Bayesian Statistical Methods for Astronomy

Part II: Markov Chain Monte Carlo

David A. van Dyk

Statistics Section, Imperial College London

Center for Astrophysics | Harvard & Smithsonian
February 2025

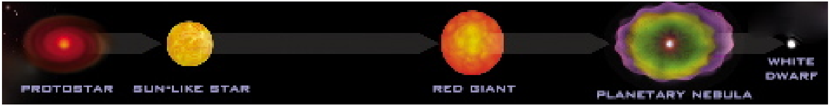
Outline

Outline

Stellar Formation

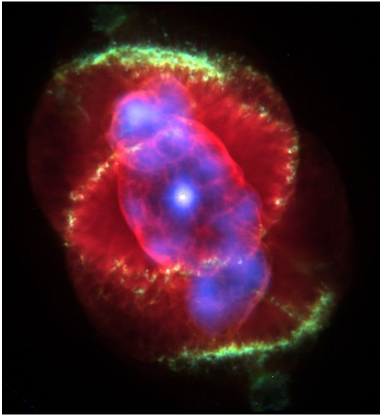
Stars form when the dense parts of a molecular cloud collapse into a ball of plasma.

Evolution of a Sun-like Star



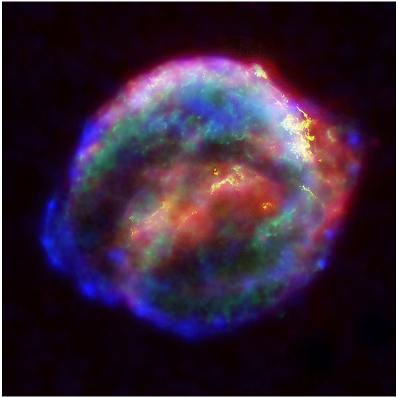
- Hydrogen fusion can last for millions or billions of years, depending on the initial stellar mass.
- When the Hydrogen in the core is depleted, the star may fuse Helium into heavier elements
- At the same time the star goes through dramatic physical changes, growing and cooling into a *red giant* star.
- Soon the star undergoes mass loss forming a *planetary nebula*.
- Eventually only the core is left, a *white dwarf star*.

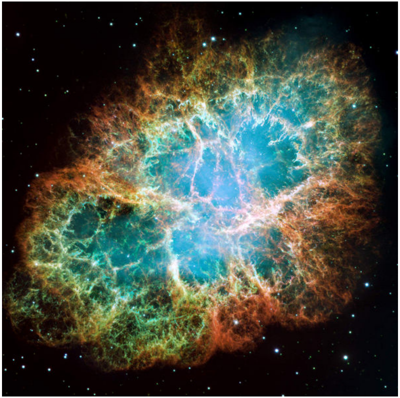
Planetary Nebula



Planetary Nebulae are the illuminated, expanding atmospheres of red giants as they lose the bulk of their mass to become white dwarfs.

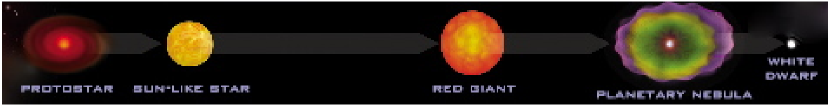
Supernovae





Supernovae are dramatically exploding Giant Stars and may result in *neutron stars* or *black holes*.

Stellar Characteristics



Six Unknown Parameters Affect a Star's Appearance as it Ages

1. More *massive* stars are denser, hotter, bluer, and burn their fuel much more quickly.

- Composition also effects the color spectrum
  2. *"Metals"* absorb more blue light.
  3. Excess *Helium* at the core reduces the efficiency of the nuclear reaction.

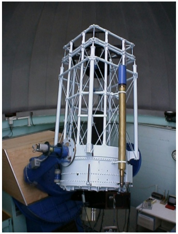
4. The spectrum of the star changes as the star *ages*.
5. Some light from a star is *absorbed* by interstellar material.
6. More *distant* stars are fainter.

Data Collection

Photometric Magnitudes

- To fit the parameters, we study light emitted by each star.
- Using filters, we measure the luminosity of a star's electro-magnetic radiation in several wide wavelength bands.
- Computationally-expensive, physics-based computer models predict magnitudes given the six parameters.

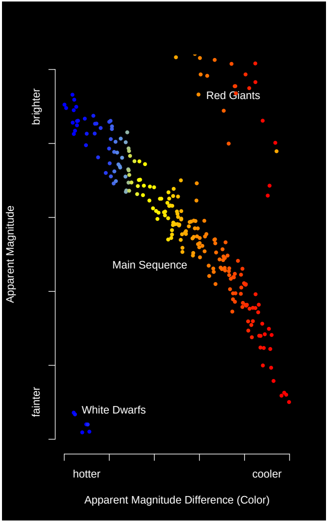
GOAL: Use data to learn about the six stellar parameters.



Stellar Clusters

- Stellar Clusters are physical groups of stars formed at the same time out of the same material.
- Cluster stars have the same *metallicity, helium abundance, age, distance,* and *absorption*.
- We call these five common parameters *cluster parameters*.
- Only the stars' *initial masses* vary.
- This significantly simplifies statistical analysis.

Color Magnitude Diagrams



Hertzsprung-Russell Diagrams

- *Apparent Magnitude* vs *Magnitude Difference (Color)*.
- Identifies stars at different stages of their lives.
- Evolution of an HR diagram.

Basic Likelihood Function

The Stellar Evolution Model as Part of a Complex Likelihood

- The computer model predicts observed magnitudes as a function of mass, M_i, and cluster parameters, $\mathbf{\Theta}$:

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

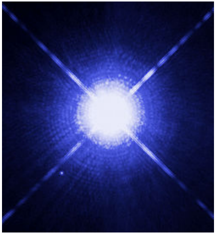
- We assume independent Gaussian errors with known variances:

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

- We use the computer model as a component of a principled statistical analysis.

Binary Star Systems

- Between 1/3 and 1/2 of stars are actually binary star systems.
- Most are unresolved.
- The luminosities of the component stars sum.



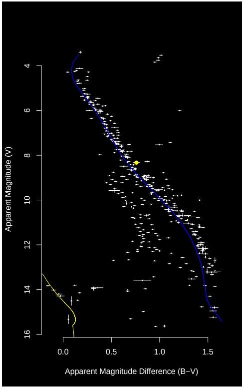
- Resulting offset on the CMD is informative for the masses.
- The expected observed magnitudes for binaries are

$$-2.5 \log_{10}\left[10^{-\boldsymbol{G}(M_{i1},\boldsymbol{\Theta})/2.5} + 10^{-\boldsymbol{G}(M_{i2},\boldsymbol{\Theta})/2.5}\right].$$

- The "secondary masses" of single stars are zero.

Field Stars

Field Stars appear in the field of view but are not part of cluster.



- Their magnitudes do not follow the pattern of the CMD.
- More distant stars are dimmer and below main sequence.
- We use a mixture model.
- Field star magnitudes are assumed uniform over the range of the data.

Likelihood Function

- The resulting Likelihood function is

$$L(\boldsymbol{M}, \Theta, \boldsymbol{Z}|\boldsymbol{X}) =$$
$$\prod_{i=1}^{N}\prod_{j=1}^{n}\left[\frac{Z_i}{\sqrt{2\pi\sigma_{ij}^2}}\exp\left(-\frac{1}{2\sigma_{ij}^2}\left\{x_{ij}+2.5\log_{10}\left[10^{\frac{-G_j(M_{i1},\Theta)}{2.5}}+10^{\frac{-G_j(M_{i2},\Theta)}{2.5}}\right]\right\}^2\right)\right.$$
$$\left. + (1-Z_i)p_{\text{field}}(\boldsymbol{X}_i)\right],$$

  where Z_i is an indicator for cluster membership for star i.
- We embed the computer models into a principled Likelihood-based analysis, rather than using “chi-by-eye”.

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. Stenning, Wagner-Kaiser, Robinson, van Dyk, von Hippel, Sarajedini, Stein, Bayesian Analysis of Stellar Populations in Galactic Globular Clusters I: Statistical and Computational Methods. *ApJ*, **826**, 41, 2016
3. Si, van Dyk, von Hippel, Robinson, Jeffery, and Stenning Bayesian Hierarchical Modelling of Initial–Final Mass Relations Across Star Clusters. *Monthly Notices of the Royal Astronomical Society*, **480**, 1300, 2018

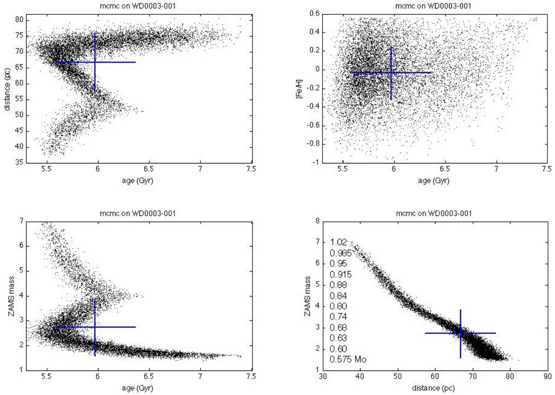
Prior Distributions

We use both informative and non-informative prior distributions:

- An informative truncated Gaussian is used on log mass, representing the population distribution of stellar masses.
- The ratio of the smaller and larger mass is uniform.
- For well studied clusters there are informative star-by-star priors on cluster membership.
- A mildly informative population-based prior is used for age.
- The remaining cluster parameters must be considered on a case-by-case basis.

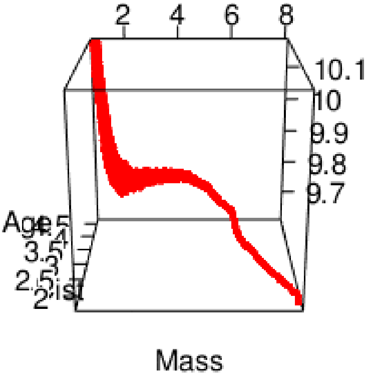
Use sophisticated computational techniques to evaluate the computer model and to fit the resulting model.

