Fundamentals of Probabilistic Data Mining - Homework 1

Iskander Gaba, Kenza Halably, Sara Rahoui, Amir Kroudir

November 21, 2020

1 Solution

Note: Code for this lab is available at this link.

1.1 Preparatory work and modelling

1.1.1 Deriving the re-estimation formula for Gaussian Mixture Model (GMM):

Let $X = (X_1, X_2, ..., X_N)$ be a random sample from the mixture distribution. Let N and K be the number of samples and the number of mixture model components, respectively. Let $Z = (Z_1, Z_2, ..., Z_N)$ latent variables representing the component k to which each sample X_n belongs to. That is: $Z_n = k$ with $n \in \{1, 2, ..., N\}$ and $k \in \{1, 2, ..., K\}$.
Let $\pi_k = P(Z_n = k)$. We call it the mixture proportion of the k-th component. We also have:

$$P(X_n = x|Z_n = k) = \mathcal{N}(X_n; \mu_k, \Sigma_k)$$

Therefore:

$$P(X_n = x) = \sum_{k=1}^{K} P(Z_n = k) \cdot \mathcal{N}(X_n; \mu_k, \Sigma_k) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(X_n; \mu_k, \Sigma_k)$$

We introduce $\theta = (\mu_1, \mu_2, ..., \mu_k, \Sigma_1, \Sigma_2, ..., \Sigma_k)$ that represents the set of parameters that we need to estimate. $\theta^{(t)}$ represents the set of parameters estimated at time-step (i.e. iteration) t. Let us first compute $\eta_{nk}^{(t)} = P(Z_n = k|X_n = x, \theta^{(t)})$. Using Bayes rule, we get:

$$\eta_{nk}^{(t)} = P(Z_n = k|X_n = x, \theta^{(t)}) = \frac{P(Z_n = k|\theta^{(t)}) \cdot P(X_n = x|Z_n = k, \theta^{(t)})}{P(X_n = x|\theta^{(t)})}$$

$$\eta_{nk}^{(t)} = P(Z_n = k|X_n = x, \theta^{(t)}) = \frac{\pi_k^{(t)} \cdot \mathcal{N}(X_n; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t)} \cdot \mathcal{N}(X_n; \mu_{k'}^{(t)}, \Sigma_{k'}^{(t)})}$$

Next, we compute $P(X, Z|\theta)$:

$$P(X, Z|\theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} (\pi_k \cdot \mathcal{N}(X_n; \mu_k, \Sigma_k))^{I(Z_n=k)}$$

where $I(x)$ being the indicator function that returns 1 if x is *true* and 0 otherwise. So we get:

$$\log(P(X, Z|\theta)) = \sum_{n=1}^{N} \sum_{k=1}^{K} I(Z_n = k) \cdot (\log(\pi_k) + \log(\mathcal{N}(X_n; \mu_k, \Sigma_k)))$$

- **E-Step**: We define $Q(\theta, \theta^{(t)})$ as:

$$Q(\theta, \theta^{(t)}) = \mathop{\mathbb{E}}_{Z|X,\theta^{(t)}}(\log(p(X, Z|\theta)))$$

$$Q(\theta, \theta^{(t)}) = \mathop{\mathbb{E}}_{Z|X,\theta^{(t)}}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} I(Z_n = k) \cdot (\log(\pi_k) + \log(\mathcal{N}(X_n; \mu_k, \Sigma_k)))\right]$$

$$Q(\theta, \theta^{(t)}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathop{\mathbb{E}}_{Z|X_n,\theta^{(t)}}[I(Z_n = k)] \cdot (\log(\pi_k) + \log(\mathcal{N}(X_n; \mu_k, \Sigma_k)))$$

$$Q(\theta, \theta^{(t)}) = \sum_{n=1}^{N}\sum_{k=1}^{K} P(Z_n = k|X_n = x, \theta^{(t)}) \cdot (\log(\pi_k) + \log(\mathcal{N}(X_n; \mu_k, \Sigma_k)))$$

$$Q(\theta, \theta^{(t)}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_{nk}^{(t)} \cdot (\log(\pi_k) + \log(\mathcal{N}(X_n; \mu_k, \Sigma_k)))$$

Knowing that:

$$\mathcal{N}(X_n; \mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}} e^{-\frac{1}{2}(X_n - u_k)^\top \Sigma_k^{-1}(X_n - \mu_k)}$$

where d is the dimension of X_n, and knowing that $\det(A) = \frac{1}{\det(A^{-1})}$, we have:

$$\log(\mathcal{N}(x_n, \mu_k, \Sigma_k)) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(X_n - \mu_k)^\top \Sigma_k^{-1}(X_n - \mu_k)$$

Hence:

$$Q(\theta, \theta^{(t)}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_{nk}^{(t)} \cdot (\log(\pi_k) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(X_n - \mu_k)^\top \Sigma_k^{-1}(X_n - \mu_k))$$

- **M-Step:** We seek to find parameters $\theta^{(t+1)}$ that maximize $Q(\theta, \theta^{(t)})$ in this step.

