Chapter 5
The Gibbs Sampler and Applications

Qing Zhou

Contents

1 The Gibbs Sampler ................................................. 2
  1.1 Algorithms ................................................. 2
  1.2 Stationary distribution and detail balance .................. 7
2 Examples of the Gibbs Sampler ............................... 8
  2.1 The slice sampler .......................................... 8
  2.2 Blocked Gibbs sampler .................................... 10
3 Missing Data Problems ......................................... 13
  3.1 Two-block Gibbs sampler .................................. 13
  3.2 Discrete data example ..................................... 14
  3.3 Gaussian data example .................................... 16

*UCLA Department of Statistics (email: zhou@stat.ucla.edu).
1. The Gibbs Sampler

The target distribution is $\pi(x) = \pi(x_1, x_2, \cdots, x_d)$, $x \in \mathbb{R}^d$. Following the notation in Chapter 4 (§5.3), define

$$x^{(t)} = (x_1^{(t)}, x_2^{(t)}, \cdots, x_d^{(t)}),$$
$$x_i^{(t)}(y) = (x_1^{(t)}, \cdots, x_{i-1}^{(t)}, y, x_{i+1}^{(t)}, \cdots, x_d^{(t)}),$$
$$x_i^{(t)}[-i] = (x_1^{(t)}, \cdots, x_{i-1}^{(t)}, x_{i+1}^{(t)}, \cdots, x_d^{(t)}).$$

1.1. Algorithms

The Gibbs sampler iteratively samples from the conditional distribution $\pi(\cdot | x_{-i}^{(t)})$ for a chosen coordinate $i \in \{1, \ldots, d\}$. There are two ways to pick a coordinate, corresponding to random-scan versus systematic-scan Gibbs sampler:

**Algorithm 1** (Random-scan Gibbs sampler). Pick an initial value $x^{(1)}$.

For $t = 1, \ldots, n$:

1. Randomly select a coordinate $i$ from $\{1, 2, \cdots, d\}$;
2. Draw $y$ from the conditional distribution $\pi(x_i | x_{-i}^{(t)})$. Let $x^{(t+1)} = x_i^{(t)}(y)$ (i.e. $x_{i}^{(t+1)} = y, x_{-i}^{(t+1)} = x_{-i}^{(t)}$).

**Algorithm 2** (Systematic-scan Gibbs sampler). Pick an initial value $x^{(1)}$.

For $t = 1, \ldots, n$: Given the current sample $x^{(t)} = (x_1^{(t)}, \cdots, x_d^{(t)})$,

for $i = 1, 2, \cdots, d,$

$$x_i^{(t+1)} \sim \pi(x_i | x_1^{(t+1)}, \cdots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \cdots, x_d^{(t)}).$$

By default, we use systematic-scan (Algorithm 2) unless noted otherwise. Given samples $\{x^{(t)} : t = 1, \ldots, n\}$ generated by the Gibbs sampler, we estimate $\mathbb{E}_{\pi} h(x)$, the expectation of $h(x)$ with respect to $\pi$, by the sample average:

$$\hat{h} = \frac{1}{n} \sum_{t=1}^{n} h(x^{(t)}). \quad (1)$$

Similar to the MH algorithm, we often throw away samples generated during the burn-in period, say the first 1000 iterations, and calculate $\hat{h}$ from post burn-in samples.

To design a Gibbs sampler for a joint distribution $\pi(x)$, the key is to derive conditional distributions $[x_i | x_{-i}]$ for all $i$. We will demonstrate how to find such conditional distributions in a few examples.
Example 1. Design a Gibbs sampler to simulate from a bivariate Normal distribution:

\[ \mathbf{X} = (X_1, X_2) \sim \mathcal{N}_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \]

i.e. the pdf of the target distribution is

\[ \pi(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp \left\{ -\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)} \right\}. \]

Use the samples to estimate \( E(X_1X_2) \) and the correlation coefficient \( \text{cor}(X_1, X_2) \).

