Bayesian Inference and Decision Theory

Unit 6: Markov Chain Monte Carlo

GEORGE MASON UNIVERSITY

Learning Objectives for Unit 6

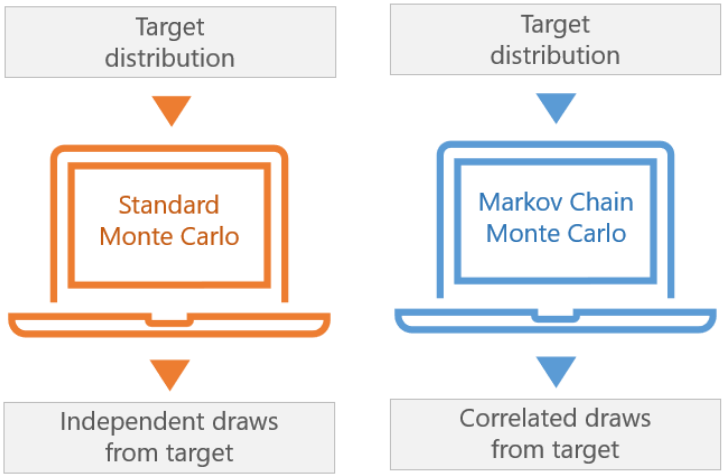
- Describe how Gibbs sampling works
- Implement a simple Gibbs sampler
- Use JAGS to perform Gibbs sampling
- Estimate posterior quantities from the output of a Gibbs sampler for the posterior distribution
- Describe some MCMC diagnostics
- Apply diagnostics to assess adequacy of MCMC sampler output
- Describe how component-wise Metropolis-Hastings sampler extends Gibbs sampler to a wider class of problems

Review: Steps in Bayesian Data Analysis

1. *Determine the question*: Use data $\underline{x} = x_1, \ldots, x_n$ to answer a question of interest to decision makers
2. *Specify the likelihood*: $f(\underline{x}|\underline{\theta})$ expresses probability distribution of data conditional on a vector of parameters
3. Specify the prior distribution: $g(\underline{\theta})$ represents beliefs about parameters prior to seeing observations $\underline{x}$
4. Find the (exact or approximate) posterior distribution:
   - For a Bayesian, the posterior distribution is everything needed to draw conclusions about $\underline{\theta}$
   - Approximation is needed when posterior distribution is intractable
5. Summarize the posterior distribution and draw conclusions:
   - We seek posterior summaries such as mean, credible interval, or predictive probabilities
   - Summaries are chosen to address the the question of interest
   - We also do analyses to check model adequacy

Step 4: Find / Approximate the Posterior Distribution

- When the prior and likelihood form a conjugate pair, we can calculate the posterior distribution and many posterior quantities exactly
  - Not all posterior quantities can be calculated exactly
  - Example: difference in independent gamma distributed defect rates for two plants
  - We can estimate some of these quantities using direct Monte Carlo



Source: https://www.statlect.com/fundamentals-of-statistics/Markov-Chain-Monte-Carlo

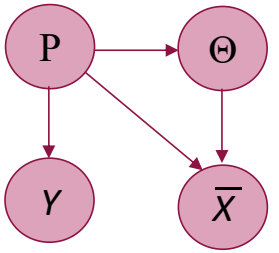
- For many interesting problems no exact posterior distribution can be found
  - We cannot use direct Monte Carlo
  - We need another way to approximate the posterior distribution
- Markov Chain Monte Carlo (MCMC) is a class of methods for taking *correlated* (not iid) draws from the posterior distribution

Markov Chain Monte Carlo (MCMC)

- General-purpose class of Monte Carlo algorithms originating in statistical physics
- Applied to problems for which both exact computation and direct Monte Carlo sampling are infeasible
- MCMC Method:
  - Construct a Markov chain with a unique stationary distribution equal to the target distribution $P(\underline{\Theta})$
  - Sample from this Markov chain
  - Use the sample to estimate $P(\underline{\Theta})$
- The most common MCMC samplers are the Gibbs sampler and a generalization called the Metropolis-Hastings sampler

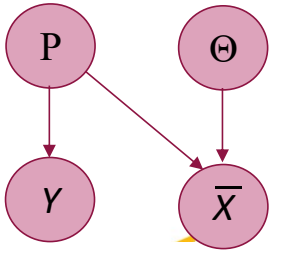
Example: Normal Random Variable with Independent Mean and Precision

- Problem: infer mean and precision of normal data
- In Unit 5 we used the normal-gamma conjugate prior
  - Prior knowledge about mean Θ and precision P are dependent
  - The greater the precision P of an observation, the more sure we are about the mean Θ
- This might not be a faithful representation of our prior information
- Consider a prior distribution in which Θ and P are independent *a priori* and:
  - P has a gamma distribution with shape α and scale β
  - Θ has a normal distribution with mean μ and standard deviation τ
- This is not a conjugate distribution
  - There is no closed-form expression for the posterior distribution



Normal-gamma conjugate prior

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad Y = \sum_{i=1}^{n}(x_i - \bar{x})^2$$



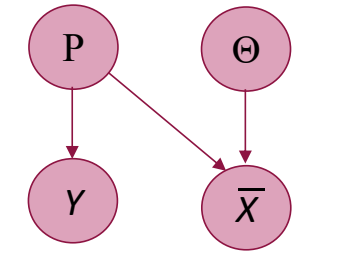
Independent normal and gamma priors (*not conjugate*)

A Semi-Conjugate Prior Distribution

- A family of distributions for two (or more) parameters is *semi-conjugate* if the distribution for each parameter given the others is conjugate
- The independent normal and gamma prior distribution is semi-conjugate
  - Observations: $X_1, \ldots, X_n \mid \Theta, \mathrm{P} \sim \text{Normal}(\Theta, 1/\sqrt{\mathrm{P}})$
  - Conditional distribution for Θ given P: (see Unit 5)
    - Prior distribution: $\Theta \sim \text{Normal}(\mu, \tau)$ is independent of P
    - Posterior distribution: $\Theta \mid \mathrm{P} = \rho, X_{1:n} \sim \text{Normal}(\mu^*, \tau^*)$

      $\tau^* = (1/\tau^2 + n\rho)^{-1/2} \quad \mu^* = \dfrac{\mu/\tau^2 + \rho \sum_i X_i}{1/\tau^2 + n\rho}$
  - Conditional distribution for P given Θ: (see next page for derivation)
    - Prior distribution: $\mathrm{P} \sim \text{Gamma}(\alpha, \beta)$ is independent of Θ
    - Posterior distribution: $\mathrm{P} \mid \Theta = \theta, X_{1:n} \sim \text{Gamma}(\alpha^*, \beta^*)$

      $\alpha^* = \alpha + n/2 \qquad \beta^* = \left(\beta^{-1} + \frac{1}{2}\sum_i (X_i - \theta)^2\right)^{-1}$



$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad Y = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Semi-Conjugate Prior Distribution (Details)

- Distribution for Θ given $\mathrm{P} = \rho$ and $X_{1:n}$ is just the case of known standard deviation from Unit 5:

  $\Theta \mid \mathrm{P} = \rho, X_{1:n} \sim$ Normal(μ^*, τ^*), where $\tau^* = (1/\tau^2 + n\rho)^{-1/2}$ and $\mu^* = \dfrac{\mu/\tau^2 + \rho \sum_i X_i}{1/\tau^2 + n\rho}$

- We find the distribution for P given $\Theta = \theta$ and $X_{1:n}$ by considering the limiting case of the normal-gamma distribution as the precision multiplier tends to infinity (i.e. the mean has infinite precision *a priori*)

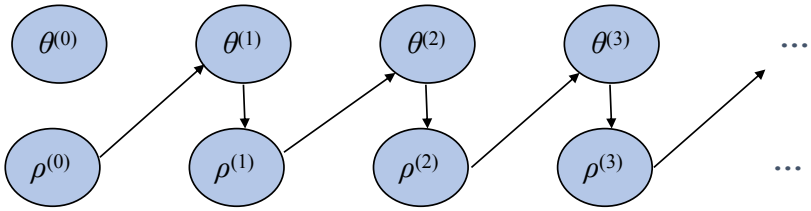
  If $X_{1:n}$ are iid Normal$(\Theta, \mathrm{P}^{-1/2})$ and prior distribution for (Θ, P) is Normal-Gamma(μ, k, α, β) then posterior distribution for P is Gamma(α^*, β^*) with

$$\alpha^* = \alpha + \frac{n}{2}$$

$$\begin{aligned}
\beta^* &= \left(\beta^{-1} + \frac{1}{2}\sum_i (x_i - \bar{x})^2 + \frac{kn}{2(k+n)}(\bar{x} - \mu)^2\right)^{-1} \\
&= \left(\beta^{-1} + \frac{1}{2}\sum_i (x_i - \bar{x})^2 + \frac{n}{2(1+n/k)}(\bar{x} - \mu)^2\right)^{-1} \\
&\to \left(\beta^{-1} + \frac{1}{2}\sum_i (x_i - \bar{x})^2 + \frac{n}{2}(\bar{x} - \mu)^2\right)^{-1} \text{ as } k \to \infty \\
&= \left(\beta^{-1} + \frac{1}{2}\sum_i (x_i - \mu)^2\right)^{-1} = \left(\beta^{-1} + \frac{1}{2}\sum_i (x_i - \theta)^2\right)^{-1}
\end{aligned}$$

Conditioning on $\Theta = \theta$ and $X_{1:n}$ means assuming that $\Theta = \theta$ is known with infinite precision ($k \to \infty$), to be equal to its prior mean μ.