Complex Posterior Distributions



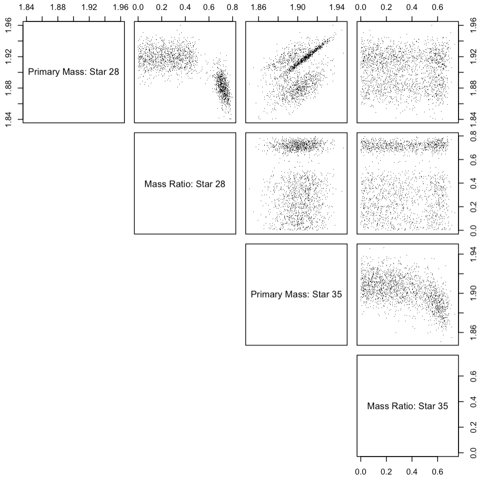
Highly non-linear relationship among stellar parameters.

Complex Posterior Distributions



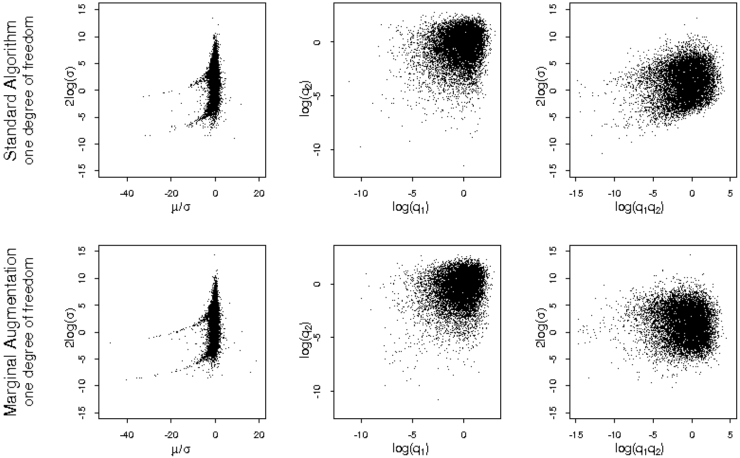
Highly non-linear relationships among stellar parameters.

Complex Posterior Distributions

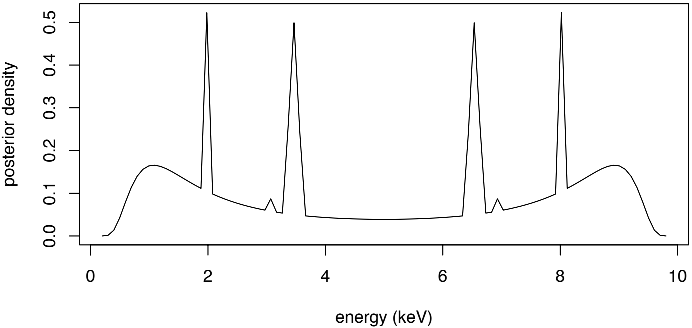


The classification of certain stars as field or cluster stars can cause multiple modes in the distributions of other parameters.

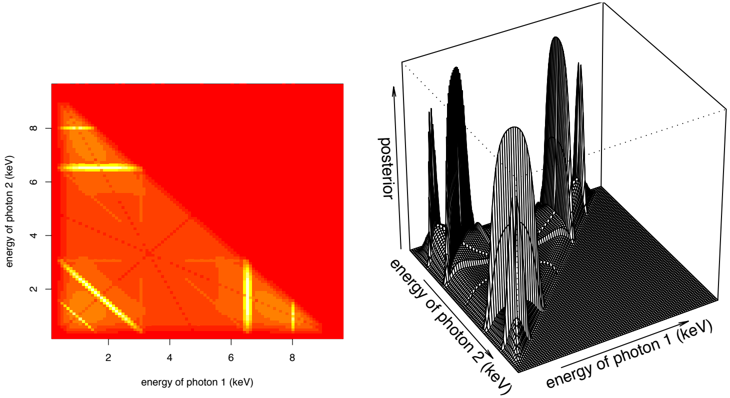
Complex Posterior Distributions



Complex Posterior Distributions

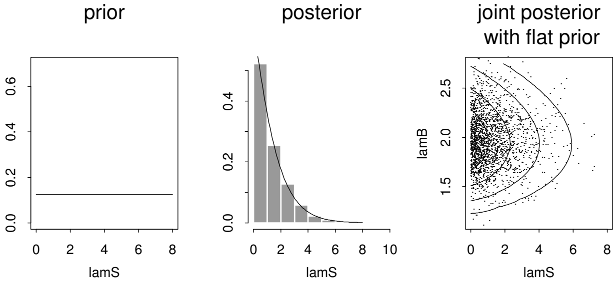


Complex Posterior Distributions



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

$$I = \mathrm{E}[g(\theta)] = \int g(\theta) f(\theta) d\theta,$$

where $f(\theta)$ is a probability density function.
If we have a sample

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

we can estimate I with

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

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 chains can be used to obtain a dependent sample.
- In a Bayesian context, we typically aim to sample the *posterior* distribution.

We first discuss an independent method:

Rejection Sampling & The Grid Method

Rejection Sampling

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

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

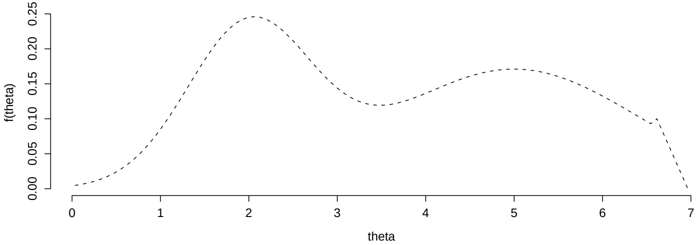
for some M.

1. Sample $\tilde{\theta} \sim g(\theta)$.
2. Sample $u \sim Unif(0, 1)$.
3. 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.

How do we compute M?

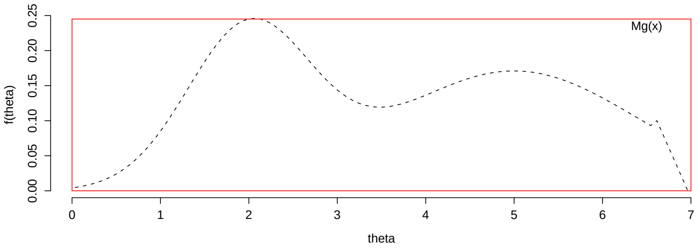
Rejection Sampling

Consider the distribution:



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

Rejection Sampling



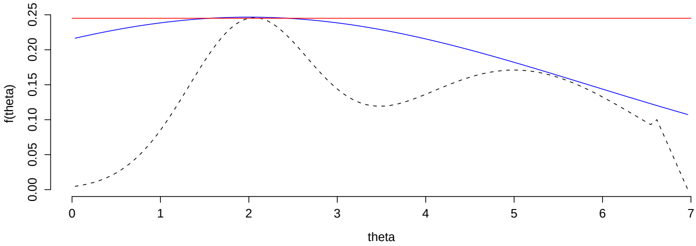
- Imagine that we sample uniformly in the red rectangle:

$$\theta \sim g(\theta) \text{ and } y = uMg(\theta)$$

- Accept samples that fall below the dashed density function.

How can we reduce the wait for acceptance??

Rejection Sampling

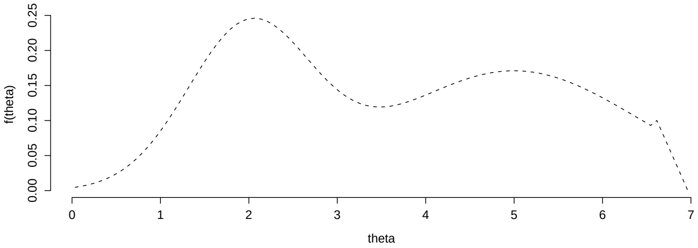


How can we reduce the wait for acceptance??

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

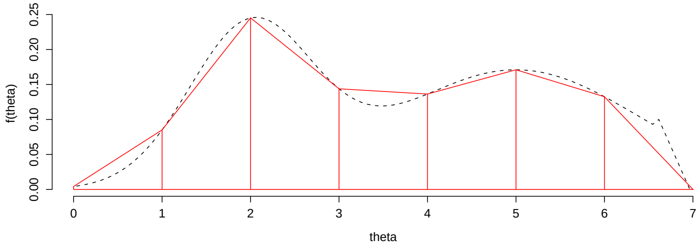
The Grid Method

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



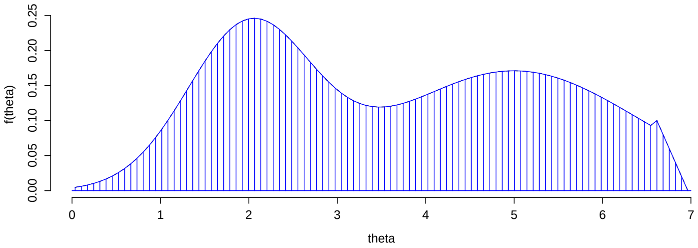
The Grid Method

1. Evaluate the density on a grid.
2. Compute the areas of the resulting trapezoids.
3. 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?

Definition

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 $U^{(1)}, U^{(2)}, \dots$ independent.

What is a Stationary Distribution?

Definition

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

1. have a sample from the stationary dist'n and
2. update the Markov chain,

then the next iterate also follows the stationary dist'n.

In practice we cannot obtain even one sample for the stationary dist'n.

What does a Markov Chain at Stationarity Deliver?

Under regularity conditions, the density at iteration t,

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

[Where F is the cummulative distributuion function with density f.]

- The Markov chain converges to its stationary distribution.
- After sufficient burn-in, we treat $\{\theta^{(t)}, t = N_0, \ldots, N\}$ as a *correlated* sample from the stationary distribution.
- This is an *approximation*: Use MCMC samples with care!
- Convergence diagnostics are critical.

We aim to find a Markov Chain with Stationary Dist'n equal to the Target Dist'n.

