Homework 5 Theory

Question 1

Assume we have a multivariate normal random variable $X = [X_1, X_2, X_3, X_4]^\top$, whose covariance matrix Σ and inverse covariance matrix Q are

$$\Sigma = \begin{bmatrix} 0.71 & -0.43 & 0.43 & 0 \\ -0.43 & 0.46 & -0.26 & 0 \\ 0.43 & -0.26 & 0.46 & 0 \\ 0 & 0 & 0 & 0.2 \end{bmatrix} \quad Q = \begin{bmatrix} 5 & 3 & -3 & 0 \\ 3 & 5 & 0 & 0 \\ -3 & 0 & 5 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

Note that Q is simply the inverse of Σ, i.e., $Q = \Sigma^{-1}$.

(a) Are X_3 and X_4 correlated?

No they are not correlated. Since Σ is symmetric, $\Sigma_{34} = \Sigma_{43}$. These both show that the covariance between X_3 and X_4 is 0.

Therefore $\Sigma_{34} = \Sigma_{43} = 0$ shows that X_3 and X_4 are uncorrelated.

(b) Are X_3 and X_4 conditionally correlated given the other variables? That is, does $\text{cov}(X_3, X_4 \mid X_1, X_2)$ equal to zero?

No.

The inverse covariance matrix Q show the conditional independence between variables. Where Q is 0, then the two X values from that row and column are conditionally independent given all other variables.

Similar to part a, looking at $Q_{34} = Q_{43} = 0$, we therefore know that X_3 and X_4 are conditionally independent and conditionally uncorrelated given X_1 and X_2.

(c) Please find the Markov blanket of X_2. Recall that the Markov blanket of X_i is the set of variables (denoted by X_{M_i}), such that

$$X_i \perp X_{\neg\{i\} \cup M_i} \mid X_{M_i}$$

where $\neg\{i\} \cup M_i$ denotes all the variables outside of $\{i\} \cup M_i$.

The precision matrix Q tell us the conditional independencies. If $Q_{ij} = 0$, then X_i and X_j are conditionally independent given all other variables.

Looking at X_2 we have $[3, 5, 0, 0]$. The second entry is just the variance of X_2 while the zeroes make up the conditional independence of X_3 and X_4. Therefore the Markov blanket of X_2 is simply X_1.

(d) [5 points] Assume that $Y = [Y_1, Y_2]^\top$ is defined by

$$Y_1 = X_1 + X_4$$
$$Y_2 = X_2 - X_4$$

Please calculate the covariance matrix of Y.

Given Y_1 and Y_2 we can define the matrix A as below knowing that $Y = AX$:

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \end{bmatrix}$$

A covariance matrix can be defined as:

$$\Sigma = A\Sigma A^\top$$

Therefore we can calculate the covariance matrix of Y using this property.

$$A\Sigma A^\top = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0.71 & -0.43 & 0.43 & 0 \\ -0.43 & 0.46 & -0.26 & 0 \\ 0.43 & -0.26 & 0.46 & 0 \\ 0 & 0 & 0 & 0.2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 0.91 & -0.63 \\ -0.63 & 0.66 \end{bmatrix} = \Sigma_Y$$

Question 2

Assume we have a dataset of two points $\left\{x^{(1)}, x^{(2)}\right\}$:

$$x^{(1)} = -1, \quad x^{(2)} = 1$$

Assume $x^{(i)}$ is drawn i.i.d. from a simple mixture distribution of two Gaussian components:

$$f\left(x \mid \mu_1, \mu_2\right) = \frac{1}{2}\phi\left(x \mid \mu_1, 1\right) + \frac{1}{2}\phi\left(x \mid \mu_2, 1\right)$$

where $\phi\left(\cdot \mid \mu_i, 1\right)$ denotes the probability density function of Gaussian distribution $\mathcal{N}\left(\mu_i, 1\right)$ with mean μ_i and unit variance. We want to estimate the unknown parameters μ_1 and μ_2.

(a) Assume we run EM starting from an initialization of $\mu_1 = -2$ and $\mu_2 = 2$. Please decide the value of μ_1 and μ_2 at the next iteration of EM algorithm. (You may find it handy to know that $1/(1 + \exp(-4)) \approx 0.98$).

We will start with the E-Step by calculating the posterior probabilities:

We define γ_{ik} as:

$$\gamma_{ik} = \frac{\frac{1}{2}\phi(x^{(i)} \mid \mu_k, 1)}{\sum_{j=1}^{2} \frac{1}{2}\phi(x^{(i)} \mid \mu_j, 1)},$$

where $\phi(x \mid \mu, 1)$ is simply the Gaussian probability density function:

$$\phi(x \mid \mu, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right).$$

For $x^{(1)} = -1$ we will use $\frac{1}{2}\phi(-1 \mid \mu_1 = +/-2, 1)$:

$$\gamma_{11} = \frac{\exp\left(-\frac{1}{2}\right)}{\exp\left(-\frac{1}{2}\right) + \exp\left(-\frac{9}{2}\right)} = \frac{1}{1 + e^{-4}} \approx 0.98)$$

$$\gamma_{12} = 1 - \gamma_{12} \approx 0.02$$

For $x^{(1)} = 1$ we will use $\frac{1}{2}\phi(1 \mid \mu_1 = +/-2, 1)$:

$$\gamma_{22} = \frac{\exp\left(-\frac{1}{2}\right)}{\exp\left(-\frac{9}{2}\right) + \exp\left(-\frac{1}{2}\right)} = \frac{1}{1 + e^{-4}} \approx 0.98$$

$$\gamma_{21} = 1 - \gamma_{22} \approx 0.02$$

Now we update the means using the posterior from above:

$$\mu_1 = \frac{\gamma_{11}x^{(1)} + \gamma_{21}x^{(2)}}{\gamma_{11} + \gamma_{21}} = \frac{0.98 \cdot (-1) + 0.02 \cdot 1}{0.98 + 0.02} = -0.96$$

$$\mu_2 = \frac{\gamma_{12}x^{(1)} + \gamma_{22}x^{(2)}}{\gamma_{12} + \gamma_{22}} = \frac{0.02 \cdot (-1) + 0.98 \cdot 1}{0.02 + 0.98} = 0.96$$

(b) Do you think EM (when initialized with $\mu_1 = -2$ and $\mu_2 = 2$) will eventually converge to $\mu_1 = -1$ and $\mu_2 = 1$ (i.e., coinciding with the two data points). Please justify your answer using either your theoretical understanding or the result of an empirical simulation.

