# Applied Bayesian Methods

Michael Escobar

University of Toronto

m.escobar@utoronto.ca

http://individual.utoronto.ca/Escobar/

# Lecture 2

Outline:

- Markov Chains
- Simple Markov Chain Monte Carlo Methods
- Exploring some difficulties

## Introduction

- In the first lecture, we looked at some simple Bayesian models.
- We found that even for these simple models, the mathematical formulas are rather complex. However, by sampling from different distributions, then it is possible to estimate various values and distributions. For example, it is possible to estimate the distribution of $1/\sqrt{\tau}$ when $\tau$ is from a gamma distribution.
- In this part of the course, a powerful computational technique for working with Bayesian models is introduced.
- This technique produces samples from the posterior distribution by sampling from a Markov chain.
- So, first we look at the the properties of a Markov chain.

## Part 1: Markov chains

- Definition of Markov chains
- Convergence of simple 2-state chains
- Infinite state chains
- Invariant distributions
- Requirements for convergence
- Path averages

## A Markov Chain

Consider a sequence of random variables $X_0, X_1, \ldots, X_{n-1}, X_n, \ldots$, which take on possible values from the set $\mathcal{S}$.

1. This sequence is a **Markov chain** if for all $n$ the conditional distribution of $X_n$ given $X_0, \ldots, X_{n-1}$ is equal to the conditional distribution of $X_n$ given $X_{n-1}$. That is, if for all $n$,

$$F(X_n|X_0, \ldots, X_{n-1}) = F(X_n|X_{n-1}).$$

2. The Markov chain has a **stationary transition probability** if for all $n$: $F(X_n|X_{n-1}) = F(X_1|X_0)$.

## A Markov Chain

Comments:

- The first property, which defines the sequence as a Markov chain, is called the Markovian property. That is, that when $X_{n-1}$ is know, then when sampling $X_n$ it does not matter what happened before $X_{n-1}$.
- For our purposes the existence of a stationary transition probability is the more important property.
- For a Markov chain with a stationary transition probability, after many iterations of the chain, the marginal distribution of $X_i$ converges to a limiting distribution.

## Example 6: Moving in and out of California†.

Suppose, each year:

- $1/10$ of people outside California move in.
- $2/10$ of people inside California move out.

So, if we start with $y_0 =$ %outside and $z_0 =$ %inside, then

$$\begin{pmatrix} y_1 \\ z_1 \end{pmatrix} = \begin{bmatrix} .9 & .2 \\ .1 & .8 \end{bmatrix} \begin{pmatrix} y_0 \\ z_0 \end{pmatrix}$$

†This is example is from Strang, G. *Linear Algebra and It Applications*, (Academic Press, Inc.: New York, 1980), pg 199-201.

## Example 6: Moving in and out of California.

Comments:

- Please note that the above is an example of a Markov chain.
- As defined (although not in real life), the probability that someone will be in California next year only depends on whether the person is in California this year.
  - If the person is in California this year, then the chance that that person will be in California next year is 8/10.
  - If the person is outside of California, then the chance that that person will be in California next year is 1/10.
  - Note that these probabilities don't depend on where the person was in before the last year.
  - Also, this chain has stationary transition probabilities since these probabilities don't change for different years.

## Transition Matrix

- Let $A$ be the matrix $\begin{bmatrix} .9 & .2 \\ .1 & .8 \end{bmatrix}$.
- The matrix $A$ is called the transition matrix.
- The columns of $A$ sum to one, since they are conditional probabilities. For example, the first column represents the probability of being inside or outside on the next step if one starts outside.

Diagonalize $A$:

$$\begin{aligned} A &= S\Lambda S^{-1} \\ &= \begin{bmatrix} 2/3 & 1/3 \\ 1/3 & -1/3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & .7 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix} \end{aligned}$$

- The columns of $S$ are the eigenvectors.
- The diagonal values of $\Lambda$ are the eigenvalues.
- The largest eigenvalue is equal to one.

So, after $k$ years:

$$\begin{aligned}
\begin{pmatrix} y_k \\ z_k \end{pmatrix} &= A^k \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \\
&= S\Lambda^k S^{-1} \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \\
&= S \begin{bmatrix} 1^k & 0 \\ 0 & (.7)^k \end{bmatrix} S^{-1} \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \\
&= (y_0 + z_0) \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix} + (y_0 - 2z_0)(.7)^k \begin{pmatrix} 1/3 \\ -1/3 \end{pmatrix} \\
&\to \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix} \quad \text{as } k \to \infty
\end{aligned}$$

Note:

- After many years, the probability converges to a particular distribution no matter what the starting value is.
- The first eigenvector is the limiting distribution[†].
- The second largest eigenvalue controls the rate of convergent[†].
- The convergent results also describe the probable location of a single person after many iterations of the Markov chain.

[†] provided that there is only one eigenvalue with absolute value one.

## Example 6: Comments

- This example demonstrates a simple 2-state Markov chain with a stationary probability distribution.
- Here the probabilities for next year depend only on the state this year. If the state is know this year, then probabilities for next year don't depend on the previous years.
- Also, we see how, after several years, the probabilities in $k$ years becomes less and less dependent on the initial state and it converges to a limiting distribution irregardless of the initial value.
- Next, these results are generalized. First, for any 2-state Markov chain and then for Markov chains with a general state space.

## The General 2-State Markov Chain

Suppose we have two states which we simple call state 1 and 2.

- Let $\alpha$ be the probability of going to state 2 on the next step if you are presently in state 1
- Let $\beta$ be the probability of going to state 1 on the next step if you are presently in state 2.
- So, the transition matrix $A$ is:

$$A = \begin{bmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{bmatrix}$$

## General $S\Lambda S^{-1}$ Form

Also, it can be shown that $A = S\Lambda S^{-1}$ can be represented by:

$$
\begin{aligned}
A &= S\Lambda S^{-1} \\
&= \begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\alpha}{\alpha+\beta} & \frac{-\alpha}{\alpha+\beta} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1-\alpha-\beta \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & \frac{-\beta}{\alpha} \end{bmatrix},
\end{aligned}
$$

## General $S\Lambda S^{-1}$ Form

Comments:

- For the matrix $A^k$, we simply replace $1 - \alpha - \beta$ with $(1 - \alpha - \beta)^k$ in the above matrix formula.
- The first column of $\mathcal{S}$ is again the first eigenvector and is represents the limiting distribution which is $(\beta/(\alpha + \beta), \alpha/(\alpha + \beta)'$.
- Also, the rate of convergence is controlled by the second eigenvalue which is $1 - \alpha - \beta$. Therefore, this will converge as long as $\alpha$ and $\beta$ are not both either 0 or 1.

## General $S\Lambda S^{-1}$ Form

- Also, we can represent the distribution at the k-th step by:

$$
\begin{aligned}
\begin{pmatrix} y_k \\ z_k \end{pmatrix} &= A^k \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} = S\Lambda^k S^{-1} \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \\
&= (y_0 + z_0) \begin{pmatrix} \beta/(\alpha+\beta) \\ \alpha/(\alpha+\beta) \end{pmatrix} \\
&\quad +(y_0 - z_0\beta/\alpha)(1-\alpha-\beta)^k \begin{pmatrix} \alpha/(\alpha+\beta) \\ -\alpha/(\alpha+\beta) \end{pmatrix} \\
&\to \begin{pmatrix} \beta/(\alpha+\beta) \\ \alpha/(\alpha+\beta) \end{pmatrix} \quad \text{as } k \to \infty
\end{aligned}
$$

## Invariant Distribution

- Invariant distribution: if you start with a sample from an invariant distribution, you still have a sample from that distribution after one iteration of the Markov chain.
- That is, let $u_*$ be a column vector of probabilities and let it be an invariant distribution of the Markov chain defined by A, then
$$Au_* = u_*.$$
- If a chain converges to a unique limiting distribution no matter what the starting value and the chain has an invariant distribution, then the limiting distribution and the invariant distribution are the same.

Note that for the values in Example 6:

$$
\begin{aligned}
\begin{bmatrix} .9 & .2 \\ .1 & .8 \end{bmatrix} \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix} &= \\
\begin{bmatrix} 2/3 & 1/3 \\ 1/3 & -1/3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & .7 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix} \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix} &= \\
\begin{bmatrix} 2/3 & 1/3 \\ 1/3 & -1/3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & .7 \end{bmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} &= \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix}
\end{aligned}
$$

The limiting distribution is an invariant distribution.

## Infinite State Markov Chain

- We will need to define a Markov Chain on the space of the parameters of the Bayesian model.
- In the simple Markov Chain model, we initially defined the Markov Chain by defining the probability of where we would be in the next step conditioning on where we were in the current step.
- Before, there were only two locations, inside or outside of California. In the Markov chain we will define for the Bayesian model, the possible locations will be defined by the possible values of the parameters of the Bayesian model.
- For example, for the simple normal model with unknown mean and precision, the parameter space is $(\mu, \tau)$ which is on the space $\Re \times \Re^+$. So, $\mathcal{S} = \Re \times \Re^+$.

## Infinite State Markov Chain

- To define the Markov chain, we just need to specify the probability of moving around the parameter space $\mathcal{S}$.
  - That is, we just need to define the probability of where we will be in the next step based on where we are in the current step.
  - For example for the simple normal model, we have $\underline{\theta} = (\mu, \tau)$.
  - So, we can define the Markov chain by defining the conditional probabilities: $p(\underline{\theta}^{(m+1)}|\underline{\theta}^{(m)})$.
  - These conditional probabilities take the place of the columns of the matrix $A$.

## Conditions for Convergence

- For finite Markov Chain, the limiting distribution exist if it is:
  - Aperiodic: the chain is not cyclic.
  - Irreducible: it is possible to get from any one state to another.
- For infinite state Markov Chains, sufficient conditions for a limiting distribution are:
  - Aperiodic
  - Irreducible
  - There exist an invariant distribution
- The invariant distribution is the limiting distribution.
- When the Markov Chains converges to a limiting distribution, this happens irregardless of the starting point.

Bad conditions:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

- Not irreducible.
- If you start in one state, you always stay in that state.

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

- Periodic: always switching states.

Potentially bad conditions:

$$A = \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix}$$

- Slow to switch groups.
- Second eigenvalue is $1 - 2\epsilon$.

## Conditions for Convergence

Comments:

- For finite state Markov chains, there are very few conditions necessary in order to insure convergence.
- For infinite state Markov chains, the existence of an invariant distribution in addition to the conditions necessary for a finite state Markov chain are sufficient to insure that the limiting distribution exists.

## Path Averages

- The purpose of looking at Markov chains is to get a method for sampling from the posterior distribution. This will enable us to know the properties of the posterior distribution.
- The usefulness would be limited if we needed to run through the chain many times and in the end we only had one sample which is approximately from the posterior distribution.
- However, instead of using only the last value from the Markov chain as an approximation of one sample from the posterior distribution, it turns out that we can use all the samples from the Markov chain. That is, the average of all the samples from the Markov chain converges to the mean of the limiting distribution. The average of all the samples in the chain is called the path average.

## Path Averages: Definition

- Let $X^{(m)}$ be the sampled value at the m-th step.
- Let $E_\infty(X)$ be average of $X$ under the limiting distribution (and we assume that $E_\infty(X)$ exists and is finite), then

$$\frac{1}{M}\sum_{m=1}^{M} X^{(m)} \xrightarrow{a.s.} E_\infty(X)$$

## Path Averages: Comments

- This technique will allow us to get more than simply the posterior mean. Most of what one wants is an expected value of some sort. One can use this technique to estimate most of the features of the limiting distribution including estimates of the density function and any moments of the posterior distribution.
- In most cases, one can simply use sampled points from the chain like one would use an independent sample of the posterior distribution.
- What will be less straight forward will be getting an estimate of the standard error of our estimates. This is discussed in more detail later in this course.

## Part 1: Markov chains

Summary:

- Under very weak conditions, the samples from a Markov chain converge to a limiting distribution regardless of the starting conditions of the chain.
- The path average of the Markov chain converges to the expected value of the limiting distribution.

## Part 2: Markov Chain Monte Carlo Methods

Outline:

- Simple example. Revisiting the normal model.
- The Gibbs Sampler: a basic MCMC algorithm.
- Example: Anova and conditional independence.
- Some other MCMC algorithms.

## MCMC: Introduction

- Before 1990, calculating the posterior distribution for the parameters in a Bayesian model was difficult for all but the simplest models.
- To compute the posteriors for models which were more complex than say the 1-way ANOVA models required some one with strong training in mathematical statistics and further specialized training in Bayesian methods. Even with this training, it was difficult to perform the need calculations.
- This contrasted with the frequentist techniques. Data from an experiment which lead to complex Bayesian calculations could be analyzed with frequentist methods by anyone with computer software and a brief introductory course in statistics which used high school level mathematics.

## MCMC: Introduction

(Continued)

- In 1990, there was the seminal paper by Gelfand and Smith which showed how to use Markov chain Monte Carlo methods in Bayesian computations.
  - Get a sample of the joint posterior distribution by sampling from a Markov chain.
  - Given a sample of the joint posterior, one can calculate desired values of the posterior.
  - This was a fundamental break through in Bayesian computations and sparked a revolution in applied Bayesian statistics.

## Example 7: Heights Revisited

- In a previous lecture, a small dataset on people's heights is looked at. Since conjugate priors are used, then the exact posterior distribution could be calculated.
- To illustrate the Markov chain Monte Carlo methods, that example is revisited here.
- As before, in this example we are interested in $\mu$ the population mean (a measure of location) and $\tau$ the precision (a measure of the spread).

## Example 7: Heights Revisited

(continued)

- For the normal model, methods are presented for the situations:

  1. $\mu$ unknown, $\tau$ known
  2. $\mu$ known, $\tau$ unknown
  3. $\mu$ unknown, $\tau$ unknown

- Since we know neither $\mu$ nor $\tau$, we used the 3rd method to get the posterior distribution in the last lecture.

- Now suppose we did not know how to do the case where neither $\mu$ nor $\tau$ are known but we did know the first two methods. That is, we only know a method when only one of $\mu$ or $\tau$ are known. Let's estimate the posterior distribution when we only know these two methods.

## Example 7: Model, Priors, and Data

- We have the following model for heights:

$$\begin{aligned} X_i|\mu,\tau &\sim N(X_i|\mu,\tau) \\ \mu|\mu_0,\tau,\theta &\sim N(\mu|\mu_0,\tau\theta) \\ \tau|\alpha,\beta &\sim \text{Gam}(\tau|\alpha,\beta) \end{aligned}$$

- Priors: $\mu_0 = 66$, $\theta = 4$, $\alpha = 1$, $\beta = 25$.
- Data: $X = (64, 73, 64, 63, 69, 71)$, $n = 6$, $\bar{X} = 67.333$, and $\sum(X_i - \bar{X})^2 = 89.33$.

## Example 7: Assume that $\tau$ is known

Suppose we know the value of $\tau$. So for the equations below, $\tau$ has a know value.

- Let the notation for the conditional distribution of $\mu$ given the other parameter be: $[\mu|\tau, \mu_0, \theta, \alpha, \beta, \bar{X}, n]$.
- From lecture 1, we know that this posterior distribution is a normal distribution with mean $\mu_0'$ and precision $\tau_0'$ where $\tau_0'$ equals $\tau(\theta + n)$ which equals $10\tau$ in this case, and where

$$\begin{aligned} \mu_0' &= \frac{\tau\mu_0 + n\tau\bar{X}}{\tau\theta + n\tau} = \frac{\mu_0 + n\bar{X}}{\theta + n} \\ &= \frac{(4)(66) + (6)(67.333)}{4 + 6} = 66.80. \end{aligned}$$

## Example 7: Assume that $\mu$ is known

Here, we assume that $\mu$ is assumed to be known and we are looking for the posterior distribution of $\tau$.

- Let the notation for the conditional distribution of $\tau$ given the other parameter be: $[\tau|\mu, \mu_0, \theta, \alpha, \beta, \bar{X}, n]$.
- In looking at the model, there are two contributions to the likelihood of $\tau$, they are the following relationships:

$$\begin{aligned} X_i|\mu, \tau &\sim N(X_i|\mu, \tau) \\ \mu|\mu_0, \tau, \theta &\sim N(\mu|\mu_0, \tau\theta) \end{aligned}$$

- The second relationship is equivalent to:

$$\sqrt{\theta}(\mu - \mu_0) \sim N(\sqrt{\theta}(\mu - \mu_0)|0, \tau).$$

## Example 7: Assume that $\mu$ is known

(continued)

- So, the posterior distribution for $\tau$ is a gamma distribution with parameters $\alpha''$ and $\beta''$ where

$$\begin{aligned} \alpha'' &= \alpha + (n+1)/2 \\ \beta'' &= \beta + \frac{1}{2}\left(\sum_{i=1}^{n}(X_i - \mu)^2 + \theta(\mu - \mu_0)^2\right) \end{aligned}$$

- Note: here $\mu$ is known and is simply ”treated” like data.

## Example 7: How to Sample

- Now the task is to simulate a sample from the posterior. If we can do this, then we can do what we did in the first lecture. The sampled values can be used to approximate desired properties and features of the posterior distribution.

- Since we know $[\mu|\tau, \underline{X}, \theta, \mu_0, \alpha, \beta]$ and $[\tau|\mu, \underline{X}, \theta, \mu_0, \alpha, \beta]$, then if we knew $\tau$ we could sample $\mu$ and if we knew $\mu$ we could sample $\tau$.

## Example 7: How to Sample

(continued)

- A solution is to first start with some values of $\mu$ and $\tau$. Call them $(\mu^{(0)}, \tau^{(0)})$
- Then, sample a new $\mu$, call it $\mu^{(1)}$, from $[\mu|\tau = \tau^{(0)}, \underline{X}, \theta, \mu_0, \alpha, \beta]$. Then sample a new $\tau$, call it $\tau^{(1)}$, from $[\tau|\mu = \mu^{(1)}, \underline{X}, \theta, \mu_0, \alpha, \beta]$.
- Continue sampling $(\mu^{(m)}, \tau^{(m)})$ given the previously sampled values $(\mu^{(m-1)}, \tau^{(m-1)})$.

# Example 7: How to Sample

Comments:

- The sequence of values, $(\mu^{(0)}, \tau^{(0)}), \ldots, (\mu^{(m)}, \tau^{(m)}), \ldots$ is a Markov chain. Also, this chain is aperiodic and irreducible.
- It can be shown that by the construction of this chain, the posterior distribution is an invariant distribution for the chain.
- The posterior distribution is the limiting distribution for this chain regardless of the starting value of the chain.
- Using the path averages, the sampled points, $(\mu^{(0)}, \tau^{(0)}), \ldots, (\mu^{(m)}, \tau^{(m)}), \ldots$, can be used to estimate features of the posterior distribution.

## Example 7: Sampling

Here are the steps which are produced from one such iteration of this algorithm.

- First starting values are picked. They can theoretically be anything. For this run of the algorithm, I choose $(\mu^{(0)}, \tau^{(0)}) = (20, .0025)$.
- Now, sample $\mu^{(1)}$ with $\tau = .0025$.
  - So, $\mu^{(1)}$ is sampled from a normal distribution with mean $\mu_0'$ and precision $\tau_0'$.
  - From the previous calculations, we know that $\mu_0' = 66.80$ and $\tau_0'$ equals $10\tau = 0.025$.
  - Drawing the sample from a normal distribution with these parameter, we get $\mu^{(1)} = 61.417$.

## Example 7: Sampling

(continued)

- Now, a value $\tau^{(1)}$ is sampled “knowing” that $\mu^{(1)} = 61.417$.
  - Again, from the previous work, we know that $\tau^{(1)}$ has a gamma distribution with parameters $\alpha''$ and $\beta''$.
  - Also, $\alpha''$ is equal to $\alpha + (n+1)/2$, so $\alpha'' = 4.5$.
  - For the parameter $\beta''$, it is equal to:

$$\beta'' = \beta + \frac{1}{2}\left(\sum_{i=1}^{n}(X_i - \mu)^2 + \theta(\mu - \mu_0)^2\right),$$

  and when $\mu = \mu^{(1)}$ with is equal to 61.417, then $\beta''$ is equal to 232.655.

## Example 7: Sampling

(continued)

- This process is continued. Below is the table of sampled values for $(\mu^{(m)}, \tau^{(m)})$. Also, in the table is the values of $\tau_0'$ and $\beta''$ used to sample the random variables $\mu$ and $\tau$ respectively. Note that for this algorithm, the parameters $\mu_0'$ and $\alpha''$ always have the values 66.8 and 4.5, respectively, in this example.

| m | $\mu$ | $\tau$ | $\tau_0'$ | $\beta''$ |
|---|---|---|---|---|
| 0 | 20.000 | 0.0025 | | |
| 1 | 61.417 | 0.0198 | 0.025 | 232.65 |
| 2 | 67.464 | 0.0352 | 0.198 | 70.60 |
| 3 | 66.974 | 0.0186 | 0.352 | 70.11 |
| 4 | 68.093 | 0.0645 | 0.186 | 74.74 |
| 5 | 67.045 | 0.1322 | 0.645 | 70.04 |
| 6 | 67.094 | 0.0607 | 1.322 | 70.01 |
| 7 | 67.767 | 0.0398 | 0.607 | 72.10 |
| 8 | 68.360 | 0.0442 | 0.398 | 77.69 |
| 9 | 66.474 | 0.0583 | 0.442 | 72.10 |

## Example 7: Sampling

(continued)

- The following series of plots show the "path" of the sampled chain. That is, it plots each of the sampled values and connects consecutive points with a line.
  - The first series is for the joint parameters.
  - The next series of plots is the "trace" plot of each sampled value of the parameter plotted against the iteration number of the chain.

