *Parametric bootstrap*: Improve asymptotic behavior of estimates for a trusted model: reduce bias of estimates, provide more accurate coverage of confidence regions
- *Nonparametric bootstrap*: Provide results that are approximately accurate with weak modeling assumptions

Most common approach uses Monte Carlo to simulate an ensemble of data sets related to the observed one, and use them to recalibrate a simple method.

Parametric bootstrap has a step producing an ensemble of estimates that looks like a set of posterior samples. Can they be thought of this way?

Coverage and Confidence Intervals

Setup

A distribution with parameters θ produces data D .

θ^* = true value of parameters producing many replicate datasets

D_{obs} = a single, actually observed dataset

Terminology

“Statistic” \equiv Function of data, $f(D)$ (i.e., θ doesn’t appear)

“Interval” \equiv Interval-valued statistic $\Delta(D)$, e.g., for 1-D parameter,

$$\Delta(D) = [l(D), u(D)]$$

Note “interval” refers both to the *statistic* (function), and to a *particular interval*, e.g., $\Delta(D_{\text{obs}})$.

Examples:

- Interval about the mean: $\Delta(D) = [\bar{x} - C, \bar{x} + C]$
- Order-statistic-based interval: $\Delta(D) = [x_{(6)}, x_{(11)}]$

“Coverage” \equiv Fraction of time interval contains θ :

$$C(\theta) = \int dD p(D|\theta) [\![\theta \in \Delta(D)]\!]$$

Monte Carlo algorithm using N simulated datasets:

$$C(\theta) \approx \frac{1}{N} \sum_{D \sim p(D|\theta)} [\![\theta \in \Delta(D)]\!]$$

1. Fix θ at some value; start a counter $n = 0$
2. Simulate a dataset from $p(D|\theta)$
3. Calculate $\Delta(D)$; increment counter if $\theta \in \Delta(D)$
4. Goto (2) for N total iterations
5. Report $C(\theta) = \frac{n}{N}$

In general the coverage *depends on θ* .

‘Plug-In’ Approximation

Problem: We don’t know θ^* (that’s why we’re doing statistics!).

When we report $\Delta(D_{\text{obs}})$, what coverage should we report?

“Confidence level” $CL \equiv$ maximum coverage over all possible values of θ , a conservative promise of coverage

For complex models, calculating $C(\theta)$ across the whole parameter space is prohibitive.

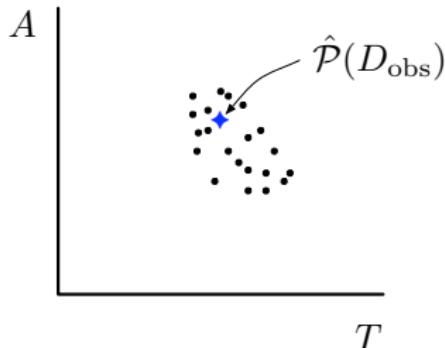
“Plug-in” approach

- Devise some estimator (a statistic!) $\hat{\theta}(D)$ for the parameters; e.g., maximum likelihood
- Calculate $\hat{C} = C(\hat{\theta}(D_{\text{obs}}))$
- Report $\Delta(D_{\text{obs}})$ with $CL \approx \hat{C}$

This gives a *parametric bootstrap* confidence interval; the term is most common when Monte Carlo simulated data sets from $p(D|\hat{\theta}(D_{\text{obs}}))$ is used to estimate \hat{C} .

Incorrect Parametric Bootstrapping

$$\mathcal{P} = (A, T)$$



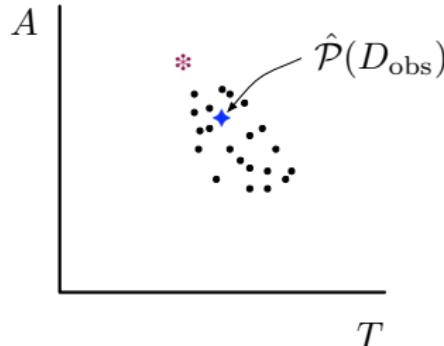
Dots show estimates found by analyzing bootstrapped data sets.

Histograms/contours of best-fit estimates from $D \sim p(D|\hat{\theta}(D_{\text{obs}}))$ provide *poor* confidence regions—no better (possibly worse) than using a least-squares/ χ^2 covariance matrix.

What's wrong with the population of $\hat{\theta}$ points for this purpose?

Incorrect Parametric Bootstrapping

$$\mathcal{P} = (A, T)$$



Dots show estimates found by analyzing bootstrapped data sets.

Histograms/contours of best-fit estimates from $D \sim p(D|\hat{\theta}(D_{\text{obs}}))$ provide *poor* confidence regions—no better (possibly worse) than using a least-squares/ χ^2 covariance matrix.

What's wrong with the population of $\hat{\theta}$ points for this purpose?

The estimates are skewed down and to the right, indicating the truth must be **up** and to the **left**.

Likelihood-Based Parametric Bootstrapping

Key idea: Use likelihood *ratios* to define confidence regions.
I.e., use $L = \ln \mathcal{L}$ or χ^2 differences to define regions.

Estimate parameter values via *maximum likelihood* ($\min \chi^2$)
 $\rightarrow L_{\max}$.

Pick a constant ΔL . Then define an interval by:

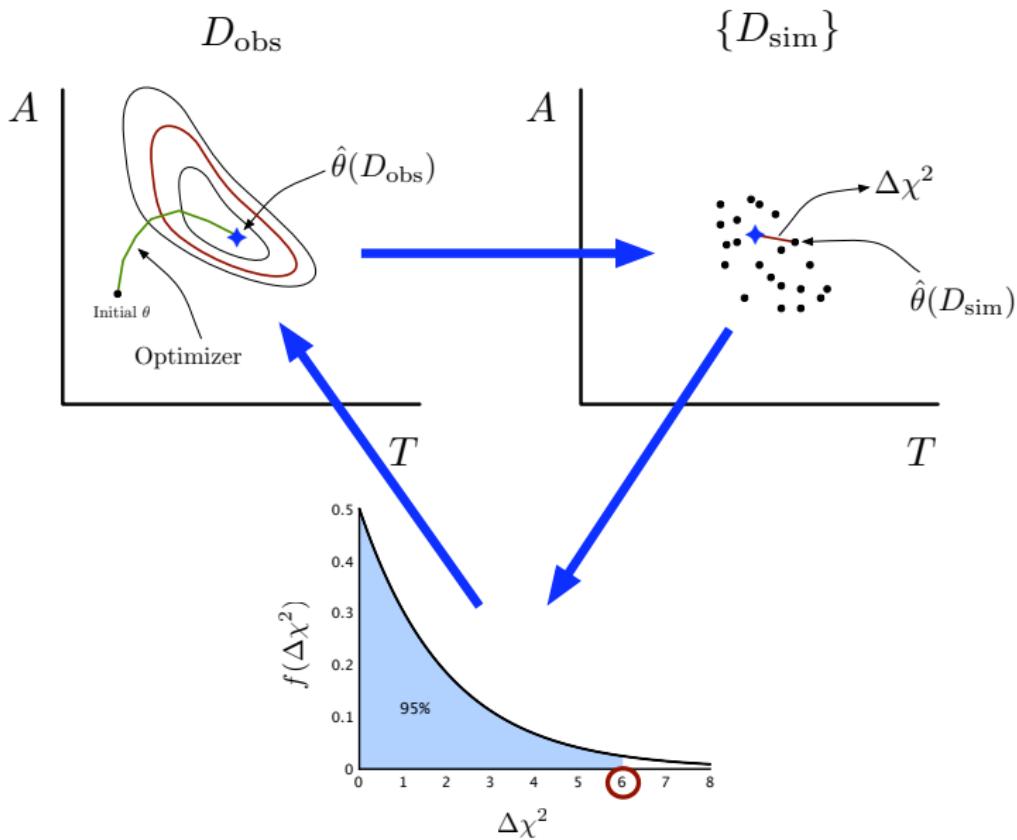
$$\Delta(D) = \{\theta : L(\theta) > L_{\max} - \Delta L\}$$

Coverage calculation:

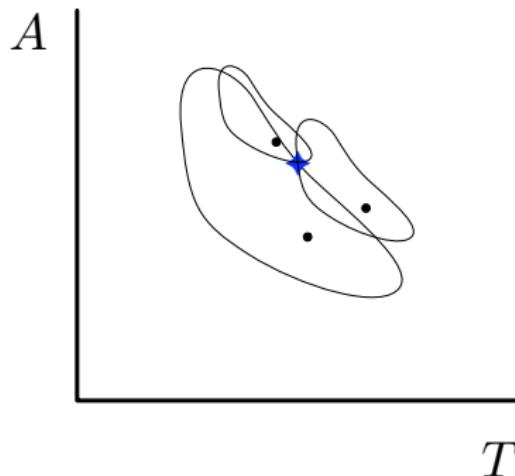
1. Fix $\theta_0 = \hat{\theta}(D_{\text{obs}})$ (plug-in approx'n)
2. Simulate a dataset from $p(D|\theta_0) \rightarrow L_D(\theta)$
3. Find maximum likelihood estimate $\hat{\theta}(D)$
4. Calculate $\Delta L = L_D(\hat{\theta}_D) - L_D(\theta_0)$
5. Goto (2) for N total iterations
6. Histogram the ΔL values to find coverage vs. ΔL
(fraction of sim'sns with smaller ΔL)

Report $\Delta(D_{\text{obs}})$ with ΔL chosen for desired approximate CL.

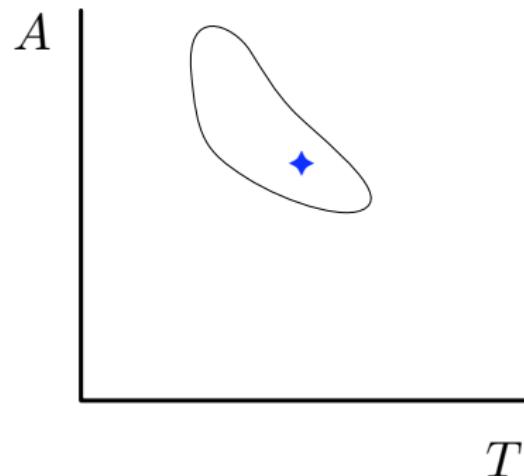
Note that CL is a property of the *function* $\Delta(D)$, not of the particular interval, $\Delta(D_{\text{obs}})$.



ΔL Calibration



Reported Region



The CL is approximate due to:

- Monte Carlo error in calibrating ΔL
- The plug-in approximation

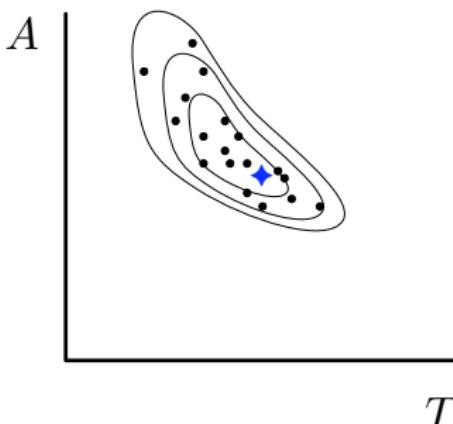
Credible Region Via Posterior Sampling

Monte Carlo algorithm for finding credible regions:

1. Create a RNG that can sample θ from $p(\theta|D_{\text{obs}})$
2. Draw N samples; record θ_i and $q_i = \pi(\theta_i)\mathcal{L}(\mu_i)$
3. Sort the samples by the q_i values
4. An HPD region of probability P is the θ region spanned by the $100P\%$ of samples with highest q_i

Note that no dataset other than D_{obs} is ever considered.

P is a property of the *particular interval* reported.



Rescuing the Bad Bootstrap

Although the best-fit parameters from bootstrapped data don't correspond to posterior samples, they are in the neighborhood of the posterior → use them to create an importance sampling distribution:

- Weighted Likelihood Bootstrap: Nonparametric bootstrap + KDE for modest-dimensional models (Newton & Raftery 1994)
- Efron (2010): Parametric bootstrap for conjugate hierarchical models (simple parameter estimates and importance weights)

Much More to Computational Bayes

Adaptive MCMC

- Single-chain adaptive proposals (use many past states)
- Population-based MCMC (e.g., differential evolution MCMC)

Model uncertainty

- Marginal likelihood computation: Thermodynamic integration, bridge sampling, nested sampling
- MCMC in *model* space: Reversible jump MCMC, birth/death
- Exploration of large (discrete) model spaces (e.g., variable selection): Shotgun stochastic search, Bayesian adaptive sampling

Sequential Monte Carlo

- Particle filters for dynamical models (posterior tracks a changing state)
- Adaptive importance sampling (“evolve” posterior via annealing or on-line processing)

This is just a small sampling!

Markov Chain Monte Carlo

David A. van Dyk

Department of Statistics, University of California, Irvine

SCMA V, June 2011

Outline

1 Background

- Monte Carlo Integration
- Markov Chains

2 Basic MCMC Jumping Rules

- Metropolis Sampler
- Metropolis Hastings Sampler
- Basic Theory

3 Practical Challenges and Advice

- Complex Posterior Distributions
- Choosing a Jumping Rule
- Transformations and Multiple Modes

4 The Gibbs Sampler and Data Augmentation

- The Gibbs Sampler
- Examples and Illustrations of Gibbs
- Data Augmentation

Outline

1 Background

- Monte Carlo Integration
- Markov Chains

2 Basic MCMC Jumping Rules

- Metropolis Sampler
- Metropolis Hastings Sampler
- Basic Theory

3 Practical Challenges and Advice

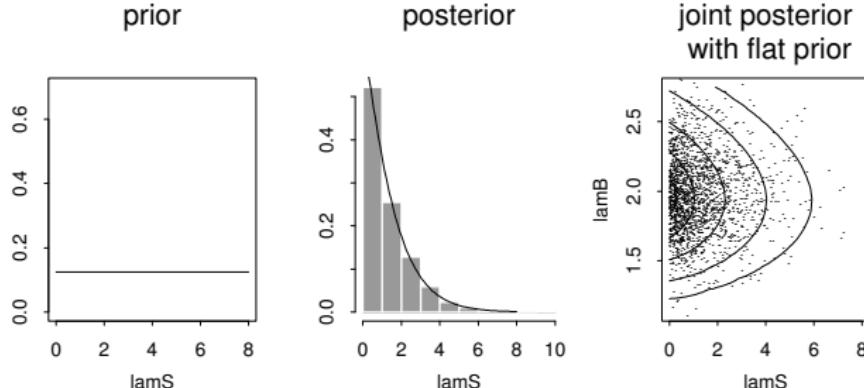
- Complex Posterior Distributions
- Choosing a Jumping Rule
- Transformations and Multiple Modes

4 The Gibbs Sampler and Data Augmentation

- The Gibbs Sampler
- Examples and Illustrations of Gibbs
- Data Augmentation

Simulating from the Posterior

- We can *simulate* or *sample* from a distribution to learn about its contours.
- With the sample alone, we can learn about the posterior.
- Here, $Y \sim \text{Poisson}(\lambda_S + \lambda_B)$ and $Y_B \sim \text{Poisson}(c\lambda_B)$.



Using Simulation to Evaluate Integrals

Suppose we want to compute

$$I = \int g(\theta) f(\theta) d\theta,$$

where $f(\theta)$ is a probability density function.

If we have a sample

$$\theta^{(1)}, \dots, \theta^{(n)} \sim f(\theta),$$

we can estimate I with

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^n g(\theta^{(i)}).$$

In this way we can compute means, variances, and the probabilities of intervals.

We Need to Obtain a Sample

Our primary goal:

Develop methods to obtain a sample from a distribution

- The sample may be independent or dependent.
- Markov chain can be used to obtain a dependent sample.
- In a Bayesian context, we typically aim to sample the *posterior* distribution.

*We first discuss independent methods:
Rejection Sampling & The Grid Method*

Rejection Sampling

Suppose we cannot sample $f(\theta)$ directly, but can find $g(\theta)$ with

$$f(\theta) \leq Mg(\theta)$$

for some M .

- ➊ Sample $\tilde{\theta} \sim g(\theta)$.
- ➋ Sample $u \sim \text{Unif}(0, 1)$.
- ➌ If

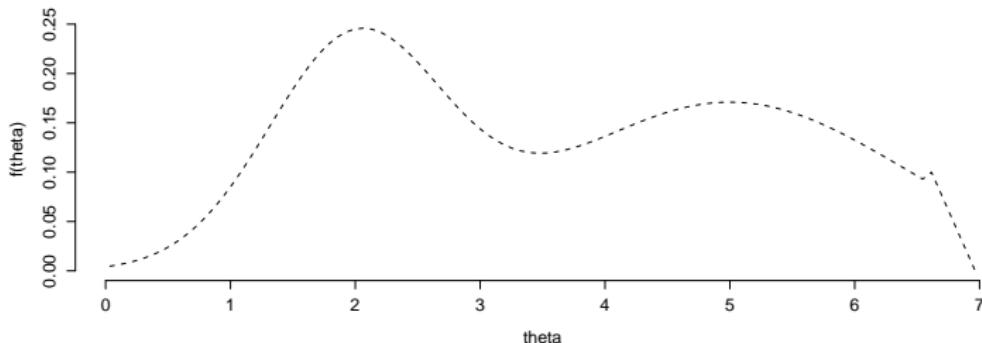
$$u \leq \frac{f(\tilde{\theta})}{Mg(\tilde{\theta})}, \text{ i.e., if } uMg(\tilde{\theta}) \leq f(\tilde{\theta})$$

accept $\tilde{\theta}$: $\theta^{(t)} = \tilde{\theta}$.

Otherwise reject $\tilde{\theta}$ and return to step 1.

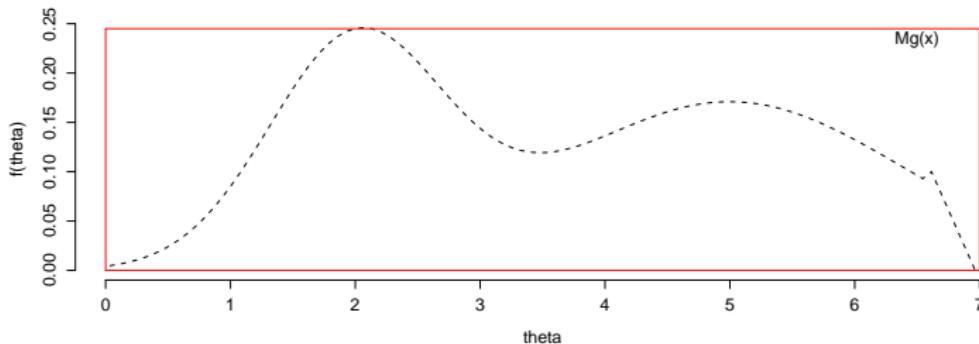
Rejection Sampling

Consider the distribution:



We must bound $f(\theta)$ with some unnormalized density, $Mg(\theta)$.

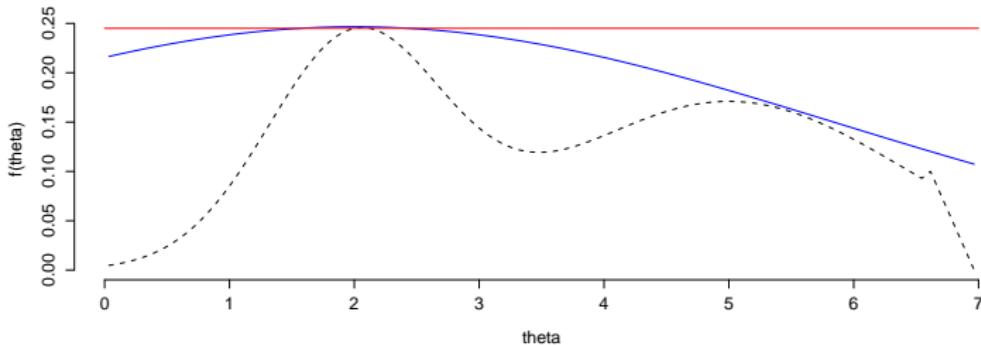
Rejection Sampling



- Imagine that we sample uniformly in the red rectangle:
 $\theta \sim g(\theta)$ and $y = uMg(\theta)$
- Accept samples that fall below the dashed density function.

How can we reduce the wait for acceptance??

Rejection Sampling

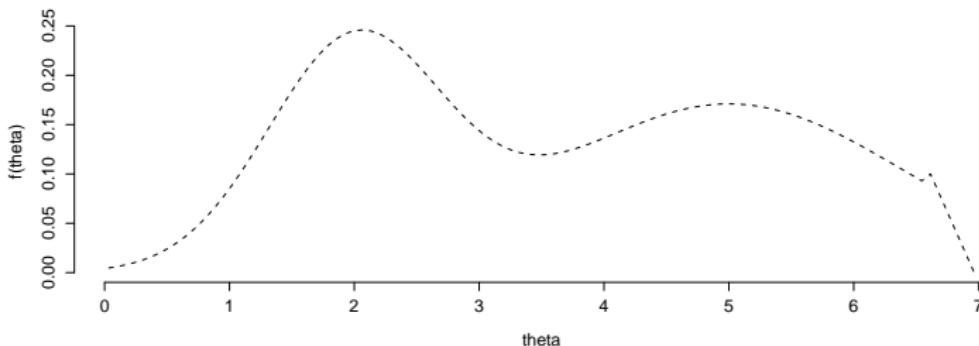


How can we reduce the wait for acceptance??

Improve $g(\theta)$ as an approximation to $f(\theta)$!!

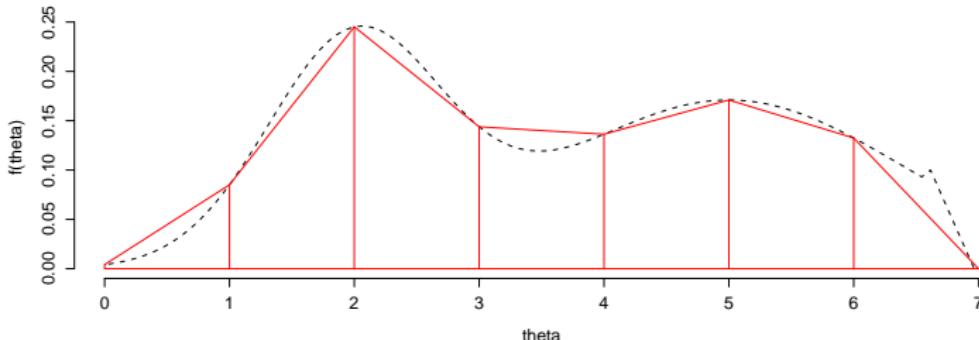
The Grid Method

The Grid method is a brute force / last resort method to sample from a density:



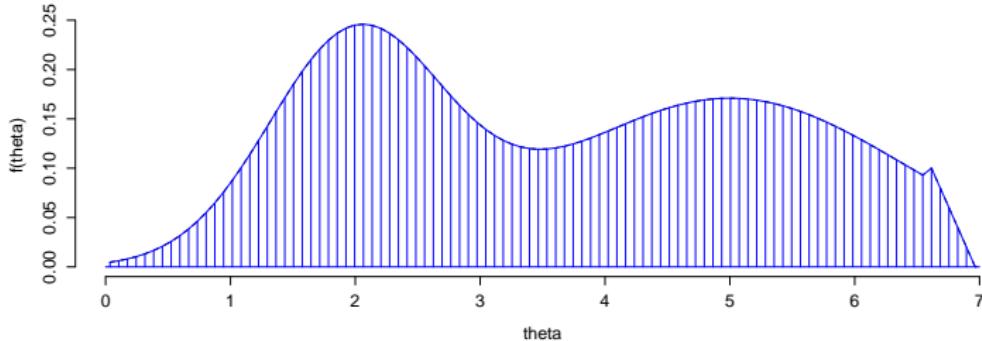
The Grid Method

- ➊ Evaluate the density on a grid.
- ➋ Compute the areas of the resulting trapezoids.
- ➌ Sample from a multinomial distribution with probabilities proportional to the areas.



How can we improve the approximation??

The Grid Method



How can we improve the approximation??

Use a finer grid!!

Limitations?

What is a Markov Chain

A Markov chain is a sequence of random variables,

$$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots$$

such that

$$p(\theta^{(t)} | \theta^{(t-1)}, \theta^{(t-2)}, \dots, \theta^{(0)}) = p(\theta^{(t)} | \theta^{(t-1)}).$$

A Markov chain is generally constructed via

$$\theta^{(t)} = \varphi(\theta^{(t-1)}, U^{(t-1)})$$

with $U^{(t)}$ independent.

What is a Stationary Distribution?

A stationary distribution is any distribution $f(x)$ such that

$$f(\theta^{(t)}) = \int p(\theta^{(t)} | \theta^{(t-1)}) f(\theta^{(t-1)}) d\theta^{(t-1)}$$

If we have a sample from the stationary dist'n and update the Markov chain, the next iterate also follows the stationary dist'n.

What does a Markov Chain at Stationarity Deliver?

Under regularity conditions, the density at iteration t ,

$$f^{(t)}(\theta | \theta^{(0)}) \rightarrow f(\theta) \quad \text{and} \quad \frac{1}{n} \sum_{t=1}^n h(\theta^{(t)}) \rightarrow E_f[h(\theta)]$$

We can treat $\{\theta^{(t)}, t = N_0, \dots, N\}$ as an approximate *correlated* sample from the stationary distribution.