As the data points are symmetric around 0, I would expect the μ_1 and μ_2 to converge to 0 and not around the data points. I would expect that maximizing the likelihood for the Gaussian mixture to converge to the point where the distance from the data is minimized which would be at 0 this can be demonstrated using a simulation in python:

```
import numpy as np

mu1 = -2
mu2 = 2
data = np.array([-1, 1])

# Conduct EM
for i in range(1000):
    # Calculate posterior probabilities
    gamma1 = np.exp(-0.5 * (data - mu1) ** 2) / (np.exp(-0.5 * (data - mu1) ** 2) + np.exp(-0.5
* (data - mu2) ** 2))
    gamma2 = 1 - gamma1

    # Update mu
    mu1_new = np.sum(gamma1 * data) / np.sum(gamma1)
    mu2_new = np.sum(gamma2 * data) / np.sum(gamma2)

    if i % 100 == 0:
      print(f"Iteration {i}: mu1 = {mu1_new:.4f}, mu2 = {mu2_new:.4f}")

    # Update means for next iteration
    mu1, mu2 = mu1_new, mu2_new
```

```
## Iteration 0: mu1 = -0.9640, mu2 = 0.9640
## Iteration 100: mu1 = -0.1211, mu2 = 0.1211
## Iteration 200: mu1 = -0.0861, mu2 = 0.0861
## Iteration 300: mu1 = -0.0704, mu2 = 0.0704
## Iteration 400: mu1 = -0.0611, mu2 = 0.0611
## Iteration 500: mu1 = -0.0546, mu2 = 0.0546
## Iteration 600: mu1 = -0.0499, mu2 = 0.0499
## Iteration 700: mu1 = -0.0462, mu2 = 0.0462
## Iteration 800: mu1 = -0.0432, mu2 = 0.0432
## Iteration 900: mu1 = -0.0408, mu2 = 0.0408
```

As you can see from the results which ran for 1000 iterations and showed the calculate μ_i at each 100th interval, the values are symmetrically converging to 0 from both of the starting values of -2 and 2 given only the two dataponts in the dataset.

(c) Please decide the fixed point of EM when we initialize it from $\mu_1 = \mu_2 = 2$.

```
import numpy as np
import matplotlib.pyplot as plt

# Initial values
mu1 = 2
mu2 = 2
data = np.array([-1, 1])

# Conduct EM
for i in range(100):
    # Calculate posterior probabilities
    gamma1 = np.exp(-0.5 * (data - mu1) ** 2) / (np.exp(-0.5 * (data - mu1) ** 2) + np.exp(-0.5
* (data - mu2) ** 2))
    gamma2 = 1 - gamma1

    # Update mu
    mu1_new = np.sum(gamma1 * data) / np.sum(gamma1)
    mu2_new = np.sum(gamma2 * data) / np.sum(gamma2)

    if i % 10 == 0:
      print(f"Iteration {i}: mu1 = {mu1_new:.4f}, mu2 = {mu2_new:.4f}")

    # Update means for next iteration
    mu1, mu2 = mu1_new, mu2_new
```

```
## Iteration 0: mu1 = 0.0000, mu2 = 0.0000
## Iteration 10: mu1 = 0.0000, mu2 = 0.0000
## Iteration 20: mu1 = 0.0000, mu2 = 0.0000
## Iteration 30: mu1 = 0.0000, mu2 = 0.0000
## Iteration 40: mu1 = 0.0000, mu2 = 0.0000
## Iteration 50: mu1 = 0.0000, mu2 = 0.0000
## Iteration 60: mu1 = 0.0000, mu2 = 0.0000
## Iteration 70: mu1 = 0.0000, mu2 = 0.0000
## Iteration 80: mu1 = 0.0000, mu2 = 0.0000
## Iteration 90: mu1 = 0.0000, mu2 = 0.0000
```

We can see that running the same simulation from the previous question, but using a different starting point for μ_1 and μ_2, where $\mu_1 = \mu_2 = 2$ results in the same convergence to 0

(d) Please decide the fixed point of $\mathbf{K}$-means when we initialize it from $\mu_1 = -2$ and $\mu_2 = 2$.

First we will start by assigning each point to the nearest centroid which are the given means in the problem statement. Then we will update the centroids given the assignments of each point and then continuosly conduct this iteration until convergence.

- For $x^{(1)} = -1$, it is obvious the closest point is $\mu_1 = -2$
- For $x^{(2)} = 1$, it is obvious the closest point is $\mu_2 = 2$

So each cluster now has one point each with -1 and 1 respectively. Now we need to update the centroids:

- For μ_1, the mean of the points is simply its only point as -1
- For μ_2, the mean of the points is simply its only point as 1

Now we would check for convergence by doing another iteration and compare results, but as we can now see that the only two centroids are centered on the one two points in the dataset, then we can see that the centroids have reached convergence at the fixed points of $\mu_1 = -1$ and $\mu_2 = 1$