Week 5: Sept 26 - Sept 30

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:

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

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 $\sigma^2 =$. A popular distribution with this property is the . Unfortunately this is not conjugate (or semi-conjugate) for the variance. It turns out that the is conjugate for the precision term $\phi =$, which many Bayesians will use. This implies that the distribution can be used as a prior for σ^2.

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

$$1/\sigma^2 \sim$$

Using this parameterization:

-

-

-

The nice thing about this parameterization is that σ_0^2 can be interpreted as from ν_0

Implementation

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

$$\begin{aligned} 1/\sigma^2 &\sim gamma(\\ \theta|\sigma^2 &\sim N(\end{aligned}$$

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:

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

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

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

It turns out that:

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

Now we seek to create draws from the **joint posterior distribution** and the **marginal posterior distributions** and . 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

2. Simulate

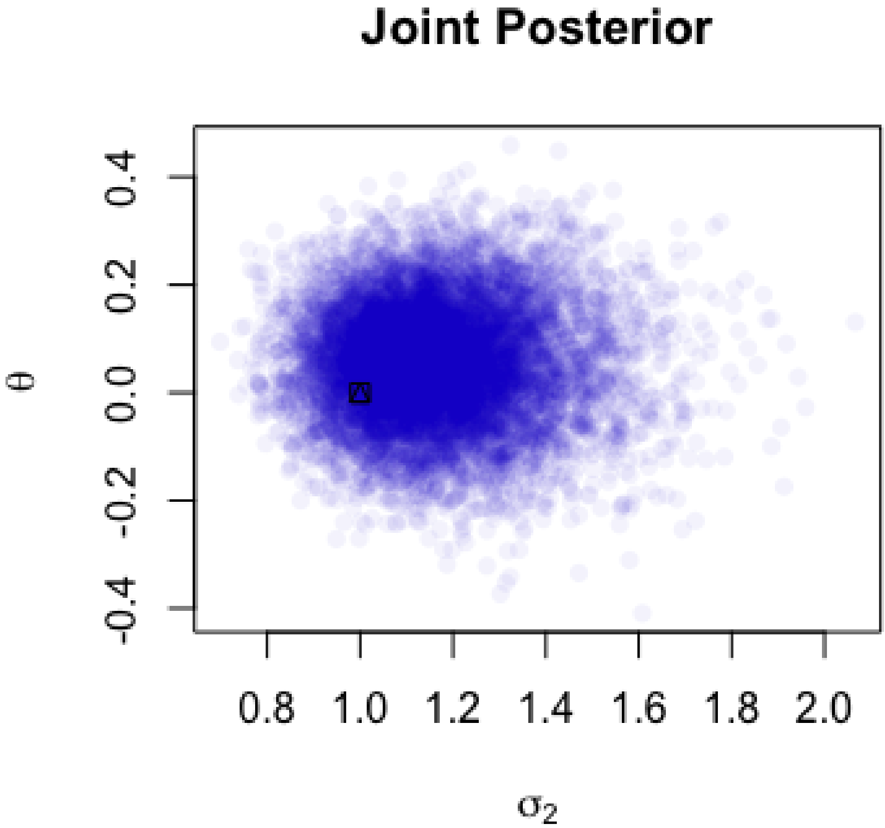
3. Repeat

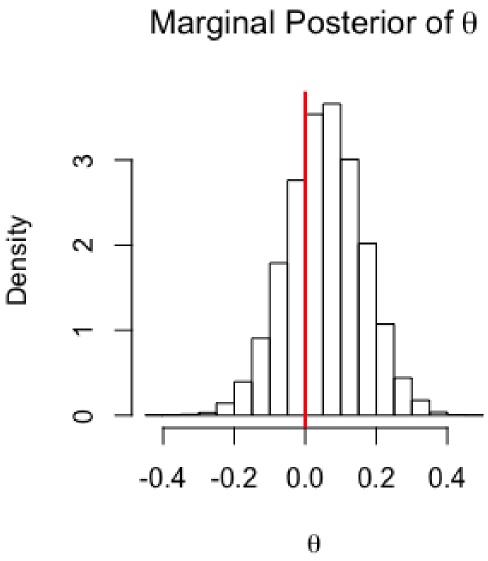
Note that each pair $\{\sigma_i^2, \theta_i\}$ is a sample from the joint posterior distibution and that $\{\sigma_1^2, \ldots, \sigma_m^2\}$ and $\{\theta_1, \ldots, \theta_m\}$ are samples from the respective marginal posterior distributions.
The R code for this follows as:

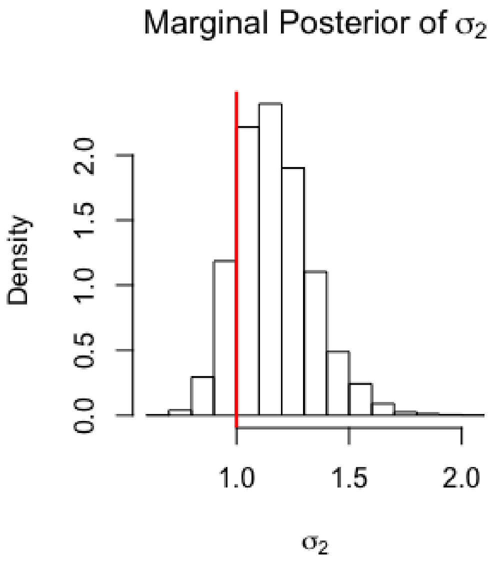
```
 #### Posterior Sampling with Normal Model
set.seed(09222016)
# true parameters from normal distribution
sigma.sq.true <- 1
theta.true <- 0

# 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 <- 1
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 <- 10000
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))
}

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')
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.


Joint Posterior
θ
0.4
0.2
0.0
-0.2
-0.4
0.8
1.0
1.2
1.4
1.6
1.8
2.0
σ_2


Marginal Posterior of θ
Density
3
2
1
0
-0.4
-0.2
0.0
0.2
0.4
θ


Marginal Posterior of σ_2
Density
2.0
1.5
1.0
0.5
0.0
1.0
1.5
2.0
σ_2

Posterior Sampling with the Gibbs Sampler

In the previous section we modeled the uncertainty in θ as a function of σ^2, where $p(\theta|\sigma^2) =$. In some situations this makes sense, but in others the uncertainty in θ may be specified independently from σ^2. Mathematically, this translates to $p(\sigma^2, \theta) =$. A common *semiconjugate* set of prior distributions is:

$$\theta \sim$$

$$1/\sigma^2 \sim$$

Note this prior on $1/\sigma^2$ is equivalent to saying $p(\sigma^2) \sim$ $(\nu_0, \nu_0, \sigma_0^2/2)$.
Now when $\{Y_1, \ldots, Y_n|\theta, \sigma^2\} \sim normal(\theta, \sigma^2)$ then $\theta|\sigma^2, y_1, \ldots, y_n \sim Normal(\mu_n, \tau_n^2)$.

$\mu_n =$ and $\tau_n^2 =$

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 $1/\sigma^2$ is not a gamma distribution or another named distribution that permits easy sampling.

Suppose that you know the value of θ. Then the conditional distribution of $\tilde{\sigma}^2 = (1/\sigma^2)$ is:

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

which is the kernel of a 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:

$$\theta^{(1)} \sim$$

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

$$\sigma^{2(2)} \sim$$

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 ,
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

2. sample

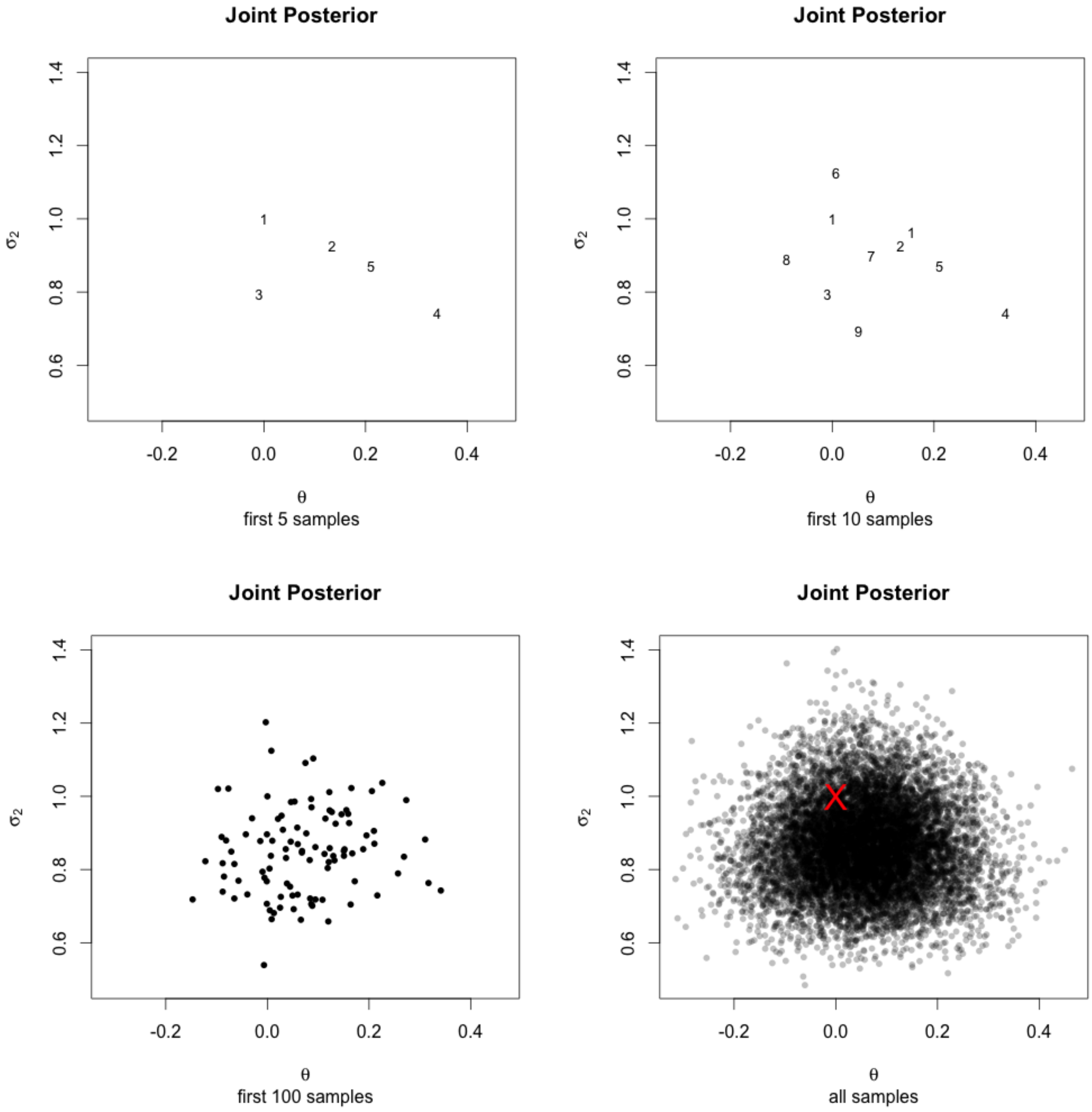
3. let

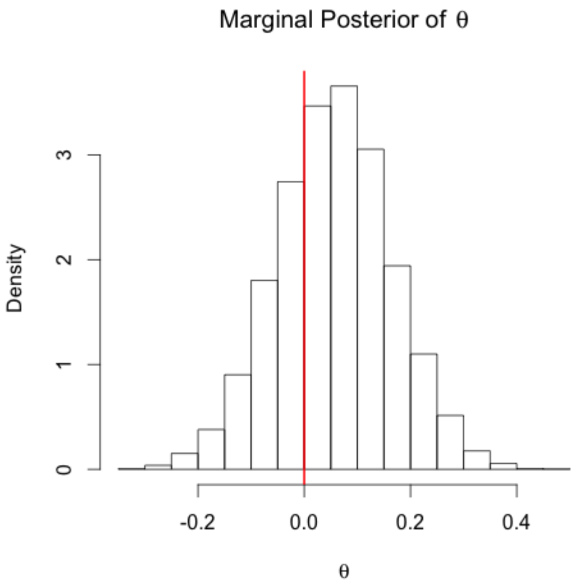
The code and R output for this follows.

