베이지안통계학 과제9 solution

Jieun Shin

Spring 2022

문제 10-1. (25점)

- 사후분포 $\pi((\theta_1,\theta_2)|(\bar{X},\bar{Y})) \sim N((\bar{X},\bar{Y})^T,\Sigma)$ 와 주변 사후분포 $\theta_1|\bar{X} \sim N(0.3,\frac{1}{30})$, $\theta_2|\bar{Y} \sim N(0.58,\frac{1}{30})$ 를 올바르게 작성 (5점, 등고선이 없는 경우, 분포가 틀린 경우 각 -1씩 감점)
- 완전 조건부 사후분포 $(\theta_1|\theta_2,\bar{X},\bar{Y}) \sim N(\bar{X}+\rho\frac{\sigma_1}{\sigma_2}(\theta_2-\bar{Y}),\sigma_1^2(1-\rho^2))$, $(\theta_2|\theta_1,\bar{X},\bar{Y}) \sim N(\bar{Y}+\rho\frac{\sigma_2}{\sigma_1}(\theta_1-\bar{X}),\sigma_2^2(1-\rho^2))$ 를 올바르게 작성 (5점, 수리적 유도과정이 전혀 없으면 -3점 감점)
- 깁스 표본 알고리즘으로 각 추정치가 참값과 비슷하게 나오면 5점
- 깁스 표본 알고리즘으로 생성한 표본의 산점도와 근사적 사후밀도함수를 그리면 5점
- 적절히 비교한 경우 (그래프의 비교 등) 5점.

(1) 이변량 정규분포의 공분산 행렬을 $\Sigma = \frac{1}{30}\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & \sigma_2^2\end{pmatrix}$ 라 하자. 이 문제에서는 공분산 행렬의 각 원소가 $\Sigma = \frac{1}{30}\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix}$ 에 해당한다. 변수 $(\bar{X},\bar{Y})$ 에 대한 가능도 함수는 $N((\theta_1,\theta_2)^T,\Sigma)$ 이다. 그리고 (θ_1,θ_2) 의 사전밀도는 $\pi(\theta_1,\theta_2)=1$ 이다.

그러면 (θ_1,θ_2) 의 사후분포는

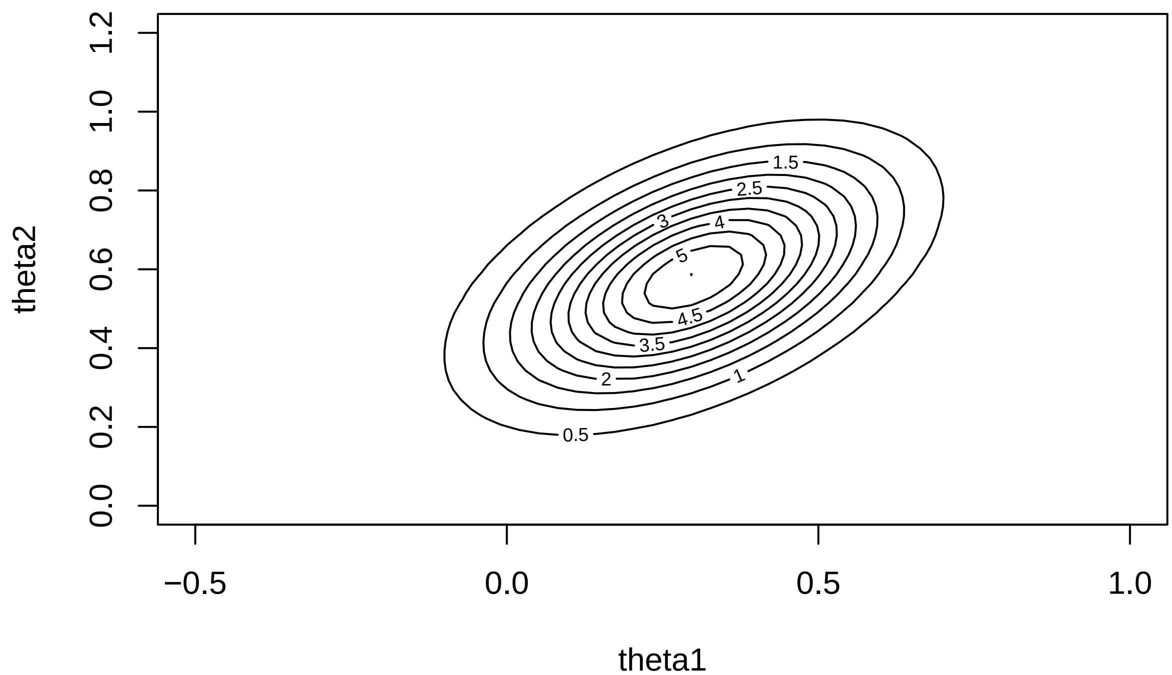
$$\begin{aligned}\pi((\theta_1,\theta_2)|(\bar{X},\bar{Y})) &\propto f((\bar{X},\bar{Y})|(\theta_1,\theta_2))\pi(\theta_1,\theta_2)\\ &= \exp\left\{-\frac{1}{2}\begin{pmatrix}\bar{X}-\theta_1\\ \bar{Y}-\theta_2\end{pmatrix}^T \Sigma^{-1}\begin{pmatrix}\bar{X}-\theta_1\\ \bar{Y}-\theta_2\end{pmatrix}\right\}\end{aligned}$$

