

STA 602 - Intro to Bayesian Statistics

Lecture 10

Li Ma

Duke University

Gaussian model with unknown mean and unknown variance

- Sampling model for n readings given the mean θ and variance σ^2 is

$$X_1, X_2, \dots, X_n | \theta, \sigma^2 \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)$$

- Prior distribution for the mean θ

$$\theta | \sigma^2 \sim N(\mu_0, \tau_0^2) \quad p(\theta | \sigma^2) = p(\theta)$$

θ is independent of σ^2

- Now let's assume that the sampling variance σ^2 is also unknown.
- So now we have a two parameter model (θ, σ^2) .
- To complete Bayesian inference, we need to specify a joint prior for (θ, σ^2) .
prior independence $\not\Rightarrow$ posterior independence.



Prior independence

The independency is reasonable, since your prior belief shouldn't be influenced by the quality of detector.

- ▶ So far we have constrained our conditional prior for θ given σ^2 to be

$$\theta | \sigma^2 \sim N(\mu_0, \tau_0^2)$$

where $\tau_0^2 = \sigma^2 / \kappa_0$. In particular, *a priori*, θ and σ^2 are *dependent*.

- ▶ Suppose instead we want θ and σ^2 to be independent *a priori*. That is,

$$p(\theta, \sigma^2) = p(\theta)p(\sigma^2).$$

- ▶ It is convenient to adopt a conjugate prior so that the resulting posterior is analytically tractable.
- ▶ Unfortunately, this is not achievable.

A “semi-conjugate” prior



prior is independent
posterior is negative correlation
(The example of air pollution reading in Durham).

- ▶ Note that given the form of the likelihood $p(x|\theta, \sigma^2)$, *a posteriori* θ and σ^2 will be dependent. That is,

$$p(\theta, \sigma^2 | \mathbf{x}) \neq p(\theta | \mathbf{x})p(\sigma^2 | \mathbf{x})$$

because

$$p(\theta, \sigma^2 | \mathbf{x}) \propto p(\theta)p(\sigma^2) \underbrace{p(\mathbf{x} | \theta, \sigma^2)}_{\text{cannot be factorized}}.$$

- ▶ For example, after observing some value of x , if θ is much smaller in absolute value than x , then σ^2 must be large. For $p(\sigma^2 | \theta, \mathbf{x})$ depends on θ .
- ▶ Question: What would small values of σ^2 imply about θ ?
- ▶ In this sense, strict conjugacy is not feasible, as the prior independence will inevitably be lost under the posterior.

$p(\theta | \sigma^2, \mathbf{x})$ conj with $p(\theta | \sigma^2)$ which can be $p(\theta)$.
 $p(\sigma^2 | \theta, \mathbf{x})$ conj with $p(\sigma^2 | \theta)$ which can be $p(\sigma^2)$.

- ▶ Let's then aim for having a posterior such that $p(\sigma^2 | \mathbf{x})$ is in the same family as $p(\sigma^2)$, and $p(\theta | \sigma^2, \mathbf{x})$ is in the same family as $p(\theta | \sigma^2)$.
- ▶ It turns out that without letting $\tau_0^2 = \sigma^2 / \kappa_0$ even this is not achievable.
- ▶ Inspired by the conjugate case, what if we use the following *Normal-inverse-Gamma* prior that is similar to what we had used previously without the constraint that $\tau_0^2 = \sigma^2 / \kappa_0$:

$$\theta | \sigma^2 \sim N(\mu_0, \tau_0^2)$$

$$\sigma^2 \sim IG(v_0/2, v_0\sigma_0^2/2)$$

or

$$\gamma = \frac{1}{\sigma^2} \sim \text{Gamma}(v_0/2, v_0\sigma_0^2/2).$$

The corresponding conditional posterior

- ▶ As before, by Bayes theorem, the conditional posterior distribution for θ given σ^2 is

$$\theta | \mathbf{x}, \sigma^2 \sim N(\mu_n, \tau_n^2)$$

where

$$\mu_n = \left(\frac{1/\tau_0^2}{1/\tau_0^2 + n/\sigma^2} \right) \mu_0 + \left(\frac{n/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \right) \bar{x}$$

and

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}.$$

- ▶ This is called the *full conditional* of θ . That is, the conditional posterior of θ given all other *random variables*—that is, all other parameters in the model and the data in this example.
- ▶ Again, note that the conditional posterior of θ given σ^2 depends on σ^2 . So there is no posterior independence.

