Gibbs Sampling (and Normal-Normal-Inverse Gamma Inference)

STAT 251, Unit 8

Overview

Review

Inference with the Normal Distribution: μ and σ^2 each unknown

Motivating Example

(Attempt at) Posterior Derivation

Complete Conditionals

Complete Conditionals of posterior distribution for μ, σ^2 in Normal-Normal-InverseGamma setting

Gibbs Sampling

Example of Gibbs Sampling to Conduct Posterior Inference

Blood Pressure Example Repeated with Different Prior (objective prior)

Bivariate Summaries

Review

Reminder for notation

n is the number of data observations.

y_i is the ith observation (that is, the ith data point), for $i = 1, 2, \ldots, n$.

μ and σ^2 are the conditional mean and variance of the distribution for the observations (that is, the population mean and population variance of all data in the *population*, if μ and σ^2 were known)

$\bar{y}$ is the sample average of the (observed) data values.

$y_i|\{\mu, \sigma^2\} \overset{iid}{\sim} N(\mu, \sigma^2)$ means that the n observations are conditionally *i*ndependent and *i*dentically *d*istributed with the *Normal*(μ, σ^2) distribution.

Review: Structure of Units 6–8

Suppose $y_i|(\mu, \sigma^2) \overset{iid}{\sim} N(\mu, \sigma^2)$.

There are two parameters to consider. We worked through the following progression:

- Unit 6: Inference with μ unknown but σ^2 known.
- Unit 7: Inference with σ^2 unknown but μ known.
- Unit 8: Inference with μ and σ^2 unknown

While the present unit, Unit 8, is the only realistic scenario among the three, we will rely on results established in Units 6 and 7.

Review: Key Unit 6 Result

If we have assumed that $y_i|\{\mu, \sigma^2\} \overset{iid}{\sim} N(\mu, \sigma^2)$, then

- if σ^2 is known and the prior distribution for μ is Normal(mean=m, variance=v), then the posterior distribution for μ is Normal(mean= $m^\star = (n\bar{y}/\sigma^2 + m/v)/(n/\sigma^2 + 1/v)$, variance=$v^\star = 1/(n/\sigma^2 + 1/v)$).

Review: Key Unit 7 Result

If we have assumed that $y_i|\{\mu, \sigma^2\} \overset{iid}{\sim} N(\mu, \sigma^2)$, then

- if μ is known and the prior distribution for σ^2 is InverseGamma(a,b), then the posterior distribution for σ^2 is InverseGamma($a^\star = a + n/2$, $b^\star = b + (1/2)\sum_{i=1}^{n}(y_i - \mu)^2$).

Inference with the Normal Distribution: μ and σ^2 each unknown

Prior Specification in the Realistic Setting

Suppose $y_i|\{\mu, \sigma^2\} \overset{iid}{\sim} N(\mu, \sigma^2)$. Typically, μ and σ^2 are unknown.

With two unknown parameters, the prior distribution is two-dimensional. That is, the prior is now $\pi(\mu, \sigma^2)$.

One convenient choice for prior specification (which we will use in this class) is to assume that our prior beliefs about μ and σ^2 are independent of each other. That is, we will assume $\pi(\mu, \sigma^2) = \pi(\mu)\pi(\sigma^2)$.

Technical note

On the previous slide, I wrote that we will assume $\pi(\mu, \sigma^2) = \pi(\mu)\pi(\sigma^2)$.

You might be confused by the use of π here to mean either the prior distribution for μ only, the prior distribution for σ^2 only, or the joint prior distribution for μ and σ^2.

Although we could be more explicit and write $\pi_{\mu,\sigma^2}(\mu, \sigma^2) = \pi_\mu(\mu)\pi_{\sigma^2}(\sigma^2)$, it is apparent by looking at the function input which distribution we are talking about.

For instance, $\pi_\mu(\mu)$ is the prior distribution for μ, but it should likewise be understood that $\pi(\mu)$ is the prior distribution for μ.

Choice of priors

As in Unit 6, we will assume $\mu \sim N(m, v)$. As in Unit 7, we will assume $\sigma^2 \sim IG(a, b)$.

With these priors on μ and σ^2, and assuming independence in the prior beliefs,

$$\begin{aligned}\pi(\mu, \sigma^2) &= \pi(\mu)\pi(\sigma^2) \\ &= \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{1}{2v}(\mu - m)^2\right] \\ &\quad \times \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)} \exp\left[-\frac{b}{\sigma^2}\right] \mathbb{1}_{\sigma^2>0}\end{aligned}$$

Motivating Example

Gibbs Sampling in Action

We'll consider a demonstration of Gibbs sampling that relies on data from Pima Indian women

Specifically, we'll focus on diastolic blood pressure for women without a diabetes diagnosis.

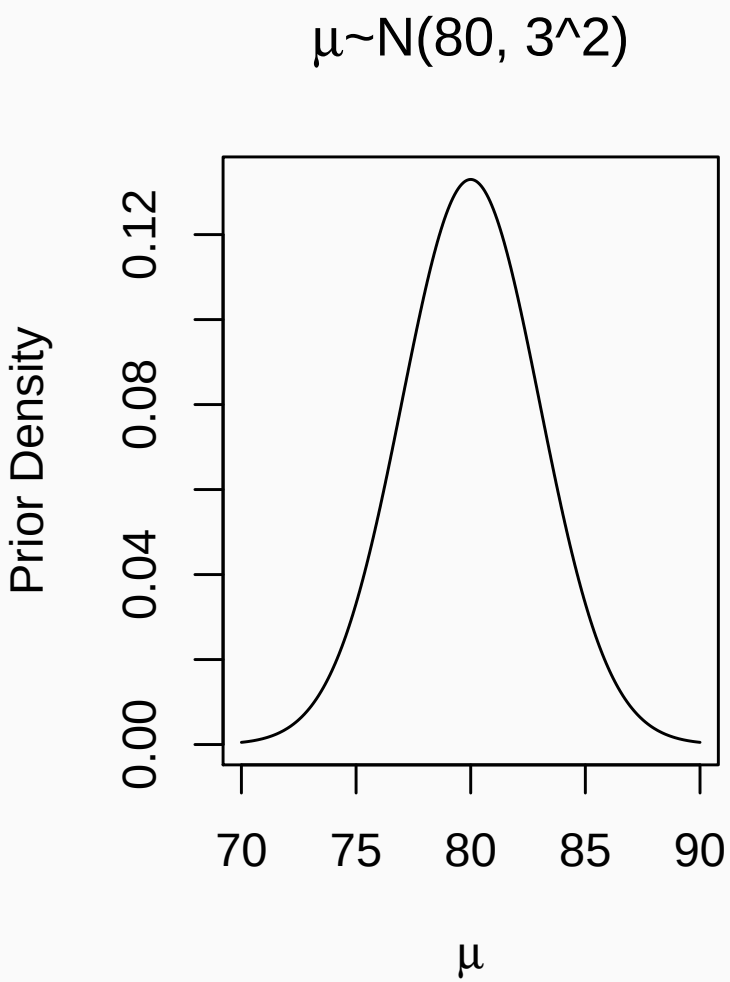
Model Specification

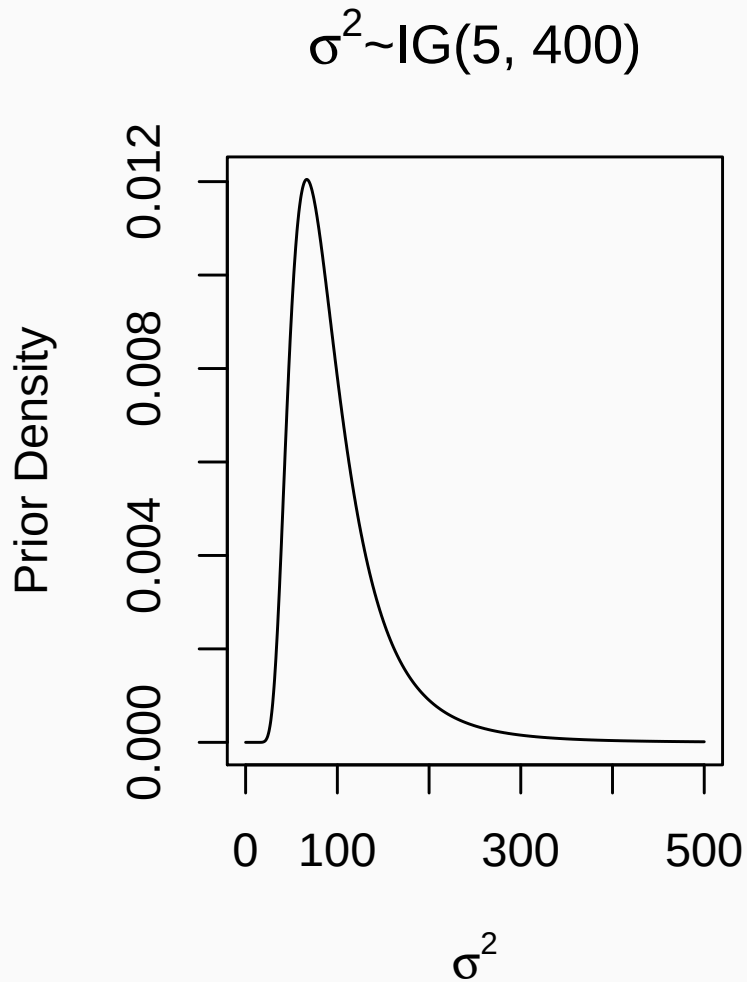
We will assume the observed diastolic blood pressures are conditionally *iid* $N(\mu, \sigma^2)$.

What do you you know about typical diastolic blood pressures in healthy adults?