Outline

The Metropolis Sampler

Metropolis Sampler

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

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

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

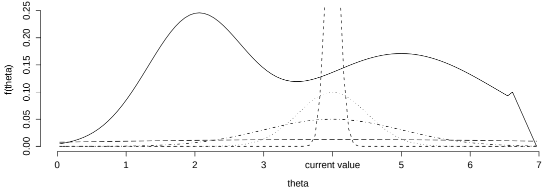
- J_t must be symmetric: $J_t(\theta^*|\theta^{(t-1)}) = J_t(\theta^{(t-1)}|\theta^*)$.
- If $p(\theta^*|y) > p(\theta^{(t-1)}|y)$, *jump!*

The Random Walk Jumping Rule

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

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

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



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

An Example: High-Energy Spectral Analysis

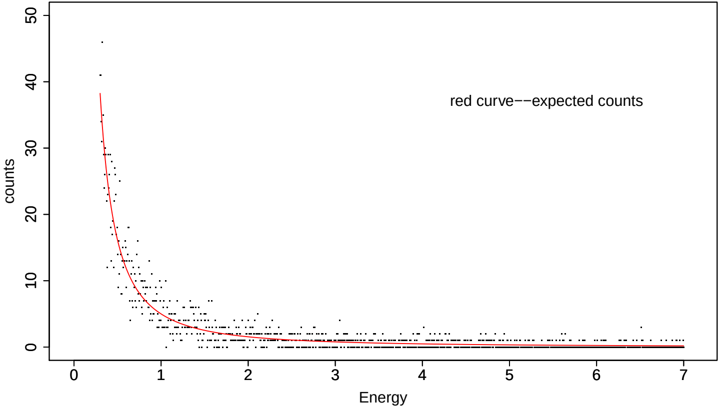
A simplified model for high-energy spectral analysis.

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

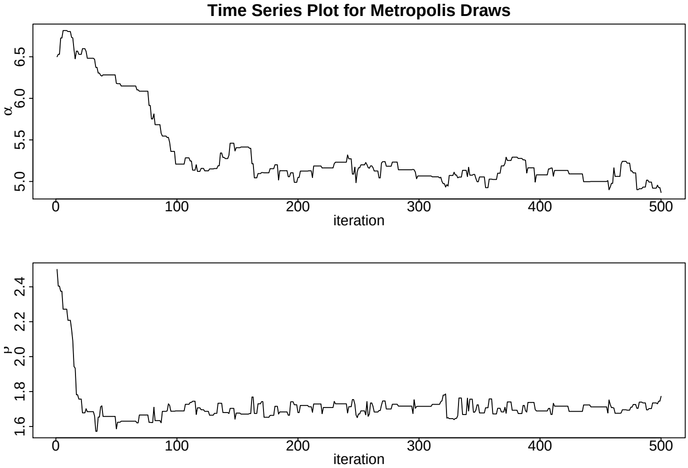
- The Sampler:
  We use a Gaussian Jumping Rule,
  - centered at the current sample, $\theta^{(t)}$
  - with standard deviations equal 0.08 and correlation zero.

Simulated Data

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

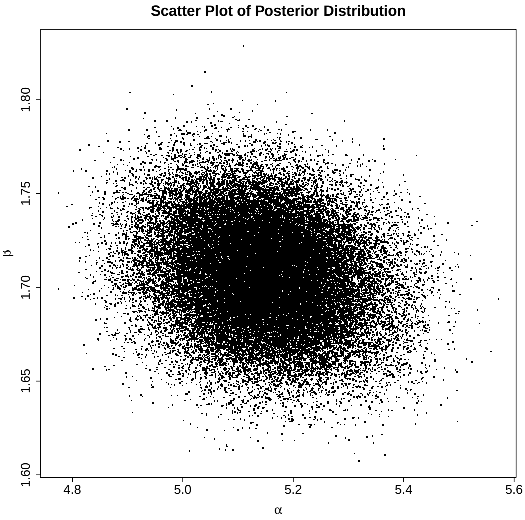


Markov Chain Trace Plots

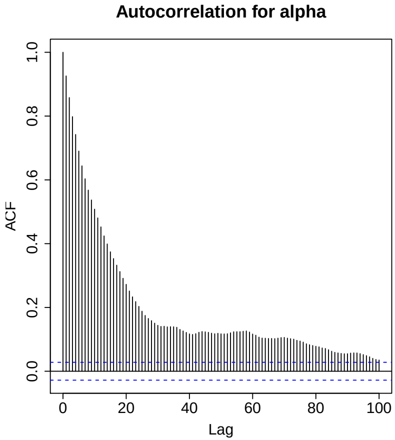


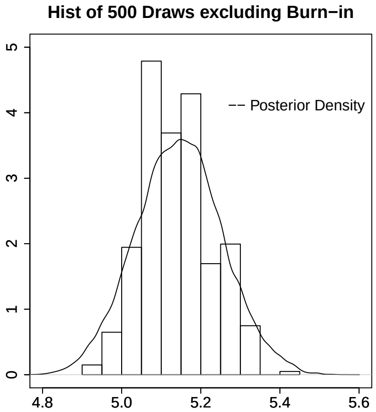
Chains "stick" at a particular draw when proposals are rejected.

The Joint Posterior Distribution



Marginal Posterior Dist'n of the Normalization

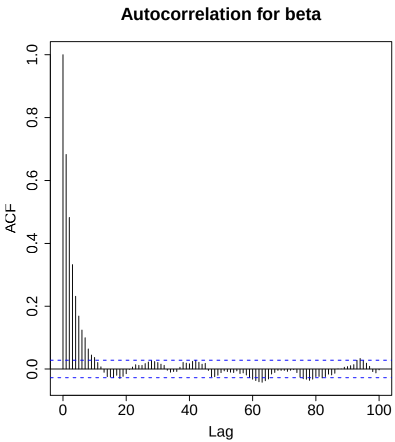


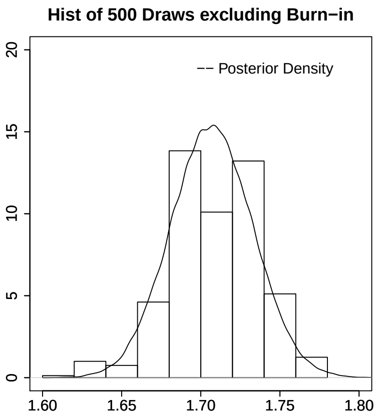


$\mathrm{E}(\alpha|Y) \approx 5.13, \ \mathrm{SD}(\alpha|Y) \approx 0.11,$ and a 95% CI is $(4.92, 5.41)$

... how does this compare with the true value? bias?

Marginal Posterior Dist'n of Power Law Param





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

... how does this compare with the true value? bias?

The Metropolis-Hastings Sampler

Metropolis-Hastings Sampler

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

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

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

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

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

Note

- *A more generic jumping rule:* 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:

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

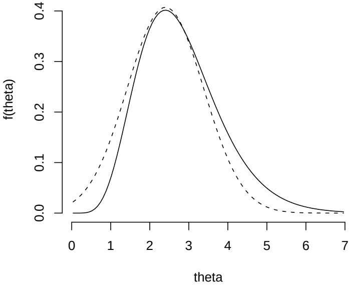
$$\text{MAP estimate} = \text{argmax}_\theta p(\theta|y)$$

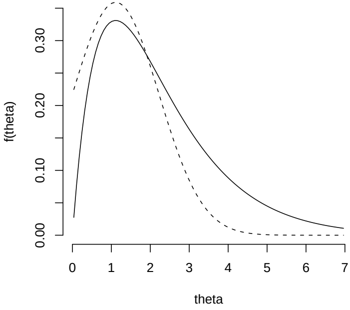
$$\text{Variance} \approx \left[-\frac{\partial^2}{\partial\theta \cdot \partial\theta} \log p(\theta|Y)\right]^{-1}$$

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`:

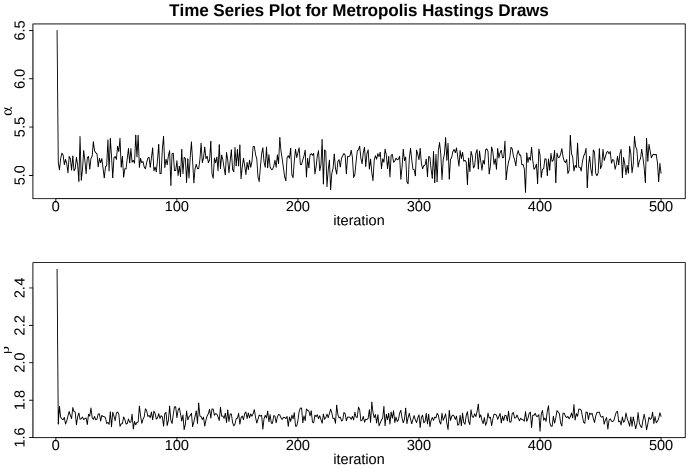
```
> glm.fit = glm( Y~I(-log(E)), family=poisson(link="log") )
> glm.fit$coef     #### best fit of (log(alpha), beta)
> vcov( glm.fit ) #### variance-covariance matrix
```

- Returns MLE of $(\log(\alpha), \beta)$ and variance-covariance matrix.

Example of Independence Sampler

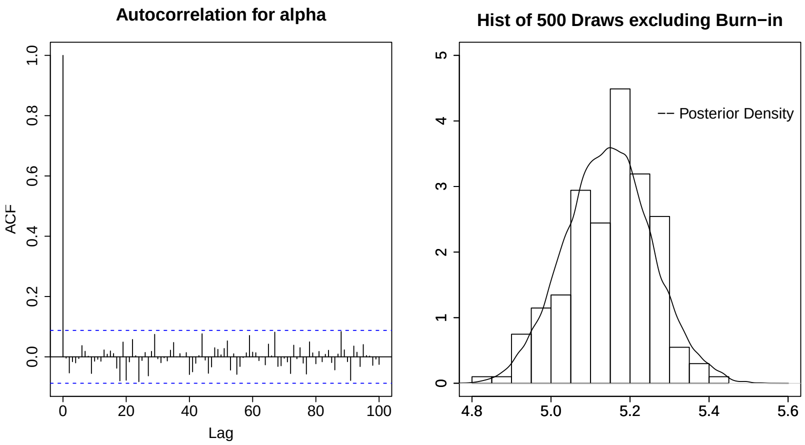
- Alternatively, we can fit (α, β) directly with a general (but less stable) mode finder.
- Requires coding likelihood, specifing starting values, etc.
- Choose parameterization to improve Gaussian approx.
  - MLE is invariant to transformations.
  - Variance matrix of transform is computed via *delta method*.
- We use the general mode finder:
  $J_t =$ Normal$_2$(MAP est, Curvature-based Variance Matrix).

