HW4 Tianbo Qiu tq2137

Problem 1

a) EM equation : $\ln p(x|\pi,\theta) = \sum_c q(c)\ln\frac{p(x,c|\pi,\theta)}{q(c)} + \sum_c q(c)\ln\frac{q(c)}{p(c|x,\pi,\theta)}$

E-Step (step 1) : $p(c|x,\pi,\theta) \propto p(x|c,\theta)p(c|\pi) \propto \prod_{i=1}^n p(x_i|c_i,\theta)p(c_i|\pi) = \frac{1}{Z}\prod_{i=1}^n p(x_i|c_i,\theta)p(c_i|\pi)$

$= \prod_{i=1}^n \frac{p(x_i|c_i,\theta)p(c_i|\pi)}{Z_i} = \prod_{i=1}^n p(c_i|x_i,\pi,\theta) = \prod_{i=1}^n q(c_i)$

, where $p(c_i=k|x,\pi,\theta) = \frac{p(x_i|\theta_k)\cdot\pi_k}{\sum_{\ell=1}^k p(x_i|\theta_\ell)\cdot\pi_\ell}$. set $q(c) = p(c|x,\pi,\theta) = \prod_{i=1}^n q(c_i)$

and we use notation $q(c_i=j) = \phi_i(j)$

E-Step (step 2) : $\mathcal{L}(\pi,\theta) = \sum_c q(c)\ln p(x,c|\pi,\theta) - \sum_c q(c)\ln q(c)$

$= \mathbb{E}_{q(c)}[\ln p(x,c|\pi,\theta)] - \mathbb{E}_{q(c)}[\ln q(c)] = \sum_{i=1}^n \mathbb{E}_{q(c_i)}[\ln p(x_i,c_i|\pi,\theta)] - \sum_{i=1}^n \mathbb{E}_{q(c_i)}[\ln q(c_i)]$

$= \sum_{i=1}^n \sum_{j=1}^k \phi_i(j)[\ln\frac{20!}{x_i!(20-x_i)!} + x_i\ln\theta_j + (20-x_i)\ln(1-\theta_j) + \ln\pi_j]$

$\underbrace{- \sum_{i=1}^n\sum_{j=1}^k \phi_i(j)\ln\phi_i(j)}_{const}$

M-Step (step 3) : maximize $\mathcal{L}(\pi,\theta)$ over π and θ.

a) $\nabla_{\theta_j}\mathcal{L} = 0$. for $j=1,2,\cdots K$. $\nabla_{\theta_j}\mathcal{L} = \sum_{i=1}^n \phi_i(j)[x_i\frac{1}{\theta_j} + (20-x_i)\frac{1}{\theta_j-1}] = 0$

$\Rightarrow \theta_j = \frac{1}{20}\cdot\frac{1}{n_j}\sum_{i=1}^n \phi_i(j)x_i$, where $n_j = \sum_{i=1}^n \phi_i(j)$

b) $\nabla_\pi\mathcal{L} = 0$, subject to $\pi_j \geq 0$ and $\sum_{j=1}^K \pi_j = 1$. $\Rightarrow \pi_j = \frac{n_j}{n}$

EM algorithm :

Input : Data $x_1, x_2, \cdots, x_n$, $x \in \{0,1,\cdots,20\}$, Number of clusters K

Output : Binomial Mixture Model parameters π, θ and cluster assignment distribution ϕ_i

1. Initialize $\pi^{(0)} = [\pi_1^{(0)}, \pi_2^{(0)}, \cdots, \pi_K^{(0)}]$, and $\theta^{(0)} = [\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_K^{(0)}]$

2. At iteration t : a) E-step : For $i=1,\cdots,n$ and $j=1,\cdots,K$, set $\phi_i^{(t)}(j) = \frac{\pi_j^{(t-1)}\cdot\text{Binomial}(x_i|20,\theta_j^{(t-1)})}{\sum_{k=1}^K \pi_k^{(t-1)}\cdot\text{Binomial}(x_i|20,\theta_k^{(t-1)})}$

3. M-step : Set $n_j^{(t)} = \sum_{i=1}^n \phi_i^{(t)}(j)$

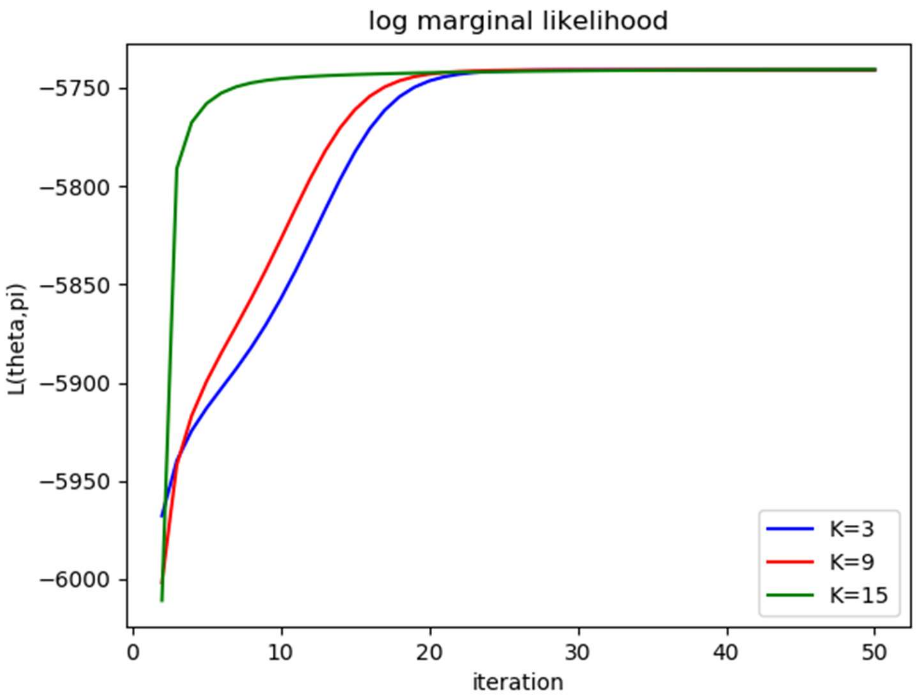
$\theta_j^{(t)} = \frac{1}{20}\cdot\frac{1}{n_j^{(t)}}\sum_{i=1}^n \phi_i^{(t)}(j)x_i$, for $j=1,2,\cdots,K$

