International Journal of Statistics and Applications 2023, 13(1): 20-26
DOI: 10.5923/j.statistics.20231301.03

Application of EM Algorithm in Classification Problem & Parameter Estimation of Gaussian Mixture Model

Felix N. Nwobi[1], Stephen I. Kalu[1,2,*]

[1]Department of Statistics, Imo State University, Owerri, Nigeria
[2]Department of Statistics, Abia State Polytechnic, Aba, Nigeria

Abstract The problem of misclassification in a *k-group* scenario which involves identifying any component from which each group elements is drawn from is studied in this paper. These problems were modeled as a Gaussian mixture model while, the expectation maximization algorithm (EM) was used in the estimation of the parameters for the identification of the group where each group elements is drawn from. Two data sets were used in this paper; the weights of 1000 students and the weights of 200 babies at birth. Results show that 70% correct classification rate, attributing 30% to misclassification using data set 1 and 74.5% correct classification rate, attributing 25.5% to misclassification using data set 2, were achieved.

Keywords Mixture Models, Estimation in mixture models, Gaussian mixture models, EM Algorithm, Classification in Gaussians

1. Introduction

According to Lindsay (1995), Mixture Models (MM) as probabilistic models are used for representing the presence of population subsets located within an overall population. Mixture models provide a much wider range of modeling possibilities and capabilities than the single component model. Some details of MM can be found in Bohning and Seidel (2003) and Bohning et al. (2007).

For estimation involving mixture models, various analytical methods have been developed for estimating $\underset{\sim}{\theta}$ (a parameter space) in finite mixture models. There are many methods of estimating the parameters of a MM, four of such methods are method of moments, minimum distance method, Bayesian method and maximum likelihood (ML) method.

Pearson's (1894) method of moments is one of the earliest methods for estimating the parameters from finite mixture models. This method held sway until the advent of modern computers to compute the maximum of the log-likelihood function. Some developments in the method of moment estimation can be found in Lindsay & Basak (1993), Furman & Lindsay (1994a, b), Lindsay (1995), Withers (1996) and Craigmile & Titherington (1998).

Minimum distance estimation methods introduced by Wolfowitz (1957) is a general method for estimating $\underset{\sim}{\theta}$ in a finite mixture, have been discussed by William et al. (1982) and Titherington et al. (1985).

* Corresponding author:
steve_karlng@yahoo.com (Stephen I. Kalu)
Received: Nov. 7, 2022; Accepted: Jan. 21, 2023; Published: Jul. 24, 2023
Published online at http://journal.sapub.org/statistics

Another method for estimating $\underset{\sim}{\theta}$ is the Bayesian method. Many reasons abound why some researchers are inclined to using Bayesian method of estimation while dealing with a finite MM (Fruhwirth-Schnatter, 2006). The reasons for these are the introduction of a suitable prior distribution or $\underset{\sim}{\theta}$ that eliminates spurious modes in the course of maximizing the log-likelihood function, and secondly, in the case where the posterior distribution for the unknown parameters is handy, this method provides reliable inference without recourse to the asymptotic normality of the distributions. These are the inherent advantages associated to this method especially when the sample size n is small $(n < 30)$, since the asymptotic theory of MLE can be implemented when n is quite large $(n \geq 30)$.

The fourth method for estimating the parameters of a finite MMs is the ML Estimation method. Likelihood based inference has enjoyed tremendous and fast developments and has contributed immensely towards resolving estimation problems involving finite MMs. Since the explicit expression for the MLE's are typically unavailable, then a numerical EM Algorithm is used for maximizing the log-likelihood function. The expectation-Maximization (EM) Algorithm is one of the methods frequently in use according to Dempster et al. (1977). We can find more details about this methods in McLachlan and Krishna (1997), McLachlan and Peel (2000), Oleszak (2020), Kuroda (2021) and Smyth (2021).

The aim of this paper is to implement classification procedure on the Gaussian mixture model involving 2 groups $(k = 2)$ and to identify the component from which each group elements probably belongs to, and as well as to estimate the respective parameters of the groups and their mixture weights. The EM Algorithm procedures were implemented in this paper using MATLAB.

