Module 8: Part III: Gibbs Sampling and Data Augmentation

Rebecca C. Steorts

Agenda

- Data Augmentation
- Dutch Example
- Other example

Data augmentation for 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.

Dutch Example

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

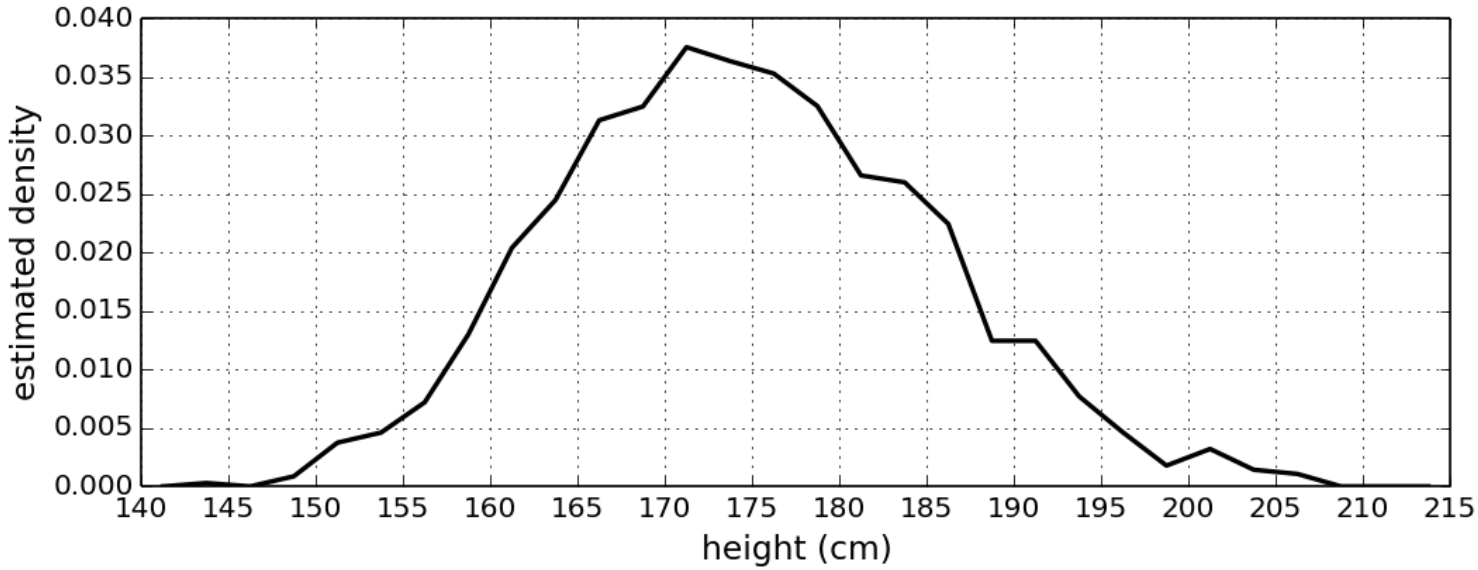


Figure 1: Heights of Dutch women and men, combined.

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

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:

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

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

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

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

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

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

Likelihood

The likelihood is

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

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.

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

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

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

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

Latent allocation variables

Recall

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

This is equivalent to the model above, since

$$p(x_i|\mu, \pi) \tag{8}$$
$$= p(x|Z_i = 0, \mu, \pi)\mathbb{P}(Z_i = 0|\mu, \pi) + p(x|Z_i = 1, \mu, \pi)\mathbb{P}(Z_i = 1|\mu, \pi) \tag{9}$$
$$= (1 - \pi)\mathcal{N}(x_i|\mu_0, \lambda^{-1}) + \pi\mathcal{N}(x_i|\mu_1, \lambda^{-1}) \tag{10}$$
$$= f(x_i|\mu, \pi), \tag{11}$$

Full conditionals

Recall

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

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

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

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

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

Full conditionals

Recall

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

- $(\mu | \cdots)$ 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:

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

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

Full conditionals

- $(z|\cdots)$

$$p(z|\mu,\pi,x) \underset{z}{\propto} p(x,z,\pi,\mu) \underset{z}{\propto} p(x|z,\mu)p(z|\pi)$$

$$= \prod_{i=1}^{n} \mathcal{N}(x_i|\mu_{z_i},\lambda^{-1}) \,\text{Bernoulli}(z_i|\pi)$$

$$= \prod_{i=1}^{n} \Big(\pi\mathcal{N}(x_i|\mu_1,\lambda^{-1})\Big)^{z_i} \Big((1-\pi)\mathcal{N}(x_i|\mu_0,\lambda^{-1})\Big)^{1-z_i}$$

$$= \prod_{i=1}^{n} \alpha_{i,1}^{z_i}\alpha_{i,0}^{1-z_i}$$

$$\underset{z}{\propto} \prod_{i=1}^{n} \text{Bernoulli}(z_i \mid \alpha_{i,1}/(\alpha_{i,0}+\alpha_{i,1}))$$

where

$$\alpha_{i,0} = (1-\pi)\mathcal{N}(x_i|\mu_0,\lambda^{-1})$$

$$\alpha_{i,1} = \pi\mathcal{N}(x_i|\mu_1,\lambda^{-1}).$$

My Factory Settings!

- $\lambda = 1/\sigma^2$ 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:

- $\pi = 1/2$ (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)

Results

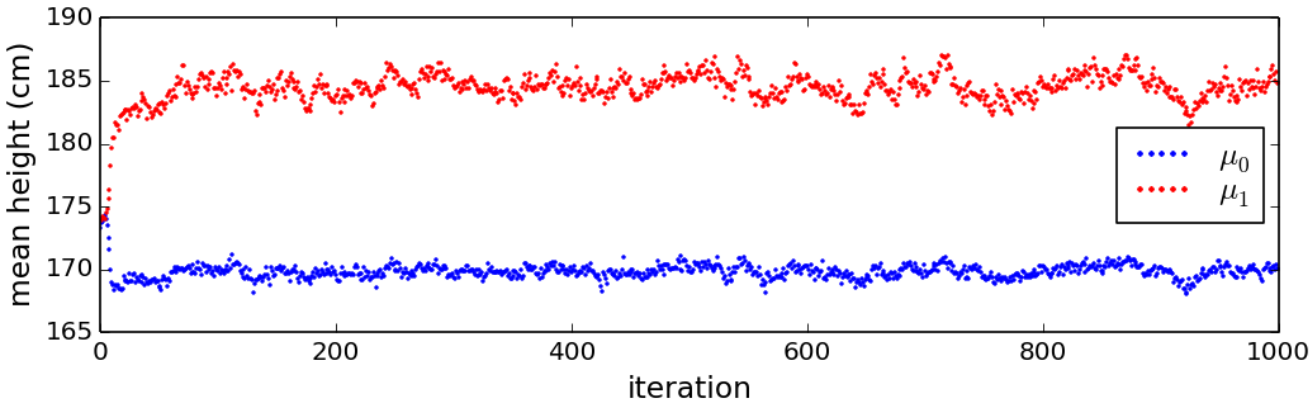


Figure 2: Traceplots of the component means, μ_0 and μ_1.

