MAST30027 Assignment 3

Name: Lang (Ron) Chen

Student ID: 1181506

Subject: MAST30027

Assignment number: 3

Tutorial time and tutor: Tues 11:00 Yidi Deng

1.

a.

$$P(x_1,\dots,x_{300}, z_1,\dots, z_{300}|\theta) = \prod_{i=1}^{300} P(x_i|z_i,\theta)\, P(z_i|\theta)$$

$$= \prod_{i=1}^{300}\prod_{k=1}^{3} \left[P(x_i|z_i=k,\theta)\, P(z_i=k|\theta)\right]^{I(z_i=k)}$$

$$\log\left[P(x_1,\dots,x_{300},z_1,\dots,z_{300}|\theta)\right] = \sum_{i=1}^{300}\sum_{k=1}^{3} I(z_i=k)\left[\log P(x_i|z_i=k,\theta)+\log P(z_i=k|\theta)\right]$$

$$Q(\theta,\theta^\circ) = E_{z|x,\theta^\circ}\left[\log P(x_1,\dots,x_{300},z_1,\dots,z_{300}|\theta)\right]$$
$$= \sum_{i=1}^{300}\sum_{k=1}^{3} P(z_i=k|x_i,\theta^\circ)\left[\log P(x_i|z_i=k,\theta)+\log P(z_i=k|\theta)\right]$$

$$= \sum_{i=1}^{300}\sum_{k=1}^{3} P(z_i=k|x_i,\theta^\circ)\Big[\log\binom{20}{x_i} + x_i\log p_k + (20-x_i)\log(1-p_k)$$
$$+\log\pi_k\Big]$$

where $\pi_3 = \pi_1 - \pi_2$

b. E-step

Let $\theta^\circ = (\pi_1^\circ, \pi_2^\circ, p_1^\circ, p_2^\circ, p_3^\circ)$

$$P(z_i=1|x_i,\theta^\circ) = \frac{P(z_i=1,x_i|\theta^\circ)}{P(x_i|\theta^\circ)}$$

$$= \frac{P(x_i|z_i=1,\theta^\circ)P(z_i=1|\theta^\circ)}{P(x_i|z_i=1,\theta^\circ)P(z_i=1|\theta^\circ)+P(x_i|z_i=2,\theta^\circ)P(z_i=2|\theta^\circ)+P(x_i|z_i=3,\theta^\circ)P(z_i=3|\theta^\circ)}$$

$$P(z_i=2|x_i,\theta^\circ) = \frac{P(z_i=2,x_i|\theta^\circ)}{P(x_i|\theta^\circ)}$$

$$= \frac{P(x_i|z_i=2,\theta^\circ)P(z_i=2|\theta^\circ)}{P(x_i|z_i=1,\theta^\circ)P(z_i=1|\theta^\circ)+P(x_i|z_i=2,\theta^\circ)P(z_i=2|\theta^\circ)+P(x_i|z_i=3,\theta^\circ)P(z_i=3|\theta^\circ)}$$

$$P(z_i=3|x_i,\theta^\circ) = 1 - P(z_i=1|x_i,\theta^\circ) - P(z_i=2|x_i,\theta^\circ)$$

where $P(x_i|z_i=k,\theta^\circ) = \binom{20}{x_i}(p_k^\circ)^{x_i}(1-p_k^\circ)^{20-x_i}$

$P(z_i=1|\theta^\circ) = \pi_1^\circ$

$P(z_i=2|\theta^\circ) = \pi_2^\circ$

$P(z_i=3|\theta^\circ) = 1-\pi_1^\circ-\pi_2^\circ$

c. $$\frac{\partial Q(\theta,\theta^0)}{\partial \pi_1} = \sum_{i=1}^{300}\left[\frac{P(z_i=1|x_i,\theta^0)}{\pi_1} - \frac{P(z_i=3|x_i,\theta^0)}{1-\pi_1-\pi_2}\right]$$

$$\frac{d}{d\pi_1}\log(1-\pi_1-\pi_2) = \frac{-1}{1-\pi_1-\pi_2}$$

$$\sum_{i=1}^{300}\left[\frac{P(z_i=1|x_i,\theta^0)}{\pi_1} - \frac{P(z_i=3|x_i,\theta^0)}{1-\pi_1-\pi_2}\right] = 0$$