A subjective prior on (μ, σ^2)

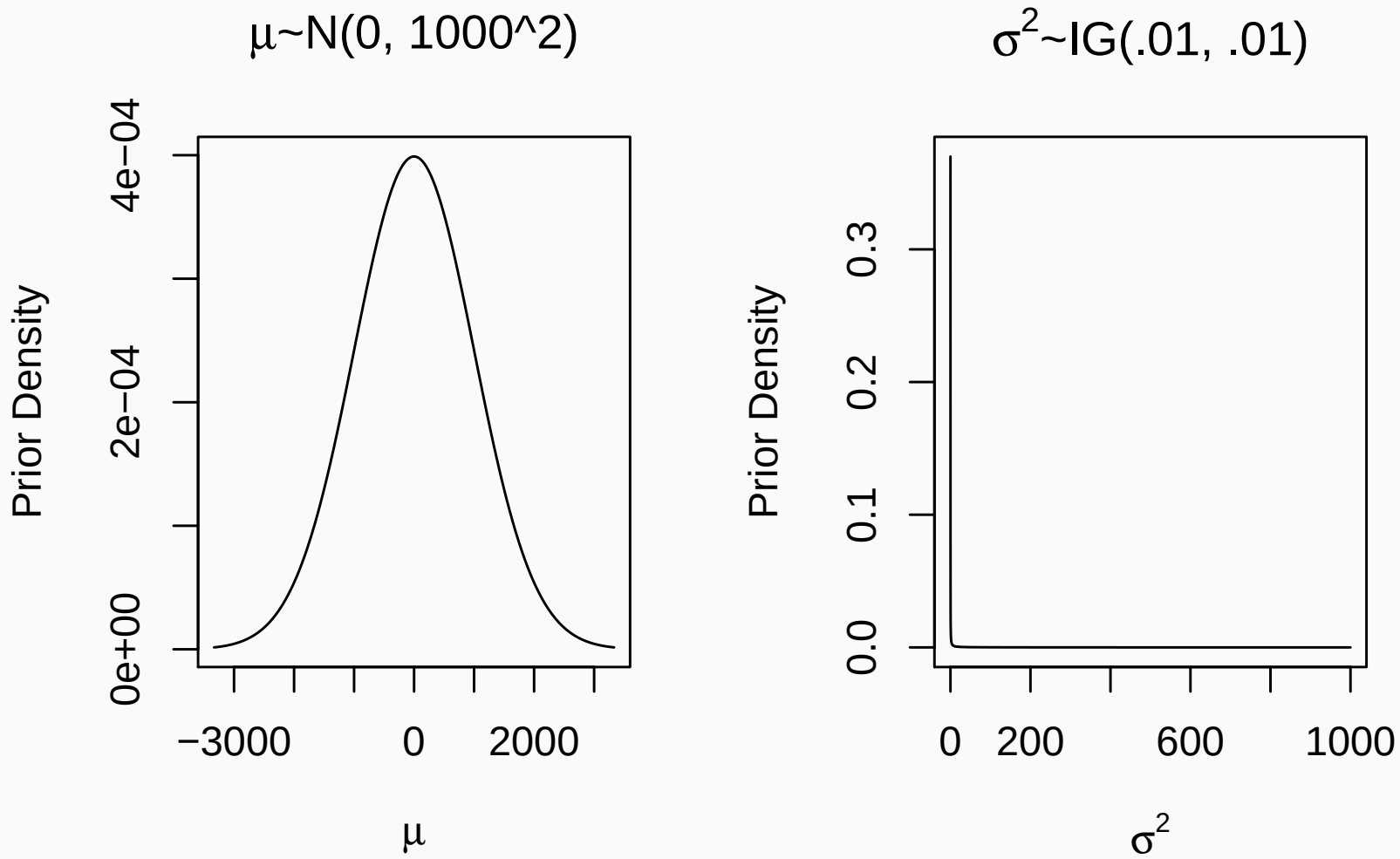
$\mu \sim N(80, 3^2)$, $\sigma^2 \sim IG(5, 400)$, with μ and σ^2 assumed *a priori* to be independent of each other.





An objective prior on (μ, σ^2)

$\mu \sim N(0, 1000^2)$, $\sigma^2 \sim IG(.01, .01)$, with μ and σ^2 assumed *a priori* to be independent of each other.



Comparing the informativeness of the priors

```
qnorm(c(.025, .975), 80, 3)
```

```
## [1] 74.12011 85.87989
```

```
qinvgamma(c(.025, .975),5, 400)
```

```
## [1]  39.05644 246.38334
```

Comparing the informativeness of the priors

```
qnorm(c(.025, .975), 80, 3)
```

```
## [1] 74.12011 85.87989
```

```
qinvgamma(c(.025, .975),5, 400)
```

```
## [1]  39.05644 246.38334
```

```
qnorm(c(.025, .975), 0, 1000)
```

```
## [1] -1959.964  1959.964
```

```
qinvgamma(c(.025, .975),.01, .01)
```

```
## [1]  2.121433e-01 2.838743e+158
```

Subjective Bayes vs. Objective Bayes

The priors for μ have the form $Normal(m, v)$, and the priors for the variances have the form $InvGamma(a, b)$.

In this setting, when v is very large and when a and b are very small, the posterior inference on the means and variances is especially driven by the likelihood.

This is why one set of priors was referred to as *objective*, whereas the other set of priors was referred to as *subjective*.

(Attempt at) Posterior Derivation

Assumptions in the NNIG setting

In what I refer to hereafter as the normal-normal-inverse-gamma setting (or NNIG), the assumptions are:

- $y_i|(\mu, \sigma^2) \overset{iid}{\sim} N(\mu, \sigma^2), \quad$ for $i = 1, 2, \ldots, n$
- $\mu \sim N(m, v)$
- $\sigma^2 \sim IG(shape = a, rate = b)$
- μ and σ^2 are *a priori* independent of each other.

These assumptions are used for the remainder of the Unit 8 notes!

Posterior distribution in NNIG setting

The posterior distribution of (μ, σ^2), namely $\pi(\mu, \sigma^2|\{y_1, y_2, \ldots, y_n\})$, is proportional to the product of the prior distribution times the likelihood.

$$\begin{aligned}
\pi(\mu, \sigma^2|\{y_1, \ldots, y_n\}) &\propto \pi(\mu, \sigma^2) f(\{y_1, \ldots, y_n\}|\mu, \sigma^2) \\
&= \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{1}{2v}(\mu - m)^2\right] \\
&\quad \times \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)} \exp\left[-\frac{b}{\sigma^2}\right] \mathbb{1}_{\sigma^2>0} \\
&\quad \times \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right]\right]
\end{aligned}$$

Unfortunately, this is not a joint distribution that we readily recognize. The posterior does not have the same form as the prior.

Posterior Distribution (cont.)

Recall the prior was the product of a normal distribution for μ and an inverse-gamma distribution for σ^2.

However, the posterior is NOT the product of a normal distribution for μ and an inverse-gamma distribution for σ^2.

In particular, even though we assumed *a priori* the mean and the variance are independent of each other, *a posteriori* the mean and variance are NOT independent of each other.

Complete Conditionals

Complete Conditionals

One possible approach to inference when the joint posterior distribution is not recognized is to consider the *complete conditional distributions*, aka, *complete conditionals*.

Definition: Suppose that in a particular setting, there are K random variables, say, $X_1, X_2, \ldots, X_K$. In general, this may be a mix of response variables (y's) and parameters (θ's). They have a joint distribution, which could be written as $f(X_1, X_2, \ldots, X_K)$.

The **complete conditional distribution** of X_k is the distribution of X_k given all of the other X's. That is, the complete conditional of X_k is

$$f(X_k | \{X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_K\})$$

Technical Note

The complete conditional distribution of X_k is the distribution of X_k given all of the other X's. That is,

$$f(X_k|\{X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_K\})$$

One alternate notation you may encounter in other classes for the complete conditional is:

$$f(X_k|\mathbf{X}_{-\mathbf{k}}),$$

where $\mathbf{X}_{-\mathbf{k}}$ is the collection of all X's except for the kth one, X_k.

Example

Suppose that there are only three random variables: X_1, X_2, X_3.

- The complete conditional distribution for X_1 is $f(X_1|\{X_2, X_3\})$
- The complete conditional distribution for X_2 is $f(X_2|\{X_1, X_3\})$
- The complete conditional distribution for X_3 is $f(X_3|\{X_1, X_2\})$

Complete Conditionals of posterior distribution for μ, σ^2 in Normal-Normal-InverseGamma setting

Complete Conditionals of posterior distribution for μ, σ^2 in Normal-Normal-InverseGamma setting

From the posterior distribution, the complete conditional of μ would be denoted as $\pi(\mu|\sigma^2, y_1, \ldots, y_n)$.

From the posterior distribution, the complete conditional of σ^2 is $\pi(\sigma^2|\mu, y_1, \ldots, y_n)$.

Okay, we have the concept of a complete conditional distribution. Now, let's consider what the complete conditionals are in our specific application setting.

Complete conditional of posterior distribution for μ

The trick is to note that because we are conditioning on all random variables save μ, anything that does not involve μ is treated as a constant for the complete conditional distribution of μ.

$$
\begin{aligned}
\pi(\mu|\sigma^2, \{y_1, \ldots, y_n\}) &\propto \pi(\mu, \sigma^2|\{y_1, \ldots, y_n\}) \\
&\propto \pi(\mu, \sigma^2) f(\{y_1, \ldots, y_n\}|\mu, \sigma^2) \\
&= \pi(\mu)\pi(\sigma^2) f(\{y_1, \ldots, y_n\}|\mu, \sigma^2) \\
&\propto \pi(\mu) f(\{y_1, \ldots, y_n\}|\mu, \sigma^2) \\
&= \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{1}{2v}(\mu - m)^2\right] \\
&\quad \times \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right]\right]
\end{aligned}
$$

Anything familiar about the last portion?

Complete conditional of posterior distribution for μ (cont.)

Refer to the previous slide. The last part of the chain of proportional equations might look familiar. *This is exactly the same as our derivation of the posterior distribution of μ when σ^2 was assumed to be known* (see Unit 6 notes).

- We had a normal prior for μ, a normal likelihood, and we assumed that σ^2 was given (because it was known).
- Moral of the story: Deriving the complete conditional for μ is equivalent to deriving the posterior distribution for μ if σ^2 were known.

In the Normal-Normal-Inverse-Gamma setting, the complete conditional distribution for μ is the Normal$(m^\star, v^\star)$ distribution, with $m^\star$ and $v^\star$ as in the Unit 6 notes.

Complete Conditional of posterior distribution for σ^2

$$
\begin{aligned}
\pi(\sigma^2|\mu, \{y_1, \ldots, y_n\}) &\propto \pi(\mu, \sigma^2|\{y_1, \ldots, y_n\}) \\
&\propto \pi(\mu, \sigma^2) f(\{y_1, \ldots, y_n\}|\mu, \sigma^2) \\
&= \pi(\mu)\pi(\sigma^2) f(\{y_1, \ldots, y_n\}|\mu, \sigma^2) \\
&\propto \pi(\sigma^2) f(\{y_1, \ldots, y_n\}|\mu, \sigma^2) \\
&= \frac{b^a}{\Gamma(a)} (\sigma^2)^{-(a+1)} \exp\left[-\frac{b}{\sigma^2}\right] \mathbb{1}_{\sigma^2>0} \\
&\qquad \times \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right]\right]
\end{aligned}
$$

Anything familiar about the last portion?

Complete Conditional of posterior distribution for σ^2 (cont.)

The last part of the chain of proportional equations should look familiar. This is exactly the same as your derivation of the posterior distribution of σ^2 when μ was assumed to be known (see Unit 7 notes).

- We had an Inverse Gamma prior for σ^2, a normal likelihood, and we assumed that μ was given (because it was known).
- Moral of the story: The derivation of the complete conditional for σ^2 is equivalent to the derivation of the posterior distribution of σ^2 if μ were known.

In the Normal-Normal-Inverse-Gamma setting, the complete conditional distribution for σ^2 is the IG($a^\star, b^\star$) distribution, with $a^\star$ and $b^\star$ as in the Unit 7 notes.

Gibbs Sampling

Significance of complete conditionals

Why do we care about the complete conditionals?

Part of the answer is because they can be used to sample from a joint distribution in an iterative fashion.