Find the conditional distribution \( [x_1 \mid x_2] \) as follows: Regarding \( x_2 \) as a constant,

\[ \pi(x_1 \mid x_2) \propto \pi(x_1, x_2) \propto \exp \left\{ -\frac{x_1^2 - 2\rho x_2 x_1}{2(1-\rho^2)} \right\}, \tag{2} \]

where any multiplicative factor that only depends on \( x_2 \) is regarded as a constant and absorbed into the proportion sign. Now complete squares:

\[ x_1^2 - 2\rho x_2 x_1 = (x_1 - \rho x_2)^2 - (\rho x_2)^2, \]

and plug it into (2),

\[ \pi(x_1 \mid x_2) \propto \exp \left\{ -\frac{(x_1 - \rho x_2)^2}{2(1-\rho^2)} \right\}, \]

which is an unnormalized density for \( \mathcal{N}(\rho x_2, 1-\rho^2) \). Thus,

\[ x_1 \mid x_2 \sim \mathcal{N}(\rho x_2, 1-\rho^2). \]

Similarly, \( x_2 \mid x_1 \sim \mathcal{N}(\rho x_1, 1-\rho^2). \)

Gibbs sampler (one iteration): Given \( \mathbf{x}^{(t)} = (x_1^{(t)}, x_2^{(t)}) \),

\[ x_1^{(t+1)\mid x_2^{(t)}} \sim \mathcal{N}(\rho x_2^{(t)}, 1-\rho^2). \tag{3} \]
\[ x_2^{(t+1)\mid x_1^{(t+1)}} \sim \mathcal{N}(\rho x_1^{(t+1)}, 1-\rho^2). \tag{4} \]
#R code: Gibbs sampler for Example 5 (bivariate normal)

rho=0.8;
n=6000;
X=matrix(0,n,2);
X[1,]=c(10,10);

for(t in 2:n)
{
    X[t,1]=rnorm(1,rho*X[t-1,2],sqrt(1-rho^2));
    X[t,2]=rnorm(1,rho*X[t,1],sqrt(1-rho^2));
}

#estimate E(X1X2)
B=1001; #post burn-in
h=X[,1]*X[,2];
acf(h)
h_hat=mean(h[B:n])

#estimate cor(X1,X2)
r=cor(X[B:n,1],X[B:n,2])

Using the post burn-in samples \( t \geq B \), the estimates of \( E(X_1X_2) \) and \( \text{cor}(X_1, X_2) \) were:

> h_hat
[1] 0.7448093
> r
[1] 0.7851272

The samples generated in the first 100 iterations and the autocorrelation plot for \( h^{(i)} = x_1^{(i)} x_2^{(i)} \) are shown below:
For this Gibbs sampler, we can use induction to work out the distribution of \( x^{(t)} \) for any \( t \geq 1 \), assuming we initialize the algorithm at \( (x^{(0)}_1, x^{(0)}_2) \):

\[
\begin{pmatrix} x^{(t)}_1 \\ x^{(t)}_2 \end{pmatrix} \sim \mathcal{N}_2 \left( \begin{pmatrix} \rho^{2t-1} x^{(0)}_2 \\ \rho^{2t} x^{(0)}_2 \end{pmatrix}, \begin{pmatrix} 1 - \rho^{4t} & \rho - \rho^{4t} \\ \rho - \rho^{4t} & 1 - \rho^{4t} \end{pmatrix} \right)
\]

\[
t \rightarrow \infty \Rightarrow \mathcal{N}_2 \left( \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} \rho \\ 1 \end{pmatrix} \right).
\]

(5)

In particular, (5) shows that the limiting distribution is indeed \( \pi(x) \).

**Example 2.** Consider a joint distribution between a discrete and a continuous random variables:

\[
\pi(x, y) \propto \binom{n}{x} y^{x+\alpha-1} (1-y)^{n-x+\beta-1}
\]

for \( x = 0, 1, \ldots, n \) and \( y \in [0, 1] \). The two conditional distributions are derived as follows:

\[
\pi(x|y) \propto \binom{n}{x} y^x (1-y)^{n-x} \Rightarrow x|y \sim \text{Bin}(n, y).
\]

\[
\pi(y|x) \propto y^{x+\alpha-1} (1-y)^{n-x+\beta-1} \Rightarrow y|x \sim \text{Beta}(x+\alpha, n-x+\beta).
\]

The pdf of the Beta(\( \alpha, \beta \)) distribution (\( \alpha > 0, \beta > 0 \)) is

\[
f(y|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1} (1-y)^{\beta-1} \quad y \in [0, 1].
\]

If \( Y \sim \text{Beta}(\alpha, \beta) \), then \( \mathbb{E}(Y) = \frac{\alpha}{\alpha+\beta} \).