```
> ######### First Gibbs Sampler
> set.seed(09222016)
> ### 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)
>
> ### initialize vectors and set starting values and priors
> num.sims <- 10000
> Phi <- matrix(0,nrow=num.sims,ncol=2)
> Phi[1,1] <- 0 # initialize theta
> Phi[1,2] <- 1 # initialize (1/sigmasq)
> mu.0 <- 0
> tausq.0 <- 1
> nu.0 <- 1
> sigmasq.0 <- 1
>
> 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 <- mean((y - Phi[i,1])^2)
+   Phi[i,2] <- rgamma(1,nu.n/2,nu.n*sigmasq.n.theta/2)
+ }
>
> # plot joint posterior
> plot(Phi[1:5,1],Phi[1:5,2],xlim=range(Phi[,1]),ylim=range(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],Phi[1:10,2],xlim=range(Phi[,1]),ylim=range(Phi[,2]),
pch=as.character(1:15),cex=.8, ylab=expression(sigma[2]),
xlab = expression(theta), main=’Joint Posterior’,sub=’first 10 samples’)
>
```

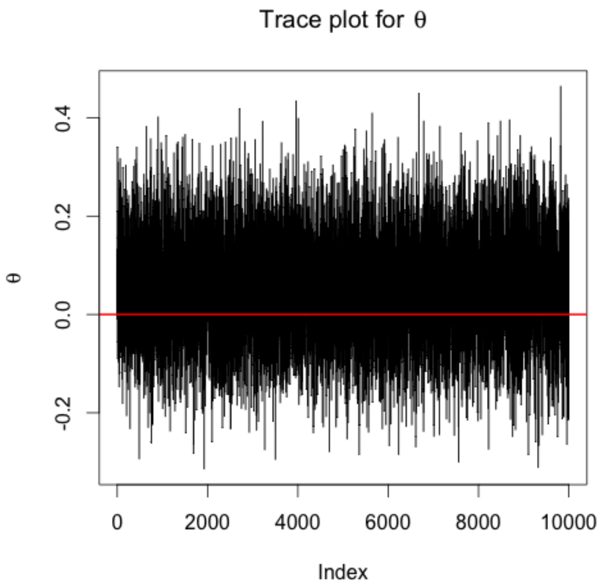
```
> plot(Phi[1:100,1],Phi[1:100,2],xlim=range(Phi[,1]),ylim=range(Phi[,2]),
pch=16,col=rgb(0,0,0,1),cex=.8,ylab=expression(sigma[2]),
xlab = expression(theta), main='Joint Posterior',sub='first 100 samples')
>
> plot(Phi[,1],Phi[,2],xlim=range(Phi[,1]),ylim=range(Phi[,2]),pch=16,
col=rgb(0,0,0,.25),cex=.8, ylab=expression(sigma[2]), xlab =
expression(theta),  main='Joint Posterior',sub='all samples')