What is **Gibbs sampling**? Iterative sampling from the complete conditionals (given current values of the quantities conditioned on) to (in the limit) obtain posterior draws from the joint posterior distribution. This technique shows up often in Bayesian computation to sample from a multivariate posterior distribution.

Overview of Gibbs Sampling in the Normal-normal-inverse-gamma setting

Goal: obtain a set of (μ, σ^2) pairs that come from the posterior distribution of (μ, σ^2)

- S0 Choose initial values for μ and σ^2; these are denoted as $\mu^{(0)}$ and $(\sigma^2)^{(0)}$
- S1 For $j = 1, \ldots, J$, where J is a LARGE number,
  - S1a Generate $\mu^{(j)}$ from the complete conditional distribution of μ assuming $\sigma^2 = (\sigma^2)^{(j-1)}$
  - S1b Generate $(\sigma^2)^{(j)}$ from the complete conditional distribution of σ^2 assuming $\mu = \mu^{(j)}$).
- S2 Assess the length of the "burn-in period" and discard this initial portion of the sequence.

Notation: the superscript (j) notation means the jth value in the sequence, NOT the jth power.

Step 0 Details, applied to NNIG setting

S0 Choose initial values $\mu^{(0)}$ and $(\sigma^2)^{(0)}$ in the support of the posterior distribution for (μ, σ^2). That is, make sure that $(\sigma^2)^{(0)} > 0$.

- Ideally, these initial values would be close to the posterior mode for μ and σ^2, respectively.
- Often reasonable to start at $\mu^{(0)} = \bar{y}$, $(\sigma^2)^{(0)} = s^2$, where $\bar{y}$ and s^2 denote the sample mean and sample variance of the observed y_i's.

Step 1/1a Details, applied to NNIG setting

S1 For $j = 1, \ldots, J$, where J is a LARGE number,

S1a Generate $\mu^{(j)}$ from the complete conditional distribution of μ, using the **current value of the variance** (i.e., $(\sigma^2)^{(j-1)}$) to finish specifying the distribution.

Step 1a Details (cont), applied to NNIG setting

From Unit 6, we saw that:

$$\mu|\{y_1, \ldots, y_n, \sigma^2\}$$
$$\sim N\left(\frac{\frac{n\bar{y}}{(\sigma^2)} + \frac{m}{v}}{\frac{n}{(\sigma^2)} + \frac{1}{v}}, \quad \frac{1}{\frac{n}{(\sigma^2)} + \frac{1}{v}}\right)$$

To generate the next value of μ, i.e., $\mu^{(j)}$, using the currently available value of the variance σ^2, i.e. $\sigma^{2(j-1)}$,

Step 1a Details (cont), applied to NNIG setting

From Unit 6, we saw that:

$$\mu|\{y_1, \ldots, y_n, \sigma^2\} \sim N\left(\frac{\frac{n\bar{y}}{(\sigma^2)} + \frac{m}{v}}{\frac{n}{(\sigma^2)} + \frac{1}{v}}, \quad \frac{1}{\frac{n}{(\sigma^2)} + \frac{1}{v}}\right)$$

To generate the next value of μ, i.e., $\mu^{(j)}$, using the currently available value of the variance σ^2, i.e. $\sigma^{2(j-1)}$, note that

$$\mu^{(j)}|\{y_1, \ldots, y_n, \sigma^2 = (\sigma^2)^{(j-1)}\} \sim N\left(\frac{\frac{n\bar{y}}{(\sigma^2)^{(j-1)}} + \frac{m}{v}}{\frac{n}{(\sigma^2)^{(j-1)}} + \frac{1}{v}}, \quad \frac{1}{\frac{n}{(\sigma^2)^{(j-1)}} + \frac{1}{v}}\right)$$

Step 1a, restated in NNIG context

Step 1a: Sample the next value of μ, $\mu^{(j)}$, by drawing one value from the
$Normal(mean = [n\bar{y}/(\sigma^2)^{(j-1)} + m/v]/[n/(\sigma^2)^{(j-1)} + 1/v]$,
$variance = 1/[n/(\sigma^2)^{(j-1)} + 1/v])$ distribution.

Step 1b Details, applied to NNIG setting

S1 cont ...

S1b Generate $(\sigma^2)^{(j)}$ from the complete conditional distribution of σ^2, using the **current value of the mean** (i.e., $\mu^{(j)}$) to finish specifying the distribution.

Step 1a details, restated in NNIG context

From Unit 7,

$$\sigma^2|\{y_1,\ldots,y_n,\mu\} \sim \mathit{InverseGamma}\left(a+\frac{n}{2}, b+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$

To generate the next value of σ^2, i.e., $(\sigma^2)^{(j)}$, using the currently available value of the mean μ, i.e. $\mu^{(j)}$,

Step 1a details, restated in NNIG context

From Unit 7,

$$\sigma^2|\{y_1,\ldots,y_n,\mu\} \sim \mathit{InverseGamma}\left(a+\frac{n}{2}, b+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$

To generate the next value of σ^2, i.e., $(\sigma^2)^{(j)}$, using the currently available value of the mean μ, i.e. $\mu^{(j)}$, note that

$$\begin{aligned}(\sigma^2)^{(j)}&|\{y_1,\ldots,y_n,\mu^{(j)}\}\\ &\sim \mathit{InverseGamma}\left(a+\frac{n}{2}, b+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu^{(j)})^2\right)\end{aligned}$$

Step 1b, restated in NNIG context

Step 1b: Sample the next value of σ^2, $(\sigma^2)^{(j)}$, by drawing one value from the InverseGamma($a^\star = a + n/2$, $b^\star = b + (1/2)\sum_{i=1}^{n}(y_i - \mu^{(j)})^2$) distribution.

Step 2, restated in NNIG context

S2 After a sufficiently large number of draws, the $(\mu^{(j)}, (\sigma^2)^{(j)})$ values form a (correlated) sequence of draws that behave as though they had been jointly sampled directly from the joint posterior distribution. Assess how large this so-called "burn-in period" is by, say, examining trace plots.

If we have a sample of values from the posterior distribution, inference can be made on the parameters by considering the sample values (Monte Carlo).

Example of Gibbs Sampling to Conduct Posterior Inference

Gibbs Sampling in Action

Recall the motivating example: posterior inference on μ and σ^2, using diastolic blood pressure data from non-diabetic Pima Indian women.

We assumed the observed diastolic blood pressures are conditionally *iid* $N(\mu, \sigma^2)$.

We'll first consider inference resulting from the *subjective* prior:

$\mu \sim N(80, 3^2)$, $\sigma^2 \sim IG(5, 400)$, with μ and σ^2 assumed *a priori* to be independent of each other.

What are the following values?

$m =$; $v =$; $a =$; $b =$.

R Functions to implement Gibbs sampling

We will define two functions: one to sample a single value from the complete conditional distribution of μ; and another to sample from the complete conditional distribution of σ^2 (in the NNIG setting)

Pseudocode

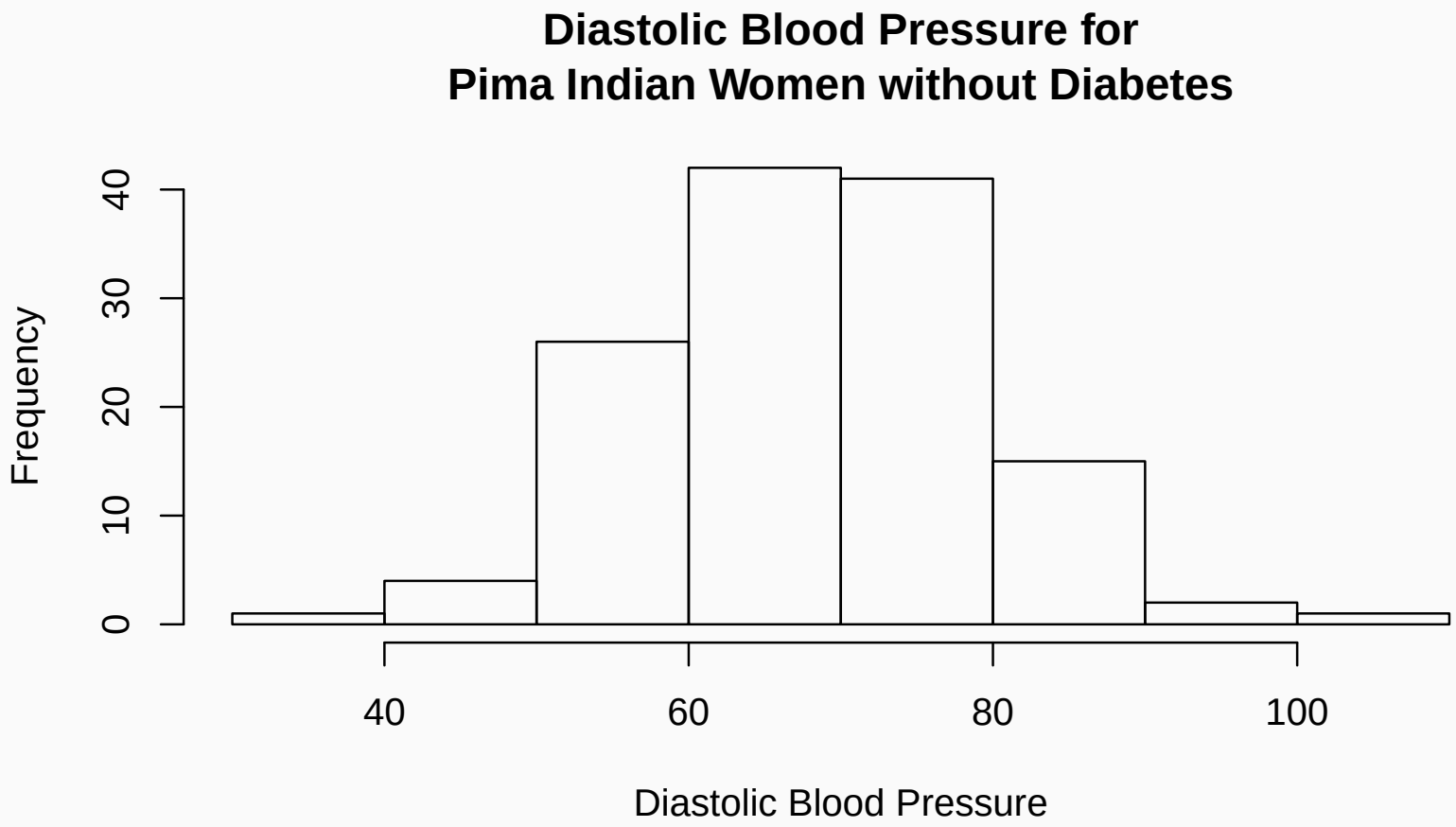
```
### the following function is to update mu given the
### variance sigma2 for conditionally iid
### normally distributed data (ys)
### assuming a Normal (m, v) prior for mu

update.mu <- function(sigma2,ys,m,v){
  ## determine n from how many elements in ys vector
  ## determine v* (i.e., vstar)
  ## Note vstar = 1/(n/sigma2 + 1/v)
  ## determine m* (i.e., mstar). Note that
  ## mstar = vstar * (n*ybar/sigma2 + m/v)
  ## return one sampled value from
  ## the Normal(mstar, vstar) distribution
}
```

