ML2024 Fall Assignment 3

Wei-Chen Chang, R12227118

Due: Nov, 8 2024

Problem 1

Problem 1 were computed by Python, the code can be seen on NTUCool.

1.

The columns are the 3 principal axes.

	PC1	PC2	PC3
1	0.617	-0.678	0.400
2	0.589	0.734	0.338
3	0.523	-0.027	-0.852

2.

The columns are the principal components of each sample.

	1	2	3	4	5	6	7	8	9	10
PC1	-7.187	-0.759	3.070	-2.608	1.823	-3.355	4.415	-3.466	2.314	5.752
PC2	-1.373	0.944	4.451	2.979	4.754	-3.919	-2.556	1.731	-6.034	-0.976
PC3	-2.251	-0.730	-3.188	-1.930	4.252	2.528	-2.140	2.278	0.204	0.977

3.

The average reconstruction error based of first two PC are 5.472032912651863.

Problem 2

1.

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \int q_\phi(\mathbf{z}|\mathbf{x}) \log p_\theta(\mathbf{x})\, dz \\
&= \int q_\phi(\mathbf{z}|\mathbf{x}) \left[\log p_\theta(\mathbf{x},\mathbf{z}) - \log p_\theta(\mathbf{z}|\mathbf{x})\right] dz \\
&= \int q_\phi(\mathbf{z}|\mathbf{x}) \left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} - \log \frac{p_\theta(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] dz \\
&= \int q_\phi(\mathbf{z}|\mathbf{x}) \left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] dz + \int q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\, dz \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \left(\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right) + \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z}|\mathbf{x})\right],
\end{aligned}$$

and $\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \left(\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right)$ is the ELBO (denote as $\mathcal{L}_{\phi,\theta}(\mathbf{x})$).

2.

Let $\mathbf{z} = g(\epsilon, \phi, \theta)$, where $\epsilon \sim p(\epsilon)$ be a r.v., which is independent of ϕ and θ.

First observed that for a function of $\mathbf{z}$, $f(\mathbf{z})$, its expectation now can be expressed as:

$$\mathbb{E}_{\mathbf{z}\sim q_\phi(\cdot|\mathbf{x})}[f(\mathbf{z})] = \mathbb{E}_{p(\epsilon)}[f(\mathbf{z})].$$

And

$$\begin{aligned}
\nabla_\phi \mathbb{E}_{\mathbf{z}\sim q_\phi(\cdot|\mathbf{x})}[f(\mathbf{z})] &= \nabla_\phi \mathbb{E}_{\epsilon\sim p(\epsilon)}[f(\mathbf{z})] \\
&= \mathbb{E}_{\epsilon\sim p(\epsilon)}[\nabla_\phi f(\mathbf{z})] \\
&\approx \nabla_\phi f(\mathbf{z}).
\end{aligned}$$

In the last line, we generate $\mathbf{z}$ from random noise ϵ to approximate expectation. Note now the expectation and gradient are interchangeable.

We apply the trick on ELBO:

$$\mathcal{L}_{\phi,\theta}(\mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] = \mathbb{E}_{\epsilon\sim p(\epsilon)}\left[\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right],$$

where $\mathbf{z} = g(\epsilon, \phi, \theta)$.

Now let

$$\tilde{\mathcal{L}}_{\phi,\theta}(\mathbf{x}) = \log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}),$$

and its gradient $\nabla_\phi \tilde{\mathcal{L}}_{\phi,\theta}(\mathbf{x})$ be the estimator of $\nabla_\phi \mathcal{L}_{\phi,\theta}(\mathbf{x})$.

