Lab 2 report

Independent mixture models

Authors:
Hussein AWALA
Amine BIROUK
Bonpagna KANN
Assem SADEK

Supervisors:
Prof. Jean-Baptiste DURAND
Prof. Xavier Alameda-Pineda

January 7, 2019

Contents

1 Independent mixture models

Independent mixture models, is one of the on going research topic. The aim is to clusters given data by finding the key distributions that can represent the data.

In this report, we will answer the questions on gaussian mixture models lab and comparing it with another related distribution called von mises distribution.

1.1 Lab work

1.1.1 Preparatory work and modelling

1. The goal is to find the parameters μ_k, Σ_k of every k multivariate Gaussian distribution in the mixture model that maximize the log-likelihood $\ln(p_\lambda(X, Z))$. But since Z is hidden variable, Thus we will first define $Q(\lambda, \lambda^{(m)})$ that we will maximize.

$$\begin{aligned} Q(\lambda, \lambda^{(m)}) &= \mathbb{E}_Z[\ln(p_\lambda(X, Z))] \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})^{(m)}[\ln(\pi_k) + \ln(\mathcal{N}(x_n|\mu_k, \Sigma_k)] \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})^{(m)}\left[\ln(\pi_k) - \frac{\ln(2\pi)}{2} - \frac{\ln(|\Sigma|)}{2} - \frac{(x_i-\mu_k)^T]\Sigma_k^{-1}(x_i-\mu_k)}{2}\right] \end{aligned} \tag{1}$$

where $\gamma(z_{nk})^{(m)} = P_{\lambda_m}(Z_i = k|X_i = x_i)$, and this is already a given value, after the E step by the old parameter $\lambda^{(m)}$.

To get the optimum parameter μ_k, Σ_k, we will take the partial derivative with respect to each one and make it equal to zero.

$$\begin{aligned} \frac{\partial Q}{\partial \mu_k} &= \sum_{n=1}^{N} \gamma(z_{nk})^{(m)}\left[\frac{-\nabla_\mu[(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)]}{2}\right] \\ &= \sum_{n=1}^{N} \gamma(z_{nk})^{(m)}\Sigma_k^{-1}(x_i-\mu_k) \end{aligned} \tag{2}$$

$$\begin{aligned} \sum_{n=1}^{N} \gamma(z_{nk})^{(m)}\Sigma_k^{-1}(x_i-\mu_k) &= 0 \\ \sum_{n=1}^{N} \gamma(z_{nk})^{(m)}(x_i-\mu_k) &= 0 \\ \sum_{n=1}^{N} \gamma(z_{nk})^{(m)}x_i &= \mu_k \sum_{n=1}^{N} \gamma(z_{nk})^{(m)} \\ \mu_k &= \frac{\sum_{n=1}^{N} \gamma(z_{nk})^{(m)}x_i}{\sum_{n=1}^{N} \gamma(z_{nk})^{(m)}} \end{aligned} \tag{3}$$

$$\begin{aligned} \frac{\partial Q}{\partial \Sigma_k} &= \sum_{n=1}^{N} \gamma(z_{nk})^{(m)}\left[-\frac{\nabla_{\Sigma_k}[\ln(|\Sigma|)]}{2} - \frac{\nabla_{\Sigma_k}[(x_i-\mu_k)^T]\Sigma_k^{-1}(x_i-\mu_k)]}{2}\right] \\ &= \sum_{n=1}^{N} \gamma(z_{nk})^{(m)}\left[-\frac{\Sigma_k^{-1}}{2} - \frac{-\Sigma_k^{-2}(x_i-\mu_k)^T(x_i-\mu_k)}{2}\right] \end{aligned} \tag{4}$$

$$\sum_{n=1}^{N} \gamma(z_{nk})^{(m)} \left[-\frac{\Sigma_k^{-1}}{2} - \frac{-\Sigma_k^{-2}(x_i - \mu_k)^T(x_i - \mu_k)}{2} \right] = 0$$

$$\sum_{n=1}^{N} \gamma(z_{nk})^{(m)} [-\Sigma_k^{-1} + \Sigma_k^{-2}(x_i - \mu_k)^T(x_i - \mu_k)] = 0$$

$$\sum_{n=1}^{N} \gamma(z_{nk})^{(m)} [-\Sigma_k + (x_i - \mu_k)^T(x_i - \mu_k)] = 0$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})^{(m)} (x_i - \mu_k)^T (x_i - \mu_k)}{\sum_{n=1}^{N} \gamma(z_{nk})^{(m)}} \tag{5}$$

