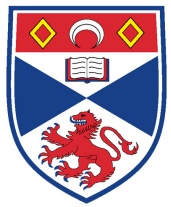
MT4531/5731: (Advanced) Bayesian Inference

Data augmentation - Auxiliary variables

Nicolò Margaritella

School of Mathematics and Statistics, University of St Andrews

–



Outline

Outline

Data augmentation (1)

- Data augmentation occurs when...
  - we have missing data, and we include them in the analysis by treating them as unknown random variables.
  - (Remember also that prediction can be put into a missing data framework, as we have seen in the Exponential-Gamma example of Lecture 12.)

Data augmentation (2)

- Data augmentation also occurs when...
  - The likelihood of the data is intractable (or difficult to calculate), but, ...
  - conditionally on a collection of unobserved data or parameters (auxiliary quantities), likelihood becomes tractable (or simpler) and,...
  - posterior calculations become simpler and quicker. We can define auxiliary variables in such a way that the resulting posterior conditional distributions are easier to sample from and/or result in better mixing.

Data augmentation (3)

- So, in simple words, data augmentation occurs when we include missing data in the analysis, or auxiliary quantities (usually model parameters) that are not an essential part of the model.
- Let $\boldsymbol{x}$ denote the observed data and $\boldsymbol{y}$ the auxiliary variables.
- Bayes' Theorem states that,

$$\pi(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{x}) \propto f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

- We use MCMC to sample from the posterior $\pi(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{x})$.
- However, we are interested in the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{x})$. This is the marginal posterior,

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \int \pi(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y}.$$

- We simply obtain a sample from the joint posterior distribution and only use the realisations of $\boldsymbol{\theta}$.

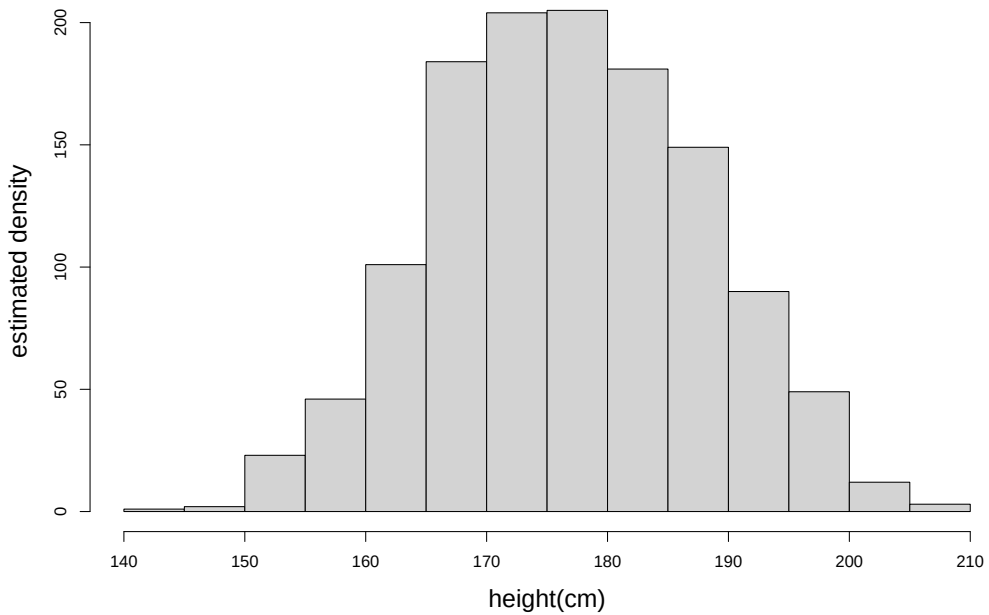
Outline

Example - mixture of normals (1)

- Consider a data set on the height of 700 women and 550 men.
- Suppose we have the list of heights, but we don't know which data points are from women and which are from men.

Example - mixture of normals (2)

- From the plot can we still infer the distribution of female and male heights?



- The answer is yes!