Below shows the unbiasedness of $\nabla_\phi \tilde{\mathcal{L}}_{\phi,\theta}(\mathbf{x})$:

$$\begin{aligned}
\mathbb{E}_{\epsilon\sim p(\epsilon)}\left[\nabla_\phi \tilde{\mathcal{L}}_{\phi,\theta}(\mathbf{x})\right] &= \mathbb{E}_{\epsilon\sim p(\epsilon)}\left[\nabla_\phi\left(\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right)\right] \\
&= \nabla_\phi\left[\mathbb{E}_{\epsilon\sim p(\epsilon)}\left[\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right]\right] \\
&= \nabla_\phi \mathcal{L}_{\phi,\theta}(x).
\end{aligned}$$

Problem 3

1.

$$\mathbf{W} = \begin{bmatrix}
0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0
\end{bmatrix}$$

2.

Since $\mathbf{W}$ is symmetric, $d_i = \sum_{i=1}^{10} W_{ij}$, the sum of the i-th row of $\mathbf{W}$,

$$\mathbf{D} = diag(3,3,2,2,2,1,2,3,2,2).$$

And

$$\mathbf{L} = \mathbf{D} - \mathbf{W} = \begin{bmatrix}
3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 3 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\
-1 & 0 & 2 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
-1 & -1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2 & -1 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & 2 & 0 & 0 & -1 \\
0 & -1 & -1 & 0 & 0 & 0 & 0 & 3 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 2
\end{bmatrix}.$$

3.

See Fig.1 for 3-D scatterplot. Detailed Python code and embeddings can be seen in NTUCool code file.

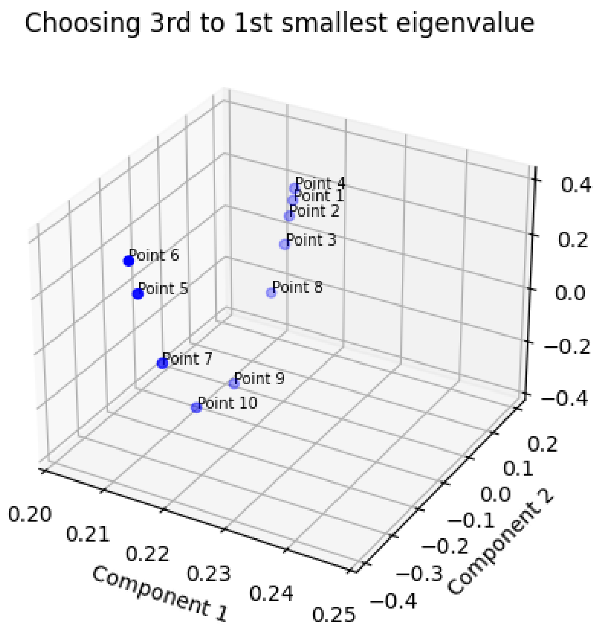

Figure 1: Scatterplot to Problem 3.3

4.

See Fig.2 for 3-D scatterplot. The output of $\text{Tr}(\mathbf{\Psi}^T\mathbf{L}\mathbf{\Psi})$ and $\mathbf{\Psi}^T\mathbf{D}\mathbf{\Psi}$ can be seen in Fig.3.

