International Journal of Innovative Engineering Research and Development
IJIERD

EXPECTATION MAXIMIZATION ALGORITHM (EM): A FOCUS ON NORMAL MIXTURE PARAMETER ESTIMATION

Kalu, S. I.

School of Science & Industrial Technology, Department of Statistics, Abia State Polytechnics, Aba

Corresponding Author: Kalu, Stephen I. (PhD), School of Science & Industrial Technology, Department of Statistics, Abia State Polytechnics, Aba. Email: monarchynwokike@gmail.com

Abstract

Given a population of two clusters, each assumed to be normally distributed. The problem is how to estimate the parameters of these clusters including their respective mixture weights. This problem is modeled as a Gaussian mixture model, and we used the recursive iterative algorithm called Expectation Maximization (EM) to estimate these parameters. Two data sets were used in this study, namely, data Set 1and data set 2. Data Set 1 are the compiled weights of 1000 first year students of Abia State Polytechnic, Aba, Nigeria, which breaks up into 500 male students and 500 female students, while Data Set 2 are the weights of 200 babies at birth, consisting of 120 males and 80 females babies at birth from the Federal Medical Centre (FMC), Owerri, Imo State, Nigeria. We used MATLAB analytical tool to implement the EM Algorithm procedures. We also presented one of the most robust estimation/classification techniques called Maximum likelihood Estimation (MLE) via the EM algorithm for discrimination on the two component mixture of Gaussian distributions We initialized the E-Step of the Algorithm using Data Set 1&2 with the following values from table1. Dataset 1 initialization values: $\pi_1 = 0.5, \pi_2 = 0.5, \mu_1 = 75.2768, \mu_2 = 67.9795, \sigma_1 = 6.3488,$ $\sigma_2 = 7.8162$. Dataset 2 initialization values: $\pi_1 = 0.5, \pi_2 = 0.5, \mu_1 = 3.1808, \mu_2 = 3.2013, \sigma_1 = 0.6216, \sigma_2 = 0.5518$. After full implementation of EM, we were able to optimize the initialized estimable values to: Dataset 1 optimizes values $\pi_1 = 0.4994, \pi_2 = 0.5006, \mu_1 = 75.9416, \mu_2 = 67.3252, \sigma_1 = 6.2114, \sigma_2 = 7.2184$ Dataset 2 optimized values $\pi_1 = 0.4998, \pi_2 = 0.5002, \mu_1 = 3.2953 = 3.0308, \sigma_1 = 0.4212, \sigma_2 =$ 0.2791.

Keywords: Mixture Models Distribution of Mixture Gaussian Mixture Model Estimation with EM Algorithm.

INTRODUCTION

Mixture Models are powerful Statistical tool for modeling complex data. According to Lindsay, (1995), Mixture Models are probabilistic models used for representing the presence of population subsets located within an overall population. There exists a long history of modeling complex data with a mixture of Gaussian; see for example (Kuroda, 2021; McLachlan & Peel, 2000a, b). Gaussian Mixture Models are used to investigate the performances of certain estimators' departures from normality (Turkey, 1960). Also as suggested by Turkey (1960), the density of an observation is taken to be a mixture of Uni-variate normal densities with the same means while the second component has a greater variance than the first i.e having; $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_{2\sim}$ $(\mu_1, k\sigma_1^2)$. Huber, (1964), considered more general forms of contamination of the Gaussian Distributions in the development of his Robust M-estimators of a location parameter. A good introduction to the theory and applications of Mixture distributions can be found in (Pearson,1894; Richard & Dean, 2002; William, 2008; Wirjanto & Xu, 2009).

Fitting a model to a set of data can be handled by any of the following techniques, which include but not limited to; the Maximum Likelihood Estimation (MLE) Method (Kiefer & Wolfowitz, 1956) and the graphical method and the method of moments. These methods were also discussed and applied by (Titherington et al., 1985). MLE method in the fitting of the mixture models due to its rich statistical resumes both in theory and in practical is the major properties of the EM algorithm has been established by (Dempster et al., 1977). The EM algorithm is a statistical tool for determining the parameters of a mixture with an *a priori* given number of components. It simplifies difficult Maximum Likelihood problems, and is of particular appeal for finite normal mixtures where closed-form expressions are possible, such as the iterative algorithm according to (Dempster et al, 1977). This paper was structured as follows.

