STA 360/602L: Module 4.6

Missing data and imputation II

Dr. Olanrewaju Michael Akande

STA 602L

Bayesian inference with missing data

- As we have seen, for MCAR and MAR, we can focus on $p(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}, \Sigma)$ in the likelihood function, when inferring $(\boldsymbol{\theta}, \Sigma)$.

- While this is great, for posterior sampling under most models (especially multivariate models), we actually do need all the $\boldsymbol{Y}$'s to update the parameters.

- In addition, we may actually want to learn about the missing values, in addition to inferring $(\boldsymbol{\theta}, \Sigma)$.

- By thinking of the missing data as **another set of parameters**, we can sample them from the "posterior predictive" distribution of the missing data conditional on the observed data and parameters:

$$p(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{\theta}, \Sigma) \propto \prod_{i=1}^{n} p(\boldsymbol{Y}_{i,mis}|\boldsymbol{Y}_{i,obs}, \boldsymbol{\theta}, \Sigma).$$

- In the case of the multivariate model, each $p(\boldsymbol{Y}_{i,mis}|\boldsymbol{Y}_{i,obs}, \boldsymbol{\theta}, \Sigma)$ is just a normal distribution, and we can leverage results on conditional distributions for normal models.

STA 602L

Gibbs sampler with missing data

At iteration $s+1$, do the following

1. Sample $\boldsymbol{\theta}^{(s+1)}$ from its multivariate normal full conditional

$$p(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(s)}, \Sigma^{(s)}).$$

2. Sample $\Sigma^{(s+1)}$ from its inverse-Wishart full conditional

$$p(\Sigma^{(s+1)}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(s)}, \boldsymbol{\theta}^{(s+1)}).$$

3. For each $i = 1, \ldots, n$, with at least one "1" value in the missingness indicator vector $\boldsymbol{R}_i$, sample $\boldsymbol{Y}_{i,mis}^{(s+1)}$ from the full conditional

$$p(\boldsymbol{Y}_{i,mis}^{(s+1)}|\boldsymbol{Y}_{i,obs}, \boldsymbol{\theta}^{(s+1)}, \Sigma^{(s+1)}),$$

   which is also multivariate normal, with its form derived from the original sampling model but with the updated parameters, that is, $\boldsymbol{Y}_i^{(s+1)} = (\boldsymbol{Y}_{i,obs}, \boldsymbol{Y}_{i,mis}^{(s+1)})^T \sim \mathcal{N}_p(\boldsymbol{\theta}^{(s+1)}, \Sigma^{(s+1)})$.

STA 602L

Gibbs sampler with missing data

- Rewrite $\boldsymbol{Y}_i^{(s+1)} = (\boldsymbol{Y}_{i,mis}, \boldsymbol{Y}_{i,obs}^{(s+1)})^T \sim \mathcal{N}_p(\boldsymbol{\theta}^{(s+1)}, \Sigma^{(s+1)})$ as

$$\boldsymbol{Y}_i = \begin{pmatrix} \boldsymbol{Y}_{i,mis} \\ \boldsymbol{Y}_{i,obs} \end{pmatrix} \sim \mathcal{N}_p \left[\begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right],$$

  so that we can take advantage of the conditional normal results.

- That is, we have

$$\boldsymbol{Y}_{i,mis} | \boldsymbol{Y}_{i,obs} = \boldsymbol{y}_{i,obs} \sim \mathcal{N}\left(\boldsymbol{\theta}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{y}_{i,obs} - \boldsymbol{\theta}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right).$$

  as the multivariate normal distribution (or univariate normal distribution if $\boldsymbol{Y}_i$ only has one missing entry) we need in step 3 of the Gibbs sampler.

- This sampling technique actually encodes MAR since the imputations for $\boldsymbol{Y}_{mis}$ depend on the $\boldsymbol{Y}_{obs}$.

- Now let's revisit the reading comprehension example again. We will add missing values to the original data and refit the model.

STA 602L

Reading example with missing data

```
Y <- as.matrix(dget("http://www2.stat.duke.edu/~pdh10/FCBS/Inline/Y.reading"))

#Add 20% missing data; MCAR
set.seed(1234)
Y_WithMiss <- Y #So we can keep the full data
Miss_frac <- 0.20
R <- matrix(rbinom(nrow(Y_WithMiss)*ncol(Y_WithMiss),1,Miss_frac),
            nrow(Y_WithMiss),ncol(Y_WithMiss))
Y_WithMiss[R==1]<-NA
Y_WithMiss[1:12,]
```

```
##       pretest posttest
##  [1,]      59       77
##  [2,]      43       39
##  [3,]      34       46
##  [4,]      32       NA
##  [5,]      NA       38
##  [6,]      38       NA
##  [7,]      55       NA
##  [8,]      67       86
##  [9,]      64       77
## [10,]      45       60
## [11,]      49       50
## [12,]      72       59
```

```
colMeans(is.na(Y_WithMiss))
```

```
##   pretest  posttest
## 0.1363636 0.2272727
```

STA 602L

Reading example with missing data

```
#ACTUAL ANALYSIS STARTS HERE!!!
#Data dimensions
n <- nrow(Y_WithMiss); p <- ncol(Y_WithMiss)

#Hyperparameters for the priors
mu_0 <- c(50,50)
Lambda_0 <- matrix(c(156,78,78,156),nrow=2,ncol=2)
nu_0 <- 4
S_0 <- matrix(c(625,312.5,312.5,625),nrow=2,ncol=2)

#Define missing data indicators
##we already know R. This is to write a more general code for when we don't
R <- 1*(is.na(Y_WithMiss))
R[1:12,]
```

```
##       pretest posttest
##  [1,]       0        0
##  [2,]       0        0
##  [3,]       0        0
##  [4,]       0        1
##  [5,]       1        0
##  [6,]       0        1
##  [7,]       0        1
##  [8,]       0        0
##  [9,]       0        0
## [10,]       0        0
## [11,]       0        0
## [12,]       0        0
```

STA 602L

Reading example with missing data

```r
#Initial values for Gibbs sampler
Y_Full <- Y_WithMiss #So we can keep the data with missing values as is
for (j in 1:p) {
Y_Full[is.na(Y_Full[,j]),j] <- mean(Y_Full[,j],na.rm=TRUE) #start with mean imputation
}

Sigma <- S_0 # can't really rely on cov(Y) because we don't have full Y

#Set null objects to save samples
THETA_WithMiss <- NULL
SIGMA_WithMiss <- NULL
Y_MISS <- NULL

#first set number of iterations and burn-in, then set seed
n_iter <- 10000; burn_in <- 0.3*n_iter
```

