STA 602 - Intro to Bayesian Statistics

Lecture 13

Li Ma

Duke University

Bayesian inference under independent Normal-inverse-Wishart priors

- Recall our sampling model

$$\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n \,|\, \boldsymbol{\theta}, \Sigma \overset{\text{iid}}{\sim} \mathrm{N}(\boldsymbol{\mu}, \Sigma).$$

- Indepednent prior specification:

$$p(\boldsymbol{\theta}, \Sigma) = p(\boldsymbol{\theta}) p(\Sigma)$$

with

$$\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\mu}_0, \Lambda_0)$$

and

$$\Sigma \sim \mathrm{IW}(\nu_0, \boldsymbol{S}_0).$$

Find the full conditionals

- Let's derive the full conditionals so we can do Gibbs sampling
- The likelihood is

$$L(\boldsymbol{\theta}, \Sigma; \mathbf{y}_1, \ldots, \mathbf{y}_n) = p(\mathbf{y}_1, \ldots, \mathbf{y}_n | \boldsymbol{\mu}, \Sigma) \propto |\Sigma|^{-n/2} e^{-\frac{1}{2}\sum_i (\mathbf{y}_i - \boldsymbol{\theta})' \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\theta})}.$$

- The prior is given by

$$p(\boldsymbol{\theta}) \propto e^{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)' \Lambda_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0)}$$

and

$$p(\Sigma) \propto |\Sigma|^{-\frac{\nu_0 + p + 1}{2}} e^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{S}_0 \Sigma^{-1})}$$

The full joint probability

- The full joint probability is then

$$
\begin{aligned}
p(\mathbf{y}_1, \ldots, \mathbf{y}_n, \boldsymbol{\theta}, \Sigma) \propto & |\Sigma|^{-n/2} e^{-\frac{1}{2}\sum_i (\mathbf{y}_i - \boldsymbol{\theta})' \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\theta})} \times \\
& e^{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)' \Lambda_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0)} \times \\
& |\Sigma|^{-\frac{\nu_0 + p + 1}{2}} e^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{S}_0 \Sigma^{-1})}
\end{aligned}
$$

The full conditional of $\boldsymbol{\theta}$

- Viewing the joint probability as a function in $\boldsymbol{\theta}$, we have

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}_1,\ldots,\mathbf{y}_n,\Sigma) &\propto e^{-\frac{1}{2}\sum_i(\mathbf{y}_i-\boldsymbol{\theta})'\Sigma^{-1}(\mathbf{y}_i-\boldsymbol{\theta})-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)'\Lambda_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)} \\ &\propto e^{-\frac{1}{2}\left[n\boldsymbol{\theta}'\Sigma^{-1}\boldsymbol{\theta}-2n\bar{y}'\Sigma^{-1}\boldsymbol{\theta}+\boldsymbol{\theta}'\Lambda_0^{-1}\boldsymbol{\theta}+2\boldsymbol{\mu}_0'\Lambda_0^{-1}\boldsymbol{\theta}\right]} \\ &\propto e^{-\frac{1}{2}\left[\boldsymbol{\theta}'(n\Sigma^{-1}+\Lambda_0^{-1})\boldsymbol{\theta}-2(n\bar{y}'\Sigma^{-1}+\boldsymbol{\mu}_0'\Lambda_0^{-1})\boldsymbol{\theta}\right]} \end{aligned}$$

- Now let Λ_n be such that

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

and

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

- Then by completion of squares we have

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}_1,\ldots,\mathbf{y}_n,\Sigma) &\propto e^{-\frac{1}{2}\left(\boldsymbol{\theta}'\Lambda_n^{-1}\boldsymbol{\theta}-2\boldsymbol{\mu}_n'\Lambda_n^{-1}\boldsymbol{\theta}\right)} \\ &\propto e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_n)'\Lambda_n^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_n)}, \end{aligned}$$

which is exactly a $\mathrm{N}(\boldsymbol{\mu}_n,\Lambda_n)$.

- Compare this with univariate case — similar interpretation.

Full conditional of Σ

- Viewing the joint probability as a function in Σ, we have

$$p(\Sigma|\mathbf{y}_1,\ldots,\mathbf{y}_n,\boldsymbol{\theta}) \propto |\Sigma|^{-(n+\nu_0+p+1)/2} e^{-\frac{1}{2}\left[\sum_i (\mathbf{y}_i-\boldsymbol{\theta})'\Sigma^{-1}(\mathbf{y}_i-\boldsymbol{\theta})+\mathrm{tr}(\boldsymbol{S}_0\Sigma^{-1})\right]}.$$

$$\mathrm{tr}\left(\sum_i (y_i-\theta)(y_i-\theta)^T\Sigma^{-1}\right)$$

- Two useful properties of $\mathrm{tr}(\cdot)$:

  (i) For two matrix $A_{k\times m}$ and $B_{m\times k}$, we can prove (by definition of trace) that

$$\mathrm{tr}(AB) = \mathrm{tr}(BA).$$

  Note that AB is $k\times k$ and BA is $m\times m$.

  (ii) Also by definition of trace, we have for two square matrix A and B,

$$\mathrm{tr}(A+B) = \mathrm{tr}(A)+\mathrm{tr}(B).$$

Full conditional of Σ

- Now note that $(\mathbf{y}_i - \boldsymbol{\theta})'\Sigma^{-1}(\mathbf{y}_i - \boldsymbol{\theta})$ is a scalar (i.e., a 1×1), and so it is equal to its trace

$$(\mathbf{y}_i - \boldsymbol{\theta})'\Sigma^{-1}(\mathbf{y}_i - \boldsymbol{\theta}) = \text{tr}\left((\mathbf{y}_i - \boldsymbol{\theta})(\mathbf{y}_i - \boldsymbol{\theta})'\Sigma^{-1}\right)$$

  by Property (i).
- Then by Property (ii)

$$\begin{aligned}\sum_i (\mathbf{y}_i - \boldsymbol{\theta})'\Sigma^{-1}(\mathbf{y}_i - \boldsymbol{\theta}) &= \sum_i \text{tr}\left((\mathbf{y}_i - \boldsymbol{\theta})(\mathbf{y}_i - \boldsymbol{\theta})'\Sigma^{-1}\right) \\ &= \text{tr}\left(\sum_i (\mathbf{y}_i - \boldsymbol{\theta})(\mathbf{y}_i - \boldsymbol{\theta})'\Sigma^{-1}\right) \\ &= \text{tr}\left(\boldsymbol{S}_{\boldsymbol{\theta}}\Sigma^{-1}\right)\end{aligned}$$

  where

$$\boldsymbol{S}_{\boldsymbol{\theta}} = \sum_i (\mathbf{y}_i - \boldsymbol{\theta})(\mathbf{y}_i - \boldsymbol{\theta})',$$

  which is the *residual sum of squared matrix* of the observations.

Full conditional of Σ

- By Property (ii) again,

$$\mathrm{tr}\left(\boldsymbol{S}_{\boldsymbol{\theta}}\Sigma^{-1}\right)+\mathrm{tr}\left(\boldsymbol{S}_0\Sigma^{-1}\right)=\mathrm{tr}\left(\left(\boldsymbol{S}_{\boldsymbol{\theta}}+\boldsymbol{S}_0\right)\Sigma^{-1}\right)$$

- Therefore, putting all pieces together, we have

$$p(\Sigma|\mathbf{y}_1,\ldots,\mathbf{y}_n,\boldsymbol{\theta})\propto|\Sigma|^{-(n+\nu_0+p+1)/2}e^{-\frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{S}_{\boldsymbol{\theta}}+\boldsymbol{S}_0\right)\Sigma^{-1}\right)},$$

  which we recognize as again an inverse-Wishart$(\nu_n,\boldsymbol{S}_0+\boldsymbol{S}_{\boldsymbol{\theta}})$ distribution where $\nu_n=\nu_0+n$.

- That is

$$\Sigma|\mathbf{y}_1,\ldots,\mathbf{y}_n,\boldsymbol{\theta}\sim\mathrm{IW}(\nu_n,\boldsymbol{S}_0+\boldsymbol{S}_{\boldsymbol{\theta}}).$$

- ν_0 is the prior degrees of freedom and $\boldsymbol{S}_0$ the prior sum of squares matrix.
- The full conditional expectation of Σ is

$$\begin{aligned}\mathrm{E}(\Sigma|\mathbf{y}_1,\ldots,\mathbf{y}_n,\boldsymbol{\theta})&=\frac{\boldsymbol{S}_0+\boldsymbol{S}_{\boldsymbol{\theta}}}{\nu_n-p-1}\qquad\text{weighted average}\\&=\frac{\nu_0-p-1}{\nu_n-p-1}\cdot\underbrace{\frac{\boldsymbol{S}_0}{\nu_0-p-1}}_{\text{prior mean of }\Sigma}+\frac{n}{\nu_n-p-1}\cdot\underbrace{\frac{\boldsymbol{S}_{\boldsymbol{\theta}}}{n}}_{\text{Empirical estimate of }\Sigma}.\end{aligned}$$

Gibbs sampling

- Initialize the chain at $(\boldsymbol{\theta}^{(0)}, \Sigma^{(0)})$.
- For $t = 1, 2, \ldots,$
  - Update $\boldsymbol{\theta}$:
    - Compute

$$\left(\Lambda_n^{(t)}\right)^{-1} = n\left(\Sigma^{(t-1)}\right)^{-1} + \Lambda_0^{-1},$$

$$\boldsymbol{\mu}_n^{(t)} = \Lambda_n^{(t)}\left(n\left(\Sigma^{(t-1)}\right)^{-1}\bar{y} + \Lambda_0^{-1}\boldsymbol{\mu}_0\right)$$

    - Draw

$$\boldsymbol{\theta}^{(t)} \sim \mathrm{N}(\boldsymbol{\mu}_n^{(t)}, \Lambda_n^{(t)})$$

  - Update Σ:
    - Compute

$$\boldsymbol{S}_{\boldsymbol{\theta}^{(t)}} = \sum_i (\mathbf{y}_i - \boldsymbol{\theta}^{(t)})(\mathbf{y}_i - \boldsymbol{\theta}^{(t)})'.$$

    - Draw

$$\Sigma^{(t)} \sim \mathrm{IW}\left(\nu_n, \boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\theta}^{(t)}}\right).$$

- Discard suitable burn-in steps.

Back to air pollutant example

- Suppose now we measure two pollutants (e.g., PM2.5 and SO2) concurrently 16 times on a day, so our data are bivariate

$$(104, 100), (105, 102), (103, 101), (102, 104), (105, 108), (107, 108),$$
$$(106, 103), (104, 104), (103, 106), (106, 107), (105, 105), (102, 101),$$
$$(102, 100), (108, 106), (105, 105), (104, 105)$$

Prior specification

- Suppose based on historical data, both pollutants have average level of 100. So we set

$$\boldsymbol{\mu}_0 = (100, 100)'.$$

- Suppose the historical distribution of PM2.5 has a standard deviation of about 10, and SO2 has a standard deviation of about 5, and their values tend to rise or fall together with a correlation of about 0.3. Turning this into our prior,

$$\Lambda_0 = \begin{pmatrix} 10^2 & 0.3 \times 10 \times 5 \\ 0.3 \times 10 \times 5 & 5^2 \end{pmatrix} = \begin{pmatrix} 100 & 15 \\ 15 & 25 \end{pmatrix}.$$

