COVID-19 Dataset Clustering based on K-Means and EM Algorithms

Youssef Boutazart[1], Hassan Satori[2], Anselme R. Affane M.[3], Mohamed Hamidi[4], Khaled Satori[5]
Computer Science Laboratory, Signals Automation and Cognition (LISAC)
Department of Mathematics and Computer Science, Faculty of Sciences, Dhar Mahraz,
Sidi Mohamed Ben Abdallah University Fez, Morocco[1,2,3,5]
Team of Modeling and Scientific Computing (LaMAO), FPN, UMP[4]

***Abstract*—In this paper, a COVID-19 dataset is analyzed using a combination of K-Means and Expectation-Maximization (EM) algorithms to cluster the data. The purpose of this method is to gain insight into and interpret the various components of the data. The study focuses on tracking the evolution of confirmed, death, and recovered cases from March to October 2020, using a two-dimensional dataset approach. K-Means is used to group the data into three categories: "Confirmed-Recovered", "Confirmed-Death", and "Recovered-Death", and each category is modeled using a bivariate Gaussian density. The optimal value for k, which represents the number of groups, is determined using the Elbow method. The results indicate that the clusters generated by K-Means provide limited information, whereas the EM algorithm reveals the correlation between "Confirmed-Recovered", "Confirmed-Death", and "Recovered-Death". The advantages of using the EM algorithm include stability in computation and improved clustering through the Gaussian Mixture Model (GMM).**

***Keywords*—COVID-19; clustering; k-means; EM algorithm; GMM**

I. Introduction

Cluster analysis involves organizing data into meaningful and valid groups [1], which are homogeneous and similar. This technique involves classifying each data point into a specific set using clustering algorithms [2], [3]. A method proposed by the authors in [4] determines the optimal number of clusters, k, which represents the inherent significant clustering structures of the dataset. K-Means and Expectation Maximization (EM) algorithms are commonly used for clustering [5]. The proposed EM algorithm, initially designed for finding maximum likelihood parameters of a statistical model, has been applied to various domains such as speech recognition [6], interactive systems [7], etc.

On the other hand, the researchers in [8] have proposed a new epidemiological mathematical model for the spread of the COVID-19 disease with a special focus on the transmissibility of individuals with severe symptoms. Recently an important report using C++ can be used to "track" the daily evolution of new confirmed cases of the COVID-19 epidemic [9]. Rizvi et al. [10] have described K-Means clustering of 79 countries has been performed for COVID-19 confirmed cases and COVID-19 death cases based on 18 feature variables.

This study presents a fresh approach to analyzing the COVID-19 dataset using clustering techniques. Specifically, we apply a standard version of K-Means and EM algorithms based on GMM to partition the local COVID-19 Moroccan dataset into three sets: "Confirmed-Recovered", "Confirmed-Death", and "Recovered-Death", with varying cluster numbers. Our primary objective is to identify the optimal classification for each data cluster.

This paper is organized as follows: Section II gives the Literature Review. The K-Means and EM algorithms is introduced in Section III. The COVID-19 pandemic is presented in Section IV. Section V describes the COVID-19 dataset. Section VI exposes the results and discussion. Finally, in Section VII we conclude this work and gives perspectives.

II. Literature Review

The COVID-19 pandemic has led to an increase in the use of data mining and machine learning techniques to understand and analyze the spread of the virus. Clustering is a popular technique used to group similar data points together. K-Means, EM Algorithm and GMM are three commonly used clustering algorithms in machine learning. Several clustering methods have been developed with the objective to find the correct number of clusters [11], [12], [13], [14], [15]. In [16] the authors focus on utilizing Probabilistic Graphical Models for detecting COVID-19, resulting in excellent detection of the disease. One potential use of the EM algorithm is to estimate the parameters of a mixture model in cases where the data is incomplete. This technique is sometimes referred to as finding the parameters of Gaussian mixture densities [17], [18]. Eva and Dharmende [19] conducted a comparison between K-Means and GMM to assess their effectiveness in representing clusters of heterogeneous resource usage in Cloud workloads. Their experiments, which utilized Google cluster trace and business critical workloads by Bitbrains, revealed that K-Means provided a more generalized representation, whereas GMM resulted in better clustering with clearly defined usage boundaries. Despite Gaussian Mixture Model's longer computation time compared to K-Means, it is preferred for more detailed workload analysis and characterization.

Appiah et al. [20] proposed a study that utilizes the EM algorithm, which is initialized by a semi-supervised K-Means clustering approach based on geodesic distance classification of crime dataset. The aim is to track changes in cluster centroids (mean), shape and orientation, volume, and predictive trends of criminal activities. In this approach, the cluster assignment obtained from K-means is assumed as the distribution of GMM. The model-based clustering algorithm is then used to estimate the parameters of the mixed model while maintaining the probabilistic assignment and multivariate nature of the

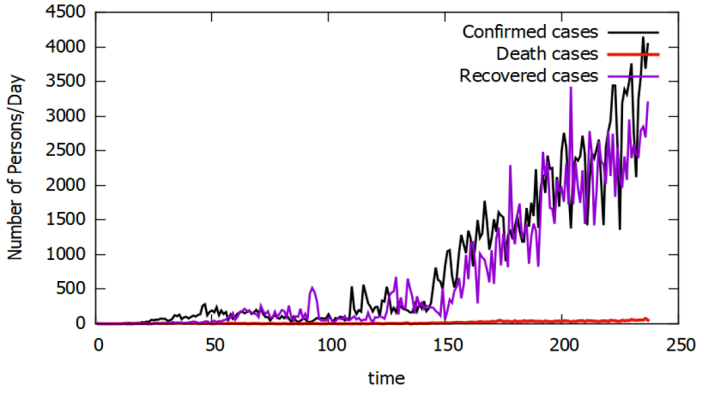

Fig. 1. COVID-19 confirmed-recovered and death cases in Morocco.

model. On the other hand, Zarikas et al. [21] developed a clustering algorithm designed specifically for grouping countries based on COVID-19 active cases, active cases per population, and per area. The results showed that countries facing similar impacts of COVID-19 also shared similar societal, economic, and other factors.

Aungkulanon et al. [22] clustered different regions of Thailand based on financial conditions and mortality differentials, revealing super-locale that are mainly urban and have a low all-cause normalized mortality proportion but a high colorectal disease-specific death rate. The study also found that deaths caused by liver cancer, diabetes, and renal diseases are common in low economic super-regions. Malav et al. [23] conducted a study to predict coronary heart disease using K-means and artificial neural networks. The combined approach led to a system with a very high accuracy rate. Another work by Singh et al. [24] used clustering and classification techniques to forecast heart diseases with high accuracy.

Isikhan et al. [25] clustered countries based on causes of deaths, health profiles, and risk factors using unsupervised K-means. The study analyzed clusters based on some financial and socio-demographic indicators and found that climate and ethnicity were more significant factors for clustering than socio-economic factors. These studies demonstrate the importance of COVID-19 dataset clustering in identifying patterns and trends associated with the virus, which can aid in developing effective strategies to combat its spread.

TABLE I. CONFIRMED-RECOVERED-DEATH COVID-19 DATASET

Day	1	2	20	30	40	50	60	90	...	237
Confirmed	1	0	22	63	74	191	102	69	...	4045
Recovered	0	0	0	10	13	23	56	141	...	3197
Death	0	0	0	3	10	2	2	2	...	50

III. K-MEANS AND EM ALGORITHMS

Given a set of observations $Y = (Y_1, ..., Y_N)$, independent and identically distributed (i.i.d) where each observation $Y_t = (y_{t1}, ..., y_{tj}, ..., y_{td}) \in R^d$ is a d-dimensional real vector. The objectives of K-Means and EM are to partition N observations into G clusters [26].

