Part 1:

question 1: M step update rules

$$p(x^{(i)} | z=k) = \prod_{j=1}^{D} p(x_j^{(i)} | z=k) = \prod_{j=1}^{D} \theta_{kj}^{x_j^{(i)}} (1-\theta_{k,j})^{1-x_j^{(i)}}$$

$$l = \sum_{i=1}^{N} \sum_{k=1}^{k} r_k^{(i)} [\log p(z^{(i)} = k) + \log P(x^{(i)} | z^{(i)} = k)] + \log p(\pi) + \log p(\theta) \quad (6)$$

$z \sim \text{Multinomial}(\pi)$

$x_j | z = k \sim \text{Bernoulli}(\theta_{k,j})$

$p(z^i = k) = \pi_k$

$p(\theta_{k,j}) \propto \theta_{k,j}^{a-1} (1-\theta_{k,j})^{b-1}$

$p(\pi) \propto \pi_1^{a_1-1} \pi_2^{a_2-1} \cdots \pi_k^{a_k-1} \quad (a_1 = a_2 \cdots a_k) = A$

$\pi = (\pi_1, \pi_2, \cdots \pi_k)$

$$\Theta = \begin{bmatrix} \theta_{11} & & \theta_{1,D} \\ & \ddots & \\ \theta_{k,1} & & \theta_{k,D} \end{bmatrix}$$

step 1: Calculate log to each part in equation (6)

$\log P(z^{(i)} = k) = \log \pi_k$

$\log P(x^{(i)} | z^{(i)} = k) = \sum_{j=1}^{D} x_j^{(i)} \log \theta_{k,j} + \sum_{j=1}^{D} (1-x_j^{(i)}) \log(1-\theta_{k,j})$

$\log p(\pi) = \log(\prod_{k=1}^{K} \pi_k^{A-1}) = (A-1) \sum_{k=1}^{K} \log \pi_k$

$\log P(\Theta) = \log\left(\prod_{k=1}^{K} \prod_{j=1}^{D} (\theta_{k,j})^{a-1} (1-\theta_{k,j})^{b-1}\right)$

$= (a-1) \sum_{k=1}^{k} \sum_{j=1}^{D} \log \theta_{k,j} + (b-1) \sum_{k=1}^{k} \sum_{j=1}^{D} \log(1-\theta_{k,j})$

$\therefore$ Eqn (6) $\Rightarrow \sum_{i=1}^{N} \sum_{k=1}^{K} r_k^{(i)} \log P(z^{(i)} = k) + \sum_{i=1}^{N} \sum_{k=1}^{K} r_k^{(i)} \log p(x^{(i)} | z^{(i)} = k) + \log p(\pi)$
$+ \log p(\Theta)$

$$\Rightarrow \sum_{i=1}^{N} \sum_{k=1}^{k} r_k^{(i)} \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{k} r_k^{(i)} \left(\sum_{j=1}^{D} x_j^{(i)} \log \theta_{k,j} + \sum_{j=1}^{D} (1-x_j^{(i)}) \log(1-\theta_{k,j}) \right)$$
$$+ (A-1) \sum_{k=1}^{k} \log \pi_k + (a-1) \sum_{k=1}^{k} \sum_{j=1}^{D} \log \theta_{k,j} + (b-1) \sum_{k=1}^{k} \sum_{j=1}^{D} \log(1-\theta_{k,j})$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{k} r_k^{(i)} \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{k} \sum_{j=1}^{D} r_k^{(i)} x_j^{(i)} \log \theta_{k,j} + \sum_{i=1}^{N} \sum_{k=1}^{k} \sum_{j=1}^{D} r_k^{(i)} (1-x_j^{(i)}) \log(1-\theta_{k,j})$$
$$+ (A-1) \sum_{k=1}^{k} \log \pi_k + \sum_{k=1}^{k} \sum_{j=1}^{D} \log \theta_{k,j} + (b-1) \sum_{k=1}^{k} \sum_{j=1}^{D} \log(1-\theta_{k,j})$$

Maximizing π_k (π_k is one of set of π_k, $k=k$)

$$\frac{\partial L}{\partial \pi_k} = \frac{1}{\pi_k}\sum_{i=1}^{N} r_k^{(i)} + A-1 \cdot \frac{1}{\pi_k}$$

