① (a) $p(X,Z \mid \mu,\Sigma,\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{n,k}} \mathcal{N}(x_n \mid \mu_k,\Sigma_k)^{z_{n,k}}$.

$$\ln p(X,Z \mid \mu,\Sigma,\pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(x_n \mid \mu_k,\Sigma_k)\}$$

Differentiating this quantity & setting it to 0, we get:

$$\frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(x_n \mid \mu_k \Sigma_k)\} = 0$$

In $\sum_k$ summation, every term except 'k' will be 0, as I.I.D.

So,

$$\sum_{n=1}^{N} z_{n,k} \cdot \left\{ 0 + \frac{1}{\cancel{\mathcal{N}(x_n \mid \mu_k,\Sigma_k)}} \cdot \cancel{\mathcal{N}(x_n \mid \mu_k,\Sigma_k)} \cdot (-\Sigma^{-1})(x-\mu) \right\} = 0$$

$$\therefore \sum_{n=1}^{N} z_{nk} (-\Sigma^{-1})(x_n - \mu_k) = 0$$

Multiply by $-\Sigma$ (Const term wrt μ_k, won't change)

$$\therefore \sum_{n=1}^{N} z_{nk} x_n - \mu_k \sum_{n=1}^{N} z_{nk} = 0$$

i.e. $$\mu_k = \frac{\sum_{n=1}^{N} z_{n,k} x_n}{\sum_{n=1}^{N} z_{n,k}}$$

Similarly, differentiating w.r.t. Σ_k, we get:

$$\sum_{n=1}^{N} z_{nk} \left\{ 0 + \left(\frac{-1}{2}\right) \frac{\partial \log|\Sigma|}{\partial \Sigma} + \frac{1}{2} \frac{\partial}{\partial \Sigma} \left((x-\mu)^T \Sigma^{-1} (x-\mu) \right) \right\} = 0$$

$$\therefore \sum_{n=1}^{N} z_{nk} \left\{ -\Sigma_k^{-1} + \Sigma_k^{-1} (x-\mu)(x-\mu)^T \Sigma_k^{-1} \right\} = 0$$

Cancelling Σ_k^{-1} by left-multiply by Σ_k

$$\sum_{n=1}^{N} z_{nk} \left(-1 + (x-\mu)(x-\mu)^T \Sigma_k^{-1} \right) = 0$$

$$\therefore \sum_{n=1}^{N} (x-\mu)(x-\mu)^T z_{nk} \Sigma_k^{-1} = \sum_{n=1}^{N} z_{nk}$$

$$\therefore \Sigma_k = \frac{\sum_{n=1}^{N} (x_n-\mu_k)(x_n-\mu_k)^T z_{nk}}{\sum_{n=1}^{N} z_{nk}}.$$

For mixing coefficients we include a Lagrange multiplier & then differentiate w.r.t. π_k.

$\therefore \ln p(\cancel{X} \mid \mu, \Sigma, \pi) + \lambda \left(\sum \pi_k - 1 \right)$ is to be maximized.

Thus,

$$\frac{\partial}{\partial \pi_k} \sum_{n=1}^{N} \sum_{k=1}^{K} Z_{nk} \{ \ln \pi_k + \ln N(x_n | \mu_k, \Sigma_k) \} + \frac{\partial}{\partial \pi_k} \lambda \left(\sum \pi_k - 1 \right)$$

$$= 0$$

Using $\nabla f = \lambda \nabla g$ condition, we get $\lambda = -N$

Plugging it above,

$$\sum_{n=1}^{N} Z_{nk} \cdot \left(\frac{1}{\pi_k} \right) + (-N)(1) = 0$$

$$\therefore \frac{1}{\pi_k} \sum_{n=1}^{N} Z_{nk} = N \text{, i.e.}$$

$$\pi_k = \frac{\sum_{n=1}^{N} Z_{nk}}{N}$$

Thus,

$$\mu_k = \frac{\sum_{n=1}^{N} Z_{n,k} x_n}{\sum_{n=1}^{N} Z_n}, \qquad \pi_k = \frac{\sum_{n=1}^{N} Z_{n,k}}{N}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} Z_{nk} (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} Z_{nk}}$$

Thus, means and covariances for each 'k' are fitted independently to that group, and mixing coefficients are fractions of points in each group.