GOAL: Markov Chain with Stationary Dist'n = Target Dist'n.

Outline

1 Background

- Monte Carlo Integration
- Markov Chains

2 Basic MCMC Jumping Rules

- Metropolis Sampler
- Metropolis Hastings Sampler
- Basic Theory

3 Practical Challenges and Advice

- Complex Posterior Distributions
- Choosing a Jumping Rule
- Transformations and Multiple Modes

4 The Gibbs Sampler and Data Augmentation

- The Gibbs Sampler
- Examples and Illustrations of Gibbs
- Data Augmentation

The Metropolis Sampler

Draw $\theta^{(0)}$ from some starting distribution.

For $t = 1, 2, 3, \dots$

Sample: θ^* from $J_t(\theta^* | \theta^{(t-1)})$

Compute: $r = \frac{p(\theta^* | y)}{p(\theta^{(t-1)} | y)}$

Set: $\theta^{(t)} = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{(t-1)} & \text{otherwise} \end{cases}$

Note

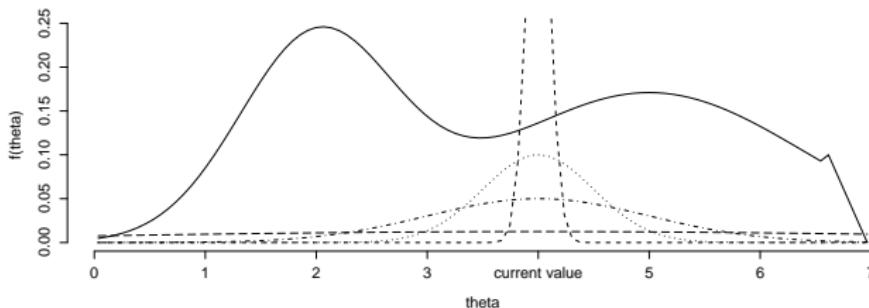
- J_t must be symmetric: $J_t(\theta^* | \theta^{(t-1)}) = J_t(\theta^{(t-1)} | \theta^*)$.
- If $p(\theta^* | y) > p(\theta^{(t-1)} | y)$, jump!

The Random Walk Jumping Rule

Typical choices of $J_t(\theta^* | \theta^{(t-1)})$ include

- Unif $(\theta^{(t-1)} - k, \theta^{(t-1)} + k)$
- Normal $(\theta^{(t-1)}, kl)$
- $t_{df}(\theta^{(t-1)}, kl)$

J_t may change, but may not depend on the history of the chain.



How should we choose k ? Replace l with M ? How?

An Example

A simplified model for high-energy spectral analysis.

- Model:

Consider a perfect detector:

- ① 1000 energy bins, equally spaced from 0.3keV to 7.0keV,
- ② $Y_i \sim \text{Poisson}(\alpha E_i^{-\beta})$, with $\theta = (\alpha, \beta)$,
- ③ E_i is the energy, and
- ④ $(\alpha, \beta) \stackrel{\text{indep.}}{\sim} \text{Unif}(0, 100)$.

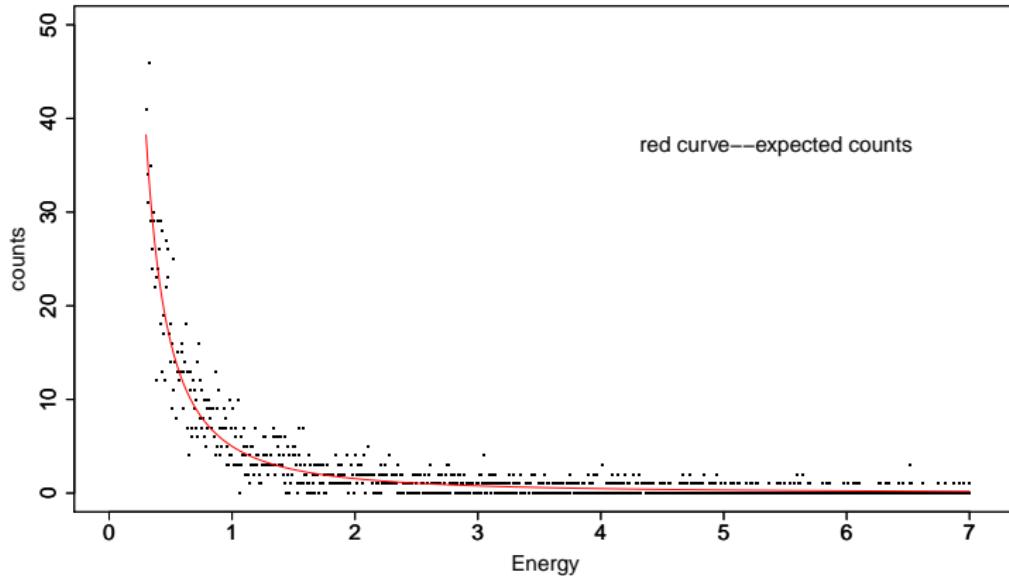
- The Sampler:

We use a Gaussian Jumping Rule,

- centered at the current sample, $\theta^{(t)}$
- with standard deviations equal 0.08 and correlation zero.

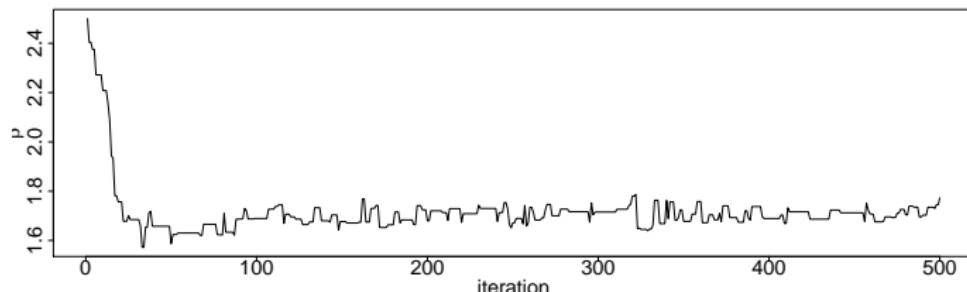
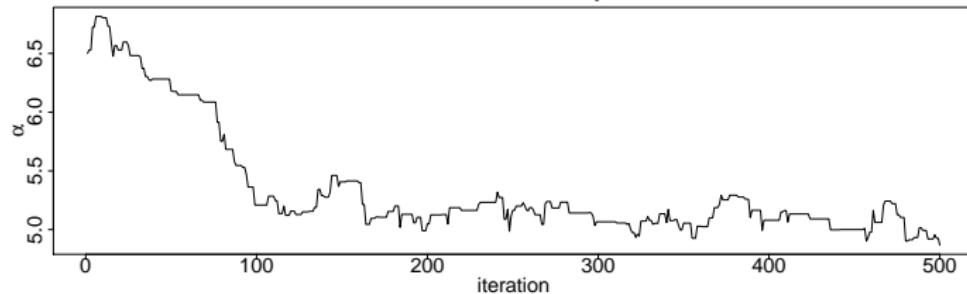
Simulated Data

2288 counts were simulated with $\alpha = 5.0$ and $\beta = 1.69$.



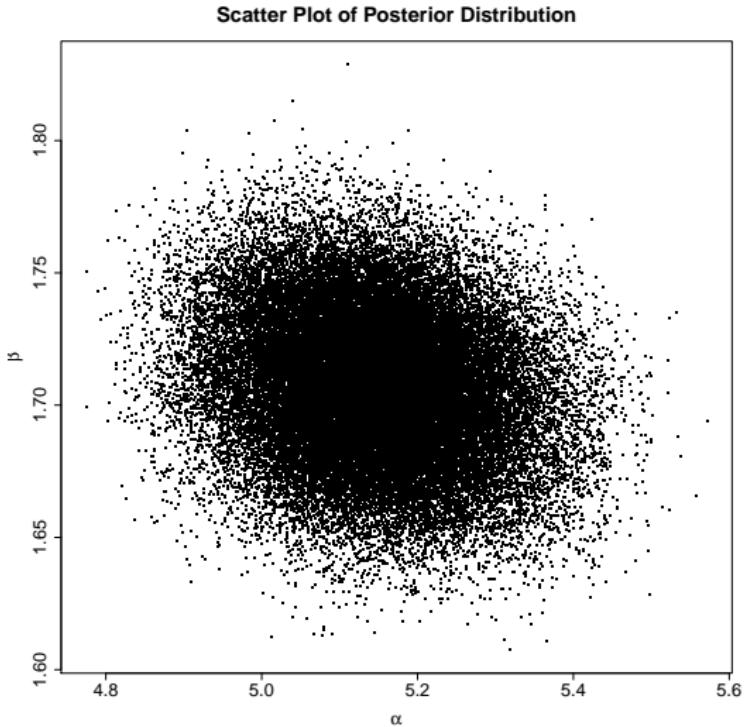
Markov Chain Trace Plots

Time Series Plot for Metropolis Draws

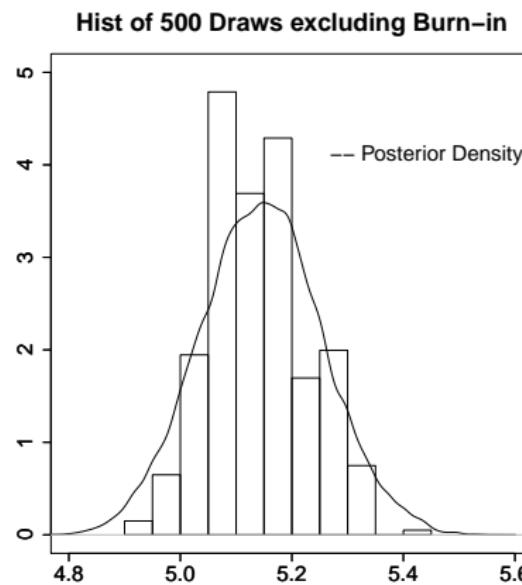
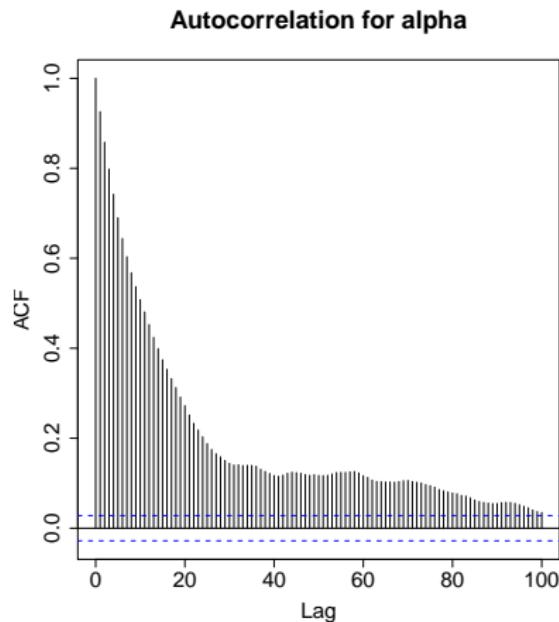


Chains “stick” at a particular draw when proposals are rejected.

The Joint Posterior Distribution

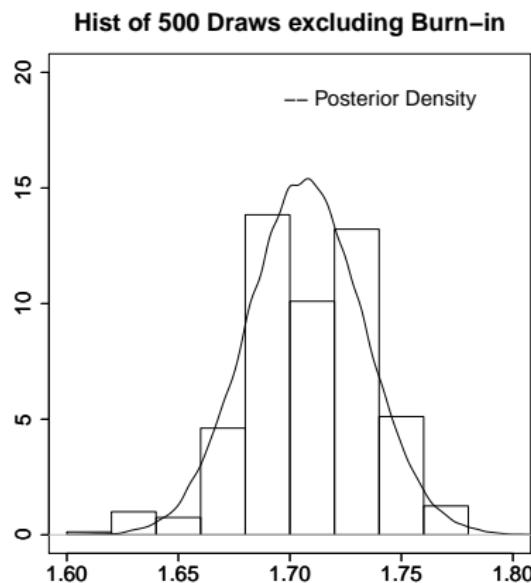
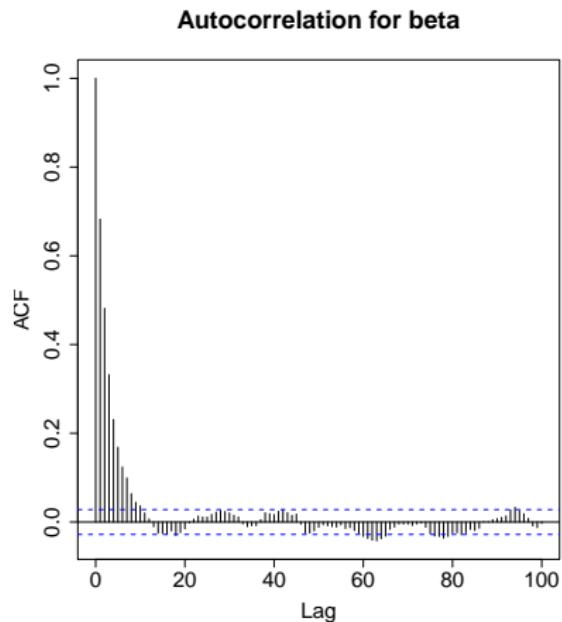


Marginal Posterior Dist'n of the Normalization



$E(\alpha|Y) \approx 5.13$, $SD(\alpha|Y) \approx 0.11$, and a 95% CI is $(4.92, 5.41)$

Marginal Posterior Dist'n of Power Law Param



$E(\beta|Y) \approx 1.71$, $SD(\beta|Y) \approx 0.03$, and a 95% CI is $(1.65, 1.76)$

The Metropolis-Hastings Sampler

A more general Jumping rule:

Draw $\theta^{(0)}$ from some starting distribution.

For $t = 1, 2, 3, \dots$

Sample: θ^* from $J_t(\theta^* | \theta^{(t-1)})$

Compute: $r = \frac{p(\theta^* | y) / J_t(\theta^* | \theta^{(t-1)})}{p(\theta^{(t-1)} | y) / J_t(\theta^{(t-1)} | \theta^*)}$

Set: $\theta^{(t)} = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{(t-1)} & \text{otherwise} \end{cases}$

Note

- J_t may be any jumping rule, it needn't be symmetric.
- The updated r corrects for bias in the jumping rule.

The Independence Sampler

Use an approximation to the posterior as the jumping rule:

$J_t = \text{Normal}_d(\text{MAP estimate}, \text{Curvature-based Variance Matrix}).$

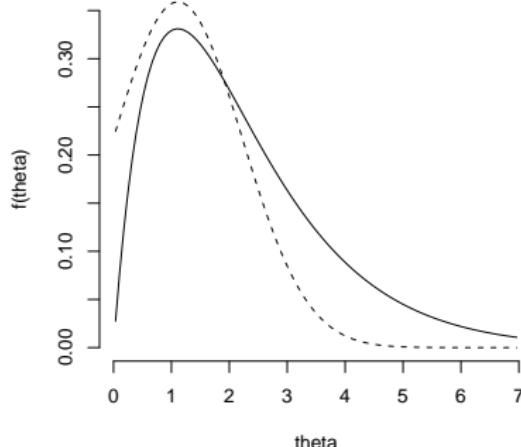
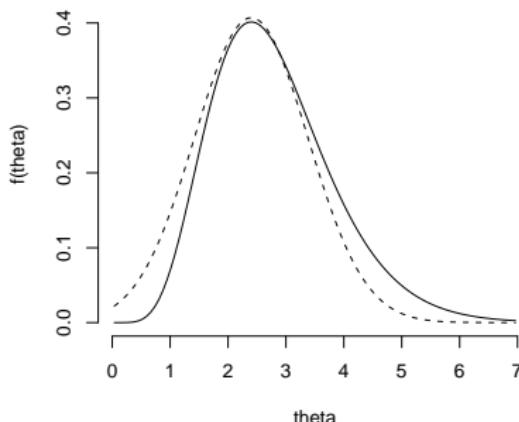
$$\text{MAP estimate} = \operatorname{argmax}_{\theta} p(\theta|y)$$

$$\text{Variance} \approx \left[-\frac{\partial^2}{\partial \theta \cdot \partial \theta} \log p(\theta|Y) \right]^{-1}$$

Note: $J_t(\theta^*|\theta^{(t-1)})$ does not depend on $\theta^{(t-1)}$.

The Independence Sampler

The Normal Approximation may not be adequate.



- We can inflate the variance.
- We can use a heavy tailed distribution, e.g., lorentzian or t .

Example of Independence Sampler

A simplified model for high-energy spectral analysis.

- We use the same model and simulated data.
- This is a simple *loglinear model*,
a special case of a *Generalized Linear Model*:

$$Y_i \sim \text{Poisson}(\lambda_i) \quad \text{with} \quad \log(\lambda_i) = \log(\alpha) - \beta \log(E_i).$$

- The model can be fit with the `glm` function in R:

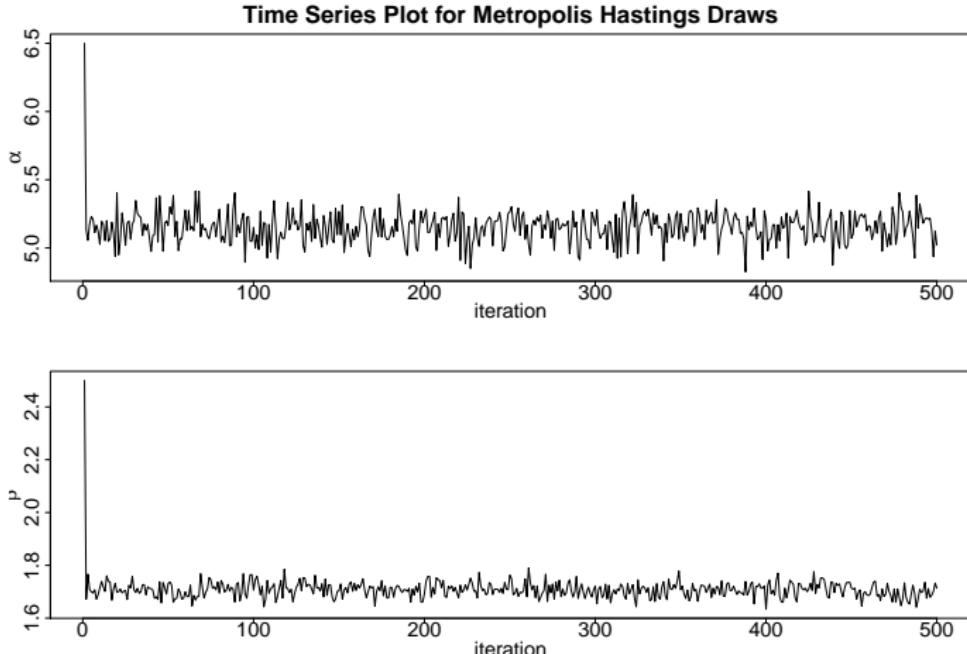
```
> glm.fit = glm( Y~I(-log(E)), family=poisson(link="log") )
> glm.fit$coef      ##### best fit of (log(alpha), beta)
> vcov( glm.fit )  ##### variance-covariance matrix
```

- Returns fit for $(\log(\alpha), \beta)$ and variance-covariance matrix.

Example of Independence Sampler

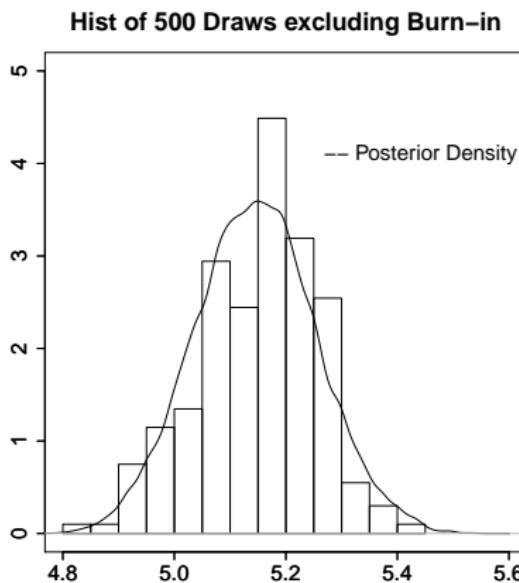
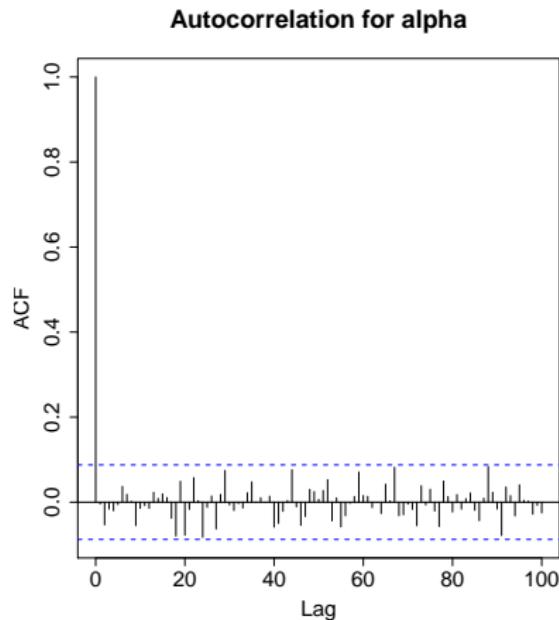
- Alternatively, we can fit (α, β) directly with a general (but less stable) mode finder.
- Requires coding likelihood, specifying starting values, etc.
- Base choice of parameter on quality of normal approx.
 - MLE is invariant to transformations.
 - Variance matrix of transform is computed via *delta method*.
- We use the general mode finder:
 $J_t = \text{Normal}_2(\text{MAP est}, \text{Curvature-based Variance Matrix}).$

Markov Chain Trace Plots



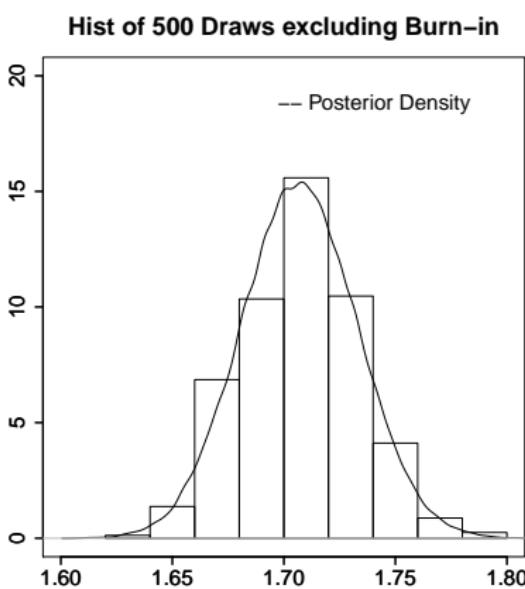
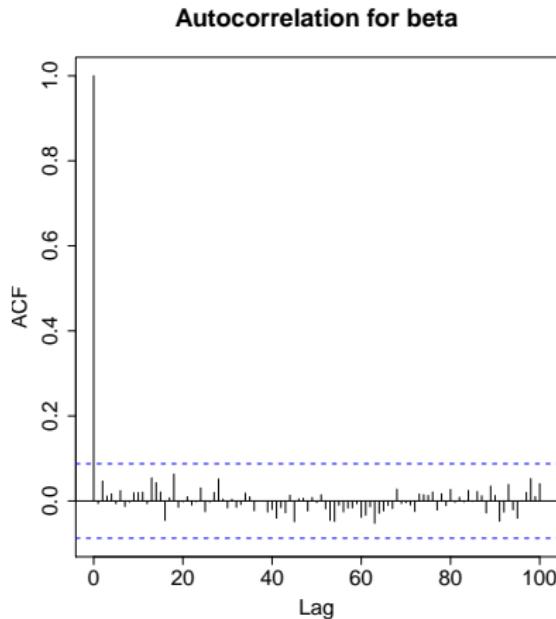
Very little “sticking” here: acceptance rate is 98.8%.

Marginal Posterior Dist'n of the Normalization



Autocorrelation is essentially zero: nearly independent sample!!

Marginal Posterior Dist'n of Power Law Param



This result depends critically on access to a very good approximation to the posterior distribution.

Convergence to Stationarity

Consider a finite state space \mathcal{S} with arbitrary elements i and j .

- Let $p_{ij}(t) = \Pr(\theta^{(t)} = j | \theta^{(0)} = i)$.
- Ergodic Theorem: If a Markov chain is *positive recurrent* and *aperiodic* then its stationary distribution is the unique distribution $\pi()$ such that

$$\sum_i p_{ij}(t)\pi(i) = \pi(j) \text{ for all } j \text{ and } t \geq 0.$$

We say the Markov chain is ergodic and the following hold:

- 1 $p_{ij}(t) \rightarrow \pi(j)$ as $t \rightarrow \infty$ for all i and j .

- 2

$$\Pr \left[\frac{1}{n} \sum_{t=1}^n h(\theta^{(t)}) \rightarrow \mathbb{E}_\pi(h(\theta)) \right] = 1$$

Convergence to Stationarity

Definitions:

- ➊ Chain is *irreducible* if for all i, j there is t with $p_{ij}(t) > 0$.

Let τ_{ii} be the time of first return, $\min\{t > 0 : \theta^{(t)} = i | \theta^{(0)} = i\}$.

