Question 1

1)

As it is mentioned in Bishop chapter 9 section 3 the relationship between these two methods are below:

- the K-means method can cluster spherical data appropriately while the EM can work well on elliptical data either.
- k-means performs hard assignment of data point to cluster in which each data is associated uniquely with one cluster but EM method makes soft assignment based on posterior in which each cluster have wight (probability) that indicate responsibility of that cluster to create data X
- **Yes on the some limits these two algorithms work similar** and the limits are below: lets first get some intuition of whats going on:

  Intuition

  as we said K-means work fine on spherical data so we will assume that data are spherical (we will see it when we write equations we will choose Σ to be a constant times Identity matrix which means data are distributed spherical around the μ). so as we know when the variance of one Gaussian model goes to zero the model shrink around the mean so it is some how the hard assignment on data like k-means and the log likelihood function increases so if we choose variance small and the data be spherical the GMM (EM) algorithm will work similar to k-means.

  limits

  Assume Gaussian Mixture model in which the co-variance matrices are multiplication of a constant to Identity matrix so they are like $\Sigma = \epsilon I$ where ϵ can be presentation of radios of sphere around mean.

  so distribution is like below:

$$p(x|\mu_k, \Sigma_k) = \frac{1}{2\pi\epsilon}^{0.5} e^{-\frac{1}{2\epsilon}|x-\mu_k|^2} \tag{1}$$

so the posterior distribution of responsibility of each component will be:

$$\gamma(Z_{n,k}) = \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j, \Sigma_j)} \tag{2}$$

by applying eq(1) we get:

$$\gamma(Z_{n,k}) = \frac{\pi_k e^{-\frac{1}{2\epsilon}|x_n-\mu_k|^2}}{\sum_j \pi_j e^{-\frac{1}{2\epsilon}|x_n-\mu_j|^2}} \tag{3}$$

as $\epsilon \longrightarrow 0$ so probability obtained from normal distribution goes to zero except for the component that $x_n \longrightarrow \mu_i$ (notice that in such case we assumed that there is a data that is equal to μ more safe to say that we choose closest data to mean as representation of mean so in such case the responsibility of that component goes to to one so it's like hard assignment in k-means method.
take the x_j closest to μ_j among other μ so we have:

$$\gamma(Z_{n,k}) = \frac{\pi_k e^{-\frac{1}{2\epsilon}|x_n-\mu_k|^2}}{\sum_j \pi_j e^{-\frac{1}{2\epsilon}|x_n-\mu_j|^2}} \tag{4}$$

$$\gamma(Z_{n,j}) = \frac{\pi_j}{\sum_k \pi_k e^{\frac{1}{2\epsilon}|x_n-\mu_j|^2-|x_n-\mu_k|^2}} = \frac{\pi_j}{\sum_k \pi_k e^{\frac{1}{2\epsilon}\Delta_{K,n}}} = \tag{5}$$

for closest to μ_k to x_n we can assume $\Delta_{K,n}$ go to zero so term is one and for others the term is zero so we have:

$$\gamma(Z_{n,j}) = \frac{\pi_j}{\pi_j \sum_{k/j} \pi_k e^{\frac{1}{2\epsilon}\Delta_{K,n}}} \tag{6}$$

$$\epsilon \longrightarrow 0 \implies \gamma(Z_{n,j}) = 1 \tag{7}$$

$$\gamma(Z_{n,k/j}) = 0 \tag{8}$$

it's like hard assignment in k-means method.
so $\gamma(Z_{n,k}) \longrightarrow r_{n,k}$ as we know the likelihood function of EM-estimator is

$$E_Z[lnP(X, Z|\mu, \Sigma, \Pi] = \sum_{n=1}^{n=N}\sum_{k=1}^{k=K} \gamma(Z_{n,k})\{ln\pi_k + lnN(x_n|\mu, \Sigma)\} \tag{9}$$

where $\gamma(Z_{n,k}) = E[Z_{n,k}]$

by replacing and applying logarithm to parameters we have :

$$MAXIMIZE E_Z[lnP(X, Z|\mu, \Sigma, \Pi] = \tag{10}$$

$$MAXIMIZE \sum_{n=1}^{n=N}\sum_{k=1}^{k=K} \gamma(Z_{n,k})\{ln\pi_k + lnN(x_n|\mu, \Sigma)\} \tag{11}$$

$$\epsilon \longrightarrow \infty \implies \tag{12}$$

$$MAXIMIZE \sum_{n=1}^{n=N}\sum_{k=1}^{k=K} r_{n,k}\{ln\pi_k + lnN(x_n|\mu, \Sigma)\} \tag{13}$$