$\pi_j^{(t)} = \frac{n_j^{(t)}}{n}$, for $j=1,2,\cdots,K$

4. Calculate $\mathcal{L}(\pi,\theta)$ to access convergence .

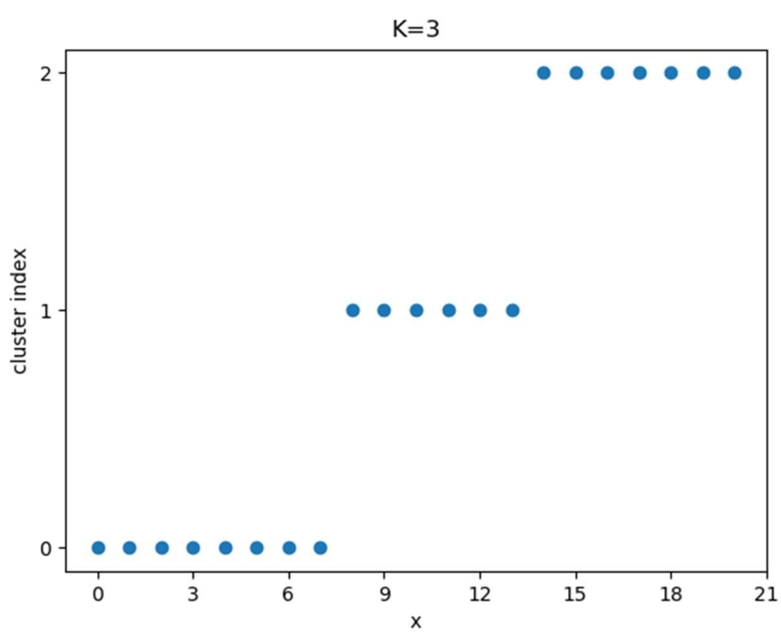
Problem 1

b)

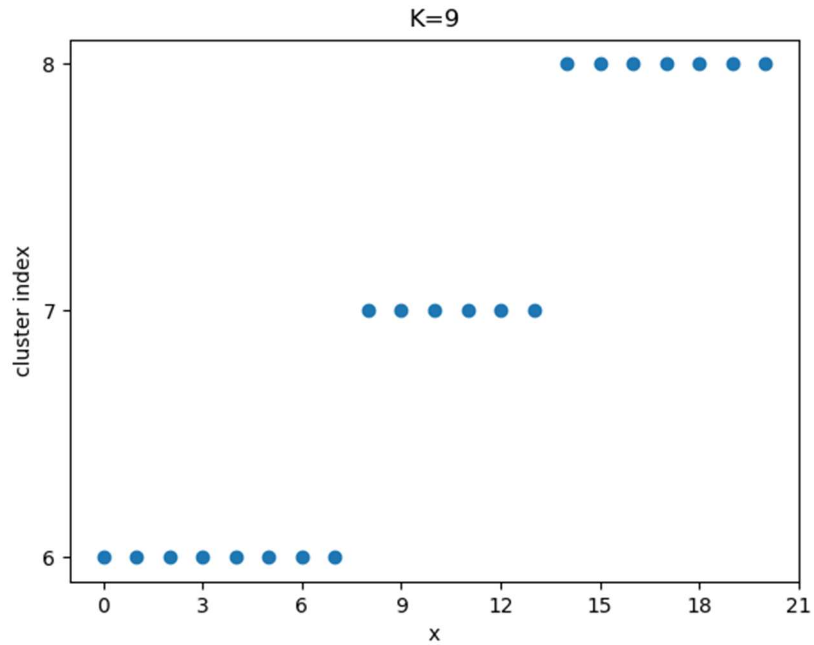


c)

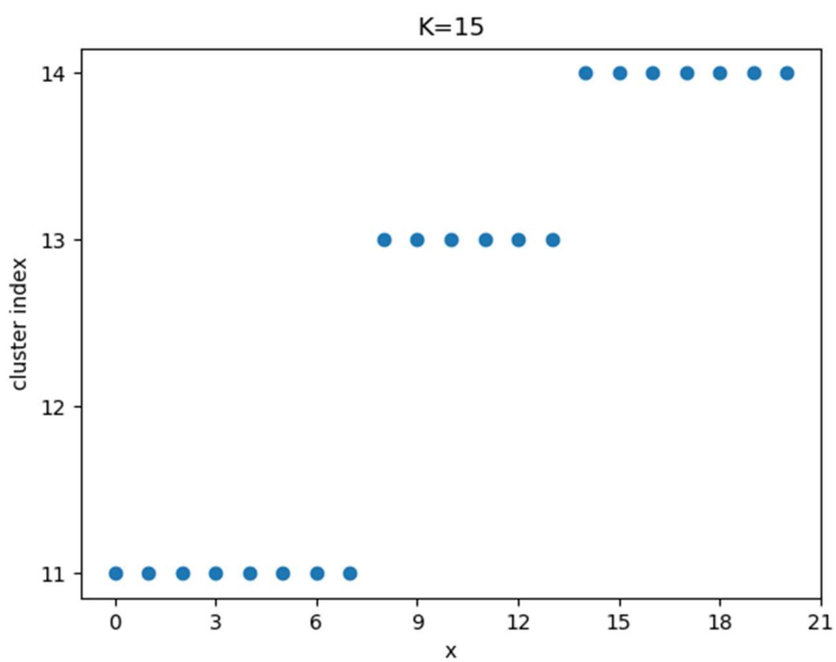
K=3:



K = 9:


K=9
cluster index
8
7
6
0
3
6
9
12
15
18
21
x

K = 15:


K=15
cluster index
14
13
12
11
0
3
6
9
12
15
18
21
x

Problem 2. a) problem set up : $X = \{x_1, \cdots, x_n\}$, $x_i \in \{0,1,2,\cdots,20\}$

$x_i | c_i \sim \text{Binomial}(20, \theta_{c_i})$ $\quad \underline{p(x_i | c_i, \theta)} = \binom{20}{x_i} \theta_{c_i}^{x_i} (1-\theta_{c_i})^{20-x_i}$

$\underline{\ln p(x_i | c_i, \theta)} = \ln\binom{20}{x_i} + x_i \ln\theta_{c_i} + (20-x_i)\ln(1-\theta_{c_i})$

$c_i \overset{iid}{\sim} \text{Discrete}(\pi)$ $\quad \underline{p(c_i = j | \pi)} = \pi_j$

$\underline{\text{priors}}$: $\pi \sim \text{Dir}(\alpha)$, $p(\pi | \alpha) = \cancel{\frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)}} \cdot \cancel{\prod \Gamma(\Sigma}$

$p(\pi | \alpha_1, \cdots \alpha_k)$

$\underline{p(\pi | \alpha_1, \cdots, \alpha_k)} = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \cdot \prod_{j=1}^k \pi_j^{\alpha_j - 1}$

$\underline{p(\theta_j)} \overset{iid}{\sim} \text{Beta}(a, b)$, $\underline{p(\theta_j | a, b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta_j^{a-1} \cdot (1-\theta_j)^{b-1}$

joint $\quad p(x, c, \pi, \theta) = p(x | c, \theta)\, p(c | \pi)\, p(\pi)\, p(\theta)$

$= \prod_{i=1}^n [p(x_i | c_i, \theta)\, p(c_i | \pi)] \cdot [\prod_{j=1}^k p(\theta_j)] \cdot p(\pi)$

$\ln p(x, c, \pi, \theta) = \ln p(x | c, \theta) + \ln p(c | \pi) + \ln p(\pi) + \ln p(\theta)$

$= \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}(c_i = j)[\ln p(x_i | \theta_j) + \ln \pi_j] + \sum_{j=1}^k \ln p(\theta_j) + \ln p(\pi)$

$p(\pi, \theta, c | x) \approx q(\pi, \theta, c) = q(\pi)[\prod_{j=1}^k q(\theta_j)][\prod_{i=1}^n q(c_i)]$

The model is a conjugate exponential family model and so each q distribution should be set to the same family as the prior.

$q(\pi)$: $\text{Dir}(\alpha_1', \cdots \alpha_k')$

$q(\theta_j) = \text{Beta}(a_j', b_j')$

$q(c_i) = \text{Multinomial}(\phi_i)$

$q(c_i = j)$ $\quad q(c_i = j) \propto \exp\{ \mathbb{E}_{q(c_{-i}, \theta, \pi)} [\ln p(x | c, \theta) + \ln p(c | \pi) + \cancel{\ln p(\pi)} + \cancel{\ln p(\theta)}] \}$

$\propto \exp\{ \mathbb{E}[\ln p(x_i | \theta_j)] + \mathbb{E}[\ln p(c_i = j | \pi)] \}$

$\propto \exp\{ \mathbb{E}_{q(\theta_j)}[\ln\binom{20}{x_i} + x_i \ln\theta_j + (20-x_i)\ln(1-\theta_j)] + \mathbb{E}_{q(\pi_j)}[\ln \pi_j] \}$

$= \alpha \exp\{ x_i \mathbb{E}_{q(\theta_j)}[\ln\theta_j] + (20-x_i) \mathbb{E}_{q(\theta_j)}[\ln(1-\theta_j)] + \mathbb{E}_{q(\pi_j)}[\ln \pi_j] \}$

$\overline{\sum_{j=1}^k \exp\{ x_i \mathbb{E}_{q(\theta_j)}[\ln\theta_j] + (20-x_i) \mathbb{E}_{q(\theta_j)}[\ln(1-\theta_j)] + \mathbb{E}_{q(\pi_j)}[\ln \pi_j] \}}$

, where $\mathbb{E}_{q(\theta_j)}[\ln\theta_j] = \psi(a_j') - \psi(a_j' + b_j')$ $\quad \mathbb{E}_{q(\pi_j)}[\ln \pi_j] = \psi(\alpha_j') - \psi(\sum_k \alpha_k')$

$\mathbb{E}_{q(\theta_j)}[\ln(1-\theta_j)] = \psi(b_j') - \psi(a_j' + b_j')$

$q(\pi)$:

$$q(\pi) \propto \exp\{\mathbb{E}_{q(c,\theta)}[\ln p(x|c,\theta)+\ln p(c|\pi)+\ln p(\pi)+\ln p(\theta)]\}$$

$$\propto \exp\{\sum_{i=1}^n \mathbb{E}_{q(c_i)} \ln p(c_i|\pi) + \ln p(\pi)\} \propto \exp\{\sum_{i=1}^n \sum_{j=1}^k \phi_i(j)\ln \pi_j\}\cdot p(\pi)$$

$$\propto \prod_{j=1}^k e^{\sum_{i=1}^n \phi_i(j)\ln\pi_j} \cdot \prod_{j=1}^k \pi_j^{\alpha_j-1} \propto \prod_{j=1}^k \cdot \pi_j^{\alpha_j+\sum_{i=1}^n \phi_i(j)-1} = Dir(\alpha')$$

