Module 7: Part III: Gibbs Sampling and Data Augmentation

Rebecca C. Steorts

Agenda

- Two Component Mixture Model
- Latent Variable Allocation (Trick for Gibbs Sampling)
- Application to the Dutch Example

Goal

The goal of this module is to introduce **mixture models**, which are commonly used in applications to networks analyses, cleaning data (entity resolution), neuroscience, and many other real case studies.

We will do this using a **latent variable**.

Structure of Module

Then we will utilize concepts that we have learned throughout the semester which include:

- a case study to Dutch heights
- hierarchical modeling
- semi-conjugacy
- Gibbs sampling
- MCMC diagnostics
- approximating posterior distributions and explaning our results
- performing sensitivity analyses

Background (Recall)

A **latent variable** is the true version of the state of a random variable that is unknown and not directly observed.

Latent Variable or Data Augmentation Approach

A commonly-used technique for designing MCMC samplers is to use *data augmentation*, which involves introducing a latent variable into the model.

- Introduce variable(s) Z that depend 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/auxiliary/hidden variables that are introduced for the purpose of simplifying/improving the sampler.

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

Goal: Sample from $p(x, y)$

Problem: We cannot sample from $p(x|y)$ and/or $p(y|x)$.

Solution:

1. Introduce a latent/hidden variable Z such that

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

   are easy to sample from.

2. Then construct a Gibbs sampler that will approximate samples from $p(x, y, z)$ using the full conditionals above.

3. The output of the three-stage Gibbs sampler is (X, Y, Z).

But, our main interest is sampling from $p(x, y)$. Thus, 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.

Dutch Example

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

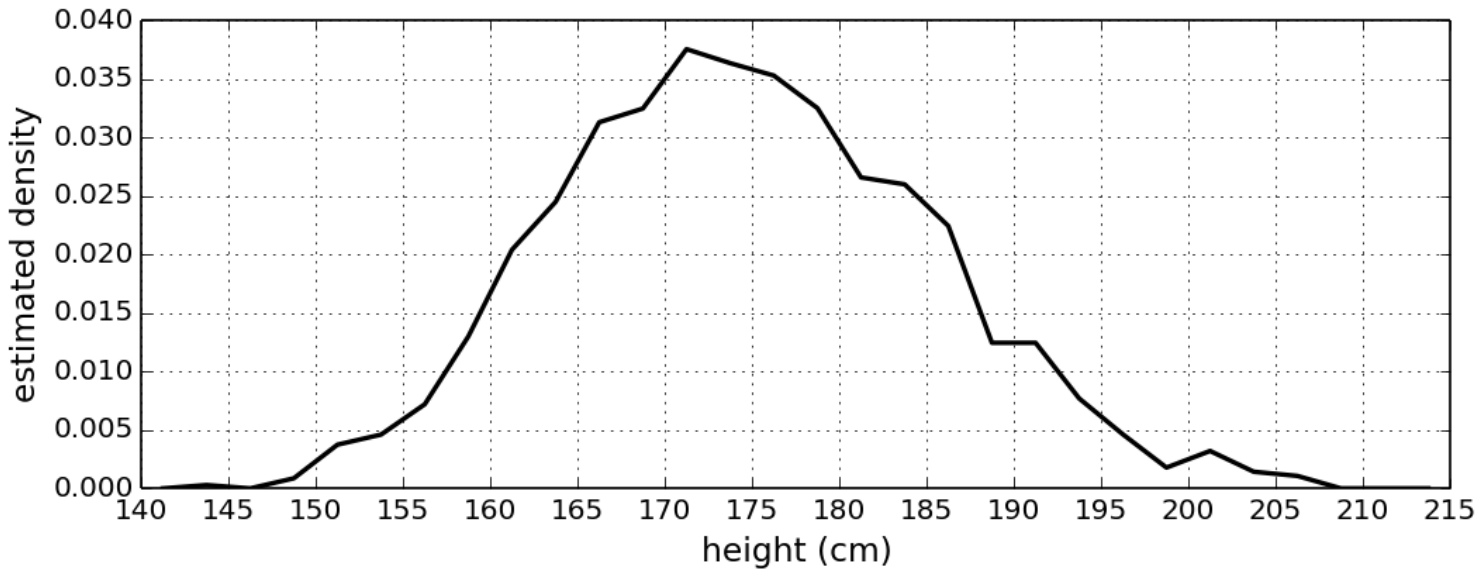


Figure 1: Heights of Dutch women and men, combined.