Prior specification on Σ

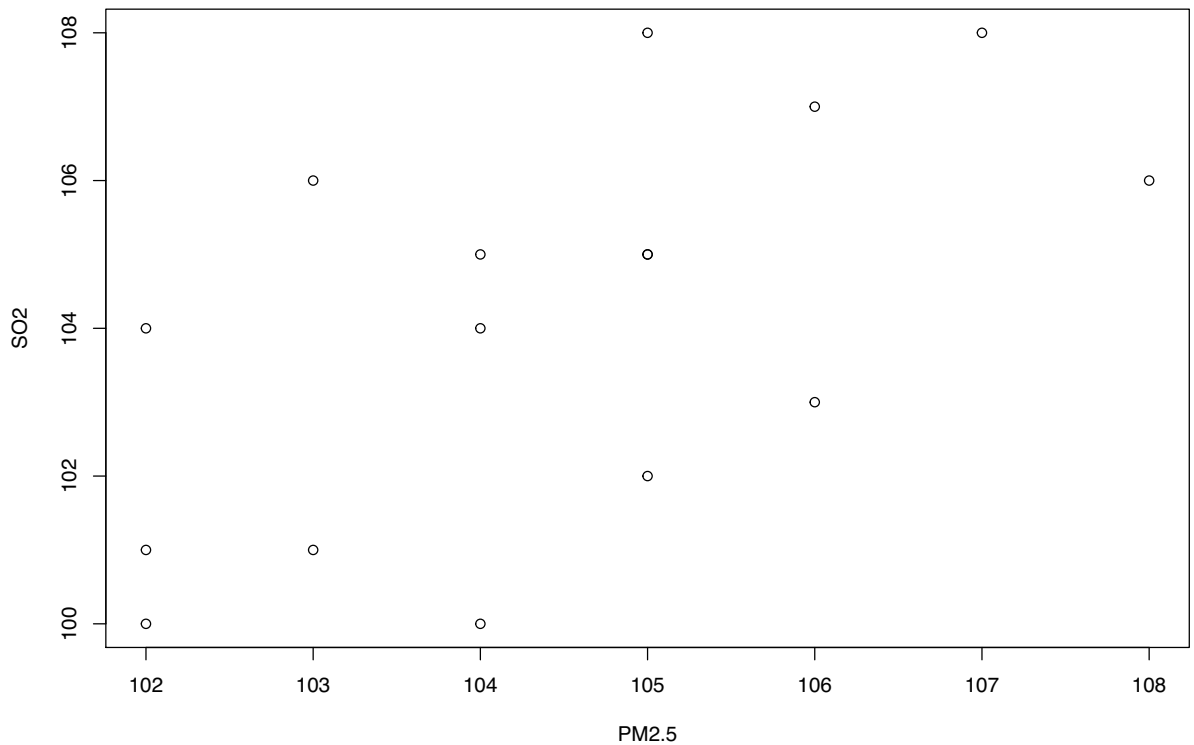
- Suppose we have very little idea about the value of Σ, so we want to choose a weak prior.
- This can be achieved by setting ν_0 to a small value, such as $\nu_0 = p+2$. (Note that the inverse-Wishart prior requires $\nu_0 > p+1$ to have a finite mean.)
- As for the prior mean $\boldsymbol{S}_0$, (e.g., $\boldsymbol{S}_0/(\nu_0 - p - 1)$), for simplicity, we use a diagonal matrix here $\nu_0 - p - 1 = 1$.

$$\boldsymbol{S}_0 = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix} = \Sigma_0$$

  where the diagonal is chosen to be 4, corresponding to marginal standard deviations of 2 for the device readings.
- The off-diagonals are set to 0 because we don't really know anything about how the device reading errors are correlated for the two pollutants.

Reading the data

```r
library(mvtnorm) # for drawing multivariate normal
library(MCMCpack) # for drawing inverse-Wishart
library(coda) # for MCMC diagnostics
options(digits=4) # tidy up print outs

y <- matrix(c(104,100,105,102,103,101,102,104,105,108,107,108,106,103,104,104,
              103,106,106,107,105,105,102,101,102,100,108,106,105,105,104,105),
            ncol=2,byrow=TRUE)
n <- nrow(y) # sample size
p <- ncol(y) # dimensionality

plot(y[,1],y[,2],xlab="PM2.5",ylab="SO2")
```



Prior specification

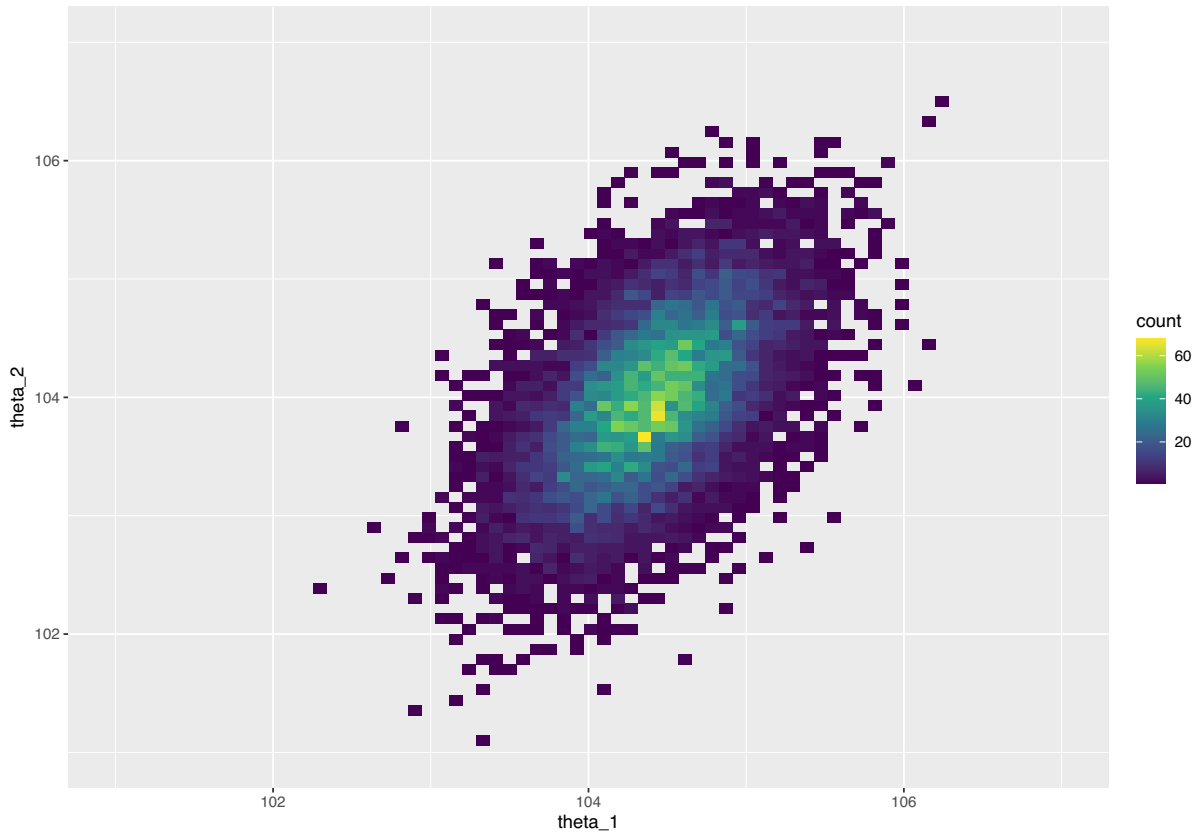
```
# prior for theta
mu.0 <- c(100,100)
Lambda.0 = matrix(c(100,15,15,25),ncol=2,byrow=TRUE)

# prior for Sigma
nu.0 <- p + 2  # a very weak prior
S0 <- (nu.0-p-1)*matrix(c(4,0,0,4),ncol=2,byrow=TRUE)

ybar <- apply(y,2,mean)
nu.n <- nu.0 + n
```

Gibbs sampling

```
niter <- 10000 # total number of iterations
nburnin <- 1000 # 1000 burn-in steps

THETA <- matrix(NA,nrow=niter,ncol=p) # matrix for storing the draws for theta
colnames(THETA) <- c("theta1","theta2")

THETA.init <- ybar # Initial values set to sample mean
THETA.curr <- THETA.init # the theta value at current iteration

SIGMA <- matrix(NA,nrow=niter,ncol=p*p) # matrix for storing the draws for Sigma
colnames(SIGMA) <- c("sigma11","sigma12","sigma21","sigma22")

SIGMA.init <- cov(y) # intial value set to sample covariance
SIGMA.curr <- SIGMA.init # the Sigma value at current iternation

### Start Gibbs sampling
for (t in 1:niter) {

  ## Update theta
  Lambda.n <- solve(n*solve(SIGMA.curr)+solve(Lambda.0))
  mu.n <- Lambda.n %*% (n*solve(SIGMA.curr,ybar)+solve(Lambda.0,mu.0))
  THETA.curr <- rmvnorm(1,mean=mu.n,sigma=Lambda.n)

  ## Update Sigma
  S.theta <- (t(y)-c(THETA.curr))%*% t(t(y)-c(THETA.curr))
  SIGMA.curr <- riwish(v=nu.n,S=S0+S.theta)

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

}
```

Histogram of MCMC draws for $\boldsymbol{\theta}$

```
ggplot(data.frame(THETA), aes(x=theta1, y=theta2) ) +
  labs(x=expression(theta_1),y=expression(theta_2)) +
  geom_bin2d(bins=70) +
  scale_fill_continuous(type = "viridis") +
  lims(x=c(101,107),y=c(101,107))
```



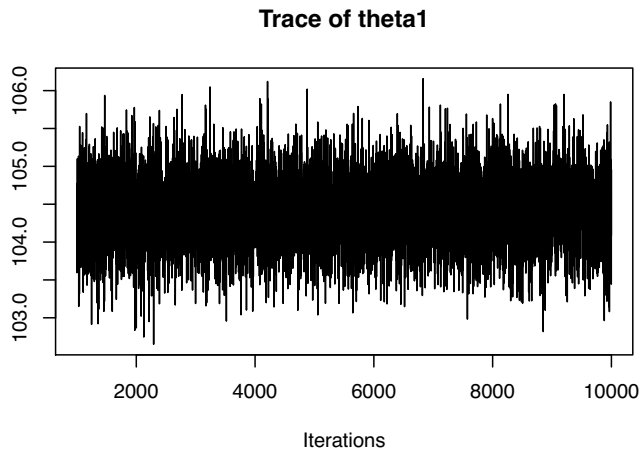
MCMC diagnostics

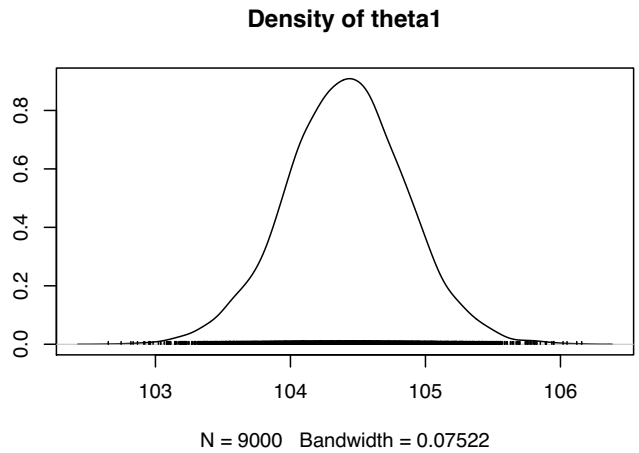
```
THETA.mcmc <- mcmc(THETA[-(1:nburnin),],start=nburnin+1)
summary(THETA.mcmc)
```

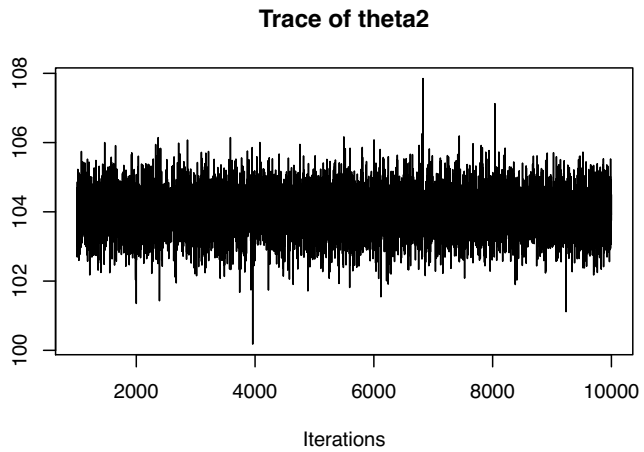
```
##
## Iterations = 1001:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 9000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##        Mean    SD Naive SE Time-series SE
## theta1  104 0.443  0.00467        0.00467
## theta2  104 0.649  0.00684        0.00697
##
## 2. Quantiles for each variable:
##
##        2.5% 25% 50% 75% 97.5%
## theta1  104 104 104 105   105
## theta2  103 104 104 104   105
```

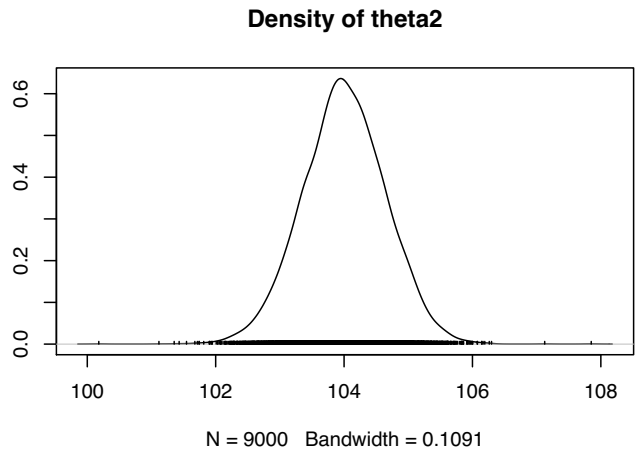
Trace plots for $\boldsymbol{\theta}$

```
plot(THETA.mcmc)
```









