Gibbs sampling

Dr. Olanrewaju Michael Akande

Feb 5, 2020

Announcements

- Homework 3 due tomorrow.

Outline

- Non-conjugate priors
- Full conditionals
- Gibbs sampling
- A simple example: bivariate normal
- In-class exercise

Bayesian inference (conjugacy recap)

- As we've seen so far, Bayesian inference is based on posterior distributions, that is,

$$p(\theta|y) = \frac{p(\theta)L(y;\theta)}{\int_\Theta p(\tilde{\theta})L(y;\tilde{\theta})\mathrm{d}\tilde{\theta}} = \frac{p(\theta)L(y;\theta)}{L(y)}$$

- Good news: we have the numerator in this expression.
- Bad news: the denominator is typically not available (may involve high dimensional integral)!
- How have we been getting by? Conjugacy! For conjugate priors, the posterior distribution of θ is available analytically.
- What if a conjugate prior does not represent our prior information well, or we have a more complex model, and our posterior is no longer in a convenient distributional form?

Some conjugate models

- We've already seen the following conjugate models.

Prior	Likelihood	Posterior
beta	binomial	beta
gamma	Poisson	gamma
gamma	exponential	gamma
normal-gamma	normal	normal-gamma

- Here are a few more we have not covered yet.

Prior	Likelihood	Posterior
beta	negative-binomial	beta
beta	geometric	beta
Dirichlet	multinomial	Dirichlet

- Clearly, we cannot restrict ourselves to conjugate models only.

Back to the normal model

- For conjugacy in the normal model, we had

$$\mu|\tau \sim \mathcal{N}\left(\mu_0, \frac{1}{\kappa_0 \tau}\right).$$
$$\tau \sim \text{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

- Suppose we wish to specify our uncertainty about μ as independent of τ, that is, we want $\pi(\mu, \tau) = \pi(\mu)\pi(\tau)$. For example,

$$\mu \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right).$$
$$\tau \sim \text{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2\tau_0}\right).$$

- When σ_0^2 is not proportional to $\frac{1}{\tau}$, the marginal density of τ is not a gamma density (or a density we can easily sample from).

- Side note: for conjugacy, the joint posterior should also be a product of two independent Normal and Gamma densities in μ and τ respectively.

Non-conjugate priors

- In general, conjugate priors are not available for generalized linear models (GLMs) other than the normal linear model.
- One can potentially rely on an asymptotic normal approximation.
- As $n \to \infty$, the posterior distribution is normal centered on MLE.
- However, even for moderate sample sizes, asymptotic approximations may be inaccurate.
- In logistic regression for example, for rare outcomes or rare binary exposures, posterior can be highly skewed.
- Appealing to avoid any reliance on large sample assumptions and base inferences on **exact posterior**.

Non-conjugate priors

- Even though we may not be able to sample from the marginal posterior of a particular parameter when using a non-conjugate prior, sometimes, we may still be able to sample from conditional distributions of those parameters given all other parameters and the data.
- These conditional distributions, known as full conditionals, will be very important for us.
- In our normal example with

$$\begin{aligned} \mu &\sim \mathcal{N}\left(\mu_0, \sigma_0^2\right). \\ \tau &\sim \text{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2\tau_0}\right), \end{aligned}$$

  even though we cannot sample easily from $\tau|Y$, turns out we will be able to sample from $\tau|\mu, Y$. That is the full conditional for τ.
- By the way, note that we already know the full conditional for μ, i.e., $\mu|\tau, Y$ (last two classes).

Full conditional distributions

- Goal: try to take advantage of those full conditional distributions (without sampling directly from the marginal posteriors) to obtain samples from the said marginal posteriors.
- In our example, with $\pi(\mu) = \mathcal{N}\left(\mu_0, \sigma_0^2\right)$, we have

$$\mu|Y, \tau \sim \mathcal{N}(\mu_n, \tau_n^{-1}),$$

where

  - $\mu_n = \dfrac{\frac{\mu_0}{\sigma_0^2} + n\tau\bar{y}}{\frac{1}{\sigma_0^2} + n\tau}$; and
  - $\tau_n = \frac{1}{\sigma_0^2} + n\tau$.

- Review results from previous two classes if you are not sure why this holds.
- Let's see if we can figure out the other full conditional $\tau|\mu, Y$.