The corresponding conditional posterior

- Without the constraint $\tau_0^2 = \sigma^2/\kappa_0$, we can no longer get a simple Gamma marginal posterior for $1/\sigma^2$ as before.
- Fortunately, we can find the *full conditional* (i.e., conditional posterior of $1/\sigma^2$ given the rest of the parameters, here θ , and the data):

$$\sigma^2 | \mathbf{x}, \theta \sim \text{IG} \left(\frac{v_0 + n}{2}, \frac{1}{2} \left(v_0 \sigma_0^2 + \sum_i (x_i - \theta)^2 \right) \right)$$

or

$$\sigma^2 | \mathbf{x}, \theta \sim \text{IG} \left(\frac{v_0 + n}{2}, \frac{1}{2} \left(v_0 \sigma_0^2 + (n-1)s^2 + n(\bar{x} - \theta)^2 \right) \right)$$

where we again use the fact that

$$\sum_i (x_i - \theta)^2 = (n-1)s^2 + n(\bar{x} - \theta)^2.$$

- Compare this to the marginal posterior of $1/\sigma^2$ in the conjugate prior.

Finding the full conditionals

- ▶ First write then the full joint density of all parameters and the data

$$\begin{aligned} p(\theta, \gamma | \mathbf{x}) &\propto p(\theta, \gamma) p(\mathbf{x} | \theta, \gamma) \\ &= p(\theta) p(\gamma) p(\mathbf{x} | \theta, \gamma) \\ &\propto e^{-\frac{(\theta - \mu_0)^2}{2\tau_0^2}} \cdot \gamma^{\frac{v_0}{2}-1} e^{-\frac{1}{2} v_0 \sigma_0^2 \gamma} \cdot \gamma^{\frac{n}{2}} e^{-\frac{1}{2} \gamma \sum_i (x_i - \theta)^2} \end{aligned}$$

- ▶ Then to find the full condition of each parameter. View this joint density as a function of that parameter, with all parameters fixed as constants.

- ▶ For example, after completion of squares (do it!), when viewed as a function of θ , it is proportional to

$$\frac{1}{\tau_n} e^{-\frac{(\theta - \mu_n)^2}{2\tau_n^2}}$$

which corresponds to $N(\mu_n, \tau_n^2)$.

- ▶ When viewed as a function of γ , it is proportional to

$$\gamma^{\frac{v_0+n}{2}-1} e^{-\frac{\gamma}{2}(v_0\sigma_0^2 + \sum_i (x_i - \theta)^2)}$$

which corresponds to $\text{Gamma}\left(\frac{v_0+n}{2}, \frac{1}{2}(v_0\sigma_0^2 + \sum_i (x_i - \theta)^2)\right)$.

- ▶ **Lesson:** In finding out the full conditional of each parameter, we can indeed treat quantities that does not involve that parameter as “constants” that can be ignored. *No integration at all!*
- ▶ This is in contrast to the calculation of *marginal posteriors* as we did earlier, in which case, the quantities that involve any of the parameters must be kept! (Recall the $\gamma^{1/2}$ term.)
- ▶ In particular, here we do not have to worry about integrating out a parameter resulting in a value that depends on the other parameters.

From full conditionals to posterior samples

- ▶ So we have the full conditionals of both θ and σ^2 , how to get the joint posterior?
- ▶ By a Monte Carlo strategy called *Gibbs sampling!*
- ▶ Drawing (correlated) samples from a multivariate distribution from its full conditionals.
- ▶ *Gibbs sampling*: Iteratively sample each parameter from its full conditional and repeat this process.