$$= \sum_{i=1}^{300}\left[\frac{(1-\pi_1-\pi_2)P(z_i=1|x_i,\theta^0) - \pi_1 P(z_i=3|x_i,\theta^0)}{\pi_1(1-\pi_1-\pi_2)}\right] = 0$$

$$\Rightarrow (1-\pi_1-\pi_2)\sum_{i=1}^{300}P(z_i=1|x_i,\theta^0) = \pi_1\sum_{i=1}^{300}P(z_i=3|x_i,\theta^0) \qquad \pi_1\text{'s equation}$$

$$\Rightarrow (1-\pi_1-\pi_2)\sum_{i=1}^{300}P(z_i=2|x_i,\theta^0) = \pi_2\sum_{i=1}^{300}P(z_i=3|x_i,\theta^0) \qquad \pi_2\text{'s equation}$$

π_1's equation + π_2's equation

$$(1-\pi_1-\pi_2)\sum_{i=1}^{300}P(z_i=1|x_i,\theta^0)+P(z_i=2|x_i,\theta^0) = (\pi_1+\pi_2)\sum_{i=1}^{300}P(z_i=3|x_i,\theta^0)$$

$$(1-\pi_1-\pi_2)\sum_{i=1}^{300}P(z_i=1|x_i,\theta^0)+P(z_i=2|x_i,\theta^0) = (\pi_1+\pi_2+\pi_3-\pi_3)\sum_{i=1}^{300}P(z_i=3|x_i,\theta^0)$$

$$(1-\pi_1-\pi_2)\sum_{i=1}^{300}P(z_i=1|x_i,\theta^0)+P(z_i=2|x_i,\theta^0) = (1-(1-\pi_1-\pi_2))\sum_{i=1}^{300}P(z_i=3|x_i,\theta^0)$$

$$(1-\pi_1-\pi_2)\sum_{i=1}^{300}[P(z_i=1|x_i,\theta^0)+P(z_i=2|x_i,\theta^0)+P(z_i=3|x_i,\theta^0)] = \sum_{i=1}^{300}P(z_i=3|x_i,\theta^0)$$

$$(1-\pi_2-\pi_1)\sum_{i=1}^{300}1 = \sum_{i=1}^{300}P(z_i=3|x_i,\theta^0)$$

$$(1-\pi_2-\pi_1) = \frac{1}{300}\sum_{i=1}^{300}P(z_i=3|x_i,\theta^0)$$

Sub this result into π_1 and π_2's equations:

$$\frac{1}{300}\cancel{\sum_{i=1}^{300}P(z_i=3|x_i,\theta^0)}\sum_{i=1}^{300}P(z_i=1|x_i,\theta^0) = \pi_1\cancel{\sum_{i=1}^{300}P(z_i=3|x_i,\theta^0)}$$

$$\Rightarrow \hat{\pi}_1 = \frac{1}{300}\sum_{i=1}^{300}P(z_i=1|x_i,\theta^0)$$

$$\Rightarrow \hat{\pi}_2 = \frac{1}{300}\sum_{i=1}^{300}P(z_i=2|x_i,\theta^0)$$

$$\Rightarrow \hat{\pi}_3 = 1-\hat{\pi}_2-\hat{\pi}_1$$

$$\frac{\partial Q(\theta,\theta^0)}{\partial p_k} = \sum_{i=1}^{300} \mathbb{P}(z_i = k \mid x_i, \theta^0)\left[\frac{x_i}{p_k} - \frac{20 - x_i}{1-p_k}\right]$$

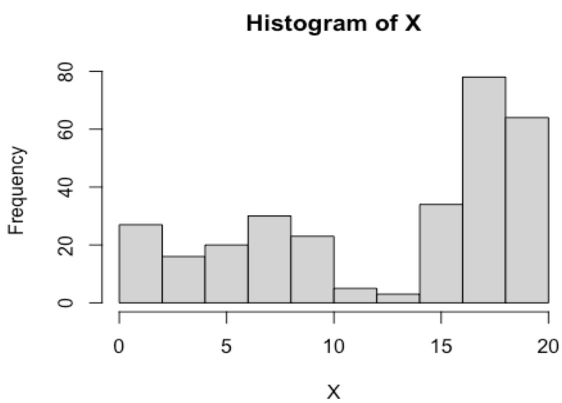
$$\Rightarrow \sum_{i=1}^{300} \mathbb{P}(z_i = k \mid x_i, \theta^0)\left[\frac{(1-p_k)x_i - p_k(20 - x_i)}{p_k(1-p_k)}\right] = 0$$