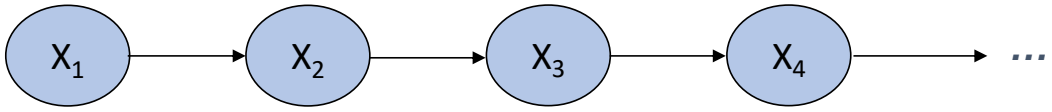
Gibbs Sampling with a Semi-Conjugate Prior



- We cannot approximate with direct Monte Carlo because we have no closed-form expression for the joint posterior distribution of (Θ, P) given $X_{1:n}$
- We can approximate using Gibbs sampling
- Gibbs sampling for the normal model with unknown mean and precision:
  - INITIALIZE: Chose arbitrary initial parameter values $\theta^{(0)}, \rho^{(0)}$
  - SAMPLE: For $k = 1, \ldots, M$
    - Sample $\theta^{(k)}$ from $g(\theta \mid \rho^{(k-1)}, x_{1:n})$ – normal distribution
    - Sample $\rho^{(k)}$ from $g(\rho \mid \theta^{(k)}, x_{1:n})$ – gamma distribution
- Facts about Gibbs sampling:
  - Successive draws are correlated: $(\theta^{(k)}, \rho^{(k)})$ depends on $(\theta^{(k-1)}, \rho^{(k-1)})$
  - The sequence $(\theta^{(1)}, \rho^{(1)})$, $(\theta^{(2)}, \rho^{(2)})$, … is a Markov chain
  - This Markov chain has a unique stationary distribution equal to the posterior distribution of (Θ, P) given $X_{1:n}$
  - We can use the samples $(\theta^{(1)}, \rho^{(1)})$, $(\theta^{(2)}, \rho^{(2)})$, … $(\theta^{(M)}, \rho^{(M)})$ to approximate posterior quantities of interest

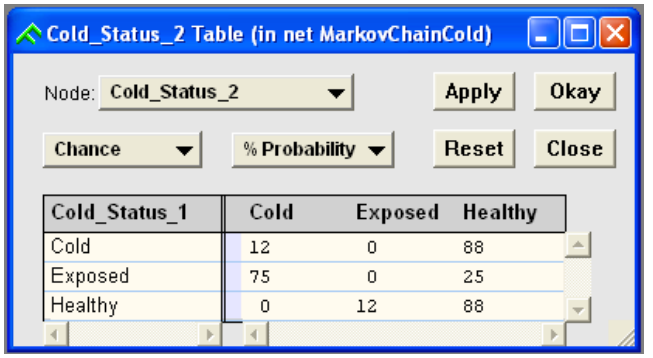
Review: Markov Chain

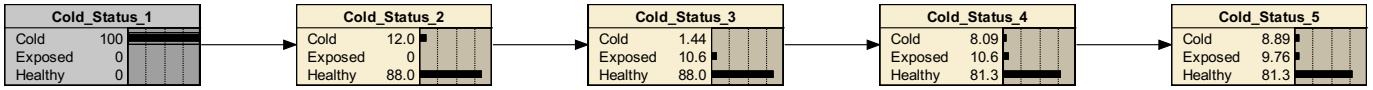
- A Markov chain (of order 1) is a sequence of random variables $X_1, X_2, \ldots$ such that X_i is independent of all lower-numbered X_j $(j < i-1)$ given X_{i-1}
  - $\Pr(X_i \mid X_1, X_2, \ldots X_{i-1}) = \Pr(X_i \mid X_{i-1})$
  - The X_i can be univariate or multivariate
- In an order k Markov chain, X_i is independent of all lower-numbered X_j $(j < i)$ given $X_{i-1}, \ldots, X_{i-k}$
- Under fairly general conditions a Markov chain has a unique stationary distribution $\pi(x)$
  - If X_i has distribution $\pi(x)$ then so does X_{i+1}

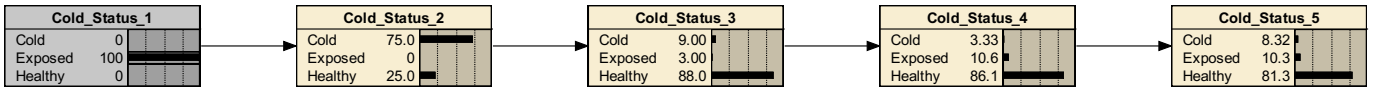


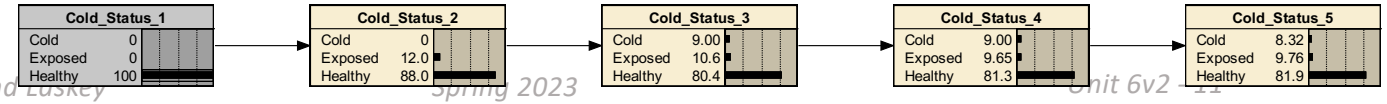
Markov Chain Example

- States: Cold, Exposed, Healthy
- Allowable transitions:
  - Cold → Cold (p=0.12)
  - Cold → Healthy (p=0.88)
  - Exposed → Cold (p=0.75)
  - Exposed → Healthy (p=0.25)
  - Healthy → Exposed (p=0.12)
  - Healthy → Healthy (p=0.88)
- Unique stationary distribution
  - P_{st}(Cold) = 0.0837; P_{st}(Exposed) = 0.0982; P_{st}(Healthy) = 0.8181
  - All initial distributions evolve to this stationary distribution





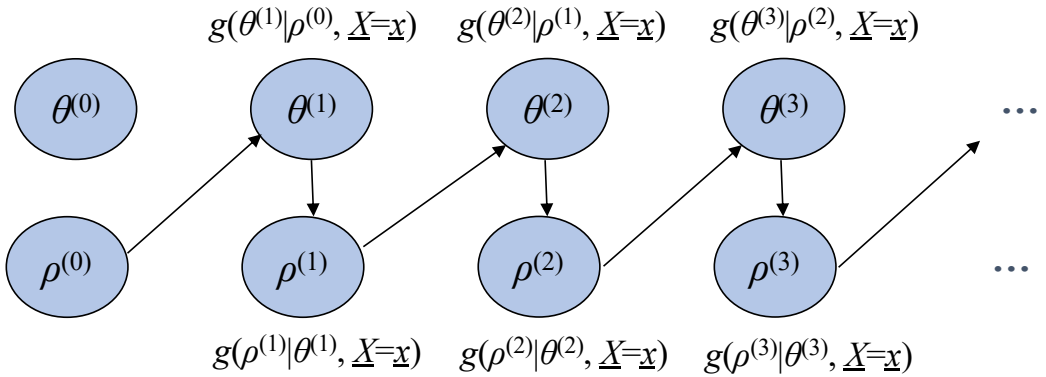




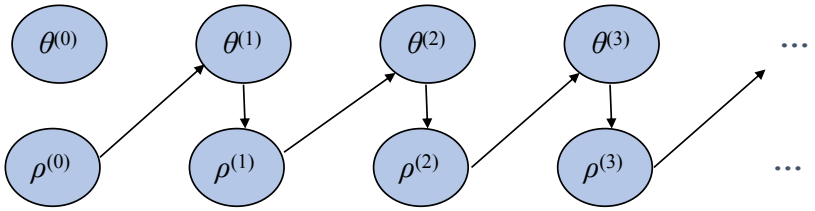
Screenshots from Netica™ tool for Bayesian networks

Gibbs Sampler for Normal Model with Semi-Conjugate Prior is a Markov Chain

- INITIALIZE: Chose arbitrary initial parameter values $\theta^{(0)}$, $\rho^{(0)}$
- SAMPLE: For k = 1, …, M
  - Sample $\theta^{(k)}$ from $g(\theta\,|\rho^{(k-1)}, \underline{X} = \underline{x})$ (Normal(μ^*, τ^*) posterior distribution of Θ given $\rho^{(k-1)}$ and $\underline{X} = \underline{x}$;
  - Sample $\rho^{(k)}$ from $g\,(\rho\,|\theta^{(k)}, \underline{X} = \underline{x})$ (Gamma(α^*, β^*) posterior distribution of P given $\theta^{(k)}$ and $\underline{X} = \underline{x}$
- This process gives a Markov chain with states ($\theta^{(k)}$, $\rho^{(k)}$)
  - ($\theta^{(k)}$, $\rho^{(k)}$) is independent of the past given ($\theta^{(k-1)}$, $\rho^{(k-1)}$)
- This Markov chain has a unique stationary distribution equal to the joint posterior distribution of (Θ, P) given $X_{1:n}$



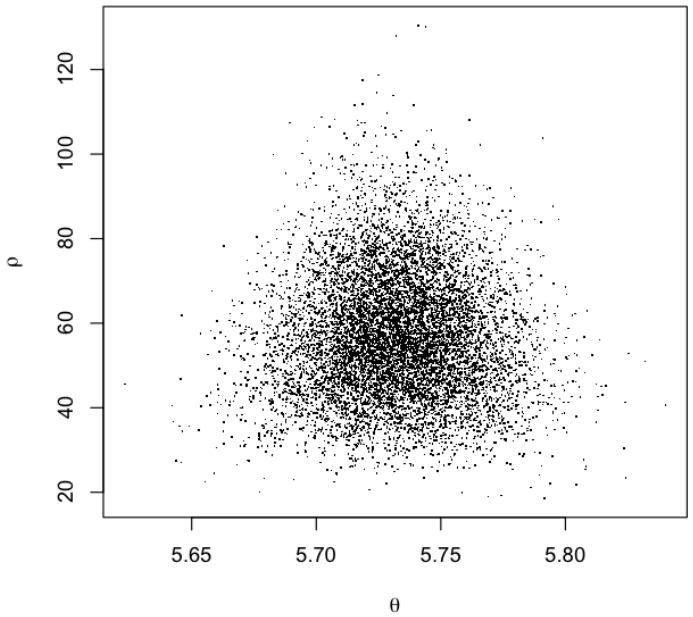
Reaction Time Example: Gibbs Sampling

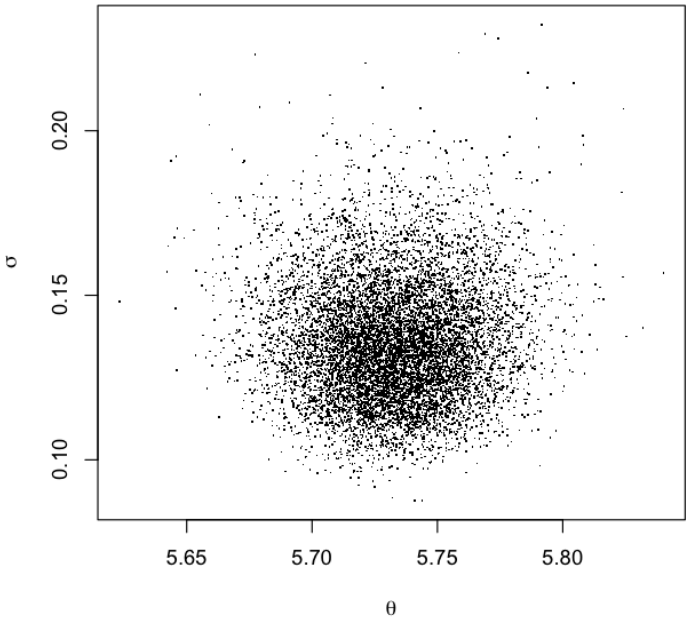


- We analyzed log reaction times using a non-informative prior distribution $g(\theta, \rho) \propto \rho^{-1}$
- This is an improper member of the conjugate family so we can find an exact posterior distribution
  - Normal-gamma(μ, k, α, β) distribution with $\mu = 0, k = 0, \alpha = -\frac{1}{2}, \beta = \infty$
  - This is actually an independent uniform distribution and a gamma distribution
- We will apply Gibbs sampling and compare to exact result
- Component-wise conditional posterior distributions:
  - Posterior distribution of Θ given $\mathrm{P} = \rho$ and $x_1, \ldots, x_n$ is normal with
    - Mean $\mu^* = (\frac{0}{\infty} + \rho n \bar{x})/(\frac{1}{\infty} + n\rho) = \bar{x} = 5.73$ and standard deviation $\tau^* = \left(\frac{1}{\infty} + n\rho\right)^{-1/2} = 1/\sqrt{30\rho}$
  - Posterior distribution of P given $\Theta = \theta$ and $x_1, \ldots, x_n$ is gamma with
    - Shape $\alpha^* = \frac{n}{2} - \frac{1}{2} = 14.5$
    - Scale $\beta^* = \left(\frac{1}{\infty} + \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right)^{-1} = \left(\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right)^{-1} = \left(\frac{1}{2}\sum_{i=1}^{n} x_i^2 - \theta\sum_{i=1}^{n} x_i + \frac{n}{2}\theta^2\right)^{-1}$
      $= (493.1 - 172.0\theta + 15\theta^2)^{-1}$
- We sample repeatedly from these distributions to find the Gibbs sampling estimate of the posterior distribution

