Multivariate normal model cont'd

Dr. Olanrewaju Michael Akande

Feb 21, 2020

Announcements

- Homework 5 will be online by 5pm today.

Outline

- Chat on survey responses
- Multivariate normal/Gaussian model
  - Inference for mean (recap)
  - Inference for covariance
  - Back to the example
  - Gibbs sampler
  - Jeffreys' prior

Multivariate normal model

Conditional inference on mean recap

- For data $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{ip})^T \sim \mathcal{N}_p(\boldsymbol{\theta}, \Sigma)$,

$$L(\boldsymbol{Y}; \boldsymbol{\theta}, \Sigma) = \prod_{i=1}^{n} (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta}) \right\}$$
$$\propto \exp\left\{ -\frac{1}{2} \boldsymbol{\theta}^T (n\Sigma^{-1}) \boldsymbol{\theta} + \boldsymbol{\theta}^T (n\Sigma^{-1} \bar{\boldsymbol{y}}) \right\}.$$

- If we assume $\pi(\boldsymbol{\theta}) = \mathcal{N}_p(\boldsymbol{\mu}_0, \Lambda_0)$, that is,

$$\pi(\boldsymbol{\theta}) \propto \exp\left\{ -\frac{1}{2} \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\theta} + \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\mu}_0 \right\}$$

- Then

$$\pi(\boldsymbol{\theta}|\Sigma, \boldsymbol{Y}) \propto \exp\left\{ -\frac{1}{2} \boldsymbol{\theta}^T \left[\Lambda_0^{-1} + n\Sigma^{-1} \right] \boldsymbol{\theta} + \boldsymbol{\theta}^T \left[\Lambda_0^{-1} \boldsymbol{\mu}_0 + n\Sigma^{-1} \bar{\boldsymbol{y}} \right] \right\} \equiv \mathcal{N}_p(\boldsymbol{\mu}_n, \Lambda_n)$$

where

$$\Lambda_n = \left[\Lambda_0^{-1} + n\Sigma^{-1} \right]^{-1}$$
$$\boldsymbol{\mu}_n = \Lambda_n \left[\Lambda_0^{-1} \boldsymbol{\mu}_0 + n\Sigma^{-1} \bar{\boldsymbol{y}} \right].$$

Conditional inference on mean recap

- As in the univariate case, we once again have that
  - Posterior precision is sum of prior precision and data precision:

$$\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}$$

  - Posterior expectation is weighted average of prior expectation and the sample mean:

$$\begin{aligned}\boldsymbol{\mu}_n &= \Lambda_n \left[\Lambda_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\boldsymbol{y}}\right] \\ &= \overbrace{\left[\Lambda_n\Lambda_0^{-1}\right]}^{\text{weight on prior mean}} \underbrace{\boldsymbol{\mu}_0}_{\text{prior mean}} + \overbrace{\left[\Lambda_n(n\Sigma^{-1})\right]}^{\text{weight on sample mean}} \underbrace{\bar{\boldsymbol{y}}}_{\text{sample mean}}\end{aligned}$$

- Compare these to the results from the univariate case to gain more intuition.

What about the covariance matrix?

- A random variable $\Sigma \sim \mathcal{IW}_p(\nu_0, \boldsymbol{S}_0)$, where Σ is positive definite and $p \times p$, has pdf

  $$p(\Sigma) \ \propto \ |\Sigma|^{\frac{-(\nu_0+p+1)}{2}} \exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{S}_0\Sigma^{-1})\right\},$$

  where

  - $\text{tr}(\cdot)$ is the **trace function** (sum of diagonal elements),
  - $\nu_0 > p - 1$ is the "degrees of freedom", and
  - $\boldsymbol{S}_0$ is a $p \times p$ positive definite matrix.

- For this distribution, $\mathbb{E}[\Sigma] = \dfrac{1}{\nu_0 - p - 1}\boldsymbol{S}_0$, for $\nu_0 > p + 1$.

- Hence, $\boldsymbol{S}_0$ is the scaled mean of the $\mathcal{IW}_p(\nu_0, \boldsymbol{S}_0)$.

Wishart distribution

- If we are very confidence in a prior guess Σ_0, for Σ, then we might set
  - ν_0, the degrees of freedom to be very large, and
  - $\boldsymbol{S}_0 = (\nu_0 - p - 1)\Sigma_0$.

  In this case, $\mathbb{E}[\Sigma] = \dfrac{1}{\nu_0 - p - 1}\boldsymbol{S}_0 = \dfrac{1}{\nu_0 - p - 1}(\nu_0 - p - 1)\Sigma_0 = \Sigma_0$, and Σ is tightly (depending on the value of ν_0) centered around Σ_0.

- If we are not at all confident but we still have a prior guess Σ_0, we might set
  - $\nu_0 = p + 2$, so that the $\mathbb{E}[\Sigma] = \dfrac{1}{\nu_0 - p - 1}\boldsymbol{S}_0$ is finite.
  - $\boldsymbol{S}_0 = \Sigma_0$

  Here, $\mathbb{E}[\Sigma] = \Sigma_0$ as before, but Σ is only loosely centered around Σ_0.

Wishart distribution

- Just as we had with the gamma and inverse-gamma relationship in the univariate case, we can also work in terms of the Wishart distribution (multivariate generalization of the gamma) instead.

- The Wishart distribution provides a conditionally-conjugate prior for the precision matrix Σ^{-1} in a multivariate normal model.

- Specifically, if $\Sigma \sim \mathcal{IW}_p(\nu_0, \boldsymbol{S}_0)$, then $\Phi = \Sigma^{-1} \sim \mathrm{W}_p(\nu_0, \boldsymbol{S}_0^{-1})$.

- A random variable $\Phi \sim \mathrm{W}_p(\nu_0, \boldsymbol{S}_0^{-1})$, where Φ has dimension $(p \times p)$, has pdf

$$f(\Phi) \ \propto \ |\Phi|^{\frac{\nu_0 - p - 1}{2}} \mathrm{exp} \left\{ -\frac{1}{2} \mathrm{tr}(\boldsymbol{S}_0 \Phi) \right\}.$$

- Here, $\mathbb{E}[\Phi] = \nu_0 \boldsymbol{S}_0$.

- Note that the textbook writes the inverse-Wishart as $\mathcal{IW}_p(\nu_0, \boldsymbol{S}_0^{-1})$. I prefer $\mathcal{IW}_p(\nu_0, \boldsymbol{S}_0)$ instead. Feel free to use either notation but try not to get confused.

Back to inference on covariance

- For inference on $\boldsymbol{\Sigma}$, we need to rewrite the likelihood a bit to match the inverse-Wishart kernel.

- First a few results from matrix algebra:

  1. $\text{tr}(\boldsymbol{A}) = \sum_{j=1}^{p} a_{jj}$, where a_{jj} is the jth diagonal element of a square $p \times p$ matrix $\boldsymbol{A}$.
  2. Cyclic property:
  $$\text{tr}(\boldsymbol{ABC}) = \text{tr}(\boldsymbol{BCA}) = \text{tr}(\boldsymbol{CAB}),$$
  given that the product $\boldsymbol{ABC}$ is a square matrix.
  3. If $\boldsymbol{A}$ is a $p \times p$ matrix, then for a $p \times 1$ vector $\boldsymbol{x}$,
  $$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \text{tr}(\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x})$$
  holds by (1), since $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$ is a scalar.
  4. $\text{tr}(\boldsymbol{A} + \boldsymbol{B}) = \text{tr}(\boldsymbol{A}) + \text{tr}(\boldsymbol{B})$.

Multivariate normal likelihood again

- It is thus convenient to rewrite $L(\boldsymbol{Y};\boldsymbol{\theta},\Sigma)$ as

$$
\begin{aligned}
L(\boldsymbol{Y};\boldsymbol{\theta},\Sigma) &\propto \underbrace{|\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{y}_i-\boldsymbol{\theta})\right\}}_{\text{no algebra/change yet}} \\
&= |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\underbrace{\mathrm{tr}\left[(\boldsymbol{y}_i-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{y}_i-\boldsymbol{\theta})\right]}_{\text{by result 3}}\right\} \\
&= |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\underbrace{\mathrm{tr}\left[(\boldsymbol{y}_i-\boldsymbol{\theta})(\boldsymbol{y}_i-\boldsymbol{\theta})^T\Sigma^{-1}\right]}_{\text{by cyclic property}}\right\} \\
&= |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\underbrace{\mathrm{tr}\left[\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\theta})(\boldsymbol{y}_i-\boldsymbol{\theta})^T\Sigma^{-1}\right]}_{\text{by result 4}}\right\} \\
&= |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{S}_\theta\Sigma^{-1}\right]\right\},
\end{aligned}
$$

where $\boldsymbol{S}_\theta = \sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\theta})(\boldsymbol{y}_i-\boldsymbol{\theta})^T$ is the residual sum of squares matrix.