$$\Rightarrow \sum_{i=1}^{300} \mathbb{P}(z_i = k \mid x_i, \theta^0)\left[(1-p_k)x_i - p_k(20 - x_i)\right] = 0$$

$$\Rightarrow \sum_{i=1}^{300} \mathbb{P}(z_i = k \mid x_i, \theta^0)\left[x_i - 20\, p_k\right] = 0$$

$$\Rightarrow \quad \hat{p}_k = \frac{\sum_{i=1}^{300} \mathbb{P}(z_i = k \mid x_i, \theta^0)\, x_i}{20 \sum_{i=1}^{300} \mathbb{P}(z_i = k \mid x_i, \theta^0)}$$

```r
# read in data

setwd("/Users/tg.chenny/Desktop/1. University/1. Undergraduate/20. Modern Applied Statistics/A
smt/Asmt 3")

ROUNDING = 20

X = scan(file="assignment3_prob1.txt", what=double())

length(X)

## [1] 300

hist(X)
```

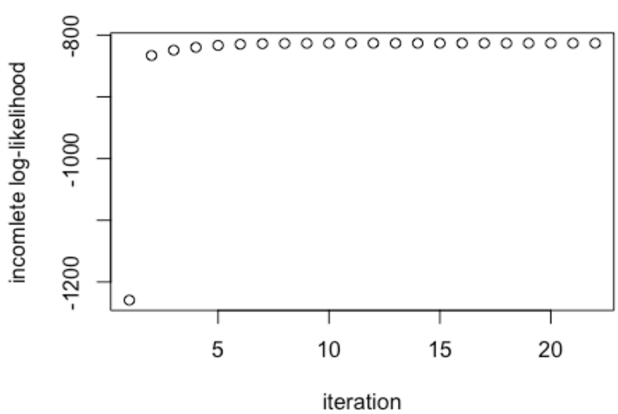


Q1 d

```r
mixture.EM = function(X, w.init, p.init, epsilon=1e-5, max.iter = 100) {

  # set initial values into current
  w.curr = w.init
  p.curr = p.init

  # setup list to store log likelihood - so to be able to plot
  log_liks = c()
  log_liks = c(log_liks, compute.log.lik(X, w.curr, p.curr)$ill)

  # set initial values for the operation of this function
  delta.ll = 1
  n.iter = 1

  # while not reach max iteration and change in log likelihood is larger than
  # epsilon
  while((delta.ll > epsilon) & (n.iter <= max.iter)) {

    # run one EM iteration
    EM.out = EM.iter(X, w.curr, p.curr)

    # put the new current values into the current
    w.curr = EM.out$w.new
    p.curr = EM.out$p.new

    # compute new log likelihood under new parameter estimates
```

```r
    # (and add to list)
    log_liks = c(log_liks, compute.log.lik(X, w.curr, p.curr)$ill)

    delta.ll = log_liks[length(log_liks)] - log_liks[length(log_liks)-1]

    n.iter = n.iter+1

  }

  return(list(w.curr=w.curr, p.curr=p.curr, log_liks=log_liks))
}


# compute incomplet log likelihood
compute.log.lik = function(X, w.curr, p.curr) {

  # for each sample, compute P(Xi, Zi=k) using helper
  prob.x.z = compute.prob.x.z(X, w.curr, p.curr)$prob.x.z

  # get incomplete log likelihoods (each instance probabilistic weighted
  # between states and then logs taken, then summed)
  ill = sum(log(rowSums(prob.x.z)))

  return(list(ill=ill))

}


# for each sample Xi, compute P(Xi, Zi=k)
compute.prob.x.z = function(X, w.curr, p.curr) {

  # for each sample Xi, compute P(Xi, Zi=k)
  L = matrix(NA, nrow=length(X), ncol=length(w.curr))

  for (k in seq_len(ncol(L))) {

    L[,k] = dbinom(X, size=20, prob = p.curr[k])*w.curr[k]

  }

  return(list(prob.x.z = L))
}


# compute one EM step (update values)
EM.iter = function(X, w.curr, p.curr) {

  # E

  # for each sample Xi, compute P(Xi, Zi=k)
  prob.x.z = compute.prob.x.z(X, w.curr, p.curr)$prob.x.z

  # compute P(Zi=k | Xi) = P(Xi, Zi=k)/P(Xi)
  P_ik = prob.x.z / rowSums(prob.x.z)


  # M
  w.new = colSums(P_ik)/sum(P_ik)
```