에 비례하므로 (θ_1,θ_2) 에 대한 식으로 보면 $\pi((\theta_1,\theta_2)|(\bar{X},\bar{Y})) \sim N((\bar{X},\bar{Y})^T,\Sigma)$ 의 분포를 따르는 것을 알 수 있다. 등고선도는 다음과 같이 그려진다.

```r
library(mvtnorm)
n = 30; xbar = 0.3; ybar = 0.58
sigma = matrix(c(1, 0.5, 0.5, 1), 2, 2) /n

theta1 = seq(-0.5, 1, length.out = 50) # theta1의 grid
theta2 = seq(0, 1.2, length.out = 50) # theta2의 grid

post = outer(theta1, theta2, function(x, y) dmvnorm(cbind(x, y), mean = c(0.3, 0.58), sigma) )
contour(theta1, theta2, post,
        xlab = "theta1", ylab = "theta2")
```

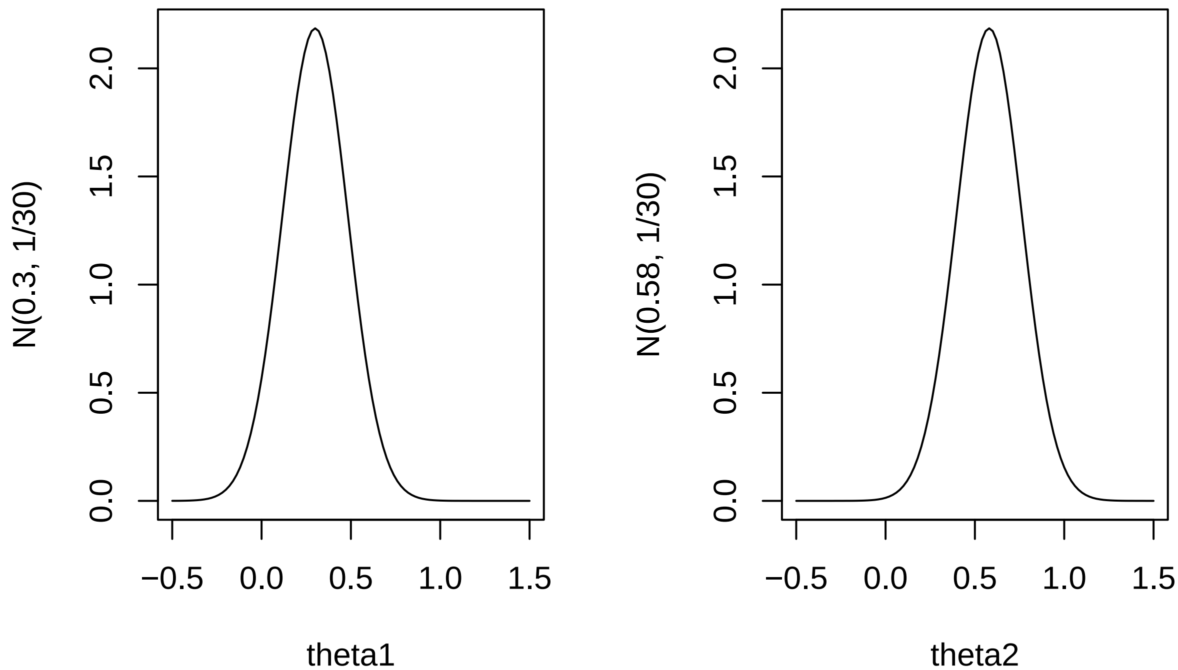


주변 사후밀도함수는 이변량 정규분포 식으로부터 다음과 같이 유도된다.

$$
\begin{aligned}
\pi(\theta_1|\bar{X}) &= \int_{-\infty}^{\infty} \pi((\theta_1,\theta_2)|(\bar{X},\bar{Y}))d\theta_2 \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(\theta_1-\bar{X})^2}{\sigma_1^2} - \frac{2\rho(\theta_1-\bar{X})(\theta_2-\bar{Y})}{\sigma_1\sigma_2} + \frac{(\theta_2-\bar{Y})^2}{\sigma_2^2}\right]\right\} d\theta_2 \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\theta_2-\bar{Y}}{\sigma_2} - \rho\frac{\bar{X}-\theta_1}{\sigma_1}\right)^2 + (1-\rho^2)\left(\frac{\bar{X}-\theta_1}{\sigma_1}\right)^2\right]\right\} d\theta_2 \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\left(\frac{\bar{X}-\theta_1}{2\sigma_1}\right)^2\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}} \exp\left\{-\frac{1}{2\sigma_2^2(1-\rho^2)}\left[\bar{Y} - \left(\theta_2 + \rho\sigma_2\frac{\bar{X}-\theta_1}{\sigma_1}\right)\right]^2\right\} d\theta_2 \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\left(\frac{\bar{X}-\theta_1}{2\sigma_1}\right)^2\right\}
\end{aligned}
$$

즉, $\theta_1|\bar{X} \sim N(0.3, \frac{1}{30})$. 마찬가지로 $\theta_2|\bar{Y} \sim N(0.58, \frac{1}{30})$를 따른다. 그림으로는 다음과 같이 그려진다.

```
par(mfrow = c(1, 2))
curve( dnorm(x, xbar, sqrt(sigma[1,1])), from = -0.5, to = 1.5,
       xlab = "theta1", ylab = "N(0.3, 1/30)")
curve( dnorm(x, ybar, sqrt(sigma[2,2])), from = -0.5, to = 1.5,
       xlab = "theta2", ylab = "N(0.58, 1/30)")
```




```
par(mfrow = c(1, 1))
```

(2) θ_2 가 주어졌을 때 θ_1 의 완전 조건부 사후분포 $\pi(\theta_1|\theta_2, \bar{X}, \bar{Y})$ 를 구해보자. 상관계수를 $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$ 라 하면, 다음과 같이 계산된다.