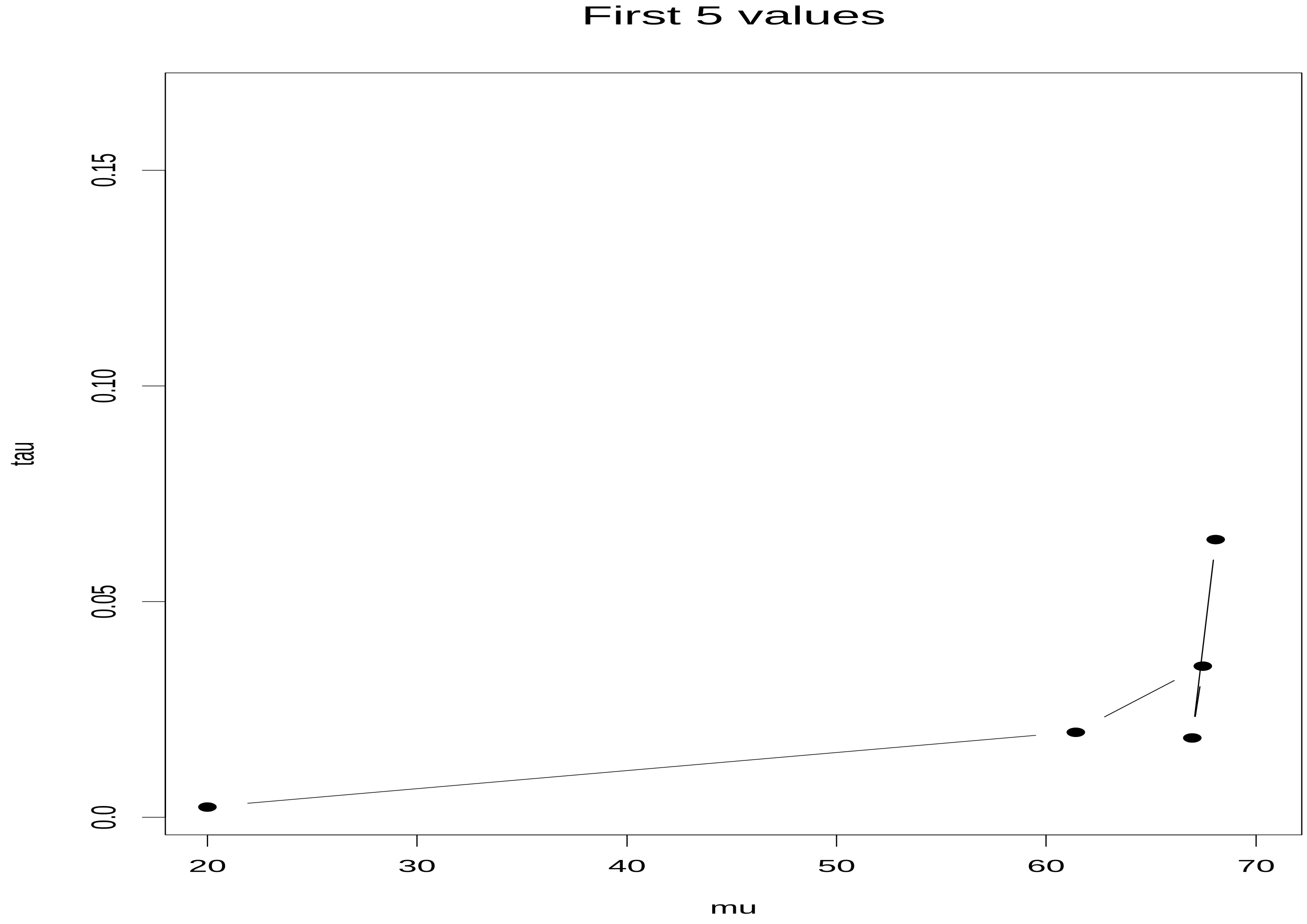

Figure 1: The sampled values from the Markov chain

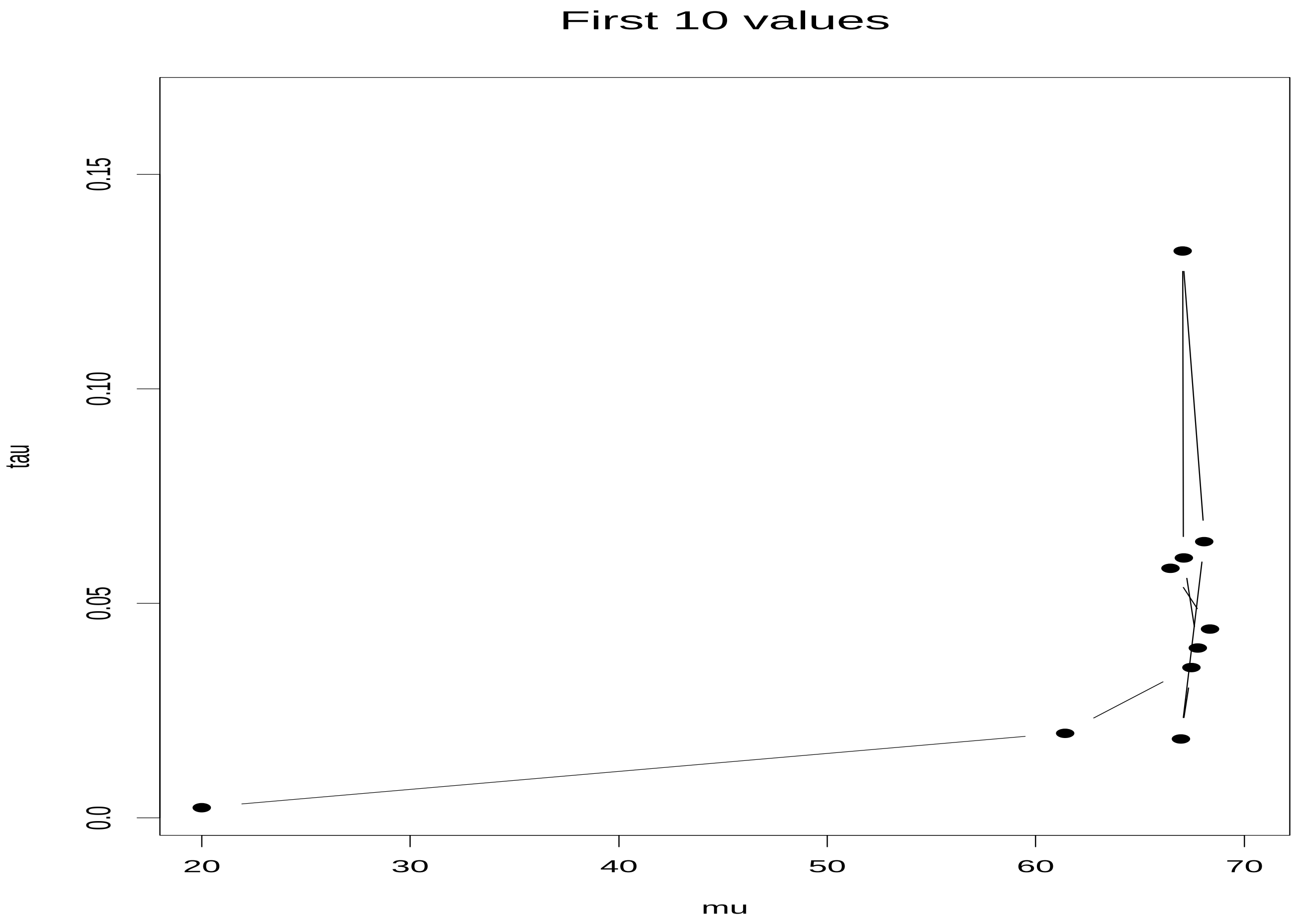

Figure 2: The sampled values from the Markov chain

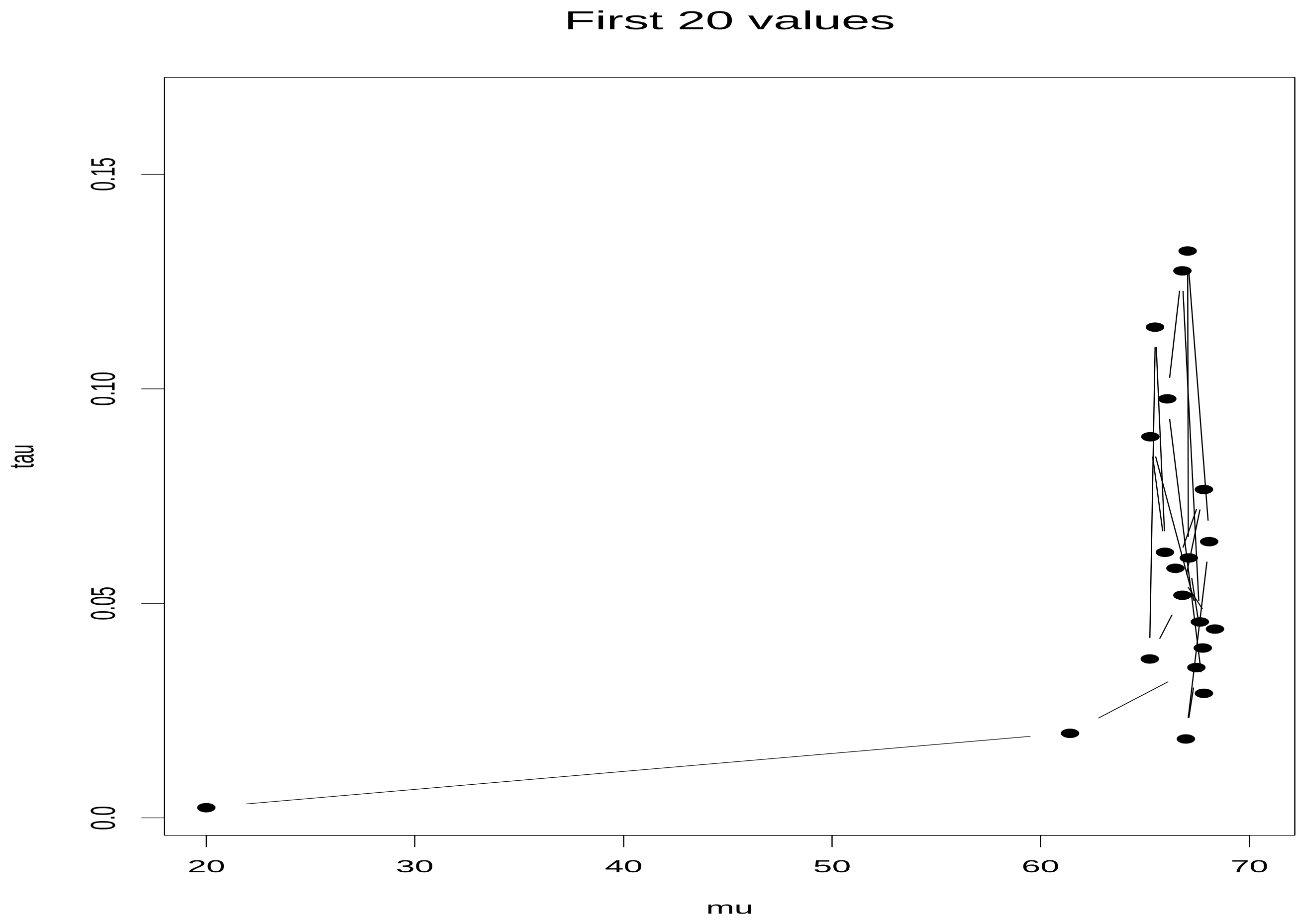

Figure 3: The sampled values from the Markov chain

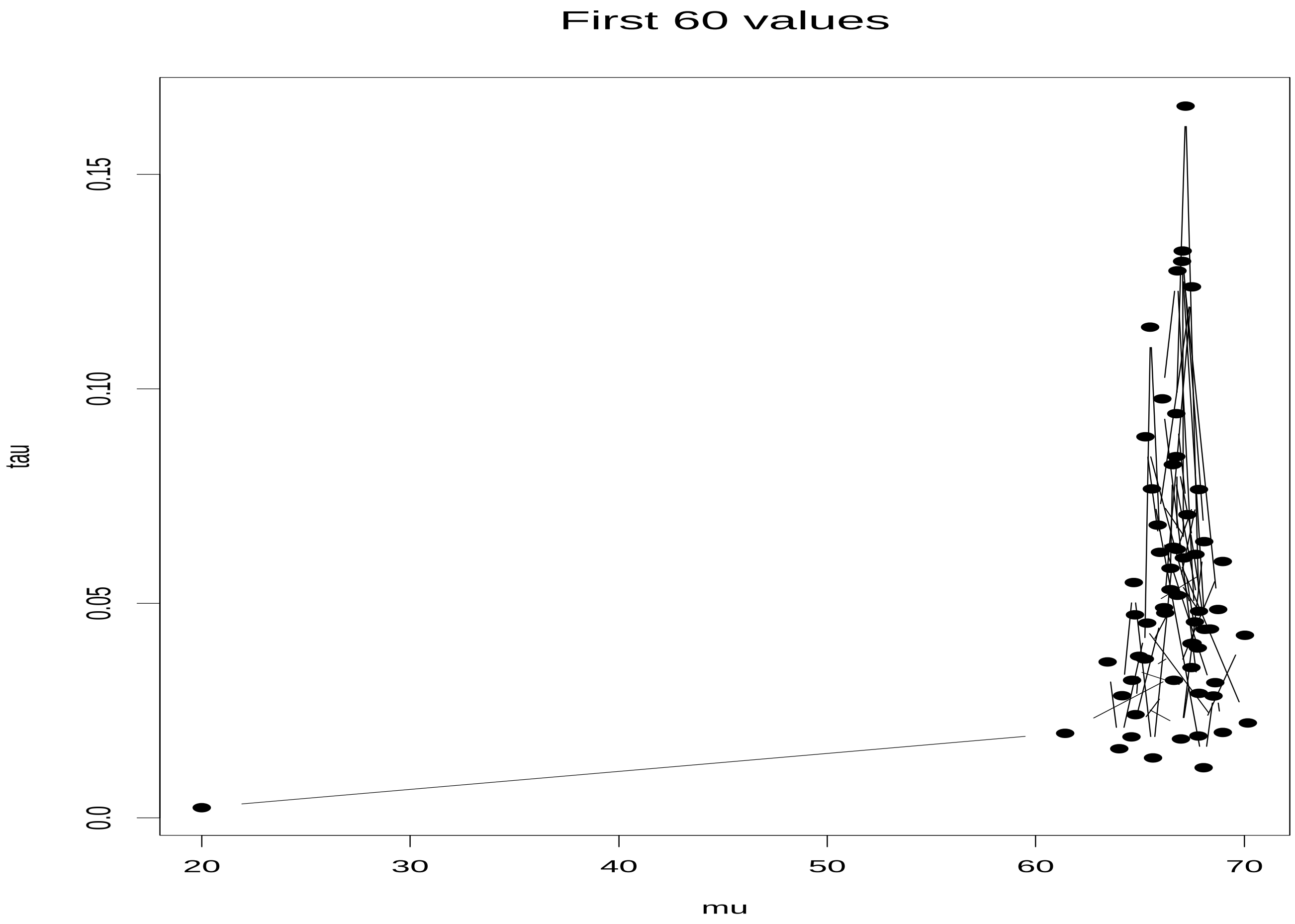

Figure 4: The sampled values from the Markov chain

## Path Plots

Some comments:

- Please note that the first point or two are somewhat far from the rest of the points. This is because it takes a while before the chain finds the area where the chain seems to settle down in.
- After running for a while, the samples are mostly located in this area.
- It is usually a good idea to "throw away" the first group of points until the chain settles down in the area of high posterior probability.
- So, in the next group of plots, the first point is discarded and we see how the chain behaves.