  - π_k: For convenience, we introduce the the variable $S_k^{(t)} = \sum_{n=1}^{N} \eta_{nk}^{(t)}$. Note that since we have to take into consideration the constraint $\sum_{k=1}^{K} \pi_k^{(t)} = 1$, we maximize the Lagrangian function $\mathcal{L}(\theta, \theta^{(t)}, \lambda)$ defined as:

$$\mathcal{L}(\theta, \theta^{(t)}, \lambda) = Q(\theta, \theta^{(t)}) - \lambda\left(1 - \sum_{k=1}^{K} \pi_k^{(t)}\right)$$

  Now we solve $\frac{d\mathcal{L}}{d\pi_k} = 0$:

$$\frac{d\mathcal{L}}{d\pi_k} = \sum_{n=1}^{N} \frac{\eta_{nk}^{(t)}}{\pi_k} - \lambda = 0$$

This means that $\pi_k^{(t+1)} = \frac{\sum_{n=1}^{N} \eta_{nk}^{(t)}}{\lambda} = \frac{S_k^{(t)}}{\lambda}$. Let us now compute λ. We have:

$$\frac{\sum_{k=1}^{K} S_k^{(t)}}{\lambda} = \sum_{k=1}^{K} \pi_k = 1$$

So:

$$\lambda = \sum_{k=1}^{K} S_k^{(t)} = \sum_{k=1}^{K} \sum_{n=1}^{N} \eta_{nk}^{(t)}$$

$$\lambda = \sum_{n=1}^{N} \sum_{k=1}^{K} P(Z_n = k | X_n = x, \theta^{(t)}) = \sum_{n=1}^{N} 1 = N$$

Therefore:

$$\pi_k^{(t+1)} = \frac{S_k^{(t)}}{N}$$

– μ_k: Let $v_k = X_n - \mu_k$

$$\frac{\partial Q}{\partial v_k} = \frac{\partial}{\partial v_k} \left[\sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk}^{(t)} \cdot (-\frac{1}{2} (X_n - \mu_k)^\top \Sigma_k^{-1} (X_n - \mu_k)) \right]$$

$$\frac{\partial Q}{\partial v_k} = -\frac{1}{2} \sum_{n=1}^{N} \eta_{nk} \frac{\partial}{\partial v_k} (v_k^T \Sigma_k^{-1} v_k)$$

As Σ_k is symmetric, so is Σ_k^{-1}. We also know that $\frac{\partial v^\top M v}{\partial v} = 2Mv$ with M symmetric. Hence:

$$\frac{\partial Q}{\partial v_k} = -\sum_{n=1}^{N} \eta_{nk} \Sigma_k^{-1} v_k$$

$$\frac{\partial Q}{\partial \mu_k} = \sum_{n=1}^{N} \eta_{nk} \Sigma_k^{-1} (X_n - \mu_k)$$

We now solve $\frac{\partial Q}{\partial \mu_k} = 0$.

$$\frac{\partial Q}{\partial \mu_k} = \sum_{n=1}^{N} \eta_{nk} \Sigma_k^{-1} (X_n - \mu_k) = 0$$

$$\Rightarrow \sum_{n=1}^{N} \eta_{nk} \Sigma_k^{-1} X_n = \sum_{n=1}^{N} \eta_{nk} \Sigma_k^{-1} \mu_k$$

$$\Rightarrow S_k^{(t)} \cdot \mu_k = \sum_{n=1}^{N} \eta_{nk} X_n$$

Therefore:

$$\mu_k^{(t+1)} = \frac{1}{S_k^{(t)}} \sum_{n=1}^{N} \eta_{nk}^{(t)} X_n$$

– Σ_k: Let us consider $\frac{\partial Q}{\partial \Sigma_k}$.

$$\frac{\partial Q}{\partial \Sigma_k} = \frac{\partial}{\partial \Sigma_k} \left[\sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk}^{(t)} \cdot (-\frac{1}{2} \log(|\Sigma_k|) - \frac{1}{2} (X_n - \mu_k)^\top \Sigma_k^{-1} (X_n - \mu_k)) \right]$$

We Introduce $A_k = \Sigma_k^{-1}$. Knowing that $\det(X) = \frac{1}{\det(X^{-1})}$, we get:

$$\frac{\partial Q}{\partial A_k} = \frac{\partial}{\partial A_k} \left[\sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk}^{(t)} \cdot (-\frac{1}{2} \log(|A_k^{-1}|) - \frac{1}{2} (X_n - \mu_k)^\top A_k (X_n - \mu_k)) \right]$$

Knowing that $\mathrm{Tr}(ABC) = \mathrm{Tr}(BCA) = \mathrm{Tr}(CAB)$:

$$\frac{\partial Q}{\partial A_k} = \sum_{n=1}^{N} \eta_{nk} \left[\frac{1}{2} \frac{\partial \log(|A_k|)}{\partial A_k} - \frac{1}{2} \frac{\partial}{\partial A_k} \mathrm{Tr} \left(A_k (X_n - \mu_k)(X_n - \mu_k)^\top \right) \right]$$

And since $\frac{\partial \log(|M|)}{\partial M} = (M^{-1})^\top$ and $\frac{\partial \mathrm{Tr}(MA)}{\partial M} = (A)^\top$, we get:

$$\frac{\partial Q}{\partial A_k} = \sum_{n=1}^{N} \eta_{nk} \left[\frac{1}{2} (A_k^{-1})^\top - \frac{1}{2} (X_n - \mu_k)(X_n - \mu_k)^\top \right]$$

$$\frac{\partial Q}{\partial A_k} = \frac{S_k^{(t)} \cdot (A_k^{-1})^\top}{2} - \frac{1}{2} \sum_{n=1}^{N} (X_n - \mu_k)(X_n - \mu_k)^\top$$

$$\frac{\partial Q}{\partial A_k} = \left(\frac{A_k^{-1}}{2} \right)^\top \cdot S_k^{(t)} - \frac{1}{2} \sum_{n=1}^{N} \eta_{n_k} (X_n - \mu_k)(X_n - \mu_k)^\top$$