Therefore, during the M-step the reestimation parameters μ_{m+1} and Σ_{m+1} will be:

$$\mu_{k_{m+1}} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})^{(m)} x_i}{\sum_{n=1}^{N} \gamma(z_{nk})^{(m)}} \tag{6}$$

$$\Sigma_{k_{m+1}} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})^{(m)} (x_i - \mu_{k_{m+1}})^T (x_i - \mu_{k_{m+1}})}{\sum_{n=1}^{N} \gamma(z_{nk})^{(m)}} \tag{7}$$

2. Figure 1 shows the result of plotting the generated data using bivariate Gaussian mixture model.

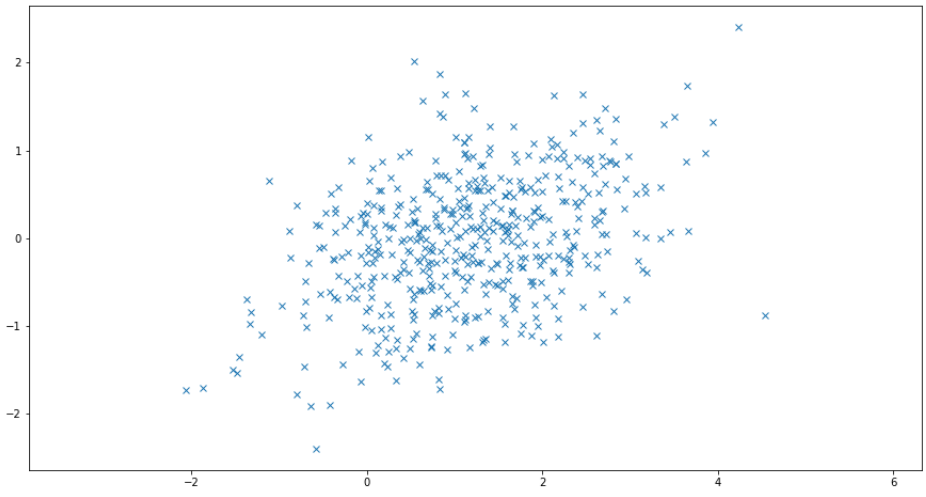

Figure 1: Plot for generated data

3. Figure 2 shows the result of plotting the Unistroke data set (letter A). It has a special circular form which could be beneficial during analysis and the interpretation of the data and modelling it.

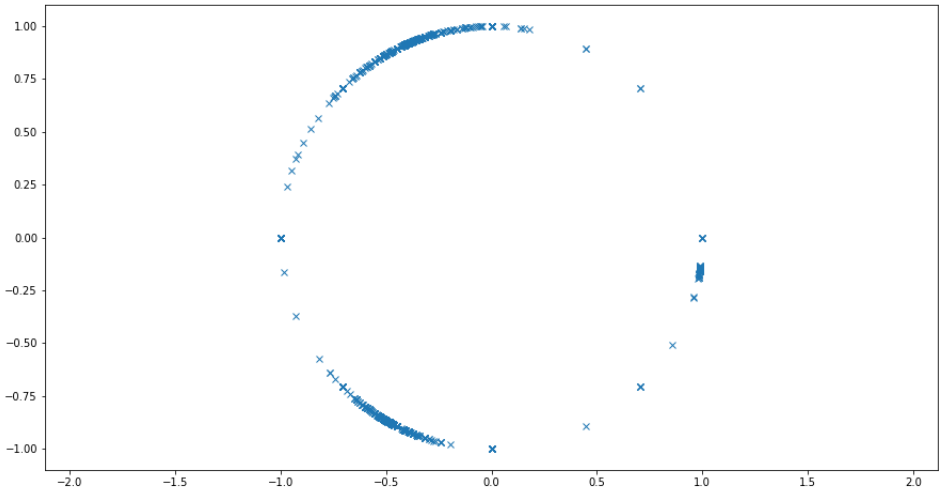

Figure 2: Plot for Unistroke data of letter 'A'

4. The 2-states Gaussian model could be appropriate for letter A, because to draw the letter A, we have in general two actions by the Unistroke : The strokes go to the up and the strokes go to the down. Thus for every action we could be modelled by a gaussian.

