Bayesian Statistical Methods for Astronomy

Part II: Markov Chain Monte Carlo

David A. van Dyk

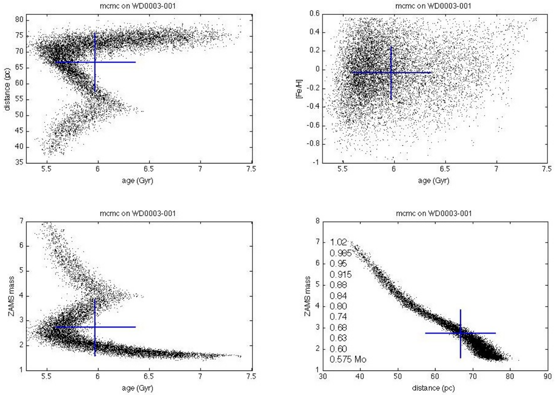
Statistics Section, Imperial College London

Stat4Astro — Autrans, France — October 2017

Outline

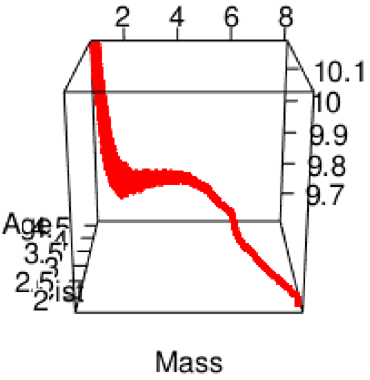
Outline

Complex Posterior Distributions



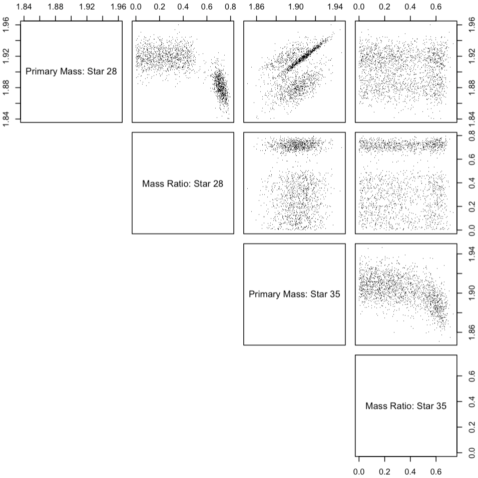
Highly non-linear relationship among stellar parameters.

Complex Posterior Distributions



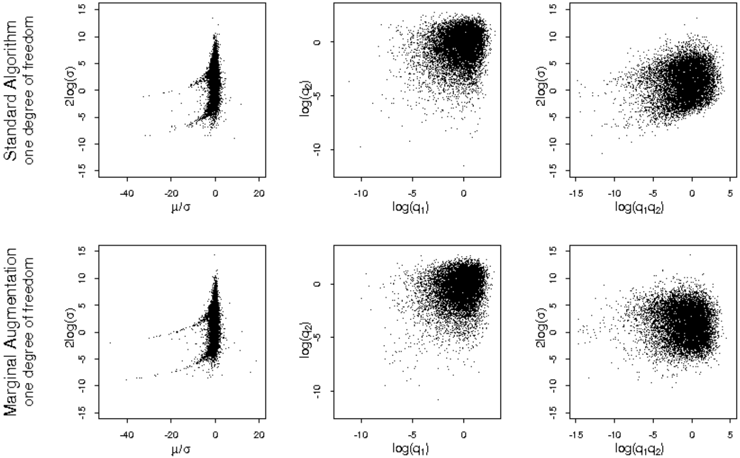
Highly non-linear relationships among stellar parameters.

Complex Posterior Distributions

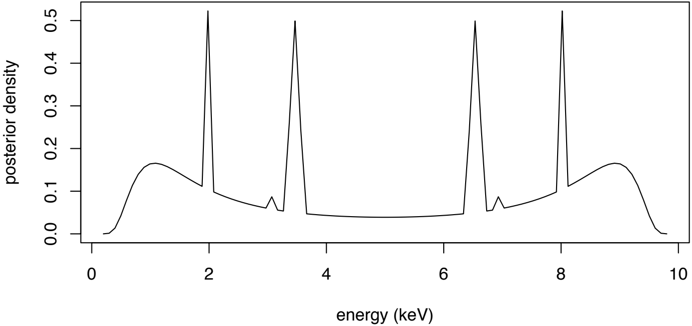


The classification of certain stars as field or cluster stars can cause multiple modes in the distributions of other parameters.

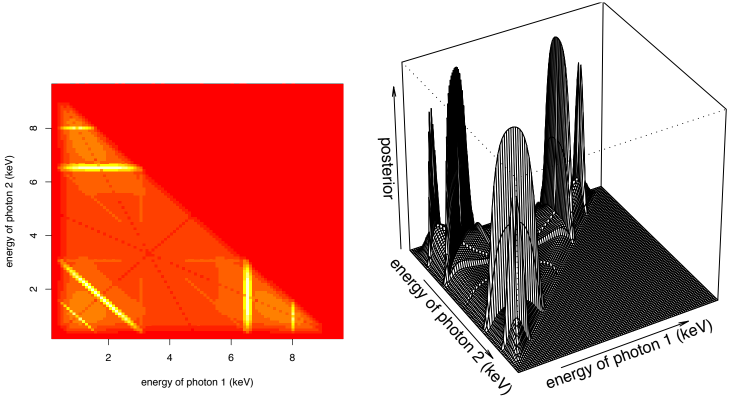
Complex Posterior Distributions



Complex Posterior Distributions

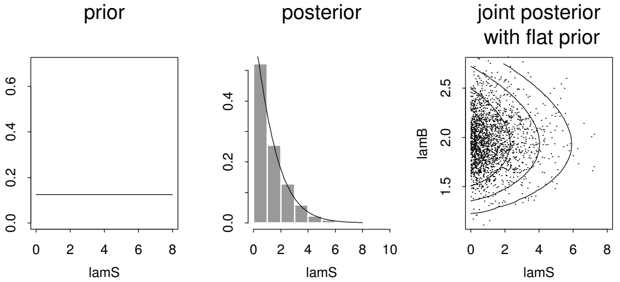


Complex Posterior Distributions



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

Rejection Sampling

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

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

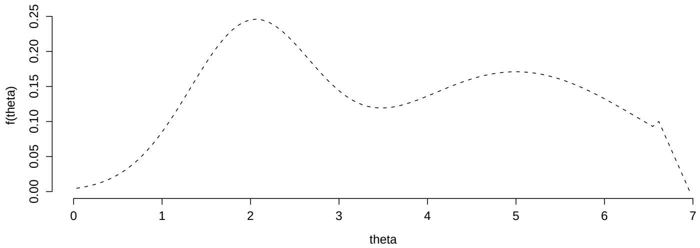
for some M.