now π_k is obtained under condition mentioned maximization is equal to : (14)

$$minimize \sum_{n=1}^{n=N}\sum_{k=1}^{k=K} r_{n,k}\{|x_n - \mu_k|^2\} \tag{15}$$

eq(15) is exactly the objective function in k-means method ✓

$$P(x_b|x_a)=\frac{P(x_a,x_b)}{P(x_a)}=\frac{P(x)}{P(x_a)}=\frac{\sum_{k=1}^{K}\pi_k P(x|k)}{P(x_a)}=\sum_{k=1}^{K}\frac{\pi_k}{P(x_a)}P(x|k)$$

$$=\sum_{k=1}^{K}\frac{\pi_k}{P(x_a)}P(x_a,x_b|k)=\sum_{k=1}^{K}\frac{\pi_k}{P(x_a)}P(x_b|x_a,k)P(x_a|k)=\sum_{k=1}^{K}\underbrace{\frac{\pi_k P(x_a|k)}{P(x_a)}}_{\pi'_k}P(x_b|x_a,k)$$

$$P(x_a|k)=\sum_{x_b}P(x_a,x_b|k)$$

$$P(x_a)=\sum_{x_b}P(x_a,x_b)=\sum_{x_b}\sum_{k=1}^{K}\pi_k P(x_a,x_b|k)=\sum_{k=1}^{K}\sum_{x_b}\pi_k P(x_a,x_b|k)=$$

$$\sum_{k=1}^{K}\pi_k\sum_{x_b}P(x_a,x_b|k)=\sum_{k=1}^{K}\pi_k P(x_a|k)$$

$$P(x_b|x_a,k)=\frac{P(x_a,x_b|k)}{P(x_a|k)}$$

$$l(\theta)=\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_k^i\log\left(\pi_k|\Sigma|^{-\frac{1}{2}}(2\pi)^{-\frac{k}{2}}\exp\left(-\frac{1}{2}(x_i-\mu'_k)^T\Sigma^{-1}(x_i-\mu'_k)\right)\right)=$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_k^i\left(\log(\pi_k)-\frac{1}{2}\log(|\Sigma|)-\frac{k}{2}\log(2\pi)-\frac{1}{2}(x_i-\mu'_k)^T\Sigma^{-1}(x_i-\mu'_k)\right)$$

$$\Rightarrow\frac{\partial l(\theta)}{\partial\Sigma}=\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_k^i\left(-\frac{1}{2}\Sigma^{-1}+\frac{1}{2}\Sigma^{-1}(x_i-\mu'_k)(x_i-\mu'_k)^T\Sigma^{-1}\right)=0$$

$$\Rightarrow -N\Sigma^{-1}+\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_k^i\Sigma^{-1}(x_i-\mu'_k)(x_i-\mu'_k)^T\Sigma^{-1}=0\Rightarrow NI=\Sigma^{-1}\sum_{i=1}^{N}\sum_{k=1}^{K}(x_i-\mu'_k)(x_i-\mu'_k)^T\gamma_k^i$$

$$\Rightarrow\boxed{\Sigma=\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_k^i(x_i-\mu'_k)(x_i-\mu'_k)^T}$$

μ'_i ، π'_j مانند حالت ماتریس کوواریانس‌های متفاوت حساب می‌شوند.

۲) ۲.۱ و ۲.۲

متغیر Z را طبق مدل Mixture تعریف می‌کنیم: احتمال اینکه داده iام به دسته jام متعلق باشد.

$$P(z_j^{(i)}=1 \mid x_i, \theta) = \gamma_j^i$$

توزیع پواسون به صورت روبرو است.

$$\text{Poisson}(X = x \mid \theta) = \frac{\theta^x e^{-\theta}}{x!}$$

E-step

$$P(z \mid X, \theta)$$

$$\gamma_j^i = P(Z_j^{(i)}=1 \mid x_i, \theta) = \frac{P(x_i \mid Z_j^{(i)}=1, \theta)\, P(Z_j^{(i)}=1)}{\sum_k P(x_i \mid Z_k^{(i)}, \theta)\, P(Z_k^{(i)}=1)} = \frac{\pi_j \, \theta_j^{x_i} e^{-\theta_j}}{\sum \pi_k \, \theta_k^{x_i} e^{-\theta_k}}$$