If two independent random variables \( X_1 \sim \text{Gamma}(\alpha, 1) \) and \( X_2 \sim \text{Gamma}(\beta, 1) \), then

\[
\frac{X_1}{X_1+X_2} \sim \text{Beta}(\alpha, \beta).
\]
Example 3 (Gibbs sampler for 1-D Ising Model). The joint distribution for the 1-D Ising model (§3.1, Ch 4) with temperature $T > 0$ is given by

$$
\pi(x) \propto \exp \left( \frac{1}{T} \sum_{i=1}^{d-1} x_i x_{i+1} \right), \quad x_i \in \{1, -1\}.
$$

To develop a Gibbs sampler for this problem, we find the conditional distribution $[x_i \mid [\neg] i]$ for each $i = 1, \ldots, d$:

$$
\pi(x_i \mid [\neg] i) \propto \pi(x_1, \ldots, x_i, \ldots, x_d)
\propto \exp \left\{ \frac{1}{T} \left( x_1 x_2 + \ldots + x_{i-1} x_i + x_i x_{i+1} + \ldots + x_{d-1} x_d \right) \right\}
\propto \exp \left\{ \frac{x_i}{T} (x_{i-1} + x_{i+1}) \right\}, \quad x_i \in \{1, -1\}.
$$

(6)

Since $x_i \in \{1, -1\}$, put

$$
Z_i = \exp \left\{ \frac{1}{T} (x_{i-1} + x_{i+1}) \right\} + \exp \left\{ -\frac{1}{T} (x_{i-1} + x_{i+1}) \right\}.
$$

We have

$$
\pi(x_i \mid [\neg] i) = \frac{1}{Z_i} \exp \left\{ \frac{x_i}{T} (x_{i-1} + x_{i+1}) \right\} \quad \text{for} \quad x_i \in \{1, -1\}.
$$

For $i = 1$ or $d$, plug in $x_0 = x_{d+1} = 0$.

Note that $\pi(x_i \mid [\neg] i) = P(X_i = x_i \mid [\neg] i), x_i \in \{1, -1\}$, is simply a binary discrete distribution. Let $\theta_1 = \pi(x_i = 1 \mid [\neg] i), \theta_2 = \pi(x_i = -1 \mid [\neg] i)$ and put $\text{theta} = (\theta_1, \theta_2)$. To sample from $[x_i \mid [\neg] i]$:

$$
x[i]=\text{sample}(c(1, -1), \text{size}=1, \text{replace}=\text{TRUE}, \text{prob}=\text{theta});
$$

where the vector $x$ stores the current sample. In fact, we do not need to normalize $\pi(x_i \mid [\neg] i)$ in the above code. Instead, we may set $\theta_1$ and $\theta_2$ by (6):

$$
\theta_1 = \exp \left\{ \frac{1}{T} (x_{i-1} + x_{i+1}) \right\}, \quad \theta_2 = \exp \left\{ -\frac{1}{T} (x_{i-1} + x_{i+1}) \right\},
$$

since the \textbf{sample} function will normalize $\text{theta}$ anyway.
1.2. Stationary distribution and detail balance

As a special case of the MH algorithm, the detail balance condition is satisfied for the Gibbs sampler, which implies that $\pi$ is a stationary distribution.

It is also easy to verify the detail balance condition directly. To do this, we regard each conditional sampling step as a one-step transition of the underlying Markov chain. Let $\mathbf{x} = (x_1, \ldots, x_d)$ and $\mathbf{y} = x_i(y)$. Then the one-step transition kernel $K(\mathbf{x}, \mathbf{y}) = \pi(y|x_{[-i]})$. Our goal is to show that $\pi(\mathbf{x})K(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})K(\mathbf{y}, \mathbf{x})$.