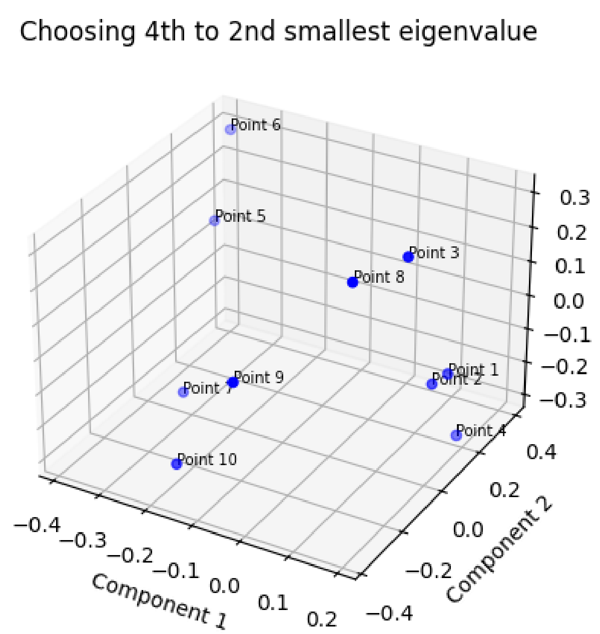

Figure 2: Scatterplot to Problem 3.4

```
The Trace of \Psi^T L \Psi:
 1.0978030751206402
The Product of \Psi^T D \Psi:
 [[1.00000000e+00 4.01623242e-16 1.31050686e-16]
 [4.01623242e-16 1.00000000e+00 1.10723734e-16]
 [1.86561837e-16 1.38479309e-16 1.00000000e+00]]
```

Figure 3: Output of $\mathrm{Tr}(\mathbf{\Psi}^T \mathbf{L} \mathbf{\Psi})$ and $\mathbf{\Psi}^T \mathbf{D} \mathbf{\Psi}$

5.

For any undirected graph, W is symmetric and $d_i = \sum_{j=1}^{N} W_{ij}$. Observe the sum over i-th row in L is:

$$d_i - \sum_{j=1}^{N} W_{ij} = 0, \quad \forall i = 1, ..., N.$$

It can be rewrite as:

$$\mathbf{D} \cdot \mathbf{1} - \mathbf{W} \cdot \mathbf{1} = \mathbf{L} \cdot \mathbf{1} = 0 \cdot \mathbf{1},$$

where $\mathbf{1} = \begin{bmatrix} 1 & 1 & ... & 1 \end{bmatrix}^T$.

This implies that $\begin{bmatrix} c & c & ... & c \end{bmatrix}^T$ is an eigenvector of $\mathbf{L}$ corresponds to eigenvalue 0.

6.

$$\begin{aligned}
\frac{1}{2} \sum_{1 \le i,j \le N} W_{ij}(f_i - f_j)^2 &= \frac{1}{2} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} f_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} f_j^2 - 2 \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} f_i f_j \right) \\
&= \frac{1}{2} \left(\sum_{i=1}^{N} f_i^2 \sum_{j=1}^{N} W_{ij} + \sum_{i=1}^{N} f_i^2 \sum_{j=1}^{N} W_{ji} - 2 \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} f_i f_j \right) \\
&= \frac{1}{2} \left(2 \sum_{i=1}^{N} f_i^2 d_i - 2 \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} f_i f_j \right) \\
&= \mathbf{f}^T \mathbf{D} \mathbf{f} - \mathbf{f}^T \mathbf{W} \mathbf{f} = \mathbf{f}^T \mathbf{L} \mathbf{f}
\end{aligned}$$

7.

If $\mathbf{f}$ is an eigenvector of $\mathbf{L}$ corresponds to eigenvalue 0, it implies that:

$$\mathbf{L}\mathbf{f} = 0 \cdot \mathbf{f} = \mathbf{0}.$$

Multiply both sides by $\mathbf{f}^T$, we have :

$$\mathbf{f}^T \mathbf{L} \mathbf{f} = 0.$$

8.

From 6. and 7., we see that if $\mathbf{f}$ is an eigenvector corresponds to 0, it satisfy that:

$$\mathbf{f}^T\mathbf{L}\mathbf{f} = \frac{1}{2}\sum_{1\le i,j\le N} W_{ij}(f_i - f_j)^2 = 0.$$

Since W_{ij} is non-zero when there's an edge between vertex i and vertex j, this condition implies that $f_i = f_j$ if i, j connected. Moreover, because the graph is connected, $f_i = f_j$ satisfied for $i, j \in \{1, ..., N\}$, $\mathbf{f}$ must be a constant vector (i.e., the eigenvector corresponds to 0 has multiplicity of 1).