Example: The normal-inverse-Gamma prior

Gibbs sampling is a MC Chain

- ▶ Choose an initial value $(\theta^{(0)}, \sigma^{2(0)})$
- ▶ For $t = 1, 2, \dots$,
 - ▶ Draw $\theta^{(t)} \sim N(\mu_n^{(t)}, \tau_n^{2(t)})$
 - ▶ Draw $\sigma^{2(t)} \sim IG\left(\frac{v_0+n}{2}, \frac{1}{2}(v_0\sigma_0^2 + \sum_i(x_i - \theta^{(t)})^2)\right)$.
- ▶ In the above,

$$\mu_n^{(t)} = \left(\frac{1/\tau_0^2}{1/\tau_0^2 + n/\sigma^{2(t-1)}} \right) \mu_0 + \left(\frac{n/\sigma^{2(t-1)}}{1/\tau_0^2 + n/\sigma^{2(t-1)}} \right) \bar{x}$$

and

$$\frac{1}{\tau_n^{2(t)}} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^{2(t-1)}}.$$

- ▶ We can use $\gamma^{(t)}$ in place of $\frac{1}{\sigma^{(t)}}$.

A few remarks

- ▶ The samples are *dependent*—so this is not vanilla Monte Carlo.
- ▶ In particular, $(\theta^{(t)}, \sigma^{2(t)})$ depends on the value of $(\theta^{(t-1)}, \sigma^{2(t-1)})$.
- ▶ The sequence of samples

$$(\theta^{(1)}, \sigma^{2(1)}), (\theta^{(2)}, \sigma^{2(2)}), \dots, (\theta^{(t-1)}, \sigma^{2(t-1)}), (\theta^{(t)}, \sigma^{2(t)}), \dots$$

form a *Markov Chain*.

Markov Chains and its convergence

- ▶ A sequence of random (possibly multi-dimensional) variables (i.e., a stochastic process)

$$\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{t-1}, \mathbf{X}_t, \mathbf{X}_{t+1}, \dots$$

is a *Markov chain* if

$$p(\mathbf{X}_{t+1} | \mathbf{X}_1, \dots, \mathbf{X}_t) = p(\mathbf{X}_{t+1} | \mathbf{X}_t).$$

- ▶ The future depends on the past only through the present.
- ▶ Check that the previous sequence of draws $(\theta^{(t)}, \sigma^2(t))$ for $t = 1, 2, \dots$ form a (homogeneous) Markov chain.

$$\mathbf{x}_t = (\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)})$$

$$\mathbf{x}_{t+1} = (\alpha^{(t+1)}, \beta^{(t+1)}, \gamma^{(t+1)})$$

Markov Chain convergence

ergodic

- Moreover, if a Markov Chain satisfies certain conditions (namely, it is *irreducible, aperiodic, and positive recurrent*), then the marginal distribution of X_t *converge* to a unique *stationary distribution p* . (Markov chain theory is beyond the scope of this class.)
- That is, under such conditions, for large enough t , the unconditional distribution *convergence*

$$\begin{aligned} X_t &\sim p \\ \Rightarrow X_{t+1} &\sim p \\ &\text{(stationary)} \end{aligned}$$

$$P(X_t \in A) \approx \int_A p(x)dx \quad \text{for large enough } t. \quad (X_t, X_{t+1}) \text{ not independent}$$

where $p(A)$ is the so-called stationary distribution, that is, if $X_{t-1} \sim p$ then $X_t \sim p$, i.e.,

$$p(y) = \int p(X_t = y | X_{t-1} = x)p(x)dx.$$

- How large a t is sufficient? *It depends on correlation, dimensionality, multimodality.*

Use Markov Chain for Monte Carlo

- ▶ In order to evaluate an integral

$$\mathbb{E}_p g = \int g(x)p(x)dx.$$

- ▶ If we cannot generate i.i.d. samples from p , but can construct a Markov Chain

$$X_1, X_2, \dots$$

that converges to the stationary distribution p .