Proof.

$$
\pi(\mathbf{x})K(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{x}) \cdot \pi(y|x_{[-i]}) = \frac{\pi(\mathbf{x}) \cdot \pi(y)}{\pi(x_{[-i]})}.
$$

$$
\pi(\mathbf{y})K(\mathbf{y}, \mathbf{x}) = \pi(\mathbf{y}) \cdot \pi(x|x_{[-i]}) = \frac{\pi(\mathbf{x}) \cdot \pi(y)}{\pi(x_{[-i]})}.
$$

\qed
2. Examples of the Gibbs Sampler

2.1. The slice sampler

Suppose we want to simulate from \( \pi(x) \propto q(x) \), where \( x \in \mathbb{R}^d \). The slice sampler simulates from a uniform distribution over the region under the surface of \( q(x) \) by the Gibbs sampler, based on the following result:

**Lemma 1.** Suppose a pdf \( \pi(x) \propto q(x), \ x \in \mathbb{R}^d \). Denote the region under the surface of \( q(x) \) by

\[
S = \{(x, y) \in \mathbb{R}^{d+1} : y \leq q(x)\}.
\]

If \( (X, Y) \sim \text{Unif}(S) \), then the marginal distribution of \( X \) is \( \pi \), i.e. \( X \sim \pi \).

\[\begin{array}{c}
\text{y} \\
\text{S} \\
\text{x}
\end{array}\]

\[q(x)\]

**Proof.** Let \( |S| \) denote the volume of \( S \):

\[
|S| = \int q(x)dx. \tag{7}
\]

Since \( (X, Y) \sim \text{Unif}(S) \), their joint pdf is

\[
f_{X,Y}(x, y) = 1/|S|, \quad (x, y) \in S.
\]

If \( X = x \), the range of \( Y \) is \((0, q(x))\). Then the marginal density at \( x \) is

\[
p_X(x) = \int_0^{q(x)} f_{X,Y}(x, y)dy = \int_0^{q(x)} \frac{1}{|S|} dy = \frac{q(x)}{|S|} = \pi(x).
\]

The last equality in the above is due to the fact that \( |S| \) is the normalizing constant for \( q(x) \) as in (7).

The slice sampler uses a Gibbs sampler to simulate from \( \text{Unif}(S) \) by iterating between \( Y \mid X \) and \( X \mid Y \). Then, according to Lemma 1, the marginal distribution of \( X \) is the target distribution \( \pi(x) \). It is easy to see that

\[
Y \mid X = x \sim \text{Unif}(0, q(x)).
\]
Let $X(y) = \{ x \in \mathbb{R}^d : q(x) \geq y \}$ be the set of $x$ with $q(x) \geq y$, i.e. a super-level set of $q(x)$. Then as shown in the following figure, $X \mid Y = y \sim \text{Unif}(X(y))$.

Consequently, one iteration of the slice sampler consists of two conditional sampling steps: Given $x^{(t)}$,

1. Draw $y^{(t+1)} \sim \text{Unif}[0, q(x^{(t)})]$ (vertical blue dashed line);
2. Draw $x^{(t+1)}$ uniformly from region $X^{(t+1)} = \{ x \in \mathbb{R}^d : q(x) \geq y^{(t+1)} \}$ (horizontal red dashed line).

Then when $t$ is large, $(x^{(t)}, y^{(t)}) \sim \text{Unif}(S)$ and $x^{(t)} \sim \pi$, achieving the goal of sampling from $\pi$.

Example 4 ($t_d$-distribution). Use slice sampler to simulate from $t$-distribution with $d$ degree of freedom:

$$
\pi(x) \propto (1 + x^2/d)^{-(d+1)/2} := q(x), \quad x \in \mathbb{R}.
$$

Suppose the sample at iteration $t$ is $x_t$. The two steps to generate $x_{t+1}$ are:

1. Draw $y_{t+1} \sim \text{Unif}[0, q(x_t)]$, where $q(x_t) = (1 + x_t^2/d)^{-(d+1)/2}$.
2. Draw $x_{t+1}$ uniformly from the interval

$$
X_{t+1} = \{ x \in \mathbb{R} : q(x) \geq y_{t+1} \} = [-b(y_{t+1}), b(y_{t+1})],
$$

where $b(y) = \sqrt{d(y^{-2/(d+1)} - 1)}$. Note that $\pm b(y)$ are the two roots of the quadratic equation $q(x) = y$. 
2.2. Blocked Gibbs sampler

Partition \( \{1, \ldots, d\} \) into two blocks, \( A \) and \( B \): \( A \cup B = \{1, \ldots, d\} \) and \( A \cap B = \emptyset \).

For \( \mathbf{x} = (x_1, \ldots, x_d) \), let \( x_A = (x_j : j \in A) \) and \( x_B = (x_j : j \in B) \) denote two subvectors with components in the sets \( A \) and \( B \), respectively.

