ML Written Homework 3

Student: R14921A13鄭皓中

October 31, 2025

1 Laplacian Eigenmaps

(a) The edge shown in the graph are: (1,2) (1,3) (1,4) (2,4) (2,8) (3,8) (5,6) (5,7) (7,10) (8,9) (9,10)

$$W = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

(b)

$$D = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

$$L = \begin{pmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\ -1 & 0 & 2 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ -1 & -1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 2 & 0 & 0 & -1 \\ 0 & -1 & -1 & 0 & 0 & 0 & 0 & 3 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 2 \end{pmatrix}$$

(c)

```python
import numpy as np
from scipy import linalg
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

W = np.array([
    [0, 1, 1, 1, 0, 0, 0, 0, 0, 0],  # x1
    [1, 0, 0, 1, 0, 0, 0, 1, 0, 0],  # x2
    [1, 0, 0, 0, 0, 0, 0, 1, 0, 0],  # x3
    [1, 1, 0, 0, 0, 0, 0, 0, 0, 0],  # x4
    [0, 0, 0, 0, 0, 1, 1, 0, 0, 0],  # x5
    [0, 0, 0, 0, 1, 0, 0, 0, 0, 0],  # x6
    [0, 0, 0, 0, 1, 0, 0, 0, 0, 1],  # x7
    [0, 1, 1, 0, 0, 0, 0, 0, 1, 0],  # x8
    [0, 0, 0, 0, 0, 0, 0, 1, 0, 1],  # x9
    [0, 0, 0, 0, 0, 0, 1, 0, 1, 0]   # x10
])

degrees = W.sum(axis=1)
D = np.diag(degrees)

L = D - W

try:
    eigenvalues, eigenvectors = linalg.eigh(L, D)
except linalg.LinAlgError:
    print("Try use pinv")
    safe_degrees = np.where(degrees == 0, 1e-6, degrees)
    D_inv_sqrt = np.diag(1.0 / np.sqrt(safe_degrees))
    L_sym = D_inv_sqrt @ L @ D_inv_sqrt
    eigenvalues, eigenvectors_sym = linalg.eigh(L_sym)
    eigenvectors = D_inv_sqrt @ eigenvectors_sym

Psi = eigenvectors[:, 1:4]

Z = Psi

print("\n--- eigenvalues ---")
print(eigenvalues)
print("\n--- Psi (Z = Psi) ---")
print(Z)

fig = plt.figure(figsize=(10, 8))
ax = fig.add_subplot(111, projection='3d')

ax.scatter(Z[:, 0], Z[:, 1], Z[:, 2],
    s=100, c=degrees, cmap='viridis', alpha=0.8)
```

```python
labels = [f'$z_{i+1}$' for i in range(Z.shape[0])]
for i, label in enumerate(labels):
    ax.text(Z[i, 0], Z[i, 1], Z[i, 2], label, size=12,
            zorder=1, color='k', ha='center', va='center')

ax.set_xlabel('Eigenvector 2', fontsize=12)
ax.set_ylabel('Eigenvector 3', fontsize=12)
ax.set_zlabel('(Eigenvector 4', fontsize=12)
ax.set_title('Laplacian Eigenmaps', fontsize=16)

plt.grid(True)
plt.show()
```

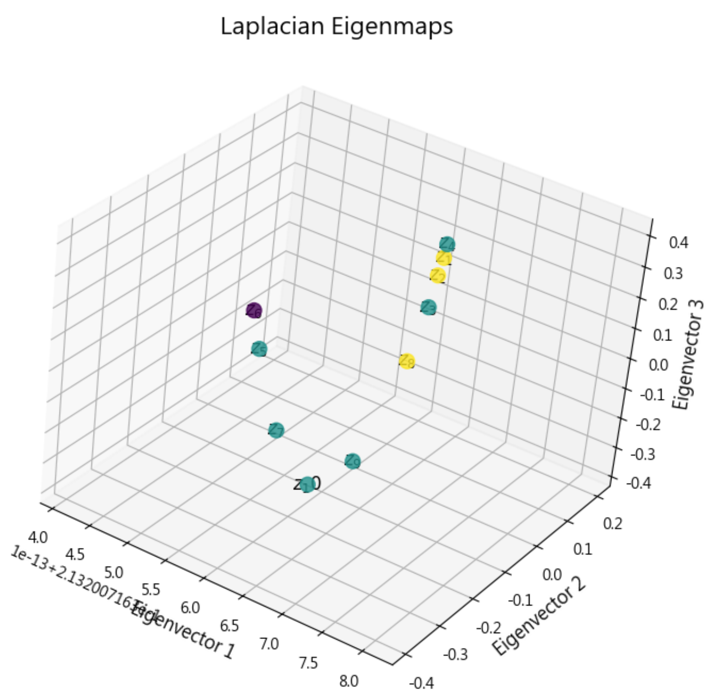

Figure 1: 3D image eigenvector 1-3

(d) This is the image of eigenvector 2-4 and the verify part

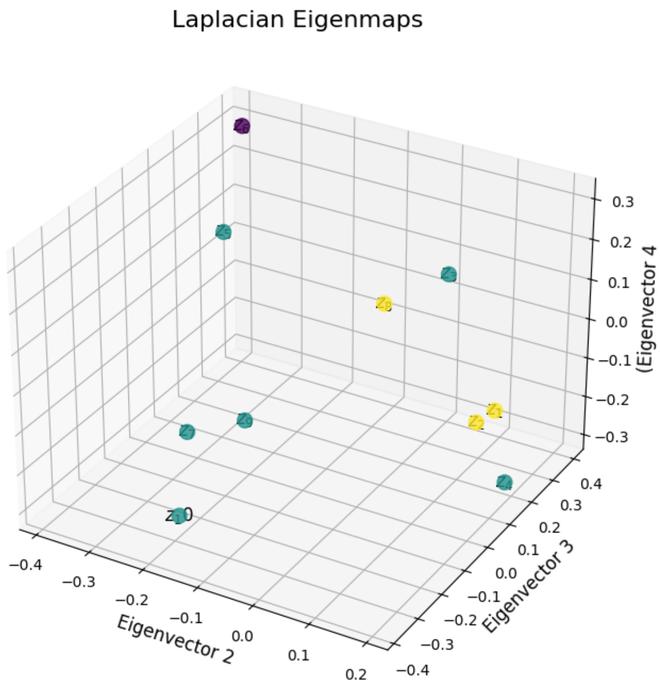

Figure 2: 3D image eigenvector 2-4

```
--- 驗證 tr(Ψ^T L Ψ) ---
tr(Ψ^T L Ψ) 的計算結果: 1.097803075120642
特徵值之和 (λ_1+λ_2+λ_3): 1.0978030751206402
此值是否約等於 1.098 (1.09779...): False

--- 驗證 Ψ^T D Ψ = I_3 ---
矩陣 Ψ^T D Ψ (應為 3x3 單位矩陣):
[[ 1.00000000e+00 -5.55111512e-17  4.16333634e-17]
 [-4.16333634e-17  1.00000000e+00 -1.28369537e-16]
 [ 4.16333634e-17 -7.28583860e-17  1.00000000e+00]]
```

Figure 3: Verify

(e) Suppose $\mathbf{c} = [c, c, c, ..., c]^T$, we want to show $L\mathbf{c} = 0$ for all L. $L\mathbf{c} = (D - W)\mathbf{c} = D\mathbf{c} - W\mathbf{c}$

$$(W\mathbf{c})_i = \sum_{j=1}^{n} W_{ij} c_j = c \sum_{j=1}^{n} W_{ij} = c \cdot d_i$$

$$(D\mathbf{c})_i = \sum_{j=1}^{n} D_{ij} c_j = c \cdot d_i$$

Hence, $L\mathbf{c} = D\mathbf{c} - W\mathbf{c} = 0$

(f)

$$\begin{aligned}
\text{RHS} &= \frac{1}{2}\sum_{i,j} w_{ij}(f_i - f_j)^2 \\
&= \frac{1}{2}\sum_{i,j} w_{ij}(f_i^2 - 2f_if_j + f_j^2) \\
&= \frac{1}{2}\left(\sum_{i,j} w_{ij}f_i^2 - \sum_{i,j} w_{ij}(2f_if_j) + \sum_{i,j} w_{ij}f_j^2\right)
\end{aligned}$$

$$\sum_{i,j} w_{ij}f_i^2 = \sum_i f_i^2\left(\sum_j w_{ij}\right) = \sum_i f_i^2 d_i$$

$$\sum_{i,j} w_{ij}f_j^2 = \sum_{j,i} w_{ji}f_i^2 = \sum_i f_i^2\left(\sum_j w_{ij}\right) = \sum_i f_i^2 d_i$$

$$\begin{aligned}
\text{RHS} &= \frac{1}{2}\left(\left(\sum_i d_if_i^2\right) - 2\sum_{i,j} w_{ij}f_if_j + \left(\sum_i d_if_i^2\right)\right) \\
&= \frac{1}{2}\left(2\sum_i d_if_i^2 - 2\sum_{i,j} w_{ij}f_if_j\right) \\
&= \sum_i d_if_i^2 - \sum_{i,j} w_{ij}f_if_j
\end{aligned}$$

$$\begin{aligned}
\text{LHS} &= f^T(D - W)f \\
&= f^TDf - f^TWf \\
&= \sum_{i,j} f_iD_{ij}f_j - \sum_{i,j} f_iW_{ij}f_j \\
&= \sum_i d_if_i^2 - \sum_{i,j} w_{ij}f_if_j
\end{aligned}$$

Hence we have LHS=RHS. QED.

(g) By the question we have

$$L \cdot f = \lambda \cdot f = 0 \cdot f = 0$$

so we have

$$f^TLf = f^T \cdot 0 = 0$$

(h) If f is the eigenvector of the eigenvalue 0, by f and g we can find that

$$f^TLf = \frac{1}{2}\sum_{i,j} w_{ij}(f_i - f_j)^2 = 0$$

That is, for any $w_{ij} > 0$ (means there is an edge between x_i and x_j), we have $f_i = f_j$. Since the graph is connect, for any two vertex x_k, x_t, there must be a connected path $x_k \to x_{p_1} \to \cdots \to x_{p_m} \to x_t$, that gives us

$$f_k = f_{p_1} = \cdots = f_{p_m} = f_t$$

Because the vertexes are arbitrarily chosen, the only eigenvector satisfied the above condition is $f = [c, c, ..., c]^T$, which means the eigenvalue 0 is only 1-dimension, then the second smallest eigenvalue λ_1 must be greater than 0 ($\lambda_1 > 0$).

2 Principal Component Analysis

(a) We want to maximize $f(u) = u^T \Sigma u$ constraint $g(u) = \|u\|_2^2 = u^T u = 1$. Suppose λ is a Lagrange multiplier, we have the Lagrange function

$$\mathcal{L}(u, \lambda) = f(u) - \lambda(g(u) - 1) = u^T \Sigma u - \lambda(u^T u - 1)$$

Then we compute the gradient and set it to 0

$$\begin{aligned} \nabla_u \mathcal{L}(u, \lambda) &= \nabla_u (u^T \Sigma u - \lambda(u^T u - 1)) \\ &= 2\Sigma u - \lambda(2u) \\ &= 2(\Sigma u - \lambda u) = 0 \end{aligned}$$

This gives us $\Sigma u = \lambda u$, so for any maximizer or minimizer u, it must satisfy the equation $\Sigma u = \lambda u$.
Then bring the maximizer u back to $f(u) = u^T \Sigma u = \lambda u^T u = \lambda$, and we have $\lambda = \max_{\|u\|_2^2 = 1} u^T \Sigma u$.

(b) $\hat{x}_i = (u^T x_i) u$ is the orthogonal projection of x_i on u, then we have $\|x_i\|_2^2 = \|\hat{x}_i\|_2^2 + \|x_i - \hat{x}_i\|_2^2$. Then $\|x_i - \hat{x}_i\|_2^2 = \|x_i\|_2^2 - \|\hat{x}_i\|_2^2$.
Now compute $\|\hat{x}_i\|_2^2$

$$\begin{aligned} \|\hat{x}_i\|_2^2 &= \|(u^T x_i) u\|_2^2 \\ &= (u^T x_i)^2 \|u\|_2^2 \\ &= (u^T x_i)^2 (1) \\ &= (u^T x_i)^2 \end{aligned}$$

So we have $\|x_i - \hat{x}_i\|_2^2 = \|x_i\|_2^2 - (u^T x_i)^2$

$$\text{LHS} = \left(\frac{1}{N} \sum_{i=1}^{N} \|x_i\|_2^2 \right) - \left(\frac{1}{N} \sum_{i=1}^{N} (u^T x_i)^2 \right)$$

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} \|x_i\|_2^2 &= \frac{1}{N} \sum_{i=1}^{N} \mathrm{tr}(x_i x_i^T) \\ &= \mathrm{tr} \left(\frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \right) \\ &= \mathrm{tr}(\Sigma) \end{aligned}$$

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} (u^T x_i)^2 &= \frac{1}{N} \sum_{i=1}^{N} u^T (x_i x_i^T) u \\ &= u^T \left(\frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \right) u \\ &= u^T \Sigma u \end{aligned}$$

Then we get

$$\frac{1}{N}\sum_{i=1}^{N}\|x_i-\hat{x}_i\|_2^2=\mathrm{tr}(\Sigma)-u^T\Sigma u$$

(c)

$$\Sigma=\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}x_{i,1}^2 & x_{i,1}x_{i,2}\\ x_{i,1}x_{i,2} & x_{i,2}^2\end{pmatrix}=\frac{1}{N}\begin{pmatrix}\sum x_{i,1}^2 & \sum x_{i,1}x_{i,2}\\ \sum x_{i,1}x_{i,2} & \sum x_{i,2}^2\end{pmatrix}$$

Define $S=\sum_{i=1}^{N}x_ix_i^T$

$$S=\begin{pmatrix}363 & -60\\ -60 & 482\end{pmatrix}$$

then $\Sigma=\frac{1}{N}S$
Solve the eigenvalue of S

$$\det(S-\lambda_S I)=0$$

$$\det\begin{pmatrix}363-\lambda_S & -60\\ -60 & 482-\lambda_S\end{pmatrix}=0$$

$$\lambda_S^2-845\lambda_S+171366=0$$

$$\lambda_S=\frac{845\pm 169}{2}=507\,\mathrm{or}\,338$$

Choose the maxima eigenvalue $\lambda_{S,1}=507$ and compute its eigenvector

$$\begin{pmatrix}363-507 & -60\\ -60 & 482-507\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}$$

$$\begin{pmatrix}-144 & -60\\ -60 & -25\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}$$

$$-144a-60b=0\implies -12(12a+5b)=0\implies 12a=-5b$$

We have the eigenvector $v_1=\begin{pmatrix}5\\ -12\end{pmatrix}$ Then we can get the unit eigenvector u_1

$$u_1=\frac{v_1}{\|v_1\|_2}=\frac{1}{13}\begin{pmatrix}5\\ -12\end{pmatrix}=\begin{pmatrix}5/13\\ -12/13\end{pmatrix}$$

So the eigenvalues and the unit eigenvector of the max eigenvalue are

$$\lambda_1=\frac{507}{N},\lambda_2=\frac{338}{N}$$

$$u_1=\begin{pmatrix}5/13\\ -12/13\end{pmatrix}$$

Now compute the total reconstruction error, by (b)

$$\frac{1}{N}\sum_{i=1}^{N}\|x_i-\hat{x}_i\|_2^2=\mathrm{tr}(\Sigma)-u^T\Sigma u\implies\sum_{i=1}^{N}\|x_i-\hat{x}_i\|_2^2=N(\mathrm{tr}(\Sigma)-u^T\Sigma u)$$

$$
\begin{aligned}
E &= N\left(\operatorname{tr}\left(\frac{1}{N}S\right) - u^T\left(\frac{1}{N}S\right)u\right)\\
&= N\left(\frac{1}{N}\operatorname{tr}(S) - \frac{1}{N}u^T S u\right)\\
&= \operatorname{tr}(S) - u^T S u
\end{aligned}
$$

Apply $S = \begin{pmatrix} 363 & -60 \\ -60 & 482 \end{pmatrix}$ and $u = \begin{pmatrix} 5/13 \\ -12/13 \end{pmatrix}$ we can calculate $\operatorname{tr}(S) = 363 + 482 = 845$
By (a)

$$
u^T S u = u^T \lambda_{S,1} u = 507 \cdot u^T u = 507
$$

So $E = \operatorname{tr}(S) - \lambda_{S,1} = \lambda_{S,2} = 845 - 507 = 338$

3 Gradient of the t-SNE Objective

1.

$$
\begin{aligned}
C &= \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}\\
&= \sum_{i \neq j} p_{ij}(\log p_{ij} - \log q_{ij})\\
&= \sum_{i \neq j} p_{ij} \log p_{ij} - \sum_{i \neq j} p_{ij} \log q_{ij}
\end{aligned}
$$

$\sum_{i \neq j} p_{ij} \log p_{ij}$: Since p_{ij} is computed by $\{\mathbf{x}_i\}$, it cannot be influenced by $\mathbf{y}_i$, so for $\mathbf{y}_i$, $\sum_{i \neq j} p_{ij} \log p_{ij}$ is a constant (constant(to $\mathbf{y}_i$). Hence, $C = \text{constant(to}\, \mathbf{y}_i) - \sum_{i \neq j} p_{ij} \log q_{ij}$

2. We want to show $\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$

$$
\begin{aligned}
\frac{\partial C}{\partial y_i} &= \frac{\partial}{\partial y_i}\left(\text{constant} - \sum_{k \neq \ell} p_{k\ell} \log q_{k\ell}\right)\\
&= -\frac{\partial}{\partial y_i}\left(\sum_{k \neq \ell} p_{k\ell} \log q_{k\ell}\right)
\end{aligned}
$$

$$
w_{ij} = (1 + \|y_i - y_j\|^2)^{-1}
$$

$$
Z = \sum_{k \neq \ell} w_{k\ell}
$$

$$
q_{ij} = \frac{w_{ij}}{Z}
$$

Use the above equation we can rewrite the formula

$$\begin{aligned}
\sum_{k\neq\ell} p_{k\ell}\log q_{k\ell} &= \sum_{k\neq\ell} p_{k\ell}\log\left(\frac{w_{k\ell}}{Z}\right) \\
&= \sum_{k\neq\ell} p_{k\ell}(\log w_{k\ell} - \log Z) \\
&= \sum_{k\neq\ell} p_{k\ell}\log w_{k\ell} - \left(\sum_{k\neq\ell} p_{k\ell}\right)\log Z \\
C &= \text{const} - \left(\sum_{k\neq\ell} p_{k\ell}\log w_{k\ell} - \log Z\right)
\end{aligned}$$

Return to the derivative of C

$$\frac{\partial C}{\partial y_i} = -\frac{\partial}{\partial y_i}\left(\sum_{k\neq\ell} p_{k\ell}\log w_{k\ell}\right) + \frac{\partial}{\partial y_i}(\log Z)$$

The first term $\frac{\partial}{\partial y_i}\left(-\sum_{k\neq\ell} p_{k\ell}\log w_{k\ell}\right)$:

$$\begin{aligned}
\frac{\partial \log w_{k\ell}}{\partial y_i} &= \frac{1}{w_{k\ell}}\frac{\partial w_{k\ell}}{\partial y_i} \\
\frac{\partial w_{k\ell}}{\partial y_i} &= \frac{\partial}{\partial y_i}(1+\|y_k-y_\ell\|^2)^{-1} \\
&= -1\cdot(w_{k\ell})^2\cdot\frac{\partial\|y_k-y_\ell\|^2}{\partial y_i}
\end{aligned}$$

If $k=i$: $\frac{\partial\|y_i-y_\ell\|^2}{\partial y_i} = 2(y_i-y_\ell)$
If $\ell=i$: $\frac{\partial\|y_k-y_i\|^2}{\partial y_i} = -2(y_k-y_i) = 2(y_i-y_k)$
Else: $\frac{\partial\|y_k-y_\ell\|^2}{\partial y_i} = 0$
Hence, the first term can be represented as

$$\begin{aligned}
\frac{\partial}{\partial y_i}\left(-\sum_{k\neq\ell} p_{k\ell}\log w_{k\ell}\right) &= \sum_{j\neq i} p_{ij}\frac{1}{w_{ij}}\underbrace{\left(-w_{ij}^2\cdot 2(y_i-y_j)\right)}_{\frac{\partial w_{ij}}{\partial y_i}} - \sum_{j\neq i} p_{ji}\frac{1}{w_{ji}}\underbrace{\left(-w_{ji}^2\cdot 2(y_i-y_j)\right)}_{\frac{\partial w_{ji}}{\partial y_i}} \\
&= \sum_{j\neq i} p_{ij}w_{ij}\cdot 2(y_i-y_j) + \sum_{j\neq i} p_{ji}w_{ji}\cdot 2(y_i-y_j) \\
&= 4\sum_{j\neq i} p_{ij}w_{ij}(y_i-y_j)
\end{aligned}$$

Now deal with the second term $\frac{\partial}{\partial y_i}(\log Z)$:

$$\begin{aligned}
\frac{\partial \log Z}{\partial y_i} &= \frac{1}{Z}\frac{\partial Z}{\partial y_i} \\
\frac{\partial Z}{\partial y_i} &= \frac{\partial}{\partial y_i}\left(\sum_{k\neq\ell} w_{k\ell}\right) \\
&= \sum_{k\neq\ell}\frac{\partial w_{k\ell}}{\partial y_i}
\end{aligned}$$

It is same as the first term, only when $k = i$ or $\ell = i$, the value is not 0.

$$\begin{aligned}\frac{\partial Z}{\partial y_i} &= \sum_{j\neq i} \frac{\partial w_{ij}}{\partial y_i} + \sum_{j\neq i} \frac{\partial w_{ji}}{\partial y_i} \\ &= \sum_{j\neq i} -w_{ij}^2 \cdot 2(y_i - y_j) + \sum_{j\neq i} -w_{ji}^2 \cdot 2(y_i - y_j) \\ &= -4\sum_{j\neq i} w_{ij}^2 (y_i - y_j)\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial y_i}(\log Z) &= \frac{1}{Z}\left(-4\sum_{j\neq i} w_{ij}^2(y_i - y_j)\right) \\ &= -4\sum_{j\neq i} \frac{w_{ij}}{Z} w_{ij}(y_i - y_j) \\ &= -4\sum_{j\neq i} q_{ij} w_{ij}(y_i - y_j)\end{aligned}$$

Combined those two terms:

$$\begin{aligned}\frac{\partial C}{\partial y_i} &= 4\sum_{j\neq i} p_{ij} w_{ij}(y_i - y_j) - 4\sum_{j\neq i} q_{ij} w_{ij}(y_i - y_j) \\ &= 4\sum_{j\neq i} (p_{ij} - q_{ij}) w_{ij}(y_i - y_j) \\ &= 4\sum_{j\neq i} (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}\end{aligned}$$

4 EM for Mixture of Multivariate t-Distributions

1. Observe the log-likelihood function $L(\theta) = \log p(Y|\theta)$, let $Z = \{z_i\}_{i=1}^N$ be the one-hot vector to show which component y_i belongs to. Then we have the relation $L(\theta) = \log p(Y|\theta) = \log\left(\frac{p(Y,Z|\theta)}{p(Z|Y,\theta)}\right)$.

$$L(\theta) = E_{Z|Y,\theta^{(t)}}[\log p(Y|\theta)]$$

$$L(\theta) = E_{Z|Y,\theta^{(t)}}\left[\log\left(\frac{p(Y,Z|\theta)}{p(Z|Y,\theta)}\right)\right]$$

$$L(\theta) = E_{Z|Y,\theta^{(t)}}[\log p(Y,Z|\theta)] - E_{Z|Y,\theta^{(t)}}[\log p(Z|Y,\theta)]$$

According to the question, we separate L to two parts Q, H

$$Q(\theta|\theta^{(t)}) = E_{Z|Y,\theta^{(t)}}[\log p(Y,Z|\theta)]$$

$$H(\theta|\theta^{(t)}) = -E_{Z|Y,\theta^{(t)}}[\log p(Z|Y,\theta)]$$

We deal with Q first, we have

$$p(Y,Z|\theta) = \prod_{i=1}^N p(y_i, z_i|\theta) = \prod_{i=1}^N p(z_i|\theta)p(y_i|z_i,\theta)$$

Use z_{ik} (if y_i belongs to $k-th$ component then $z_{ik}=1$ else 0):

$$p(y_i, z_i|\theta) = \prod_{k=1}^{K}[\pi_k \cdot tp(y_i; \mu_k, \Sigma_k, \nu_k)]^{z_{ik}}$$

So we can rewrite $p(Y,Z|\theta)$:

$$p(Y, Z|\theta) = \prod_{i=1}^{N}\prod_{k=1}^{K}[\pi_k \cdot tp(y_i; \mu_k, \Sigma_k, \nu_k)]^{z_{ik}}$$

$$\log p(Y, Z|\theta) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}(\log \pi_k + \log tp(y_i; \mu_k, \Sigma_k, \nu_k))$$

Define $\delta_{i,k}^{(t)} := E[z_{ik}|y_i, \theta^{(t)}] = P(z_i = k|y_i, \theta^{(t)})$ to be the posterior expected value of z_{ik}, then by Bayes' theorem

$$\begin{aligned}
\delta_{i,k}^{(t)} &= \frac{P(z_i = k|\theta^{(t)})p(y_i|z_i = k, \theta^{(t)})}{\sum_{\ell=1}^{K} P(z_i = \ell|\theta^{(t)})p(y_i|z_i = \ell, \theta^{(t)})} \\
&= \frac{\pi_k^{(t)} tp(y_i; \mu_k^{(t)}, \Sigma_k^{(t)}, \nu_k^{(t)})}{\sum_{\ell=1}^{K} \pi_\ell^{(t)} tp(y_i; \mu_\ell^{(t)}, \Sigma_\ell^{(t)}, \nu_\ell^{(t)})}
\end{aligned}$$

$$\begin{aligned}
Q(\theta|\theta^{(t)}) &= E\left[\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}(\log \pi_k + \log tp(y_i; \mu_k, \Sigma_k, \nu_k))\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} E[z_{ik}|y_i, \theta^{(t)}](\log \pi_k + \log tp(y_i; \mu_k, \Sigma_k, \nu_k)) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \delta_{i,k}^{(t)}(\log \pi_k + \log tp(y_i; \mu_k, \Sigma_k, \nu_k))
\end{aligned}$$

Now deal with H

$$p(Z|Y, \theta) = \prod_{i=1}^{N} p(z_i|y_i, \theta) = \prod_{i=1}^{N}\prod_{k=1}^{K}[P(z_i = k|y_i, \theta)]^{z_{ik}}$$

$$\log p(Z|Y, \theta) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \log P(z_i = k|y_i, \theta)$$

where $E[z_{ik}|y_i, \theta^{(t)}] = P(z_i = k|y_i, \theta^{(t)}) = \delta_{i,k}^{(t)}$

$$\begin{aligned}
E[\log p(Z|Y, \theta)] &= E\left[\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \log P(z_i = k|y_i, \theta)\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} E[z_{ik}|y_i, \theta^{(t)}] \log P(z_i = k|y_i, \theta) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \delta_{i,k}^{(t)} \log P(z_i = k|y_i, \theta)
\end{aligned}$$

Hence, we get H

$$H(\theta|\theta^{(t)}) = -\sum_{i=1}^{N}\sum_{k=1}^{K}\delta_{i,k}^{(t)}\log\left(\frac{\pi_k tp(y_i;\mu_k,\Sigma_k,\nu_k)}{\sum_{\ell=1}^{K}\pi_\ell tp(y_i;\mu_\ell,\Sigma_\ell,\nu_\ell)}\right)$$

2. In this question, we are asked to find the posterior probability $p(u_{i,k}|y_i, z_i = k)$.
From Bayes' theorem

$$p(u_{i,k}|y_i, z_i = k) \propto p(y_i|u_{i,k}, z_i = k)\cdot p(u_{i,k}|z_i = k)$$

(a) Likelihood: $p(y|u, z = k)$ For given $(u,\ y \sim N_p(\mu, \Sigma/u))$, its PDF can be computed by

$$p(y|u) = \frac{1}{(2\pi)^{p/2}|\Sigma/u|^{1/2}}\exp\left(-\frac{1}{2}(y-\mu)^T(\Sigma/u)^{-1}(y-\mu)\right)$$

$$p(y|u) = \frac{|\Sigma|^{-1/2}u^{p/2}}{(2\pi)^{p/2}}\exp\left(-\frac{u}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right)$$

Define Mahalanobis distance $d = (y-\mu)^T\Sigma^{-1}(y-\mu)$

$$p(y|u) \propto u^{p/2}\exp\left(-\frac{d}{2}u\right)$$

(b) Prior probability: $p(u|z = k)$ Given $z = k$, $u \sim \text{Gamma}(\nu/2, \nu/2)$, the PDF is:

$$p(u) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}u^{\nu/2-1}\exp\left(-\frac{\nu}{2}u\right)$$

$$p(u) \propto u^{\nu/2-1}\exp\left(-\frac{\nu}{2}u\right)$$

(c) Posterior probability: $p(u|y, z = k)$

$$p(u|y) \propto \left[u^{p/2}\exp\left(-\frac{d}{2}u\right)\right]\cdot\left[u^{\nu/2-1}\exp\left(-\frac{\nu}{2}u\right)\right]$$

$$p(u|y) \propto u^{p/2+\nu/2-1}\cdot\exp\left(-\frac{d}{2}u - \frac{\nu}{2}u\right)$$

$$p(u|y) \propto u^{\frac{\nu+p}{2}-1}\cdot\exp\left(-\left[\frac{\nu+d}{2}\right]u\right)$$

From the relation we can find that $u|y$ is a Gamma distribution with $\alpha_{\text{new}} = \frac{\nu+p}{2}$, $\beta_{\text{new}} = \frac{\nu+d}{2}$
Now put i, k, t back to the relation, we have

$$u_{i,k}|(y_i, z_i = k; \theta^{(t)}) \sim \text{Gamma}\left(\frac{\nu_k^{(t)}+p}{2}, \frac{\nu_k^{(t)}+d_{i,k}^{(t)}}{2}\right)$$

where $d_{i,k}^{(t)} = (y_i - \mu_k^{(t)})^T(\Sigma_k^{(t)})^{-1}(y_i - \mu_k^{(t)})$

3. We need to compute the moments of $X \sim \text{Gamma}(\alpha, \beta)$:$E[X]$ and $E[\log X]$. The standard expected value $E(X)$ of the Gamma distribution is $E[X] = \frac{\alpha}{\beta}$. And for the expected value $E[\log X]$ can be calculated by

$$E[\log X] = \psi(\alpha) - \log(\beta)$$

with $\psi(\alpha) = \frac{d}{d\alpha} \log \Gamma(\alpha)$, $\alpha = \frac{\nu_k^{(t)}+p}{2}$, $\beta = \frac{\nu_k^{(t)}+d_{i,k}^{(t)}}{2}$
So we can rewrite the moments $w_{i,k}^{(t)} := E[u_{i,k}|y_i, z_i = k; \theta^{(t)}]$ and $\ell_{i,k}^{(t)} := E[\log u_{i,k}|y_i, z_i = k; \theta^{(t)}]$

$$\begin{aligned}
w_{i,k}^{(t)} &= \frac{\alpha}{\beta} \\
&= \frac{(\nu_k^{(t)} + p)/2}{(\nu_k^{(t)} + d_{i,k}^{(t)})/2} \\
&= \frac{\nu_k^{(t)} + p}{\nu_k^{(t)} + d_{i,k}^{(t)}}
\end{aligned}$$

$$\begin{aligned}
\ell_{i,k}^{(t)} &= \psi(\alpha) - \log(\beta) \\
&= \psi\left(\frac{\nu_k^{(t)} + p}{2}\right) - \log\left(\frac{\nu_k^{(t)} + d_{i,k}^{(t)}}{2}\right)
\end{aligned}$$

4. We use lagrange multiplier λ to find $\pi_k^{(t+1)}$

$$\mathcal{L}(\pi, \lambda) = Q(\pi) + \lambda\left(1 - \sum_{k=1}^{K} \pi_k\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} \delta_{i,k}^{(t)} \log \pi_k + \lambda\left(1 - \sum_{k=1}^{K} \pi_k\right)$$

Let $N_k = \sum_{i=1}^{N} \delta_{i,k}^{(t)}$

$$\mathcal{L}(\pi, \lambda) = \sum_{k=1}^{K} N_k \log \pi_k + \lambda\left(1 - \sum_{k=1}^{K} \pi_k\right)$$

Set its derivative to 0

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \frac{N_k}{\pi_k} - \lambda = 0 \implies \pi_k = \frac{N_k}{\lambda}$$

$$\sum_{k=1}^{K} \frac{N_k}{\lambda} = 1 \implies \frac{1}{\lambda}\sum_{k=1}^{K} N_k = 1$$

$$\sum_{k=1}^{K} N_k = \sum_{k=1}^{K}\sum_{i=1}^{N} \delta_{i,k}^{(t)} = \sum_{i=1}^{N}\sum_{k=1}^{K} \delta_{i,k}^{(t)} = \sum_{i=1}^{N} 1 = N$$

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

Then we set the partial derivative of μ to 0 to find $\mu_k^{(t+1)}$

$$Q(\mu_k) = \sum_{i=1}^{N} \delta_{i,k}^{(t)} \left(-\frac{w_{i,k}^{(t)}}{2}(y_i - \mu_k)^T \Sigma_k^{-1}(y_i - \mu_k)\right)$$

$$\frac{\partial Q(\mu_k)}{\partial \mu_k} = \sum_{i=1}^{N} \delta_{i,k}^{(t)} \left(-\frac{w_{i,k}^{(t)}}{2} \cdot 2\Sigma_k^{-1}(\mu_k - y_i) \right) = 0$$

$$\sum_{i=1}^{N} \delta_{i,k}^{(t)} w_{i,k}^{(t)} \Sigma_k^{-1}(y_i - \mu_k) = 0$$

Since Σ^{-1} is invertible, we can remove it from the equation

$$\sum_{i=1}^{N} \delta_{i,k}^{(t)} w_{i,k}^{(t)} (y_i - \mu_k) = 0$$

$$\sum_{i=1}^{N} \delta_{i,k}^{(t)} w_{i,k}^{(t)} y_i - \sum_{i=1}^{N} \delta_{i,k}^{(t)} w_{i,k}^{(t)} \mu_k = 0$$

$$\left(\sum_{i=1}^{N} \delta_{i,k}^{(t)} w_{i,k}^{(t)} \right) \mu_k = \sum_{i=1}^{N} \delta_{i,k}^{(t)} w_{i,k}^{(t)} y_i$$

$$\boxed{\mu_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{i,k}^{(t)} w_{i,k}^{(t)} y_i}{\sum_{i=1}^{N} \delta_{i,k}^{(t)} w_{i,k}^{(t)}}}$$

Last, we use the new $\mu_k^{(t+1)}$ and $S_k = \Sigma_k^{-1}$ to find $\Sigma_k^{(t+1)}$

$$Q(S_k) = \sum_{i=1}^{N} \delta_{i,k}^{(t)} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{w_{i,k}^{(t)}}{2} (y_i - \mu_k^{(t+1)})^T S_k (y_i - \mu_k^{(t+1)}) \right)$$

Note that $\log |\Sigma_k| = -\log |\Sigma_k^{-1}| = -\log |S_k|$ and let $N_k = \sum_{i=1}^{N} \delta_{i,k}^{(t)}$ again, so we can rewrite $Q(S_k)$

$$Q(S_k) = \frac{N_k}{2} \log |S_k| - \sum_{i=1}^{N} \frac{\delta_{i,k}^{(t)} w_{i,k}^{(t)}}{2} (y_i - \mu_k^{(t+1)})^T S_k (y_i - \mu_k^{(t+1)})$$

For a scalar a, we have $\mathrm{Tr}(a) = a$, so we can write $\frac{\delta_{i,k}^{(t)} w_{i,k}^{(t)}}{2} (y_i - \mu_k^{(t+1)})^T S_k (y_i - \mu_k^{(t+1)}) =$ its trace, and for the trace, we have the property $\mathrm{Tr}(u^T S u) = \mathrm{Tr}(S u u^T)$.

$$Q(S_k) = \frac{N_k}{2} \log |S_k| - \frac{1}{2} \mathrm{Tr} \left(S_k \cdot \sum_{i=1}^{N} \delta_{i,k}^{(t)} w_{i,k}^{(t)} (y_i - \mu_k^{(t+1)})(y_i - \mu_k^{(t+1)})^T \right)$$

Let $A_k = \sum_{i=1}^{N} \delta_{i,k}^{(t)} w_{i,k}^{(t)} (y_i - \mu_k^{(t+1)})(y_i - \mu_k^{(t+1)})^T$

$$Q(S\_k) = \frac{N_k}{2} \log |S_k| - \frac{1}{2} \mathrm{Tr}(S_k A_k)$$

Set the derivative to 0

$$\frac{\partial Q(S_k)}{\partial S_k} = \frac{N_k}{2} (S_k^{-1})^T - \frac{1}{2} A_k^T = 0$$

$$\frac{N_k}{2} S_k^{-1} - \frac{1}{2} A_k = 0 \implies N_k S_k^{-1} = A_k \implies S_k^{-1} = \frac{A_k}{N_k} = \Sigma_k$$

$$\boxed{\Sigma_k^{(t+1)} = \frac{A_k}{N_k} = \frac{\sum_{i=1}^{N} \delta_{i,k}^{(t)} w_{i,k}^{(t)} (y_i - \mu_k^{(t+1)})(y_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{N} \delta_{i,k}^{(t)}}}$$

5. Write Q with the parameter ν_k

$$Q(\nu_k) = \sum_{i=1}^{N} \delta_{i,k}^{(t)} \left(\frac{\nu_k}{2} \log(\frac{\nu_k}{2}) - \log \Gamma(\frac{\nu_k}{2}) + (\frac{\nu_k}{2} - 1)\ell_{i,k}^{(t)} - \frac{\nu_k}{2} w_{i,k}^{(t)} + \frac{p}{2} \ell_{i,k}^{(t)} \right)$$

$$Q(\nu_k) = N_k \left(\frac{\nu_k}{2} \log(\frac{\nu_k}{2}) - \log \Gamma(\frac{\nu_k}{2}) \right) + \frac{\nu_k}{2} \sum_{i=1}^{N} \delta_{i,k}^{(t)} (\ell_{i,k}^{(t)} - w_{i,k}^{(t)}) + \text{const}$$

$$\frac{\partial}{\partial \nu_k} \left(\frac{\nu_k}{2} \log(\frac{\nu_k}{2}) \right) = \frac{\partial}{\partial(\nu_k/2)} \left(\frac{\nu_k}{2} \log(\frac{\nu_k}{2}) \right) \cdot \frac{\partial(\nu_k/2)}{\partial \nu_k} = \left[\log(\frac{\nu_k}{2}) + 1 \right] \cdot \frac{1}{2}$$

$$\frac{\partial}{\partial \nu_k} \left(\log \Gamma(\frac{\nu_k}{2}) \right) = \psi(\frac{\nu_k}{2}) \cdot \frac{1}{2}$$

Now compute the derivative and set it to 0

$$\frac{\partial Q(\nu_k)}{\partial \nu_k} = N_k \left[\frac{1}{2} \left(\log(\frac{\nu_k}{2}) + 1 \right) - \frac{1}{2} \psi(\frac{\nu_k}{2}) \right] + \frac{1}{2} \sum_{i=1}^{N} \delta_{i,k}^{(t)} (\ell_{i,k}^{(t)} - w_{i,k}^{(t)}) = 0$$

$$N_k \left(\log(\frac{\nu_k}{2}) + 1 - \psi(\frac{\nu_k}{2}) \right) + \sum_{i=1}^{N} \delta_{i,k}^{(t)} (\ell_{i,k}^{(t)} - w_{i,k}^{(t)}) = 0$$

$$\log(\frac{\nu_k}{2}) + 1 - \psi(\frac{\nu_k}{2}) + \frac{1}{N_k} \sum_{i=1}^{N} \delta_{i,k}^{(t)} (\ell_{i,k}^{(t)} - w_{i,k}^{(t)}) = 0$$

$$\boxed{\log(\frac{\nu_k}{2}) - \psi(\frac{\nu_k}{2}) + 1 + \frac{1}{\sum_{i=1}^{N} \delta_{i,k}^{(t)}} \sum_{i=1}^{N} \delta_{i,k}^{(t)} (\ell_{i,k}^{(t)} - w_{i,k}^{(t)}) = 0}$$

6. From the result we get from last question, let $f(\nu_k) = \log(\frac{\nu_k}{2}) - \psi(\frac{\nu_k}{2}) + 1 + C_k$ with $C_k = \frac{1}{\sum_{i=1}^{N} \delta_{i,k}^{(t)}} \sum_{i=1}^{N} \delta_{i,k}^{(t)} (\ell_{i,k}^{(t)} - w_{i,k}^{(t)})$.
Find the derivative of f

$$\begin{aligned} f'(\nu_k) &= \frac{d}{d\nu_k} \left(\log(\frac{\nu_k}{2}) - \psi(\frac{\nu_k}{2}) + 1 + C_k \right) \\ &= \frac{1}{\nu_k} - \frac{1}{2} \psi'(\frac{\nu_k}{2}) \end{aligned}$$

Then we find the one step update

$$\boxed{\nu_k^{\text{new}} \leftarrow \nu_k - \frac{\log(\frac{\nu_k}{2}) - \psi(\frac{\nu_k}{2}) + 1 + C_k}{\frac{1}{\nu_k} - \frac{1}{2}\psi'(\frac{\nu_k}{2})}}$$

5 Gradient Descent Convergence

1. Observe the log-likelihood function $L(\theta) = \log \prod_{i=1}^{N} p(x_i|y_i, \theta) = \sum_{i=1}^{N} \log p(x_i|y_i, \theta)$. Consider two set $L = \{i|y_i \neq 0\}$ and $U = \{i|y_i = 0\}$ to represent the labeled and unlabeled data. Then $L(\theta)$ can be rewritten as

$$
\begin{aligned}
L_{obs}(\theta) &= \sum_{i \in L} \log p(x_i|y_i, \theta) + \sum_{i \in U} \log p(x_i|y_i = 0, \theta) \\
&= \sum_{i: y_i \neq 0} \log \mathcal{N}(x_i|\mu_{y_i}, \Sigma_{y_i}) + \sum_{i: y_i = 0} \log \sum_{k=1}^{N} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)
\end{aligned}
$$

Let z_i be the one-hot label for unlabeled x_i, that is, $z_{i,k} = 1$ if x_i is from the component k.

$$
\begin{aligned}
L_c(\theta) &= \sum_{i \in L} \log p(x_i|y_i, \theta) + \sum_{i \in U} \log p(x_i, z_i|\theta) \\
&= \sum_{i \in L} \log p(x_i|y_i, \theta) + \sum_{i \in U} \log \prod_{k=1}^{K} [\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)]^{z_{i,k}} \\
&= \sum_{i \in L} \log p(x_i|y_i, \theta) + \sum_{i \in U} \sum_{k=1}^{K} z_{i,k} (\log \pi_k + \log \mathcal{N}(x_i|\mu_k, \Sigma_k))
\end{aligned}
$$

Compute Q

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) &= E_{Z|D_{obs}, \theta^{(t)}}[L_c(\theta)] \\
&= \sum_{i \in L} \log p(x_i|y_i, \theta) + \sum_{i \in U} \sum_{k=1}^{K} E[z_{i,k}|x_i, \theta^{(t)}] (\log \pi_k + \log \mathcal{N}(x_i|\mu_k, \Sigma_k))
\end{aligned}
$$

Define $\delta_{i,k}^{(t)} = E[z_{i,k}|x_i, \theta^{(t)}]$, then we have

$$
Q(\theta|\theta^{(t)}) = \sum_{i \in L} \log p(x_i|y_i, \theta) + \sum_{i \in U} \sum_{k=1}^{K} \delta_{i,k}^{(t)} (\log \pi_k + \log \mathcal{N}(x_i|\mu_k, \Sigma_k))
$$

First we use lagrange multiplier λ to find π_k

$$
Q_\pi(\pi) = \sum_{i \in U} \sum_{k=1}^{K} \delta_{i,k}^{(t)} \log \pi_k
$$

$$
\begin{aligned}
\mathcal{L}(\pi, \lambda) &= \sum_{k=1}^{K} \sum_{i \in U} \delta_{i,k}^{(t)} \log \pi_k + \lambda \left(\sum_{k=1}^{N} \pi_k - 1 \right) \\
\frac{\partial \mathcal{L}}{\partial \pi_k} &= \frac{\sum_{i \in U} \delta_{i,k}^{(t)}}{\pi_k} + \lambda = 0 \implies \pi_k = -\frac{\sum_{i \in U} \delta_{i,k}^{(t)}}{\lambda} \\
\sum_{k=1}^{N} \pi_k &= 1 \implies \sum_{k=1}^{N} -\frac{\sum_{i \in U} \delta_{i,k}^{(t)}}{\lambda} = 1
\end{aligned}
$$

Since $\sum_{k=1}^{K} \delta_{i,k}^{(t)} = 1$ we have

$$-\frac{\sum_{i\in U}\sum_{k=1}^{N} \delta_{i,k}^{(t)}}{\lambda} = 1 \implies \lambda = -\sum_{i\in U} 1 = -\sum_{i:y_i=0} 1$$

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

Then calculate μ

$$Q_k(\mu_k, \Sigma_k) = \sum_{i:y_i=k} \log \mathcal{N}(x_i|\mu_k, \Sigma_k) + \sum_{i:y_i=0}\sum_{k=1}^{K} \delta_{i,k}^{(t)} \log \mathcal{N}(x_i|\mu_k, \Sigma_k)$$

We only need the term with μ

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

$$\nabla_{\mu_k} L = \sum_{i:y_i=k} -2\Sigma_k^{-1}(x_i - \mu_k) + \sum_{i:y_i=0} -2\delta_{i,k}^{(t)}\Sigma_k^{-1}(x_i - \mu_k) = 0$$

Since Σ is invertible, we have

$$\sum_{i:y_i=k} (x_i - \mu_k) + \sum_{i:y_i=0} \delta_{i,k}^{(t)}(x_i - \mu_k) = 0$$

$$\sum_{i:y_i=k} x_i + \sum_{i:y_i=0} \delta_{i,k}^{(t)} x_i = \left(\sum_{i:y_i=k} 1 + \sum_{i:y_i=0} \delta_{i,k}^{(t)}\right) \mu_k$$

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

Last we can use the result from the last problem

$$\Sigma_k^{(t+1)} == \frac{\sum_{i=1}^{N} \text{weight}_i (y_i - \mu_k^{(t+1)})(y_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{N} \text{weight}_i}$$

In this equation, $\text{weight}_i = 1$ if $y_i = k$ else $\delta_{i,k}^{(t)}$, so we have

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

2. Compute $\delta_{i,k}^{(t)}$

$$\delta_{i,k}^{(t)} = E[z_{i,k}|x_i, y_i = 0, \theta^{(t)}] = P(z_{i,k} = 1|x_i, y_i = 0, \theta^{(t)})$$

$$P(z_{i,k} = 1|x_i, y_i = 0, \theta^{(t)}) = \frac{P(z_{i,k} = 1|y_i = 0, \theta^{(t)}) \cdot P(x_i|z_{i,k} = 1, y_i = 0, \theta^{(t)})}{P(x_i|y_i = 0, \theta^{(t)})}$$

We have

$$P(z_{i,k} = 1|y_i = 0, \theta^{(t)}) = \pi_k^t$$

$$P(x_i|z_{i,k} = 1, y_i = 0, \theta^{(t)}) = \mathcal{N}(x_i; \mu_k, \Sigma_k)$$

$$P(x_i|y_i = 0, \theta^{(t)}) = \sum_{j=1}^{K} P(z_{i,j} = 1|y_i = 0, \theta^{(t)}) \cdot P(x_i|z_{i,j} = 1, y_i = 0, \theta^{(t)}) = \sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}(x_i; \mu_j, \Sigma_j)$$

Hence, we get $\delta_{i,k}^{(t)}$

$$\delta_{i,k}^{(t)} = \frac{\pi_k^t \mathcal{N}(x_i; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j^t \mathcal{N}(x_i; \mu_j, \Sigma_j)}$$