Lecture 5 - Key

The Normal Model

A random variable Y is said to be normally distributed with mean θ and variance σ^2 if the density of Y is:

$$p(y|\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2\right]$$

Key points about the normal distribution:

- The distribution is symmetric about θ, *and the mode, median and mean are all equal to* θ

- about 95% of the mass resides within *two standard deviations of the mean*

- if $X \sim N(\mu,\tau^2)$ and $Y \sim N(\theta,\sigma^2)$ and X and Y are independent, then $aX + bY* \sim N(a\mu + b\theta, a^2\tau^2 + b^2\sigma^2)*$

- the `rnorm, dnorm, pnorm` and `qnorm` commands in `R` are very useful, but they take σ as their argument not σ^2, so be careful.

Inference for θ, conditional on σ^2

When sigma is known, we seek the posterior distribution of $p(\theta|y_1,....,y_n,\sigma^2)$. A conjugate prior, $p(\theta|\sigma^2)$ is of the form:

$$\begin{aligned} p(\theta|y_1,...,y_n,\sigma^2) &\propto p(\theta|\sigma^2) \times \exp\left[-\frac{1}{2\sigma^2}\sum (y_i-\theta)^2\right] \\ &\propto \exp\left[c_1\left(\theta - c_2\right)^2\right] \end{aligned}$$

thus a conjugate prior for $p(\theta|y_1,...,y_n,\sigma^2)$ *is from the normal family of distributions.*

Now consider a prior distribution $p(\theta|\sigma^2) \sim N(\mu_0, \tau_0^2)$ and compute the posterior distribution.

$$\begin{aligned} p(\theta|y_1,\ldots,y_n,\sigma^2) &\propto p(y_1,\ldots,y_n|\theta,\sigma^2)p(\theta|\sigma^2) \\ &\propto \exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\} \times \exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\theta)^2\right\} \\ &\quad \text{now combine terms with the powers of } \theta \\ &\propto \exp\left\{-\frac{1}{2}\left[\theta^2\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right) - 2\theta\left(\frac{\mu_0}{\tau_0^2}+\frac{\sum y_i}{\sigma^2}\right) + c(y_1,\ldots,y_n,\mu,\sigma^2,\tau_0^2)\right]\right\} \\ &\quad \text{Note from here we could complete the square} \end{aligned}$$

However, there is a shortcut here that you will probably do *approximately* 50 times over the course of this class. Note if $\theta \sim N(E, V)$ then

$$*p(\theta) \propto \exp\left[-\frac{1}{2V}(\theta-E)^2\right] \tag{1}$$

$$* \propto \exp\left[-\frac{1}{2}\left(\frac{\theta^2}{V} - \frac{2\theta E}{V} + c(E,V)\right)\right] * \tag{2}$$

Hence from above, the variance of the distribution is the reciprocal of the term with θ^2. That is:

$$V[\theta] = *\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1} *$$

Similarly the term associated with 2θ is *E/V, so the expectation is this term times the variance. So the expectation is calculated as:*

$$E[\theta] = *\left(\frac{\mu_0}{\tau_0^2}+\frac{\sum y_i}{\sigma^2}\right)\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1} *$$

Notes about the posterior and predictive distributions

It is common to reparameterize the variance using the inverse, which is known as the precision. Then:

- $\tilde{\sigma}^2 = 1/\sigma^2 =$ sampling precision

- $\tilde{\tau}_0^2 = 1/\tau_0^2 =$ prior precision

- $\tilde{\tau}_n^2 = 1/\tau_n^2 =$ posterior precision, where τ_n^2 is the posterior variance

Now the posterior precision (i.e. how close the data are to θ) is a function of the prior precision and information from the data: $\tilde{\tau}_n^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2$

The posterior mean can be reparameterized as a weighted average of the prior mean and the sample mean.

$$*\mu_n = \frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\mu_0 + \frac{n\tilde{\sigma}^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\bar{y}, *$$

where μ_n is the posterior mean and $\bar{y}$ is the sample mean.