Markov Chain Trace Plots



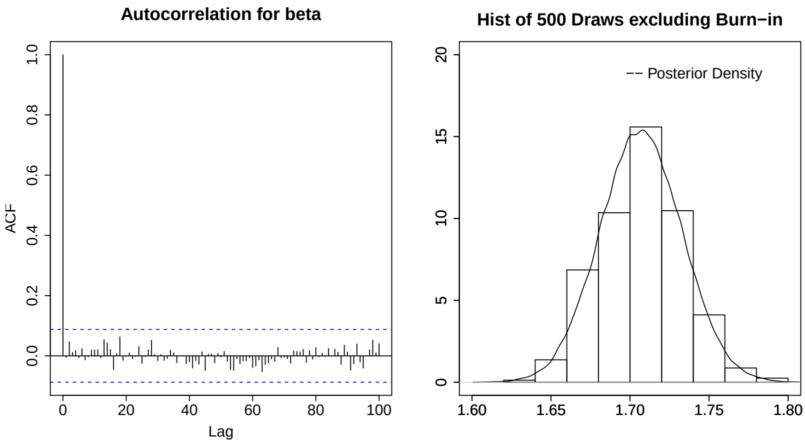
Very little “sticking” here: acceptance rate is 98.8%.

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

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

In this case, we say the Markov chain is ergodic and:

1. $p_{ij}(t) \to \pi(j)$ as $t \to \infty$ for all i and j.
2. $$\Pr\left[\frac{1}{n}\sum_{t=1}^{n} h(\theta^{(t)}) \to \mathrm{E}_\pi(h(\theta))\right] = 1$$

Convergence to Stationarity

Definitions:

1. 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, $\min\{t > 0 : \theta^{(t)} = i | \theta^{(0)} = i\}$.

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

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

So we need our chain to

1. be irreducible,
2. be aperiodic, and
3. have the posterior distribution as a stationary distribution.

Outline

Has this Chain Converged?

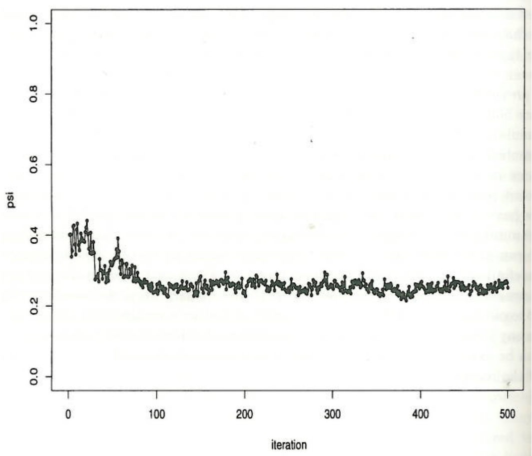
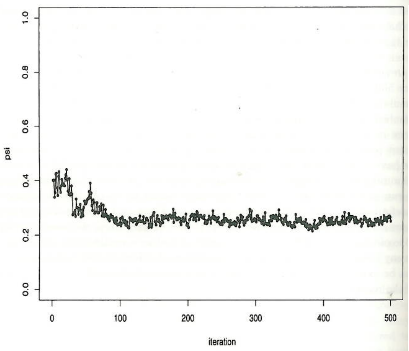


Image credit: Gelman (1995) In "MCMC in Practice" (Editors: Gilks, Richardson, and Spiegelhalter).

Has this Chain Converged?



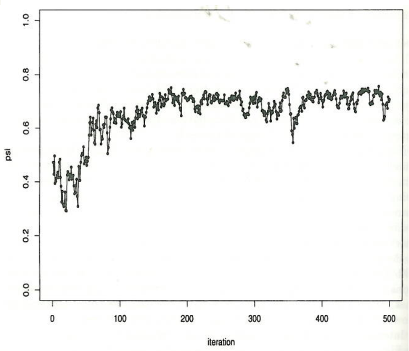
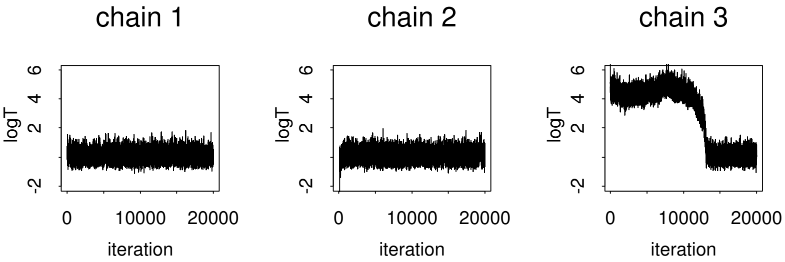


Image credit: Gelman (1995) In “MCMC in Practice” (Editors: Gilks, Richardson, and Spiegelhalter).

Comparing multiple chains can be informative!

Using Multiple Chains



- Compare results of multiple chains to check convergence.
- Start the chains from distant points in parameter space.
- Run until they appear to give similar results
  - ... or they find different solutions (multiple modes).

The Gelman and Rubin "R hat" Statistic

Consider M chains of length N: $\{\psi_{nm}, n = 1, \ldots, N\}$.

$$B = \frac{N}{M-1} \sum_{m-1}^{M} (\bar{\psi}_{\cdot m} - \bar{\psi}_{\cdot\cdot})^2$$

$$W = \frac{1}{M} \sum_{m=1}^{M} s_m^2 \quad \text{where} \quad s_m^2 = \frac{1}{N-1} \sum_{n=1}^{N} (\psi_{nm} - \bar{\psi}_{\cdot m})^2$$

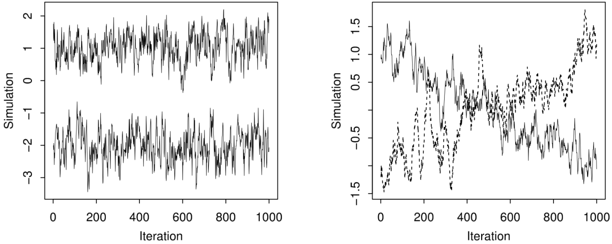
Two estimates of $\mathrm{Var}(\psi \mid Y)$:

1. W: under estimate of $\mathrm{Var}(\psi \mid Y)$ for any finite N.
2. $\widehat{\mathrm{var}}^{+}(\psi \mid Y) = \frac{N-1}{N} W + \frac{1}{N} B$: over estimate of $\mathrm{Var}(\psi \mid Y)$.

$$\hat{R} = \sqrt{\frac{\widehat{\mathrm{var}}^{+}(\psi \mid Y)}{W}} \downarrow 1 \quad \text{as the chains converge.}$$

Compute with `coda` package in R: http://cran.r-project.org/web/packages/coda/index.html

The Split “R hat” Statistic



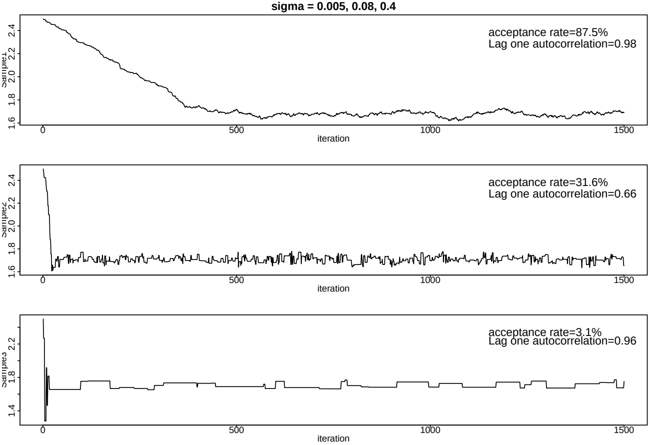
[Image Credit: Gelman et al., Bayesian Data Analysis (2013), Fig 11.3.]

In Practice:

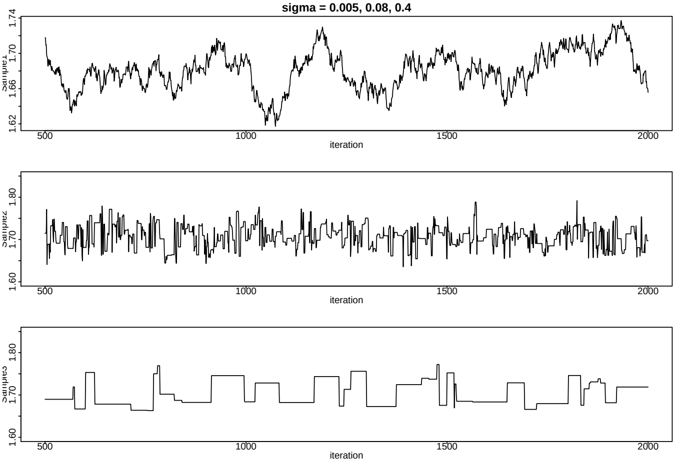
- Run several chains (e.g., 4) with dispersed starting values.
- Discard the first half of each – for burn in.
- Split the second half of each into two (e.g., for a total of $M = 8 = 2 \times 4$ chains, length is 25% of full run).
- Splitting helps identify problems like the right panel.