Results

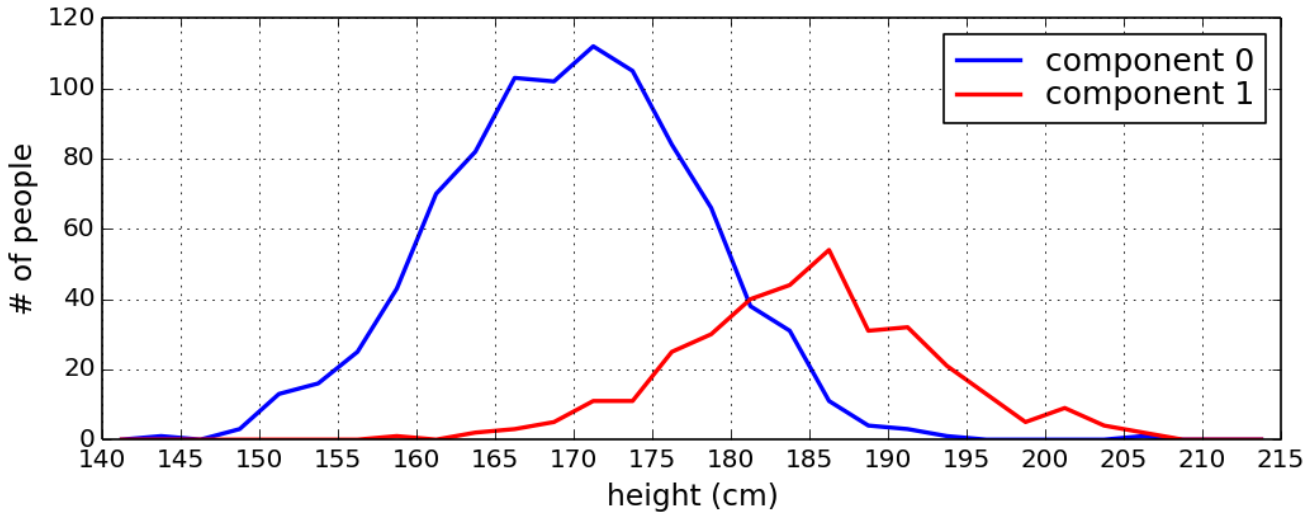


Figure 3: Histograms of the heights of subjects assigned to each component, according to $z_1, \ldots, z_n$, in a typical sample.

Results from two runs of the mixture model

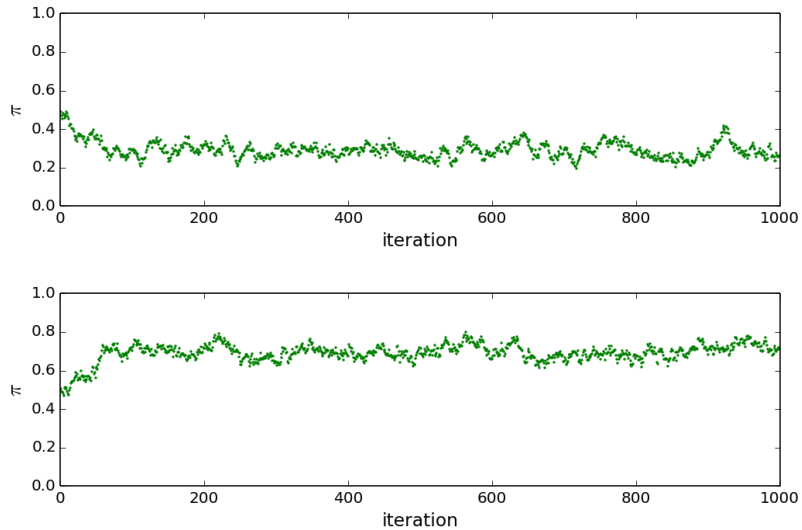


Figure 4: Traceplots of the mixture weight, π.

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 4.
- 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.

Takeaway from example

- 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).

Three component mixture model

- Consider a three component mixture of normal distribution with a common prior on the mixture component means, the error variance and the variance within mixture component means.
- The prior on the mixture weights w is a three component Dirichlet distribution.

$$
\begin{aligned}
p(Y_i|\mu_1, \mu_2, \mu_3, w_1, w_2, w_3, \varepsilon^2) &= \sum_{j=1}^{3} w_i N(\mu_j, \varepsilon^2) \\
\mu_j|\mu_0, \sigma_0^2 &\sim N(\mu_0, \sigma_0^2) \\
\mu_0 &\sim N(0, 3) \\
\sigma_0^2 &\sim \text{InverseGamma}(2, 2) \\
(w_1, w_2, w_3) &\sim \text{Dirichlet}(1, 1, 1) \\
\varepsilon^2 &\sim \text{InverseGamma}(2, 2),
\end{aligned}
$$

for $i = 1, \ldots n$.