, where $\alpha_j' = \alpha_j + \sum_{i=1}^n \phi_i(j) = \alpha_j + n_j$, $n_j = \sum_{i=1}^n \phi_i(j)$

for $j = 1, 2, \cdots, K$

$q(\theta_j)$:

$$q(\theta_j) \propto \exp\{\mathbb{E}_{q(\theta_{-j},c,\pi)}[\ln p(x|c,\theta) + \cancel{\ln p(c|\pi)} + \cancel{\ln p(\pi)} + \ln p(\theta)]\}$$

$$\propto \exp\{\mathbb{E}_{q(c)}[\sum_{i=1}^n \sum_{j=1}^k \mathbb{1}(c_i=j)[\ln p(x_i|\theta_j)] + \ln p(\theta_j)]\}$$

$$\propto \exp\{\sum_{i=1}^n \phi_i(j)[\ln\binom{20}{x_i} + x_i\ln\theta_j + (20-x_i)\ln(1-\theta_j)\}\cdot p(\theta_j)$$

$$\propto \prod_{i=1}^n \exp\{\phi_i(j)[x_i\ln\theta_j + (20-x_i)\ln(1-\theta_j)]\cdot\theta_j^{a-1}(1-\theta_j)^{b-1}$$

$$\propto \theta_j^{\sum_{i=1}^n \phi_i(j)x_i}(1-\theta_j)^{\sum_{i=1}^n\phi_i(j)(20-x_i)}\cdot\theta_j^{a-1}\cdot(1-\theta_j)^{b-1}$$

$= Beta(a_j', b_j')$, where $a_j' = a + \sum_{i=1}^n \phi_i(j)x_i$

$b_j' = b + \sum_{i=1}^n \phi_i(j)(20-x_i)$

VI objective function. $\mathcal{L} = \mathbb{E}_{q(\pi,\theta,c)}[\ln p(x,c,\pi,\theta)] - \mathbb{E}_{q(\pi,\theta,c)}[\ln q(\pi,c,\theta)]$

$$\mathcal{L} = \mathbb{E}_{q(\pi,\theta,c)}[\ln p(x|c,\theta) + \ln p(c|\pi) + \ln p(\pi) + \ln p(\theta)] - \mathbb{E}_{q(\pi,\theta,c)}[\ln q(\pi) + \ln q(\theta) + \ln q(c)]$$

$$= \mathbb{E}_{q(\theta,c,\pi)}[\sum_{i=1}^n\sum_{j=1}^k \mathbb{1}(c_i=j)[\ln p(x_i|\theta_j)+\ln\pi_j] + \ln p(\pi) + \sum_{j=1}^k \ln p(\theta_j)] - \mathbb{E}_{q(\pi)}\ln q(\pi) - \sum_{j=1}^k\mathbb{E}_{q(\theta_j)}\ln q(\theta_j) - \sum_{i=1}^n \mathbb{E}_{q(c_i)}\ln q(c_i)$$

$$= \underbrace{\mathbb{E}_{q(\theta,c,\pi)}[\sum_{i=1}^n\sum_{j=1}^k \mathbb{1}(c_i=j)[\ln p(x_i|\theta_j)+\ln\pi_j]]}_{①} + \underbrace{\mathbb{E}_{q(\pi)}\ln p(\pi)}_{②} + \underbrace{\sum_{j=1}^k \mathbb{E}_{q(\theta_j)}[\ln p(\theta_j)]}_{③}$$

$$\underbrace{-\mathbb{E}_{q(\pi)}\ln q(\pi)}_{④} \underbrace{-\sum_{j=1}^k \mathbb{E}_{q(\theta_j)}\ln q(\theta_j)}_{⑤} \underbrace{-\sum_{i=1}^n \mathbb{E}_{q(c_i)}\ln q(c_i)}_{⑥}$$

① $= \sum_{i=1}^{n} \sum_{j=1}^{k} \phi_i(j) \left[\mathbb{E}_{q(\theta_j)}[\ln p(x_i|\theta_j)] + \mathbb{E}_{q(\pi)}[\ln \pi_j] \right]$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \phi_i(j) \left[\ln\left(\frac{20!}{x_i!(20-x_i)!}\right) + x_i \mathbb{E}_{q(\theta_j)}[\ln \theta_j] + (20-x_i) \cdot \mathbb{E}_{q(\theta_j)}[\ln(1-\theta_j)] + \mathbb{E}_{q(\pi)}[\ln \pi_j] \right]$$

② $= \mathbb{E}_{q(\pi)}[\ln p(\pi)] = \mathbb{E}_{q(\pi)}\left[\ln\left(\frac{\Gamma(\sum_{j=1}^{k} \alpha_j)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \right) + \sum_{j=1}^{k} (\alpha_j - 1) \ln \pi_j \right]$

$$= \ln \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - \sum_{j=1}^{k} \ln \Gamma(\alpha_j) + \sum_{j=1}^{k} (\alpha_j - 1) \mathbb{E}_{q(\pi)}[\ln \pi_j]$$

③ $= \sum_{j=1}^{k} \mathbb{E}_{q(\theta_j)}[\ln p(\theta_j)] = \sum_{j=1}^{k} \mathbb{E}_{q(\theta_j)} \left[\ln \Gamma(a+b) - \ln \Gamma(a) - \ln \Gamma(b) + (a-1)\ln \theta_j + (b-1)\ln(1-\theta_j) \right]$

$$= \sum_{j=1}^{k} \left[\ln \Gamma(a+b) - \ln \Gamma(a) - \ln \Gamma(b) + (a-1) \mathbb{E}_{q(\theta_j)}[\ln \theta_j] + (b-1) \mathbb{E}_{q(\theta_j)}[\ln(1-\theta_j)] \right]$$

④ : $-\mathbb{E}_{q(\pi)} \ln q(\pi) = \ln \frac{\prod_{j=1}^{k} \Gamma(\alpha_j')}{\Gamma(\sum_{j=1}^{k} \alpha_j')} + \left(\sum_{j=1}^{k} \alpha_j' - k\right) \psi\left(\sum_{j=1}^{k} \alpha_j'\right) - \sum_{j=1}^{k} (\alpha_j' - 1) \psi(\alpha_j')$