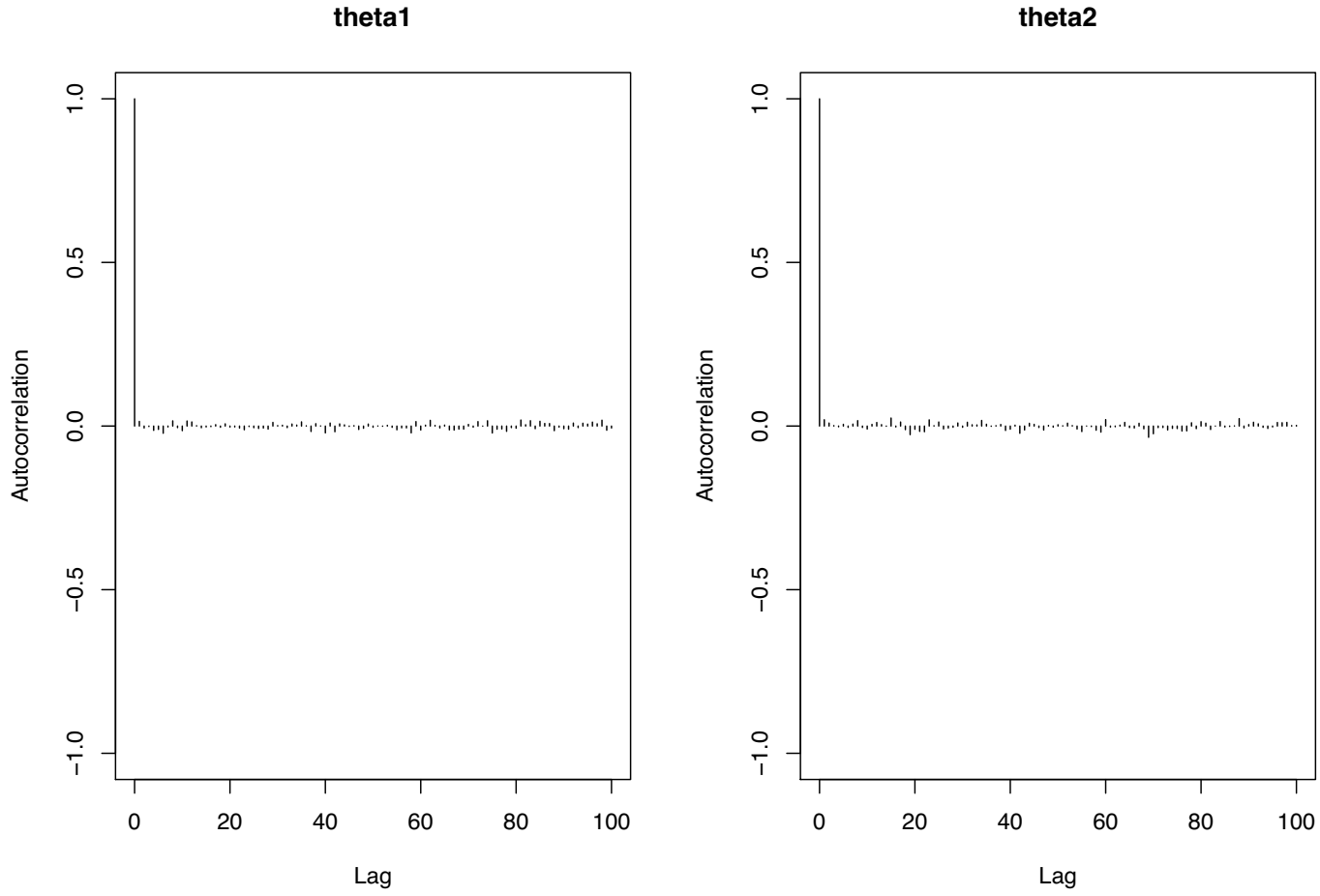
ESS for $\boldsymbol{\theta}$

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

```
## theta1 theta2
##   9000   8658
```

Autocorrelation plot for $\boldsymbol{\theta}$

```
autocorr.plot(THETA.mcmc,lag.max=100)
```



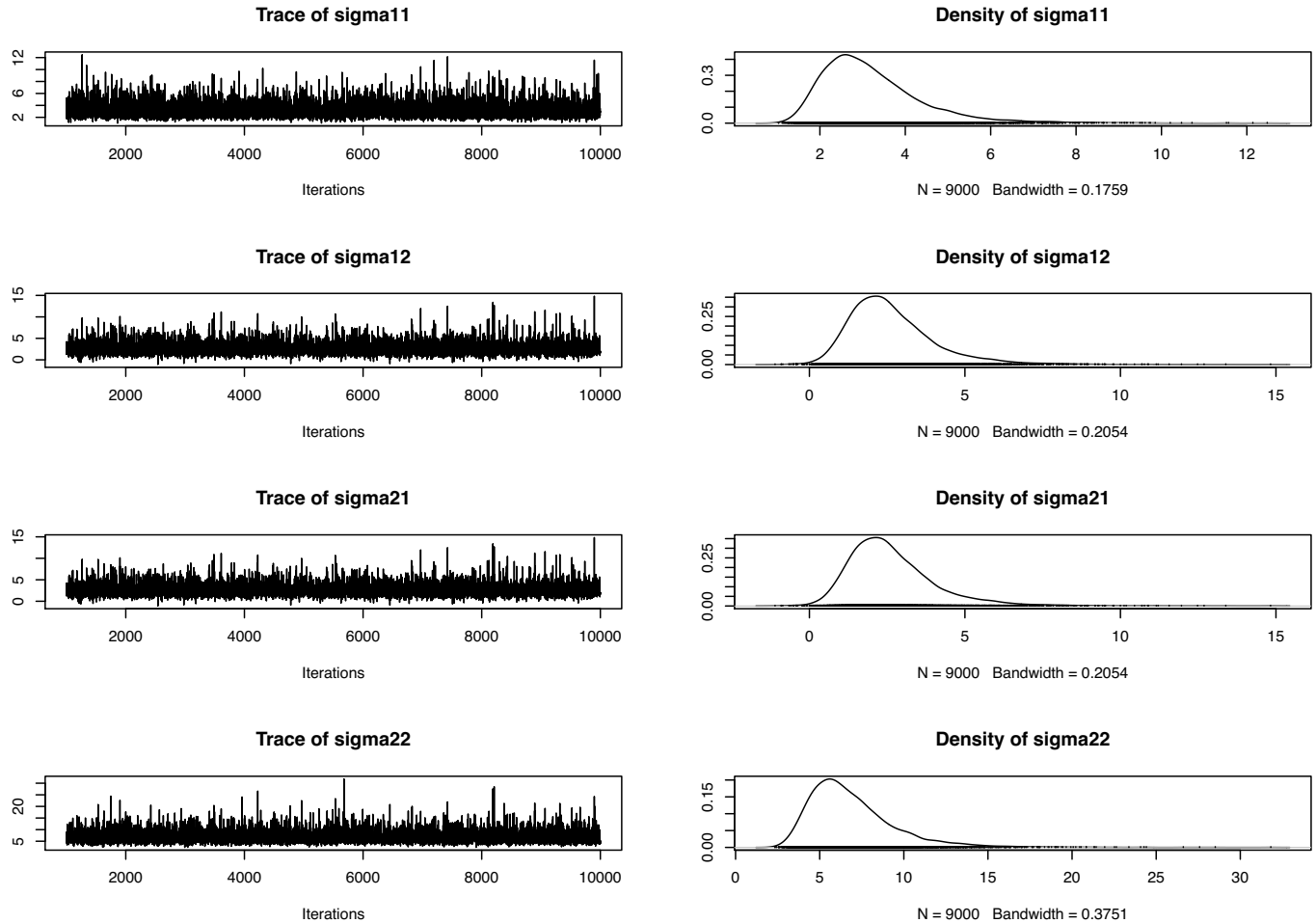
MCMC diagnostics

```
SIGMA.mcmc <- mcmc(SIGMA[-(1:nburnin),],start=nburnin+1)
summary(SIGMA.mcmc)
```

```
##
## Iterations = 1001:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 9000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##          Mean   SD Naive SE Time-series SE
## sigma11 3.22 1.20   0.0127         0.0132
## sigma12 2.64 1.41   0.0149         0.0160
## sigma21 2.64 1.41   0.0149         0.0160
## sigma22 6.90 2.59   0.0273         0.0295
##
## 2. Quantiles for each variable:
##
##          2.5%  25%  50%  75% 97.5%
## sigma11 1.635 2.39 2.98 3.77  6.29
## sigma12 0.652 1.69 2.39 3.30  6.10
## sigma21 0.652 1.69 2.39 3.30  6.10
## sigma22 3.521 5.13 6.37 8.06 13.39
```

Trace plots for Σ

```
plot(SIGMA.mcmc)
```



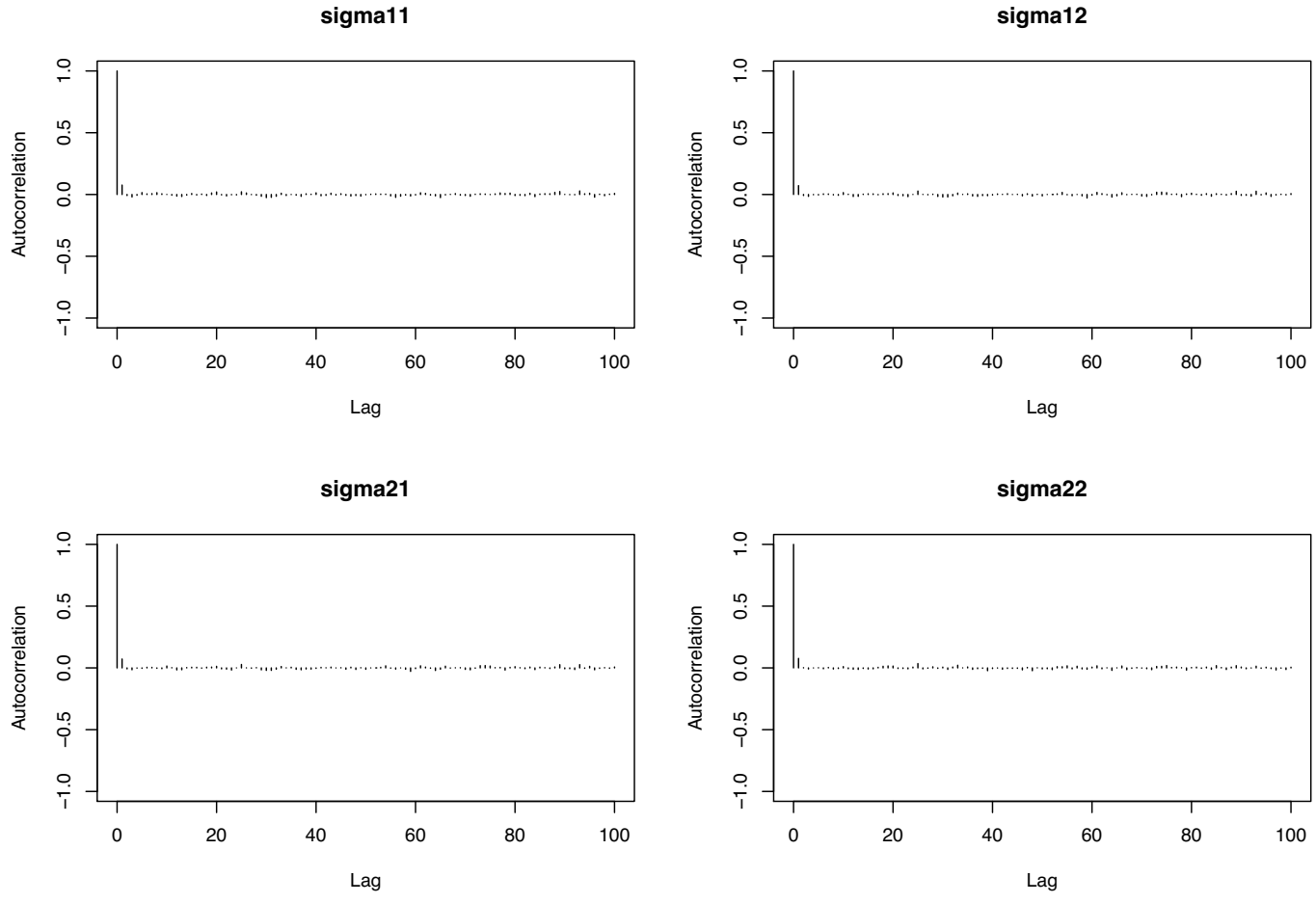
ESS for Σ

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

```
## sigma11 sigma12 sigma21 sigma22
##    8239    7791    7791    7710
```