- ▶ It turns out that under certain conditions (called *ergodicity*) of the MC, there is a corresponding law of large number for Markov Chains that ensures

Monte-Carlo error will decrease if the MC sample size increases. Due to the stickiness of draws, the decreasing is slow.

$$\frac{1}{S} \sum_{i=1}^S g(X_i) \rightarrow \mathbb{E}_p g \quad \text{as } S \rightarrow \infty.$$

- ▶ In practice, people will **discard** the X_i for small i (the “burn-in”). For example, draw 20000 X_i s but only use the last 10000 values.

The Gaussian example

- ▶ The Markov chain formed $(\theta^{(t)}, \sigma^{2(t)})$ formed by iteratively sampling from the full conditionals of θ and σ^2 for $t = 1, 2, \dots$ satisfies the conditions to ensure eventual convergence to the target distribution

$$p(\theta, \sigma^2 | \mathbf{x}).$$

and the convergence of Monte Carlo estimates.

- ▶ How fast is the convergence? It depends on how strongly correlated θ and σ^2 are under the target distribution.

Gibbs sampling, more generally

- ▶ Suppose we are interested in sampling d -dimensional random vectors $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_d)$ from a multivariate probability distribution $p(\boldsymbol{\theta})$.
- ▶ Suppose we are able to sample from all of the d full conditionals. That is, we know how to sample from

$$p(\theta_i | \boldsymbol{\theta}_{-i})$$

for all $i = 1, 2, \dots, d$, where for notational simplicity, we let $\boldsymbol{\theta}_{-i} = (\theta_1, \theta_2, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_d)$.

- ▶ In Bayesian inference problems, the target distribution is usually the posterior distribution of $\boldsymbol{\theta}$ given data, $p(\boldsymbol{\theta} | \mathbf{x})$.
 - ▶ In this case the full conditionals are $p(\theta_i | \boldsymbol{\theta}_{-i}, \mathbf{x})$.
 - ▶ For example, we have applied this algorithm to a bivariate parameter (θ, σ^2) , and the target distribution is $p(\theta, \sigma^2 | \mathbf{x})$.

Gibbs sampling, more generally

- ▶ The Gibbs sampler proceeds by
 - ▶ Initialize the parameter values at $\boldsymbol{\theta}^{(0)}$.
 - ▶ For $t = 1, 2, \dots,$
 - ▶ For $i = 1, 2, \dots, d$, update the i th parameter θ_i while keeping all other parameters $\boldsymbol{\theta}_{-i}$ fixed.

$$\theta_i^{(t)} \sim p(\theta_i | \boldsymbol{\theta}_{-i}^{(t)}).$$

where $\boldsymbol{\theta}_{-i}^{(t)}$ is the latest values of all other parameters, that is,

$$\boldsymbol{\theta}_{-i}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t)}, \dots, \theta_d^{(t)}).$$

- ▶ For large enough t , the draws $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_d^{(t)})$ will have the desired marginal distribution $p(\boldsymbol{\theta})$.

The speed of convergence to the target distribution

- ▶ The speed of such convergence depends on the correlation among the d parameters under $p(\boldsymbol{\theta})$ (or $p(\boldsymbol{\theta}|\mathbf{x})$ for Bayesian inference).
- ▶ In the special case where all parameters are independent under the target, then it takes one iteration (i.e., $t = 1$) to converge!
- ▶ In the other extreme if all parameters are perfectly correlated, then the chain will never move after $t = 1$ and so will never converge!