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?

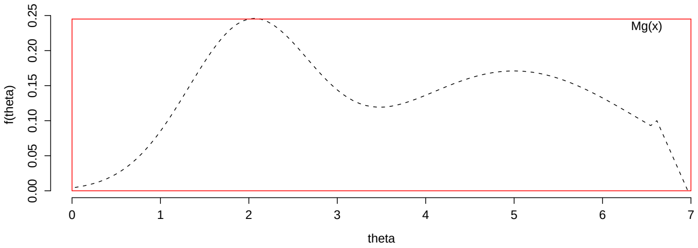
Rejection Sampling

Consider the distribution:



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

Rejection Sampling



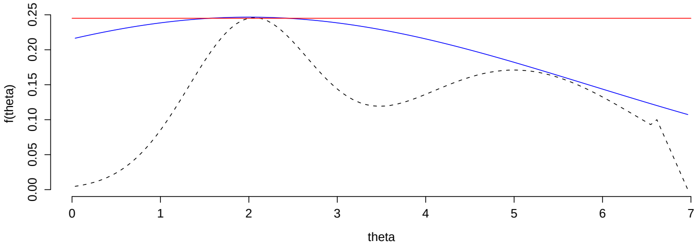
- Imagine that we sample uniformly in the red rectangle:

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

- 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 $f(\theta)$!!

What is a Markov Chain

Definition

A Markov chain is a sequence of random variables,

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

such that

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

A Markov chain is generally constructed via

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

with $U^{(1)}, U^{(2)}, \ldots$ 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,

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

- 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

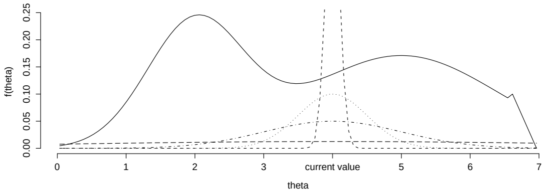
- J_t must be symmetric: $J_t(\theta^*|\theta^{(t-1)}) = J_t(\theta^{(t-1)}|\theta^*)$.
- If $p(\theta^*|y) > p(\theta^{(t-1)}|y)$, *jump!*

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

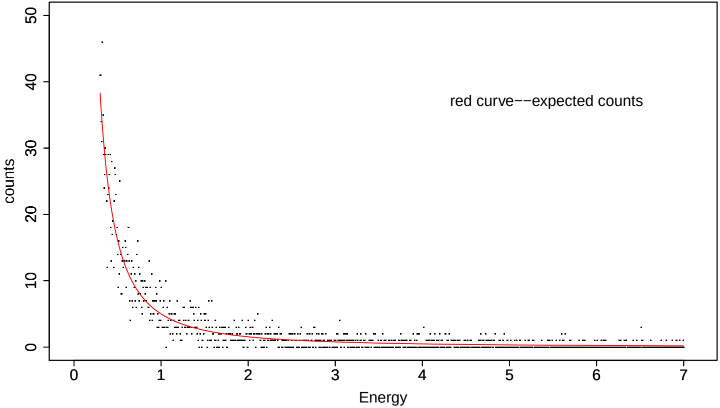
A simplified model for high-energy spectral analysis.

- 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. $(\alpha, \beta) \overset{\text{indep.}}{\sim} \text{Unif}(0, 100)$.

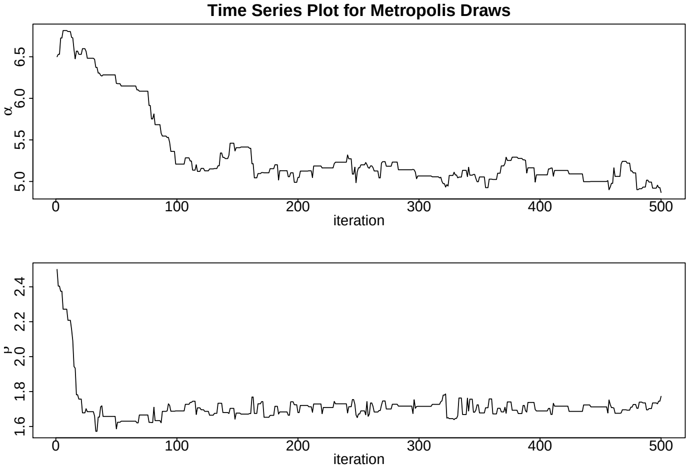
- The Sampler:
  We use a Gaussian Jumping Rule,
  - centered at the current sample, $\theta^{(t)}$
  - with standard deviations equal 0.08 and correlation zero.

Simulated Data

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

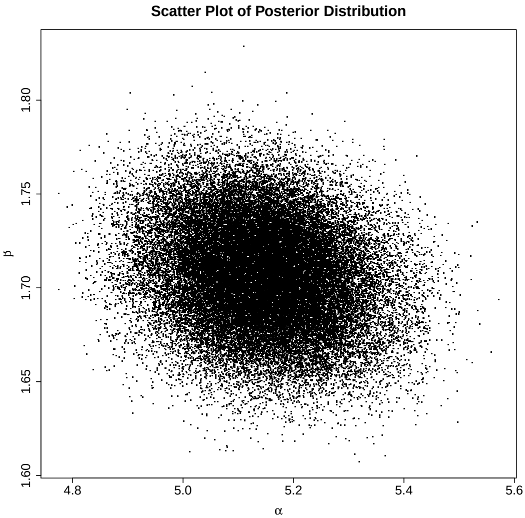


Markov Chain Trace Plots

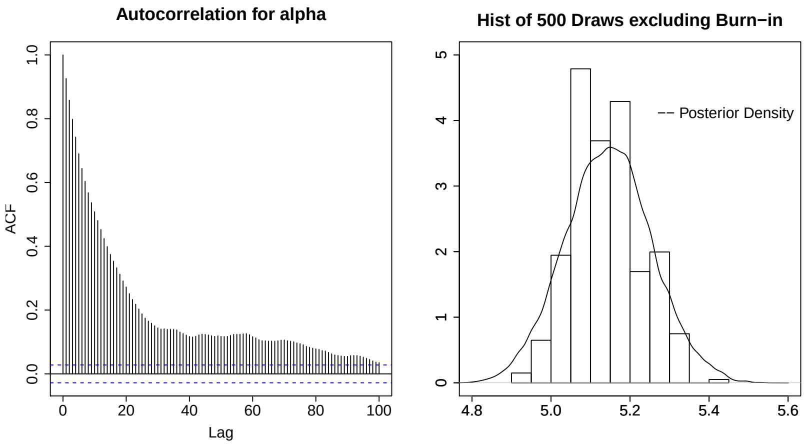


Chains “stick” at a particular draw when proposals are rejected.