Full conditional distributions

$$\pi(\tau|\mu, Y) = \frac{\Pr[\tau, \mu, Y]}{\Pr[\mu, Y]} = \frac{L(y; \mu, \tau)\pi(\mu, \tau)}{\Pr[\mu, Y]}$$

$$= \frac{L(y; \mu, \tau)\pi(\mu)\pi(\tau)}{\Pr[\mu, Y]}$$

$$\propto L(y; \mu, \tau)\pi(\tau)$$

$$\propto \underbrace{\tau^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\tau\sum_{i=1}^{n}(y_i - \mu)^2\right\}}_{\propto L(Y;\mu,\tau)} \times \underbrace{\tau^{\frac{\nu_0}{2}-1}\exp\left\{-\frac{\tau\nu_0}{2\tau_0}\right\}}_{\propto \pi(\tau)}$$

$$= \underbrace{\tau^{\frac{\nu_0+n}{2}-1} \exp\left\{-\frac{1}{2}\tau\left[\frac{\nu_0}{\tau_0} + \sum_{i=1}^{n}(y_i - \mu)^2\right]\right\}}_{\text{Gamma Kernel}}.$$

Full conditional distributions

$$\pi(\tau|\mu, Y) \propto \underbrace{\tau^{\frac{\nu_0+n}{2}-1} \exp\left\{-\frac{1}{2}\tau\left[\frac{\nu_0}{\tau_0} + \sum_{i=1}^{n}(y_i-\mu)^2\right]\right\}}_{\text{Gamma Kernel}}$$

$$= \text{Gamma}\left(\frac{\nu_n}{2}, \frac{\nu_n}{2\tau_n(\mu)}\right) \quad \text{OR} \quad \text{Gamma}\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2(\mu)}{2}\right),$$

where

$$\nu_n = \nu_0 + n$$

$$\sigma_n^2(\mu) = \frac{1}{\nu_n}\left[\frac{\nu_0}{\tau_0} + \sum_{i=1}^{n}(y_i-\mu)^2\right] = \frac{1}{\nu_n}\left[\frac{\nu_0}{\tau_0} + ns_n^2(\mu)\right]$$

$$\text{OR} \quad \tau_n(\mu) = \frac{\nu_n}{\left[\frac{\nu_0}{\tau_0} + \sum_{i=1}^{n}(y_i-\mu)^2\right]} = \frac{\nu_n}{\left[\frac{\nu_0}{\tau_0} + ns_n^2(\mu)\right]};$$

$$\text{with} \quad s_n^2(\mu) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Iterative scheme

- Now we have two full conditional distributions but what we really need is to sample from $\pi(\tau|Y)$.
- Actually, if we could sample from $\pi(\mu, \tau|Y)$, we already know that the draws for μ and τ will be from the two marginal posterior distributions. So, we just need a scheme to sample from $\pi(\mu, \tau|Y)$.
- Suppose we had a single sample, say $\tau^{(1)}$ from the marginal posterior distribution $\pi(\tau|Y)$. Then we could sample

$$\mu^{(1)} \sim p(\mu|\tau^{(1)}, Y).$$

- This is what we did in the last class, so that the pair $\{\mu^{(1)}, \tau^{(1)}\}$ is a sample from the joint posterior $\pi(\mu, \tau|Y)$.
- $\Rightarrow$ $\mu^{(1)}$ can be considered a sample from the marginal distribution of μ, which again means we can use it to sample

$$\tau^{(2)} \sim p(\tau|\mu^{(1)}, Y),$$

and so forth.

Gibbs sampling

- So, we can use two **full conditional distributions** to generate samples from the **joint distribution**, once we have a starting value $\tau^{(1)}$.
- Formally, this sampling scheme is known as Gibbs sampling.
  - Purpose: Draw from a joint distribution, say $p(\mu, \tau|Y)$.
  - Method: Iterative conditional sampling
    - Draw $\tau^{(1)} \sim p(\tau|\mu^{(0)}, Y)$
    - Draw $\mu^{(1)} \sim p(\mu|\tau^{(1)}, Y)$
  - Purpose: Full conditional distributions have known forms, with sampling from the full conditional distributions fairly easy.
- More generally, we can use this method to generate samples of $\theta = (\theta_1, \ldots, \theta_p)$, the vector of p parameters of interest, from the joint density.