Three component mixture model

Specifically,

- w_1, w_2 and w_3 are the mixture weight of mixture components 1,2 and 3 respectively
- μ_1, μ_2 and μ_3 are the means of the mixture components
- ε^2 is the variance parameter of the error term around the mixture components.

Dirichlet

A Dirichlet distribution is a distribution of the K-dimensional probability simplex

$$\triangle_K = \{(\pi_1, \ldots, \pi_k) : \pi_k \geq 0, \sum_k \pi_k = 1\}$$

We say that $(\pi_1, \ldots, \pi_k)$ is Dirichlet distributed:

$$(\pi_1, \ldots, \pi_k) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$$

if

$$p(\pi_1, \ldots, \pi_k) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_{k-1}}$$

Dirichlet distribution

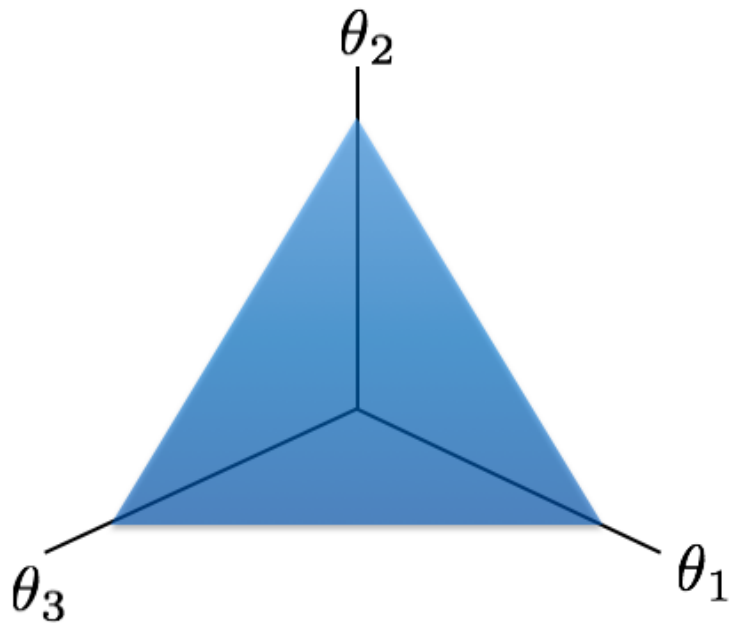
Let

$$\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$$

where the probability density function is

$$p(\theta \mid \alpha) \propto \prod_{k=1}^{m} \theta_k^{\alpha_k - 1},$$

where $\sum_k \theta_k = 1, \theta_i \geq 0$ for all i



Dirichlet distribution

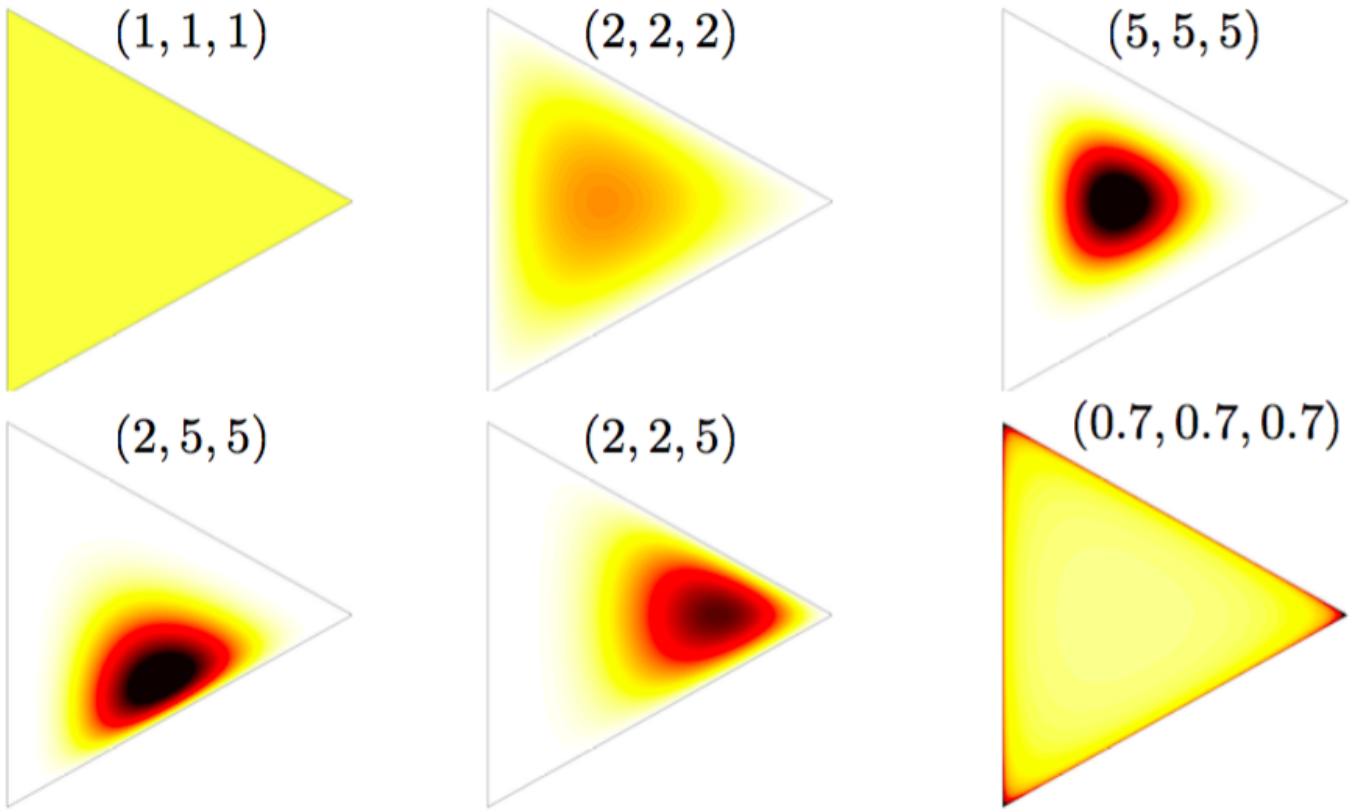


Figure 6: Far left: We get a uniform prior on the simplex. Moving to the right we get things unimodal. On the bottom, we get distributions that are multimodal at the corners.