2. Methods

Let X be a random variable from a normal population. Let also $x_1, x_2, ..., x_N$ be a random sample from X that constitute two groups, such that

$x_1, x_2, ..., x_n \in X_1; x_{n+1}, x_{n+2}, ..., x_{N-1}, x_N \in X_2$ where $X = (X_1, X_2)$.

According to Wirjanto (2009), we assume that $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$. The Gaussian distribution is defined as

$$f_X(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad (1)$$

$-\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$, so that the probability density function of mixture of Gaussians will be given as

$$f(x_j) = \sum_{i=1}^{k} \pi_i \phi(X_j; \mu_i; \sigma_i^2) \quad (2)$$

where $\sum_{i=1}^{k} \pi_i = 1$ and $i = 1, ..., k,$ π_i are the mixing parameter weights.

In two component Gaussian mixture models, $k = 2$ and $\phi(X_j; \mu_i, \sigma_i^2)$ is the PDF of a normal distribution with finite mean μ_i and finite variance σ_i^2. The number of estimable parameters of the Gaussian mixture distributions is given by the formula $3k - 1$ so that if $k = 2$, as in our case, $\underline{\theta} = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi\}$.

2.1. Implementation Procedure of EM Algorithm

Dempster et al (1977) and Wu (1985) have proposed the use of initial guesses for the parameters $\hat{\pi}, \hat{\mu}_j, \hat{\sigma}_j^2$ ($j = 1, ..., k$). Soderlind (2010) in line with Dempster et al (1977) and Wu (1985) suggested using $\mu_1 = x_1, \mu_2 = x_2$. Both σ_1^2 and σ_2^2 can be set equal to the overall sample variance $\sum_{i=1}^{n} \frac{\left(x_j - \bar{x}\right)^2}{n}$. Here, we take $\mu_2 = \sum X_j / n_2$, similarly for S_i^2. We initialize the mixing proportion at $\pi_i = \frac{1}{2}$.

(a) e-step:

Following Huber (1964), for a given number of an observed data, we can evaluate the corresponding posterior probabilities, called responsibilities as follows

$$\gamma_i = \frac{\pi_i \phi\left(x_j \mid \mu_i, \sigma_i^2\right)}{\sum_{i=1}^{k} \pi_i \phi\left(x_j \mid \mu_i, \sigma_i^2\right)} \quad (3)$$

for $i = 1, ..., k$ and $j = 1, ..., k$

(b) m-step:

Now, we compute the weighted means and variances using the obtained responsibilities in (3) as

$$\text{(i)} \quad \hat{\mu}_j = \frac{\sum_{j=1}^{k} (\hat{\gamma}_i) x_j}{\sum_{j=1}^{k} \hat{\gamma}_j}, \ j = 1, ..., k(groups) \text{ and } i = 1, 2 \quad (4)$$

$$\text{(ii)} \quad \hat{\sigma}_j^2 = \frac{\sum_{i=x_1}^{x_n} \hat{\gamma}_i (x_j - \hat{\mu}_1)^2}{\sum_{i=1}^{N} \hat{\gamma}_i}, \ j = 1, i = x_1, ..., x_n \quad (5)$$

$$\text{(iii)} \quad \hat{\sigma}_j^2 = \frac{\sum_{i=x_{n+1}}^{x_N} (1 - \hat{\gamma}_i)(x_j - \hat{\mu}_2)^2}{\sum_{i=x_{n+1}}^{x_N} (1 - \hat{\gamma}_i)} \quad j = 2, i = x_{n+1}, ..., x_N \quad (6)$$

The mixing probability is computed as;

$$\text{(iv)} \quad \pi_i = \frac{n_i}{N}, i = 1, 2. \quad (7)$$

Naturally, iteration continues until convergence is achieved. Convergence is generally achieved by evaluating the log-likelihood after performing each iteration and stop further iteration when it appears that the log-likelihood is not changing in a significant manner from one iterative step to another. In Kiefer & Wolfowitz (1956), under the independent and identically distributed (iid) assumption, the log-likelihood is defined as follows:

$$\log l(\theta) = \sum_{i=1}^{N} \log P(x_j \mid \theta) = \sum_{i=1}^{N} (\log \sum_{k=1}^{K} \pi_i \phi(x_j; \mu_i, {\sigma_i}^2)) \quad (8)$$

The up-dated mixture mean and variance are obtained if we have x_j as a random variable with a two component Gaussian mixture as follows: Representing the mixture mean weight as μ_m and that of the mixture variance weight σ_m^2, then the respective mixture mean and variance weights can be estimated respectively as

$$\mu_m = \pi_1 \mu_1 + \pi_2 \mu_2 \quad (9)$$

$$\sigma_m^2 = \pi_1(\sigma_1^2 + \mu_1^2) + \pi_2(\sigma_2^2 + \mu_2^2) - \mu_m^2 \quad (10)$$

where μ_m and σ_m^2 is the up-dated mean and variance of the Gaussian mixture after each iterative step of the EM Algorithm.