Surprisingly, the answer is yes!

Dutch example

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 data point, 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

Assume that both mixture components (female and male) have the same precision, λ, which is fixed and known. Assume below that a, b, m, ℓ are known. Let π be the prior probability that a subject comes from component 1.

Then the two-component Normal mixture model is:

$$X_1, \ldots, X_n \mid \mu, \pi \sim F(\mu, \pi) \tag{1}$$

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

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

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

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

Likelihood

The likelihood is

$$
\begin{aligned}
p(x_{1:n}|\mu,\pi) &= \prod_{i=1}^{n} f(x_i|\mu,\pi) \\
&= \prod_{i=1}^{n} \Big[(1-\pi)\mathcal{N}(x_i \mid \mu_0, \lambda^{-1}) + \pi\mathcal{N}(x_i \mid \mu_1, \lambda^{-1})\Big].
\end{aligned}
$$

Likelihood

What do you notice about the likelihood function?

$$
\begin{aligned}
p(x_{1:n}|\mu,\pi) &= \prod_{i=1}^{n} f(x_i|\mu,\pi) \\
&= \prod_{i=1}^{n} \Big[(1-\pi)\mathcal{N}(x_i \mid \mu_0, \lambda^{-1}) + \pi\mathcal{N}(x_i \mid \mu_1, \lambda^{-1})\Big].
\end{aligned}
$$

Likelihood

The **likelihood** is very complicated function of μ and π.

This makes the posterior difficult to sample from directly.

Thus, we will rewrite the likelihood using **latent variables**.

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.

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

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

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

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

How can we check that the latent allocation model is equivalent to our original model? This is left as a practice exercise for Exam II. (Solution is at the end of the module.)

Full conditionals

Recall

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

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

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

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

- $(\pi | \cdots)$ 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 $k \in \{0, 1\}$.

That is, $n_o = \sum_{i=1}^{n} \mathbb{1}(z_i = 0)$; $n_1 = \sum_{i=1}^{n} \mathbb{1}(z_i = 1)$.

Full conditionals

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

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

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

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

- $(\mu | \cdots)$ 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:

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

$$\boldsymbol{\mu}_1 | \mu_0, x, z, \pi \sim \mathcal{N}(M_1, L_1^{-1}), \text{where}$$

$$L_0 = \ell + n_0\lambda; L_1 = \ell + n_1\lambda$$

$$M_0 = \frac{\ell m + \lambda \sum_{i:z_i=0} x_i}{\ell + n_0\lambda}.$$

Full conditionals

- $(z|\cdots)$

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

where

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

[1] Not multiplication here due to dependence of z_i

My Factory Settings!

We initialize the sampler at the same settings that we did when we looked at this application before. Let's review them below.

- $\lambda = 1/\sigma^2$ where $\sigma = 8$ cm (≈ 3.1 inches) (σ = 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 the same settings that we did when we looked at this application before. Let's review them below.