Pseudocode

```
### the following function is to update sigma2 given mu
### for conditionally iid normally distributed data (ys)
### and an inverse gamma (a,b) prior for sigma2

update.var <- function(mu,ys,a,b){
  ## determine n from how many elements in ys vector
  ## Find a* (i.e., astar) = a + n/2
  ## Find b* (i.e., bstar) = b + 0.5*sum((ys-mu)^2)
  ## Return one sampled value from the IG(astar, bstar)
  ## distribution. This can be done in two equivalent
  ## ways: the rinvgamma function of the invgamma
  ## package, or inverting a sampled value from the
  ## Gamma(astar, bstar) distribution.
}
```

Data Prep in R

```r
library(MASS)                    ### load the MASS package
?Pima.tr                         ### learn about the data

### Extract the diastolic blood pressure (bp) for
### Pima Indian women without diabetes
nondiabetic <- Pima.tr$bp[Pima.tr$type=="No"]

par(mfrow=c(1,1))
hist(nondiabetic, xlab="Diastolic Blood Pressure",
     main=paste("Diastolic Blood Pressure for",
                "Pima Indian Women without Diabetes",
                sep="\n"))
```


Diastolic Blood Pressure for
Pima Indian Women without Diabetes
Frequency
0
10
20
30
40
40
60
80
100
Diastolic Blood Pressure

Preliminary Notes for Gibbs Sampler

We need to:

0. Initialize the sequences of μ and σ^2 draws
1. For $j = 1, 2, \ldots, J$, with J large, say, 1000):
   1a. Update each μ from its complete conditional, given current σ^2
   1b. Update each σ^2 from its complete conditional, given current μ
2. Diagnose at which point the (μ, σ^2) pairs of draws have converged to their desired distribution (the posterior), and make inference on the post-convergence pairs.

Step 0

```
set.seed(243767)  # for reproducibility

### Create vectors that will store draws.
### Notice each has same length
mu.nd <- rep(NA, 1001)
sigma2.nd <- rep(NA, 1001)

##  Step 0: set the initial values   ##
## Reasonable initial value for mu_nd
mu.nd[1] <- mean(nondiabetic)
## Reasonable initial value for sigma2_nd
sigma2.nd[1] <- var(nondiabetic)
```

Step 1

```
#  Step 1: Sequential sampling from comp. conditionals #
for (j in 2:1001){
  # Step 1a: get the next mu value in the sequence
  mu.nd[j] <- update.mu(sigma2.nd[j-1],
                        nondiabetic, 80, 9)

  # Step 1b: get the next sigma2 value in the sequence
  sigma2.nd[j] <- update.var(mu.nd[j],
                             nondiabetic, 5, 400)
}
```

Step 1 (selected output)

To see the first three pairs in the (μ, σ^2) sequence of draws:

```
cbind(mu.nd, sigma2.nd)[1:3,]
```

```
##          mu.nd sigma2.nd
## [1,] 69.54545  122.8452
## [2,] 70.38306  124.9963
## [3,] 70.81245  134.5026
```

To see the last two pairs:

```
cbind(mu.nd, sigma2.nd)[1000:1001,]
```

```
##          mu.nd sigma2.nd
## [1,] 68.72226  130.5261
## [2,] 72.00950  112.7755
```

Step 2: Assessing Convergence and Mixing

Note: it is not typical to call this a step in the Gibbs sampling algorithm. But I want to emphasize this needs to be performed before making inference.

Two key features to consider in regards to the "chain" of (μ, σ^2) draws:

- Convergence of the chain
- Mixing of the chain.

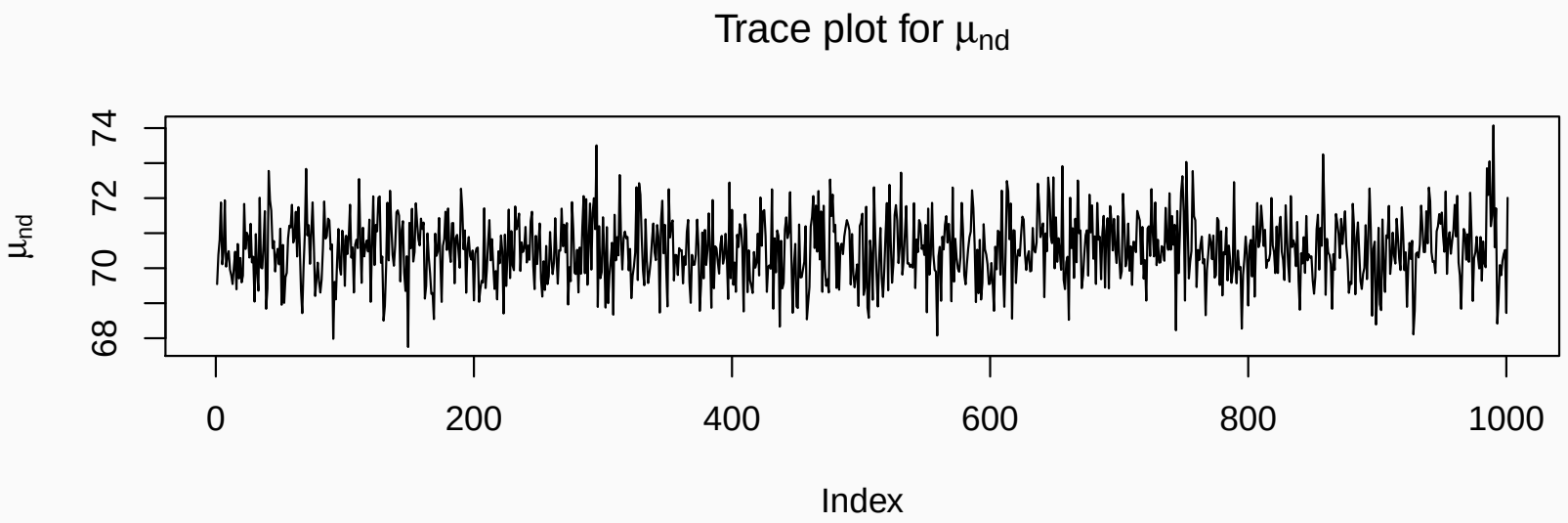
What is meant by the Gibbs sampling chain converging?

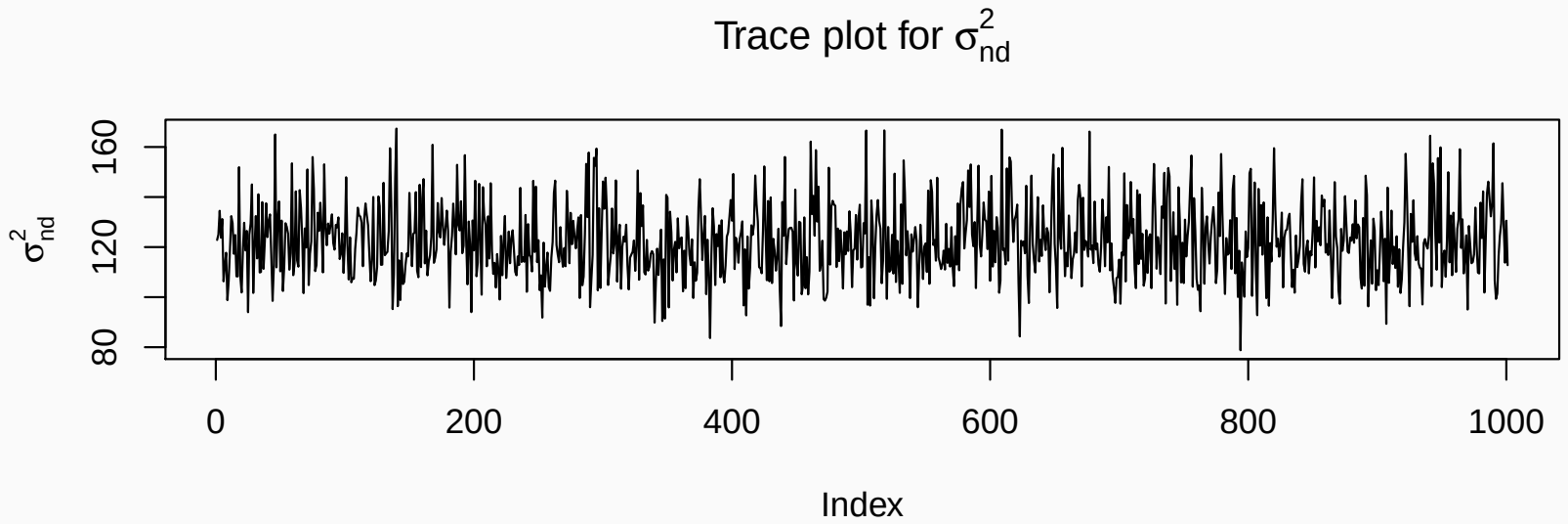
The chain (i.e., the sequence of sampled values) is said to have **converged** when the sampled values become consistent with being drawn from the correct *stationary distribution*. In particular, the mean and variance must have *stopped systematically changing* across iterations.

Common to use a trace plot to assess convergence.

A **trace plot** is a line plot that shows the sequence of sampled values, in the order they were sampled. (It is a special case of a time series plot.)