Choice of Jumping Rule with Random Walk Metropolis

Spectral Analysis: effect of jumping rule on 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: $\bar{\psi}_{\cdot\cdot} = \frac{1}{NM}\sum \psi^{(t)}$ *(estimate of E(ψ|data)!!)*
- At stationarity:

$$\lim_{N\to\infty} MN \,\mathrm{var}\left(\bar{\psi}_{\cdot\cdot}\right) = \left(1 + 2\sum_{t=1}^{\infty} \rho_t\right) \mathrm{var}(\psi \mid y)$$

- $\rho_t =$ lag t autocorrelation
- If draws were indep't, var $\left(\bar{\psi}_{\cdot\cdot}\right)$ would be $\frac{1}{NM}\mathrm{var}(\psi \mid y)$.
- Thus the *effective sample size* is

$$n_{\mathrm{eff}} = \frac{MN}{1 + 2\sum_{t=1}^{\infty} \rho_t}.$$

- Autocorrelations computed from chains (without burnin).
- The infinite sum is truncated when $\hat{\rho}_t < 0.05$ or first becomes negative.

...all computed with `coda` in R.

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?

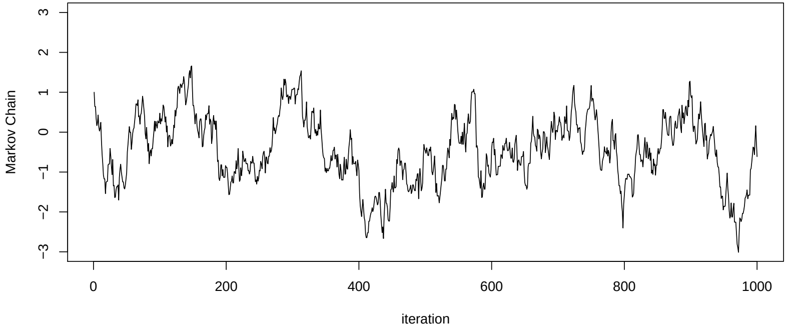


Illustration of the Effective Sample Size

What is the Effective Sample Size here? and σ?

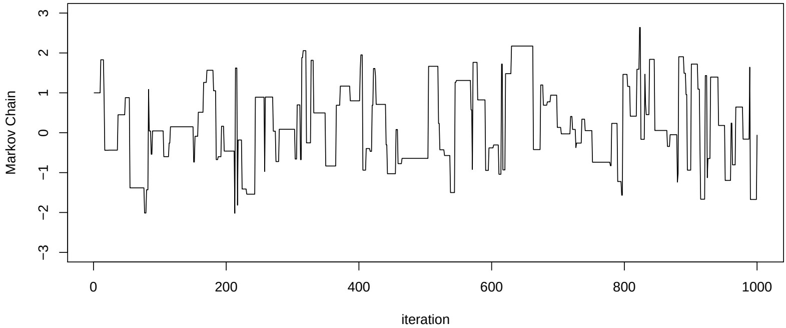


Illustration of the Effective Sample Size

What is the Effective Sample Size here? and σ?

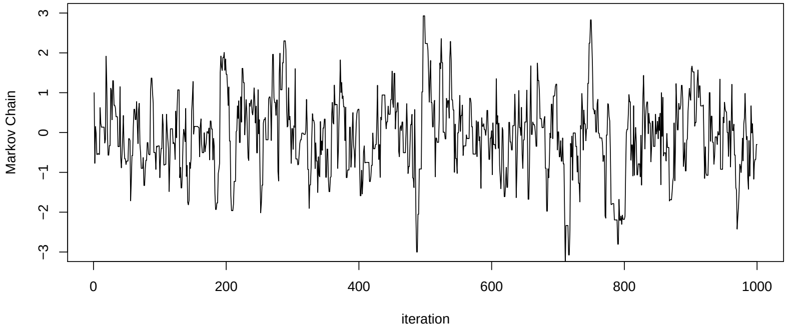


Illustration of the Effective Sample Size

What is the Effective Sample Size here? and σ?

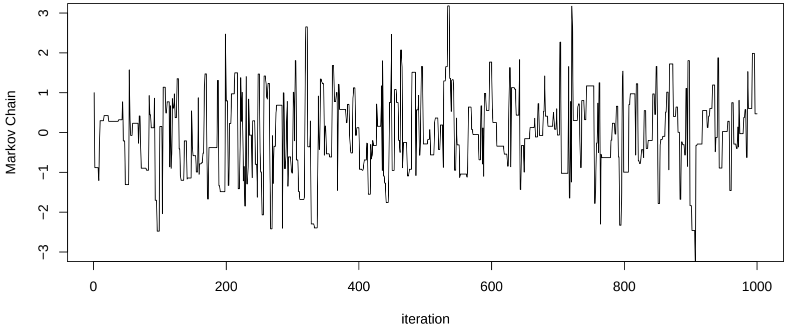


Illustration of the Effective Sample Size

What is the Effective Sample Size here? and σ?

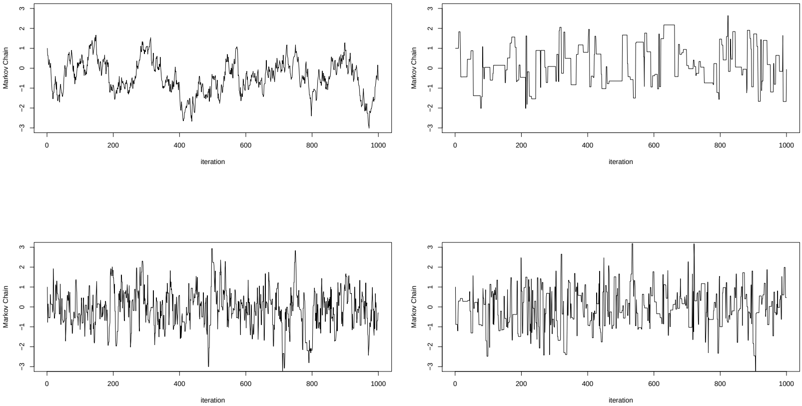
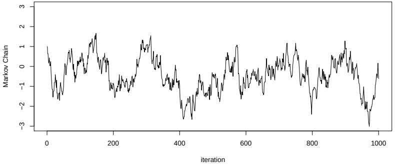
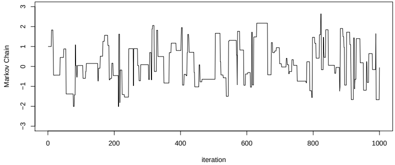


Illustration of the Effective Sample Size

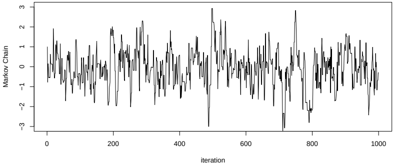
Effective Sample $= 20$; $\sigma = 0.25$.



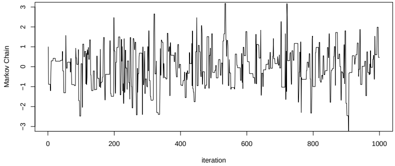
Effective Sample $= 75$; $\sigma = 10$.



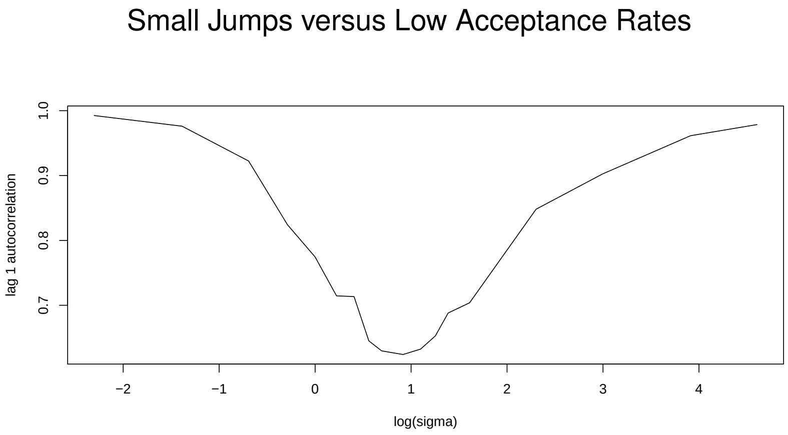
Effective Sample $= 100$; $\sigma = 1$.



Effective Sample $= 216$; $\sigma = 3.5$.

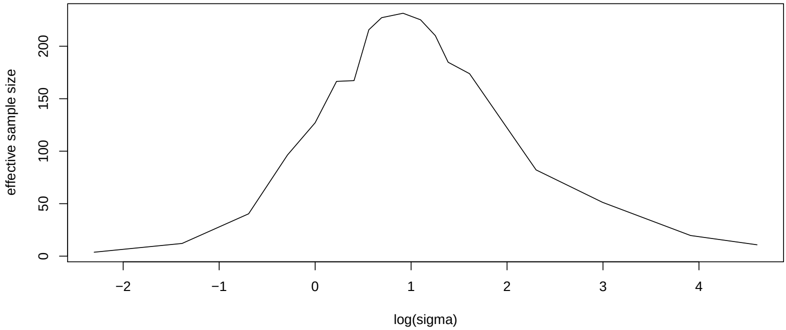


Lag One Autocorrelation



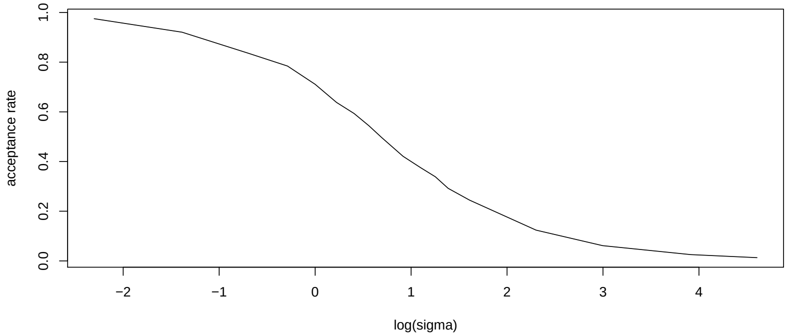
Effective Sample Size

Balancing the Trade-Off



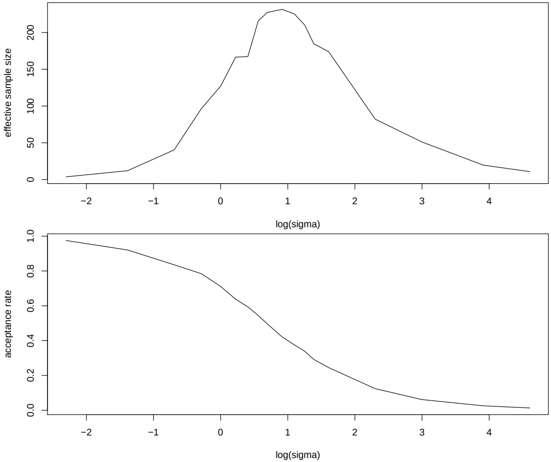
Acceptance Rate

Bigger is not always Better!!