- $\pi = 1/2$ (equal probability for each component)
- $z_1, \ldots, z_n$ sampled i.i.d. from 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)

Traceplots of the component means

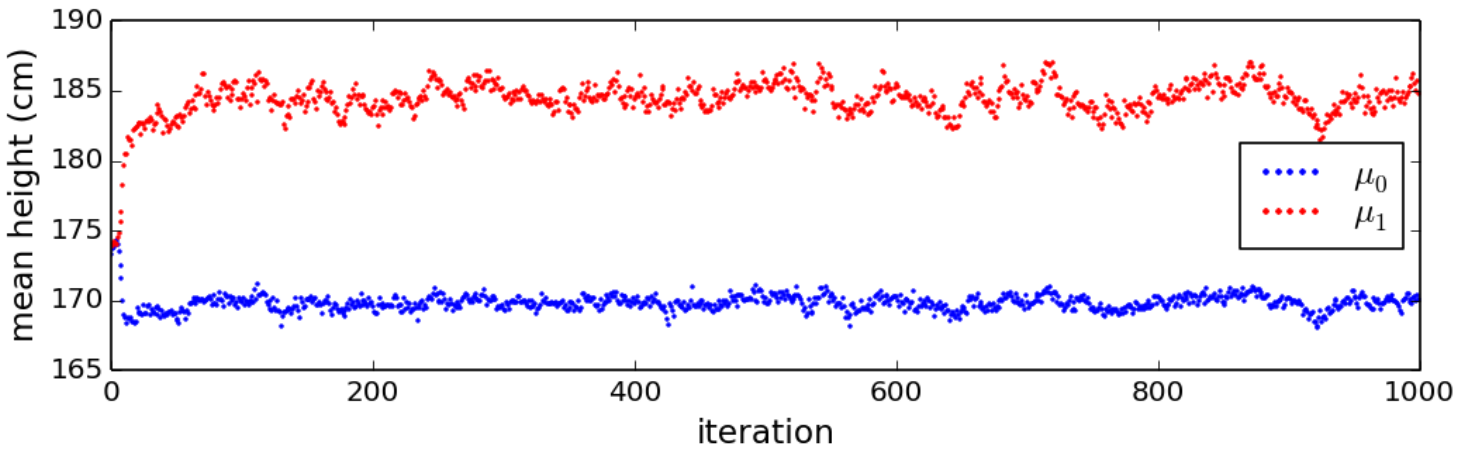


Figure 2: Traceplots of the component means, μ_0 and μ_1. The blue component quickly settles to have a mean of around 168–170 cm and the other (red) to a mean of around 182–186 cm.

Recall we're not using the true assignments of subjects to components (we don't know whether they are male or female). Note that this is fairly close to the sample averages: 168.0 cm (5 feet 6.1 inches) for females, and 181.4 cm (5 feet 11.4 inches) for males.