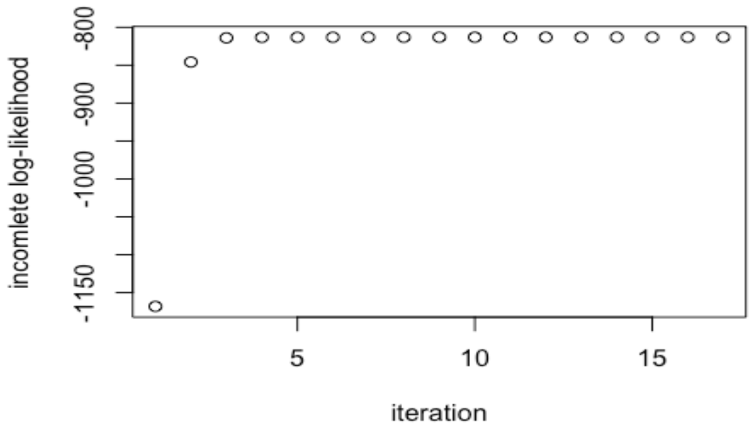
```
  p.new = colSums(P_ik*X)/(colSums(P_ik)*20)

  return(list(w.new=w.new, p.new=p.new))
}

EM1_1 = mixture.EM(X, w.init=c(0.3, 0.3, 0.4), p.init=c(0.2, 0.5, 0.7),
                   epsilon = 1e-5, max.iter=100)

print(paste("Estimate pi = (", round(EM1_1$w.curr[1],ROUNDING), ",",
            round(EM1_1$w.curr[2],ROUNDING),
            ",", round(EM1_1$w.curr[3],ROUNDING), ")", sep=""))

## [1] "Estimate pi = (0.119153640188926,0.284710722972906,0.596135636838168)"

print(paste("Estimate p = (", round(EM1_1$p.curr[1],ROUNDING), ",",
            round(EM1_1$p.curr[2],ROUNDING),
            ",", round(EM1_1$p.curr[3],ROUNDING), ")", sep=""))

## [1] "Estimate p = (0.0899609440376146,0.378959408458842,0.889988205308517)"

formatC(EM1_1$log_liks[length(EM1_1$log_liks)], digits = ROUNDING)

## [1] "-813.0135902939779271"

plot(EM1_1$log_liks, ylab='incomlete log-likelihood', xlab = 'iteration')
```




```
EM1_2 = mixture.EM(X, w.init=c(0.1, 0.2, 0.7), p.init=c(0.1, 0.3, 0.7),
                   epsilon = 1e-5, max.iter=100)

print(paste("Estimate pi = (", round(EM1_2$w.curr[1],ROUNDING), ",",
            round(EM1_2$w.curr[2],ROUNDING),
            ",", round(EM1_2$w.curr[3],ROUNDING), ")", sep=""))

## [1] "Estimate pi = (0.119074013242143,0.284787730895773,0.596138255862084)"

print(paste("Estimate p = (", round(EM1_2$p.curr[1],ROUNDING), ",",
            round(EM1_2$p.curr[2],ROUNDING),
            ",", round(EM1_2$p.curr[3],ROUNDING), ")", sep=""))

## [1] "Estimate p = (0.089888138447484,0.378906212317849,0.889987313569237)"

formatC(EM1_2$log_liks[length(EM1_2$log_liks)], digits = ROUNDING)
```

```
## [1] "-813.01359001659068326"
```

```r
plot(EM1_2$log_liks, ylab='incomlete log-likelihood', xlab = 'iteration')
```



Using the EM algorithm, the parameters which gave the highest incomplete log-likelihood were: pi = (0.119074013242143, 0.284787730895773, 0.596138255862084), p = (0.0898881384474844, 0.378906212317849, 0.889987313569237).

```
```

2.

a.