First, we offered the case study and provide necessary discussion on how mixture distribution mingles well with non-normality situations and asymmetry nature of the data sets. We then present the most frequently and commonly used Maximum Likelihood Estimation (MLE) through the EM algorithm to fit the two component mixture of Gaussian distribution. We also demonstrate the use of data sets in construction of the clusters and density plots of the mixtures. Note that we applied simulation on the weights of babies at birth since the numbers of male and female babies at birth are not the same. This strategy is to enable us put up a smooth shaped density plots mimicking normal distribution curves for the weights of babies at birth. Finally, we offered conclusion based on our findings.

Case-Study

Data: The data sets we used in this paper include; (a) the weights of one thousand (1000) Individuals consisting of male and female students, drawn from Abia State Polytechnic, Aba. These are among the students admitted into various programmes of the Institution for the 2020/2021 academic session into level 100 programmes. Here, the number of males and females are the same ie $n_1 = n_2$ = 500. (b) The weights of two hundred (200) babies delivered at birth, obtain from the Federal Medical Centre (FMC), Owerri, Imo State recorded within a 5-year period (2018-2022). The dataset consists of one hundred and twenty (120) males' weights and eighty (80) female weights. Here, the numbers of male and female babies at birth, selected are not the same i.e $n_1 \neq n_2$. *MATLAB version R2007b package was used for all our empirical analyses. * The MATLAB source codes are available on request.

Description of the empirical weight properties of the students

The indices of the weights of students as outlined in table 1, which provides descriptive statistics of the weights of students. We can see that the male-Index and the mixture are negatively skewed,

while the female-Index is positively skewed. Now, a look at the Kurtosis (which equal to 3 for mesokurtic or normal distribution), indicates values, each far less than 3 for the three respective indices, hence, indicating lighter tails and lower peak than the normal distribution. Hence, we reject the null hypothesis of normality for each of the three (3) indices displayed in figure1

Table 1: Summary Statistics for Students and babies weights

Indices of the weights of students				Indices of the weights of babies at birth			
Statistics	Male-Index	Female-Index	Mixture	Statistics	Male-Index	Female-Index	Mixture
Mean	75.2768	67.9795	71.6282	Mean	3.18	3.2013	3.1885
St. Error	0.2839	0.3496	0.2529	St. Error	0.05681	0.06169	0.04199
Median	75.4817	67.7409	72.1667	Median	3.2	3.2	3.2
Mode	71.7263	47.4214	73.5364	Mode	3.5	3	3
St.Dev.	6.3488	7.8162	7.9985	St.Dev.	0.62236	0.55175	0.59379
Samp. Var.	40.3076	61.0925	63.9755	Samp. Var.	0.387	0.304	0.353
Kurtosis	-0.326	-0.0348	-0.2455	Kurtosis	2.125	0.205	1.593
Skewness	-0.0546	0.1536	-0.1647	Skewness	-0.639	-0.248	-0.521
Range	35.03	48.4573	48.4573	Range	3.9	2.9	3.9
Minimum	57.691	47.4214	47.4214	Minimum	1	1.8	1
Maximum	92.721	95.8787	95.8787	Maximum	4.9	4.7	4.9
Sum	37638.4	33989.75	71628.18	Sum	381.6	256.1	637.7
Count	500	500	1000	Count	120	80	200

Distributions of Mixture

Distributions of mixture or rather mixture of distributions can be thought of as an alternative way of explaining distribution(s) unconformity or departure from normality, (Wirjanto & Xu, 2009). Also in (Wirjanto & Xu, 2009), given a set of an independently and identically distributed random variable X drawn from different normal distributions having probability p_k by specifying the component's probability density function thus;

$$f(X;\theta) = \sum_{i=1}^{k} p_k N(\mu_k, \sigma_k^2) \quad (1)$$

here, k is the number of components.

