Part 1

1. $f(\pi, \theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} r_k^{(i)} \left[\log P(z^{(i)} = k) + \log p(x^{(i)} \mid z^{(i)} = k) \right] + \log P(\pi) + \log P(\theta)$

①

$$\sum_{i=1}^{N} \sum_{k=1}^{K} r_k^{(i)} \left[\log P(z^{(i)} = k) + \log P(x^{(i)} \mid z^{(i)} = k) \right]$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} r_k^{(i)} \left[\log \pi_k + \log \left(\prod_{j=1}^{D} \theta_{k,j}^{x_j^{(i)}} (1 - \theta_{k,j})^{1 - x_j^{(i)}} \right) \right]$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} r_k^{(i)} \left[\log \pi_k + \sum_{j=1}^{D} x_j^{(i)} \log \theta_{k,j} + \sum_{j=1}^{D} (1 - x_j^{(i)}) \log (1 - \theta_{k,j}) \right]$$

② $\log P(\pi)$

$$= \log \left(\pi_1^{a_1 - 1} \pi_2^{a_2 - 1} \cdots \pi_k^{a_k - 1} \right) = \sum_{j=1}^{K} \log \pi_j^{a_j - 1}$$

③ $\log P(\theta)$

$$= \log \Big(P(\theta_{1,1}) \cdot P(\theta_{1,2}) \cdots P(\theta_{1,D}) \cdot P(\theta_{2,1}) \cdot P(\theta_{2,2}) \cdots P(\theta_{2,D}) \cdot$$
$$P(\theta_{k,1}) \cdot P(\theta_{k,2}) \cdots P(\theta_{k,D}) \Big)$$

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

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

$$\Rightarrow f(\pi, \theta) = ① + ② + ③$$

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

$$\Rightarrow \sum_{k=1}^{K} \pi_k - 1 = 0 = g(\pi)$$

$$\Rightarrow L(\theta, \pi, \lambda) = f(\theta, \pi) - \lambda g(\pi)$$

$$= ① + ② + ③ + \lambda - \lambda \sum_{k=1}^{K} \pi_k$$

Let take arbitrary k, j

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

$$\frac{\partial f}{\partial \theta_{k,j}} = \sum_{i=1}^{N} r_k^{(i)}\left[\left(x_j^{(i)}\frac{1}{\theta_{k,j}}\right) - \left(1 - x_j^{(i)}\right)\frac{1}{1-\theta_{k,j}}\right] +$$

$$(a-1)\frac{1}{\theta_{k,j}} + (b-1)\cdot\frac{1}{1-\theta_{k,j}}\cdot(-1)$$

$$\Rightarrow \begin{cases} \dfrac{\partial L}{\partial \pi_k} = \dfrac{\partial f}{\partial \pi_k} - \lambda = 0 & (i) \\ \dfrac{\partial L}{\partial \theta_{k,j}} = \dfrac{\partial f}{\partial \theta_{k,j}} = 0 & (ii) \\ \dfrac{\partial L}{\partial \lambda} = 1 - \sum_{k'=1}^{k} \pi_{k'} = 0 & (iii) \end{cases}$$

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

with (iii)

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

$$\frac{1}{\lambda} \sum_{k=1}^{k} \sum_{i=1}^{N} r_k^{(i)} + a_k' - 1 = 1$$

$$\Rightarrow \quad \lambda = \sum_{k=1}^{k} \sum_{i=1}^{N} r_k^{(i)} + a_k' - 1$$

$$\Rightarrow \quad \hat{\pi}_k = \frac{1}{\sum_{k=1}^{k} \sum_{i=1}^{N} r_{k'}^{(i)} + a_{k'} - 1} \left(\sum_{i=1}^{N} r_k^{(i)} + a_k - 1\right)$$

from (ii)

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

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

$$\Rightarrow (1-\theta_{kj})\cdot M = \theta_{kj} N$$

$$\Rightarrow \theta_{kj} = \frac{M}{M+N}$$

where $M = \sum_{i=1}^{N} r_k^{(i)} x_j^{(i)} + a - 1$

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

$$\Rightarrow \theta_{kj} = \frac{\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)} + a - 1}{\sum_{i=1}^{N} r_k^{(i)} + a + b - 2} \qquad \square$$

2.

```
/Library/Frameworks/Python.framework/Versions/3.6/bin/python3.6 /Users/jerry/PycharmProjects/CSC411/hw6/mixture.py
pi[0] 0.08499999999999992
pi[1] 0.12999999999999987
theta[0, 239] 0.6427106227106232
theta[3, 298] 0.46573612495845823

Process finished with exit code 0
```

Part 2

1. $P(z=k \mid x_{obs})$

$$= \frac{P(z=k) \cdot P(x_{obs} \mid z=k)}{P(x_{obs})}$$

$$= \frac{P(z=k) \cdot P(x_{obs} \mid z=k)}{\sum_{k'=1}^{K} P(x_{obs} \mid z=k')}$$

$$= \frac{\pi_k \cdot \prod_{j}^{D_{obs}} P(x_j \mid z=k)}{\sum_{k'=1}^{K} \prod_{j}^{D_{obs}} P(x_j \mid z=k')}$$

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

where D_{obs} is the set of observed dimension.

2.

```
/Library/Frameworks/Python.framework/Versions/3.6/bin/python3.6 /Users/jerry/PycharmProjects/CSC411/hw6/mixture.py
R[0, 2] 0.1748895149211729
R[1, 0] 0.6885376761092292
P[0, 183] 0.6516151998131037
P[2, 628] 0.4740801724913301

Process finished with exit code 0
```

Part 3.

1. If $a=b=1$, the MAP result is equal MLE result.

The reason of the problem is that if a pixel in all training set are all 0, then the learnt θs related to that pixel will be all 0 which will lead 0 probability allocation in test set.

2.

Both supervised model and unsupervised model try to optimize the EM lower bound of likelihood. But there are mismatch for each row of R matrix where there is uncertainty in unsupervised one, one-hot vector in supervised one which lead average log probabilities difference.

3. This doesn't mean more 1 will be generated than 8. The reason is that there are more variation in 8's image than 1's image.