Example - mixture of normals (3)

- We will assume that both mixture components (female and male) have the same precision s, which is fixed and known.
- The model is

$$\begin{aligned} X_1, \ldots, X_n | \boldsymbol{\mu}, \pi &\sim \mathrm{g}(x|\boldsymbol{\mu}, \pi); \\ \boldsymbol{\mu} \equiv (\mu_1, \mu_2) &\sim \mathrm{N}(\phi, \lambda^{-1}); \\ \pi &\sim \mathrm{Beta}(a, b), \end{aligned} \tag{1}$$

where the pdf of $\mathrm{g}(x|\boldsymbol{\mu}, \pi)$ is $\pi \mathrm{N}(\mu_1, s^{-1}) + (1-\pi)\mathrm{N}(\mu_2, s^{-1})$

Example - mixture of normals (4)

- The likelihood is

$$f(x_1, \ldots, x_n|\boldsymbol{\mu}, \pi) = \prod_{i=1}^{n} g(x_i|\boldsymbol{\mu}, \pi);$$
$$= \prod_{i=1}^{n} \left[\pi \mathrm{N}(\mu_1, s^{-1}) + (1-\pi)\mathrm{N}(\mu_2, s^{-1})\right],$$

which is a complicated function of $\boldsymbol{\mu}$ and π and it will produce a difficult posterior to sample from!

Example - mixture of normals (5)

$\Rightarrow$ Define an allocation variable $Z_{i=1,\ldots,n}$ which identifies to what mixture component each data point comes from; i.e. Z_i identifies whether observation i is male or female.

- Using this auxiliary variable, we can rewrite our model as follows:

$$\begin{aligned} X_1, \ldots, X_n | \mu, \pi &\sim \mathrm{N}(\mu_{Z_i}, s^{-1}); \\ Z_1, \ldots, Z_n &\sim \text{Bernoulli}(\pi); \\ (\mu_1, \mu_2) &\sim \mathrm{N}(\phi, \lambda^{-1}); \\ \pi &\sim \text{Beta}(a, b), \end{aligned} \tag{2}$$

which is the same model as (1)!

Example - mixture of normals (5)

- In fact we can rewrite (using what probability law?)

$$p(x_i|\mu,\pi) = \sum_{j=1}^{2} p(x_i|Z_i = j,\mu,\pi)p(Z_i = j|\mu,\pi)$$

and then

$$\begin{aligned} p(x_i|\mu,\pi) &= p(x_i|Z_i = 1,\mu,\pi)p(Z_i = 1|\mu,\pi) + \\ &+ p(x_i|Z_i = 2,\mu,\pi)p(Z_i = 2|\mu,\pi), \\ &= \mathrm{N}(\mu_1, s^{-1})\pi + \mathrm{N}(\mu_2, s^{-1})(1-\pi), \\ &= g(x_i|\boldsymbol{\mu},\pi). \end{aligned}$$

Example - mixture of normals (6)

Let's look at the full conditionals now:

(1) $p(\pi|\boldsymbol{\mu}, \mathbf{Z}, x)$ - Conditional on $\mathbf{Z}$, π is independent of the other parameters and the conditional distribution reduces to a Beta-Bernoulli model and we obtain

$$p(\pi|\boldsymbol{\mu}, \mathbf{Z}, x) = p(\pi|\mathbf{Z}) = \text{Beta}(a + n_1, b + n_2),$$

where $n_j = \sum_i \mathbb{1}(Z_i = j)$ with $j = \{1, 2\}$

Example - mixture of normals (7)

Let's look at the full conditionals now:

(2) $p(\boldsymbol{\mu}|\mathbf{Z}, \pi, x)$ - Conditional on $\mathbf{Z}$, we know from which component each observation comes from! Therefore, conditionally on z, the conditional distribution reduces to two independent Normal-Normal models.

$$\mu_1|\mu_2, \mathbf{Z}, \pi, x \sim \mathrm{N}(M_1, L_1^{-1});$$
$$\mu_2|\mu_1, \mathbf{Z}, \pi, x \sim \mathrm{N}(M_2, L_2^{-1})$$

where $n_j = \sum_i \mathbb{1}(Z_i = j)$, $L_j = \lambda + n_j s$,
$M_j = \dfrac{\lambda\phi + s\sum_{i:z_i=j} x_i}{\lambda + n_j s}$ and $j = \{1, 2\}$.

