UMBC, Department of Mathematics and Statistics

Project Report: Explaining the Gibbs Sampler

Ji Li

1. Introduction

This project is to give a simple explanation of how and why the Gibbs sampler works and reproduce the simulations in Casella and George (1992). The Gibbs sampler is a technique for generating random variables from a marginal distribution indirectly, without having to calculate the density. We can illustrate the application of the Gibbs sampler in bivariate situations and some higher dimensional cases. However, the Gibbs sampler doesn't always converge. The report will contain 3 examples and 5 figures to show the histograms of samples from Gibbs sampling and make some comparisons with the true marginal densities.

By using the Gibbs sampler, we are able to avoid difficult calculations, replacing them with a sequence of easier calculations. Gelfand and Smith (1990) state that it also has potential in a wide variety of conventional statistical problems.

2. Description of Gibbs Sampler and Examples

2.1 Gibbs Sampler and Gibbs Sequence

Suppose we are interested in obtaining characteristics of the marginal density $f(x) = \int \dots \int f(x, y_1, \dots, y_p)\, dy_1 \dots dy_p$ of its joint density $f(x, y_1, \dots, , y_p)$. But sometimes $f(x)$ is extremely difficult to get by calculating integration. In such cases, the Gibbs sampler allows

us effectively to generate a sample $X_1, \ldots, X_m \sim f(x)$ without requiring $f(x)$. Then based on simulations, the mean of $f(x)$, for example, $\lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} X_i = \int_{-\infty}^{\infty} x f(x) dx = EX$.

To understand how the Gibbs sampler works, we first take a two-variable case as an example. Starting with a pair of random variables (X, Y), the Gibbs sampler generates a sample from $f(x)$ by sampling instead from the conditional distributions $f(x \mid y)$ and $f(y \mid x)$, distributions that are often known in statistical models.

First the initial value $Y_0' = y_0'$ is specified, and then $X_0', Y_1', X_1'...$ are obtained from the conditional distributions

$$\begin{aligned} X_j' &\sim f\left(x \mid Y_j' = y_j'\right) \\ Y_{j+1}' &\sim f\left(y \mid X_j' = x_j'\right) \end{aligned} \tag{1}$$

The sequence of variables $Y_0', X_0', Y_1', X_1', Y_2', X_2', \ldots, Y_k', X_k'$ is called Gibbs sequence. We refer to the generation of Gibbs sequence as Gibbs sampling. It turns out that under reasonably general conditions, the distribution of X_k' converges to $f(x)$. Thus, as $k \to \infty$, the final observation in Gibbs sequence $X_k' = x_k'$, is effectively a sample point from $f(x)$.

Gibbs sampling can be used to estimate the density itself by averaging the final conditional densities from each Gibbs sequence. Just as the values $X_k' = x_k'$ yield a realization of $X_1, \cdots, X_m \sim f(x)$, the values $Y_k' = y_k'$ yield a realization of $Y_1, \cdots, Y_m \sim f(y)$. The average of the conditional densities $f\left(x \mid Y_k' = y_k'\right)$ will be a close approximation to $f(x)$ and we can estimate $f(x)$ with

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} f\left(x \mid y_i\right) \tag{2}$$

2.2 Two-variable case

2.2.1 Example1: Beta-Binomial distribution

This example is to compare two histograms of samples from the Beta-Binomial distribution. One was obtained using Gibbs sampling, and the other was generated directly from the true marginal. The following are joint distribution of X and Y and the conditional distributions.

$$\begin{aligned} &f(x,y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}, x = 0, 1, \ldots n, \quad 0 \le y \le 1 \\ &f(x \mid y) \text{ is Binomial } (n, y) \\ &f(y \mid x) \text{ is Beta } (x+\alpha, n-x+\beta) \end{aligned} \tag{3}$$

Actually, $f(x)$ can be obtained directly from

$$f(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}, \quad x = 0, 1, \ldots, n \tag{4}$$

So this example is useful for illustrative purposes.