> points(0,1,pch='X',col='red',cex=2)
>
> # plot marginal posterior of theta
> hist(Phi[,1],xlab=expression(theta),main=
expression('Marginal Posterior of ' ~ theta),probability=T)
> abline(v=0,col='red',lwd=2)
> # plot marginal posterior of sigmasq
> hist(1/Phi[,2],xlab=expression(sigma[2]),main=expression(
'Marginal Posterior of ' ~ sigma[2]),probability=T)
> abline(v=1,col='red',lwd=2)
>
> # plot trace plots
> plot(Phi[,1],type='l',ylab=expression(theta), main=expression('
Trace plot for ' ~ theta))
> abline(h=0,lwd=2,col='red')
> plot(Phi[,2],type='l',ylab=expression(sigma[2]), main=expression('
Trace plot for ' ~ sigma[2]))
> abline(h=1,lwd=2,col='red')
>
> # compute posterior mean and quantiles
> colMeans(Phi)
[1] 0.05603744 0.86535524
> apply(Phi,2,quantile,probs=c(.025,.975))
            [,1]      [,2]
2.5%  -0.1559274 0.6440163
97.5%  0.2691388 1.1200151
```

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).

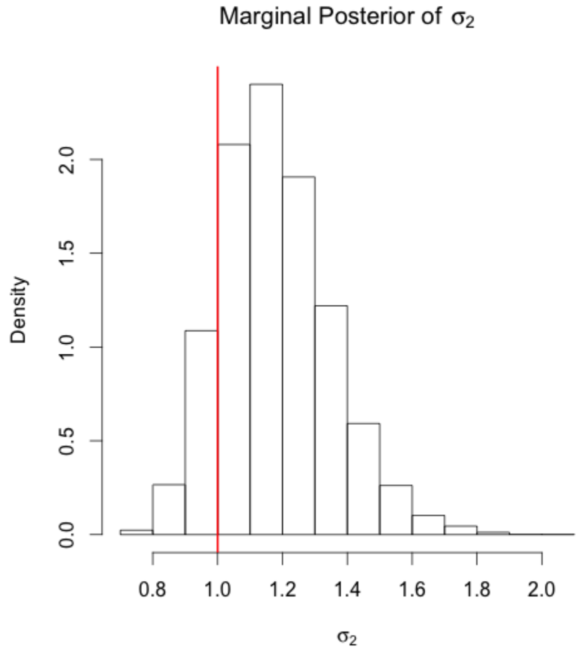

Joint Posterior
σ2
1.4
1.2
1.0
0.8
0.6
-0.2
0.0
0.2
0.4
θ
first 5 samples
1
2
5
3
4
Joint Posterior
6
1
1
2
8
7
5
3
4
9
θ
first 10 samples
Joint Posterior
θ
first 100 samples
Joint Posterior
X
θ
all samples



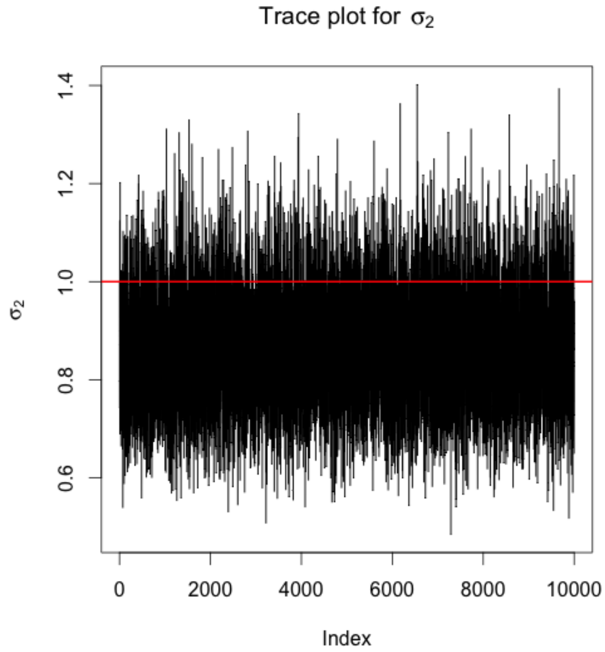
Marginal Posterior of θ



Trace plot for θ



Marginal Posterior of σ_2



Trace plot for σ_2