CS 391L: Machine Learning **Spring 2024**

Homework 5 - Theory - Solutions

Lecture: Prof. Qiang Liu

1. (a) No, X_3 and X_4 are not correlated since $cov(X_3, X_4) = \Sigma_{3,4} = 0$ as given by Σ.

   (b) No, X_3, X_4 are not conditionally correlated given X_1, X_2, because $Q_{34} = 0$, and hence $cov(X_3, X_4 \mid X_1, X_2) = 0$.

   As an alternative way to see this, denote by $\tilde{\Sigma}$ and $\tilde{Q}$ the covariance and the inverse covariance matrix of the conditional distribution $X_3, X_4|X_1, X_2$. We know that Q simply equals the submatrix of Q on $[3:4]$, that is,

$$\tilde{\Sigma} = \tilde{Q}^{-1} = Q^{-1}_{3:4,3:4} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{5} \end{bmatrix}.$$

   and see that $\tilde{\Sigma}_{1,2} = cov(X_3, X_4 \mid X_1, X_2) = 0$, and hence X_3 and X_4 are conditionally independent conditioning on $[X_1, X_2]$.

   (c) The Markov blanket of X_2 is $\{X_1\}$. To show this, let $\bar{\Sigma}$ be the covariance matrix of $X_2, X_3, X_4|X_1$. We have that

$$\bar{\Sigma} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{5} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{5} \end{bmatrix}$$

   Thus, $\mathrm{cov}(X_2, X_3|X_1) = 0$ and $\mathrm{cov}(X_2, X_4|X_1) = 0$, and so, X_2 is independent of X_3 and X_4 conditioned on X_1.

   (d) By the definition, we have

$$Y = AX,$$

   where

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

   We have $Cov(Y) = A\Sigma A^\top$, which gives

$$Cov(Y) = A\Sigma A^\top = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0.71 & -0.43 & 0.43 & 0 \\ -0.43 & 0.46 & -0.26 & 0 \\ 0.43 & -0.26 & 0.46 & 0 \\ 0 & 0 & 0 & 0.2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.71 & -0.43 & 0.43 & 0.2 \\ -0.43 & 0.46 & -0.26 & -0.2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.91 & -0.63 \\ -0.63 & 0.66 \end{bmatrix}.$$

Alternatively, we can calculate the individual elements of $Cov(Y)$ one by one.

$$\begin{aligned}
var(Y_1) &= var(X_1 + X_4) = var(X_1) + 2cov(X_1, X_4) + var(X_4) \\
&= 0.71 + 2 * (0) + 0.2 = 0.91 \\
var(Y_2) &= var(X_2 - X_4) = var(X_2) - 2cov(X_2, X_4) + var(X_4) \\
&= 0.46 - 2 * (0) + 0.2 = 0.66 \\
cov(Y_1, Y_2) &= cov(X_1 + X_4, X_2 - X_4) = cov(X_1, X_2) - cov(X_1, X_4) + cov(X_4, X_2) - cov(X_4, X_4) \\
&= -0.43 - 0 + 0 - 0.2 = -0.63.
\end{aligned}$$

Therefore,

$$Cov(Y) = \begin{bmatrix} var(Y_1) & cov(Y_1, Y_2) \\ cov(Y_1, Y_2) & var(Y_2) \end{bmatrix} = \begin{bmatrix} 0.91 & -0.63 \\ -0.63 & 0.66 \end{bmatrix}.$$

2. (a) We have that

$$\gamma_1(x^{(1)}) = \gamma_1(-1) = \frac{\frac{1}{2}\phi(-1| - 2, 1)}{\frac{1}{2}\phi(-1| - 2, 1) + \frac{1}{2}\phi(-1|2, 1)} = \frac{\exp(-(-1+2)^2/2)}{\exp(-(-1+2)^2/2) + \exp(-(-1-2)^2/2)}$$

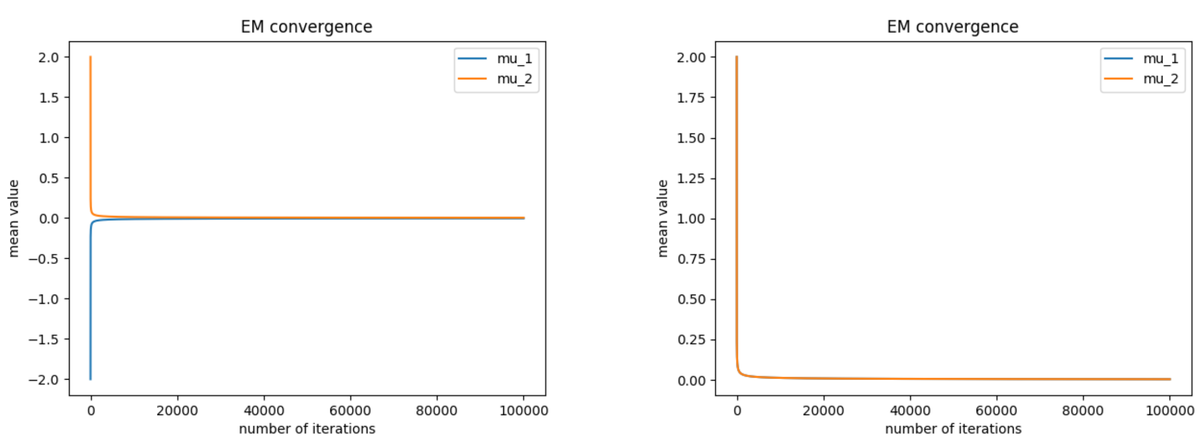
$$= \frac{1}{1 + \exp(-4)} \approx 0.98$$

Similarly, we have that $\gamma_2(-1) \approx 0.02, \gamma_1(1) \approx 0.02, \gamma_2(1) \approx 0.98$
Thus, in the next iteration,

$$\mu_1 = 0.98 * (-1) + 0.02 * 1 = -0.96$$

$$\mu_2 = 0.02 * (-1) + 0.98 * 1 = 0.96$$

(b) No, it does not converge to $\mu_1 = -1$ and $\mu_2 = 1$. Instead, both μ_1 and μ_2 converges to zero as we see from the plot in Figure 1(a).



(a) (b)

Figure 1: (a): EM trajectory when initiated from $\mu_1 = -2$ and $\mu_2 = 2$. (b): EM trajectory when initiated from $\mu_1 = \mu_2 = 2$.

Alternatively, we can show that $\mu_1 = -1$ and $\mu_2 = 1$ can not be a fixed point of EM theoretically and hence does not converge to it. To see it, assume we have reached $\mu_1 = -1$ and $\mu_2 = 1$ at some point. In this case, the EM update will be

$$\mu_1^{new} \leftarrow \gamma_1(x^{(1)})x^{(1)} + \gamma_1(x^{(2)})x^{(2)}$$
$$\mu_2^{new} \leftarrow \gamma_2(x^{(1)})x^{(1)} + \gamma_2(x^{(2)})x^{(2)},$$

where $\gamma_i(x^{(j)})$ are the posterior weights. Because all of the weights $\gamma_i(x^{(j)})$ are strictly larger than zero, the value of μ_1^{new} and μ_2^{new} will lie between $x^{(1)}$ $x^{(2)}$, that is,

$$x^{(1)} < \mu_1^{new} < x^{(2)}, \qquad\qquad x^{(1)} < \mu_2^{new} < x^{(2)}.$$

Therefore, μ_1^{new} and μ_2^{new} can not strictly equal to $x^{(1)}$ and $x^{(2)}$.

If you are still not convinced, we can calculate the new values μ_1^{new} and μ_2^{new} explicitly. Note that

$$\gamma_1(x^{(1)}) = \frac{\phi(x^{(1)} \mid \mu_1, 1)}{\phi(x^{(1)} \mid \mu_1, 1) + \phi(x^{(1)} \mid \mu_2, 1)} = \frac{\phi(-1 \mid -1, 1)}{\phi(-1 \mid -1, 1) + \phi(-1 \mid 1, 1)} = \frac{1}{1 + \exp(-2)} \approx 0.88.$$

and similarly $\gamma_1(x^{(2)}) = 0.12$, $\gamma_2(x^{(1)}) = 0.12$, $\gamma_2(x^{(2)}) = 0.88$. Therefore,

$$\mu_1^{new} \leftarrow \gamma_1(x^{(1)})x^{(1)} + \gamma_1(x^{(2)})x^{(2)} = 0.88 \times (-1) + 0.12 \times (1) = -0.76$$
$$\mu_2^{new} \leftarrow \gamma_2(x^{(1)})x^{(1)} + \gamma_2(x^{(2)})x^{(2)} = 0.12 \times (-1) + 0.88 \times (1) = 0.76.$$

This shows that $\mu_1^{new} \neq -1$ and $\mu_2^{new} \neq 1$, and hence ($\mu_1 = -1$, $\mu_2 = 1$) is not a fixed point.

(c) As we see from the plot in Figure 1(b), the fixed point is $\mu_1 = \mu_2 = 0$ in this case.

Alternatively, we also can infer the fixed point theoretically. Because we are initializing μ_1 and μ_2 to be the same value in this case, their updates will be identical all the time and hence μ_1 and μ_2 will keep equal all the time. In this case, we are effectively fitting the data with a simple Gaussian distribution, whose mean should be obviously zero. Therefore, $\mu_1 = \mu_2 = 0$ should be the fixed point in this case.

To further confirm the result, we can verify that $\mu_1 = \mu_2 = 0$ is indeed a fixed point of EM. Note that in this case, we have $\gamma_1(x^{(1)}) = \gamma_1(x^{(2)}) = \gamma_2(x^{(1)}) = \gamma_2(x^{(2)}) = 1/2$, and so,

$$\mu_1 = 0.5 * (-1) + 0.5 * 1 = 0$$
$$\mu_2 = 0.5 * (-1) + 0.5 * 1 = 0$$

Also, our mixing weights remain unchanged, and so, this is a fixed point.

(d) The fixed point is $\mu_1 = -1$, $\mu_2 = 1$. This is because during the first iteration, $x^{(1)}$ will be assigned to the cluster with center $\mu_1 = -2$ and $x^{(2)}$ will be assigned to the cluster with center $\mu_2 = 2$. Since each cluster has only one point assigned to it, our updated μ_1 is -1, and our updated μ_2 is 1. After this, point $x^{(1)}$ will be assigned again to μ_1, and $x^{(2)}$ will be assigned to μ_2, and so again μ_1 and μ_2 will not change.