Step 2 (R Code to create these trace plots on next slide)

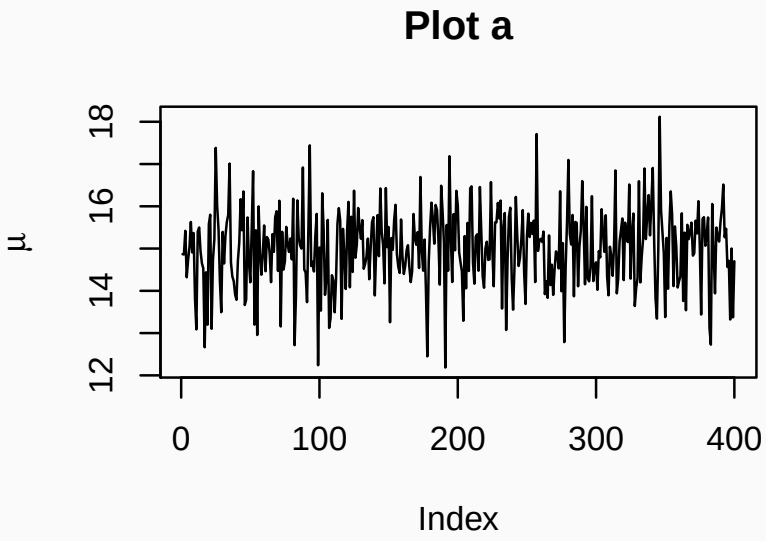


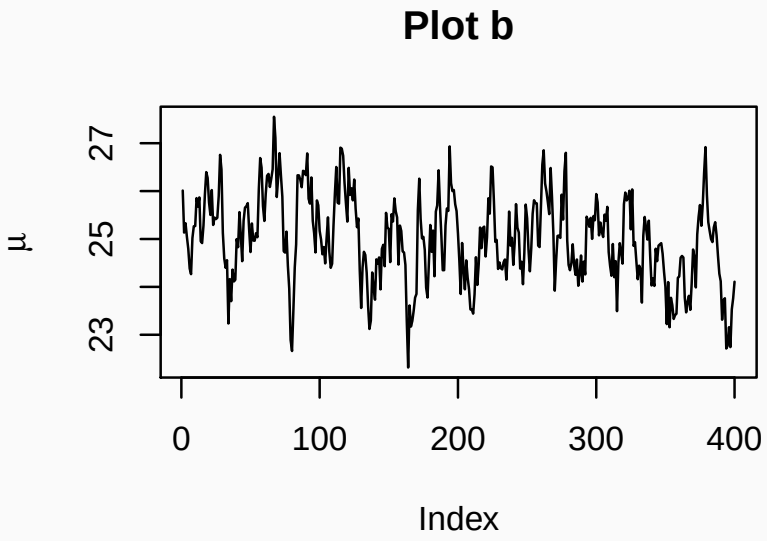


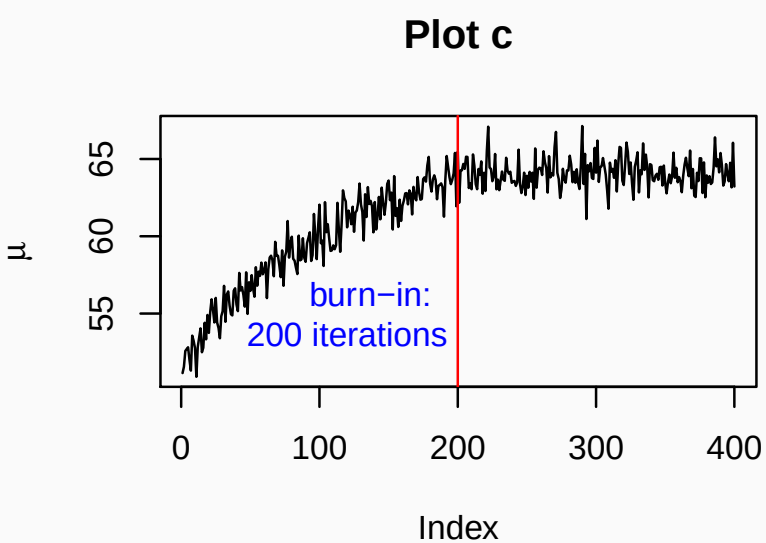
Step 2

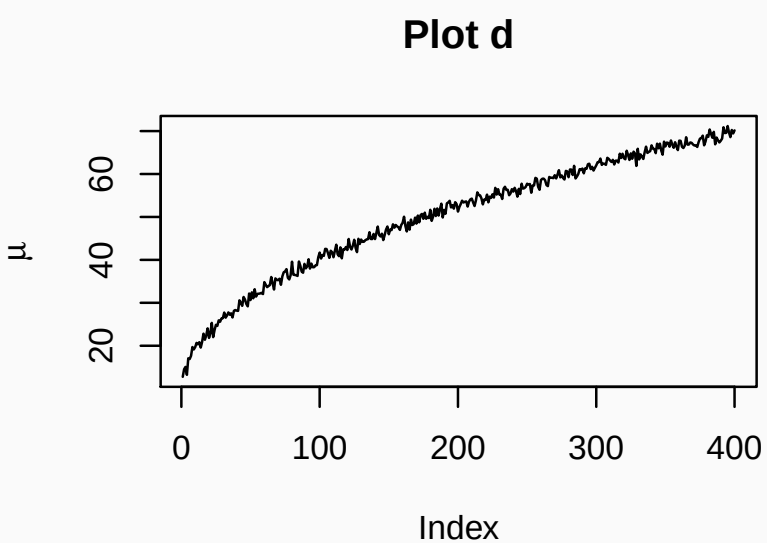
```
##############################################
##   Step 2: Assessing Convergence          ##
##############################################

## Trace plot of the mu_nd draws
plot(mu.nd, type="l", ylab=expression(mu[nd]),
   main=expression(paste("Trace plot for ",mu[nd])))

## Trace plot of the sigma2_nd draws
plot(sigma2.nd, type="l", ylab="",
     main=expression(paste(
   "Trace plot for ",sigma[nd]^2)))
mtext(expression(sigma[nd]^2), side=2, line=2.5 )
```

Examples of Immediate Convergence (a and b), Delayed Convergence (c), and Nonconvergence (d)









Interpretation of trace plots

Although trace plots have shortcomings in that they don't provide *proof* of convergence, they might demonstrate a lack of convergence.

If the chain has not converged, we should run the algorithm much longer and/or choose better initial values for the parameters.

Because the burn-in iterations will be removed for inferential purposes, make sure the chain is long enough to have many post-burn-in iterations.

What is meant by how well the Gibbs sampling chain mixes?

The chain (i.e., the sequence of sampled values) is said to **mix** well when the draws are able to quickly move around the posterior distribution's space. This happens when the depedendence between consecutive draws quickly diminishes.

A chain that **mixes** slowly has persistent dependence in the sequence and takes a long time to explore the posterior. *Ceteris paribus* (all else being equal), we prefer chains that mix quickly because the resultant Monte Carlo inferences are more precise.

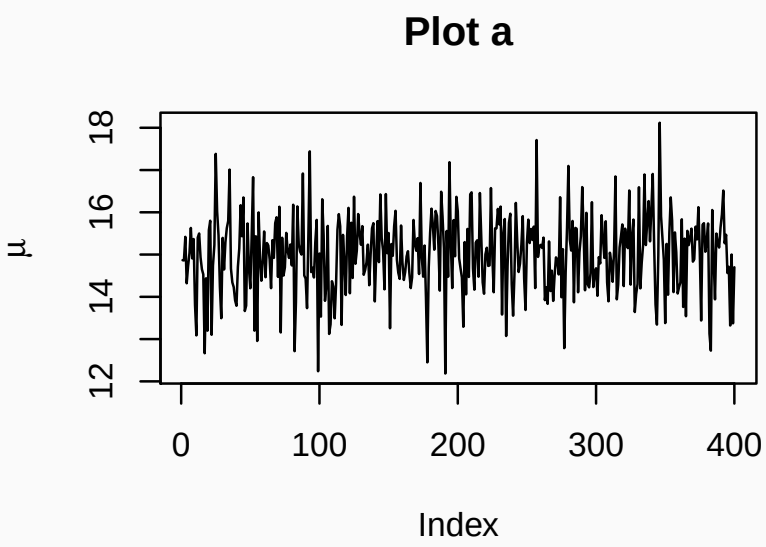
Assessing mixing with the Autocorrelation Function

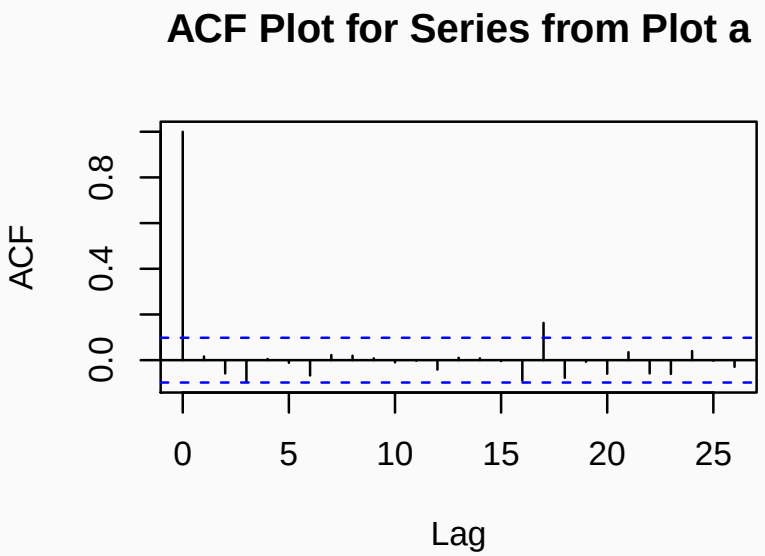
Common to use an autocorrelation function (acf) plot to assess mixing. In doing so, **first remove the burn-in period**—the pre-convergence values in the sequence.

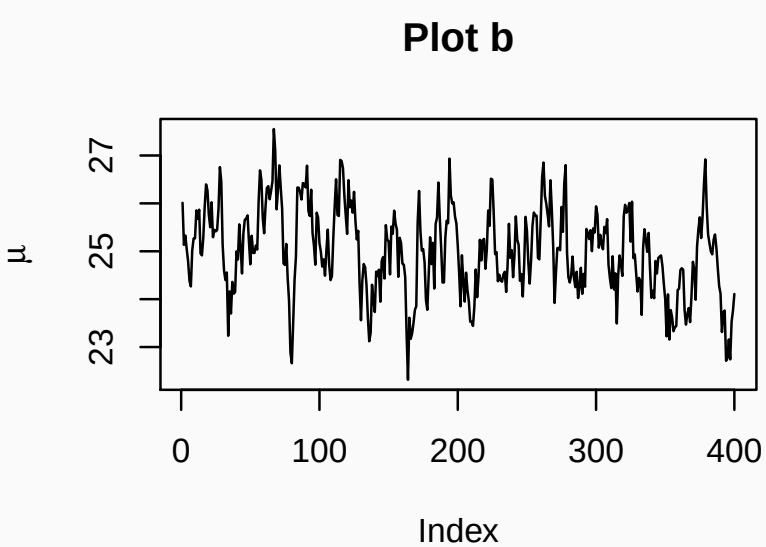
> An **ACF plot** shows the estimated autocorrelation in a sequence of values as a function of the lag (the lag measures how far apart the observations are in the sequence).

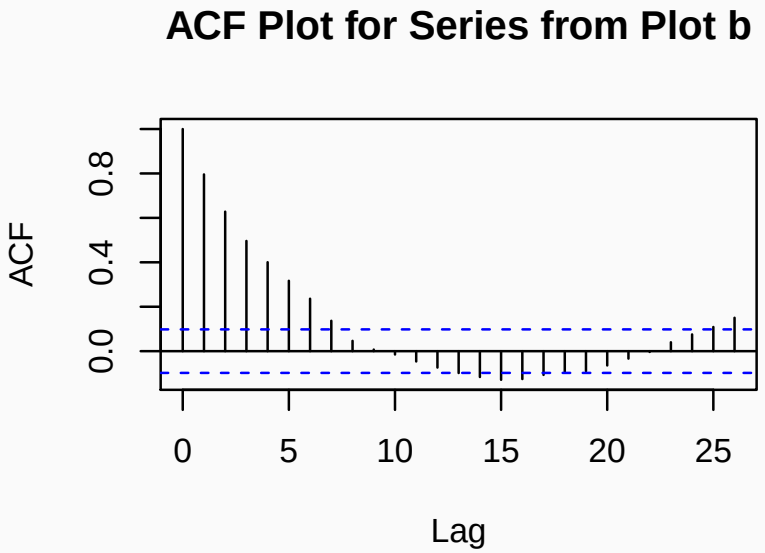
The R function is simply *acf(vals)*, where *vals* is the sequence of (post-burn-in) values from the chain.