Scatterplots for Gibbs Sampler Output

- Scatterplots for 10,000 samples from Gibbs sampler for posterior distribution given 30 observations on the first non-schizophrenic subject
  - Left: joint distribution of mean and precision
  - Right: joint distribution of mean and standard deviation

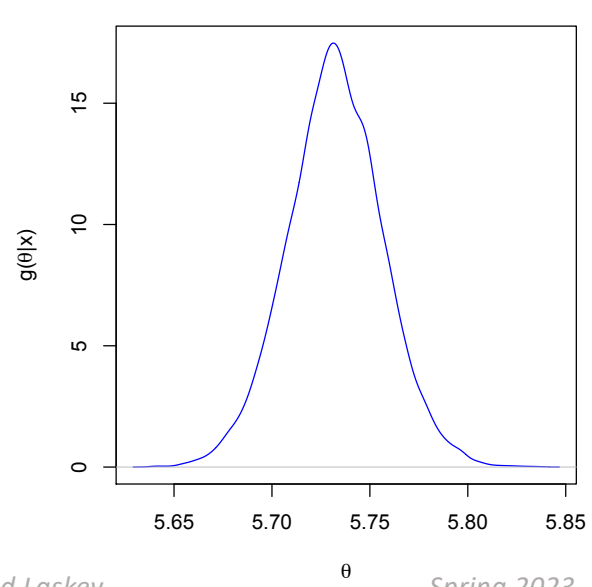


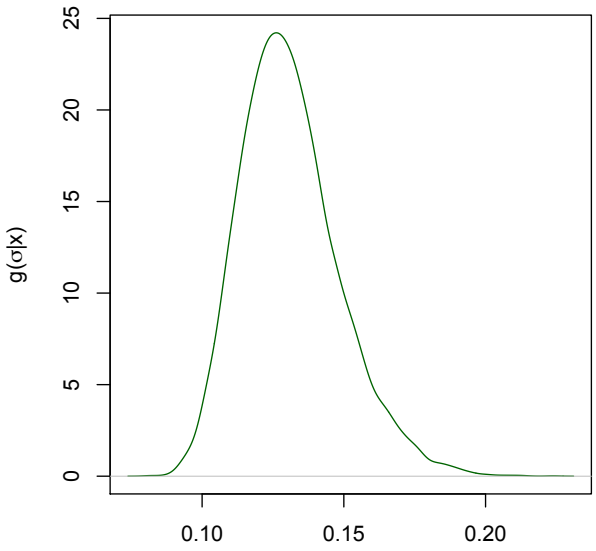


Results: Gibbs Sampling for Reaction Times with Semi-Conjugate Prior

- 10,000 samples were drawn from the Gibbs sampler for the posterior distribution given the 30 reaction time observations:
  - 95% credible interval for Θ: [5.68, 5.78]
  - 95% credible interval for Σ: [0.102, 0.171]

Kernel Density Plots for Marginal Posterior Densities of Θ and Σ



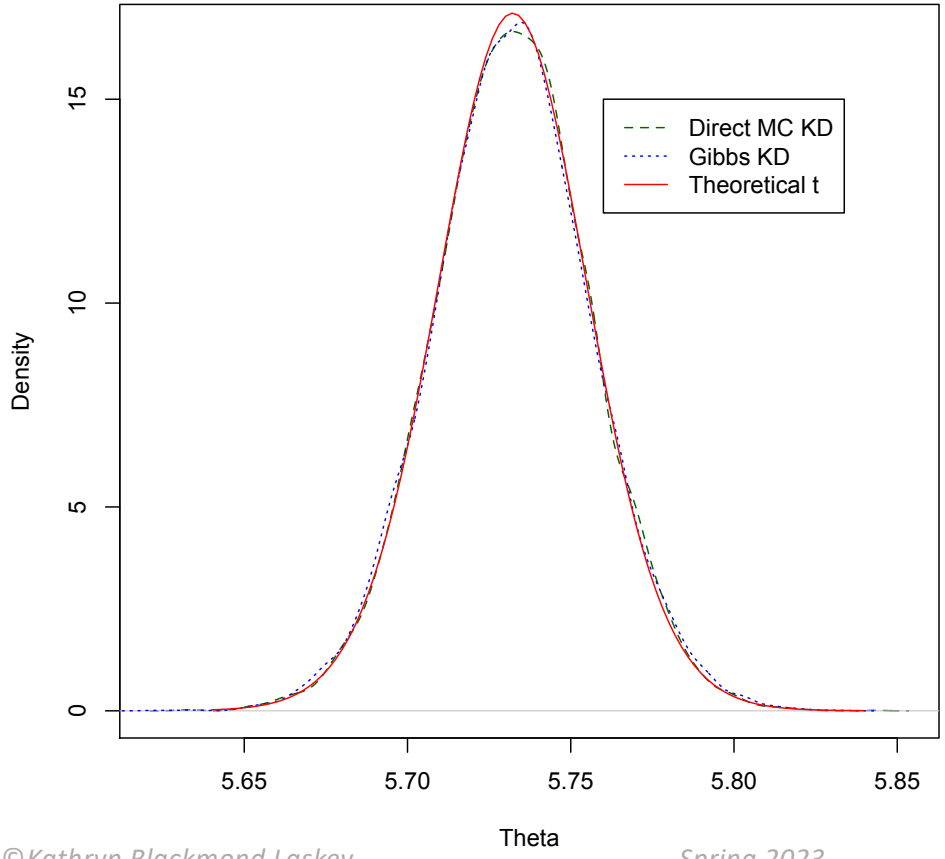


R code is available on Blackboard

Comparison: Exact, Direct MC and Gibbs for Normal Model

- In Unit 5 we used normal-gamma conjugate updating to find a posterior distribution for (Θ, P)
  - Marginal distribution of P is gamma
  - Marginal distribution of Θ is nonstandard t
- We can approximate this distribution via direct Monte Carlo
  - Simulate precision ρ from its posterior gamma distribution, and simulate θ from its posterior normal distribution with precision that depends on ρ
- We can also approximate this distribution by Gibbs sampling
  - Sample $\theta^{(m)}$ from a normal distribution that depends on the observations and $\rho^{(m-1)}$
  - Sample $\rho^{(m)}$ from a gamma distribution that depends on the observations and $\theta^{(m)}$
- Because of the improper prior, we have an exact posterior, but we can use Gibbs sampling when we do not have an exact posterior

Comparison of Posterior Density Estimates



- Plot compares exact and approximate posterior density functions for mean of reaction time distribution
  - Dashed green line shows kernel density estimate from 10,000 direct Monte Carlo samples
  - Dotted blue line shows kernel density estimate from 10,000 Gibbs samples
  - Solid red line shows posterior t density with center μ^* = 5.732, spread $1/(k^*\alpha^*\beta^*)^{1/2}$ = 0.0231, and degrees of freedom $2\alpha^*$ = 29

R code is available on Blackboard

©Kathryn Blackmond Laskey

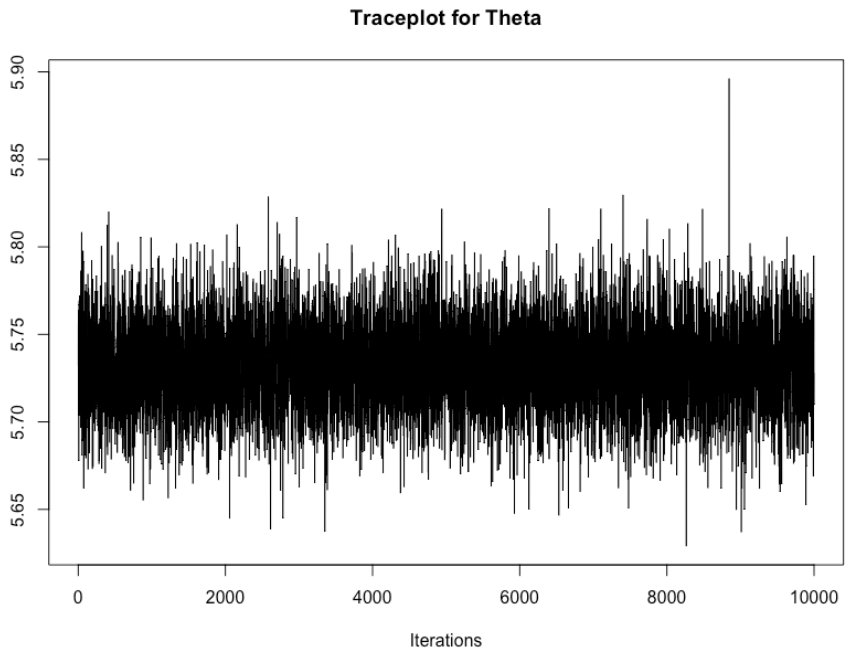
Review: Markov Chain Monte Carlo (MCMC)

- General-purpose class of Monte Carlo algorithms
- Method: estimate $P(X)$ using samples from a Markov Chain constructed to have $P(X)$ as its unique stationary distribution
- MCMC takes *correlated* samples so is inherently less efficient than direct Monte Carlo
  - Sampler can get "stuck" in regions near a local mode of the distribution, yielding a poor approximation to the target distribution
- We need to use MCMC when we cannot find an exact posterior distribution
  - MCMC is used to estimate posterior distribution with a semi-conjugate prior distribution

MCMC Diagnostics

- MCMC diagnostics help us to identify problems with the sampler and to assess whether we have collected enough samples to get a good approximation to the target distribution
- Examples:
  - *Traceplot* – plots a parameter against iteration number to help diagnose whether the sampler is getting stuck in a local region
  - *Sample autocorrelation function (acf)* – evaluates correlation between elements of the sequence as a function of the time separation
  - *Effective sample size* – Uses acf to estimate the number of independent MC draws needed to achieve same precision as the MCMC samples
  - *Convergence diagnostics* – Run parallel MCMC chains and evaluate convergence using within and between chain variance