The predictive distribution of $p(y* | \sigma^2, y_1, \ldots, y_n) \sim N(\mu_n, \tau_n^2 + \sigma)$. This will be a homework problem.

Joint inference for mean and variance in normal model

Thus far we have focused on Bayesian inference for settings with one parameter. Dealing with multiple parameters is not fundamentally different as we use a joint prior $p(\theta_1, \theta_2)$ and use the same mechanics with Bayes rule.

In the normal case we seek the posterior:

$$p(\theta, \sigma^2|y_1, \ldots, y_n) \propto p(y_1, \ldots, y_n|\theta, \sigma^2)p(\theta, \sigma^2)$$

Recall that $p(\theta, \sigma^2)$ can be expressed as $p(\theta|\sigma^2)p(\sigma^2)$. For now, let the prior on θ the mean term be:

$$*p(\theta|\sigma^2) \sim N(\mu_0, \sigma^2/\kappa_0).*$$

Then μ_0 can be interpreted as the mean and κ_0 corresponds to the 'hypothetical' number of prior observations.

A prior on σ^2 is still needed, a required property for this prior is the the support of $\sigma^2 = (0, \infty)$. A popular distribution with this property is the Gamma distribution. Unfortunately this is not conjugate (or semi-conjugate) for the variance. It turns out that the gamma distribution is conjugate for the precision term $\phi = 1/\sigma^2$, which many Bayesians will use. This implies that the inverse gamma distribution can be used as a prior for σ^2.

For now, set the prior on the precision term $(1/\sigma^2)$ to a gamma distribution. For interpretability this is parameterized as:

$$1/\sigma^2 \sim gamma(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2)$$

Using this parameterization:

- $E[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2-1}$
- $\text{mode}[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2+1}$,
- $Var[\sigma^2]$ is decreasing in ν_0

The nice thing about this parameterization is that σ_0^2 can be interpreted as the sample variance from ν_0 prior samples.

Implementation

Use the following prior distributions for θ and σ^2:

$$\begin{aligned} 1/\sigma^2 &\sim gamma(\nu_0/2, \nu_0\sigma_0)^2/2) \\ \theta|\sigma^2 &\sim N(\mu_0, \sigma^2/\kappa_0) \end{aligned}$$

and the sampling model for Y

$$Y_1, \ldots, Y_n|\theta, \sigma^2 \sim i.i.d.\ \ normal(\theta, \sigma^2).$$

Now the posterior distribution can also be decomposed in a similar fashion to the prior such that:

$$p(\theta, \sigma^2|y_1, \ldots, y_n) = p(\theta|\sigma^2, y_1, \ldots, y_n)p(\sigma^2|y_1, \ldots, y_n).$$

Using the results from the case where σ^2 was known, we get that:

$$*\theta|y_1, \ldots, y_n, \sigma^2 \sim normal(\mu_n, \sigma^2\kappa_n), *$$

where $\kappa_n = \kappa_0 + n$ *and* $\mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}$. *Note that this distribution still depends on* σ^2 *which we do not know.*

The *marginal posterior* distribution of σ^2 integrates out θ

$$\begin{aligned} p(\sigma^2|y_1, \ldots, y_n) &\propto p(\sigma^2)p(y_1, \ldots, y_n|\sigma^2) \\ &= p(\sigma^2)\int p(y_1, \ldots, y_n|\theta, \sigma^2)p(\theta|\sigma^2)d\theta \end{aligned}$$

It turns out (HW?) that:

$$1/\sigma^2|y_1, \ldots, y_n \sim gamma(\nu_n/2, \nu_n\sigma_n^2/2),$$

where $\nu_n = \nu_0 + n$, $\sigma_n^2 = \frac{1}{\nu_n}\left\{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2\right\}$, and $s^2 = \frac{\sum_i (y_i - \bar{y})^2}{n-1}$. Again the interpretation is that ν_0 is the prior sample size for σ_0^2.