2.2. Classification with Gaussians

We used Bayes' Theorem for our problem to relate the probability density function of the data, x_j, given the class to the posterior probability or the class given the data. Considering our univariate data consisting of a set of random variable X_j, whose PDFs, given k, are Gaussians with means μ_i and variances σ_i^2. Using Bayes' theorem, we specify the component probability density function as;

$$P(k \mid X_j) = \pi_i P(X_j \mid k) = \pi_i \phi(X_j; \mu_i, \sigma_i^2) \qquad (11)$$

where $P(X_j \mid k)$ is the likelihood of class k given observation X_j. Probability of misclassification is a measure of the likelihood that individuals or objects are classified wrongly. We have two types of misclassification error as;

1. To classify into population π_i given that it is actually from population π_j, $i \neq j$.
2. To classify into population π_j given that it is actually from population π_i.

Following Richard et al. (2007), we classify the classification against the true group by creating a Statistician's Confusion Matrix as in Table 1. Apparent error rate can be defined as the measure of performance in classification that does not depend on the form of the parent population. This rate is the fraction of observed values in the training sample that are misclassified by the sample classification function and can be calculated for any classification procedure. The confusion matrix is of the form

Table 1

		Predicted membership π_1	Predicted membership π_2	
Actual membership	π_1	n_1c	$n_1m = n_1 - n_1c$	n_1
	π_2	$n_2m = n_2 - n_2c$	n_2c	n_2

where

$n_1c =$ Number of π_1 objects that are correctly classified as π_1 objects

$n_1m =$ Number of π_1 objects misclassified as π_2 objects

$n_2c =$ Number of π_2 objects that are correctly classified into π_2 objects

$n_2m =$ Number of π_2 objects that are misclassified into π_1 objects.

The formula error rate is given as

$$APER = \left(\frac{n_1m + n_2m}{n_1 + n_2} \right) \times 100 \qquad (12)$$

See Appendix G for the MATLAB Source codes of obtaining the relevant quantities in Table 1.

3. Results and Discussions

Two data sets were used. Data Set 1are weights of male and female first year students of Abia State Polytechnic, Aba. Data Set 2 consists of weights of 120 males and 80 female babies at birth from Federal Medical Centre (FMC), Owerri, Imo State. See data on Appendix A, B and C. In this section, we used MATLAB source codes to implement the EM Algorithm procedures and as well, carry out the data classification analysis.

3.1. Result of the Expectation Step

Using Data set 1, taking initial values for $\mu_1 = 75.2768, \mu_2 = 67.9795, \sigma_1 = 6.3488, \sigma_2 = 7.8162$. We also take the overall mean $\mu_{m_1} = 71.6282$. X_j $(j = 1)$ is an $N \times 1$ column vector of the combined weights of male female students. We generate the probabilities for each of the data points. This step helps us in allocating the distinct mixture observations to previously defined groups. See equations (1) - (7) in page 3-4. Using Data set 2, the Algorithm was initialized with the following parameter values, $\pi_1 = 0.5, \pi_2 = 0.5, \mu_1 = 3.1808, \mu_2 = 3.2013, \sigma_1 = 0.6216$, $\sigma_2 = 0.5518$, taking its overall initial mean as $\mu_{m_2} = 3.1885$. In this case, X_j $(j = 2)$ is a $N \times 1$ column vector of the combined baby weights at birth. Data set 1 consist of $n_1 = n_2 = 500$ while in Data set 2, $n_1 \neq n_2$ since, $n_1 = 120$ and $n_2 = 80$.