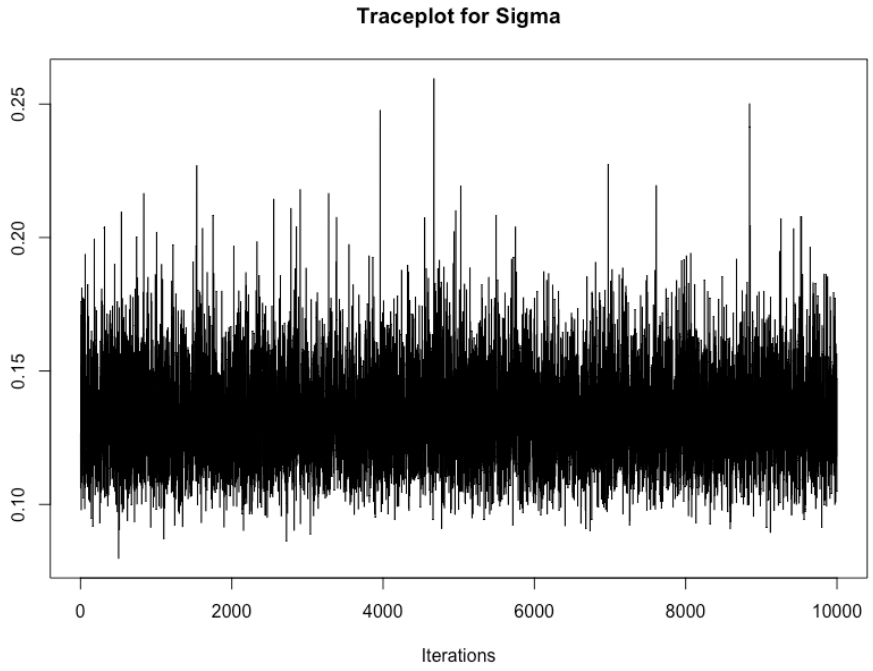
Traceplots for Reaction Time Data



Effective sample size 10,000

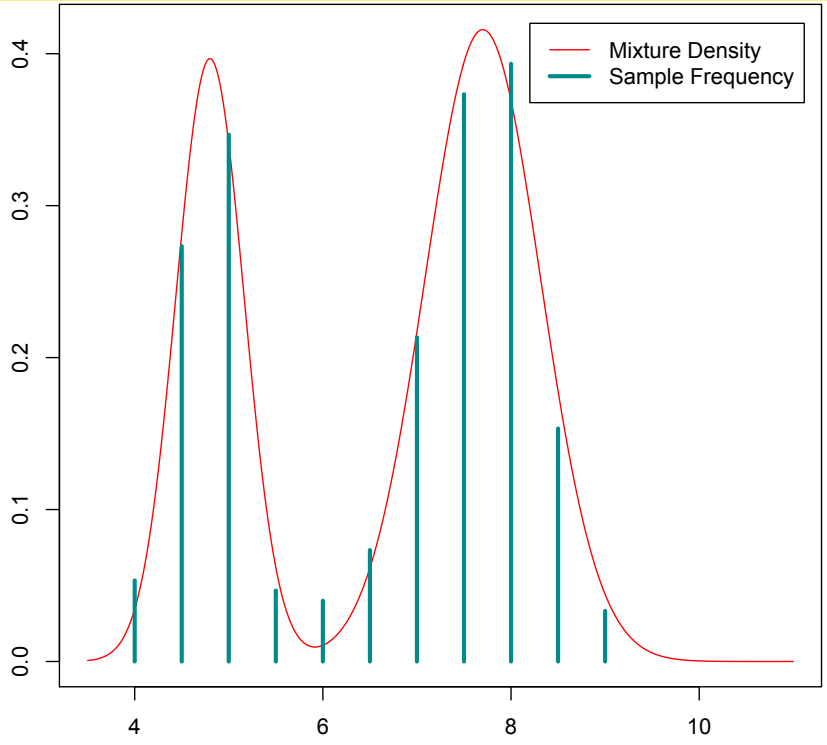
- Plot MCMC sample versus iteration number
- Helps to assess convergence of sampler

Effective sample size 9479



What Can Go Wrong: Weaver Ants Example

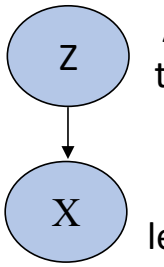
- Body lengths of weaver ant workers show a bimodal distribution
  - Minor workers are a little more than half the size of major workers
  - There is very little overlap in the size distributions
- A mixture of two normal distributions provides a good model for the body length data
  - Minor workers (36%)
    - Mean 4.8 mm
    - Std dev 0.36 mm
  - Major workers (64%)
    - Mean 7.7 mm
    - Std dev 0.61 mm



Source: Weber, NA (1946). Dimorphism in the African Oecophylla worker and an anomaly (Hym.: Formicidae). *Annals of the Entomological Society of America 39*: pp. 7–10. http://antbase.org/ants/publications/10434/10434.pdf

R Code for Ant Mixture Density Plot

```
#Ants example
#Data from Weber, 1946, "Dimorphism in the African Oecophylla Worker and an Anomaly
#http://antbase.org/ants/publications/10434/10434.pdf

# Mixture model
mu1 <- 4.8
sd1 <- 0.36
mu2 <- 7.7
sd2 <- 0.61
pr1 <- 0.36
pr2 <- (1-pr1)

xvals <- 350:1100/100
mixDens <- pr1*dnorm(xvals,mu1,sd1)+pr2*dnorm(xvals,mu2,sd2)

# Sample Frequencies and Mixture Density Plot
plot(xvals,mixDens,col="red",type="l",main="",ylab="",xlab="Body Length (mm)")
```



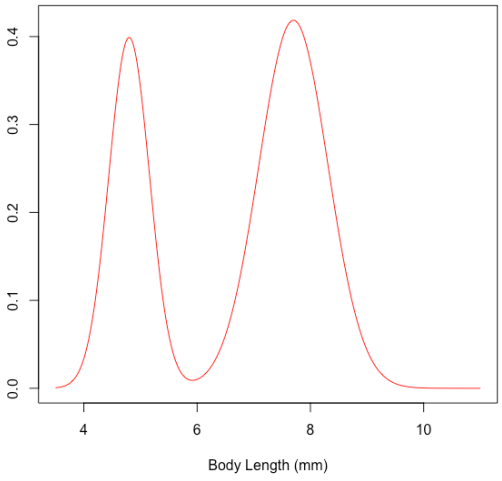
Ant type

$$Z = \begin{cases} 1 & w.p.\ 0.36 \\ 2 & w.p.\ 0.64 \end{cases}$$

Ant length

$$X|Z = 1 \sim Normal(4.8, 0.36)$$

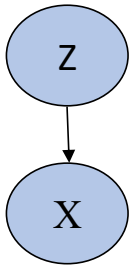
$$X|Z = 2 \sim Normal(7.7, 0.61)$$



Marginal density of X

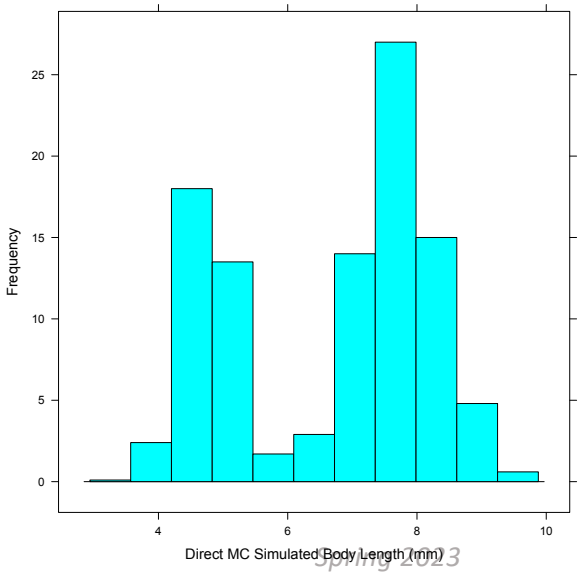
Direct Monte Carlo for Ant Lengths

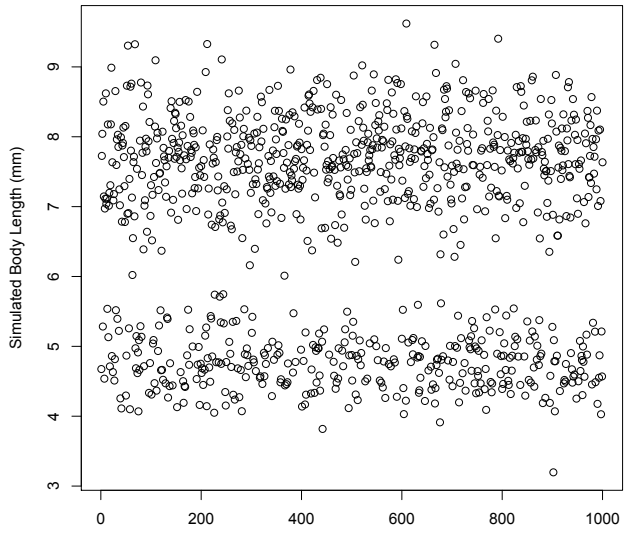
- The distribution for ant lengths can be simulated directly:
  - Simulate z = 1 with probability 0.36 and z=2 with probability 0.64
  - If z=1 simulate length from Normal(4.8, 0.36)
  - If z=2 simulate length from Normal(7.7, 0.61)
- This produces an iid sample of ant lengths



Graphical probability model for ant lengths

Histogram of 1000 Direct MC Values

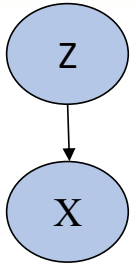




Trace Plot of 1000 Direct MC Values

Gibbs Sampling for Ant Lengths (not a good idea!)