We let $\theta = (p_i, \mu_i, \sigma_i^2)$, where, $i = 1, \dots, k$.

Where p, μ and σ is each $k \times 1$ vector? This implies that we have $3k$ unknown parameters to estimate from (1) if parameter q is included. However, when $k = 2$, only p is estimated, while $q = 1 - p$ is obtained from the relationship $p + q = 1$**.** The total numbers of parameters to be estimated is $3k - 1$**;** which implies that since$k = 2$**,** we are expected to estimate five (5) parameters values from the given Gaussian distributions.

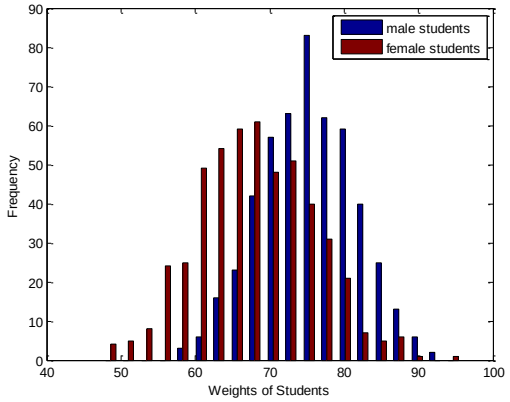


Figure1: Bar-Chart of the students' weights

Figure1 displayed the average weight of the entire students is approximately 70 (that is the intersection point of the male and female distribution curves).From the figure, it could also be seen by visual inspection that the average weights of male and female students are about 75 and 68 respectively.

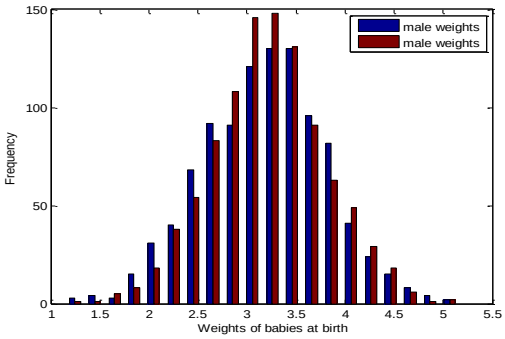


Figure 2: The Multiple Bar-Chart of the weights of babies at birth.

From figure 2, it can be clearly deduced that the average age of both male and female weights of babies at birth does not differ significantly as is the case with the student weights.This implies that weights tends to differ as age increases. By visual inspection,the average weights of male and female babies at birth are 3.18 and 3.20 respectively, almost the same value by intrinsic approximation

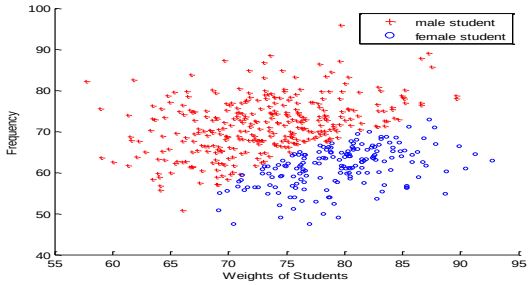


Figure 3: Cluster plots of the weights of students

By visual inspection from figure 3,suppose that the underlying parameter values are unknown,we can derive from this cluster plots the following information;

· That there are two components.

· The variances between the clusters are not the same.

. The males weights has lesser densities compared to the females weights (since males weights floats on top of females weights as can be seen in the diagram above).

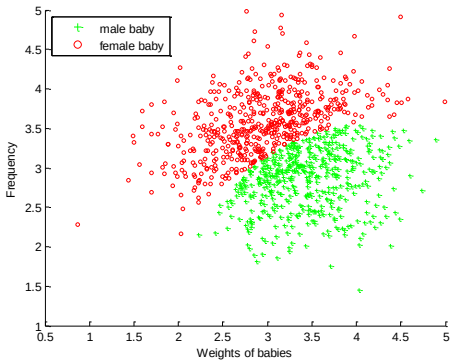


Figure 4: Cluster plots of babies' weights at birth.

