1) Part 1

Gaussian distribution:

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-m)^T \Sigma^{-1} (x-m)\right)$$

Mixture of models:

$$p(x) = \sum_{i=1}^{K} \pi_i \mathcal{N}(x \mid \bar{m}_i, \Sigma_i)$$

Likelihood of seen data:

$$p(X) = \prod_{i=1}^{N} p(x_i) = \prod_{i=1}^{N} \sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \bar{m}_j, \Sigma_j)$$

$$NLL = -\sum_{i=1}^{N} \log\left(\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \bar{m}_j, \Sigma_j)\right)$$

The optimization problem is:

$$\begin{cases} \min & -\sum_{i=1}^{N} \log\left(\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \bar{m}_j, \Sigma_j)\right) \\ \text{s.t.} & \sum_{j=1}^{K} \pi_j = 1, \quad \pi_j \geq 0 \ \forall j \end{cases}$$

Lagrangian:

$$\mathcal{L} = -\sum_{i=1}^{N} \log\left(\sum_{j=1}^{K} \pi_j \mathcal{N}\left(x_i \mid \bar{m}_j, \Sigma_j\right)\right) + \lambda\left(\sum_{j=1}^{K} \pi_j - 1\right) - \langle \lambda, \pi \rangle$$

a)

$$\frac{\partial \mathcal{L}}{\partial \bar{m}_k} = \sum_{i=1}^{N} \underbrace{\frac{\pi_k \cdot \mathcal{N}\left(x_i \mid \bar{m}_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(x_i \mid \bar{m}_j, \Sigma_j\right)}}_{\gamma(z_{ik})} \cdot \Sigma_k^{-1}\left(\bar{m}_k - x_i\right) = 0$$

$$\Sigma_k^{-1} \left(\overbrace{\sum_{i=1}^{N} \gamma(z_{ik})}^{\in \mathbb{R}}\right) \bar{m}_k = \Sigma_k^{-1} \left(\sum_{i=1}^{N} \gamma(z_{ik})\, x_i\right)$$

$$\boxed{\bar{m}_k = \frac{\sum_{i=1}^{N} \gamma(z_{ik})\, x_i}{\sum_{i=1}^{N} \gamma(z_{ik})}}$$

b)

$$\frac{\partial \mathcal{L}}{\partial \left(\Sigma_k^{-1}\right)} = -\sum_{i=1}^{N} \left[\frac{\pi_k \mathcal{N}\left(x_i \mid \bar{m}_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(x_i \mid \bar{m}_j, \Sigma_j\right)}\right] \left(-\frac{1}{2}\left(x_i - \bar{m}_k\right)\left(x_i - \bar{m}_k\right)^T\right)$$

way easier to calculate

$$+ \frac{\pi_K \cdot \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \frac{1}{2}\sqrt{\det \Sigma^{-1}} \cdot \Sigma \cdot \exp\left(-\frac{1}{2}(x_i - m_k)^T \Sigma^{-1} (x_i - m_k)\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i | \bar{m}_j, \Sigma_j)} \Bigg) =$$

$$// \quad \frac{\partial \sqrt{\det M}}{\partial M} = \frac{(\det M) M^{-1}}{2\sqrt{\det M}} = \frac{1}{2}\sqrt{\det M}\, M^{-1} \quad //$$

$$= -\sum_{i=1}^{N} \left[\gamma(z_{ik}) \cdot \left(\left(-\frac{1}{2}\right)(x_i - m_k)(x_i - m_k)^T + \frac{1}{2}\Sigma_k \right) \right] = 0$$

$$\left(\sum_{i=1}^{N} \gamma(z_{ik}) \right) \Sigma_k = \sum_{i=1}^{N} \gamma(z_{ik}) (x_i - m_k)(x_i - m_k)^T \Leftrightarrow$$