Traceplots of the mixture weight, π

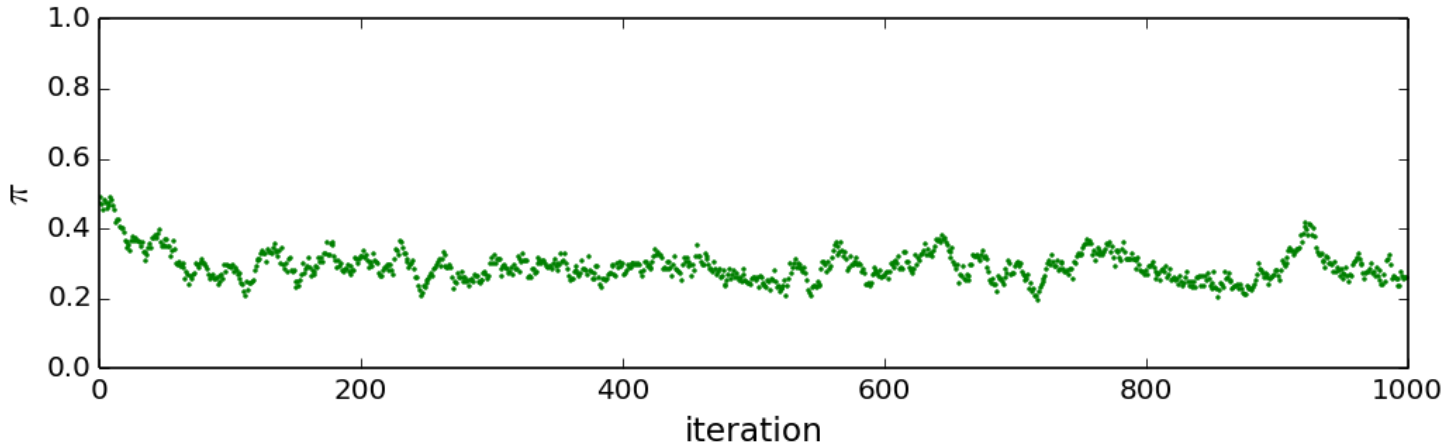


Figure 3: Traceplots of the mixture weight, π (prior probability that a subject comes from component 1).

The traceplot of π indicates that the sampler is exploring values of around 0.2 to 0.4—that is, the proportion of people coming from group 1 is around 0.2 to 0.4.

Looking at the actual labels (female and male), the empirical proportion of males is $562/(695 + 562) \approx 0.45$, so this is slight off.

How might we fix this in our model?

Comceptual Question

Recall

$$X_i \mid \mu, Z \sim \mathcal{N}(\mu_{Z_i}, \lambda^{-1}) \text{ independently for } i = 1, \ldots, n.$$
$$Z_1, \ldots, Z_n | \mu, \pi \overset{iid}{\sim} \text{Bernoulli}(\pi)$$
$$\mu = (\mu_0, \mu_1) \overset{iid}{\sim} \mathcal{N}(m, \ell^{-1})$$
$$\pi \sim \text{Beta}(a, b),$$

where a, b, m, ℓ are fixed.

Hint: Is it reasonable to think that we should have the same λ for the data? Consider

$$\lambda = (\lambda_0, \lambda_1)$$

How would you proceed next in a Bayesian hierarchical manner?