- ➋ Chain is *recurrent* if $\Pr[\tau_{ii} < \infty] = 1$ for all i .
- ➌ Chain is *positive recurrent* if $E[\tau_{ii}] < \infty$ for all i .

Fact: Irreducible chain with a stationary dist'n is pos recurrent.

So we need our chain to

- ➊ be irreducible,
- ➋ be aperiodic, and
- ➌ have the posterior distribution as a stationary distribution.

Outline

1 Background

- Monte Carlo Integration
- Markov Chains

2 Basic MCMC Jumping Rules

- Metropolis Sampler
- Metropolis Hastings Sampler
- Basic Theory

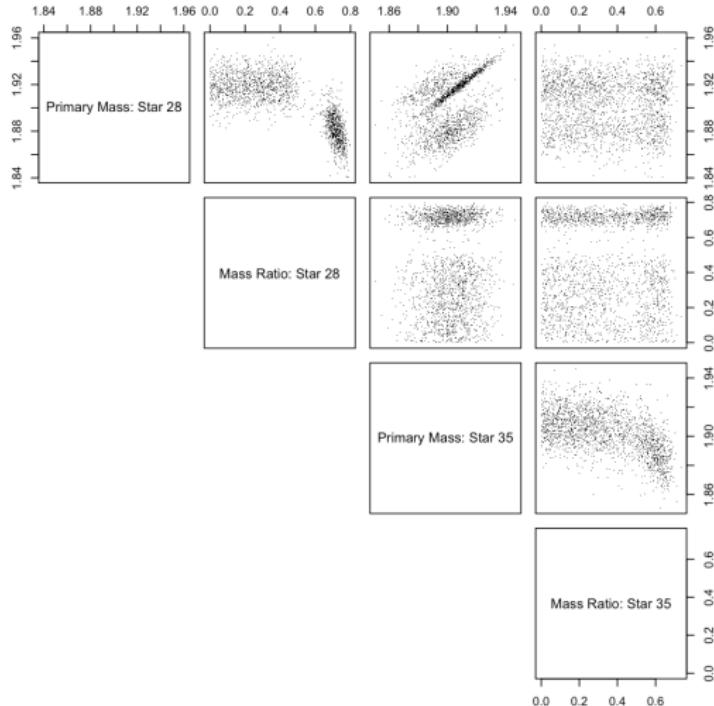
3 Practical Challenges and Advice

- Complex Posterior Distributions
- Choosing a Jumping Rule
- Transformations and Multiple Modes

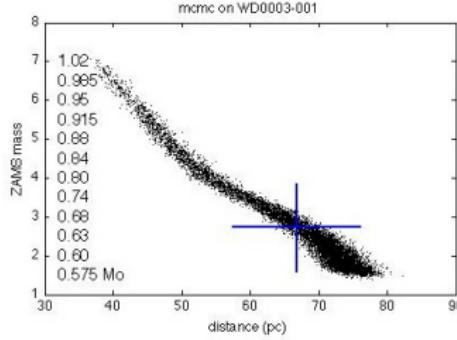
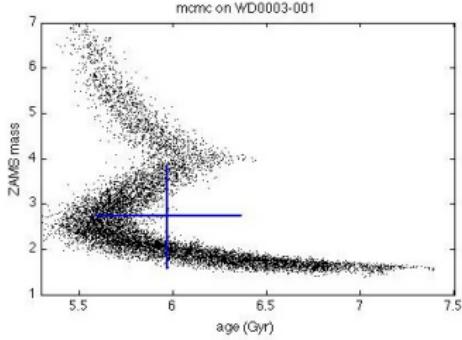
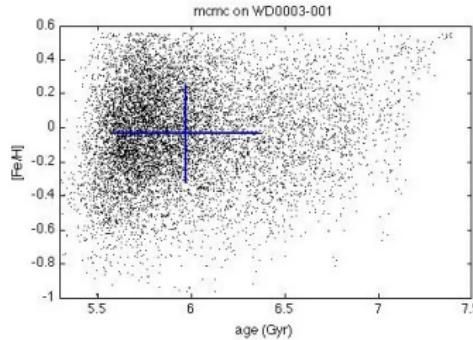
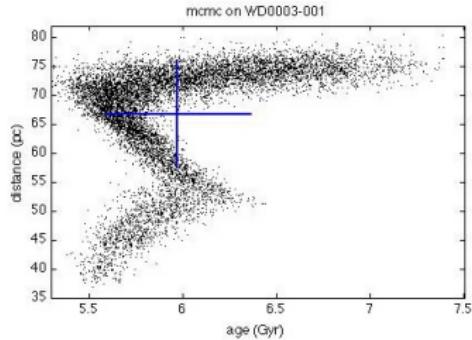
4 The Gibbs Sampler and Data Augmentation

- The Gibbs Sampler
- Examples and Illustrations of Gibbs
- Data Augmentation

Complex Posterior Distributions

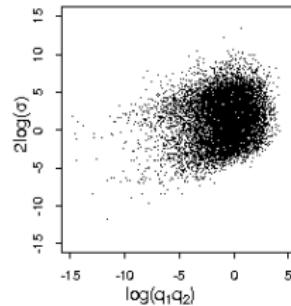
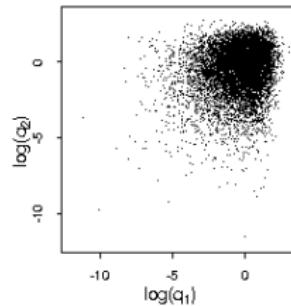
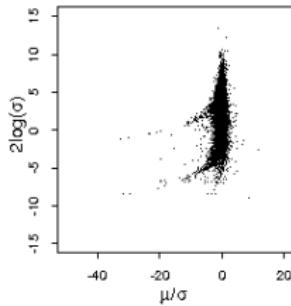


Complex Posterior Distributions

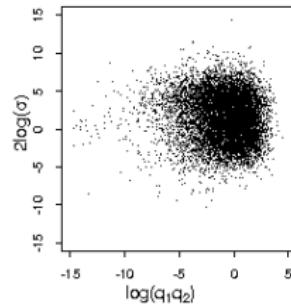
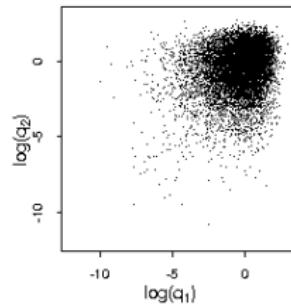
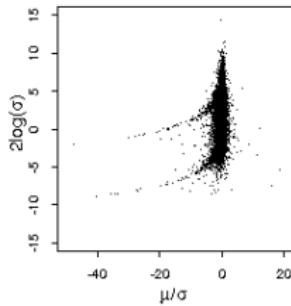


Complex Posterior Distributions

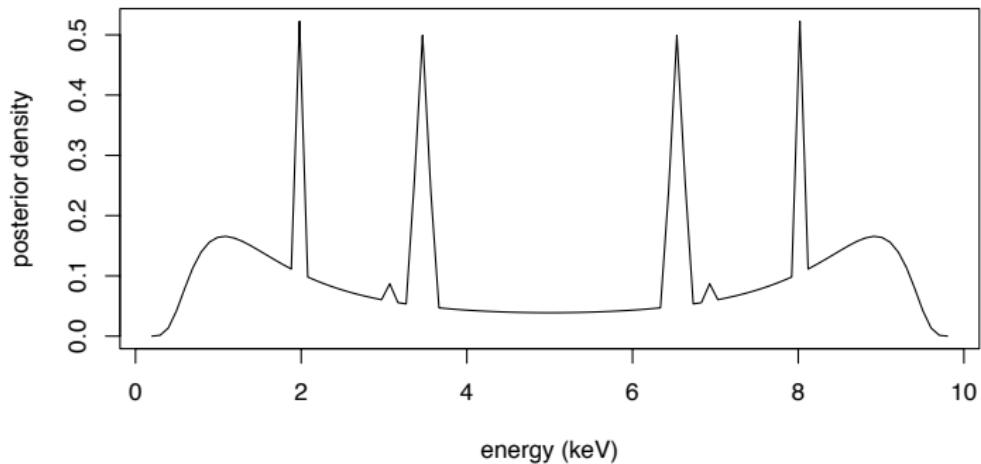
Standard Algorithm
one degree of freedom



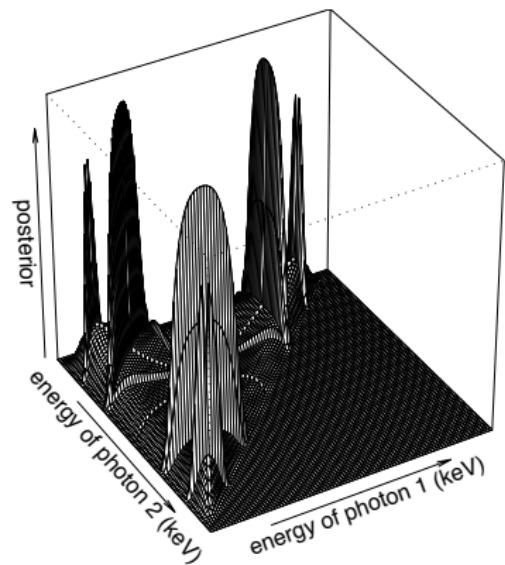
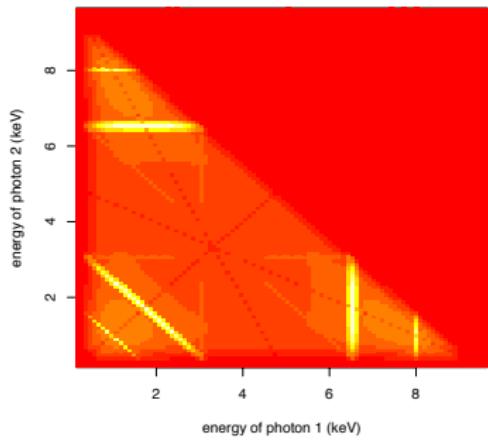
Marginal Augmentation
one degree of freedom



Complex Posterior Distributions

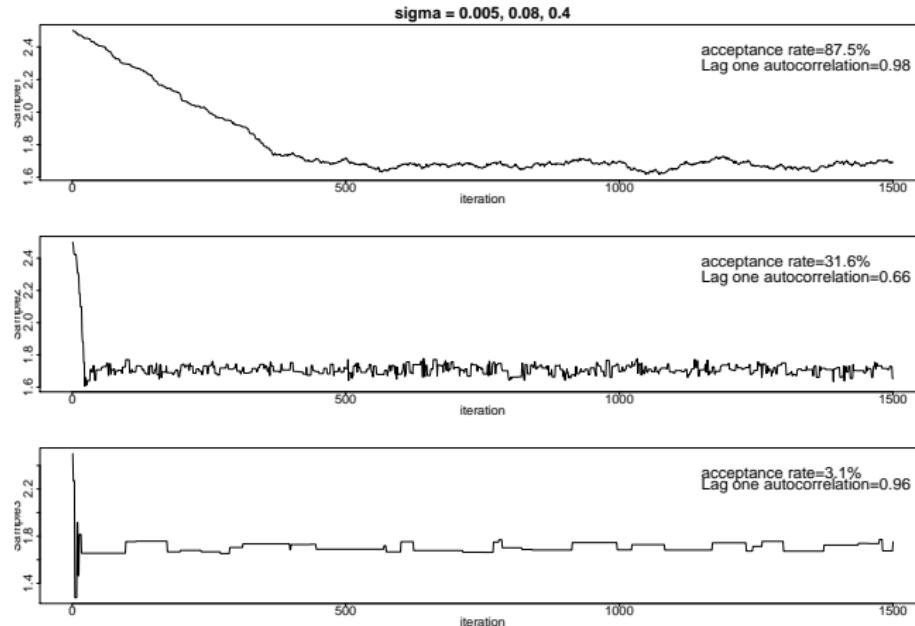


Complex Posterior Distributions

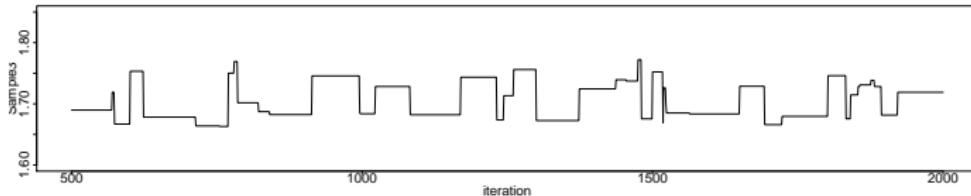
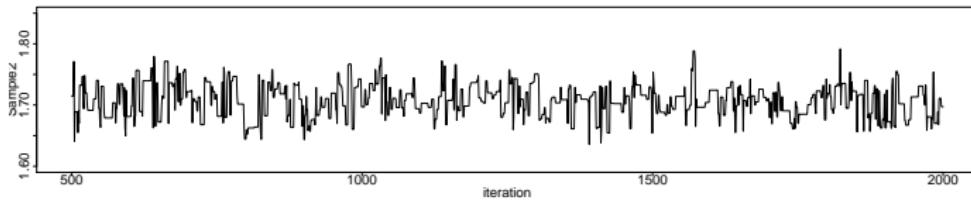
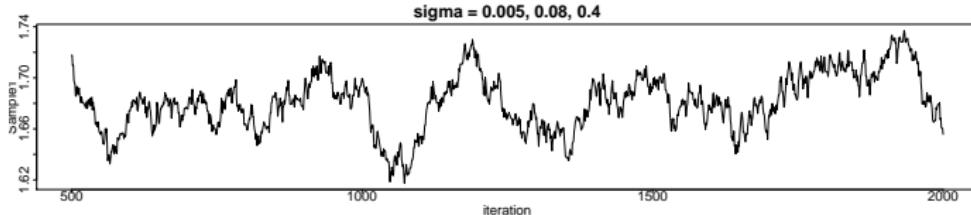


Choice of Jumping Rule with Random Walk Metropolis

Spectral Analysis: effect on burn in of power law parameter



Higher Acceptance Rate is not Always Better!



Aim for 20% (vectors) - 40% (scalars) acceptance rate

Statistical Inference and Effective Sample Size

- Point Estimate: $\bar{h}_n = \frac{1}{n} \sum h(\theta^{(t)})$
- Variance Estimate $\text{Var}(\bar{h}_n) \approx \frac{\sigma^2}{n} \frac{1+\rho}{1-\rho}$ with
 $\sigma^2 = \text{Var}(h(\theta))$ estimated by $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{t=1}^n [h(\theta^{(t)}) - \bar{h}_n]^2$,

$\rho = \text{corr}[h(\theta^{(t)}), h(\theta^{(t-1)})]$ estimated by

$$\hat{\rho} = \frac{1}{n-1} \frac{\sum_{t=2}^n [h(\theta^{(t)}) - \bar{h}_n][h(\theta^{(t-1)}) - \bar{h}_n]}{\sqrt{\sum_{t=1}^{n-1} [h(\theta^{(t)}) - \bar{h}_n]^2 \sum_{t=2}^n [h(\theta^{(t)}) - \bar{h}_n]^2}}$$

- Interval Estimate: $\bar{h}_n \pm t_d \sqrt{\text{Var}(\bar{h}_n)}$ with $d = n \frac{1-\rho}{1+\rho} - 1$
 The *effective sample size* is $n \frac{1-\rho}{1+\rho}$.

Illustration of the Effective Sample Size

Sample from $N(0, 1)$

with random walk Metropolis with $J_t = N(\theta^{(t)}, \sigma)$.

What is the Effective Sample Size here? and σ ?

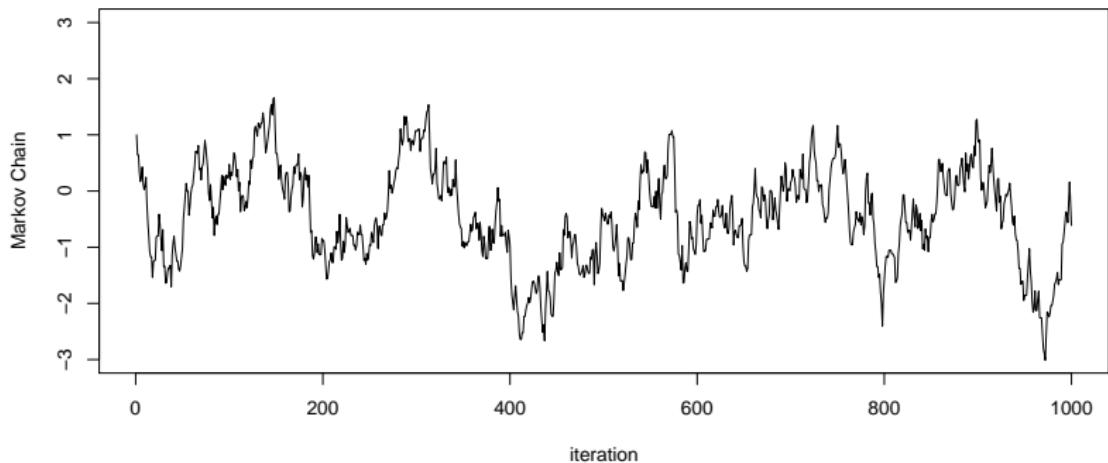


Illustration of the Effective Sample Size

What is the Effective Sample Size here? and σ ?

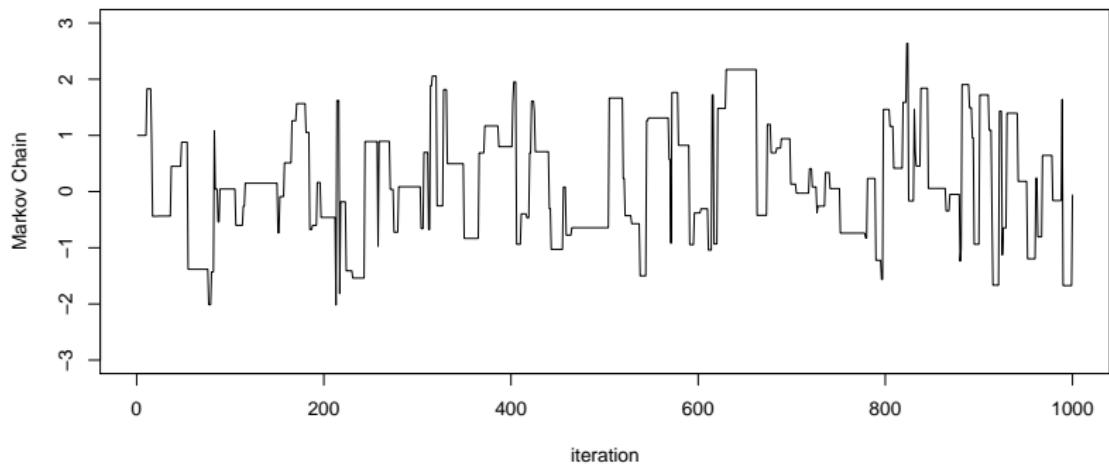


Illustration of the Effective Sample Size

What is the Effective Sample Size here? and σ ?

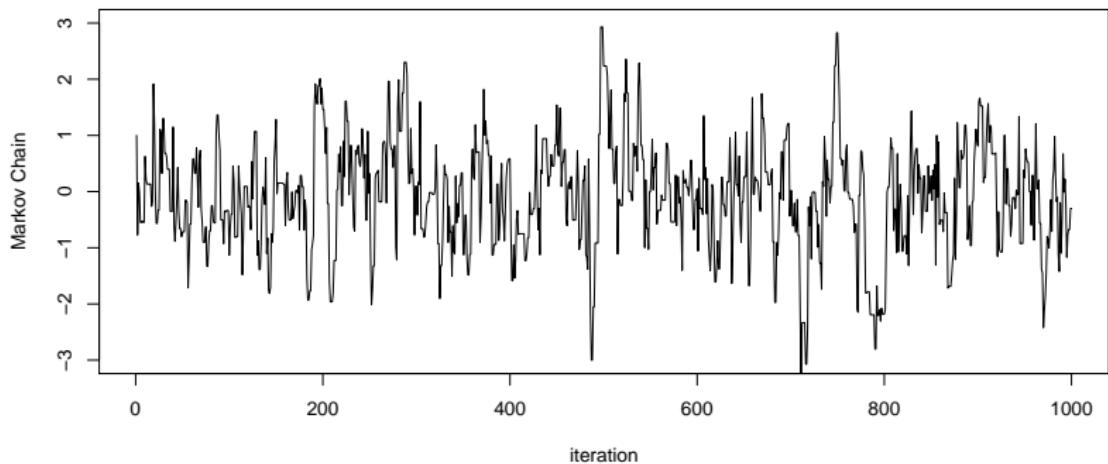
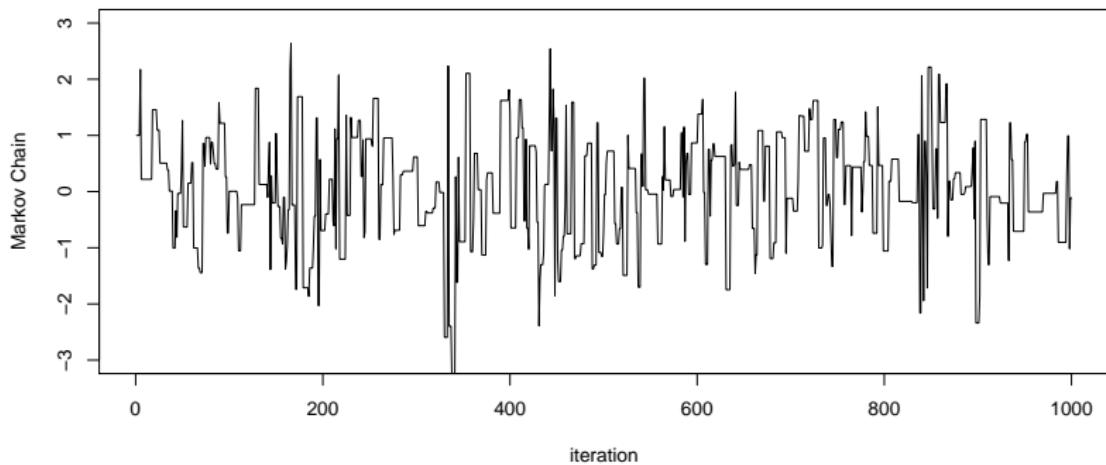


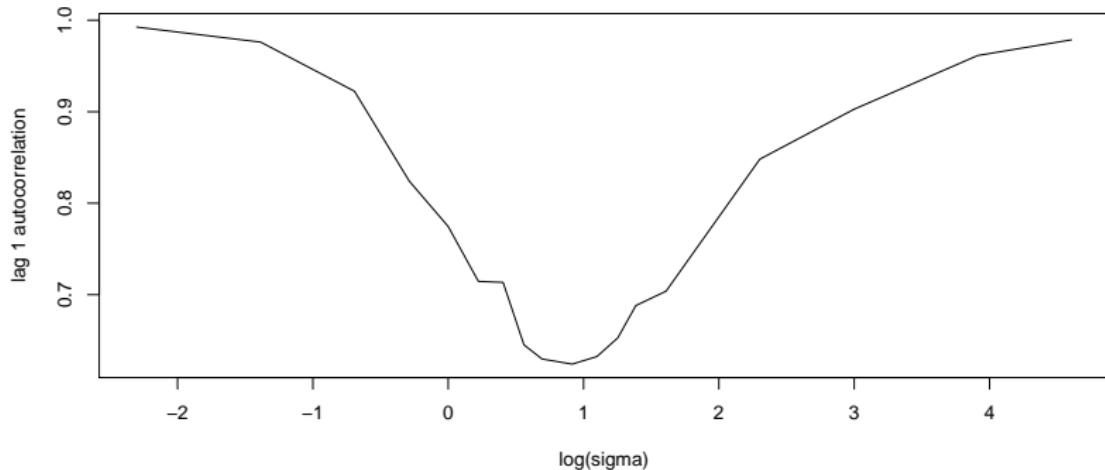
Illustration of the Effective Sample Size

What is the Effective Sample Size here? and σ ?