$\therefore$ we need set $\frac{1}{\pi_k}\sum_{i=1}^{N} r_k^{(i)} + \frac{A-1}{\pi_k} = 0$

$$\frac{1}{\pi_k}\cdot\left(\sum_{i=1}^{N} r_k^{(i)} + A-1\right) = 0$$

$$\sum_{k=1}^{k} \pi_k = 1$$

we need to set the partial derivative to zero, and then it cannot solve the problem.
From lecture resource, I will add the $\lambda\left(1-\sum_{k=1}^{k}\pi_k\right)$ to eqn (b)
Then, the partial derivative equation becomes

$$\frac{1}{\pi_k}\cdot\left(\sum_{i=1}^{N} r_k^{(i)} + A-1\right) = \lambda$$

$$\therefore \pi_k = \frac{1}{\lambda}\left(\sum_{i=1}^{N} r_k^{(i)} + A-1\right)$$

$$\because \sum_{k=1}^{k} \pi_k = 1$$

$$\therefore \sum_{k=1}^{k} \frac{1}{\lambda}\left(\sum_{i=1}^{N} r_k^{(i)} + A-1\right) = 1$$

$$\therefore \lambda = \sum_{k=1}^{k}\left(\sum_{i=1}^{N} r_k^{(i)} + A-1\right)$$

$$\therefore \pi_k = \frac{\sum_{i=1}^{N} r_k^{(i)} + A-1}{\sum_{k=1}^{k}\left(\sum_{i=1}^{N} r_k^{(i)} + A-1\right)}$$

$$\boxed{= \frac{\sum r_k^{(i)} + A-1}{\left(\sum_{k=1}^{k}\sum_{i=1}^{N} r_k^{(i)}\right) + kA - K}}$$

Maximizing $\theta_{k,j}$

$$\frac{\partial L}{\partial \theta_{k,j}} = \sum_{i=1}^{N}\frac{r_k^{(i)} x_j^{(i)}}{\theta_{k,j}} - \sum_{i=1}^{N}\frac{r_k^{(i)}(1-x_j^{(i)})}{1-\theta_{k,j}} + \frac{a-1}{\theta_{kj}} - \frac{(b-1)}{(1-\theta_{kj})} = 0$$

$$(1-\theta_{kj})\left(\sum_{i=1}^{N} r_k^{(i)}\cdot x_j^{(i)} + a-1\right) = \theta_{kj}\left(\sum_{i=1}^{N} r_k^{(i)}(1-x_j^{(i)}) + (b-1)\right)$$

$$\boxed{\theta_{k,j} = \frac{\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)} + (a-1)}{\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)} + (a-1) + \sum_{i=1}^{N} r_k^{(i)}(1-x_j^{(i)}) + (b-1)}}$$

2:

```
pi[0] 0.085
pi[1] 0.13
theta[0, 239] 0.6427106227106227
theta[3, 298] 0.465736124958458
```

Part 2

1. $Pr(z=k \mid x) = ?$

Answer:

$$p(z=k \mid x) = \frac{p(z=k) \cdot p(x \mid z=k)}{p(x)}$$

from Part 1, we know $\begin{cases} p(x \mid z=k) = \prod_{j=1}^{p} (\theta_{kj})^{x_j} (1-\theta_{kj})^{(1-x_j)} \\ p(z=k) = \pi_k \end{cases}$

$$p(x) = \sum_{k=1}^{K} p(x, z=k) = \sum_{k=1}^{K} p(z=k)\, p(x \mid z=k)$$

$$= \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} (\theta_{kj})^{x_j} (1-\theta_{kj})^{(1-x_j)}$$

$$\therefore \quad Pr(z=k \mid x) = \frac{p(z=k)\, p(x \mid z=k)}{p(x)}$$

$$= \frac{\pi_k \prod_{j=1}^{p} (\theta_{kj})^{x_j} (1-\theta_{kj})^{(1-x_j)}}{\sum_{k}^{K} \pi_k \prod_{j=1}^{D} (\theta_{kj})^{x_j}\, \theta_{kj}^{(1-x_j)}}$$

3:

```
R[0, 2] 0.17488951492117252
R[1, 0] 0.6885376761092291
P[0, 183] 0.6516151998131037
P[2, 628] 0.4740801724913304
```

Part 3.

1) From question 1:

$$\theta_{k,j} = \frac{\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)} + (a-1)}{\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)} + (a-1) + \sum_{j=1}^{N} r_k^{(i)} (1 - x_j^{(i)}) + (b-1)}$$

when $a = b = 1$

$$\theta_{kj} = \frac{\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)}}{\sum_{j=1}^{N} r_k^{(i)} x_j^{(i)} + \sum_{i=1}^{N} r_k^{(i)} [1 - x_j^{(i)}]}$$

Therefore, if a pixel is always 0 in the training set, which means the x_j will always be zero. Then $\theta_{k,j}$ will always be zero.
From equation (2): $p(x^{(i)} | z=k) = \prod_{j=1}^{D} p(x_j^{(i)} | z=k)$ will always be zero when $\theta_{k,j}$ is equal to zero.

2) Even through we have only 10 latent components, we are able to get the learned θ_{kj}, and π_k from M step. Then it can use to assign the responsibilities (r_k) from E step. The Expeuation Maximum algorithm still works even if we have few latent components.

3) No, I don't think the model thinks 1's are far more common than 8's. The higher log probability only means the model is more confident about predicting an image when it would be 1's than it would be 8's. The reason could be that 8's is harder to predict. Because, 5 and 3 looks similar to 8. I feel the model can predict 1 easily because the shape of 1 is unique.