High acceptance rates only come with small steps!!

Finding the Optimal Acceptance Rate

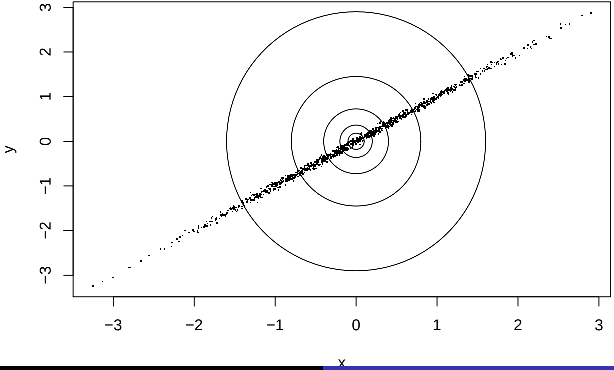


Random Walk Metropolis with High Correlation

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

Tradeoff between high autocorrelation and high rejection rate:

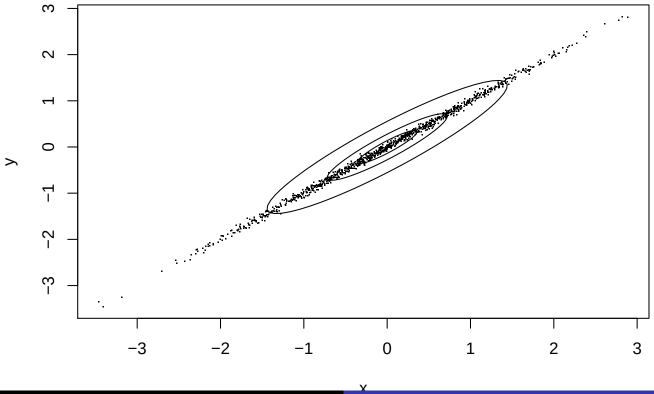
- more acute with high posterior correlations
- more acute with high dimensional parameter



Random Walk Metropolis with High Correlation

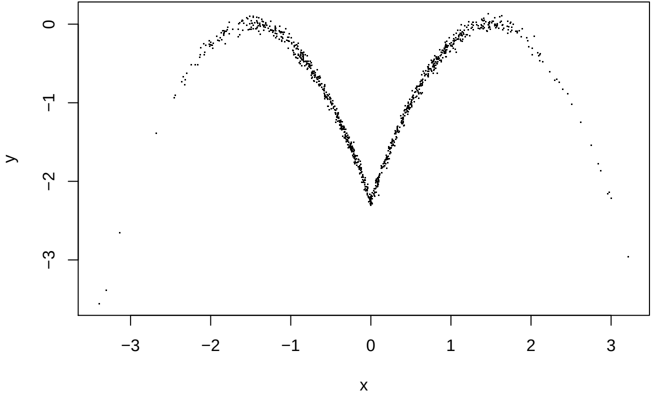
In principle we can use a correlated jumping rule, but

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



Random Walk Metropolis with High Correlation

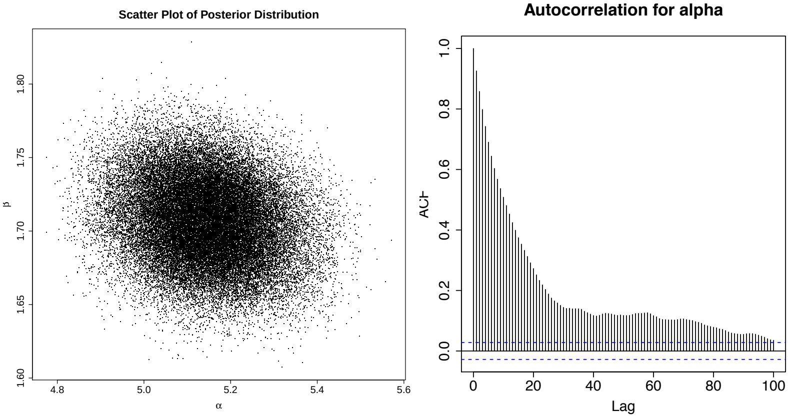
What random walk jumping rule would you use here?



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

Parameters on Different Scales

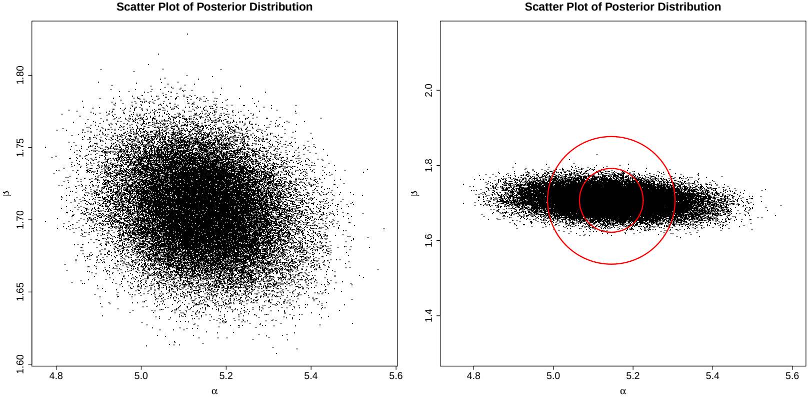
Random Walk Metropolis for Spectral Analysis:



Why is the Mixing SO Poor?!??

Parameters on Different Scales

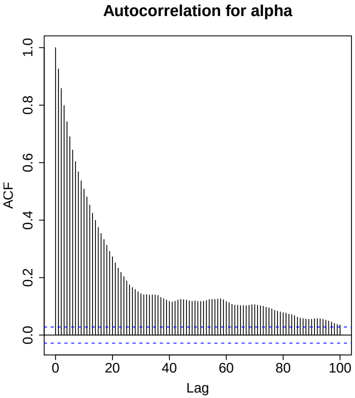
Consider the Scales of α and β:

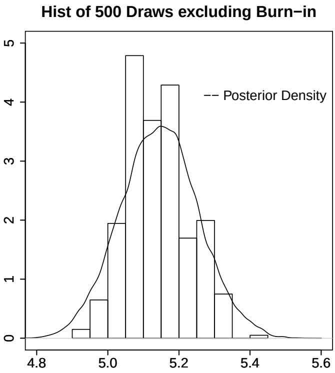


A new jumping rule: std dev for $\alpha = 0.110$, for $\beta = 0.026$, and corr $= -0.216$.

Improved Convergence

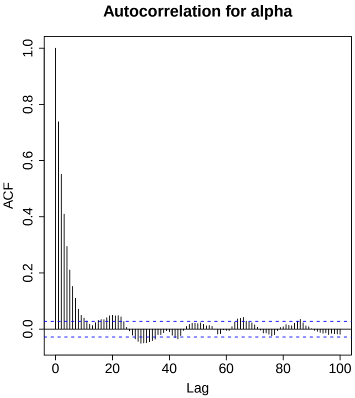
Original Jumping Rule:

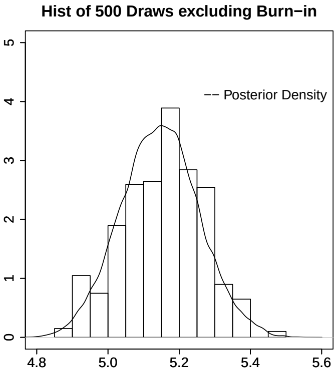




Improved Convergence

Improved Jumping Rule:





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

Parameters on Different Scales

With Jumping Rule: $\textsc{Norm}(\theta^{(t-1)}, kM)$, or better $t_{\mathrm{df}}(\theta^{(t-1)}, kM)$.

Try:

1. Using the variance-covariance matrix from a standard fitted model for M

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

2. Adaptive methods that allow the jumping rule to evolve on the fly.[1]

Always: Adjust k and/or M to aim for acceptance rate of

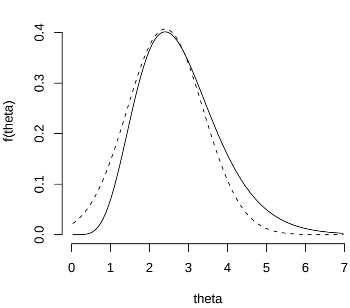
$\sim$20% (multivariate update) or $\sim$40% (univariate update).

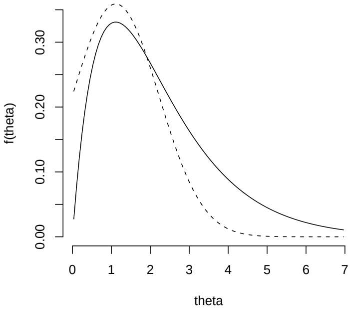
[1] E.g., "Optimal proposal distributions and adaptive MCMC" by JS Rosenthal in Handbook of Markov Chain Monte Carlo (CRC Press, 2011).

Transforming to Normality

Parameter transformations can greatly improve MCMC.

Recall the Independence Sampler:

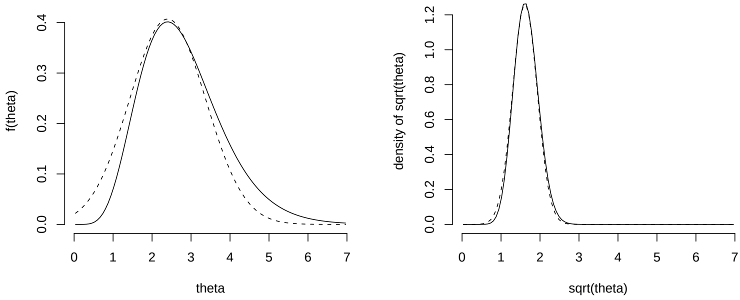




The normal approximation is not as good as we might hope...