From the Simulated scatter plots for clusters of babies weights at birth in figure 4, we can derive the followings by visual inspection that;

(a) That there are two component clusters.

(b) The between cluster variances are not the same.

(c) The weights of females have higher densities than that of males.

Methods

The advantages of mixture of normals includes that they maintain the tractibility of individual normal properties and have finite higher order moments and can capture the kurtosis and skewness of given normal distribution(s), (McLachlan & Krishna, 1997). Furthermore, other merits of mixture of normals are that it can capture the structural change in the distribution variances as well as distribution means, and can be asymmetric, according to (Oleszak, 2020).

(Wu, 1985) gave the general form of a normal mixture thus

$$F(x) = \sum_{i=1}^{K} \pi_i \, \phi\left(\frac{x-\mu_i}{\sigma_i}\right) \qquad (1)$$

ϕ is the cumulative density function of $N(0,1)$, the probability density function of a mixture of Gaussian is given as follows

$$f(x) = \sum_{i=1}^{K} \pi_i \, \emptyset(x; \mu_i, \sigma_i) \quad (2)$$

Where

$$\emptyset(x; \mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}$$

$$\sum_{i=1}^{K} \pi_i = 1 \;\&\; 0 \leq \pi_i \leq 1$$

i = 1, 2,..., K. Hence, in mixture of Gaussians, the distribution of weights is approximated by a Normal mixture, each having unique mean μ_i and variance σ_i and mixing probabilities π_i .

In (Huber, 1964) a mixture of two Gaussian distributions model is given by

$$f(x_j) = \pi_1 \varphi(x_j ; \mu_1, \sigma_1^2) + \pi_2 \varphi(x_j ; \mu_2, k\sigma_2^2) \quad (3)$$

Where k is the number of component, and $\varphi(x_j;\mu_j, \sigma_j^2)$ is the pdf of a normal distribution with finite mean μ_j and finite varianceσ_j^2. Here, j = 1, 2. The five estimable parameters of the mixture distribution which are$\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ allows for a very flexible definition of asymmetry and non-normality of the distributions. The weight is the mixing probabilityπ, where π_2= 1- π_1 are weights summing to unity or global one, implying that the first regime occurs with probability π_1 and the second regime withπ_2. By using mixture distributions, we can have densities the characteristics of lower peaks and lighter tails than normal distribution as can be seen in this paper.

Fitting a Gaussian mixture model to our data using EM algorithm for two component mixture

According to (Titherington et al., 1985), EM algorithm is seemingly the most popular technique currently utilized to evaluate the parameter of a mixture with an "a priori" given number of

components. EM is of great particular appeal for finite mixture where closed-form expressions are possible.

On the basis of the current estimate for the parameters, the EM algorithm will evaluate the expected value as well as the parameter estimates at the E step and maximizes the resulting expectation at the M step. Repeating the E step and M step, convergence is detected by computing the value of the log-likelihood after each iteration is implemented, and we stop when it appears the optimization is no longer changing in a significant manner from one unique solution to the next. In this section, we explain a mixture model of the simplest type for density estimation and also offer the associated EM algorithm for implementing maximum likelihood estimation.

Estimation of the mixture normal probability density function (pdf) with two mixture components

Given two mixture components, the likelihood function is given as

$$(\theta; x) = \frac{p_1}{\sqrt{2\pi\sigma_1^2}}\ exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] + \frac{1-p_2}{2\pi\sigma_2^2}\ exp\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right] \quad (4)$$

& the log likelihood is expressed as:

$$l(\theta; x) = \sum_{i=1}^{n} ln\left(\frac{p_1}{\sqrt{2\pi\sigma_1^2}}\ \left[-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right] + \frac{1-p_2}{\sqrt{2\pi\sigma_2^2}}\ exp\ \left[-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right]\right) \quad (5)$$

The log likelihood function is maximized by taking the partial derivatives of (5) w.r.t. θ, and equating to zero. That is $\frac{\partial\ ln(\theta;x)}{\partial\theta} = 0$.