Example: Gibbs sampling for a bivariate normal

- ▶ Suppose we want to draw samples of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ from a bivariate normal distribution

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

- ▶ Of course, we can draw independent samples in R from this distribution directly. Here let's practice Gibbs sampling.
- ▶ The full conditionals are

$$\theta_1 | \theta_2 \sim N(\mu_1 + \rho(\theta_2 - \mu_2), 1 - \rho^2)$$

$$\theta_2 | \theta_1 \sim N(\mu_2 + \rho(\theta_1 - \mu_1), 1 - \rho^2)$$

Example: Gibbs sampling for a bivariate normal

- ▶ We initialize the Gibbs sampler at $\boldsymbol{\theta}^{(0)} = (0, 0)$.
- ▶ Then for $t = 1, 2, \dots$
 - ▶ Draw

$$\theta_1^{(t)} | \theta_2^{(t-1)} \sim N(\mu_1 + \rho(\theta_2^{(t-1)} - \mu_2), 1 - \rho^2)$$

and

$$\theta_2^{(t)} | \theta_1^{(t)} \sim N(\mu_2 + \rho(\theta_1^{(t)} - \mu_1), 1 - \rho^2)$$

In a general notion, marginal distribution \rightarrow stationary distribution

In a specific context of MCMC, * convergence: chain has moved into a high density region.

* mixing: chain is able to make long-range movement, explore vast majority of probability mass.

Gibbs sampling code

```
mu <- c(2,2)
rho <- 0.5 # change this to different values and see how it works
S <- 100
theta.mc <- matrix(0,nrow=S,ncol=2)
theta <- c(0,0) # initial value

# Gibbs sampling
for (t in 1:S) {
  theta[1] <- rnorm(1,mean=mu[1]+rho*(theta[2]-mu[2]),sd=sqrt(1-rho^2))
  theta[2] <- rnorm(1,mean=mu[2]+rho*(theta[1]-mu[1]),sd=sqrt(1-rho^2))
  theta.mc[t,] <- theta
}
```

```
library(mvtnorm)
mu <- c(2,2)
rho <- 0.5 # change this to different values
Sigma <- matrix(c(1,rho,rho,1),ncol=2);
ylim <- xlim <- c(-1,5)
x <- seq(xlim[1],xlim[2],length=200)
y <- x
xy.grid <- expand.grid(x,y)
den.grid <- matrix(dmvnorm(xy.grid,mean=mu,sigma=Sigma),nrow=length(x))

contour(x,y,den.grid,xlim=xlim,ylim=ylim,
        xlab=expression(theta[1]),ylab=expression(theta[2]))
```

```

S <- 100
theta.mc <- matrix(0,nrow=S,ncol=2)
theta <- c(0,0) # initial value

points(x=theta[1],y=theta[2],col="red4",pch=16); Sys.sleep(1)
theta.prev <- theta
for (t in 1:S) {
  theta[1] <- rnorm(1,mean=mu[1]+rho*(theta[2]-mu[2]),sd=sqrt(1-rho^2))
  if (t<20) Sys.sleep(1)
  segments(theta.prev[1],theta.prev[2],theta[1],theta[2],col="gray") # gray line segments
  theta[2] <- rnorm(1,mean=mu[2]+rho*(theta[1]-mu[1]),sd=sqrt(1-rho^2))
  if (t<20) Sys.sleep(1)
  segments(theta[1],theta.prev[2],theta[1],theta[2],col="gray") # gray line segments
  theta.mc[t,] <- theta

  if(t < 20) {
    points(x=theta[1],y=theta[2],col="red4",pch=16);
  } else {
    points(x=theta[1],y=theta[2],col="green4",pch=16); Sys.sleep(0.1)
  }

  theta.prev <- theta
}

}

```

Markov Chain Monte Carlo

- ▶ Gibbs sampling is an example of a Markov Chain Monte Carlo (MCMC) algorithm.
- ▶ The idea is to sample $\boldsymbol{\theta}$ through constructing a Markov Chain. As such, the draws

$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$$

are dependent and form a Markov chain. So each $\boldsymbol{\theta}^{(i)}$ is generated based on only the value of $\boldsymbol{\theta}^{(i-1)}$.