For converged values, indicators of mixing. (a) has excellent mixing, (b) mixes more slowly

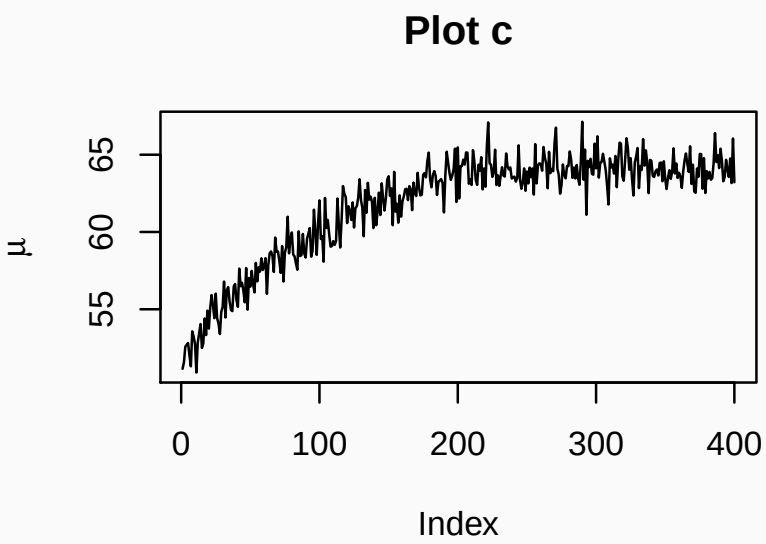


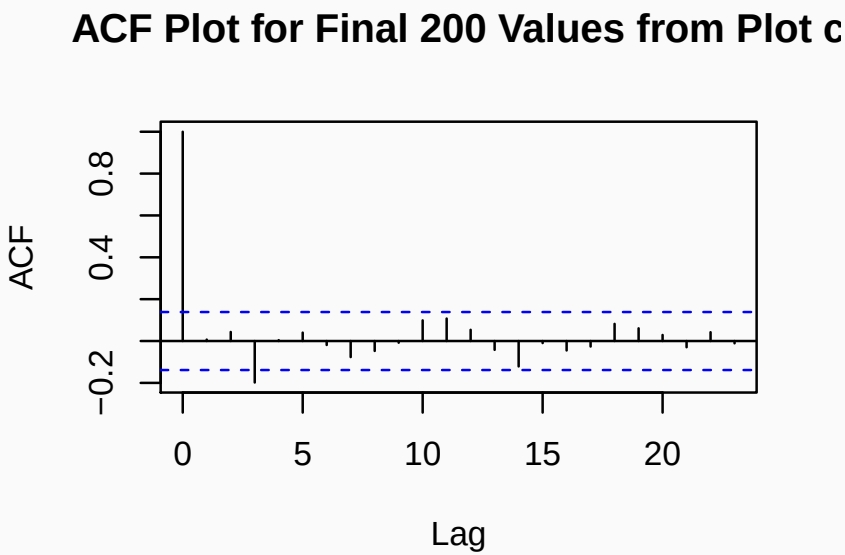


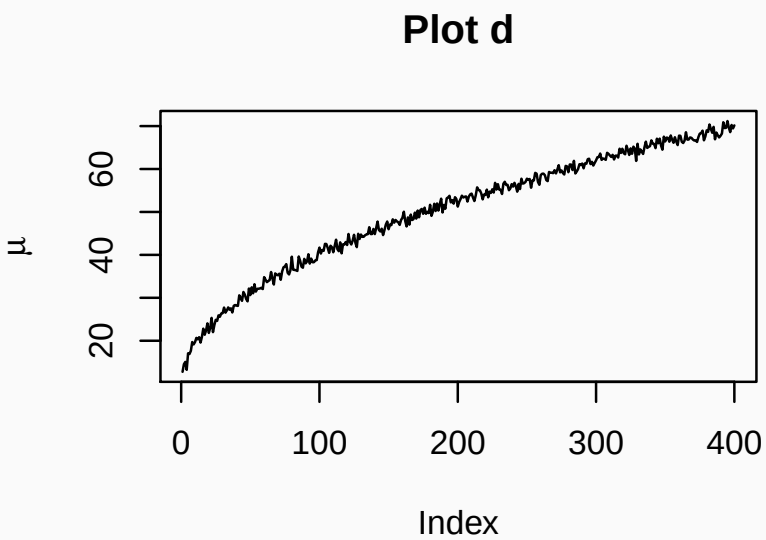




For converged values, indicators of mixing. (c) has excellent post-burn-in mixing; (d) not assessed because not converged







I don't create the ACF plot
for the chain depicted in Plot d:
sequence still has not converged!

Assessing Mixing via acf function:

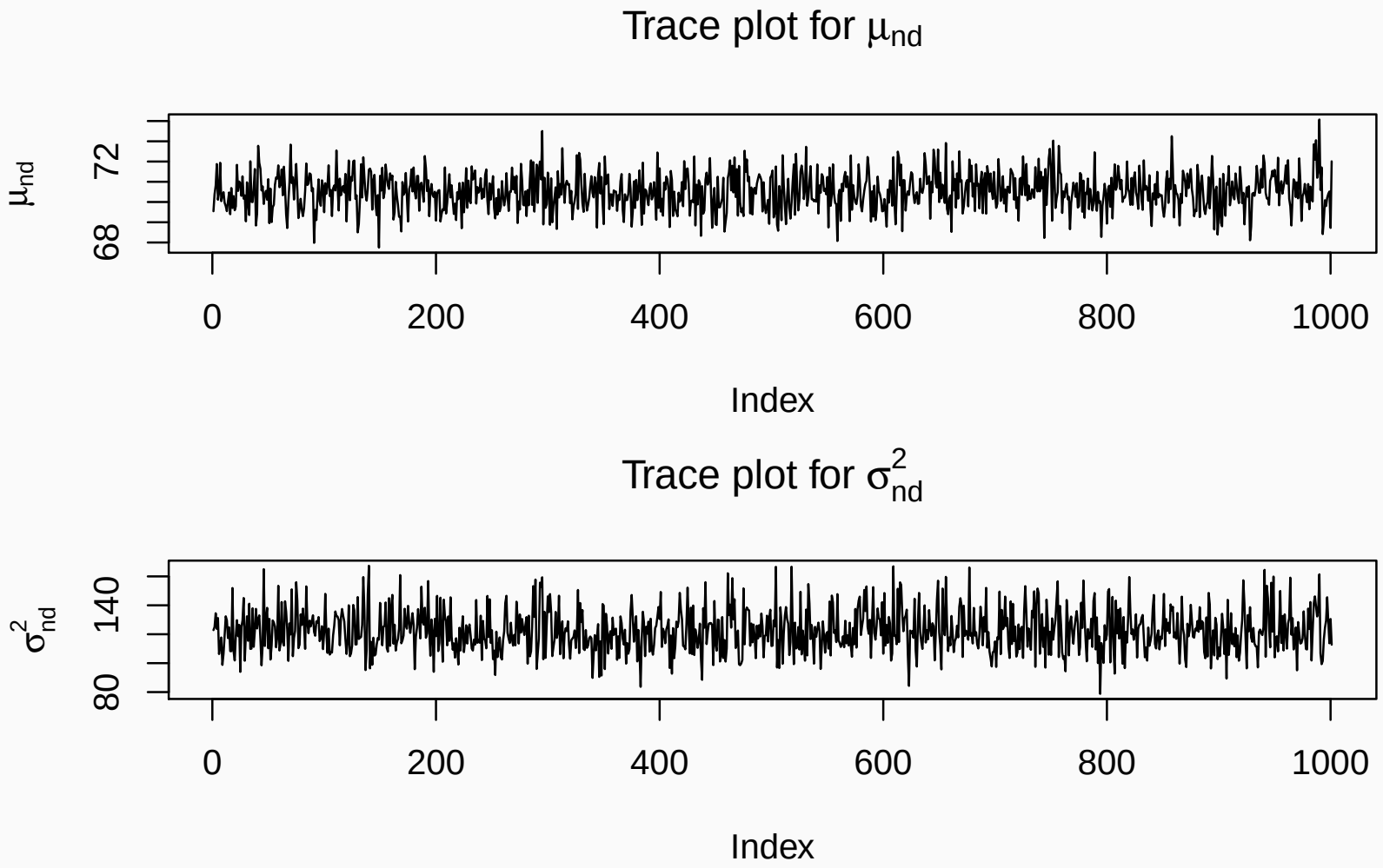
The key: consider the sequence of draws in the Gibbs sampler as constituting a time series. If the autocorrelation is high, consecutive values tend to be so similar that a subsequent value does not contain much "new" information. Rather, it tends to be like the more recent values.

This is bad news in a Gibbs sampler, because it increases the number of draws you need in the chain to get good Monte Carlo estimates.

> High ACF (at any lag ≥ 1) can be indicative of a chain not mixing well. If ACF is close to 1 at, say, 10 or more lags, the chain's mixing is particularly bad.

Blood Pressure Example (cont)

Err on the side of caution: I will remove the first 40 values in each chain as the burn-in period before computing posterior summaries.

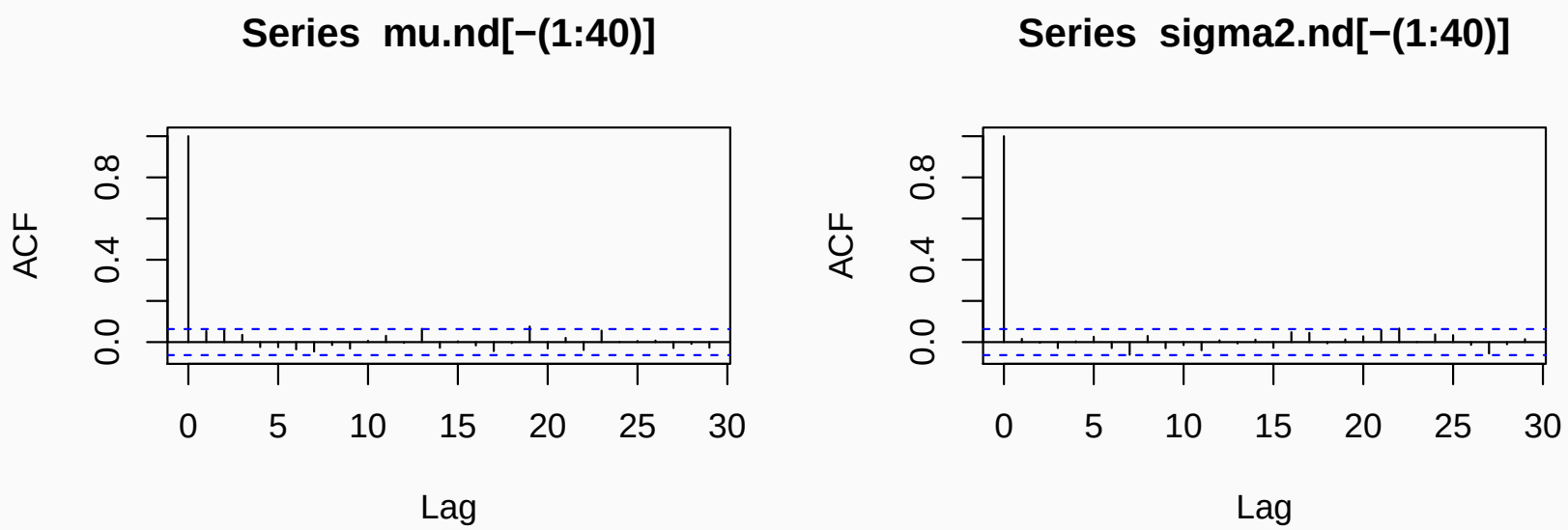


Assessing Mixing for Blood Pressure Example

Estimated autocorrelations, looking at draws "lag" units apart.

