Use of Gibbs Sampling to Approximate the Posterior

As we discussed in Chapter 4, it is often not known how to efficiently sample from the posterior distribution, especially when the number of parameters is large.

Gibbs sampling provides a means of sampling from a distribution when *it is known how to sample from each one of the full conditional distributions*.

What are the full conditionals?

Suppose that $f(x_1, \ldots, x_p)$ is the joint density from which we want to draw samples. The *full conditionals* are

$$f_1(x_1|x_2, \ldots, x_p),\ f_2(x_2|x_1, x_3, \ldots, x_p), \ldots,$$

$$f_i(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p), \ldots$$

$$f_p(x_p|x_1, \ldots, x_{p-1}).$$

To illustrate the idea behind Gibbs sampling, let $p = 3$, and suppose that (a, b, c) has been randomly selected from the joint density of (X_1, X_2, X_3).

1. We know that (b, c) has been randomly selected from the marginal of (X_2, X_3). (Recall point 1 in the digression on p. 84N.)

2. Now suppose that we randomly select a value from $f_1(\cdot|b, c)$, and call the selected value g.

3. It follows that (g, b, c) is a random selection from the density of (X_1, X_2, X_3), and so (g, c) is a selection from the marginal of (X_1, X_3).

4. Randomly select a value from $f_2(\cdot|g, c)$, and call this value h.

5. As before, (g, h, c) is a random selection from the density of (X_1, X_2, X_3), and so (g, h) is a selection from the marginal of (X_1, X_2).

6. Finally select a value from $f_3(\cdot|g, h)$ and call the selected value i. The value (g, h, i) is a draw from the joint density of (X_1, X_2, X_3).

It's great that we can use this method to draw samples from a joint density, but note that *the draws (a, b, c) and (g, h, i) are not independent*.

This is different from the situation we're used to where different draws from a distribution are independent.

Actually, (a, b, c), (g, b, c), (g, h, c) and (g, h, i) are all draws from the joint density of (X_1, X_2, X_3), but in Gibbs sampling only (a, b, c) and (g, h, i) are retained.

The draws (a, b, c) and (g, b, c), for example, are *more* highly related with each other than are (a, b, c) and (g, h, i), and so (g, b, c) is not retained.

Pages 137-138N provide an essentially complete description of the Gibbs sampling algorithm.

Once (g, h, i) has been obtained, the same process can be repeated to get a third draw, and so on.

There remains the question of how to get a starting value.

Typically, one tries to select a value near the center of the distribution.

A good starting value is not critical, because once enough values are generated, the effect of the starting value is negligible.

One can simply discard the first m values generated, and retain the rest. The first m values are often referred to as a *burnin period*.

Often a good choice for m is evident from looking at plots of the generated observations.

The key question is whether or not the observations have *converged* to a stationary phase.

An example will help illustrate these ideas.

Example 12 Generating observations from a trivariate normal distribution using the Gibbs sampler

Suppose (X_1, X_2, X_3) has a trivariate normal distribution with mean vector $(0, 0, 0)$ and covariance matrix $\boldsymbol{\Sigma}$ with (i, j) element σ_{ij}.

The conditional distribution of X_1 given $X_2 = x_2$ and $X_3 = x_3$ is $N(\mu_{\boldsymbol{x}}, v)$, where

$$\mu_{\boldsymbol{x}} = (\sigma_{12}, \sigma_{13})\boldsymbol{V}^{-1}(x_2, x_3)^T,$$

$$v = \sigma_{11} - (\sigma_{12}, \sigma_{13})\boldsymbol{V}^{-1}(\sigma_{12}, \sigma_{13})^T$$

and

$$\boldsymbol{V} = \begin{bmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{23} & \sigma_{33} \end{bmatrix}.$$

The other two full conditionals are also normal, and their means and variances can be inferred from the previous page. How?

In R I will generate observations in the case where

$$\Sigma = \begin{bmatrix} 9 & 0.80(3)(4) & 0.64(3)(5) \\ 0.80(3)(4) & 16 & 0.80(4)(5) \\ 0.64(3)(5) & 0.80(4)(5) & 25 \end{bmatrix}.$$

Each number in blue is a correlation between two variables.

My Gibbs sampler code may be found at the course website.

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be the sequence of observations generated using the Gibbs sampler.

We saw evidence of two things in Example 12:

- The generated observations are correlated with each other, although the correlation between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ dies out the larger $|i-j|$ is.

- In spite of the correlation, sample averages still resulted in good estimates of the corresponding population averages.

