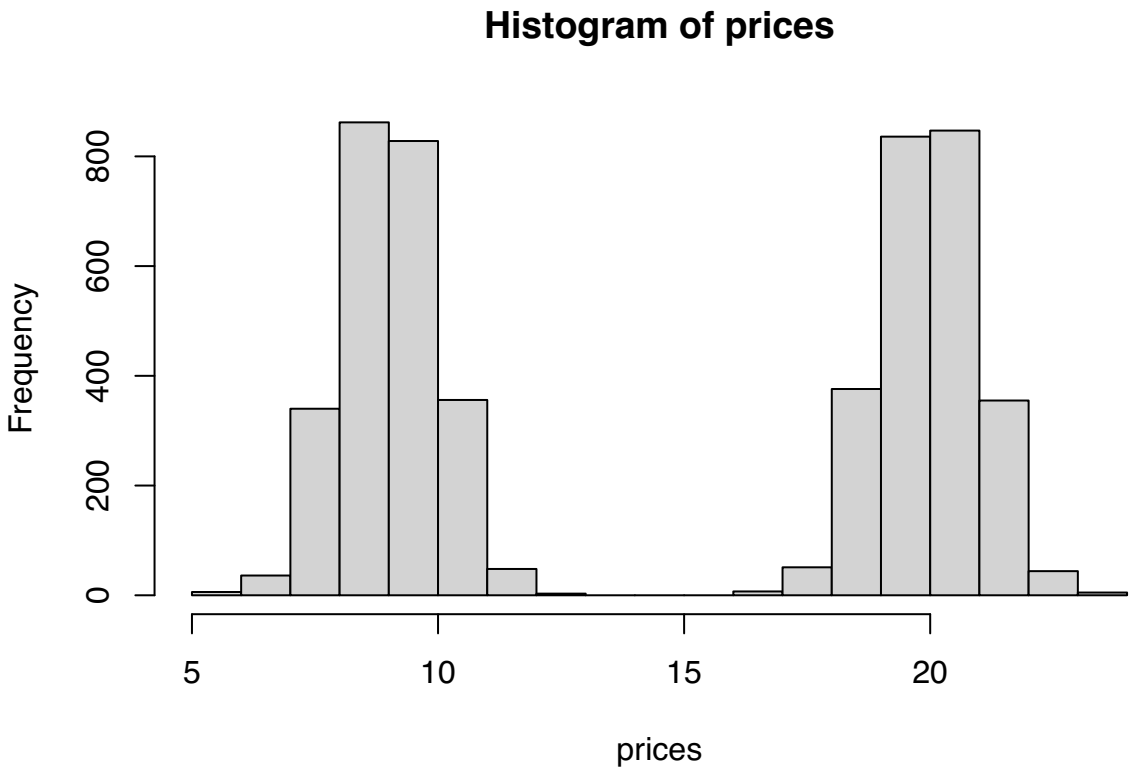
Mixture model examples

Matthew Stephens and Heejung Shim

Simulating the price of a randomly chosen book

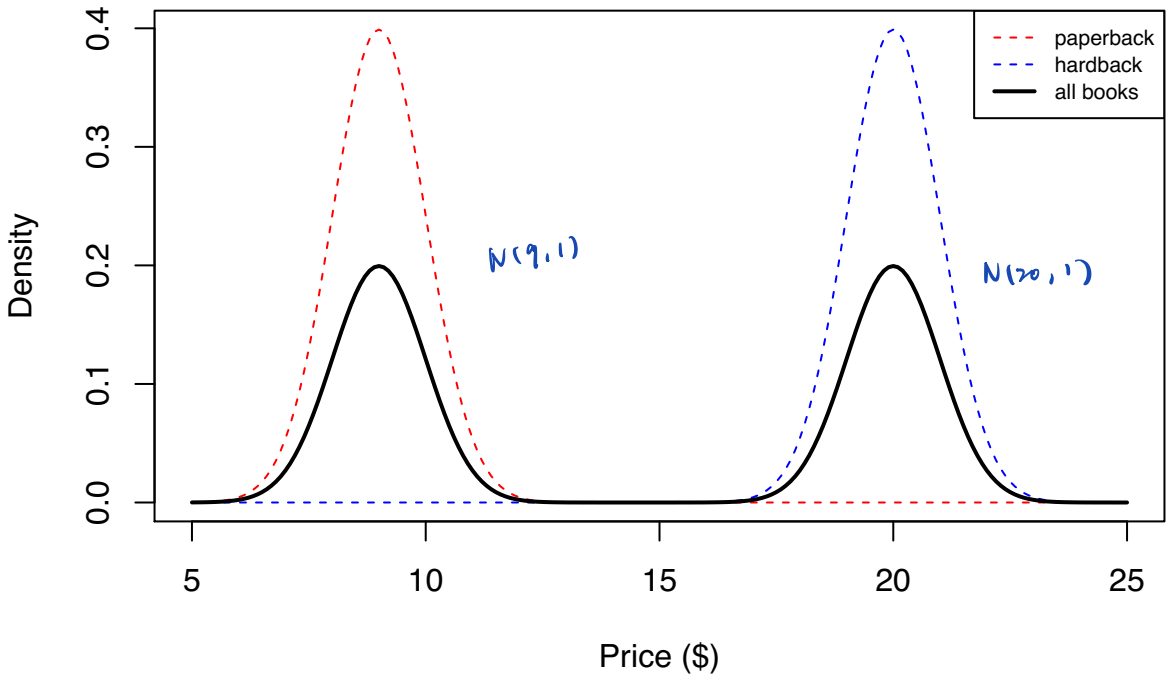
```
NUM.SAMPLES <- 5000
prices      <- rep(NA, NUM.SAMPLES)
for(i in seq_len(NUM.SAMPLES)) {
  z.i <- rbinom(1,1,0.5)
  if(z.i == 0) prices[i] <- rnorm(1, mean = 9, sd = 1)
  else prices[i] <- rnorm(1, mean = 20, sd = 1)
}
hist(prices)
```



We see that our histogram is bimodal. Indeed, even though the **mixture components** are each normal distributions, the distribution of a randomly chosen book is not. We illustrate the true densities below:

```
mu.pb   <- 9
sd.pb   <- 1
mu.hb   <- 20
sd.hb   <- 1