⑤ : $-\sum_{j=1}^{k} \mathbb{E}_{q(\theta_j)} \ln q(\theta_j) = \cancel{\ln \frac{\Gamma(a_j')\Gamma(b_j')}{\Gamma(a_j'+b_j')}} \sum_{j=1}^{k} \left[\ln \frac{\Gamma(a_j')\Gamma(b_j')}{\Gamma(a_j'+b_j')} - (a_j'-1)\psi(a_j') - (b_j'-1)\psi(b_j') + (a_j'+b_j'-2)\psi(a_j'+b_j') \right]$

⑥ : $-\sum_{i=1}^{n} \mathbb{E}_{q(c_i)} \ln q(c_i) = -\sum_{i=1}^{n} \sum_{j=1}^{k} q(c_i=j) \ln q(c_i=j) = -\sum_{i=1}^{n} \sum_{j=1}^{k} \phi_i(j) \ln \phi_i(j)$

$\mathcal{L}$ = ① + ② + ③ + ④ + ⑤ + ⑥

VI algorithm for the Binomial Mixture Model

Input: Data. $x_1, \cdots, x_n$, $x_i \in \{0,1,2,\cdots,20\}$, Number of cluster K

Output: Parameters for $q(\pi)$, $q(\theta_j)$ and $q(c_i)$

1. Initialize $(\alpha_1'^{(0)}, \alpha_2'^{(0)}, \cdots, \alpha_K'^{(0)}, a_1'^{(0)}, a_2'^{(0)}, a_3'^{(0)} \cdots, a_K'^{(0)}, b_1'^{(0)}, b_2'^{(0)}, b_3'^{(0)} \cdots, b_K'^{(0)})$

2. At iteration t.

(a) Update $q(c_i)$, for $i=1,2,\cdots,n$.

$$\phi_i^{(t)}(j) = \frac{\exp\{x_i \mathbb{E}_{q(\theta_j)_{(t-1)}}[\ln\theta_j] + (20-x_i)\mathbb{E}_{q(\theta_j)_{(t-1)}}[\ln(1-\theta_j)] + \mathbb{E}_{q(\pi)_{(t-1)}}[\ln\pi_j]\}}{\sum_{k=1}^{K} \exp\{x_i \mathbb{E}_{q(\theta_k)_{(t-1)}}[\ln\theta_k] + (20-x_i)\mathbb{E}_{q(\theta_k)_{(t-1)}}[\ln(1-\theta_k)] + \mathbb{E}_{q(\pi)_{(t-1)}}[\ln\pi_k]\}}$$

~~(b) Set~~, where $\mathbb{E}_{q(\theta_j)_{(t-1)}}[\ln\theta_j] = \psi(a_j') - \psi(a_j'+b_j') \Big|_{a'_{j(t-1)},\, b'_{j(t-1)}}$

$$\mathbb{E}_{q(\theta_j)_{(t-1)}}[\ln(1-\theta_j)] = \psi(b_j') - \psi(a_j'+b_j') \Big|_{a'_{j(t-1)},\, b'_{j(t-1)}}$$

$$\mathbb{E}_{q(\pi)_{(t-1)}}[\ln\pi_j] = \psi(\alpha_j') - \psi(\textstyle\sum_k \alpha_k') \Big|_{\alpha'_{j(t-1)}}.$$

(b) Set $n_j^{(t)} = \sum_{i=1}^n \phi_i^{(t)}(j)$, for $j=1,\cdots,K$

(c) Update $q(\pi)$ by setting: $\alpha_j'^{(t)} = \alpha_j + n_j^{(t)}$, for $j=1,\cdots,K$

(d) Update $q(\theta_j)$ by setting: $a_j'^{(t)} = a + \sum_{i=1}^n \phi_i^{(t)}(j) x_i$

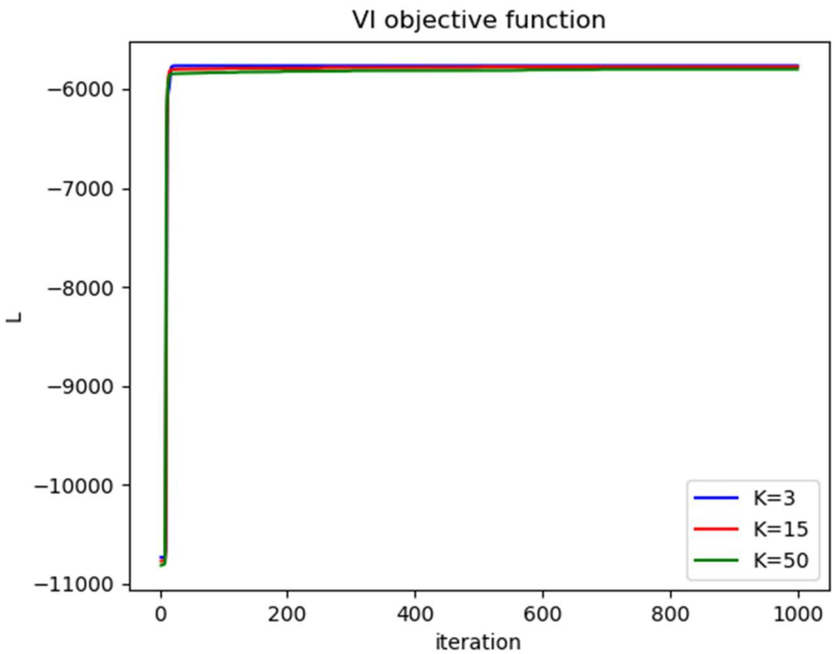
$$b_j'^{(t)} = b + \sum_{i=1}^n \phi_i^{(t)}(j)(20-x_i)$$

(e) Calculate the variational objective function $\mathcal{L}$ to assess convergence.

