

STAT 238 - Bayesian Statistics

Lecture Twenty Six

Spring 2026, UC Berkeley

Aditya Guntuboyina

03 April 2026

1 The Gibbs Sampler

Let us introduce some notation. The goal is to obtain samples from a target distribution with density π (in our applications, π will be the posterior density). π represents the density on \mathbb{R}^d of a random vector Θ of size $d \times 1$:

$$\Theta \sim \pi.$$

We assume that Θ can be decomposed as:

$$\Theta = (\Theta_{(1)}, \dots, \Theta_{(k)})$$

for k subvectors $\Theta_{(1)}, \dots, \Theta_{(k)}$. The j^{th} subvector is $\Theta_{(j)}$. We shall also denote the vector obtained by dropping the j^{th} subvector $\Theta_{(j)}$ from Θ by $\Theta_{-(j)}$. For example,

$$\Theta_{-(1)} = (\Theta_{(2)}, \dots, \Theta_{(k)}) \quad \text{and} \quad \Theta_{-(2)} = (\Theta_{(1)}, \Theta_{(3)}, \dots, \Theta_{(k)}).$$

Gibbs sampling works according to the following algorithm.

1. Initialize with an arbitrary value $\theta^{(0)}$.
2. Repeat the following for $t = 1, \dots, N$ (for a large number N):
 - a) Pick J uniformly at random from $1, \dots, k$. Suppose J turned out to be j .
 - b) Set $\theta_{-(j)}^{(t+1)} = \theta_{-(j)}^{(t)}$ i.e., $\theta_{(i)}^{(t+1)} = \theta_{(i)}^{(t)}$ for $i \neq j$.
 - c) $\theta_{(j)}^{(t+1)}$ is randomly drawn according to the conditional distribution of $\Theta_{(j)}$ given $\Theta_{-(j)} = \theta_{-(j)}^{(t)}$.

This algorithm assumes that we have some way of generating observations from the conditional distribution of $\Theta_{(j)}$ given $\Theta_{-(j)}$ for each j .

2 Examples of the Gibbs Sampler

In the first two examples below, the posterior can be written in closed form so MCMC is not really necessary. These are still useful for evaluating the performance of the Gibbs sampler. The third example is more non-trivial and illustrates the usefulness of the Gibbs sampler.

2.1 Bivariate Normal

We want to sample (θ_1, θ_2) from the following bivariate distribution:

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \bigg| \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N \left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

Here $y = (y_1, y_2)^T$ is fixed. One can, of course, directly sample from this normal distribution without any MCMC. We can use this example to illustrate the performance of the Gibbs sampler. To use the Gibbs sampler, observe that

$$\theta_1 | \theta_2, y \sim N(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2) \quad \text{and} \quad \theta_2 | \theta_1, y \sim N(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

The Gibbs sampler algorithm then would be the following:

1. Initialize at arbitrary $\theta_1^{(0)}$ and $\theta_2^{(0)}$.
2. Repeat the following for $t = 1, \dots, N$:
 - a) Pick J to be either 1 or 2 with probability 0.5.
 - b) If $J = 1$, take $\theta_2^{(t+1)} = \theta_2^{(t)}$ and generate $\theta_1^{(t+1)}$ from the normal distribution with mean $y_1 + \rho(\theta_2^{(t)} - y_2)$ and variance $1 - \rho^2$.
 - c) If $J = 2$, take $\theta_1^{(t+1)} = \theta_1^{(t)}$ and generate $\theta_2^{(t+1)}$ from the normal distribution with mean $y_2 + \rho(\theta_1^{(t)} - y_1)$ and variance $1 - \rho^2$.

This example is also useful for illustrating one potential problem with the Gibbs sampler. Suppose $\rho = 1$. In this case $\theta_1 - y_1 \sim N(0, 1)$ and $\theta_2 - y_2 = \theta_1 - y_1$. The Gibbs iterates in this case will be

$$\theta_1^{(t+1)} - y_1 = \theta_2^{(t)} - y_2 \quad \text{if } J = 1$$

and

$$\theta_2^{(t+1)} - y_2 = \theta_1^{(t)} - y_1 \quad \text{if } J = 2$$

This results in the chain not moving after the first iterate. For example, if the first step is:

$$\begin{pmatrix} \theta_1^{(0)} \\ \theta_2^{(0)} \end{pmatrix} \rightarrow \begin{pmatrix} y_1 - y_2 + \theta_2^{(0)} \\ \theta_2^{(0)} \end{pmatrix},$$

then the chain will stay at $\begin{pmatrix} y_1 - y_2 + \theta_2^{(0)} \\ \theta_2^{(0)} \end{pmatrix}$ forever. On the other hand, if the first step is

$$\begin{pmatrix} \theta_1^{(0)} \\ \theta_2^{(0)} \end{pmatrix} \rightarrow \begin{pmatrix} \theta_1^{(0)} \\ y_2 - y_1 + \theta_1^{(0)} \end{pmatrix},$$

then the chain will stay forever at $\begin{pmatrix} \theta_1^{(0)} \\ y_2 - y_1 + \theta_1^{(0)} \end{pmatrix}$. This chain will obviously not converge unless the initial iterate already has the right distribution. A similar problem arises when $\rho = -1$.

The lesson to be drawn from this is that the Gibbs sampler has issues of convergence when there is high correlation between the variables.