Gibbs sampling

- Procedure:
  - Start with initial value $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$.
  - For iterations $t = 1, \ldots, T$,
    1. Sample $\theta_1^{(t)}$ from the conditional posterior distribution

$$\pi(\theta_1 | \theta_2 = \theta_2^{(t-1)}, \ldots, \theta_p = \theta_p^{(t-1)}, Y)$$

    2. Sample $\theta_2^{(t)}$ from the conditional posterior distribution

$$\pi(\theta_2 | \theta_1 = \theta_1^{(t)}, \theta_3 = \theta_3^{(t-1)}, \ldots, \theta_p = \theta_p^{(t-1)}, Y)$$

    3. Similarly, sample $\theta_3^{(t)}, \ldots, \theta_p^{(t)}$ from the conditional posterior distributions given current values of other parameters.
  - This generates a **dependent** sequence of parameter values.

MCMC

- Gibbs sampling is one of several flavors of Markov chain Monte Carlo (MCMC).
  - Markov chain: a stochastic process in which future states are independent of past states conditional on the present state.
  - Monte Carlo: simulation.
- MCMC provides an approach for generating samples from posterior distributions.
- From these samples, we can obtain summaries (including summaries of functions) of the posterior distribution for θ, our parameter of interest.

How does MCMC work?

- Let $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_p^{(t)})$ denote the value of the $p \times 1$ vector of parameters at iteration t.
- Let $\theta^{(0)}$ be an initial value used to start the chain (*should not be sensitive*).
- MCMC generates $\theta^{(t)}$ from a distribution that depends on the data and potentially on $\theta^{(t-1)}$, but not on $\theta^{(1)}, \ldots, \theta^{(t-2)}$.
- This results in a Markov chain with **stationary distribution** $\pi(\theta|Y)$ under some conditions on the sampling distribution.
- The theory of Markov Chains (structure, convergence, reversibility, detailed balance, stationarity, etc) is well beyond the scope of this course so we will not dive into it.
- If you are interested, consider taking STA 531/831 or courses on stochastic process.

Properties

- **Note**: Our Markov chain is a collection of draws of θ that are (slightly we hope!) dependent on the previous draw.
- The chain will wander around our parameter space, only remembering where it had been in the last draw.
- We want to have our MCMC sample size, T, big enough so that we can
  - Move out of areas of low probability into regions of high probability (convergence)
  - Move between high probability regions (good mixing)
  - Know our Markov chain is stationary in time (the distribution of samples is the same for all samples, regardless of location in the chain)
- At the start of the sampling, the samples are **not** from the posterior distribution. It is necessary to discard the initial samples as a burn-in to allow convergence. We'll talk more about that in the next class.

Different flavors of MCMC

- The most commonly used MCMC algorithms are:
  - Metropolis sampling (Metropolis et al., 1953).
  - Metropolis-Hastings (MH) (Hastings, 1970).
  - Gibbs sampling (Geman & Geman, 1984; Gelfand & Smith, 1990).
- Overview of Gibbs - Casella & George (1992, The American Statistician, 46, 167-174). the first two
- Overview of MH - Chib & Greenberg (1995, The American Statistician).
- We will get to Metropolis and Metropolis-Hastings later in the course.

Example: bivariate normal

- Consider

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

  where ρ is known (and is the correlation between θ_1 and θ_2).

- We will review details of the multivariate normal distribution very soon but for now, let's use this example to explore Gibbs sampling.