Conditional posterior for covariance

- Assuming $\pi(\Sigma) = \mathcal{IW}_p(\nu_0, \boldsymbol{S}_0)$, the conditional posterior (full conditional) $\Sigma|\boldsymbol{\theta}, \boldsymbol{Y}$, is then

$$\begin{aligned}
\pi(\Sigma|\boldsymbol{\theta}, \boldsymbol{Y}) &\propto L(\boldsymbol{Y}; \boldsymbol{\theta}, \Sigma) \cdot \pi(\boldsymbol{\theta}) \\
&\propto \underbrace{|\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{S}_\theta \Sigma^{-1}\right]\right\}}_{L(\boldsymbol{Y};\boldsymbol{\theta},\Sigma)} \cdot \underbrace{|\Sigma|^{\frac{-(\nu_0+p+1)}{2}} \exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{S}_0\Sigma^{-1})\right\}}_{\pi(\boldsymbol{\theta})} \\
&\propto |\Sigma|^{\frac{-(\nu_0+p+n+1)}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{S}_0\Sigma^{-1} + \boldsymbol{S}_\theta\Sigma^{-1}\right]\right\}, \\
&\propto |\Sigma|^{\frac{-(\nu_0+n+p+1)}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[\left(\boldsymbol{S}_0 + \boldsymbol{S}_\theta\right)\Sigma^{-1}\right]\right\},
\end{aligned}$$

which is $\mathcal{IW}_p(\nu_n, \boldsymbol{S}_n)$, or using the notation in the book, $\mathcal{IW}_p(\nu_n, \boldsymbol{S}_n^{-1})$, with

  - $\nu_n = \nu_0 + n$, and
  - $\boldsymbol{S}_n = [\boldsymbol{S}_0 + \boldsymbol{S}_\theta]$

Conditional posterior for covariance

- We once again see that the "posterior sample size" or "posterior degrees of freedom" ν_n is the sum of the "prior degrees of freedom" ν_0 and the data sample size n.

- $\boldsymbol{S}_n$ can be thought of as the "posterior sum of squares", which is the sum of "prior sum of squares" plus "sample sum of squares".

- Recall that if $\Sigma \sim \mathcal{IW}_p(\nu_0, \boldsymbol{S}_0)$, then $\mathbb{E}[\Sigma] = \dfrac{1}{\nu_0 - p - 1}\boldsymbol{S}_0$.

- $\Rightarrow$ the conditional posterior expectation of the population covariance is

$$
\begin{aligned}
\mathbb{E}[\Sigma|\boldsymbol{\theta}, \boldsymbol{Y}] &= \frac{1}{\nu_0 + n - p - 1}\left[\boldsymbol{S}_0 + \boldsymbol{S}_\theta\right] \\
&= \underbrace{\frac{\nu_0 - p - 1}{\nu_0 + n - p - 1}}_{\text{weight on prior expectation}} \overbrace{\left[\frac{1}{\nu_0 - p - 1}\boldsymbol{S}_0\right]}^{\text{prior expectation}} + \underbrace{\frac{n}{\nu_0 + n - p - 1}}_{\text{weight on sample estimate}} \overbrace{\left[\frac{1}{n}\boldsymbol{S}_\theta\right]}^{\text{sample estimate}},
\end{aligned}
$$

which is a weighted average of prior expectation and sample estimate.

Reading comprehension example again

- Vector of observations for each student: $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})^T$.
  - Y_{i1}: pre-instructional score for student i.
  - Y_{i2}: post-instructional score for student i.
- Questions of interest:
  - Do students improve in reading comprehension on average?
  - If so, by how much?
  - Can we predict post-test score from pre-test score? How correlated are they?
  - If we have students with missing pre-test scores, can we predict the scores? (Will defer this till next week!)

Reading comprehension example

- Since we have bivariate continuous responses for each student, and test scores are often normally distributed, we can use a bivariate normal model.
- Model the data as $\boldsymbol{y_i} = (y_{i1}, y_{i2})^T \sim \mathcal{N}_2(\boldsymbol{\theta}, \Sigma)$, that is,

$$\boldsymbol{Y} = \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathcal{N}_2 \left[\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \right].$$

- We can answer the first two questions of interest by looking at the posterior distribution of $\theta_2 - \theta_1$.
- The correlation between Y_1 and Y_2 is

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2},$$

so we can answer the third question by looking at the posterior distribution of ρ, which we have once we have posterior samples of Σ.

Reading example: prior on mean

- Clearly, we first need to set the hyperparameters $\boldsymbol{\mu}_0$ and Λ_0 in $\pi(\boldsymbol{\theta}) = \mathcal{N}_2(\boldsymbol{\mu}_0, \Lambda_0)$, based on prior belief.

- For this example, both tests were actually designed *apriori* to have a mean of 50, so, we can set $\boldsymbol{\mu}_0 = (\mu_{0(1)}, \mu_{0(2)})^T = (50, 50)^T$.

- $\boldsymbol{\mu}_0 = (0, 0)^T$ is also often a common choice when there is no prior guess, especially when there is enough data to "drown out" the prior guess.

- Next, we need to set values for elements of

$$\Lambda_0 = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix}$$

- It is quite reasonable to believe *apriori* that the true means will most likely lie in the interval $[25, 75]$ with high probability (perhaps 0.95?), since individual test scores should lie in the interval $[0, 100]$.

- Recall that for any normal distribution, 95% of the density will lie within two standard deviations of the mean.

Reading example: prior on mean

- Therefore, we can set

$$\mu_{0(1)} \pm 2\sqrt{\lambda_{11}} = (25, 75) \quad \Rightarrow \quad 50 \pm 2\sqrt{\lambda_{11}} = (25, 75)$$
$$\Rightarrow \quad 2\sqrt{\lambda_{11}} = 25 \quad \Rightarrow \quad \lambda_{11} = \left(\frac{25}{2}\right)^2 \approx 156.$$

- Similarly, set $\lambda_{22} \approx 156$.

- Finally, we expect some correlation between $\mu_{0(1)}$ and $\mu_{0(2)}$, but suppose we don't know exactly how strong. We can set the prior correlation to 0.5.

$$\Rightarrow 0.5 = \frac{\lambda_{12}}{\sqrt{\lambda_{11}}\sqrt{\lambda_{22}}} = \frac{\lambda_{12}}{156} \quad \Rightarrow \quad \lambda_{12} = 156 \times 0.5 = 78.$$

- Thus,

$$\pi(\boldsymbol{\theta}) = \mathcal{N}_2\left(\boldsymbol{\mu}_0 = \begin{pmatrix} 50 \\ 50 \end{pmatrix}, \Lambda_0 = \begin{pmatrix} 156 & 78 \\ 78 & 156 \end{pmatrix}\right).$$

Reading example: prior on covariance

- Next we need to set the hyperparameters ν_0 and $\boldsymbol{S}_0$ in $\pi(\Sigma) = \mathcal{IW}_2(\nu_0, \boldsymbol{S}_0)$, based on prior belief.

- First, let's start with a prior guess Σ_0 for Σ.

- Again, since individual test scores should lie in the interval $[0, 100]$, we should set Σ_0 so that values outside $[0, 100]$ are highly unlikely.

- Just as we did with Λ_0, we can use that idea to set the elements of Σ_0

$$\Sigma_0 = \begin{pmatrix} \sigma_{11}^{(0)} & \sigma_{12}^{(0)} \\ \sigma_{21}^{(0)} & \sigma_{22}^{(0)} \end{pmatrix}$$

- The identity matrix is also often a common choice for Σ_0 when there is no prior guess, especially when there is enough data to "drown out" the prior guess.

Reading example: prior on covariance

- Therefore, we can set

$$\mu_{0(1)} \pm 2\sqrt{\sigma_{11}^{(0)}} = (0,100) \;\; \Rightarrow \;\; 50 \pm 2\sqrt{\sigma_{11}^{(0)}} = (0,100)$$
$$\Rightarrow \;\; 2\sqrt{\sigma_{11}^{(0)}} = 50 \;\; \Rightarrow \;\; \sigma_{11}^{(0)} = \left(\frac{50}{2}\right)^2 \approx 625.$$

- Similarly, set $\sigma_{22}^{(0)} \approx 625.$

- Again, we expect some correlation between Y_1 and Y_2, but suppose we don't know exactly how strong. We can set the prior correlation to 0.5.

$$\Rightarrow 0.5 = \frac{\sigma_{12}^{(0)}}{\sqrt{\sigma_{11}^{(0)}}\sqrt{\sigma_{22}^{(0)}}} = \frac{\sigma_{12}^{(0)}}{625} \;\; \Rightarrow \;\; \sigma_{12}^{(0)} = 625 \times 0.5 = 312.5.$$

- Thus,

$$\Sigma_0 = \begin{pmatrix} 625 & 312.5 \\ 312.5 & 625 \end{pmatrix}$$

Reading example: prior on covariance

- Recall that if we are not at all confident on a prior value for Σ, but we have a prior guess Σ_0, we can set
  - $\nu_0 = p + 2$, so that the $\mathbb{E}[\Sigma] = \dfrac{1}{\nu_0 - p - 1}\boldsymbol{S}_0$ is finite.
  - $\boldsymbol{S}_0 = \Sigma_0$

  so that Σ is only loosely centered around Σ_0.
- Thus, we can set
  - $\nu_0 = p + 2 = 2 + 2 = 4$
  - $\boldsymbol{S}_0 = \Sigma_0$

  so that we have

$$\pi(\Sigma) = \mathcal{IW}_2\left(\nu_0 = 4, \Sigma_0 = \begin{pmatrix} 625 & 312.5 \\ 312.5 & 625 \end{pmatrix}\right).$$

Reading example: data

Now, to the data (finally!)

