Module 8: Advanced Gibbs Sampling and Missing Data

Rebecca C. Steorts

Agenda

- Gibbs sampling (multi-stage sampler)
- Missing data application

Multi-stage Gibbs sampler

Assume d random variables, with joint pmf or pdf $p(v^1, \ldots, v^d)$.

At each iteration $(1, \ldots, M)$ of the algorithm, we sample from

$$
\begin{aligned}
v^1 &\mid v^2, v^3, \ldots, v^d \\
v^2 &\mid v^1, v^3, \ldots, v^d \\
&\vdots \\
v^d &\mid v^1, v^2, \ldots, v^{d-1}
\end{aligned}
$$

always using the most recent values of all the other variables.

The conditional distribution of a variable given all of the others is referred to as the *full conditional* in this context, and for brevity denoted $v^i \mid \cdots$.

Example: Censored data

In many real-world data sets, some of the data is either missing altogether or is partially obscured.

One way in which data can be partially obscured is by *censoring*, which means that we know a data point lies in some particular interval, but we do not observe it.

Medical data censoring

Suppose 6 patients participate in a cancer trial, however, patients 1, 2 and 4 leave the trial early.

Then we know when they leave the study, but we don't know information about them as the trial continues.

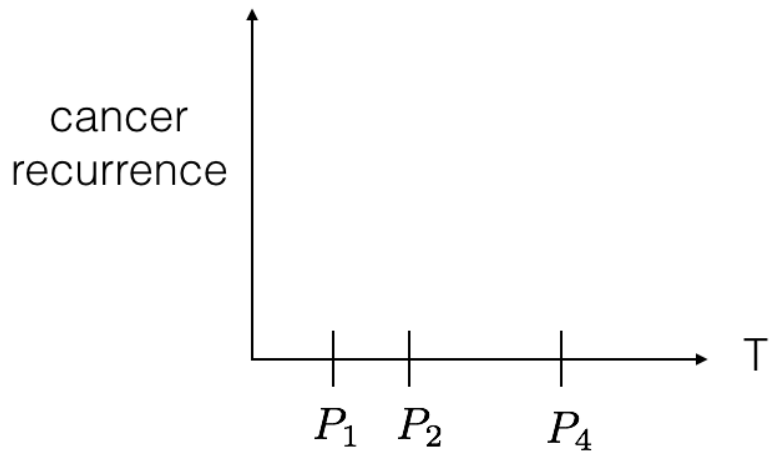

Figure 1: Example of censoring for medical data.

This is a certain type of missing data.

Heart Disease (Censoring) Example

- Researchers are studying the length of life (lifetime) following a particular medical intervention, such as a new surgical treatment for heart disease.
- The study consists of 12 patients.
- The number of years before death for each is

$$3.4, 2.9, 1.2+, 1.4, 3.2, 1.8, 4.6, 1.7+, 2.0+, 1.4+, 2.8, 0.6+$$

where the $+$ indicates that the patient was alive after x years, but the researchers lost contact with the patient after that point in time.

Model

$$X_i = \begin{cases} Z_i & \text{if } Z_i \leq c_i \\ c_i & \text{if } Z_i > c_i \end{cases} \tag{1}$$

$$Z_1, \ldots, Z_n | \theta \overset{iid}{\sim} \text{Gamma}(r, \theta) \tag{2}$$

$$\theta \sim \text{Gamma}(a, b) \tag{3}$$

where a, b, and r are known.

- c_i is the censoring time for patient i, which is fixed, but known only if censoring occurs.
- X_i is the observation
  - if the lifetime is less than c_i then we get to observe it ($X_i = Z_i$),
  - otherwise all we know is the lifetime is greater than c_i ($X_i = c_i$).
- θ is the parameter of interest—the rate parameter for the lifetime distribution.
- Z_i is the lifetime for patient i, however, this is not directly observed.

Posterior inference

Goal: find $p(\theta, z_{1:n}|x_{1:n})$?

1. Straightforward approaches that are in closed form do not work (think about these on your own). Instead we turn to Gibbs!

