The Multi Stage Gibbs Sampling: Data Augmentation Dutch Example

Rebecca C. Steorts
Bayesian Methods and Modern Statistics: STA 360/601

Module 8

Example: Data augmentation / Auxiliary variables

- A commonly-used technique for designing MCMC samplers is to use *data augmentation*, also known as *auxiliary variables*.
- Introduce variable(s) Z that depends on the distribution of the existing variables in such a way that the resulting conditional distributions, with Z included, are easier to sample from and/or result in better mixing.
- Z's are latent/hidden variables that are introduced for the purpose of simplifying/improving the sampler.

Idea: Create Z's and throw them away at the end!

- Suppose we want to sample from $p(x, y)$, but $p(x|y)$ and/or $p(y|x)$ are complicated.

- Choose

$$p(z|x, y)$$

such that $p(x|y, z)$, $p(y|x, z)$, and $p(z|x, y)$ are easy to sample from.

- Then construct a Gibbs sampler to sample all three variables (X, Y, Z) from $p(x, y, z)$.

- Then we just throw away the Z's and we will have samples (X, Y) from $p(x, y)$.

Dutch Example

Consider a data set on the heights of 695 Dutch women and 562 Dutch men.

Suppose we have the list of heights, but we don't know which data points are from women and which are from men.

From Figure 1 can we still infer the distribution of female heights
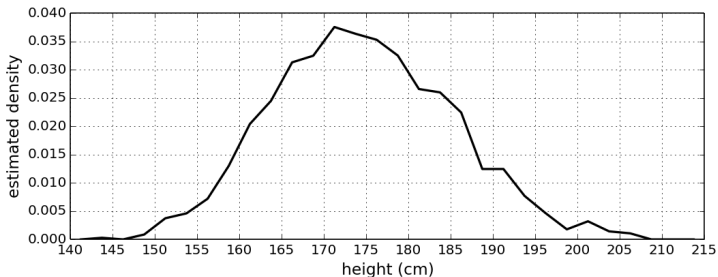and male heights?

Figure 1: Heights of Dutch women and men, combined.

From Figure 1 can we still infer the distribution of female heights
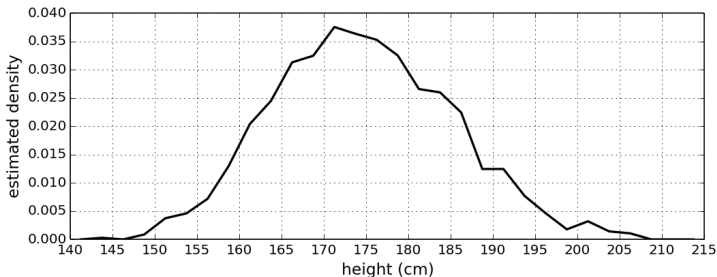and male heights?

Figure 1: Heights of Dutch women and men, combined.

Surprisingly, the answer is yes!

What's the magic trick?

The reason is that this is a two-component mixture of Normals, and there is an (essentially) unique set of mixture parameters corresponding to any such distribution.

We'll get to details soon. Be patient!

Constructing a Gibbs sampler

To construct a Gibbs sampler for this situation:

- Common to introduce an auxiliary variable Z_i for each datapoint, indicating which mixture component it is drawn from.
- In this example, Z_i indicates whether subject i is female or male.
- This results in a Gibbs sampler that is easy to derive/implement.

Two component Mixture Model

Let's assume that both mixture components (female and male) have the same precision, say λ, and that λ is fixed and known.

Then the usual two-component Normal mixture model is:

$$X_1, \ldots, X_n | \mu, \pi \overset{\text{iid}}{\sim} F(\mu, \pi) \tag{1}$$

$$\mu := (\mu_0, \mu_1) \overset{\text{iid}}{\sim} \mathcal{N}(m, \ell^{-1}) \tag{2}$$

$$\pi \sim \text{Beta}(a, b) \tag{3}$$

where $F(\mu, \pi)$ is the distribution with p.d.f.

$$f(x | \mu, \pi) = (1 - \pi) \mathcal{N}(x \mid \mu_0, \lambda^{-1}) + \pi \mathcal{N}(x \mid \mu_1, \lambda^{-1})$$

and $\mu = (\mu_0, \mu_1)$.

