Part 1

1.

update π

Lagrangian for π function:

$$\mathcal{L}_\pi = \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} \log \pi_k + \lambda\left(1 - \sum_{k=1}^{K}\pi_k\right) + \sum_{k=1}^{K}(a_k - 1)\log \pi_k \quad (1)$$

Setting derivation to 0:

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \sum_{i=1}^{N} r_k^{(i)} - \lambda + \frac{a_k - 1}{\pi_k} = 0$$

$$\lambda = \frac{a_k - 1 + \sum_{i=1}^{N} r_k^{(i)}}{\pi_k} \quad (2) \quad \text{and} \quad \pi_k = \frac{a_k - 1 + \sum_{i=1}^{N} r_k^{(i)}}{\lambda} \quad (3)$$

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

$$\lambda = \sum_{k=1}^{K}\left(a_k - 1 + \sum_{i=1}^{N} r_k^{(i)}\right) \rightarrow \quad \text{by equation (2) we can get:}$$

$$\frac{a_k - 1 + \sum_{i=1}^{N} r_k^{(i)}}{\pi_k} = \sum_{k'=1}^{K}\left(a_{k'} - 1 + \sum_{i=1}^{N} r_{k'}^{(i)}\right)$$

$$\pi_k = \frac{a_k - 1 + \sum_{i=1}^{N} r_k^{(i)}}{\sum_{k'=1}^{K}\left(a_{k'} - 1 + \sum_{i=1}^{N} r_{k'}^{(i)}\right)} \quad (\text{update rule for } \pi_k)$$

thus:

$$\pi_k \leftarrow \frac{a_k - 1 + \sum_{i=1}^{N} r_k^{(i)}}{\sum_{k'=1}^{K}\left(a_{k'} - 1 + \sum_{i=1}^{N} r_{k'}^{(i)}\right)}$$

update θ:

Lagrangian for θ function:

$$\mathcal{L}_\theta = \sum_{i=1}^{N} r_k^{(i)} \log \theta_{k,j}^{x_j} \log(1 - \theta_{k,j})^{1 - x_j} + \log \theta_{kj}^{a-1} \log(1 - \theta_{k,j})^{b-1}$$

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

Calculating partial derivative of θ:

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

re-arranging:

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

$$\left(a + b - 2 + \sum_{i=1}^{N} r_k^{(i)}\right) \theta_{k,j} = a - 1 + \sum_{i=1}^{N} r_k^{(i)} x_j^{(i)}$$

$$\theta_{k,j} = \frac{a - 1 + \sum_{i=1}^{N} r_k^{(i)} x_j^{(i)}}{a + b - 2 + \sum_{i=1}^{N} r_k^{(i)}} \qquad \text{(update rules for } \theta\text{)}$$

thus:

$$\theta_{k,j} \leftarrow \frac{a - 1 + \sum_{i=1}^{N} r_k^{(i)} x_j^{(i)}}{a + b - 2 + \sum_{i=1}^{N} r_k^{(i)}}$$

2.

pi[0] 0.085
pi[1] 0.13
theta[0, 239] 0.642710622711
theta[3, 298] 0.465736124958

Part 2

1.

$$Pr(z=k \mid x^{(i)}) = \frac{P(z=k, x^{(i)})}{P(x^{(i)})}$$

$$= \frac{P(z=k)\, P(m^{(i)}, x^{(i)} \mid z=k)}{\sum_{k'=1}^{K} P(z=k')\, P(m^{(i)}, x^{(i)} \mid z=k')}$$

$$= \frac{\pi_k \prod_{j=1}^{D} P(m_j^{(i)}, x_j^{(i)} \mid z=k)}{\sum_{k'=1}^{K} \pi_{k'} \prod_{j=1}^{D} P(m_j^{(i)}, x_j^{(i)} \mid z=k')}$$

$$= \frac{\pi_k \prod_{j=1}^{D} \theta_{k,j}^{m_j^{(i)} x_j^{(i)}} (1-\theta_{k,j})^{m_j^{(i)}(1-x_j^{(i)})}}{\sum_{k'=1}^{K} \pi_{k'} \prod_{j=1}^{D} \theta_{k',j}^{m_j^{(i)} x_j^{(i)}} (1-\theta_{k',j})^{m_j^{(i)}(1-x_j^{(i)})}}$$

2. See in mixture.py

3.

R[0, 2] 0.174889514921
R[1, 0] 0.688537676109
P[0, 183] 0.651615199813
P[2, 628] 0.474080172491

Part 3.

1. if $a = b = 1$:

$$\theta_{k,j} = \frac{\sum_{i=1}^{N} r_k^{(i)} x_j^{(i)} + a - 1}{\sum_{i=1}^{N} r_k^{(i)} + a + b - 2}$$

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

if a pixel is always 0 in the training set, the numerator will be 0 and θ will be 0 or 1. This means pixel will be assigned a probability 0 or 1. Therefore, it will cause data sparsity and it is not a good model design.

2.

Part1 model has only 10 cases. It is not sufficient to model the variation. Part 2 model has 10 times more cases than part1 model. Part 2 model is more accurate and has better performance.

3.

No. Since we are only given top half of each image to predict the digit, There is no digit has similar shape to the number 1. However, the number 8's top half is very much similar to the number 9. The number 8 might predict as the number 9. Thus, the log probability of this image being 8 or 9 are very close but it is smaller than the log probability of 1