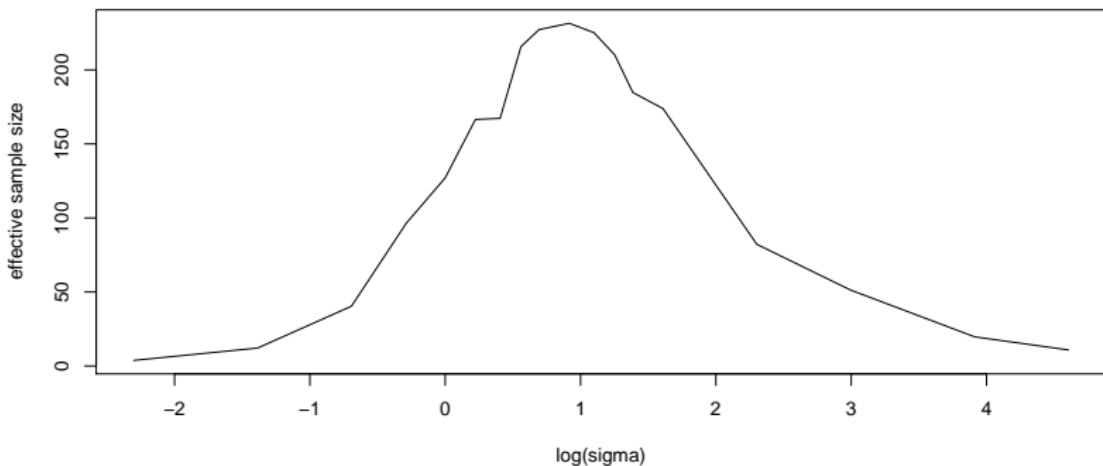
Lag One Autocorrelation

Small Jumps versus Low Acceptance Rates



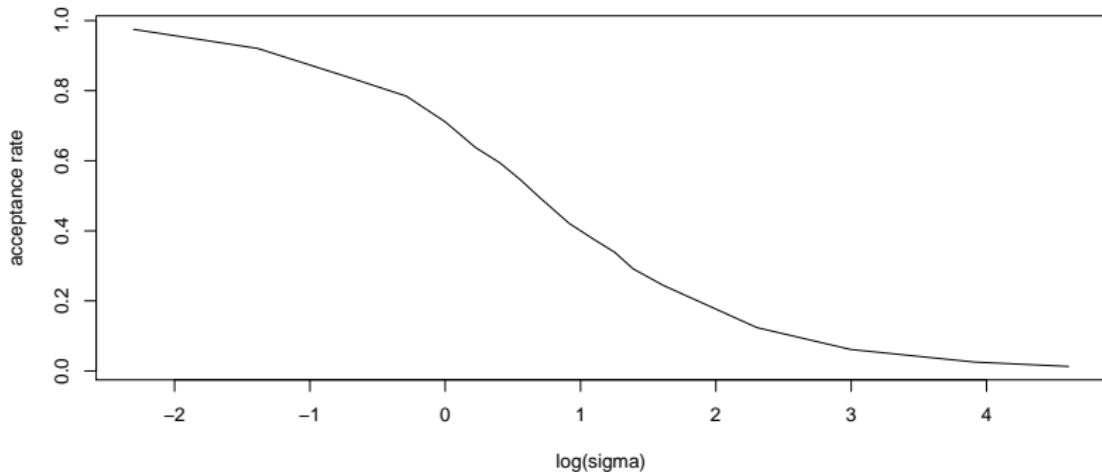
Effective Sample Size

Balancing the Trade-Off

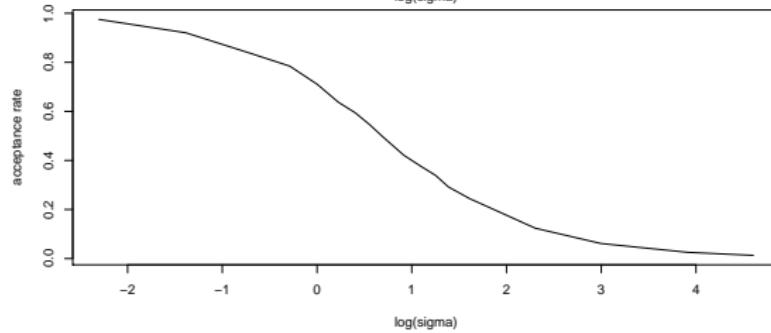
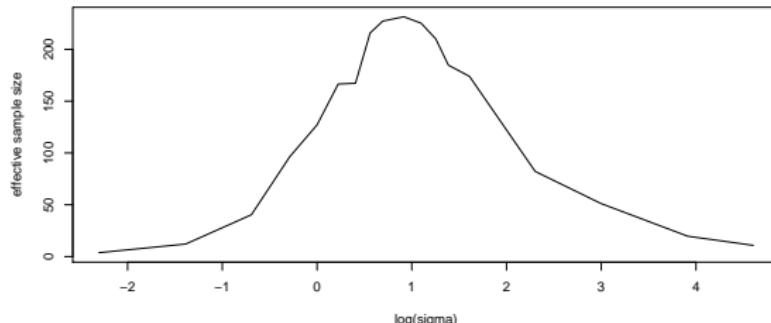


Acceptance Rate

Bigger is not always Better!!



Finding the Optimal Acceptance Rate

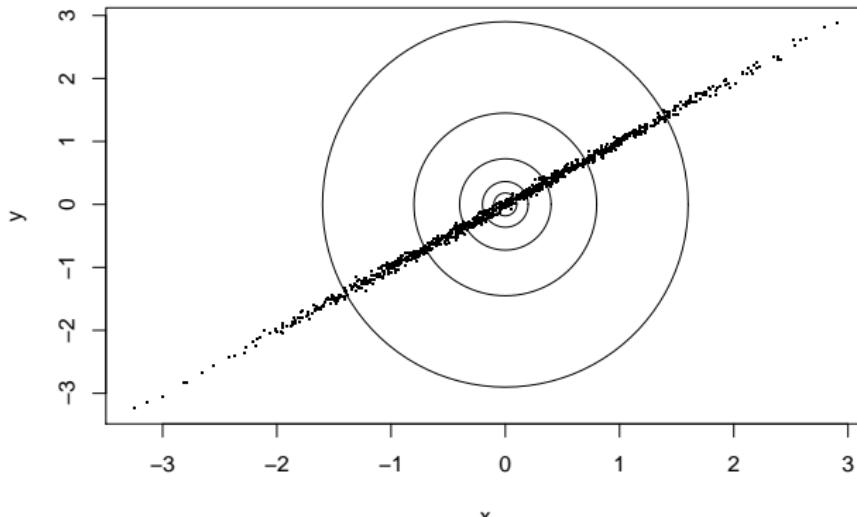


Random Walk Metropolis with High Correlation

A whole new set of issues arise in higher dimensions...

Tradeoff between high autocorrelation and high rejection rate:

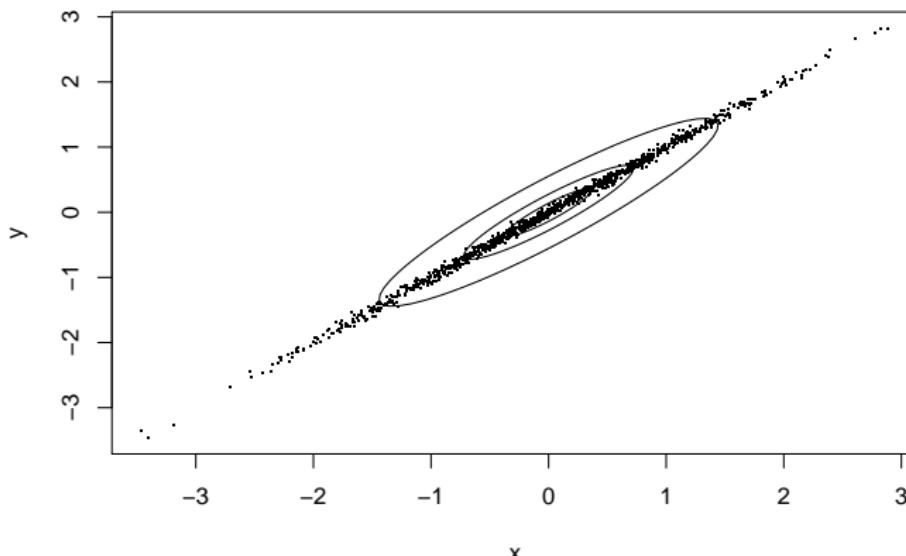
- more acute with high posterior correlations
- more acute with high dimensional parameter



Random Walk Metropolis with High Correlation

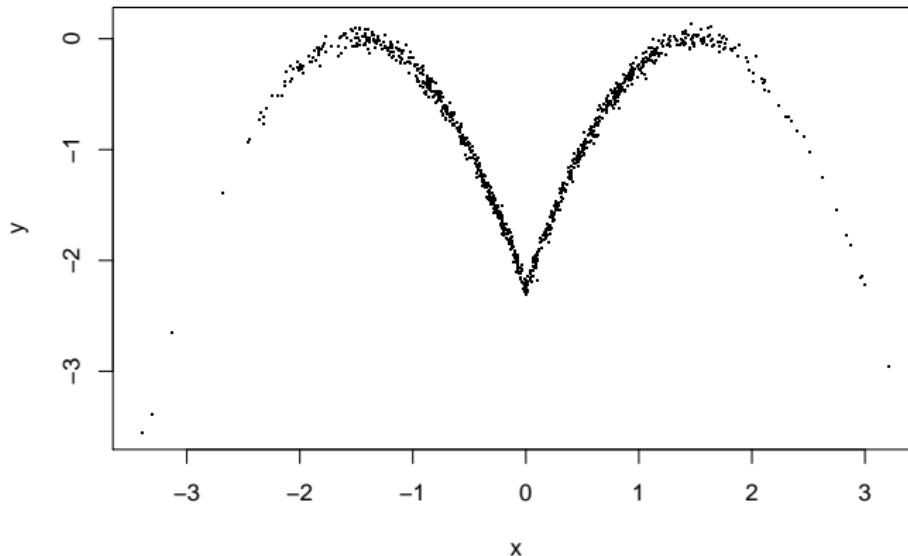
In principle we can use a correlated jumping rule, but

- the desired correlation may vary, and
- is often difficult to compute in advance.



Random Walk Metropolis with High Correlation

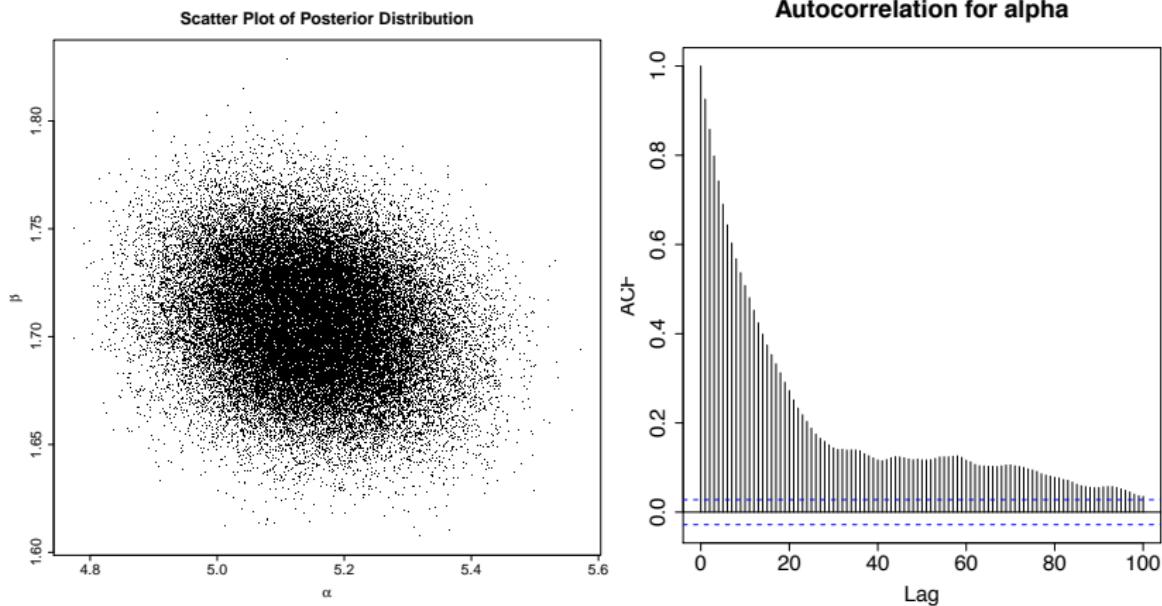
What random walk jumping rule would you use here?



Remember: you don't get to see the distribution in advance!

Parameters on Different Scales

Random Walk Metropolis for Spectral Analysis:

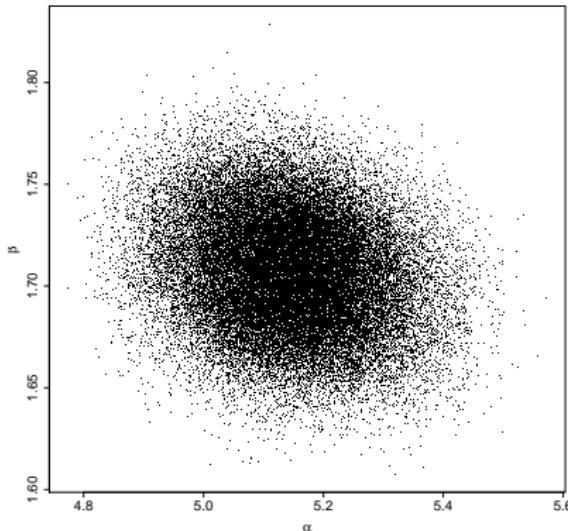


Why is the Mixing SO Poor?!??

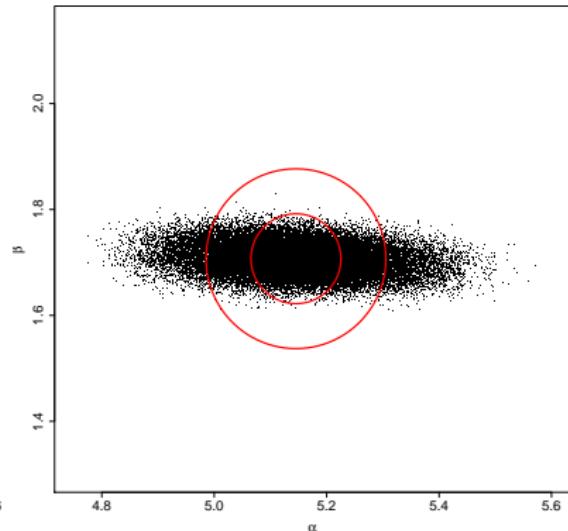
Parameters on Different Scales

Consider the Scales of α and β :

Scatter Plot of Posterior Distribution



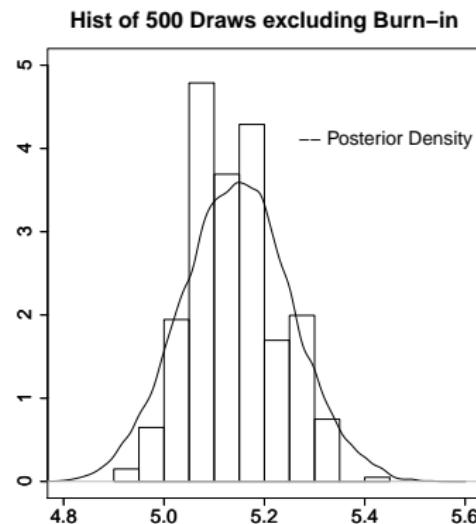
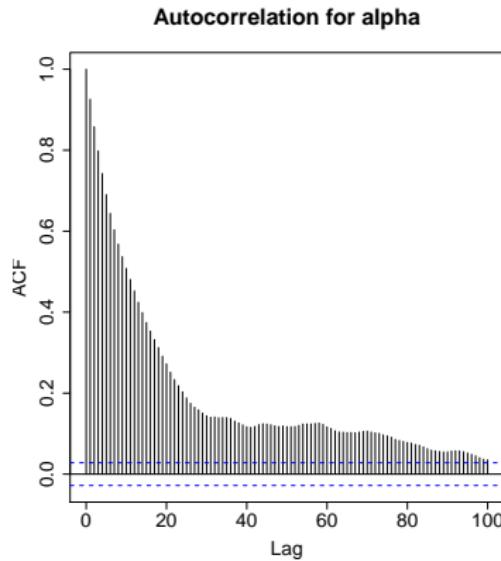
Scatter Plot of Posterior Distribution



A new jumping rule: std dev for $\alpha = 0.110$, for $\beta = 0.026$, and corr = -0.216 .

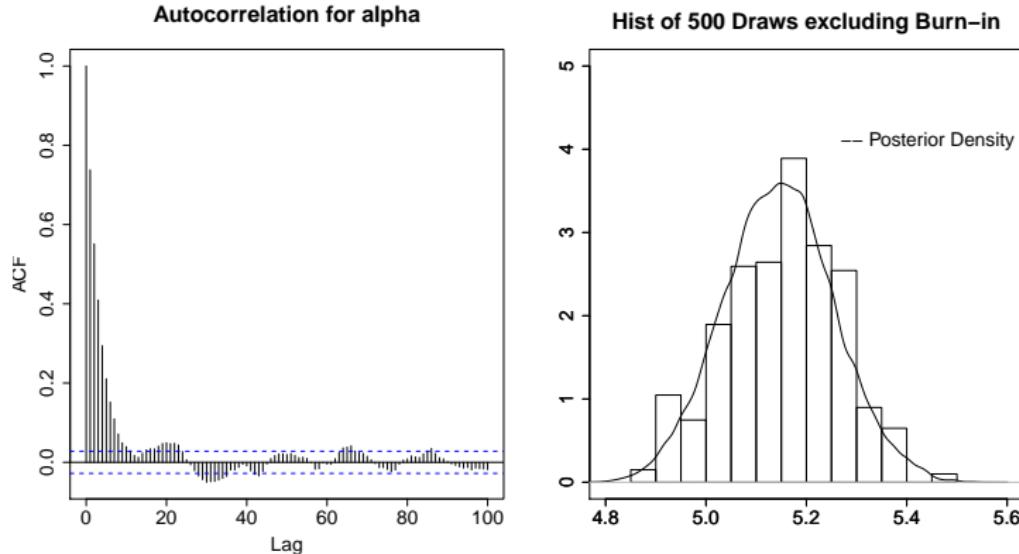
Improved Convergence

Original Jumping Rule:



Improved Convergence

Improved Jumping Rule:



Original Eff Sample Size = 19, Improved Eff Sample Size = 75, with $n = 500$.

Parameters on Different Scales

Strategy: When using

- Normal ($\theta^{(t-1)}, kM$) or better yet
- $t_{\text{df}}(\theta^{(t-1)}, kM)$

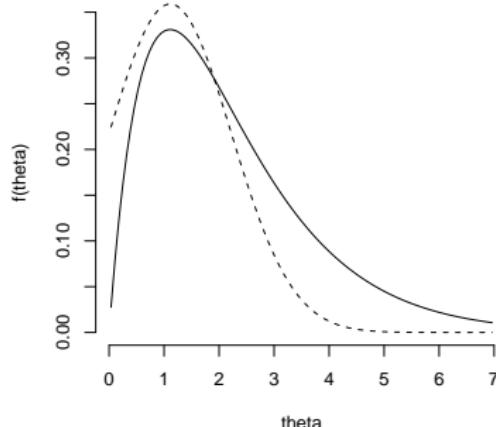
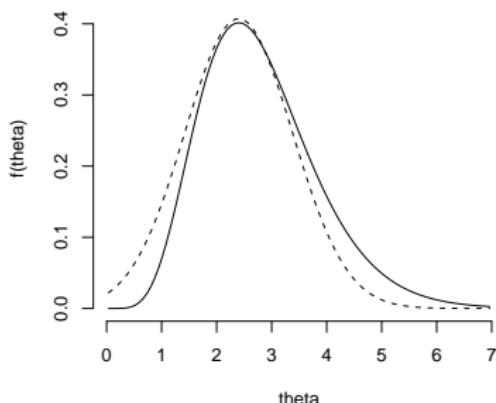
try using the variance-covariance matrix from a standard fitted model for M

... at least when there is standard mode-based model-fitting software available.

Transforming to Normality

Parameter transformations can greatly improve MCMC.

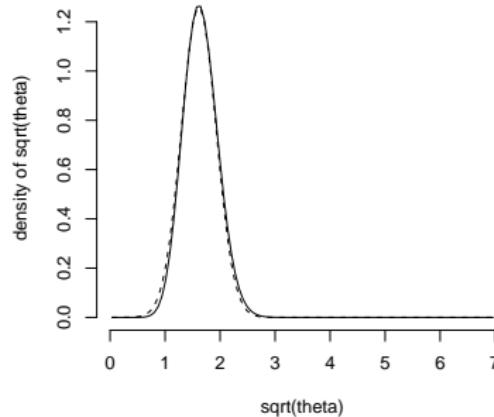
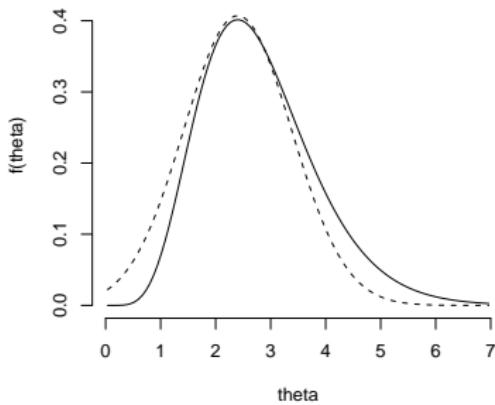
Recall the Independence Sampler:



The normal approximation is not as good as we might hope...

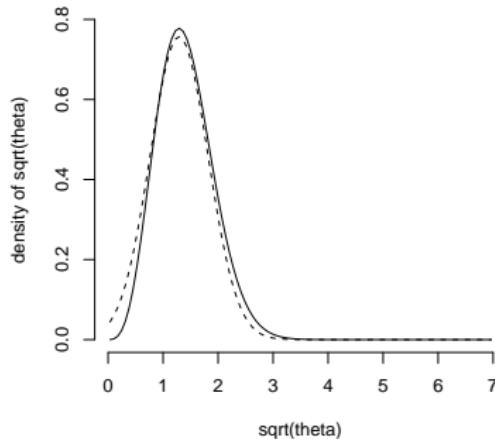
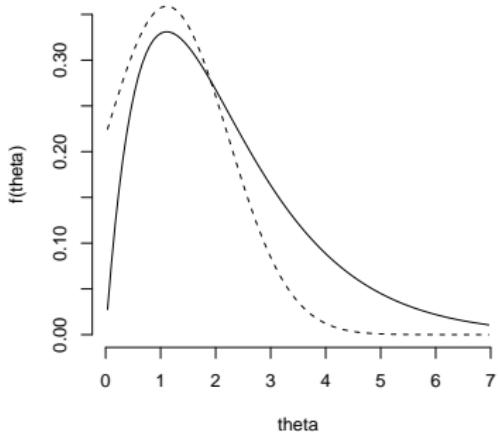
Transforming to Normality

But if we use the square root of θ :



Transforming to Normality

And...



The normal approximation is much improved!

Transforming to Normality

Working with Gaussian or symmetric distributions leads to more efficient Metropolis and Metropolis Hastings Samplers.

General Strategy:

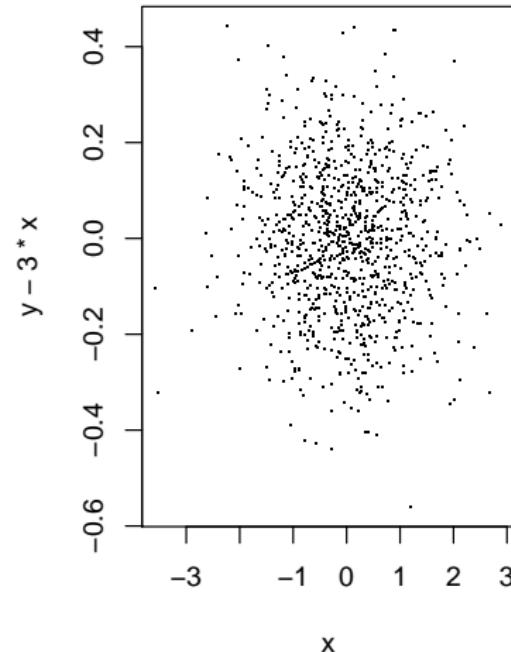
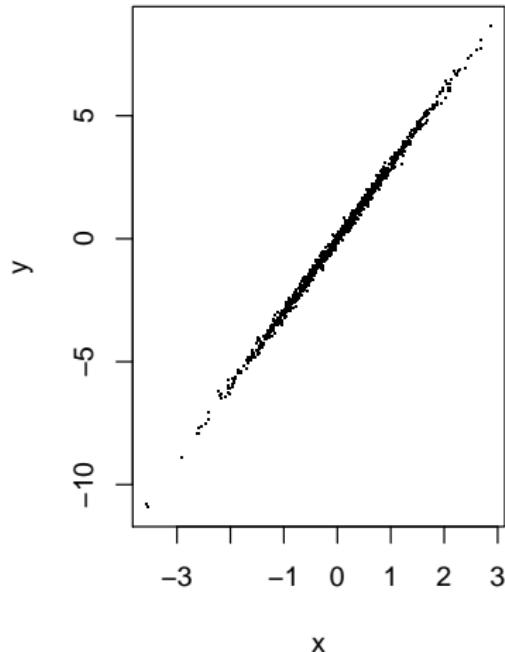
- Transform to the Real Line.
- Take the log of positive parameters.
- If the log is “too strong”, try square root.
- Probabilities can be transformed via the logit transform:

$$\log(p/(1 - p)).$$

- More complex transformations for other quantities.
- Statistical advantages to using normalizing transforms.