$$\boxed{\Sigma_k = \frac{\sum_{i=1}^{N} \gamma(z_{ik}) (x_i - m_k)(x_i - m_k)^T}{\sum_{i=1}^{N} \gamma(z_{ik})}}$$

c) $$\frac{\partial L}{\partial \pi_k} = -\sum_{i=1}^{N} \frac{\mathcal{N}_k(x_i | m_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}_j (x_i | \bar{m}_j, \Sigma_j)} + \gamma - \lambda_k = 0$$

$$\Updownarrow$$

$$-\sum_{i=1}^{N} \gamma(z_{ik}) + \nu \pi_k - \lambda_k \pi_k = 0 \qquad (*)$$

Summing up above equality over all k, we get:

$$-\sum_{k=1}^{K} \sum_{i=1}^{N} \gamma(z_{ik}) + \nu \sum_{k=1}^{K} \pi_k - \langle \lambda, \pi \rangle = 0$$

Complementary slackness gives: $\langle \lambda, \pi \rangle = 0$.
Moreover, $\sum \pi_k = 1$. Hence,

$$-\sum_{k=1}^{K} \sum_{i=1}^{N} \gamma(z_{ik}) + \nu = 0 \Rightarrow \nu = \sum_{k=1}^{K} \sum_{i=1}^{N} \gamma(z_{ik})$$

Since we presume number of classes is fixed, $\lambda = \bar{0}$.
From $(*)$ we get

$$-\sum_{i=1}^{N} \gamma(z_{ik}) + \nu \pi_k = 0 \Rightarrow$$

$$\boxed{\pi_k = \frac{\sum_{i=1}^{N} \gamma(z_{ik})}{\sum_{k=1}^{K} \sum_{i=1}^{N} \gamma(z_{ik})}}$$

EM-algorithm

1) E-step: estimate all $\gamma(z_{ik})$
2) M-step: recompute m_k, Σ_k, π_k, using derived formulas

Part 2

Let us look at gaussian distribution again:

$$\mathcal{N}(x \mid m, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\bar{x}-\bar{m})\Sigma^{-1}(\bar{x}-\bar{m})\right)$$

Let $\Sigma = \varepsilon I$, where $\varepsilon > 0$ is „close" to zero.

Then,

$$\mathcal{N}(x \mid \bar{m}_k) = \frac{1}{(2\pi\varepsilon)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\varepsilon}\|\bar{x}-\bar{m}_k\|^2\right)$$

And $\gamma(z_{ik})$ becomes

$$\gamma(z_{ik}) = \frac{\pi_k \exp\left(-\frac{\|\bar{x}_i - m_k\|^2}{2\varepsilon}\right)}{\sum_{j=1}^{K} \pi_j \exp\left(-\frac{\|x_i - m_j\|^2}{2\varepsilon}\right)} =$$

$$= \frac{1}{\sum_{\substack{j=1,\\ j\neq k}}^{K} \frac{\pi_j}{\pi_k} \exp\left(\frac{\|\bar{x}_i - m_k\|^2 - \|x_i - m_j\|^2}{2\varepsilon}\right) + 1} \longrightarrow$$

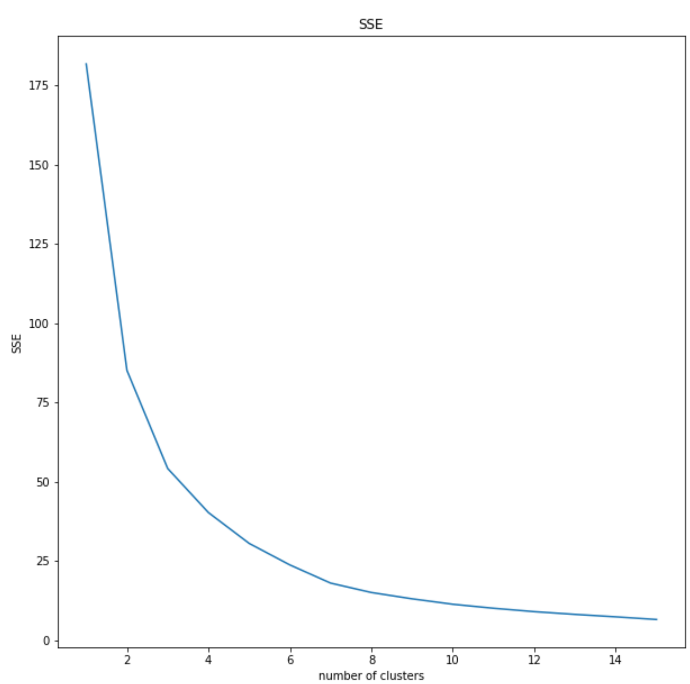
$$\longrightarrow \begin{cases} 1, & \text{if} \quad \underset{j}{\operatorname{argmin}} \|\bar{x}_i - m_j\|^2 = k, \\ 0, & \text{otherwise} \end{cases}$$

when $\varepsilon \to 0$.

That's why „soft" assignment transforms to hard one. We derive k means algorithm.

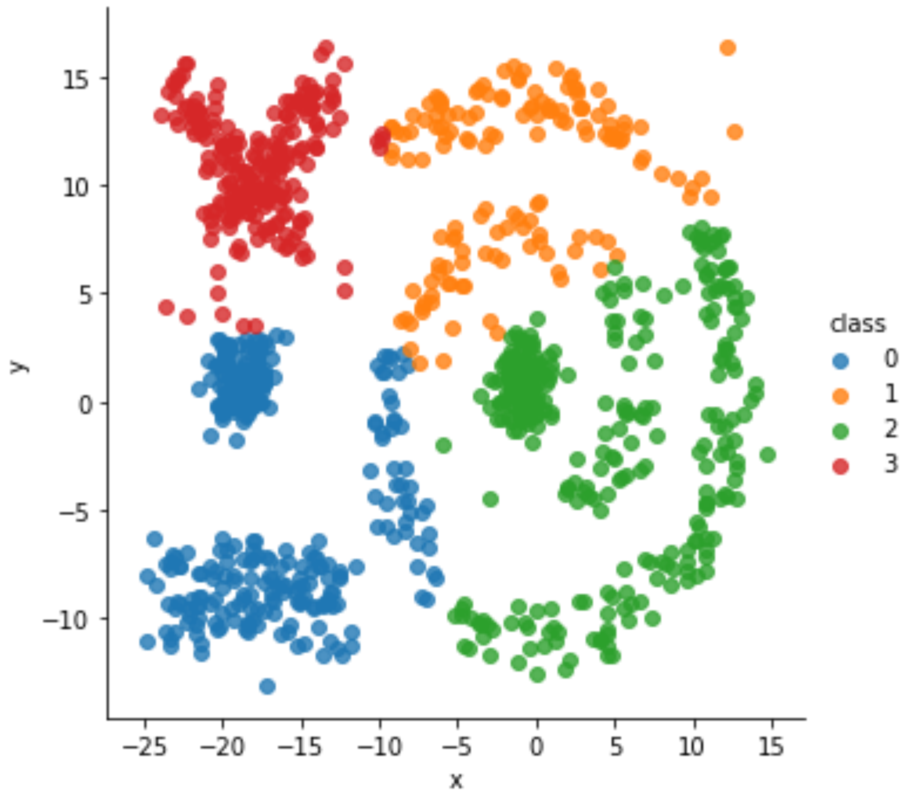
Part 3

1,2) Having followed the task, we get the following plot:



Clearly, SSE plot is of no use in a clustering problem. If number of clusters is equal to # of points, SSE is zero.

3)



Best k was chosen to be 4.

a) As can be seen from the plot, kmeans fails to choose "correct" clusters (one in right half of image, 3 in left half).

b) I think mixture of gaussians may perform better (if covariances of left 3 clusters turn out to be small, then EM may define clusters correctly).