STA 602L

Gibbs sampler with missing data

```
#library(mvtnorm) for multivariate normal
#library(MCMCpack) for inverse-Wishart

Lambda_0_inv <- solve(Lambda_0) #move outside sampler since it does not change

for (s in 1:(n_iter+burn_in)){
  ##first we must recalculate ybar inside the loop now since it changes every iteration
  ybar <- apply(Y_Full,2,mean)

  ##update theta
  Sigma_inv <- solve(Sigma) #invert once
  Lambda_n <- solve(Lambda_0_inv + n*Sigma_inv)
  mu_n <- Lambda_n %*% (Lambda_0_inv%*%mu_0 + n*Sigma_inv%*%ybar)
  theta <- rmvnorm(1,mu_n,Lambda_n)

  ##update Sigma
  S_theta <- (t(Y_Full)-c(theta))%*%t(t(Y_Full)-c(theta))
  S_n <- S_0 + S_theta
  nu_n <- nu_0 + n
  Sigma <- riwish(nu_n, S_n)
```

STA 602L

Gibbs sampler with missing data

```
##update missing data using updated draws of theta and Sigma
  for(i in 1:n) {
    if(sum(R[i,]>0)){
       obs_index <- R[i,]==0
       mis_index <- R[i,]==1
       Sigma_22_obs_inv <- solve(Sigma[obs_index,obs_index]) #invert just once
       Sigma_12_Sigma_22_obs_inv <- Sigma[mis_index,obs_index]%*%Sigma_22_obs_inv

       Sigma_cond_mis <- Sigma[mis_index,mis_index] -
         Sigma_12_Sigma_22_obs_inv%*%Sigma[obs_index,mis_index]

       mu_cond_mis <- theta[mis_index] +
         Sigma_12_Sigma_22_obs_inv%*%(t(Y_Full[i,obs_index])-theta[obs_index])

      Y_Full[i,mis_index] <- rmvnorm(1,mu_cond_mis,Sigma_cond_mis)
      }
    }

  #save results only past burn-in
  if(s > burn_in){
  THETA_WithMiss <- rbind(THETA_WithMiss,theta)
  SIGMA_WithMiss <- rbind(SIGMA_WithMiss,c(Sigma))
  Y_MISS <- rbind(Y_MISS, Y_Full[R==1] )
  }
}

colnames(THETA_WithMiss) <- c("theta_1","theta_2")
colnames(SIGMA_WithMiss) <- c("sigma_11","sigma_12","sigma_21","sigma_22") #symmetry in s
```

STA 602L

Diagnostics

```r
#library(coda)
THETA_WithMiss.mcmc <- mcmc(THETA_WithMiss,start=1); summary(THETA_WithMiss.mcmc)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##          Mean    SD Naive SE Time-series SE
## theta_1 45.64 3.012  0.03012        0.03276
## theta_2 54.15 3.453  0.03453        0.03939
##
## 2. Quantiles for each variable:
##
##          2.5%   25%   50%   75% 97.5%
## theta_1 39.60 43.65 45.62 47.64 51.55
## theta_2 47.31 51.91 54.17 56.45 61.08
```

STA 602L

Diagnostics

```
SIGMA_WithMiss.mcmc <- mcmc(SIGMA_WithMiss,start=1); summary(SIGMA_WithMiss.mcmc)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean    SD Naive SE Time-series SE
## sigma_11 194.8 62.89   0.6289         0.6063
## sigma_12 152.1 60.58   0.6058         0.6910
## sigma_21 152.1 60.58   0.6058         0.6910
## sigma_22 247.7 83.55   0.8355         0.9659
##
## 2. Quantiles for each variable:
##
##            2.5%   25%   50%   75% 97.5%
## sigma_11 108.30 151.2 182.5 224.4 348.6
## sigma_12  64.76 110.3 141.9 182.0 299.6
## sigma_21  64.76 110.3 141.9 182.0 299.6
## sigma_22 133.33 189.3 231.8 289.0 450.8
```

STA 602L

Compare to inference from full data

With missing data:

```
apply(THETA_WithMiss,2,summary)
```

```
##          theta_1  theta_2
## Min.    32.64839 41.13748
## 1st Qu. 43.65457 51.90859
## Median  45.61740 54.16720
## Mean    45.63740 54.14929
## 3rd Qu. 47.64129 56.45068
## Max.    58.65830 70.26826
```

Based on true data:

```
apply(THETA,2,summary)
```

```
##          theta_1  theta_2
## Min.    35.50314 37.80999
## 1st Qu. 45.35465 51.53327
## Median  47.36177 53.68602
## Mean    47.29978 53.68529
## 3rd Qu. 49.22875 55.82192
## Max.    60.94924 69.92354
```

Very similar for the most part.

Compare to inference from full data

With missing data:

```
apply(SIGMA_WithMiss,2,summary)
```

```
##            sigma_11  sigma_12  sigma_21  sigma_22
## Min.       74.61274 -10.83674 -10.83674  82.55346
## 1st Qu. 151.17000 110.33973 110.33973 189.31667
## Median  182.49663 141.85462 141.85462 231.76447
## Mean    194.75107 152.14494 152.14494 247.72255
## 3rd Qu. 224.42867 181.98838 181.98838 288.99033
## Max.    712.33562 600.36262 600.36262 960.62283
```

Based on true data:

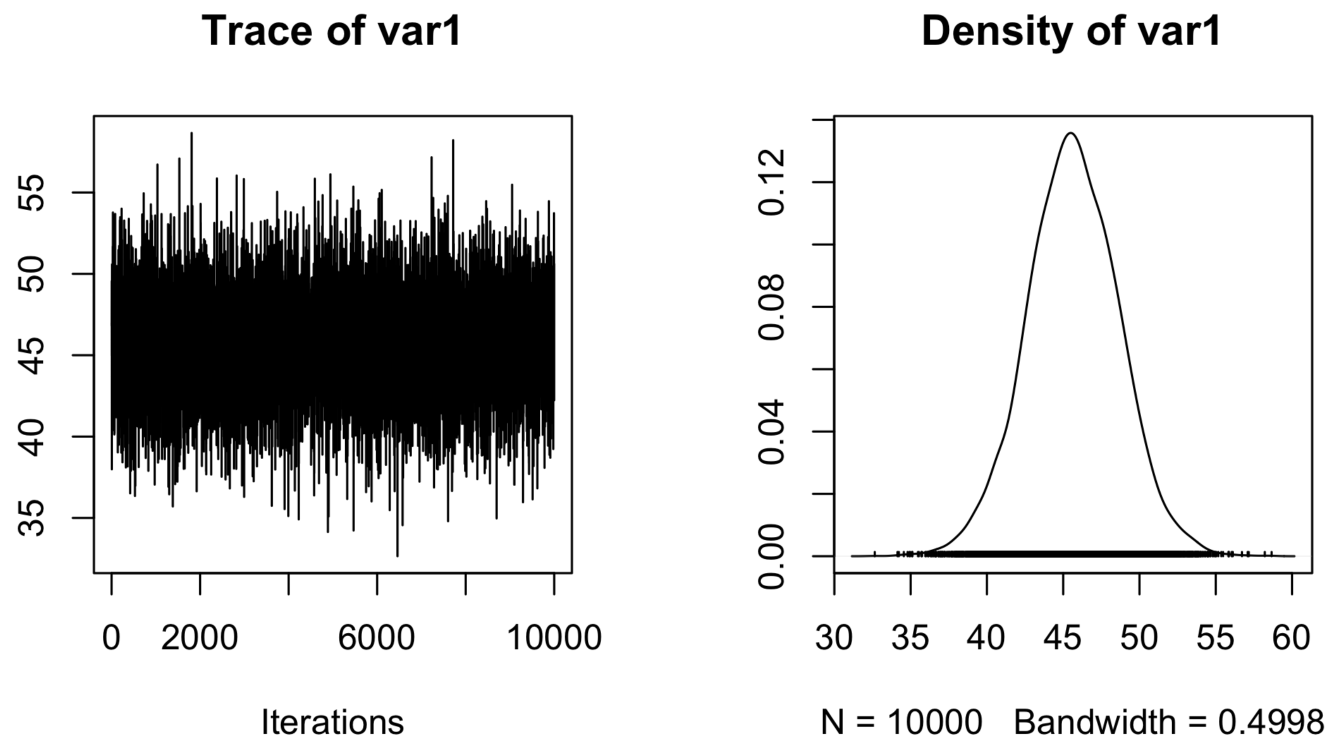
```
apply(SIGMA,2,summary)
```

```
##            sigma_11  sigma_12  sigma_21  sigma_22
## Min.       79.44258  11.41663  11.41663  93.65776
## 1st Qu. 158.21469 113.23258 113.23258 203.21138
## Median  190.77854 144.74881 144.74881 244.56334
## Mean    202.34721 155.33355 155.33355 260.07072
## 3rd Qu. 234.77319 186.50429 186.50429 300.90761
## Max.    671.16538 613.88088 613.88088 947.39333
```

Also very similar. A bit more uncertainty in dimension of Y_{i2} because we have more missing data there.

Diagnostics: trace plots

```
plot(THETA_WithMiss.mcmc[,"theta_1"])
```

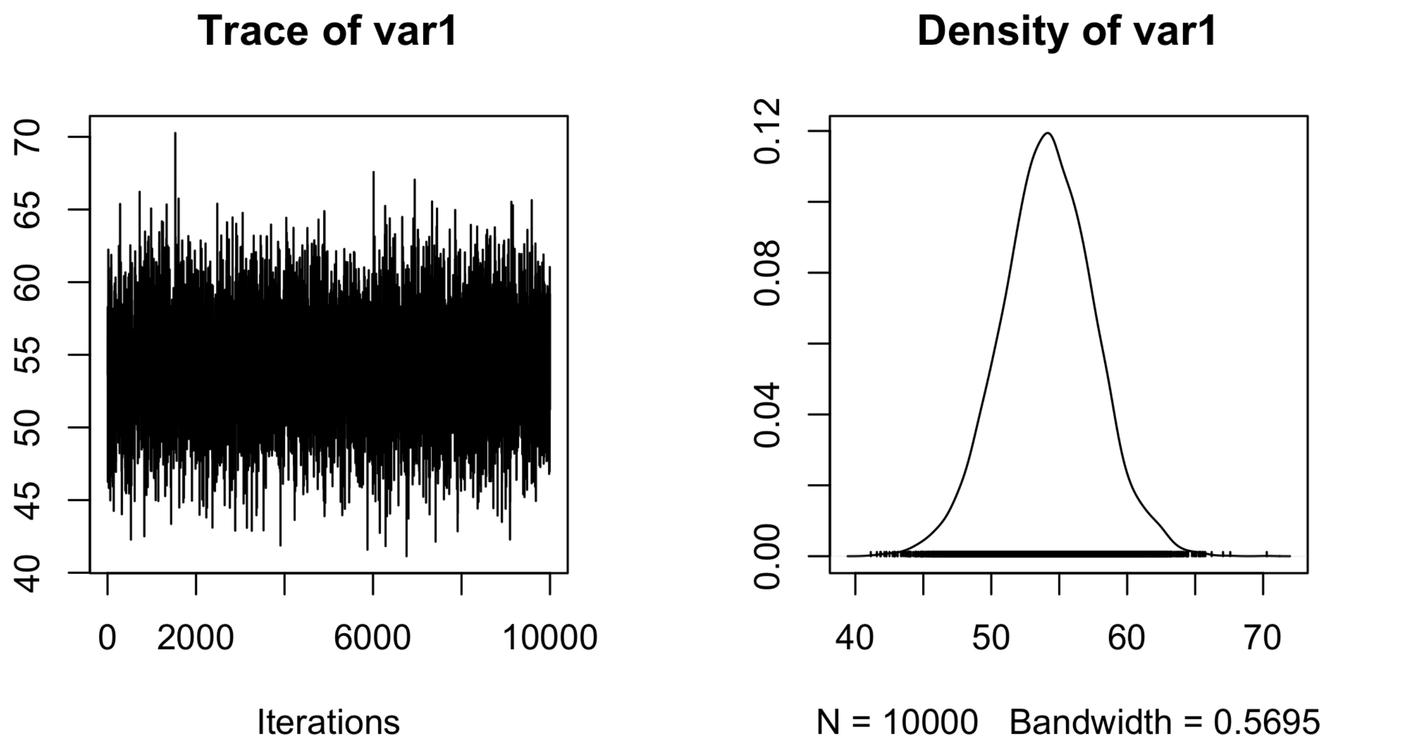


Looks good!