A. K-Means Algorithm

In this part, the objective is to find values for z_{tk} and μ_k the mean so as to minimize D. Let $\Phi = \mu = \{\mu_1, ..., \mu_G\}$ be the set represents the mean of each cluster c_k, where $C_k \in \{C_1, ..., C_G\}$ the set of G clusters, and let $Z = (z_1, z_2, ..., z_N)$ the set of binary indicator variables.

$$D(\Phi, Z) = \sum_{t=1}^{N} \sum_{k=1}^{G} z_{tk} \|Y_t - \mu_k\|^2 \tag{1}$$

Where $z_{tk} = 1$ when Y_t is a member of C_k , otherwise $z_{tk} = 0$. Or more exactly $\arg\min_k D(\Phi, Z)$. when D achieved minimal value, sum of $\|Y_t - \mu_k\|^2$ is minimal [27].

$$d(Y_t, \mu_k) = \sqrt{\sum_{j=1}^{d} (y_{tj} - \mu_{kj})^2} \tag{2}$$

by Euclidean distance. We can do this through an iterative procedure in which each iteration involves two successive steps corresponding to successive optimizations with respect to z_{tk} and μ_k. We initialize the class centers $\left\{\mu_1^{(0)}, ..., \mu_G^{(0)}\right\}$ for the $\{C_1, ..., C_G\}$ set of clusters; by some initial values called seed-points, using methodically sampling.

Step 1:

We minimize D and we update z_{tk}, keeping the μ_k fixed.

Step 2:

We minimize D and we update μ_k, keeping the z_{tk} fixed.

$$\mu_k^{(m+1)} = \frac{\sum_{t=1}^{N} z_{tk}^{(m)} . Y_t}{\sum_{t=1}^{N} z_{tk}^{(m)}} \tag{3}$$

(m) being the current iteration. This two-stage optimization is then repeated until convergence.

B. Expectation Maximization Algorithm

In this work, EM algorithm is used to complete the missing COVID-19 data. We introduce the latent variable Z. Y_t can describe the mix "Confirmed cases-Recovered cases". The same study for the mixture of confirmed cases - death cases and recovered cases - death cases. We will assume that the observations Y_t are i.i.d and the observations from different clusters have correlated Bivariate Gaussian Density. If data t belongs to cluster C_k (denoted by $t \in C_k$) then:

$$Y_t \setminus t \in C_k \sim f(y_t/\mu_k, \Sigma_k) \tag{4}$$

$$f(y_t/\mu_k, \Sigma_k) = \frac{1}{2\Pi^{\frac{d}{2}} \sqrt{|\Sigma_k|}}$$
$$\exp \frac{-1}{2}[(y_t - \mu_k)^t \Sigma_k^{-1} (y_t - \mu_k)] \tag{5}$$

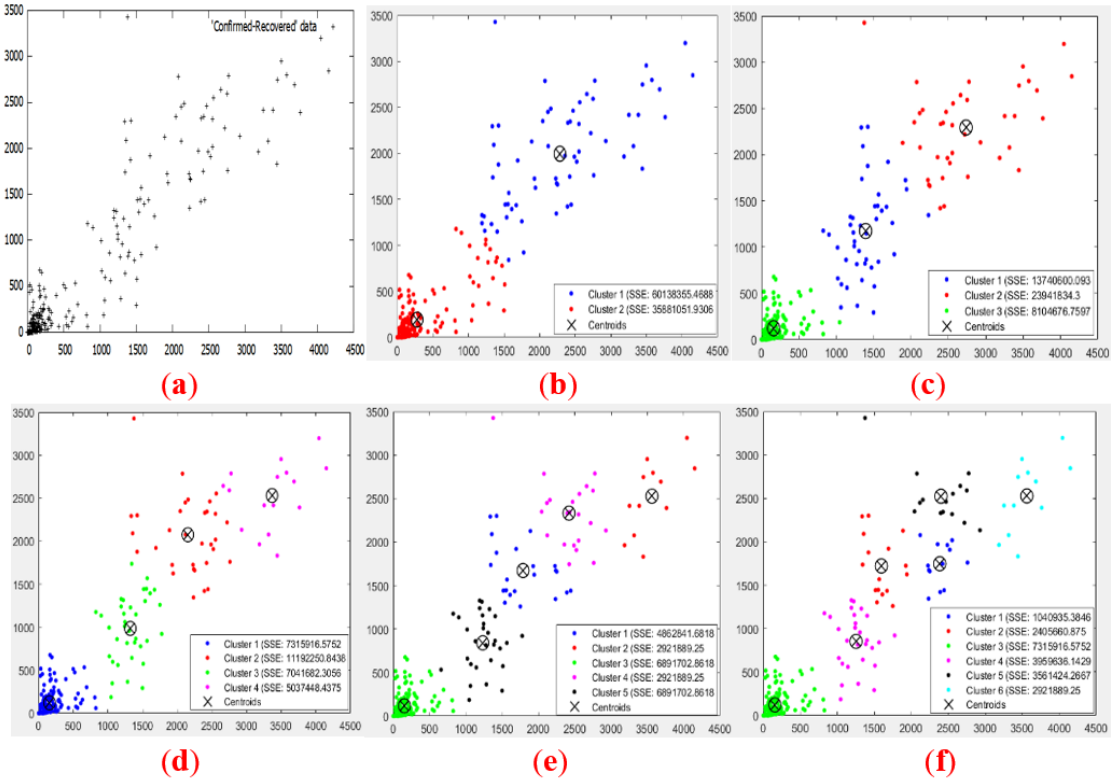

Fig. 2. (a) Two-dimensional input "Confirmed-Recovered" data; with no clustering. (b), (c), (d), (e) and (f) K-means partitions, respectively with k = 2, k = 3, k = 4, k = 5 and k = 6. The centroids are marked with a cross.

μ_k and Σ_k denote the mean vector and covariance matrix. Assign a data point to a nearest cluster, with calculate the following likelihood [28].:

$$\gamma(z_{tk}) = \mathbf{E}(z_{tk}) = \mathbf{P}(z_{tk} = 1/y) = \frac{\mathbf{P}(z_{tk} = 1)\mathbf{P}(y_t/z)}{\mathbf{P}(y)} \tag{6}$$

Where $\mathbf{P}(z_{tk} = 1/y)$ is a posterior probability of $y_t \in C_k$ the k^{th} – classes and z_t correspond to the Gaussian identity which generated an entry y_t.

Step 1 (Expectation): Given the current estimates, $[\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}, \boldsymbol{\Pi_k}]$

$$\boldsymbol{\gamma(z_{tk})} = \frac{\boldsymbol{\Pi_k f_k(y_t/\mu_k, \Sigma_k)}}{\sum_{j=1}^{G} \boldsymbol{\Pi_j f_j(y_t/\mu_j, \Sigma_j)}} \tag{7}$$

Step 2 (Maximization): Compute the parameters that maximize the likelihood of the data set $\mathbf{P}(\boldsymbol{Y/\mu_k, \Sigma_k, \Pi_k, z_{tk}})$ which is the probability of all of the data under the GMM. Find the probability $\mathbf{P}(\boldsymbol{Y})$ that generated the COVID-19 dataset. Maximizing this with respect to each of the parameters can be done in closed form:

$$\boldsymbol{\Pi_k^{new\ ite}} = \frac{\sum_{t=1}^{N} \boldsymbol{\gamma(z_{tk})}}{\boldsymbol{N}} \tag{8}$$

$$\boldsymbol{\mu_k^{new\ ite}} = \frac{\sum_{t=1}^{N} \boldsymbol{\gamma(z_{tk})} \boldsymbol{y_t}}{\sum_{t=1}^{N} \boldsymbol{\gamma(z_{tk})}} \tag{9}$$

$$\boldsymbol{\Sigma_k^{new\ ite}} = \frac{\sum_{t=1}^{N} \boldsymbol{\gamma(z_{tk})((y_t - \mu_k^{new\ ite}) \otimes (y_t - \mu_k^{new\ \ ite})^t}}{\sum_{t=1}^{N} \boldsymbol{\gamma(z_{tk})}} \tag{10}$$

1) Re-estimation of mixed weights: To find the parameter we using a Lagrange multipliers [29] with constraint $\sum_{i=1}^{G} \boldsymbol{\Pi_i} = \mathbf{1}$ and maximizing the following quantity:

$$\boldsymbol{L(l(\Phi), \lambda) = l(\Phi) + \lambda(\sum_{k=1}^{G} \Pi_k - 1)} \tag{11}$$

$$\boldsymbol{Where \frac{\partial L(l(\Phi), \lambda)}{\partial \Pi_k} = 0}$$

Then we obtain

$$\sum_{t=1}^{N} \frac{\boldsymbol{\Pi_k f(y_t/\mu_k, \Sigma_k)]}}{\sum_j [\boldsymbol{\Pi_j f(y_t/\mu_j, \Sigma_j)]}} + \boldsymbol{\lambda \Pi_k} = \mathbf{0}$$

and we have new estimation for $\boldsymbol{\Pi_k}$ (see Eq. 8).