We model X as a mixture of two normal distributions:

$Model\ 1 = X_1 \sim N(\mu_1, \sigma_1^2)$,

$Model\ 2 = X_2 \sim N(\mu_2, \sigma_2^2)$,

$X = \Delta X_1 + (1-\Delta) X_2$ (6)

$\Delta \in \{0,1\}$ having Pr.(Δ=1)= p

This implies that we should generate a $\Delta \in \{0,1\}$ with probability p, then depending on the outcome, we deliver X_1 *or* X_2 .

If we let $\varphi_\theta(y)$ symbolizes the normal density with parameters $\theta = (\mu, \sigma^2)$. Then this implies that the density of the random variable X can be expressed thus

$$G_x(X_k) = p\varphi_{\theta 1} + (1-p)\varphi_{\theta 2}$$

Which is the same as $\frac{p}{\sqrt{2\pi\sigma_1^2}}\left[-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right] + \frac{1-p}{\sqrt{2\pi\sigma_2^2}}\ exp\left[-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right]$ (7)

In fitting this model to our data using maximum likelihood estimation method, the expected parameters are stated as follows

$\theta = (p, \theta_1, \theta_2) = (p, \mu_1, \sigma_1^2, \sigma_2^2)$.

The log likelihood function is maximized by taking the partial derivatives of (5) with respect to $(p, \mu_1, \sigma_1^2, \sigma_2^2)$ and equating each to zero. That is $\frac{\partial \ln(\theta;x)}{\partial\theta} = 0$.

For the respective five (5) equations to be solved explicitly for their parameters, we use EM algorithm. The logic behind this is because more often than not, the stated system of equations cannot be resolved explicitly. Implying that the stated Equations has no unique roots, hence, warranting the use of EM algorithm to make estimations possible. We use this procedure for the special case of finite Gaussian mixtures, (Titherington et al., 1985). According to the approach adopted by (Wu, 1985) given that each component k_i , $i = 1,2$ is a multivariate Gaussian density with parameters $\underline{\mu}_k$ and Σ_k , where $\underline{\mu}_k = \mu_1, \mu_2$ and $\Sigma_k = \sigma_1^2, \sigma_2^2$. Then$\theta_{k=}\{p_k;\ \underline{\mu}_k, \Sigma_k\}$

$$p_k(\underline{x}/\theta_k) = \frac{1}{(2\pi)^{d/2}|\sum_k|^{1/2}} \; exp\left[-\frac{1}{2}(\underline{x}-\underline{u}_k)^t \sum_k^{-1}(\underline{x}-\underline{u}_k)\right] \quad (8)$$

The EM algorithm for Gaussian mixture is an iteration process that commences from some initial estimate θ until convergence. Each iteration consists of an E-Step and an M-Step. At the E-Step, with the initial guesses for the parameters, the latent variable for each data point from each distribution is computed by calculating expectation values of the membership weight for each data point. While at the M-Step, we now use the computed latent weight together with the observed data to calculate the new parameter values. After which we go back and start re-computing the membership weights in the E-Step and keep updating the parameters this way until convergence.

To derive the full E-M update for Gaussian mixtures. The E-Step simply consists of computing the posterior probabilities, as in

$$r_k^{(i)} \leftarrow \Pr\left(z^{(i)} = k | x^{(i)}\right. \quad (9)$$

$$= \frac{\Pr\left(z^{(i)}=k\right) p(x^{(i)}|z^{(i)}=k)}{\sum k^1 \Pr\left(z^{(i)}=k^1\right) P(x^{(i)}|z^{(i)}=k^1)} \quad (10)$$

$$= \frac{\pi_{k\,.Gaussian(x^{(i)};\mu_k,\sigma_k)}}{\sum k^1 \pi_{k^1}.Gaussian(x^{(i)};\mu_{k^1},\sigma_{k^1})} \quad (11)$$