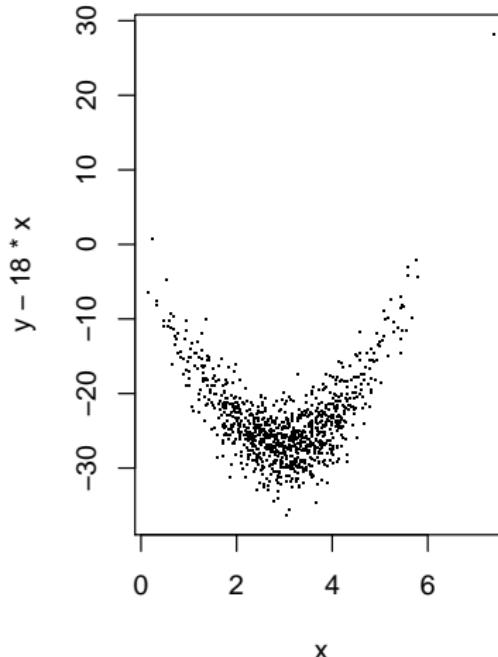
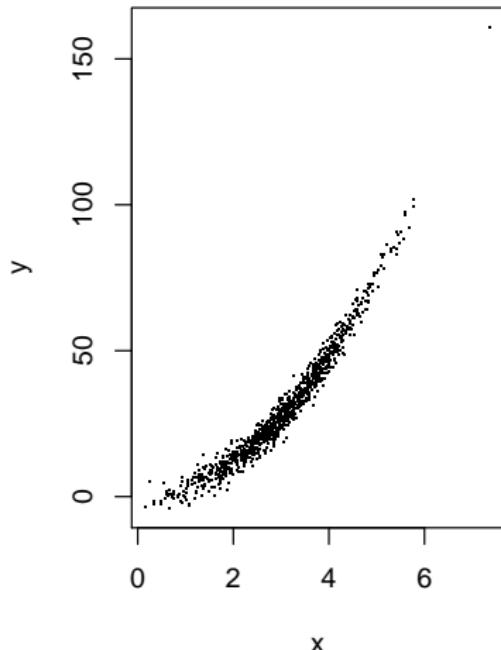
Removing Linear Correlations

Linear transformations can remove linear correlations



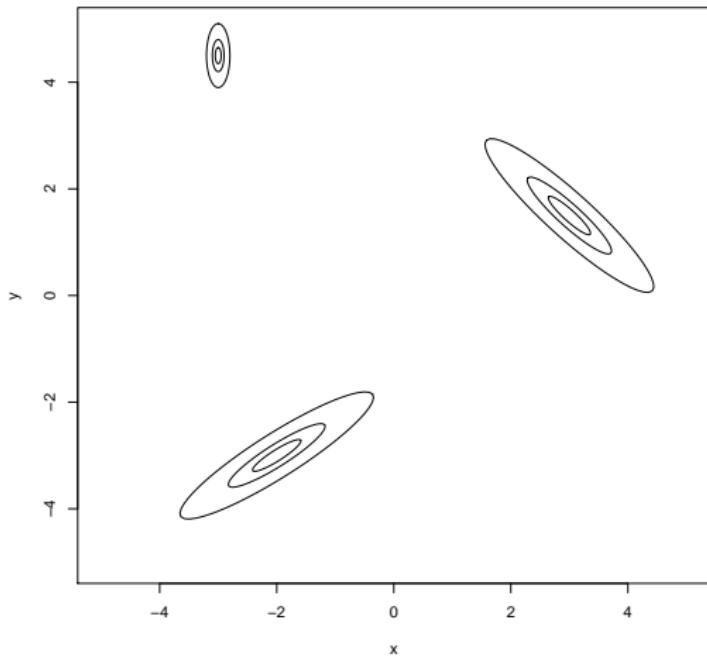
Removing Linear Correlations

... and can help with non-linear correlations.



Multiple Modes

- Scientific meaning of multiple modes.
- Do not focus only on the major mode!
- “Important” modes.
- Computational challenges for Bayesian and Frequentist methods.
- Consider Metropolis & Metropolis Hastings.



Multiple Modes

- ① Use a mode finder to “map out” the posterior distribution.
 - ① Design a jumping rule that accounts for all of the modes.
 - ② Run separate chains for each mode.
- ② Use one of several sophisticated methods tailored for multiple modes.
 - ① Adaptive Metropolis Hastings. Jumping rule adapts when new modes are found (van Dyk & Park, MCMC Hdbk 2011).
 - ② Parallel Tempering.
 - ③ Many other specialized methods.

Outline

1 Background

- Monte Carlo Integration
- Markov Chains

2 Basic MCMC Jumping Rules

- Metropolis Sampler
- Metropolis Hastings Sampler
- Basic Theory

3 Practical Challenges and Advice

- Complex Posterior Distributions
- Choosing a Jumping Rule
- Transformations and Multiple Modes

4 The Gibbs Sampler and Data Augmentation

- The Gibbs Sampler
- Examples and Illustrations of Gibbs
- Data Augmentation

Breaking a Complex Problem into Simpler Pieces

- Ideally we sample directly from $p(\theta|Y)$ without Metropolis.
- This only works in the simplest problems.
- **BUT** in some cases we can split $\theta = (\theta_1, \theta_2)$ so that

$$p(\theta_1|\theta_2, Y) \text{ and } p(\theta_2|\theta_1, Y)$$

are both easy to sample although $p(\theta|Y)$ is not.

- The *Two-Step Gibbs Sampler*, starting with some $\theta^{(0)}$,

For $t = 1, 2, 3, \dots$

Draw: $\theta_1^{(t)} \sim p(\theta_1|\theta_2^{(t-1)}, Y)$
Draw: $\theta_2^{(t)} \sim p(\theta_2|\theta_1^{(t)}, Y)$

An Example

Recall Simple Spectral Model: $Y_i \sim \text{Poisson}(\alpha E_i^{-\beta})$.

Using $p(\alpha, \beta) \propto 1$,

$$\begin{aligned} p(\theta | Y) &\propto \prod_{i=1}^n e^{-[\alpha E_i^{-\beta}]} [\alpha E_i^{-\beta}]^{Y_i} \\ &= e^{-\alpha \sum_{i=1}^n E_i^{-\beta}} \alpha^{\sum_{i=1}^n Y_i} \prod_{i=1}^n E_i^{-\beta Y_i} \end{aligned}$$

So that

$$\begin{aligned} p(\alpha | \beta, Y) &\propto e^{-\alpha \sum_{i=1}^n E_i^{-\beta}} \alpha^{\sum_{i=1}^n Y_i} \\ &= \text{Gamma} \left(\sum_{i=1}^n Y_i + 1, \sum_{i=1}^n E_i^{-\beta} \right) \end{aligned}$$

Embedding Other Samplers within Gibbs

In this case $p(\beta|\alpha, Y)$ is not a standard distribution:

$$p(\beta|\alpha, Y) \propto e^{-\alpha \sum_{i=1}^n E_i^{-\beta}} \prod_{i=1}^n E_i^{-\beta Y_i}$$

- We can use a Metropolis or Metropolis-Hastings step to update β within the Gibbs sampler.
- The result is known as Metropolis within Gibbs Sampler.
- **Advantage:** Metropolis tends to perform poorly in high dimensions. Gibbs reduces the dimension.
- **Disadvantage:** Case-by-case probabilistic calculations.

The General Gibbs Sampler

- ① In general we break θ into P subvectors $\theta = (\theta_1, \dots, \theta_P)$.
- ② The Complete Conditional Distributions are given by

$$p(\theta_p | \theta_1, \dots, \theta_{p-1}, \theta_{p+1}, \dots, \theta_P, Y), \text{ for } p = 1, \dots, P$$

- ③ The *Gibbs Sampler*, starting with some $\theta^{(0)}$,

For $t = 1, 2, 3, \dots$

Draw 1: $\theta_1^{(t)} \sim p(\theta_1 | \theta_2^{(t-1)}, \dots, \theta_P^{(t-1)}, Y)$

⋮

Draw p: $\theta_p^{(t)} \sim p(\theta_p | \theta_1^{(t)}, \dots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \dots, \theta_P^{(t-1)}, Y)$

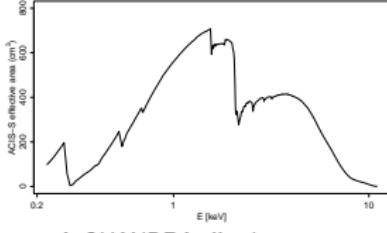
⋮

Draw P: $\theta_P^{(t)} \sim p(\theta_P | \theta_1^{(t)}, \dots, \theta_{P-1}^{(t)}, Y)$

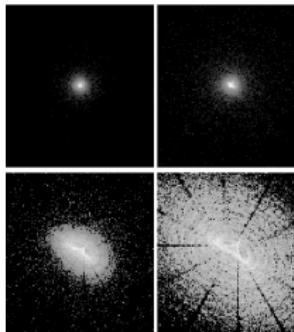
- ④ Determining the partition of θ is a matter of skill and art.

Example: Calibration Uncertainty in High Energy Astrophysics

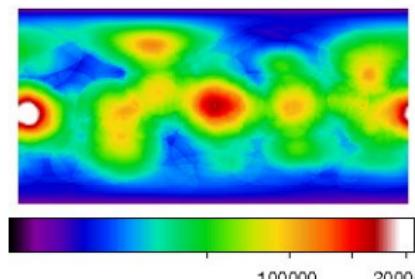
- Analysis is highly dependent on *Calibration Products*:
 - Effective area records sensitivity as a function of energy
 - Energy redistribution matrix can vary with energy/location
 - Point Spread Functions can vary with energy and location
 - Exposure Map shows how effective area varies in an image



A CHANDRA effective area.



Sample Chandra psf's
(Karovska et al., ADASS X)

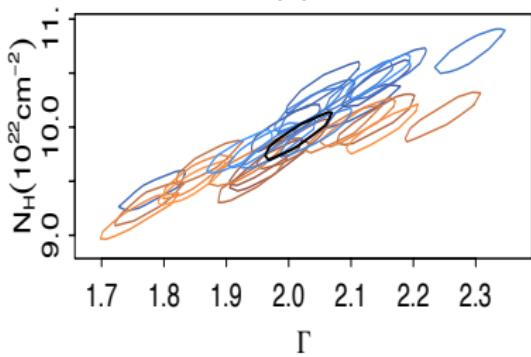
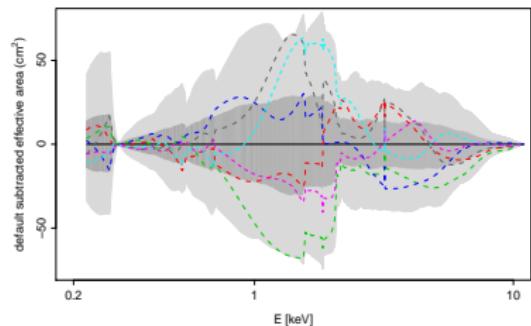


EGERT exposure map
(area \times time)

Example: Calibration Uncertainty

Derivation of Calibration Products

- Prelaunch ground-based and post-launch space-based empirical assessments.
- Aim to capture deterioration of detectors over time.
- Complex computer models of subassembly components.
- Calibration scientists provide a sample representing uncertainty



Example: Calibration Uncertainty

We wish to incorporate uncertainty represented in Calibration sample a Fully Bayesian Analysis.

- **PyBLoCXS (Python Bayesian Low Count X-ray Spectral):** provides a MCMC output for spectral analysis with *known* calibration products.
- Can we leverage PyBLoCXS for calibration uncertainty?
- Gibbs Sampler:
 - Draw 1: Update A (effective area) given θ (parameter).
 - Draw 2: Update θ given A with PyBLoCXS.

Power of Gibbs Sampling: breaks a problem into easier parts.

How do we draw A ?

We have only a calibration sample, not a formal model.

We use Principal Component Analysis to represent uncertainty:

$$A \sim A_0 + \bar{\delta} + \sum_{j=1}^m e_j r_j \mathbf{v}_j,$$

A_0 : default effective area,

$\bar{\delta}$: mean deviation from A_0 ,

r_j and \mathbf{v}_j : first m principle component eigenvalues & vectors,

e_j : independent standard normal deviations.

Capture 95% of variability with $m = 6 - 9$.

A Prototype Fully Bayesian Sampler

An MH within Gibbs Sampler:

STEP 1: $e \sim \mathcal{K}(e|e', \theta')$ via MH with limiting dist'n $p(e|\theta, Y)$

STEP 2: $\theta \sim \mathcal{K}(\theta|e', \theta')$ via MH with limiting dist'n $p(\theta|e, Y)$

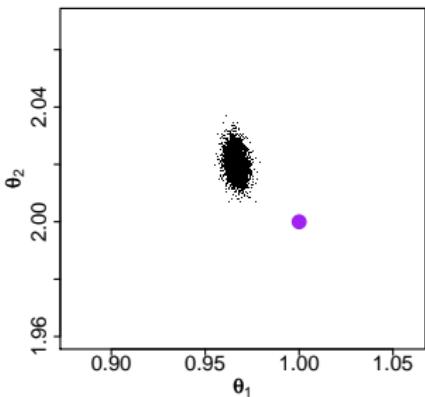
- STEP 1: Gaussian Metropolis jumping rule centered at e' .
- STEP 2: Simplified pyBLoCXS (no rmf or background).

A Simulation.

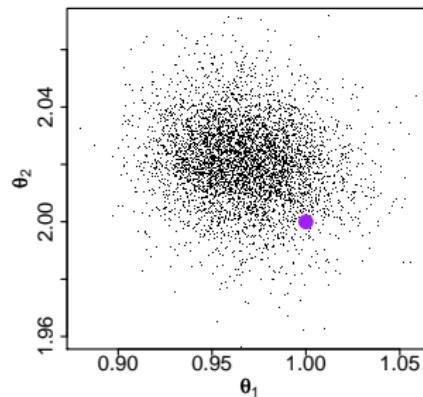
- Sampled 10^5 counts from a power law spectrum: e^{-2E} .
- A_{true} is 1.5σ from the center of the calibration sample.

Sampling From the Full Posterior

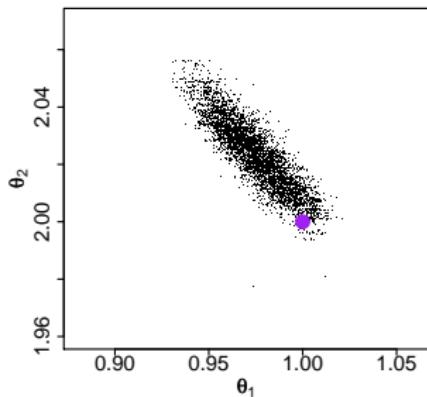
Default Effective Area



Pragmatic Bayes



Fully Bayes

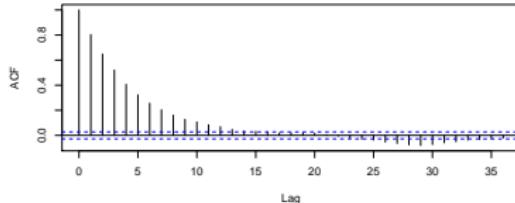
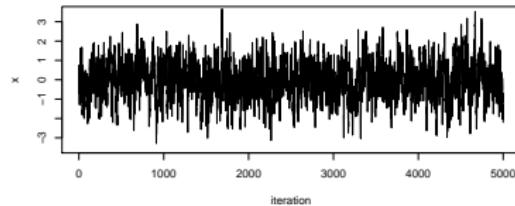
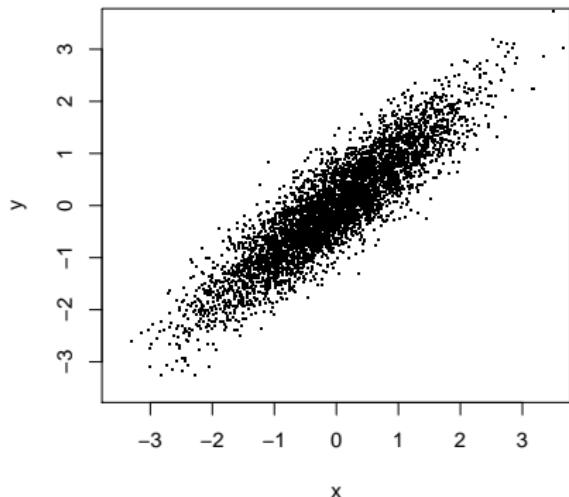


θ_1 = normalization, θ_2 = power law param, purple bullet = truth

See Poster: van Dyk, Xu, Kashyap, Siemiginowska, and Connors for real analyses.

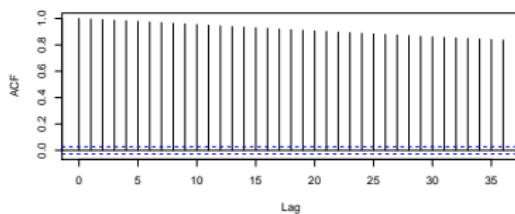
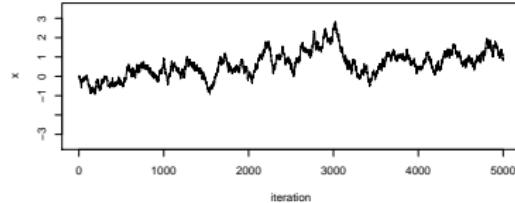
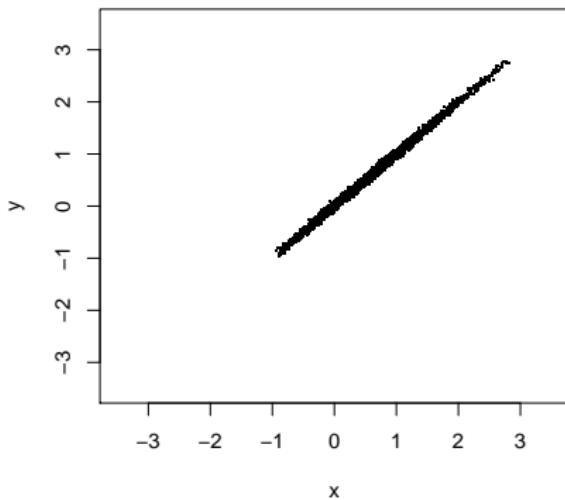
Citation: Lee, H., Kashyap, V. L., van Dyk, D. A., Connors, A., Drake, J. J., Izem, R., Meng, X. L., Min, S., Park, T., Ratzlaff, P., Siemiginowska, A., and Zelas, A. (2011). Accounting for Calibration Uncertainties in X-ray Analysis: Effective Areas in Spectral Fitting. *The Astrophysical Journal*, **731**, 126–144.

When Will Gibbs Sampling Work Well?



autocorrelation = 0.81, effective sample size = 525

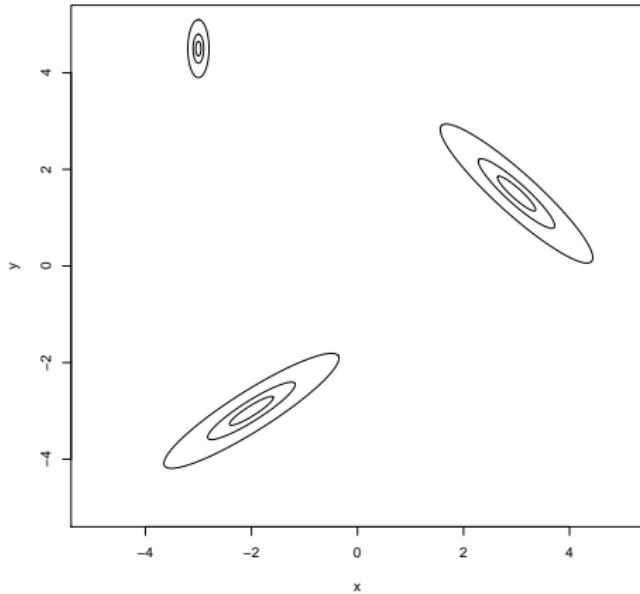
When Will Gibbs Sampling Work Poorly?



autocorrelation = 0.998, effective sample size = 5

High Posterior Correlations are Always Problematic.

Multiple Modes



How will the Gibbs Sampler Handle Multiple modes?

Example: Transformations are Key

Fitting Computer Models for Stellar Evolution

- A complex computer model predicts observed *photometric magnitudes* of a stellar cluster as a function of

M_i : stellar masses, and

Θ : cluster composition, age, distance, and absorption:

$$\mathbf{G}(M_i, \Theta)$$

- We assume indep Gaussian errors with known variances:

$$L_0(\mathbf{M}, \Theta | \mathbf{X}) = \prod_{i=1}^N \left(\prod_{j=1}^n \left[\frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(-\frac{(x_{ij} - G_j(M_{i1}, \Theta))^2}{2\sigma_{ij}^2}\right) \right] \right).$$

Example: Stellar Evolution

Model Extensions:

- Binary stars: The luminosities of component stars sum.
- Field stars: Contaminate the data and magnitudes don't follow the pattern of the cluster.
- Initial Final Mass Relation is fit to combine stellar evolution models for the main sequence and for white dwarfs.
- A combination of informative and non-informative priors.

Citations:

- 1 van Dyk, D. A., DeGennaro, S., Stein, N., Jeffreys, W. H., von Hippel, T. Statistical Analysis of Stellar Evolution. *The Annals of Applied Statistics* **3**, 117-143, 2009.
- 2 DeGennaro, S., von Hippel, T., Jeffreys, W., Stein, N., van Dyk, D., and Jeffery, E. Inverting Color-Magnitude Diagrams to Access Precise Cluster Parameters: A New White Dwarf Age for the Hyades. *The Astrophysical Journal*, **696**, 12–23, 2009.
- 3 Jeffery, E., von Hippel, T., DeGennaro, S., van Dyk, D., Stein, N., and Jeffreys, W. H., The White Dwarf Age of NGD 2477. *The Astrophysical Journal*, **730**, 35–44, 2011.

Stellar Evolution: MCMC Strategy

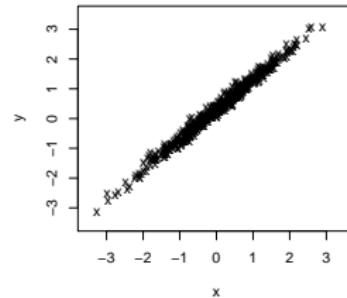
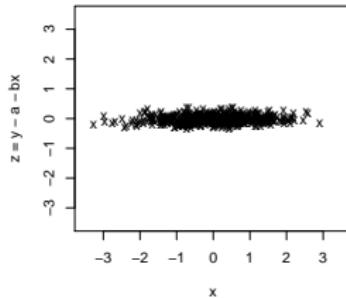
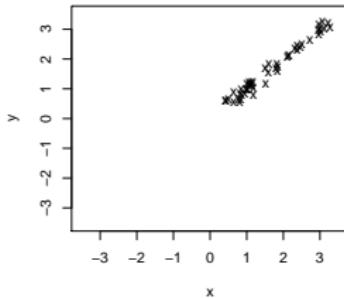
Metropolis within Gibbs Sampling

- $3N + 5$ parameters, none with closed form update.
- Strong posterior correlations among the parameters.

Strong Linear and Non-Linear Correlations Among Parameters

- Static and/or dynamic (power) transformations remove non-linear relationships.
- A series of preliminary runs is used to evaluate and remove linear correlations.
- We tune a linear transformation to the correlations of the posterior distribution on the fly.
- Results in a dramatic improvement in mixing.

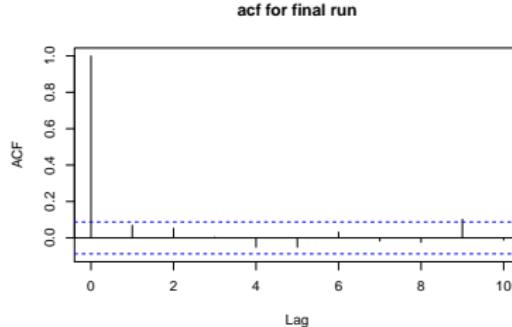
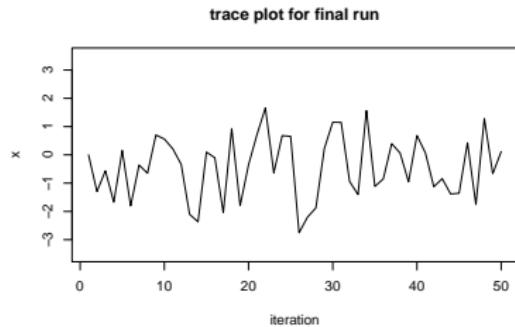
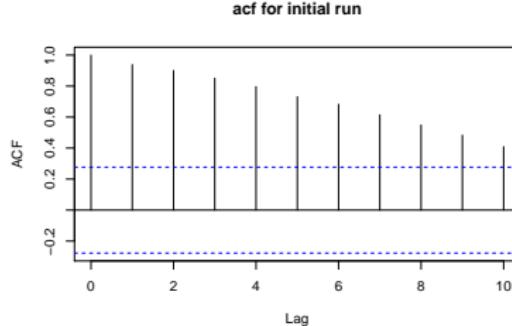
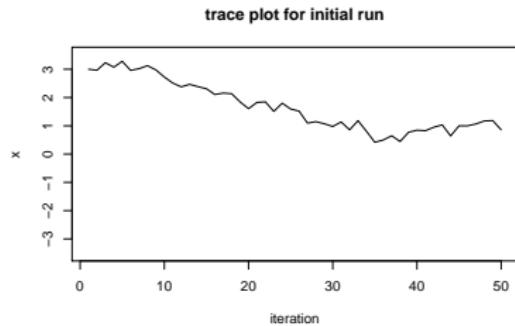
Dynamic transformations



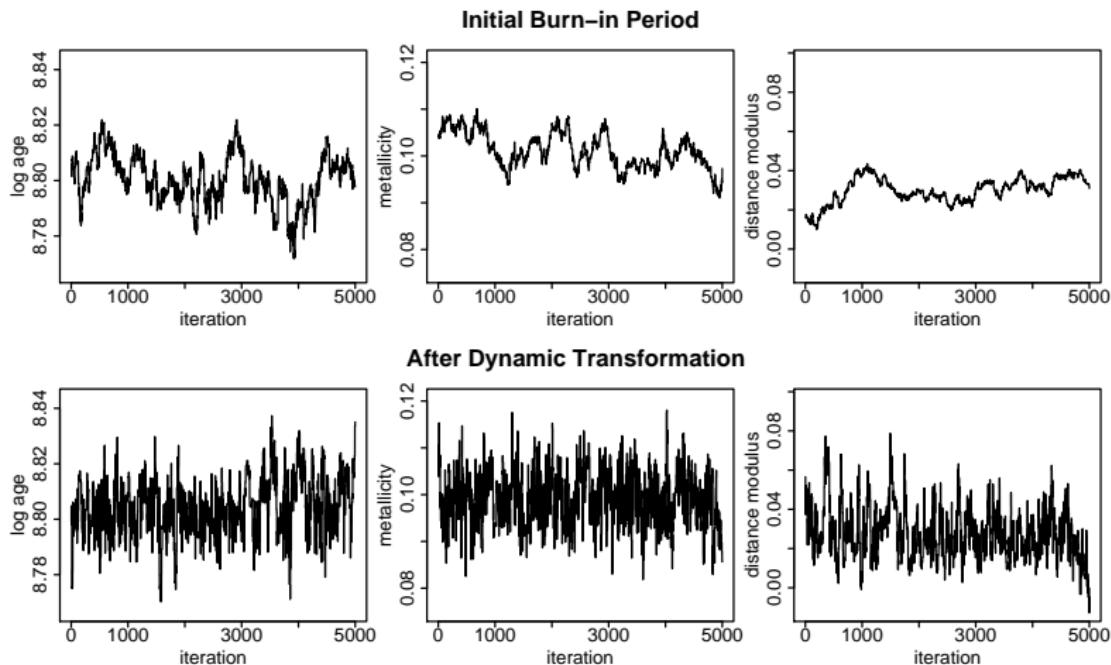
A toy example:

- ① Initial Gibbs run shows high autocorrelation, panel 1.
- ② Fit $y = \alpha + \beta x$ and transform $Z = Y - \hat{\alpha} - \hat{\beta}X$.
- ③ Rerun Gibbs, but sampling $p(X|Z)$ and $p(Z|X)$, panel 2.
- ④ Transform back to X, Y , panel 3.