Autocorrelation plot for Σ

```
autocorr.plot(SIGMA.mcmc,lag.max=100)
```



With a stronger prior on Σ with $\nu_0 = 50$

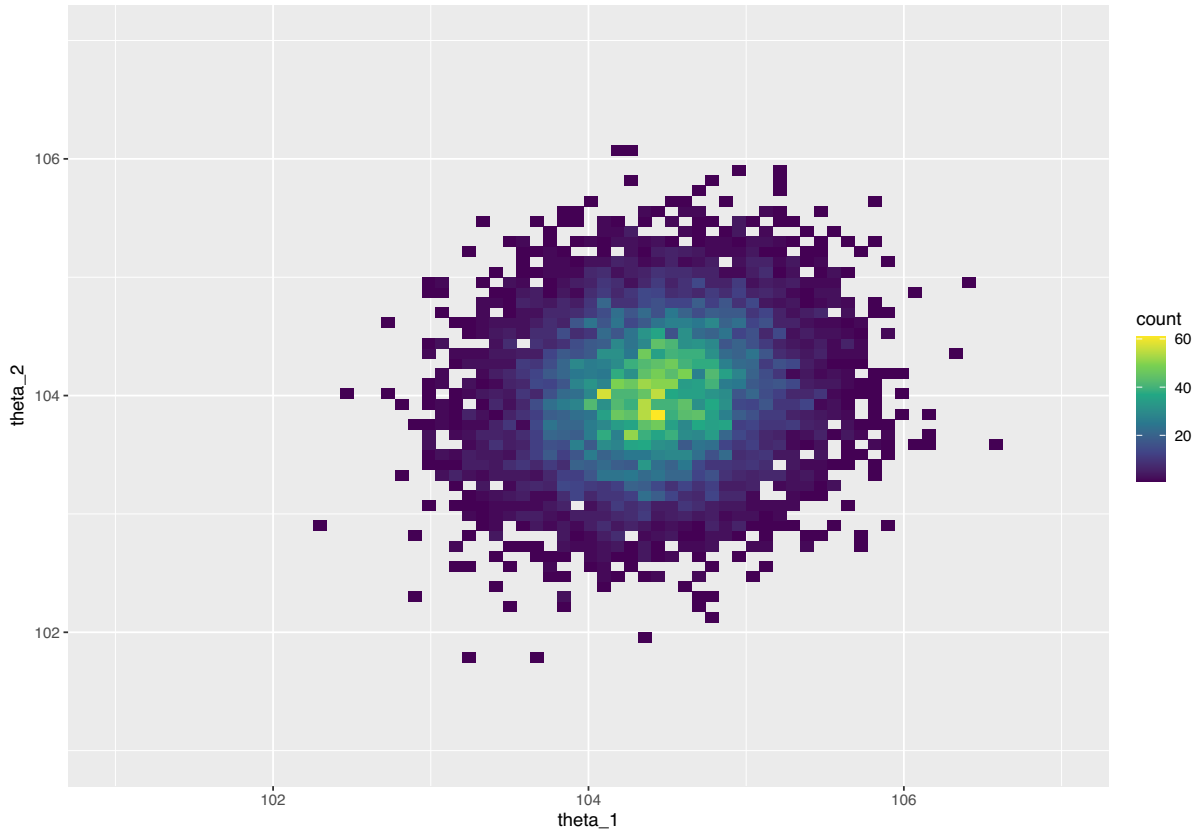
- Suppose we have a better idea about the value of Σ, so we want to choose a stronger prior.
- For example we set $\nu_0 = 50$.
- Now the prior mean $\boldsymbol{S}_0/(\nu_0 - p - 1) = \boldsymbol{S}_0/47$. Thus if we want the same prior mean as before, we set

$$\boldsymbol{S}_0 = 47 \cdot \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix} = \cdot \begin{pmatrix} 188 & 0 \\ 0 & 188 \end{pmatrix}$$

```
# prior for theta
mu.0 <- c(100,100)
Lambda.0 = matrix(c(100,15,15,25),ncol=2,byrow=TRUE)

# prior for Sigma
nu.0 <- 50  # a stronger prior on covariance now
S0 <- (nu.0-p-1) * matrix(c(4,0,0,4),ncol=2,byrow=TRUE) # maintaining the same prior mean

ybar <- apply(y,2,mean)
nu.n <- nu.0 + n
```

Histogram of MCMC draws for $\boldsymbol{\theta}$

```
ggplot(data.frame(THETA), aes(x=theta1, y=theta2) ) +
  labs(x=expression(theta_1),y=expression(theta_2)) +
  geom_bin2d(bins=70) +
  scale_fill_continuous(type = "viridis") +
  lims(x=c(101,107),y=c(101,107))
```



- How does this compare to the case with $\nu_0 = 4$?
- Why is the posterior spread of $\boldsymbol{\theta}$ smaller?

a tight prior of Σ propogates to the tight posterior of θ. $(\Lambda_n^{-1} = n\Sigma^{-1} + \Lambda_0^{-1})$

(c) Suppose we decided to adopt the following Normal-inverse-Wishart prior on $(\boldsymbol{\theta}, \Sigma)$. That is

$$p(\boldsymbol{\theta}, \Sigma) = p(\boldsymbol{\theta})p(\Sigma)$$

where

$$\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\mu}_0, \Lambda_0) \quad \text{and} \quad \Sigma \sim \mathrm{IW}(\nu_0, \boldsymbol{S}_0).$$

with $\boldsymbol{\mu}_0 = (100, 100)'$, $\Lambda_0 = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$, and $\boldsymbol{S}_0 = (\nu_0 - p - 1)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. How does increasing ν_0, the prior degree of freedom, affect the posterior variance of $\boldsymbol{\theta}$? Explain your reasoning. Hint: You can explain this using the full conditional and/or Gibbs update for $\boldsymbol{\theta}$.

(c)

The full conditionals are

$$\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\Sigma} \sim \mathrm{N}(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n)$$

$$\boldsymbol{\Sigma}|\boldsymbol{x}, \boldsymbol{\theta} \sim \mathrm{inverseWishart}(\nu_n, \boldsymbol{S}_n)$$

where $\boldsymbol{\Lambda}_n = (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}, \mu_n = \boldsymbol{\Lambda}_n(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}}), \nu_n = \nu_0 + n, \boldsymbol{S}_n = \boldsymbol{S}_0 + \boldsymbol{S}_\theta, \boldsymbol{S}_\theta = \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\theta})(\boldsymbol{x}_i - \boldsymbol{\theta})^T$.

Since $\mathrm{E}(\boldsymbol{\Sigma}|\boldsymbol{x}, \boldsymbol{\theta}) = (\nu_0 + n - p - 1)^{-1}\boldsymbol{S}_n = \frac{\nu_0 - p - 1}{\nu_0 + n - p - 1}\boldsymbol{I} + \frac{n}{\nu_0 + n - p - 1}(\boldsymbol{S}_\theta/n)$, which is a weighted average of $\boldsymbol{I}$ and $(\boldsymbol{S}_\theta/n)$, then as ν_0 increases, if $(\boldsymbol{S}_\theta/n)$ is much more precise than $\boldsymbol{I}$, the posterior variance of $\boldsymbol{\theta}$ will increase. Otherwise the posterior variance will be reduced.

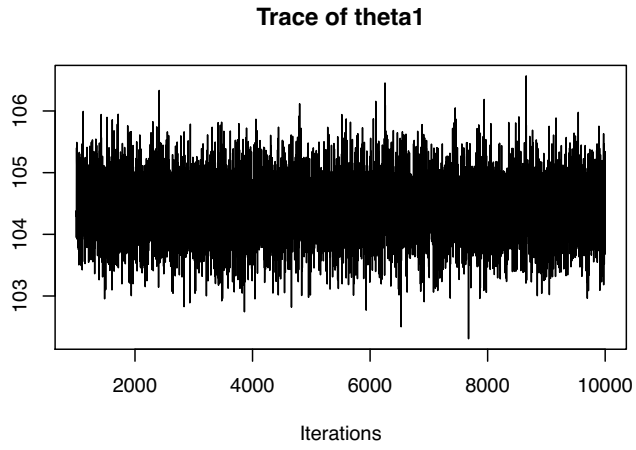
MCMC diagnostics

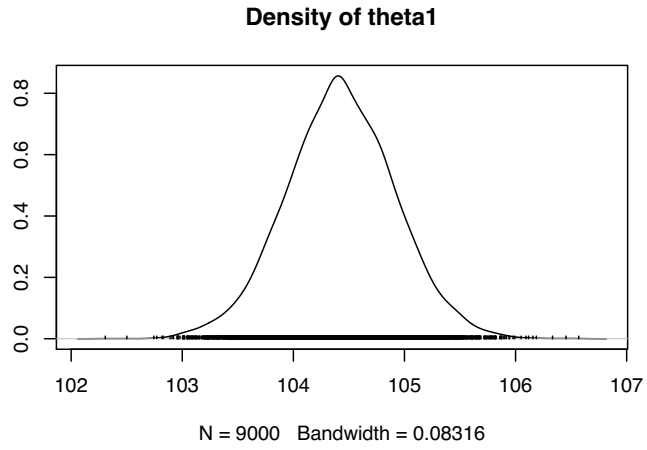
```
THETA.mcmc <- mcmc(THETA[-(1:nburnin),],start=nburnin+1)
summary(THETA.mcmc)
```

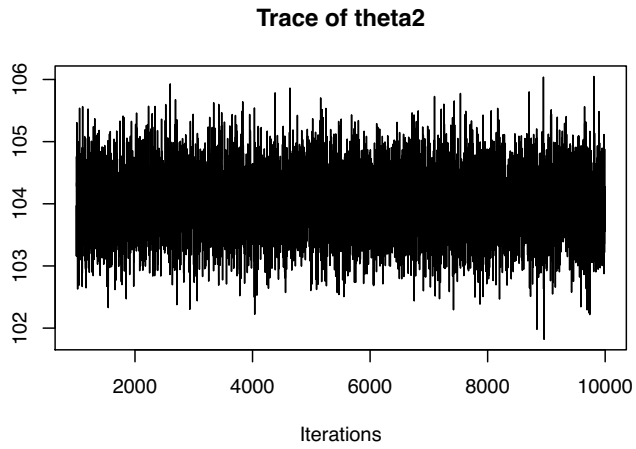
```
##
## Iterations = 1001:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 9000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##        Mean    SD Naive SE Time-series SE
## theta1  104 0.490  0.00517        0.00528
## theta2  104 0.531  0.00560        0.00571
##
## 2. Quantiles for each variable:
##
##        2.5% 25% 50% 75% 97.5%
## theta1  103 104 104 105   105
## theta2  103 104 104 104   105
```

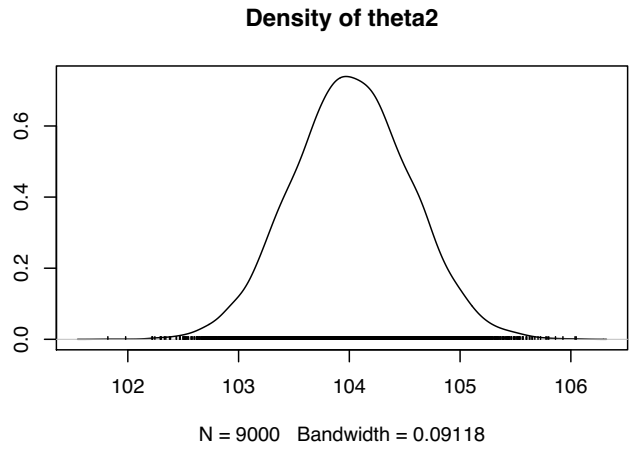
Trace plots for $\boldsymbol{\theta}$

```
plot(THETA.mcmc)
```