2. To sample from $p(\theta, z_{1:n}|x_{1:n})$, we cycle through each of the full conditional distributions,

$$
\begin{aligned}
\theta \mid z_{1:n}, x_{1:n} \\
z_1 \mid \theta, z_{2:n}, x_{1:n} \\
z_2 \mid \theta, z_1, z_{3:n}, x_{1:n} \\
\vdots \\
z_n \mid \theta, z_{1:n-1}, x_{1:n}
\end{aligned}
$$

sampling from each in turn, always conditioning on the most recent values of the other variables.

Likelihood

Recall the model is:

$$X_i = \begin{cases} Z_i & \text{if } Z_i \leq c_i \\ c_i & \text{if } Z_i > c_i \end{cases} \tag{4}$$

$$Z_1, \ldots, Z_n | \theta \overset{iid}{\sim} \text{Gamma}(r, \theta) \tag{5}$$

$$\theta \sim \text{Gamma}(a, b) \tag{6}$$

The pdf associated with this random variable is rather strange, as it consists of two point masses: one at Z_i and one at c_i. The formula is

$$p(x_i|z_i) = \mathbf{1}(x_i = z_i)\mathbf{1}(z_i \leq c_i) + \mathbf{1}(x_i = c_i)\mathbf{1}(z_i > c_i).$$

.

Full conditionals

The full conditionals are easy to calculate. Let's start with $\theta|\cdots$

- Since $\theta \perp x_{1:n} \mid z_{1:n}$ (i.e., θ is conditionally independent of $x_{1:n}$ given $z_{1:n}$),

$$p(\theta|\cdots) = p(\theta|z_{1:n}, x_{1:n}) = p(\theta|z_{1:n}) \tag{7}$$

$$= \text{Gamma}\left(\theta \,\middle|\, a + nr,\ b + \textstyle\sum_{i=1}^{n} z_i\right) \tag{8}$$

  using the fact that the prior on θ is conjugate.

Full conditionals

Now we can easily find the full conditionals.

- Note that z_i is conditionally independent of z_j given θ for $i \neq j$.
- This implies that x_i is conditionally independent of x_j given z_i for $i \neq j$.

Now we have

$$\begin{aligned} p(z_i|z_{-i}, x_{1:n}, \theta) &= p(z_i|x_i, \theta) \\ &\underset{z_i}{\propto} p(z_i, x_i, \theta) \\ &= p(\theta)p(z_i|\theta)p(x_i|z_i, \theta) \\ &\underset{z_i}{\propto} p(z_i|\theta)p(x_i|z_i, \theta) \\ &= p(z_i|\theta)p(x_i|z_i). \end{aligned}$$

Full conditionals (continued)

There are now two cases to consider.

1. If $x_i \neq c_i$, then $p(z_i|\theta)p(x_i|z_i)$ is only non-zero when $z_i = x_i$.

   - The density devolves to a point mass at x_i.

2. If $x_i = c_i$, then the density becomes $p(x_i|z_i) = \mathbf{1}(z_i > c_i)$, so

$$p(z_i|\ldots) \propto p(z_i|\theta)\mathbf{1}(z_i > c_i),$$

   which is a truncated Gamma.

Sampling from the truncated Gamma

We sample from the truncated gamma using a modified version of the inverse CDF method.

For the censored values of Z_i we know c_i.

If we know θ (which we will in a Gibbs' sampler), we know the distribution of $Z_i|\theta \sim Gamma(r, \theta)$.

Let F be the CDF of this distribution.

Suppose we truncate this distribution to (c, ∞). The new CDF is

$$P(Z_i < z) = \frac{F(z) - F(c)}{1 - F(c)}.$$

Therefore Y is a sample from the truncated Gamma.

Remark: when we implement the GS, we don't sample the observed values. We impute the censored values using the method just outlined.

Homework 8

1. Code up your own multi-stage GS in R. Be sure to use efficient functions. (You will have assistance on this in lab next week).
2. Use the censored data

$$3.4, 2.9, 1.2+, 1.4, 3.2, 1.8, 4.6, 1.7+, 2.0+, 1.4+, 2.8, 0.6+$$

. Specifically, give (a) give traceplots of all unknown paramaters from the G.S. (b) a running average plot, (c) the estimated density of $\theta \mid \cdots$ and $z_9 \mid \cdots$. Be sure to give brief explanations of your results and findings.