Jairo Pena

Solutions to 5106 HW6

1. Solution:

Let $X = (x_1, \ldots, x_4) \stackrel{d}{=} Mult(n; 0.25\theta, 0.25(2+\theta), 0.5(1-2\theta), 0.5\theta)$.

Suppose that $x_2 = y_1 + y_2$, where $(y_1, y_2) \stackrel{d}{=} Bin(x_2; \frac{2}{2+\theta}, \frac{\theta}{2+\theta})$.

Then, the likelihood function of X is given by

$$f(X|\theta) = \binom{n}{x_1, y_1, y_2, x_3, x_4}(0.25\theta)^{x_1} 0.5^{y_1} (0.25\theta)^{y_2} (0.5(1-2\theta))^{x_3} (0.5\theta)^{x_4}$$

And the log-likelihood function is given by

$$\begin{aligned} l(X|\theta) &= \ln \binom{n}{x_1, y_1, y_2, x_3, x_4} + x_1 \ln(0.25\theta) + y_1 \ln(0.5) + y_2 \ln(0.25\theta) + x_3 \ln(0.5(1-2\theta)) + x_4 \ln(0.5\theta) \\ &= (x_1 + y_2 + x_4) \ln(\theta) + x_3 \ln(1-2\theta) + C \end{aligned}$$

E step:

Let $Q(\theta|\theta_k, x) = E_y(l(X|\theta)|\theta_k, x) = (x_1 + E_y(y_2|\theta_k, x) + x_4) \ln(\theta) + x_3 \ln(1-2\theta)$.

Since y_2 is a binomial distribution, its expected value is given by the following sum:

$$\begin{aligned} E_y(y_2|\theta_k, x) &= \sum_{y_2=0}^{x_2} y_2 \binom{x_2}{y_2} \left(\frac{2}{2+\theta_k}\right)^{x_2-y_2} \left(\frac{\theta_k}{2+\theta_k}\right)^{y_2} \\ &= x_2 \left(\frac{\theta_k}{2+\theta_k}\right) \sum_{y_2=1}^{x_2} \binom{x_2-1}{y_2-1} \left(\frac{2}{2+\theta_k}\right)^{x_2-1-(y_2-1)} \left(\frac{\theta_k}{2+\theta_k}\right)^{y_2-1} \\ &= x_2 \left(\frac{\theta_k}{2+\theta_k}\right) \cdot 1 \end{aligned}$$

And we get $Q(\theta|\theta_k, x) = \left(x_1 + x_2 \left(\frac{\theta_k}{2+\theta_k}\right) + x_4\right) \ln(\theta) + x_3 \ln(1-2\theta)$.

M step:

We are now concerned with finding the argument $\theta \in (0, 0.5)$ that maximizes $Q(\theta|\theta_k, x)$. Notice first and second order conditions guarantee the existence of a maximizing argument:

Let $a = x_1 + \frac{x_2\theta_k}{2+\theta_k} + x_4$, so that $Q(\theta|\theta_k, x) = a \ln(\theta) + x_3 \ln(1-2\theta)$

$$\frac{\partial Q}{\partial \theta} = \frac{a}{\theta} - \frac{2x_3}{1-\theta} = 0 \Rightarrow \theta = \theta_{k+1} := \frac{a}{a + 2x_3} = \frac{x_1 + \frac{x_2\theta_k}{2+\theta_k} + x_4}{x_1 + \frac{x_2\theta_k}{2+\theta_k} + x_4 + 2x_3}$$

$$\frac{\partial^2 Q}{\partial \theta^2} = -\frac{a}{\theta^2} - \frac{4x_3}{(1-\theta)^2} < 0$$

Matlab Code:

```
x1 = 6; x2 = 52; x3 = 28; x4 = 14;
t=@(x)(x1+x2*x/(2+x)+x4)/(x1+x2*x/(2+x)+x4+2*x3);
eps = 10^(-6); theta = [0.25]; i = 1;
theta = [theta,t(theta(i))];
while(abs(theta(i+1)-theta(i))>eps)
    theta = [theta,t(theta(i+1))];
    i = i+1;
end
```

Python Code:

```
import numpy as np
x1 = 6;x2 = 52;x3 = 28;x4 = 14;theta = {};theta[0] = 0.4;
t = lambda x: (x1+x2*x/(2+x)+x4)/(x1+x2*x/(2+x)+x4+2*x3);
eps = 10**(-6); i = 1; theta[1] = t(theta[0]);
while abs(theta[i]-theta[i-1])>eps:
    theta[i+1] = t(theta[i]);
    i += 1;
```

Results:

Figure 1: MATLAB results for $\theta_0 = 0.25$

```
>> theta'

ans =

   0.250000000000000
   0.315217391304348
   0.325949367088608
   0.327626764244642
   0.327886792452830
   0.327927050195790
   0.327933281691736
   0.327934246235389
```

Figure 2: Python results for $\theta_0 = 0.4$

```
In [6]: theta
Out[6]:
{0: 0.4,
 1: 0.33858267716535434,
 2: 0.3295711060948081,
 3: 0.3281874825630057,
 4: 0.32797358601991783,
 5: 0.32794048455930785,
 6: 0.3279353611231583,
 7: 0.327934568098935}
```

EM algorithm takes 7 steps to find an estimation of θ from both a starting value of 0.25 and 0.4.