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

```
## [1] 22  2
```

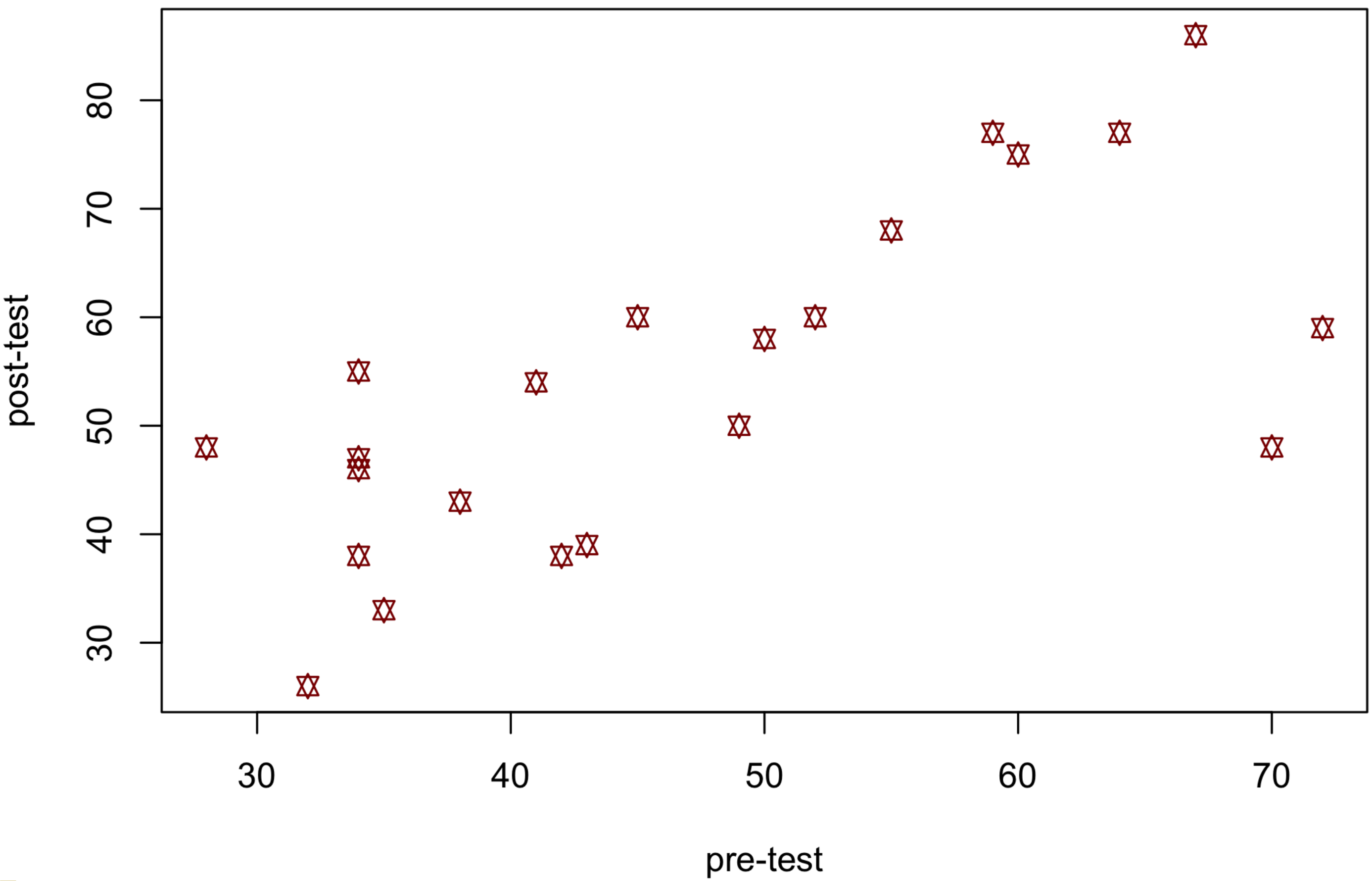
```
head(Y)
```

```
##      pretest posttest
## [1,]      59       77
## [2,]      43       39
## [3,]      34       46
## [4,]      32       26
## [5,]      42       38
## [6,]      38       43
```

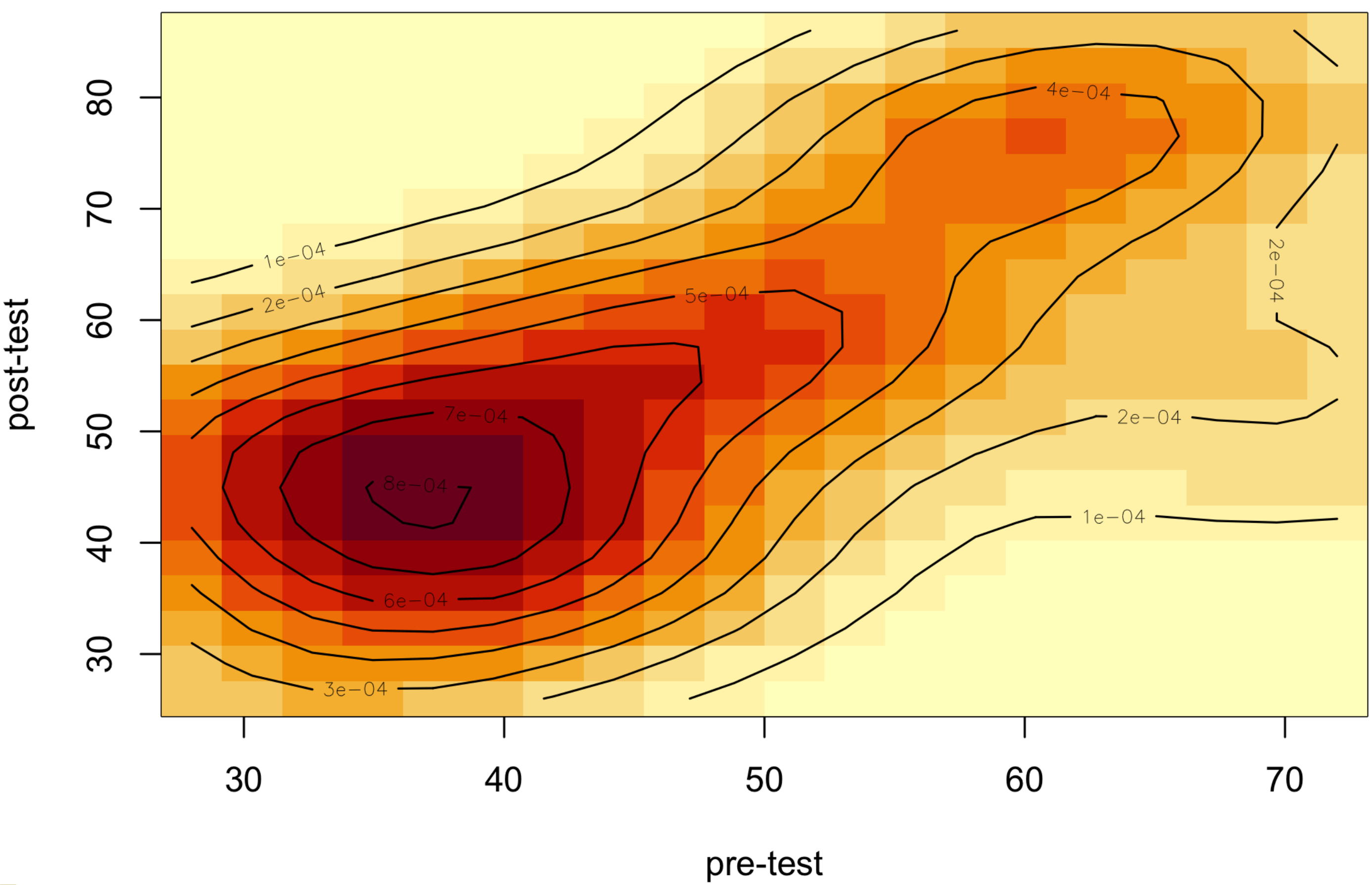
```
summary(Y)
```

```
##     pretest         posttest
##  Min.   :28.00   Min.   :26.00
##  1st Qu.:34.25   1st Qu.:43.75
##  Median :44.00   Median :52.00
##  Mean   :47.18   Mean   :53.86
##  3rd Qu.:58.00   3rd Qu.:60.00
##  Max.   :72.00   Max.   :86.00
```