- To illustrate how MCMC can go wrong we use Gibbs sampling to simulate ant length distribution (this is *not* a good way to approach this problem!)
  - Initialize length x_0
  - For k from 1 to desired sample size:
    - Calculate $L_1 = f(x^{(k-1)} | 4.8, 0.36)$ (normal density with mean 4.8, sd 0.36)
    - Calculate $L_2 = f(x^{(k-1)} | 7.7, 0.61)$ (normal density with mean 7.7, sd 0.61)
    - Calculate $p_1 = 0.36L_1 / (0.36L_1 + 0.64L_2)$
    - Simulate $z^{(k)}$= 1 or 2, with probabilities p_1 and 1- p_1 respectively
    - If $z^{(k)}$= 1 simulate length $x^{(k)}$ from Normal(4.8, 0.36)
    - If $z^{(k)}$= 2 simulate length $x^{(k)}$ from Normal(7.7, 0.61)

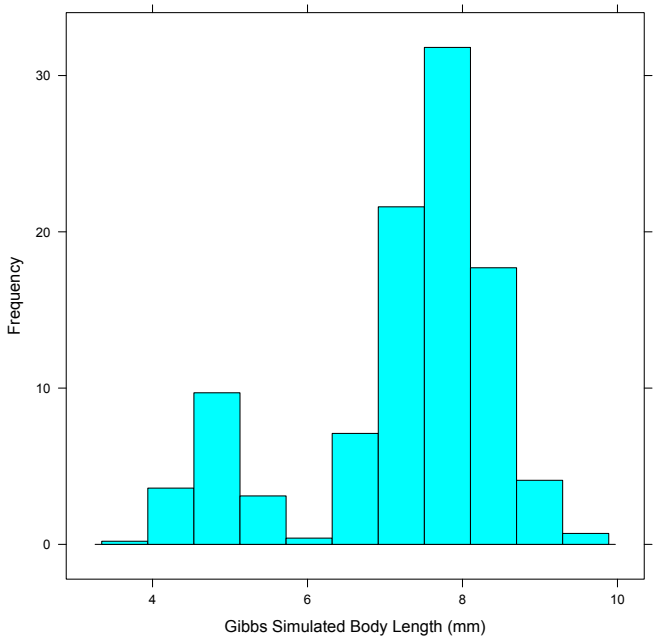


Likelihood of $x^{(k-1)}$ if $Z^{(k-1)} = 1$
Likelihood of $x^{(k-1)}$ if $Z^{(k-1)} = 2$
Probability $Z^{(k-1)} = 1$ given $x^{(k-1)}$

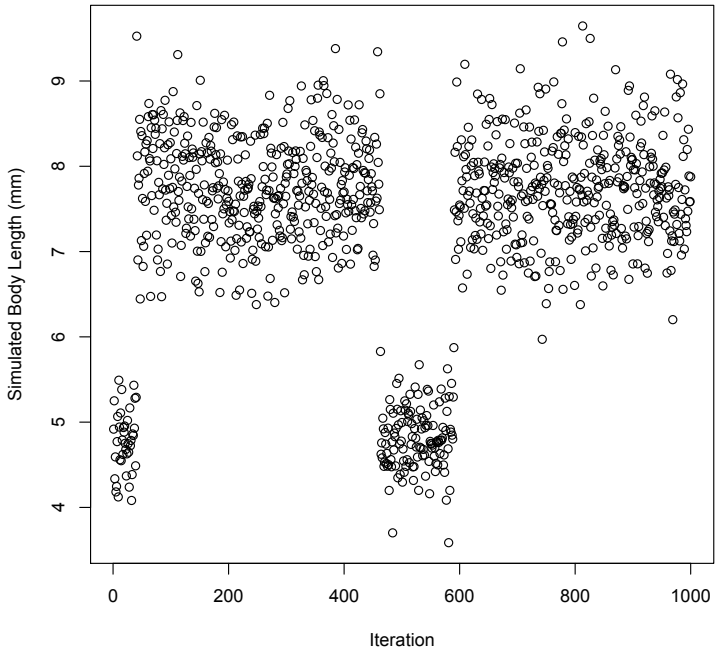
- This produces a sample of ant lengths
  - Consecutive observations are correlated
  - This is a Markov chain with stationary distribution equal to the target mixture distribution
  - Due to high correlation between successive samples, this is a very inefficient way to simulate samples from the target distribution

Gibbs Sampling Results (1000 Samples)

Histogram of 1000 Gibbs Samples

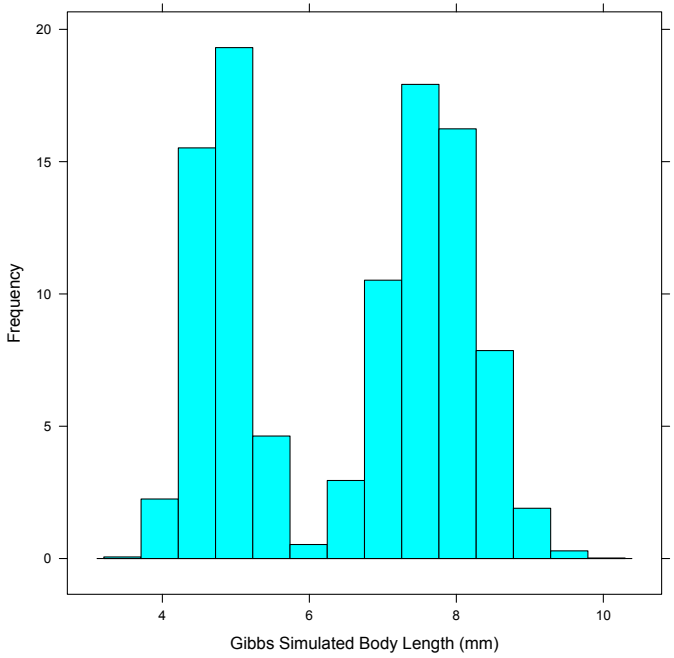


Trace Plot of 1000 Gibbs Samples

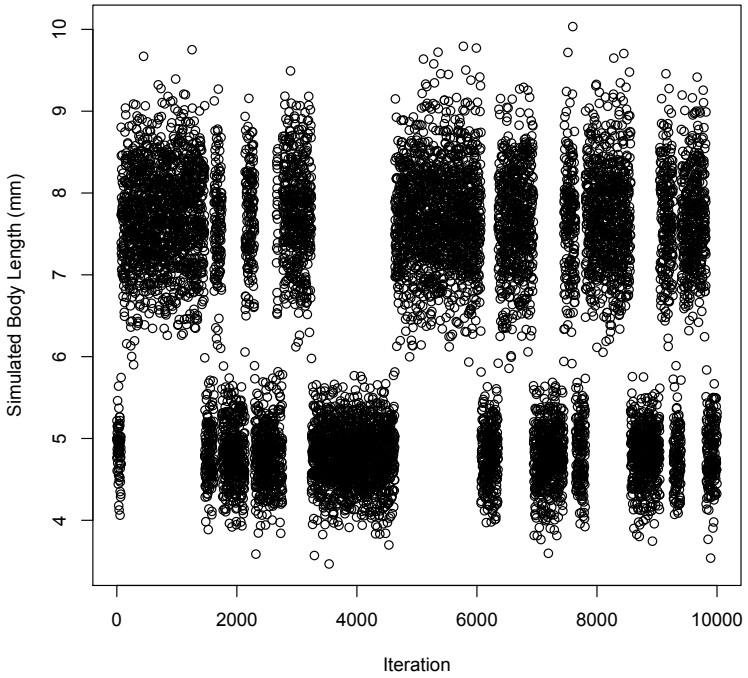


Gibbs Sampling Results (10,000 Samples)

Histogram of 10,000 Gibbs Samples



Trace Plot of 10,000 Gibbs Samples

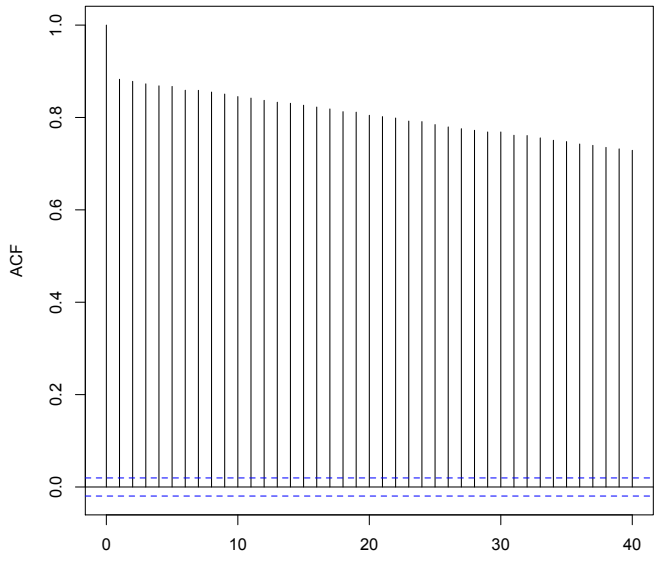


Autocorrelation Function

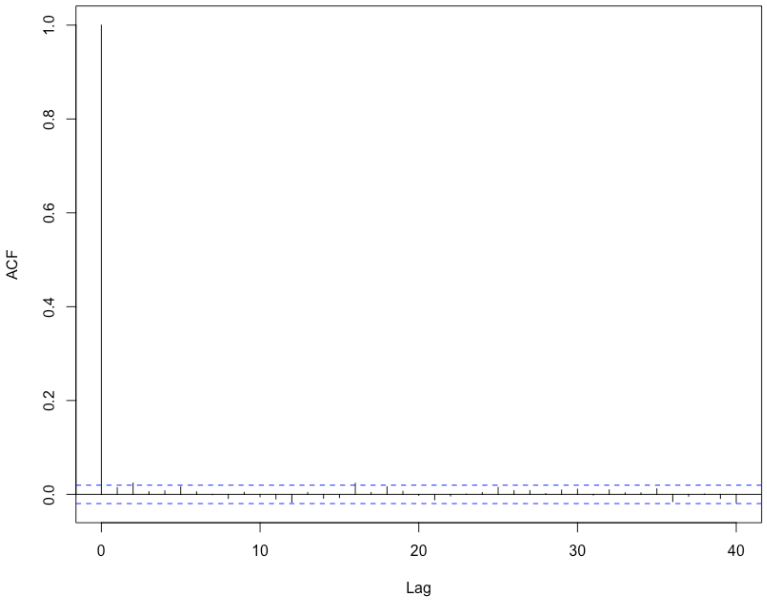
- The lag-k autocorrelation function (acf) estimates correlation between observations k steps apart
- The lag-1 autocorrelation for the 10,000 Gibbs simulations of ant length is 0.882
- The lag-1 autocorrelation for the 10,000 direct MC simulations of ant length is 0.014

R command:
`acf(x)`

ACF for 10,000 Gibbs Samples for Ant Length



ACF for 10000 Direct MC Samples for Ant Length



©Kathryn Blackmond Laskey

Effective Sample Size

- We can use the autocorrelation function to estimate the number of independent draws we would need to give the same precision as our MCMC sample
- The `effectiveSize` function in R calculates such an estimate
  - To use this function, you must load the "`coda`" package
  - This package provides output analysis and diagnostics for MCMC samples
- Effective sample sizes for reaction time and ant length simulations
  - For 10,000 Gibbs samples of reaction time mean and standard deviation, effective size was 10,000 for the mean and 9,479 for the standard deviation
  - For 10,000 direct MC samples of ant lengths, the effective size was 10,000
  - For 10,000 Gibbs samples of ant lengths, the effective size was **28.5**