Posterior Sampling from Normal

Now we seek to create draws from the *joint posterior distribution* $p(\theta, \sigma^2|y_1, \ldots, y_n)$ and the *marginal posterior distributions* $p(\theta|y_1, \ldots, y_n)$ and $p(\sigma^2|y_1, \ldots y_n)$ Note the marginal posterior distributions would be used to calculate quantities such as $Pr[\theta > 0|y_1, \ldots, y_n]$.

Using a Monte Carlo procedure, we can simulate samples from the joint posterior using the following algorithm:

1. *Simulate $\sigma^{2(i)} \sim IG(\nu_n/2, \sigma_n^2\nu_n/2)$

   *
2. *Simulate $\theta^{(i)} \sim N(\mu_n, \sigma^{2(i)}/\kappa_n)$

   *
3. *Repeat m times

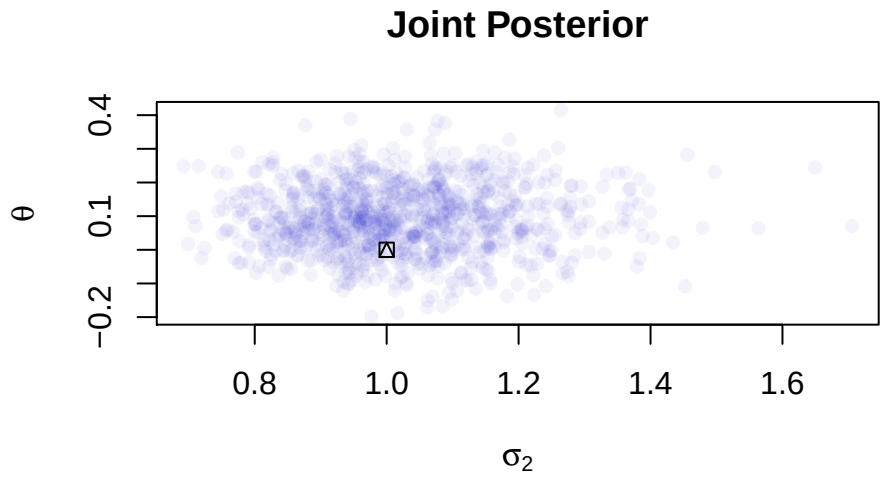
   *

Note that each pair $\{\sigma^{2(i)}, \theta^{(i)}\}$ is a sample from the joint posterior distibution and that $\{\sigma_1^2, \ldots, \sigma_m^2\}$ and $\{\theta_1, \ldots, \theta_m\}$ are samples from the respective marginal posterior distributions.