M-step

$$Q(\theta) = \mathbb{E}_{Z \sim p(Z \mid X, \theta)} [\lg P(X, Z \mid \theta)] = \sum_{i=1}^{n} \sum_{j=1}^{I} \gamma_j^i \lg \left[\pi_j \, \theta_j^{x_i} e^{-\theta_j} \frac{1}{x_i!}\right]$$

$$\frac{\partial Q}{\partial \theta_K} = \sum_{i=1}^{n} \sum_{j=1}^{I} \gamma_j^i \frac{\partial}{\partial \theta_K}\left[\lg \pi_j + x_i \lg \theta_j - \theta_j - \lg x_i!\right] = 0$$

فقط جمله kام در سیگما دوم مؤثر است.

$$\sum_{i=1}^{n} \gamma_K^i \left(\frac{x_i}{\theta_K} - 1\right) = 0 \longrightarrow \theta_K \sum_{i=1}^{n} \gamma_K^i = \sum_{i=1}^{n} \gamma_K^i x_i$$

$$\longrightarrow \boxed{\theta_K = \frac{\sum_{i=1}^{n} \gamma_K^i x_i}{\sum_{i=1}^{n} \gamma_K^i}}$$

یافتن مقدار بهینه برای π_Kها

$$L(\theta, \lambda) = Q(\theta) - \lambda \left(\sum_{j=1}^{I} \pi_j - 1\right)$$

تابع لاگرانژ

$$\frac{\partial L}{\partial \pi_K} = \sum_{i=1}^{n} \sum_{j=1}^{I} \gamma_j^i \frac{\partial}{\partial \pi_K}\left[\lg \pi_j + x_i \lg \theta_j - \theta_j - \lg x_i!\right] - \lambda = 0$$

فقط جمله kام در سیگما دوم.

$$\sum_{i=1}^{n} \gamma_K^i \frac{1}{\pi_K} - \lambda = 0 \longrightarrow \text{برای } K \text{ ثابت:} \quad \lambda = \frac{\sum_{i=1}^{n} \gamma_K^i}{\pi_K} \rightarrow \pi_K = \frac{\sum \gamma_K^i}{\lambda} \quad (*)$$

$$\xrightarrow{\sum_{K=1}^{I} \pi_K = 1} \sum_{K=1}^{I} \pi_K = 1 = \frac{\sum_K \sum_i \gamma_K^i}{\lambda} = \frac{\sum_i \sum_K \gamma_K^i}{\lambda} = \frac{n}{\lambda}$$

$$\longrightarrow \lambda = n \xrightarrow{(*)} \boxed{\pi_K = \frac{\sum_{i=1}^{n} \gamma_K^i}{n}}$$

Question 3)

1)

As it's binomial distribution it's easy to obtain:

$$I)\ \ p(w_t, m_t|k) = \binom{m_t}{w_t} p_k^{w_t}(1-p_k)^{m_t-w_t}$$

$$II)\ \ p(w_t, m_t|ct) = \binom{m_t}{w_t} p_{ct}^{w_t}(1-p_{ct})^{m_t-w_t}$$

2)

As we know from previous question for E-step we have :

$$Q(k) = p(k|x,\theta) \ \propto \ p(x|ct=k,\theta)p(ct=k) \propto \binom{m_t}{x} p_k^x(1-p_k)^{m_t-x}\pi_k$$

$$\sum_K Q(k) = 1 \implies \sum_K \binom{m_t}{x} p_k^x(1-p_k)^{m_t-x}\pi_k = 1$$

As we know the equation above is propto so we need to normalize it that it adds up to one:

$$Q_t[k]^i = \frac{\binom{m_t}{x} p_k^x(1-p_k)^{m_t-x}\pi_k}{\sum_K \binom{m_t}{x} p_j^x(1-p_j)^{m_t-x}\pi_j}$$

3)

For writing M-step we have to maximize equation below:

$$L = \sum_N \sum_K Q_t[k]^i log(p(w_t|p_k)) = \sum_N \sum_K Q_t[k]^i log(\binom{m_t}{w_t} p_k^{w_t}(1-p_k)^{m_t-w_t}) =$$

$$\sum_N \sum_K Q_t[k]^i \{log(\binom{m_t}{w_t}) + w_t log(p_k) + (m_t - w_t) log(1-p_k)\}$$

$$\frac{\partial L}{\partial p_j} = \sum_N Q_t[j]^i \{\frac{w_t}{p_j} - \frac{m_t - w_t}{1-p_j}\} = \sum_N Q_t[j]^i \{\frac{w_t - m_t p_j}{p_j - p_j^2}\} = 0 \implies$$

$$\sum_N Q_t[j]^i \{w_t - m_t p_j\} = 0 \implies p_j = \frac{\sum_{t=1}^N Q_t[j]^i w_t}{\sum_{t=1}^N Q_t[j]^i m_t} \checkmark$$