Other MCMC Diagnostics

- Potential scale reduction (Gelman and Rubin, 1992)
  - Compares within and between variance components for multiple MCMC chains
  - Chains are started at overdispersed starting points; convergence occurs when output of all chains is indistinguishable
  - Large values (above 1.1 or 1.2) suggest chain has not converged
- Geweke (1992) z-score
  - Test for equality of means of first and last part of a Markov chain
  - "Burn-in" period may be discarded – but more than half the chain should be retained
- Heidelberger and Welch (1983) diagnostic
  - Uses Cramer-von-Mises statistic to test null hypothesis that sampled values come from a stationary distribution.
- Raftery and Lewis (1992) diagnostic
  - Use on a short pilot run of the chain
  - Provides information on number of iterations needed to obtain estimate of given accuracy

These diagnostics are available as part of the "coda" package in R. Documentation is available at http://cran.r-project.org/web/packages/coda/coda.pdf

Remarks

- Direct MC is preferable to Gibbs sampling when feasible
- For many problems, MCMC is necessary because we cannot sample directly from the target distribution of interest
  - We must be careful to draw enough samples for reliable inference
  - We must be on the lookout for problems such as multimodality
- In hard problems it can be very difficult to assess whether realizations of a MCMC sampler provide an adequate approximation to the target distribution
- Although MCMC diagnostics are helpful, it is possible for a chain to be "stuck" in a local optimum without being detected by MCMC diagnostics
- For highly multi-modal problems it may be that the best we can do is find a "good" local mode of the posterior distribution

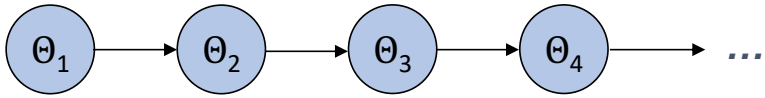
The General Gibbs Sampler

- The Gibbs sampler is used to estimate $g(\underline{\theta}|\underline{x}) = g(\theta_1, \theta_2, \ldots, \theta_p|x_1, x_2, \ldots, x_n)$ for problems in which we can sample from each of the "full conditional" distributions $g(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p, \underline{x})$
- Gibbs sampling proceeds as follows:
  - INITIALIZE: Chose initial parameter values, $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)}$
  - SAMPLE: For m = 1, …, M
    - Sample $\theta_1^{(m)}$ from $g(\theta_1|\theta_2^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})$
    - Sample $\theta_2^{(m)}$ from $g(\theta_2|\theta_1^{(m)}, \theta_3^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})$
    - …
    - Sample $\theta_i^{(m)}$ from $g(\theta_i|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_{i+1}^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})$
    - …
    - Sample $\theta_p^{(m)}$ from $g(\theta_p|\theta_1^{(m)}, \ldots, \theta_{p-1}^{(m)}, \underline{x})$
- Because the distribution of $\theta_1^{(m)}, \ldots, \theta_p^{(m)}$ is independent of previous samples given $\theta_1^{(m-1)}, \ldots, \theta_p^{(m-1)}$, this process samples from a Markov chain
- Under fairly general conditions $g(\underline{\theta}|\underline{x})$ is the unique stationary distribution

Generalizing the Gibbs Sampler

- Gibbs sampler can be used only when we have a semi-conjugate prior distribution
  - Conditional distribution for each parameter given all the others is a conjugate family
  - This means we can find exact conditional posterior distributions for all parameters given the other parameters and the data
- The component-wise Metropolis-Hastings sampler generalizes the Gibbs sampler to problems for which we do not have a semi-conjugate distribution

Basic Metropolis Sampler

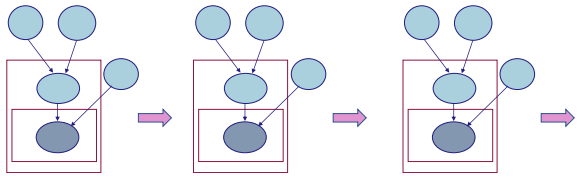


- Goal: simulate from posterior distribution $g(\theta|\underline{X} = \underline{x})$
- Assume: We can calculate prior density $g(\theta)$ and likelihood $f(\underline{x}|\theta)$
- Procedure:
  - Initialize: Set initial value $\theta^{(0)}$; set width δ
  - Sample: For $m = 1, \ldots, M$
    - Simulate trial value θ_{trial} from uniform distribution on $(\theta^{(m-1)} - \delta, \theta^{(m-1)} + \delta)$
    - Calculate acceptance ratio
      $$R = \frac{f(\underline{x}|\theta_{trial})g(\theta_{trial})}{f(\underline{x}|\theta^{(m-1)})g(\theta^{(m-1)})}$$
    - If $R \geq 1$, accept and set $\theta^{(m)} = \theta_{trial}$
    - If $R < 1$, reject with probability $1 - R$, i.e., set $\theta^{(m)} = \begin{cases} \theta_{trial} & \text{with probability } R \\ \theta^{(m-1)} & \text{with probability } 1 - R \end{cases}$
- The sequence $(\theta^{(0)}, \ldots, \theta^{(M)})$ is a Markov chain with (under mild regularity conditions) a unique stationary distribution $g(\theta|\underline{X} = \underline{x})$

Extensions of Basic Metropolis Sampler

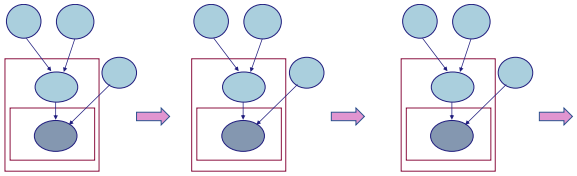
- The Hastings correction allows sampling from a distribution other than uniform
  - This is called the Metropolis-Hastings sampler
- The component-wise Metropolis (or Metropolis-Hastings) sampler samples one random variable at a time from a multivariate state $(\theta_1, \theta_2, \ldots, \theta_p)$
- The Gibbs sampler is a special case of the component-wise Metropolis-Hastings sampler that samples from the "full conditional" distributions $g(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p, \underline{x})$ of each component given the others

Review: Gibbs Sampler



- Objective: estimate $g(\underline{\theta}|\underline{x}) = g(\theta_1, \theta_2, \ldots, \theta_p | x_1, x_2, \ldots, x_n)$
- Assume: we cannot sample directly from $g(\underline{\theta}|\underline{x})$, but we can sample from each of the "full conditional" distributions $g(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p, \underline{x})$
- Gibbs sampling procedure:
  - INITIALIZE: Chose initial values, $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)}$
  - SAMPLE: For m = 1, …, M
    - Sample $\theta_1^{(m)}$ from $g(\theta_1|\theta_2^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})$
    - …
    - Sample $\theta_i^{(m)}$ from $g(\theta_i|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_{i+1}^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})$
    - …
    - Sample $\theta_p^{(m)}$ from $g(\theta_p|\theta_1^{(m)}, \ldots, \theta_{p-1}^{(m)}, \underline{x})$
- This sampling process is a Markov chain, because the distribution of $\theta_1^{(m)}, \ldots, \theta_p^{(m)}$ is independent of the past given $\theta_1^{(m-1)}, \ldots, \theta_p^{(m-1)}$
- Under fairly general conditions $g(\underline{\theta}|\underline{x})$ is the unique stationary distribution

Component-Wise Metropolis-Hastings Sampler

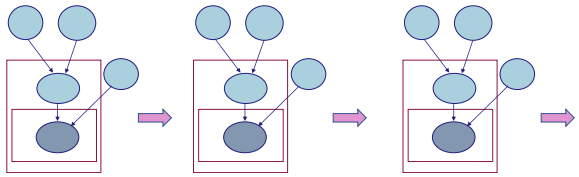


- For many inference problems:
  - We cannot compute the full conditional distributions $g(\theta_i|\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p, \underline{x})$
  - For any two values θ_i and θ_i' we can easily compute the ratio

$$\frac{g(\theta_i|\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p, \underline{x})}{g(\theta_i'|\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p, \underline{x})} = \frac{f(\underline{x}|\theta_1, \dots, \theta_{i-1}, \theta_i, \theta_{i+1}, \dots, \theta_p)g(\theta_1, \dots, \theta_{i-1}, \theta_i, \theta_{i+1}, \dots, \theta_p)}{f(\underline{x}|\theta_1, \dots, \theta_{i-1}, \theta_i', \theta_{i+1}, \dots, \theta_p)g(\theta_1, \dots, \theta_{i-1}, \theta_i', \theta_{i+1}, \dots, \theta_p)}$$

- For such problems we can use component-wise MH sampler
- The process is similar to Gibbs sampling but samples with a different distribution
  - Sample a trial value not from the full conditional distribution but from some other tractable proposal distribution
  - Either accept the newly sampled value as our new value or keep the old value with probability that depends on the ratio of full conditional distributions
  - Acceptance probability is chosen to achieve the intended stationary distribution

Details: Component-Wise MH Procedure



- Objective: estimate $g(\underline{\theta}|\underline{x}) = g(\theta_1, \theta_2, \ldots, \theta_p|x_1, x_2, \ldots, x_n)$
- MH sampling procedure with proposal distributions $h_i(\theta_i|\theta_1, \theta_2, \ldots, \theta_p, \underline{x})$:
  - INITIALIZE: Chose initial values, $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)}$
  - SAMPLE: For $m = 1, \ldots, M$

    For $i = 1, \ldots, p$
    - Sample trial value θ_{trial} from proposal distribution $h_i(\theta_i|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_i^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})$
    - Compute acceptance ratio:

$$R = \frac{g(\theta_{trial}|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_{i+1}^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x}) h_i(\theta_i^{(m-1)}|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_i^{(m-1)} \ldots, \theta_p^{(m-1)}, \underline{x})}{g(\theta_i^{(m-1)}|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_{i+1}^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x}) h_i(\theta_{trial}|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_i^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})}$$

    - Accept with probability $\min\{1, R\}$ and set $\theta_i^{(m)} = \theta_{trial}$
    - Reject with probability $1 - \min\{1, R\}$ and set $\theta_i^{(m)} = \theta_i^{(m-1)}$

If R>1, accept. If R<1, accept with probability R and reject with probability 1-R

- This sampling process is a Markov chain, because the distribution of $\theta_1^{(m)}, \ldots, \theta_p^{(m)}$ is independent of the past given $\theta_1^{(m-1)}, \ldots, \theta_p^{(m-1)}$
- Under fairly general conditions $g(\underline{\theta}|\underline{x})$ is the unique stationary distribution

Interpreting the MH Acceptance Probability

- The probability of accepting the proposed new state is given by:

$$A_i\left(\underline{\theta}_{trial} \middle| \underline{\theta}^{(m)}\right) = \min\{1, R\} = \min\left\{1, \frac{g(\theta_{trial}|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_{i+1}^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})}{g(\theta_i^{(m)}|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_{i+1}^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})} \frac{h_i(\theta_i^{(m)}|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_i^{(m-1)} \ldots, \theta_p^{(m-1)}, \underline{x})}{h_i(\theta_{trial}|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_i^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})}\right\}$$

- The acceptance ratio R consists of two factors:

$$\frac{g(\theta_{trial}|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_{i+1}^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})}{g(\theta_i^{(m-1)}|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_{i+1}^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})}$$