- For this density, turns out that we have

$$\theta_1 | \theta_2 \sim \mathcal{N}\left(\rho\theta_2, 1 - \rho^2\right)$$

  and

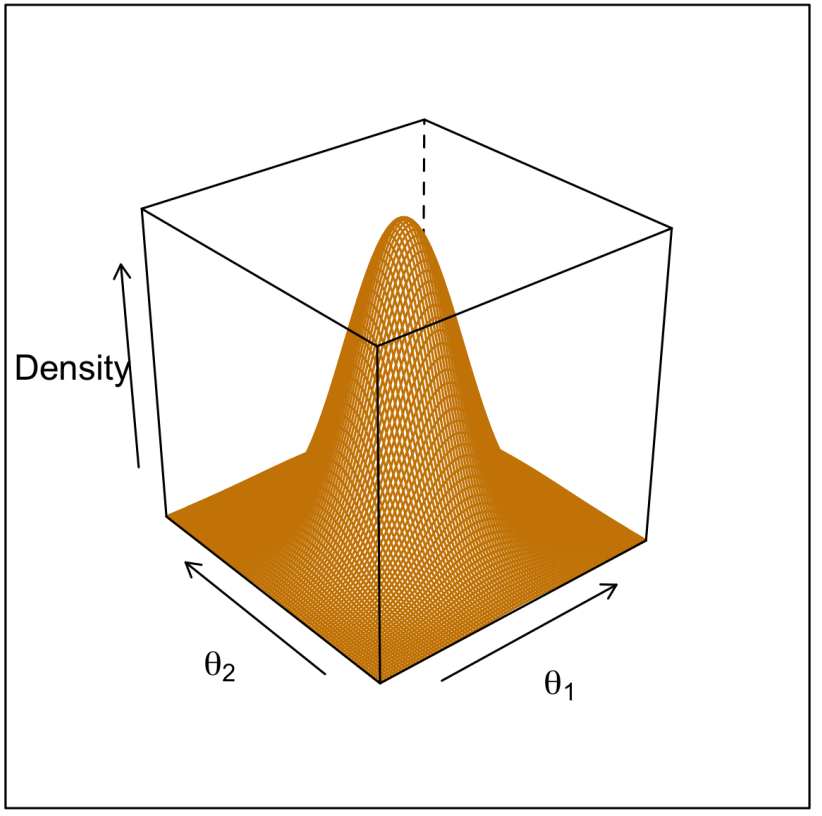
$$\theta_2 | \theta_1 \sim \mathcal{N}\left(\rho\theta_1, 1 - \rho^2\right)$$

- While we can easily sample directly from this distribution (using the `mvtnorm` or `MASS` packages in R), let's instead use the Gibbs sampler to draw samples from it.

Bivariate normal

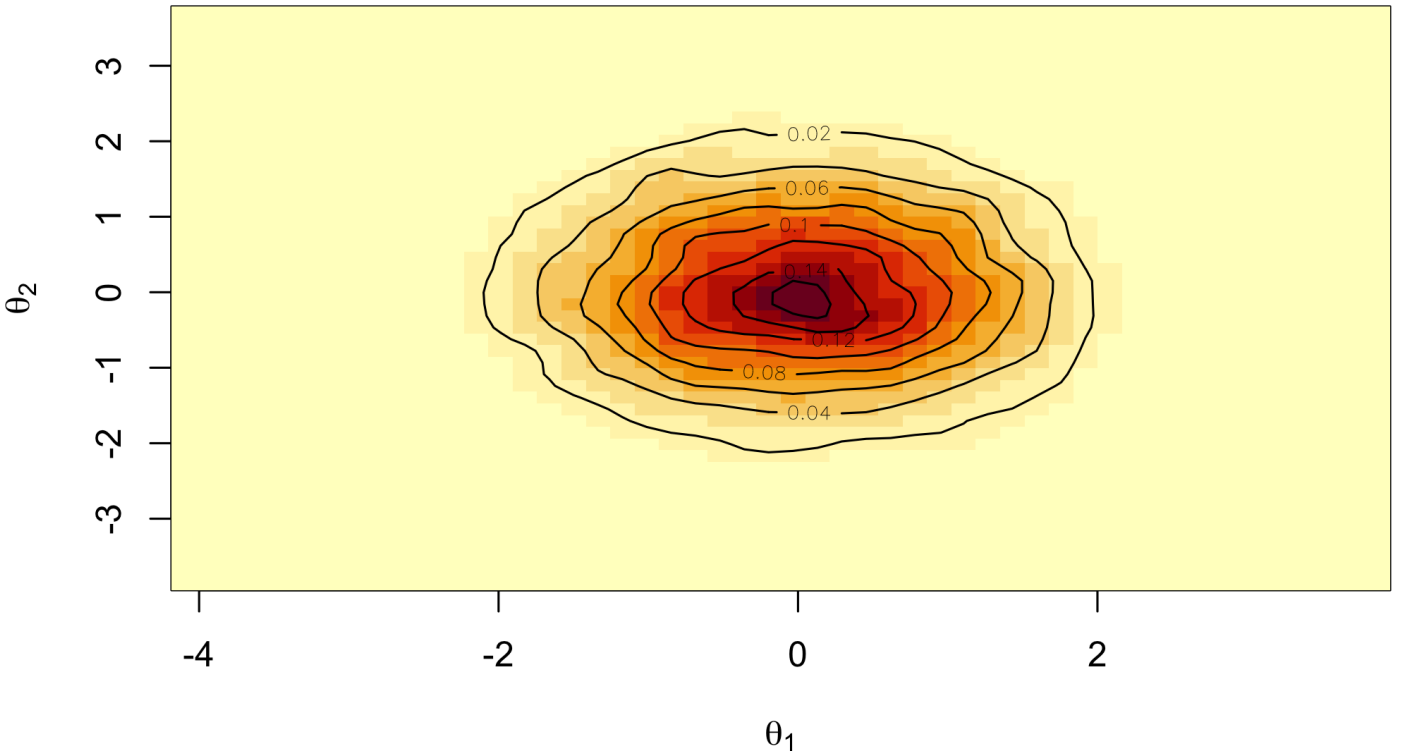
First, a few examples of the bivariate normal distribution.