sample.pts     <- seq(5, 25, by=0.1) #grid for x values
```

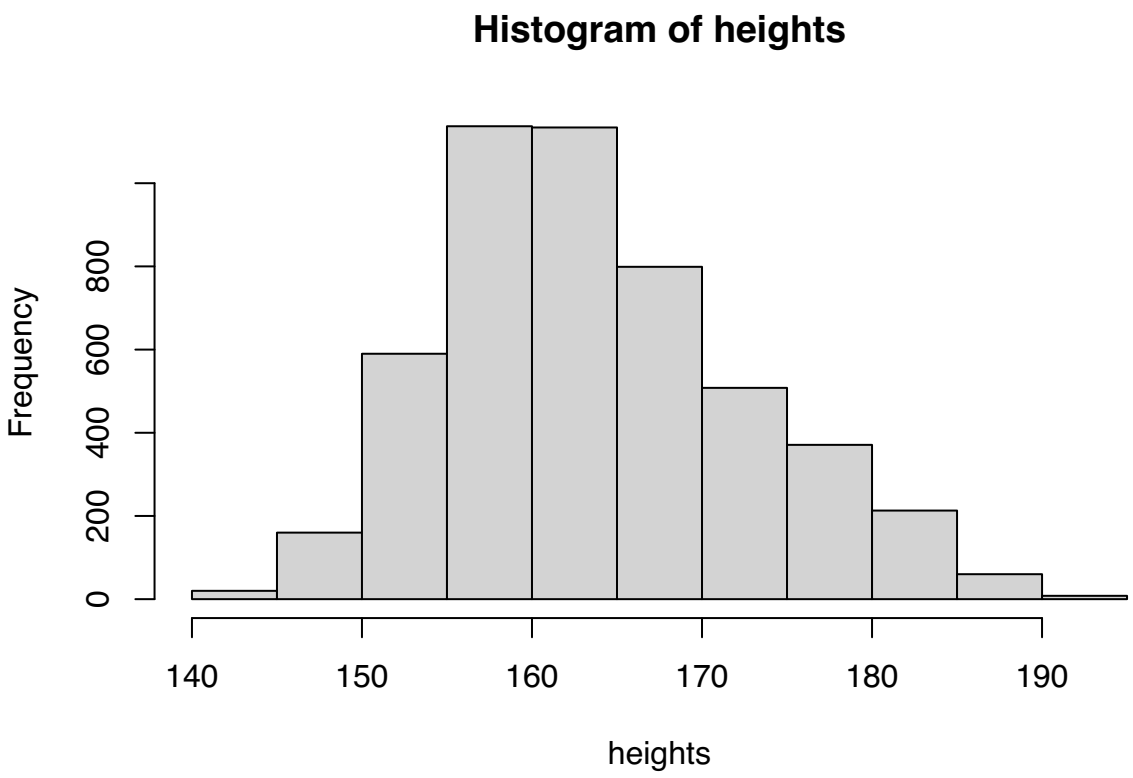
```r
density_pb   <- dnorm(sample.pts, mean=mu.pb, sd=sd.pb)
density_hb <- dnorm(sample.pts, mean=mu.hb, sd=sd.hb)

# plot distribution of paperback book price
plot(sample.pts, density_pb, col='red', type='l', xlab="Price ($)", ylab="Density", lty=2)
# plot distribution of hardback book price
lines(sample.pts, density_hb, col='blue', type='l', lty=2)
# plot distribution of randomly chosen book price
lines(sample.pts, .5*density_hb + .5*density_pb, col='black', type='l', lwd=2)

legend('topright', c('paperback', 'hardback', 'all books'), col=c('red', 'blue', 'black'), lty=c(2,2,1)
```