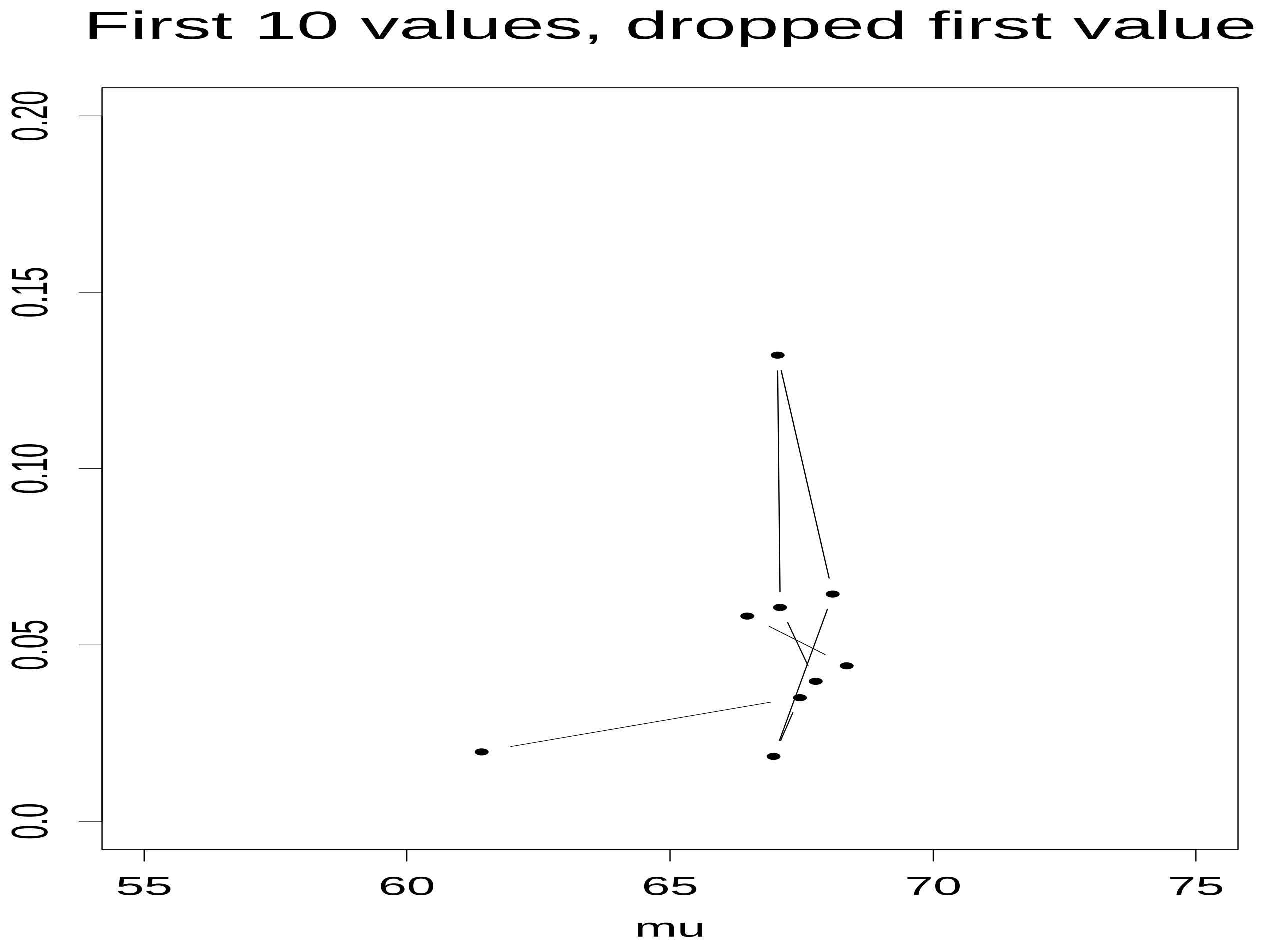

Figure 5: The sampled values from the Markov chain

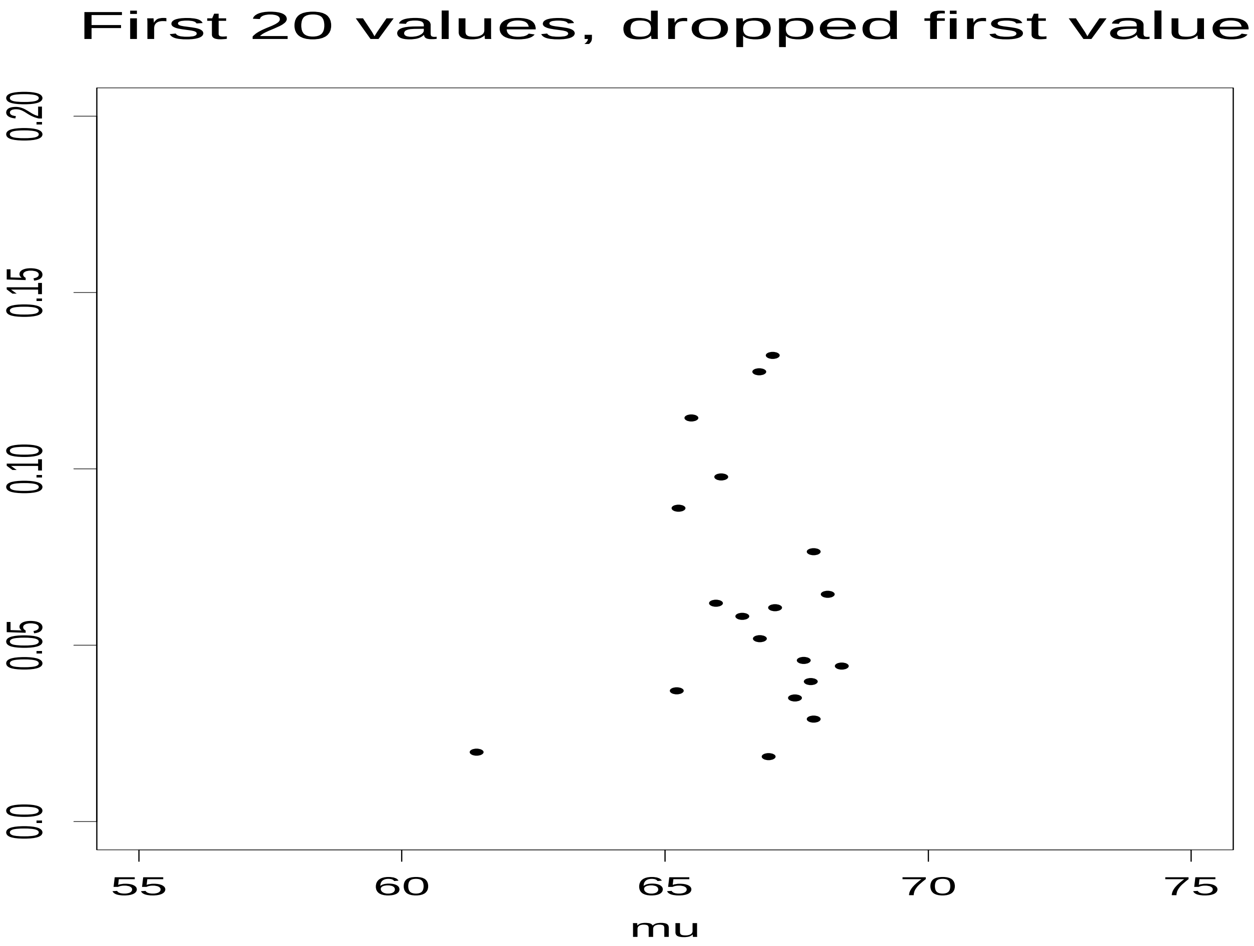

Figure 6: The sampled values from the Markov chain

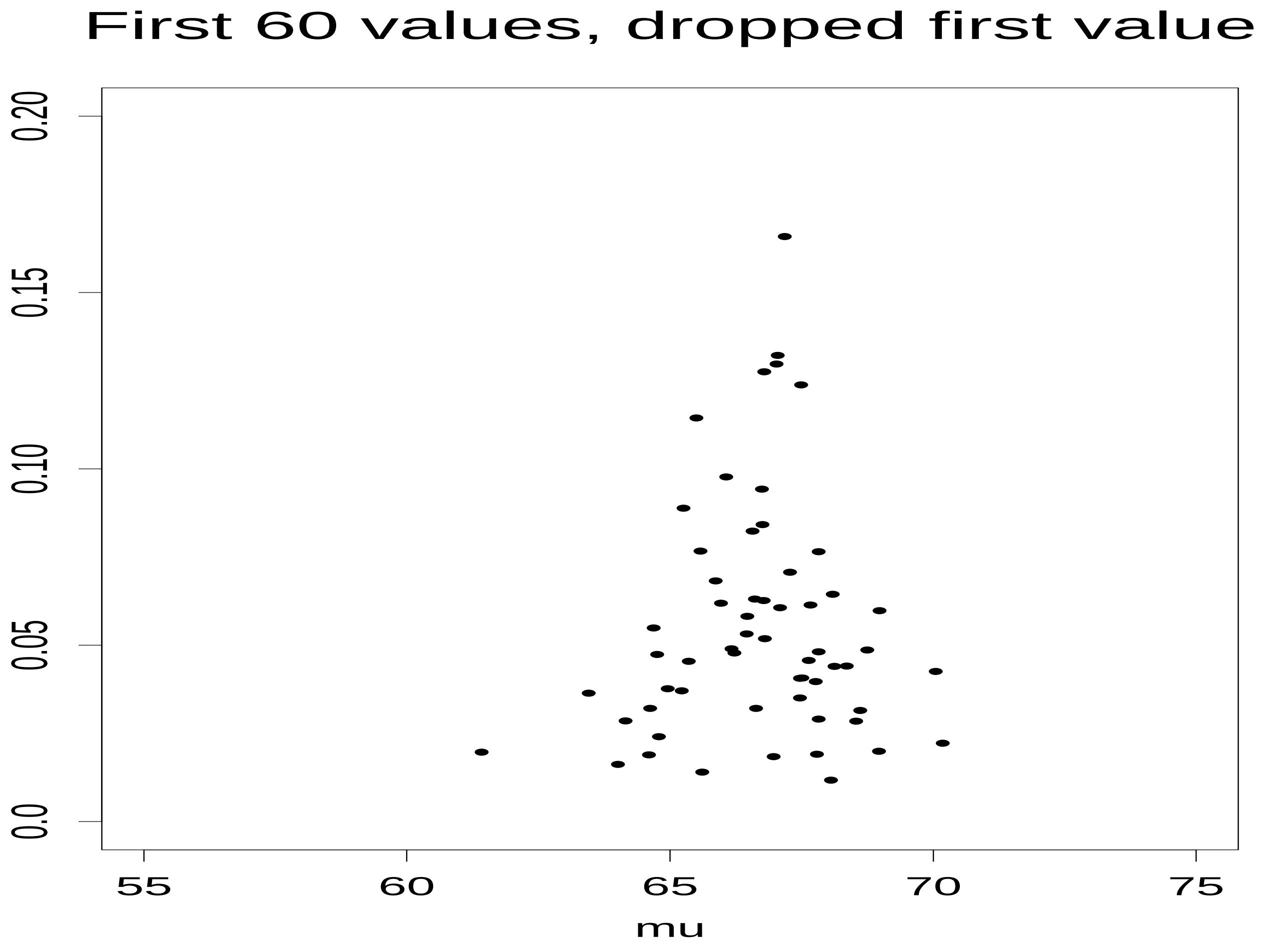

Figure 7: The sampled values from the Markov chain

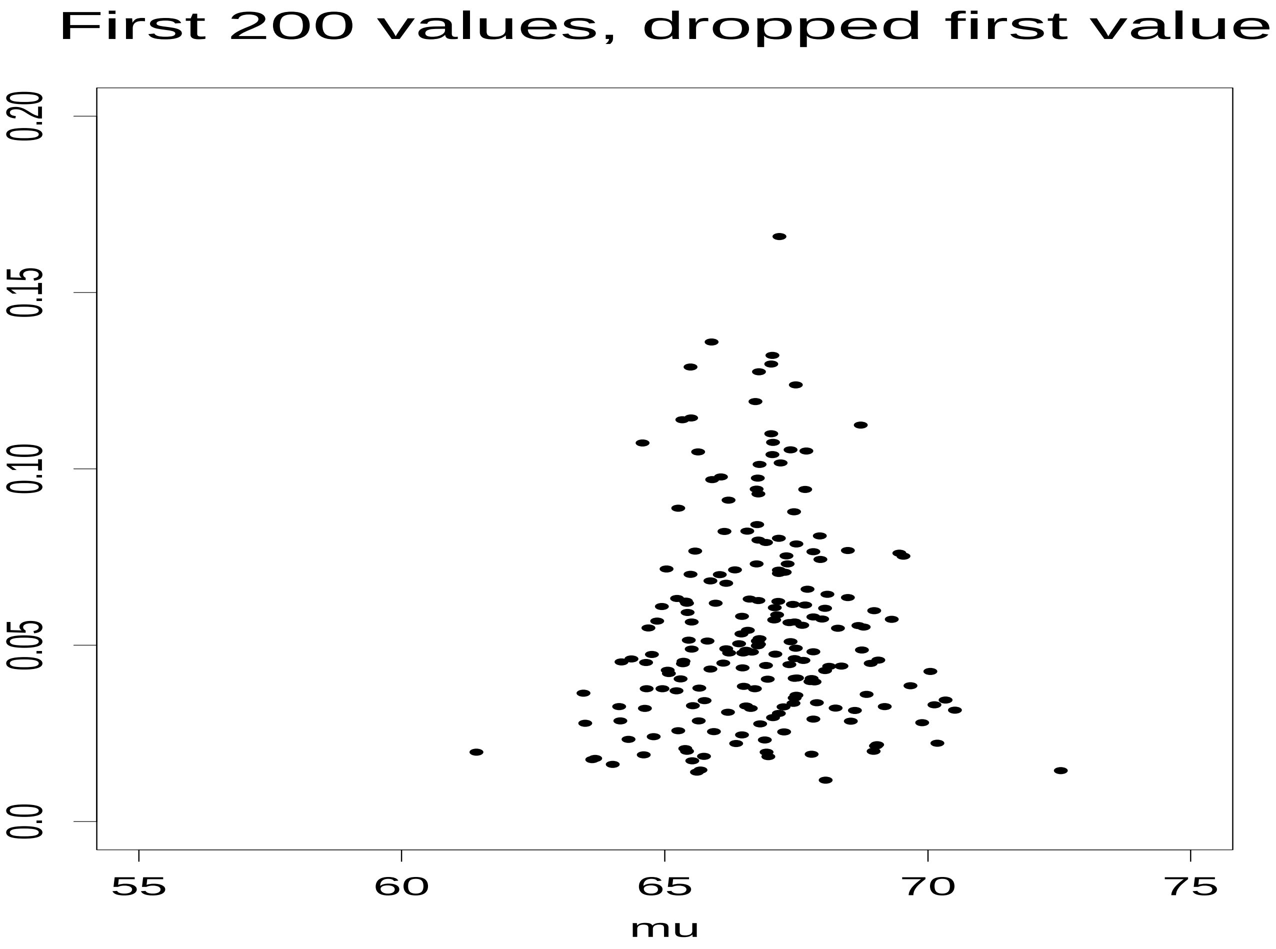

Figure 8: The sampled values from the Markov chain

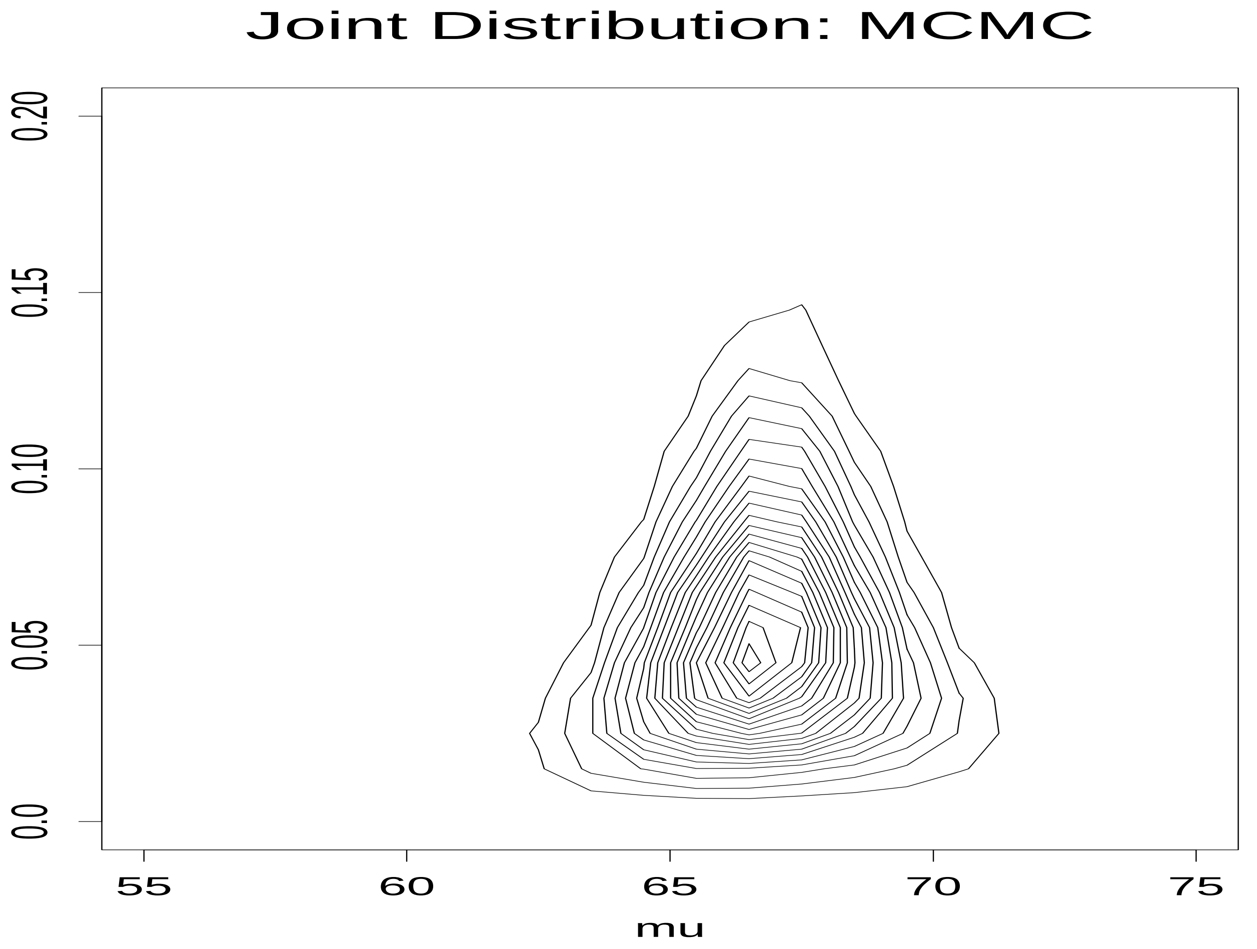

Figure 9: The sampled values from the Markov chain

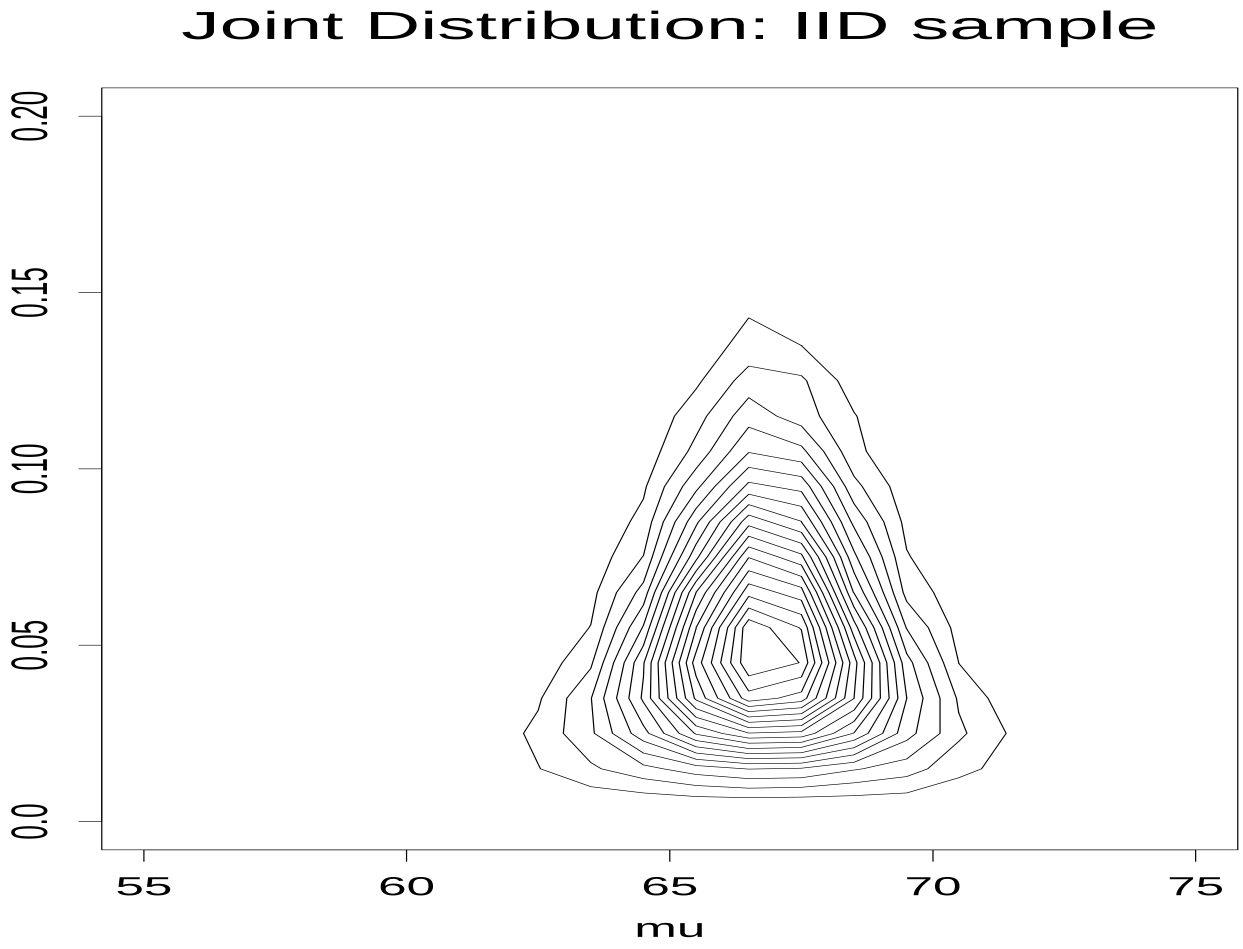

Figure 10: The sampled values from an independent sample.

## Path Plots

(continued)

- After observing the path of the samples, we see that the points are being mostly sampled from the area of highest posterior probability.
- These sampled points can be used to estimate the joint posterior distribution.
- Estimating the joint posterior from the sampled points results in an estimate of the joint posterior distribution which is similar to the estimate of the joint posterior distribution obtained from the independent samples from the posterior distribution.

# Trace Plots

- The next series of plots plot one of the parameters against the iteration number. They are usually called trace plots.
- These plots show how quickly the sampled parameter values move around the sample space. With one of these plots, one can sometimes detect autocorrelation. If the plots show that the sequentially sampled parameter values seem to jump over the high probability region, then this shows that the Markov chain “mixes” well.
- Also, these plots can be helpful in seeing how long it takes the chain to “forget” the initial value of the chain. After a while, the chain seems to find the area of high probability. This initial portion of the chain where the chain wanders around “looking” for the area of high probability is sometimes called the “burn-in”.

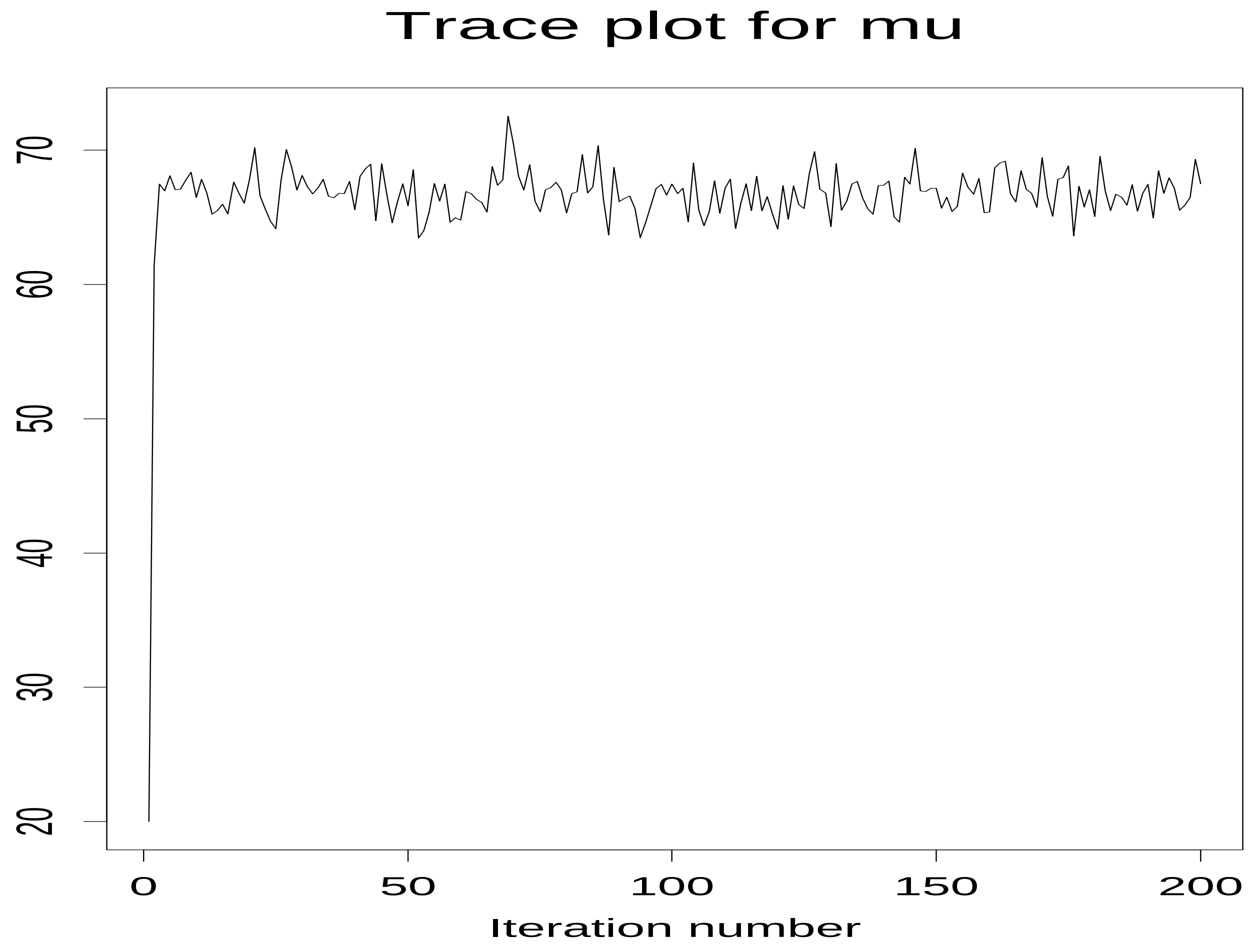

Figure 11: Plot of $\mu^{(i)}$ versus $i$.