1.1.2 Data analysis: Gaussian model

1. After fitting the Gaussian mixture on the data using scikit-learn (SKlearn.mixture), we got for the estimated parameters the following results:

   - The means for class 1: $\mu_{1_x} = -0.48162 \quad , \quad \mu_{1_y} = -0.838$
   - The means for class 2: $\mu_{2_x} = -0.20488 \quad , \quad \mu_{2_y} = 0.76291$
   - The covariance for class 1:

   $$\Sigma_1 = \begin{bmatrix} 0.03313 & -0.02664 \\ -0.02664 & 0.03266 \end{bmatrix}$$

   - The covariance for class 2:

   $$\Sigma_2 = \begin{bmatrix} 0.23341 & -0.13776 \\ -0.13776 & 0.14257 \end{bmatrix}$$

   We can notice from this result that the two gaussian models have very close mean on the variable x. This means that the two gaussian have no big difference if we tried to project the data on the x variable only we will get very close distribution that could be not relevant. On the other side, there's is a relevant difference in the means for the y variable. This could give an insight of the importance of the variable y.

2. Figure 3 shows the result of plotting the Unistroke data set (letter A). But this time labelled with the inference of the Gaussian mixture. It's obvious again that the labelling decision is highly affected by the y position more than the x position.

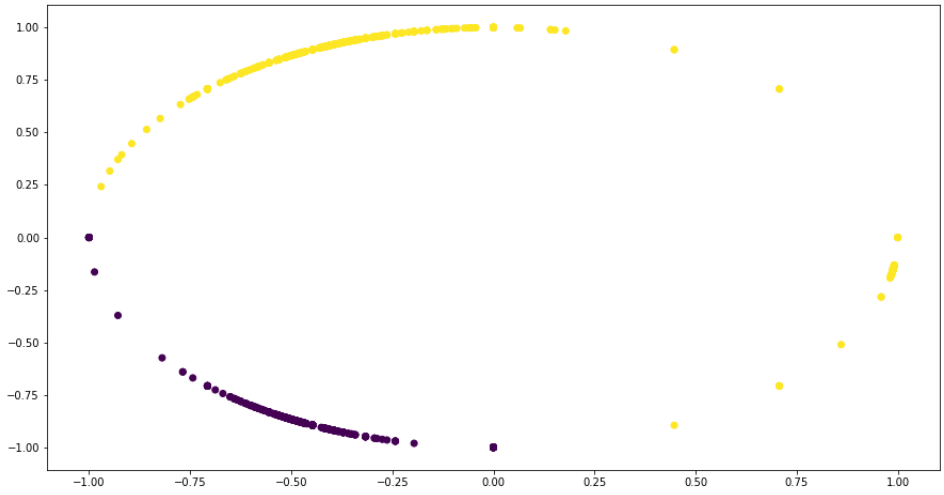

Figure 3: Unistroke data labelled

3. Marginal Histograms:

   (a) Figure 4 illustrates what we interpreted in the previous question (The mixture models have very close means), that the mixture model doesn't fit very well on the x variable. This is normal since the histogram of x show that the data on x can be encapsulated with one single distribution, while the mixture model is composed of two gaussian model. Thus, this variable doesn't contribute much in the inference.
   On the contrary, figure 5 shows that the variable y separate the data very well in two different area in the histogram and each of the two gaussian model are isolated in the same way. This variable could contribute more in prediction.