Now, we turn to the M- Step. Considering the mixing proportions π_k, we go through the derivation and guess what the solution should be. A good estimate of the probability of an outcome is the proportion of times it appears, i.e. $\pi_k = \frac{N_k}{N}$, where N_K is the count for outcome k, and N is the number of observations. If we replace N_K with the sum of the responsibilities (which can be thought of as "fractional counts"), we should then end up with;

$$\pi_k \leftarrow \frac{1}{N}\sum_{i=1}^{N} r_k^{(i)}. \quad (12)$$

To see if we were right, we can observe that the mixing proportions doesn't affect the second term inside the sum in Equation (12), hence, we only need to worry about the first term, $\log \Pr(z^{(i)} = k)$. Maximizing the equation, we have;

$$\sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}\, logPr\left(z^{(i)} = k\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)}\, log\pi_k \,. \qquad (13)$$

Subject to the normalization constraint $\sum_k \pi_k = 1$. we now do this by computing the Lagrange and setting its partial derivatives to zero:

$$\mathrm{l} = \sum_{i=1}^{N}\sum_{k=1}^{K} r_k^{(i)} log\pi_k + \lambda\left(1 - \sum_{k=1}^{K}\pi_k\right) \qquad (14)$$

$$\frac{\partial \mathrm{l}}{\partial \pi_k} = \frac{\sum_{i=1}^{N} r_k^{(i)}}{\pi_k} - \lambda \qquad (15)$$

Setting the partial derivative to zero, we have

$$\lambda = \frac{\sum_{i=1}^{N} r_k^{(i)}}{\pi_k} = 0 \qquad (16)$$

Initialization and Convergence Issues for EM algorithm

The EM algorithm (Dempster et al., 1977) starts by initializing the algorithm with a set of initial parameters and then applying E-step, or starting with a set of initial weights followed by doing a first M-step. We can choose the initial parameters or weights randomly (i.e select k random points as initial means and select the covariance matrix from the whole set of data for each of the initial k covariance matrices).

Technically, the idea of EM is to iteratively define a lower bound on the log-likelihood and to maximize this lower bound. The major advantages of the EM algorithm over other parameter estimation techniques like *gradient* and *sampling based methods* are (i) no parameters that are set

as initial guesses; influence the optimization algorithms, and (ii) its ease of implementation. Its biggest drawback, like every other local-optimization methods, is the sensitivity of the found solution to the initial parameter values. This is because EM is sensitive to initial conditions and might converge to a local optimum instead of the global one.

Convergence is detected generally by computing the value of the log-likelihood after each iteration is implemented, and we stop when it appears the optimization is no longer changing in a significant manner from one unique solution to the other or $/\theta_i - \theta_{i+1}/ \leq 0.0001$.

Summary of EM Algorithm for two components Gaussian mixture.

Following (Wu, 1985)**:**

Take initial guesses for the parameters $\hat{p}$; $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2$.

Both σ_1^2 and σ_2^2 can be set equal to the overall sample variance $\sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{N}$. The mixing proportion π_k can be initialized at the value of 0.5.

E- STEP: Compute the posterior probabilities

$$\hat{\gamma}_i = \frac{\hat{p}\varphi_{\theta_2}(x_i)}{(1-\hat{p})\varphi_{\theta_1}(x_i) + \hat{p}\varphi_{\theta_2}(x_i)} \, , i = 1,2, \dots, n$$

M- STEP:

Compute the weighted means and variances

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N}(1-\hat{\gamma}_i)x_i}{\sum_{i=1}^{N}(1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^{N}(1-\hat{\gamma}_i)(x_i-\hat{\mu}_1)^2}{\sum_{i=1}^{N}(1-\hat{\gamma}_i)}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{N}\hat{\gamma}_i x_i}{\sum_{i=1}^{N}\hat{\gamma}_i}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^{N}\hat{\gamma}_i(x_i-\hat{\mu}_1)^2}{\sum_{i=1}^{N}\hat{\gamma}_i}$$

And recall, the mixing probability is given as

$$\pi_k = \frac{n_k}{N}, k = 1, 2$$

Iterate Steps 2 and 3 until convergence is achieved when all $|\theta_{i+1} - \theta_i| \leq 0.0001$.