We now solve $\frac{\partial Q}{\partial A_k} = 0$.

$$\frac{\partial Q}{\partial A_k} = 0 \Rightarrow (A_k^{-1})^\top = \frac{1}{S_k^{(t)}} \sum_{n=1}^{N} \eta_{nk} (X_n - \mu_k)(X_n - \mu_k)^\top$$

This means that:

$$\Sigma_k^{(t+1)} = \frac{1}{S_k^{(t)}} \sum_{n=1}^{N} \eta_{nk} (X_n - \mu_k^{(t+1)})(X_n - \mu_k^{(t+1)})^\top$$

1.1.2 Bivariate GMM Simulation

To simulate the model, we first choose the class Z of each data point (i.e. the component to which it belongs) from the probability distribution defined by $P(Z = 1) = 0.7$ and $P(Z = 2) = 0.3$. After that, we plug the data point into the equation of the Gaussian component it belongs to and plot the result. The distribution of the Gaussian Mixture Model is represented in Figure 1.

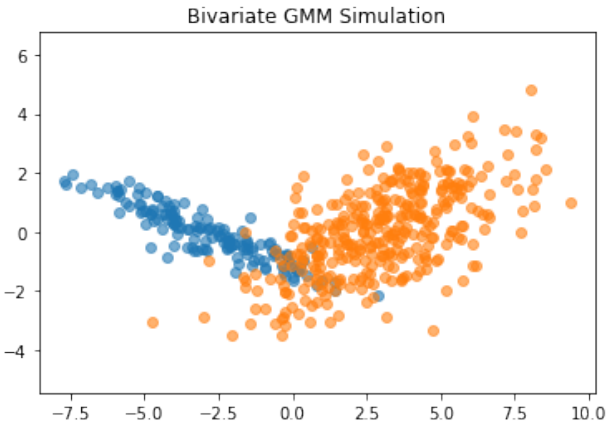


Figure 1: Bivariate GMM

1.1.3 Data Pre-processing

We downloaded the Unistroke dataset, processed the data for the letter A as per the instructions of the joint README file, and plotted the result in Figure 2.

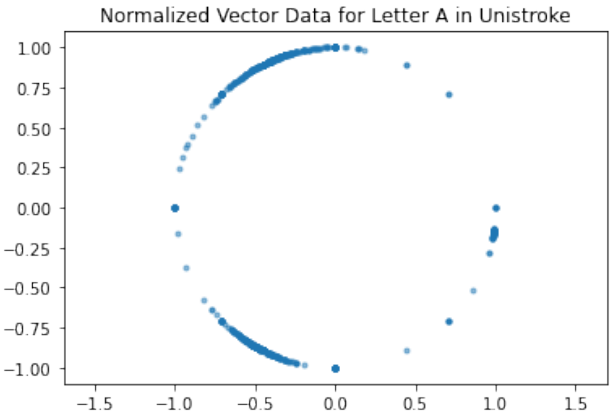


Figure 2: Unistroke Data for the Letter A

1.1.4 Could a 2-components GMM be appropriate for letter A? Why?

Yes, we think a that a 2-components GMM could be appropriate for letter A to a certain extent. This is because the letter A is composed of two unique strokes (i.e. drawn like a "Λ") and assuming that the data for each stroke is normally distributed, a 2-components GMM should be appropriate to recognize its pattern. To illustrate this, we have a plot of the combined raw data for the letter A in Figure 3. As you can see, the data represent clear two distinct strokes. Please note that the

representation in Figure 3 (and all the figures using raw data) are not used in the actual fitment of models and are only approximate representation for the purpose of visualizing the strokes and better understanding.

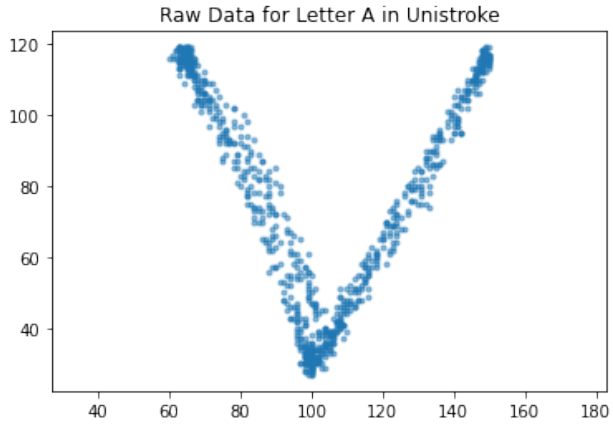


Figure 3: Approximate Representation of Raw Unistroke data for the Letter A

However, we need to keep in mind that the data Figure 2 may suggest a third GMM component, albeit with a small contribution as there is a third smaller cluster of points that can be considered. This does make sense as well considering that the Letter A has three points when drawn like a Λ.