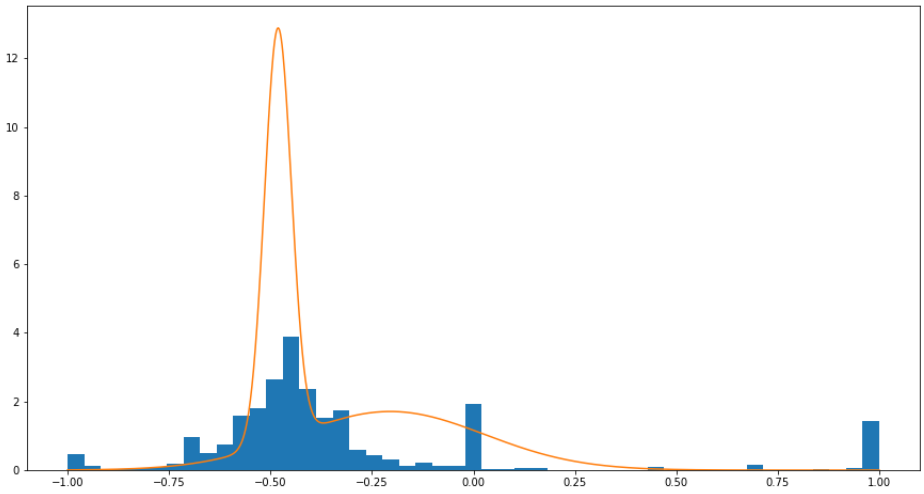

Figure 4: Marginal histogram with the projected mixture model on the variable x

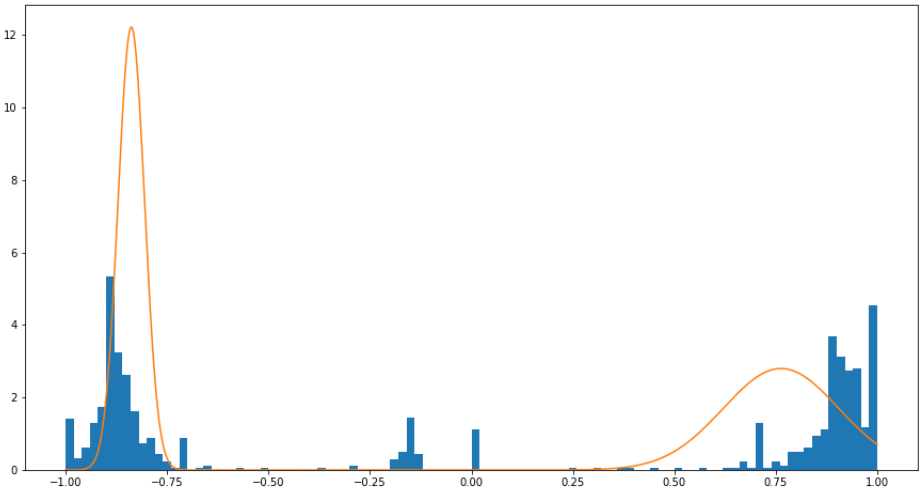

Figure 5: Marginal histogram with the projected mixture model on the variable y

(b) Figure 6 illustrates again that after we labelled the data, the two clusters are fused together and we can get information on the separation of the two clusters, because the separation here is so week. Going to figure 7, we can see obviously that the two clusters are well separated and the two mixture models fit very well when they are projected on the y.

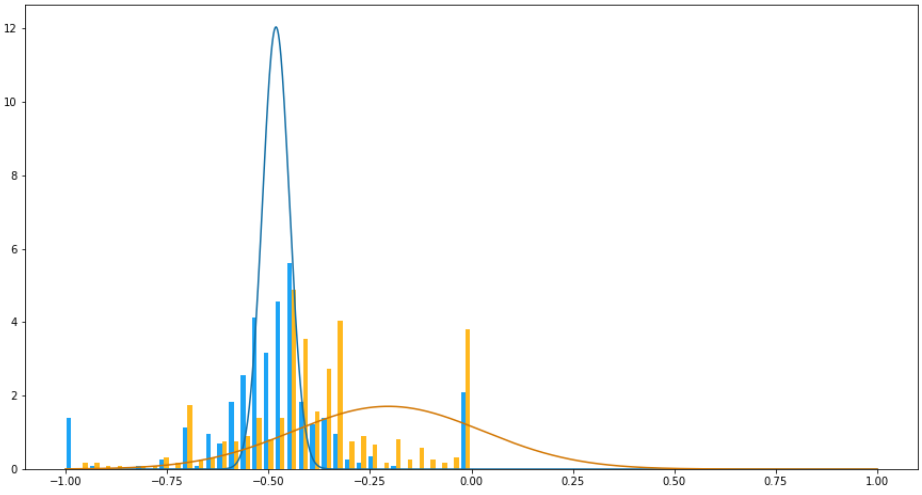

Figure 6: Marginal histogram with the projected mixture model on the variable x with the predicted labels