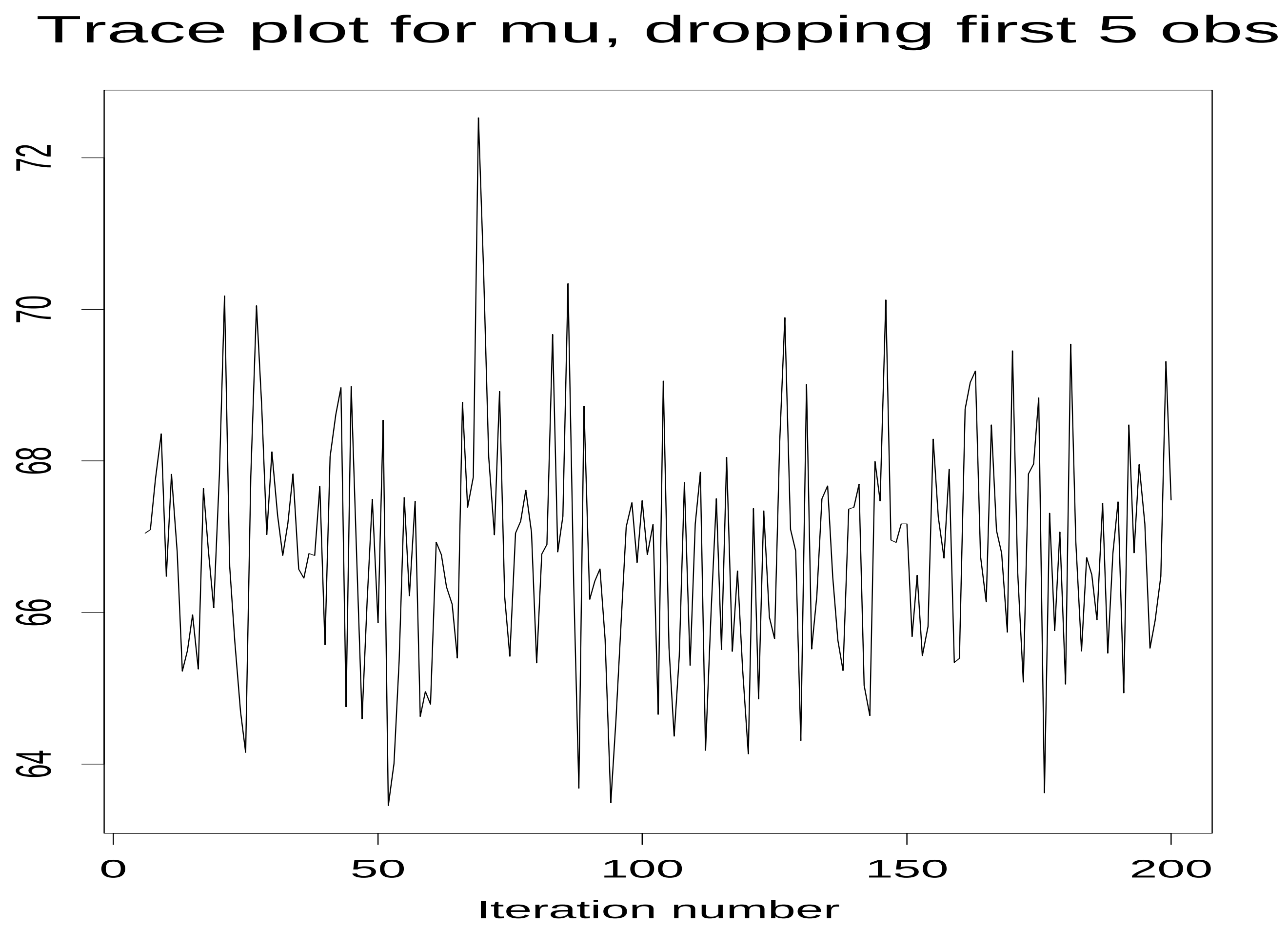

Figure 12: Plot of $\mu^{(i)}$ versus $i$.

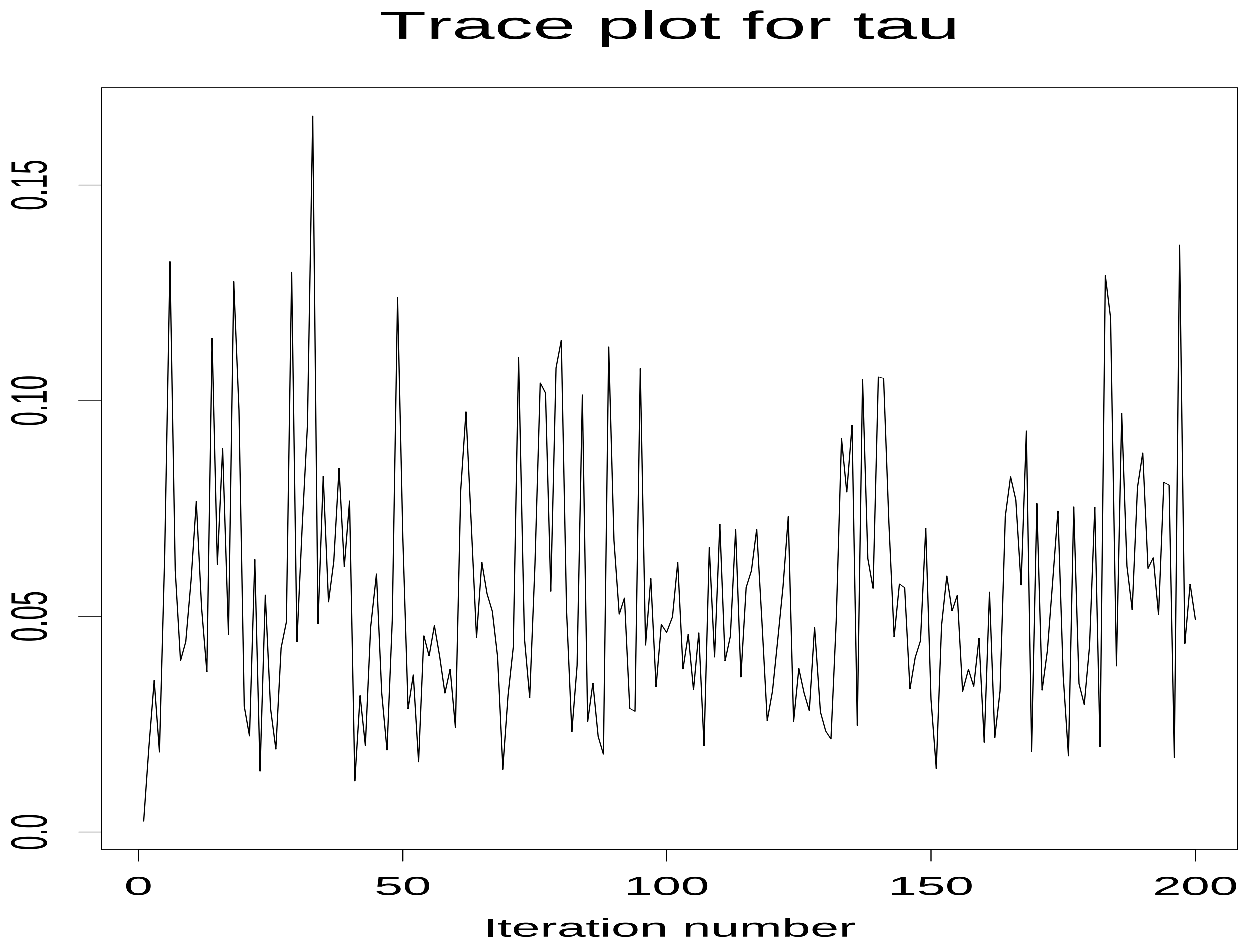

Figure 13: Plot of $\tau^{(i)}$ versus $i$.

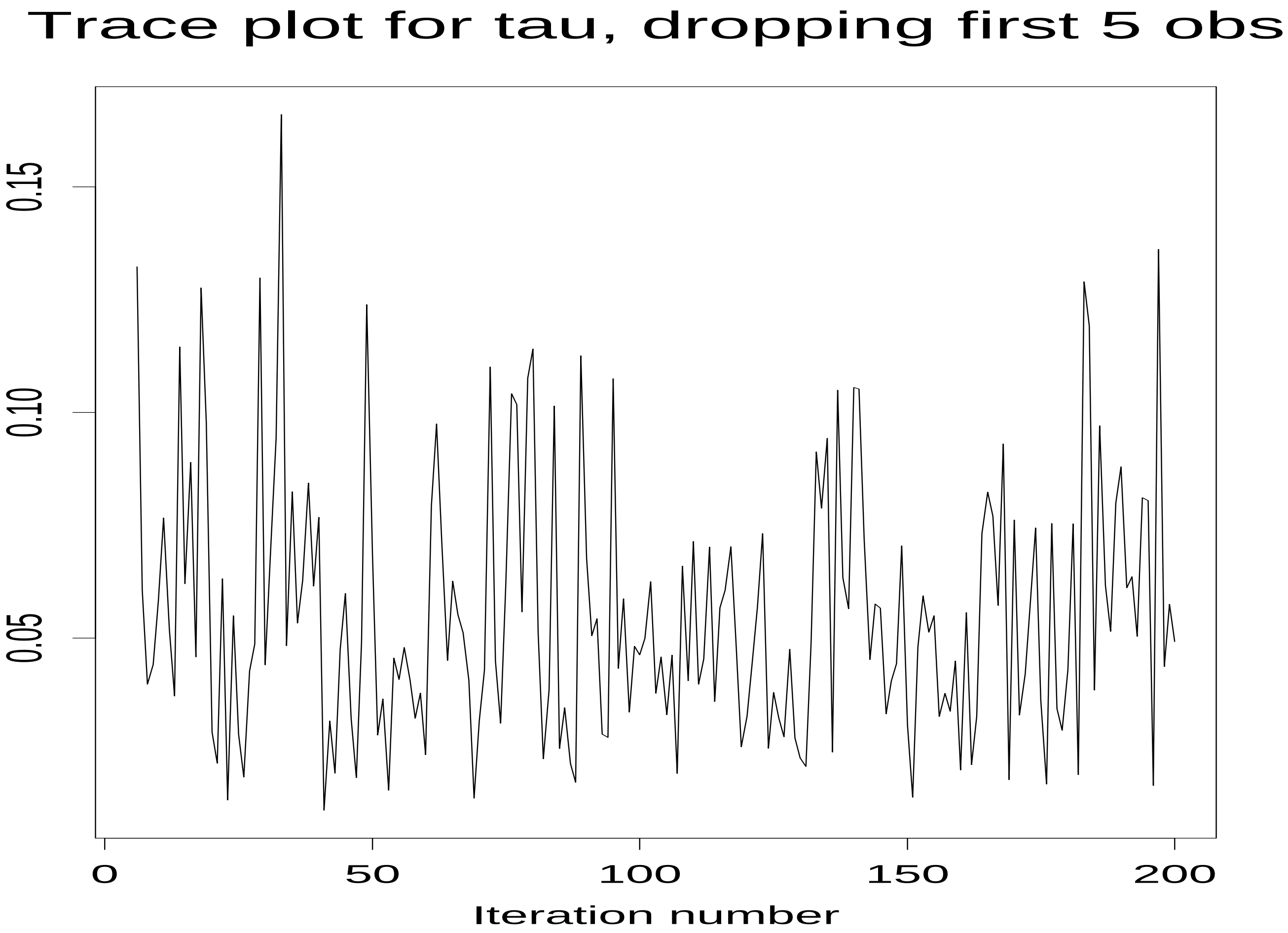

Figure 14: Plot of $\tau^{(i)}$ versus $i$.

## Trace Plots

(continued)

- In these plots, we see that the chain quickly finds the area of high probability. Because of this, the first few points are discarded. (In the preceding plots, the first 5 points are discarded).
- After the first few points are removed, the plots of each parameter seem to show that the parameter values quickly move over the entire high probability area.

## Estimating the Posterior Distribution

- From the Markov chain, 50 000 samples are obtained.
- These samples are used to estimate the joint posterior distribution in Figure 9 and in the proceeding plots. These estimates are compared to the estimates of the posterior distribution obtained from independent samples from the posterior distribution.
- In the lectures later in this course, there is a discussion as to number of samples which should be obtained to make estimates.

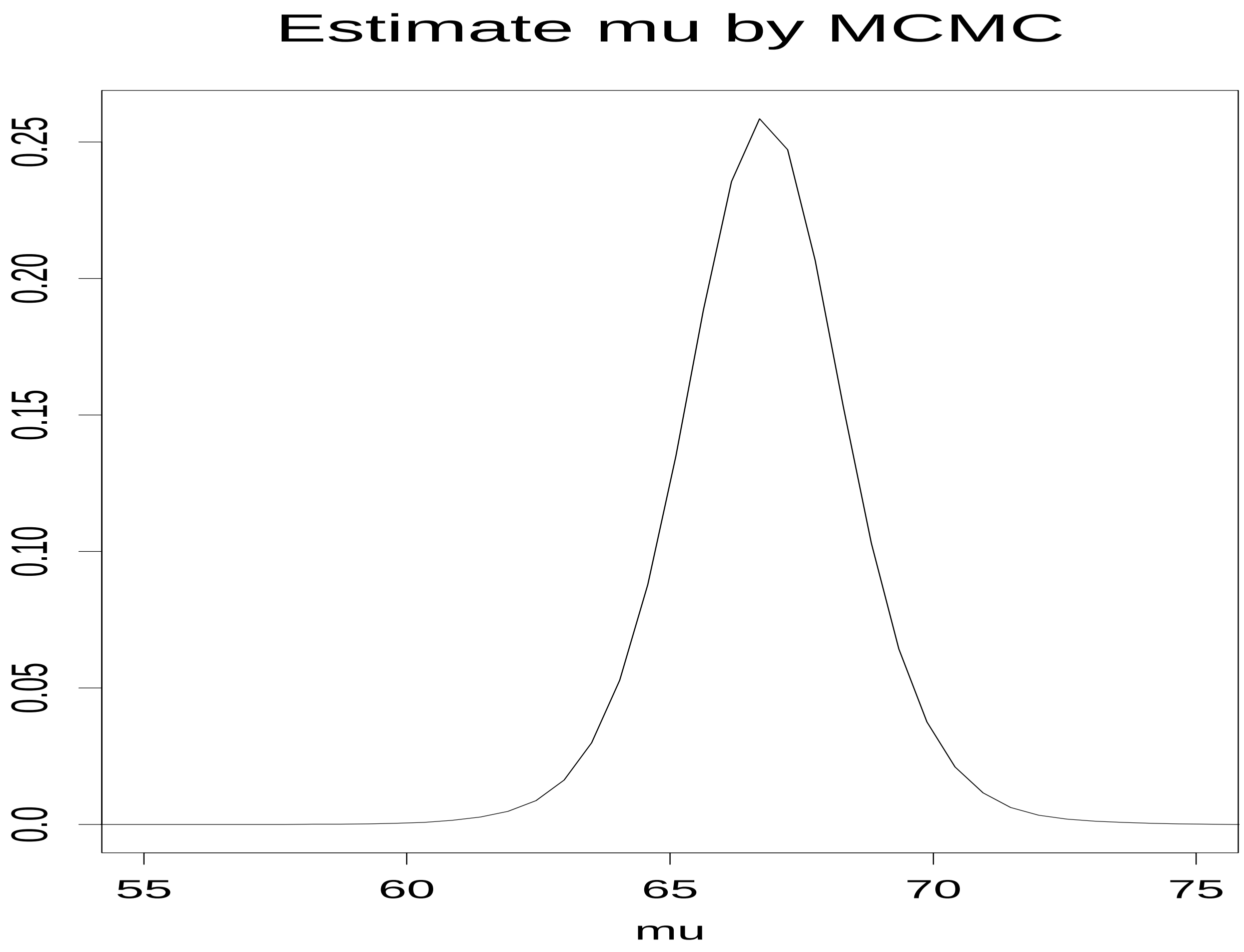

Figure 15: Estimate of the posterior density of $\mu$.

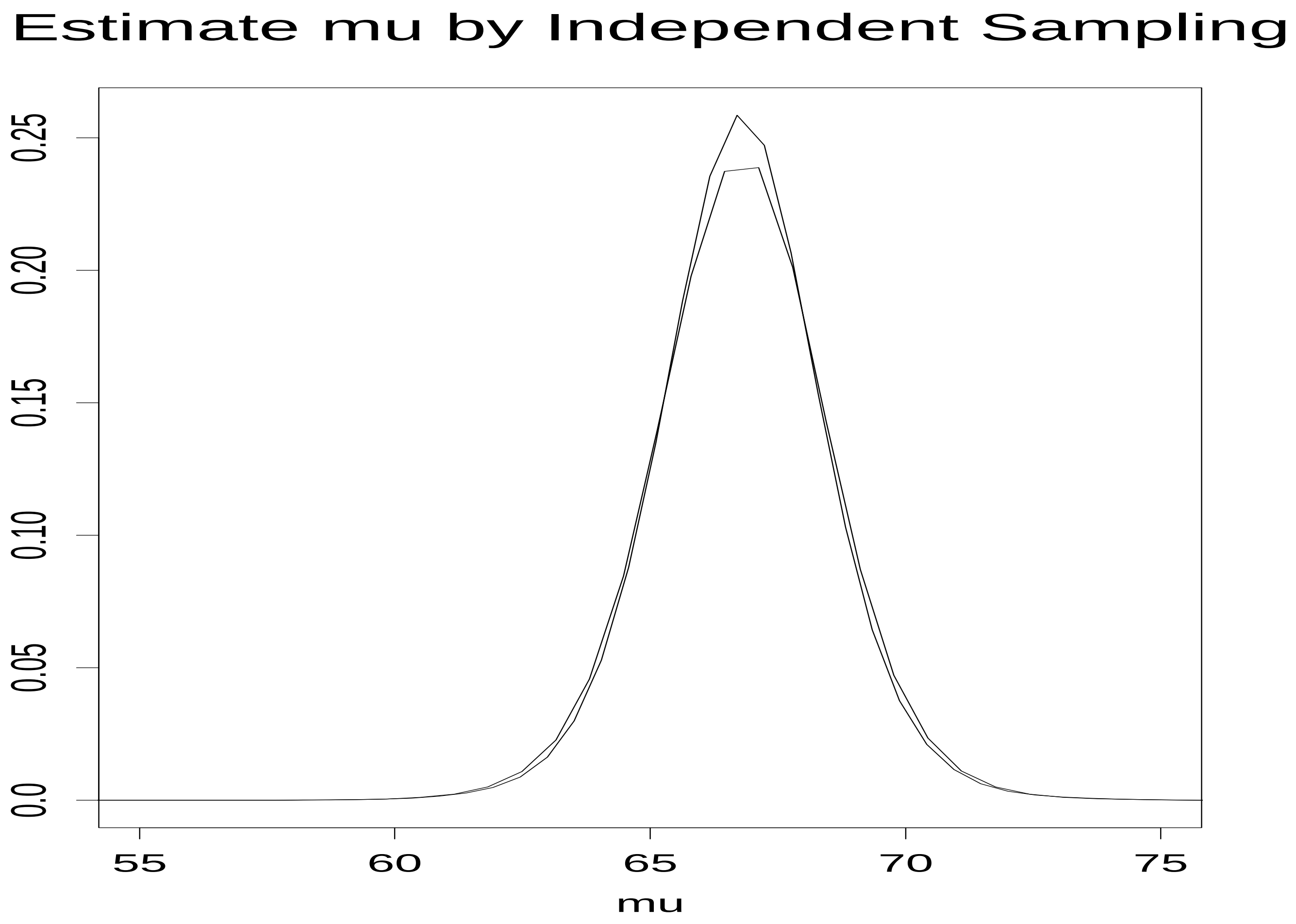

Figure 16: Estimate of the posterior density of $\mu$.

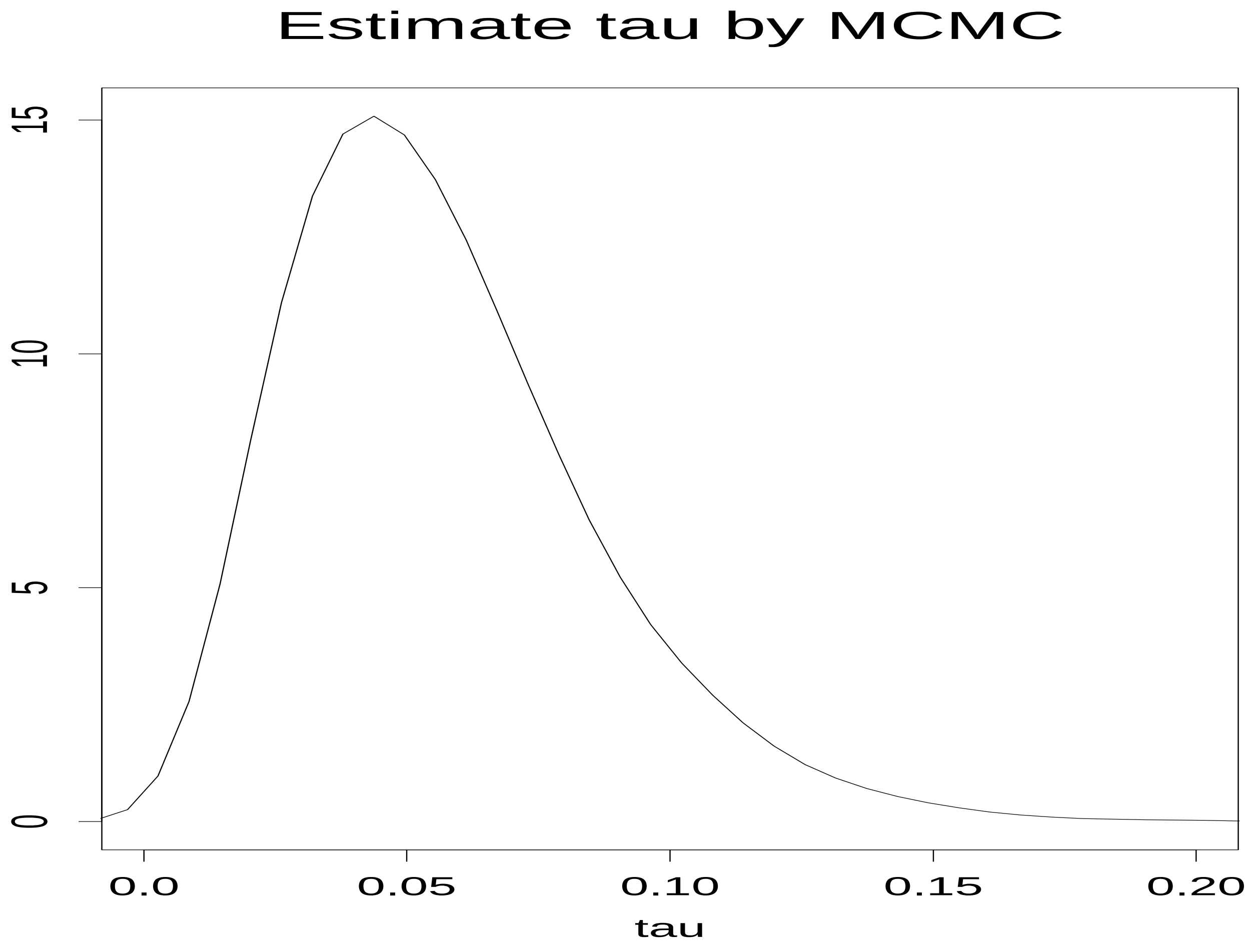

Figure 17: Estimate of the posterior density of $\tau$.