The likelihood is

$$p(x_{1:n}|\mu, \pi) = \prod_{i=1}^{n} f(x_i|\mu, \pi)$$
$$= \prod_{i=1}^{n} \left[(1 - \pi)\mathcal{N}(x_i \mid \mu_0, \lambda^{-1}) + \pi\mathcal{N}(x_i \mid \mu_1, \lambda^{-1}) \right]$$

which is a complicated function of μ and π, making the posterior difficult to sample from directly.

Latent allocation variables to the rescue!

Define an equivalent model that includes latent "allocation" variables $Z_1, \ldots, Z_n$

These indicate which mixture component each data point comes from–that is, Z_i indicates whether subject i is female or male.

$$X_i \sim \mathcal{N}(\mu_{Z_i}, \lambda^{-1}) \text{ independently for } i = 1, \ldots, n. \quad (4)$$

$$Z_1, \ldots, Z_n | \mu, \pi \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi) \quad (5)$$

$$\mu = (\mu_0, \mu_1) \overset{\text{iid}}{\sim} \mathcal{N}(m, \ell^{-1}) \quad (6)$$

$$\pi \sim \text{Beta}(a, b) \quad (7)$$

Recall

$$X_i \sim \mathcal{N}(\mu_{Z_i}, \lambda^{-1}) \text{ independently for } i = 1, \ldots, n.$$

$$Z_1, \ldots, Z_n | \mu, \pi \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi)$$

$$\mu = (\mu_0, \mu_1) \overset{\text{iid}}{\sim} \mathcal{N}(m, \ell^{-1})$$

$$\pi \sim \text{Beta}(a, b)$$

This is equivalent to the model above, since

$$p(x_i | \mu, \pi) \tag{8}$$

$$= p(x | Z_i = 0, \mu, \pi) \mathbb{P}(Z_i = 0 | \mu, \pi) + p(x | Z_i = 1, \mu, \pi) \mathbb{P}(Z_i = 1 | \mu, \pi) \tag{9}$$

$$= (1 - \pi) \mathcal{N}(x_i | \mu_0, \lambda^{-1}) + \pi \mathcal{N}(x_i | \mu_1, \lambda^{-1}) \tag{10}$$

$$= f(x_i | \mu, \pi), \tag{11}$$

and thus it induces the same distribution on $(x_{1:n}, \mu, \pi)$. The latent model is considerably easier to work with, particularly for Gibbs sampling.

Full conditionals

Derive the full conditionals below as an exercise.

Denote $x = x_{1:n}$ and $z = z_{1:n}$.

- Find $(\pi | \cdots)$
- Find $(\mu | \cdots)$
- Find $(z | \cdots)$

As usual, each iteration of Gibbs sampling proceeds by sampling from each of these conditional distributions, in turn.

They are in this module, so you can check your results after completion.

Full conditionals derived

- $(\pi | \cdots)$ Given z, π is independent of everything else, so this reduces to a Beta–Bernoulli model, and we have

$$p(\pi|\mu, z, x) = p(\pi|z) = \text{Beta}(\pi \mid a + n_1,\, b + n_0)$$

where $n_k = \sum_{i=1}^{n} \mathbb{1}(z_i = k)$ for $k \in \{0, 1\}$.

- $(\mu|\cdots)$ Given z, we know which component each data point comes from.