2) Re-estimation of the means vectors: We assume $\boldsymbol{\gamma(z_{tk})}$ fixed. We derive this equation with respect to the means $\boldsymbol{\mu_k}$ at zero, we obtain:

$$\boldsymbol{l(\Phi)} = \sum_{t=1}^{N} \ln[\sum_{k=1}^{G} \frac{\boldsymbol{\Pi_k}}{\boldsymbol{2\Pi\sqrt{|\Sigma_k|}}} \exp \frac{-1}{2}[(y_t - \mu_k)^t \Sigma_k^{-1}(y_t - \mu_k)]] \tag{12}$$

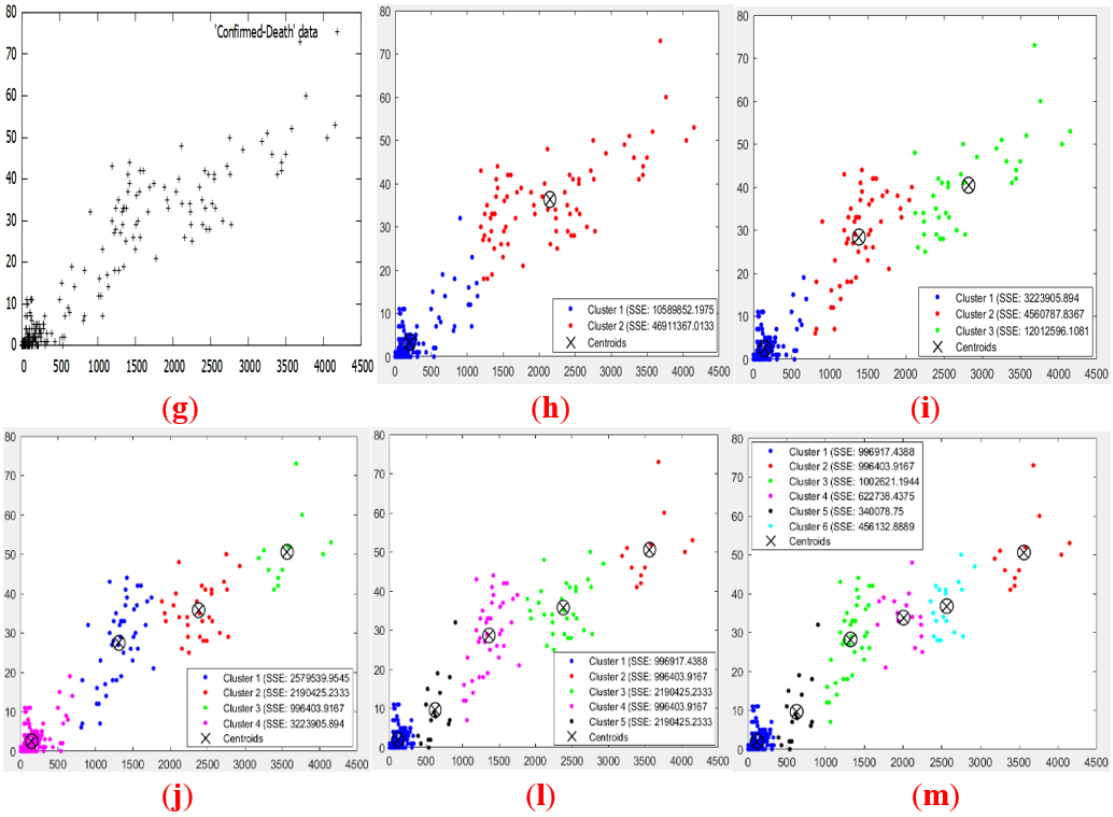

Fig. 3. (g) Two-dimensional input "Confirmed-Death" data; with no clustering. (h), (i), (j), (l) and (m) K-means partitions, respectively with k = 2, k = 3, k = 4, k = 5 and k = 6. The centroids are marked with a cross.

$\boldsymbol{Where}\ \frac{\partial l(\Phi)}{\partial \mu_k} = \mathbf{0}$, Then we find:

$$\sum_{t=1}^{N} \frac{\boldsymbol{\Pi_k f(y_t/\mu_k, \Sigma_k)}]}{\sum_j [\boldsymbol{\Pi_j f(y_t/\mu_j, \Sigma_j)}]} \boldsymbol{\Sigma_k^{-1}(y_t - \mu_k)} = \mathbf{0}$$

The new $\boldsymbol{\mu_k}$ is gives in (Eq. 9)

3) Re-estimation of the covariance matrix: In the same way we derive $\boldsymbol{l(\Phi)}$ with respect to $\boldsymbol{\Sigma_k}$ Where $\frac{\partial l(\Phi)}{\partial \Sigma_k} = \mathbf{0}$, then we obtain new values of covariance matrix (see Eq. 10).

IV. COVID-19 Pandemic

Later in 2019, in the city of Wuhan, in China, a new discovered version of coronavirus was detected as the principal reason for a strange aspect of pneumonia cluster. Local scientists react by isolating the SARS-CoV-2 into a patient on the earlier of January 2020, which led to the genome sequence of the SARS-CoV-2 [30].

According to the authors of sequencing, phylogenetic analysis this genome has made it possible to establish that the initial host of this virus is an animal sold on the market in Wuhan. Several studies have suggested bats could be at the origin of SARS-CoV-2 [31]. The virus was referred to as 2019-nCoV before the COVID-19 name. It is defined as a severe acute respiratory syndrome coronavirus number 2 (SARS-CoV-2). The WHO declares that the first the infection as a pandemic on March 11, 2020. It rapidly spread, followed by an increase in the number of infected cases around the globe. To this disease of August 16, 2020, the world has had 21.294.845 total confirmed cases, and 761.779 total deaths cases [32].

V. COVID- 19 Dataset Description

In the present study, we use public data from the COVID-19 outbreak in Morocco to estimate the evolution of this epidemic. The data is received through the official website created by the Moroccan Ministry of Health. For this disease, Morocco has had 194461 total confirmed cases, said the Director of epidemiology and disease control at the Ministry of Health as of October 24,2020 the total number of deaths has increased to 3255; and 160372 total cured cases (see Fig. 1) [33].

The training dataset is composed of the real COVID-19 cases daily collected Confirmed, Recovered, and Death patterns. The clustering is done with two-dimensional dataset "Confirmed – Recovered", "Confirmed – Death" and "Recovered – Death" features of 237 samples. The Table I below shows a part of the complete data.

The recording of the $\mathbf{237}^{th}$ COVID-19 cases are store in the Table I. Each feature is a combination of two parameters, the Confirmed recorder and the Death cases, then the Confirmed recorder and the Recovered cases and the Recovered recorder and the Death cases, respectively.