A two-block Gibbs sampler iteratively sample from \([x_A | x_B]\) and \([x_B | x_A]\) in each iteration of Algorithm 2: Given the current sample \((x_A^{(t)}, x_B^{(t)})\),

\[
\begin{align*}
\text{draw } x_A^{(t+1)} &\sim \pi(x_A | x_B^{(t)}), \\
\text{draw } x_B^{(t+1)} &\sim \pi(x_B | x_A^{(t+1)}).
\end{align*}
\]

Consider the Ising model on a graph \( G = (V, E) \), where \( V = \{1, \ldots, d\} \) is the vertex set and \( E \subseteq V \times V \) is the edge set of the graph \( G \): There is an edge between two vertices \( i, j \) if and only if \((i, j) \in E\). Given \( G \), define a Boltzmann distribution for \((X_1, \ldots, X_d)\) at temperature \( T > 0\):

\[
\pi(x_1, \ldots, x_d) \propto \exp \left\{ \frac{1}{T} \sum_{(i,j) \in E} x_i x_j \right\}, \quad x_i \in \{1, -1\}. \quad (8)
\]

**Definition 1.** For three random vectors \( X, Y, Z \), we say \( X \) is *conditionally independent* of \( Z \) given \( Y \), denoted by \( X \perp \perp Z \mid Y \), if

\[
\mathbb{P}(X \in A \mid Y, Z) = \mathbb{P}(X \in A \mid Y)
\]

for any set \( A \) in the sample space of \( X \). That is, the conditional distribution of \([X \mid Y, Z]\) does *not* depend on \( Z \).

This joint distribution (8) implies the following conditional independence statements among \( X_1, \ldots, X_d \):

**Theorem 1.** Let \( N_i = \{j : (i, j) \in E\} \) be the set of neighbors of vertex \( i \) in the graph \( G \). If \( k \notin N_i \) and \( k \neq i \), then

\[ X_i \perp \perp X_k \mid \{X_j : j \in N_i\}. \]

**Proof.** It follows from (8) that the conditional density of \( X_i \) given \( X_{[-i]} \) is

\[
\pi(x_i \mid x_{[-i]}) \propto \exp \left( \frac{1}{T} \sum_{j \in N_i} x_j \right) = \pi(x_i \mid x_j, j \in N_i),
\]

which only depends on \( x_j, j \in N_i \). \( \square \)
This theorem shows that the graph $G$ (the neighborhoods of vertices) encodes conditional independence statements among the random variables.

A few common examples of graphs $G$:

- Chain, $E = \{(1, 2), (2, 3), \ldots, (d − 1, d)\}$: 1-D Ising model (Example 3).
- Complete graph, $E = \{(i, j) : i < j\}$, i.e. there is an edge between every pair of nodes $i, j$. For example, a complete graph over four nodes ($d = 4$):

  \[ \begin{array}{cccc}
  X_1 & \cdots & \cdots & X_d \\
  \vdots & \ddots & \ddots & \vdots \\
  X_d & \cdots & \cdots & X_1
  \end{array} \]

- Star topology, $E = \{(1, i) : i = 2, \ldots, d\}$: $X_1$ is the hub node (vertex) and is the only neighbor of all other nodes $X_2, \ldots, X_d$.

  \[ X_i \perp \perp X_j \mid X_1 \quad \text{for all } i \neq j \in \{2, \ldots, d\}. \quad (9) \]

  \[ \begin{array}{cccc}
  X_1 & X_2 & \cdots & X_d \\
  \vdots & \cdots & \ddots & \vdots \\
  X_d & \cdots & \cdots & X_1
  \end{array} \]

**Example 5.** If $G$ has a star topology, we can develop a two-block Gibbs sampler to sample from (8) by letting $A = \{1\}$ and $B = \{2, \ldots, d\}$. The two conditional sampling steps in one iteration of the Gibbs sampler are:

1. Sample from $[x_A \mid x_B] = [x_1 \mid x_{[-1]}]$:

   \[ \pi(x_1 \mid x_2, \ldots, x_d) \propto \exp \left[ \frac{1}{T} (x_2 + \ldots + x_d) x_1 \right], \]

   for $x_1 \in \{1, -1\}$, after normalization we have

   \[ \pi(x_1 \mid x_2, \ldots, x_d) = \frac{\exp \left[ \frac{1}{T} (x_2 + \ldots + x_d) x_1 \right]}{\exp \left[ \frac{1}{T} (x_2 + \ldots + x_d) \right] + \exp \left[ -\frac{1}{T} (x_2 + \ldots + x_d) \right]}, \]

   $x_1 \in \{1, -1\}$.