Results for Toy Example



Results for Stellar Evolution Model



Data Augmentation

- We can sometimes simplify computation by including other unknown quantities in the model.
- Canonical Examples: *Missing Data* in Sample Surveys.
- If we had *Complete Data* analysis would be easier.
- More generally: there may quantities that we never *expected to observe*, but had we observed them, data analysis would be easier.

We call such quantities *Augmented Data* and their use in statistical computation *The Method of Data Augmentation*.

Handling Background with DA

Simple Example: Backgd contamination in single bin detector.

- Contaminated source counts: $Y = Y_S + Y_B$
- Background counts: X
- Background exposure is 24 times the source exposure.
- We observe Y and X .

A Poisson Multi-Level Model:

LEVEL 1: $Y|Y_B, \lambda_S \sim \text{Poisson}(\lambda_S) + Y_B$.

LEVEL 2: $Y_B|\lambda_B \sim \text{Pois}(\lambda_B)$ and $X|\lambda_B \sim \text{Pois}(24\lambda_B)$.

LEVEL 3: Specify a prior distribution on λ_B and λ_S .

Handling Background with DA

A Poisson Multi-Level Model:

LEVEL 1: $Y|Y_B, \lambda_S \sim \text{Poisson}(\lambda_S) + Y_B$.

LEVEL 2: $Y_B|\lambda_B \sim \text{Pois}(\lambda_B)$ and $X|\lambda_B \sim \text{Pois}(24\lambda_B)$.

LEVEL 3: Specify a prior distribution on λ_B and λ_S .

Data Augmentation

- Formulate model in terms of “missing data”.
- If Y_B were known.
- If λ_B and λ_S were known.

With Y_B we simplify the relationships among the quantities.

The Data Augmentation Sampler

A Two-Step Gibbs Sampler:

STEP 1: Sample Y_B given (λ_S, λ_B) , X , and Y .

$$Y_B \sim \text{Binomial}\left(Y, \frac{\lambda_B}{\lambda_S + \lambda_B}\right)$$

STEP 2: Sample (λ_S, λ_B) given X , Y_B , and Y_S .

$$\lambda_B \sim \text{Gamma}(X + Y_B + 1, 24 + 1)$$

$$\lambda_S \sim \text{Gamma}(Y_S + 1, 1)$$

The power of data augmentation is that it separates a complex problem into a series of simpler parts.

Details of STEP 1

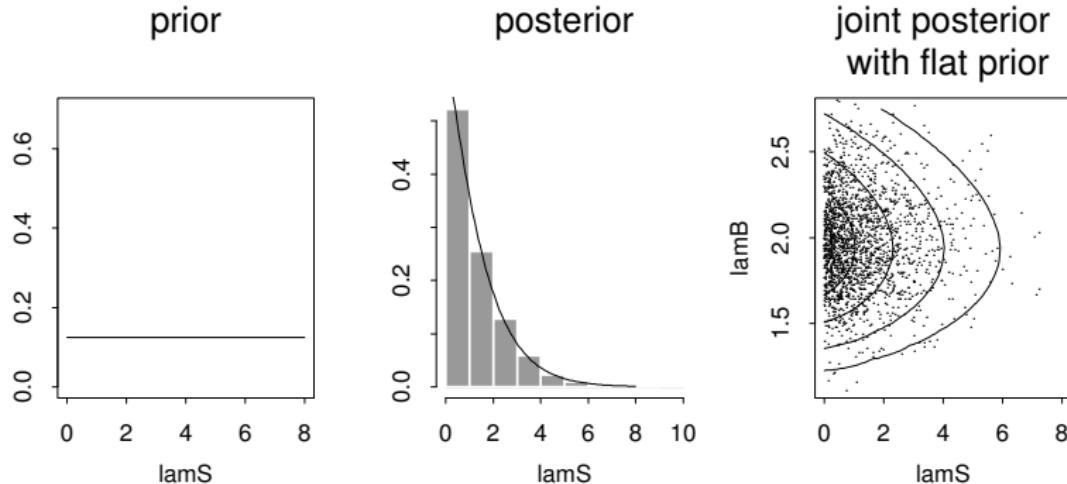
$$\begin{aligned}
 p(Y_B, |\lambda_B, \lambda_S, Y) &\propto p(Y_B, Y|\lambda_B, \lambda_S) \\
 &= p(Y|\lambda_B, \lambda_S, Y_B) \times p(Y_B|\lambda_B, \lambda_S) \\
 &= \frac{e^{-\lambda_S} \lambda_S^{Y - Y_B}}{(Y - Y_B)!} \times \frac{e^{-\lambda_B} \lambda_B^{Y_B}}{Y_B!} \\
 &\propto \frac{1}{(Y - Y_B)! Y_B!} \lambda_S^{Y - Y_B} \lambda_B^{Y_B} \\
 &\propto \frac{Y!}{(Y - Y_B)! Y_B!} \left(\frac{\lambda_S}{\lambda_S + \lambda_B}\right)^{Y - Y_B} \left(\frac{\lambda_B}{\lambda_S + \lambda_B}\right)^{Y_B} \\
 &= \text{Binomial}\left(Y, \frac{\lambda_B}{\lambda_S + \lambda_B}\right)
 \end{aligned}$$

Requires case-by-case probability calculations.

Details of STEP 2

$$\begin{aligned}
 p(\lambda_S, \lambda_B, | Y_B, Y, X) &= p(\lambda_S, \lambda_B, | Y_S, Y_B, X) \\
 &\propto p(Y_S, Y_B, X | \lambda_B, \lambda_S) \\
 &= p(Y_S | \lambda_S) p(Y_B | \lambda_B) p(X | \lambda_B) \\
 &= \frac{e^{-\lambda_S} \lambda_S^{Y_S}}{Y_S!} \frac{e^{-\lambda_B} \lambda_B^{Y_B}}{Y_B!} \frac{e^{-24\lambda_B} (24\lambda_B)^X}{X!} \\
 &\propto \left(e^{-\lambda_S} \lambda_S^{Y_S} \right) \times \left(e^{-(24+1)\lambda_B} \lambda_B^{Y_B+X} \right) \\
 &\propto \gamma(Y_S + 1, 1) \times \gamma(X + Y_B + 1, 24 + 1)
 \end{aligned}$$

Results



Here $Y = 1$ and $X = 48$.

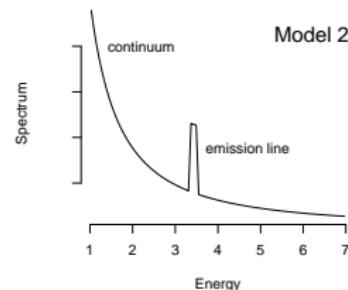
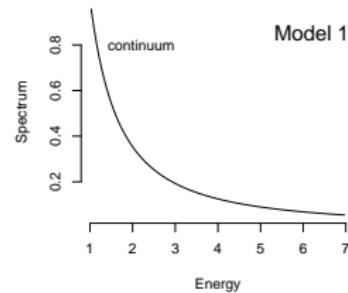
Handling a Spectral Emission Line

Recall the Power Law Spectral Model:

- $Y_i \sim \text{Poisson}(\alpha E_i^{-\beta})$.

Add a Spectral Emission Line:

- ① $Y_i \sim \text{Poisson}(\alpha E_i^{-\beta} + \gamma I\{i \in \mathcal{L}(\delta)\})$.
- ② $I\{i \in \mathcal{L}(\delta)\}$ is one if $i \in \mathcal{L}(\delta)$, otherwise it is zero.
- ③ $\mathcal{L}(\delta) = \{\delta - 1, \delta, \delta + 1\}$
- ④ $\theta_2 = (\alpha, \beta, \gamma, \delta)$



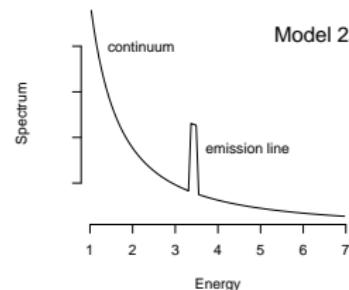
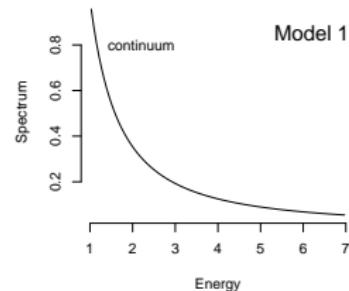
Handling a Spectral Emission Line

Continuum + Emission Line Model:

- ① $Y_i \sim \text{Poisson} \left(\alpha E_i^{-\beta} + \gamma I\{i \in \mathcal{L}(\delta)\} \right)$.
- ② An example of a *finite mixture model*.
- ③ Let Z_i in count in bin i due to line.
- ④ $Z_i | (Y_i, \theta_2) \sim$

$$\text{Binomial} \left(Y_i, \frac{\gamma I\{i \in \mathcal{L}(\delta)\}}{\gamma I\{i \in \mathcal{L}(\delta)\} + \alpha E_i^{-\beta}} \right)$$

- ⑤ Update α, β, γ , and δ given Z_i and $X_i = Y_i - Z_i$?



A Metropolis within Gibbs Sampler

A Two-Step Sampler:

STEP 1: Sample Z_i given (θ_2, Y_i) , for $i = 1, \dots, n$.

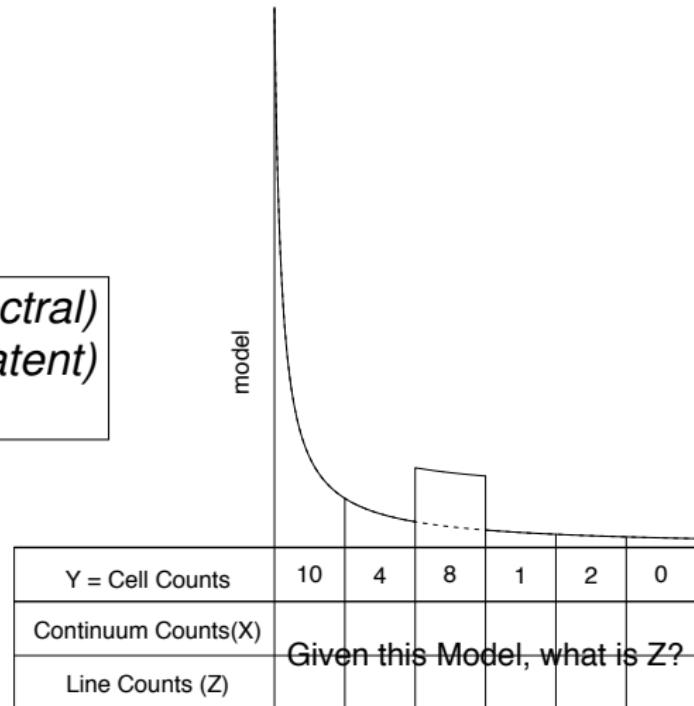
$$Z_i | (Y_i, \theta_2) \sim \text{Binomial} \left(Y_i, \frac{\gamma I\{i \in \mathcal{L}(\delta)\}}{\gamma I\{i \in \mathcal{L}(\delta)\} + \alpha E_i^{-\beta}} \right)$$

STEP 2: $p(\alpha, \beta, \gamma, \delta | X, Z) = p(\alpha, \beta | X)p(\gamma, \delta | Z)$
 $= p(\alpha, \beta | X)p(\gamma | \delta, Z)p(\delta | Z)$

- ① Sample $p(\alpha, \beta | X)$ using Metropolis or MH.
- ② $\gamma | (\delta, X) \sim \text{gamma}(\sum Z_i, 3)$
- ③ Updating δ given X is tricky.

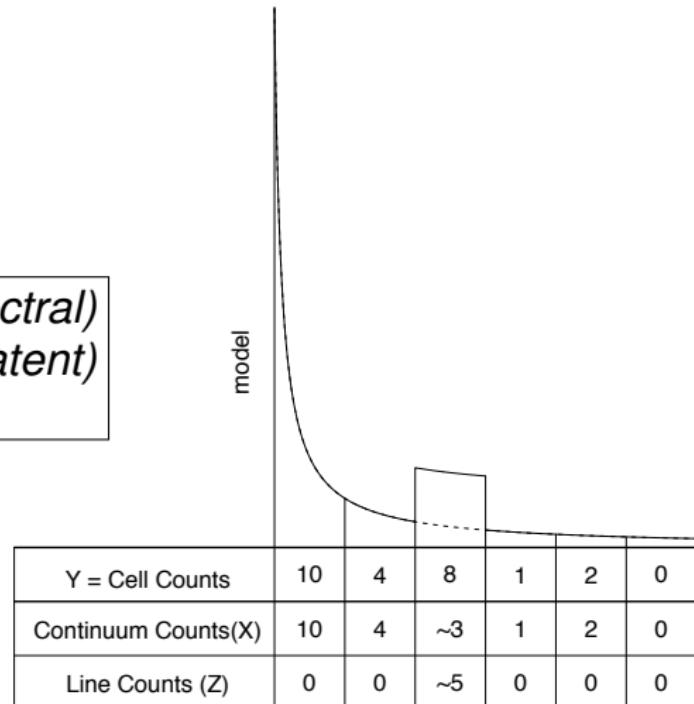
When Data Augmentation Fails

Consider a simple (spectral) model with the given (latent) cell counts.



When Data Augmentation Fails

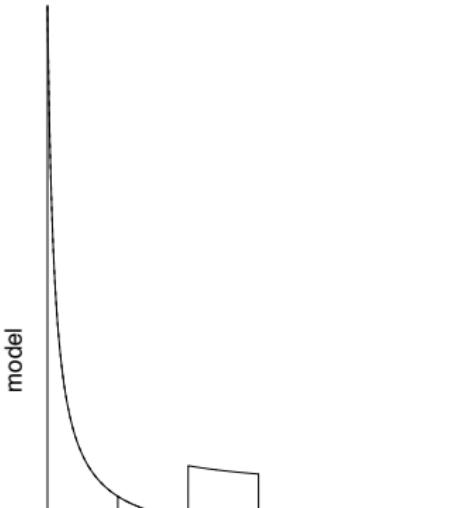
Consider a simple (spectral) model with the given (latent) cell counts.



When Data Augmentation Fails

Consider a simple (spectral) model the with given (latent) cell counts.

Given Z what is the location of the emission line??



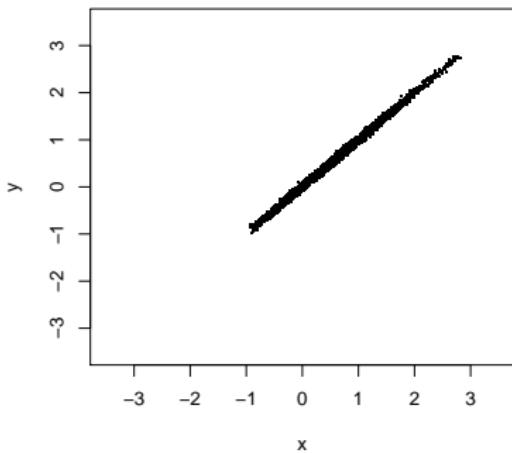
| | | | | | | |
|---------------------|----|---|----|---|---|---|
| Y = Cell Counts | 10 | 4 | 8 | 1 | 2 | 0 |
| Continuum Counts(X) | 10 | 4 | ~3 | 1 | 2 | 0 |
| Line Counts (Z) | 0 | 0 | ~5 | 0 | 0 | 0 |

Handling a Spectral Emission Line

What Went Wrong?

*High Posterior Correlations
Are Always Problematic*

- Here Z and δ are highly correlated. In fact $\text{Var}(\delta|Z) = 0$.
- Given Z , δ will not change from iteration to iteration.



SOLUTION: Sample Z and δ in the same step.

An Improved Metropolis within Gibbs Sampler

A Two-Step Sampler:

STEP 1: Sample (Z, δ) given $(\alpha, \beta, \gamma, Y)$:

- ① Sample δ given Y, α, β, γ using grid method:

$$p(\delta | \alpha, \beta, \gamma, Y) \propto p(Y | \theta_2).$$

- ② For $i = 1, \dots, n$,

$$Z_i | (Y_i, \theta_2) \sim \text{Binomial} \left(Y_i, \frac{\gamma I\{i \in \mathcal{L}(\delta)\}}{\gamma I\{i \in \mathcal{L}(\delta)\} + \alpha E_i^{-\beta}} \right)$$

STEP 2: $p(\alpha, \beta, \gamma | \delta, X, Z) = p(\alpha, \beta | X)p(\gamma | \delta, Z)$:

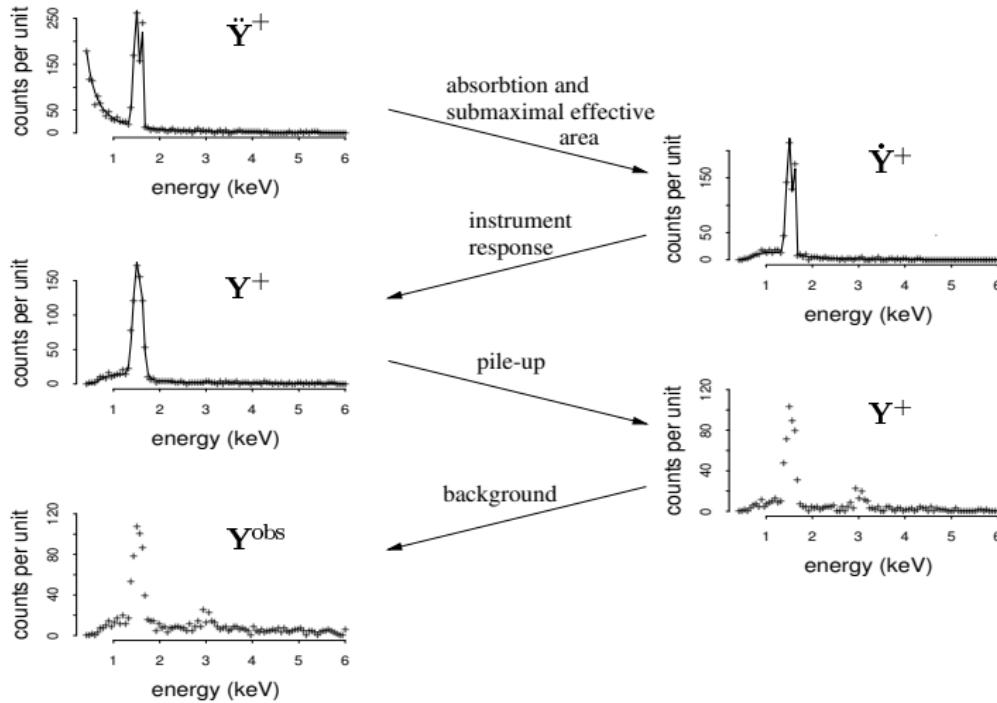
- ① Sample $p(\alpha, \beta | X)$ using Metropolis or MH.
- ② $\gamma | (\delta, X) \sim \text{gamma}(\sum Z_i, 3)$

Strategies for Implementing Gibbs Samplers

How we set up the complete conditional distributions can have a big impact on the performance of a Gibbs Sampler.

- ➊ We have seen the potential effect of the choice of subsets:
 - $p(\vartheta|\varphi, \varsigma)$ and $p(\varphi, \varsigma|\vartheta)$ versus
 - $p(\vartheta, \varphi|\varsigma)$ and $p(\varsigma|\vartheta, \varphi)$
- ➋ Combining steps into a single joint step is called *blocking*. This generally improves convergence:
 - $p(\vartheta|\varphi, \varsigma)$, $p(\varphi|\vartheta, \varsigma)$, and $p(\varsigma|\vartheta, \varphi)$ versus
 - $p(\vartheta, \varphi|\varsigma)$ and $p(\varsigma|\vartheta, \varphi)$
- ➌ Removing a variable from the chain is called *collapsing*. This is also generally helpful:
 - $p(\vartheta, \varphi|\varsigma)$ and $p(\varsigma|\vartheta, \varphi)$ versus
 - $p(\vartheta|\varsigma)$ and $p(\varsigma|\vartheta)$
- ➍ *Partial Collapsing* encompasses blocking and collapsing.

Example: Using DA for Spectral Analysis

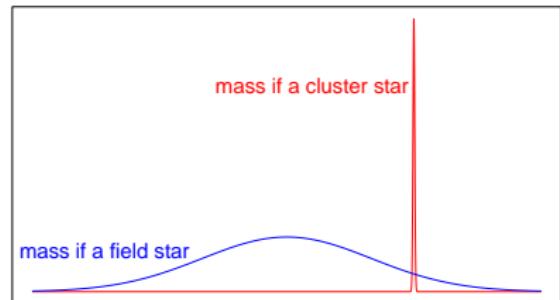


Stellar Evolution: Correlation Reduction via Prior Dist'n

Field/Cluster Indicator is Highly Correlated with Masses

- Data are uninformative for the masses of field stars.
- Data are highly informative for cluster star masses.
- Cannot easily jump from field to cluster star designation.

*Solution: Replace prior for
masses given field star
membership by approximation
of the posterior given cluster
star membership.*



Does not effect statistical inference & enables efficient mixing.

Overview of Recommended Strategy

(Adopted from *Bayesian Data Analysis*, Section 11.10, Gelman et al. (2005), Second Edition)

- ① Start with a crude approximation to the posterior distribution, perhaps using a mode finder.
- ② Simulate directly, avoiding MCMC, if possible.
- ③ If necessary use MCMC with one parameter at a time updating or updating parameters in batches.
- ④ Use Gibbs draws for closed form complete conditionals.
- ⑤ Use metropolis jumps if complete conditional is not in closed form. Tune variance of jumping distribution so that acceptance rates are near 20% (for vector updates) or 40% (for single parameter updates).

Overview of Recommended Strategy- Con't

- ⑥ To improve convergence, use transformations so that parameters are approximately independent.
- ⑦ Check for convergence using multiple chains. (Topic for later today.)
- ⑧ Compare inference based on crude approximation and MCMC. If they are not similar, check for errors before believing the results of the MCMC.