Figure 1 displays histograms of two samples $x_1, \ldots x_m$ of size $m = 500$ from the beta-binomial distribution of (1) with $n = 16$, $\alpha = 2$, and $\beta = 4$. We choose the length of Gibbs sequence $k = 10$ in this case. The two histograms will be more similar when we take lager sample size m. And the maximum of differences between the true $f(x)$ and its estimate is 0.124. So we may claim that the Gibbs scheme for random variable generation is indeed generating variables from the marginal distribution.

In any bivariate situation, the Gibbs sampler is really not needed when the joint distribution can be calculated. But in situations where $f(x, y)$ and $f(x)$ cannot be calculated, Gibbs sampling may be indispensable.

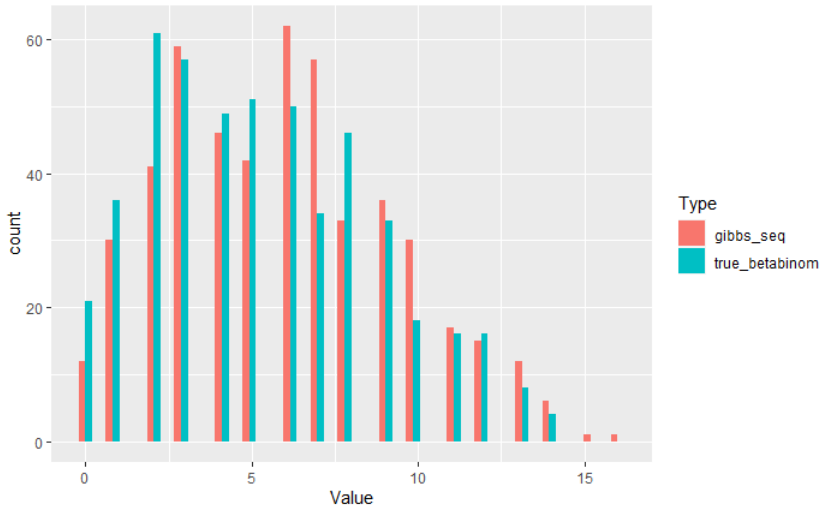

Figure 1: The red histogram sample was obtained using Gibbs sampling with length of Gibbs sequence k = 10 and sample size m=500. The blue histogram sample was generated directly from the beta-binomial distribution.

2.2.2 Example 2: Exponential distributions

This example uses (2) to estimate marginal distribution. X and Y both have conditional distributions that are exponential distributions. The form of this marginal is not easily calculable, so the Gibbs sampler can be applied to the conditionals to obtain the characteristic of $f(x)$.

$$\begin{aligned} f(x \mid y) &\propto ye^{-yx}, \quad 0 < x < B < \infty \\ f(y \mid x) &\propto xe^{-xy}, \quad 0 < y < B < \infty \end{aligned} \tag{5}$$

As explained in Casella and George (1992) Section 4.1, $f(x \mid y) = ye^{-yx}/\left(1 - e^{-By}\right), 0 < x < B$, with a similar expression for $f(y \mid x)$. Substituting these functions into (4.1) yields the solution $f(x) \propto \left(1 - e^{-Bx}\right)/x$.

Three things are compared in one figure: histogram for X of a sample from the pair of conditional distributions, an estimate of the marginal density obtained from equation $\hat{f}(x) = \frac{1}{m}\sum_{i=1}^{m} f\left(x \mid y_i\right)$, and the true marginal density.

It seems that both a sample from the pair of conditional distributions and an estimate of

the marginal density obtained from equation (2) are close to the true marginal density. But Figure 2 from Casella and George (1992) shows that an estimate from (2) is more accurate.