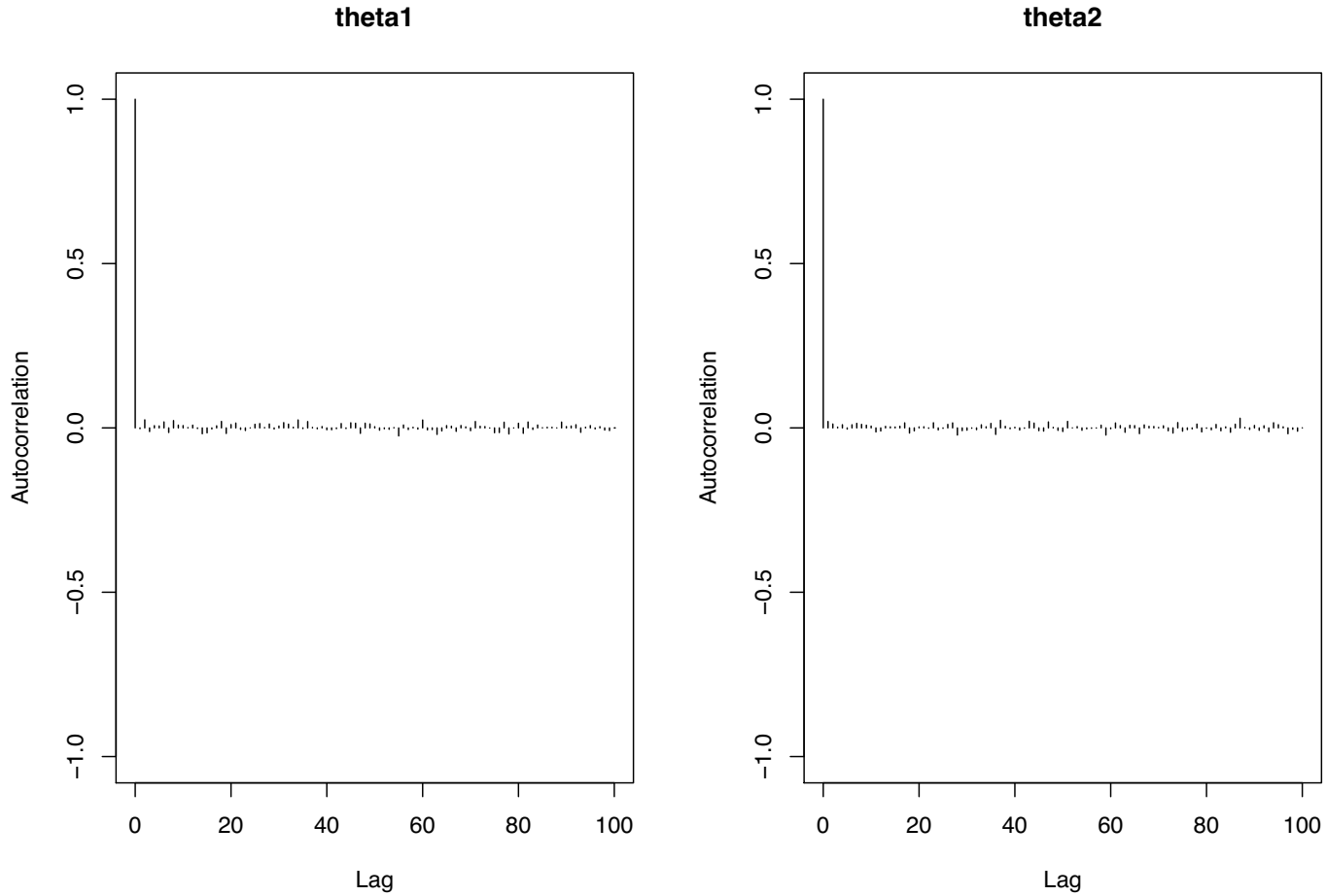
ESS for $\boldsymbol{\theta}$

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

```
## theta1 theta2
##   8607   8654
```

Autocorrelation plot for $\boldsymbol{\theta}$

```
autocorr.plot(THETA.mcmc,lag.max=100)
```



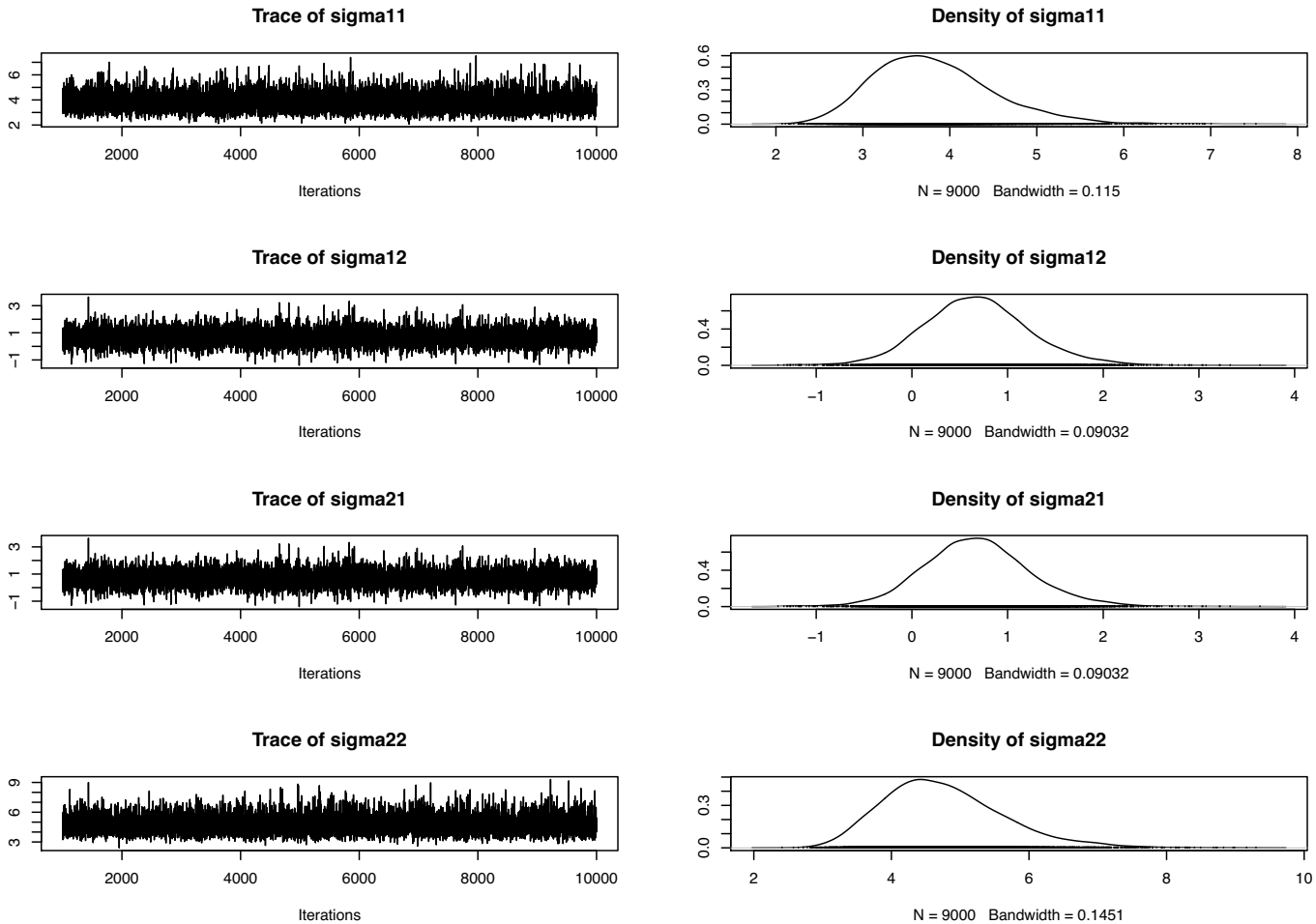
MCMC diagnostics

```
SIGMA.mcmc <- mcmc(SIGMA[-(1:nburnin),],start=nburnin+1)
summary(SIGMA.mcmc)
```

```
##
## Iterations = 1001:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 9000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##          Mean    SD Naive SE Time-series SE
## sigma11 3.818 0.692  0.00730        0.00748
## sigma12 0.678 0.559  0.00589        0.00590
## sigma21 0.678 0.559  0.00589        0.00590
## sigma22 4.761 0.871  0.00918        0.00933
##
## 2. Quantiles for each variable:
##
##           2.5%   25%   50%  75% 97.5%
## sigma11  2.682 3.323 3.743 4.22  5.38
## sigma12 -0.374 0.314 0.662 1.02  1.86
## sigma21 -0.374 0.314 0.662 1.02  1.86
## sigma22  3.354 4.142 4.663 5.28  6.78
```

Trace plots for Σ

```
plot(SIGMA.mcmc)
```



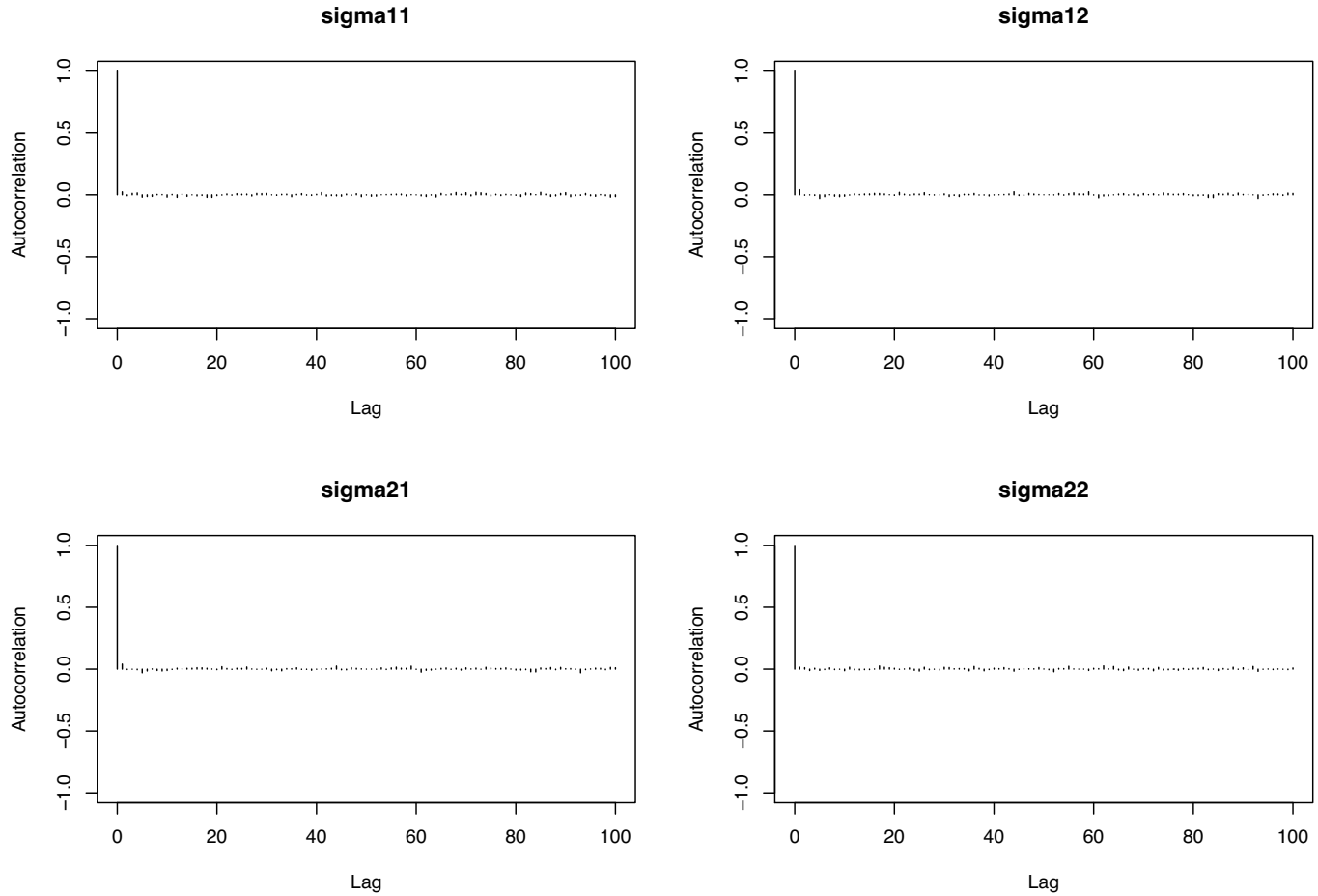
ESS for Σ

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

```
## sigma11 sigma12 sigma21 sigma22
##    8572    8977    8977    8715
```

Autocorrelation plot for Σ

```
autocorr.plot(SIGMA.mcmc,lag.max=100)
```



Stronger prior on $\boldsymbol{\theta}$

prior:

* $\theta \sim N(\mu_0, \Lambda_0)$
  $\Sigma \sim IG(\nu_0, S_0)$

* $x_1, \ldots, x_n \mid \theta, \Sigma \sim N(\theta, \Sigma)$

posterior:

* $\theta \mid \Sigma, \underline{x} \sim N\left[(\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}),\ (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}\right]$
  $\Sigma \mid \theta, \underline{x} \sim IG(\nu_0 + n,\ S_0 + S_\theta)$

- Suppose the historical distribution of PM2.5 has a standard deviation of about *1* (vs 10 before), and SO2 has a standard deviation of about *0.5* vs (5 before), still with a correlation of about 0.3. Turning this into our prior,

$$\Lambda_0 = \begin{pmatrix} 1^2 & 0.3 \times 1 \times 0.5 \\ 0.3 \times 1 \times 0.5 & 0.5^2 \end{pmatrix} = \begin{pmatrix} 1 & 0.15 \\ 0.15 & 0.25 \end{pmatrix}.$$

With weak prior on Σ: $\nu_0 = p + 2$

```
ggplot(data.frame(THETA), aes(x=theta1, y=theta2) ) +
  labs(x=expression(theta_1),y=expression(theta_2)) +
  geom_bin2d(bins=70) +
  scale_fill_continuous(type = "viridis") +
  lims(x=c(99,105),y=c(99,105))
```

- Why is there a lot *more* shrinkage toward prior mean?