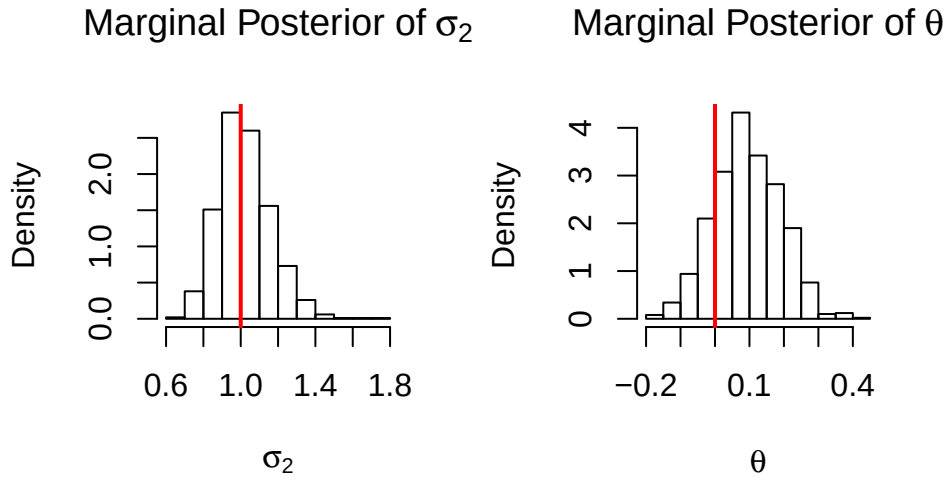
The R code for this follows as:

```
#### Posterior Sampling with Normal Model
set.seed(09182017)
# true parameters from normal distribution
sigma.sq.true <- 1
theta.true <- 0

# generate data
num.obs <- 100
y <- rnorm(num.obs,mean = theta.true, sd = sqrt(sigma.sq.true))

# specify terms for priors
nu.0 <- 1
sigma.sq.0 <- 10
mu.0 <- 0
kappa.0 <- 1

# compute terms in posterior
kappa.n <- kappa.0 + num.obs
nu.n <- nu.0 + num.obs
s.sq <- var(y) #sum((y - mean(y))^2) / (num.obs - 1)
sigma.sq.n <- (1 / nu.n) * (nu.0 * sigma.sq.0 + (num.obs - 1) * s.sq +
(kappa.0*num.obs)/kappa.n * (mean(y) - mu.0)^2)
mu.n <- (kappa.0 * mu.0 + num.obs * mean(y)) / kappa.n

# simulate from posterior
#install.packages("LearnBayes")
library(LearnBayes) # for rigamma
num.sims <- 1000
sigma.sq.sims <- theta.sims <- rep(0,num.sims)
for (i in 1:num.sims){
  sigma.sq.sims[i] <- rigamma(1,nu.n/2,sigma.sq.n*nu.n/2)
  theta.sims[i] <- rnorm(1, mu.n, sqrt(sigma.sq.sims[i]/kappa.n))
}
```

```r
## Plots
library(grDevices) # for rgb
plot(sigma.sq.sims,theta.sims,pch=16,col=rgb(.1,.1,.8,.05),ylab=expression(theta),
     xlab=expression(sigma[2]),main='Joint Posterior')
points(1,0,pch=14,col='black')
```




```r
par(mfcol=c(1,2))
hist(sigma.sq.sims,prob=T,main=expression('Marginal Posterior of' ~ sigma[2]),
     xlab=expression(sigma[2]))
abline(v=1,col='red',lwd=2)
hist(theta.sims,prob=T,main=expression('Marginal Posterior of' ~ theta),xlab=expression(theta))
abline(v=0,col='red',lwd=2)
```



It is important to note that the prior structure is very specific in this case, where $p(\theta|\sigma^2)$ is a function of σ^2. In most prior structures this type of conditional sampling scheme is not as easy as this case and we need to use Markov Chain Monte Carlo methods.

Posterior Sampling with the Gibbs Sampler

In the previous section we modeled the uncertainty in θ as a function of σ^2, where $p(\theta|\sigma^2) = N(\mu_0, \sigma^2/\kappa_0)$. In some situations this makes sense, but in others the uncertainty in θ may be specified independently from σ^2 *Mathematically, this translates to* $p(\sigma^2, \theta) = p(\theta) \times p(\sigma^2)$

A common *semiconjugate* set of prior distributions is:

$$\begin{aligned} *\theta &\sim N(\mu_0, \tau_0^2) \\ *\sigma^2 &\sim IG(\nu_0, \nu_0\sigma_0^2/2)* \end{aligned}$$

Now when $Y_1, \ldots, Y_n|\theta, \sigma^2 \sim N(\theta, \sigma^2)$ then $\theta|\sigma^2, y_1, \ldots, y_n \sim N(\mu_n, \tau_n^2)$.

$$\mu_n = \frac{\mu_0/\tau_0^2 + n\bar{y}/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \qquad \text{and} \qquad \tau_n^2 = \left(\frac{1}{\tau_0^2} = \frac{n}{\sigma^2}\right)^{-1}$$

In the conjugate case where τ_0^2 was proportional to σ^2, samples from the joint posterior can be taken using the Monte Carlo procedure demonstrated before. However, when τ_0^2 is not proportional to σ^2 the marginal density of σ^2 is not an inverse gamma distribution or another named distribution that permits easy sampling.