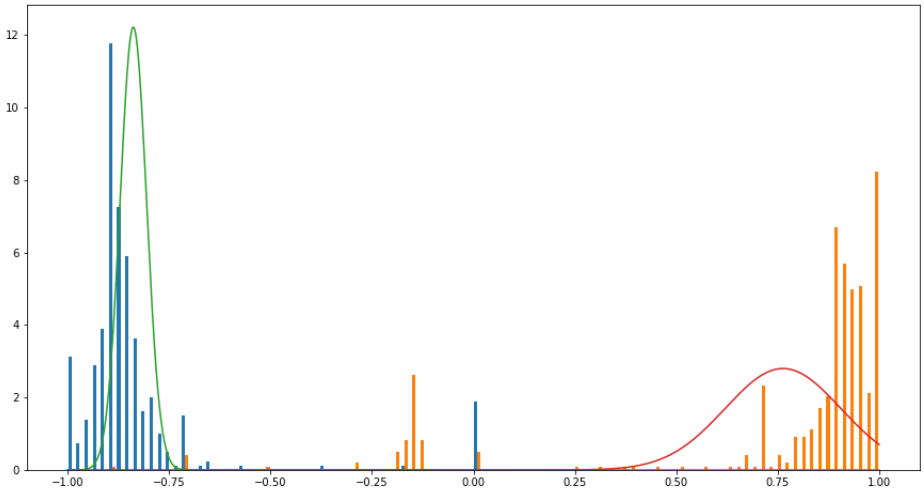

Figure 7: Marginal histogram with the projected mixture model on the variable y with the predicted labels

4. After analyzing the marginal histograms and compare them with the mixture model, we can suggest that a univariate gaussian mixture will be more informative and good give better results on this data. This is normal fact because we can see that the y variable contribute more in differentiating between the data. Also, we have decided in the first section that the main two classes are the up strokes and the down strokes. Thus, the classes' choice was mainly based on the vertical feature of the stroke. All this arguments enforce the idea of using univariate gaussian model.

5. Figure 8 illustrates one important issue that one from the two clusters are dominated unfairly. This means that the Gaussian distributions are not will fitted, because there are some points that are misclassified or the cluster 1 has more effect on the data that the other cluster (if we strongly consider that the variable y has the most contribution). One thing we can notice that this disambiguity appear on the extremes of the two arcs (we assume that the clusters will have the shape of arcs). This means that the gaussian model is limited in front of this kind of shape of the data.

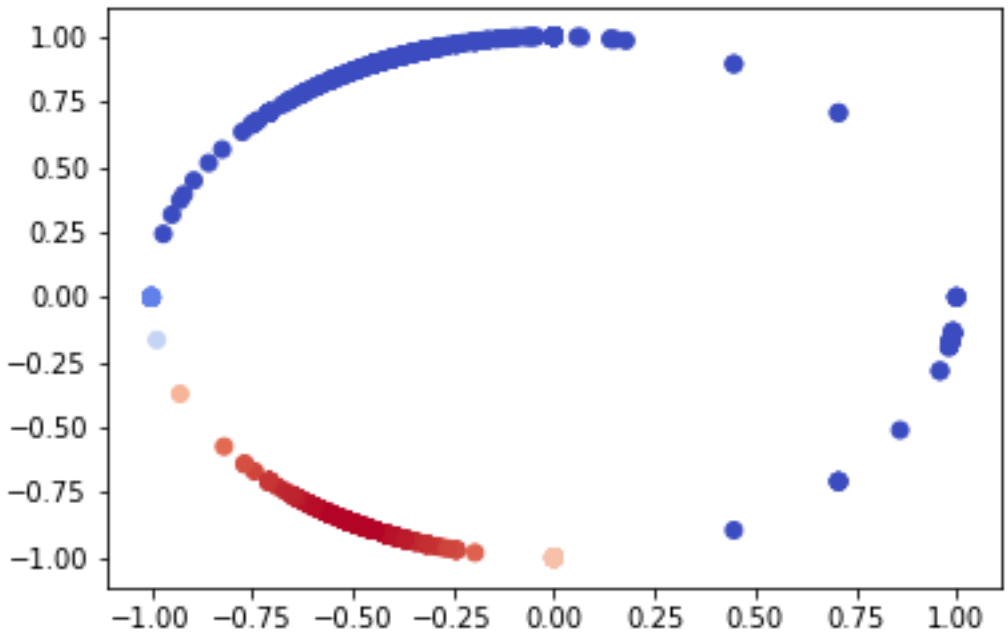

Figure 8: Plot for generated data

1.2 Mandatory additional questions

1. Figure 9 shows the histogram of he angular data. It illustrates a high correlation with the histogram of the y variable. This correlation support the argument that the most important information of the data that could help to separate the clusters can be encapsulated within one variable.