1.2 Data Analysis: Gaussian Model

1.2.1 2-components Bivariate GMM Estimation on The Letter A

After fitting the data to a 2-component GMM, we get the following parameters:

$$\mu_1 \approx \begin{pmatrix} -0.39 \\ 0.87 \end{pmatrix}, \ \mu_2 \approx \begin{pmatrix} -0.27 \\ -0.78 \end{pmatrix} \text{ and } \Sigma_1 \approx \begin{pmatrix} 0.06 & 0.04 \\ 0.04 & 0.04 \end{pmatrix}, \ \Sigma_2 \approx \begin{pmatrix} 0.26 & 0.11 \\ 0.11 & 0.06 \end{pmatrix}$$

and weights:

$$Z \approx (0.503, 0.497)$$

Notice that the weights are almost equal which is to be expected of the way the two strokes of the letter A are drawn.

1.2.2 Label the data using the estimated model and show the pdf of the estimated GMM

The desired PDF plot is shown in Figure 4.

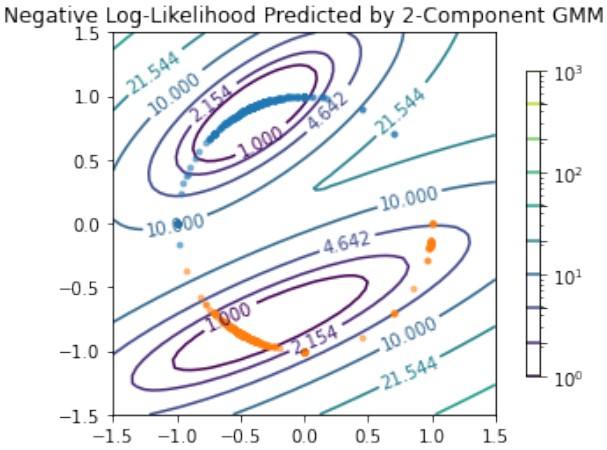


Figure 4: Labeled Unistroke Data of Letter A using 2-Component GMM

What happens if you use more components?
When we fit a 3-component GMM, we get the following parameters:

$$\mu_1 \approx \begin{pmatrix} -0.37 \\ 0.90 \end{pmatrix}, \ \mu_2 \approx \begin{pmatrix} -0.48 \\ -0.83 \end{pmatrix}, \ \mu_3 \approx \begin{pmatrix} 0.95 \\ -0.17 \end{pmatrix}$$

$$\text{and } \Sigma_1 \approx \begin{pmatrix} 0.04 & 0.02 \\ 0.02 & 0.01 \end{pmatrix}, \ \Sigma_2 \approx \begin{pmatrix} 0.03 & -0.03 \\ -0.03 & 0.03 \end{pmatrix}, \ \Sigma_3 \approx \begin{pmatrix} 0.01 & 0.01 \\ 0.01 & 0.06 \end{pmatrix}$$

and weights:

$$Z \approx (0.48, 0.45, 0.07)$$

You can notice here that the third Gaussian component has a very small weight of 0.07 making its contribution to the pattern recognition a very small one.

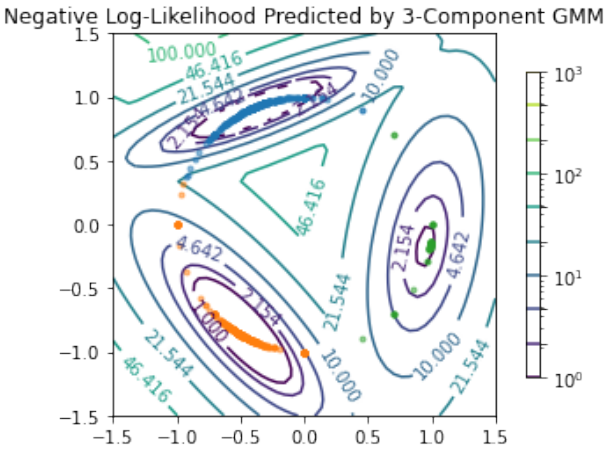
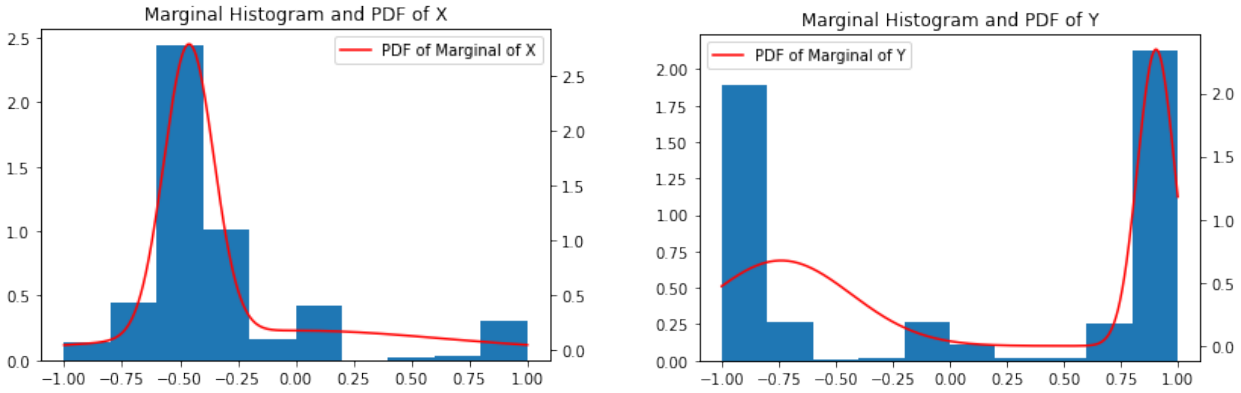


Figure 5: Labeled Unistroke Data of Letter A using 3-Component GMM

One can also see from Figure 5 that the amount of data belonging to the third Gaussian component (shown in green) is very small as expected earlier.