Second, observe that for any eigenvector $\mathbf{f}$ of $\mathbf{L}$ that corresponds to eigenvalue λ:

$$\mathbf{f}^T\mathbf{L}\mathbf{f} = \mathbf{f}^T\lambda\mathbf{f} = \lambda\|\mathbf{f}\|^2 = \frac{1}{2}\sum_{1\le i,j\le N} W_{ij}(f_i - f_j)^2 \ge 0.$$

This implies $\lambda \ge 0$, which means the all other eigenvalue of L are greater than 0. Therefore 0 is the smallest eigenvalue, and the second smallest one has nonzero value.

Problem 4

First, express the likelihood function of as the product of the likelihood for labeled data ($y_i \neq 0$) and the likelihood for unlabeled data ($y_i = 0$):

$$p_\theta(\{\mathbf{x_i}, y_i\}_{i=1}^N) = \prod_{i:y_i\neq 0} p_\theta(\mathbf{x_i}, y_i) \prod_{i:y_i=0} p_\theta(\mathbf{x_i}),$$

and the log-likelihood function is:

$$\begin{aligned}
\log p_\theta(\{\mathbf{x_i}, y_i\}_{i=1}^N) &= \sum_{y_i\neq 0} \log p_\theta(\mathbf{x_i}, y_i) + \sum_{i:y_i=0} \log p_\theta(\mathbf{x_i}) \\
&= \sum_{i:y_i\neq 0} \log p_\theta(\mathbf{x_i}, y_i) + \sum_{i:y_i=0}\sum_{k=1}^{K} p_\theta(y_i = k|\mathbf{x_i}) \log p_\theta(\mathbf{x_i}) \\
&= \sum_{i:y_i\neq 0} \log p_\theta(\mathbf{x_i}, y_i) + \sum_{i:y_i=0}\sum_{k=1}^{K} p_\theta(y_i = k|\mathbf{x_i}) \left[\log p_\theta(\mathbf{x_i}, y_i = k) - \log p_\theta(y_i = k|\mathbf{x_i})\right] \\
&= \sum_{i:y_i\neq 0}\sum_{k=1}^{K} \mathbf{1}_{y_i=k} \log p_\theta(\mathbf{x_i}, y_i = k) + \sum_{i:y_i=0}\sum_{k=1}^{K} p_\theta(y_i = k|\mathbf{x_i}) \log p_\theta(\mathbf{x_i}, y_i = k) \\
&\quad - \sum_{i:y_i=0}\sum_{k=1}^{K} p_\theta(y_i = k|\mathbf{x_i}) \log p_\theta(y_i = k|\mathbf{x_i}),
\end{aligned}$$

where $\mathbf{1}_{y_i=k}$ is the indicator function.

Now define $Q(\theta\|\theta^{(t)})$ as:

$$Q(\theta\|\theta^{(t)}) = \sum_{i:y_i\neq 0}\sum_{k=1}^{K} \mathbf{1}_{y_i=k} \log p_\theta(\mathbf{x_i}, y_i = k) + \sum_{i:y_i=0}\sum_{k=1}^{K} p_{\theta^{(t)}}(y_i = k|\mathbf{x_i}) \log p_\theta(\mathbf{x_i}, y_i = k).$$

In the E-step, we write down the explicit form of $Q(\theta\|\theta^{(t)})$:

$$\begin{aligned}
Q(\theta\|\theta^{(t)}) &= \sum_{i=1}^{N}\sum_{k=1}^{K} \left[\mathbf{1}_{y_i=k} + \mathbf{1}_{y_i=0} p_{\theta^{(t)}}(y_i = k|\mathbf{x_i})\right] \log p_\theta(\mathbf{x_i}, y_i = k) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \left[\mathbf{1}_{y_i=k} + \mathbf{1}_{y_i=0} \frac{p_{\theta^{(t)}}(y_i = k, \mathbf{x_i})}{\sum_{k=1}^{K} p_{\theta^{(t)}}(y_i = k, \mathbf{x_i})}\right] \log p_\theta(\mathbf{x_i}, y_i = k) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \xi_{i,k}^{(t)} \log \pi_k \mathcal{N}(\mathbf{x_i}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \xi_{i,k}^{(t)} \left(\log \pi_k - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma_k}| - \frac{1}{2}(\mathbf{x_i} - \boldsymbol{\mu_k})^T {\boldsymbol{\Sigma_k}}^{-1}(\mathbf{x_i} - \boldsymbol{\mu_k})\right),
\end{aligned}$$

where

$$\xi_{i,k}^{(t)} = \mathbf{1}_{y_i=k} + \mathbf{1}_{y_i=0} \frac{p_{\theta^{(t)}}(y_i = k, \mathbf{x_i})}{\sum_{k=1}^{K} p_{\theta^{(t)}}(y_i = k, \mathbf{x_i})} = \mathbf{1}_{y_i=k} + \mathbf{1}_{y_i=0} \frac{\pi_k^{(t)} \mathcal{N}(\mathbf{x_i}; {\boldsymbol{\mu_k}}^{(t)}, {\boldsymbol{\Sigma_k}}^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} \mathcal{N}(\mathbf{x_i}; {\boldsymbol{\mu_k}}^{(t)}, {\boldsymbol{\Sigma_k}}^{(t)})} = \mathbf{1}_{y_i=k} + \mathbf{1}_{y_i=0}\ \delta_{i,k}^{(t)}.$$

For the M-step, the goal is to maximize $Q(\theta\|\theta^{(t)})$ w.r.t. $\theta = \{\pi_k, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}\}_{k=1}^{K}$.

First we optimize over $\boldsymbol{\mu_k}$:

$$\frac{\partial}{\partial \boldsymbol{\mu_k}} Q(\theta\|\theta^{(t)}) = -\sum_{i=1}^{N} \xi_{i,k}^{(t)} \left(-{\boldsymbol{\Sigma_k}}^{-1}\mathbf{x_i} + {\boldsymbol{\Sigma_k}}^{-1}\boldsymbol{\mu_k}\right)$$

Since the second derivative over μ_i is $-\sum_{i=1}^{N} \xi_{i,k}^{(t)} < 0$, set the first derivative to 0 guarantees the maximum.

$$\sum_{i=1}^{N} \xi_{i,k}^{(t)} \left({\boldsymbol{\Sigma_k}}^{-1}\mathbf{x_i} - {\boldsymbol{\Sigma_k}}^{-1}\boldsymbol{\mu_k}\right) = 0$$

Solving the equation, we have:

$$\boldsymbol{\mu_k} = \frac{\sum_{i=1}^{N} \xi_{i,k}^{(t)} \mathbf{x_i}}{\sum_{i=1}^{N} \xi_{i,k}^{(t)}} = \frac{\sum_{i=1}^{N} \mathbf{1}_{y_i=k}\mathbf{x_i} + \mathbf{1}_{y_i=0}\ \delta_{i,k}^{(t)} \mathbf{x_i}}{\sum_{i=1}^{N} \mathbf{1}_{y_i=k} + \mathbf{1}_{y_i=0}\ \delta_{i,k}^{(t)}}$$

Hence we update ${\boldsymbol{\mu_k}}^{(t+1)}$ by

$${\boldsymbol{\mu_k}}^{(t+1)} = \frac{\sum_{i:y_i=k} \mathbf{x_i} + \sum_{i:y_i=0} \delta_{i,k}^{(t)} \mathbf{x_i}}{N_k + \sum_{i:y_i=0} \delta_{i,k}^{(t)}}$$

Second, we optimize over $\boldsymbol{\Sigma_k}$, which is equivalent to optimize over $\boldsymbol{\Sigma_k}^{-1}$. Let $\boldsymbol{\Sigma_k}^{-1} = [a_{ij}^k]$, we have,

$$
\begin{aligned}
\frac{\partial}{\partial a_{i'j}^k} Q(\theta \| \theta^{(t)}) &= \frac{\partial}{\partial a_{i'j}^k} \sum_{i=1}^{N} \sum_{k=1}^{K} \xi_{i,k}^{(t)} \left(\log \pi_k - \frac{m}{2} \log(2\pi) + \frac{1}{2} \log |\boldsymbol{\Sigma_k}^{-1}| - \frac{1}{2} (\mathbf{x_i} - \boldsymbol{\mu_k})^T \boldsymbol{\Sigma_k}^{-1} (\mathbf{x_i} - \boldsymbol{\mu_k}) \right) \\
&= \sum_{i=1}^{N} \xi_{ik}^{(t)} \left[\frac{1}{2} e_j^T \boldsymbol{\Sigma_k} e_{i'} - \frac{1}{2} \frac{\partial}{\partial a_{i'j}^k} \operatorname{tr} \left((\mathbf{x_i} - \boldsymbol{\mu_k})(\mathbf{x_i} - \boldsymbol{\mu_k})^T \boldsymbol{\Sigma_k}^{-1} \right) \right] \\
&= \sum_{i=1}^{N} \xi_{ik}^{(t)} \left[\frac{1}{2} e_j^T \boldsymbol{\Sigma_k} e_{i'} - \frac{1}{2} e_j^T \left((\mathbf{x_i} - \boldsymbol{\mu_k})(\mathbf{x_i} - \boldsymbol{\mu_k})^T \right) e_{i'} \right] \\
&= \frac{1}{2} \sum_{i=1}^{N} \xi_{ik}^{(t)} \left[e_j^T \left(\boldsymbol{\Sigma_k} - (\mathbf{x_i} - \boldsymbol{\mu_k})(\mathbf{x_i} - \boldsymbol{\mu_k})^T \right) e_{i'} \right]
\end{aligned}
$$

Set the first derivative to 0 and solving the equation, we have,

$$
\begin{aligned}
\sum_{i=1}^{N} \xi_{ik}^{(t)} \boldsymbol{\Sigma_k} &= \sum_{i=1}^{N} \xi_{ik}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \\
\boldsymbol{\Sigma}_k &= \frac{\sum_{i=1}^{N} \xi_{ik}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{N} \xi_{ik}^{(t)}} \\
&= \frac{\sum_{i=1}^{N} \mathbf{1}_{y_i=k} (\mathbf{x_i} - \boldsymbol{\mu_k})(\mathbf{x_i} - \boldsymbol{\mu_k})^T + \mathbf{1}_{y_i=0}\ \delta_{i,k}^{(t)} (\mathbf{x_i} - \boldsymbol{\mu_k})(\mathbf{x_i} - \boldsymbol{\mu_k})^T}{\sum_{i=1}^{N} \mathbf{1}_{y_i=k} + \mathbf{1}_{y_i=0}\ \delta_{i,k}^{(t)}}
\end{aligned}
$$

Hence we update $\boldsymbol{\Sigma_k}^{(t+1)}$ by

$$
\boldsymbol{\Sigma_k}^{(t+1)} = \frac{\sum_{i:y_i=k} (\mathbf{x_i} - \boldsymbol{\mu_k}^{(t+1)})(\mathbf{x_i} - \boldsymbol{\mu_k}^{(t+1)})^T + \sum_{i:y_i=0}\ \delta_{i,k}^{(t)} (\mathbf{x_i} - \boldsymbol{\mu_k})(\mathbf{x_i} - \boldsymbol{\mu_k}^{(t+1)})^T}{N_k + \sum_{i:y_i=0} \delta_{i,k}^{(t)}}
$$

Lastly, for π_k, we solve the Lagrange multiplier:

$$
\frac{\partial}{\partial \pi_k} \left(Q(\theta \| \theta^{(t)}) - \lambda \sum_{k=1}^{K} \pi_k \right) = 0.
$$

We have:

$$
\pi_k = \frac{1}{\lambda} \sum_{i=1}^{N} \xi_{i,k}^{(t)}.
$$

The constraint then can be rewrite as:

$$
\sum_{k=1}^{K} \pi_k = \frac{1}{\lambda} \sum_{i=1}^{N} \sum_{k=1}^{K} \xi_{i,k}^{(t)} = 1,
$$

Since $\sum_{i=1}^N \sum_{k=1}^K \xi_{i,k}^{(t)} = N$, we have $\lambda = N$. Therefore, we update $\pi_k^{(t+1)}$ as follows:

$$\pi_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^N \xi_{i,k}^{(t)} = \frac{N_k + \sum_{i:y_i=0}\delta_{i,k}^{(t)}}{N}$$

Problem 5

First, the log-likelihood of this model is:

$$\begin{aligned}
\sum_{i=1}^N \log p(y_i|\mathbf{x}_i;\theta) &= \sum_{i=1}^N\sum_{k=1}^K p(z=k|y_i,\mathbf{x}_i;\theta^{(t)})\log p(y_i|\mathbf{x}_i;\theta) \\
&= \sum_{i=1}^N\sum_{k=1}^K p(z=k|y_i,\mathbf{x}_i;\theta^{(t)})\left[\log p(y_i,z=k|\mathbf{x}_i;\theta) - \log p(z=k|\mathbf{x}_i;\theta)\right] \\
&= Q(\theta\|\theta^{(t)}) - \sum_{i=1}^N\sum_{k=1}^K p(z=k|y_i,\mathbf{x}_i;\theta^{(t)})\log p(z=k|\mathbf{x}_i;\theta),
\end{aligned}$$

where $Q(\theta\|\theta^{(t)}) = \sum_{i=1}^N\sum_{k=1}^K p(z=k|y_i,\mathbf{x}_i;\theta^{(t)})\log p(y_i,z=k|\mathbf{x}_i;\theta)$.

In each iteration of EM, first the E step is to compute $Q(\theta\|\theta^{(t)})$, then the M step is to maximize $Q(\theta\|\theta^{(t)})$ w.r.t $\theta^{(t)}$.

For the E-step, we write down the explicit form of $Q(\theta\|\theta^{(t)})$:

$$\begin{aligned}
Q(\theta\|\theta^{(t)}) &= \sum_{i=1}^N\sum_{k=1}^K p(z=k|y_i,\mathbf{x}_i;\theta^{(t)})\log p(y_i,z=k|\mathbf{x}_i;\theta) \\
&= \sum_{i=1}^N\sum_{k=1}^K \frac{p(y_i,z=k|\mathbf{x}_i;\theta^{(t)})}{p(y_i|\mathbf{x}_i;\theta^{(t)})}\log p(y_i,z=k|\mathbf{x}_i;\theta) \\
&= \sum_{i=1}^N\sum_{k=1}^K \delta_{ik}^{(t)}\log p(y_i,z=k|\mathbf{x}_i;\theta) \\
&= \sum_{i=1}^N\sum_{k=1}^K \delta_{ik}^{(t)}\left[\log\pi_k + \log\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\log\rho_k - \frac{1}{2\rho_k}\left(y_i - \mathbf{w}_k^T\mathbf{x}_i\right)^2\right],
\end{aligned}$$

where

$$\delta_{ik}^{(t)} = \frac{p(y_i,z=k|\mathbf{x}_i;\theta^{(t)})}{p(y_i|\mathbf{x}_i;\theta^{(t)})} = \frac{\pi_k^{(t)}\mathcal{N}(y_i;f_k(\mathbf{x}_i,\theta^{(t)}),\rho_k^{(t)})}{\sum_{k=1}^K\pi_k^{(t)}\mathcal{N}(y_i;f_k(\mathbf{x}_i,\theta^{(t)}),\rho_k^{(t)})}.$$

For the M-step, to maximize $Q(\theta\|\theta^{(t)})$, we need to set the partial derivatives w.r.t $\theta = ((\pi_k,\mathbf{w}_k,\rho_k))_{k=1}^K$ to 0.

First optimizing $\mathbf{w}_k$:

$$\frac{\partial}{\partial\mathbf{w}_k}Q(\theta\|\theta^{(t)}) = \sum_{i=1}^N \delta_{ik}^{(t)}\left[\frac{1}{2\rho_k}\left(y_i - \mathbf{w}_k^T\mathbf{x}_i\right)\mathbf{x}_i\right],$$

Since $\rho_k > 0$, as w_k increasing, the first derivative decreases, the maximum happens when we set first derivative to 0. Hence,

$$\sum_{i=1}^{N} \delta_{ik}^{(t)} \left[\frac{1}{2\rho_k} \left(y_i - \mathbf{w}_k^T \mathbf{x}_i \right) \mathbf{x}_i \right] = 0.$$

Solving the equation:

$$\begin{aligned}
\sum_{i=1}^{N} \delta_{ik}^{(t)} y_i \mathbf{x}_i &= \sum_{i=1}^{N} \delta_{ik}^{(t)} \left(\mathbf{w}_k^T \mathbf{x}_i \right) \mathbf{x}_i \\
\sum_{i=1}^{N} \delta_{ik}^{(t)} y_i \mathbf{x}_i &= \left(\sum_{i=1}^{N} \delta_{ik}^{(t)} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{w}_k \\
\mathbf{w}_k^{(t+1)} &= \left(\sum_{i=1}^{N} \delta_{ik}^{(t)} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{N} \delta_{ik}^{(t)} y_i \mathbf{x}_i.
\end{aligned}$$

Next, optimizing ρ_k leads to :

$$\frac{\partial}{\partial \rho_k} Q(\theta \| \theta^{(t)}) = \sum_{i=1}^{N} \delta_{ik}^{(t)} \left[-\frac{1}{2\rho_k} + \frac{1}{2\rho_k^2} \left(y_i - \mathbf{w}_k^T \mathbf{x}_i \right)^2 \right],$$

As ρ_k increasing, the first derivative decreases, the maximum also happens when we set first derivative to 0. Hence,

$$\sum_{i=1}^{N} \delta_{ik}^{(t)} \left[-\frac{1}{2\rho_k} + \frac{1}{2\rho_k^2} \left(y_i - \mathbf{w}_k^T \mathbf{x}_i \right)^2 \right] = 0$$

$$\frac{1}{\rho_k} \sum_{i=1}^{N} \delta_{ik}^{(t)} \left(y_i - \mathbf{w}_k^T \mathbf{x}_i \right)^2 = \sum_{i=1}^{N} \delta_{ik}^{(t)}$$

$$\rho_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} \left(y_i - \mathbf{w}_k^{(t+1)^T} \mathbf{x}_i \right)^2}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}$$

Lastly, for π_k, we need to optimize under constraint $\sum_{k=1}^{K} \pi_k = 1$. Consider a Lagrange multiplier:

$$\frac{\partial}{\partial \rho_k} \left(Q(\theta \| \theta^{(t)}) - \lambda \sum_{k=1}^{K} \pi_k \right) = 0, \quad \forall k = 1, ...K.$$

$$\sum_{i=i}^{N} \delta_{ik}^{(t)} \frac{1}{\pi_k} - \lambda = 0$$

$$\pi_k = \frac{1}{\lambda} \sum_{i=i}^{N} \delta_{ik}^{(t)}.$$

Since the constraint $\sum_{k=1}^{K} \pi_k = \sum_{k=1}^{K} \frac{1}{\lambda} \sum_{i=i}^{N} \delta_{ik}^{(t)} = 1$, and $\sum_{k=i}^{K} \delta_{ik}^{(t)} = 1$, we have $\lambda = N$. Hence,

$$\pi_k{}^{(t+1)} = \frac{1}{N} \sum_{i=i}^{N} \delta_{ik}^{(t)}.$$

The M step is finished.