2. Sample from $[x_B \mid x_A] = [x_{[-1]} \mid x_1]$:

   Since $X_2, \ldots, X_d$ are independent
given $X_1 = x_1$ by (9), we have

$$\pi(x_2, \ldots, x_d \mid x_1) = \prod_{j=2}^{d} \pi(x_j \mid x_1)$$

$$\propto \prod_{j=2}^{d} \exp\left(\frac{x_1 x_j}{T}\right), \quad x_j \in \{1, -1\}.$$ 

Draw $x_j$ from $[x_j \mid x_1]$ for each $j = 2, \ldots, d$ independently according to:

$$\pi(x_j \mid x_1) = \frac{\exp\left(\frac{1}{T} x_1 x_j\right)}{\exp\left(\frac{x_j}{T}\right) + \exp\left(-\frac{x_j}{T}\right)}, \quad x_j \in \{1, -1\}.$$
3. Missing Data Problems

Suppose we have data
\[ y_1, y_2, \ldots, y_n \sim i.i.d. f(y | \theta), \]
where each data point \( y_i = (y_{i1}, y_{i2}, \ldots, y_{ip}) \in \mathbb{R}^p \). Put them into a data matrix \( Y = (y_{ij})_{n \times p} \). However, some data points contain missing elements, shown as ‘?’ in the following table, such as \( y_2 \) and \( y_n \).

<table>
<thead>
<tr>
<th>( Y )</th>
<th>1</th>
<th>2</th>
<th>\ldots</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y_2 )</td>
<td>?</td>
<td>?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ldots )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y_n )</td>
<td>?</td>
<td>?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

?: missing value (e.g. \( y_{22}, y_{2p}, \ldots, y_{n2} \))

\( Y_{\text{obs}} \): observed elements of \( Y \) (observed data).
\( Y_{\text{mis}} \): missing elements of \( Y \) (missing data).
\( Y = (Y_{\text{obs}}, Y_{\text{mis}}) \): complete data.

Denote by \( Y_{\text{obs}} \) the observed elements of \( Y \) and \( Y_{\text{mis}} \) the missing elements of \( Y \). We call \( Y_{\text{obs}} \) the observed data, \( Y_{\text{mis}} \) the missing data, and \( Y = (Y_{\text{obs}}, Y_{\text{mis}}) \) the complete data. Our goal is to estimate the model parameter \( \theta \) based on the observed data \( Y_{\text{obs}} \).

3.1. Two-block Gibbs sampler

Bayesian inference for missing data problems (1) estimates \( \theta \) and (2) predicts missing data \( Y_{\text{mis}} \) based on the joint posterior distribution of \( (\theta, Y_{\text{mis}}) \):
\[
p(\theta, Y_{\text{mis}} | Y_{\text{obs}}) \propto p(\theta)p(Y_{\text{obs}}, Y_{\text{mis}} | \theta),
\]
where \( p(\theta) \) is the prior for \( \theta \) and
\[
p(Y_{\text{obs}}, Y_{\text{mis}} | \theta) = p(Y | \theta) = \prod_i f(y_i | \theta)
\]
is the complete-date likelihood.

Usually there are no closed-form formulas for posterior mean or quantiles of the posterior distribution of \( \theta \):
\[
p(\theta | Y_{\text{obs}}) \propto p(\theta)p(Y_{\text{obs}} | \theta)
= p(\theta) \int p(Y_{\text{obs}}, Y_{\text{mis}} | \theta) dY_{\text{mis}},
\]
which involves marginalization over the missing data \( Y_{\text{mis}} \). We need to draw samples of \( (\theta, Y_{\text{mis}}) \) from the joint posterior distribution \( \{\theta, Y_{\text{mis}} | Y_{\text{obs}}\} \) to perform Bayesian inference. To do that, we develop a two-block Gibbs sampler, one iteration of which contains two conditional sampling steps:
1. Given $\theta^{(t)}$, draw $Y_{mis}^{t+1} \sim p(Y_{mis} \mid Y_{obs}, \theta^{(t)})$;

2. Given $Y_{mis}^{t+1}$, draw $\theta^{(t+1)} \sim p(\theta \mid Y_{obs}, Y_{mis}^{t+1}) = p(\theta \mid Y^{t+1})$, where $Y^{t+1} = (Y_{obs}, Y_{mis}^{t+1})$ is a complete data matrix with missing values imputed as $Y_{mis}^{t+1}$.