$$\mathbb{P}(x_1,\dots,x_n, z_1,\dots, z_n|\theta) = \prod_{i=1}^{300}\left[\mathbb{P}(x_i|z_i,\theta)\,\mathbb{P}(z_i|\theta)\right]\prod_{j=301}^{400}\mathbb{P}(x_j|\theta)$$

$$= \prod_{i=1}^{300}\prod_{k=1}^{3}\mathbb{P}(x_i|z_i=k,\theta)\,\mathbb{P}(z_i=k|\theta)]^{I(z_i=k)}\left[\prod_{j=301}^{400}\mathbb{P}(x_j|\theta)\right]$$

$$\log\left[\mathbb{P}(x_1,\dots,x_n,z_1,\dots,z_n|\theta)\right] = \sum_{i=1}^{n}\sum_{k=1}^{3} I(z_i=k)\left[\log\mathbb{P}(x_i|z_i=k,\theta)+\log\mathbb{P}(z_i=k|\theta)\right] + \sum_{j=301}^{400}\log(\mathbb{P}(x_j|\theta))$$

$$Q(\theta,\theta^\circ) = E_{z|x,\theta^\circ}\left[\log\mathbb{P}(x_1,\dots,x_n,z_1,\dots,z_n|\theta)\right]$$

$$= \sum_{i=1}^{300}\sum_{k=1}^{3}\mathbb{P}(z=k|x_i,\theta^\circ)\left[\log\mathbb{P}(x_i|z_i=k,\theta)+\log\mathbb{P}(z_i=k|\theta)\right] + \log\prod_{j=301}^{400}\mathbb{P}(x_j|\theta)$$

$$= \sum_{i=1}^{300}\sum_{k=1}^{3}\mathbb{P}(z_i=k|x_i,\theta^\circ)\left[\log\binom{20}{x_i}+x_i\log p_k+(20-x_i)\log(1-p_k) + \log\pi_k\right] + \log\prod_{j=301}^{400}\binom{20}{x_j}p_1^{x_j}(1-p_1)^{20-x_j}$$

$$= \sum_{i=1}^{300}\sum_{k=1}^{3}\mathbb{P}(z_i=k|x_i,\theta^\circ)\left[\log\binom{20}{x_i}+x_i\log p_k+(20-x_i)\log(1-p_k) + \log\pi_k\right] + \sum_{j=301}^{400}\left[\log\binom{20}{x_j}+x_j\log p_1+(20-x_j)\log(1-p_1)\right]$$

where $\pi_3 = \pi_1 - \pi_2$

b.

E step

Let $\theta^\circ = (\pi_1^\circ, \pi_2^\circ, p_1^\circ, p_2^\circ, p_3^\circ)$

$$\mathbb{P}(z_i=1|x_i,\theta^\circ) = \frac{\mathbb{P}(z_i=1,x_i|\theta^\circ)}{\mathbb{P}(x_i|\theta^\circ)}$$

$$= \frac{\mathbb{P}(x_i|z_i=1,\theta^\circ)\,\mathbb{P}(z_i=1|\theta^\circ)}{\mathbb{P}(x_i|z_i=1,\theta^\circ)\,\mathbb{P}(z_i=1|\theta^\circ)+\mathbb{P}(x_i|z_i=2,\theta^\circ)\,\mathbb{P}(z_i=2|\theta^\circ)+\mathbb{P}(x_i|z_i=3,\theta^\circ)\,\mathbb{P}(z_i=3|\theta^\circ)}$$

$$\mathbb{P}(z_i=2|x_i,\theta^\circ) = \frac{\mathbb{P}(z_i=2,x_i|\theta^\circ)}{\mathbb{P}(x_i|\theta^\circ)}$$

$$= \frac{\mathbb{P}(x_i|z_i=2,\theta^\circ)\,\mathbb{P}(z_i=2|\theta^\circ)}{\mathbb{P}(x_i|z_i=1,\theta^\circ)\,\mathbb{P}(z_i=1|\theta^\circ)+\mathbb{P}(x_i|z_i=2,\theta^\circ)\,\mathbb{P}(z_i=2|\theta^\circ)+\mathbb{P}(x_i|z_i=3,\theta^\circ)\,\mathbb{P}(z_i=3|\theta^\circ)}$$

$$\mathbb{P}(z_i=3|x_i,\theta^\circ) = 1-\mathbb{P}(z_i=1|x_i,\theta^\circ)-\mathbb{P}(z_i=2|x_i,\theta^\circ)$$

where $\mathbb{P}(x_i \mid z_i = k, \theta^\circ) = \binom{20}{x_i} (p_k^\circ)^{x_i} (1-p_k^\circ)^{20-x_i}$