- ▶ By ensuring that the Markov chain that satisfies the necessary conditions to ensure its convergence to a desired stationary distribution, typically a posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$, as well as the convergence of Monte Carlo estimates, we are able to
 - ▶ draw these dependent samples from $p(\boldsymbol{\theta}|\mathbf{x})$;
 - ▶ evaluate integrals with respect to $p(\boldsymbol{\theta}|\mathbf{x})$.
- ▶ Thus based on this sample we can proceed with Bayesian inference as usual.

Some distinctions from sampling independent samples

- ▶ Unlike standard Monte Carlo where we draw independent samples *exactly* from some target p ,
 - ▶ MCMC samples will only converge to the target p after a sufficient number of steps. So we will typically discard the first steps in the chain, which is called the *burn-in* stage.
 - ▶ The sampled values are dependent, and each is drawn based on the previous value, and so the *effective sample size* depends on how effectively that chain of values move around the parameter space to explore the distribution.

Two conditions necessary for an effective MCMC chain

- ▶ ***Fast convergence***: The Markov chain needs to be long enough so that it has moved into a high-probability region of the target $p(\boldsymbol{\theta} | \mathbf{x})$.
 - ▶ This ensures that we are actually sampling from the desired posterior after the *burn-in*.
- ▶ ***Good mixing***: The Markov chain should be moving across the high probability regions of $p(\boldsymbol{\theta} | \mathbf{x})$, rather than being stuck in one or a few regions, while move across to another only occasionally.
 - ▶ This ensures that the effective sample size is large so that the Monte Carlo error is small.