Three component mixture model (Task 1 and 2)

Derive the full conditionals for all the parameters up to a normalizing constant.

Task 1 and 2

Specifically, you should derive the following conditional distributions below:

- $p(w_1, w_2, w_3 | \mu_1, \mu_2, \mu_3, \varepsilon^2, Y_1, ..., Y_N) \propto$
- $p(\mu_1 | \mu_2, \mu_3, w_1, w_2, w_3, Y_1, ..., Y_N, \varepsilon^2, \mu_0, \sigma_0^2) \propto$
- $p(\mu_2 | \mu_1, \mu_3, w_1, w_2, w_3, Y_1, ..., Y_N, \varepsilon^2, \mu_0, \sigma_0^2) \propto$
- $p(\mu_3 | \mu_1, \mu_2, w_1, w_2, w_3, Y_1, ..., Y_N, \varepsilon^2, \mu_0, \sigma_0^2) \propto$
- $p(\varepsilon^2 | \mu_1, \mu_2, \mu_3, Y_1, ..., Y_N) \propto$
- $p(\mu_0 | \mu_1, \mu_2, \mu_3, \sigma_0^2) \propto$
- $p(\sigma_0^2 | \mu_0, \mu_1, \mu_2, \mu_3) \propto$

Task 1 (Solution)

We start by deriving the full conditional kernels.

$$p(\mu_0|\mu_1, \mu_2, \mu_3, \varepsilon^2, \sigma_0^2) \propto \text{Normal-Normal mean update} \tag{12}$$

$$p(\sigma_0^2|\mu_1, \mu_2, \mu_3, \mu_0) \propto \text{Normal-InverseGamma variance update} \tag{13}$$