Simulating the heights of students at the University of Melbourne

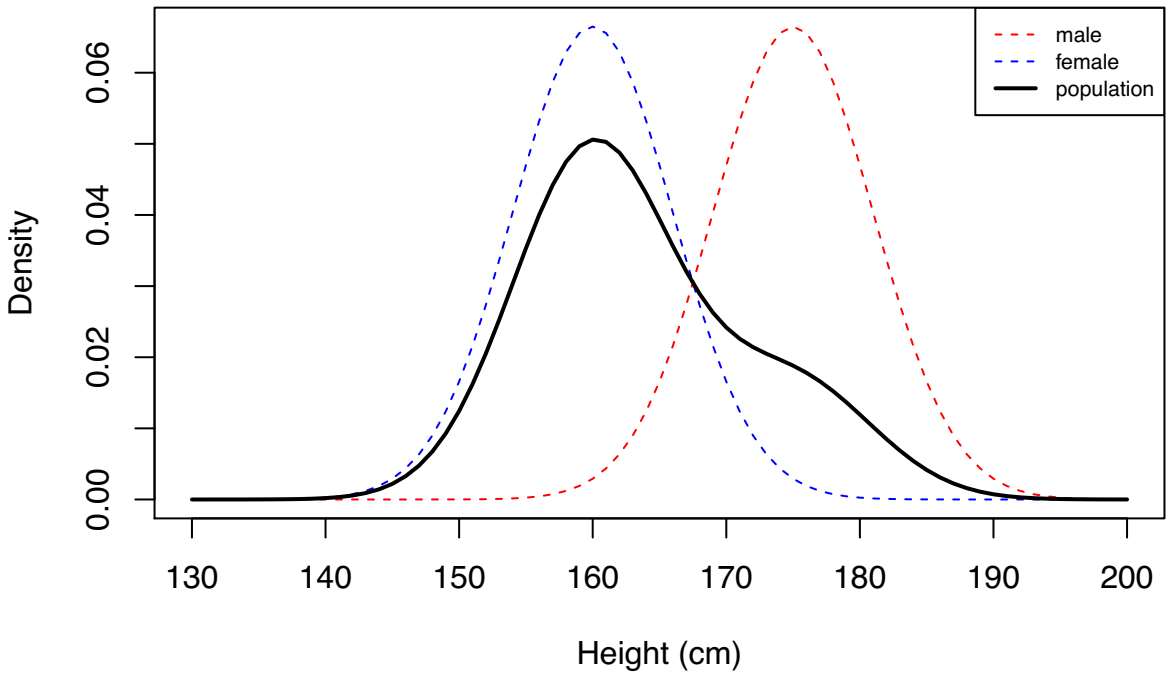
```r
NUM.SAMPLES <- 5000
heights      <- rep(NA, NUM.SAMPLES)
for(i in seq_len(NUM.SAMPLES)) {
  z.i <- rbinom(1,1,0.75)
  if(z.i == 0) heights[i] <- rnorm(1, mean = 175, sd = 6)
  else heights[i] <- rnorm(1, mean = 160, sd = 6)
}
hist(heights)
```