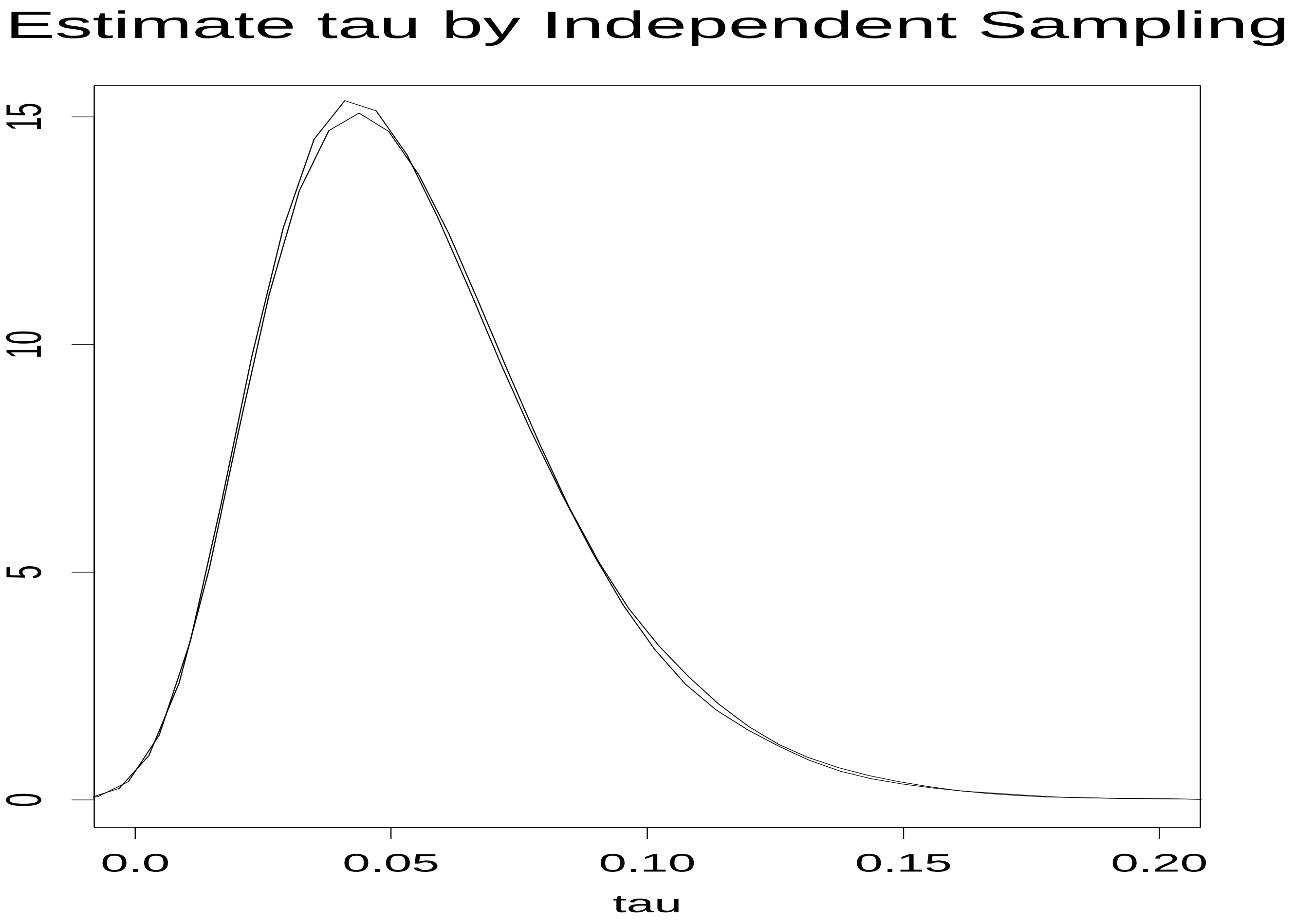

Figure 18: Estimate of the posterior density of $\tau$.

## Example 7: Comments

- This example illustrates how a Markov chain can be created to obtain samples from the posterior distribution.
- These sampled points can then be used to learn about the posterior distribution.
- Next, some of the basic steps to create a Markov chain are discussed.

## The Gibbs Sampler

- References: Geman and Geman (1984) and Gelfand and Smith (1990).
- Two of the oldest types of Markov Chain Monte Carlo algorithms are the Gibbs sampler and the Metropolis algorithm. Here the Gibbs Sampler is explained. Later in the course, the Metropolis algorithm is discussed.

## The Gibbs Sampler

The basic Gibbs Sampler:

- Let $\underline{\theta}$ be the parameters for the Bayesian model. Without loss of generality, let $\underline{\theta}$ be 3-dimensional.
- The algorithm:
  - Start with a set of initial values: $\underline{\theta}^{(0)}$.
  - Given the m-th update, get the (m+1)-th sample by:

$$\begin{aligned}
\theta_1^{(m+1)} &\sim [\theta_1|\theta_2 = \theta_2^{(m)}, \theta_3 = \theta_3^{(m)}] \\
\theta_2^{(m+1)} &\sim [\theta_2|\theta_1 = \theta_1^{(m+1)}, \theta_3 = \theta_3^{(m)}] \\
\theta_3^{(m+1)} &\sim [\theta_3|\theta_1 = \theta_1^{(m+1)}, \theta_2 = \theta_2^{(m+1)}]
\end{aligned}$$

## The Gibbs Sampler

(continued)

- The joint posterior is invariant; therefore, the posterior is the limiting distribution and

$$\frac{1}{M}\sum_{m=1}^{M} g(\underline{\theta}^{(m)}) \xrightarrow{a.s.} E[g(\underline{\theta})|\text{data}]$$

# Getting Results of Interest

What do we want to do:

- Posterior moments
- Posterior distribution (graphing the shape)
- Posterior distribution of functions of the parameters

How to get:

- Path averages

Note: If you give a statistician an unlimited number of samples from a population, then the statistician will be able to figure out what the population parameters are.

## Other MCMC Methods

- Gibbs Sampling is only one of the MCMC algorithms.
- Another of the more common algorithms is the Metropolis Algorithm. In this algorithm, you start with an algorithm chain which allows the chain to wander over the entire sample space. Then, an adjustment is made to the algorithm (similar to an importance sampling step) which results in the new algorithm having the correct limiting distribution.
- Please note that this is an active area of statistical research and different and improved algorithms are continuing to be developed.
- In order to build your own algorithms, one needs to know some mathematical statistics (and some calculus). This is beyond the level of this course.

## Part 2: MCMC algorithms

Summary

- Basic idea: get a sample of the posterior distribution and use the path averages to estimate the desired features of the posterior distribution.
- Since a Markov Chain is used, the sampled points are correlated.

## Part 3: Unusual Markov Chain Samplers

Outline:

- Introduction
- Correlated Normals
- 3-block model
- Witch’s hat
- Comments

## Unusual Markov Chain Samplers

- Theoretically, there are very weak requirements needed to insure that the algorithm converges to an answer.
- Also, the example of the heights data shows how the algorithm works in practice.
- Still, it is important to explore what might go wrong in practice. So, the following examples shows some of situations which might cause problems for an MCMC algorithm.

## Part 3: Unusual Markov Chain Samplers

(continued)

- The three distributions below describe joint distributions on the 2 dimensional plan. Call one axis $X$ and the other $Y$. All these distributions are symmetrical in $X$ and $Y$.
- To run the Markov Chain, you need to know the conditional distribution of $X$ given $Y$ (and $Y$ given $X$). The model given below are specified by giving the joint distribution.
- To run the algorithm, the chain is started in the support of $(X, Y)$ and then sampled from the conditional distributions $X|Y$ and $Y|X$.

## Correlated Normal distributions

- The joint distribution or $X$ and $Y$ is a bivariate normal with correlation $\rho$, with $\rho$ having values .999, .90, and 0.
- Let the random variable $X$ have mean $\mu_X$ and standard deviation $\sigma_X$, and $Y$ has mean $\mu_Y$ and standard deviation $\sigma_Y$.
- to get a sample of $Y$ given $X$, first sample a random variable $Z$ which is normally distributed with mean 0 and precision 1 and $Z$ is independent of $X$. Then, let $Y$ be the following:

$$Y = \left[ \frac{(X - \mu_X)}{\sigma_X} \rho + Z(1 - \rho^2)^{\frac{1}{2}} \right] \sigma_Y^2 + \mu_Y .$$

- Draw $X$ given $Y$ by a similar formula.
- The next set of plots looks at samples drawn from this chain when $\rho = .999$.

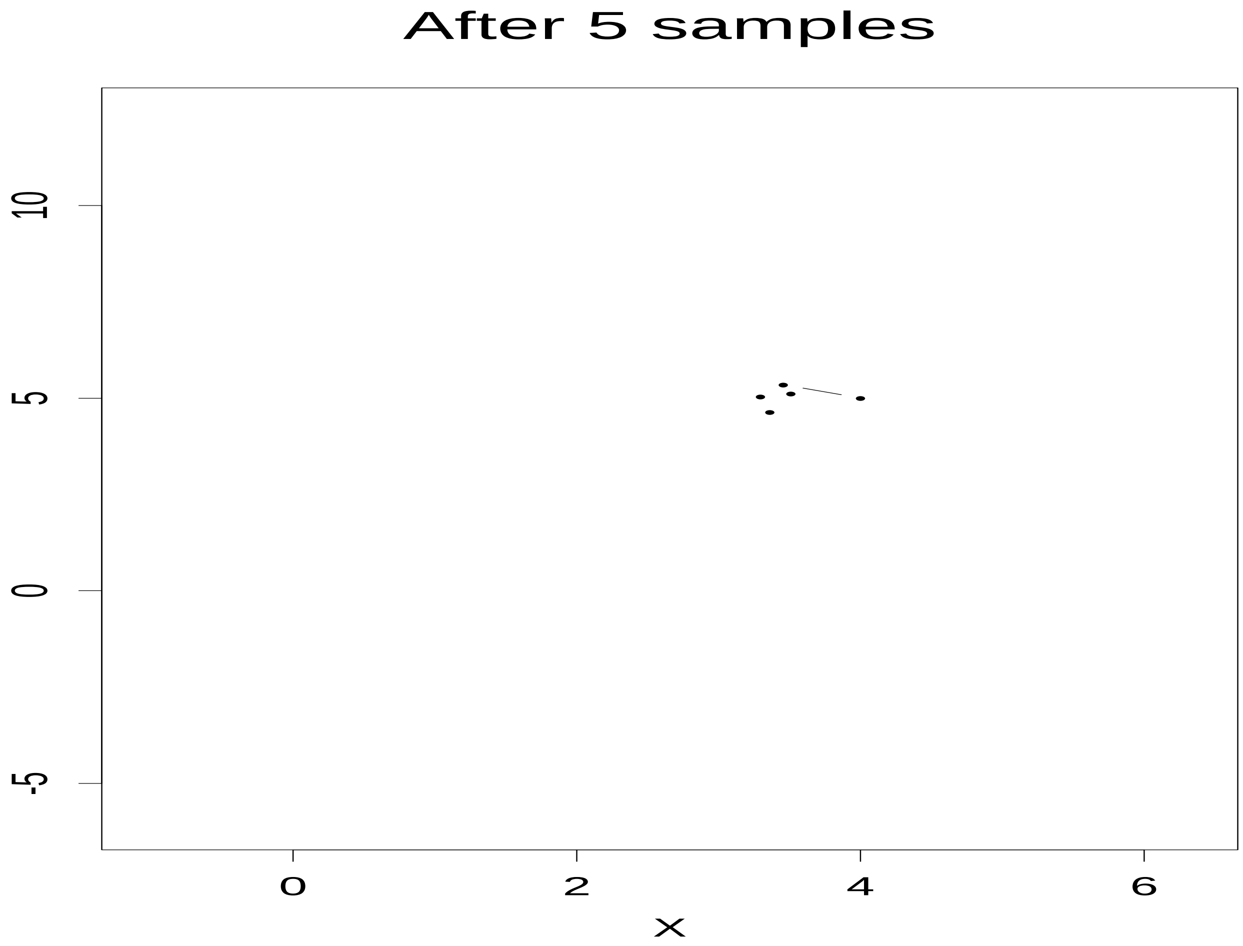

Figure 19: This is the path plot of X values sampled from a multivariate normal with a correlation of .999.

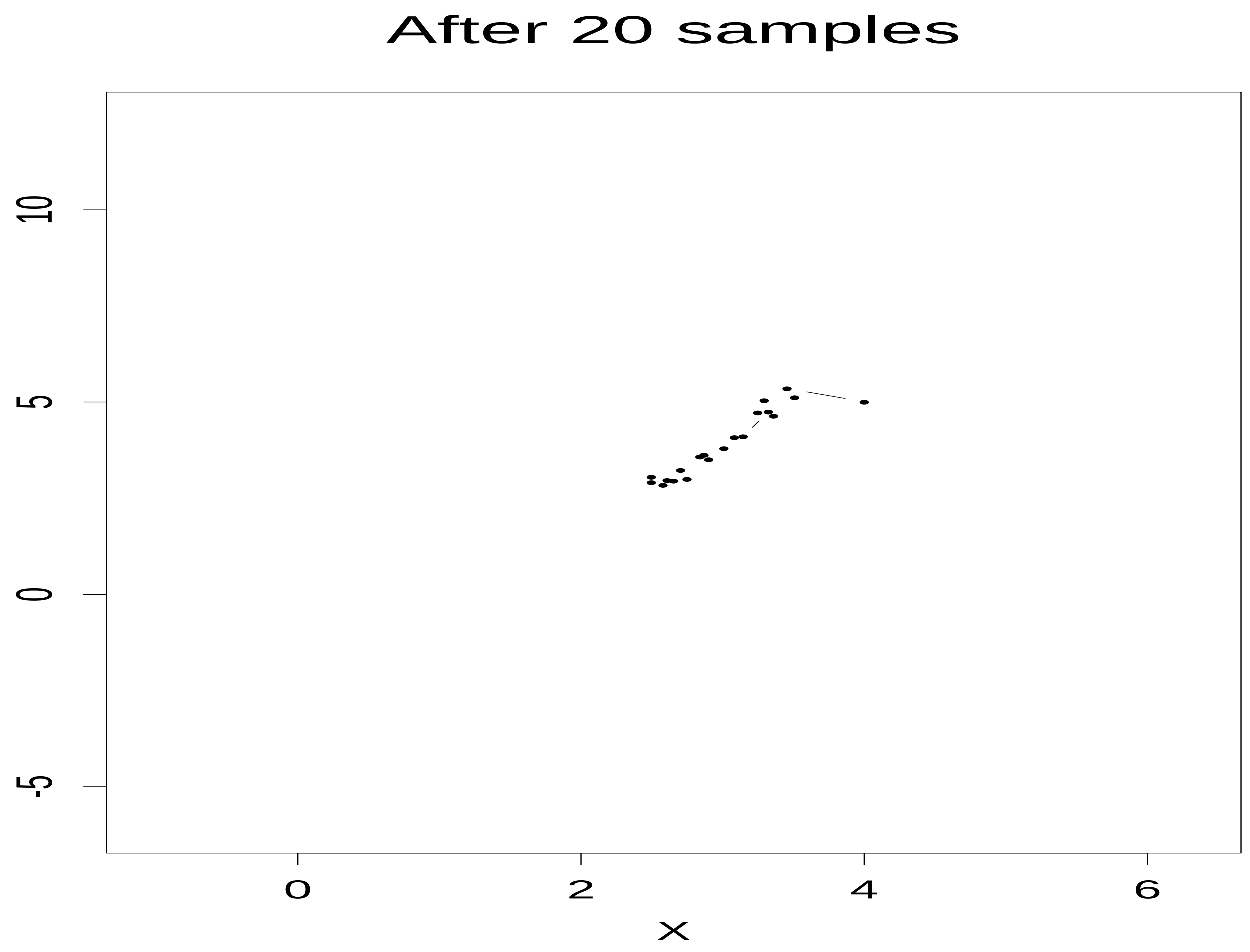

Figure 20: This is the path plot of X values sampled from a multivariate normal with a correlation of .999.

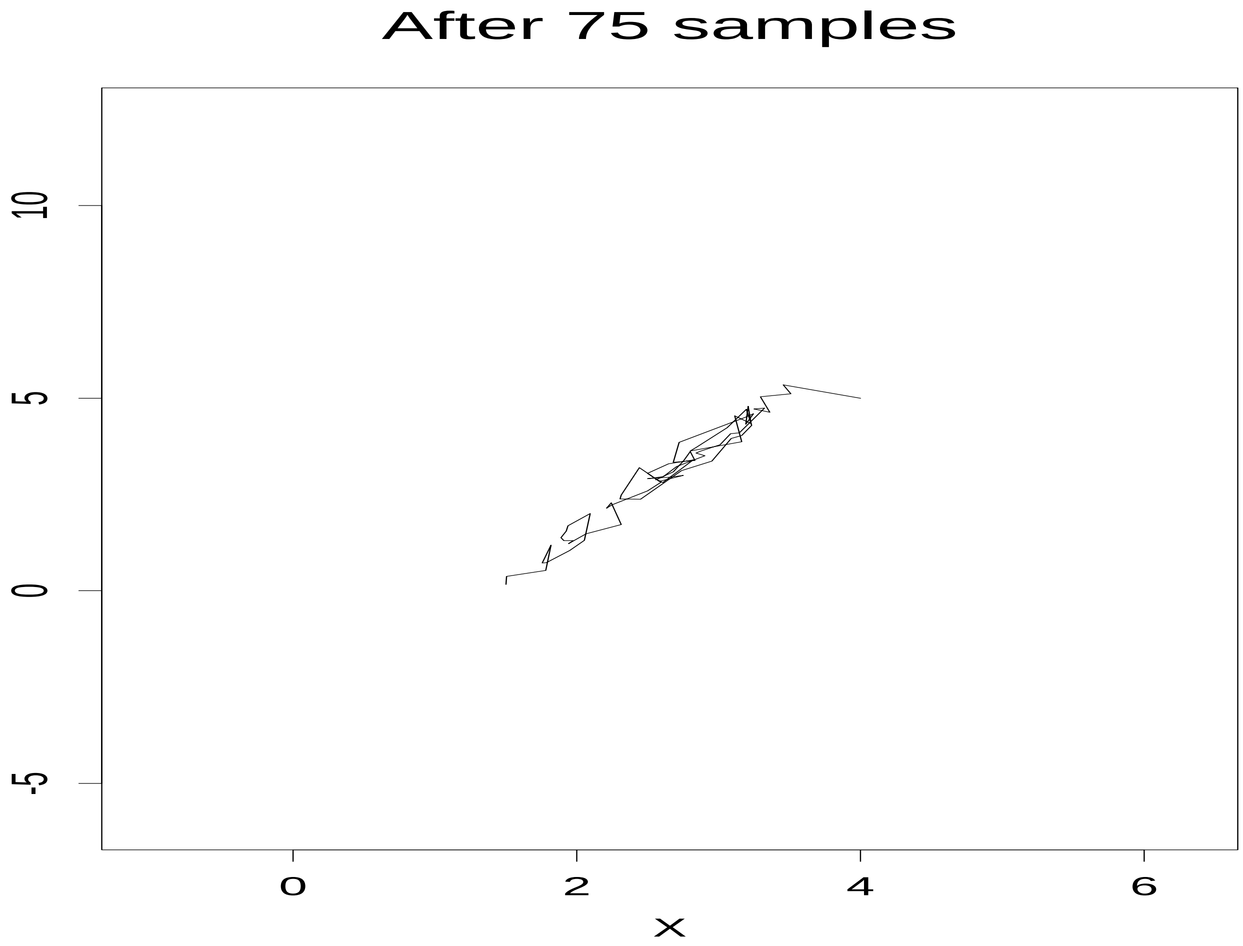

Figure 21: This is the path plot of X values sampled from a multivariate normal with a correlation of .999.

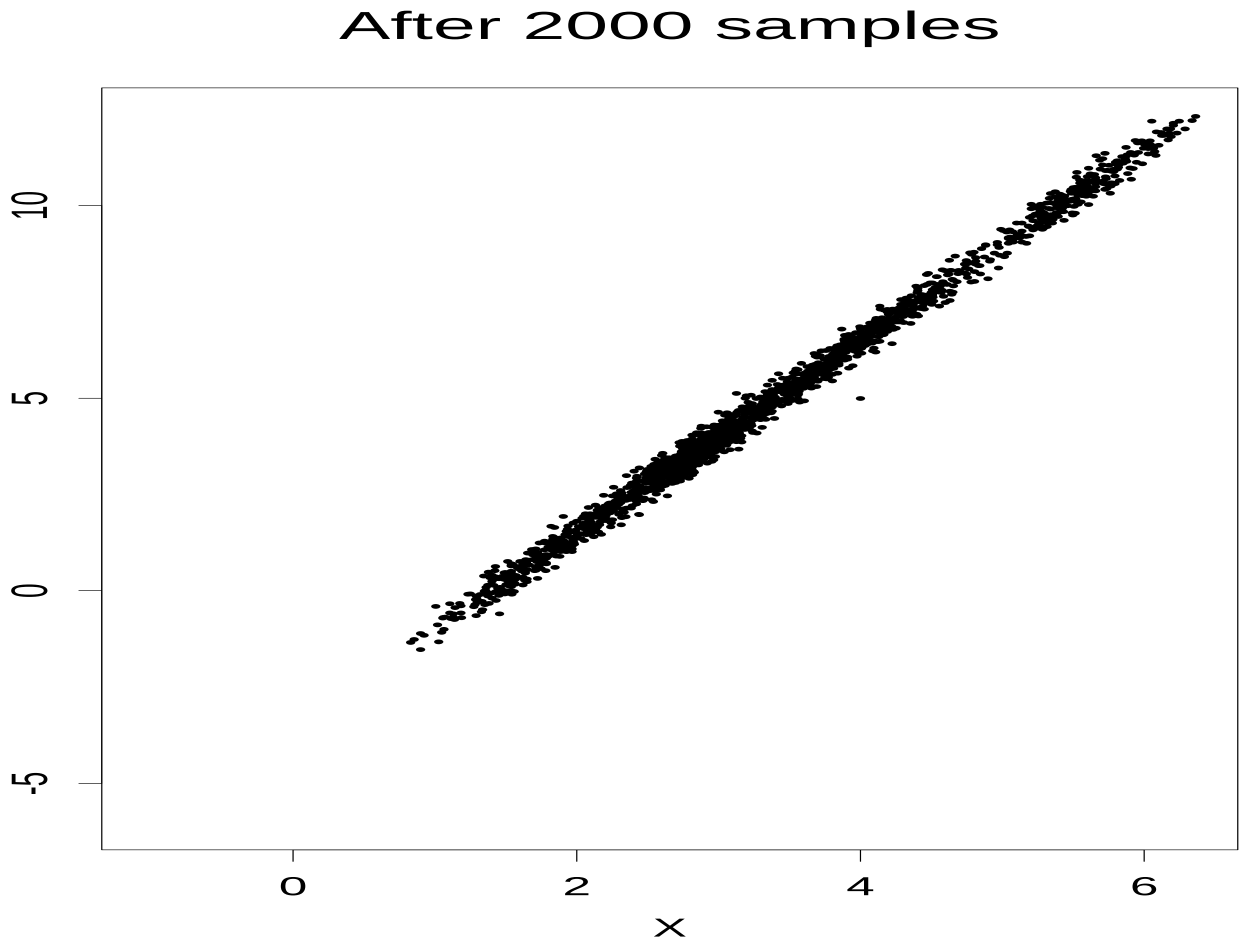

Figure 22: This is the path plot of X values sampled from a multivariate normal with a correlation of .999.