Task 1 (Solution)

$$p(\mu_k | Y_1, ..., Y_N, \sigma_0^2, \varepsilon^2, w_1, w_2, w_3) \tag{14}$$

$$\propto \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{\sigma_0^2}(\mu_k - \mu_0)^2} \prod_{i=1}^{N} \left(\sum_{j=1}^{3} w_j \frac{1}{\sqrt{2\pi\varepsilon^2}} e^{-\frac{1}{2\varepsilon^2}(Y_i - \mu_j)^2} \right) \tag{15}$$

$$\propto? \tag{16}$$

Task 1 (Solution)

$$p(\varepsilon^2 | Y_1, ..., Y_N, \mu_1, \mu_2, \mu_3, w_1, w_2, w_3) \tag{17}$$

$$\propto (\varepsilon^2)^{-3} e^{-\frac{2}{\varepsilon^2}} \prod_{i=1}^{N} \left(\sum_{j=1}^{3} w_j \frac{1}{\sqrt{2\pi\varepsilon^2}} e^{-\frac{1}{2\varepsilon^2}(Y_i - \mu_j)^2} \right) \tag{18}$$

$$\propto? \tag{19}$$

Task 1 (Solution)

$$p(w_1, w_2, w_3 | Y_1, ..., Y_N, \mu_1, \mu_2, \mu_3, \varepsilon^2) \tag{20}$$

$$\propto \prod_{i=1}^{N} \left(\sum_{j=1}^{3} w_j \frac{1}{\sqrt{2\pi\varepsilon^2}} e^{-\frac{1}{2\varepsilon^2}(Y_i - \mu_j)^2} \right) \tag{21}$$

$$\propto ? \tag{22}$$

Note that everything that involves the likelihood includes the products of sums, and becomes exceedingly painful. Thus, let us look at the full conditionals under data augmentation.

Data augumentation scheme

- Neither the joint posterior nor any of the full conditionals involving the likelihood are of a form that's easy to sample from.

Solution: introduce an additional set of random variables $\{Z_i\}_{i=1}^N$ that assign each observation to one of the mixture components with the proabilitiy of assignment being the respective mixture weight.

If we condition on Z_i we can then write the likelihood of Y_i as

$$p(Y_i|Z_i, \mu_1, \mu_2, \mu_3, \varepsilon^2) = \sum_{j=1}^{3} N(\mu_j, \varepsilon^2)\delta_j(Z_i) = N(\mu_{Z_i}, \varepsilon^2)$$

$$P(Z_i = j) = w_j.$$

Data augmentation (continued)

- Conditional on Z_i we no longer have a sum of Normal pdfs in our likelihood, resulting in a significant simplification.
- Conditional on the $\{Z_i\}$ updates will be straightforward, only depending on the mixture component that any given Y_i is currently assigned to.
- The drawback is that we also have to update $\{Z_i\}_{i=1}^N$ as well, introducing extra steps into our sampler.

The updated model

The model is now

$$
\begin{aligned}
Y_i \mid Z_i, \mu_1, \mu_2, \mu_3, \epsilon^2 &\sim N(\mu_{Z_i}, \epsilon^2) \\
\mu_j \mid \mu_0, \sigma_0^2 &\sim N(\mu_0, \sigma_0^2) \\
Z_i \mid w_1, w_2, w_3 &\sim \text{Cat}(3, \boldsymbol{w}) \\
\boldsymbol{w} = (w_1, w_2, w_3) &\sim \text{Dirichlet}(1, 1, 1) \\
\mu_0 &\sim N(0, 3) \\
\sigma_0^2 &\sim IG(2, 2) \\
\epsilon^2 &\sim IG(2, 2)
\end{aligned}
$$

$i = 1, \ldots, n$ $j = 1, \ldots, 3$

Multinomial-Dirichlet

In order to proceed with the lab, we'll need to learn about the Multinomial or Categorical distribution.

Multinomial or Categorical distribution

- $\theta = (\theta_1, \ldots, \theta_m)$,
- $X_i \in \{1, \ldots, m\}$,
- $\sum_i \theta_i = 1$.