Reading example: data



Reading example: data



Posterior computation

- To recap, we have

$$\pi(\boldsymbol{\theta}|\Sigma, \boldsymbol{Y}) = \mathcal{N}_2(\boldsymbol{\mu}_n, \Lambda_n)$$

where

$$\Lambda_n = \left[\Lambda_0^{-1} + n\Sigma^{-1}\right]^{-1}$$

$$\boldsymbol{\mu}_n = \Lambda_n \left[\Lambda_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\boldsymbol{y}}\right],$$

$$\boldsymbol{\mu}_0 = (\mu_{0(1)}, \mu_{0(2)})^T = (50, 50)^T$$

$$\Lambda_0 = \begin{pmatrix} 156 & 78 \\ 78 & 156 \end{pmatrix}$$

Posterior computation

- We also have

$$\pi(\Sigma|\boldsymbol{\theta}\boldsymbol{Y}) = \mathcal{IW}_2(\nu_n, \boldsymbol{S}_n)$$

or using the notation in the book, $\mathcal{IW}_2(\nu_n, \boldsymbol{S}_n^{-1})$, where

$$\begin{aligned}
\nu_n &= \nu_0 + n \\
\\
\boldsymbol{S}_n &= [\boldsymbol{S}_0 + \boldsymbol{S}_\theta] \\
&= \left[\boldsymbol{S}_0 + \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{\theta})^T\right].
\end{aligned}$$

$$\nu_0 = p + 2 = 4$$

$$\Sigma_0 = \begin{pmatrix} 625 & 312.5 \\ 312.5 & 625 \end{pmatrix}$$

Posterior computation

```r
#Data summaries
n <- nrow(Y)
ybar <- apply(Y,2,mean)

#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)

#Initial values for Gibbs sampler
#No need to set initial value for theta, we can simply sample it first
Sigma <- cov(Y)

#Set null matrices to save samples
THETA <- SIGMA <- NULL
```

Next, we need to write the code for the Gibbs sampler.

Posterior computation

```r
#Now, to the Gibbs sampler
#library(mvtnorm) for multivariate normal
#library(MCMCpack) for inverse-Wishart

#first set number of iterations and burn-in, then set seed
n_iter <- 10000; burn_in <- 0.3*n_iter
set.seed(1234)

for (s in 1:(n_iter+burn_in)){
##update theta using its full conditional
Lambda_n <- solve(solve(Lambda_0) + n*solve(Sigma))
mu_n <- Lambda_n %*% (solve(Lambda_0)%*%mu_0 + n*solve(Sigma)%*%ybar)
theta <- rmvnorm(1,mu_n,Lambda_n)

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

#save results only past burn-in
if(s > burn_in){
  THETA <- rbind(THETA,theta)
  SIGMA <- rbind(SIGMA,c(Sigma))
  }
}
colnames(THETA) <- c("theta_1","theta_2")
colnames(SIGMA) <- c("sigma_11","sigma_12","sigma_21","sigma_22") #symmetry in sigma
```

Note that the text also has a function to sample from the Wishart distribution.

Diagnostics

```r
#library(coda)
THETA.mcmc <- mcmc(THETA,start=1); summary(THETA.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 47.30 2.956  0.02956        0.02956
## theta_2 53.69 3.290  0.03290        0.03290
##
## 2. Quantiles for each variable:
##
##          2.5%   25%   50%   75% 97.5%
## theta_1 41.55 45.35 47.36 49.23 53.08
## theta_2 47.08 51.53 53.69 55.82 60.13
```

```r
effectiveSize(THETA.mcmc)
```

```
## theta_1 theta_2
##   10000   10000
```

Diagnostics

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

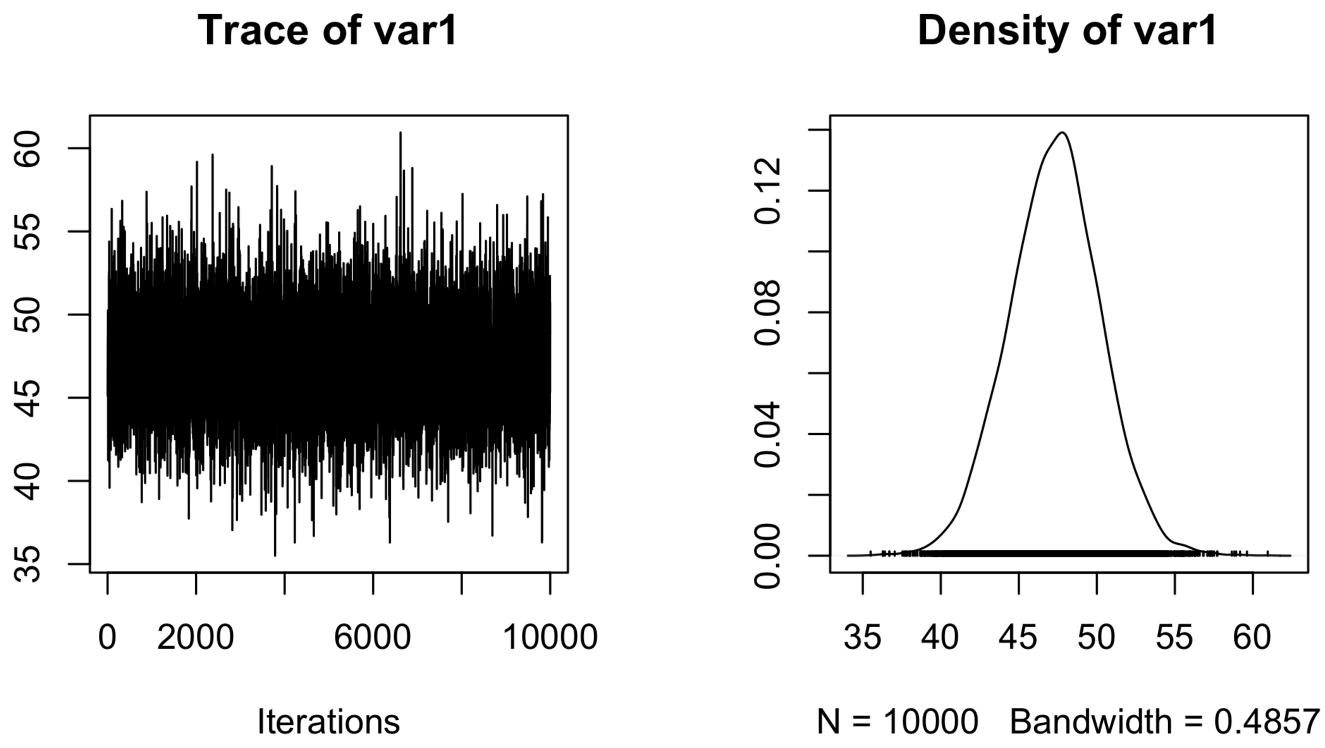
```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean    SD Naive SE Time-series SE
## sigma_11 202.3 63.39   0.6339         0.6511
## sigma_12 155.3 60.92   0.6092         0.6244
## sigma_21 155.3 60.92   0.6092         0.6244
## sigma_22 260.1 81.96   0.8196         0.8352
##
## 2. Quantiles for each variable:
##
##            2.5%   25%   50%   75% 97.5%
## sigma_11 113.50 158.2 190.8 234.8 357.3
## sigma_12  67.27 113.2 144.7 186.5 305.4
## sigma_21  67.27 113.2 144.7 186.5 305.4
## sigma_22 145.84 203.2 244.6 300.9 461.0
```

```
effectiveSize(SIGMA.mcmc)
```

```
## sigma_11 sigma_12 sigma_21 sigma_22
## 9478.710 9517.989 9517.989 9629.352
```