The Joint Posterior Distribution

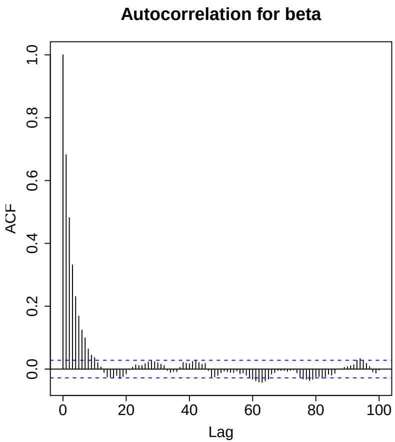


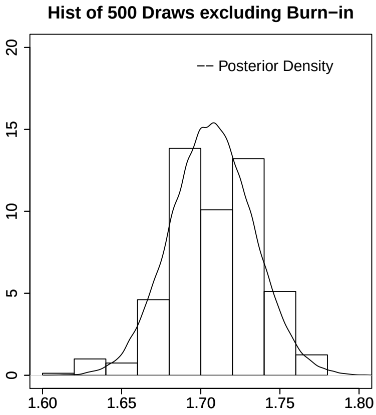
Marginal Posterior Dist'n of the Normalization



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

Marginal Posterior Dist'n of Power Law Param





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

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:

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

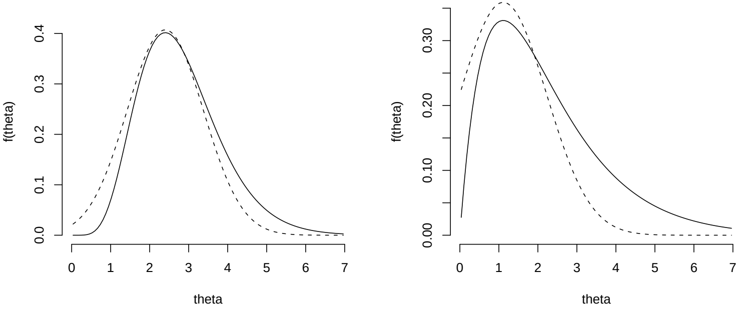
$$\text{MAP estimate} = \text{argmax}_\theta\, p(\theta|y)$$

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

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`:

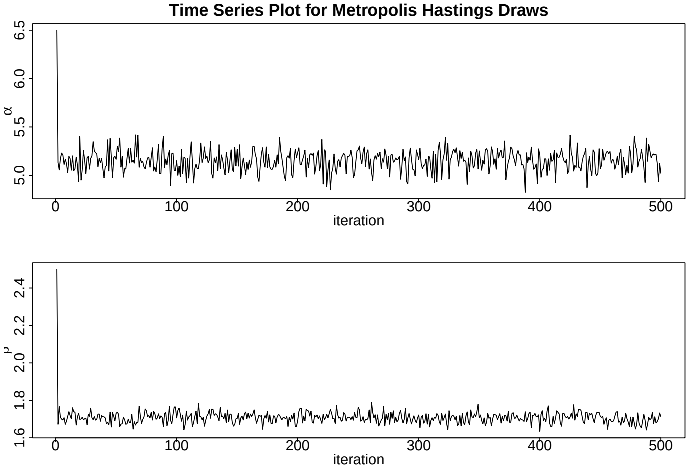
```
> glm.fit = glm( Y~I(-log(E)), family=poisson(link="log") )
> glm.fit$coef     #### best fit of (log(alpha), beta)
> vcov( glm.fit ) #### variance-covariance matrix
```

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

Example of Independence Sampler

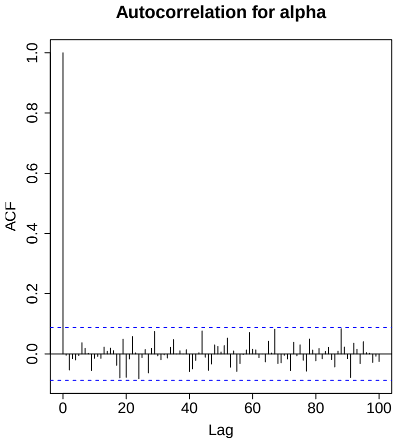
- Alternatively, we can fit (α, β) directly with a general (but less stable) mode finder.
- Requires coding likelihood, specifing starting values, etc.
- Choose parameterization to improve Gaussian approx.
  - MLE is invariant to transformations.
  - Variance matrix of transform is computed via *delta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

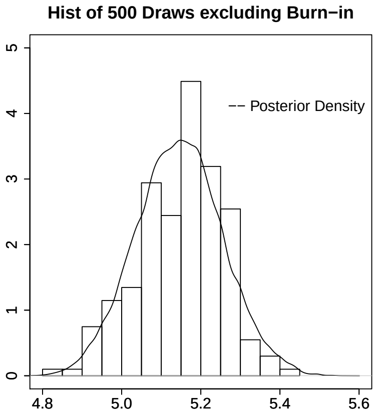
Markov Chain Trace Plots



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

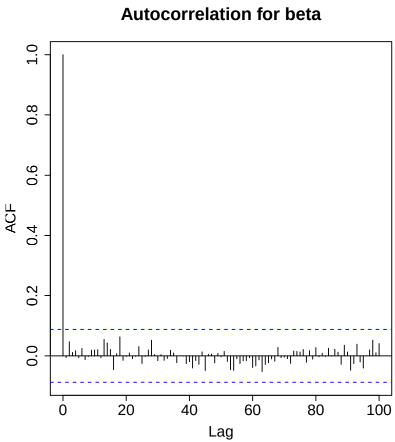
Marginal Posterior Dist'n of the Normalization

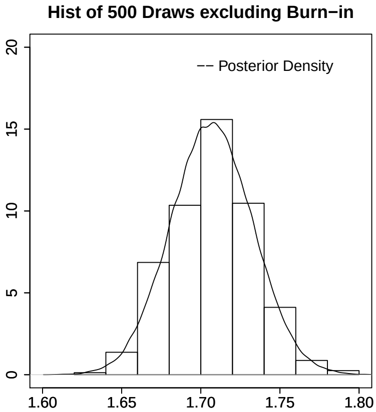




Autocorrelation is essentially zero: nearly independent sample!!

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

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

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

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?