$z_i = \begin{cases} 0 & 0.25 \leftarrow \text{male} \quad N(175, 6^2) \\ 1 & 0.75 \leftarrow \text{female} \quad N(160, 6^2) \end{cases}$



Now we see that histogram is unimodal. Are heights normally distributed under this model? We plot the corresponding densities below:

```
mu.male   <- 175
sd.male   <- 6
mu.female <- 160
sd.female <- 6

sample.pts     <- seq(130, 200, by=1)
density_male   <- dnorm(sample.pts, mean=mu.male, sd=sd.male)
density_female <- dnorm(sample.pts, mean=mu.female, sd=sd.female)

plot(sample.pts, density_male, col='red', type='l', xlab="Height (cm)", ylab="Density", lty=2)
lines(sample.pts, density_female, col='blue', type='l', lty=2)
lines(sample.pts, .75*density_female + .25*density_male, col='black', type='l', lwd=2)

legend('topright', c('male', 'female', 'population'), col=c('red', 'blue', 'black'), lty=c(2,2,1), lwd=
```



Here we see that the Gaussian mixture model is unimodal because there is so much overlap between the two densities. In this example, you can see that the population density is not symmetric, and therefore not normally distributed.

Computing the probability of each book being paperback or hardback

Simulate the data.

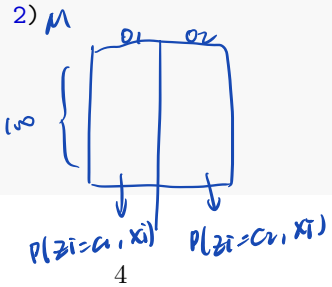
```
set.seed(30027)
NUM.SAMPLES <- 100
X       <- rep(NA, NUM.SAMPLES)
Z       <- rep(NA, NUM.SAMPLES)
for(i in seq_len(NUM.SAMPLES)) {
  Z[i] <- rbinom(1,1,0.5)
  if(Z[i] == 0) X[i] <- rnorm(1, mean = 9, sd = 1)
  else X[i] <- rnorm(1, mean = 20, sd = 1)
}
```

$Z = \begin{cases} 0 & \text{paperback} \\ 1 & \text{hardback} \end{cases}$

X contains simulated book prices and Z contains true book types. We pretend that we just obtain X (a list of book prices) without Z (book type information). Using the observed book prices, can we cluster book prices into two clusters by using GMM? We assume we have already obtained MLE of the parameters in the model. We will use true values of parameters as MLE in this analysis ($\hat{\pi}_1$=0.5, $\hat{\mu}_1 = 9$, $\hat{\sigma}_1 = 1$, $\hat{\mu}_2 = 20$, $\hat{\sigma}_2 = 1$). In practice, you should estimate parameters (e.g., obtaining MLE using EM algorithm).