Diagnostics: trace plots

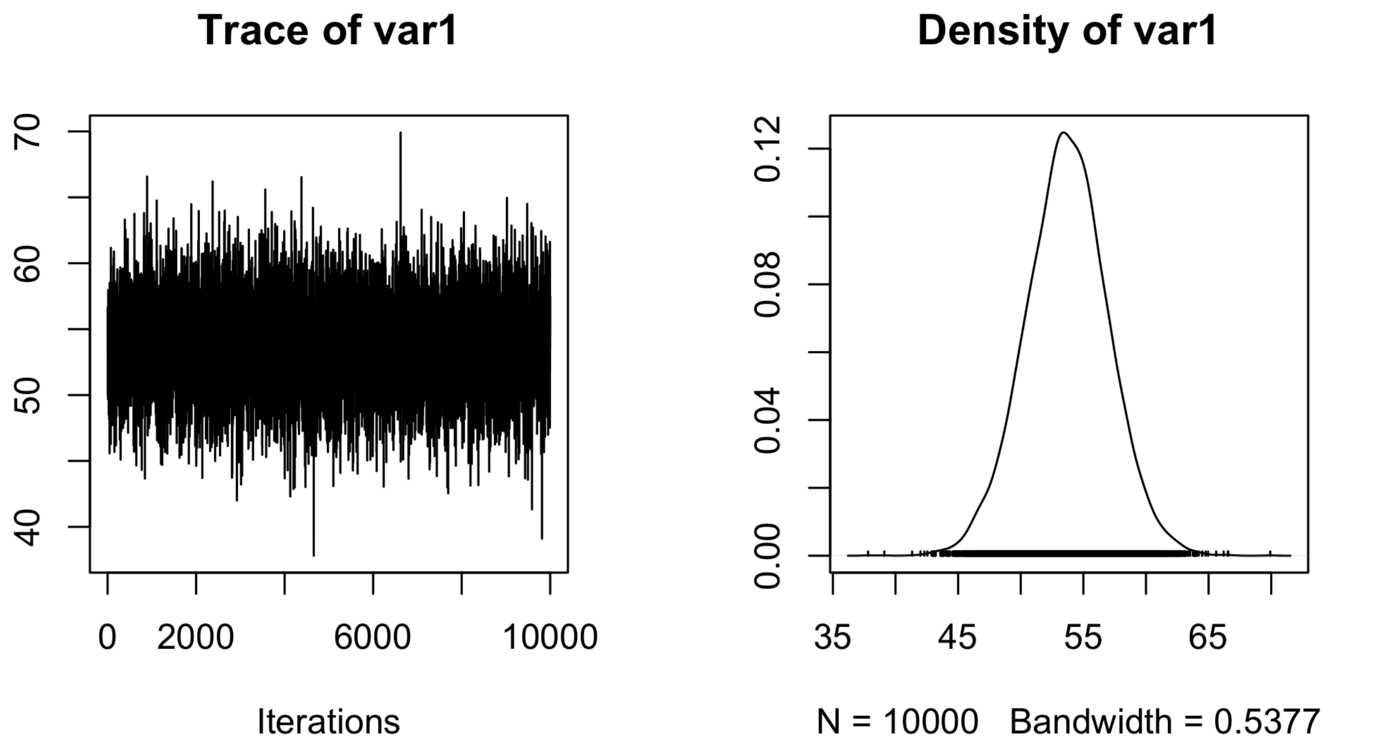
```
plot(THETA.mcmc[,"theta_1"])
```



Looks good!

Diagnostics: trace plots

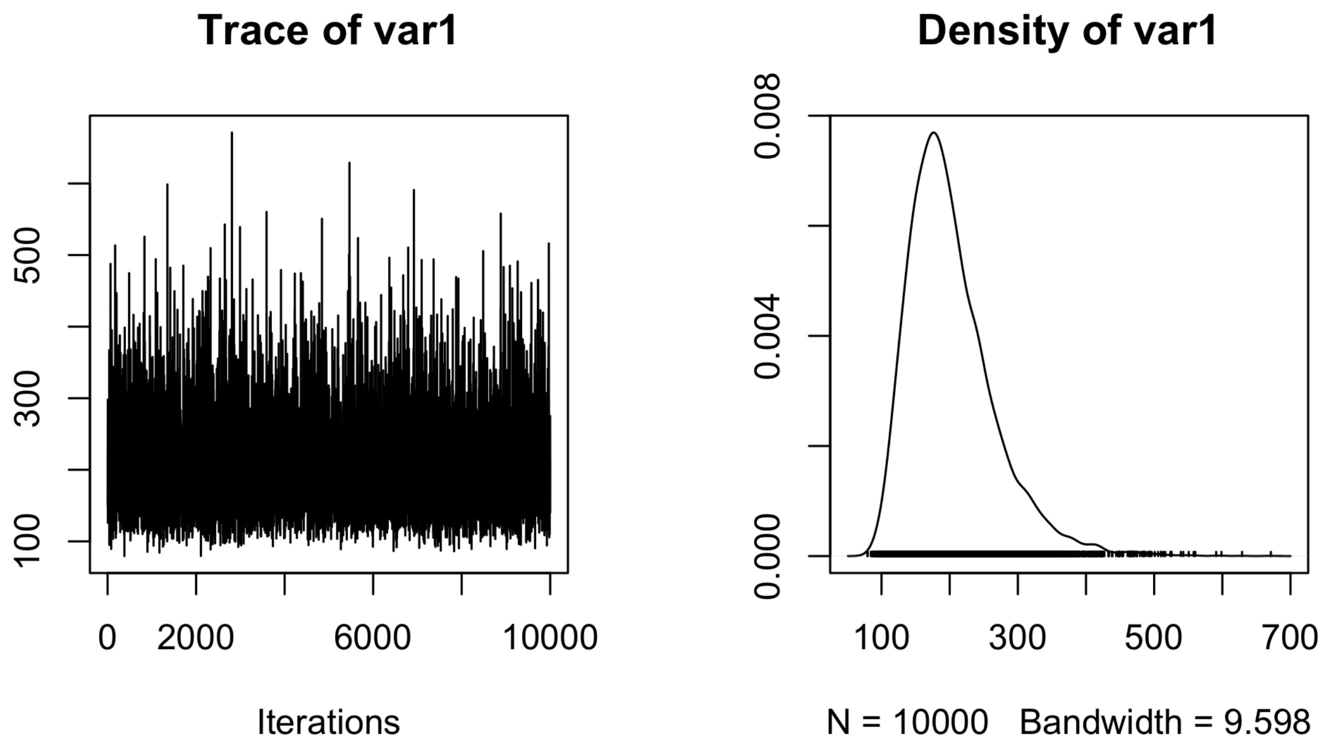
```
plot(THETA.mcmc[,"theta_2"])
```



Looks good!

Diagnostics: trace plots

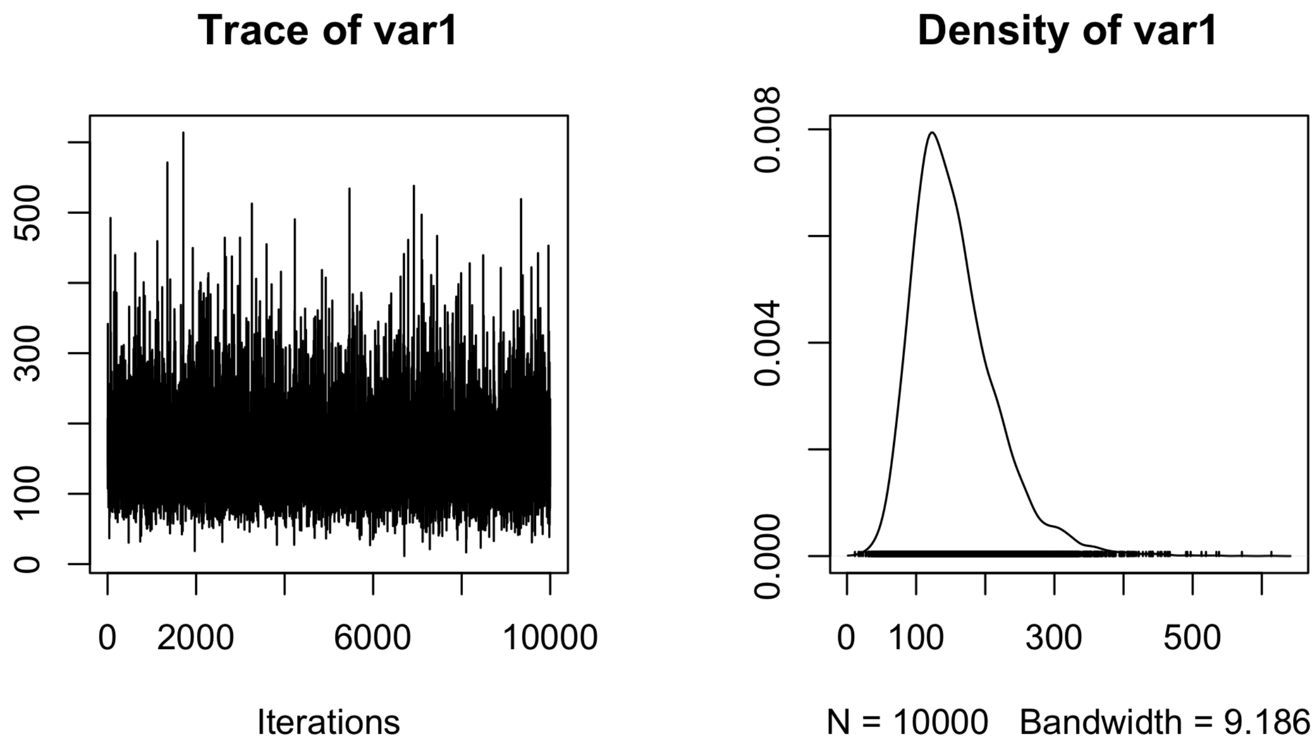
```
plot(SIGMA.mcmc[,"sigma_11"])
```



Looks good!