Example - mixture of normals (8)

Let's look at the full conditionals now:

(3) $p(\mathbf{Z}|\boldsymbol{\mu}, \pi, x)$ - First observe that

$$p(\mathbf{Z}|\boldsymbol{\mu}, \pi, x) \propto p(\mathbf{Z}, \boldsymbol{\mu}, \pi, x) \propto p(x|\mathbf{Z}, \boldsymbol{\mu})p(\mathbf{Z}|\pi),$$

hence,

$$\begin{aligned} p(\mathbf{Z}|\boldsymbol{\mu}, \pi, x) &= \prod_{i=1}^{n} \mathrm{N}(x_i|\mu_{z_i}, s^{-1})\mathrm{Bernoulli}(z_i|\pi), \\ &= \prod_{i=1}^{n} \left(\mathrm{N}(x_i|\mu_1, s^{-1})\pi\right)^{z_i} \left(\mathrm{N}(x_i|\mu_2, s^{-1})(1-\pi)\right)^{1-z_i}, \\ &\propto \prod_{i=1}^{n} \mathrm{Bernoulli}\left(z_i|\frac{\theta_{i,1}}{\theta_{i,1}+\theta_{i,2}}\right) \end{aligned}$$

with $\theta_{i,1} = \mathrm{N}(x_i|\mu_1, s^{-1})\pi$ and $\theta_{i,2} = \mathrm{N}(x_i|\mu_2, s^{-1})(1-\pi)$

Outline

Example - NIMBLE analysis (1)

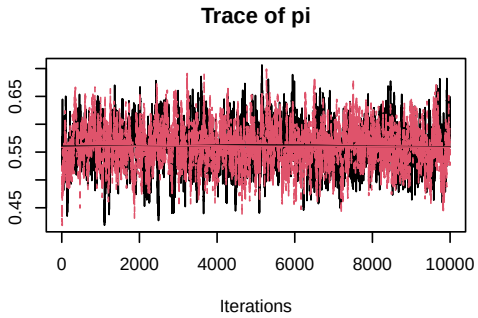
- Model settings:
  - $s = 1/\sigma^2$ where $\sigma = 8$cm is the SD of subjects' hight within each component.
  - $a = 1, b = 1$ for the Beta prior correspond to a Uniform prior on [0,1].
  - $\phi = 177$cm is the mean of the prior on the component means.
  - $1/\sqrt{\lambda} = 15$cm is the SD of the prior on the component means.

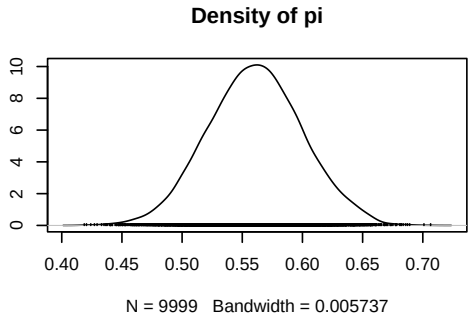
Example - NIMBLE analysis (2)

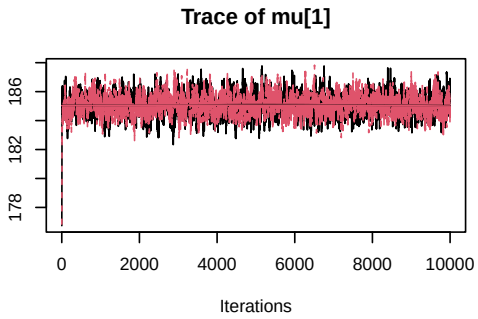
- Let's take a look at the BUGS code for the 2-component mixture model:

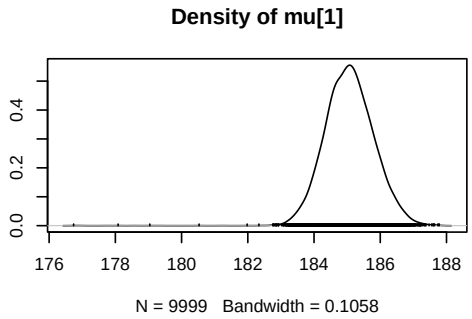
```
MixtureCode <- nimbleCode({
# Specify the likelihood:
for (i in 1:N){
x[i] ~ dnorm(mu[t[i]],s)
t[i] <- z[i]+1
z[i] ~ dbin(pi,1)# auxiliary variable
}
# Prior specification:
mu[1] ~ dnorm(phi,lambda)
mu[2] ~ dnorm(phi,lambda)
pi ~ dbeta(1,1)
})
```

