Part 1.

1.

$$f = \sum\sum r_k^{(i)} [\log \Pr(z^i = k) + \log p(x^i \mid z^i = k)] + \log p(\pi) + \log p(\theta)$$

$$= \sum\sum r_k^{(i)} [\log(\pi_k) + \log(\prod_{j=1}^{D} \theta_{k,j}^{x_j^i} (1 - \theta_{k,j})^{1-x_j^{(i)}})] + \log p(\pi) + \log p(\theta)$$

$$= \sum\sum r_k^i (\log(\pi_k) + \sum x_j^i \log(\theta_{k,j}) + \sum (1-x_j^i) \log(1-\theta_{k,j})) + \log p(\pi) + \log p(\theta)$$

$$\frac{\partial f}{\partial \pi_k} = \frac{\sum\sum r_k^i \log(\pi_k) + \log p(\pi)}{\partial \pi_k}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} r_k^i \frac{1}{\pi_k} + \frac{\sum^{K} (\alpha_k - 1)}{\pi_k}.$$

$$\frac{\partial f}{\partial} \quad g(\pi) = (\sum_{k=1}^{K} \pi_k) - 1 = 0$$

by lagrang thm, $\exists \lambda \in \mathbb{R}$ $\frac{\partial f}{\partial \pi_k} = \lambda \frac{\partial g}{\partial \pi_k}$

$$\frac{\sum r_k^i}{\pi_k} + \frac{\alpha_k - 1}{\pi_k} = \lambda$$

$$\therefore \pi_k = \frac{\alpha_k - 1 + \sum r_k^i}{\sum(\alpha_{k'} - 1 + \sum r_{k'}^i)}$$

$$\frac{\partial f}{\partial \theta_{k,j}} = \frac{\sum\sum r_k^{(i)} (\sum x_j \log(\theta_{k,j}) + \sum(1-x_j^i) \log(1-\theta_{k,j}) + \log p(\theta)}{\partial \theta}.$$

$$0 = \frac{\sum r_k^i x_j}{\theta_{k,j}} - \frac{\sum r_k^i (1-x_j^i)}{1-\theta_{k,j}} + \frac{a-1}{\theta_{k,j}} - \frac{b-1}{1-\theta_{k,j}}$$

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

2.

[9.1 11.8   9.1 11.8   9.1   9.1 10.   9.1   9.1 11.8]

pi[0] 0.08499999999999992

pi[1] 0.1299999999999999

theta[0, 239] 0.6427106227106232

theta[3, 298] 0.46573612495845823

part 2

1.

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

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

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

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

3.

[5925. 6744. 5960. 6133. 5844. 5423. 5920. 6267. 5853. 5951.]

R[0, 2] 1.0415281095257362e-14

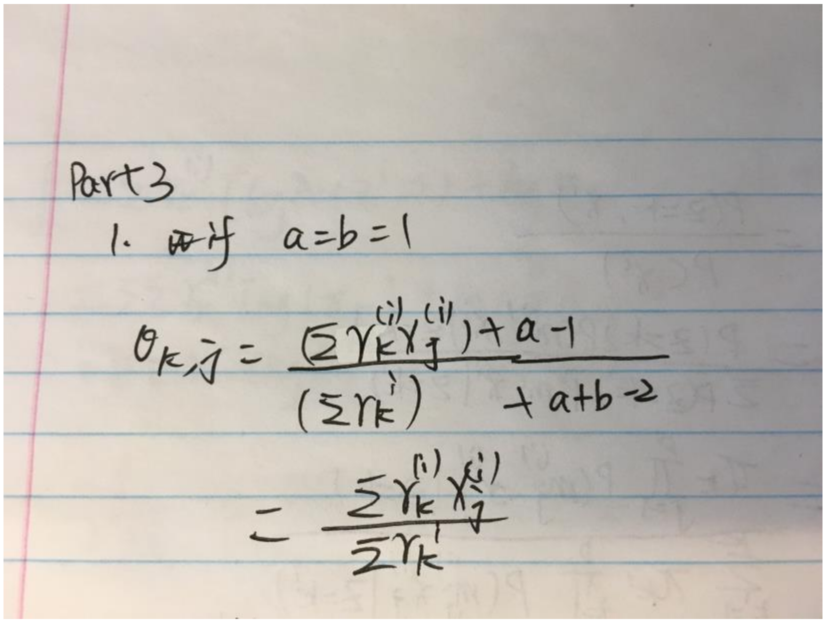
R[1, 0] 1.0

P[0, 183] 0.7432744485494311

P[2, 628] 0.2426125001677637

Part3

1.



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.