Part 1

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

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

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

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

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

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

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

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

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

$$\frac{\partial f}{\partial \theta_{kj}} \quad \frac{\sum\sum r_k^{(i)} (\sum x_j \log(\theta_{kj}) + \sum(1-x_j^i)\log(1-\theta_{kj}) + \log p(\theta)}{\partial \theta}.$$

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

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

Part 2

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

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

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

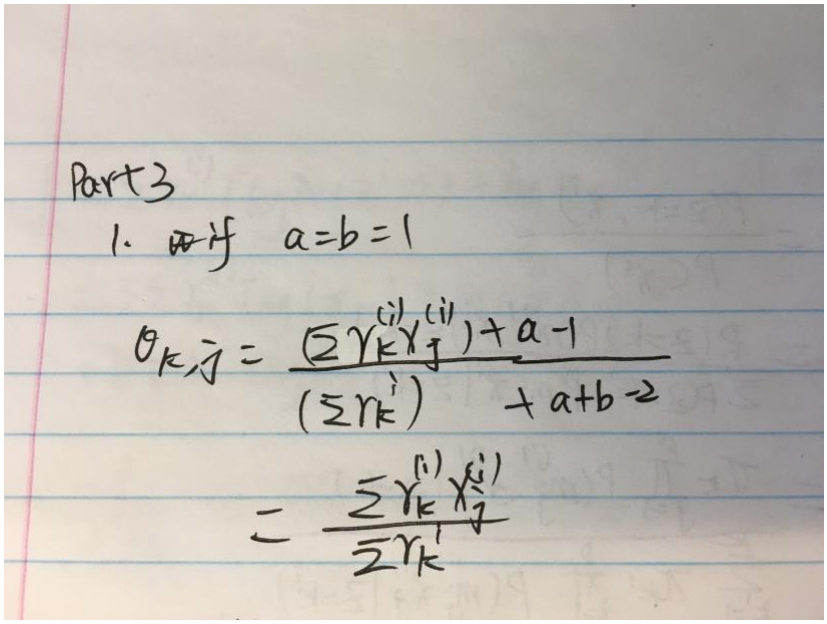
$$= \frac{\pi_k \prod_{j=1}^{D} \theta_{k,j}^{m_j x_j} (1-\theta_{k,j})^{m_j (1-x_j^{(i)})}}{\sum_{k'=1}^{K} \pi_{k'} \prod_{j=1}^{D} \theta_{k',j}^{m_j x_j} (1-\theta_{k',j})^{m_j^{(i)}(1-x_j^{(i)})}}$$

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. This means pixel will be assigned a probability 0 of being a 1. Therefore, the uniform prior get poor probability estimates in this case.

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. The different between the average log probabilities can be explained better by the similarity between different digits. 1 is very much look like 7 but there are not very much digit look like 8. So it does not mean that it will generate far more 1's than 8's. This means 1 is assigned higher log probability due to many digits that look like 1. Whereas, there is no much digit look like 8 and it is assigned lower log probability.