Histograms

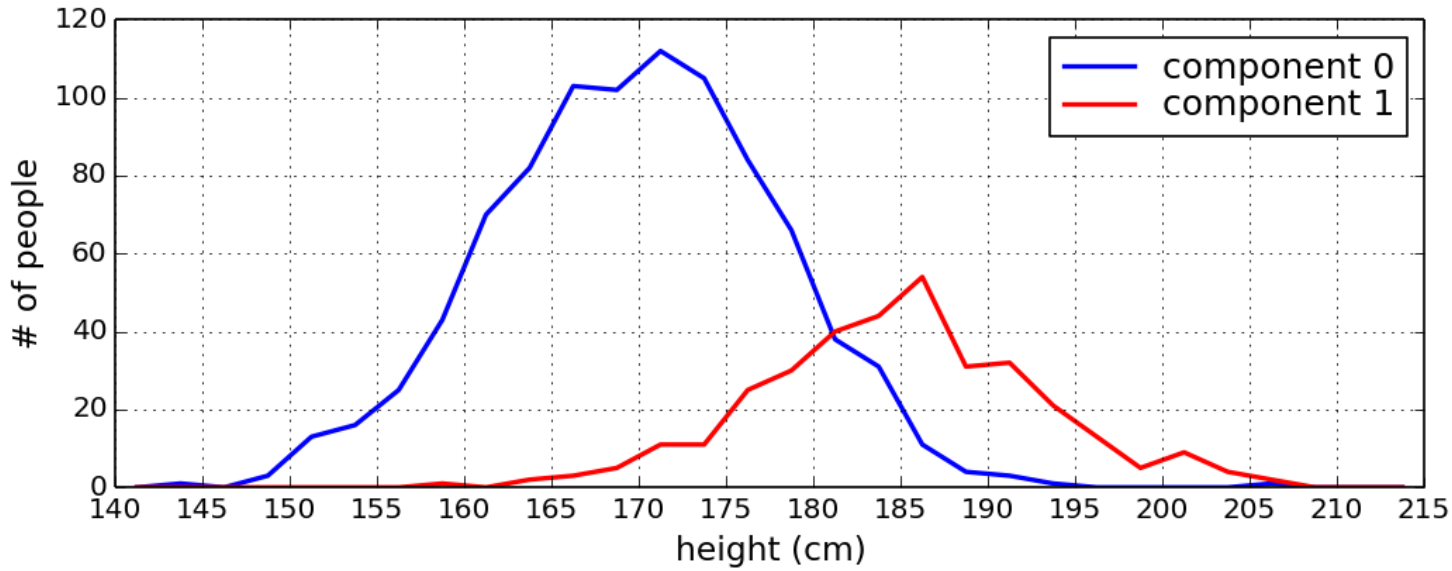


Figure 4: Histograms of the heights of subjects assigned to each component, according to $z_1, \ldots, z_n$, in a typical sample.

Question

Why are females assigned to component 0 and males assigned to component 1? Why not the other way around? Let's investigate this and see what happens if we look at another choice of starting values.