Consider:

* $\theta \mid \Sigma, \underline{x} \sim N\left[(\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}),\ (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}\right]$

  $\Sigma \mid \theta, \underline{x} \sim IG(\nu_0 + n,\ S_0 + S_\theta)$, where $S_\theta = \sum_{i=1}^{n}(x_i - \theta)(x_i - \theta)^T$

Reason 1:

Here Λ_0 decreases, so the weight $\frac{\Lambda_0^{-1}}{\Lambda_0^{-1} + n\Sigma^{-1}}$ of prior mean increases.

Reason 2:

θ is pulled more away from $\underline{x}$. So $S_\theta \uparrow$, $\mathbb{E}[\Sigma \mid \theta, \underline{x}] = \frac{S_0 + S_\theta}{\nu_0 + n - p - 1} \uparrow$

This makes the weight $\frac{n\Sigma^{-1}}{\Lambda_0^{-1} + n\Sigma^{-1}}$ of empirical mean decreases.

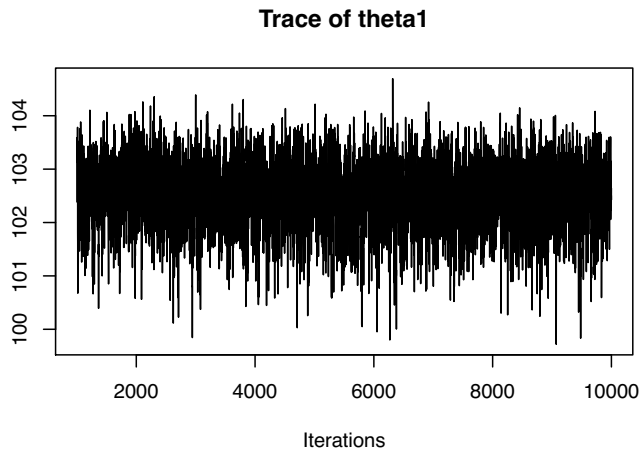
MCMC diagnostics

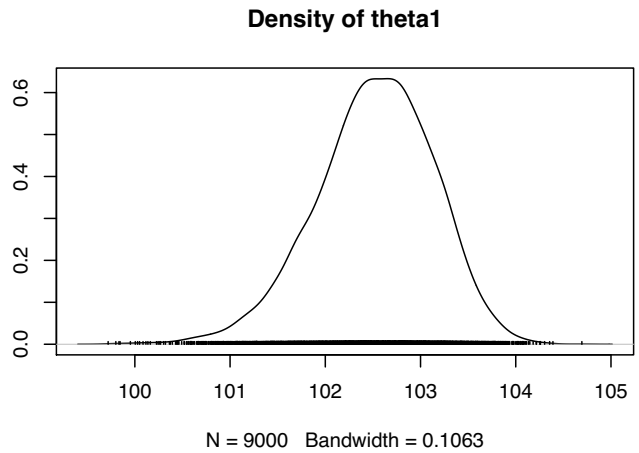
```
THETA.mcmc <- mcmc(THETA[-(1:nburnin),],start=nburnin+1)
summary(THETA.mcmc)
```

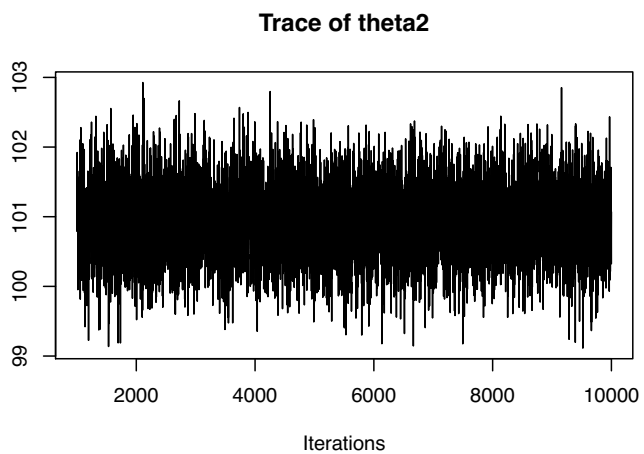
```
##
## Iterations = 1001:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 9000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##        Mean    SD Naive SE Time-series SE
## theta1  103 0.622  0.00656        0.01209
## theta2  101 0.511  0.00539        0.00716
##
## 2. Quantiles for each variable:
##
##         2.5% 25% 50% 75% 97.5%
## theta1 101.2 102 103 103   104
## theta2  99.9 101 101 101   102
```

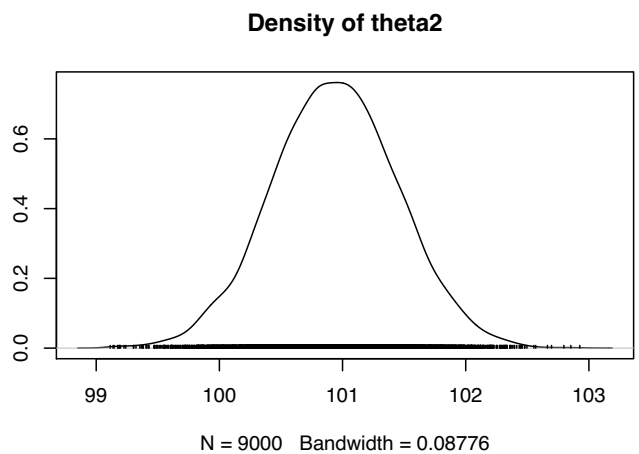
Trace plots for $\boldsymbol{\theta}$

```
plot(THETA.mcmc)
```









ESS for $\boldsymbol{\theta}$

The full conditional densities are:

* $\theta \mid \Sigma, \underline{x} \sim N\left[(\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}),\ (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}\right]$

  $\Sigma \mid \theta, \underline{x} \sim IG(\nu_0 + n,\ S_0 + S_\theta)$, where $S_\theta = \sum_{i=1}^{n}(x_i - \theta)(x_i - \theta)^T$

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

```
## theta1 theta2
##   2647   5109
```

→ Here $\Lambda_0 \downarrow$, θ are pulled away from $\underline{x}$.

- This stronger prior induced much stronger autocorrelation and smaller ESS.
- Why?
- The prior concentrated near a value far from the empirical estimate.
- Stronger posterior dependency between $\boldsymbol{\theta}$ and Σ.

As a result, $\theta^{(t)}$ highly depends on $\theta^{(t-1)}$.

It's the same for $\Sigma^{(t)}$ and $\Sigma^{(t-1)}$.

Explanation more intuitive:

* If the prior is concentrated, then the posterior distribution is also concentrated.
* If you have a θ that's far away from $\bar{x}$,

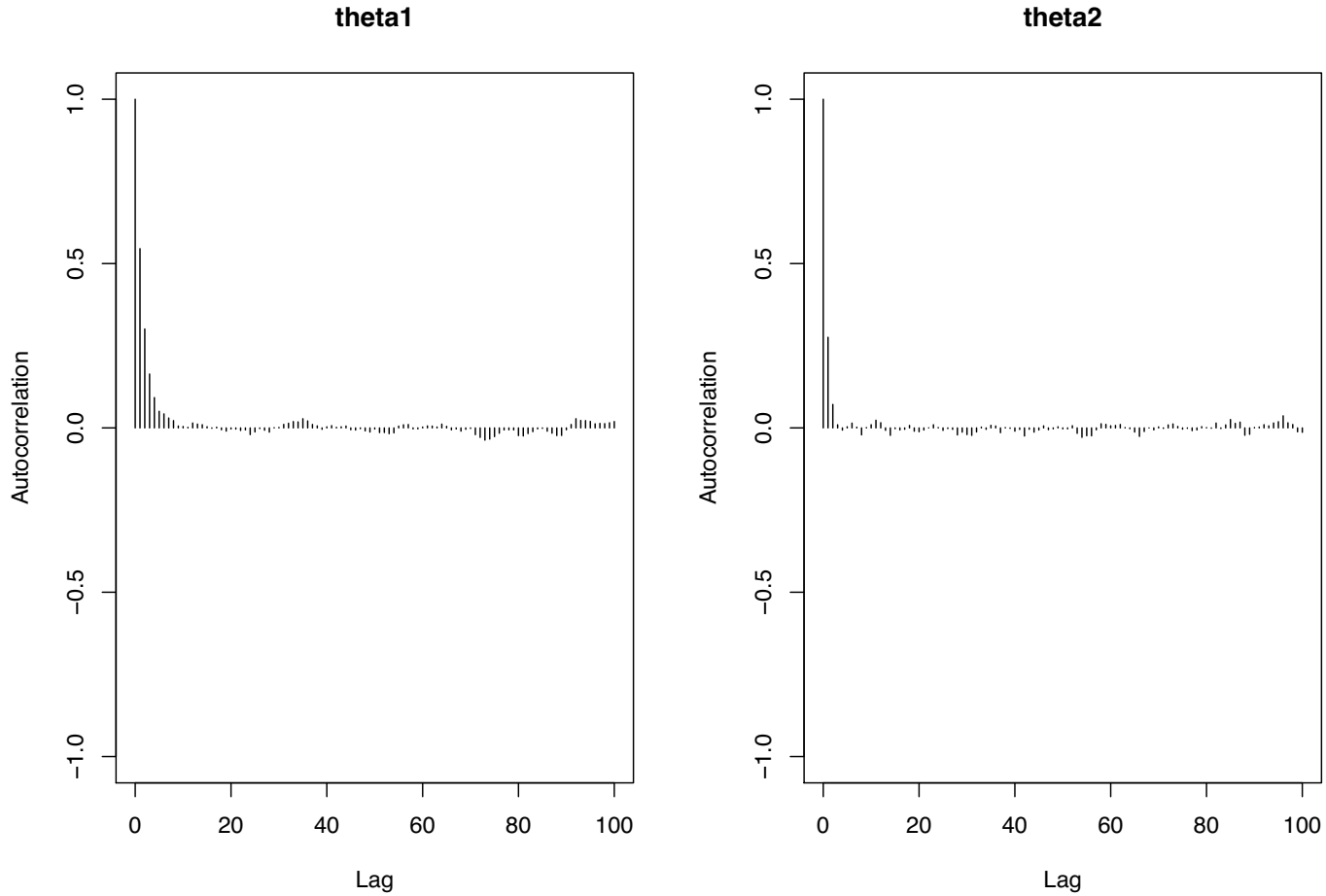
  then Σ must be large to allow you to occassionally have high measurement errors.

  $\Rightarrow$ If you're close, then Σ is small

  $\Rightarrow$ θ and Σ are strongly correlated.