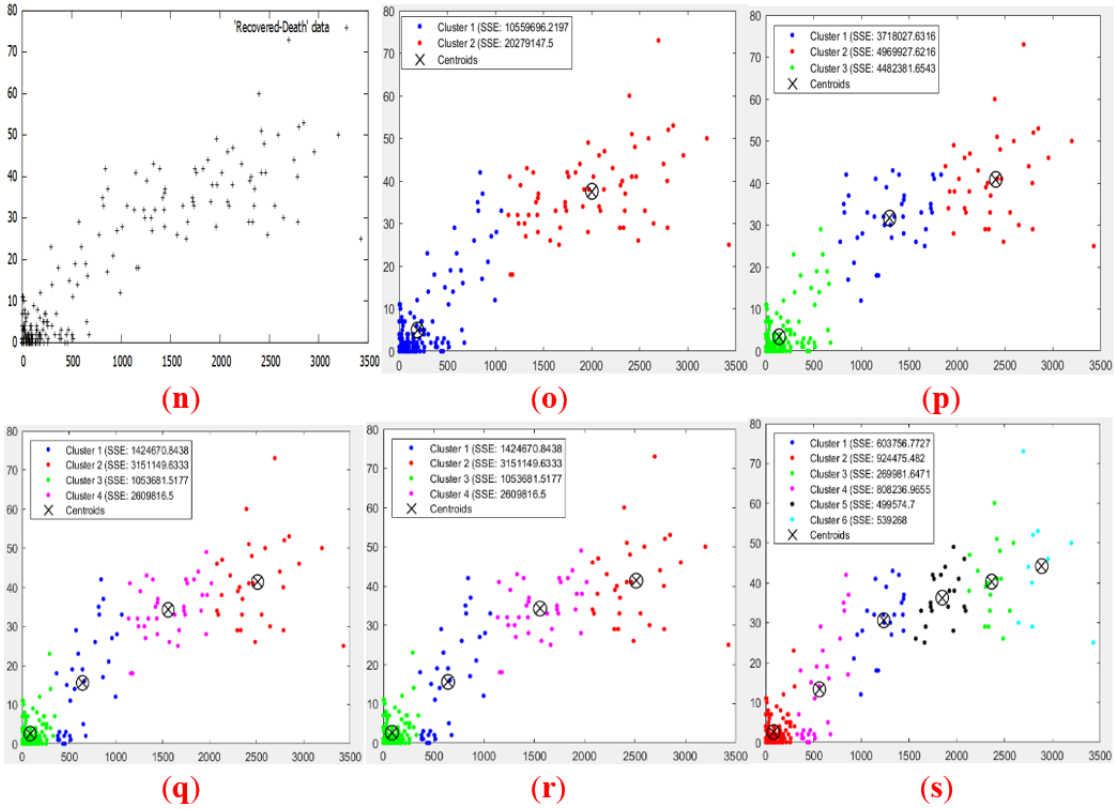

Fig. 4. (n) Two-dimensional input "Recovered - Death" data; with no clustering. Data to illustrate the K-means procedure. (o), (p), (q), (r) and (s) K-Means partitions, respectively with k = 2, k = 3, k = 4, k = 5 and k = 6. The red dots represent the centroid of each cluster.

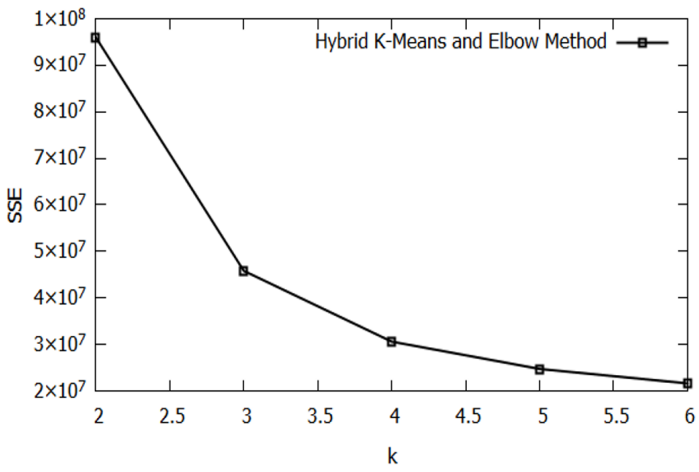

Fig. 5. Graph of sum square error depending on the number k for 'Confirmed-Recovered' data.

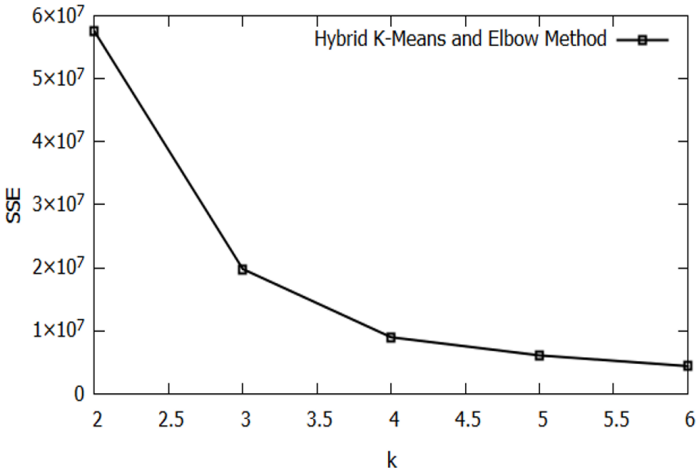

Fig. 6. Graph of sum square error depending on the number k for 'Confirmed-Death' data.

VI. Results and Discussion

In this work, we selected k-points as the primary group ranks as the points are calculated in order. The total number of initial points c_k is $237/k$ for all groups, then we define the initial centroids μ_k. The test data consists of three groups, "Confirmed – Recovered" (see Fig. 2), "Confirmed – Death" (see Fig. 3) and "Recovered– Death" (see Fig. 4). After divided each group into $k = 2$ to $k = 6$. The hybrid of K-Means algorithm and the Elbow method is been used to determine the best clustering as in [34].

Each data point is classified by computing the distance between that point and each group center, and then classifying the point to be in group whose center is closest to it. The results of sum square error calculations of each cluster have experienced the greatest decrease in k = 4 for groups "Confirmed-Recovered" and "Confirmed-Death", k=3

TABLE II. THE INITIAL VALUES

	K-Means	EM GMM		
'Confirmed-Recovered' data				
C_1	$\mu_1 = [73.24, 15.73]^t$	$\mu_1 = [73.24, 15.73]^t$	$\Sigma_1 =$	$\begin{vmatrix} 4974.83 & 763.34 \\ 763.34 & 729.72 \end{vmatrix}$
C_2	$\mu_2 = [127.88, 132.76]^t$	$\mu_2 = [127.88, 132.69]^t$	$\Sigma_2 =$	$\begin{vmatrix} 11465.05 & -1746.47 \\ -1746.47 & 10188.45 \end{vmatrix}$
C_3	$\mu_3 = [722.90, 501.69]^t$	$\mu_3 = [722.90, 501.69]^t$	$\Sigma_3 =$	$\begin{vmatrix} 24987.30 & 122031.60 \\ 122031.60 & 133911.50 \end{vmatrix}$
C_4	$\mu_4 = [2332.72, 2041.35]^t$	$\mu_4 = [2332.72, 2041.35]^t$	$\Sigma_4 =$	$\begin{vmatrix} 606823,60 & 256480.40 \\ 256480.40 & 321025.90 \end{vmatrix}$
'Confirmed-Death' data				
C_1	$\mu_1 = [73.24, 2.86]^t$	$\mu_1 = [73.24, 2.86]^t$	$\Sigma_1 =$	$\begin{vmatrix} 4974.83 & 95.79 \\ 95.79 & 11.27 \end{vmatrix}$
C_2	$\mu_2 = [127.88, 0.88]^t$	$\mu_2 = [127.88, 0.88]^t$	$\Sigma_2 =$	$\begin{vmatrix} 11465.05 & 1.02 \\ 1.02 & 15.10 \end{vmatrix}$
C_3	$\mu_3 = [722.90, 12.58]^t$	$\mu_3 = [722.90, 12.58]^t$	$\Sigma_3 =$	$\begin{vmatrix} 249875.30 & 5027.91 \\ 5027.91 & 133.19 \end{vmatrix}$
C_4	$\mu_4 = [2332.72, 38.37]^t$	$\mu_4 = [2332.72, 38.37]^t$	$\Sigma_4 =$	$\begin{vmatrix} 606823.60 & 4415.75 \\ 4415.75 & 81.86 \end{vmatrix}$
'Recovered-Death' data				
C_1	$\mu_1 = [49.38, 2.46]^t$	$\mu_1 = [49.38, 2.46]^t$	$\Sigma_1 =$	$\begin{vmatrix} 4480.11 & -31.63 \\ -31.63 & 9.34 \end{vmatrix}$
C_2	$\mu_2 = [211.65, 3.32]^t$	$\mu_2 = [211.65, 3.32]^t$	$\Sigma_2 =$	$\begin{vmatrix} 30352.05 & 396.53 \\ 396.53 & 20.27 \end{vmatrix}$
C_3	$\mu_3 = [1724.95, 35.56]^t$	$\mu_3 = [1724.95, 35.56]^t$	$\Sigma_3 =$	$\begin{vmatrix} 485690.30 & 3820.91 \\ 3820.91 & 102.70 \end{vmatrix}$