, where $\mathcal{L} = ① + ② + ③ + ④ + ⑤ + ⑥$

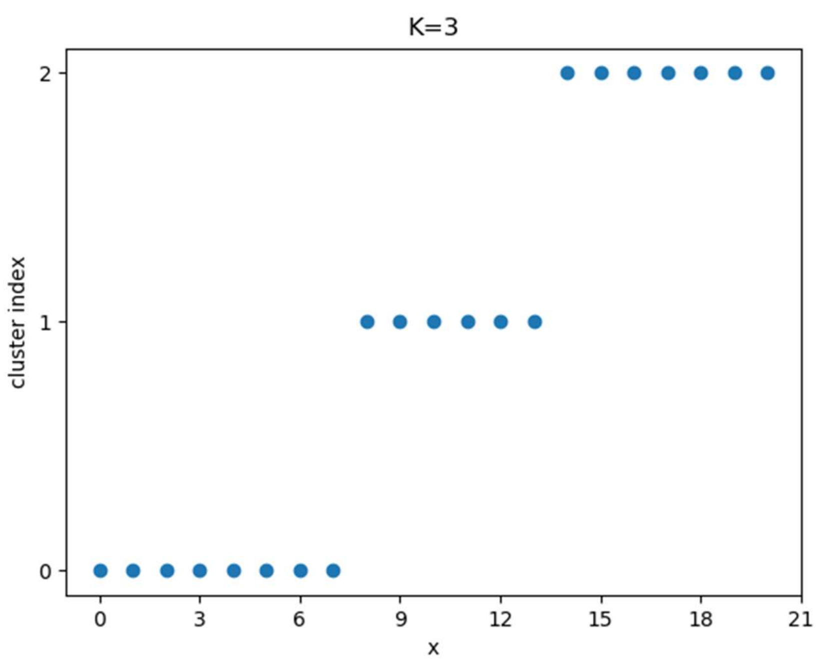
Problem 2

b)

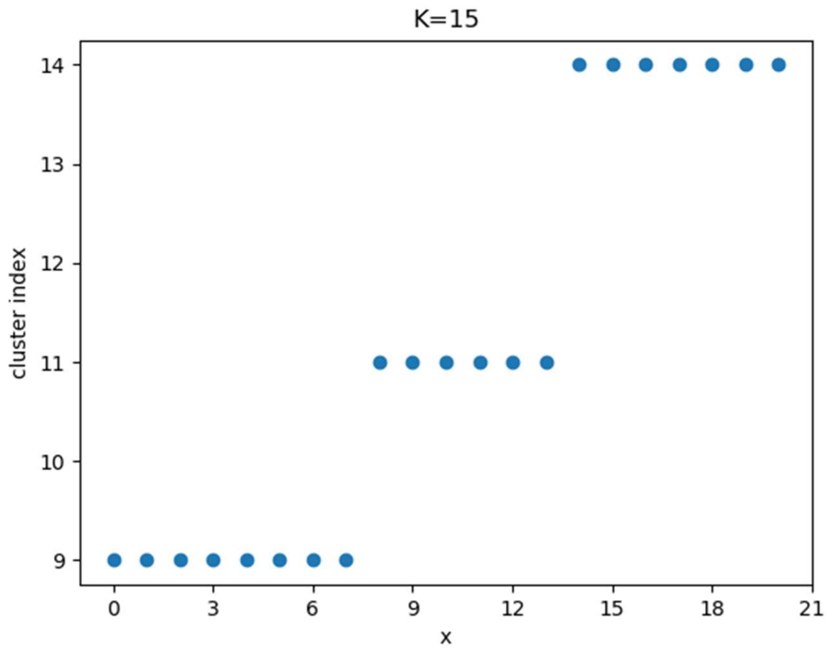


c)

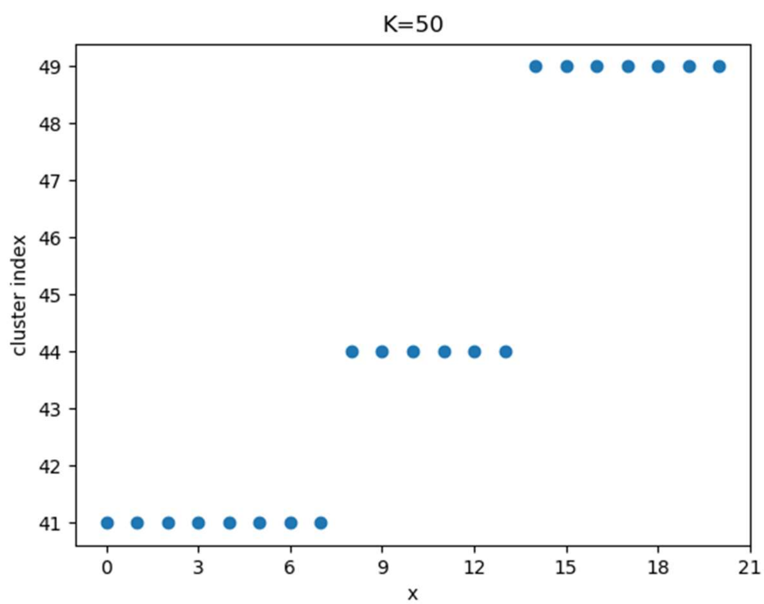
K = 3



K = 15


K=15
cluster index
x
14
13
12
11
10
9
0
3
6
9
12
15
18
21

K = 50


K=50
cluster index
x
49
48
47
46
45
44
43
42
41
0
3
6
9
12
15
18
21

Problem 3

a) Model : $X = \{x_1, \cdots, x_n\}$, $x_i \in \{0,1,2,\cdots,20\}$. $x_i | C_i \sim \text{Binomial}(20, \theta_{C_i})$. $p(x_i|C_i,\theta) = \binom{20}{x_i}\theta_{C_i}^{x_i}(1-\theta_{C_i})^{20-x_i}$