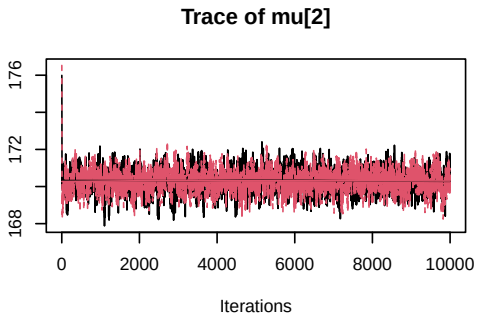
Example - NIMBLE analysis (3)

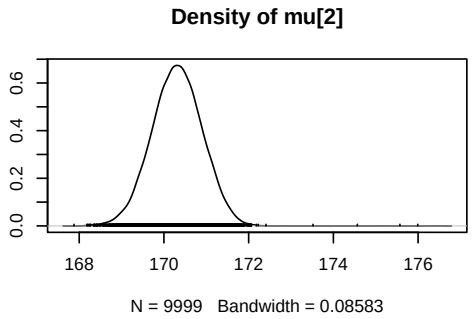






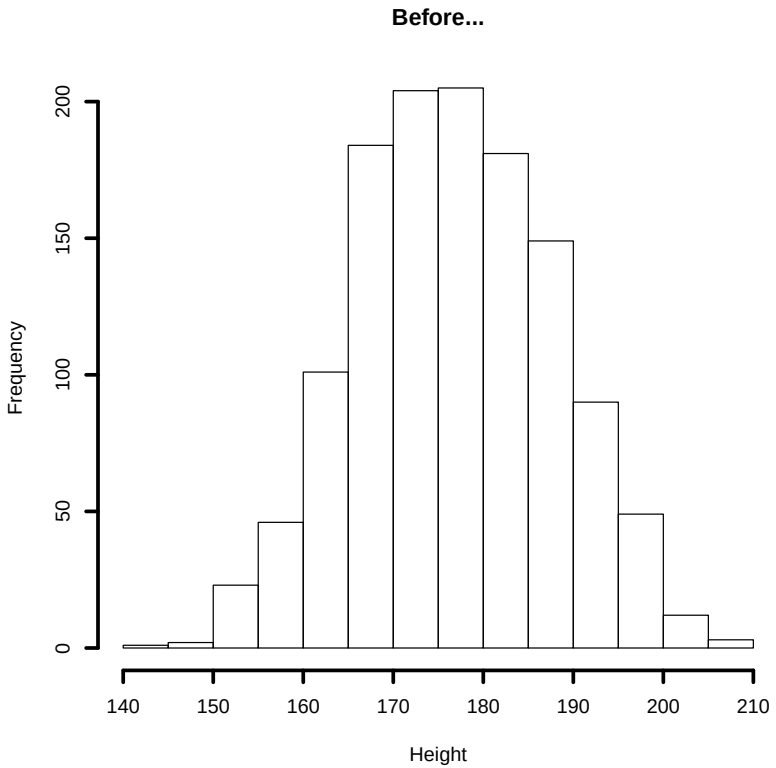


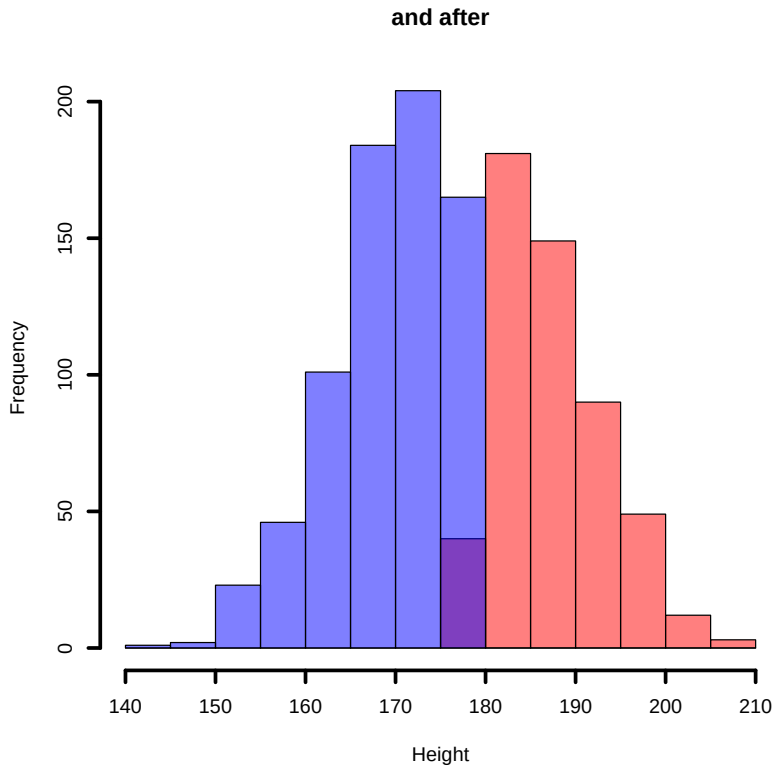




Example - NIMBLE analysis (4)