This two-block Gibbs sampler is illustrated by the following diagram:

For many commonly used models, both conditional sampling steps are easy to implement, as shown by the following examples.

### 3.2. Discrete data example

**Example 6.** Suppose $x_1, x_2, \cdots, x_n \overset{\text{iid}}{\sim} \text{Discrete}(\theta_1, \theta_2, \theta_3)$:

$$\mathbb{P}(x_i = k) = \theta_k, \ k = 1, 2, 3.$$  

As shown in the following table, the data is coarsened, in which $x_1, x_2, x_3$ are only partially classified: $x_1 \in \{2, 3\}$, $x_2 \in \{1, 3\}$ and $x_3 \in \{1, 2\}$, while the other data points are fully classified: $x_4 = 1, \ldots, x_n = 2$.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>$x_2$</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>$x_3$</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>$x_4$</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_5$</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$x_n$</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Prior: $\theta \sim \text{Dir}(\alpha_1, \alpha_2, \alpha_3)$, \quad $(\theta_1 + \theta_2 + \theta_3 = 1)$

$$p(\theta_1, \theta_2, \theta_3) \propto \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \theta_3^{\alpha_3-1}.$$
Missing data in $x_1, x_2, x_3,$ and $Y_{obs} = (x_1 \neq 1, x_2 \neq 2, x_3 \neq 3, x_4, \ldots, x_n)$.

$$p(\theta, x_1, x_2, x_3 | x_4, \ldots, x_n) \propto p(\theta)p(x_1, x_2, x_3, x_4, \ldots, x_n | \theta)$$

$$\propto \left( \prod_{j=1}^{3} \theta_{j}^{\alpha_{j} - 1} \right) \left( \prod_{i=1}^{3} p(x_i | \theta) \right) \left( \prod_{j=1}^{3} \theta_{j}^{C_{j}^{obs}} \right)$$

$$\propto \left( \prod_{j=1}^{3} \theta_{j}^{C_{j}^{obs} + \alpha_{j} - 1} \right) \left( \prod_{i=1}^{3} p(x_i | \theta) \right),$$

where $C_{j}^{obs} = \sum_{i=4}^{n} I(x_i = j)$: observed counts for the $j$th category from $x_4$ to $x_n$.

1. Given $\theta = (\theta_1, \theta_2, \theta_3)$, $P(x_1 = j | \theta) \propto \theta_j$ for $j = 1, 2, 3$,

$$\Rightarrow P(x_1 = j | x_1 \neq 1, \theta) = \frac{\theta_j}{\theta_2 + \theta_3}, j = 2, 3.$$  

Similarly,

$$P(x_2 = j | x_2 \neq 2, \theta) = \frac{\theta_j}{\theta_1 + \theta_3}, j = 1, 3.$$  

$$P(x_3 = j | x_3 \neq 3, \theta) = \frac{\theta_j}{\theta_1 + \theta_2}, j = 1, 2.$$  

Draw $x_1, x_2, x_3$ independently according to the above conditional probabilities.

2. Given $(x_1, x_2, x_3)$, $C_{j}^{mis} = \sum_{i=1}^{3} I(x_i = j)$,

then $p(\theta | x_1, \ldots, x_n) \propto \prod_{j=1}^{3} \theta_{j}^{C_{j}^{Obs} + C_{j}^{mis} + \alpha_{j} - 1}$. Draw $\theta$ from

$$\theta | x \sim Dir(C_{1}^{obs} + C_{1}^{mis} + \alpha_1, C_{2}^{obs} + C_{2}^{mis} + \alpha_2, C_{3}^{obs} + C_{3}^{mis} + \alpha_3),$$

where $x = (x_1, \ldots, x_n)$ is complete data.

Iterate between steps 1 and 2 to generate $(\theta^{(t)}, x_{1,2,3}^{(t)})$ for $t = 1, \ldots, m$.