$$
\begin{aligned}
\pi(\theta_1|\theta_2, \bar{X}, \bar{Y}) &= \frac{\pi((\theta_1, \theta_2)|(\bar{X}, \bar{Y}))}{\pi(\theta_1|\bar{X})} \\
&= \frac{\frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(\theta_1-\bar{X})^2}{\sigma_1^2} - \frac{2\rho(\theta_1-\bar{X})(\theta_2-\bar{Y})}{\sigma_1\sigma_2} + \frac{(\theta_2-\bar{Y})^2}{\sigma_2^2}\right]\right\}}{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\frac{1}{2}\left(\frac{\theta_2-\bar{Y}}{\sigma_2}\right)^2\right\}} \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(\theta_1-\bar{X})^2}{\sigma_1^2} - \frac{2\rho(\theta_1-\bar{X})(\theta_2-\bar{Y})}{\sigma_1\sigma_2} + \frac{(\theta_2-\bar{Y})^2}{\sigma_2^2} - \frac{(\theta_2-\bar{Y})^2}{\sigma_2^2}(1-\rho^2)\right]\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_1^2}\left[(\theta_1-\bar{X})^2 - \frac{2\sigma_1\rho(\theta_1-\bar{X})(\theta_2-\bar{Y})}{\sigma_2} + \frac{\sigma_1^2\rho^2(\theta_2-\bar{Y})^2}{\sigma_2^2}\right]\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_1^2}\left[\theta_1 - \left(\bar{X} + \rho\frac{\sigma_1}{\sigma_2}(\theta_2-\bar{Y})\right)\right]^2\right\}.
\end{aligned}
$$

따라서 $(\theta_1|\theta_2, \bar{X}, \bar{Y}) \sim N(\bar{X} + \rho\frac{\sigma_1}{\sigma_2}(\theta_2 - \bar{Y}), \sigma_1^2(1-\rho^2))$ 이다. 마찬가지로 θ_2 가 주어졌을 때 θ_1 의 완전 조건부 사후분포도 같은 방법으로 계산하면 $(\theta_2|\theta_1, \bar{X}, \bar{Y}) \sim N(\bar{Y} + \rho\frac{\sigma_2}{\sigma_1}(\theta_1 - \bar{X}), \sigma_2^2(1-\rho^2))$ 을 따르는 것을 알 수 있다.

(3) θ_1과 θ_2의 사후표본을 생성하기 위한 깁스 표본기법 알고리즘은 다음과 같다.

1. 초기치 $\theta_1^{(0)}, \theta_2^{(0)}$을 정한다.

2. 다음의 과정을 $k = 1, \ldots, K$번 반복한다.

    1) $\theta_1^{(k)} \sim N(\bar{X} + \rho\frac{\sigma_1}{\sigma_2}(\theta_2^{(k-1)} - \bar{Y}), \sigma_1^2(1-\rho^2))$

    2) $\theta_2^{(k)} \sim N(\bar{Y} + \rho\frac{\sigma_2}{\sigma_1}(\theta_1^{(k)} - \bar{X}), \sigma_2^2(1-\rho^2))$

```
theta1_sam = runif(1)  # 초기치 부여
theta2_sam = runif(1)

rho = sigma[1,2] / sqrt( (sigma[1,1] * sigma[2,2]) )
K = 5000  # 5000개의 샘플 추출

for(k in 1:K){
  theta1_sam[k+1] = rnorm(1, xbar + rho * sigma[1,1]/sigma[2,2] * (theta2_sam[k] - ybar),
                          sqrt(sigma[1,1] * (1 - rho^2)))
  theta2_sam[k+1] = rnorm(1, ybar + rho * sigma[2,2]/sigma[1,1] * (theta1_sam[k+1] - xbar),
                          sqrt(sigma[2,2] * (1 - rho^2)))
}

theta1_sam = theta1_sam[-c(1:500)]  # burn-in 기간 제외
theta2_sam = theta2_sam[-c(1:500)]

mean(theta1_sam); mean(theta2_sam)  # 사후기대치
```

```
## [1] 0.292191
```

```
## [1] 0.5699201
```

```
var(theta1_sam); var(theta2_sam)    # 사후분산
```

```
## [1] 0.0338381
```

```
## [1] 0.03409603
```