1.2.3 To validate the assumption of bivariate Gaussian mixture

- **(a) Plot each marginal histogram (in X and Y) and add the estimated mixture of univariate Gaussian pdfs to the figure.**

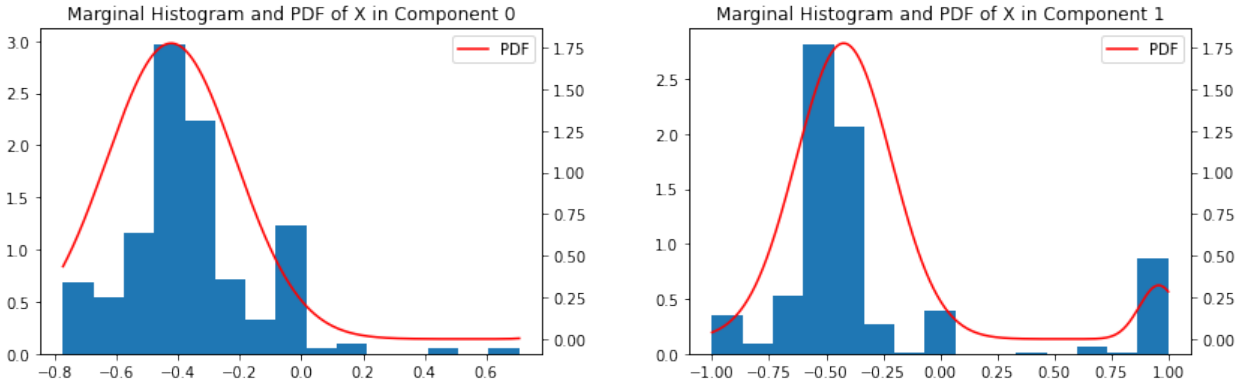


(a) X (b) Y

Figure 6: Marginal Histograms and PDFs of X and Y

- **(b) For each marginal, provide separate histograms of each cluster and add the estimated univariate Gaussian pdf to the figure.**

  Results are shown in Figure 7 and Figure 8.

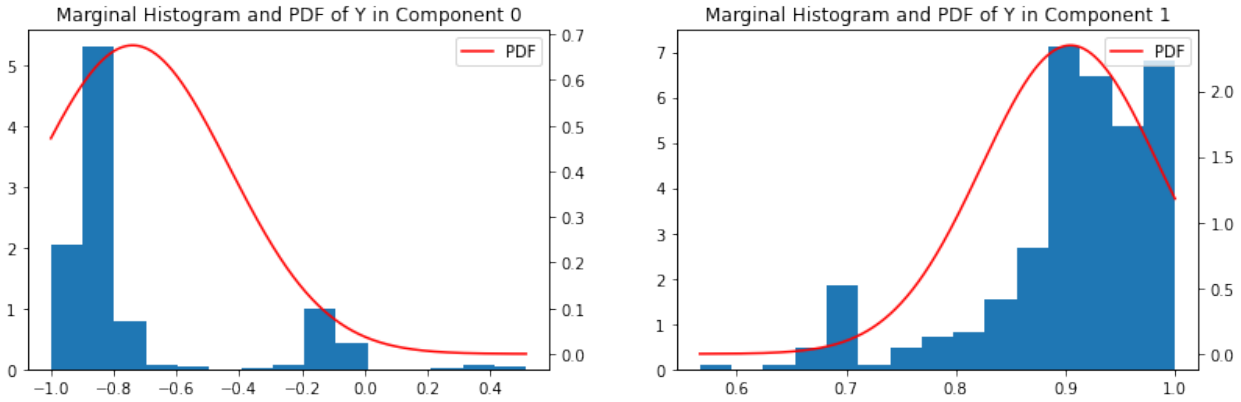


(a) Class 0 (b) Class 1

Figure 7: Marginal Histograms and PDFs of X by Class

1.2.4 Comment the results of questions 3 (a) and (b). What to think about the bivariate Gaussian mixture assumption? Why?

We can see in Figure 6 that the marginal histogram of X contains only one peak whereas that of Y contains two peaks. This can be explained by the fact the Y vectors in stroke 1 are moving in the opposite direction of Y vectors in stroke 2 with about the same magnitude whereas the vectors for X are moving in the same direction for both strokes, with about the same magnitude. When we break the marginals by class, we notice that pdfs for X in both classes are centered approximately around the same mean (see Figure 7) while those of Y are clearly separated as it is shown in Figure