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right]$$



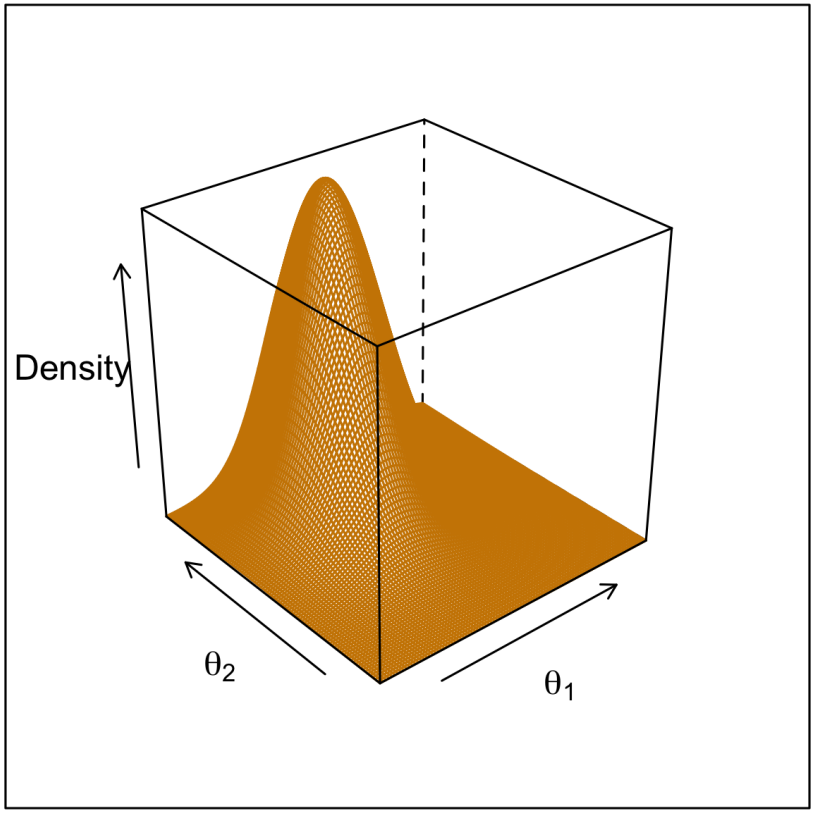
Bivariate normal

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right]$$



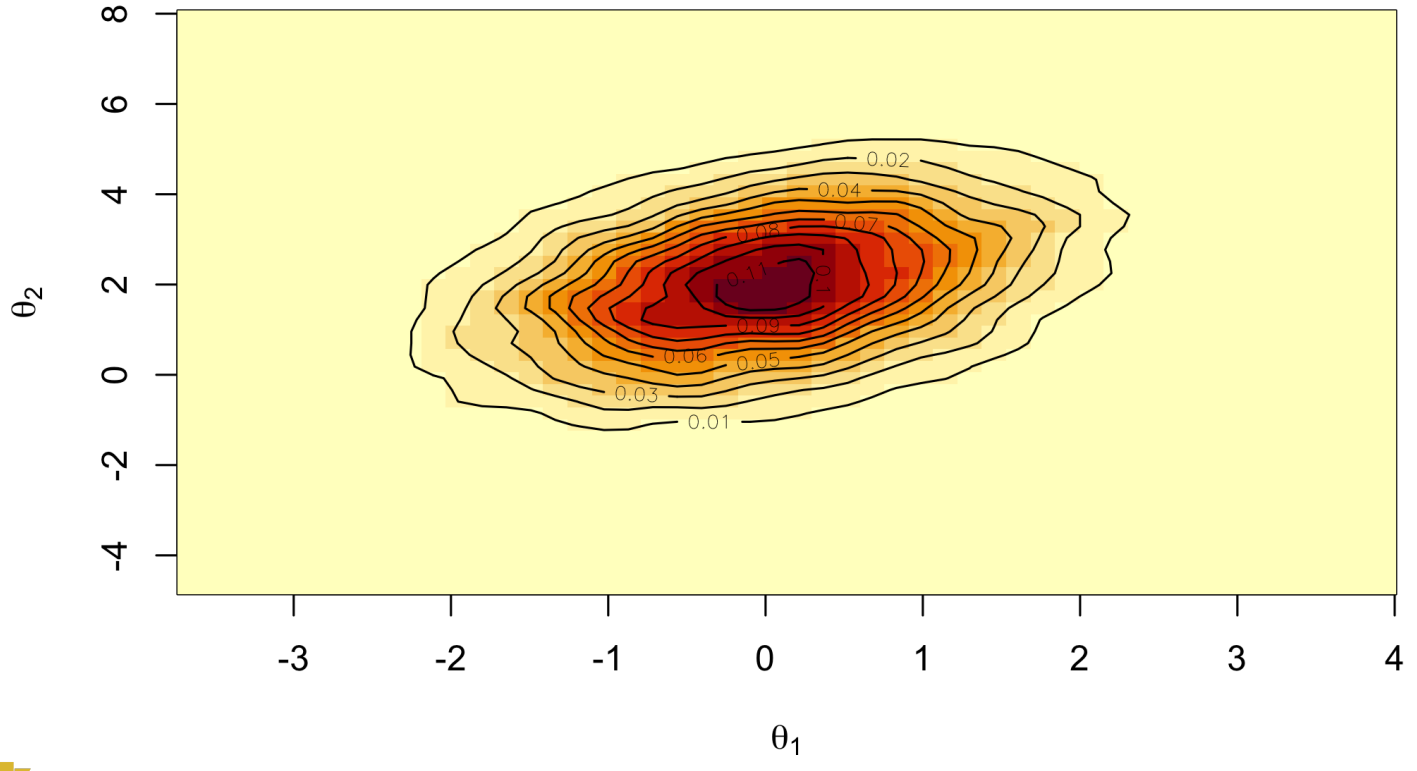
Bivariate normal

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix}0\\ 2\end{pmatrix}, \begin{pmatrix}1 & 0.5\\ 0.5 & 2\end{pmatrix}\right]$$