$C_i \overset{iid}{\sim} \text{Discrete}(\pi)$. $p(C_i = j|\pi) = \pi_j$

Priors : $\pi \sim \text{Dir}(\underbrace{\frac{\alpha}{K}, \frac{\alpha}{K}, \cdots, \frac{\alpha}{K}}_{K \text{ times}})$ for $K \to \infty$, $\theta_k \overset{iid}{\sim} \text{Beta}(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$

Sample θ_j : The conditional posterior $p(\theta_j|\theta_{-j}, c, X)$ only depends on the data X that has been assigned to cluster j . We can write this as : $p(\theta_j|X,C) = p(\theta_j|\{x_i : C_i = j\}) \propto p(X|\theta_j, C)p(\theta_j)$

However, now since there are an infinite number of θ_j , we can break this process down into two categories

$\propto \left[\prod_{i=1}^{n} p(x_i|\theta_j)^{\mathbb{1}(C_i=j)}\right] p(\theta_j)$ ~~$\propto \theta_j^{\sum_{i=1}^n \mathbb{1}(C_i=j)} (1-\theta_j)^{20 - \sum_{i=1}^n \mathbb{1}(C_i=j)}$~~ $\propto \theta_j^{\sum_{C_i=j} x_i}(1-\theta_j)^{\sum_{C_i=j}(20-x_i)}\theta_j^{a-1}(1-\theta_j)^{b-1}$

1. $n_j > 0$ There can be only finite number of j such that this is the case. Therefore, we need to sample from the ~~f~~ conditional posterior.

$\text{Beta}(a', b')$ where . $a' = a + \sum_{C_i=j} x_i$

$b' = b + \sum_{C_i=j}(20-x_i)$

2. $n_j = 0$ We sample ~~in~~ from the prior $p(\theta)$

Sample ~~θ_i~~ C_i : By Bayes rule, $p(C_i=j|X,\theta,C_{-i}) \propto p(X|C_i=j,\theta)p(C_i=j|C_{-i})$

$p(C_i=j|C_{-i}) = \int p(C_i=j|\pi)p(\pi|C_{-i})d\pi = \int \pi_j \cdot \text{Dir}(\pi|\frac{\alpha}{K}+n_1^{(-i)}, \cdots \frac{\alpha}{K}+n_K^{(-i)})d\pi$, where $n_j^{(-i)} = \sum_{s \neq i}\mathbb{1}(C_s=j)$

$= \frac{\Gamma(\alpha+n-1)}{\Gamma(\alpha+n)} \cdot \frac{\Gamma(\frac{\alpha}{K}+n_j^{(-i)}+1)}{\Gamma(\frac{\alpha}{K}+n_j^{(-i)})} = \frac{\alpha/K + n_j^{(-i)}}{\alpha+n-1}$

As a result , $p(C_i=j|X,\theta,C_{-i}) \propto p(x_i|\theta_j)\left(\frac{\alpha/K + n_j^{(-i)}}{\alpha+n-1}\right)$ ~~for~~

~~$C_i = \begin{cases} j & \text{w.p} \propto p(x_i|\theta_j)\, n_j^{(-i)}/(\alpha+n-1) \quad \text{if } n_j^{(-i)} > 0 \\ \text{a new index} \end{cases}$~~

Case : $n_j^{(-i)} > 0$: $K \to \infty$. $p(C_i=j|X,\theta,C_{-i}) \propto p(x_i|\theta_j)\frac{n_j^{(-i)}}{\alpha+n-1}$

Case : $n_j^{(-i)} = 0$: $K \to \infty$ $p(C_i = new|X,\theta,C_{-i}) = \sum_{j:n_j^{(-i)}=0} p(C_i=j|X,\theta,C_{-i}) \propto \lim_{K\to\infty}\frac{\alpha}{\alpha+n-1}\sum_{j:n_j^{(-i)}=0}\frac{p(x_i|\theta_j)}{K}$

$\propto \frac{\alpha}{\alpha+n-1}\int p(x_i|\theta)p(\theta)d\theta$

To summarize, we have shown that

$$C_i = \begin{cases} j . & \text{w.p} \propto p(x_i|\theta_j)\frac{n_j^{(-i)}}{\alpha+n-1} \quad \text{if } n_j^{(-i)} > 0 \\ \text{a new index} & \text{w.p.} \propto \frac{\alpha}{\alpha+n-1}\int p(x_i|\theta)p(\theta)d\theta . \end{cases}$$

, where $p(x_i|\theta_j) = \binom{20}{x_i}\theta_j^{x_i}(1-\theta_j)^{20-x_i}$, $p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$

$p(x_i|\theta)p(\theta) \propto \theta^{a+x_i-1}(1-\theta)^{b+(20-x_i)-1} \propto \text{Beta}(a', b')$, $a' = a + x_i$, $b' = b + (20-x_i)$

$\int p(x_i|\theta)p(\theta)d\theta = \frac{\Gamma(a')\Gamma(b')}{\Gamma(a'+b')}$, Therefore . $p(C_i = new|X,\theta,C_{-i}) = \frac{\alpha}{\alpha+n-1} \cdot \frac{\Gamma(a')\Gamma(b')}{\Gamma(a'+b')}$

Gibbs sampling algorithm:

Initialize in some way

At iteration t. 1. For $i = 1, 2, \cdots, n$

(a) For all j such that $n_j^{(-i)} > 0$. set $\hat{\phi}_i(j) = p(x_i|\theta_j) n_j^{(-i)} / \alpha+n-1 = \binom{20}{x_i} \theta_j^{x_i} (1-\theta_j)^{20-x_i} \cdot n_j^{(-i)} / \alpha+n-1$

(b) For ~~all~~ a new value j', set $\hat{\phi}_i(j) = \frac{\alpha}{\alpha+n-1} \cdot \frac{\Gamma(a+x_i)\Gamma(b+20-x_i)}{\Gamma(a+b+20)}$

(c) Normalize $\hat{\phi}_i$ and sample the index c_i from a discrete distribution with this parameter.