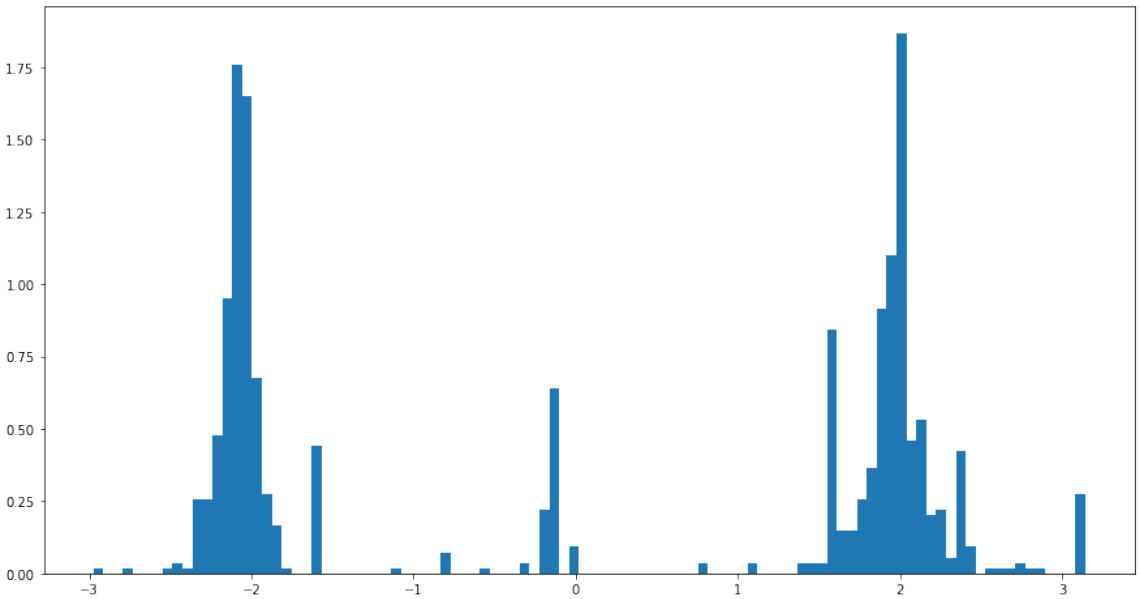

Figure 9: Histogram of the angular data

2. Von Mises distribution is continuous distribution that is given by:

$$p(x) = \frac{exp(\kappa cos(x-a))}{2\pi I_0(\kappa)}$$

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

where $\theta \in [0, 2\pi]$, I_0 is the modified Bessel function of order zero. The parameters of Von Mises is the circular mean α and κ is the measure of the concentration. For a mixture of K independent Von Mises distribution, we have:

$$p(x) = \sum_{k=1}^{K} \pi_k \nu(x|\mu_k, \kappa_k)$$

3. From Von Mises distributions definition and the nature form of the data, we estimate that von Mises distributions would be more adequate than Gaussian mixtures, because the distribution of the data is circular, and the most influential component of the data is the angle which is a circular variable $\in [0, 2\pi]$). von Mises deals better with circular data since it's a circular version of the Gaussian distribution.