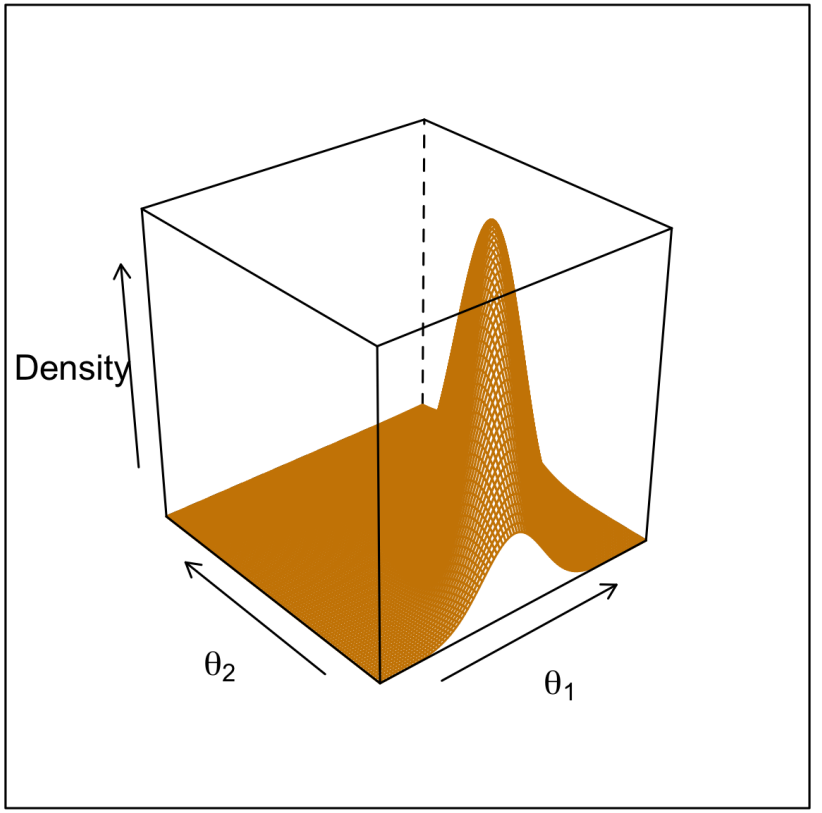
Bivariate normal

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix}0\\ 2\end{pmatrix}, \begin{pmatrix}1 & 0.5\\ 0.5 & 2\end{pmatrix}\right]$$