Traceplots of the component means

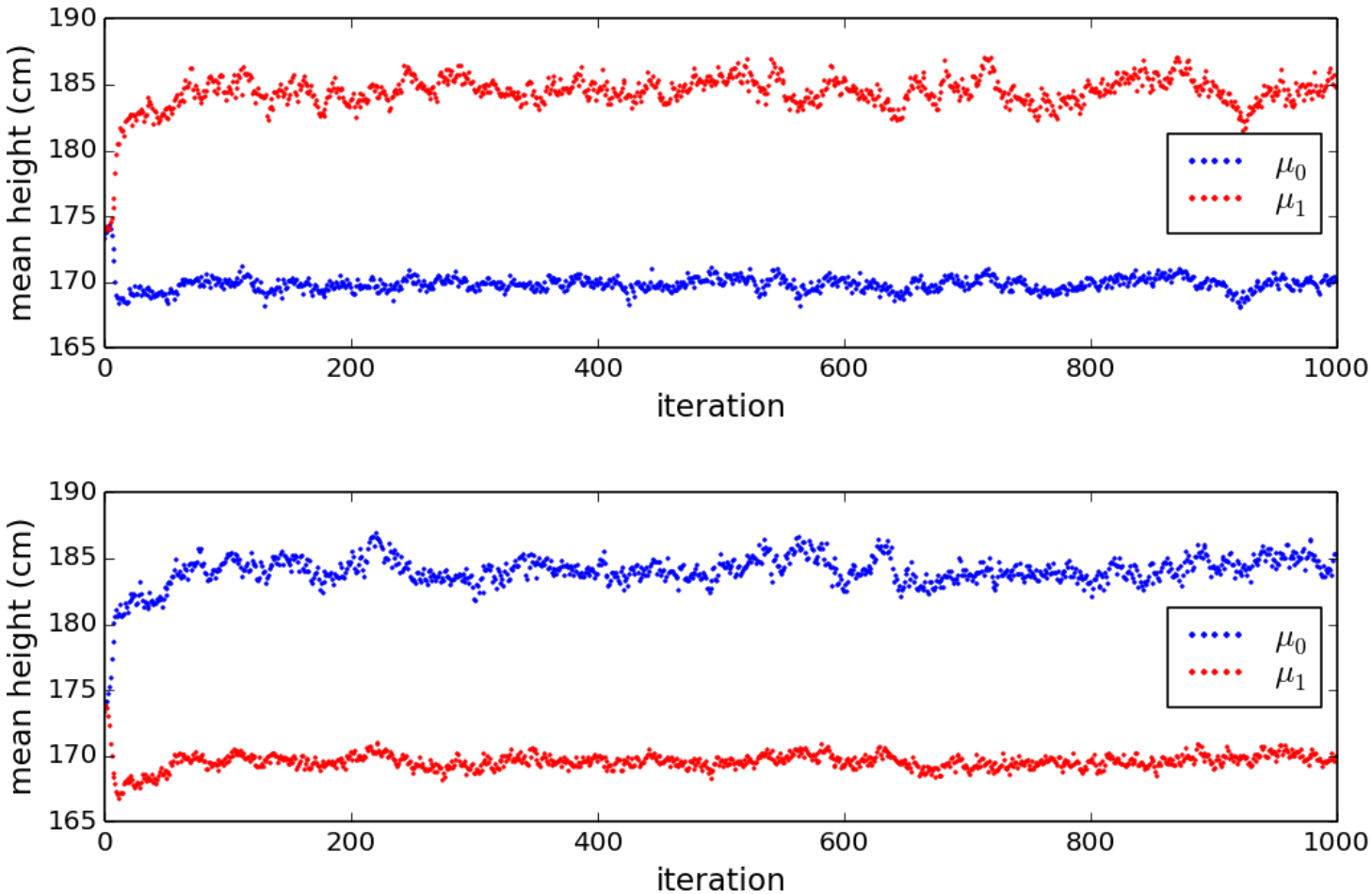


Figure 5: Traceplots of the component means, μ_0 and μ_1.