3.2. Results of the Maximization Step

For us to obtain the log-likelihood and the mixing proportions using Data sets 1 & 2, we applied the MATLAB code in Appendix F. This approach, maximizes the E-Step and outputs the optimal mixing weights using data set 1 as $\pi_1 = 0.4994$, $\pi_2 = 0.5006$ and component means as $\mu_1 = 75.9416$, $\mu_2 = 67.3252$ & component variances as $\sigma_1^2 = 38.5815$, $\sigma_2^2 = 52.1053$. Using data set 2 in the same manner, we obtained the final requisite parameter up-dates as $\pi_1 = 0.4998$, $\pi_2 = 0.5002$, component means as $\mu_1 = 3.2953$, $\mu_2 = 3.0308$, and component variances as $\sigma_1^2 = 0.4212$, $\sigma_2^2 = 0.2791$. See equations (9) - (10) in page 4 of this paper.

3.3. Results of the EM Algorithm

Data set 1

After our implementation of the EM algorithm 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 for implementation.

Data set 2

Also, having implemented the EM algorithm using data set 2, the iterated maximum likelihood estimates for the parameters are contained in Table 3.

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.

3.4. Description of the Iteration Procedure Using Data Set 1&2

To implement the EM Algorithm iterative procedure to our data sets, we applied the following sequence of operation

INITIALIZATION

Data set 1 initial values:

$$\pi_1 = 0.5, \pi_2 = 0.5, \mu_1 = 75.2768, \mu_2 = 67.9795,$$
$$\sigma_1 = 6.3488, \sigma_2 = 7.8162$$

Data set 2 initial values:

$$\pi_1 = 0.5, \pi_2 = 0.5, \mu_1 = 3.1808, \mu_2 = 3.2013,$$
$$\sigma_1 = 0.6216, \sigma_2 = 0.5518$$

Expectation step:

1. Input: (Slot in the initial values for either data set 1 or 2 into the source codes of Appendix D)
   Output: A set of $k-groups$ with weights that maximizes the log-likelihood function of equation (8) will be generated.

Maximization step:

2. Update the mixture model parameters with the computed output weights from E-step using the MATLAB source codes in Appendix E.
3. Stopping criteria: If stopping rule are satisfied (convergence of parameters and log-likelihood) then we stop, else we set $j = j+1$ and go back to step 1 and input the updated parameter values for the next iteration. $j = 0,1,...,m$. Where m is the n^{th} number at which convergence or optimality conditions was achieved. Stopping Rule: When all the epsilon $(|eps|)$ values are less than or equal to 0.0001 $(i.e\,|eps| \le 0.0001)$, or the values of all $(|eps|)$ appears not to be changing significantly from one iterative step to another, then we assume that the EM solution is optimal at that point and cannot be improved upon further. The values of the complete iterations are contained in Table 3 and Table 4 of this paper.

Table 2. Results of estimation of model parameters using data set 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 estimation of model parameters using data set 2

Iterations	π_1	π_2	μ_1	μ_2	σ_1^2	σ_2^2
0	0.5000	0.5000	3.1808	3.2013	0.3864	0.3045
1	0.4998	0.5002	3.1625	3.2145	0.4212	0.2791
2	0.4974	0.5026	3.1433	3.2333	0.4571	0.2416
3	0.4998	0.5002	3.1625	3.2145	0.4212	0.2791
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
83	0.4997	0.5003	3.2952	3.0211	0.4160	0.2912
84	0.4998	0.5002	3.2972	3.0307	0.4210	0.2667
85	0.4998	0.5002	3.2952	3.0308	0.4210	0.2791
86	0.4998	0.5002	3.2953	3.0308	0.4212	0.2791

3.5. Result of the Classification Using Data Set 1

After implementing the classification procedures on the maximized generated posterior probabilities using equation (11) & (12), we obtained the following confusion matrix. See Appendix G for the MATLAB's source codes for deriving the discriminant values of Table 4 and Table 5.