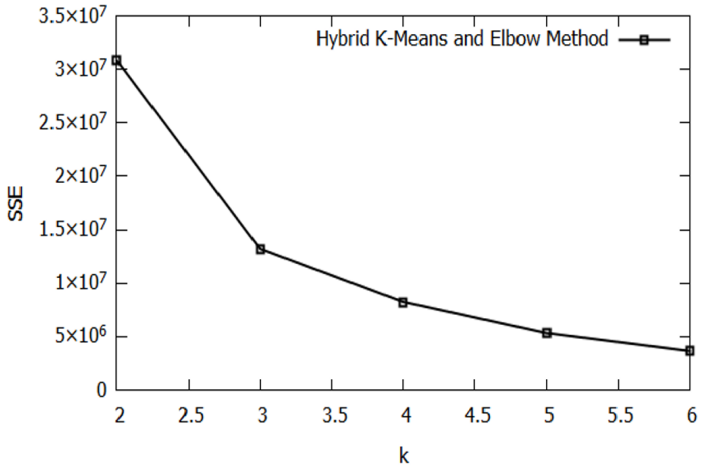

Fig. 7. Graph of sum square error depending on the number k for 'Recovered-Death' data.

for "Recovered-Death" data can be seen in (Fig. 5), (Fig. 6) and (Fig. 7).

We used the hybrid K-Means algorithm and Elbow method, which gave best clustering with 4 and 3 clusters. This result is exploited in the EM classification based on GMM, we notice that the "Confirmed-Recovered", "Confirmed-Death" and "Recovered-Death" can be divided into 4, 4 and 3 subsets, respectively. We analyze the correlation of feature variables for COVID-19, Correlation matrix is used to find the relationship between two variables "Confirmed-Recovered", "Confirmed-Death" and "Recovered-Death". Correlation Coefficient r is used to calculate the strength of this relationship between two quantitative variables $\boldsymbol{Y_i}$ and $\boldsymbol{Y_j}$ by using the formula given in (Eq. 13):

$$r = \frac{(\boldsymbol{Y_i} - \boldsymbol{\mu_i})^t (\boldsymbol{Y_j} - \boldsymbol{\mu_j})}{\sqrt{\|\boldsymbol{Y_i} - \boldsymbol{\mu_i}\|^2 \|\boldsymbol{Y_j} - \boldsymbol{\mu_j}\|^2}} \tag{13}$$

$\boldsymbol{i}$ and $\boldsymbol{j}$ = Confirmed, Recovered , Death r, the correlation coefficient is a unitless value between -1 and 1.

In Table II, we have the initial parameters of the different groups. To start the K-Means and EM algorithms, we use the same means values and the same coefficients of the found covariance matrix.

In this part, we aim to implement selected $\boldsymbol{C^{++}}$ object from [35] using K-Means algorithm is to partition the first, the second and the third group into four, four and three

TABLE III. VALUES AT CONVERGENCE FOR K-MEANS AND EM ALGORITHM

	K-Means	EM GMM		
'Confirmed-Recovered' data				
	Number of iterations: 86	Number of iterations: 445		
C_1	$\mu_1 = [157.36, 120.93]^t$	$\mu_1 = [17.81, 0.48]^t$	$\Sigma_1 =$	$\begin{bmatrix} 503.74 & 11.23 \\ 11.23 & 0.55 \end{bmatrix}$
C_2	$\mu_2 = [1323.33, 985.14]^t$	$\mu_2 = [135.36, 95.57]^t$	$\Sigma_2 =$	$\begin{bmatrix} 5532.70 & 1657.50 \\ 1657.50 & 5640.30 \end{bmatrix}$
C_3	$\mu_3 = [2152.97, 2074.56]^t$	$\mu_3 = [891.17, 660.67]^t$	$\Sigma_3 =$	$\begin{bmatrix} 320600.00 & 174070.00 \\ 174070.00 & 194950.00 \end{bmatrix}$
C_4	$\mu_4 = [3366.81, 2530.50]^t$	$\mu_4 = [2470.80, 2176.30]^t$	$\Sigma_4 =$	$\begin{bmatrix} 572850,00 & 175290.00 \\ 175290.00 & 249060.00 \end{bmatrix}$
'Confirmed-Death' data				
	Number of iterations: 88	Number of iterations: 456		
C_1	$\mu_1 = [148.60, 2.49]^t$	$\mu_1 = [88.88, 2.72]^t$	$\Sigma_1 =$	$\begin{bmatrix} 830.59 & 18.17 \\ 18.17 & 8.50 \end{bmatrix}$
C_2	$\mu_2 = [1315.52, 27.48]^t$	$\mu_2 = [114.66, 1.38]^t$	$\Sigma_2 =$	$\begin{bmatrix} 7680.60 & 72.23 \\ 72.23 & 2.21 \end{bmatrix}$
C_3	$\mu_3 = [2380.27, 35.77]^t$	$\mu_3 = [1124.20, 23.04]^t$	$\Sigma_3 =$	$\begin{bmatrix} 182120.00 & 4896.50 \\ 4896.50 & 180.76 \end{bmatrix}$
C_4	$\mu_4 = [3562.50, 50.58]^t$	$\mu_4 = [2599.90, 38.21]^t$	$\Sigma_4 =$	$\begin{bmatrix} 494200.00 & 5960.00 \\ 5960.00 & 116.10 \end{bmatrix}$
'Recovered-Death' data				
	Number of iterations: 77	Number of iterations: 219		
C_1	$\mu_1 = [140.29, 3.40]^t$	$\mu_1 = [59.08, 0.67]^t$	$\Sigma_1 =$	$\begin{bmatrix} 3067.70 & 10.24 \\ 10.24 & 0.61 \end{bmatrix}$
C_2	$\mu_2 = [1294.61, 31.66]^t$	$\mu_2 = [175.72, 4.22]^t$	$\Sigma_2 =$	$\begin{bmatrix} 26438.00 & -225.24 \\ -225.24 & 9.16 \end{bmatrix}$
C_3	$\mu_3 = [2403.27, 40.86]^t$	$\mu_3 = [1636.30, 33.23]^t$	$\Sigma_3 =$	$\begin{bmatrix} 602620.00 & 6191.30 \\ 6191.30 & 140.61 \end{bmatrix}$

TABLE IV. THE CORRELATION COEFFICIENTS OF THE COVID-19 DATA; FOR INITIAL VALUES WITH DIFFERENT CLUSTERS

	Confirmed cases					Confirmed cases				Recovered cases			
RC	0.40	0	0	0	DC	0.25	0	0	DC	0.40	0	0	0
RC	0	−0.16	0	0	DC	0	0.81	0	DC	0	0.14	0	0
RC	0	0	0.67	0	DC	0	0	0.69	DC	0	0	0.87	0
RC	0	0	0	0.58					DC	0	0	0	0.63