$\mathbb{P}(z_i = 1 \mid \theta^\circ) = \pi_1^\circ$

$\mathbb{P}(z_i = 2 \mid \theta^\circ) = \pi_2^\circ$

$\mathbb{P}(z_i = 3 \mid \theta^\circ) = 1 - \pi_1^\circ - \pi_2^\circ$

M step

$$\frac{\partial Q(\theta, \theta^\circ)}{\partial \pi_1} = \sum_{i=1}^{300} \left[\frac{\mathbb{P}(z_i=1 \mid x_i, \theta^\circ)}{\pi_1} - \frac{\mathbb{P}(z_i=3 \mid x_i, \theta^\circ)}{1-\pi_1-\pi_2} \right]$$

$$\frac{d}{d\pi_1} \log(1-\pi_1-\pi_2) = \frac{-1}{1-\pi_1-\pi_2}$$

$$\sum_{i=1}^{300} \left[\frac{\mathbb{P}(z_i=1 \mid x_i, \theta^\circ)}{\pi_1} - \frac{\mathbb{P}(z_i=3 \mid x_i, \theta^\circ)}{1-\pi_1-\pi_2} \right] = 0$$

$$\Rightarrow \sum_{i=1}^{300} \left[\frac{(1-\pi_1-\pi_2)\mathbb{P}(z_i=1 \mid x_i, \theta^\circ) - \pi_1 \mathbb{P}(z_i=3 \mid x_i, \theta^\circ)}{\pi_1(1-\pi_1-\pi_2)} \right] = 0$$

$$\Rightarrow \sum_{i=1}^{300} \left[\mathbb{P}(z_i=1 \mid x_i, \theta^\circ) - \pi_1 (\mathbb{P}(z_i=1 \mid x_i, \theta^\circ) + \mathbb{P}(z_i=3 \mid x_i, \theta^\circ)) - \pi_2 \mathbb{P}(z_i=1 \mid x_i, \theta^\circ) \right] = 0$$

$$\Rightarrow \hat{\pi}_1 = \frac{1}{300} \sum_{i=1}^{300} \mathbb{P}(z_i=1 \mid x_i, \theta^\circ)$$

$$\Rightarrow \hat{\pi}_2 = \frac{1}{300} \sum_{i=1}^{300} \mathbb{P}(z_i=2 \mid x_i, \theta^\circ)$$

$$\Rightarrow \hat{\pi}_3 = 1 - \hat{\pi}_2 - \hat{\pi}_1$$

$k=1$

$$\frac{\partial Q(\theta,\theta^0)}{\partial p_1} = \sum_{i=1}^{300} \mathbb{P}(z_i=1 \mid x_i,\theta^0)\left[\frac{x_i}{p_1} - \frac{20-x_i}{1-p_1}\right] + \sum_{i=301}^{400}\left[\frac{x_i}{p_1} + \frac{20-x_i}{1-p_1}\right]$$

$$\Rightarrow \sum_{i=1}^{300} \mathbb{P}(z_i=1 \mid x_i,\theta^0)\left[\frac{x_i}{p_1} - \frac{20-x_i}{1-p_1}\right] + \sum_{i=301}^{400}\left[\frac{x_i}{p_1} + \frac{20-x_i}{1-p_1}\right] = 0$$

$$\Rightarrow \sum_{i=1}^{300} \mathbb{P}(z_i=1 \mid x_i,\theta^0)\left[(1-p_1)x_i - p_1(20-x_i)\right] + \sum_{i=301}^{400}\left[(1-p_1)x_i - p_1(20-x_i)\right] = 0$$

$$\Rightarrow \sum_{i=1}^{300} \mathbb{P}(z_i=1 \mid x_i,\theta^0)\left[x_i - 20p_1\right] + \sum_{i=301}^{400}\left[x_i - 20p_1\right] = 0$$

$$\Rightarrow \hat{p}_1 = \frac{\sum_{i=1}^{300} \mathbb{P}(z_i=1 \mid x_i,\theta^0)x_i + \sum_{i=301}^{400} x_i}{20\sum_{i=1}^{300} \mathbb{P}(z_i=1 \mid x_i,\theta^0) + 2000}$$