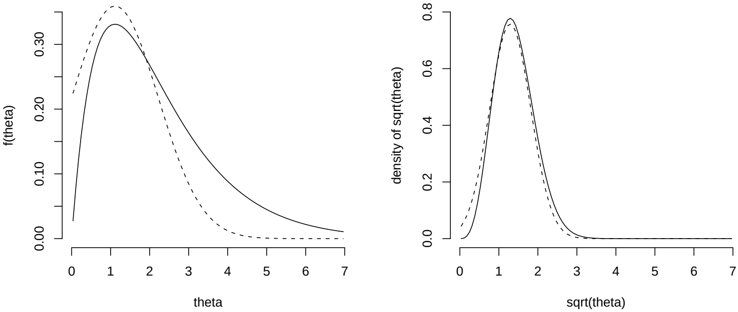
Transforming to Normality

But if we use the square root of θ:



Transforming to Normality

And...



The normal approximation is much improved!

Transforming to Normality

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

General Strategy:

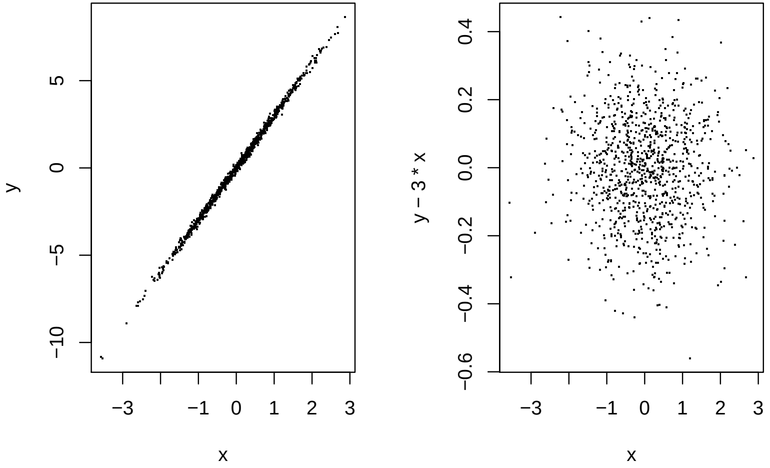
- Transform to the Real Line.
- Take the log of positive parameters.
- If the log is “too strong”, try square root.
- Probabilities can be transformed via the logit transform:

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

- More complex transformations for other quantities.
- *Try out various transformations using an initial MCMC run.*
- Statistical advantages to using normalizing transforms.

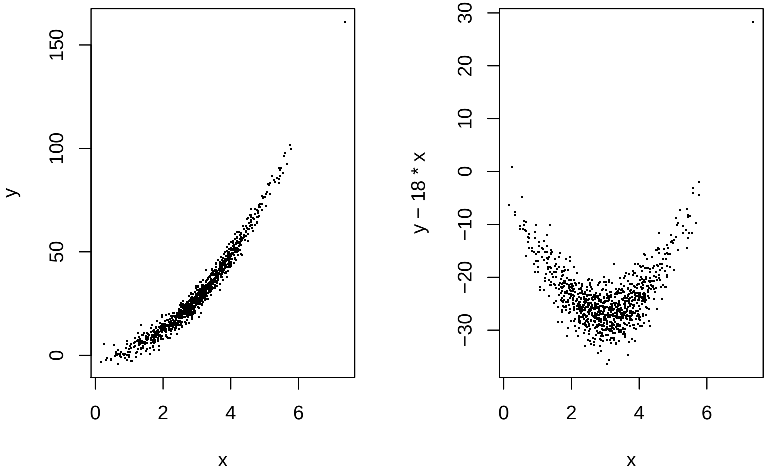
Removing Linear Correlations

Linear transformations can remove linear correlations



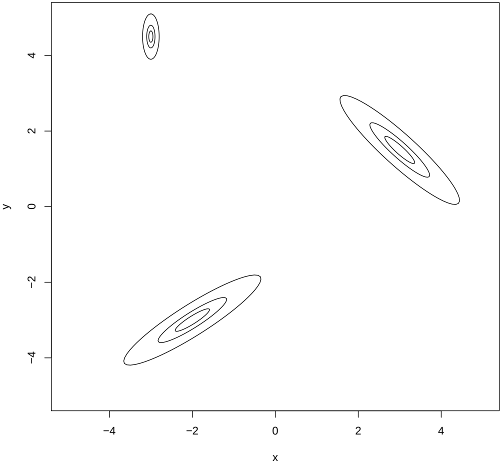
Removing Linear Correlations

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



Multiple Modes

- Scientific meaning of multiple modes.
- Do not focus only on the biggest mode!
- “Important” modes.
- Challenging for Bayesian and Frequentist methods.
- Consider Metropolis & Metropolis Hastings.
- Value of excess dispersion and multiple starting values.



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. Nested Sampling (Skilling, 2006, *Bayesian Analysis*)
   4. Many other specialized methods.

Outline

Breaking a Complex Problem into Simpler Pieces

- Ideally we sample directly from $p(\theta|Y)$ without Metropolis.
- This may not work in complex 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.

Two-Step Gibbs Sampler,

Starting with some $\theta^{(0)}$, for $t = 1, 2, 3, \ldots$

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

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

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

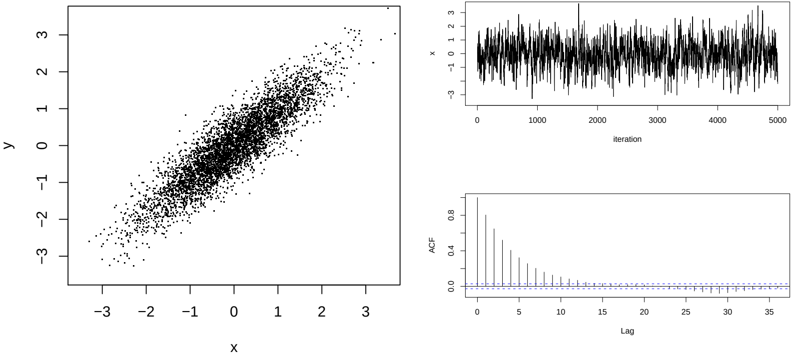
Embedding Other Samplers within Gibbs

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

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

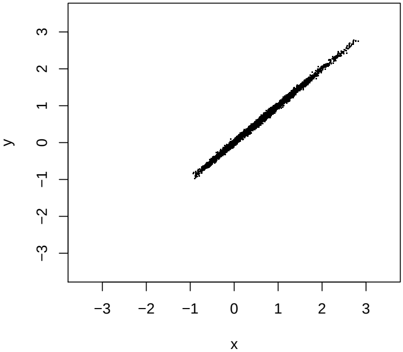
- We can use a Metropolis or Metropolis-Hastings step to update β within the Gibbs sampler.
- The result is known as *Metropolis within Gibbs Sampler*.
- **Advantage:** Metropolis tends to preform poorly in high dimensions. Gibbs reduces the dimension.
- **Disadvantage:** Case-by-case probabilistic calculations.
  (But case-by-case algorithmic development and tuning always helps)

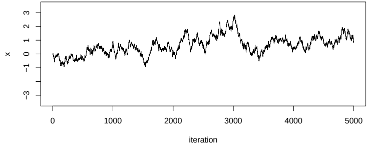
When Will Gibbs Sampling Work Well?

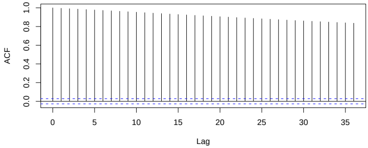


autocorrelation = 0.81, effective sample size = 525

When Will Gibbs Sampling Work Poorly?



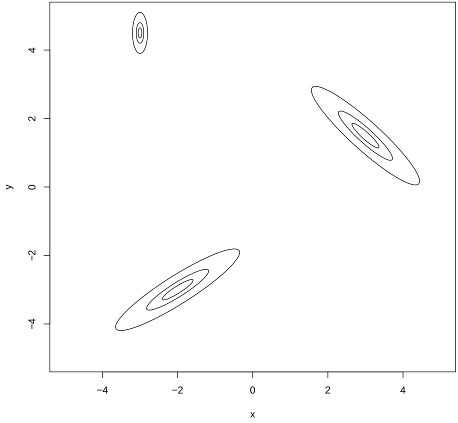




autocorrelation = 0.998, effective sample size = 5

High Posterior Correlations are Always Problematic.

Multiple Modes



How will the Gibbs Sampler Handle Multiple modes?

The General Gibbs Sampler

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

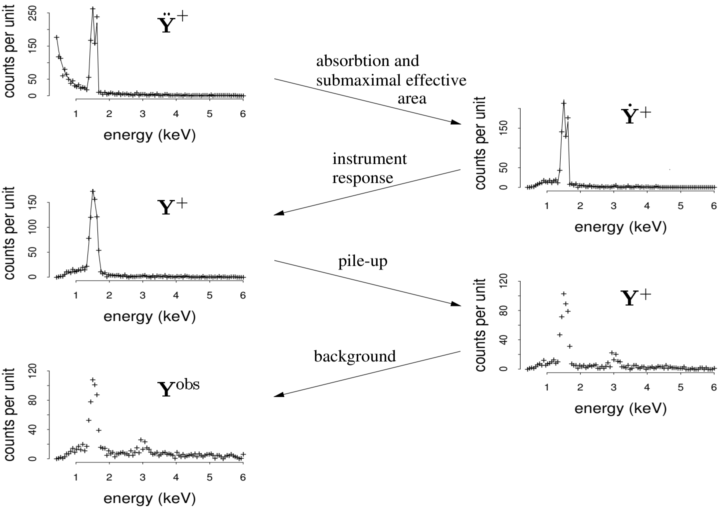
$$p(\theta_p | \theta_1, \ldots, \theta_{p-1}, \theta_{p+1}, \ldots, \theta_P, Y), \text{ for } p = 1, \ldots, P$$

Gibbs Sampler

Starting with some $\theta^{(0)}$, for $t = 1, 2, 3, \ldots$

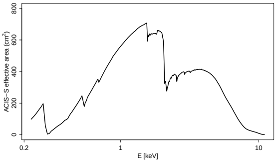
$$\begin{aligned}
\text{Draw 1:}\quad & \theta_1^{(t)} \sim p(\theta_1 | \theta_2^{(t-1)}, \ldots, \theta_P^{(t-1)}, Y) \\
& \vdots \\
\text{Draw p:}\quad & \theta_p^{(t)} \sim p(\theta_p | \theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_P^{(t-1)}, Y) \\
& \vdots \\
\text{Draw P:}\quad & \theta_P^{(t)} \sim p(\theta_P | \theta_1^{(t)}, \ldots, \theta_{P-1}^{(t)}, Y)
\end{aligned}$$

Example: Calibration Uncertainty

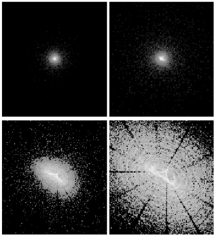


Calibration Products

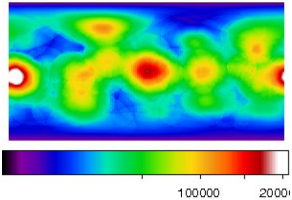
- Analysis is highly dependent on *Calibration Products:*
  - Effective area records sensitivity as a function of energy
  - Energy redistribution matrix can vary with energy/location
  - Point Spread Functions can vary with energy and location
  - Exposure Map shows how effective area varies in an image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



A *CHANDRA* effective area.