TABLE V. THE CORRELATION COEFFICIENTS OF THE COVID-19 DATA; FOR VALUES AT CONVERGENCE WITH DIFFERENT CLUSTERS

	Confirmed cases					Confirmed cases				Recovered cases			
RC	0.67	0	0	0	DC	0.24	0	0	DC	0.22	0	0	0
RC	0	0.30	0	0	DC	0	−0.40	0	DC	0	0.55	0	0
RC	0	0	0.70	0	DC	0	0	0.67	DC	0	0	0.85	0
RC	0	0	0	0.51					DC	0	0	0	0.78

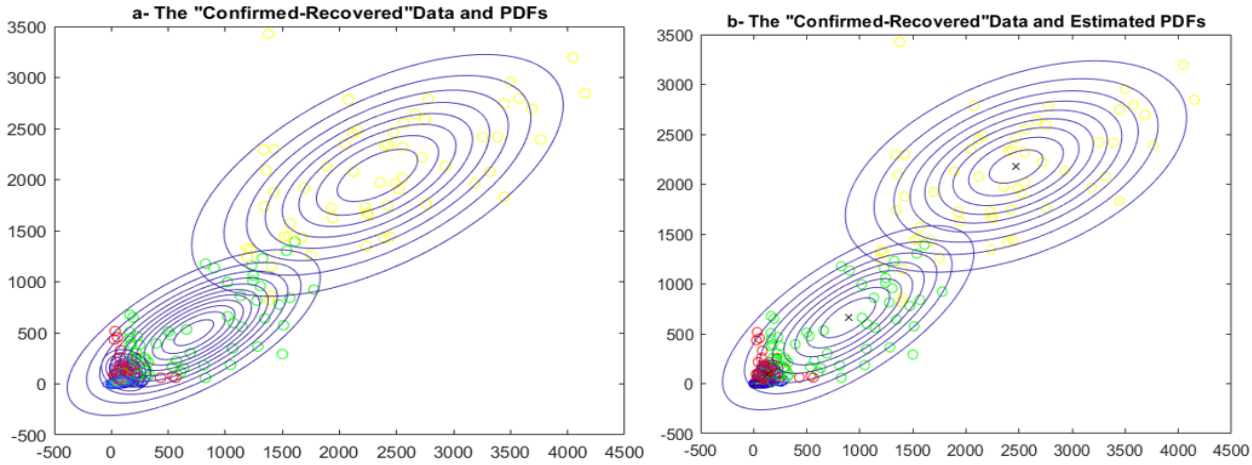

Fig. 8. Experiments result after implementation EM clustering for 'Confirmed-Recovered' 2-dimensional data generated by GMM with four mixture components. (a)-Graphs at initials values (b)-Graphs at convergences values.

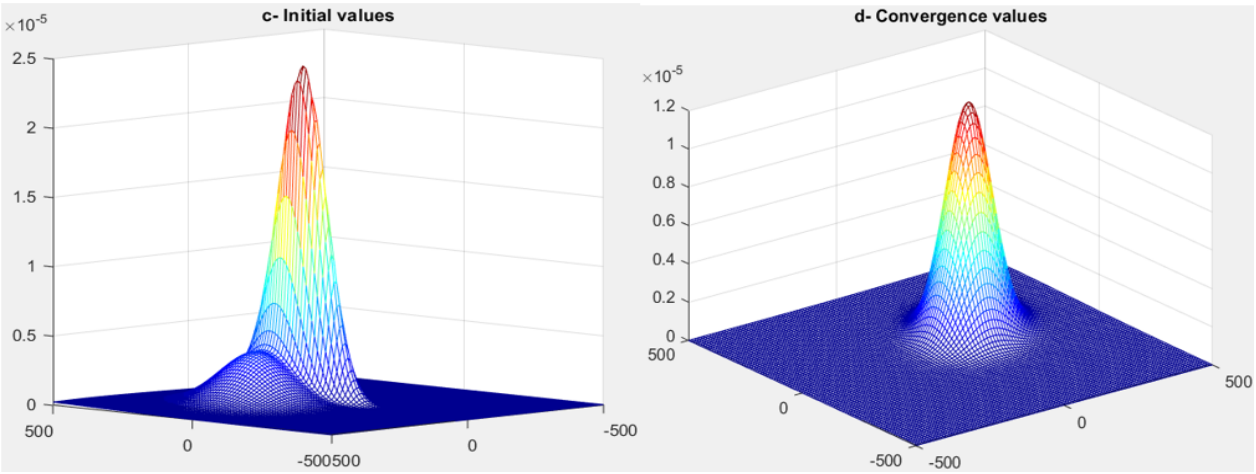

Fig. 9. Contours of probability density function (PDF) with four mixture components of "Confirmed-Recovered" data for (c) and (d) figures.

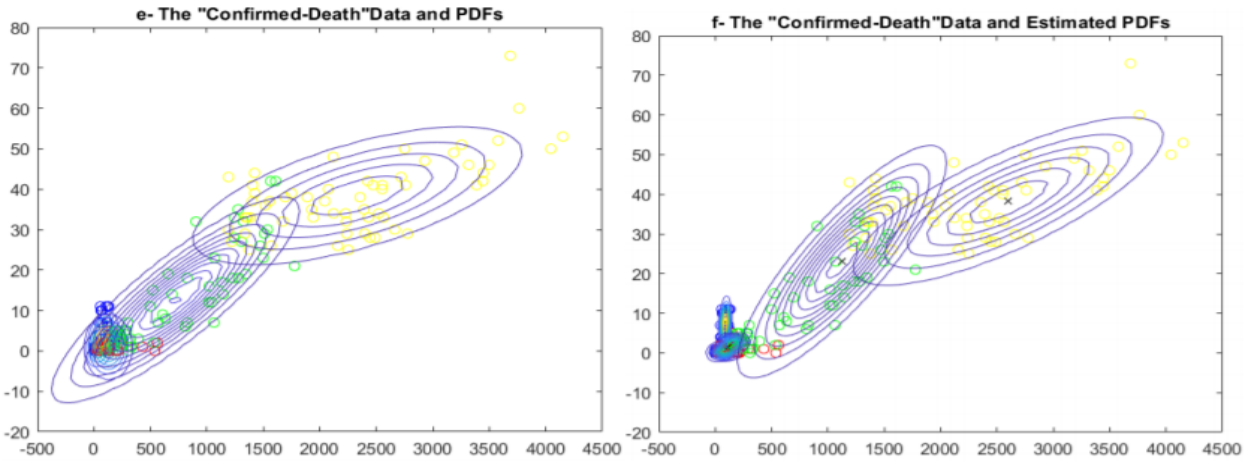

Fig. 10. Experiments result after implementation EM clustering for "Confirmed-Death" 2-dimensional data generated by GMM, with four mixture components. (e)-Graphs at initials values (f)-Graphs at convergences values.

clusters, respectively. Also, we apply EM by using GMM based on Matlab for all three groups 'Confirmed – Recovered'