$k=2,3$

$$\frac{\partial Q(\theta,\theta^0)}{\partial p_k} = \sum_{i=1}^{300} \mathbb{P}(z_i=k \mid x_i,\theta^0)\left[\frac{x_i}{p_k} - \frac{20-x_i}{1-p_k}\right]$$

$$\Rightarrow \sum_{i=1}^{300} \mathbb{P}(z_i=k \mid x_i,\theta^0)\left[\frac{(1-p_k)x_i - p_k(20-x_i)}{p_k(1-p_k)}\right] = 0$$

$$\Rightarrow \sum_{i=1}^{300} \mathbb{P}(z_i=k \mid x_i,\theta^0)\left[(1-p_k)x_i - p_k(20-x_i)\right] = 0$$

$$\Rightarrow \sum_{i=1}^{300} \mathbb{P}(z_i=k \mid x_i,\theta^0)\left[x_i - 20p_k\right] = 0$$

$$\Rightarrow \hat{p}_k = \frac{\sum_{i=1}^{300} \mathbb{P}(z_i=k \mid x_i,\theta^0)x_i}{20\sum_{i=1}^{300} \mathbb{P}(z_i=k \mid x_i,\theta^0)}$$

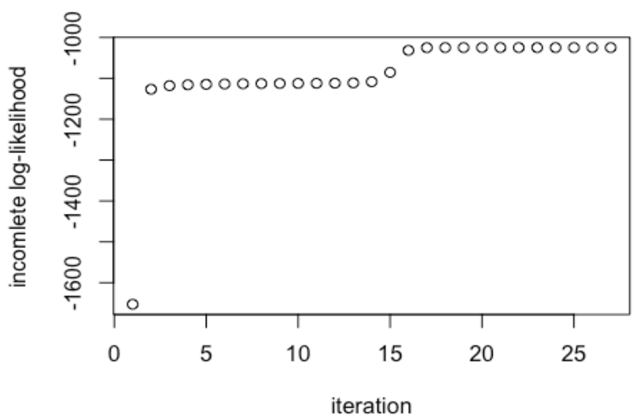
```
X.more = scan(file="assignment3_prob2.txt", what=double())
X.join = c(X, X.more)
```