Results of the EM algorithm

Table 2: Results of model parameter estimation for dataset 1

Iterations	π_1	π_2	μ_1	μ_2	σ_1^2	σ_2^2
0	0.5000	0.5000	75.2768	67.9795	40.3073	61.0930
1	0.4981	0.5019	75.3350	67.9493	41.0496	55.4302
2	0.4983	0.5017	75.3848	67.8974	41.3783	58.3558
3	0.4986	0.5014	75.4320	67.8452	41.4363	57.5626
4	0.4989	0.5011	75.4786	67.7949	41.3410	56.9285
5	0.4990	0.5010	75.5248	67.7465	41.1599	56.3836
6	0.4991	0.5009	75.5705	67.6995	40.9318	55.8891
7	0.4992	0.5008	75.6155	67.6536	40.6802	55.4236
8	0.4992	0.5008	75.6598	67.6088	40.4165	54.9748
9	0.4993	0.5007	75.7033	67.5650	40.1474	54.5382
10	0.4993	0.5007	75.7458	67.5222	39.8780	54.1107
11	0.4993	0.5007	75.7730	67.4805	39.6107	53.6923
12	0.4993	0.5007	75.8277	67.4399	39.3455	53.2812
13	0.4994	0.5006	75.8669	67.4004	39.0850	52.8791
14	0.4994	0.5006	75.9049	67.3622	38.8303	52.4871
15	0.4994	0.5006	75.9415	67.3253	38.5816	52.1052
16	0.4994	0.5006	75.9416	67.3252	38.5815	52.1053

Table 3: Results of model parameter estimation for dataset 2

Iterations	π_1	π_2	μ_1	μ_2	σ_1^2	σ_2^2
0	0.5000	3.1808	3.2013	3.2013	0.3864	0.3045
1	0.4998	3.1625	3.2145	3.2145	0.4212	0.2791
2	0.4974	3.1433	3.2333	3.2333	0.4571	0.2491
⋮	⋮	⋮	⋮	⋮	⋮	⋮
85	0.4998	3.2952	3.2952	3.0308	0.4210	0.2791
86	0.4998	3.2953	3.2953	3.0308	0.3212	0.2791

Discussion

We have extensively described an EM Algorithm for fitting a finite Gaussian mixture model. MATLAB source codes designed and used for the Algorithm can be obtained from the author strictly on request.

Using data Set 1

After implementation of the EM algorithm source codes using data set 1, the iterated maximum likelihood estimates for the parameters are contained in Table 2.

To achieve convergence, 17 iterations were required for optimality criterion, with log-likelihood = -3492.46. See source codes in appendix F (unattached) for implementation.

Dataset 1 initialization values: $\pi_1 = 0.5,\ \pi_2 = 0.5, \mu_1 = 75.2768, \mu_2 = 67.9795,\ \sigma_1 = 6.3488,$ $\sigma_2 = 7.8162$. Dataset 1 optimize values $\pi_1 = 0.4994, \pi_2 = 0.5006,\ \mu_1 = 75.9416, \mu_2 = 67.3252, \sigma_1 = 6.2114, \sigma_2 = 7.2184$.

Using data set 2

Also, having implemented the EM algorithm source codes using data set 2, the iterated maximum likelihood estimates for the parameters are contained in Table 3 of this paper. To iteratively achieve convergence, 87 iterations were required for optimality criterion to be met, with log-likelihood $= -170.968$. See the derivation source codes in appendix F (unattached).

Dataset 2 initialization values: $\pi_1 = 0.5, \pi_2 = 0.5, \mu_1 = 3.1808, \mu_2 = 3.2013, \sigma_1 = 0.6216,$ $\sigma_2 = 0.5518$

Dataset 2 optimized values $\pi_1 = 0.4998, \pi_2 = 0.5002, \mu_1 = 3.2953 = 3.0308, \sigma_1 = 0.4212, \sigma_2 =$ 0.2791.