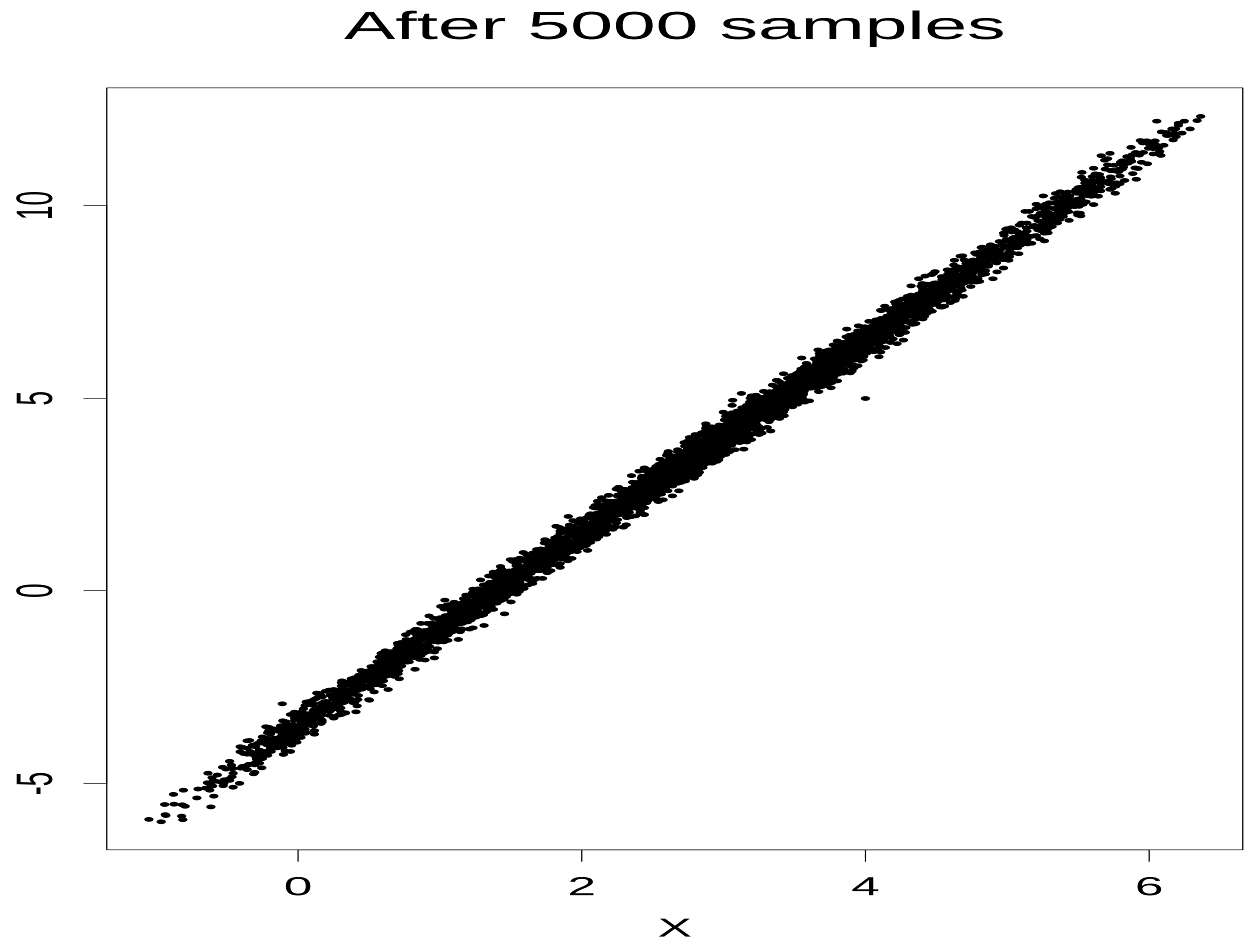

Figure 23: This is the path plot of X values sampled from a multivariate normal with a correlation of .999.

- Please note that the sequential samples of $(X, Y)$ are fairly close to each other. The sampler takes time to go from one end of the main probability mass to the other end.
- After about 2000 samples, it appears to some that the $X$ values are concentrated in the range of values between 1 and 6.
- However, when 3000 more samples are observed, then we clearly see that there is a high probability for $X$ in the interval between 0 and 1. If we stopped observing values at the 2000th iteration of the Markov chain, then we would not have known this.
- Now, let us look at the trace plots for another look at what is happening with the $X$ values.

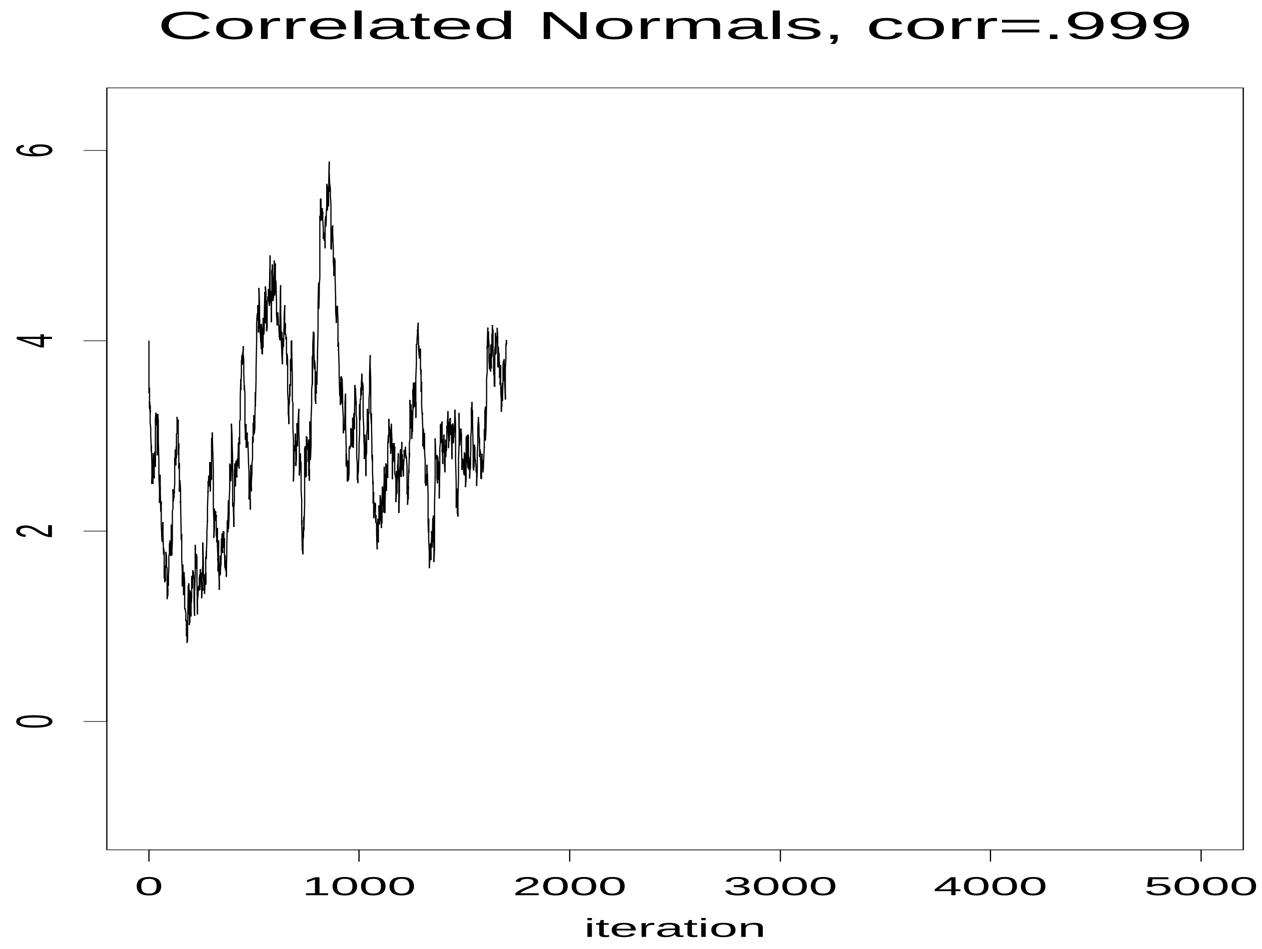

Figure 24: This is the trace plot of X values sampled from a multivariate normal with a correlation of .999.

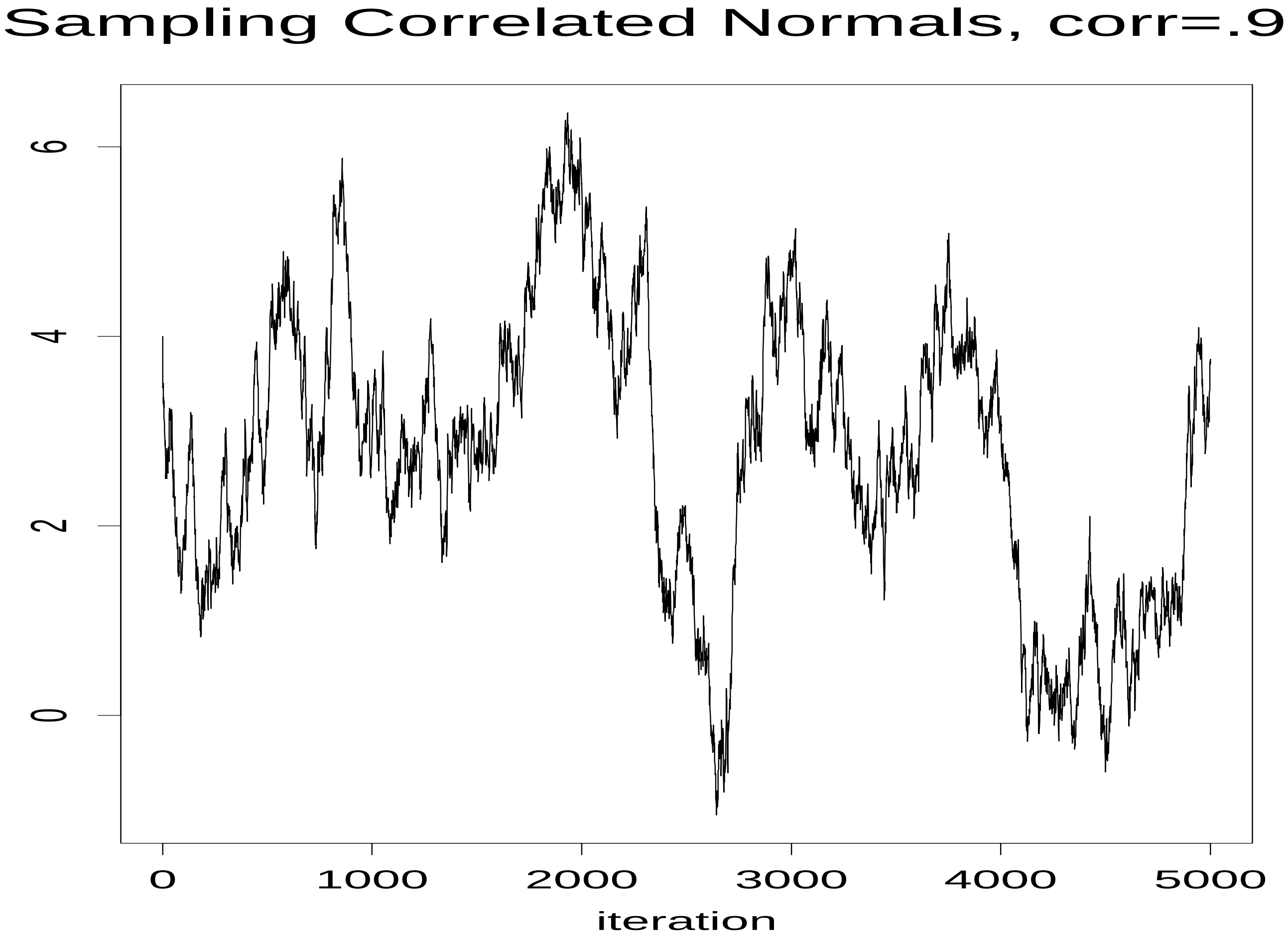

Figure 25: This is the trace plot of X values sampled from a multivariate normal with a correlation of .999.

## Correlated Normals: Path Diagrams

- It looks like the chain is very happy sampling values between 1 and 6 for the first part of the chain. Then, at around the 2000th sample, the chain suddenly drops down and begins to sample points below 1 before it jumps back up between 2 and 4.
- Now let us consider correlations of .90 and 0. (Note that when the correlation is zero, then the two normals are independent.)

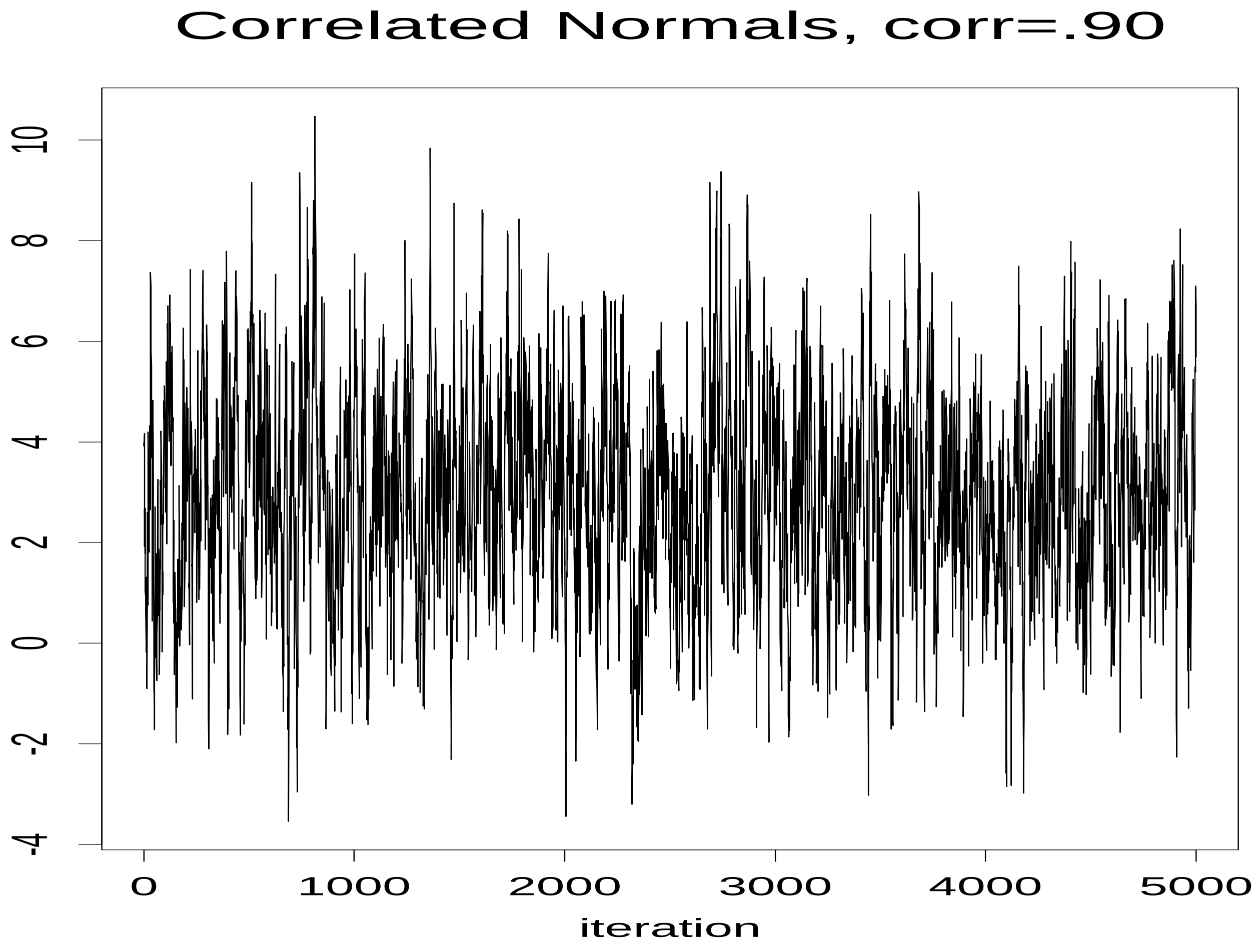

Figure 26: This is the trace plot of X values sampled from a multivariate normal with a correlation of .90.

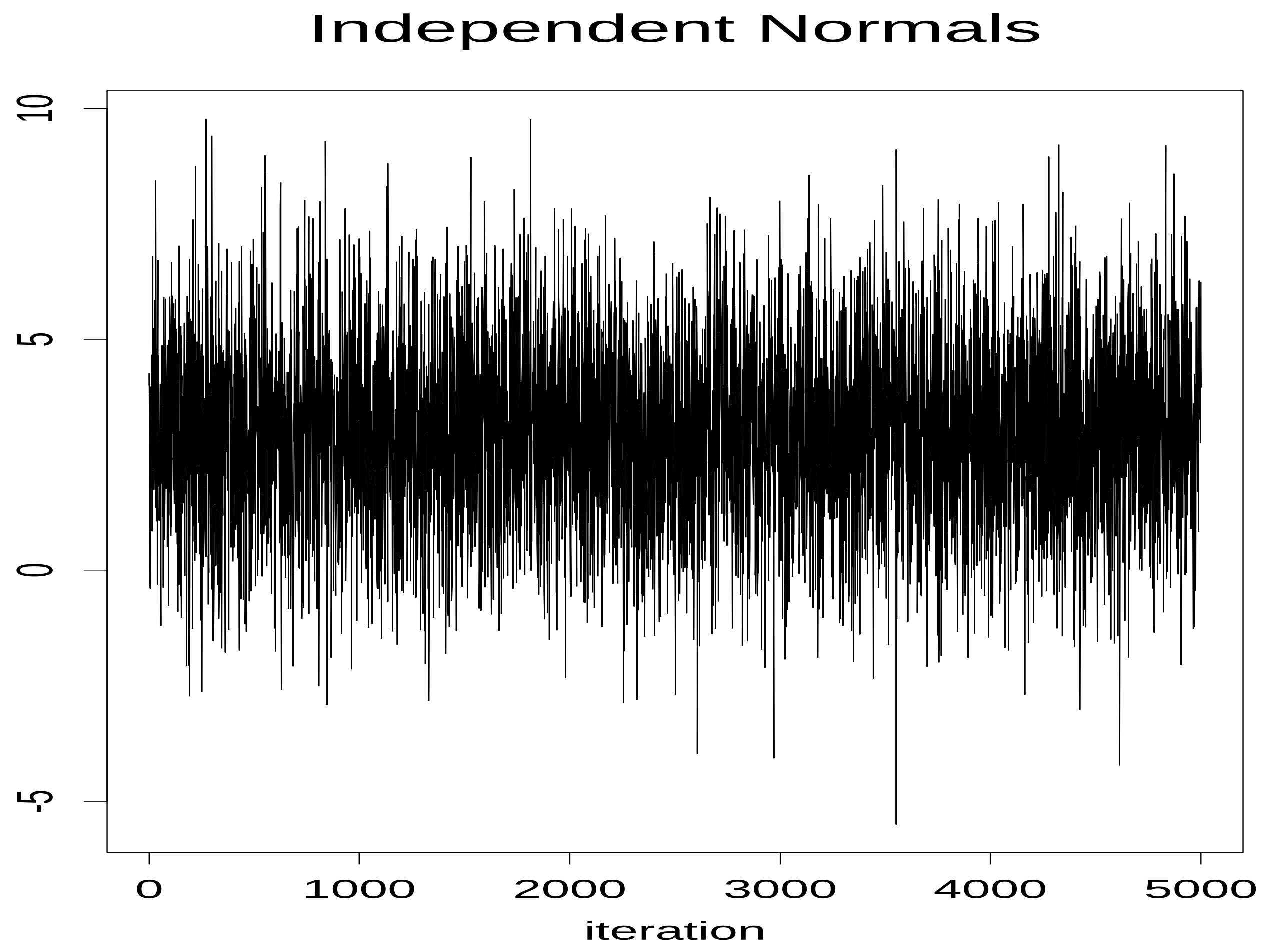

Figure 27: This is the trace plot of X values sampled from a multivariate normal with a correlation of 0.

## Correlated Normals

- With correlations of .90 and 0, the trace plots show that the sampled $S$ values quickly move over the range of the high probability mass. So, here, the estimated parameters for the limiting distribution will be fairly good.

- An alternative way to sample the $(X, Y)$ with correlation .999 is to sample some other chain, call it $(U, V)$, and then transform $(U, V)$ to $(X, Y)$. For example, if $(U, V)$ are the sampled from a bivariate normal with zero correlation (like we already did), then one could define $(X, Y)$ as the following:

$$\begin{aligned} X &= U * \sigma_X + \mu_X \\ Y &= (U * \rho + V * \sqrt{1-\rho^2})\sigma_Y + \mu_Y. \end{aligned}$$

- The next figures show the path of $(X, Y)$ for the first 15 samples and the trace plot for $X$ for the first 100 samples.