Bayesian estimates: $\hat{\theta}_B = \frac{1}{m} \sum_{t} \theta^{(t)}$ and histogram of $\theta_j^{(t)}$. 
3.3. Gaussian data example

Example 7. \( y_1, y_2, \ldots, y_n \overset{iid}{\sim} N_2(\mu, \Sigma) \), \( y_i = (y_{i1}, y_{i2}) \).

\[
\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma_{\text{known}} = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}.
\]

<table>
<thead>
<tr>
<th>( Y_1 )</th>
<th>( Y_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_1 )</td>
<td>?</td>
</tr>
<tr>
<td>( y_2 )</td>
<td>( \checkmark )</td>
</tr>
<tr>
<td>( y_3 )</td>
<td>( \checkmark )</td>
</tr>
<tr>
<td>( y_4 )</td>
<td>( \checkmark )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>( y_n )</td>
<td>( \checkmark )</td>
</tr>
</tbody>
</table>

?: missing value, \( \checkmark \): observed value.

Improper flat prior: \( p(\mu) \propto 1 \).

Missing data \( Y_{\text{mis}} = (y_{11}, y_{22}) \) and observed data \( Y_{\text{obs}} = (y_{12}, y_{21}, y_3, \ldots, y_n) \).

Data augmentation for this problem:

1. Given \( \mu \), sample \( y_{11} \) and \( y_{22} \), \( [y_{11}|y_{12}, \mu, \Sigma] \sim? \) Recall \( y_1 = (y_{11}, y_{12}) \).

\[
p(y_{11}|y_{12}, \mu, \Sigma) \propto p(y_{11}, y_{12}|\mu, \Sigma) \propto \exp \left[-\frac{1}{2}(y_{11} - \mu)^T \Sigma^{-1}(y_{11} - \mu)\right]
\]

\[
= \exp \left\{-\frac{1}{2(1-\rho^2)} \left[ \frac{(y_{11} - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(y_{11} - \mu_1)(y_{12} - \mu_2)}{\sigma_1 \sigma_2} + \frac{(y_{12} - \mu_2)^2}{\sigma_2^2} \right] \right\}
\]

\[
\times \exp \left\{-\frac{1}{2(1-\rho^2)\sigma_1^2} \left[ (y_{11} - \mu_1)^2 - \frac{2\rho \sigma_1}{\sigma_2} (y_{12} - \mu_2)(y_{11} - \mu_1) \right] \right\}
\]

\[
= \exp \left\{-\frac{1}{2(1-\rho^2)\sigma_1^2} \left[ y_{11} - \mu_1 - \frac{\rho \sigma_1}{\sigma_2} (y_{12} - \mu_2) \right]^2 + C \right\}.
\]

\( \therefore y_{11}|y_{12}, \mu, \Sigma \sim N \left( \mu_1 + \frac{\rho \sigma_1}{\sigma_2} (y_{12} - \mu_2), (1-\rho^2)\sigma_1^2 \right) \).

Similarly, \( y_{22}|y_{21}, \mu, \Sigma \sim N \left( \mu_2 + \frac{\rho \sigma_2}{\sigma_1} (y_{21} - \mu_1), (1-\rho^2)\sigma_2^2 \right) \).

Given \( \mu \), draw \( y_{11} \) and \( y_{22} \) independently from the two normal distributions.
2. Given $y_{11}$ and $y_{22}$, sample $\mu$?

$$p(\mu|y_1, y_2, \ldots, y_n, \Sigma) \propto p(y_1, \ldots, y_n|\mu, \Sigma)$$

$$= (2\pi\Sigma)^{-\frac{n}{2}} \exp \left( -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \right)$$

$$\propto \exp \left( -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \right).$$

Let $\bar{y} = \sum_i y_i/n$.

$$\sum_i (\mu - y_i)^T \Sigma^{-1} (\mu - y_i)$$

$$= \sum_i (\mu - \bar{y} + \bar{y} - y_i)^T \Sigma^{-1} (\mu - \bar{y} + \bar{y} - y_i)$$

$$= \sum_i \left[ (\mu - \bar{y})^T \Sigma^{-1} (\mu - \bar{y}) + 2(\mu - \bar{y})^T \Sigma^{-1} (\bar{y} - y_i) + (\bar{y} - y_i)^T \Sigma^{-1} (\bar{y} - y_i) \right]$$

$$= n(\mu - \bar{y})^T \Sigma^{-1} (\mu - \bar{y}) + C.$$

Therefore, $\mu|y_1, \ldots, y_n \sim \mathcal{N}_2(\bar{y}, \frac{1}{n} \Sigma)$.

Iterate between steps 1 and 2.