STA 602L

Diagnostics: trace plots

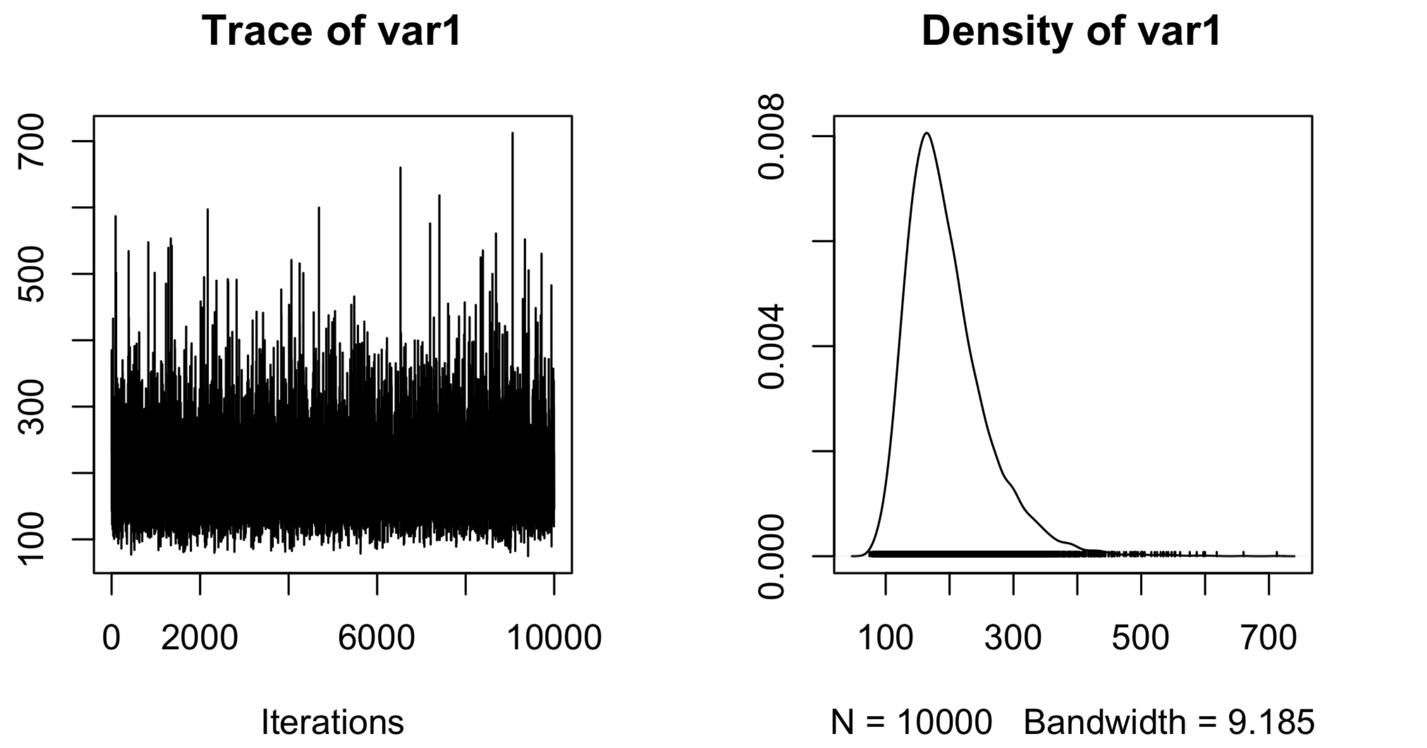
```
plot(THETA_WithMiss.mcmc[,"theta_2"])
```



Looks good!

Diagnostics: trace plots

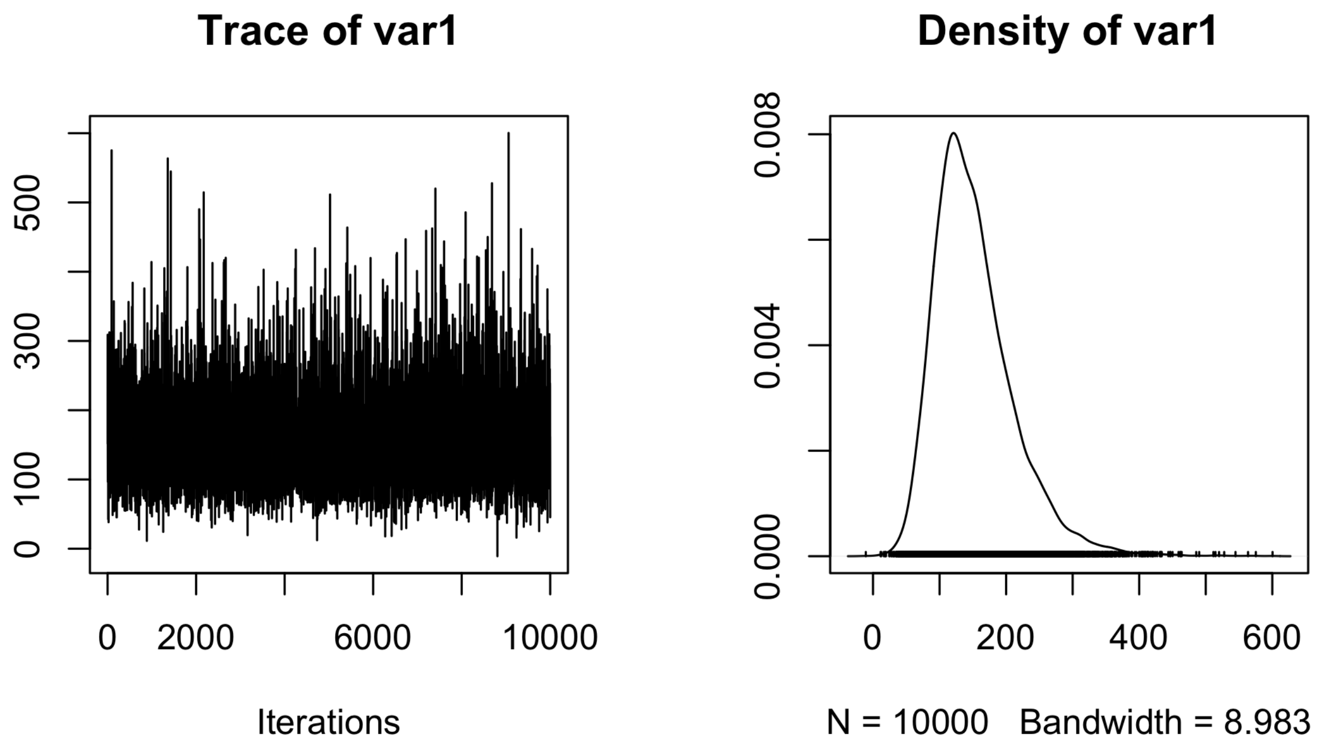
```
plot(SIGMA_WithMiss.mcmc[,"sigma_11"])
```



Looks good!