(b) To obtain a hard clustering, we assign to each point the cluster for which the associated probability π_j turns out to be the maximum. The mixture density can, thus, be used to obtain a hard clustering.

Thus, $k = \{ \arg\max \; \gamma_k(x)$

↳ posterior probability, corresponds to a point belonging in that cluster.

{ Further, we can set $\Sigma = \epsilon I$ & $\epsilon \to 0$ for all covariances and show that that case corresponds to k-means }

② $\ln p(X,Z|\mu,\Sigma,\pi) = \sum_{n=1}^{N}\sum_{k=1}^{K} \cancel{\pi_k} Z_{nk}\{\ln\pi_k + \ln N(x_n|\mu_k,\Sigma_k)\}$

$E_Z(\ln p(X,Z|\mu,\Sigma,\pi))$

$= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(Z_{nk})\{\ln\pi_k + \ln N(x_n|\mu_k,\Sigma_k)\}$

If all Σ_i have covariance ϵI, i.e. $\forall j\ \Sigma_j = \epsilon I$,

$N(x_n|\mu_k,\Sigma_k) = \frac{1}{\sqrt{2\pi\epsilon}} e^{-\frac{1}{2\epsilon}\|x_n-\mu_k\|^2}$

Thus,

$\lim_{\epsilon\to 0} E_Z[\ln p(X,Z|\mu,\Sigma,\pi)]$

$= \lim_{\epsilon\to 0}\sum_{n=1}^{N}\sum_{k=1}^{K} \underbrace{\frac{\pi_k e^{-\frac{\|x_n-\mu_k\|^2}{2\epsilon}}}{\sum_j \pi_j e^{-\frac{\|x_n-\mu_j\|^2}{2\epsilon}}}}_{\gamma(Z_n,k)} \{\ln\pi_k + \ln\left(\frac{1}{\sqrt{2\pi\epsilon}} e^{-\frac{1}{2\epsilon}\|x_n-\mu_k\|^2}\right)\}$

Consider $\gamma(Z_{n,k})$. As $\epsilon\to 0$, $-\frac{1}{2\epsilon}\to -\infty$ &

$e^{-\frac{1}{2\epsilon}\|x_n-\mu_j\|^2} \to 0$. However, the term in the denominator with smallest $\|x_n-\mu_j\|^2$ will become 0 most slowly, so $\gamma(Z_{n,k})$ will be 1 for that

For the others the denominator will [illegible] a different value (not as slow) making $\gamma(z_{nk})$ 0 for that.

Thus, $\gamma(z_{n,k}) \to r_{n,k}$ for $\epsilon \to 0$

Thus,

$$\lim_{\epsilon \to 0} E[\ln p(X, Z \mid \mu, \Sigma, \pi)] = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\{ \ln \pi_k + \ln \left[\frac{1}{\sqrt{2\pi\epsilon}} e^{-\frac{1}{2}\|x-\mu\|^2} \right] \right\}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\{ \ln \pi_k - \frac{1}{2} \lg(2\pi\epsilon) + \frac{(-1)}{2} \|x_n - \mu_k\|^2 \right\}$$

$$= -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 + \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\{ \ln \pi_k - \frac{1}{2} \lg 2\pi\epsilon \right\}$$

$$\lg\left(\frac{1}{\epsilon}\right) = -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 + \text{const.} \quad (\text{as } \epsilon \to 0)$$

Thus, in the case of ϵI covariance & $\epsilon \to 0$, maximizing the complete data log-likelihood is equivalent to minimizing the distortion function of k-means.
(as it is $\propto -J$).