Diagnostics: trace plots

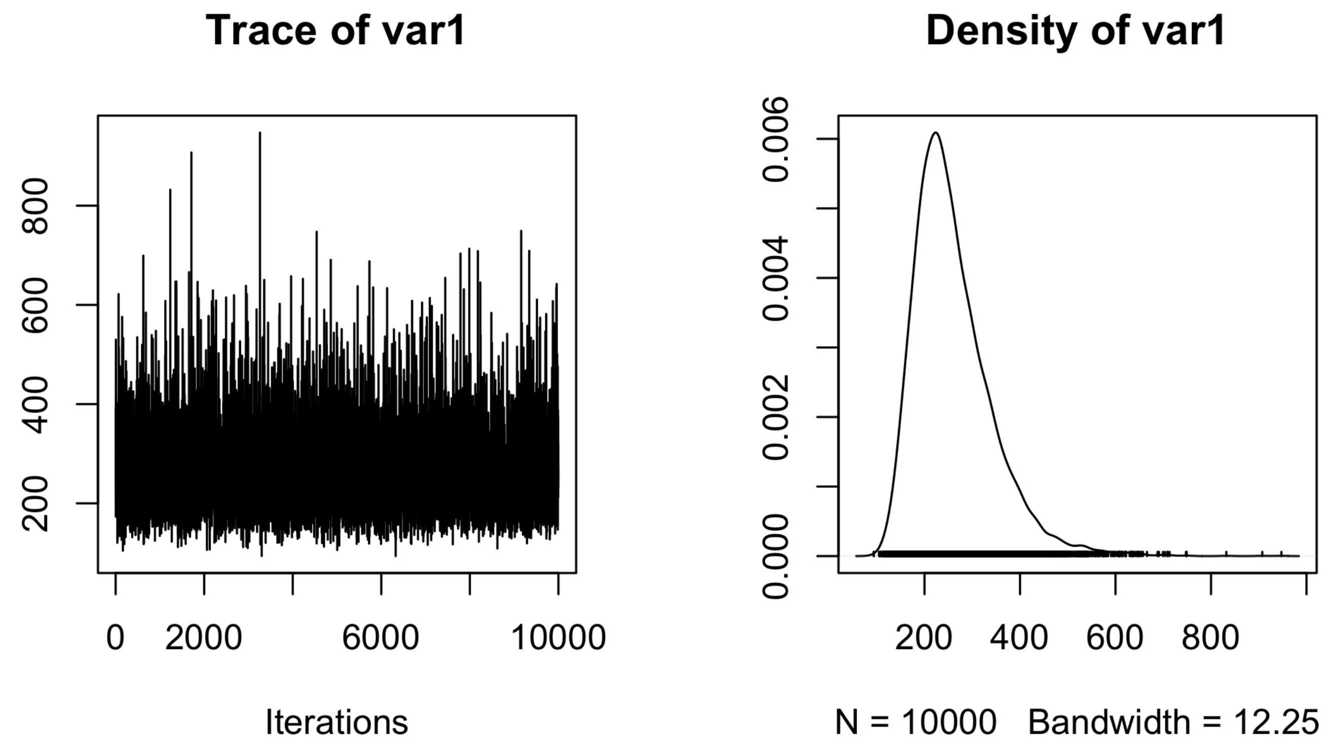
```
plot(SIGMA.mcmc[,"sigma_12"])
```



Looks good!

Diagnostics: trace plots

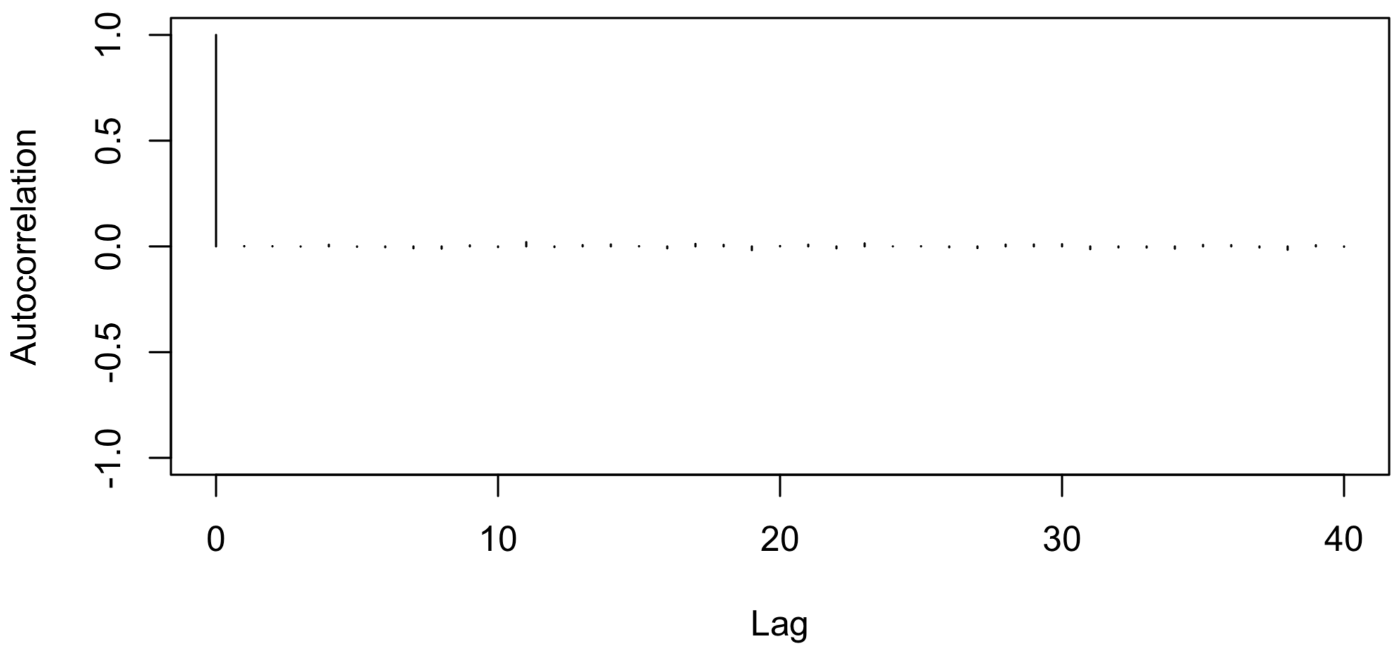
```
plot(SIGMA.mcmc[,"sigma_22"])
```



Looks good!

Diagnostics: autocorrelation

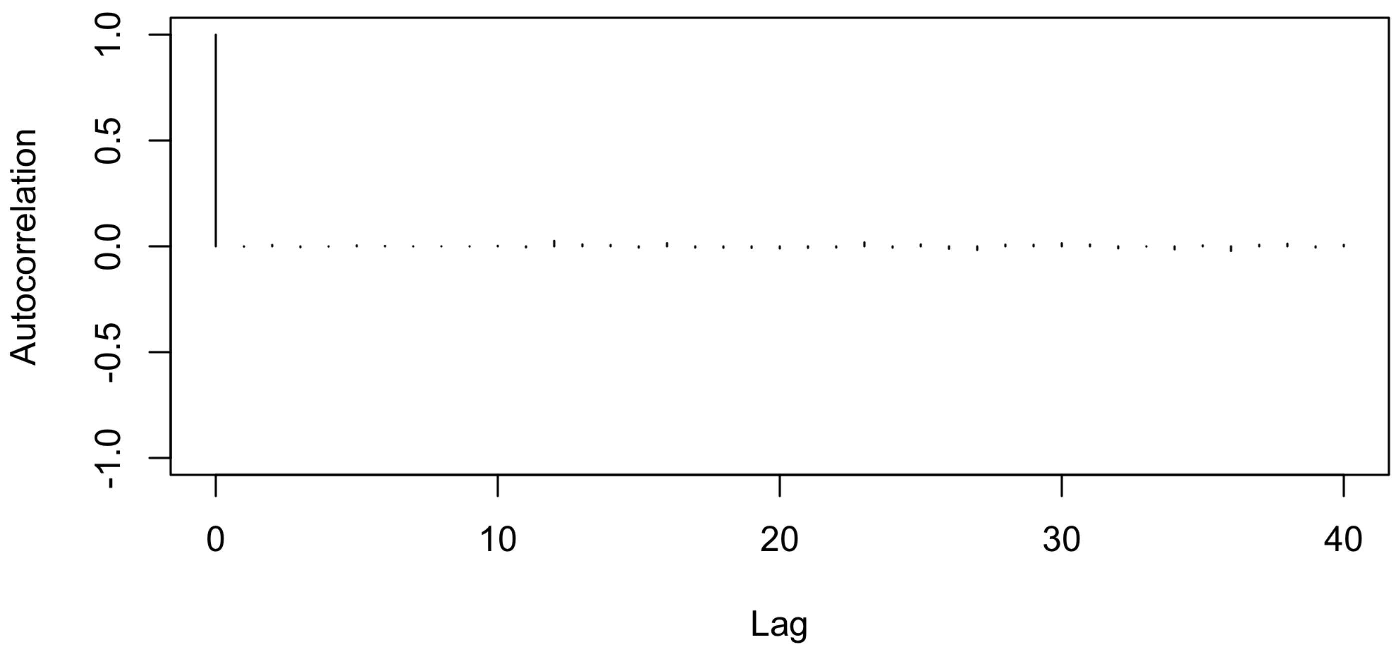
```
autocorr.plot(THETA.mcmc[,"theta_1"])
```



Looks good!