2.Solution:

a)

Data-set hw6_2_data1 (Figure 3) shows two disjoint bell shaped distributions. It appears the left distribution corresponds to 33% of the sample, its centered at 0 and has a standard deviation of 1. The right one would correspond to 66% of the sample, appears to be centered at 5 and has a standard deviation of 0.5.

On the other hand, data-set hw6_2_data2 (Figure 4) doesn't show two bell shaped distributions as clearly as the previous data-set. If there were two distributions superimposed to each other, the left distribution would correspond roughly to 15% of the sample, it would be centered at 0 and would have a a standard deviation of 0.5. The right one would correspond to 85% of the sample, it would be centered at 3 and it would have a standard deviation of 1.

Figure 3: Histogram for hw6_2_data1

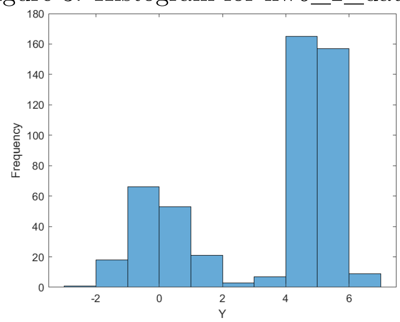
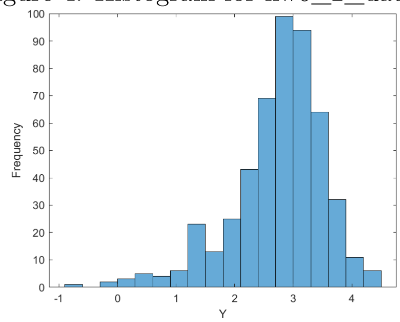


Figure 4: Histogram for hw6_2_data2



b)

Using the previous guesses as initial values, the EM algorithm appeared to converge after two iterations for data-set hw6_2_data1 (judging by stability in the log-likelihood evolution). The results were:

$$(\alpha_1, \mu 1, \sigma_1, \alpha_2, \mu 2, \sigma_2) = (0.3241, 0.0269, 0.9195, 0.6759, 4.9724, 0.5178)$$

As for data-set hw6_2_data2, EM algorithm appeared to converge after 30 iterations (judging by stability in the log-likelihood evolution). The results were

$$(\alpha_1, \mu 1, \sigma_1, \alpha_2, \mu 2, \sigma_2) = (0.0508, 0.7017, 0.5923, 0.9492, 2.8661, 0.6359)$$

c)

The next figures show how stability in the EM algorithm was achieved after a few iterations.

Figure 5: Histogram for hw6_2_data1

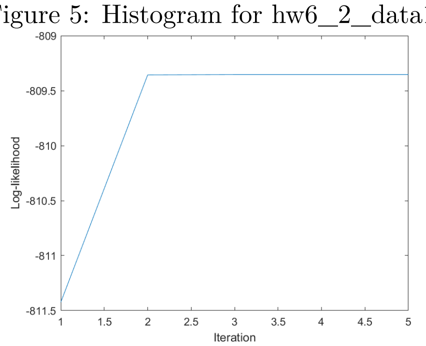
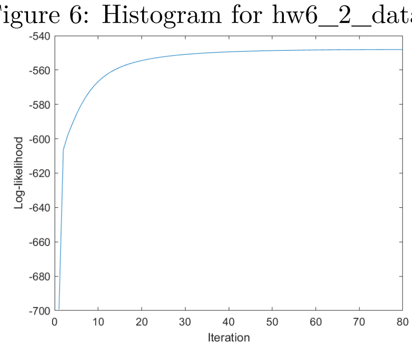


Figure 6: Histogram for hw6_2_data2



Matlab Code:

```
load hw6_2_data1 %load hw6_2_data1
a1 = [0.3]; a2 = [1-a1]; %a1 = [0.15]; a2 = [1-a1];
mu1 = [0]; mu2 = [5]; % mu1 = [0]; mu2 = [3];
s1 = [1]; s2 = [0.5]; % s1 = [0.5]; s2 = [1];
lf = [];
for i = 1:5
    f1 = normpdf(Y,mu1(i),s1(i));
    f2 = normpdf(Y,mu2(i),s2(i));
    p1 = a1(i)*f1./(a1(i)*f1+a2(i)*f2);
    p2 = a2(i)*f2./(a1(i)*f1+a2(i)*f2);
    lf = [lf,sum(log(a1(1)*f1+a2(1)*f2))];
    a1 = [a1,mean(p1)]; a2 = [a2,mean(p2)];
    mu1 = [mu1,sum(Y.*p1)/sum(p1)];
    mu2 = [mu2,sum(Y.*p2)/sum(p2)];
    s1=[s1,sqrt(sum(p1.*(Y-mu1(i)).*(Y-mu1(i)))/sum(p1))];
    s2=[s2,sqrt(sum(p2.*(Y-mu2(i)).*(Y-mu2(i)))/sum(p2))];
end
histogram(Y); xlabel('Y'); ylabel('Frequency');
plot(lf); xlabel('Iteration'); ylabel('Log-likelihood');
```