Problem 5: Justification of the K-means Algorithm (10 pts)

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ denote the expression levels of n genes in p samples, with x_{ij} indicating the expression of gene i in sample j. Let $C_1, \ldots, C_K$ denote the K non-overlapping clusters, each containing a subset of $\{1, \ldots, n\}$, with $\cup_{k=1}^{K} C_k = \{1, \ldots, n\}$. Let $|C_k|$ denote the size of cluster k and $m_k = (m_{k1}, \ldots, m_{kp})'$ be the center of cluster k. The objective function to minimize is

$$f(C_1, \ldots, C_K, m_1, \ldots, m_K) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - m_{kj})^2 = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2.$$

1. In the first step of the $(t+1)$-th iteration of the algorithm ($t = 0, 1, \ldots$), given the clusters from the t-th iteration $C_1^{(t)}, \ldots, C_K^{(t)}$. Show that updating the cluster centers as

$$m_{kj}^{(t+1)} = \frac{1}{|C_k^{(t)}|} \sum_{i \in C_k^{(t)}} x_{ij}, \; j = 1, \ldots, p$$

satisfies that

$$f\left(C_1^{(t)}, \ldots, C_K^{(t)}, m_1^{(t+1)}, \ldots, m_K^{(t+1)}\right) \leq f\left(C_1^{(t)}, \ldots, C_K^{(t)}, m_1^{(t)}, \ldots, m_K^{(t)}\right).$$

① From the problem statement

$$f(C_1, \ldots, C_K, m_1, \ldots, m_K) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - m_{kj})^2 = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Thus,

$$f(C_1^{(t)}, \ldots, C_K^{(t)}, m_1^{(t)}, \ldots, m_K^{(t)}) = \sum_{k=1}^{K} \sum_{i \in C_k^{(t)}} \sum_{j=1}^{p} (x_{ij} - m_{kj}^{(t)})^2 = \sum_{k=1}^{K} \frac{1}{|C_k^{(t)}|} \sum_{i,i' \in C_k^{(t)}} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Rearranging the order of the summation

$$= \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i \in C_k^{(t)}} (x_{ij} - m_{kj}^{(t)})^2$$

To minimize the sum of the squared distances between a group of points & its cluster center at time t, we must minimize:

$$\sum_{i \in C_k^{(t)}} (x_{ij} - m_{kj}^{(t)})^2$$

We take the partial derivative w.r.t. ar current cluster center $m_{kj}^{(t)}$ & set it equal to 0

$$\frac{d}{dm_{kj}^{(t)}} \sum_{i \in C_k^{(t)}} (x_{ij} - m_{kj}^{(t)})^2 = 0$$

We move differentiation inside as a linear operator

$$\sum_{i \in C_k^{(t)}} \frac{d}{dm_{kj}^{(t)}} (x_{ij} - m_{kj}^{(t)})^2 = \sum_{i \in C_k^{(t)}} 2(x_{ij} - m_{kj}^{(t)})(-1)$$

chain rule

Move constants outside

Distribute the summation

$$= -2 \sum_{i \in C_k^{(t)}} (x_{ij} - m_{kj}^{(t)}) = -2 \sum_{i \in C_k^{(t)}} x_{ij} + 2 \sum_{i \in C_k^{(t)}} m_{kj}^{(t)}$$

$$2 \sum_{i \in C_k^{(t)}} m_{kj}^{(t)} = 2 \sum_{i \in C_k^{t}} x_{ij}$$

Doesn't depend on i, so $\sum_{i \in C_k^{(t)}} m_{kj}^{(t)} = m_{kj}^{(t)} + \ldots + m_{kj}^{(t)}$ ($C_k^{(t)}$ times)

$$|C_k^{(t)}|\, m_{kj}^{(t)} = \sum_{i \in C_k^t} x_{ij}$$

$$m_{kj}^{(t)} = \frac{1}{|C_k^{(t)}|} \sum_{i \in C_k^t} x_{ij}$$

Thus,

$$m_{kj}^{(t+1)} = \frac{1}{|C_k^{(t)}|} \sum_{i \in C_k^t} x_{ij}$$

minimizes the sum of the squared distances between a group of points & its cluster center & consequently $FCC(C_1^{(t)}, \dots, C_K^{(t)}, m_1^{(t)}, \dots, m_K^{(t)})$. $m_{kj}^{(t+1)}$ satisfies

$$FCC(C_1^{(t)}, \dots, C_K^{(t)}, m_1^{(t)}, \dots, m_K^{(t+1)}) \leq FCC(C_1^{(t)}, \dots, C_K^{(t)}, m_1^{(t)}, \dots, m_K^{(t)})$$

2. In the second step of the $(t+1)$-th iteration of the algorithm, given the cluster centers from the first step $m_1^{(t+1)}, \dots, m_K^{(t+1)}$. Show that if we update the cluster membership of gene i as

$$c(i)^{(t+1)} = \underset{k \in \{1,\dots,K\}}{\arg\min} \sum_{j=1}^{p} \left(x_{ij} - m_{kj}^{(t+1)}\right)^2,$$

the resulting updated clusters

$$C_k^{(t+1)} = \{i : c(i)^{(t+1)} = k\},\ k = 1, \dots, K$$

satisfy that

$$f\left(C_1^{(t+1)}, \dots, C_K^{(t+1)}, m_1^{(t+1)}, \dots, m_K^{(t+1)}\right) \leq f\left(C_1^{(t)}, \dots, C_K^{(t)}, m_1^{(t+1)}, \dots, m_K^{(t+1)}\right).$$

② From problem 5.1, we have shown that the new cluster centers are computed as

$$m_{kj}^{(t+1)} = \frac{1}{|C_k^{(t)}|} \sum_{i \in C_k^t} x_{ij}$$

Thus, the new loss function becomes

$$FCC(C_1^{(t)}, \dots, C_K^{(t)}, m_1^{(t+1)}, \dots, m_K^{(t+1)}) = \sum_{k=1}^{K} \sum_{i \in C_k^{(t)}} \sum_{j=1}^{p} (x_{ij} - m_{kj}^{(t+1)})^2$$

From the problem statement,

$$c(i)^{(t+1)} = \underset{k \in \{1, \dots, K\}}{\operatorname{argmin}} \sum_{j=1}^{p} (x_{ij} - m_{kj}^{(t+1)})^2$$

Thus,

$$\sum_{j=1}^{p} (x_{C(i)^{t+1} j} - m_{kj}^{(t+1)}) \leq \sum_{j=1}^{p} (x_{ij} - m_{kj}^{(t+1)})^2$$

as $C(i)^{t+1}$ is the argument that minimizes $\sum_{j=1}^{p} (x_{ij} - m_{kj}^{(t+1)})^2$

And the resulting updated clusters

$$C_k^{(t+1)} = \{i : c(i)^{(t+1)} = k\},\ k = 1, \dots, K$$

satisfy

$$FCC(C_1^{(t+1)}, \dots, C_K^{(t+1)}, m_1^{(t+1)}, \dots, m_K^{(t+1)}) \leq FCC(C_1^{(t)}, \dots, C_K^{(t)}, m_1^{(t+1)}, \dots, m_K^{(t+1)})$$

Problem 6: Practice of the Hierarchical Clustering (10 pts)

Suppose that we have four observations, for which we have a dissimilarity matrix

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.9 \\ 0.7 & 0.8 & 0.9 & \end{bmatrix}.$$

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

1. On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchical clustering on these four observations using complete linkage. Be sure to indicate on the plot the height at which each merging occurs, as well as the observations corresponding to each leaf in the dendrogram. Cut the dendrogram to obtain two clusters. Which observations are in each cluster?

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.9 \\ 0.7 & 0.8 & 0.9 & \end{bmatrix} \end{array}$$

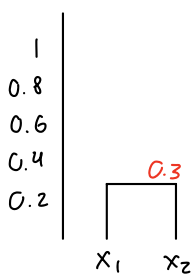
① Complete linkage is given as

$$d_{CL}(A, C) = \max_{x_i \in A,\, x_j \in C} d_{ij}$$

$d(x_1, x_2) = 0.3$ $\quad d(x_2, x_3) = 0.5$ $\quad d(x_3, x_4) = 0.9$
$d(x_1, x_3) = 0.4$ $\quad d(x_2, x_4) = 0.8$
$d(x_1, x_4) = 0.7$

For hierarchical clustering, we start with n datapoints
→ Find the two closest data points
→ Merge
→ Repeat

x_1 & x_2 are the closest datapoints with the lowest dissimilarity / highest similarity of 0.3

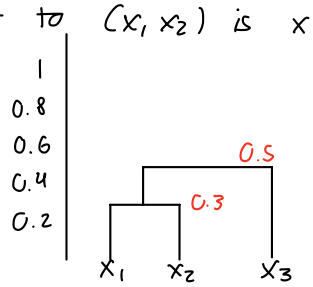


Since this is complete linkage, we find the max distance between clusters & points

$$d_{CL}((x_1, x_2), (x_3)) = \max(d(x_1, x_3), d(x_2, x_3)) = \max(0.4, 0.5) = 0.5$$

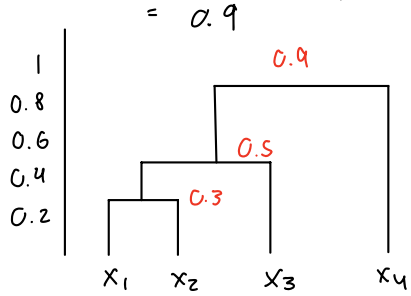
$$d_{CL}((x_1, x_2), (x_4)) = \max(d(x_1, x_4), d(x_2, x_4)) = \max(0.7, 0.8) = 0.8$$

The closest point to (x_1, x_2) is x_3 with a dissimilarity of 0.5



We calculate the distance of x_4 to the rest of the points, taking the max as our splitting threshold

$$d_{cl}((x_1,x_2,x_3),(x_4)) = \max(d(x_1,x_4), d(x_2,x_4), d(x_3,x_4))$$
$$= \max(0.7, 0.8, 0.9)$$
$$= 0.9$$



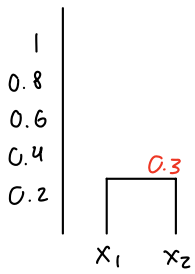
If we want to cut the dendrogram into two clusters, x_1, x_2, & x_3 belong in one cluster while x_4 is its own cluster.

2. Repeat 1, this time using single linkage.

② Single linkage is given as

$$d_{SL}(A,C) = \min_{x_i \in A, x_j \in C} d_{ij}$$

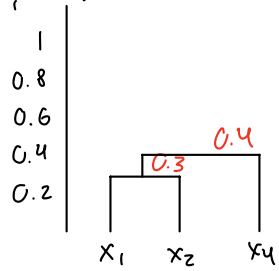
x_1 & x_2 are the closest datapoints with the lowest dissimilarity/highest similarity of 0.3



Since this is single linkage, we find the min distance between clusters & points

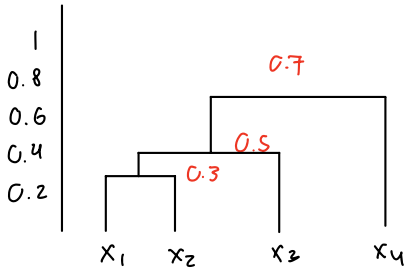
$$d_{cl}((x_1,x_2),(x_3)) = \min(d(x_1,x_3), d(x_2,x_3))$$
$$= \min(0.4, 0.5)$$
$$= 0.4$$

$$d_{cl}((x_1,x_2),(x_4)) = \min(d(x_1,x_4), d(x_2,x_4))$$
$$= \min(0.7, 0.8)$$
$$= 0.7$$



We calculate the distance of x_4 to the rest of the points, taking the min as our splitting threshold

$$d_{cl}((x_1,x_2,x_3),(x_4)) = \min(d(x_1,x_4), d(x_2,x_4), d(x_3,x_4))$$
$$= \min(0.7, 0.8, 0.9)$$
$$= 0.7$$



Problem 7: EM Algorithm for the Gaussian Mixture Model (20 pts)

In the following Gaussian Mixture Model

$$X_i|Z_i = 0 \sim N(\mu_0, \sigma_1^2);$$

$$X_i|Z_i = 1 \sim N(\mu_1, \sigma_2^2);$$

$$Z_i \sim Bernoulli(\gamma),\ i = 1, \ldots, n,$$

where X_i's are observable random variables, and Z_i's are hidden random variables.

Given observed data points $x_1, \ldots, x_n$, derive the EM algorithm for estimating μ_0, μ_1, σ_1^2, σ_2^2 and γ in the following steps .

1. Write down the complete log-likelihood $\ell(\mu_0, \mu_1, \sigma_1^2, \sigma_2^2, \gamma)$ in terms of $x_1, \ldots, x_n$ and $Z_1, \ldots, Z_n$.

The problem statement defines

$$X_i | Z_i = 0 \sim N(\mu_0, \sigma_1^2)$$
$$X_i | Z_i = 1 \sim N(\mu_1, \sigma_2^2)$$
$$Z_i \sim \text{Bernoulli}(\gamma),\ i = 1, \ldots, n$$

Where γ is the probability of $Z_i = 1$

To find the log likelihood, we first find the complete likelihood

Note: Probability chain rule $P(A, B, C) = P(a | b, c)\, p(b, c)$

$$L(x_1, z_1, x_2, z_2, \ldots, x_n, z_n, \mu_0, \mu_1, \sigma_1^2, \sigma_2^2, \gamma) = \prod_{i=1}^{n} P(X_i, Z_i, \mu_0, \mu_1, \sigma_1^2, \sigma_2^2, \gamma)$$
$$= \prod_{i=1}^{n} P(X_i | Z_i, \mu_0, \mu_1, \sigma_1^2, \sigma_2^2, \gamma)\, P(Z_i, \mu_0, \mu_1, \sigma_1^2, \sigma_2^2, \gamma)$$

Note: The PDF of $X \sim N(\mu, \sigma^2)$ & $Z \sim \text{Bernoulli}(\gamma)$ is given as $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$; $f(z) = p^{\gamma}(1-p)^{1-\gamma}$

$$= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma_1} \cdot e^{-\frac{(x_i-\mu_0)^2}{2\sigma_1^2}}\right]^{1-z_i} (1-\gamma)^{1-z_i} \cdot \left[\frac{1}{\sqrt{2\pi}\sigma_2} \cdot e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}}\right]^{z_i} \gamma^{z_i}$$
$$= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma_1} \cdot e^{-\frac{(x_i-\mu_0)^2}{2\sigma_1^2}} \cdot (1-\gamma)\right]^{1-z_i} \cdot \left[\frac{1}{\sqrt{2\pi}\sigma_2} \cdot e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \cdot \gamma\right]^{z_i}$$

$$l_c(x_1, z_1, x_2, z_2, \ldots, x_n, z_n, \mu_0, \mu_1, \sigma_1^2, \sigma_2^2, \gamma) = \log(L(x_1, z_1, x_2, z_2, \ldots, x_n, z_n, \mu_0, \mu_1, \sigma_1^2, \sigma_2^2, \gamma))$$
$$= \log\left(\prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma_1} \cdot e^{-\frac{(x_i-\mu_0)^2}{2\sigma_1^2}} \cdot (1-\gamma)\right]^{1-z_i} \cdot \left[\frac{1}{\sqrt{2\pi}\sigma_2} \cdot e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \cdot \gamma\right]^{z_i}\right)$$
$$= \sum_{i=1}^{n} \log\left(\left[\frac{1}{\sqrt{2\pi}\sigma_1} \cdot e^{-\frac{(x_i-\mu_0)^2}{2\sigma_1^2}} \cdot (1-\gamma)\right]^{1-z_i} \cdot \left[\frac{1}{\sqrt{2\pi}\sigma_2} \cdot e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \cdot \gamma\right]^{z_i}\right)$$
$$= \sum_{i=1}^{n} \log\left(\left[\frac{1}{\sqrt{2\pi}\sigma_1} \cdot e^{-\frac{(x_i-\mu_0)^2}{2\sigma_1^2}} \cdot (1-\gamma)\right]^{1-z_i}\right) + \log\left(\left[\frac{1}{\sqrt{2\pi}\sigma_2} \cdot e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \cdot \gamma\right]^{z_i}\right)$$
$$= \sum_{i=1}^{n} (1-z_i) \log\left(\left[\frac{1}{\sqrt{2\pi}\sigma_1} \cdot e^{-\frac{(x_i-\mu_0)^2}{2\sigma_1^2}} \cdot (1-\gamma)\right]\right) + z_i \log\left(\left[\frac{1}{\sqrt{2\pi}\sigma_2} \cdot e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \cdot \gamma\right]\right)$$

$$\log\left(\left[\frac{1}{\sqrt{2\pi}\sigma_1} \cdot e^{-\frac{(x_i-\mu_0)^2}{2\sigma_1^2}} \cdot (1-\gamma)\right]\right) = \log\left(\frac{1}{\sqrt{2\pi}\sigma_1}\right) + \log\left(e^{-\frac{(x_i-\mu_0)^2}{2\sigma_1^2}}\right) + \log(1-\gamma)$$
$$= \log\left(\frac{1}{\sqrt{2\pi}\sigma_1}\right) - \frac{(x_i-\mu_0)}{2\sigma_1^2} + \log(1-\gamma)$$

$$\log\left(\left[\frac{1}{\sqrt{2\pi}\sigma_2} \cdot e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \cdot \gamma\right]\right) = \log\left(\frac{1}{\sqrt{2\pi}\sigma_1}\right) + \log\left(e^{-\frac{(x_i-\mu_0)^2}{2\sigma_1^2}}\right) + \log(\gamma)$$
$$= \log\left(\frac{1}{\sqrt{2\pi}\sigma_2}\right) - \frac{(x_i-\mu_1)}{2\sigma_2^2} + \log(\gamma)$$

$$l_c(\mu_0, \mu_1, \sigma_1^2, \sigma_2^2, \gamma) = \sum_{i=1}^{n} (1-z_i)\left(\log\left(\frac{1}{\sqrt{2\pi}\sigma_1}\right) - \frac{(x_i-\mu_0)}{2\sigma_1^2} + \log(1-\gamma)\right) + z_i\left(\log\left(\frac{1}{\sqrt{2\pi}\sigma_2}\right) - \frac{(x_i-\mu_1)}{2\sigma_2^2} + \log(\gamma)\right)$$

2. In the E-step of the $(t+1)$-th iteration $(t = 0, 1, 2, \ldots)$, derive the conditional expectation of Z_i given x_i and the current parameter estimates $\left(\widehat{\mu}_0^{(t)}, \widehat{\mu}_1^{(t)}, (\widehat{\sigma}_1^{(t)})^2, (\widehat{\sigma}_2^{(t)})^2, \widehat{\gamma}^{(t)}\right)$:

$$\tau_i^{(t+1)} = E\left[Z_i | x_i, \widehat{\mu}_0^{(t)}, \widehat{\mu}_1^{(t)}, (\widehat{\sigma}_1^{(t)})^2, (\widehat{\sigma}_2^{(t)})^2, \widehat{\gamma}^{(t)}\right] .$$

$$\tau_i^{(t+1)} = E\left[z_i \mid x_i; \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)} \right]$$

Note: The expectation of a discrete random variable X is given as

$$E[X] = \sum_{x \in R_X} x\, p(x)$$

$$\tau_i^{(t+1)} = 1 \cdot P\left(z_i = 1 \mid x_i; \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)}\right] +$$

$$\cancel{0 \cdot P\left(z_i = 0 \mid x_i; \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)}\right]}$$

Note: Bayes Rule & the law of total probability states

$$P(B_J \mid A) = \frac{P(A \mid B_J)\, P(B_J)}{\sum_{i=1}^{n} P(A \mid B_i) P(B_i)}$$

$$P\left(z_i = 1 \mid x_i, \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)}\right]$$

$$= \frac{P(x_i \mid z_i = 1, x_i, \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})\, P(z_i = 1; x_i, \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})}{\sum_{z_i \in \{0,1\}} P(x_i \mid z_i = z_i, x_i, \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})\, P(z_i = z_i; x_i, \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})}$$

$$= \frac{P(x_i \mid z_i = 1, x_i, \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})\, P(z_i = 1; \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})}{P(x_i \mid z_i = 0, x_i, \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})\, P(z_i = 0; \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)}) + P(x_i \mid z_i = 1, x_i, \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})\, P(z_i = 1; \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})}$$

From problem 5.1

$$P(x_i \mid z_i = 0, x_i, \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})\, P(z_i = 1; \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})$$

$$= \left[\frac{1}{\sqrt{2\pi}\,\sigma_1} \cdot e^{-\frac{(x_i - \mu_0)^2}{2\sigma_1^2}} \cdot (1-\gamma)\right]^{1-0} \cdot \left[\frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \cdot \gamma\right]^{0}$$

$$= \left[\frac{1}{\sqrt{2\pi}\,\sigma_1} \cdot e^{-\frac{(x_i - \mu_0)^2}{2\sigma_1^2}} \cdot (1-\gamma)\right]$$

$$P(x_i \mid z_i = 1, x_i, \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})\, P(z_i = 1; \hat{\mu}_0^{(t)}, \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^{(t)})^2, (\hat{\sigma}_2^{(t)})^2, \hat{\gamma}^{(t)})$$

$$= \left[\frac{1}{\sqrt{2\pi}\,\sigma_1} \cdot e^{-\frac{(x_i - \mu_0)^2}{2\sigma_1^2}} \cdot (1-\gamma)\right]^{1-1} \cdot \left[\frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \cdot \gamma\right]^{1}$$

$$= \left[\frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \cdot \gamma\right]^{1}$$

$$\tau_i^{(t+1)} = \frac{\frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \cdot \gamma}{\frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \cdot \gamma + \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x_i-\mu_0)^2}{2\sigma_1^2}} \cdot (1-\gamma)}$$

3. In the M-step of the $(t+1)$-th iteration, derive the updated parameter estimates based on $x_1, \ldots, x_n$ and $\tau_1^{(t+1)}, \ldots, \tau_n^{(t+1)}$.

$$\left(\widehat{\mu}_0^{(t+1)}, \widehat{\mu}_1^{(t+1)}, (\widehat{\sigma}_1^{(t+1)})^2, (\widehat{\sigma}_2^{(t+1)})^2, \widehat{\gamma}^{(t+1)}\right) .$$

Using the formulas given in the notes sheet & as derived in class,

$$\hat{\gamma}^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_i^{(t)}$$

$$\hat{\mu}_0^{(t+1)} = \frac{\sum_{i=1}^{n} x_i (1 - \tau_i^{(t)})}{\sum_{i=1}^{n} (1 - \tau_i^{(t)})} \qquad \hat{\mu}_1^{(t+1)} = \frac{\sum_{i=1}^{n} x_i \tau_i^{(t)}}{\sum_{i=1}^{n} \tau_i^{(t)}}$$

$$(\sigma_1^2)^{(t+1)} = \frac{\sum_{i=1}^{n} (x_i - \mu_0^{(t+1)})^2 (1 - \tau_i^{(t+1)})}{\sum_{i=1}^{n} (1 - \tau_i^{(t+1)})} \qquad (\sigma_2^2)^{(t+1)} = \frac{\sum_{i=1}^{n} (x_i - \mu_0^{(t+1)})^2 \tau_i^{(t+1)}}{\sum_{i=1}^{n} \tau_i^{(t+1)}}$$