Diagnostics: trace plots

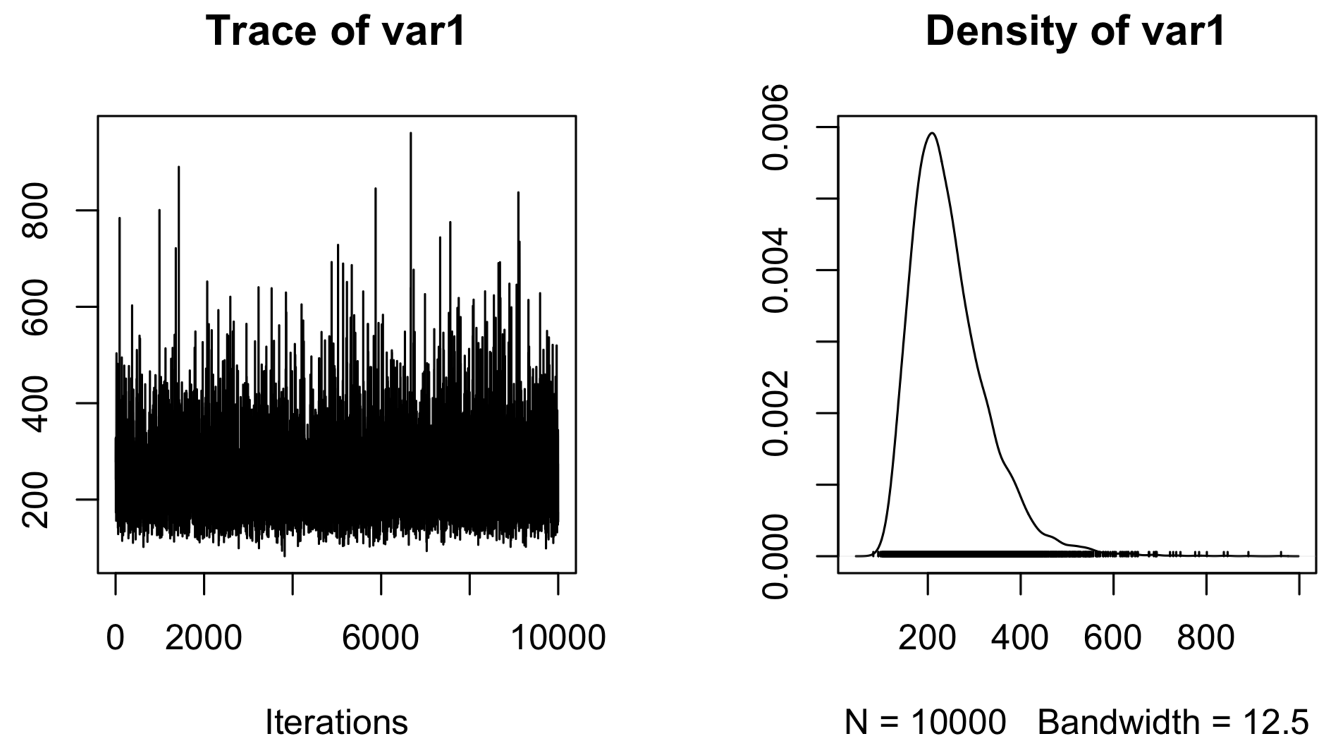
```
plot(SIGMA_WithMiss.mcmc[,"sigma_12"])
```



Looks good!

Diagnostics: trace plots

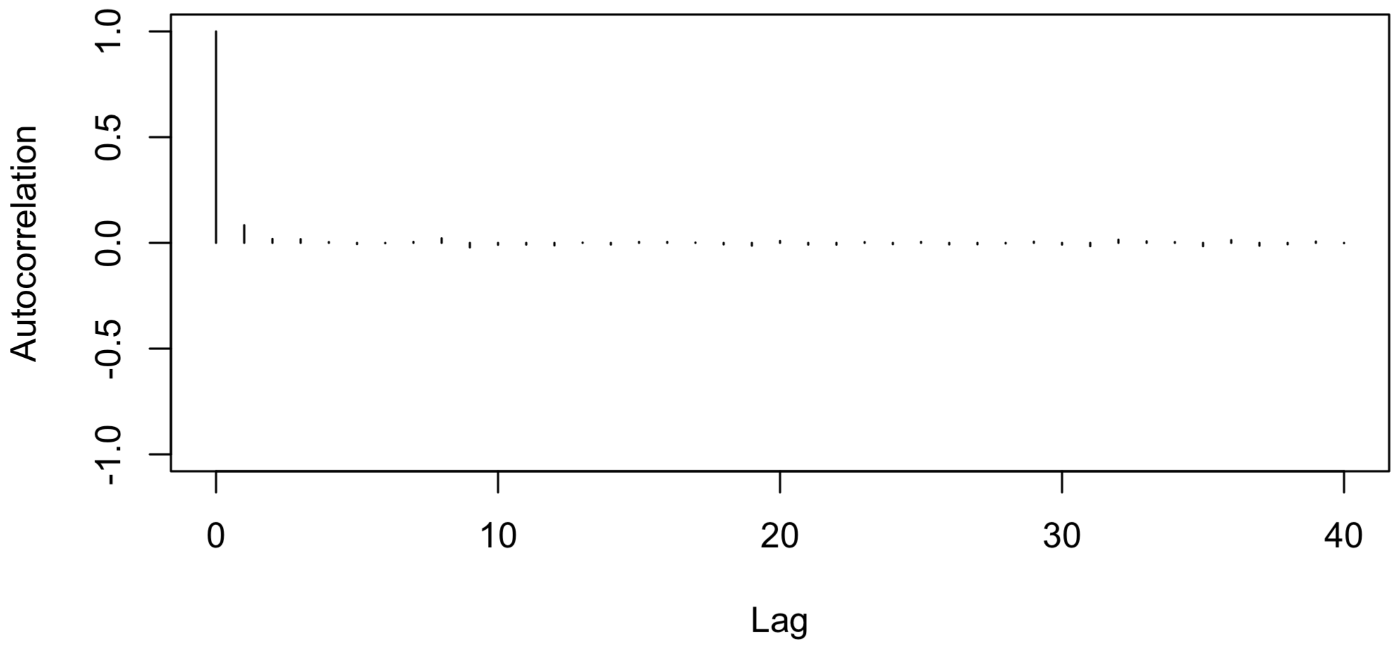
```
plot(SIGMA_WithMiss.mcmc[,"sigma_22"])
```