Suppose that you know the value of θ. Then the conditional distribution of σ^2 is:

$$\begin{aligned} p(\sigma^2|\theta, y_1, \ldots y_n) &\propto p(y_1, \ldots, y_n|\theta, \sigma^2)p(\sigma^2) \\ &\propto (\sigma^2)^{-n/2} \exp[-1/\sigma^2 \sum (y_i - \theta)^2/2] \times (\sigma^2)^{\frac{\nu_0}{2}-1} \exp[-1/\sigma^2 \nu_0\sigma_0^2/2] \\ &\propto (\sigma^2)^{(\frac{\nu_0+n}{2})-1} \exp[-\frac{1}{\sigma^2}\frac{\nu_0\sigma_0^2 + \sum(y_i - \theta)^2)}{2}] \end{aligned}$$

which is the kernel of an inverse gamma distribution. So $\sigma^2|\theta, y_1, \ldots, y_n \sim InvGamma(\nu_n/2, \nu_n\sigma_n^2(\theta)/2)$, where $\nu_n = \nu_0 + n$, $\sigma_n^2(\theta) = \frac{1}{\nu_n}[\nu_0\sigma_0^2 + ns_n^2(\theta)]$ and $s_n^2(\theta) = \sum(y_i - \theta)^2/n$ the unbiased estimate of σ^2 if θ were known.

Now can we use the full conditional distributions to draw samples from the joint posterior?

Suppose we had $\sigma^{2(1)}$, a single sample from the marginal posterior distribution $p(\sigma^2|y_1, \ldots, y_n)$. Then we could sample:

$$\theta^{(1)} \sim p(\theta|y_1, \ldots, y_n, \sigma^{2(1)})$$

and $\{\theta^{(1)}, \sigma^{2(1)}\}$ would be a sample from the joint posterior distribution $p(\theta, \sigma^2|y_1, \ldots, y_n)$. Now using $\theta^{(1)}$ we can generate another sample of σ^2 from

$$\sigma^{2(2)} \sim p(\sigma^2|y_1, \ldots, y_n, \theta^{(1)})$$

This sample $\{\theta^{(1)}, \sigma^{2(2)}\}$ would also be a sample from the joint posterior distribution. This process follows iteratively. However, we don't actually have $\sigma^{2(1)}$.

Gibbs Sampler

The distributions $p(\theta|y_1,\ldots,y_n,\sigma^2)$ and $p(\sigma^2|y_1,\ldots,y_n,\theta)$ are known as the full conditional distributions, that is they condition on all other values and parameters. The Gibbs sampler uses these full conditional distributions and the procedure follows as:

1. sample $\theta^{(i+1)}$ from $^*p(\theta|y_1,\ldots,y_n,\sigma^{2(i)})$

   *

2. sample $\sigma^{2(i+1)}$ from $^*p(\sigma^2|y_1,\ldots,y_n,\theta^{(i+1)})$

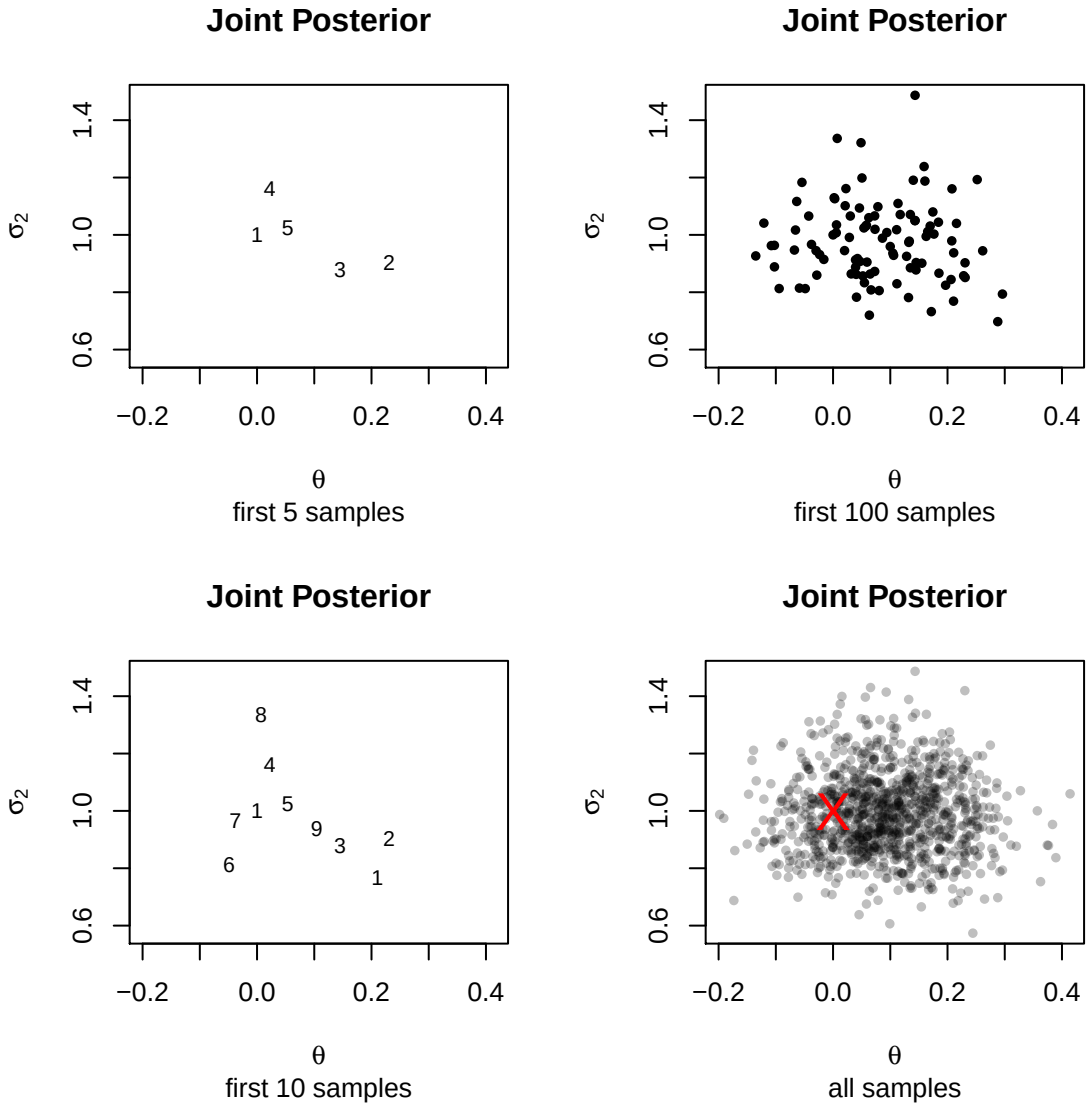
   *

3. let $^*\phi^{(i+1)} = \{\theta^{(i+1)},\sigma^{2(i+1)}\}$

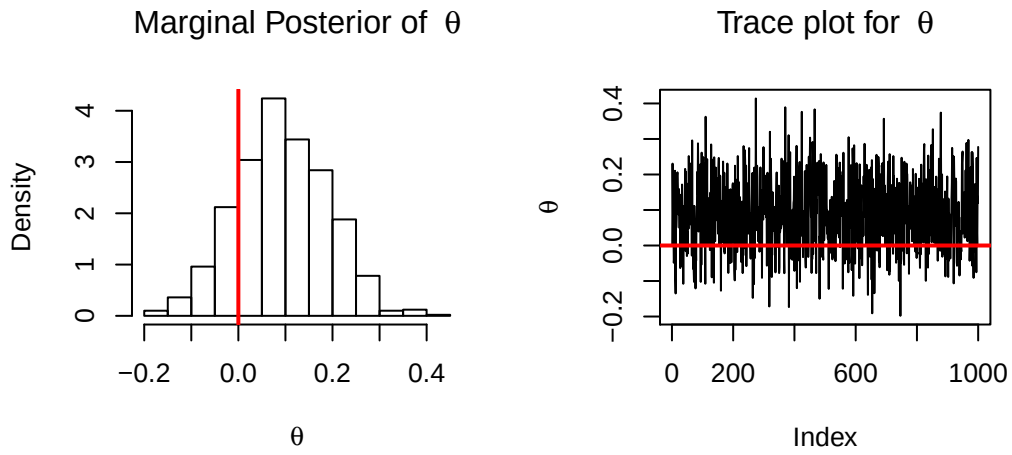
   *

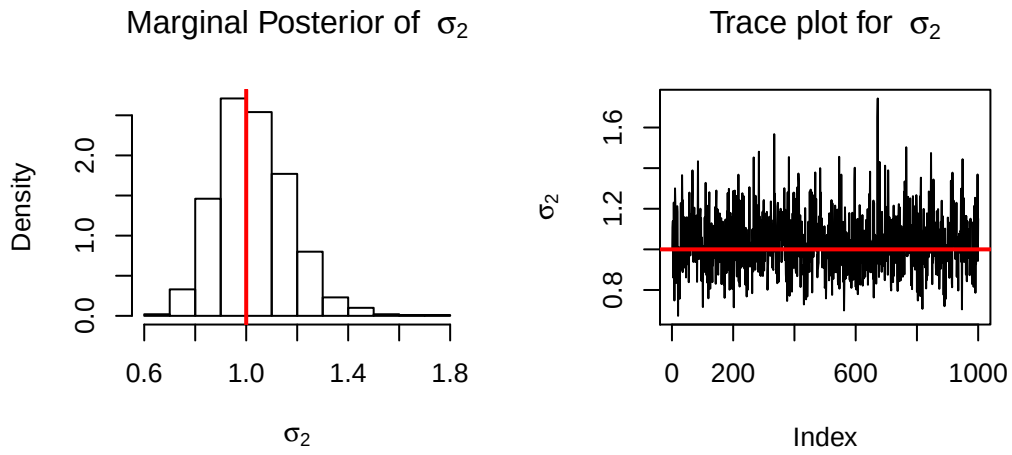
```
######### First Gibbs Sampler
set.seed(09182017)
### simulate data
num.obs <- 100
mu.true <- 0
sigmasq.true <- 1
y <- rnorm(num.obs,mu.true,sigmasq.true)
mean.y <- mean(y)
var.y <- var(y)
library(LearnBayes) # for rigamma
### initialize vectors and set starting values and priors
num.sims <- 1000
Phi <- matrix(0,nrow=num.sims,ncol=2)
Phi[1,1] <- 0 # initialize theta
Phi[1,2] <- 1 # initialize (sigmasq)
mu.0 <- 0
tausq.0 <- 1
nu.0 <- 1
sigmasq.0 <- 10
for (i in 2:num.sims){
  # sample theta from full conditional
  mu.n <- (mu.0 / tausq.0 + num.obs * mean.y / Phi[(i-1),2]) / (1 / tausq.0 + num.obs / Phi[(i-1),2] )
  tausq.n <- 1 / (1/tausq.0 + num.obs / Phi[(i-1),2])
  Phi[i,1] <- rnorm(1,mu.n,sqrt(tausq.n))

  # sample (1/sigma.sq) from full conditional
  nu.n <- nu.0 + num.obs
  sigmasq.n.theta <- 1/nu.n*(nu.0*sigmasq.0 + sum((y - Phi[i,1])^2))
  Phi[i,2] <- rigamma(1,nu.n/2,nu.n*sigmasq.n.theta/2)
}
```

```r
par(mfcol=c(2,2))
# plot joint posterior
plot(Phi[1:5,1],1/Phi[1:5,2],xlim=range(Phi[,1]),ylim=range(1/Phi[,2]),pch=c('1','2','3','4','5'),
     cex=.8, ylab=expression(sigma[2]), xlab = expression(theta),
     main='Joint Posterior',sub='first 5 samples')