  The model (conditionally on z) is just two independent Normal–Normal models, as we have seen before:

  $$\boldsymbol{\mu}_0|\mu_1, x, z, \pi \sim \mathcal{N}(M_0, L_0^{-1})$$
  $$\boldsymbol{\mu}_1|\mu_0, x, z, \pi \sim \mathcal{N}(M_1, L_1^{-1})$$

  where for $k \in \{0, 1\}$,

  $$n_k = \sum_{i=1}^{n} \mathbb{1}(z_i = k)$$
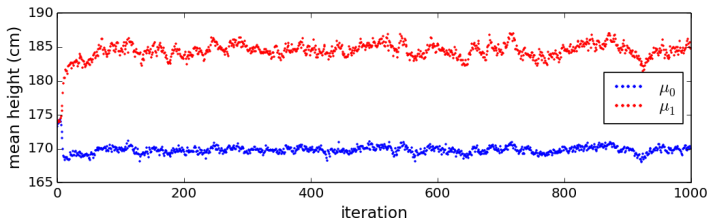 $$L_k = \ell + n_k \lambda$$
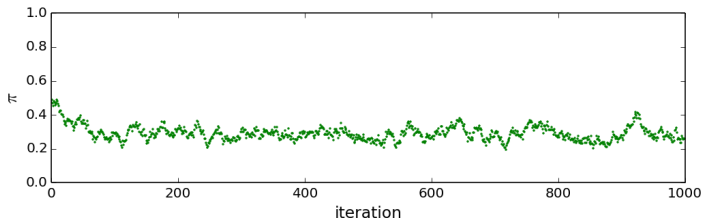 $$M_k = \frac{\ell m + \lambda \sum_{i:z_i=k} x_i}{\ell + n_k \lambda}.$$

- $(z|\cdots)$

$$
\begin{aligned}
p(z|\mu, \pi, x) &\underset{z}{\propto} p(x, z, \pi, \mu) \underset{z}{\propto} p(x|z, \mu)p(z|\pi) \\
&= \prod_{i=1}^{n} \mathcal{N}(x_i|\mu_{z_i}, \lambda^{-1}) \, \text{Bernoulli}(z_i|\pi) \\
&= \prod_{i=1}^{n} \left(\pi \mathcal{N}(x_i|\mu_1, \lambda^{-1}) \right)^{z_i} \left((1-\pi)\mathcal{N}(x_i|\mu_0, \lambda^{-1}) \right)^{1-z_i} \\
&= \prod_{i=1}^{n} \alpha_{i,1}^{z_i} \alpha_{i,0}^{1-z_i} \\
&\underset{z}{\propto} \prod_{i=1}^{n} \text{Bernoulli}(z_i \mid \alpha_{i,1}/(\alpha_{i,0} + \alpha_{i,1}))
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha_{i,0} &= (1-\pi)\mathcal{N}(x_i|\mu_0, \lambda^{-1}) \\
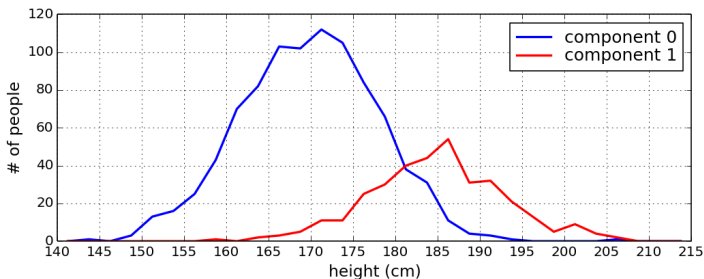\alpha_{i,1} &= \pi \mathcal{N}(x_i|\mu_1, \lambda^{-1}).
\end{aligned}
$$

15

(a) Traceplots of the component means, μ_0 and μ_1.

(b) Traceplot of the mixture weight, π (prior probability that a subject comes from component 1).

Figure 2: Results from one run of the mixture example.

(a) Histograms of the heights of subjects assigned to each component, according to $z_1, \ldots, z_n$, in a typical sample.

Figure 3: Results from one run of the mixture example. One way of visualizing the allocation/assignment variables $z_1, \ldots, z_n$ is to make histograms of the heights of the subjects assigned to each component.