Looks good!

Diagnostics: autocorrelation

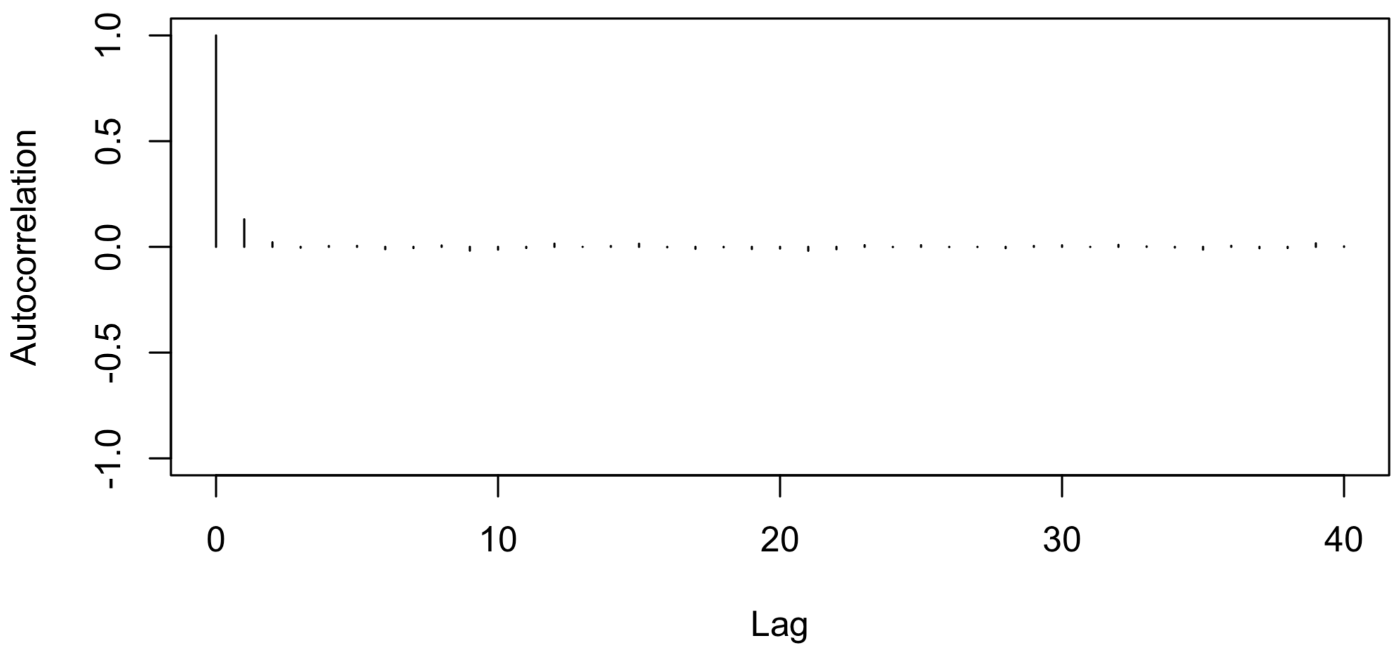
```
autocorr.plot(THETA_WithMiss.mcmc[,"theta_1"])
```



Looks good!

Diagnostics: autocorrelation

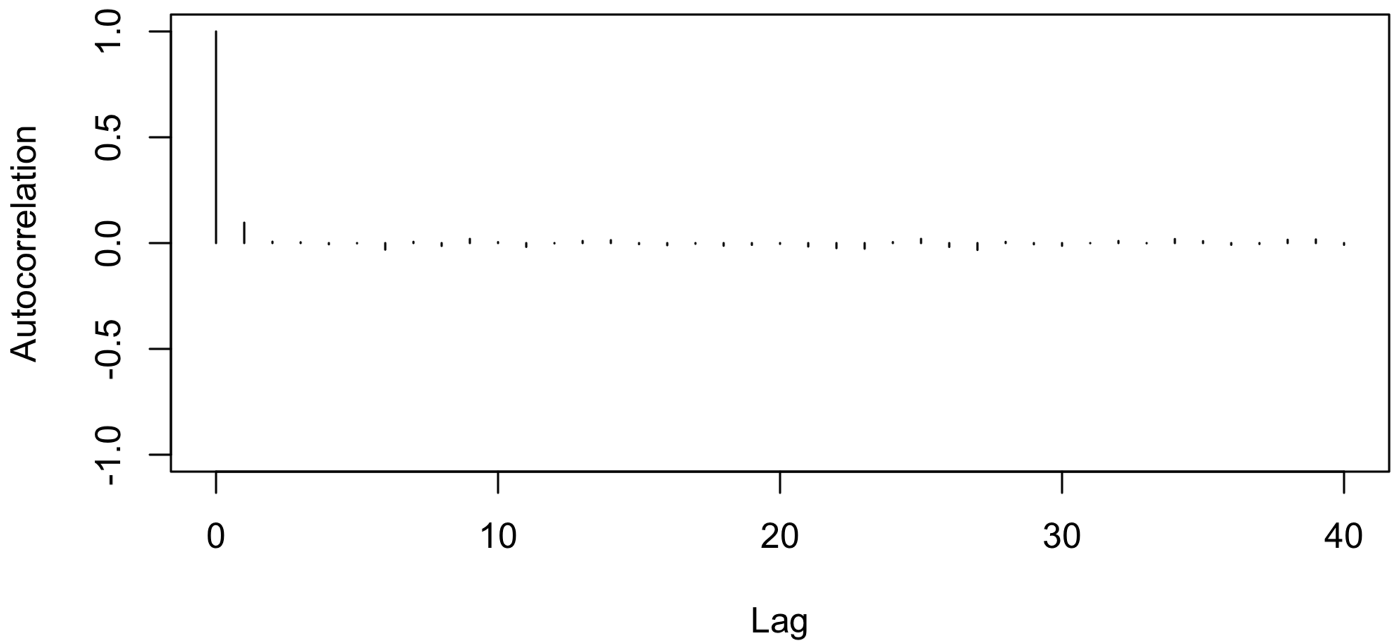
```
autocorr.plot(THETA_WithMiss.mcmc[,"theta_2"])
```



Looks good!