These obtained optimized values for Data 1 & 2 are the optimal improvements of the initial parameter values, and hence cannot be statistically improved upon, and can now be used to conduct classification on the original data sets.

Summary

We have been able to use GMM to identify the two Components of a population of interest in this Study. The EM algorithm was used to estimate parameters of the GMM. The Practical example used in this study is the weights of Students at Abia State Polytechnic, Aba. Here, the weights of the male and female students were pooled before analysis**.**

Conclusion

We hereby conclude that we can indeed Apply EM algorithm in the estimation of the parameters of Gaussian Mixture Model (GMM.).In this Study, it was powered by the use of statistical package.

Recommendations

We make the following Recommendations;

- Study of this nature should be extended to other distributions that are not normally distributed e.g are Poisson distribution, geometric distribution, Weibull distribution, Bernoulli distribution, Binomial distribution, Uniform distribution ,etc
- Comparative study should be carried out to the use of other parameter estimation techniques like, (i) the step-size for gradient based algorithms, and (ii) sampling based methods, however these are not treated within this paper. These will enable researchers,

through comparing and contrasting, to logically decipher which of these parameter estimation methods will give the best result estimates.

References

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum-Likelihood from incomplete data via the EM Algorithm, *Journal of the Royal Society, Series B* 39(1): 1-38.

Huber, P.J. (1964).Robust Estimation of a location parameter, *Annals of Mathematical Statistics*, 35, 73-101

Kiefer, J. &Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Annals of Mathematical Statistics*, 32, 1350-1360

Kuroda, M. (2021). Fast computation of the EM Algorithm for Mixture Models. Computational Statistics and Applications, DOI: 10.5772/intechopen.101249.

Lindsay, B.G. (1995). Mixture Models theory, geometry and applications, Institute for Mathematical Statistics, Hayward.

McLachlan, G.J. & Krishna, T. (1997). The EM Algorithm and Extensions, *Wiley and sons Inc*

McLachlan, G.J. & Peel, D. (2000a). Finite Mixture Models, inference and applications to clustering, *Marker Dekker NY*

McLachlan, G.J. & Peel, D. (2000b). Construction of Gaussian Mixture distribution using MATLAB, *Wiley and sons Inc*

Oleszak, P. (2020). Estimating the economy with finite mixture models and the EM Algorithm, DOI: d975d280dbc6

Pearson, K. (1894). Construction to the theory of mathematical evolution, A186, 71-110

Richard, A.J. and Dean, W.W. (2007). Applied Multivariate Statistical Analysis of Finite Mixture Distributions, *Wiley and sons Inc*

Titherington, D.M., Smith, A.F.M., Markov, U.E. (1985). Statistical Analysis of Finite Mixture Distributions, *Wiley and sons Inc*

Turkey, J. (1960). A survey of sampling from contaminated distributions to probability and statistics, *Wiley and sons Inc*

William, H.P. (2008). Computational Statistics with application to Bio-informatics, Unit H: Gaussian mixture models and EM methods using MATLAB.

Wirjanto, T.S., & Xu D. (2009). The applications of Mixture of Normal Distributions in Empirical Finance, A Selected Survey, Chinese Economic Society (CES)

Wu, C.F. (1985). On the convergence properties of the EM Algorithm, *Annals of Mathematics and Statistics*, 11, 95-103

Author Biography

Dr. Kalu, Stephen Iheanyichukwu is a Statistics lecturer, in the Department of Statistics, Abia State Polytechnic Aba, Nigeria. He holds a PhD in Statistics and is currently a Principal Instructor by rank. His specialized realms of scientific research are “Statistical distribution theory mixture modeling optimization and classification in artificial intelligence (AI) via deep machine learning (DML)”. Email: monarchynwokike@gmail.com, Orcid number: https://orcid.org/0009-0005-3563-2352