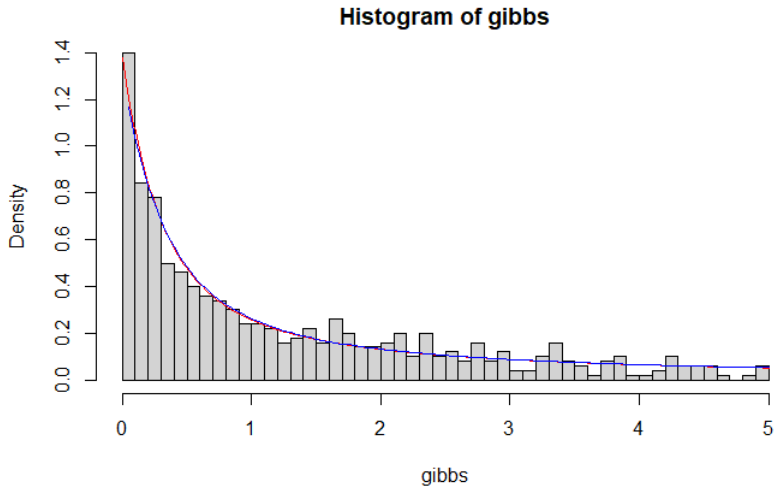

Figure 2: The histogram sample was obtained using Gibbs sampling with length of the Gibbs sequence k = 15, the upper boundary B=5 and sample size m=500. The blue curve was the true marginal distribution and the red curve was an estimate of the marginal density obtained from equation (2).

2.2.3 Example 1 (continued)

Analogous to (2), we can also estimate the marginal probabilities of X in Example 1 using

$$\hat{P}(X = x) = \frac{1}{m} \sum_{i=1}^{m} P\left(X = x \mid Y_i = y_i\right). \tag{6}$$

The maximum of differences between the true $f(x)$ and its estimate is 0.006621522, which is much smaller than the one in Example 1.

$f\left(x \mid y_1\right), \cdots, f\left(x \mid y_m\right)$, calculated using the simulated values $y_1, \cdots, y_m$, carry more information about $f(x)$ than $x_1, \cdots, x_m$ alone, and will yield better estimates. For example, an estimate of the mean$(1/m) \sum_{i=1}^{m} E\left(X \mid y_i\right)$is a better estimate than $f(x)$ is $(1/m) \sum_{i=1}^{m} x_i$, as long as these conditional expectations are obtainable. The intuition behind this feature

is the Rao-Blackwell theorem (illustrated by Gelfand and Smith (1990), and established analytically by Liu, Wong, and Kong (1991).

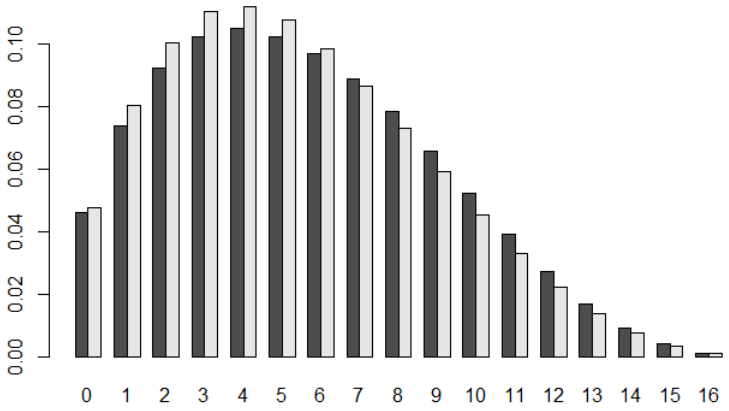

Figure 3: The black histogram represents estimates of the marginal distribution of X using Equation (6), based on a sample of Size m = 500 from the pair of conditional distributions in (3). The Gibbs sequence had length k = 10. The white histogram represents the true beta- binomial probabilities with $n = 16$, $\alpha = 2$, and $\beta = 4$.

2.2.4 Fixed point integral equation

We can consider Gibbs sampling in this way: if we know the conditional distributions, it is sufficient to determine a joint distribution. It can be illustrated in the bivariate case.

$$\begin{aligned} f_X(x) &= \int f_{XY}(x,y)dy \\ &= \int f_{X|Y}(x \mid y) f_Y(y) dy \\ &= \int f_{X|Y}(x \mid y) \int f_{Y|X}(y \mid t) f_X(t) dt dy \\ &= \int \left[\int f_{X|Y}(x \mid y) f_{Y|X}(y \mid t) dy \right] f_X(t) dt \\ &= \int h(x,t) f_X(t) dt \end{aligned} \tag{7}$$

This defines a fixed point integral equation for which $f_X(t)$ is a solution. The fact that

it is a unique solution is explained by Gelfand and Smith (1990).