Note I have removed a conservative burn-in number of values (40).

```
par(mfrow=c(1,2))
acf(mu.nd[-(1:40)])
acf(sigma2.nd[-(1:40)])
```



Only very weak autocorrelations in the sequence of draws—the chain *appears* to mix well.

Recap

We have many (960) post-burn-in values for μ and σ^2 that were sampled from the posterior distribution via Gibbs sampling. The Gibbs sampler appeared to mix well (relatively low autocorrelation).

We can now make Monte-Carlo based posterior inference. E.g.,

```
### Some posterior summaries of mu_d:
mean(mu.nd[-(1:40)])

## [1] 70.53998

quantile(mu.nd[-(1:40)], c(.05, .95))

##       5%      95%
## 68.96943 72.08989
```

Recap (cont)

Monte-Carlo based inference on $\sigma^2|\{y_1, y_2, \ldots, y_n\}$:

```
### Some posterior summaries of sigma_nd^2:
# Estimated Posterior mean of mu
mean(sigma2.nd[-(1:40)])
```

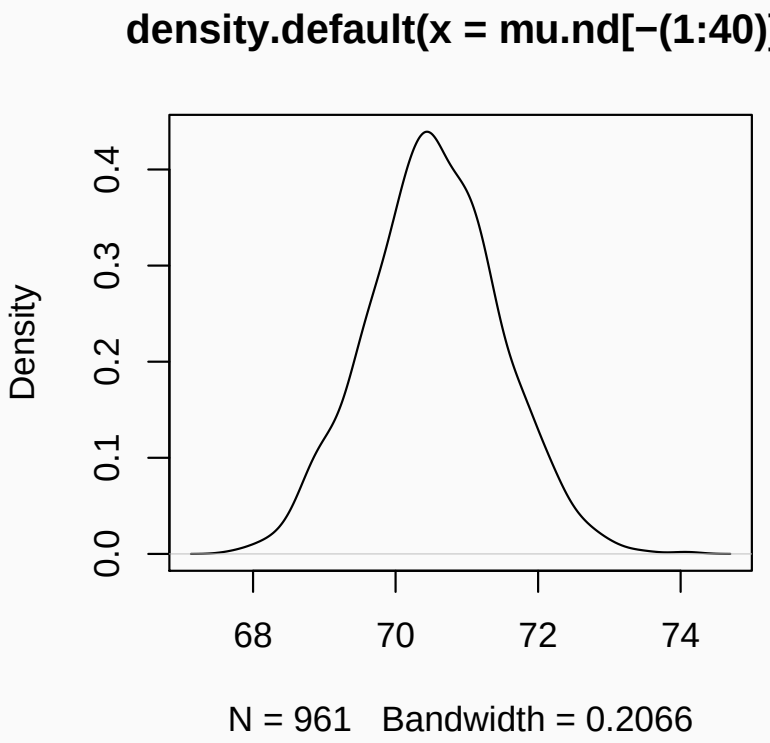
```
## [1] 122.5549
```

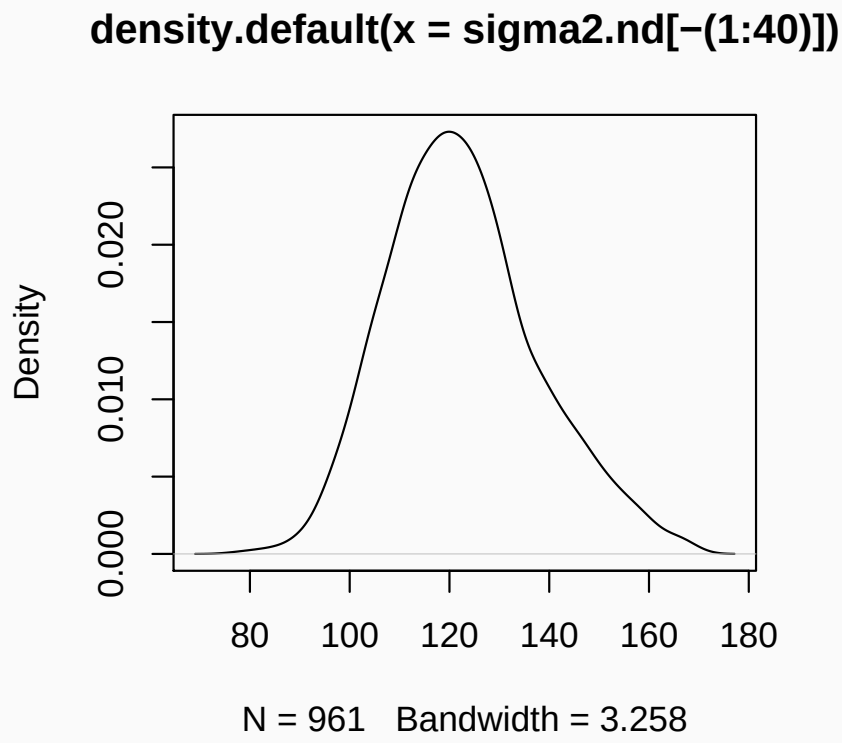
```
# Estimated 90% Credible interval of sigma_nd^2
quantile(sigma2.nd[-(1:40)], c(.05, .95))
```

```
##         5%        95%
##  99.79127 150.60609
```

Estimated Posterior Densities

```
par(mfrow=c(1,2))
plot(density(mu.nd[-(1:40)]))
plot(density(sigma2.nd[-(1:40)]))
```





Blood Pressure Example Repeated with Different Prior (objective prior)

Gibbs sampling with the objective prior

Only difference is that we have different values for m, v, a and b in the prior specification.

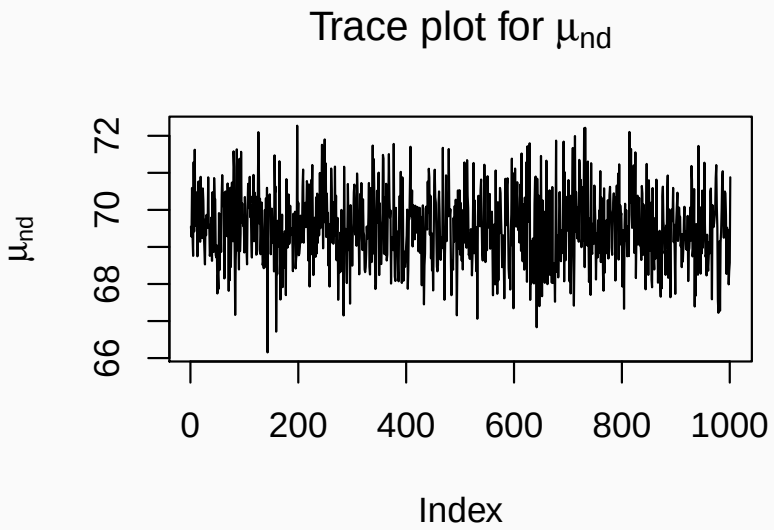
The subjective prior I chose had $m = 80, v = 9, a = 5, b = 400$.
The objective prior I chose had
$m = 0, v = 1000000, a = .01, b = .01$.

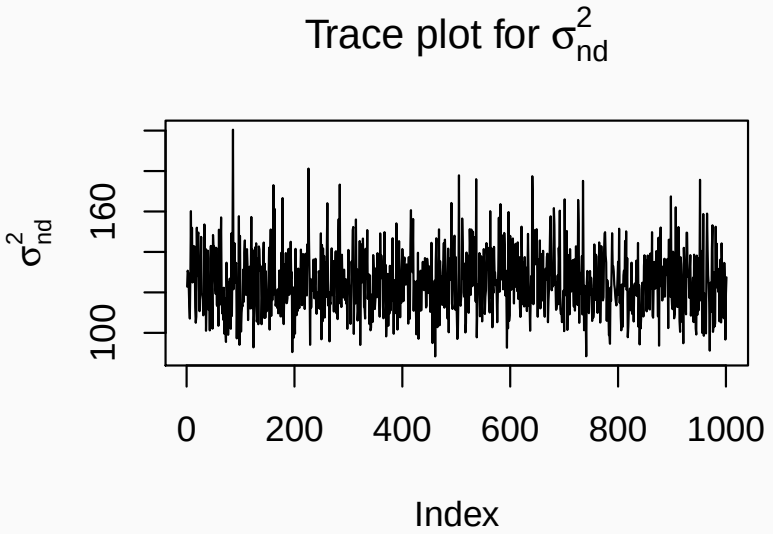
Steps 0, 1 with the objective prior

```
set.seed(23431)  # for reproducibility

mu.nd <- rep(NA, 1001)  # to store the mu draws
sigma2.nd <- rep(NA, 1001) # to store the sigma2 draws

mu.nd[1] <- mean(nondiabetic)  #initialize
sigma2.nd[1] <- var(nondiabetic) # initialize

for (j in 2:1001){
  mu.nd[j] <- update.mu(sigma2.nd[j-1],
                        nondiabetic, m=0, v=1e6)
  sigma2.nd[j] <- update.var(mu.nd[j],
                             nondiabetic, a=.01, b=.01)
}
```

Step 2 with the objective prior

```
par(mfrow=c(1,2))
plot(mu.nd, type="l", ylab=expression(mu[nd]),
   main=expression(paste("Trace plot for ",mu[nd])))
plot(sigma2.nd, type="l", ylab="",
     main=expression(paste(
   "Trace plot for ",sigma[nd]^2)))
mtext(expression(sigma[nd]^2), side=2, line=2.5 )
```