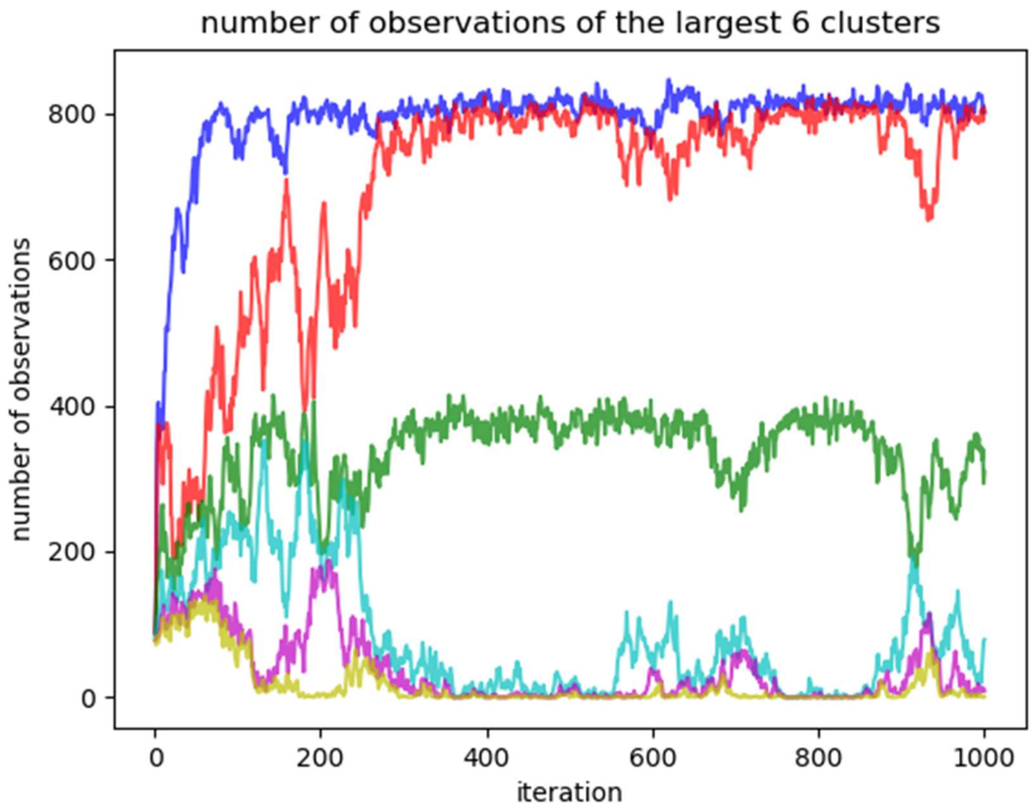
(d) If $c_i = j'$. generate $\theta_j' \sim p(\theta | x_i) = \text{Beta}(\theta | a+x_i, b+20-x_i)$

2. For $j = 1, 2, \cdots, K^{(t)}$. generate $\theta_j \sim p(\theta | \{x_i : c_i = j\}) = \text{Beta}(\theta | a + \sum_{x_i : c_i = j} x_i, b + \sum_{x_i : c_i = j} (20 - x_i))$

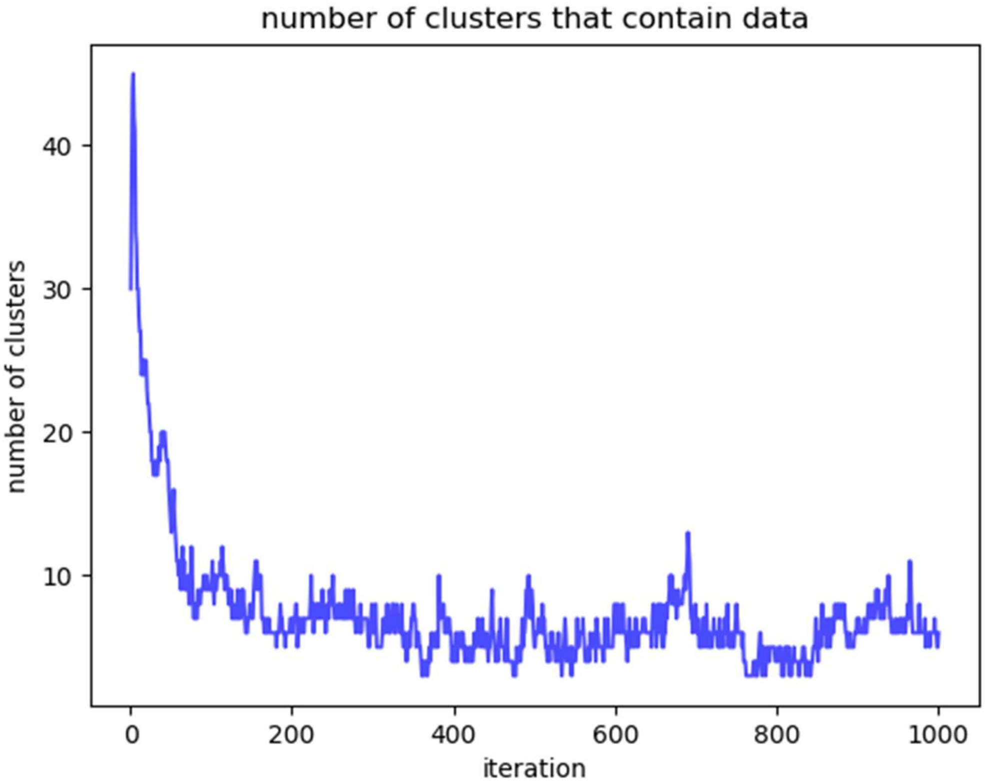
$K^{(t)}$ ~~K_t~~ = # non-zero clusters that are reindexed after completing step 1.

Problem 3

b)


number of observations of the largest 6 clusters
number of observations
0
200
400
600
800
iteration
0
200
400
600
800
1000

c)


number of clusters that contain data
number of clusters
40
30
20
10
0
200
400
600
800
1000
iteration