Probability density under target distribution of new state / Probability density under target distribution of old state

$$\frac{h_i(\theta_i^{(m-1)}|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_i^{(m-1)} \ldots, \theta_p^{(m-1)}, \underline{x})}{h_i(\theta_{trial}|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_i^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x})}$$

Probability density of proposing to come back to old state from new state / Probability density of proposing to go to new state from old state

- Acceptance probability is larger when:
  - Posterior probability of proposed new state is higher
  - Transition back to old state is more likely

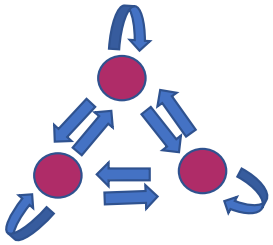
Gibbs and Component-Wise MH samplers

- Target distribution is:
  - $g(\theta_1, \theta_2, \dots, \theta_p | x_1, x_2, \dots, x_n)$
- On each iteration, sample from full conditional distributions for each variable in turn:
  - $g(\theta_i | \theta_1^{(m)}, \dots, \theta_{i-1}^{(m)}, \theta_{i+1}^{(m-1)}, \dots, \theta_p^{(m-1)}, \underline{x})$
- Continue for many iterations
- This Markov chain has the target distribution as its stationary distribution (given regularity conditions)

- Target distribution is:
  - $g(\theta_1, \theta_2, \dots, \theta_p | x_1, x_2, \dots, x_n)$
- On each iteration, sample from proposal distributions for each variable in turn:
  - $h_i(\theta_i | \theta_1^{(m)}, \dots, \theta_{i-1}^{(m)}, \theta_i^{(m-1)}, \dots, \theta_p^{(m-1)}, \underline{x})$
  - Calculate acceptance ratio R
  - Accept sampled value with probability $\min\{1, R\}$; otherwise reject
- Continue for many iterations
- This Markov chain has the target distribution as its stationary distribution (given regularity conditions)

Detailed Balance

- The probability of transitioning from state $\underline{\theta}$ to state $\underline{\theta}' \neq \underline{\theta}$ is $T(\underline{\theta}'|\underline{\theta}) = h(\underline{\theta}'|\underline{\theta})A(\underline{\theta}'|\underline{\theta})$
  - $h(\underline{\theta}'|\underline{\theta})$ is the probability of proposing $\underline{\theta}'$ when currently in state $\underline{\theta}$
  - $A(\underline{\theta}'|\underline{\theta})$ is the probability of accepting $\underline{\theta}'$ if it is proposed from state $\underline{\theta}$



- The probability of staying at state $\underline{\theta}$ (not changing state) is $T(\underline{\theta}|\underline{\theta}) = h(\underline{\theta}|\underline{\theta}) + \sum_{\underline{\theta}' \neq \underline{\theta}} h(\underline{\theta}'|\underline{\theta})\left(1 - A(\underline{\theta}'|\underline{\theta})\right)$

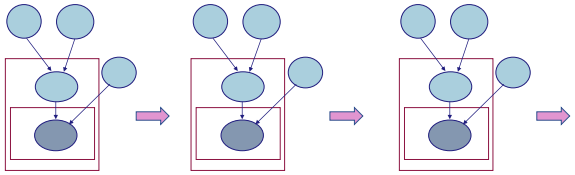
(This is the formula for discrete distributions; a similar analysis works for continuous distributions

- Transition probability $T(\underline{\theta}'|\underline{\theta})$ satisfies *detailed balance*:
  - $T(\underline{\theta}'|\underline{\theta})g(\underline{\theta}) = T(\underline{\theta}|\underline{\theta}')g(\underline{\theta}')$

At the equilibrium distribution $g(\cdot)$, transitions from $\underline{\theta}$ to $\underline{\theta}'$ balance transitions from $\underline{\theta}$ to $\underline{\theta}'$

- Detailed balance is a special case of the general balance condition for equilibrium in Markov chains (balance is satisfied for every pair of states)

Comments on MH Sampler and Gibbs Sampler



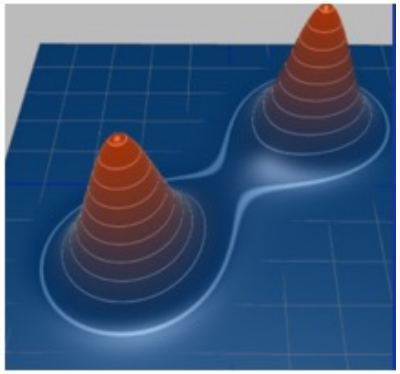
- The Metropolis-Hastings sampler has the same stationary distribution as the Gibbs sampler
- Metropolis-Hastings sampler is same as Gibbs sampler when we sample from the full conditional distribution:

  $h_i(\theta|\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \theta_i^{(m-1)}, \ldots, \theta_p^{(m-1)}, \underline{x}) = g(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p, \underline{x})$

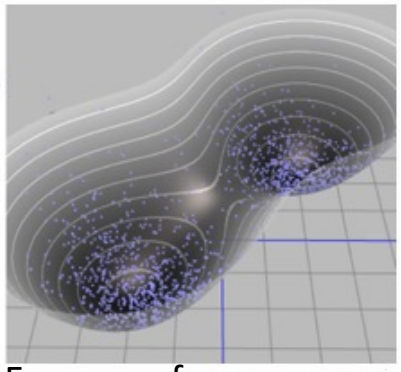
  In this case the acceptance probability is 1
- Under regularity conditions, the MH stationary distribution is unique and satisfies a central limit theorem
  - Proposal distributions $h_i(\cdot)$ and acceptance probabilities $A_i(\cdot)$ do not change as sampling progresses and any $(\theta_1, \ldots, \theta_i, \ldots, \theta_p)$ is reachable in finite expected time from any other state $(\theta_1, \ldots, \theta_i', \ldots, \theta_p)$
- Even when these conditions are satisfied, Gibbs and component-wise MH may suffer from the problem of getting stuck for long periods in local peaks of the posterior distribution
- We may want to thin the MH sample to avoid successive samples having identical values
  - Thinning with interval k means keeping only every k^{th} sample in the chain

The Physics-Information Analogy

- For any probabilistic system we can construct an equivalent physical system
  - The *Boltzmann distribution* (also known as the *Gibbs distribution*) for a given temperature is the probability the system will be in a given state at thermodynamic equilibrium at that temperature: $p_s \propto \exp\left\{-\frac{E_s}{kT}\right\}$
    - E_s = energy of the system
    - k = Boltzmann's constant
    - T = temperature of the system
  - We can exploit this relationship by setting $kT = 1$ and setting energy equal to the negative logarithm of the probability density we want to estimate
- In physics, Gibbs sampling and other MCMC algorithms are used to simulate physical systems at thermodynamic equilibrium, thereby estimating the Gibbs distribution
- Bayesian statistics uses the mapping between log probability and energy to apply MCMC to simulating from posterior distributions



pdf

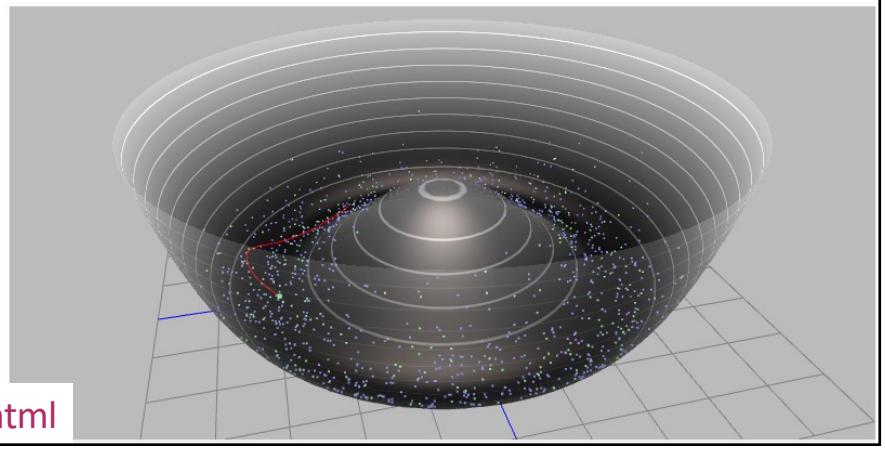


Energy surface

https://arogozhnikov.github.io/2016/12/19/markov_chain_monte_carlo.html

Hamiltonian Monte Carlo (HMC)

- The Metropolis-Hastings sampler can be very inefficient especially in high dimensions
  - Need to propose very small changes to get high enough acceptance probability but then successive states will be very highly correlated
- HMC exploits the physical analogy to achieve a more mobile and therefore more efficient sampler (only for continuous random variables)
  - Add “momentum” variables to state of the sampler
  - Follow (approximate) Hamiltonian dynamical equations to propose (approximately) energy conserving next state
  - Use Metropolis-Hastings step to correct for numerical approximation errors
- There are several variations that improve efficiency
  - NUTS (No U-Turn Sampler) adaptively optimizes HMC algorithm



https://arogozhnikov.github.io/2016/12/19/markov_chain_monte_carlo.html

Computing MCMC Estimates

- We constructed a Gibbs sampler ”by hand” for the reaction time data
- Building a MCMC sampler can be tedious and error-prone for models with many parameters related in complex ways
- Several software toolkits have been developed to support MCMC sampling
  - Functions for MCMC sampling for many commonly used statistical models
  - Plots and output diagnostics
  - High-level language for specifying model and automatic construction of sampler
- These toolkits are still evolving and a good deal of know-how is needed to use them effectively
  - MCMC computation is still an active area of research

Free Software for MCMC Computation

R packages: https://CRAN.R-project.org/package=MCMCpack - functions to perform posterior MCMC simulation for a number of commonly used statistical models

https://CRAN.R-project.org/package=mcmc - called by MCMCpack; simulates continuous distributions using several MCMC methods

https://cran.r-project.org/package=coda - output analysis and diagnostics for MCMC

MATLAB tools: https://mjlaine.github.io/mcmcstat/ - functions for MCMC simulation in MATLAB

Python tools: https://pypi.python.org/pypi/pymc - functions for MCMC simulation and output analysis in Python

Bayesian Inference Using Gibbs Sampling (BUGS) and its descendants provide high-level language for specifying model and methods for constructing and running sampler from high-level specification

WinBUGS http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/

OpenBUGS http://www.openbugs.net/

JAGS http://mcmc-jags.sourceforge.net/

Stan http://mc-stan.org Uses NUTS HMC algorithm

BUGS and Its Descendants

- BUGS is:
  - A high-level language for defining Bayesian models
  - A library of sampling routines
  - An interface for running the sampler
  - An output processor for processing and interpreting results
- BUGS is intended to free the modeler to focus on the problem without worrying about details of inference implementation
- Incarnations of BUGS:
  - Classic BUGS developed 1995, cross-platform, not maintained
  - WinBUGS Windows-only GUI, creates coda files for input to R, last version was 2007
  - OpenBUGS Open-source version of BUGS for Windows and Linux, interfaces with R
  - JAGS cross-platform, interfaces to R with rjags and R2jags and Python with pyJAGS
- Stan is a newer statistical modeling platform that is under active development
  - User specifies model in probabilistic programming language
  - Stan performs inference
  - Stan interfaces with R, Python, MATLAB and others
- We provide examples of using JAGS from within R to do MCMC inference