Autocorrelation plot for $\boldsymbol{\theta}$

```
autocorr.plot(THETA.mcmc,lag.max=100)
```



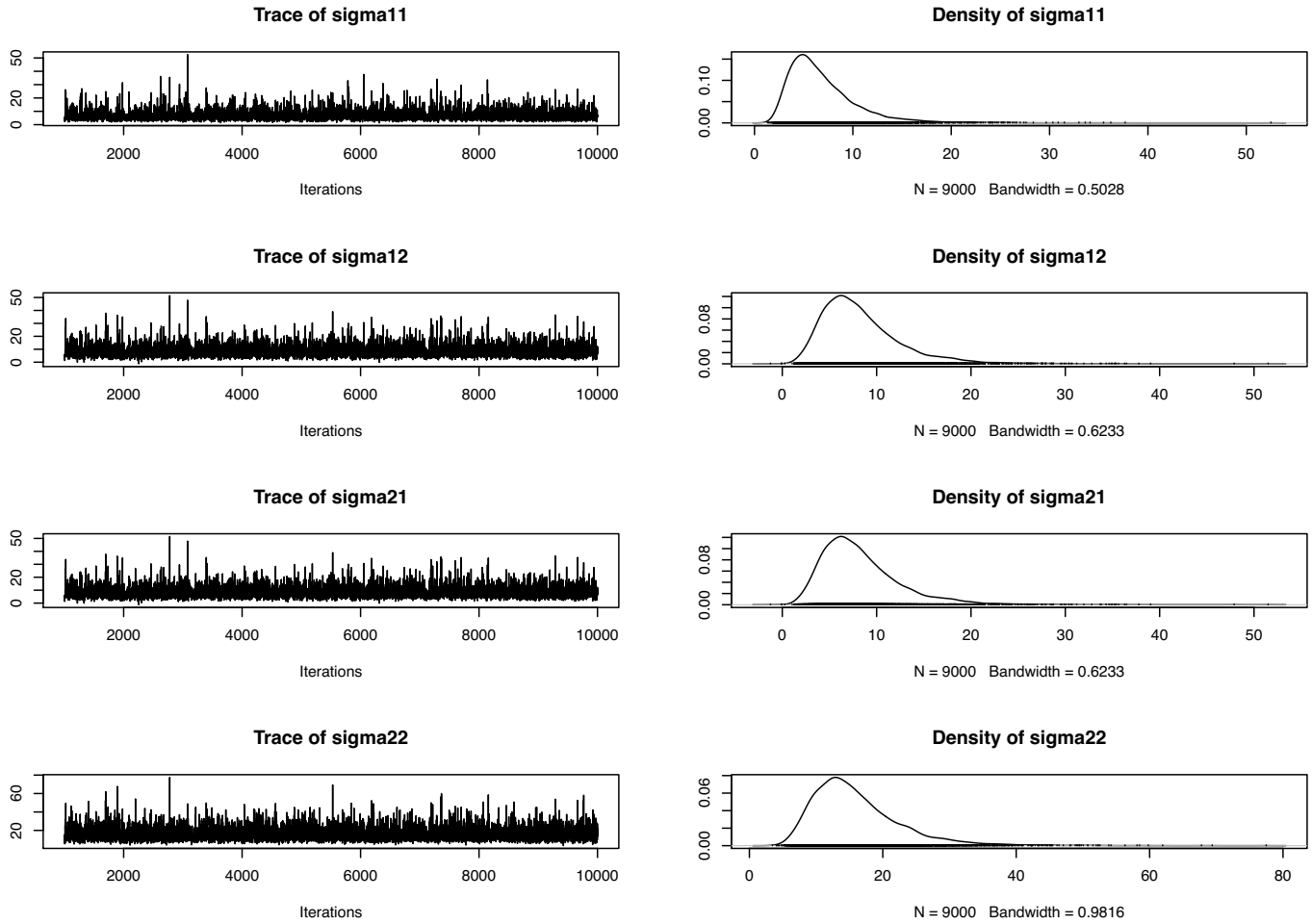
MCMC diagnostics

```
SIGMA.mcmc <- mcmc(SIGMA[-(1:nburnin),],start=nburnin+1)
summary(SIGMA.mcmc)
```

```
##
## Iterations = 1001:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 9000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##          Mean   SD Naive SE Time-series SE
## sigma11  6.91 3.63   0.0383         0.0654
## sigma12  8.33 4.34   0.0457         0.0693
## sigma21  8.33 4.34   0.0457         0.0693
## sigma22 16.03 6.68   0.0705         0.0888
##
## 2. Quantiles for each variable:
##
##         2.5%   25%   50%   75% 97.5%
## sigma11 2.62  4.47  6.05  8.39  16.2
## sigma12 2.69  5.38  7.47 10.24  19.0
## sigma21 2.69  5.38  7.47 10.24  19.0
## sigma22 7.15 11.42 14.69 19.09  32.6
```

Trace plots for Σ

```
plot(SIGMA.mcmc)
```



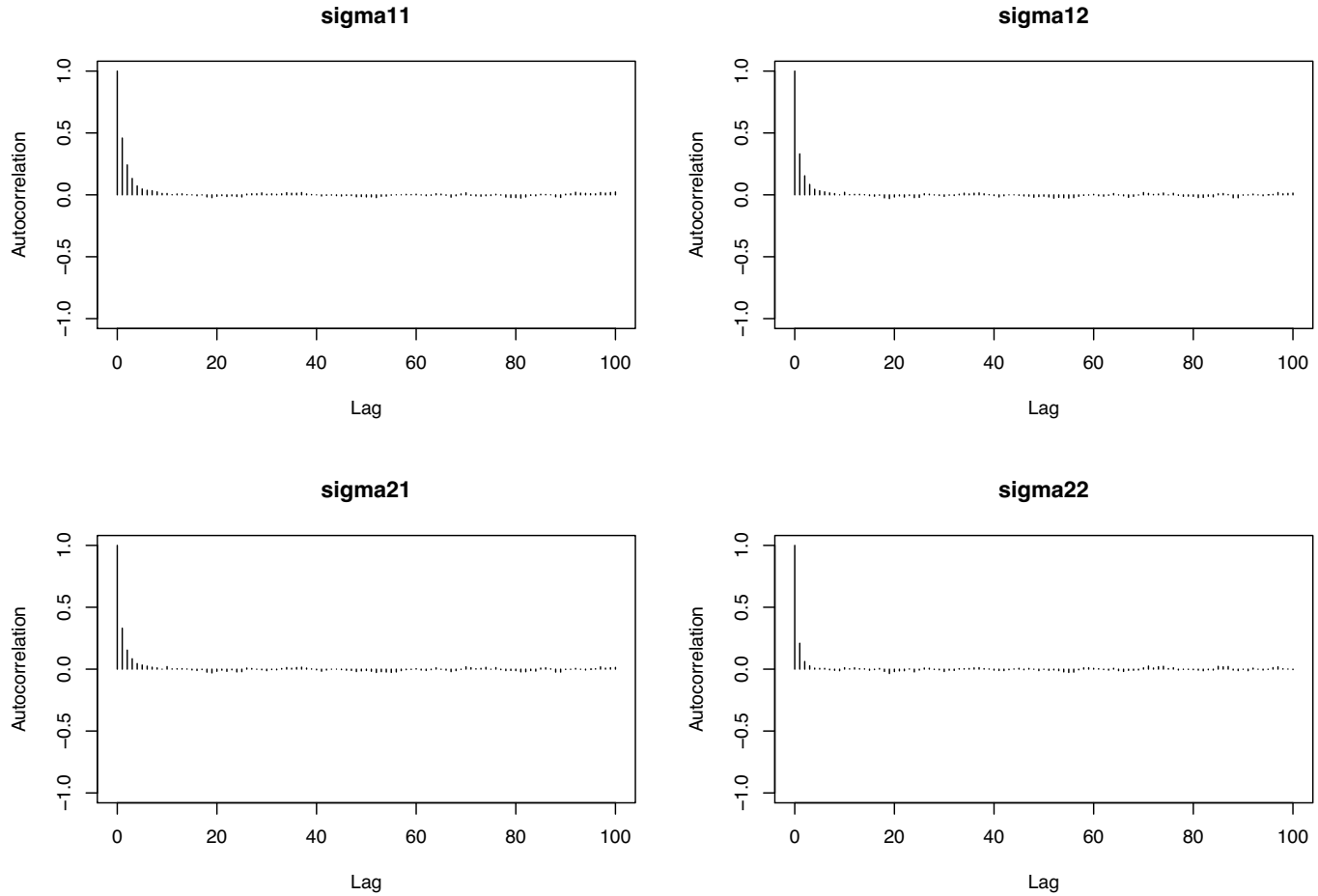
ESS for Σ

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

```
## sigma11 sigma12 sigma21 sigma22
##    3081    3919    3919    5664
```

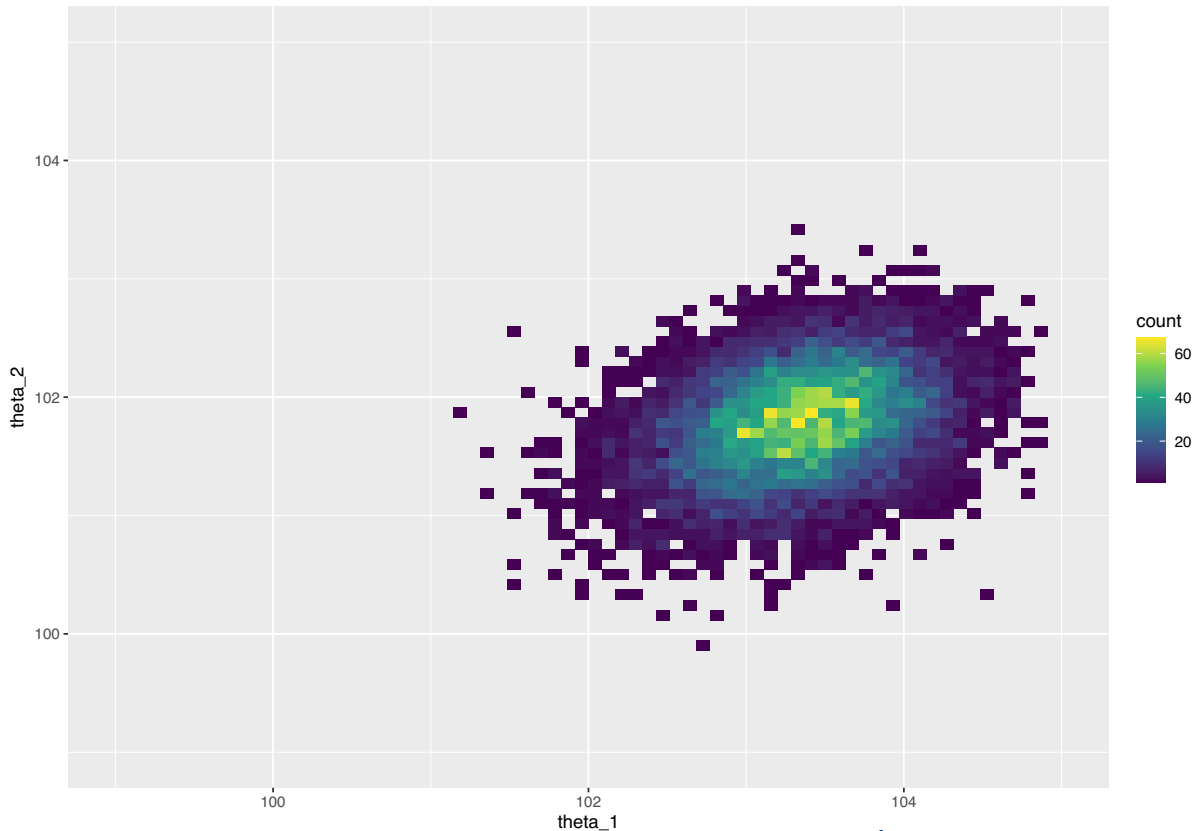
Autocorrelation plot for Σ

```
autocorr.plot(SIGMA.mcmc,lag.max=100)
```



With a strong prior on Σ: $\nu_0 = 50$

```
ggplot(data.frame(THETA), aes(x=theta1, y=theta2) ) +
  labs(x=expression(theta_1),y=expression(theta_2)) +
  geom_bin2d(bins=70) +
  scale_fill_continuous(type = "viridis") +
  lims(x=c(99,105),y=c(99,105))
```



- Why is there a lot *less* shrinkage toward prior mean?