Both of these things are generally true.

The output from a Gibbs sampler is a *Markov chain*, meaning that

$$P(\boldsymbol{X}_t \in A | \boldsymbol{X}_0 = \boldsymbol{x}_0, \ldots, \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1}) =$$

$$P(\boldsymbol{X}_t \in A | \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1}),$$

or, in words, $\boldsymbol{X}_t$ depends on the whole history of the process only through the most recent value.

So, the observations *are* correlated, but in a fairly simple and well-understood way.

It is also true that the output is eventually *stationary*, in the sense that, regardless of starting value, *the density of $\boldsymbol{X}_t$ is approximately f for all t sufficiently big*, where f is the density whose full conditionals are used to generate the data.

Because the output is a stationary Markov chain, sample averages converge to population averages (assuming the moments exist).

So, for continuous functions g,

$$\frac{1}{n}\sum_{t=1}^{n} g(\boldsymbol{X}_t) \xrightarrow{p} \int g(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}$$

as $n \to \infty$.

Read carefully the part of Section 6.5 in Hoff with the subtitle *Distinguishing parameter estimation from posterior approximation.*

We can summarize this part of the text as follows:

- All the information about the sampled population is contained in the posterior.

- Sampling more and more observations from the posterior only gives more information about the *posterior, but doesn't add to the information in the data and prior.*

MCMC diagnostics

The two most relevant issues when considering output from Gibbs sampling, or any other MCMC method, are

- How quickly does the marginal distribution of the chain converge?

- Once the chain has converged, how rapidly does it *mix*?

Convergence is mostly a function of the starting value. Typically, convergence takes a long time only when the chain starts out in a region of low probability.

Mixing refers to how quickly the chain moves around the support of the random vector. Good mixing is exemplified by data that are independent.

Poor mixing (of a scalar component) is indicated by the chain staying on the same side of the population median for several iterations in a row.

We said previously that sample averages of Gibbs sampling output converge to the corresponding population averages as we obtain more and more samples.

However, when mixing is poor, *these averages do not converge as quickly as they would if the data were independent.*

Poor mixing is a manifestation of *positive correlation* among the generated data.

Let $X_1, X_2, \ldots, X_n$ be identically distributed but *not* independent observations. We may judge how rapidly the sample mean $\bar{X}$ converges to $E(X_i)$ by considering Var(X_i).

We have

$$\begin{aligned} \mathrm{Var}(\bar{X}) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j) \qquad (8) \\ &= \frac{\mathrm{Var}(X_1)}{n} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{Cov}(X_i, X_j). \end{aligned}$$

Notice that $\mathrm{Var}(X_1)/n$ is the variance we get for $\bar{X}$ when the observations are independent.

Now, in Gibbs sampling the covariance terms are usually positive, which means that $\mathrm{Var}(\bar{X})$ will be larger than if the observations were independent.

A good diagnostic for mixing (or lack thereof) is the *sample autocorrelation function (ACF)*. For the sequence $X_1, X_2, \ldots, X_n$, the sample ACF is

$$\widehat{\rho}(t) = \frac{\sum_{i=1}^{n-t}(X_i - \bar{X})(X_{i+t} - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

We use $\widehat{\rho}(t)$ to estimate the correlation between X_i and X_{i+t}, i.e., observations that are t iterations apart.

Assessing the variance of a sample mean

Obviously it's important to be able to estimate $\mathrm{Var}(\bar{X})$. There are two main ways we can deal with the effect of correlation on this variance.

- Determining *effective number of independent observations.*

- The use of *thinning*.

The effective number of independent observations is n_{eff}, which is defined by

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_i)}{n_{\mathrm{eff}}},$$

where $\text{Var}(\bar{X})$ is as given in (8). Typically n_{eff} is smaller than the actual number (n) of generated observations.

For n_{eff} large, an approximate $(1-\alpha)100\%$ confidence interval for $E(X_i)$ would be

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{S}{\sqrt{n_{\text{eff}}}},$$

where S^2 is the sample variance of all n observations.

Thinning

As we observed previously, values far enough apart are usually not highly correlated. This leads to the practice of *thinning*, wherein we only use every kth observation.

The idea is to choose k large enough that the thinned data effectively form an i.i.d. sequence.

Then we can apply the *usual* ways of assessing the variance of a sample mean to the thinned data.

The sample ACF may be used to determine a good choice for k.

Suppose that $\widehat{\rho}(t) \approx 0$ for all $t \geq k$. Then to a good approximation observations k lags apart are independent.