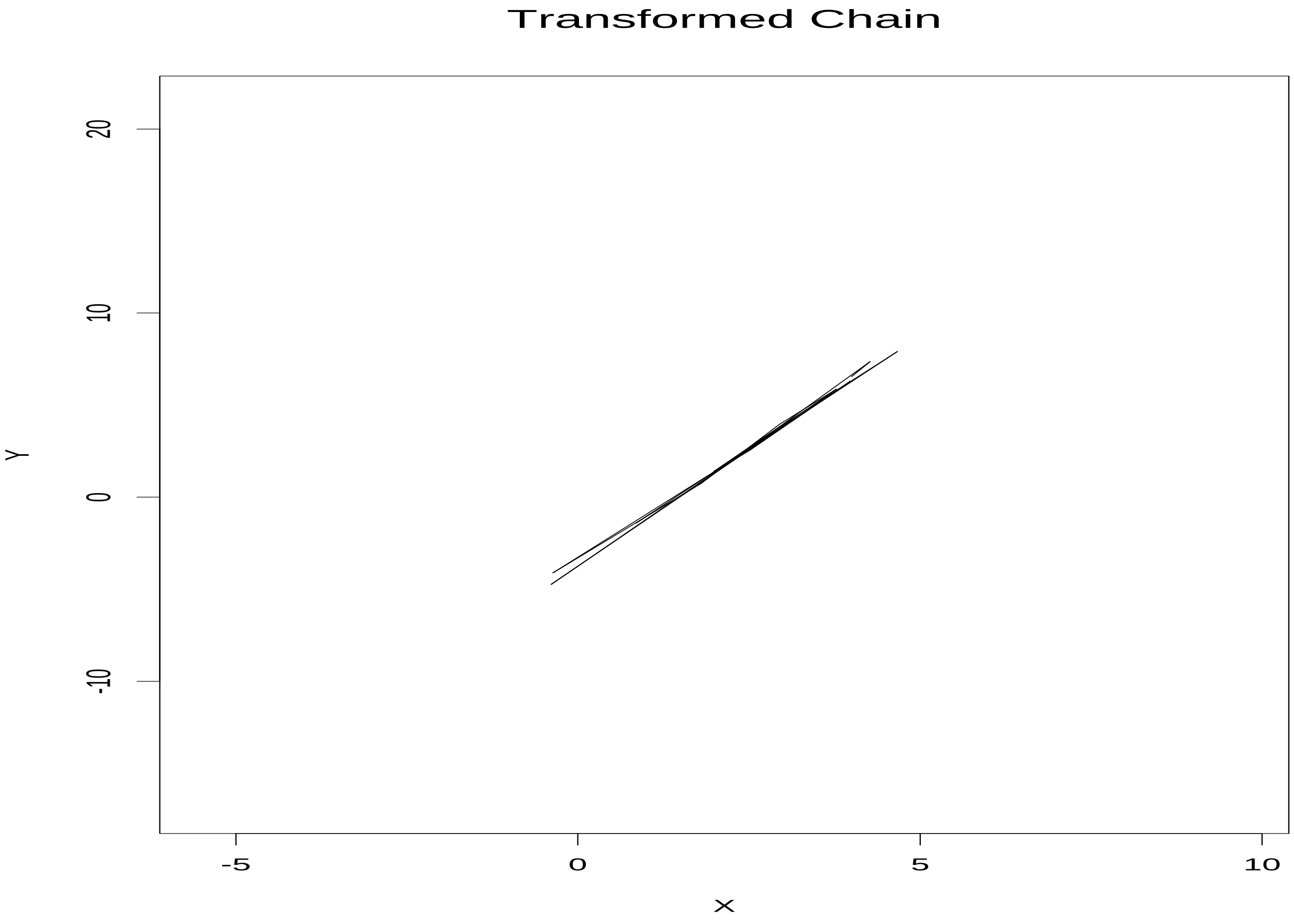

Figure 28: This is the path plot of $(X, Y)$. The random variables are transformed from another chain which is better behaved. This is the first 15 values.

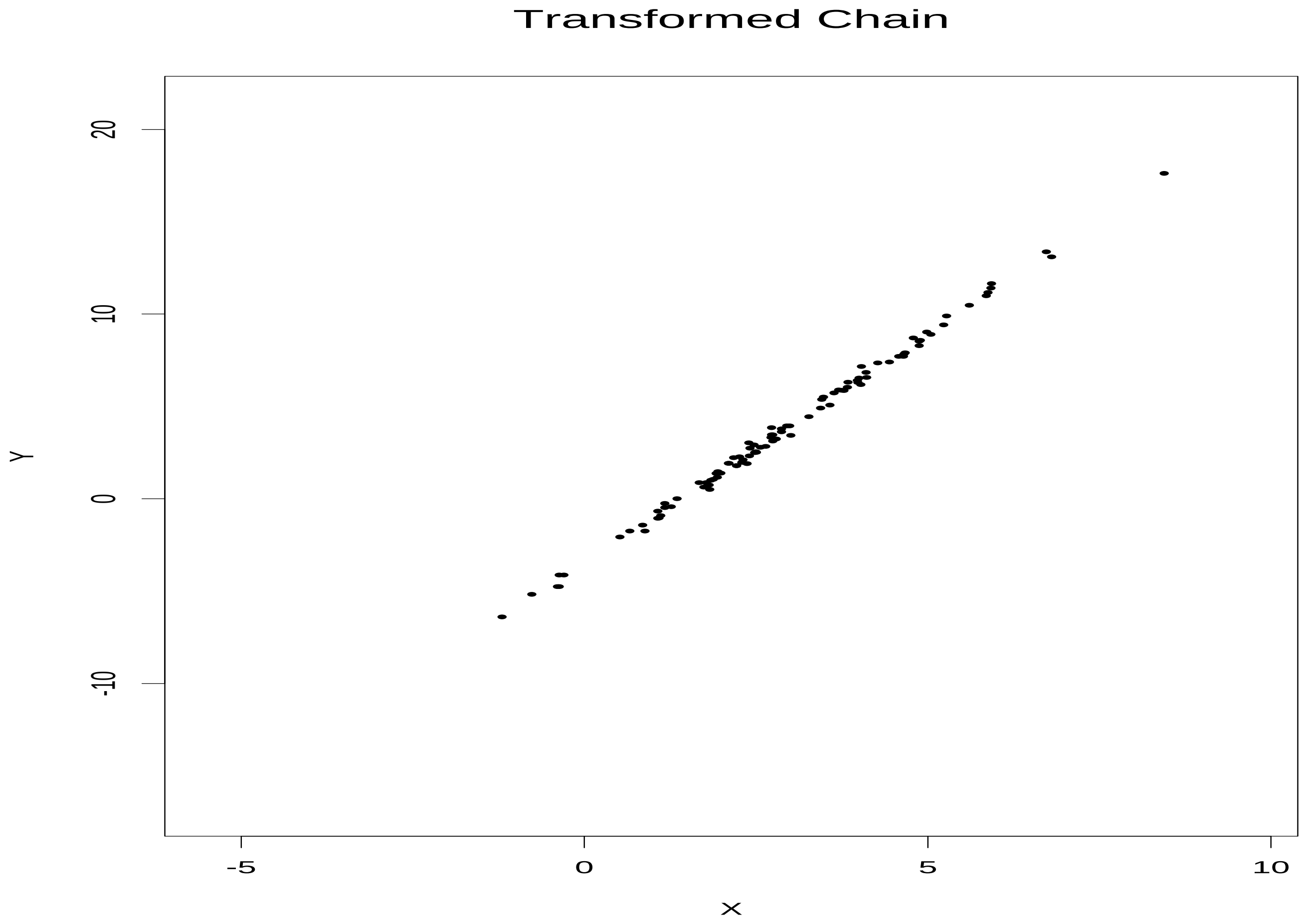

Figure 29: This is the path plot of $(X, Y)$. The random variables are transformed from another chain which is better behaved. This is the first 100 values.

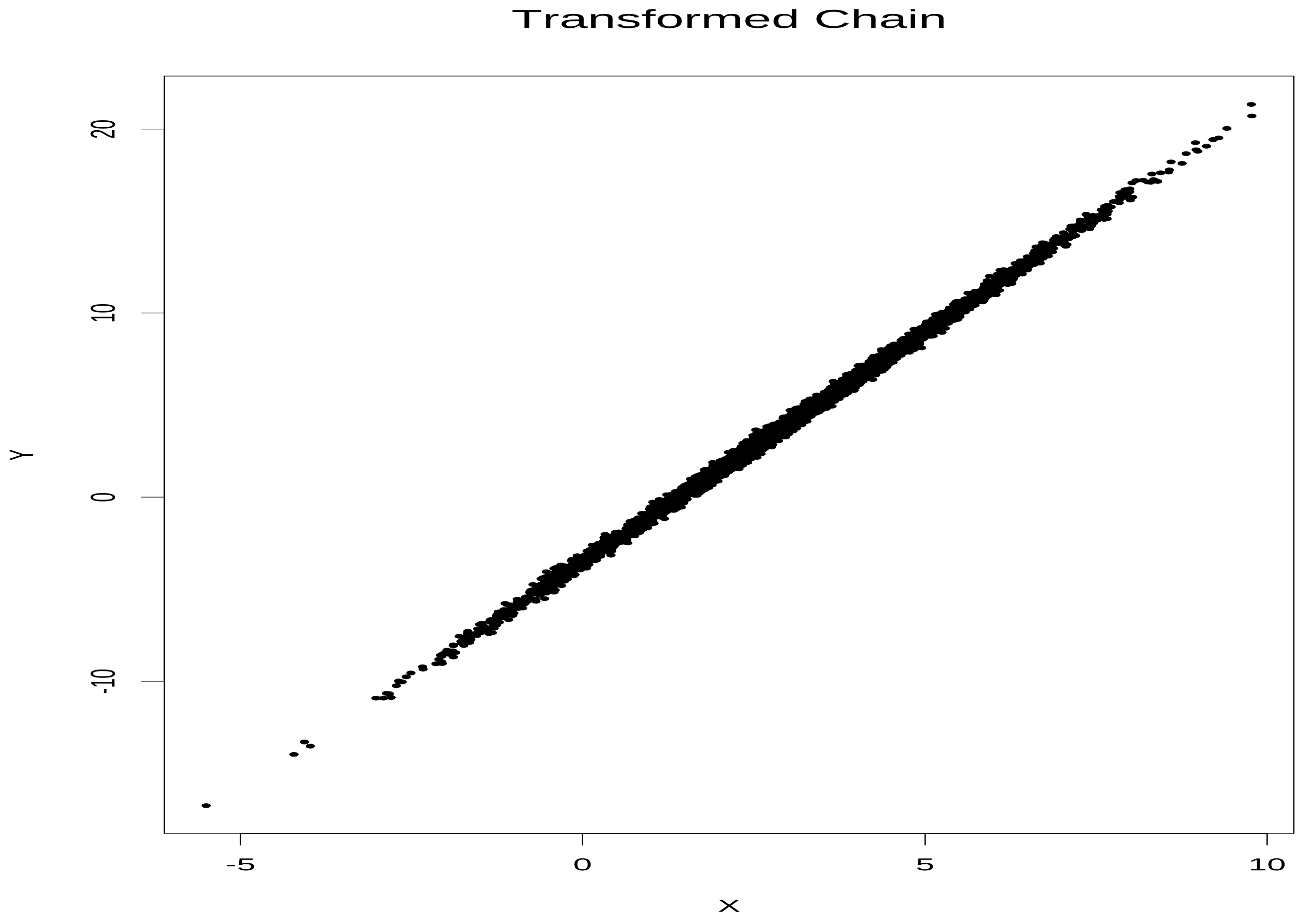

Figure 30: This is the path plot of $(X, Y)$. The random variables are transformed from another chain which is better behaved. This is the first 5000 values.

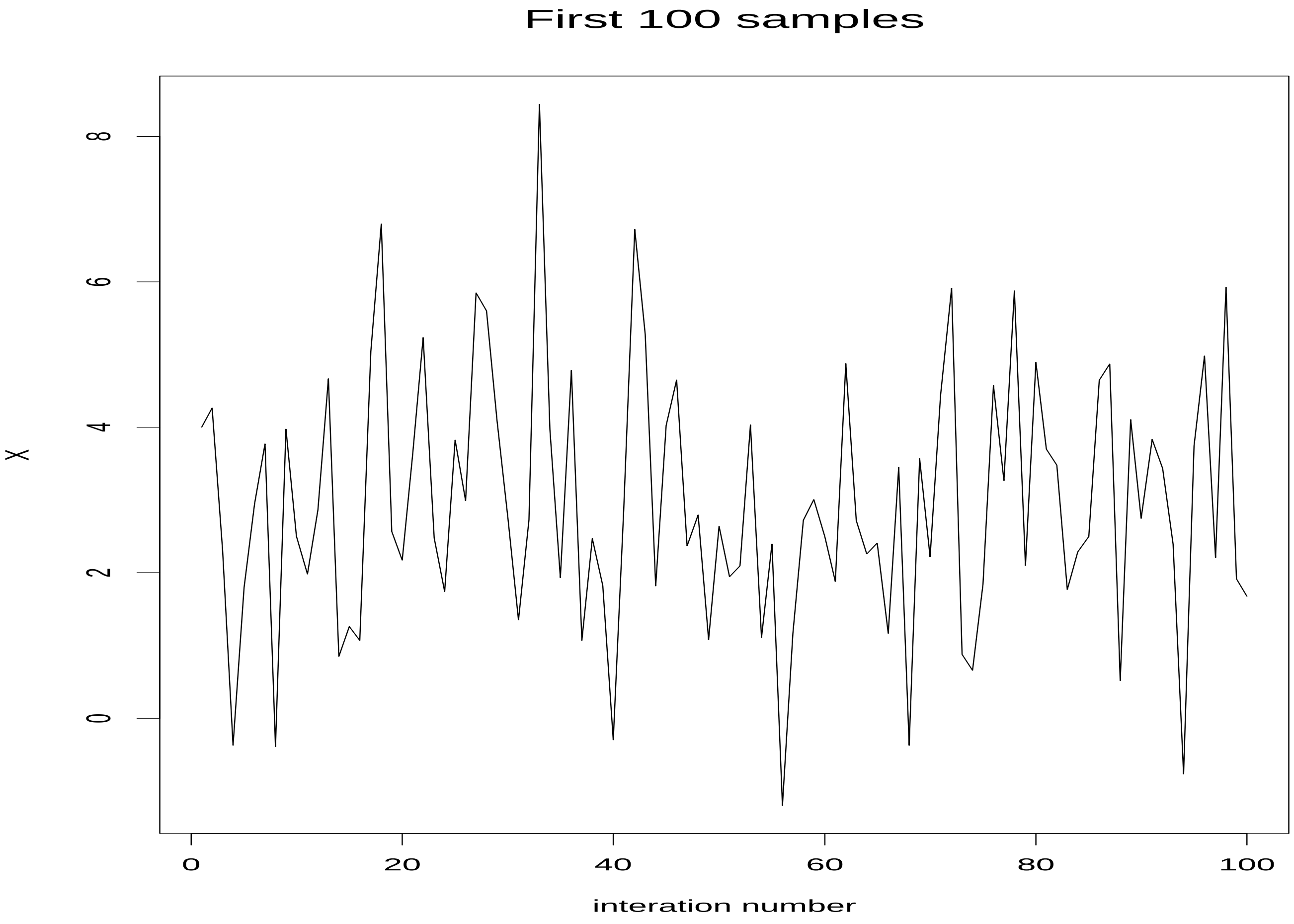

Figure 31: This is the trace plot of $(X, Y)$. The random variables are transformed from another chain which is better behaved. This is the first 100 sampled values of $X$.

## Correlated Normals: Comments

- When there is a lot of cross correlation between $X$ and $Y$, then it may take awhile for the sampler to travel across the region of high probability and therefore it may need more observations to accurately estimate the limiting distribution. (Or, more formally, one could say that the effective sample size is smaller for compared to the same number of samples which are sampled independent and identically distributed.)
- If the sampler moves slowly, then there is the possibility that one might be fooled into thinking that the sampler has already visited the entire area of high probability. The danger here is that one might stop the sampler too soon.

## Correlated Normals: Comments

(continued)

- Sometimes a transformation of the variables results in a sampler which “mixes” quicker. That is, it results in a sampler which quickly moves throughout the values of the limiting distribution.
- This is equivalent to a $2 \times 2$ Markov chain with a very high second eigenvalue. Therefore, in such cases, the convergence is slower.
- Also, with many variables, the sampling space will be much larger (since there are more dimensions). If the Markov chain moves slowly in such a case, then there is the increased chance that one might mistake the slow movement for a chain which has “converged”.

## 3 Blocks Distribution

The joint distribution for this chain is a mixture of three uniform distributions on different squares. Call the three blocks A, B, and C. Figure 32 shows a picture of these blocks.

- For the A block, both X and Y are between -14.5 and -4.5.
- For block B, both X and Y are between -5 and 5. For block C, both X and Y are between 4.5 and 14.5.
- In this chain, the relative weights of the three different blocks can be changed.
- For the trace plot in figure 35 the relative weights of A, B, and C were 50, 1, and 50.

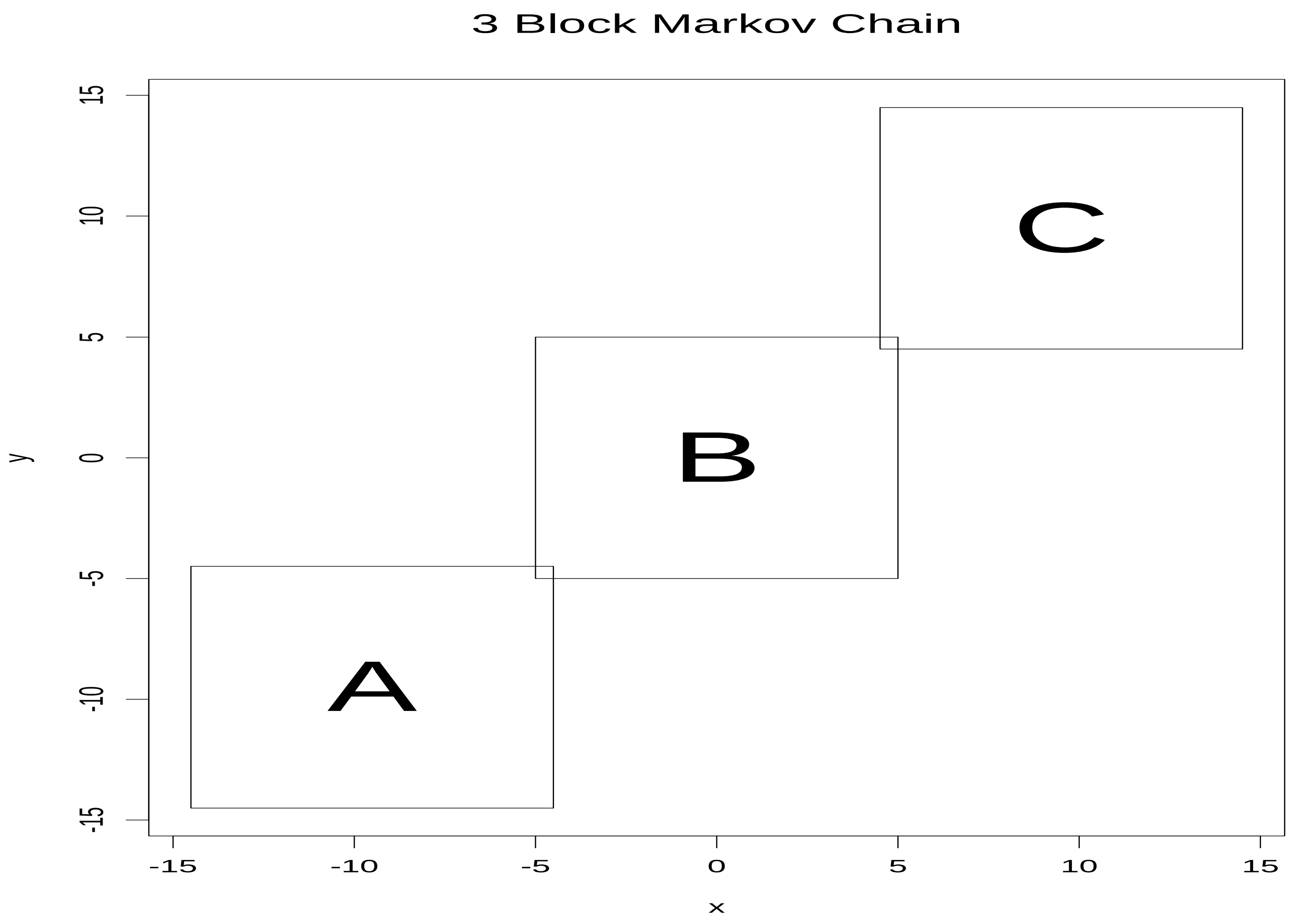

Figure 32: The above plot shows the three blocks which contain the support for the 3 blocks distribution.