Compute $P(X_i|Z_i = C_1)$ and $P(X_i|Z_i = C_2)$. We store these values in the columns of L:

```
L = matrix(NA, nrow=length(X), ncol= 2)

# L[, 1] : P(X_i | Z_i= C_1)
L[, 1] = dnorm(X, mean=9, sd = 1)

# L[, 2] : P(X_i | Z_i= C_2)
```



```
L[, 2] = dnorm(X, mean=20, sd = 1)
```

Compute $P(X_i, Z_i = C_1)$ and $P(X_i, Z_i = C_2)$. We store these values in the columns of M:

```
M = matrix(NA, nrow=length(X), ncol= 2)

# M[, 1] : P(X_i, Z_i= C_1)
M[,1] = L[, 1]*0.5

# M[, 2] : P(X_i, Z_i= C_2)
M[,2] = L[, 2]*0.5
```

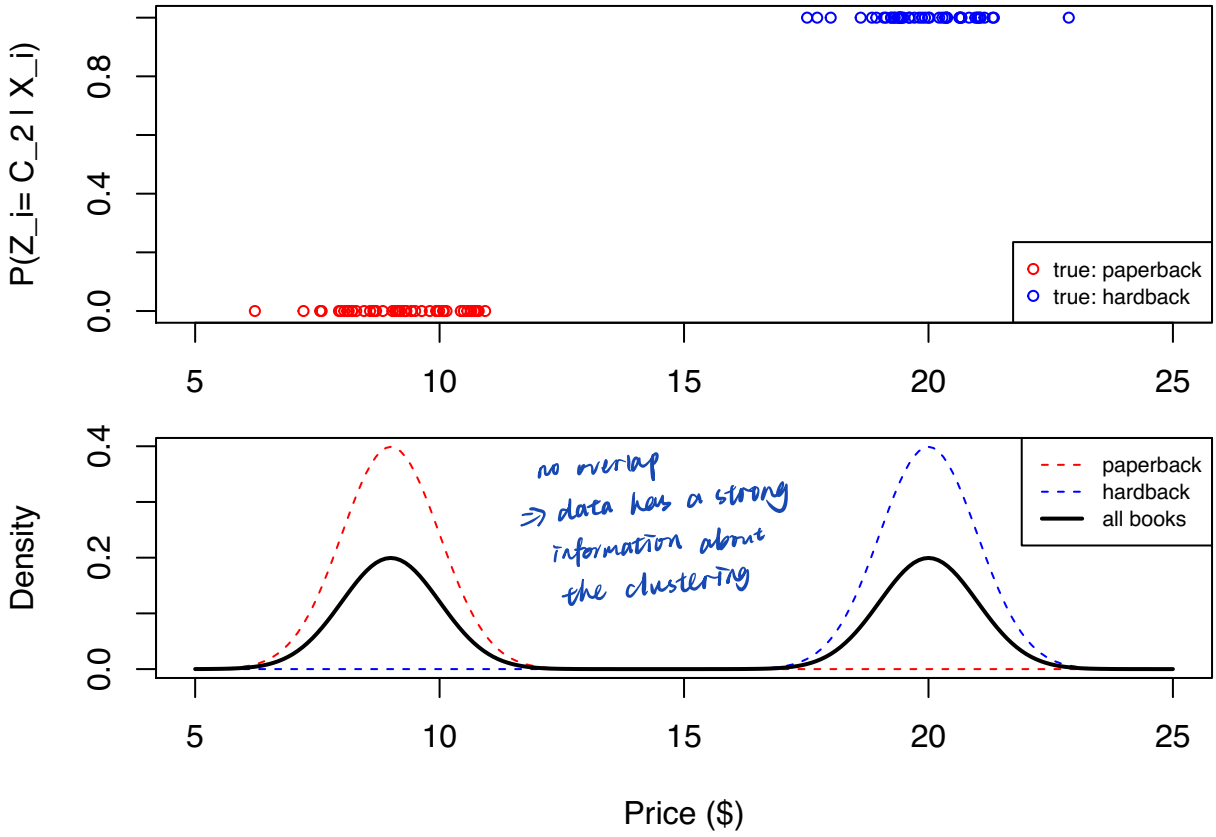
Compute $P(Z_i = C_1|X_i)$ and $P(Z_i = C_2|X_i)$. We store these values in the columns of CP:

```
CP = M/rowSums(M)
```

if $P(z_i = C_1 | x_i) > 0.5$
$\Leftrightarrow P(z_i = C_1|x_i) > P(z_i = C_2|x_i) \Leftrightarrow x_i \Rightarrow C_2$
if $P(z_i = C_1|x_i) < 0.5 \Rightarrow x_i \Rightarrow C_1$