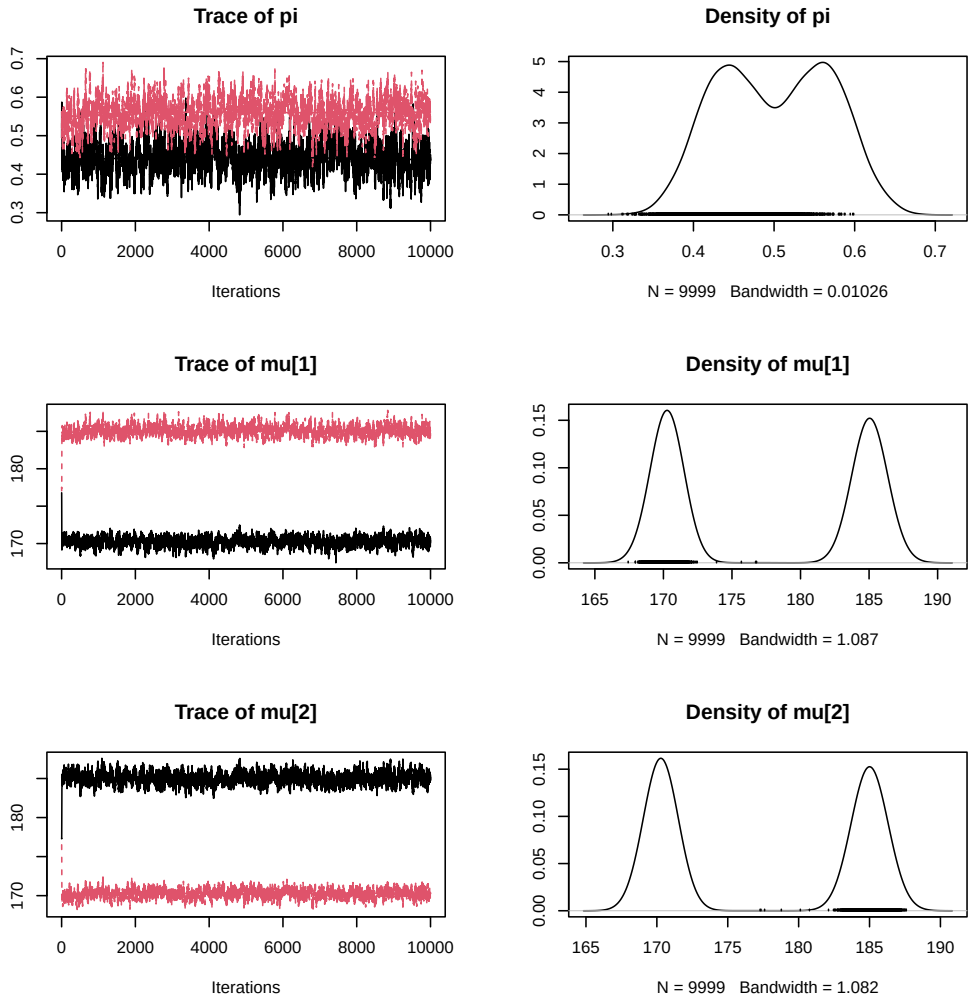
- Histograms of the heights of subjects assigned to each component, according to the posterior mode of z_j (right panel)





Watch out for multiple modes!

- If we run the sampler multiple times starting from the same initial values, sometimes it will settle on [females = 1, males = 2] and sometimes on [females = 2, males = 1].
- If the sampler were behaving properly, it would move back and forth between these two modes. But it gets stuck in one mode and stays there.
- Nevertheless, all parameters show good rate of convergence and mixing and we could just relabel the groups and then calculate the posterior summaries of interest



Try this at home

- **Task**: There is another interesting example of data augmentation in Section 2.5 of the lecture notes (Genetic Linkage) for you to read. This will also be useful to answer question 2 of tutorial 7.