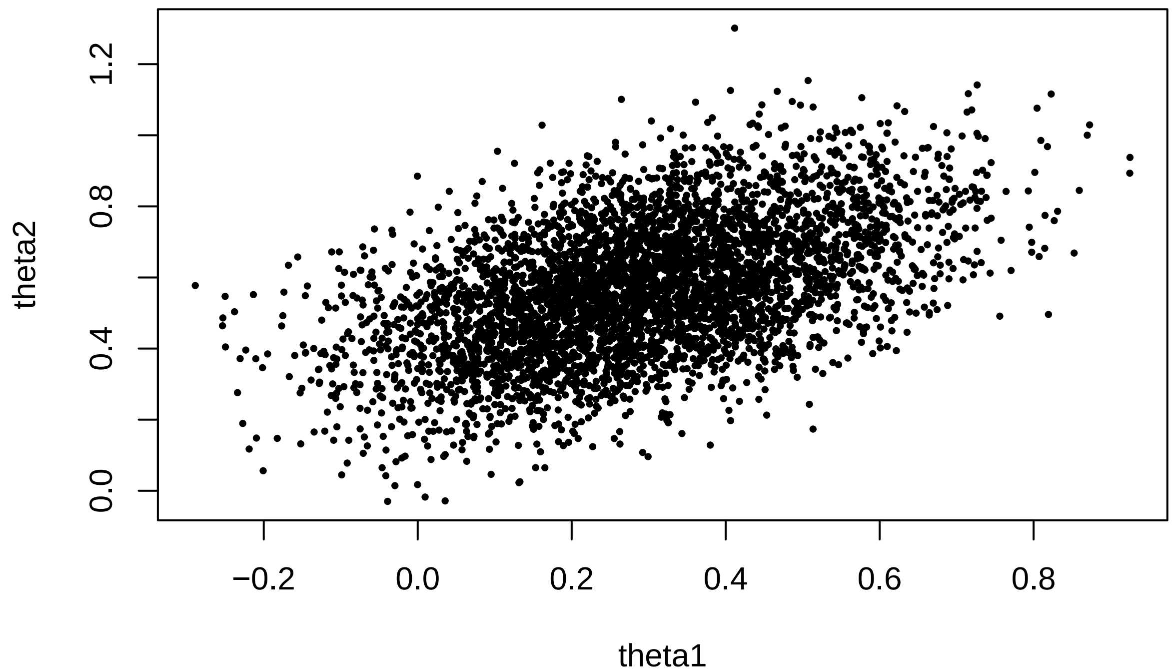
```
cor(theta1_sam, theta2_sam)         # 상관계수
```

```
## [1] 0.5047504
```

깁스 표본기법으로부터 생성한 표본의 기댓값은 약 0.3, 0.58, 그리고 분산은 약 0.33, 두 표본의 상관계수는 약 0.5로 참값과 유사하다.

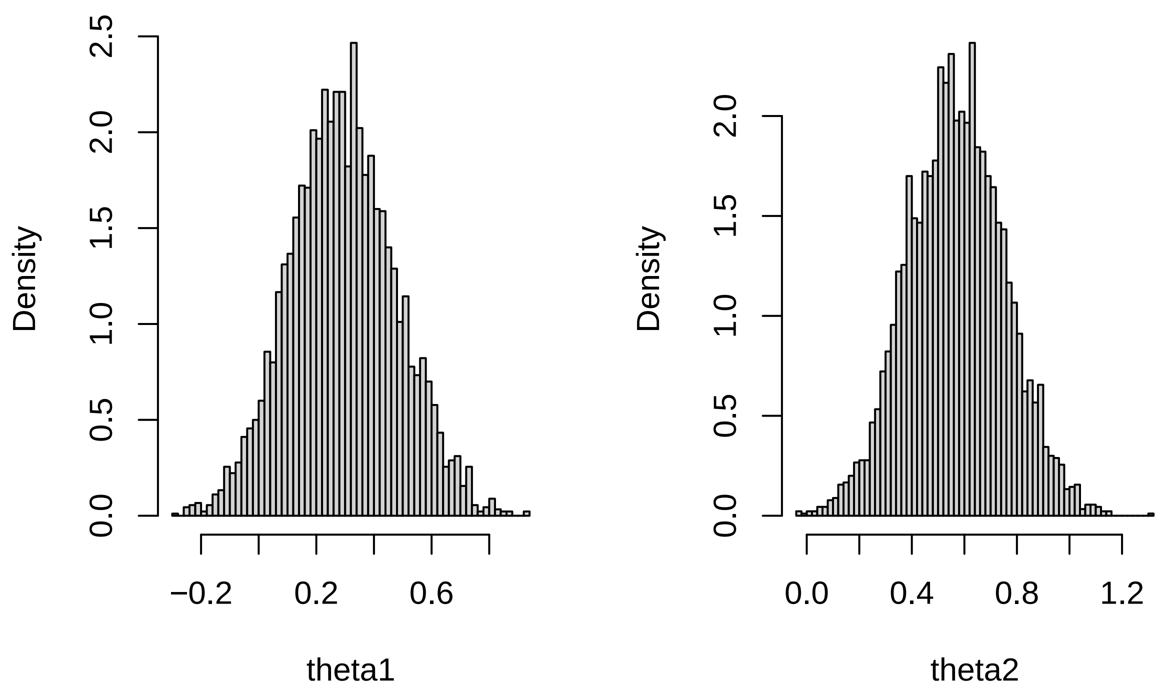
(4) 깁스 표본기법으로부터 구한 (θ_1, θ_2)의 산점도는 다음과 같이 그려진다.

```r
plot(theta1_sam, theta2_sam, pch = 16, cex = 0.5, xlab = "theta1", ylab = "theta2")
```