Traceplots of the mixture weight, π

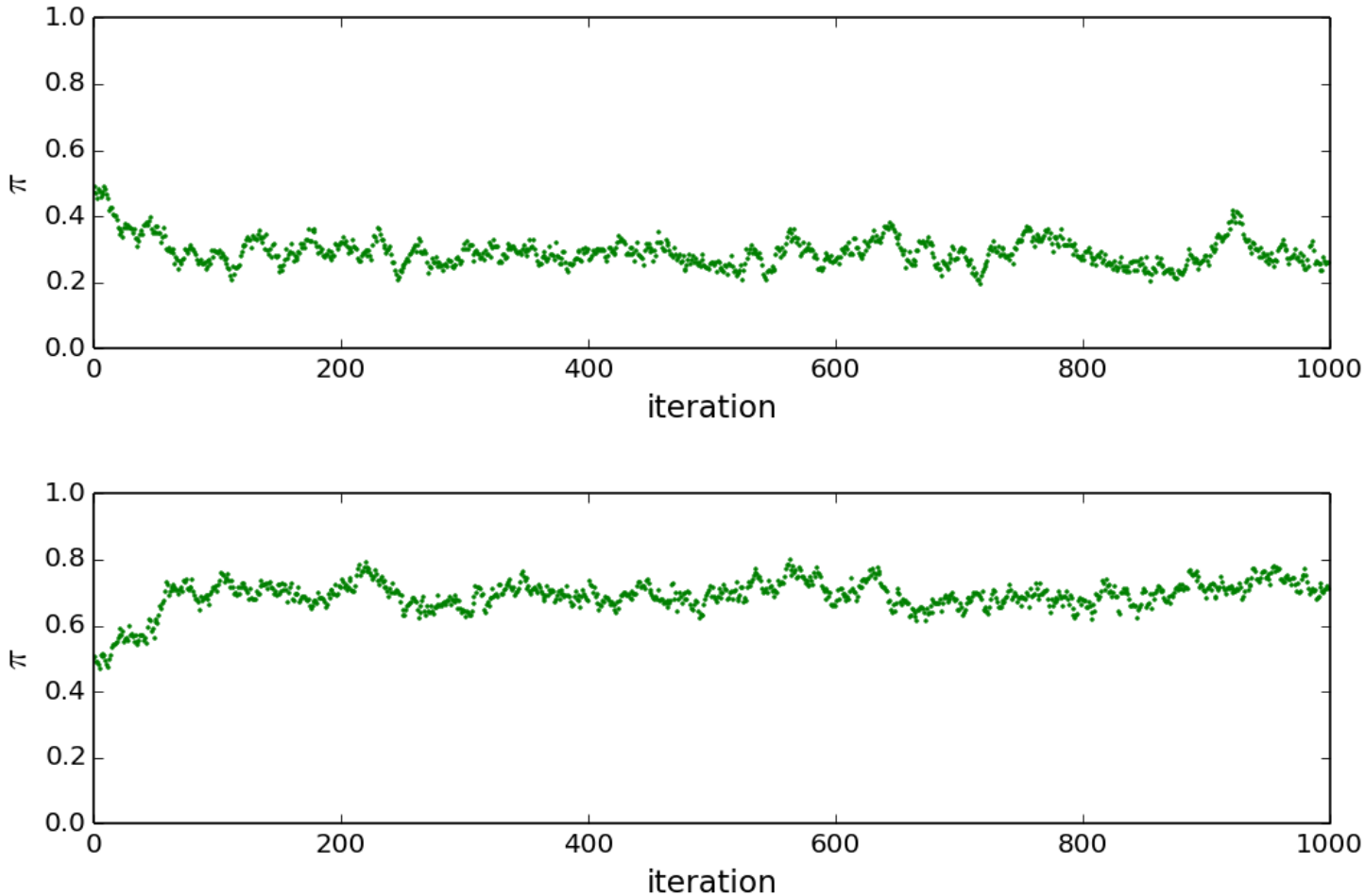
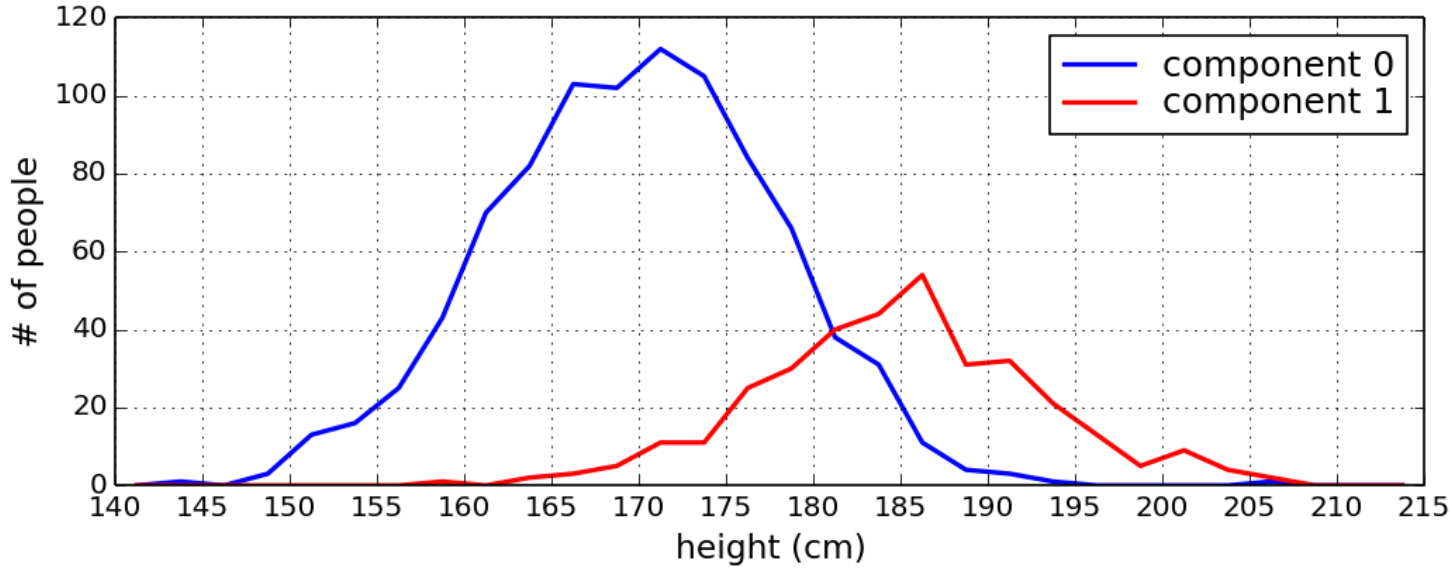
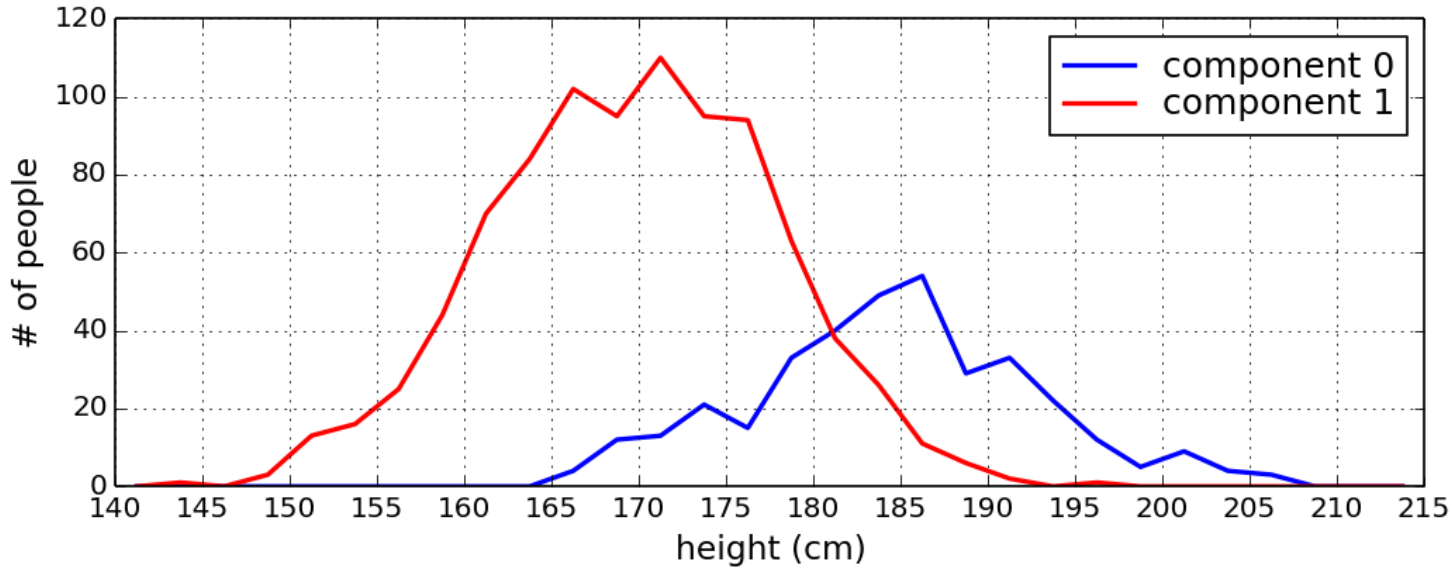


Figure 6: Traceplots of the mixture weight, π (prior probability that a subject comes from component 1).

Histograms


component 0
component 1
# of people
height (cm)
0
20
40
60
80
100
120
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215


component 0
component 1
# of people
height (cm)
0
20
40
60
80
100
120
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215

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 6.
- 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.

Detailed Takeaways

- Latent variable
- Mixture models
- Model for Dutch case study
- Why we need to use a latent variable
- Deriving conditional distributions
- Understanding the Gibbs sampler updates
- Understanding conceptual questions that relate to the case study
- What are the main take aways?
- What are benefits of this model? How could you improve the model?

Takeaways from Dutch Application

- 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, others).
- To learn more about invariance and what quantities remain unchanged, see Theory of Point Estimation, Lehmann and Casella, Chapter 1 for a formal treatment. (For more advanced reading, see Chapter 3.) Feel free to see me in OH if you want to chat about this!

Equivalence of both models (Practice Exercise for Exam II)

Recall

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

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

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

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

This is equivalent to the model above, since

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

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

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

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

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.