MCMC Diagnostics

Tom Loredo

Dept. of Astronomy, Cornell University

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

Bayesian Computation Tutorials — 11-12 June 2011

MCMC Diagnostics

- ① My MCMC misadventure
- ② Markov chain behavior
- ③ Posterior sample diagnostics
- ④ Joint distribution diagnostics
- ⑤ Closing advice

MCMC Diagnostics

- ① My MCMC misadventure
- ② Markov chain behavior
- ③ Posterior sample diagnostics
- ④ Joint distribution diagnostics
- ⑤ Closing advice

Modeling SN 1987A Neutrino Data, ca. 1991

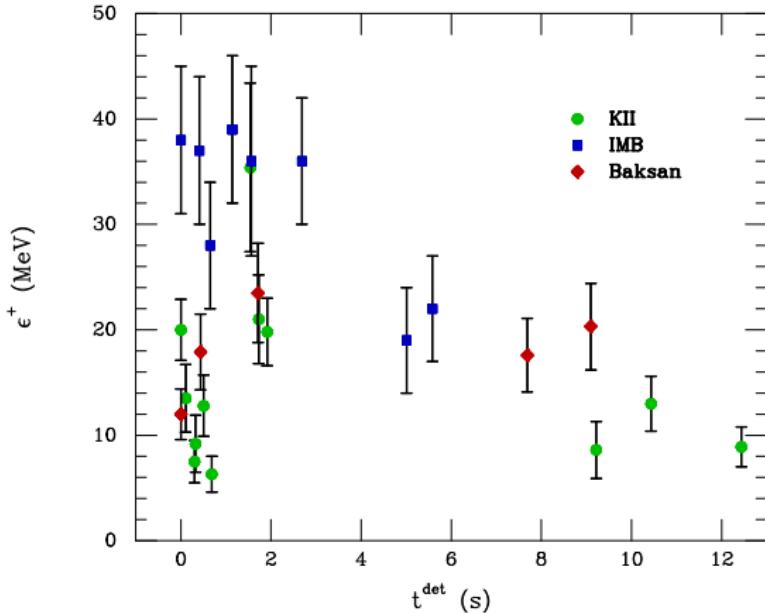
Before 1987



Feb 1987



Arrival times and energies for SN 1987A ν s



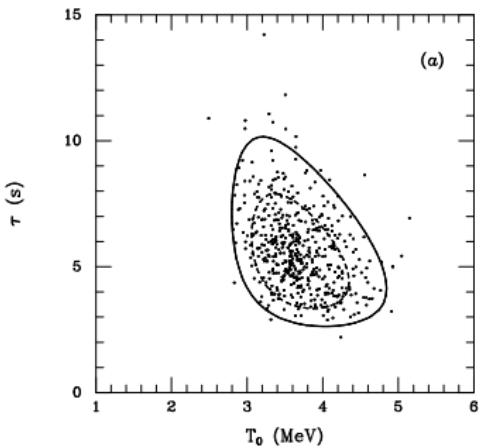
- Do we understand the basics of core collapse?
- Can we discriminate between prompt and delayed shock models?
- How does the nascent NS compare with predictions?
- What is the $\bar{\nu}_e$ rest mass?

Bayesian dynamic spectroscopy

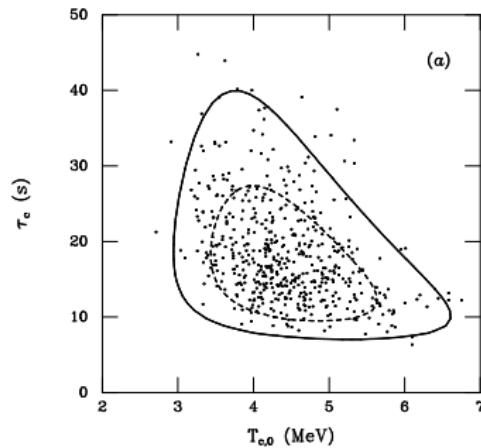
- Consider ~ 10 parametric models for flux vs. (ϵ, t) : prompt shock and delayed shock scenarios, 6 – 10 parameters
- Model data as thinned marked Poisson point process with measurement error—a multilevel model
- Computations: Accept/reject posterior sampling with Student- $t \times \Gamma$ envelope for parameter estimation; subregion-adaptive cubature for Bayes factors

Signal parameter inferences

Prompt shock

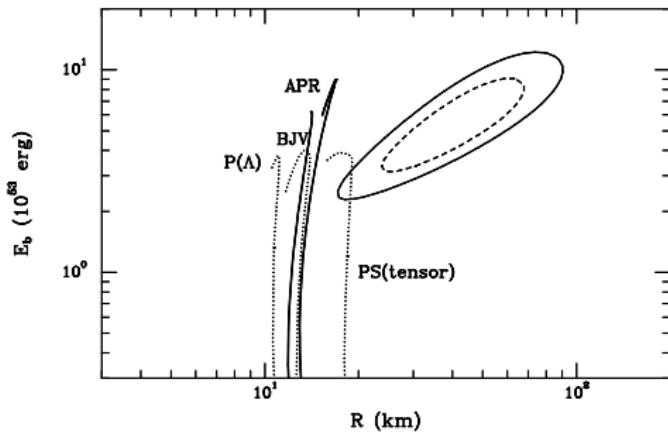


Delayed shock; $B \approx 125$

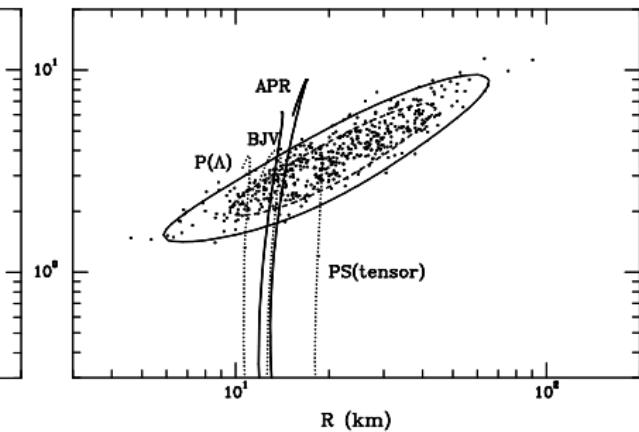


Derived nascent neutron star properties

Prompt shock



Delayed shock



Adding NS mass/radius prior $\rightarrow B \approx 2500$

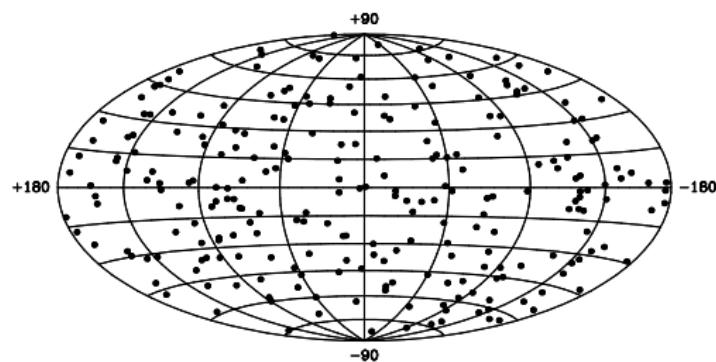
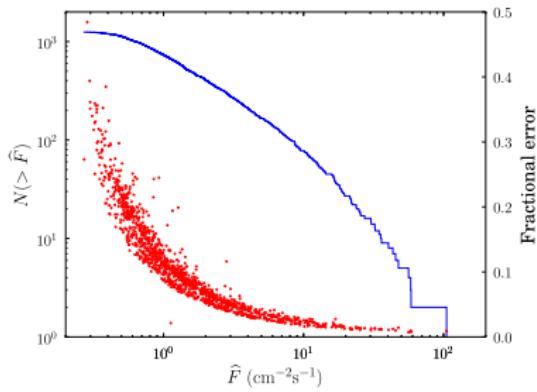
Why not MCMC?

- Tried Metropolis sampler and hybrid (Hamiltonian) Monte Carlo
- Chains mixed slowly
- Brownian bridge convergence test indicated convergence *to an incorrect distribution*
- Sound chains had to be run so long they weren't competitive with accept/reject, considering added complexity of output analysis

For models of modest dimension
MCMC is not a panacea

Hundreds of Parameters... Without MCMC

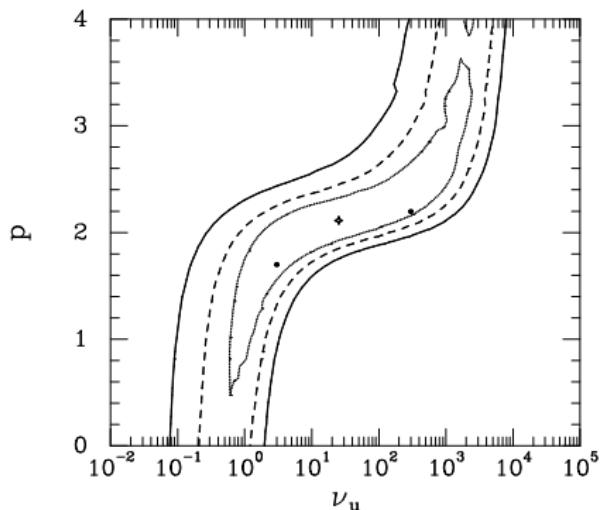
Peak fluxes and directions of GRBs from 4B catalog



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

Modeling

- Marked Poisson point process with measurement error & thinning (MLM)
- Few-parameter GRB luminosity functions (top level)
- Latent flux & direction parameters for 279 or 463 GRBs—*conditionally independent*
- 1-D or 3-D cubatures for latents; adaptive cubature for top-level



Inevitability of MCMC

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*

MCMC Diagnostics

- ➊ My MCMC misadventure
- ➋ Markov chain behavior
- ➌ Posterior sample diagnostics
- ➍ Joint distribution diagnostics
- ➎ Closing advice

The Good News

The Metropolis-Hastings algorithm enables us to draw a few time series realizations $\{\theta_t\}$, $t = 0$ to N , from a Markov chain with a specified stationary distribution $p(\theta)$

The marginal distribution at each time is $p_t(\theta)$

- *Stationarity:* If $p_0(\theta) = p(\theta)$, then $p_t(\theta) = p(\theta)$
- *Convergence:* If $p_0(\theta) \neq p(\theta)$, eventually

$$\|p_t(\theta), p(\theta)\| < \epsilon$$

for an appropriate norm between distributions

- *Ergodicity:*

$$\bar{g} \equiv \frac{1}{N} \sum_i g(\theta_i) \rightarrow \langle g \rangle \equiv \int d\theta g(\theta) p(\theta)$$

The Bad News

- We never have $p_0(\theta) = p(\theta)$: we have to figure out how to initialize a realization, and we are always in the situation where $p_t(\theta) \neq p(\theta)$
- “Eventually” means $t < \infty$; that’s not very comforting!
- After convergence at time $t = c$, $p_t(\theta) \approx p(\theta)$, but θ values at different times are dependent; the Markov chain CLT says

$$\bar{g} \sim N(\langle g \rangle, \sigma^2/N)$$

$$\sigma^2 = \text{var}[g(\theta_c)] + 2 \sum_{k=1}^{\infty} \text{cov}[g(\theta_c), g(\theta_{c+k})]$$

- We have to learn about $p_t(\theta)$ from just a few time series realizations (maybe just one)

MCMC Diagnostics

- ➊ My MCMC misadventure
- ➋ Markov chain behavior
- ➌ Posterior sample diagnostics
- ➍ Joint distribution diagnostics
- ➎ Closing advice

Posterior sample diagnostics

Posterior sample diagnostics use single or multiple chains, $\{\theta_t\}$, to diagnose:

- **Convergence:** How long until starting values are forgotten? (Discard as “burn-in,” or run long enough so averages “forget” initialization bias.)
- **Mixing:** How long until we have fairly sampled the full posterior? (Make finite-sample Monte Carlo uncertainties small.)

Two excellent R packages with routines, descriptions, references:

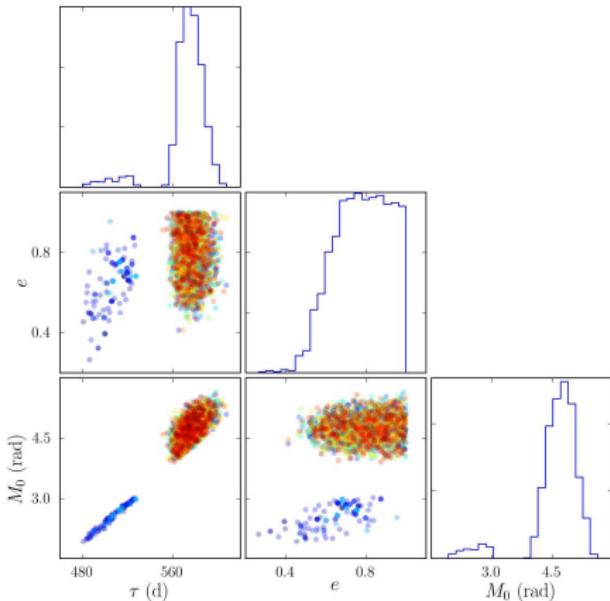
- **boa**
<http://cran.r-project.org/web/packages/boa/index.html>
- **coda**
<http://cran.r-project.org/web/packages/coda/index.html>

They also supply output analysis: estimating means, variances, marginals, HPD regions...

Diagnosing convergence

Qualitative

- Trace plots—trends?
- Diagnostic plots; e.g., running mean
- Color-coded pair plots



Exoplanet parameter
estimation using RV data
from HD 222582

Quantitative

- Gelman-Rubin R : multiple chains, within/between-chain variance (Alan's lecture)
- Geweke: single chain, consistency of early/late means
- Heidelberger & Welch: single chain, checks for brownian motion signature of stationarity, estimates burn-in
- Fan-Brooks-Gelman score statistic:

$$U_k(\theta) = \frac{\partial \log p(\theta)}{\partial \theta_{(k)}}$$

Uses $\langle U_k \rangle_p = 0$

Use diagnostics for all quantities of interest!

Diagnosing mixing

Qualitative

- Trace plots—does chain get stuck?
- Diagnostic plots; e.g., running mean, autocorrelation function

Quantitative

- Batch means (Murali's CASt summer school lab)
- AR and spectral analysis estimators (David's lecture)

MCMC Diagnostics

- ➊ My MCMC misadventure
- ➋ Markov chain behavior
- ➌ Posterior sample diagnostics
- ➍ Joint distribution diagnostics
- ➎ Closing advice

Bayesian Inference and the Joint Distribution

Recall that Bayes's theorem comes from the *joint distribution for data and hypotheses* (parameters/models):

$$\begin{aligned} p(\theta, D|M) &= p(\theta|M) p(D|\theta, M) \\ &= p(D|M) p(\theta|D, M) \end{aligned}$$

Bayesian inference takes $D = D_{\text{obs}}$ and solves RHS for the posterior:

$$\rightarrow p(\theta|D_{\text{obs}}, M) = \frac{p(\theta|M)p(D_{\text{obs}}|\theta, M)}{p(D_{\text{obs}}|M)}$$

MCMC is nontrivial technology for building RNGs to sample θ values from the *intractable posterior*, $p(\theta|D_{\text{obs}}, M)$.

Posterior sampling is hard, but sampling from the other distributions is often easy:

- Often easy to draw θ^* from $\pi(\theta)$
- Typically easy to draw D_{sim} from $p(D|\theta, M)$
- Sample the joint for (θ, D) by sequencing:

$$\theta^* \sim \pi(\theta)$$

$$D_{\text{sim}} \sim p(D|\theta^*, M)$$

- $\{D_{\text{sim}}\}$ from above are samples from

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

Now note that $\{D_{\text{sim}}, \theta\}$ with $\theta \sim p(\theta|D_{\text{sim}}, M)$ are also samples from the joint distribution

Joint distribution methods check the consistency of these two joint samplers to validate a posterior sampler

Example: “Calibration” of credible regions

How often may we expect an HPD region with probability P to include the true value if we analyze many datasets? I.e., what's the frequentist coverage of an interval rule $\Delta(D)$ defined by calculating the Bayesian HPD region each time?

Suppose we generate datasets by picking a parameter value from $\pi(\theta)$ and simulating data from $p(D|\theta)$.

The fraction of time θ will be in the HPD region is:

$$Q = \int d\theta \pi(\theta) \int dD p(D|\theta) \llbracket \theta \in \Delta(D) \rrbracket$$

Note $\pi(\theta)p(D|\theta) = p(\theta, D) = p(D)p(\theta|D)$, so

$$Q = \int dD \int d\theta p(\theta|D) p(D) \llbracket \theta \in \Delta(D) \rrbracket$$

$$\begin{aligned}
Q &= \int dD \int d\theta p(\theta|D) p(D) \llbracket \theta \in \Delta(D) \rrbracket \\
&= \int dD p(D) \int d\theta p(\theta|D) \llbracket \theta \in \Delta(D) \rrbracket \\
&= \int dD p(D) \int_{\Delta(D)} d\theta p(\theta|D) \\
&= \int dD p(D) P \\
&= P
\end{aligned}$$

The HPD region includes the true parameters $100P\%$ of the time.

This is exactly true for any problem, even for small datasets.

Keep in mind it involves drawing θ from the prior; credible regions are “calibrated with respect to the prior.”

A Tangent: Average Coverage

Recall the original Q integral:

$$\begin{aligned} Q &= \int d\theta \pi(\theta) \int dD p(D|\theta) \llbracket \theta \in \Delta(D) \rrbracket \\ &= \int d\theta \pi(\theta) C(\theta) \end{aligned}$$

where $C(\theta)$ is the (frequentist) coverage of the HPD region when the data are generated using θ .

This indicates Bayesian regions have accurate *average coverage*.

The prior can be interpreted as quantifying how much we care about coverage in different parts of the parameter space.

Basic Bayesian Calibration Diagnostics

Encapsulate your sampler: Create an MCMC posterior sampling algorithm for model M that takes data D as input and produces posterior samples $\{\theta_i\}$, and a $100 P\%$ credible region $\Delta_P(D)$.

Initialize counter $Q = 0$.

Repeat $N \gg 1$ times:

- ① Sample a “true” parameter value θ^* from $\pi(\theta)$
- ② Sample a dataset D_{sim} from $p(D|\theta^*)$
- ③ Use the encapsulated posterior sampler to get $\Delta_P(D_{\text{sim}})$ from $p(\theta|D_{\text{sim}}, M)$
- ④ If $\theta^* \in \Delta_P(D)$, increment Q

Check that $Q/N \approx P$

Easily extend the idea to check *all* credible region sizes:

Initialize a list that will store N probabilities, P .

Repeat $N \gg 1$ times:

- ① Sample a “true” parameter value θ^* from $\pi(\theta)$
- ② Sample a dataset D_{sim} from $p(D|\theta^*)$
- ③ Use the encapsulated posterior sampler to get $\{\theta_i\}$ from $p(\theta|D_{\text{sim}}, M)$
- ④ Find P so that θ^* is on the boundary of $\Delta_P(D)$; append to list
[$P = \text{fraction of } \{\theta_i\} \text{ with } q(\theta_i) > q(\theta^*)$]

Check that the P s follow a uniform distribution on $[0, 1]$

Other Joint Distribution Tests

- Geweke 2004: Calculate means of scalar functions of (θ, D) two ways; compare with z statistics
- Cook, Gelman, Rubin 2006: Posterior quantile test, expect $p[g(\theta) > g(\theta^*)] \sim \text{Uniform}$ (HPD test is special case)

What Joint Distribution Tests Accomplish

Suppose the prior and sampling distribution samplers are well-validated.

- **Convergence verification:** If your sampler is bug-free but was not run long enough → unlikely that inferences will be calibrated
- **Bug detection:** An incorrect posterior sampler implementation will not converge to the correct posterior distribution → unlikely that inferences will be calibrated, even if the chain converges

Cost: Prior and data sampling is often cheap, but posterior sampling is often expensive, and joint distribution tests require you run your MCMC code *hundreds* of times

Compromise: If MCMC cost grows with dataset size, running the test with smaller datasets provides a good bug test, and *some* insight on convergence

MCMC Diagnostics

- ➊ My MCMC misadventure
- ➋ Markov chain behavior
- ➌ Posterior sample diagnostics
- ➍ Joint distribution diagnostics
- ➎ Closing advice

The Experts Speak

All the methods can fail to detect the sorts of convergence failure they were designed to identify. We recommend a combination of strategies... it is not possible to say with certainty that a finite sample from an MCMC algorithm is representative of an underlying stationary distribution.

— Cowles & Carlin review of 13 diagnostics

In more than, say, a dozen dimensions, it is difficult to believe that a few, even well-chosen, scalar statistics give an adequate picture of convergence of the multivariate distribution.

— Peter Green 2002

Handbook of Markov Chain Monte Carlo (2011)

Your humble author has a dictum that the least one can do is to make an overnight run. What better way for your computer to spend its time? In many problems that are not too complicated, this is millions or billions of iterations. If you do not make runs like that, you are simply not serious about MCMC. Your humble author has another dictum (only slightly facetious) that one should start a run when the paper is submitted and keep running until the referees' reports arrive. This cannot delay the paper, and may detect pseudo-convergence.

— Charles Geyer

When all is done, compare inferences to those from simpler models or approximations. Examine discrepancies to see whether they represent programming errors, poor convergence, or actual changes in inferences as the model is expanded.

— Gelman & Shirley

Bayesian Computation: Beyond the Basics

Tom Loredo

Dept. of Astronomy, Cornell University

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

Bayesian Computation Tutorials — 11-12 June 2011

Beyond the Basics

① Avoiding random walks in MCMC

Auxiliary variables

Annealing and parallel tempering

Population-based MCMC

Exact sampling

② Computation for model uncertainty

Bayes factors via trans-dimensional MCMC

Marginal likelihood computation

Beyond the Basics

① Avoiding random walks in MCMC

Auxiliary variables

Annealing and parallel tempering

Population-based MCMC

Exact sampling

② Computation for model uncertainty

Bayes factors via trans-dimensional MCMC

Marginal likelihood computation

Random Walks

Metropolis random walk (MRW) and Gibbs sampler updates execute a *random walk* through parameter space:

- Moves are local, with a characteristic scale ℓ
- Total distance traversed over time $t \propto \sqrt{t}$

This is a relatively slow (albeit steady) rate of exploration

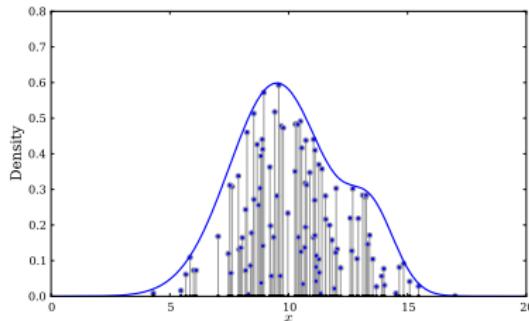
Multimodality → even slower exploration; only rare large jumps can move between modes

We need methods designed to make large moves

Auxiliary variables

The accept/reject method for sampling a d -D density:

- Sample from a *uniform* $(d + 1)$ -D density (with a complicated boundary):



- Report the marginal samples for the d original dimensions

A paradoxical notion motivating some advanced MCMC methods is that making the problem “harder” (higher-dimensional) may actually make it *easier*

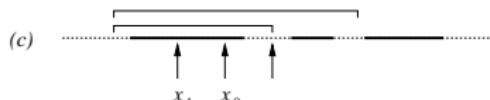
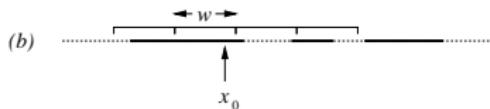
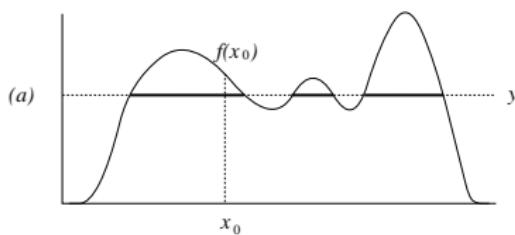
Data augmentation is an example: explicitly adding “missing data” variables → tractable posterior sampling

Slice Sampling

Add a vertical dimension (like rejection), and make a chain that samples uniformly under $p(\theta)$:

- Sample y uniformly over $[0, p(\theta_i)]$ (y given θ)
- Sample θ_{i+1} from dist'n for θ given y

Latter is done sampling uniformly over $\{\theta : y < p(\theta)\}$



Hybrid (Hamiltonian) Monte Carlo

Alan's previous lecture!

Double the dimensionality!