4.1

Bayes' Theorem: $P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$

$\Rightarrow \ln P(\theta|X) = \ln P(X|\theta) + \ln P(\theta) - \ln P(X)$

$\Rightarrow \ln P(\theta|X) \propto \ln\left(\sum_z P(X,z|\theta)\right) + \ln P(\theta) = \ln\left(\sum_z P(X,z|\theta)P(\theta)\right)$

پس تنها تفاوت با حالت اصلی $P(\theta)$ است که اضافه شده است. پس مشابها در مرحله E باید $P(z|X,\theta^{old})$ را حساب کنیم و در M تابع را برحسب θ بیشینه کنیم.

E-Step: calculate $P(z|X,\theta^{old})$

M-Step: $\theta^{new} = \underset{\theta}{\text{argmax}} \sum_z P(z|X,\theta^{old}) \ln\left[P(X,z|\theta)P(\theta)\right]$

marginalizing

$= \underset{\theta}{\text{argmax}} \underbrace{\sum_z P(z|X,\theta^{old}) \ln P(X,z|\theta)}_{\text{همان تابع حالت قبل}} + \ln P(\theta)$

4.2 ابتا باید X (متغیر observed) و Z (متغیر latent) را تعریف کنیم.

با توجه به داده‌های مثال، به این صورت تعریف می‌شوند:

$X \in \{A \text{ or } C, B, D\}$ $\quad$ $Z \in \{A, B, C, D\}$

با توجه به داده‌های مثال، X تنها ۳ حالت دارد در حالی که برحسب اصلی ۴ حالت دارد.

حالا با توجه به قسمت اول مثال، معادلات EM را می‌نویسیم.

E-Step: $\mathbb{P}\{Z=A \mid X=A \text{ or } C\} = \dfrac{1}{1+2\theta}$

$$\mathbb{P}\{Z=C \mid X=A \text{ or } C\} = \frac{2\theta}{1+2\theta}$$

به غیراز ۲ حالت بالا، $\mathbb{P}\{Z \mid X\}$ برابر با صفر یا یک است و به سادگی تعیین می‌شود.

M-Step: $\theta^{new} = \underset{\theta}{\arg\max} \sum_{i=1}^{n} \sum_{Z} \mathbb{P}\{Z \mid X^{(i)}; \theta^{old}\} \ln \mathbb{P}\{X^{(i)}, Z; \theta\} + \ln \mathbb{P}\{\theta\}$

$$= \underset{\theta}{\arg\max}\ (n_a+n_c)\left(\frac{1}{1+2\theta^{old}} \ln\frac{1}{3} + \frac{2\theta^{old}}{1+2\theta^{old}} \ln \frac{2}{3}\theta\right) + n_b \ln\frac{1}{3}(1-\theta)$$

$$+ n_d \ln\frac{1}{3}(1-\theta) + n(\nu_1-1)\ln\theta + n(\nu_2-1)\ln(1-\theta)$$

$$= \underset{\theta}{\arg\max}\ (n_a+n_c)\frac{2\theta^{old}}{1+2\theta^{old}} \ln\theta + (n_b+n_d)\ln(1-\theta) + n(\nu_1-1)\ln\theta + n(\nu_2-1)\ln(1-\theta)$$

$$\Rightarrow \frac{2\theta^{old}(n_a+n_c)}{(1+2\theta^{old})\theta} + \frac{n(\nu_1-1)}{\theta} = \frac{n_b+n_d+n(\nu_2-1)}{1-\theta}$$

$$\Rightarrow \frac{2\theta^{old}(n_a+n_c) + n(1+2\theta^{old})(\nu_1-1)}{(1+2\theta^{old})\theta} = \frac{n_b+n_d+n(\nu_2-1)}{1-\theta}$$

$$\Rightarrow \theta^{new} = \frac{\frac{2\theta^{old}}{1+2\theta^{old}}(n_a+n_c) + n(\nu_1-1)}{\frac{2\theta^{old}}{1+2\theta^{old}}(n_a+n_c) + n(\nu_1-1) + n_b + n_d + n(\nu_2-1)}$$

دقت کنید $n_a+n_c = n - n_b - n_d$