Assume that

$$X \mid \theta \overset{ind}{\sim} \text{Multinomial}(\theta)$$

or

$$X \mid \theta \overset{ind}{\sim} \text{Categorical}(\theta)$$

$$P(X_i = j \mid \theta) = \theta_j$$

Conjugate prior (Dirichlet)

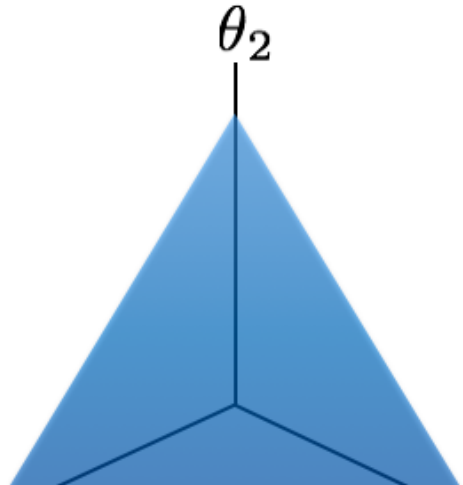
$$\theta \sim \text{Dirichlet}(\alpha)$$

Recall the density of the Dirichlet is the following:

$$p(\theta \mid \alpha) \propto \prod_{j=1}^{m} \theta_j^{\alpha_j - 1},$$

where $\sum_j \theta_j = 1, \theta_i \geq 0$ for all i

\begin{figure}[htbp]



Likelihood

Define the data as $D = (x_1, \ldots, x_n)$, $x_i \in \{1, \ldots m\}$. Consider

$$p(D \mid \theta) = \prod_{i=1}^{n} P(X_i = x_i \mid \theta) \tag{23}$$

$$= \prod_{i=1}^{n} \theta_{x_i} \tag{24}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{m} \theta_j^{I(x_i=j)} \tag{25}$$

$$= \prod_{j=1}^{m} \theta_j^{\sum_i I(x_i=j)} \tag{26}$$

$$= \prod_{j=1}^{m} \theta_j^{c_j} \tag{27}$$

where $c = (c_1, \ldots c_m)$, $c_j = \#\{i : x_i = j\}$.

Likelihood, Prior, and Posterior

$$p(D \mid \theta) = \prod_{j=1}^{m} \theta_j^{c_j}$$

$$P(\theta) \propto \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} I(\sum_j \theta_j = 1, \theta_i \geq 0 \forall i)$$

Then

$$P(\theta \mid D) \propto \prod_{j=1}^{m} \theta_j^{c_j} \times \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} I(\sum_j \theta_j = 1, \theta_i \geq 0 \forall i) \tag{28}$$

$$\propto \prod_{j=1}^{m} \theta_j^{c_j + \alpha_j - 1} I(\sum_j \theta_j = 1, \theta_i \geq 0 \forall i) \tag{29}$$

This implies

$$\theta \mid D \sim \text{Dirichlet}(c + \alpha).$$

Takeaways

1. Dirichlet is conjugate for Categorical or Multinomial.[1]
2. Useful formula:

$$\prod_i \text{Multinomial}(x_i \mid \theta) \times \text{Dir}(\theta \mid \alpha) \propto \text{Dir}(\theta \mid c + \alpha).$$

[1] The word Categorical seems to be used in CS and ML. The word Multinomial seems to be used in Statistics and Mathematics. I have no idea what is used in other sciences.

Task 3

Where necessary, (re)derive the full conditionals under the data augmentation scheme.

(See the lab solutions).

Task 4

In task 3 you derived all the full conditionals, and due to data augmentation scheme they are all in a form that is easy to sample. Use these full conditionals to implement Gibbs sampling using the data from "Lab8Mixture.csv".

Task 5

- Show traceplots for all estimated parameters
- Show means and 95% credible intervals for the marginal posterior distributions of all the parameters

Now suppose you re-run the sampler using 3 different starting values, are your results in a,b the same? Justify your reasoning with visualizations.

Sample code

Partial code for this problem can be found on Sakai!