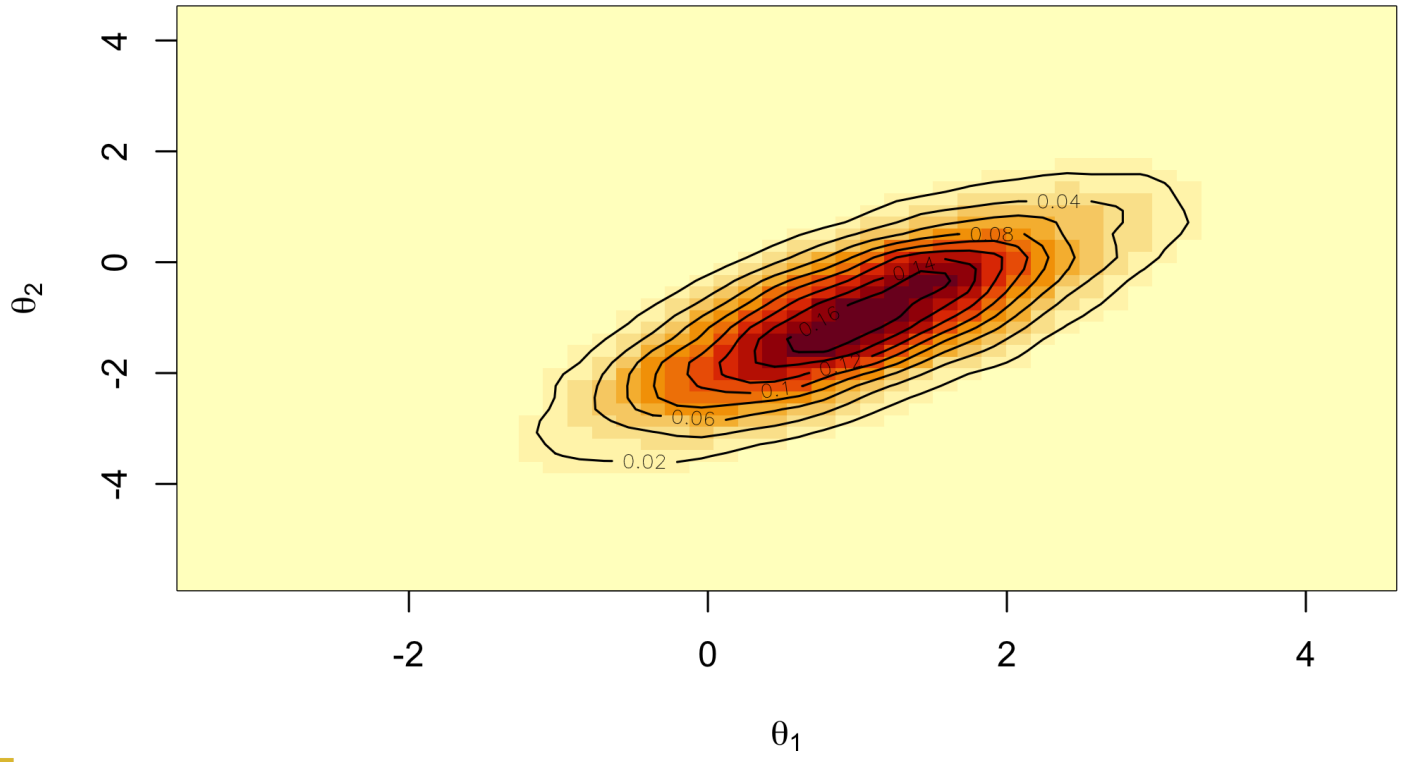
Bivariate normal

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix}1\\ -1\end{pmatrix}, \begin{pmatrix}1 & 0.9\\ 0.9 & 0.5\end{pmatrix}\right]$$



Bivariate normal

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix}1\\ -1\end{pmatrix}, \begin{pmatrix}1 & 0.9\\ 0.9 & 0.5\end{pmatrix}\right]$$



Back to the example

- Again, we have

$$\theta_1|\theta_2 \sim \mathcal{N}\left(\rho\theta_2, 1-\rho^2\right); \quad \theta_2|\theta_1 \sim \mathcal{N}\left(\rho\theta_1, 1-\rho^2\right)$$

- Here's a code to do Gibbs sampling using those full conditionals:

```
rho <- #set correlation
S <- #set number of MCMC samples
thetamat <- matrix(0,nrow=S,ncol=2)
theta <- c(10,10) #initialize values of theta
for (s in 1:S) {
theta[1] <- rnorm(1,rho*theta[2],sqrt(1-rho^2)) #sample theta1
theta[2] <- rnorm(1,rho*theta[1],sqrt(1-rho^2)) #sample theta2
thetamat[s,] <- theta
}
```

- Here's a code to do sample directly instead:

```
library(mvtnorm)
rho <- #set correlation; no need to set again once you've used previous code
S <- #set number of MCMC samples; no need to set again once you've used previous code
Mu <- c(0,0)
Sigma <- matrix(c(1,rho,rho,1),ncol=2)
thetamat_direct <- rmvnorm(S, mean = Mu,sigma = Sigma)
```

Participation exercise

- You will work in groups of three. Work with the three students closest to you.
- For $S \in \{50, 250, 500\}$ and $\rho \in \{0.1, 0.5, 0.95\}$, do the following:
  1. Generate S samples using the two methods.
  2. Make a scatter plot of the samples from each method (plot the samples from the Gibbs sampler first) and compare them.
- How do the results differ between the two methods for the different combinations of S and ρ?
- Discuss within your teams, document your team findings and submit.
- You can have one person document the findings but make sure to write the name of all three members at the top of the sheet.

More code

See how the chain actually evolves with an overlay on the true density:

```
rho <- #set correlation
Sigma <- matrix(c(1,rho,rho,1),ncol=2); Mu <- c(0,0)
x.points <- seq(-3,3,length.out=100)
y.points <- x.points
z <- matrix(0,nrow=100,ncol=100)
for (i in 1:100) {
  for (j in 1:100) {
    z[i,j] <- dmvnorm(c(x.points[i],y.points[j]),mean=Mu,sigma=Sigma)
  }
}
contour(x.points,y.points,z,xlim=c(-3,10),ylim=c(-3,10),"orange2",
        xlab=expression(theta[1]),ylab=expression(theta[2]))

S <- #set number of MCMC samples;
thetamat <- matrix(0,nrow=S,ncol=2)
theta <- c(10,10)
points(x=theta[1],y=theta[2],col="black",pch=2)
for (s in 1:S) {
  theta[1] <- rnorm(1,rho*theta[2],sqrt(1-rho^2))
  theta[2] <- rnorm(1,rho*theta[1],sqrt(1-rho^2))
  thetamat[s,] <- theta
  if(s < 20){
    points(x=theta[1],y=theta[2],col="red4",pch=16); Sys.sleep(1)
  } else {
    points(x=theta[1],y=theta[2],col="green4",pch=16); Sys.sleep(0.1)
  }
}
```