Table 4

	π_1	π_2
π_1	380	120
π_2	180	320
Total	$n_1 = 500$	$n_2 = 500$

From the obtained confusion matrix in Table 4, we can now compute the apparent error rate to determine the probability of misclassification considering the weights of students. Using equation (12). The apparent error rate is computed as follows:

$$APER = \frac{n_1 m + n_2 m}{n_1 + n_2} = \left(\frac{120+180}{500+500}\right) \times 100 = 30\%$$

3.6. Result of the Classification Using Data Set 2

Table 5

	π_1	π_2
π_1	88	19
π_2	32	61
Total	$n_1 = 120$	$n_2 = 80$

Likewise from Table 5, we compute the APER to determine the probability of misclassification considering the babies weights at birth.

$$APER = \frac{n_1 m + n_2 m}{n_1 + n_2} = \left(\frac{19+32}{120+80}\right) \times 100 = 25.5\%$$

4. Conclusions

From this paper, we explained the intricacies of how to estimate the parameters of two combined Gaussian models as well as their classifications using EM Algorithm procedure. We also explained exhaustively the actual estimation of these model parameters using sets of sample data, namely data set 1 for the weights of students and data set 2 for the weights of babies at birth.

Table 2 of our analysis displayed the estimated maximized values of the relevant parameters for the Gaussian mixture model using data set 1, while Table 3 showed the estimated maximized values of the parameters for the mixture model using data set 2. Having a look at the both ends of Table 2 & Table 3, revealed that convergences have been achieved, since the parameter estimates stopped changing significantly at those points and $(|eps| \le 0.0001)$. Table 4 presents the result of classification using data set 1, whose interpretation implies that we misclassified data set 1 by about 30% failure rate, attributing 70% to correct classification rate. Table 5 at the other hand showed that we misclassified the data set 2 by about 25.5%, having a classification success rate of 74.5%. However, the overall classification efficiency of the Algorithm based on the output of sample data can be considered high, since 70% of Data set 1 and 74.5% of Data set 2 were correctly classified by the Algorithm. These validations achieved by the sample data procedures suggest that EM Algorithm may be useful as early warning statistical tool for parameter estimation and for predicting and classifying the mixture of Gaussians.

Appendix D

The E-step: See William (2008) for Appendix D, E and F.

$>> mu = [\ \mu_1\ \ \mu_2\ ];$

$>> sigma = [\ \sigma_1\ \ \sigma_2\ ];$

$>> pr = @(X)\exp(-0.5*((X - mu)./sigma).\wedge 2)./sigma;$

$>> pr(\mu_m);$ % probability of the entire data average

$>> prn = @(X)\,pr(X)./sum(pr(X));$ % non-normalized probability values whose sum is $\neq 1$.

$>> pr(\mu_m);$ % normalized probability values whose sum is equal to 1.

$>> prns = ([prod(size(X)),1]);$

$>> for\ \ j = 1: prod(size(X)); prns(j,:) = prn(X(j)); end;$

$>> prns(1:10,:)$ % outputs the normalized posterior probabilities from 1 to 10. Hence, in our Algorithm, we replaced 10 with 1000 or 200 as the case may require. Note: X is the entire data for student weights or baby weights at birth. $pr(\mu_m)$ is the entire mean of either data set 1 or data set 2 as the case may also demand.

Appendix E

The M-step source codes

$>> mu = sum(prns.*repmat(X,[1,2]./sum(prns,1);$

$>> Xmmu = repmat(X,[1,2]) - repmat(mu,[prod(size(X)),1]);$

$>> sigma = sqrt(sum(prns.*Xmmu.\wedge 2,1)./sum(prns,1))$

$>> pop = sum(prns,1)/prod(size(X))$

Appendix F

MATLAB's Source codes for the Log-likelihood and the mixing proportions.

$>> options = statset('Display','iter');$

$>> obj = gmdistribution.Fit(X,2,'Options',options)$

Appendix G

MATLAB's Source codes for the classification/Discriminant Analysis:

For Data Set 1

$>> prns(1:1000,:);$ % generates the entire probabilities from 1 to 1000.

$>> A = prns(1:500,:);$% Extract the probability values from 1 to 500 and assign it to A.

$>> A_1 = A(:,[1]);$% Extract the first column probability values from 1 to 500 and assign to A_1.

$>> B = prns(501:1000,:);$% The probability values from 501 to 1000 and assign to B.

$>> B_1 = B(:,[1]);$% the first column probability values from 501 to 1000 and assign to B_1.

>> Number of $n_1c = length(find(A_1 >= 0.5))$% Number of correct classification into π_1.

>> Number of $n_1m = length(find(A_1 < 0.5))$% Number of misclassification into π_2.

>> Number of $n_2c = length(find(B_1 >= 0.5))$% Number of correct classification into π_2.