Diagnostics: autocorrelation

```
autocorr.plot(SIGMA_WithMiss.mcmc[,"sigma_11"])
```

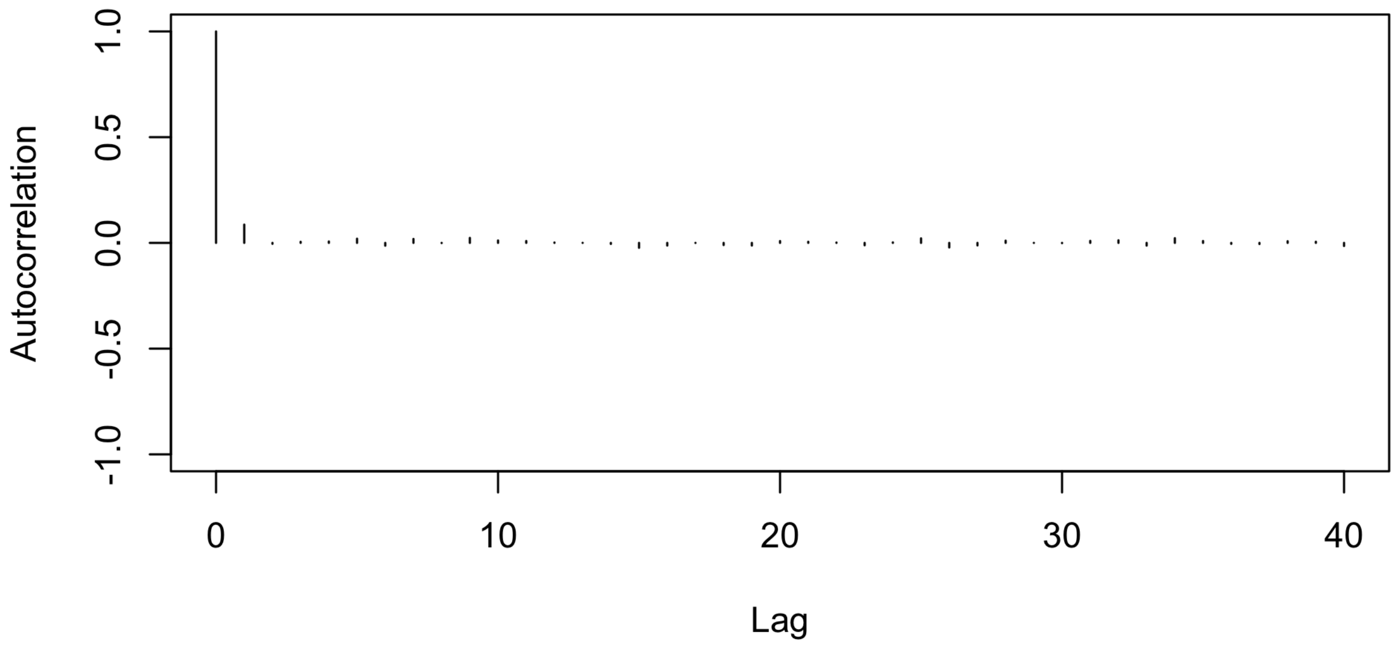


Looks good!

STA 602L

Diagnostics: autocorrelation

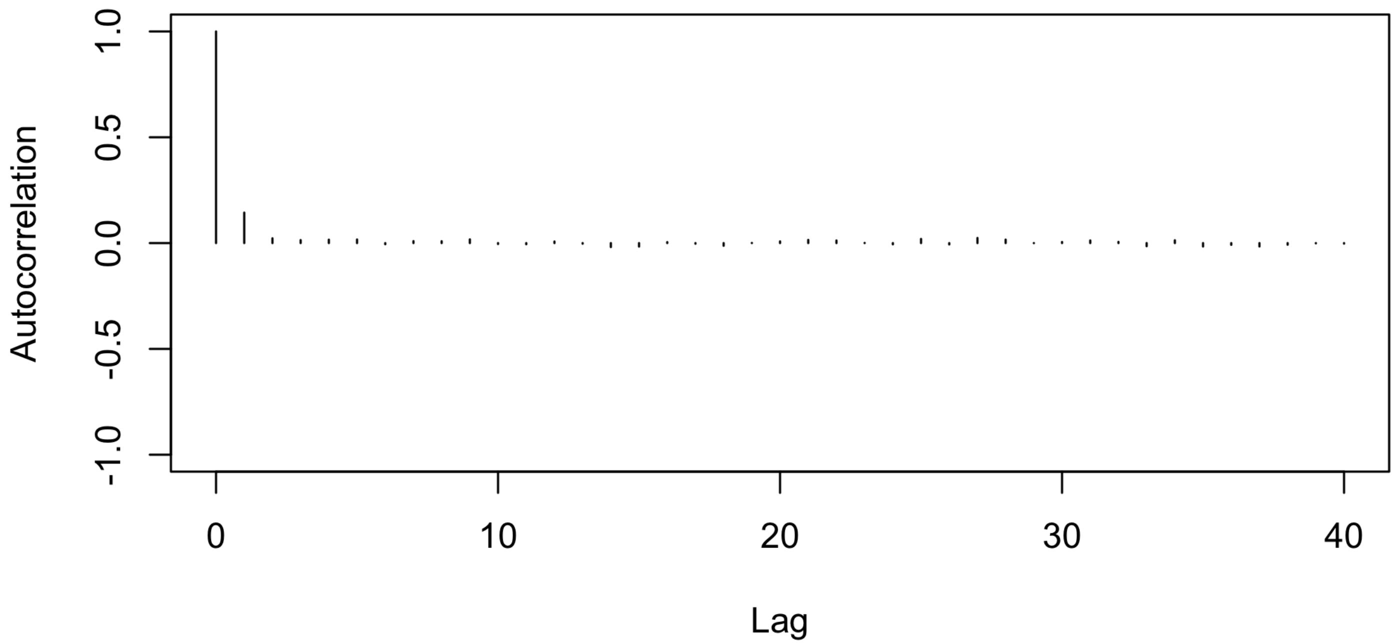
```
autocorr.plot(SIGMA_WithMiss.mcmc[,"sigma_12"])
```



Looks good!