DIAGNOSTICS: AUTOCORRELATION

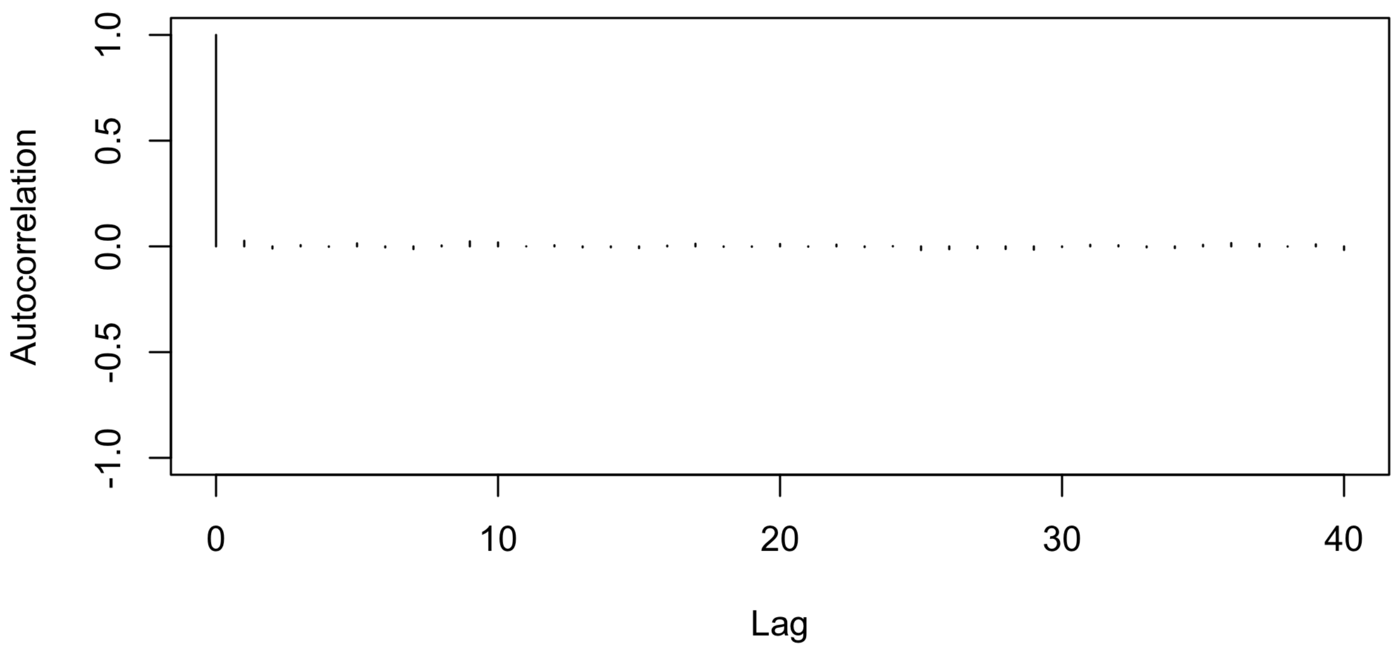
```
autocorr.plot(THETA.mcmc[,"theta_2"])
```



Looks good!

DIAGNOSTICS: AUTOCORRELATION

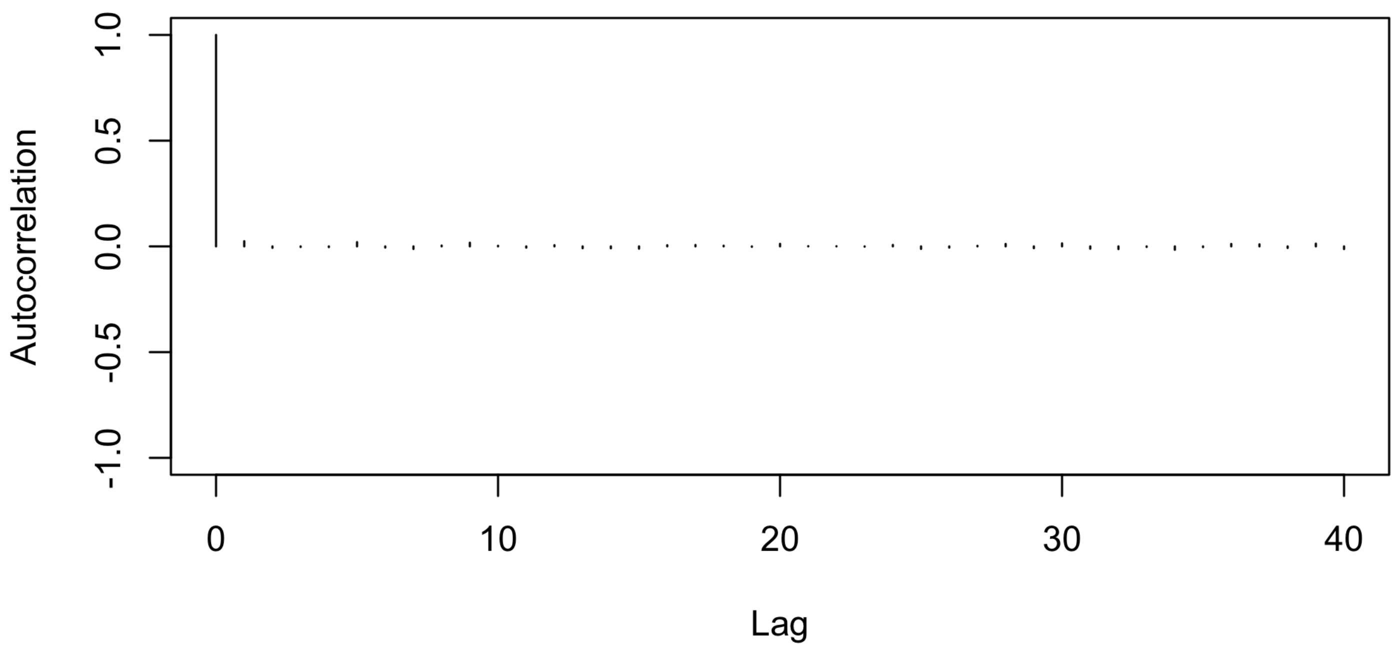
```
autocorr.plot(SIGMA.mcmc[,"sigma_11"])
```



Looks good!

DIAGNOSTICS: AUTOCORRELATION

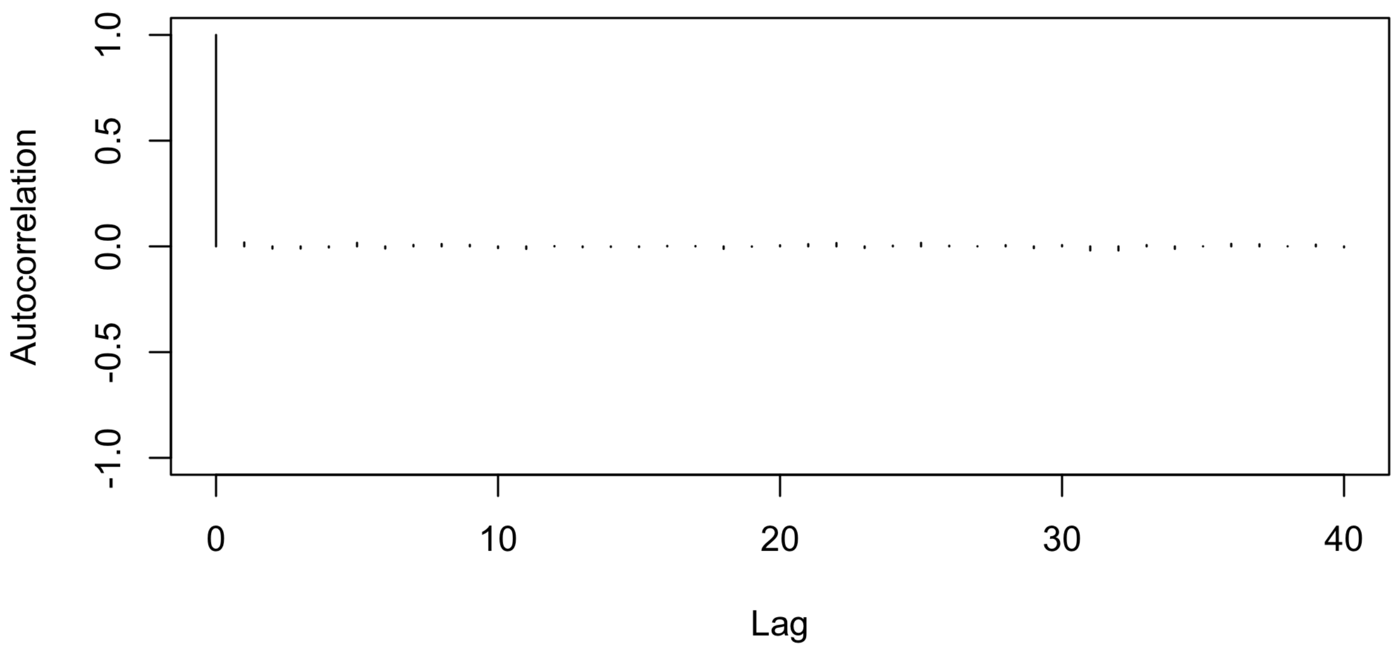
```
autocorr.plot(SIGMA.mcmc[,"sigma_12"])
```



Looks good!