2.2.5 Example 2 (continued)

This example shows that proper conditional distributions will not always determine a proper marginal distribution. In this case, when the Gibbs sampler is applied to the conditional densities, convergence breaks down.

Suppose that $B = \infty$ in example 2, so that X and Y have the conditional densities

$$f(x \mid y) = ye^{-yx}, \quad 0 < x < \infty \qquad f(y \mid x) = xe^{-xy}, \quad 0 < y < \infty \tag{8}$$

Applying (7),

$$\begin{aligned} f_X(x) &= \left[\int ye^{-yx}te^{-ty}dy\right] f_X(t)dt \\ &= \int \left[\frac{t}{(x+t)^2}\right] f_X(t)dt \end{aligned}$$

$\therefore f_X(x) = \frac{1}{x}$, which is not a density function.

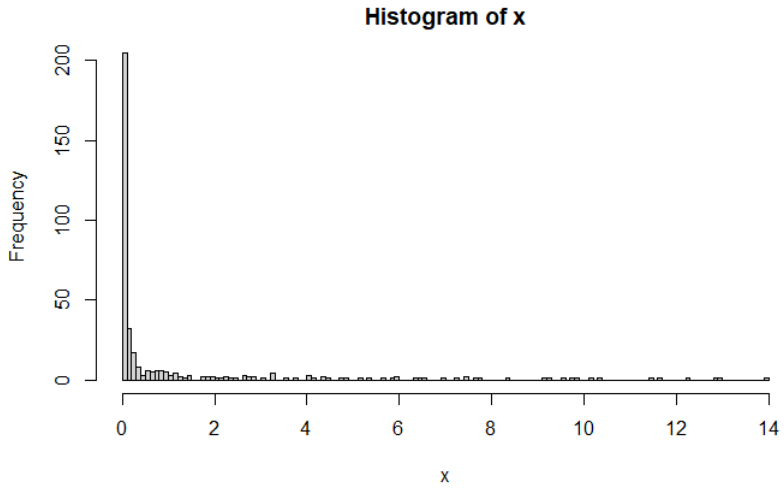

Figure 4: Histogram of a sample of size m = 500 from the conditional Distributions in example2, using Gibbs Sampling With length of Gibbs sequence k = 10.

When the Gibbs sampler is applied in this example, convergence breaks down. It does

not approximate to $1/x$.

2.3 More Than Two Variables

Similar as two varibles case, the Gibbs sampler would sample from $f_{X|YZ}, f_{Y|XZ}$, and $f_{Z|XY}$. The jth iteration is like

$$\begin{aligned} X'_j &\sim f\left(x \mid Y'_j = y'_j, Z'_j = z'_j\right) \\ Y'_{j+1} &\sim f\left(y \mid X'_j = x'_j, Z'_j = z'_j\right) \\ Z'_{j+1} &\sim f\left(z \mid X'_j = x'_j, Y'_{j+1} = y'_{j+1}\right) \end{aligned}$$

Gibbs sequence is:

$$Y'_0, Z'_0, X'_0, Y'_1, Z'_1, X'_1, Y'_2, Z'_2, X'_2, \ldots$$

Example 3 is to calculate the marginal distribution in a problem with more than two random variables X, Y, and Z. The iteration will also solve the fixed-point equation like example 2 (continued).