$* \quad \theta \mid \Sigma, \underline{x} \sim N\left[(\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}),\ (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}\right]$

$\Sigma \mid \theta, \underline{x} \sim IG(\nu_0 + n,\ S_0 + S_\theta)$

B/C Σ decreases:

$\mathbb{E}[\Sigma \mid \theta, \underline{x}) = \frac{S_0 + S_\theta}{\nu_0 + n - p - 1}$ ↓ with ν_0 ↑ and S_θ ↓ $\left(S_\theta = \sum_{i=1}^{n}(x_i - \theta)(x_i - \theta)^T\right)$

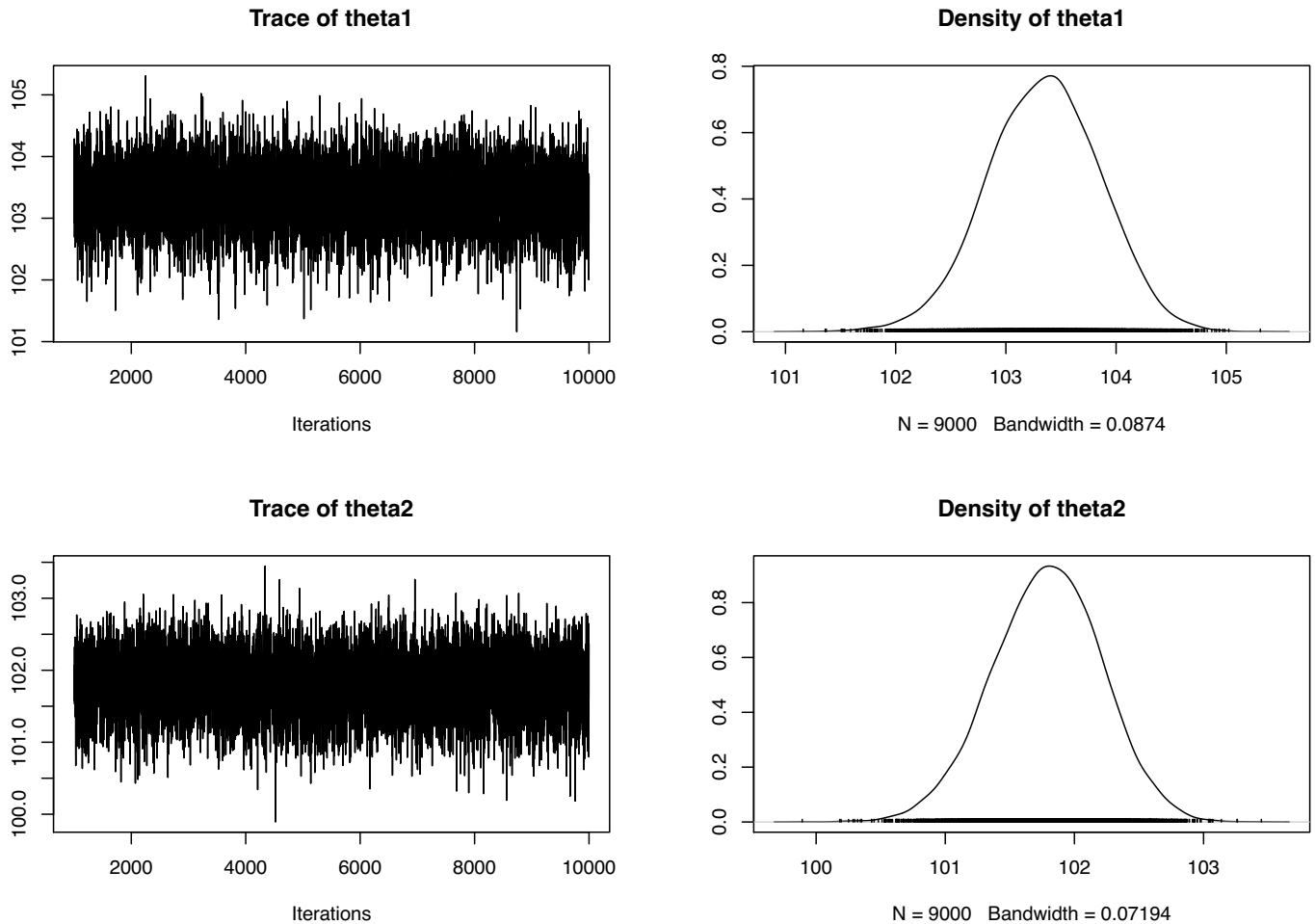
MCMC diagnostics

```
THETA.mcmc <- mcmc(THETA[-(1:nburnin),],start=nburnin+1)
summary(THETA.mcmc)
```

```
##
## Iterations = 1001:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 9000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##        Mean    SD Naive SE Time-series SE
## theta1  103 0.509  0.00537        0.00677
## theta2  102 0.419  0.00442        0.00534
##
## 2. Quantiles for each variable:
##
##        2.5% 25% 50% 75% 97.5%
## theta1  102 103 103 104   104
## theta2  101 102 102 102   103
```

Trace plots for $\boldsymbol{\theta}$

```
plot(THETA.mcmc)
```



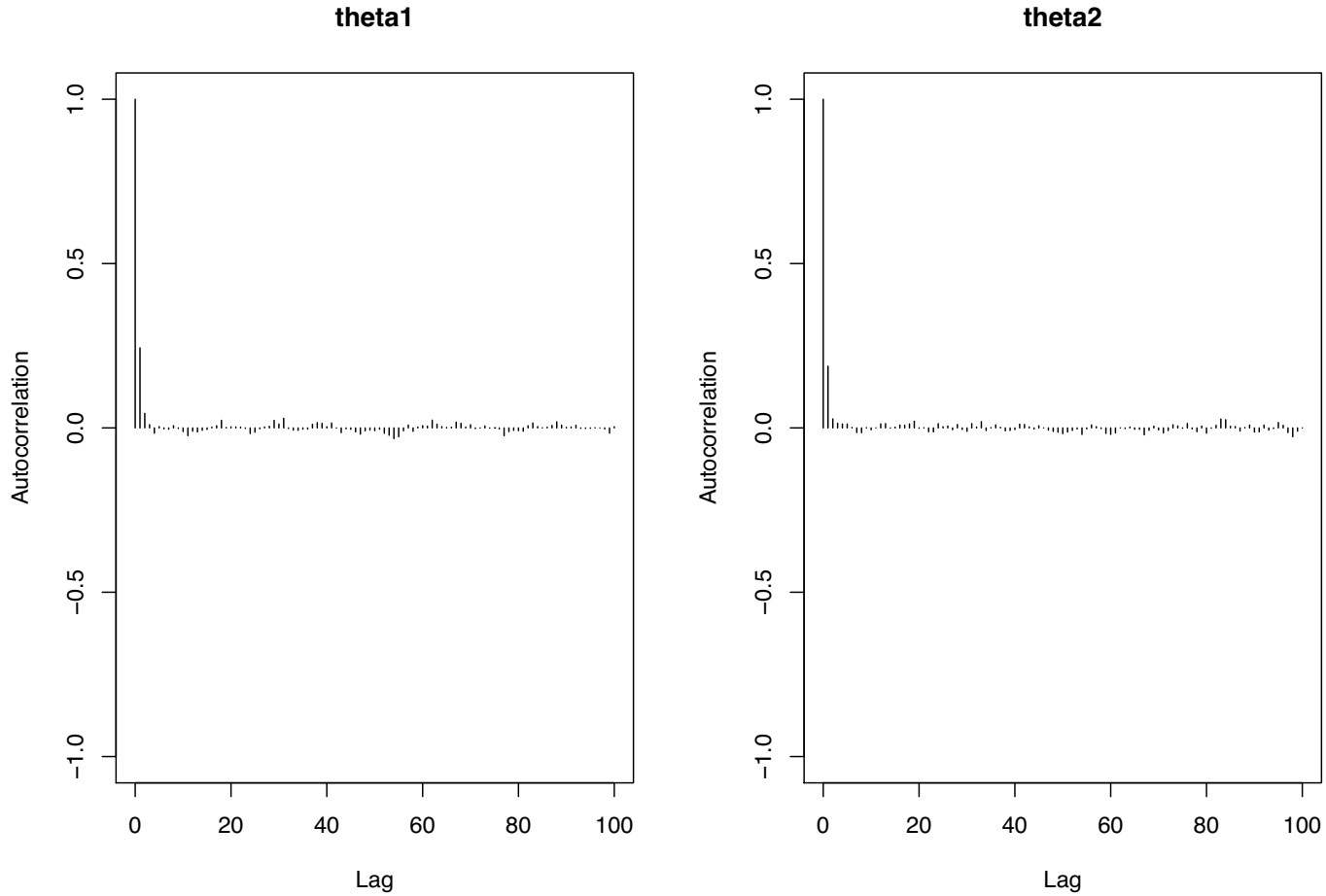
ESS for $\boldsymbol{\theta}$

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

```
## theta1 theta2
##   5654   6155
```

Autocorrelation plot for $\boldsymbol{\theta}$

```
autocorr.plot(THETA.mcmc,lag.max=100)
```



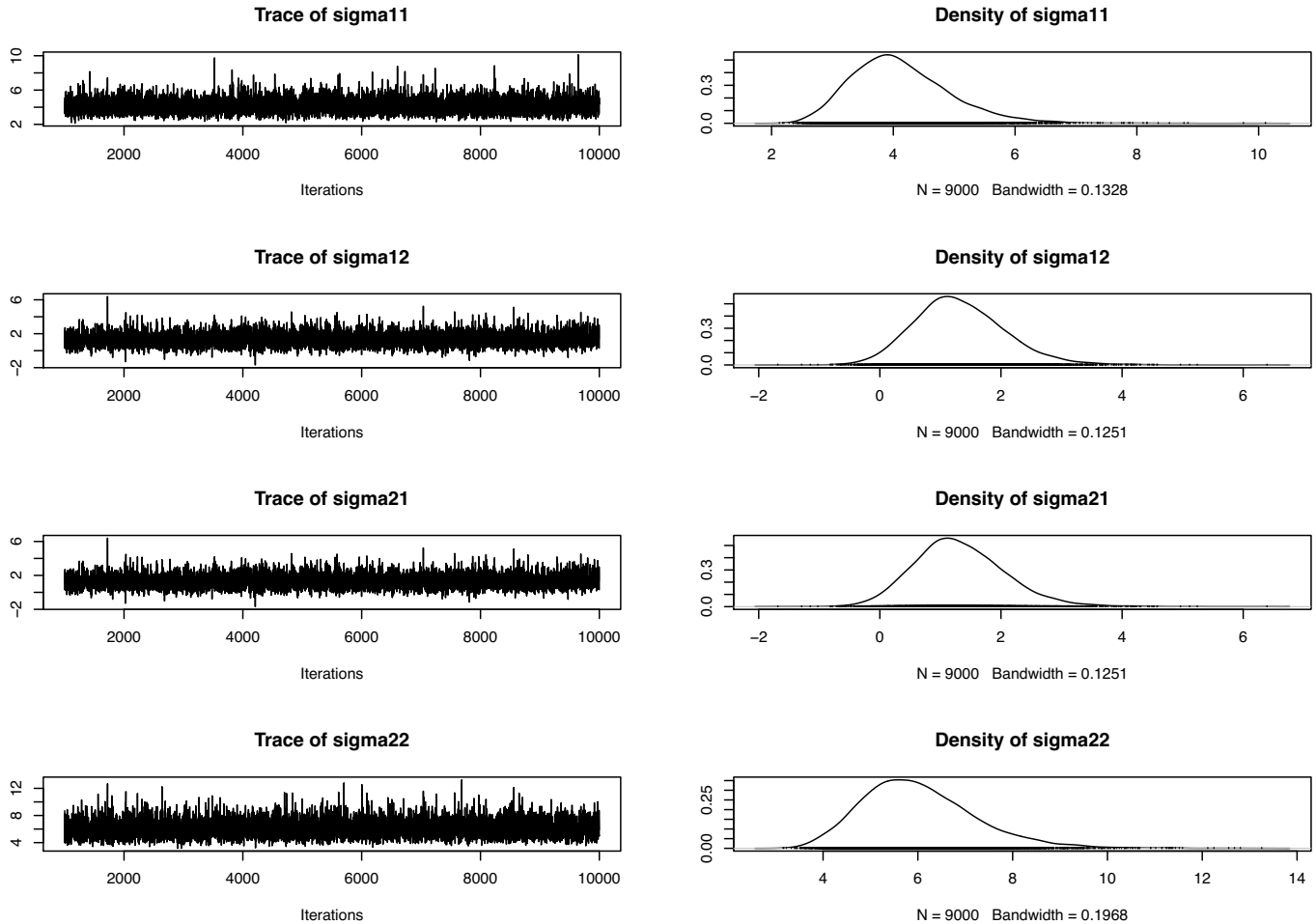
MCMC diagnostics

```
SIGMA.mcmc <- mcmc(SIGMA[-(1:nburnin),],start=nburnin+1)
summary(SIGMA.mcmc)
```

```
##
## Iterations = 1001:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 9000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##          Mean    SD Naive SE Time-series SE
## sigma11 4.12 0.808  0.00852        0.00992
## sigma12 1.32 0.754  0.00795        0.01009
## sigma21 1.32 0.754  0.00795        0.01009
## sigma22 6.02 1.196  0.01261        0.01484
##
## 2. Quantiles for each variable:
##
##             2.5%   25%  50%  75% 97.5%
## sigma11  2.81740 3.551 4.02 4.59  5.96
## sigma12 -0.00367 0.805 1.26 1.78  2.94
## sigma21 -0.00367 0.805 1.26 1.78  2.94
## sigma22  4.08475 5.179 5.88 6.72  8.71
```

Trace plots for Σ

```
plot(SIGMA.mcmc)
```



ESS for Σ

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

```
## sigma11 sigma12 sigma21 sigma22
##    6631    5586    5586    6498
```

Autocorrelation plot for Σ

```
autocorr.plot(SIGMA.mcmc,lag.max=100)
```