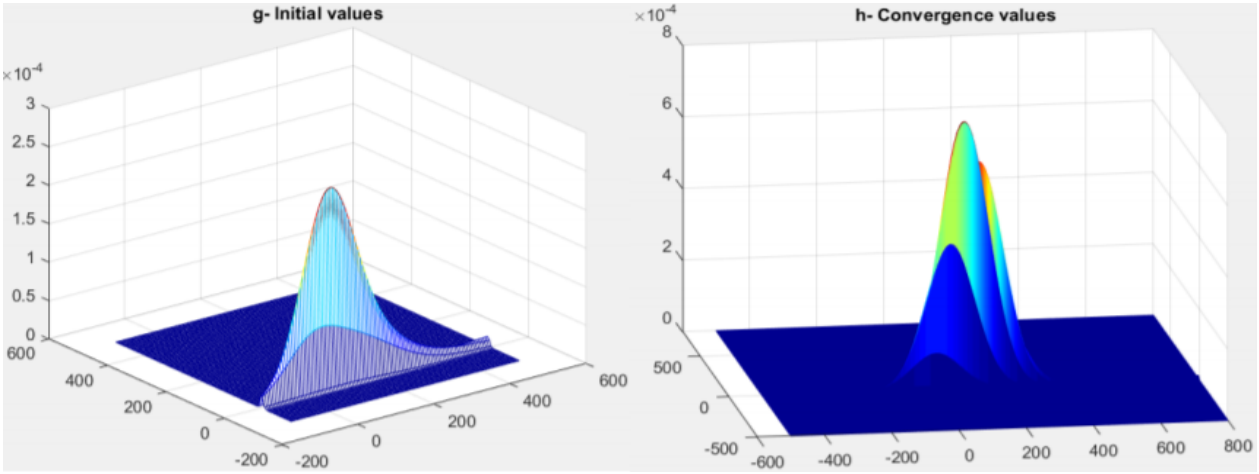

Fig. 11. Contours of probability density function (PDF) with four mixture components of "Confirmed-Death" data for (g) and (h) figures.

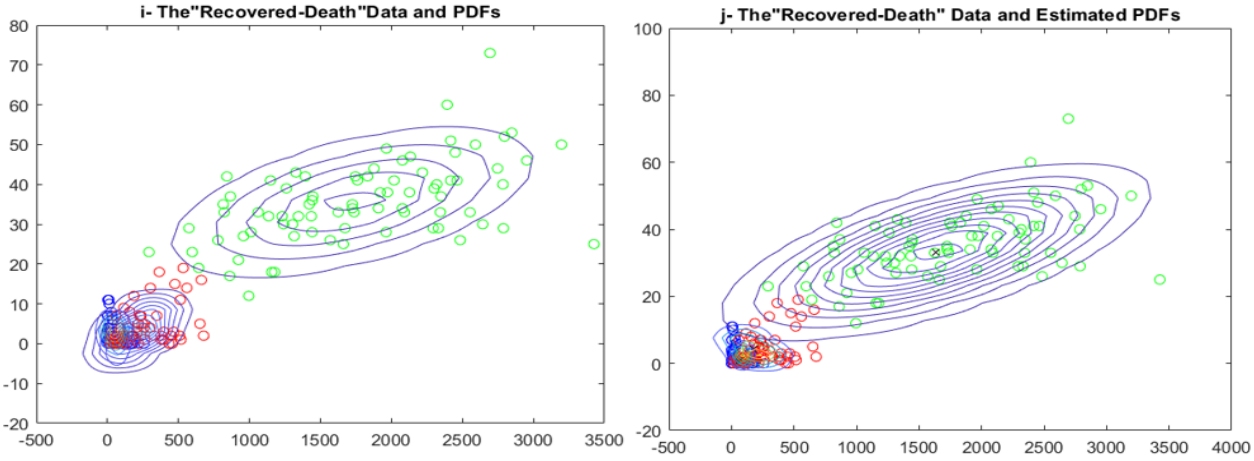

Fig. 12. Experiments result after implementation EM clustering for "Confirmed-Recovered" 2-dimensional Data generated by a GMM, with four mixture components. (i)-Graphs at initials values (j)-Graphs at convergences values.

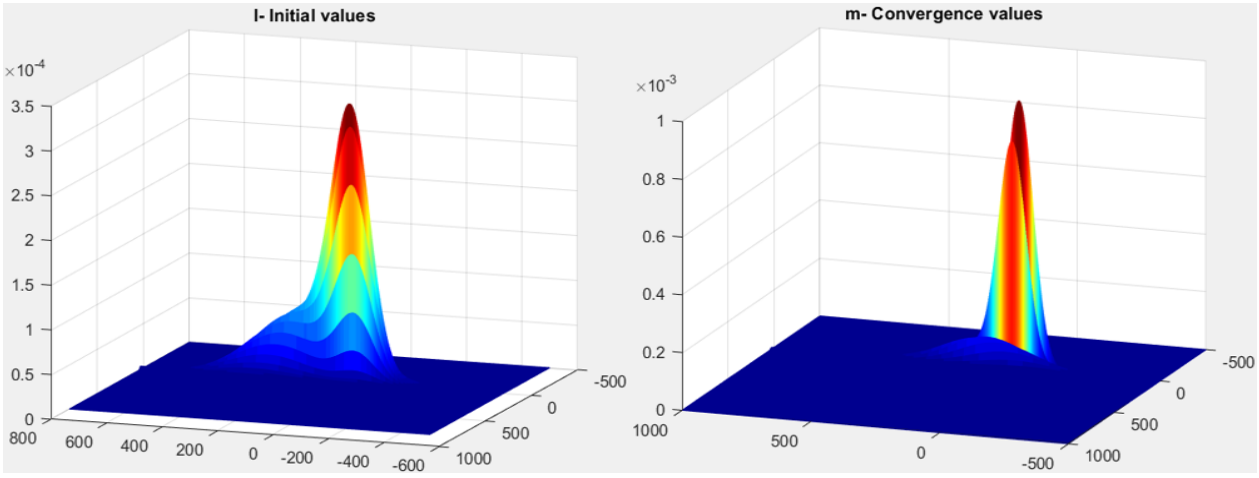

Fig. 13. Contours of probability density function (PDF) with four mixture components of "Confirmed-Recovered" data for (l) and (m) figures.

(see Fig. 8, 9), 'Confirmed – Death' (see Fig. 10, 11), and 'Confirmed – Recovered' (see Fig. 12, 13).

We obtain values at convergence by using K-Means algorithm and EM algorithm (see Table III).

The correlation matrix of the initial values (see Table IV), and values at convergences (see Table V) for features "Confirmed – Recovered – Death" cases using (Eq. 13):

Positive values of r indicate a positive correlation, as the values of the two variables tend to increase together. Negative values of r indicate a negative correlation when the values of one variable tend to increase and the values of the other variable decrease. In the data mining of COVID-19 in Morocco, K-Means is a simple and fast algorithm for solving clustering issues, but it requires clarification in advance the exact number of clusters k, which is often difficult.

The "Confirmed – Recovered", "Confirmed – Death" and "Recovered – Death" groups are Mixtures Models of four and three two-dimensional Gaussians. K-Means algorithm only considers the mean to update the new centroids nevertheless EM based GMM takes into account the mean value as well as the covariance matrix of this data groups. We use this partitioning to start K-Means and EM. We start from a real model with correlated covariance matrices, the values at convergence are of the same nature. It can be interpreted that high positive correlation exists, in the third phase of the epidemic's spread, between Confirmed cases and Recovered cases (0.70), Confirmed cases and Death cases (0.67) and Recovered cases and Death cases (0.85) [10]. To evaluate clusters "Confirmed – Recovered" and "Recovered – Death"; values are in forms four categories (low, lower-middle, uppermiddle, and high), on the order hand "Confirmed – Death" data is in forms three phase (low, medium, and high).

We notice a clear difference between means of the K-Means algorithm and the means of the GMM. The EM based GMM has higher computation time than K-Means; because K-Means does not account for variance. The findings are in according with those of [19]. The Data membership points to clusters in GMM is probabilistic as versus the non-probabilistic, hard clustering K-Means process, thus resolving the membership vagueness that may appear in overlapping clusters. The analysis exposes a more meaningful workloads clustering with GMM than with K-Means, enabling a detailed characterization of resource usage needs of Cloud workload. As a comparison, the clustering by using K-Means algorithm is faster than Gaussian Mixture Models method.