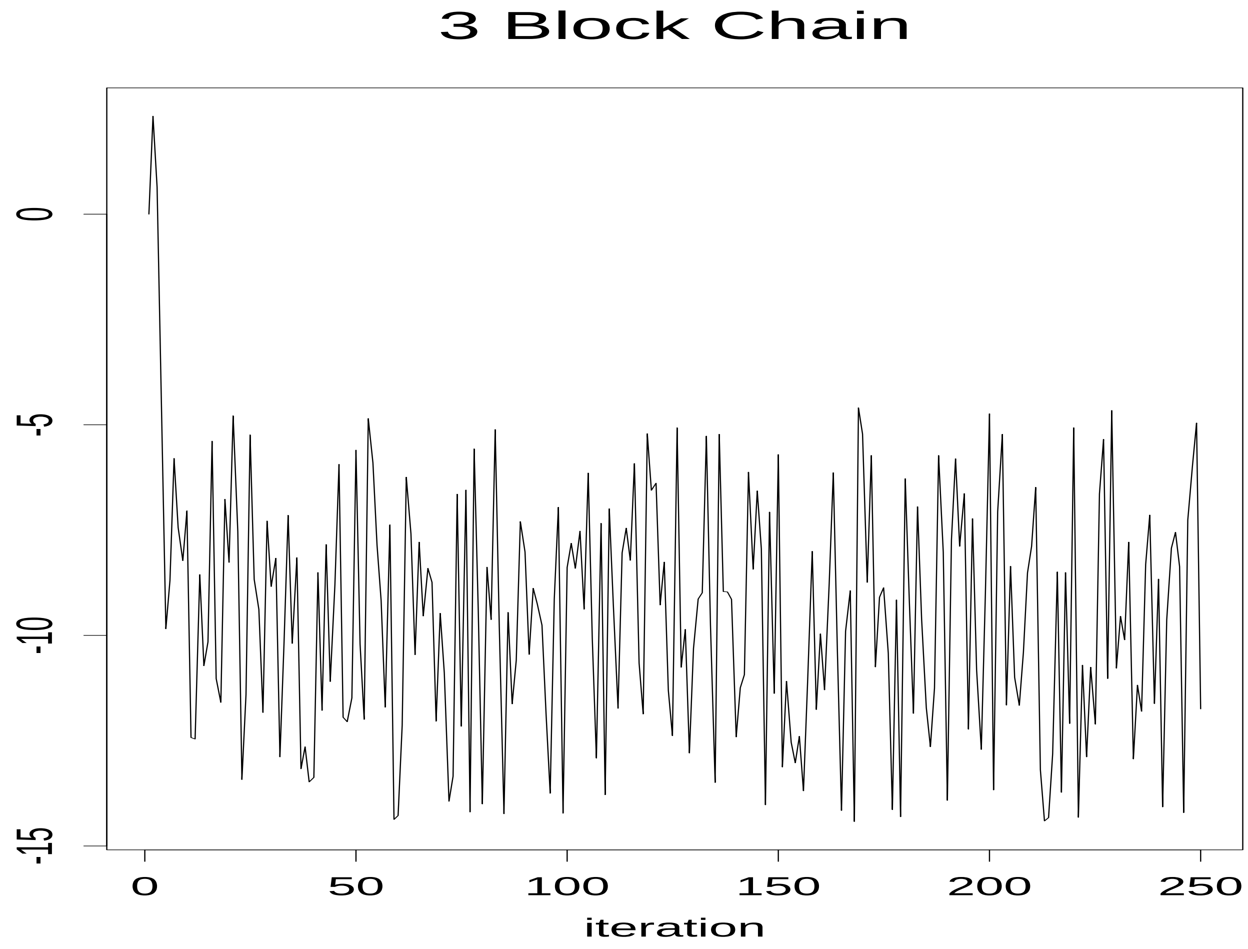

Figure 33: This is the trace plot of X values sampled from the 3 blocks distribution.

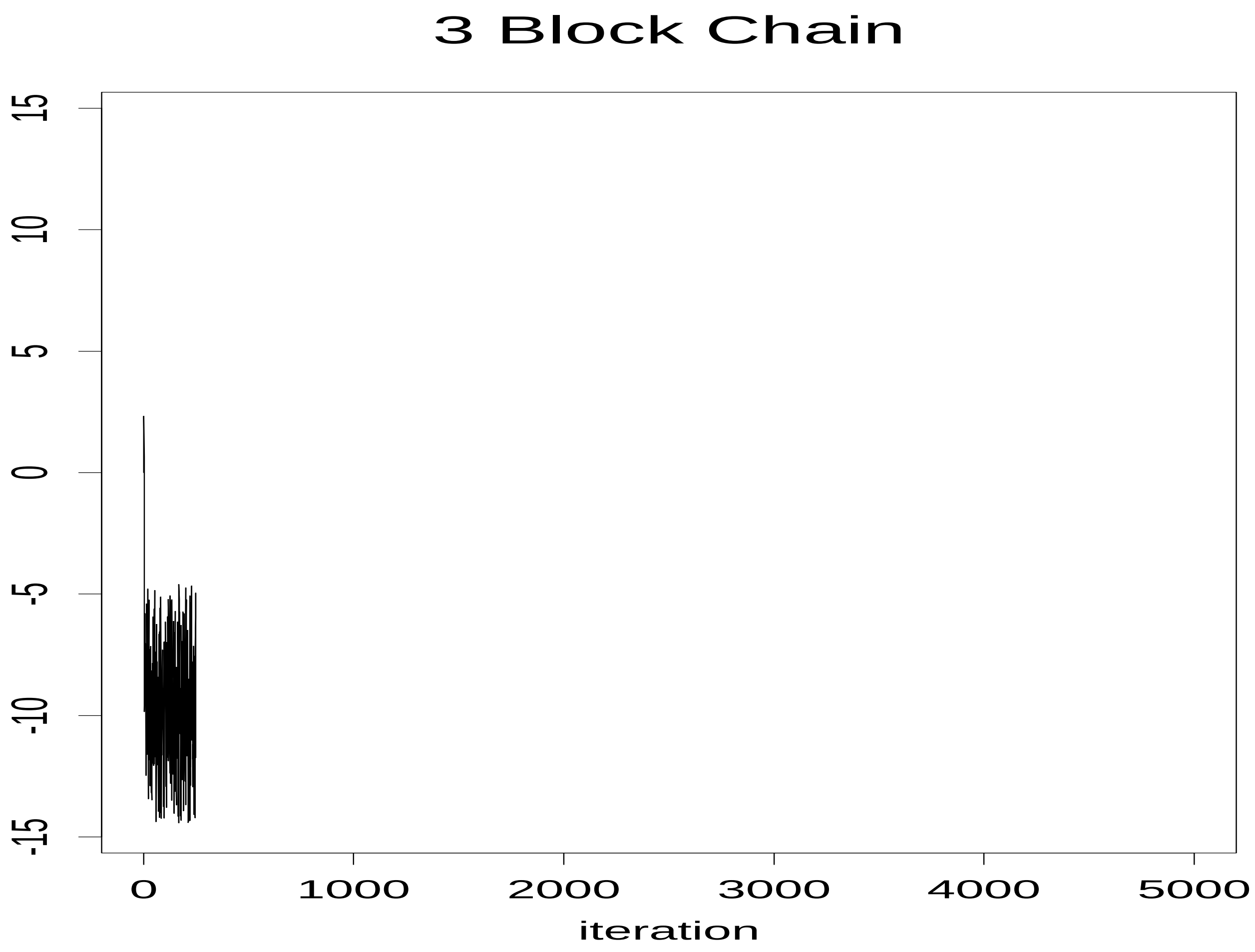

Figure 34: This is the trace plot of X values sampled from the 3 blocks distribution.

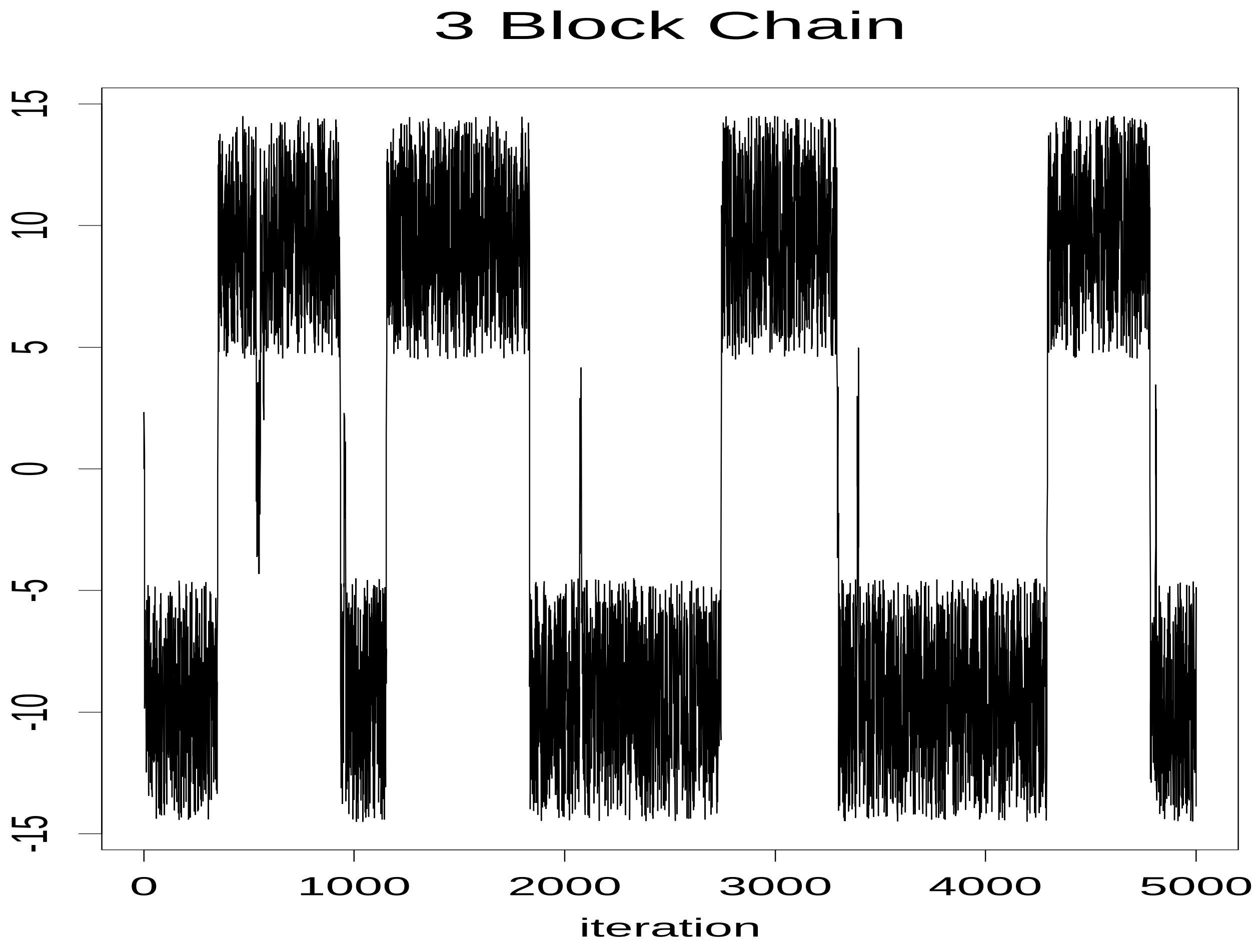

Figure 35: This is the trace plot of X values sampled from the 3 blocks distribution.

## 3 Block Model: Comment

- When we only look at the first 250 iterations, it appears that after a brief “burn-in”, the chain quickly settles down and mixes nicely. The chain quickly cycles through values between -5 and -15.

- However, after iteration number 300 or so, the chain jumps up to the area of block C (values between the range 4.5 and 14.5).

- Basically, there are two large areas of probability with a small bridge formed by block B. When considering the conditions for convergence, one might say that this chain is barely irreducible.

## Witch's Hat

The joint distribution is a mixture of two distribution.

- The first is a uniform on square where both X and Y are between 0 and 20.
- The second is a uniform distribution on the square where X and Y are between 9.995 and 10.005.
- The mixing weight of the small square to the large square is given a value. In this run, the weight is .999.

The figure 36 is the connected plot of sequential values of X which are sampled.

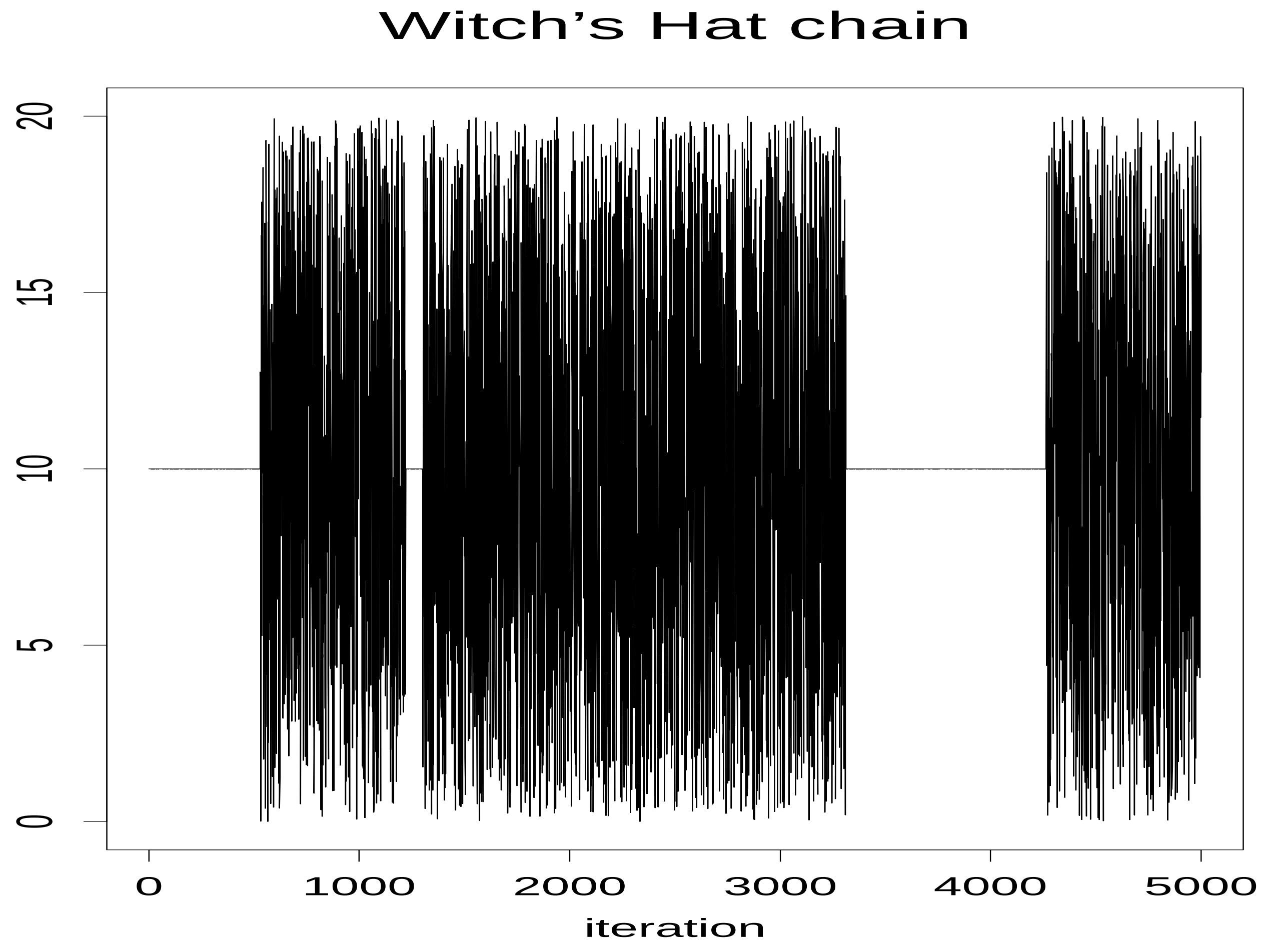

Figure 36: This is the trace plot of X values sampled from the Witch's hat.

## Witch's Hat: Comment

- When a chain has two or more main areas of probability mass, then it is common to refer to these masses as modes.
- the curious thing about the witch's hat distribution is that one of these modes is "inside" the other mode.
- One strategy to find unknown modes when monitoring a chain is to try different starting values around the support for the random variables in the chain. This strategy is sometimes successful against chains like the 3 block chain. However, such a strategy would not work with the witch's hat chain. Here the spike of mass is inside the flatter mode. Also, it is the case that the limiting distribution contains both masses, the "peaked" part of the hat as well as the "brim" of the witch's hat.

## Part 3: Unusual Markov Chain Samplers

Some specific comments:

- First note that after running each of these chains for about 5000 iterations allows one to learn about the limiting distribution. So, there really is no problem, per se, with these chains.
- However, these chains do show that it would be a serious mistake to stop the chains too soon.

## Part 3: Unusual Markov Chain Samplers

(Continued)

- What is worrisome is that one can not tell when to stop by simply looking at the empirical output from the chain. Also, by changing some of the parameters in the model (such as the correlation in the correlated normal problem or the relative weights in the 3 block model), one can control features of the chain like how slow the convergence is or how often the chain shifts from say block A to block C. Therefore, one can not be assured that some arbitrary large number is sufficient to know that nothing unusual is going on.

## Part 3: Unusual Markov Chain Samplers

Some more general comments:

Historically, there has been a bit of a collective manic-depressive personality among Bayesians over MCMC methods.

Historically, in the Bayesian community there was a big emotional high when MCMC methods were first introduced to the community. However, afterwards there was a bit of a let down when Bayesians discovered that there were some limits. Superman seemed to have found his kryptonite...

- When MCMC methods first came out, there was general euphoria. Also, there was the basic attitude that with these new methods Bayesians could now conquer the world. Not only could Bayesians finally work on practical problems like the frequentist, but with these new methods they could easily solve problems which the frequentist found quite difficult.

- Then came the let down stage. There was the concern that we did not know when the chain would converge. In the research effort to find diagnostic procedures for MCMC chains, there seemed to be the hope among some that there existed a perfect procedure which would flag that a chain had not converged. As we can begin to see with the three unusual chains, there is little hope of producing a universal diagnostic procedure based solely on the observed values from the sampler.

- However, the situation is similar to all numeric techniques which are used to explore a mathematical surface. If there are severe ridges in the surface or if there are unknown modes, then one is going to have trouble with the numerical technique.

## Part 3: Unusual Markov Chain Samplers

(Continued)

- As all applied statisticians know, you simply cannot throw numbers into a black box procedure. You need to know your data and the type of model that you are using. The better you know your data and your model, then the better the analysis will be. One does not need Bayesian theory with MCMC algorithms to produce dangerous and misguided results. It can be done quite easily by someone who does not know what they are doing and who are using a simple t-test or trying to interpret a histogram. Even a linear regression can be numerically unstable.

## Part 3: Unusual Markov Chain Samplers

(continued)

- Please note that for many models that a Bayesian might wish to look at, there is usually a long history of research on the shape of the likelihood function for these models (by frequentist statisticians).
- Also, note that the posterior surface which we need to understand is formed by combining the likelihood with the prior.
- So, one can usually use the information about the likelihood as a beginning to understanding the basic shape of the posterior.

## Part 3: Unusual Markov Chain Samplers

(continued)

So, in the next few slides the following is discussed:

- The likelihood bad, but the prior helps. So, the posterior is well behaved.
- The likelihood at first looks bad, but reparameterization helps. This shows how the posterior can be improved by reparameterization.
- A quick look at some of the other scenarios.

## Helpful Priors

When there have been problems with the likelihood surface, it is common to make some kind of correction. Sometimes these corrections involve either adding prior information or adding what looks to a Bayesian like prior information. Some examples include

- In a simple linear regression problem, when the covariates are highly correlated, then the least squares/maximum likelihood estimate is not very stable. One solution to this problem is to use a ridge regression. This solution basically adds an informative prior to the parameters for the coefficients of the covariates.

## Helpful Priors

(continue)

- When doing categorical analysis (such as a logistic regression) one can sometimes have unstable parameters when there are zero counts in some of the table values. Many of the common adjustments are similar to the addition of prior information. (See for example, Anderson, Appl. Stat., 1974, 397-404.)
- Penalized likelihood models add a penalty term to the log-likelihood. Often this penalty function can be interpreted as a type of log prior function.

## Transforming the Likelihood

- Sometimes when the likelihood is not well behaved, the likelihood surface is better behaved after the parameters are transformed. For example, in a linear regression model, the likelihood surface could have a severe ridge if some of the covariates are highly correlated. If the covariates (and associated coefficients) are transformed so that the new covariates are no longer correlated, then the ridge would be removed.
- Since the posterior distribution is a combination of the likelihood and the prior, sometimes the ridge in the likelihood will also appear in the surface of the posterior distribution. Then, it is often the case that if the ridge is removed in the likelihood through a transformation, the ridge in the posterior will also be removed.

## Other Likelihood/Posterior Scenarios

The posterior is a combination of the prior and the likelihood. In using information about the likelihood to provide information on the general shape of the posterior, the following is a list of possible scenarios:

- *The likelihood is well behaved and the posterior is also well behaved.* When the likelihood is well behaved, it is not uncommon that the posterior is also well behaved.
- *The likelihood is not well behaved but the posterior is well behaved.* Sometimes the prior distribution will lead to an improvement. This scenario has already been discussed a few slides back.
- *Neither the likelihood nor the posterior are well behaved, but they are improved with a transformation.* Again, this scenario has just been described.

## Other Likelihood/Posterior Scenarios

(continued)

- *The likelihood is well behaved, but the prior will lead to a posterior which is not well behaved.* Well this can happen. For example suppose the likelihood is a t-distribution on 1 degree of freedom (a Cauchy) and so is the prior. Also assume that they have the same precision (say it is one) and the prior is centered on zero. Now, suppose that we observe the value 10. In this case the posterior will be bimodal with one mode near 0 and the other near 10. If we start our sampler near either 0 or 10, then it might be a while before we find the other mode.

## Other Likelihood/Posterior Scenarios

(continued)

- *Unknown likelihoods.* Because MCMC methods allow one to attempt models that have rarely been attempted before, there is the possibility that one might try models (including likelihood and prior parts) which are not fully understood. In such cases, one needs to be very careful what one is doing.

Big message:

Understand Your Data

Understand Your Model

The model includes the prior and the likelihood.

## Summary of Second Lecture

Basic Markov Chain

- Under weak conditions, when continuously sampling from a Markov chain, the marginal distribution of the samples approach some limiting distribution.
- The average of these sampled values converges to the expected value under the limiting distribution.

Bayesian Analysis with MCMC methods

- With just knowledge of the conditional distribution one can construct a Markov chain to get approximate samples from the posterior distribution.
- In addition, the path averages of these samples can be used to get an estimate of the features of the posterior distribution.

# Summary of Second Lecture

(continued)

Unusual Markov Chains

- When running the chains, one wishes that the chain frequently visits the main areas of probability mass.
- As with most numerical methods, severe ridges and unknown modes could cause problems. However, this means that one can not simply rely on empirical methods to diagnose convergence, but an applied statistician must understand their model.