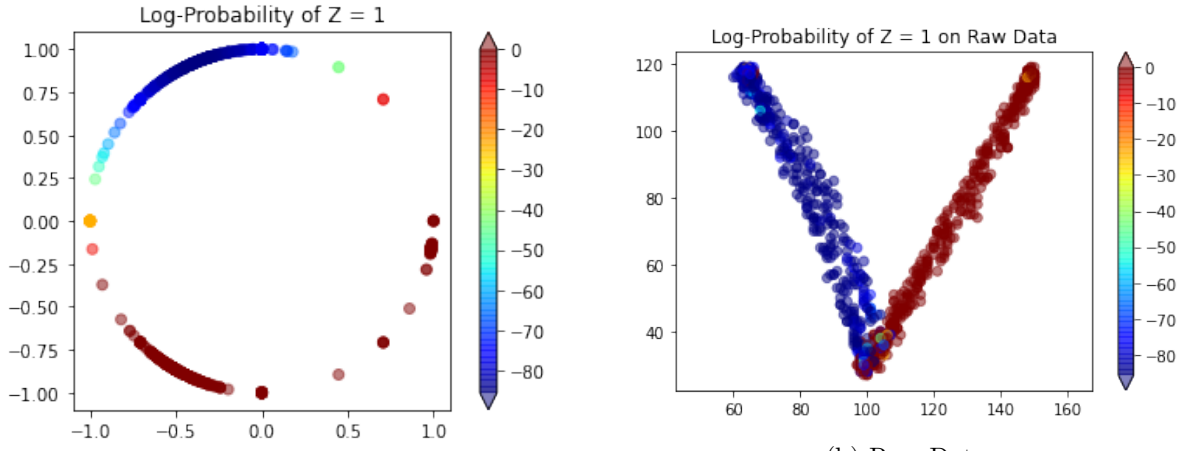
(a) Class 0 (b) Class 1

Figure 8: Marginal Histograms and PDFs of Y by Class

8. Furthermore, we can see in Figure 7 that there is a smaller outlying peak to the right, suggesting that the smaller outlier cluster we referred to earlier maybe influencing it.
The observations above suggest that, for the letter A, only Y marginal values do clearly belong to two different Gaussian components. The x values seem to belong to one Gaussian component only, rendering our bivariate assumption weaker as it appears that only Y vectors that are making the clearest distinction between each stroke.

1.2.5 Interpretation of the colormap plots

Figure 9a represents the log-probability of data points belonging to the second component of the GMM. Figure 9b represents the log-probability of raw data points for the letter A belonging to the second component of the GMM.



(a) Processed Data (b) Raw Data

Figure 9: Log-Probability of $P(Z_i = 1|X)$ Represented by Colormap

Please note that we used the log scale to better emphasize colors. We can see that while there are no hard boundaries and despite the weakness of our bivariate assumption, the model does recognize each of the strokes fairly well as it is shown in Figure 9b.

1.3 Mandatory additional questions:

1.3.1 Transform the Unistroke data to angular data. Plot the histogram of angles and comment

The histogram in Figure 10 represents the histogram of the angles of the data vectors in the range $[-\pi, \pi]$. We can clearly see from Figure 10 that the data is mostly concentrated around two extreme peaks. It is clear therefore that only the angle information is needed to be able to capture each of the two strokes that make up the letter A. We can also see a small peak concentrated at around $0°$. These probably represent the three dots making the start, the end, and the turning points of the drawing of Λ as change in them could initially be very minimal sometimes.

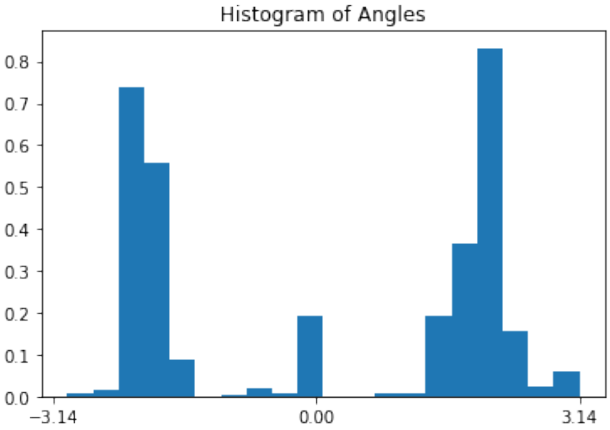


Figure 10: Histogram of Angles of the Data Vectors

1.3.2 Define von Mises and mixtures of von Mises distributions

- **von Mises Distribution:** The von Mises probability density function for the angle x is given by:

$$f(x \mid \mu, \kappa) = \text{vM}(x; \mu, \kappa) = \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)}$$

where $I_0(\kappa)$ is the modified Bessel function of order 0:

$$I_0(\kappa) = \frac{1}{\pi} \int_0^{\pi} e^{\kappa \cos(\theta)} d\theta$$

μ and $\frac{1}{\kappa}$ are analogous to μ and σ^2 in the normal distribution. κ is a measure of concentration.

- **von Mises Mixture Model:**

$$p(x_n) = \sum_{k=1}^{k} \pi_k \cdot \text{vM}(x_n; \mu_k, \kappa_k) = \sum_{k=1}^{k} \pi_k \cdot \frac{e^{\kappa_d (x_n - \mu_k)}}{2\pi I_0(\kappa_k)}$$

where:

- π_k: the prior probability.
- κ_k: the number of mixture component.

1.3.3 A priori, would a mixture of von Mises distributions be more or less adequate than Gaussian mixtures on the real data set of part 1.1? Why?

A priori, von Mises Mixture models can be seen as more adequate to the problem at hand since it can be reduced to the estimation of one parameter which is the angle towards which the unitary gradient vector is heading. From the histogram of the angular data, we can see that there are two major peaks. This makes sense as the transcription of the letter A requires two strokes where the angle of each one is about the inverse of the other.
On the other hand, very small vector changes can be challenging to classify within the right GMM component as the angle information is not considered. In the particular case of the letter A, if two vectors, each located at a different stroke, are located very close to pointy tip of the letter, it can be hard to classify each in the correct GMM component. von Mises Mixture on the other hand will have less trouble doing that as the angles of each of the vectors are opposite in sign in the range $[-\pi, \pi]$.
Note that we can even use a 3-component von Mises misture if we wanted to capture the smaller cluster represented by the small peak around 0 radian degrees. We believe that these points with very minimal degree changes represent few points at the start and end of drawing the letter A, and in between the two main strokes.