Q2 c

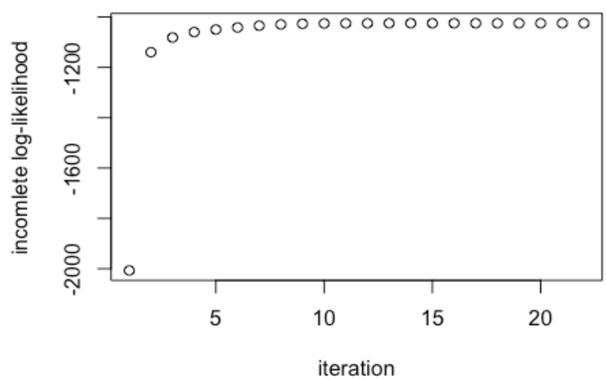
```
mixture.EM2 = function(X, X.more, w.init, p.init, epsilon=1e-5, max.iter=100) {

  # set initial values into current
  w.curr = w.init
  p.curr = p.init

  # setup list to store log likelihood - so to be able to plot
  log_liks = c()
  log_liks = c(log_liks, compute.log.lik2(X, X.more, w.curr, p.curr)$ill)

  # set initial values for the operation of this function
  delta.ll = 1
  n.iter = 1

  # while not reach max iteration and change in log likelihood
  # is larger than epsilon
  while((delta.ll > epsilon) & (n.iter <= max.iter)) {

    # run one EM iteration
    EM.out = EM.iter2(X, X.more, w.curr, p.curr)

    # put the new current values into the current
    w.curr = EM.out$w.new
    p.curr = EM.out$p.new

    # compute new log likelihood under new parameter estimates
    # (and add to list)
    log_liks = c(log_liks, compute.log.lik2(X, X.more, w.curr, p.curr)$ill)

    delta.ll = log_liks[length(log_liks)] - log_liks[length(log_liks)-1]

    n.iter = n.iter+1

  }

  return(list(w.curr=w.curr, p.curr=p.curr, log_liks=log_liks))
}


# for each sample Xi, compute P(Xi, Zi=k) - altered for purpose of Q2
compute.prob.x.z2 = function(X, X.more, w.curr, p.curr) {

  # for each sample Xi, compute P(Xi, Zi=k)
  L = matrix(NA, nrow=length(X), ncol=length(w.curr))

  for (k in seq_len(ncol(L))) {

    L[,k] = dbinom(X, size=20, prob = p.curr[k])*w.curr[k]

  }

  # for each sample Xi, compute P(Xi, Zi=1) - change made for X301-400
  L.more = matrix(NA, nrow=length(X.more), ncol=1)
```

```
  L.more[,1] = dbinom(X.more, size=20, prob = p.curr[1])


  return(list(prob.x.z = L, prob.x.z.more = L.more))
}


# compute incomplete log likelihood
compute.log.lik2 = function(X, X.more, w.curr, p.curr) {

  # for each sample, compute P(Xi, Zi=k) using helper
  prob.x.z2 = compute.prob.x.z2(X, X.more, w.curr, p.curr)

  prob.x.z = prob.x.z2$prob.x.z
  prob.x.z.more = prob.x.z2$prob.x.z.more # change made for X301-400

  # get incomplete log likelihoods (each instance probabilistic weighted
  # between states and then logs taken, then summed)
  ill = sum(log(rowSums(prob.x.z)))
  ill.more = sum(log(rowSums(prob.x.z.more))) # change made for X301-400

  ill = ill+ill.more

  return(list(ill=ill))

}


# compute one EM step (update values)
EM.iter2 = function(X, X.more, w.curr, p.curr) {

  # E

  # for each sample Xi, copute P(Xi, Zi=k)
  prob.x.z2 = compute.prob.x.z2(X, X.more, w.curr, p.curr)

  prob.x.z = prob.x.z2$prob.x.z
  prob.x.z.more = prob.x.z2$prob.x.z.more # change made for X301-400


  # compute P(Zi=k | Xi) = P(Xi, Zi=k)/P(Xi)
  P_ik = prob.x.z / (rowSums(prob.x.z))

  # M
  indicator = c(1,0,0) # only activate for k=1, ignore for k=2;3

  w.new = colSums(P_ik)/(sum(P_ik)) # X301-400 d.n. affect weights
  p.new = (colSums(P_ik*X)+(indicator * sum(X.more))) /
    (colSums(P_ik)*20 + (indicator*2000)) # change made for X301-400

  return(list(w.new=w.new, p.new=p.new))
}

EM2_1 = mixture.EM2(X, X.more, w.init=c(0.3, 0.3, 0.4),
                    p.init=c(0.2, 0.5, 0.7), epsilon = 1e-5, max.iter=100)

print(paste("Estimate pi = (", round(EM2_1$w.curr[1],ROUNDING), ",",
            round(EM2_1$w.curr[2],ROUNDING),
            ",", round(EM2_1$w.curr[3],ROUNDING), ")", sep=""))
```

```
## [1] "Estimate pi = (0.278620486737536,0.125841600718041,0.595537912544423)"

print(paste("Estimate p = (", round(EM2_1$p.curr[1],ROUNDING), ",",
            round(EM2_1$p.curr[2],ROUNDING),
            ",", round(EM2_1$p.curr[3],ROUNDING), ")", sep=""))

## [1] "Estimate p = (0.393549436273744,0.0955640373232472,0.8901867983395)"

formatC(EM2_1$log_liks[length(EM2_1$log_liks)], digits = ROUNDING)

## [1] "-1024.8063167941156735"

plot(EM2_1$log_liks, ylab='incomlete log-likelihood', xlab = 'iteration')
```




```
EM2_2 = mixture.EM2(X, X.more, w.init=c(0.1, 0.2, 0.7),
                    p.init=c(0.1, 0.3, 0.7), epsilon = 1e-5, max.iter=100)

print(paste("Estimate pi = (", round(EM2_2$w.curr[1],ROUNDING), ",",
            round(EM2_2$w.curr[2],ROUNDING),
            ",", round(EM2_2$w.curr[3],ROUNDING), ")", sep=""))

## [1] "Estimate pi = (0.278550191216451,0.125912787721757,0.595537021061791)"

print(paste("Estimate p = (", round(EM2_2$p.curr[1],ROUNDING), ",",
            round(EM2_2$p.curr[2],ROUNDING),
            ",", round(EM2_2$p.curr[3],ROUNDING), ")", sep=""))

## [1] "Estimate p = (0.39357074021193,0.0956312030153907,0.890187071845526)"

formatC(EM2_2$log_liks[length(EM2_2$log_liks)], digits = ROUNDING)

## [1] "-1024.8063166078052291"

plot(EM2_2$log_liks, ylab='incomlete log-likelihood', xlab = 'iteration')
```



Using the EM algorithm, the parameters which gave the highest incomplete log-likelihood were: pi = (0.278550191216451, 0.125912787721757, 0.595537021061791), p = (0.39357074021193, 0.0956312030153907, 0.890187071845526).