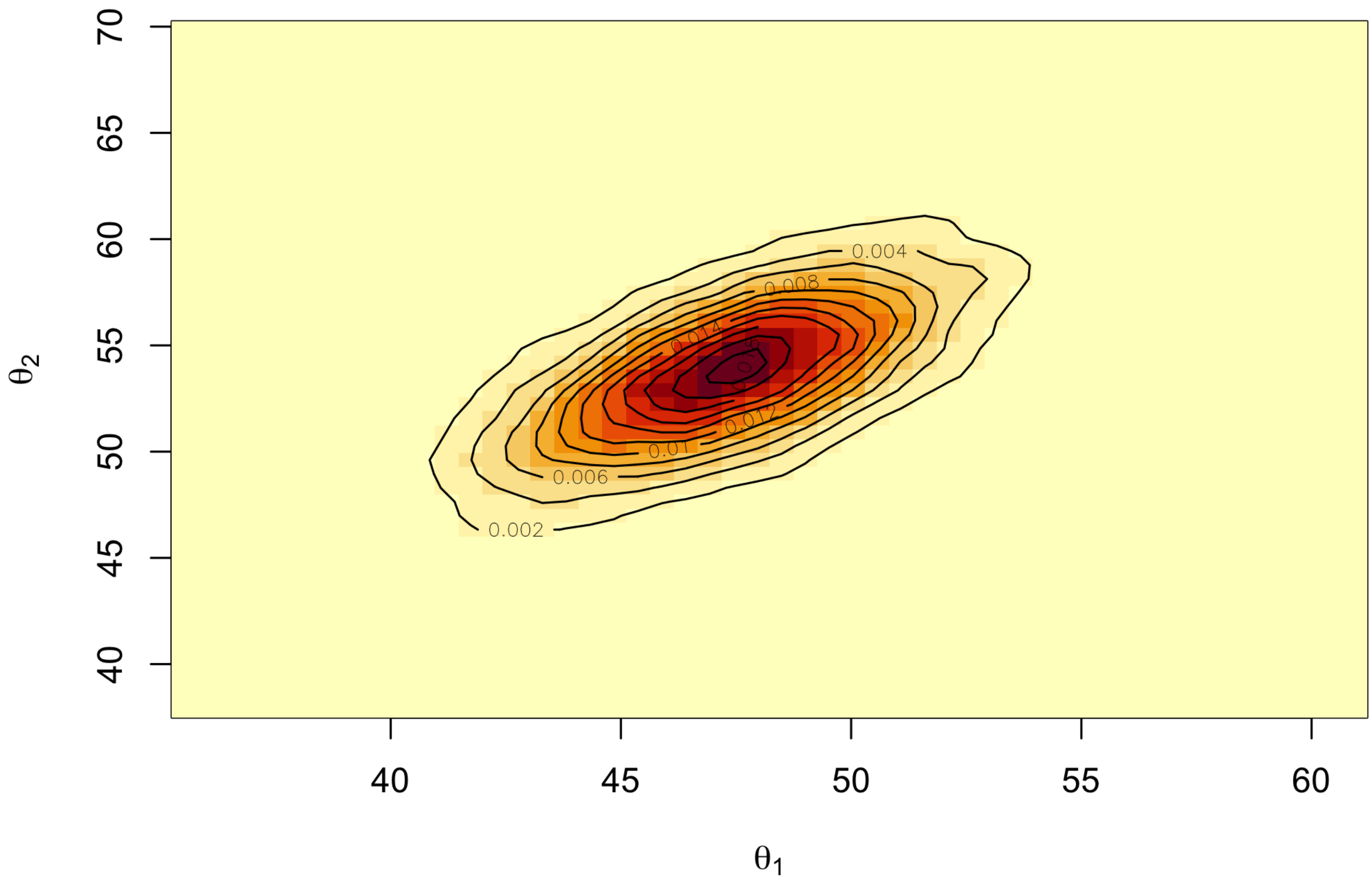
Diagnostics: autocorrelation

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



Looks good!

Posterior distribution of the mean



Answering questions of interest

- Questions of interest:
  - Do students improve in reading comprehension on average?
- Need to compute $\Pr[\theta_2 > \theta_1 | \boldsymbol{Y}]$. In R,

```
mean(THETA[,2]>THETA[,1])
```

```
## [1] 0.992
```

- That is, posterior probability > 0.99 and indicates strong evidence that test scores are higher in the second administration.

Answering questions of interest

- Questions of interest:
  - If so, by how much?
- Need posterior summaries $\Pr[\theta_2 - \theta_1 | \boldsymbol{Y}]$. In R,

```
mean(THETA[,2] - THETA[,1])
```

```
## [1] 6.385515
```

```
quantile(THETA[,2] - THETA[,1], prob=c(0.025, 0.5, 0.975))
```

```
##      2.5%       50%     97.5%
##  1.233154  6.385597 11.551304
```

- Mean (and median) improvement is ≈ 6.39 points with 95% credible interval (1.23, 11.55).

Answering questions of interest

- Questions of interest:
  - How correlated (positively) are the post-test and pre-test scores?
- We can compute $\Pr[\sigma_{12} > 0|\boldsymbol{Y}]$. In R,

```
mean(SIGMA[,2]>0)
```

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

- Posterior probability that the covariance between them is positive is basically 1.

Answering questions of interest

- Questions of interest:
  - How correlated (positively) are the post-test and pre-test scores?
- We can also look at the distribution of ρ instead. In R,

```
CORR <- SIGMA[,2]/(sqrt(SIGMA[,1])*sqrt(SIGMA[,4]))
quantile(CORR,prob=c(0.025, 0.5, 0.975))
```

```
##      2.5%       50%     97.5%
## 0.4046817 0.6850218 0.8458880
```

- Median correlation between the 2 scores is 0.69 with a 95% quantile-based credible interval of (0.40, 0.85)
- Because density is skewed, we may prefer the 95% HPD interval, which is (0.45, 0.88).

```
#library(hdrcde)
hdr(CORR,prob=95)$hdr
```

```
##          [,1]      [,2]
## 95% 0.4468522 0.8761174
```

Jeffreys' prior

- Clearly, there's a lot of work to be done in specifying the hyperparameters (two or which are $p \times p$ matrices).
- What if we want to specify the priors so that we put in as little information as possible?
- We already know how to do that somewhat with Jeffreys' priors.
- For the multivariate normal model, turns out that the Jeffreys' rule for generating a prior distribution on $(\boldsymbol{\theta}, \Sigma)$ gives

$$\pi(\boldsymbol{\theta}, \Sigma) \propto |\Sigma|^{-\frac{(p+2)}{2}}.$$

- Can we derive the full conditionals under this prior?
- **To be done on the board.**

JEFFREYS' PRIOR

- We can leverage previous work. For the likelihood we have both

$$L(\boldsymbol{Y}; \boldsymbol{\theta}, \Sigma) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T(n\Sigma^{-1})\boldsymbol{\theta} + \boldsymbol{\theta}^T(n\Sigma^{-1}\bar{\boldsymbol{y}})\right\}$$

and

$$L(\boldsymbol{Y}; \boldsymbol{\theta}, \Sigma) \propto |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{S}_\theta\Sigma^{-1}\right]\right\},$$

where $\boldsymbol{S}_\theta = \sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{\theta})^T$.

- Also, we can rewrite any $\mathcal{N}_p(\boldsymbol{\mu}_0, \Lambda_0)$ as

$$p(\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\mu}_0\right\}.$$

- Finally, $\Sigma \sim \mathcal{IW}_p(\nu_0, \boldsymbol{S}_0)$,

$$\Rightarrow \quad p(\Sigma) \;\propto\; |\Sigma|^{\frac{-(\nu_0+p+1)}{2}} \exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{S}_0\Sigma^{-1})\right\}.$$