My Factory Settings!

- $\lambda = 1/\sigma^2$ where $\sigma = 8$ cm (≈ 3.1 inches) ($\sigma =$ standard deviation of the subject heights within each component)

- $a = 1$, $b = 1$ (Beta parameters, equivalent to prior "sample size" of 1 for each component)

- $m = 175$ cm (≈ 68.9 inches) (mean of the prior on the component means)

- $\ell = 1/s^2$ where $s = 15$ cm (≈ 6 inches) ($s =$ standard deviation of the prior on the component means)
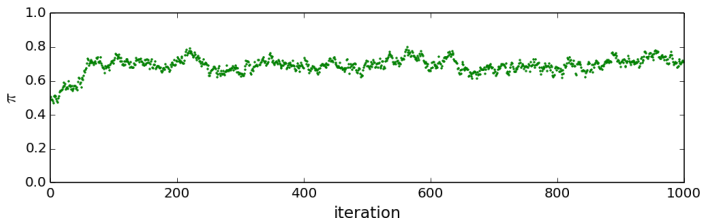
My Factory Settings!

We initialize the sampler at:

- $\pi = 1/2$ (equal probability for each component)
- $z_1, \ldots, z_n$ sampled i.i.d. from $\text{Bernoulli}(1/2)$ (initial assignment to components chosen uniformly at random)
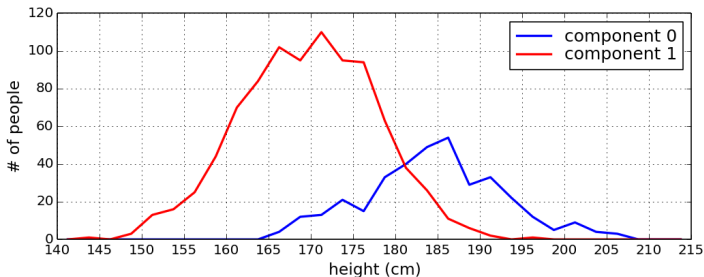- $\mu_0 = \mu_1 = m$ (component means initialized to the mean of their prior)
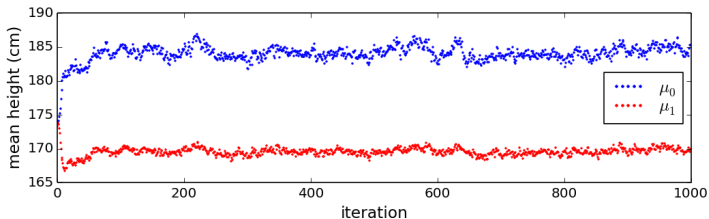
(a) Traceplots of the component means, μ_0 and μ_1.

(b) Traceplot of the mixture weight, π (prior probability that a subject comes from component 1).

Figure 4: Results from another run of the mixture example.

(a) Histograms of the heights of subjects assigned to each component, according to $z_1, \ldots, z_n$, in a typical sample.

Figure 5: Results from another run of the mixture example.

Caution – watch out for modes

Example illustrates a big thing that can go wrong with MCMC (although fortunately, in this case, the results are still valid if interpreted correctly).

- ▶ Why are females assigned to component 0 and males assigned to component 1? Why not the other way around?
- ▶ In fact, the model is symmetric with respect to the two components, and thus the posterior is also symmetric.
- ▶ If we run the sampler multiple times (starting from the same initial values), sometimes it will settle on females as 0 and males as 1, and sometimes on females as 1 and males as 0 — see Figure 5.
- ▶ Roughly speaking, the posterior has two modes.
- ▶ If the sampler were behaving properly, it would move back and forth between these two modes.
- ▶ But it doesn't—it gets stuck in one and stays there.

This is a very common problem with mixture models. Fortunately, however, in the case of mixture models, the results are still valid if we interpret them correctly. Specifically, our inferences will be valid as long as we only consider quantities that are invariant with respect to permutations of the components (e.g. symmetry about the mean).