4. According to Calderara et al [1] the Steps of the EM algorithms for mixed von Mises distribution are as follow:

   (a) The E-step:

$$\begin{aligned} P_{\lambda^{(m)}}(z_n = k|x_n) &= \frac{\pi_k^{(m)} \nu(x_n|\mu_k^{(m)}, \kappa_k^{(m)})}{\sum_i \pi_i^{(m)} \nu(x_n|\mu_i^{(m)}, \kappa_i^{(m)})} \\ &= \gamma_{nk}^{(m)} \end{aligned} \tag{8}$$

   (b) While the M-Step is:

$$\pi_k^{(m+1)} = \frac{1}{N}\sum_{n=1}^{N} \gamma_{nk}^{(m)} \tag{9}$$

$$\alpha_k^{(m+1)} = arctan\left(\frac{\sum_{n=1}^{N} \gamma_{nk}^{(m)} sin(x_i)}{\sum_{n=1}^{N} \gamma_{nk}^{(m)} cos(x_i)}\right) \tag{10}$$

$$A(\kappa_k^{(m+1)}) = \frac{\sum_{n=1}^{N} \gamma_{nk}^{(m)} sin(x_i - \alpha_k^{(m)})}{\sum_{n=1}^{N} \gamma_{nk}^{(m)}} \tag{11}$$

5. We used the model VonMisesFisherMixture from the library is spherecluster. After we used the library to run the new mixture model on the data. We can see in figure 9 that von Mises appreciated more than circular behaviour of the data. it had manipulated the data in a way that like converting the data in a variable that can deal with in a circular way. Also the colour map in figure 12 shows that there exists this time a fairness in the extremity of the clusters without being biased to a certain cluster.

   A key aspect of von Mises is illustrated in figure 11, we can see in the right extreme of the figure that the data has been assigned to the blue cluster and not the orange one. This makes sense, because has we said the data is circular, thus this extremes should be assigned circulary to the blue cluster and not to the red cluster. Again von Mises shows how it figures the circular aspect of the data.