Visualise results.

```
par(mfrow=c(2,1))

par(mar=c(2, 4,1,1)) # margin: bottom, left, top, right

col_list = c("red", "blue")[Z+1]  # Z=1 (hardback) => blue  Z=0 (paperback) => red
plot(X,CP[,2], col=col_list , xlim=c(5,25), ylab = "P(Z_i= C_2 | X_i)", pch =1, cex = 0.7)
legend('bottomright', c('true: paperback', 'true: hardback'), col=c('red', 'blue'), pch = 1, cex=0.7)


par(mar=c(4, 4,1,1))
sample.pts     <- seq(5, 25, by=0.1)
plot(sample.pts, density_pb, col='red', type='l', xlab="Price ($)", ylab="Density", lty=2, xlim=c(5,25))
# plot distribution of hardback book price
lines(sample.pts, density_hb, col='blue', type='l', lty=2)
# plot distribution of randomly chosen book price
lines(sample.pts, .5*density_hb + .5*density_pb, col='black', type='l', lwd=2)
legend('topright', c('paperback', 'hardback', 'all books'), col=c('red', 'blue', 'black'), lty=c(2,2,1))
```



Computing the probability of each student being female or male

Simulate the data.

```
set.seed(30027)
NUM.SAMPLES <- 100
X <- rep(NA, NUM.SAMPLES)
Z <- rep(NA, NUM.SAMPLES)
for(i in seq_len(NUM.SAMPLES)) {
    Z[i] <- rbinom(1,1,0.75)
  if(Z[i] == 0) X[i] <- rnorm(1, mean = 175, sd = 6)
  else X[i] <- rnorm(1, mean = 160, sd = 6)
}
```

X contains simulated heights and Z contains true gender information. We pretend that we just obtain X (a list of heights) without Z (gender information). Using the observed heights, can we cluster student's heights into two clusters by using GMM? We assume we have already obtained MLE of the parameters in the model. We will use true values of parameters as MLE in this analysis ($\hat{\pi}_1$=0.25, $\hat{\mu}_1 = 175$, $\hat{\sigma}_1 = 6$, $\hat{\mu}_2 = 160$, $\hat{\sigma}_2 = 6$). In practice, you should estimate parameters (e.g., obtaining MLE using EM algorithm).

Compute $P(X_i|Z_i = C_1)$ and $P(X_i|Z_i = C_2)$. We store these values in the columns of L:

```
L = matrix(NA, nrow=length(X), ncol= 2)

# L[, 1] : P(X_i | Z_i= C_1)
L[, 1] = dnorm(X, mean=175, sd = 6)

# L[, 2] : P(X_i | Z_i= C_2)
```

```
L[, 2] = dnorm(X, mean=160, sd = 6)
```

Compute $P(X_i, Z_i = C_1)$ and $P(X_i, Z_i = C_2)$. We store these values in the columns of M:

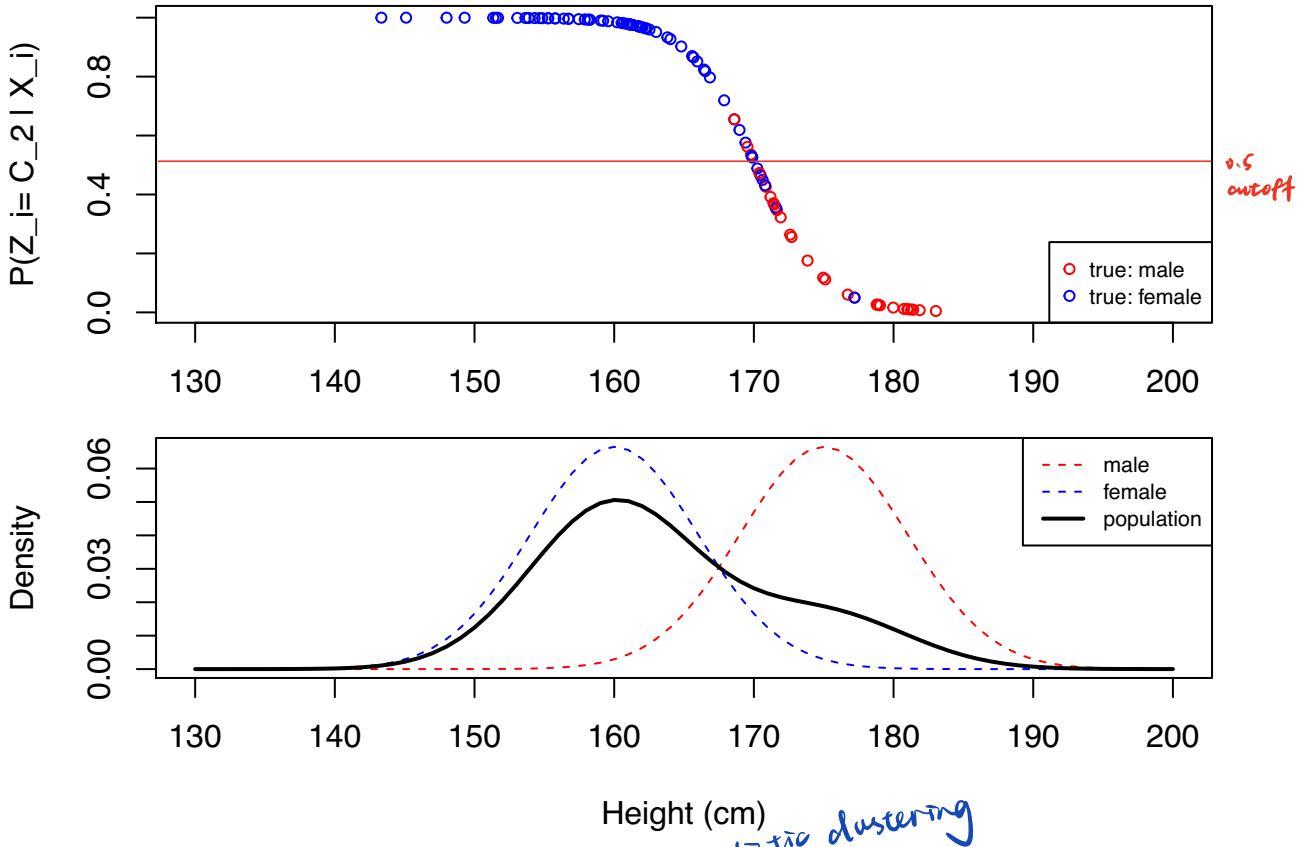
```
M = matrix(NA, nrow=length(X), ncol= 2)

# M[, 1] : P(X_i, Z_i= C_1)
M[,1] = L[, 1]*0.25

# M[, 2] : P(X_i, Z_i= C_2)
M[,2] = L[, 2]*0.75
```

Compute $P(Z_i = C_1|X_i)$ and $P(Z_i = C_2|X_i)$. We store these values in the columns of CP:

```
CP = M/rowSums(M)
```

Visualise results.

```
par(mfrow=c(2,1))

par(mar=c(2, 4,1,1)) # margin: bottom, left, top, right

col_list = c("red", "blue")[Z+1]  # Z=1 (female) => blue  Z=0 (male) => red
plot(X,CP[,2], col=col_list , xlim=c(130,200), ylab = "P(Z_i= C_2 | X_i)", pch =1, cex = 0.7)
legend('bottomright', c('true: male', 'true: female'), col=c('red', 'blue'), pch = 1, cex=0.7)

par(mar=c(4, 4,1,1))
sample.pts     <- seq(130, 200, by=1)
plot(sample.pts, density_male, col='red', type='l', xlab="Height (cm)", ylab="Density", lty=2)
lines(sample.pts, density_female, col='blue', type='l', lty=2)
lines(sample.pts, .75*density_female + .25*density_male, col='black', type='l', lwd=2)
legend('topright', c('male', 'female', 'population'), col=c('red', 'blue', 'black'), lty=c(2,2,1), lwd=
```



probabilistic clustering

$P(z_i = c_2 | x)$

also can be represent as confidence

0.999 → sure

0.655 → not much sure