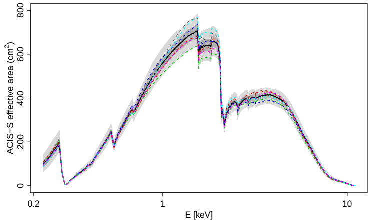
Sample Chandra psf's
(Karovska et al., ADASS X)

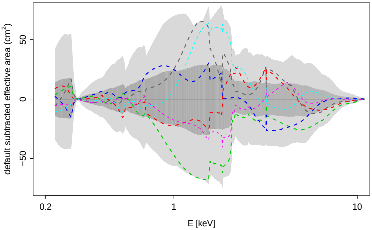


EGERT exposure map
(area × time)

Derivation of Calibration Products

- Prelaunch ground-based and post-launch space-based empirical assessments.
- Complex computer models of subassembly components.
- Calibration scientists provide a sample representing uncertainty
- *Calibration Sample* is typically of size $\approx$ 1000.
- This is a sample from the prior distribution for A.





Example: Calibration Uncertainty

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

- **PyBLoCXS** (**Py**thon **B**ayesian **Lo**w **C**ount **X**-ray **S**pectral): 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.

Two Possible Models

We consider inference under:

A PRAGMATIC BAYESIAN MODEL: $\pi_0(A, \theta) = p(A)p(\theta|A, Y)$.

THE FULLY BAYESIAN POSTERIOR: $\pi(A, \theta) = p(A|Y)p(\theta|A, Y)$.

Concerns:

Statistical Fully Bayesian model is "correct".

Cultural Astronomers have concerns about letting the current data influence calibration products.

Computational Both models pose challenges, but pragmatic Bayesian is easier to fit.

Practical How different are $p(A)$ and $p(A|Y)$?

How do we draw A?

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

We use *Principal Component Analysis* to represent uncertainly:

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

A_0: default effective area,
$\bar{\delta}$: mean deviation from A_0,
r_j and $\boldsymbol{v}_j$: first m principle component eigenvalues & vectors,
e_j: independent standard normal deviations.

Effectively, we are placing a degenerate MV Normal prior on A.

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

Gibbs Samplesr for Calibration Uncertainty

An MH within Gibbs Sampler (Fully Bayes):

DRAW 1: Update e via MH with limiting dist'n $p(e|\theta, Y)$

DRAW 2: Update θ via MH with limiting dist'n $p(\theta|e, Y)$

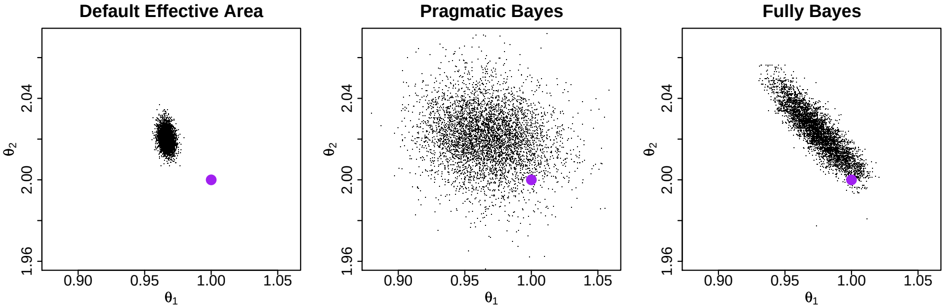
Fully Bayesian Approach:

- DRAW 1: Gaussian Metropolis jumping rule centered at e'.
- DRAW 2: pyBLoCXS

Pragmatic Bayesian Approach:

- Replace Draw 1 with a sample from prior distribution on A.

Sampling From the Full Posterior



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

Citations:

1. Lee, Kashyap, van Dyk, Connors, Drake, Izem, Meng, Min, et al. (2011). Accounting for Calibration Uncertainties in X-ray Analysis: Effective Areas in Spectral Fitting. *The Astrophysical Journal*, **731**, 126.
2. Xu, van Dyk, Kashyap, Siemiginowska, Connors, Drake, Meng, et al. (2014). A Fully Bayesian for Jointly Fitting Instrumental Calibration and X-ray Spectral Models. *The Astrophysical Journal*, to appear.
3. Chen, Meng, Wang, van Dyk, Marshall, and Kashyap (2019). Calibration Concordance for Astronomical Instruments via Multiplicative Shrinkage. *Journal of the American Statistical Association*, **114**, 1018–1037.

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:

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

- We assume indep Gaussian errors with known variances:

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

Stellar Evolution: MCMC Strategy

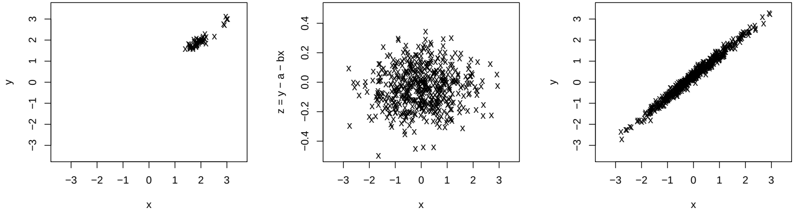
Metropolis within Gibbs Sampling

- $3N + 5$ 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.

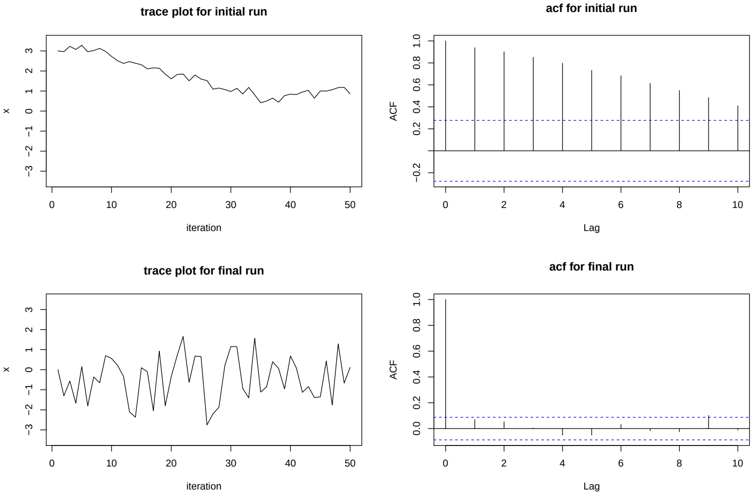
Dynamic transformations



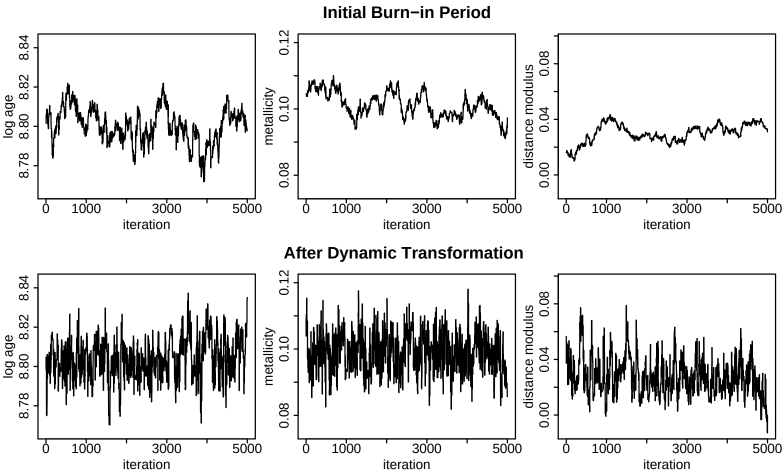
A toy example:

1. Initial Gibbs run shows high autocorrelation, panel 1.
2. Fit $y = \alpha + \beta x$ and transfrom $Z = Y - \hat{\alpha} - \hat{\beta}X$.
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



Overview of Recommended Strategy for MCMC

1. Start with a crude approximation to the posterior distribution, perhaps using a mode finder.
2. Use the approximation to setup the jumping rule of an initial sampler (e.g., Gibbs, MH, etc.): update one parameter at a time or update parameters in batches.
3. Use Gibbs draws for closed form complete conditionals.
4. Use metropolis jumps if complete conditional is not in closed form.
5. Run with multiple chains
6. After an initial run, update the jumping rule using the variance-covariance matrix of the initial sample, rescaling so that acceptance rates are near 20% (for vector updates) or 40% (for single parameter updates).

Overview of Recommended Strategy- Con't

7. To improve convergence, use transformations so that parameters are approximately independent and/or approximately Gaussian.
8. Check for convergence using multiple chains.
9. Compute effective sample size to be sure you have sufficiently long chains.
10. Compare inference based on crude approximation and MCMC. If they are not similar, check for errors before believing the results of the MCMC.