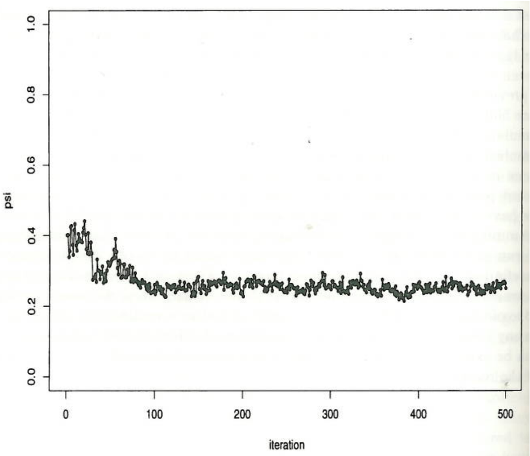
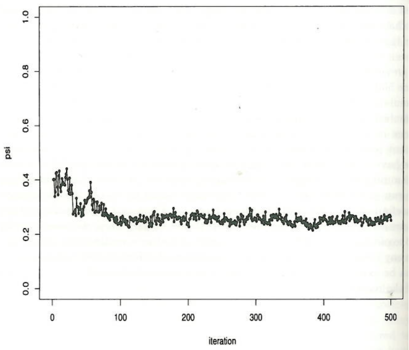


Image credit: Gelman (1995) In “MCMC in Practice” (Editors: Gilks, Richardson, and Spiegelhalter).

Has this Chain Converged?



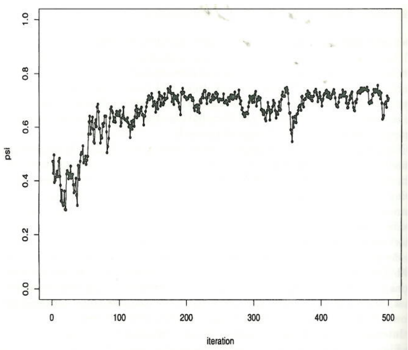
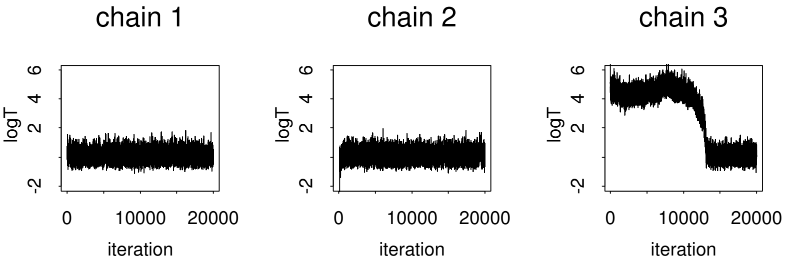


Image credit: Gelman (1995) In “MCMC in Practice” (Editors: Gilks, Richardson, and Spiegelhalter).

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

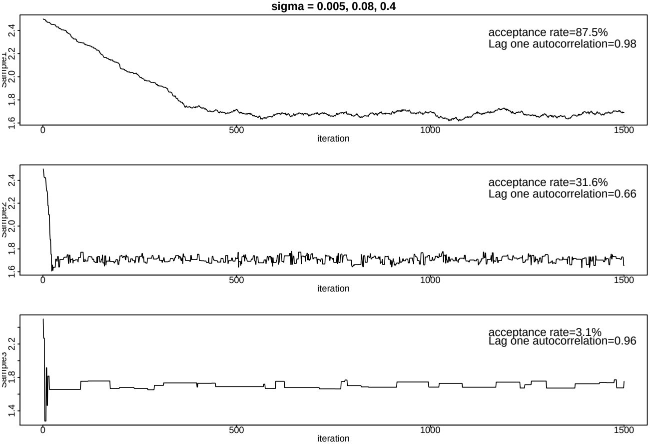
1. W: under estimate of $\mathrm{Var}(\psi \mid Y)$ for any finite N.
2. $\widehat{\mathrm{var}}^{+}(\psi \mid Y) = \frac{N-1}{N} W + \frac{1}{N} B$: over estimate of $\mathrm{Var}(\psi \mid Y)$.

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

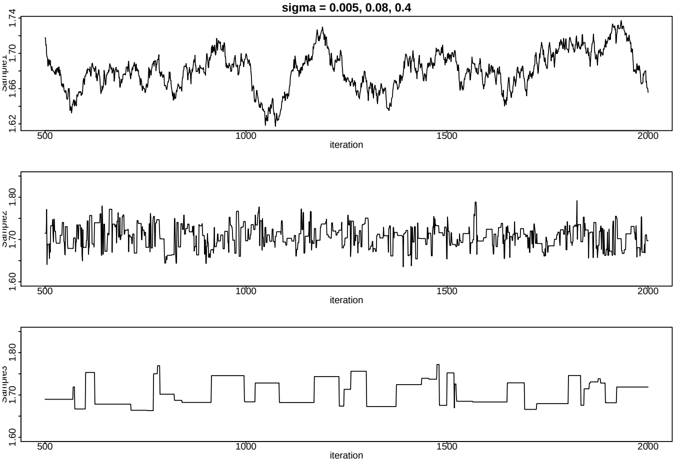
Compute with `coda` package in R: http://cran.r-project.org/web/packages/coda/index.html

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: $\bar{h}_n = \frac{1}{n}\sum h(\theta^{(t)})$ *(estimate of E(h(θ)|x)!!)*
- Variance Estimate: $\text{Var}(\bar{h}_n) \approx \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}$ with *(not var(θ)!!)*

  $\sigma^2 = \text{Var}(h(\theta))$ estimated by $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{t=1}^{n}[h(\theta^{(t)}) - \bar{h}_n]^2$,

  $\rho = \text{corr}\left[h(\theta^{(t)}), h(\theta^{(t-1)})\right]$ 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 $n\frac{1-\rho}{1+\rho}$...

...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?

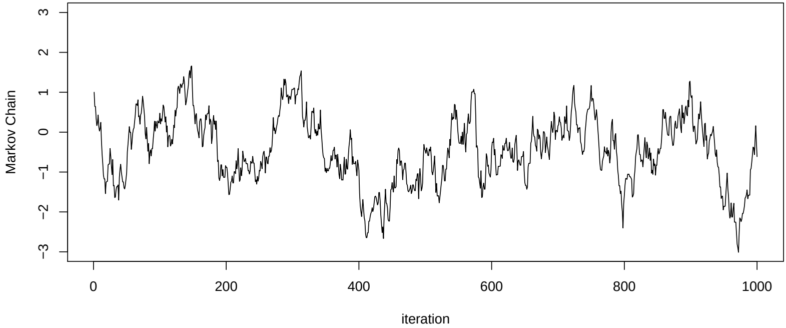


Illustration of the Effective Sample Size

What is the Effective Sample Size here? and σ?