Quick Guide to Installing and Running JAGS

- JAGS runs on Linux, MacOS, and Windows and interfaces with R through the rjags and R2jags packages.
- To install JAGS and set it up to be used from R:
  - If necessary Download and install R and potentially a user interface to R like R Studio (see here for tips on getting started withR).
  - Download and install JAGS as per operating system requirements.
  - Install additional R packages: e.g.,
    - rjags to interface with JAGS
    - R2jags to call JAGS from R (depends on rjags)
    - coda to process MCMC output
    - superdiag for MCMC convergence diagnostics

Source: http://www.r-bloggers.com/getting-started-with-jags-rjags-and-bayesian-modelling/

JAGS Example: Reaction Times (1 of 2)

1. Specify model in BUGS language and save as .jags file

2. Run the model from R (working directory should be set to location of .jags file)

```
model {
  for(i in 1:n) {
    reaction.times[i]~dnorm(theta,rho)  # mean theta precision rho
  }
  theta~dnorm(0,0.0001)         # mean 0, very small precision
  rho~dgamma(0.01,0.0001)       # very small shape, very small rate (large scale)
}
```

```
# The data
reaction.times=c(5.743, 5.606, 5.858, 5.656, 5.591, 5.793, 5.697, 5.875, 5.677, 5.73,
                 5.69, 5.919, 5.981, 5.996, 5.635, 5.799, 5.537, 5.642, 5.858, 5.793,
                 5.805, 5.73, 5.677, 5.553, 5.829, 5.489, 5.724, 5.793, 5.684, 5.606)
n <- length(reaction.times)

rt.data  <- list("reaction.times", "n")

rt.params <- c("theta", "rho")

rt.inits <- function(){
  list("theta"=c(mean(reaction.times)), "rho"=c(1/var(reaction.times)))
}

# Fit the JAGS model
# Make sure working directory is set to location of file "reaction.time.model.jags"
reaction.times.fit <- jags(data=rt.data, inits=rt.inits,
 rt.params, n.chains=4, n.iter=9000, n.burnin=1000,
 model.file="reaction.time.model.jags")
```

JAGS Example: Reaction Times (2 of 2)

3. Analyze output using coda package

```
reaction.times.fit.mcmc<-as.mcmc(reaction.times.fit) # convert to R MCMC object
summary(reaction.times.fit.mcmc)     # Summary table of MCMC outputs
plot(reaction.times.fit.mcmc)        # Plots of MCMC results
```

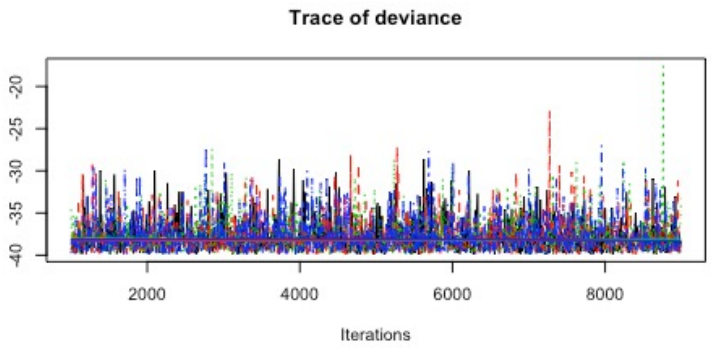
4. Review summary table

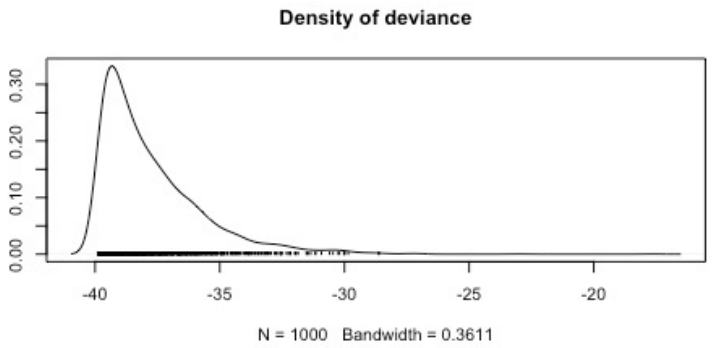
```
Iterations = 1001:8993
Thinning interval = 8
Number of chains = 4
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

            Mean       SD  Naive SE Time-series SE
deviance -37.743  2.06077 0.0325837      0.0325828
rho       62.449 16.66568 0.2635075      0.2634465
theta      5.732  0.02449 0.0003872      0.0003873

2. Quantiles for each variable:

            2.5%     25%     50%     75%  97.5%
deviance -39.805 -39.256 -38.354 -36.858 -32.23
rho       34.381  50.582  60.673  72.877  99.16
theta      5.685   5.716   5.732   5.748   5.78
```

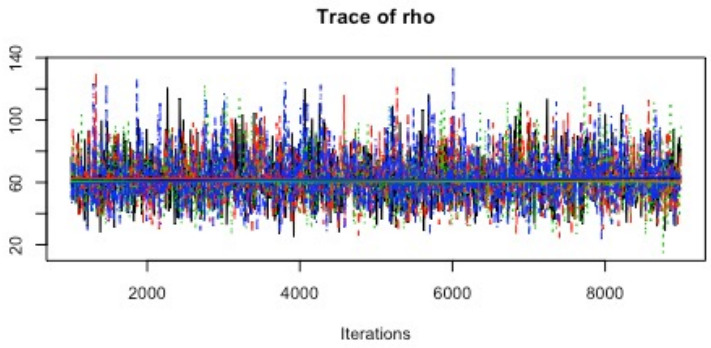
It is common to "thin" the chain by keeping only every k^{th} observation. This reduces the autocorrelation.

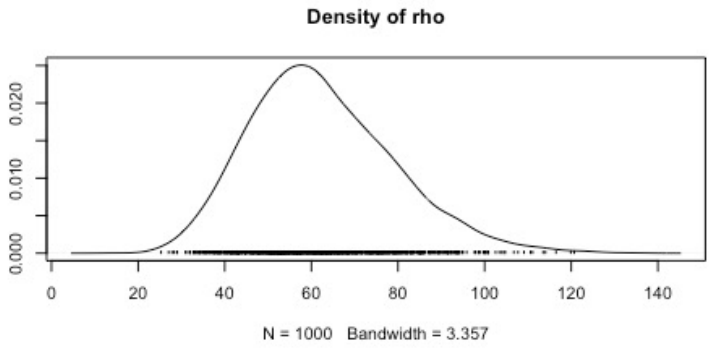
Deviance = -2log $(x|\theta,\rho)$ is a measure of how well the observations fit the model

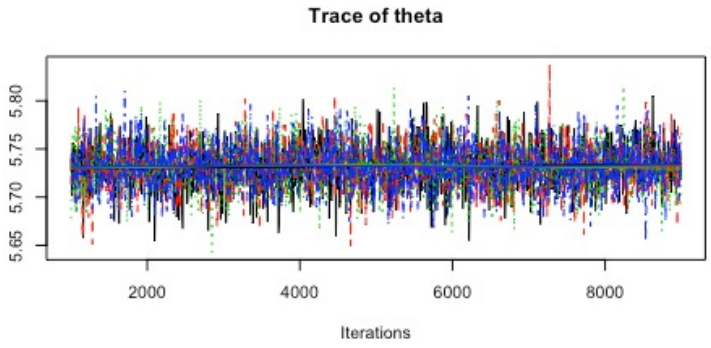
5. Review plots and diagnostics

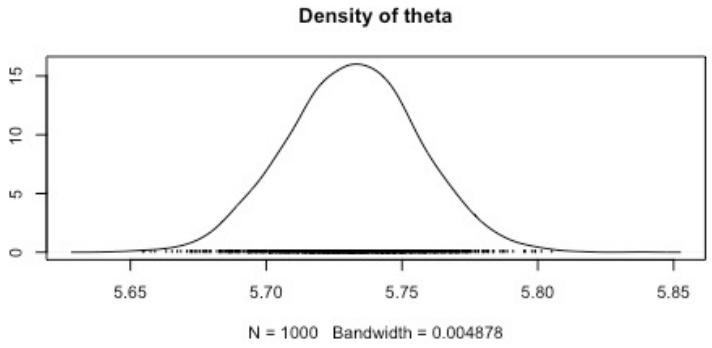












©Kathryn Blackmond Laskey

Summary and Synthesis

- MCMC estimates a target distribution $P(X)$ by constructing a Markov chain with $P(X)$ as a stationary distribution
- The two most popular MCMC methods are Gibbs and Metropolis-Hastings sampling
  - Gibbs sampling can be used when it is possible to sample from the "full conditional" distributions of each target variable given all the others
  - Metropolis-Hastings generalizes Gibbs sampling to problems where ratios of "full conditional" distributions are available.
- MCMC methods yields correlated draws
  - Diagnostic tools can help assess the severity of autocorrelation and evaluate whether enough samples have been collected
  - Although these diagnostics are useful, they can be deceptive
  - For very hard problems, it may be infeasible to obtain an accurate estimate of the posterior distribution, and "good" local optima are the best that can be achieved
- MCMC algorithms have proven very useful for otherwise intractable estimation problems
- Software tools are available for constructing MCMC samplers
  - We examined JAGS, which uses Gibbs sampling when possible; otherwise MH sampling

References

- JAGS
  - http://www.johnmyleswhite.com/notebook/2010/08/20/using-jags-in-r-with-the-rjags-package/
  - http://www.r-bloggers.com/getting-started-with-jags-rjags-and-bayesian-modelling/
  - http://martynplummer.wordpress.com/
  - http://sousalobo.com/aom2011pdw/jags_tutorial_lobo.pdf
- MCMC Diagnostics
  - Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation using Multiple Sequences. Statistical Science, 7, 457–472.
  - Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bayesian Statistics 4, Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds.), 169-193. Oxford: Oxford University Press.
  - Heidelberger P and Welch PD. (1983). Simulation run length control in the presence of an initial transient. Opns Res., 31, 1109-44
  - Raftery, Adrian E.; Lewis, Steven M. (1992). [Practical Markov Chain Monte Carlo]: Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. Statist. Sci. 7, no. 4, 493--497.
- Hamiltonian Monte Carlo
  - Betancourt, M. (2018) A Conceptual Introduction to Hamiltonian Monte Carlo. https://arxiv.org/pdf/1701.02434.pdf