Suppose the joint distribution is

$$\begin{aligned} &f(x, y, n) \propto \left(\begin{array}{c} n \\ x \end{array} \right) y^{x+\alpha-1}(1-y)^{nx+\beta-1} e^{-\lambda} \frac{\lambda^n}{n!} \\ &x = 0, 1, \ldots, n, 0 \leq y \leq 1, n = 1, 2, \ldots \end{aligned} \tag{9}$$

There are three conditional densities:

$$\begin{aligned} &f(x \mid y, n) \text{ is binomial } (n, y) \\ &f(y \mid x, n) \text{ is beta } (x + \alpha, n - x + \beta) \\ &f(n \mid x, y) \propto e^{-(1-y)\lambda} \frac{[(1-y)\lambda]^{n-x}}{(n-x)!}, n = x, x+1, \cdots \end{aligned} \tag{10}$$

Now apply the iteration to the distributions in (10)to generate a sequence $X_1, X_2, \ldots, X_m$

from $f(x)$ and use this sequence to estimate the desired characteristic. This is done and is given in Figure 5, which shows the probabilities of the Marginal Distribution of X Using Equation (6).

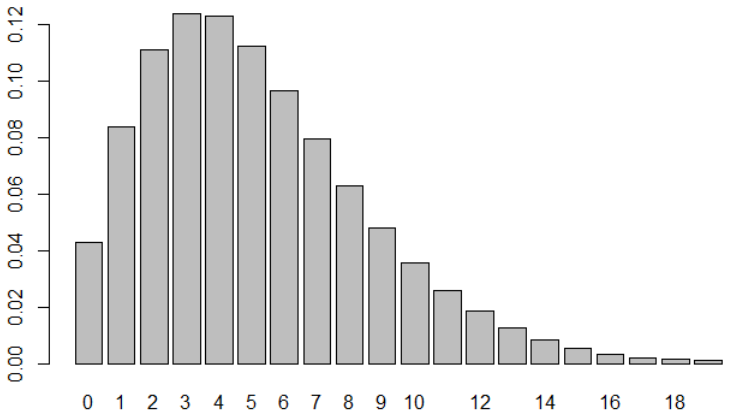

Figure 5: Estimates of probabilities of the marginal distribution of X, based on a sample of size $m = 500$ from the conditional distributions in (10) with $\lambda = 16, \alpha = 2$, and $\beta = 4$. The Gibbs sequences had length $k = 10$.

3. Discussion

In general, this project introduces the concept of Gibbs sampler and applied it to some examples in two variables and three variables cases. In these examples, we know the true joint distributions or the true marginal distributions $f_X(x)$ so that we can evaluate if the Gibbs scheme is indeed generating variables from the marginal distribution. The figures show that the Gibbs sampler can estimate $f(x)$ very well. What's more, it would be better to estimate $f(x)$ with equation $\hat{f}(x) = \frac{1}{m}\sum_{i=1}^{m} f\left(x \mid y_i, z_i, \ldots\right)$ rather than use $x_1, \ldots, x_m$ alone. Example 2 (continued) also shows that we cannot always generate the marginal distribution from the conditional distributions. And if we apply Gibbs sampler, it would not converge.

There are some important issues in Gibbs sampling surround the implementation and comparison of the various approaches to extracting information from the Gibbs sequence.

For instance, to sample from $f(x)$, one popular approach is to selecting some large value for k, and then to using any X'_j for $j \geq k$. A general strategy for choosing such k is to monitor the convergence of some aspect of the Gibbs sequence. For example, Gelfand and Smith (1990) and Gelfand, Hills, Racine-Poor, and Smith (1990) suggest monitoring density estimates from m independent Gibbs sequences, and choosing k to be the first point at which these densities appear to be the same under a "felt-tippen test." An alternative may be to choose k based on theoretical considerations, as in Raftery and Banfield (1990). M.T. Wells (personal communication) has suggested a connection between selecting k and the cooling parameter in simulated annealing.

Approaches to sampling the Gibbs sequence is also a topic to research. We can generate one long Gibbs sequence and then extract every r-th observation or all the observations after the burn in period. For large, computationally expensive problems, a less wasteful approach to exploiting the Gibbs sequence is to use all realizations of X'_j for $j \leq k$, as in George and McCulloch (1991). Although the resulting data will be dependent, it will still be the case that the empirical distribution of X'_j converges to $f(x)$.