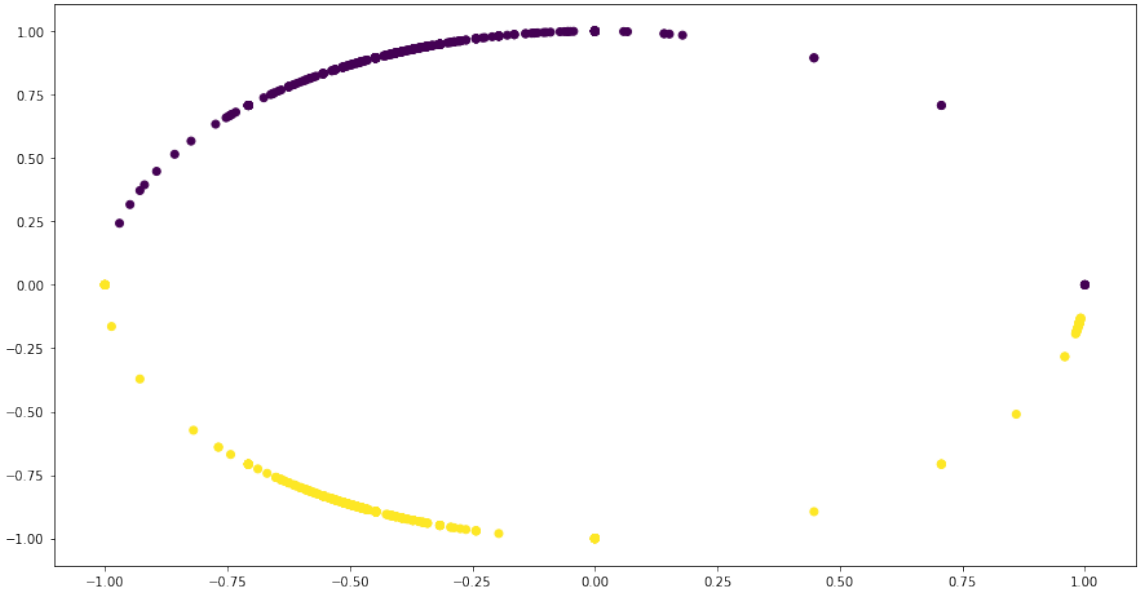

Figure 10: Plot for generated data

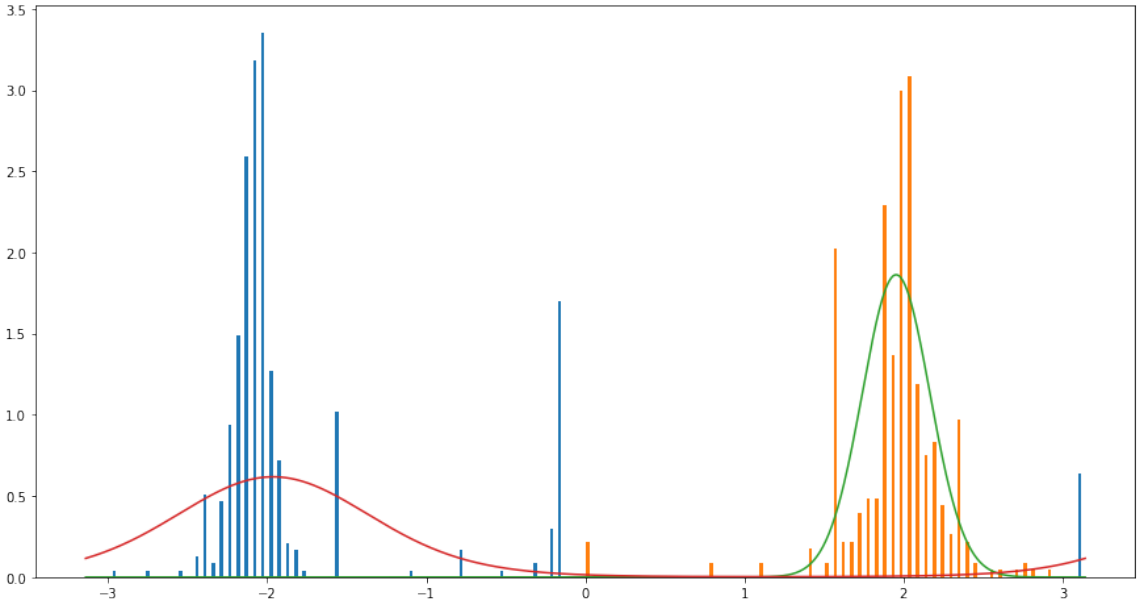

Figure 11: Plot for generated data

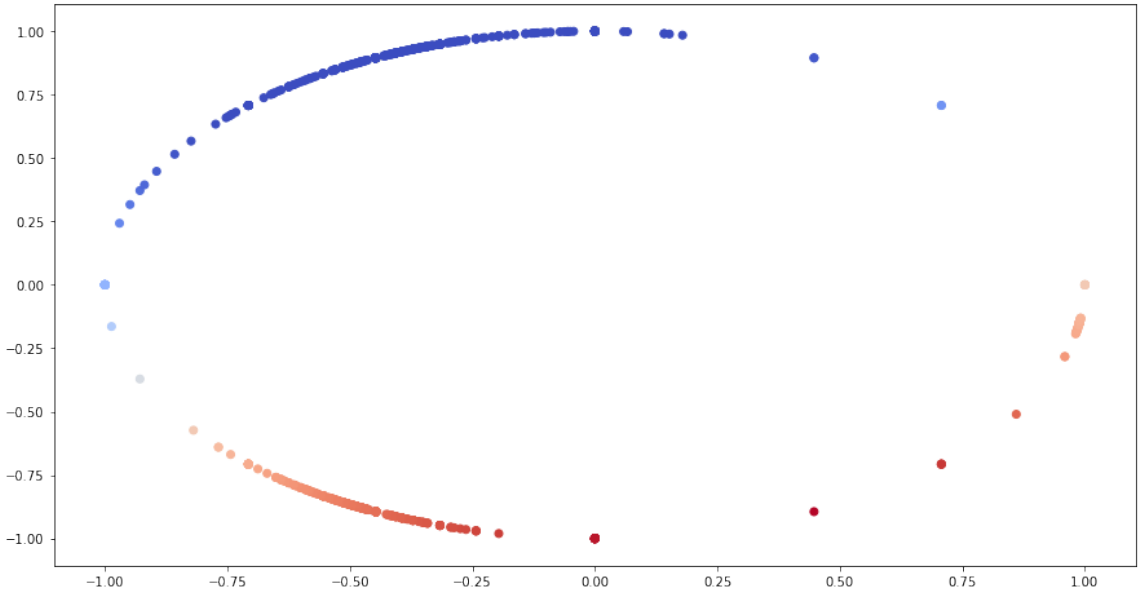

Figure 12: Plot for generated data

References

[1] S. Calderara et al. (2011) *Mixtures of von Mises Distributions for People Trajectory Shape Analysis.*

[2] C. Bishop. (2006) *Pattern Recognition and Machine Learning.*