>> Number of $n_2m = length(find(B_1 < 0.5))$% Number of misclassification into π_1.

For Data Set 2

$>> prns(1:200,:);$ % generates the entire probabilities from 1 to 200.

$>> C = prns(1:120,:);$% Extract the probability values from 1 to 120 and assign it to C.

$>> C_1 = C(:,[1]);$% Extract the first column probability values from 1 to 120 and assign to C_1.

$>> D = prns(121:200,:);$% The probability values from 121 to 200 and assign to B.

$>> D_1 = D(:,[1]);$% the first column probability values from 121 to 200 and assign to D_1.

>> Number of $n_1c = length(find(C_1 >= 0.5))$% Number of correct classification into π_1.
>> Number of $n_1m = length(find(C_1 < 0.5))$% Number of misclassification into π_2.
>> Number of $n_2c = length(find(D_1 >= 0.5))$% Number of correct classification into π_2.
>> Number of $n_2m = length(find(D_1 < 0.5))$% Number of misclassification into π_1.

REFERENCES

[1] Bohning, D. and Seidel, W. (2003). Editorial: Recent developments in mixture models, computational statistics & Data Analysis 41: 349-357.

[2] Bohning, D., Seidel W., Alfo, M., Garel, B., Patilea, V., and Walter, G. (2007). Editorial: Advances in mixture models, Computational Statistics and Data Analysis 51: 5205-5210.

[3] Craigmile, P.F. & Titherington, D.M (1998). Parameter Estimation for finite mixtures of uniform distributions, Communication in Statistics-Theory and methods 26: 1981-1995.

[4] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum-likelihood from incomplete data via the EM algorithm". Journal of the Royal Statistical Society, Series B 39 (1): 1-38 (https://www.jstor.org/stable/2984875).

[5] Fruhwirth-Schnatter, S. (2006). Finite Mixture and Markov Switching Models. New York/Berlin/Heidelberg: Springer.

[6] Furman, W.D. and Lindsay, B.G. (1994a). Testing for the number of components in a mixture of normal distributions using moment estimators. Computational Statistics and Data Analysis 17:473-492.

[7] Furman, W.D. and Lindsay, B.G. (1994b). Measuring the effectiveness of moment estimators as starting values in maximizing mixture likelihoods. Computational Statistics and Data Analysis 17. 493-507.

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

[9] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Annals of Mathematical Statistics 20, 1350-1360.

[10] Kuroda, M. (2021). Fast computation of the EM Algorithm for mixture Models. Computational Statistics and Application. DOI:10.5772/intechOpen.101249.

[11] Lindsay, B.G. (1995). Mixture models. Theory, geometry and applications. Institute for mathematical statistics, Hayward, Califonia (Download: Zipped pdf).

[12] Lindsay, B.G. and Basak, P. (1993). Multivariate Normal Mixtures: a fast consistent method of moments. Journal of the American Statistical Association 86: 96-107.

[13] McLachlan, G.J. and Krishna, T. (1997). The EM Algorithm and Extensions. N.Y: Wiley and sons Inc.

[14] McLachlan, G.J. and Peel, D. (2000). Finite mixture models, inference and application to clustering. Marcel dekker, New York. (a)

[15] MacLachlan, G.J and Peel, D (2000). Construction of Gaussian mixture distribution using MATLAB – Infinite mixture models. Hoboken, NJ: John Wiley & Sons, Inc. (b)

[16] Oleszak, P. (2020). Estimating the economy with finite mixture models and the EM Algorithm. Htt:/towardsdatascience.com/estimating-the-state-of-the-economy-with-finite-mixture-models-and-the-em-algorithm-d975d280dbc6.

[17] Pearson, K. (1894). Contribution to the theory of mathematical evolution, Phil.Trans. Roy. Sco. London A 186, 71-110.

[18] Richard, A, J. and Dean, W.W. (2007). Applied Multivariate Statistical Analysis, 16th Edition, Pearson education Inc., New Jersey, USA.

[19] Smyth, P. (2021). Mixture Models and EM Algorithm. htts://www.ics.uci.edu/~smyth/couses/cs274/notes/mixture_models_EM.pdf.

[20] Soderlind, P. (2010): Lecture Notes On Empirical Finance (PhD). Return Distributions. University of St. Gallen.

[21] Titherington, D.M., