# Homework 4 STOR 767: Theory Part

# Due November 12, 2020

Student Name: Leo Li (PID: 730031954)

**Instructions**

- Edit this LaTeX file with your solutions and generate a PDF file from it. Upload **both the tex and the pdf** file to Sakai.
- Use proper fonts for a clear presentation:
  $x$ for an observed value; $X$ for a random variable; $\boldsymbol{X}$ for a vector; $\mathbf{X}$ for a matrix.
- You are allowed to work with other students but homework should be in your own words. Identical solutions will receive a **0** in grade and will be investigated.

**1.** Consider the Gaussian mixture model for clustering with latent $Z \sim Multinomial(1, (\eta_1, \ldots, \eta_K))$ and $\boldsymbol{X}|Z = j \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \forall j = 1, \ldots, K$. Develop the E-step and M-step for the parameters $\eta_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$'s.

Solution.

In this question, we would like to derive the EM algorithm to obtain the parameters in the Gaussian mixture model for clustering. We start by letting $Y = (\boldsymbol{X}_1, Z_1), \ldots, (\boldsymbol{X}_n, Z_n)$ be the full data, and $Y_{obs} = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be the observed data. The full data likelihood function is derived as,

$$L_n = \prod_{i=1}^{n} \prod_{j=1}^{K} \eta_j^{I(Z_i=j)} P_{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j}(\boldsymbol{X}_i)^{I(Z_i=j)}.$$

Then the log-likelihood function is derived as,

$$l_n = \sum_{i=1}^{n} \sum_{j=1}^{K} I(Z_i = j)(\log(\eta_j) + \log(P_{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j}(\boldsymbol{X}_i))).$$

We also initialize $\psi^{(0)} = (\eta_1^{(0)}, \ldots, \eta_K^{(0)}, \boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}, \boldsymbol{\Sigma}_1^{(0)}, \ldots, \boldsymbol{\Sigma}_K^{(0)})$, and for $t = 0, 1, 2, \ldots$, the E-step and M-step are developed as follow:

- E-step:

  We calculate the conditional expectation of the full data log-likelihood function given the observed data and $\psi^{(t)}$:

$$\begin{aligned} E[l_n(Y;\psi)|Y_{obs}, \psi^{(t)}] &= E[\sum_{i=1}^{n}\sum_{j=1}^{K} I(Z_i = j)(\log(\eta_j) + \log(P_{\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j}(\boldsymbol{X}_i)))|\{\boldsymbol{X}_i\}_1^n, \psi^{(t)}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{K} \gamma_{ij}^{(t)}(\log(\eta_j) + \log(P_{\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j}(\boldsymbol{X}_i))), \end{aligned}$$

  where,

$$\begin{aligned} \gamma_{ij}^{(t)} &= E[I(Z_i = j)|\{\boldsymbol{X}_i\}_1^n, \psi^{(t)}] \\ &= E[I(Z_i = j)|\boldsymbol{X}_i, \psi^{(t)}] \\ &= P(Z_i = j|\boldsymbol{X}_i, \psi^{(t)}) \\ &= \frac{P(\boldsymbol{X}_i|Z_i = j, \psi^{(t)})P(Z_i = j|\psi^{(t)})}{P(\boldsymbol{X}_i|\psi^{(t)})} \\ &= \frac{\eta_j^{(t)} P_{\boldsymbol{\mu}_j^{(t)},\boldsymbol{\Sigma}_j^{(t)}}(\boldsymbol{X}_i)}{\sum_{j=1}^{K}\eta_j^{(t)} P_{\boldsymbol{\mu}_j^{(t)},\boldsymbol{\Sigma}_j^{(t)}}(\boldsymbol{X}_i)}. \end{aligned}$$

- M-step

  At the M-step, we would like to maximize the conditional expectation that we obtained in the E-step to find new estimates for the parameters. Specifically, we would like to obtain the updated estimate $\psi^{(t+1)}$ such that,

$$\begin{aligned} \psi^{(t+1)} &= argmax_\psi E[\sum_{i=1}^{n}\sum_{j=1}^{K} I(Z_i = j)(\log(\eta_j) + \log(P_{\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j}(\boldsymbol{X}_i)))|\{\boldsymbol{X}_i\}_1^n, \psi^{(t)}] \\ &= argmax_\psi\{\sum_{i=1}^{n}\sum_{j=1}^{K} \gamma_{ij}^{(t)}(\log(\eta_j) + \log(P_{\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j}(\boldsymbol{X}_i)))\}. \end{aligned}$$

  Then, we can find the estimate $\boldsymbol{\mu}_j^{(t+1)}$ by differentiating the objective function and setting the derivative to zero,

$$\frac{\partial}{\partial \boldsymbol{\mu}_j}\{\sum_{i=1}^{n}\sum_{j=1}^{K} \gamma_{ij}^{(t)}(\log(\eta_j) + \log(P_{\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j}(\boldsymbol{X}_i)))\} = \sum_{i=1}^{n} \gamma_{ij}^{(t)} \boldsymbol{\Sigma}_j^{-}(\boldsymbol{X}_i - \boldsymbol{\mu}_j) = \mathbf{0}.$$

By solving the above equation, we have that,

$$\boldsymbol{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^n \gamma_{ij}^{(t)} \boldsymbol{X}_i}{\sum_{i=1}^n \gamma_{ij}^{(t)}}.$$

Similarly, we can find the estimate $\boldsymbol{\Sigma}_j^{(t+1)}$ by using the same technique,

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}_j^-}\{\sum_{i=1}^n \sum_{j=1}^K \gamma_{ij}^{(t)}(\log(\eta_j) + \log(P_{\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j}(\boldsymbol{X}_i)))\} \\
=&\frac{1}{2}\sum_{i=1}^n \gamma_{ij}^{(t)} \boldsymbol{\Sigma}_j - \frac{1}{2}\sum_{i=1}^n \gamma_{ij}^{(t)}(\boldsymbol{X}_i - \boldsymbol{\mu}_j)(\boldsymbol{X}_i - \boldsymbol{\mu}_j)^T \\
=&\mathbf{0}.
\end{aligned}$$

By solving the above equations and using the invariance property, we have that,

$$\boldsymbol{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^n \gamma_{ij}^{(t)}(\boldsymbol{X}_i - \boldsymbol{\mu}_j^{(t+1)})(\boldsymbol{X}_i - \boldsymbol{\mu}_j^{(t+1)})^T}{\sum_{i=1}^n \gamma_{ij}^{(t)}}.$$

Finally, to find the estimate $\eta_j^{(t+1)}$, we would need to maximize the objective function with respect to $\eta_j$ subject to the constraint such that,

$$\sum_{j=1}^K \eta_j^{(t+1)} = 1$$

By using the Lagrange Multiplier, we have that,

$$\frac{\partial}{\partial \eta_j}\{\sum_{i=1}^n \sum_{j=1}^K \gamma_{ij}^{(t)}(\log(\eta_j) + \log(P_{\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j}(\boldsymbol{X}_i)))\} = \sum_{i=1}^n \frac{\gamma_{ij}^{(t)}}{\eta_j} = \lambda,$$

so that,

$$\eta_j = \sum_{i=1}^n \frac{\gamma_{ij}^{(t)}}{\lambda},$$

$$\sum_{j=1}^K \eta_j^{(t+1)} = \frac{\sum_{i=1}^n \sum_{j=1}^K \gamma_{ij}^{(t)}}{\lambda} = \frac{n}{\lambda} = 1,$$

$$\lambda = n.$$

Therefore, the estimate $\eta_j^{(t+1)}$ is obtained by solving

$$\sum_{i=1}^n \frac{\gamma_{ij}^{(t)}}{\eta_j} = n,$$

which leads to,

$$\eta_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\gamma_{ij}^{(t)}.$$

To sum up, at the M-step, we have that,

$$\eta_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\gamma_{ij}^{(t)},$$

$$\boldsymbol{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n}\gamma_{ij}^{(t)}\boldsymbol{X}_i}{\sum_{i=1}^{n}\gamma_{ij}^{(t)}},$$

$$\boldsymbol{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^{n}\gamma_{ij}^{(t)}(\boldsymbol{X}_i - \boldsymbol{\mu}_j^{(t+1)})(\boldsymbol{X}_i - \boldsymbol{\mu}_j^{(t+1)})^T}{\sum_{i=1}^{n}\gamma_{ij}^{(t)}}.$$

$\square$

**2.** Consider the Gaussian graphical model $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with precision matrix $\Theta = \boldsymbol{\Sigma}^{-1}$. Show that $\Theta_{jk} = 0$ if and only if $X_j \perp\!\!\!\perp X_k | \{X_1, \ldots, X_d\} \backslash \{X_j, X_k\}$.

Proof.

For the Gaussian graphical model $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if we partition the random vector $\boldsymbol{X}$ into two components, $\boldsymbol{X}_A \in \mathbb{R}^a$ and $\boldsymbol{X}_B \in \mathbb{R}^b$ such that $a + b = d$, and let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be partitioned accordingly, i.e.,

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{bmatrix},$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{A,A} & \boldsymbol{\Sigma}_{A,B} \\ \boldsymbol{\Sigma}_{B,A} & \boldsymbol{\Sigma}_{B,B} \end{bmatrix},$$

then,

- The marginal distribution of $\boldsymbol{X}_A$ is $N(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_{A,A})$.
- The conditional distribution of $\boldsymbol{X}_A | \boldsymbol{X}_B = \boldsymbol{x}_B$ is $N(\boldsymbol{\mu}_{A|B}, \boldsymbol{\Sigma}_{A|B})$, where

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{A,B} \boldsymbol{\Sigma}_{B,B}^{-1} (\boldsymbol{x}_B - \boldsymbol{\mu}_B),$$

$$\boldsymbol{\Sigma}_{A|B} = \boldsymbol{\Sigma}_{A,A} - \boldsymbol{\Sigma}_{A,B} \boldsymbol{\Sigma}_{B,B}^{-1} \boldsymbol{\Sigma}_{B,A}.$$

Now, we can start by proving the following claim:

- **Claim:** Under the setting of $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for $j, k \in \{1, \ldots, d\}$ with $j \neq k$, we have that, $X_j \perp\!\!\!\perp X_k$ if and only if $\sigma_{jk} = 0$, where $X_j$ and $X_k$ are $j^{th}$ and $k^{th}$ components of $\boldsymbol{X}$, and $\sigma_{jk}$ is the $(j,k)^{th}$ element of $\boldsymbol{\Sigma}$.

- **Proof:** We know that the marginal distribution of $(X_j, X_k)^T$ is $N((\mu_j, \mu_k)^T, \boldsymbol{\Sigma}_{\{j,k\}})$, where,

$$\boldsymbol{\Sigma}_{\{j,k\}} = \begin{bmatrix} \sigma_{jj} & \sigma_{jk} \\ \sigma_{kj} & \sigma_{kk} \end{bmatrix}.$$

Then the marginal distribution of $X_j$ is $N(\mu_j, \sigma_{jj})$, and the conditional distribution of $X_j|X_k = x_k$ is,

$$N(\mu_j + \sigma_{jk}\sigma_{kk}^{-1}(x_k - \mu_k), \sigma_{jj} - \sigma_{jk}\sigma_{kk}^{-1}\sigma_{kj}).$$

Then, $X_j \perp\!\!\!\perp X_k$ if and only if the marginal distribution of $X_j$, $f_{X_j}(x_j)$, equals to the conditional distribution of $X_j|X_k = x_k$, $f_{X_j|X_k=x_k}(x_j)$ (by the definition of independence), which happens if and only if $\sigma_{jk} = 0$ (both directions are very obvious), as desired.

Once we have proved the above claim, we move onto the original statement that we would like to prove. Then we have that,

- $X_j \perp\!\!\!\perp X_k|\{X_1, \ldots, X_d\}\backslash\{X_j, X_k\}$ if and only if the conditional covariance matrix, $\mathbf{\Sigma}_{\{j,k\}|[d]\backslash\{j,k\}}$, is diagonal (by the claim that we have proved earlier).
- $\mathbf{\Sigma}_{\{j,k\}|[d]\backslash\{j,k\}}$ is diagonal if and only if its inverse, $\mathbf{\Sigma}^{-1}_{\{j,k\}|[d]\backslash\{j,k\}}$, is diagonal (because the inverse of a non-singular diagonal matrix is also diagonal).
- $\mathbf{\Sigma}^{-1}_{\{j,k\}|[d]\backslash\{j,k\}} = \Theta_{\{j,k\}}$, where
$$\Theta_{\{j,k\}} = \begin{bmatrix} \Theta_{jj} & \Theta_{jk} \\ \Theta_{kj} & \Theta_{kk} \end{bmatrix}$$
(this can be verified by deriving the inverse of block matrix).
- $\Theta_{\{j,k\}}$ is diagonal matrix is equivalent to $\Theta_{jk} = 0$.

Therefore, $\Theta_{jk} = 0$ if and only if $X_j \perp\!\!\!\perp X_k|\{X_1, \ldots, X_d\}\backslash\{X_j, X_k\}$, as desired. $\square$