Back to air pollutant example

```
x <- c(104,105,103,102,105,107,106,104,103,106) # the data
n <- length(x) # sample size

S <- 10000
xbar <- mean(x)
s2 <- var(x)
n <- length(x)

THETA <- matrix(NA,nrow=S,ncol=2,dimnames=list(1:S,c("theta","sigma2")))
THETA.init <- c(xbar,s2) # Initial values set to the MLE
THETA.curr <- THETA.init # the parameter values at the current iteration
mu.0 <- 100; nu.0 <- 1; tau2.0 <- 25; sigma2.0 <- 4

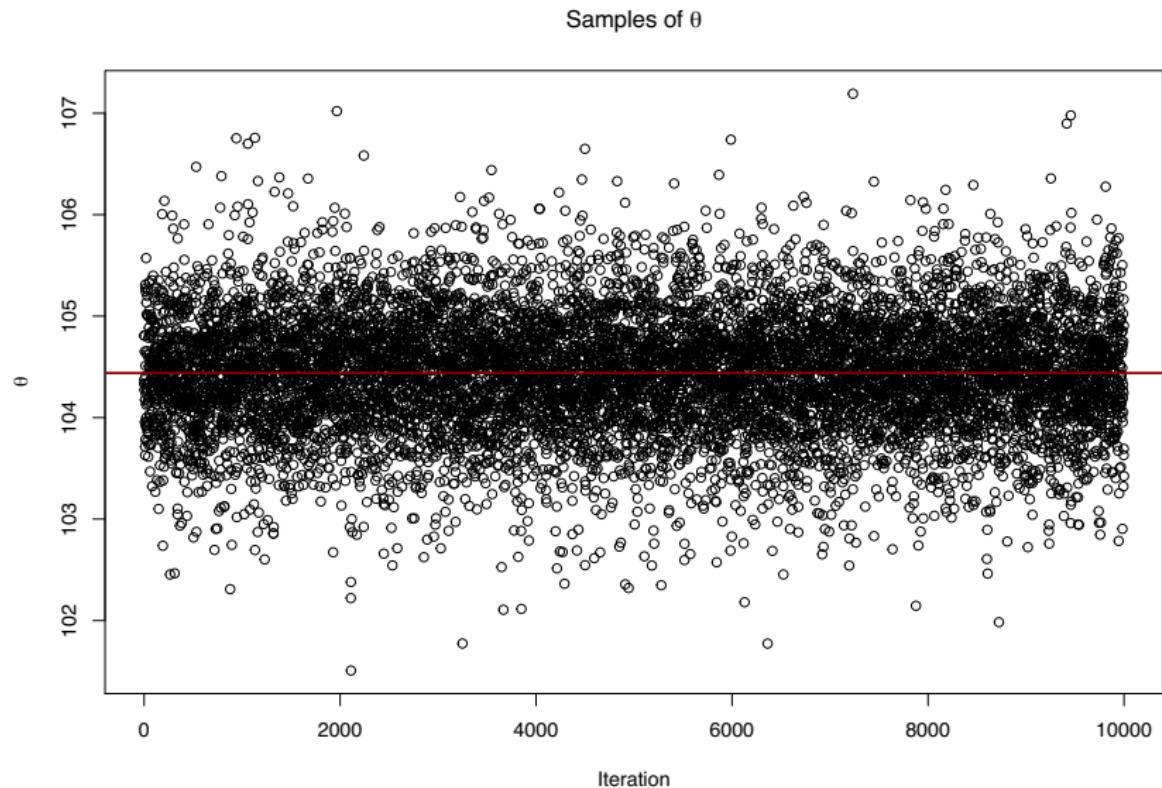
### Start Gibbs sampling
for (t in 1:S) {
  tau2.n <- 1/(1/tau2.0 + n/THETA.curr[2])
  mu.n <- (mu.0/tau2.0 + xbar*n/THETA.curr[2])/(1/tau2.0 + n/THETA.curr[2])

  ## Update theta
  THETA.curr[1] <- rnorm(1,mean=mu.n,sd=sqrt(tau2.n))

  ## Update sigma2
  THETA.curr[2] <- 1/rgamma(1,shape=(nu.0+n)/2,
    rate=1/2*(nu.0*sigma2.0+sum((x-THETA.curr[1])^2)))

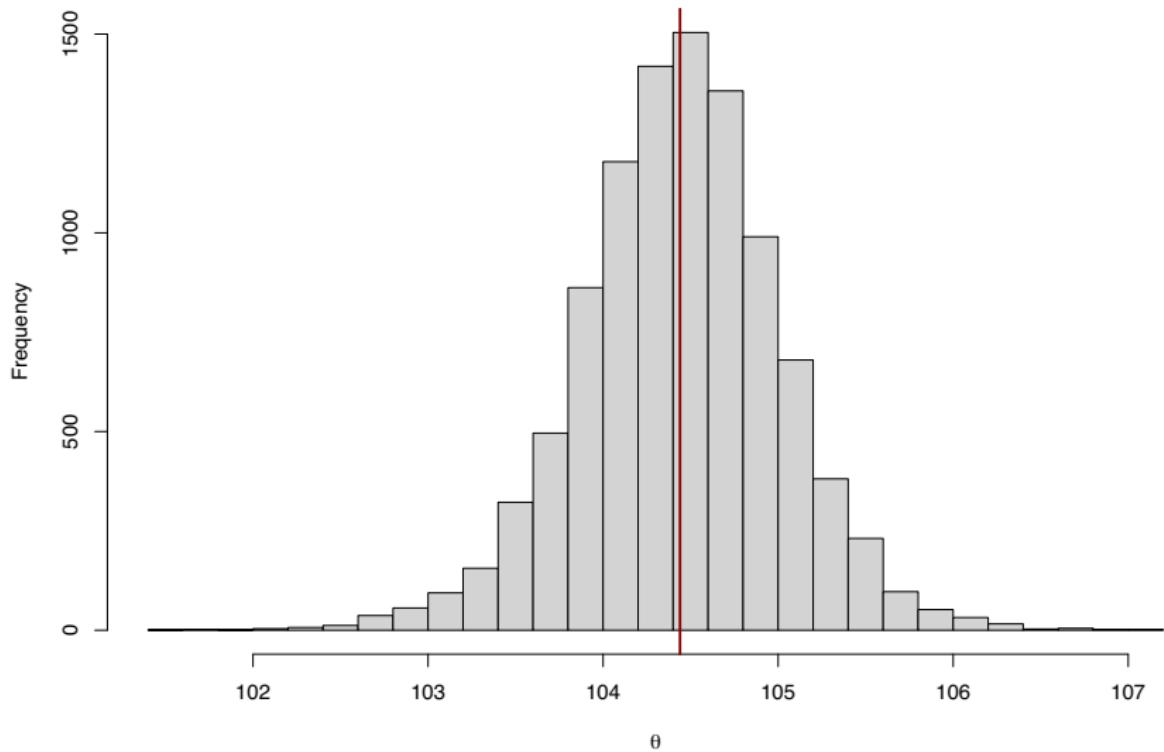
  ## Save the current iteration
  THETA[t,] <- THETA.curr
}
```