1.3.4 Equations for the E-step and M-step of the EM algorithm for mixtures of von Mises distributions

Let us first recall the definition of vM, the von Mises distribution:

$$\mathrm{vM}(x_n; \mu_k, \kappa_k) = \frac{e^{\kappa_k(x_n - \mu_k)}}{2\pi I_0(\kappa_k)}$$

The probability density function:

$$P(x_n) = \sum_{k=1}^{k} P(Z_n = k) \cdot \mathrm{vM}(x_n; \mu_k, \kappa_k) = \sum_{k=1}^{k} \pi_k \cdot \mathrm{vM}(x_n; \mu_k, \kappa_k)$$

First, we compute $\eta_{nk}^{(t)} = P(Z_n = k | x_n, \theta^{(t)})$:

$$P(Z_n = k | x_n, \theta^{(t)}) = \frac{p(x_n | Zn = k, \theta^{(t)}) \cdot P(Z_n = k)}{p(x_n | \theta^{(t)})}$$

$$P(Z_n = k | x_n, \theta^{(t)}) = \frac{\pi_k \cdot \mathrm{vM}(x_n; \mu_k, \kappa_k)}{\sum_{k=1}^{k} \pi_k \cdot \mathrm{vM}(x_n; \mu_k, \kappa_k)} = \eta_{nk}^{(t)}$$

Next, we compute $\log(p(x, Z|\theta))$:

$$p(x, Z|\theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} (\pi_k \cdot \mathrm{vM}(x_n; \mu_k, \kappa_k)^{I(Z_n = k)}$$

where $I(x)$ being the indicator function that returns 1 if x is $true$ and 0 otherwise.

$$\log(p(x, Z|\theta)) = \sum_{n=1}^{N}\sum_{k=1}^{K} I(Z_n = k) \cdot (\log(\pi_k) + \log(\text{vM}(x_n; \mu_k, \kappa_k)))$$

Therefore, we have the same form as in GMM.

- **E-Step:** We will first need to compute $Q(\theta, \theta^{(t)})$:

$$Q(\theta, \theta^{(t)}) = \mathop{\mathbb{E}}_{Z|x,\theta^{(t)}}(\log(p(x, Z|\theta)))$$

$$Q(\theta, \theta^{(t)}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_{nk}^{(t)} \cdot (\log(\pi_k^{(t)}) + \log(\text{vM}(x_n; \mu_k^{(t)}, \kappa_k^{(t)})))$$

$$Q(\theta, \theta^{(t)}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_{nk}^{(t)} \cdot (\log(\pi_k^{(t)}) + \kappa_k \cdot \cos(x_n - \mu_k^{(t)}) - \log(2\pi) - \log(I_0(\kappa_k^{(t)})))$$

- **M-Step:** In this step we need to maximize the $Q(\theta, \theta^{(t)})$ with respect to the parameters θ.
  - π_k: Like in GMM, we do not need to know anything about the distribution of x and we get the same update rule for π_k. We introduce the the variable $S_k^{(t)} = \sum_{n=1}^{N} \eta_{nk}^{(t)}$. We can then conclude that:
  $$\pi_k^{(t+1)} = \frac{S_k^{(t)}}{N}$$
  - μ_k: We consider solving the equation $\frac{\partial Q}{\partial \mu_k} = 0$ for μ_k:
  $$\frac{\partial Q}{\partial \mu_k} = \sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot \kappa_k^{(t)} \sin(x_n - \mu_k^{(t)}) = 0$$
  Since $\kappa_k^{(t)} > 0$, we can divide both sides of the equation by κ_k and get:
  $$\frac{\partial Q}{\partial \mu_k} = \sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot \sin(x_n - \mu_k^{(t)}) = 0$$
  Next, we simplify the equation further:
  $$\frac{\partial Q}{\partial \mu_k} = \sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot (\sin(x_n)\cos(\mu_k^{(t)}) - \cos(x_n)\sin(\mu_k^{(t)})) = 0$$
  If $\cos(\mu_k) \neq 0$, then by dividing both sides by $\cos(\mu_k)$, we get:
  $$\frac{\partial Q}{\partial \mu_k} = \sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot (\sin(x_n) - \cos(x_n)\tan(\mu_k^{(t)})) = 0$$
  $$\iff \frac{\partial Q}{\partial \mu_k} = \sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot \sin(x_n) - \tan(\mu_k^{(t)}) \sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot \cos(x_n) = 0$$

Hence:

$$\tan^{(t+1)}(\mu_k) = \frac{\sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot \sin(x_n)}{\sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot \cos(x_n)}$$

Thus:

$$\mu_k^{(t+1)} = \tan^{-1}\left(\frac{\sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot \sin(x_n)}{\sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot \cos(x_n)}\right)$$

If $\cos(\mu_k) = 0$ (and therefore $\sin(\mu_k) = 1$), then we have:

$$\frac{\partial Q}{\partial \mu_k} = \sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot \cos(x_n) = 0$$

This means that $\tan^{(t+1)}(\mu_k)$ can be any value in $\mathbb{R}$, therefore the update rule works in the general case.