K-Means clustering faces a major challenge in determining the optimal number of clusters, especially when working with COVID-19 data. Depending on the type of data being analyzed, the number of clusters may vary, and selecting the correct number of clusters is crucial for obtaining meaningful results. Furthermore, K-Means clustering relies on the Euclidean distance metric, which may not be suitable for all COVID-19 data. Other distance metrics, such as cosine distance, may be necessary to accurately capture the similarity between data points. Another clustering algorithm, EM clustering, is also sensitive to the initial conditions of the algorithm. Different initial conditions may result in different cluster assignments, leading to inconsistent results. Additionally, EM clustering may struggle to converge to a solution when working with high-dimensional data or complex probability distributions. Preprocessing and tuning of the algorithm may be necessary to ensure reliable results.

VII. Conclusion

This study focuses on analyzing the COVID-19 situation in Morocco using K-Means and EM clustering algorithms. The dataset includes daily Confirmed, Death, and Recovered cases from March 2 to October 24, 2020. For the k-means algorithm, discovering intra-cluster similarity in complex nonlinear models using Euclidean distance is difficult. The EM algorithm is more computationally intensive and requires larger sample sizes for accurate parameter estimates. The results indicate that the EM-based GMM method is the preferred clustering method as it yields smaller classification error rates. The K-Means generated clusters provide limited information, and the best clustering was found with four and three clusters. Furthermore, the EM algorithm demonstrates the correlation between "Confirmed-Recovered", "Confirmed-Death", and "Recovered-Death". The number of clusters corresponds to the number of phases of the epidemic propagation, as determined by the process of identifying the optimal number of clusters. In the future work, we will be focused on the enhancement of our model clustering for multi-dimensional datasets with several features.

References

[1] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[2] G. Gan and E. A. Valdez, "Data clustering with actuarial applications," *North American Actuarial Journal*, vol. 24, no. 2, pp. 168–186, 2020.

[3] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[4] K. Chowdhury, D. Chaudhuri, A. K. Pal, and A. Samal, "Seed selection algorithm through k-means on optimal number of clusters," *Multimedia Tools and Applications*, vol. 78, no. 13, pp. 18 617–18 651, 2019.

[5] H. Jiang and E. Arias-Castro, "$\boldsymbol{k}$-means and gaussian mixture modeling with a separation constraint," *arXiv preprint arXiv:2007.04586*, 2020.

[6] M. Hamidi, H. Satori, O. Zealouk, and K. Satori, "Amazigh digits through interactive speech recognition system in noisy environment," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 101–109, 2020.

[7] M. Hamidi, H. Satori, O. Zealouk, K. Satori, and N. Laaidi, "Interactive voice response server voice network administration using hidden markov model speech recognition system," in *2018 Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*. IEEE, 2018, pp. 16–21.

[8] M. Amouch and N. Karim, "Modeling the dynamic of covid-19 with different types of transmissions," *Chaos, Solitons & Fractals*, p. 111188, 2021.

[9] A. Xavier Jr, "A c++ code for predicting covid-19 cases by least-squares fitting of the logistic model," *Pre-print available on Research Gate (https://www. researchgate. net/). DOI*, vol. 10, 2020.

[10] A. Rizvi, M. Umair, and M. A. Cheema, "Clustering of countries for covid-19 cases based on disease prevalence, health systems and environmental indicators," *medRxiv*, 2021.

[11] J. P. F. Arocutipa, J. J. Huallpa, G. C. Navarro, and L. D. B. Peralta, "Clustering k-means algorithms and econometric lethality model by covid-19, peru 2020," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021.

[12] P. V. Sagar, T. P. Kumar, G. K. Chaitanya, and M. N. Rao, "Covid-19 transmission risks assessment using agent-based weighted clustering approach," 2020.

[13] M. Zubair, M. Asif Iqbal, A. Shil, E. Haque, M. Moshiul Hoque, and I. H. Sarker, "An efficient k-means clustering algorithm for analysing covid-19," in *Hybrid Intelligent Systems: 20th International Conference on Hybrid Intelligent Systems (HIS 2020), December 14-16, 2020*. Springer, 2021, pp. 422–432.

[14] B. A. Hassan, T. A. Rashid, and H. K. Hamarashid, "A novel cluster detection of covid-19 patients and medical disease conditions using improved evolutionary clustering algorithm star," *Computers in biology and medicine*, vol. 138, p. 104866, 2021.

[15] R. Kurniawan, S. N. H. S. Abdullah, F. Lestari, M. Z. A. Nazri, A. Mujahidin, and N. Adnan, "Clustering and correlation methods for predicting coronavirus covid-19 risk analysis in pandemic countries," in *2020 8th International Conference on Cyber and IT Service Management (CITSM)*. IEEE, 2020, pp. 1–5.

[16] E. Alsuwat, S. Alzahrani, and H. Alsuwat, "Detecting covid-19 utilizing probabilistic graphical models," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[18] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.

[19] E. Patel and D. S. Kushwaha, "Clustering cloud workloads: K-means vs gaussian mixture model," *Procedia Computer Science*, vol. 171, pp. 158–167, 2020.

[20] S. K. Appiah, K. Wirekoh, E. N. Aidoo, S. D. Oduro, and Y. D. Arthur, "A model-based clustering of expectation–maximization and k-means algorithms in crime hotspot analysis," *Research in Mathematics*, vol. 9, no. 1, p. 2073662, 2022.

[21] V. Zarikas, S. G. Poulopoulos, Z. Gareiou, and E. Zervas, "Clustering analysis of countries using the covid-19 cases dataset," *Data in brief*, vol. 31, p. 105787, 2020.

[22] S. Aungkulanon, V. Tangcharoensathien, K. Shibuya, K. Bundhamcharoen, and V. Chongsuvivatwong, "Post universal health coverage trend and geographical inequalities of mortality in thailand," *International journal for equity in health*, vol. 15, no. 1, pp. 1–12, 2016.

[23] A. Malav, K. Kadam, and P. Kamat, "Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy," *International Journal of Engineering and Technology*, vol. 9, no. 4, pp. 3081–3085, 2017.

[24] R. Singh and E. Rajesh, "Prediction of heart disease by clustering and classification techniques prediction of heart disease by clustering and classification techniques," *International Journal of Computer Sciences and Engineering*, 2019.

[25] S. Y. Işikhan and D. Güleç, "The clustering of world countries regarding causes of death and health risk factors," *Iranian Journal of Public Health*, vol. 47, no. 10, p. 1520, 2018.

[26] Y. G. Jung, M. S. Kang, and J. Heo, "Clustering performance comparison using k-means and expectation maximization algorithms," *Biotechnology & Biotechnological Equipment*, vol. 28, no. sup1, pp. S44–S48, 2014.

[27] I. B. Mohamad and D. Usman, "Standardization and its effects on k-means clustering algorithm," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, no. 17, pp. 3299–3303, 2013.

[28] C. M. Bishop, "Pattern recognition," *Machine learning*, vol. 128, no. 9, 2006.

[29] R. O. Duda, P. E. Hart *et al.*, *Pattern classification*. John Wiley & Sons, 2006.

[30] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu *et al.*, "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding," *The lancet*, vol. 395, no. 10224, pp. 565–574, 2020.

[31] D. Paraskevis, E. G. Kostaki, G. Magiorkinis, G. Panayiotakopoulos, G. Sourvinos, and S. Tsiodras, "Full-genome evolutionary analysis of the novel corona virus (2019-ncov) rejects the hypothesis of emergence as a result of a recent recombination event," *Infection, Genetics and Evolution*, vol. 79, p. 104212, 2020.

[32] W. H. Organization, "Coronavirus disease (covid-19): situation report, 209," 2020.

[33] M. H. Ministry, "http://www.covidmaroc.ma/ (last accessed: November 30 2020, 17:00 gmt)," 2020.

[34] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP Conference Series: Materials Science and Engineering*, vol. 336, no. 1. IOP Publishing, 2018, p. 012017.

[35] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes with Source Code CD-ROM 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 2007.