Give samples “momentum” so moves tend to go in the same direction a while; use derivatives to guide the evolution → suppress random walks

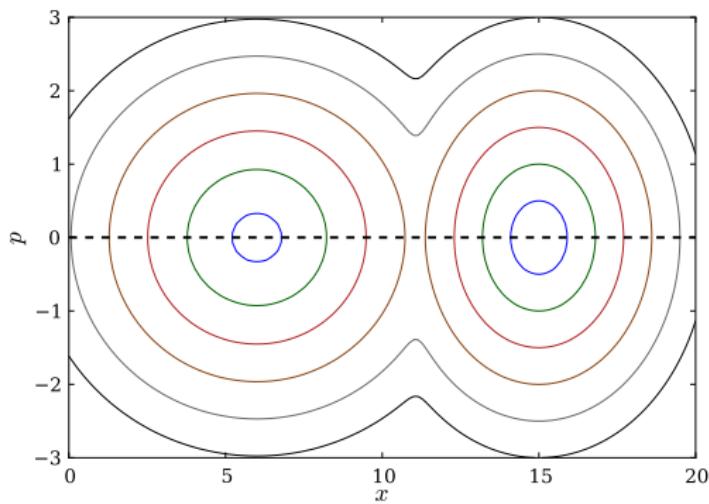
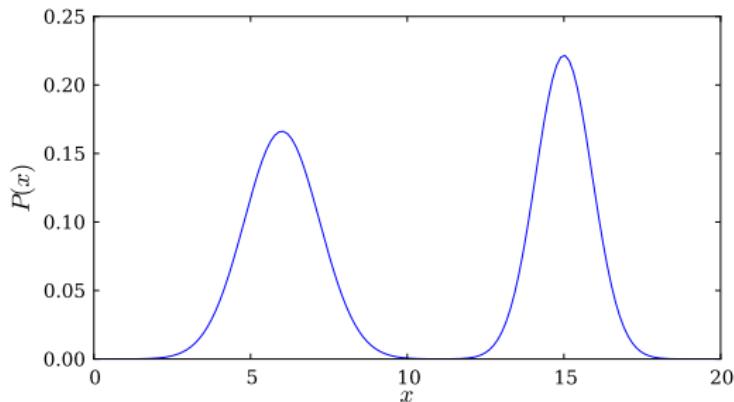
Adds d additional variables, P , with a joint Gaussian dist'n:

$$\log p(\theta, P) = - \left[H(\theta) + \frac{1}{2} P^2 \right]; \quad H(\theta) \equiv -\log p(\theta)$$

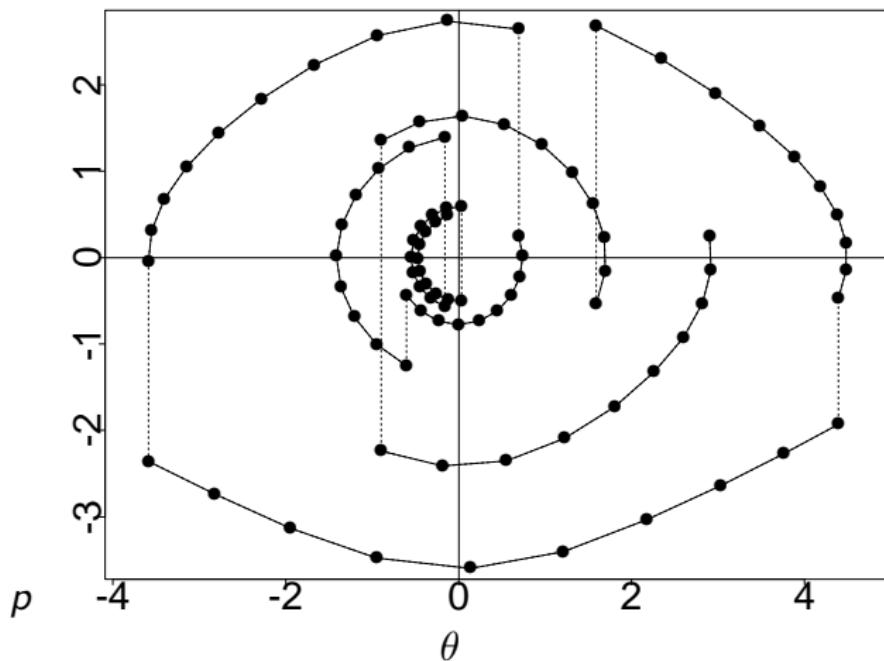
Sample P from a Gaussian, and use it to generate proposals via

$$\dot{\theta} = P; \quad \dot{P} = -\frac{\partial H}{\partial \theta}$$

Hamiltonian dynamics → reversible, preserves volume, keeps p constant (proposals always accepted)



Sampling a 1-D Student- t dist'n with dof= 5

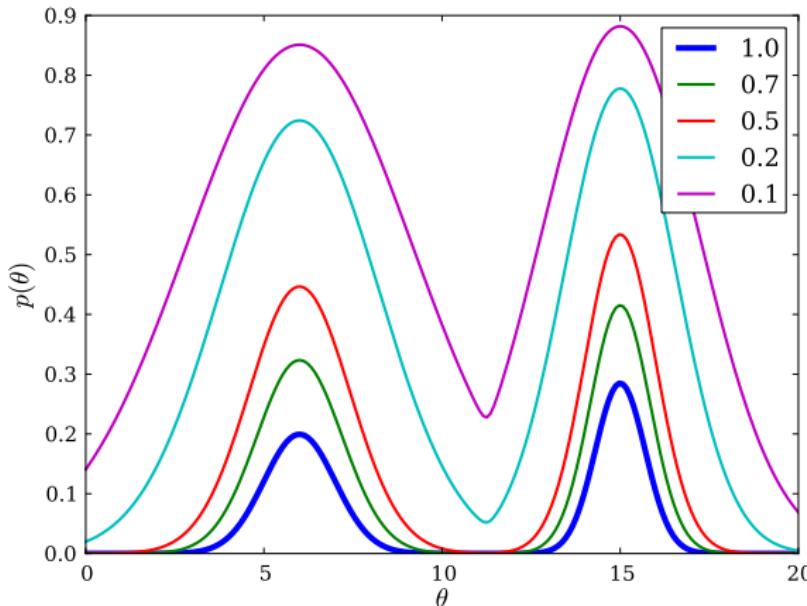


Annealing and Parallel Tempering

PT, aka Metropolis-coupled MCMC

To enable large jumps, *anneal* or *temper* the posterior:

$$q_\beta(\theta) = [q(\theta)]^\beta, \quad \beta \in [0, 1]$$



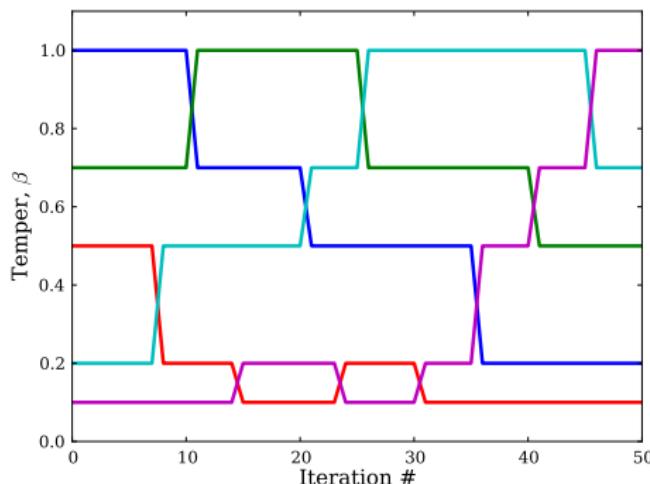
Consider a set of tempers (“inverse temperatures”) $\{\beta_i\}$

Think of each $q_i = q_{\beta_i}$ as its own “model” with its own parameters, and construct a sampler for the joint distribution

$$p(\theta_1, \dots, \theta_m) = \prod_i q_i(\theta_i)$$

Alternate within-temper proposals and swap proposals between adjacent tempers

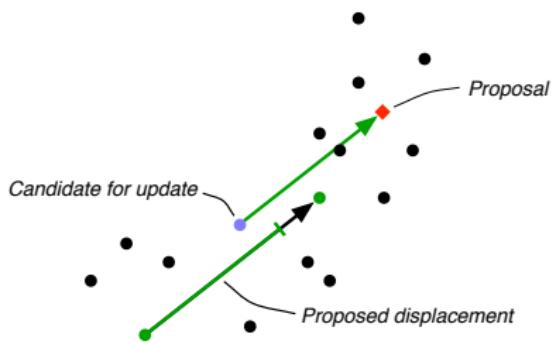
Swaps between tempered chains



Differential Evolution MCMC

Combine evolutionary computing & MCMC (ter Braak 2006)

Follow a *population* of states, where a randomly selected state is considered for updating via the (scaled) vector difference between two other states.



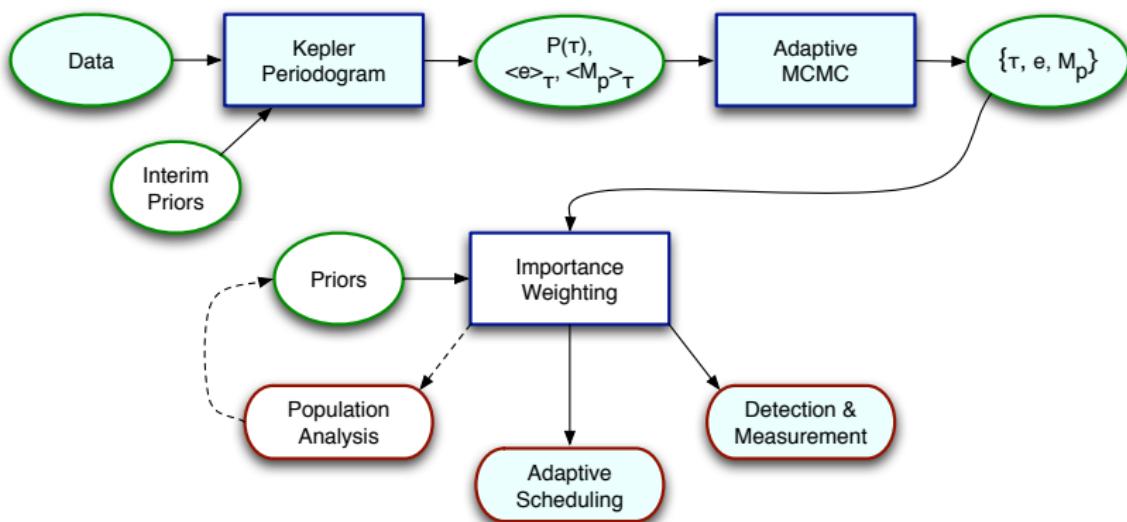
Behaves roughly like RWM, but with a proposal distribution that automatically adapts to shape & scale of posterior

Step scale: Optimal $\gamma \approx 2.38/\sqrt{2d}$, but occasionally switch to $\gamma = 1$ for mode-swapping

Original DE-MCMC uses these simple moves and pop'n size
 $N \sim 3d$; works well if given a “smart start”

Later version (ter Braak & Vrugt 2008) adds new moves and can sample effectively with just $N = 3$ in up to a few dozen dimensions, without a smart start

Bayesian Exoplanet RV Data Analysis Pipeline



Exact Sampling

aka Perfect Sampling

Consider chain as a random mapping: $\theta_{t+1} = F(\theta_t, u_t)$ for input uniform random draws u_t .

Run *coupled* chains: Start a chain in *each* state, and evolve them all with the *same* random numbers. This chain will coalesce at a random time, T . It has forgotten its starting point.

The coalesced point is *not* a fair sample; e.g., if there is no way to get from state 3 to state 4, the chain will never coalesce in state 4, though the coalesced chain will eventually pass through state 4 (via other states).

Propp & Wilson showed how to cure this bias by running the chain *from the past*: Coupling from the past. Others have extended and generalized this (Fill's algorithm, etc.).

Exact sampling can produce *independent* samples (though costly).

Only applies readily to spaces with special structure.

Beyond the Basics

① Avoiding random walks in MCMC

Auxiliary variables

Annealing and parallel tempering

Population-based MCMC

Exact sampling

② Computation for model uncertainty

Bayes factors via trans-dimensional MCMC

Marginal likelihood computation

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

Several approaches: 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)

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!

Marginal likelihood computation

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 “marginal likelihood identity”):

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

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

Marginal likelihood computation

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

- ① Get posterior samples
- ② Use a density estimator to estimate $p(\theta|D, M)$ at some θ^*
(probably near the mode is best)
- ③ Evaluate the formula $\rightarrow \hat{Z}$

Fails in more than very few dimensions because of the curse of dimensionality for density estimation (yesterday's lecture)

(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

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 ϕ_0 = value of ϕ 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:

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

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

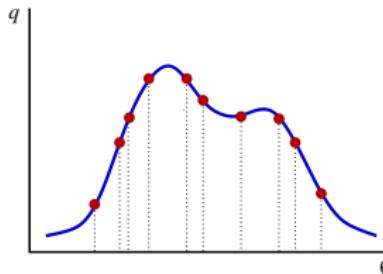
$$\begin{aligned} \mathcal{L}(M_1) &= \int d\theta f(\theta) g(\phi_0) \mathcal{L}_2(\theta, \phi_0) \times \frac{1}{g(\phi_0)} \\ &= \frac{p(\phi_0|D, M_2)}{p(\phi_0|M_2)} \mathcal{L}(M_2) \\ \rightarrow B_{21} &= \frac{p(\phi_0|M_2)}{p(\phi_0|D, M_2)} \sim \begin{cases} \text{small for broad } \phi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

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

Adaptive Simplex Cubature

Another bad method!

Motivation: Use MCMC sample locations *and densities*

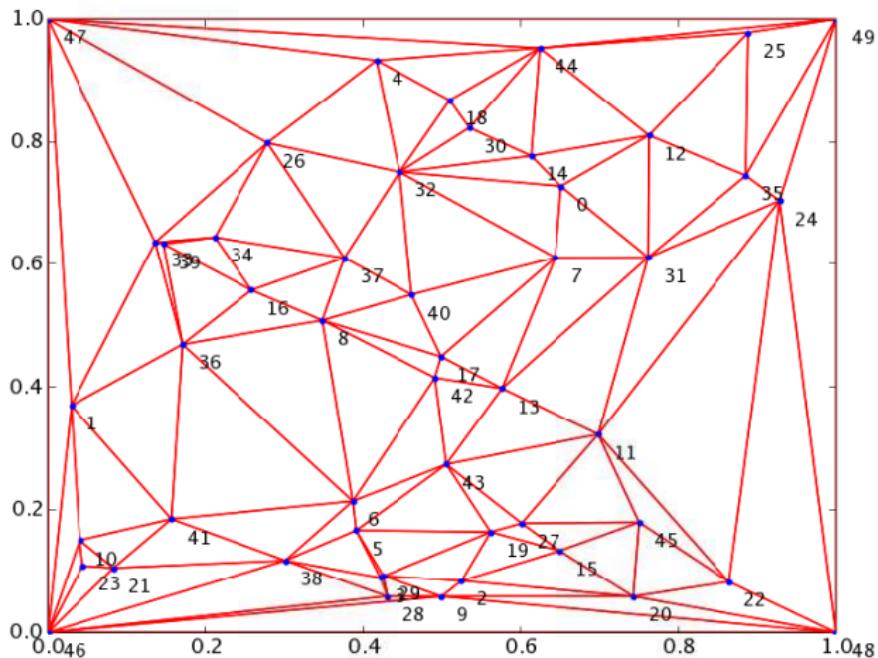


Suppose you were given $\{\theta_i, q_i\}$ and told to estimate $Z = \int d\theta q(\theta)$ for this 1-d q .

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 → area. Make the triangles via *Delaunay triangulation*.



Higher dimensions: Combine n -d Delaunay triangulation and n -d simplex trapezoidal rules of Lyness & Genz

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

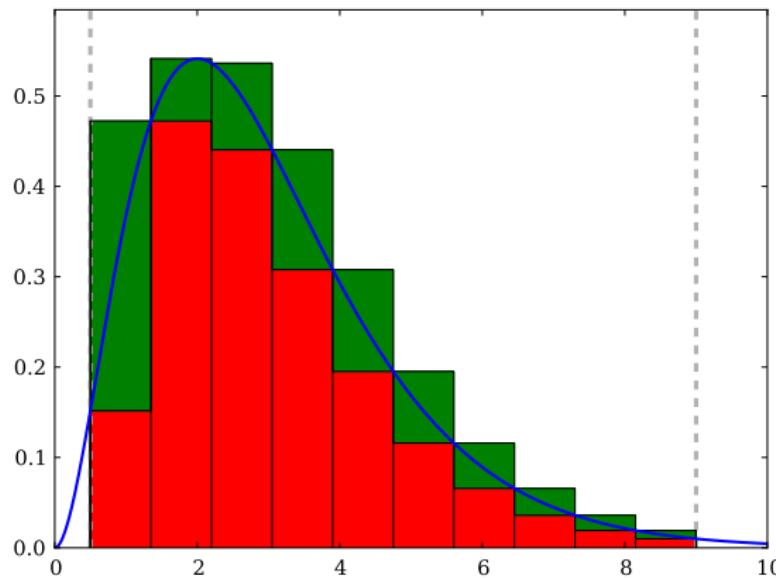
- Tempered/derivative-weighted resampling (seems to work to 6- or 7-D)
- Non-optimal triangulations

Lebesgue Integration and Nested Sampling

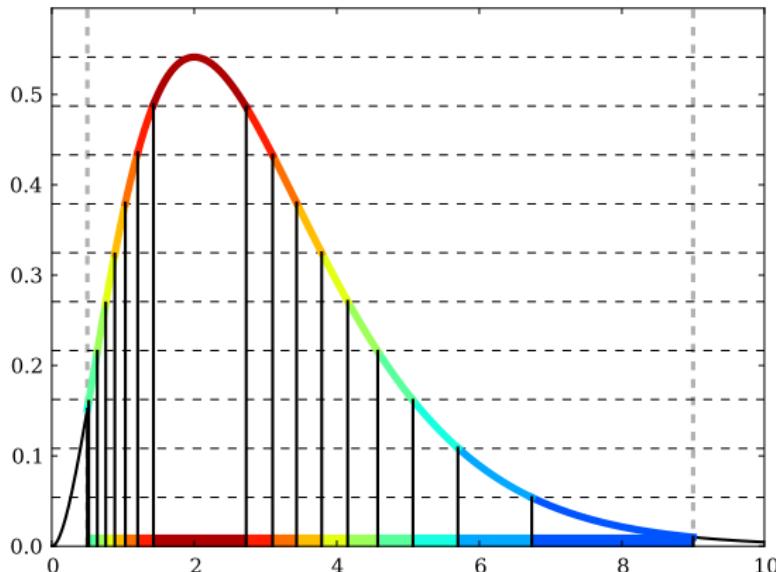
Adaptive simplex quadrature implements a Riemann integral in d -D

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

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

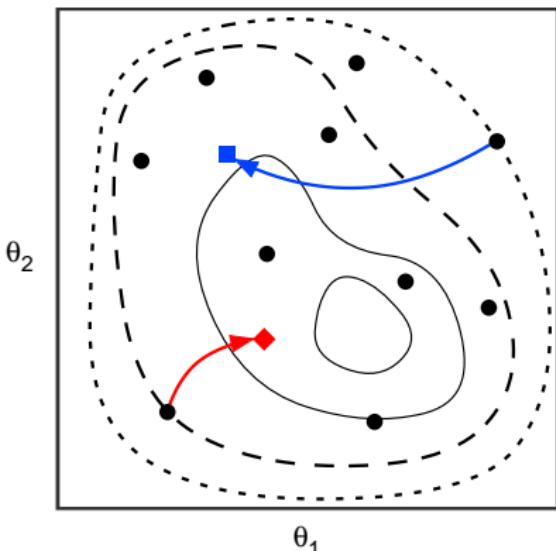
$$Z \approx \sum_i f_i \mu_i(x, y)$$

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*

Many Other Methods

- Thermodynamic integration: Based on annealing and expectations
- Posterior expectation methods: Chib; Gelfand & Dey; “harmonic mean” (bad!) — use posterior samples and identities
- Adaptive importance sampling (Liu⁺ 2011)

Tools for Computational Bayes

Astronomer/Physicist Tools

- **BIE** <http://www.astro.umass.edu/~weinberg/BIE/>
Bayesian Inference Engine: General framework for Bayesian inference, tailored to astronomical and earth-science survey data. Built-in database capability to support analysis of terabyte-scale data sets. Inference is by Bayes via MCMC.
- **CIAO/Sherpa** <http://cxc.harvard.edu/sherpa/>
On/off marginal likelihood support, and Bayesian Low-Count X-ray Spectral (BLoCXS) analysis via MCMC via the **pyblocxs** extension
<https://github.com/brefsdal/pyblocxs>
- **XSpec** <http://heasarc.nasa.gov/xanadu/xspec/>
Includes some basic MCMC capability
- **CosmoMC** <http://cosmologist.info/cosmomc/>
Parameter estimation for cosmological models using CMB, etc., via MCMC
- **MultiNest** <http://ccpforge.cse.rl.ac.uk/gf/project/multinest/>
Bayesian inference via an approximate implementation of the nested sampling algorithm
- **ExoFit** <http://www.homepages.ucl.ac.uk/~ucapola/exofit.html>
Adaptive MCMC for fitting exoplanet RV data
- **extreme-deconvolution**
<http://code.google.com/p/extreme-deconvolution/>
Multivariate density estimation with measurement error, via a multivariate normal finite mixture model; partly Bayesian; Python & IDL wrappers

Astronomer/Physicist Tools, cont'd...

- **root/RooStats** <https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome>
Statistical tools for particle physicists; Bayesian support being incorporated
- **CDF Bayesian Limit Software**
http://www-cdf.fnal.gov/physics/statistics/statistics_software.html
Limits for Poisson counting processes, with background & efficiency uncertainties
- **SuperBayeS** <http://www.superbayes.org/>
Bayesian exploration of supersymmetric theories in particle physics using the MultiNest algorithm; includes a MATLAB GUI for plotting
- **CUBA** <http://www.feynarts.de/cuba/>
Multidimensional integration via adaptive cubature, adaptive importance sampling & stratification, and QMC (C/C++, Fortran, and Mathematica; R interface also via 3rd-party R2Cuba)
- **Cubature** <http://ab-initio.mit.edu/wiki/index.php/Cubature>
Subregion-adaptive cubature in C, with a 3rd-party R interface; intended for low dimensions (< 7)
- **APEMoST** <http://apemost.sourceforge.net/doc/>
Automated Parameter Estimation and Model Selection Toolkit in C, a general-purpose MCMC environment that includes parallel computing support via MPI; motivated by asteroseismology problems
- **Inference** Forthcoming at <http://inference.astro.cornell.edu/>
Several self-contained Bayesian modules; Parametric Inference Engine

Python

- **PyMC** <http://code.google.com/p/pymc/>
A framework for MCMC via Metropolis-Hastings; also implements Kalman filters and Gaussian processes. Targets biometrics, but is general.
- **SimPy** <http://simpy.sourceforge.net/>
SimPy (rhymes with "Blimpie") is a process-oriented public-domain package for discrete-event simulation.
- **RSPython** <http://www.omegahat.org/>
Bi-directional communication between Python and R
- **MDP** <http://mdp-toolkit.sourceforge.net/>
Modular toolkit for Data Processing: Current emphasis is on machine learning (PCA, ICA...). Modularity allows combination of algorithms and other data processing elements into "flows."
- **Orange** <http://www.ailab.si/orange/>
Component-based data mining, with preprocessing, modeling, and exploration components. Python/GUI interfaces to C++ implementations. Some Bayesian components.
- **ELEFANT** <http://rubis.rsise.anu.edu.au/elefant>
Machine learning library and platform providing Python interfaces to efficient, lower-level implementations. Some Bayesian components (Gaussian processes; Bayesian ICA/PCA).

R and S

- **CRAN Bayesian task view**

<http://cran.r-project.org/web/views/Bayesian.html>

Overview of many R packages implementing various Bayesian models and methods; pedagogical packages; packages linking R to other Bayesian software (BUGS, JAGS)

- **Omega-hat** <http://www.omegahat.org/>

RPython, RMatlab, R-Xlisp

- **BOA** <http://www.public-health.uiowa.edu/boa/>

Bayesian Output Analysis: Convergence diagnostics and statistical and graphical analysis of MCMC output; can read BUGS output files.

- **CODA**

<http://www.mrc-bsu.cam.ac.uk/bugs/documentation/coda03/cdaman03.html>

Convergence Diagnosis and Output Analysis: Menu-driven R/S plugins for analyzing BUGS output

- **R2Cuba**

<http://w3.jouy.inra.fr/unites/miaj/public/logiciels/R2Cuba/welcome.html>

R interface to Thomas Hahn's Cuba library (see above) for deterministic and Monte Carlo cubature

Java

- **Hydra** <http://research.warnes.net/projects/mcmc/hydra/>
HYDRA provides methods for implementing MCMC samplers using Metropolis, Metropolis-Hastings, Gibbs methods. In addition, it provides classes implementing several unique adaptive and multiple chain/parallel MCMC methods.
- **YADAS** <http://www.stat.lanl.gov/yadas/home.html>
Software system for statistical analysis using MCMC, based on the multi-parameter Metropolis-Hastings algorithm (rather than parameter-at-a-time Gibbs sampling)
- **Omega-hat** <http://www.omegahat.org/>
Java environment for statistical computing, being developed by XLisp-stat and R developers

C/C++/Fortran

- **BayeSys 3** <http://www.inference.phy.cam.ac.uk/bayesys/>
Sophisticated suite of MCMC samplers including transdimensional capability, by the author of MemSys
- **fbm** <http://www.cs.utoronto.ca/~radford/fbm.software.html>
Flexible Bayesian Modeling: MCMC for simple Bayes, nonparametric Bayesian regression and classification models based on neural networks and Gaussian processes, and Bayesian density estimation and clustering using mixture models and Dirichlet diffusion trees
- **BayesPack, DCUHRE**
<http://www.sci.wsu.edu/math/faculty/genz/homepage>
Adaptive quadrature, randomized quadrature, Monte Carlo integration
- **BIE, CDF Bayesian limits, CUBA** (see above)

Other Statisticians' & Engineers' Tools

- **BUGS/WinBUGS** <http://www.mrc-bsu.cam.ac.uk/bugs/>
Bayesian Inference Using Gibbs Sampling: Flexible software for the Bayesian analysis of complex statistical models using MCMC
- **OpenBUGS** <http://mathstat.helsinki.fi/openbugs/>
BUGS on Windows and Linux, and from inside the R
- **JAGS** <http://www-fis.iarc.fr/~martyn/software/jags/>
"Just Another Gibbs Sampler;" MCMC for Bayesian hierarchical models
- **XLisp-stat** <http://www.stat.uiowa.edu/~luke/xls/xlsinfo/xlsinfo.html>
Lisp-based data analysis environment, with an emphasis on providing a framework for exploring the use of dynamic graphical methods
- **ReBEL** <http://choosh.csee.ogi.edu/rebel/>
Library supporting recursive Bayesian estimation in Matlab (Kalman filter, particle filters, sequential Monte Carlo).