그리고 근사적 주변 사후밀도함수는 다음과 같이 그려진다.

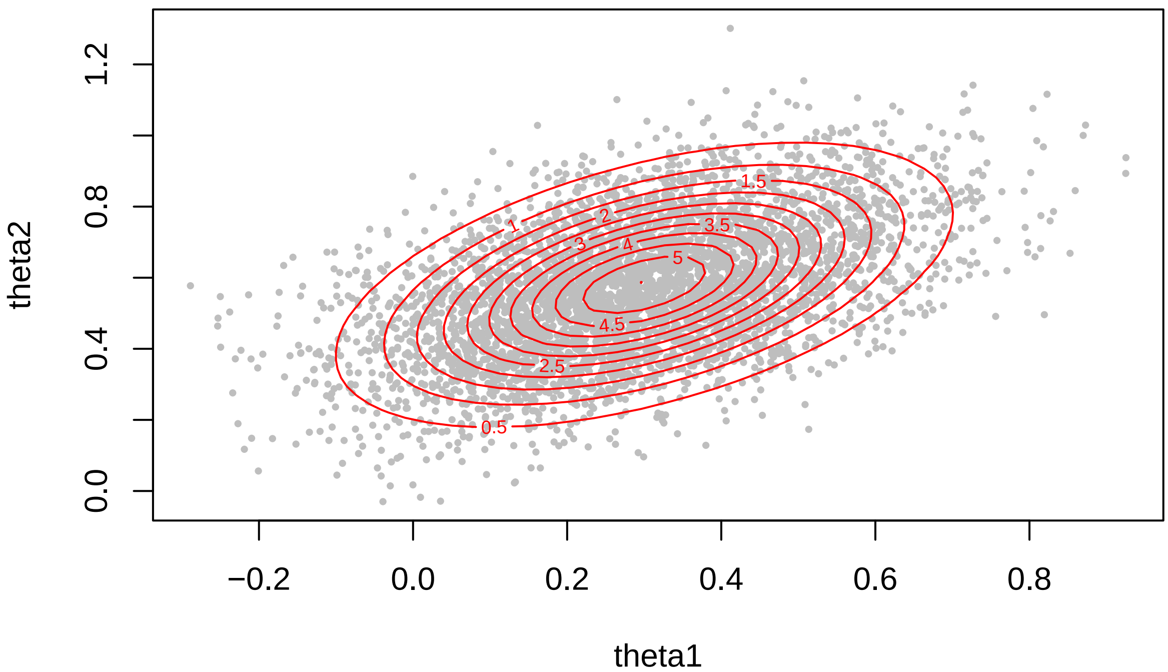
```r
par(mfrow = c(1, 2))
hist(theta1_sam, breaks = 50, freq = F, xlab = "theta1", main = "")
hist(theta2_sam, breaks = 50, freq = F, xlab = "theta2", main = "")
```




```
par(mfrow = c(1, 1))
```