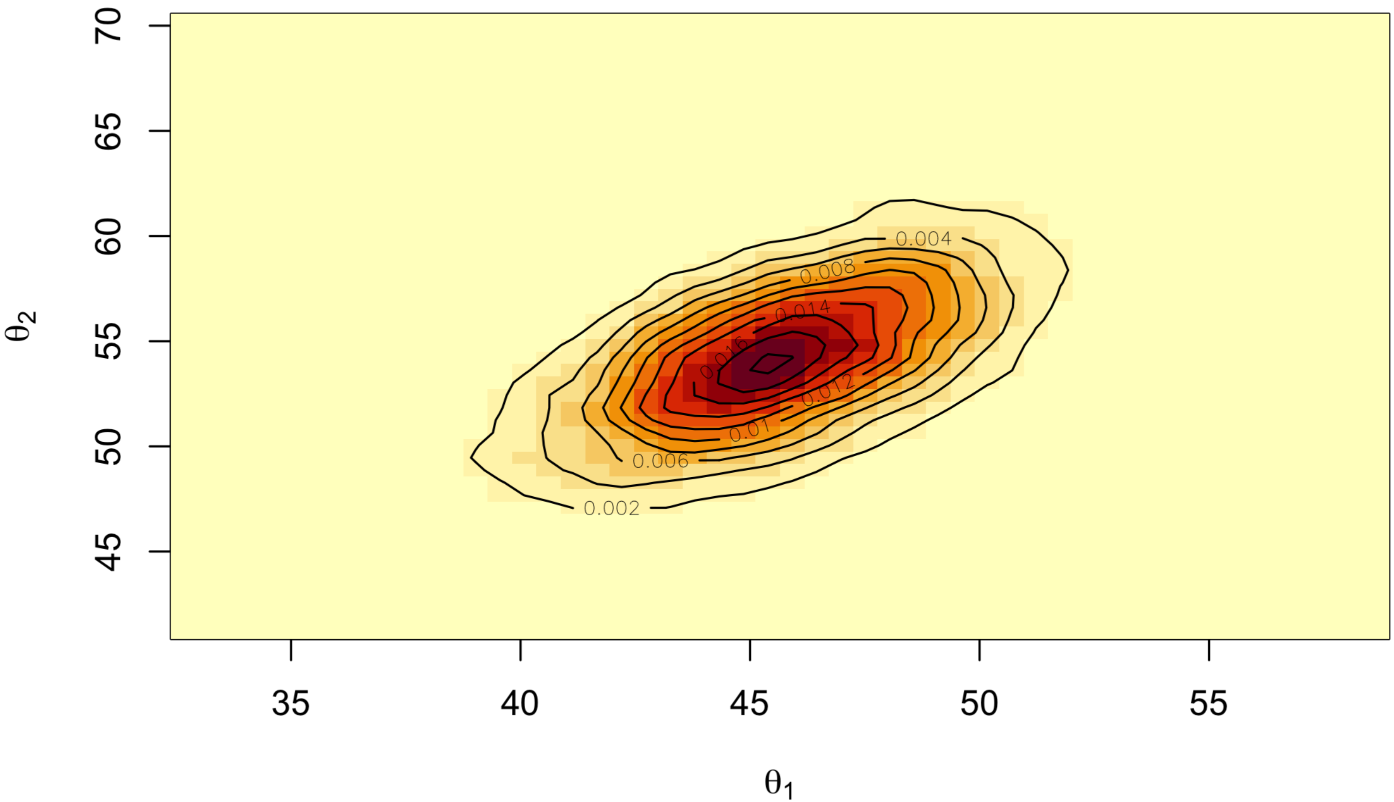
Diagnostics: autocorrelation

```
autocorr.plot(SIGMA_WithMiss.mcmc[,"sigma_22"])
```



Looks good!

Posterior distribution of the mean



STA 602L

MISSING DATA VS PREDICTIONS FOR NEW OBSERVATIONS

- How about predictions for completely new observations?
- That is, suppose your original dataset plus sampling model is $\boldsymbol{y_i} = (y_{i,1}, y_{i,2})^T \sim \mathcal{N}_2(\boldsymbol{\theta}, \Sigma)$, $i = 1, \ldots, n$.
- Suppose now you have $n^\star$ new observations with $y_2^\star$ values but no $y_1^\star$.
- How can we predict $y_{i,1}^\star$ given $y_{i,2}^\star$, for $i = 1, \ldots, n^\star$?
- Well, we can view this as a "train $\rightarrow$ test" prediction problem rather than a missing data problem on an original data.

STA 602L

Missing data vs predictions for new observations

- That is, given the posterior samples of the parameters, and the test values for $y_{i2}^{\star}$, draw from the posterior predictive distribution of $(y_{i,1}^{\star}|y_{i,2}^{\star}, \{(y_{1,1}, y_{1,2}), \ldots, (y_{n,1}, y_{n,2})\})$.

- To sample from this predictive distribution, think of compositional sampling.

- That is, for each posterior sample of $(\boldsymbol{\theta}, \Sigma)$, sample from $(y_{i,1}|y_{i,2}, \boldsymbol{\theta}, \Sigma)$, which is just from the form of the sampling distribution.

- In this case, $(y_{i,1}|y_{i,2}, \boldsymbol{\theta}, \Sigma)$ is just a normal distribution derived from $(y_{i,1}, y_{i,2}|\boldsymbol{\theta}, \Sigma)$, based on the conditional normal formula.

- No need to incorporate the prediction problem into your original Gibbs sampler!

What's next?

Move on to the readings for the next module!

STA 602L