plot(Phi[1:10,1],1/Phi[1:10,2],xlim=range(Phi[,1]),ylim=range(1/Phi[,2]),pch=as.character(1:15),
     cex=.8,ylab=expression(sigma[2]), xlab = expression(theta),
     main='Joint Posterior',sub='first 10 samples')

plot(Phi[1:100,1],1/Phi[1:100,2],xlim=range(Phi[,1]),ylim=range(1/Phi[,2]),pch=16,col=rgb(0,0,0,1),
     cex=.8,ylab=expression(sigma[2]), xlab = expression(theta),
     main='Joint Posterior',sub='first 100 samples')

plot(Phi[,1],1/Phi[,2],xlim=range(Phi[,1]),ylim=range(1/Phi[,2]),pch=16,col=rgb(0,0,0,.25),
     cex=.8, ylab=expression(sigma[2]), xlab = expression(theta),
     main='Joint Posterior',sub='all samples')
points(0,1,pch='X',col='red',cex=2)
```



```r
par(mfcol=c(2,2))
# plot marginal posterior of theta
hist(Phi[,1],xlab=expression(theta),main=expression('Marginal Posterior of ' ~ theta),probability=T)
abline(v=mu.true,col='red',lwd=2)
# plot marginal posterior of sigmasq
hist(Phi[,2],xlab=expression(sigma[2]),main=expression('Marginal Posterior of ' ~ sigma[2]),probability=
abline(v=sigmasq.true^2,col='red',lwd=2)

# plot trace plots
plot(Phi[,1],type='l',ylab=expression(theta), main=expression('Trace plot for ' ~ theta))
abline(h=mu.true,lwd=2,col='red')
plot(Phi[,2],type='l',ylab=expression(sigma[2]), main=expression('Trace plot for ' ~ sigma[2]))
abline(h=sigmasq.true^2,lwd=2,col='red')
```






```r
# compute posterior mean and quantiles
colMeans(Phi)
```

```
## [1] 0.09207189 1.03132560
```

```r
apply(Phi,2,quantile,probs=c(.025,.975))
```

```
##              [,1]      [,2]
## 2.5%  -0.09821121 0.7841903
## 97.5%  0.27699990 1.3534315
```