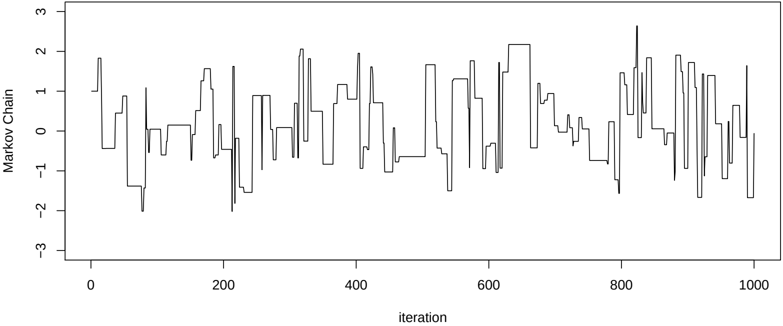


Illustration of the Effective Sample Size

What is the Effective Sample Size here? and σ?

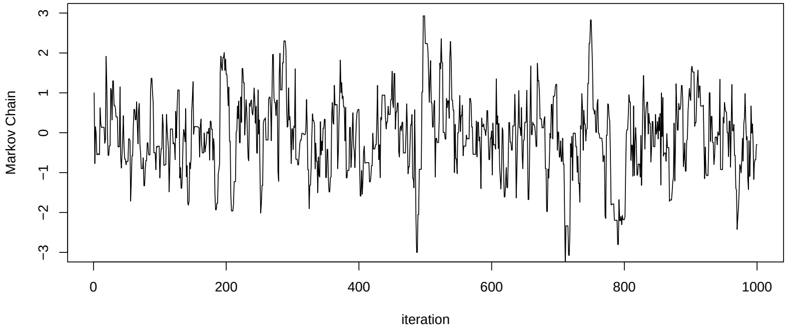


Illustration of the Effective Sample Size

What is the Effective Sample Size here? and σ?

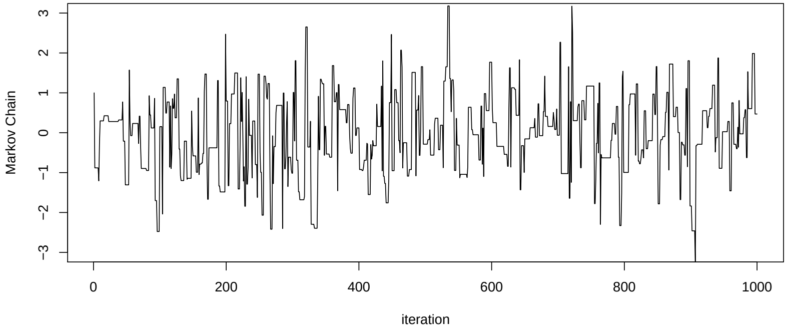


Illustration of the Effective Sample Size

What is the Effective Sample Size here? and σ?

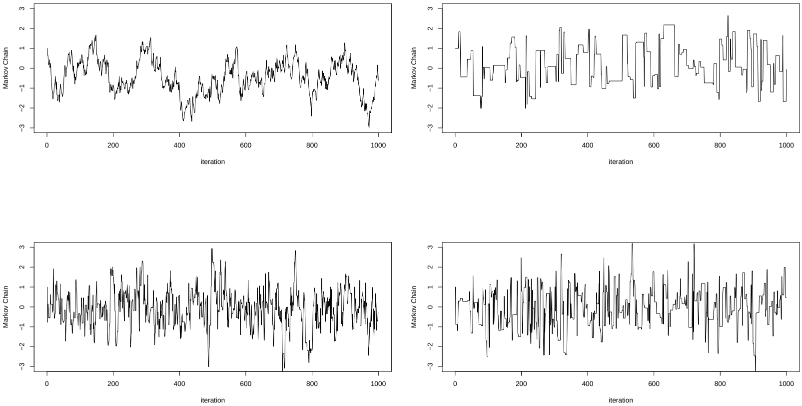
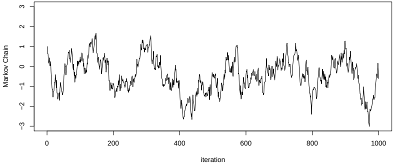
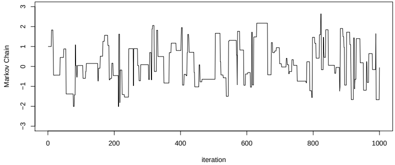


Illustration of the Effective Sample Size

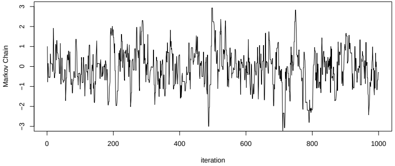
Effective Sample $= 20$; $\sigma = 0.25$.



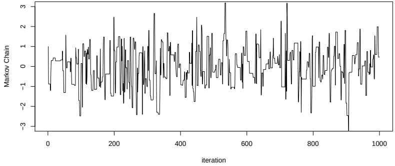
Effective Sample $= 75$; $\sigma = 10$.



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

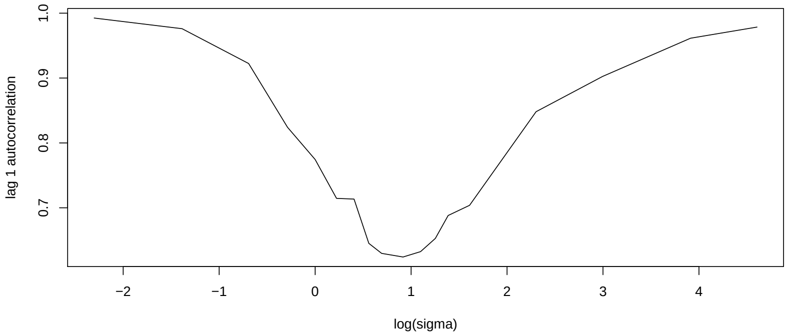


Effective Sample $= 216$; $\sigma = 3.5$.