– κ_k: in the attempt to find the update rule for κ_k, we need first to compute $\frac{dI_0(x)}{dx}$:

$$\frac{dI_0(x)}{dx} = I_1(x) = \frac{1}{\pi}\int_0^{\pi} \cos(\theta) \cdot e^{x\cos(\theta)} d\theta$$

We then consider solving $\frac{\partial Q}{\partial \kappa_k} = 0$ for κ_k:

$$\frac{\partial Q}{\partial \kappa_k} = \sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot \left(\cos(x_n - \mu_k) - \frac{I_1(\kappa_k)}{I_0(\kappa_k)}\right) = 0$$

Using the update rule for μ_k, we get:

$$\frac{I_1(\kappa_k^{(t+1)})}{I_0(\kappa_k^{(t+1)})} = \frac{1}{S_k^{(t)}} \sum_{n=1}^{N} \eta_{nk}^{(t)} \cdot \cos(x_n - \mu_k^{(t+1)})$$

Due to the nature of Bassel function, the last equation will have to be numerically computed.

1.3.5 Fitting the Data in a 2-components von Mises Mixture Model

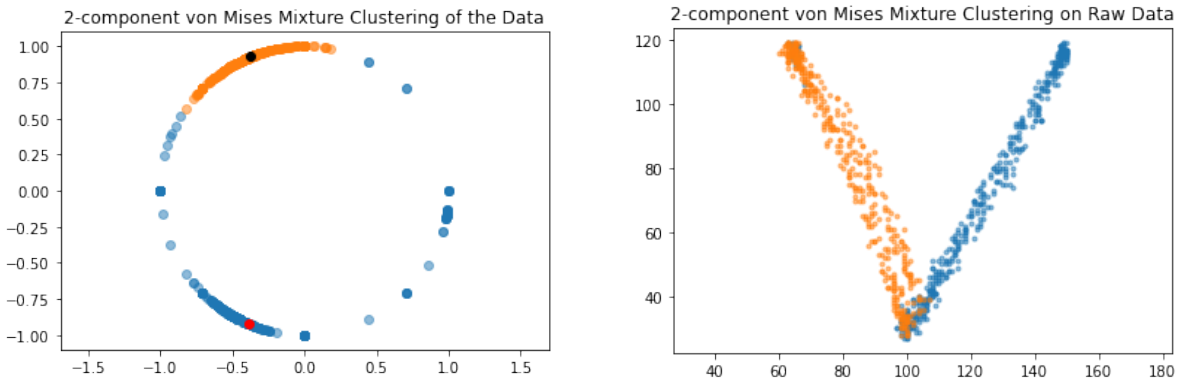
After fitting the data to a 2-component von Mises Mixture Model (VMM), we get the following parameters:

$$\mu_1 \approx \begin{pmatrix} -0.38 \\ -0.92 \end{pmatrix},\ \mu_2 \approx \begin{pmatrix} -0.37 \\ 0.92 \end{pmatrix} \text{ and } \kappa_1 \approx 2.71,\ \kappa_2 \approx 22.07$$

and weights:

$$Z \approx (0.53, 0.47)$$

The results are shown in Figure 11. We can see that each stroke of the letter is very well detected by the VMM.



(a) Processed Data

(b) Raw Data

Figure 11: Data Clustering Using 2-component VMM

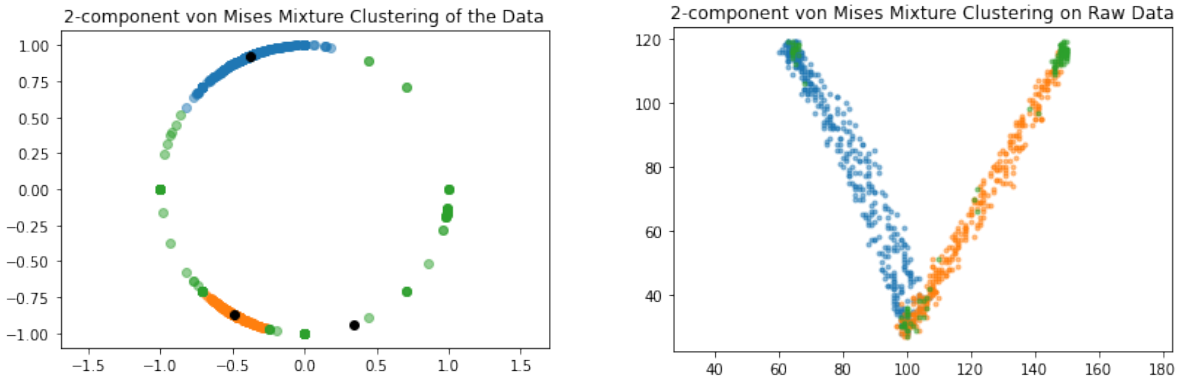
We also fit the data to a 3-component VMM. We get the following parameters:

$$\mu_1 \approx \begin{pmatrix} -0.38 \\ -0.92 \end{pmatrix},\ \mu_2 \approx \begin{pmatrix} -0.49 \\ -0.87 \end{pmatrix},\ \mu_3 \approx \begin{pmatrix} 0.34 \\ -0.94 \end{pmatrix} \text{ and } \kappa_1 \approx 23.48,\ \kappa_2 \approx 120.13,\ \kappa_3 \approx 0.6$$

and weights:

$$Z \approx (0.46, 0.36, 0.18)$$

Results are shown in Figure 12. As shown, the third component (represented in green) manages to capture points in the three corner areas of the letter A (drawn as Λ).



(a) Processed Data

(b) Raw Data

Figure 12: Data Clustering Using 3-component VMM