Trace plot for θ

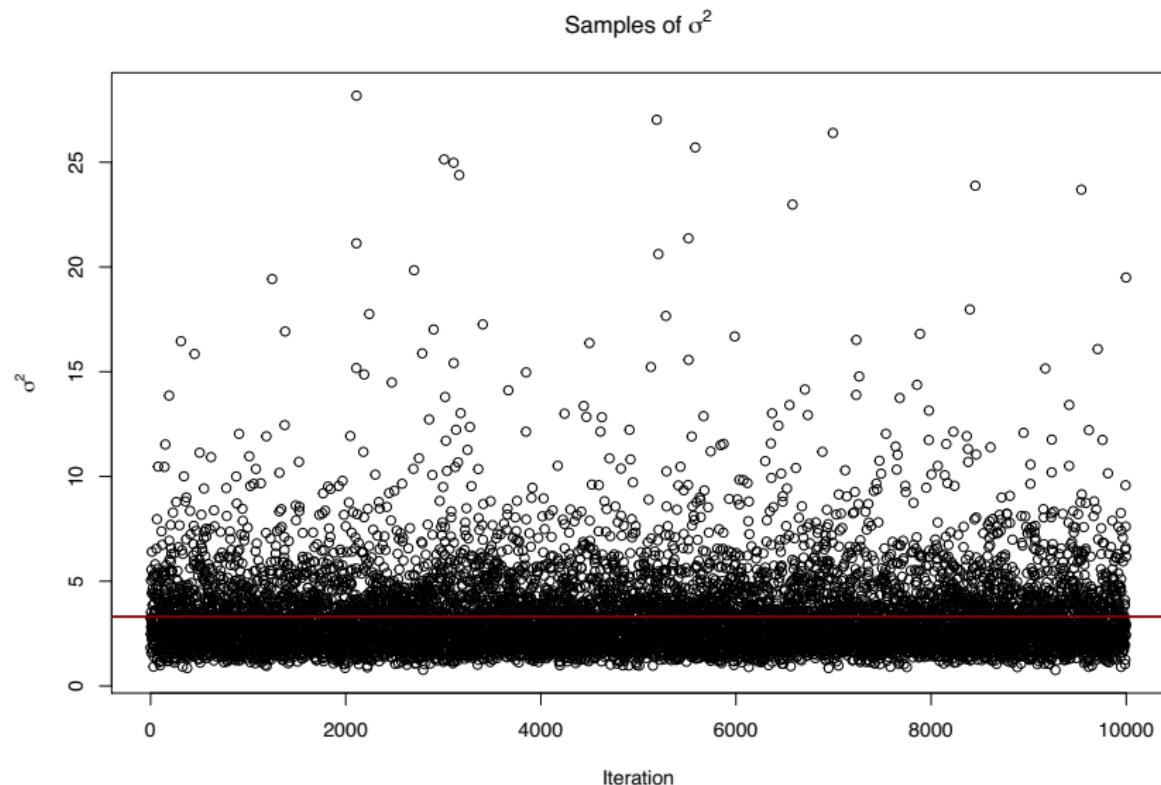


Histogram for θ

Histogram of sampled θ



Trace plot for σ^2



Histogram for σ^2

Histogram of sampled σ^2