2.2 Linear Regression

Consider the linear regression model that we discussed previously. We observe data $(y_1, x_1), \dots, (y_n, x_n)$ with y_1, \dots, y_n real-valued and $x_1, \dots, x_n \in \mathbb{R}^p$. We treat x_1, \dots, x_n as non-random and y_1, \dots, y_n as random. The parameter is $\theta \in \mathbb{R}^{p+1}$ with $\theta = (\beta, \sigma)$ (here $\beta \in \mathbb{R}^p$ represents the vector of regression coefficients and σ is the noise standard deviation). Consider the model

$$y_i \sim N(x_i' \beta, \sigma^2)$$

with y_1, \dots, y_n being independent conditional on θ along with the (improper) prior:

$$\pi(\beta, \sigma) \propto \frac{1}{\sigma} I\{\sigma > 0\}.$$

The exact joint posterior is given by

$$\pi(\beta, \sigma | \text{data}) \propto \sigma^{-(n+1)} \exp\left(\frac{-1}{2\sigma^2} \|Y - X\beta\|^2\right) I\{\sigma > 0\} \quad (1)$$

We can directly sample from this posterior (e.g., by first sampling from the marginal posterior of σ and then from the conditional posterior of β given σ). Alternatively, one can first sample from the marginal posterior of β and then from the conditional posterior of σ given β).

Now let us use the Gibbs sample to simulate from (1). Note that:

$$\pi(\beta | \sigma, \text{data}) \propto \exp\left(\frac{-1}{2\sigma^2} \|Y - X\beta\|^2\right) \propto \exp\left(\frac{-1}{2\sigma^2} \|X\beta - X\hat{\beta}\|^2\right) = \exp\left(\frac{-1}{2\sigma^2} (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})\right)$$

which gives

$$\beta | \sigma, \text{data} \sim N_p(\hat{\beta}, \sigma^2 (X'X)^{-1}).$$

Here $\hat{\beta}$ is the least squares estimator. For the conditional posterior of σ given β , note that

$$\pi(\sigma | \beta, \text{data}) \propto \sigma^{-(n+1)} \exp\left(\frac{-1}{2\sigma^2} \|Y - X\beta\|^2\right) I\{\sigma > 0\}.$$

Using the formula $f_{1/\sigma^2}(x) = f_\sigma(x^{-1/2})x^{-3/2}$, it is now easy to check that the density of $1/\sigma^2$ conditional on β and the data is proportional to

$$x^{(n-2)/2} \exp\left(\frac{-x}{2} \|Y - X\beta\|^2\right) I\{x > 0\}$$

which means that

$$\frac{1}{\sigma^2} \Big| \beta, \text{data} \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2} \|Y - X\beta\|^2\right).$$

The Gibbs sampler algorithm is thus:

1. Initialize at arbitrary $\beta^{(0)}$ and $\sigma^{(0)}$.
2. Repeat the following for $t = 1, \dots, N$:
 - a) Pick J to be either 1 or 2 with equal probability.
 - b) If $J = 1$, take $\beta^{(t+1)} \sim N_p(\hat{\beta}, (\sigma^{(t)})^2 (X'X)^{-1})$ and $\sigma^{(t+1)} = \sigma^{(t)}$.
 - c) If $J = 2$, take $\beta^{(t+1)} = \beta^{(t)}$. Generate an observation x from the Gamma distribution with parameters $n/2$ and $\|Y - X\beta^{(t)}\|^2/2$. Then take $\sigma^{(t+1)} = x^{-1/2}$.

2.3 Probit Regression

Our next application of the Gibbs sampler is to Probit Regression which was introduced in Homework Four. The model is given by the following. We observe data $(y_1, x_1), \dots, (y_n, x_n)$ with y_1, \dots, y_n binary and $x_1, \dots, x_n \in \mathbb{R}^p$. x_1, \dots, x_n are treated non-random and y_1, \dots, y_n random. The parameter is $\beta \in \mathbb{R}^p$. The probit regression model assumes that y_1, \dots, y_n are independent given β with

$$y_i | \beta \sim \text{Bernoulli}(\Phi(x_i' \beta))$$

where $\Phi(\cdot)$ is the standard normal cdf. Let $\pi(\cdot)$ be our prior for β . The posterior is then given by:

$$\pi(\beta | \text{data}) \propto \pi(\beta) \prod_{i=1}^n (\Phi(x_i' \beta))^{y_i} (1 - \Phi(x_i' \beta))^{1-y_i}.$$

Suppose we take the (improper) uniform prior for β so that the posterior becomes

$$\pi(\beta | \text{data}) \propto \prod_{i=1}^n (\Phi(x_i' \beta))^{y_i} (1 - \Phi(x_i' \beta))^{1-y_i}.$$

It is not actually clear how to use the Gibbs sampler on the above posterior. We can try decomposing β as $(\beta_1, \dots, \beta_p)$ and then attempt to obtain samples from the conditional:

$$\pi(\beta_j | \beta_{-j}, \text{data}) \propto \prod_{i=1}^n (\Phi(x_i' \beta))^{y_i} (1 - \Phi(x_i' \beta))^{1-y_i}.$$

I am not sure how to obtain samples from the above conditional.

There is a trick that people use to run the Gibbs sampler for probit regression. Let us introduce random variables w_1, \dots, w_n which, conditional on β , are independent with

$$w_i | \beta \sim N(x_i' \beta, 1).$$

We then let

$$y_i = I\{w_i > 0\}.$$

It is then clear that y_1, \dots, y_n defined thus have the same likelihood as in the probit regression model. Let $w = (w_1, \dots, w_n)$. The trick is to use the Gibbs sampler with (w, β) . We shall see how this is done in the next class.