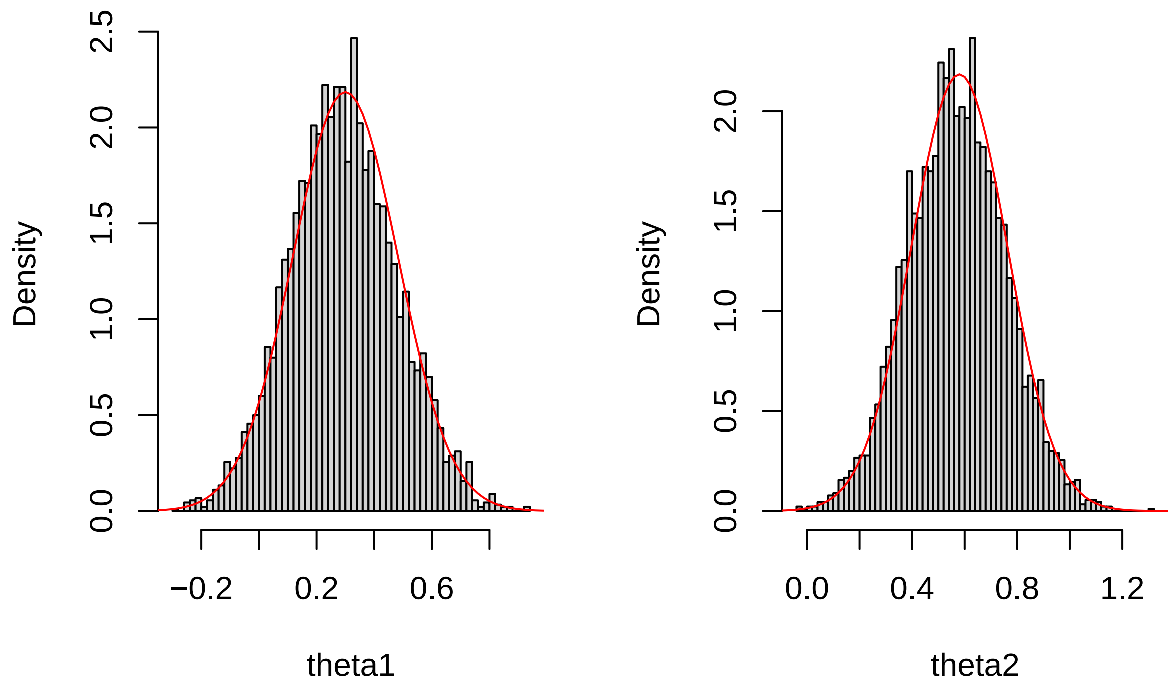
(5) (1)번과 (4)번의 그래프를 겹쳐 그리면 다음과 같다.

```
plot(theta1_sam, theta2_sam, col = "gray", pch = 16, cex = 0.5, xlab = "theta1", ylab = "theta2")
contour(theta1, theta2, post, col = "red",
        xlab = "theta1", ylab = "theta2", add = T)
```




```
par(mfrow = c(1, 2))
hist(theta1_sam, breaks = 50, freq = F, xlab = "theta1", main = "")
curve( dnorm(x, xbar, sqrt(sigma[1,1])), from = -0.5, to = 1.5, main = "theta1",
       col = "red", add = T)

hist(theta2_sam, breaks = 50, freq = F, xlab = "theta2", main = "")
curve( dnorm(x, ybar, sqrt(sigma[2,2])), from = -0.5, to = 1.5, main = "theta2",
       col = "red", add = T)
```




```r
par(mfrow = c(1, 1))
```

따라서 수리적으로 유도한 사후분포와 깁스 표본기법으로 생성한 사후분포가 유사함을 확인할 수 있다.