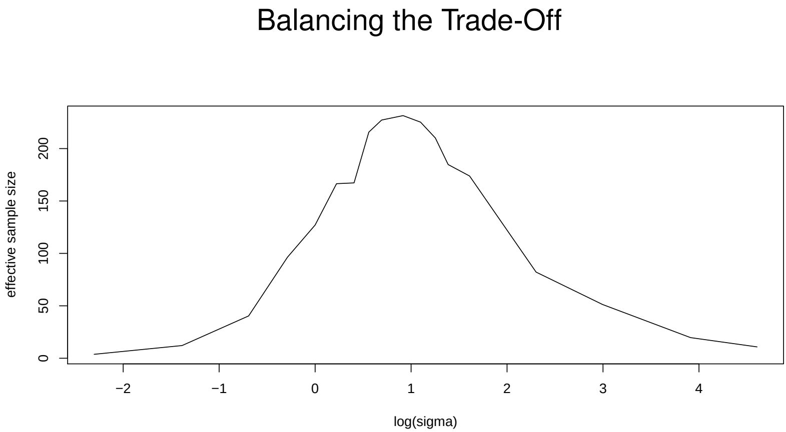


Lag One Autocorrelation

Small Jumps versus Low Acceptance Rates

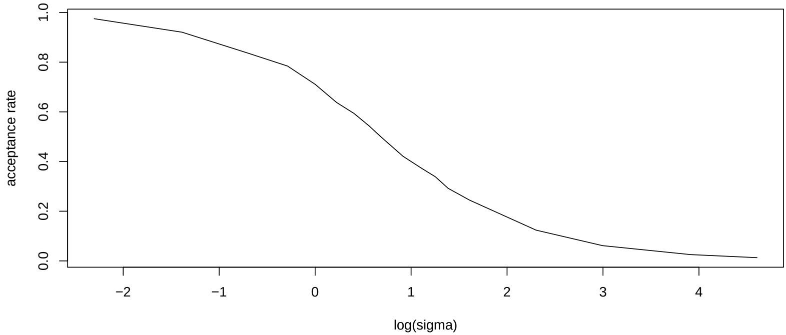


Effective Sample Size



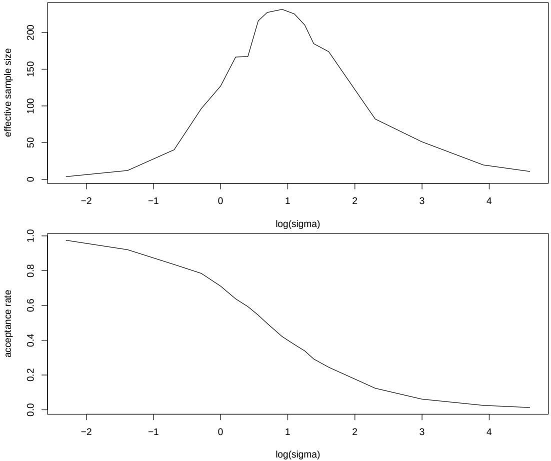
Acceptance Rate

Bigger is not always Better!!



High acceptance rates only come with small steps!!

Finding the Optimal Acceptance Rate

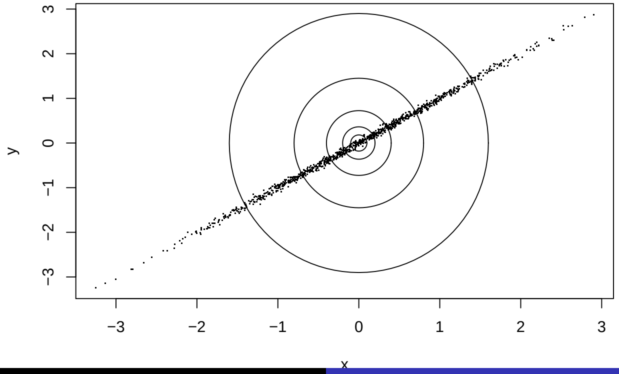


Random Walk Metropolis with High Correlation

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

Tradeoff between high autocorrelation and high rejection rate:

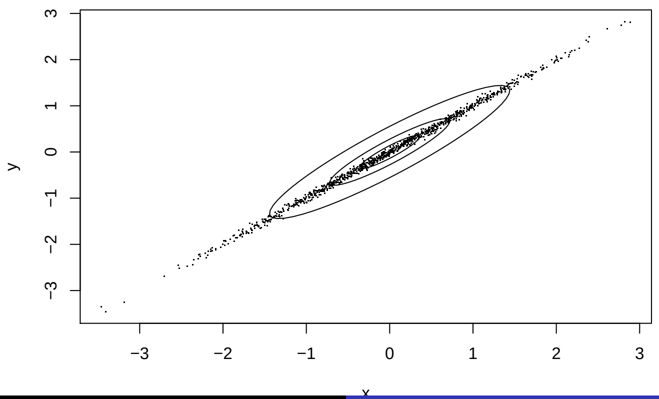
- more acute with high posterior correlations
- more acute with high dimensional parameter



Random Walk Metropolis with High Correlation

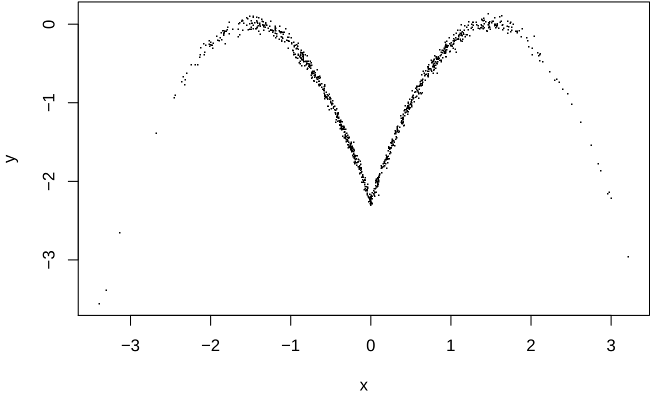
In principle we can use a correlated jumping rule, but

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



Random Walk Metropolis with High Correlation

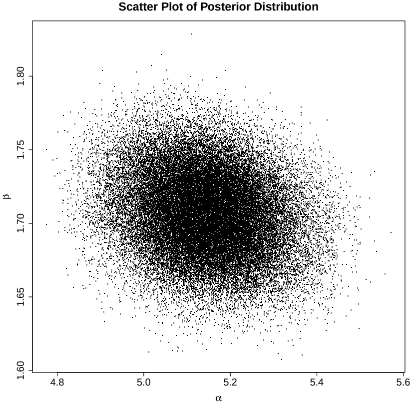
What random walk jumping rule would you use here?

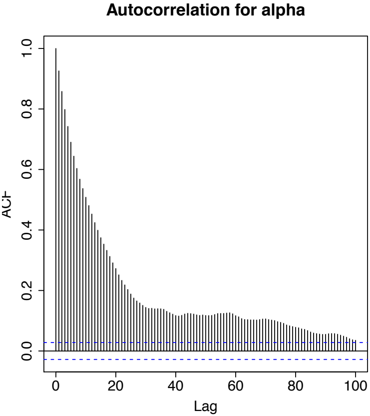


Remember: you don't get to see the distribution in advance!