So what do we do about the starting point? We will see that given a reasonable starting point the algorithm will converge to the true posterior distribution. Hence the first (few) iterations are regarded as the burn-in period and are discarded (as they have not yet reached the true posterior).

More on the Gibbs Sampler

The algorithm previously detailed is called the *Gibbs Sampler* and generates a dependent sequence of parameters $\{\phi_1, \phi_2, \ldots, \phi_n\}$. This is in contrast to the Monte Carlo procedure we previously detailed, including the situation where $p(\theta|\sigma^2) \sim N(\mu_0, \sigma^2/\kappa_0)$.

The Gibbs Sampler is a basic Markov Chain Monte Carlo (MCMC) algorithm. A Markov chain is a stochastic process where the current state only depends on the previous state. Formally

$$Pr(X_n = x_n|X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots) = Pr(X_n = x_n|X_{n-1} = x_{n-1})$$

Depending on the class interests, we may return to talk more about the theory of MCMC later in the course, but the basic ideas are:

$$Pr(\phi^{(i)} \in A) \rightarrow \int_A p(\phi)d(\phi) \text{ as } j \rightarrow \infty$$

That is the sampling distribution of the draws from the MCMC algorithm approach the desired target distribution (generally a posterior in Bayesian statistics) as the number of samples j goes to infinity. The is not dependent on the starting values of $\phi^{(0)}$, but poor starting values will take longer for convergence. Note this will be more problematic when we consider another MCMC algorithm, the Metropolis-Hastings sampler.

Given the equation above, for most functions $g(.)$:

$$\frac{1}{I}\sum_{i=1}^{I} g(\phi) \rightarrow E[g(\phi)] = \int g(\phi)p(\phi)d\phi \text{ as } I \rightarrow \infty$$

Thus we can approximate expectations of functions of ϕ using the sample average from the MCMC draws, similar to our Monte Carlo procedures presented earlier.