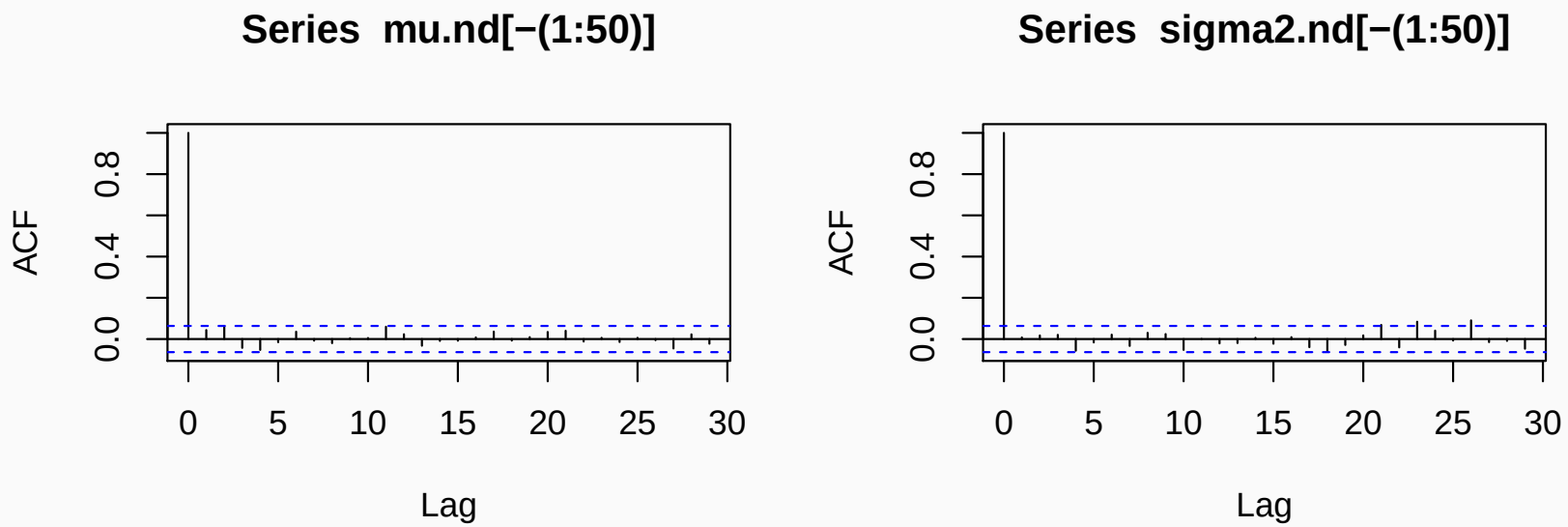


Step 2 with the objective prior (cont)

Estimated autocorrelations, looking at draws "lag" units apart.

Note I have removed a conservative burn-in number of values (50).

```
par(mfrow=c(1,2))
acf(mu.nd[-(1:50)])
acf(sigma2.nd[-(1:50)])
```



The chain *appears* to mix well. Proceed with posterior inference (ignoring first 50 pairs in chain).

Bivariate Summaries

Bivariate Summaries

In Units 4–7, there was only one unknown parameter on which to make inference.

With two (or more) parameters, we can consider additional summaries.

E.g., posterior correlation between μ and σ^2. (First remove the burn-in period)

```
cor(mu.nd[-(1:50)], sigma2.nd[-(1:50)])
```

```
## [1] 0.03430953
```

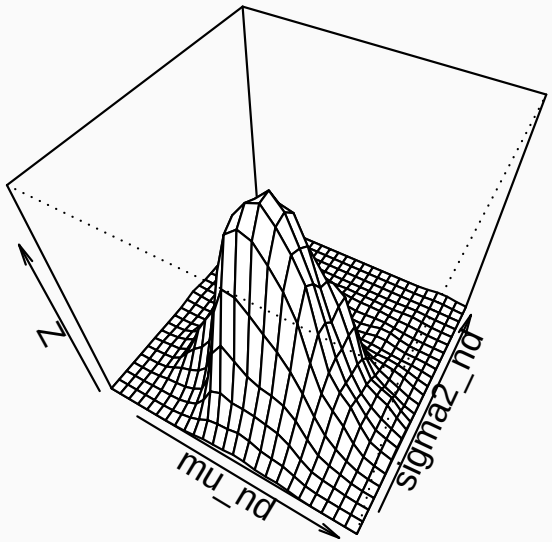
Estimated corelation is quite small, indicating the posterior beliefs about μ and σ^2 are not strongly (linearly) related in this specific example.

Perspective Plots

A *perspective plot* can depict the estimated posterior density of two unknown parameters. (First create the bivariate kernel density estimate with the *kde2d* function of the *MASS* package.)

```
### Two-dimenionsal kernel density estimate,
### with burn-in removed.
library(MASS)
par(mfrow=c(1,2))
persp(kde2d(mu.nd[-(1:50)], sigma2.nd[-(1:50)]),
      phi=50,theta=30, ## Options for vantage viewpoint
      xlab="mu_nd", ylab="sigma2_nd")
persp(kde2d(mu.nd[-(1:50)], sigma2.nd[-(1:50)]),
      phi=80, theta=35, ## different viewpoint
      xlab="mu_nd", ylab="sigma2_nd")
```

Perspective Plots (R Code on previous slide)



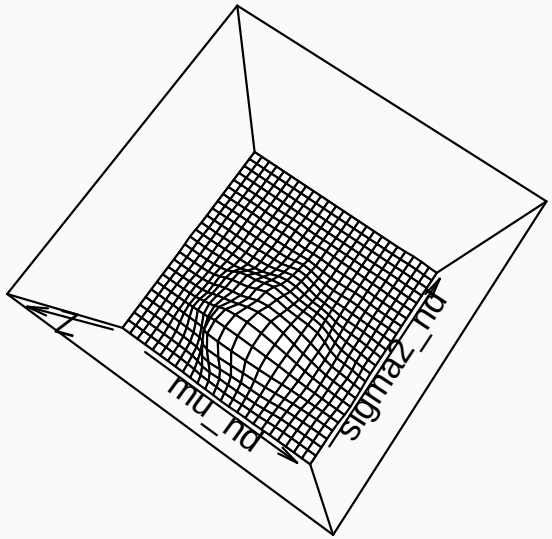


Image Plots

Another useful plot to show the bivariate relationship is an *image plot*. Again, first create the two-dimensional kernel density estimate with the *kde2d* function of the *MASS* package.

Image plots convey information about bivariate densities. Regions that are red have relatively low density, and as the region becomes more yellow and eventually white, the density is increasing.

Contours can be added, also, if this is desired.

Image Plots (cont); R Code on Subsequent Slide

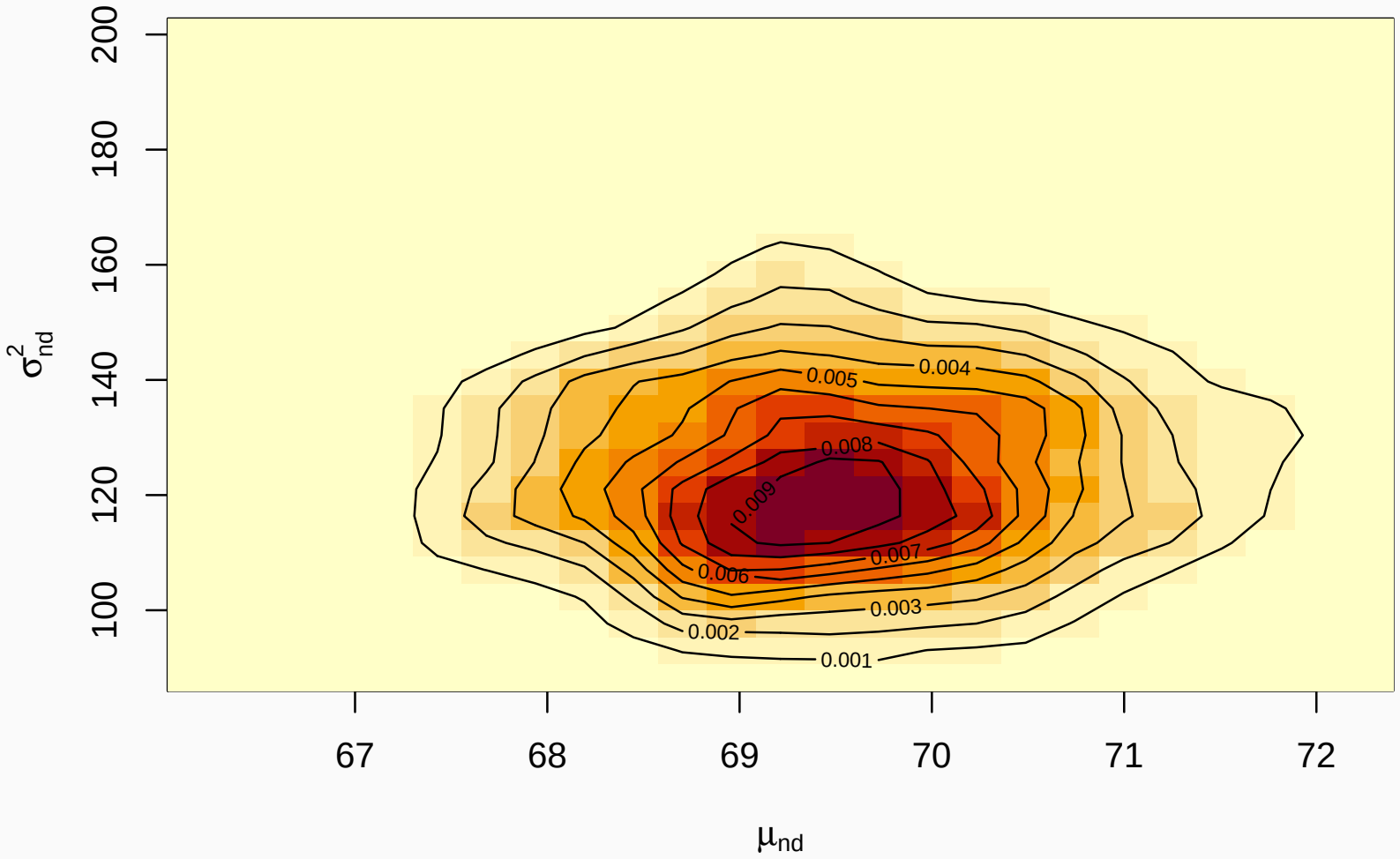


Image Plots (cont)

```
### Two-dimensional kernel density estimate
### with burn-in removed.
library(MASS)
image(kde2d(mu.nd[-(1:50)], sigma2.nd[-(1:50)]),
      xlab=expression(mu[nd]), ylab="")
mtext(expression(sigma[nd]^2),side=2, line=2.5)
### add contour lines to the preceding image plot ###
contour(kde2d(mu.nd[-(1:50)], sigma2.nd[-(1:50)]),
        add=TRUE)
```

Number of Iterations for Gibbs Sampler

Getting 950 post-initialization iterations did not take particularly long.

Obtaining more draws reduces the Monte Carlo error that is introduced by taking samples from the posterior, as opposed to numerically calculating quantities directly from the posterior distribution.

I reran the Gibbs sampler to have 101,000 values in the sequence, where only the last 100,000 were used (generous burn-in period of 1000). The image graph of the previous slide was then applied to this larger sample from the posterior distribution.

Image Plot with 100000 post-burn-in values