Parameters on Different Scales

Random Walk Metropolis for Spectral Analysis:

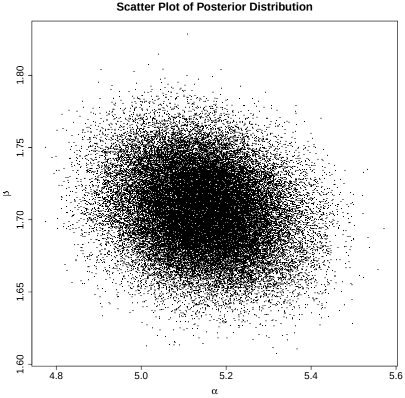


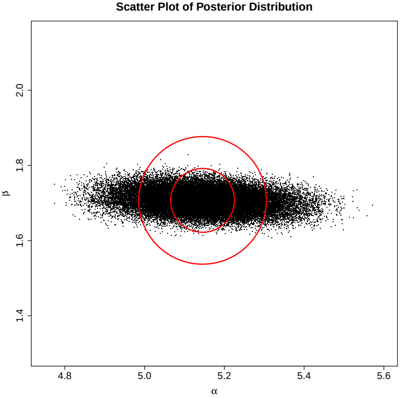


Why is the Mixing SO Poor?!??

Parameters on Different Scales

Consider the Scales of α and β:

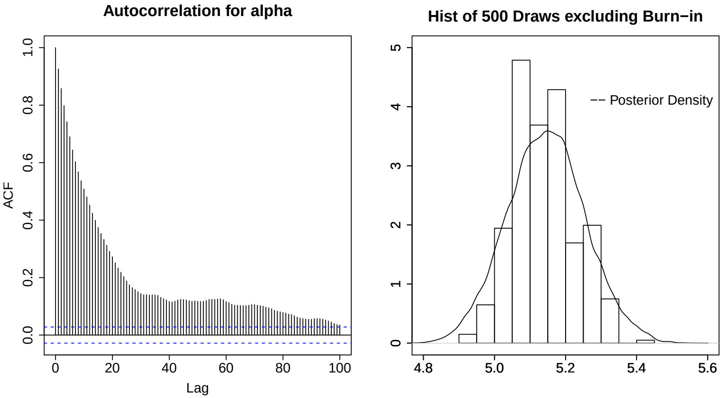




A new jumping rule: std dev for $\alpha = 0.110$, for $\beta = 0.026$, and corr $= -0.216$.

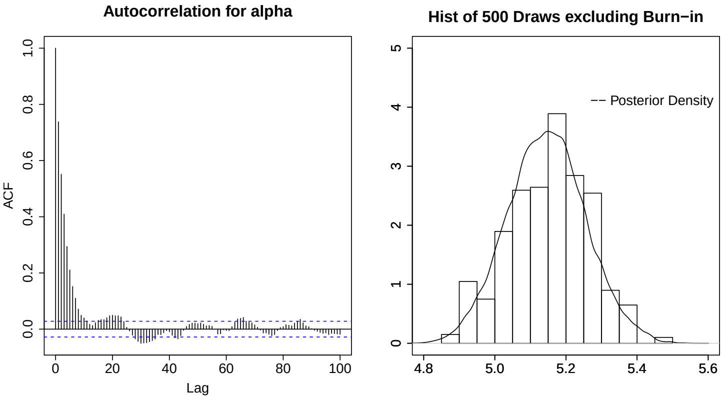
Improved Convergence

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 $t_{\rm df}(\theta^{(t-1)}, kM)$.

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: Aim for acceptance rate of

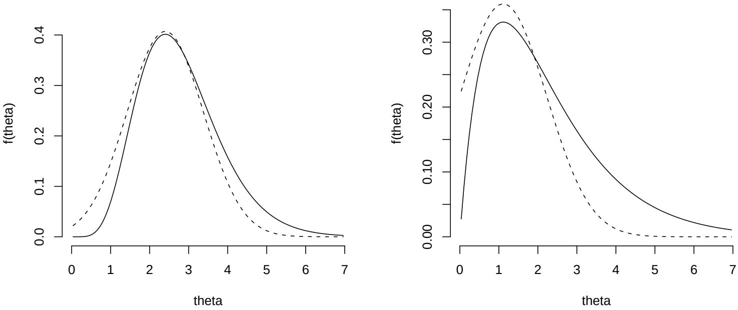
$\sim$20% (multivariate update) or $\sim$40% (univariate update).

[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.

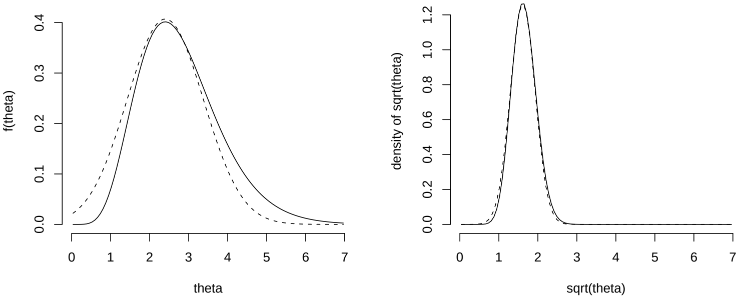
Recall the Independence Sampler:



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

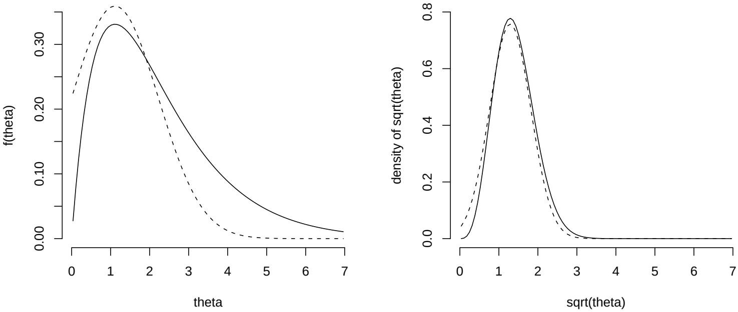
Transforming to Normality

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:

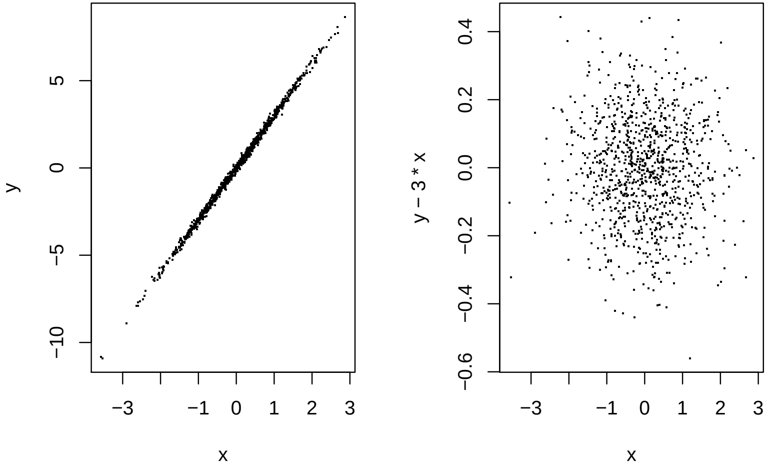
- Transform to the Real Line.
- Take the log of positive parameters.
- If the log is “too strong”, try square root.
- Probabilities can be transformed via the logit transform:

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

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

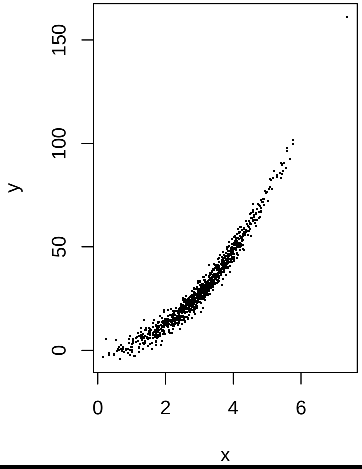
Removing Linear Correlations

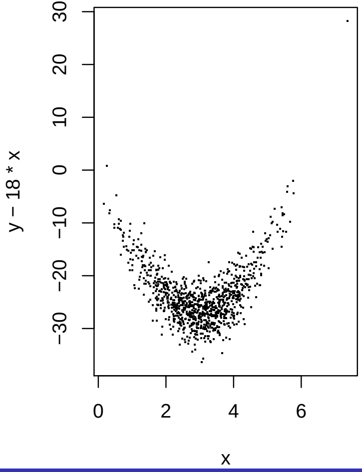
Linear transformations can remove linear correlations



Removing Linear Correlations

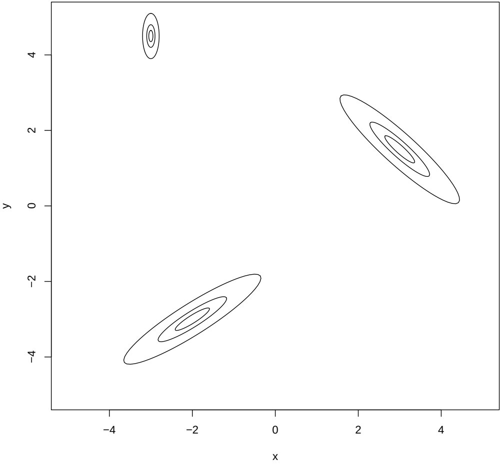
... and can help with non-linear correlations.





Multiple Modes

- Scientific meaning of multiple modes.
- Do not focus only on the major 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.

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

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

$$= e^{-\alpha \sum_{i=1}^{n} E_i^{-\beta}} \alpha^{\sum_{i=1}^{n} Y_i} \prod_{i=1}^{n} E_i^{-\beta Y_i}$$

So that

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

$$= \text{Gamma}\left(\sum_{i=1}^{n} Y_i + 1, \sum_{i=1}^{n} E_i^{-\beta}\right)$$

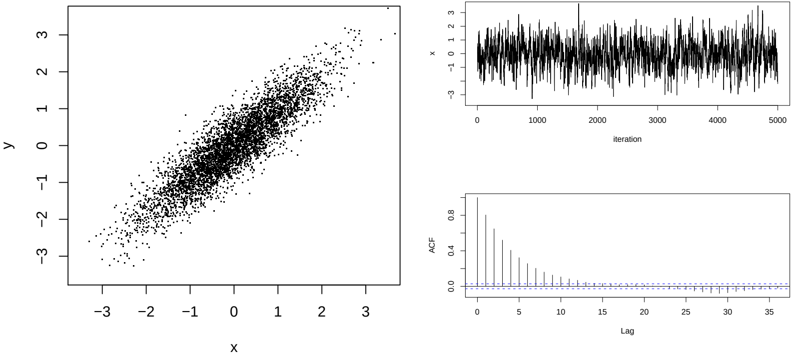
Embedding Other Samplers within Gibbs

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

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

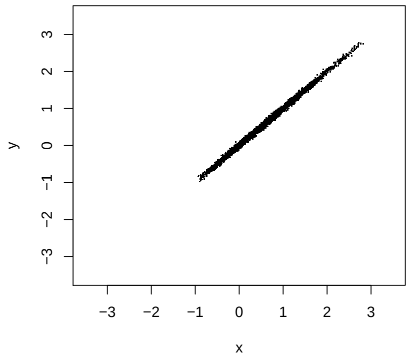
- We can use a Metropolis or Metropolis-Hastings step to update β within the Gibbs sampler.
- The result is known as *Metropolis within Gibbs Sampler*.
- **Advantage:** Metropolis tends to preform poorly in high dimensions. Gibbs reduces the dimension.
- **Disadvantage:** Case-by-case probabilistic calculations.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

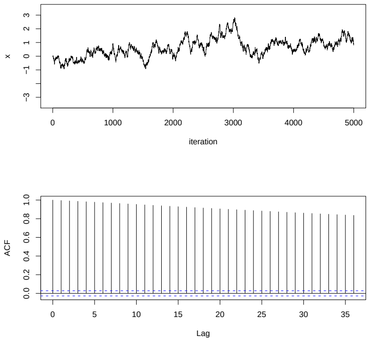
When Will Gibbs Sampling Work Well?



autocorrelation = 0.81, effective sample size = 525

When Will Gibbs Sampling Work Poorly?

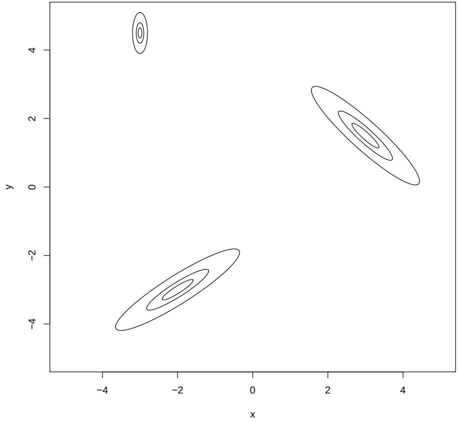




autocorrelation = 0.998, effective sample size = 5

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

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

$\vdots$

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

$\vdots$

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

Outline

Overview of Recommended Strategy

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. Compare inference based on crude approximation and MCMC. If they are not similar, check for errors before believing the results of the MCMC.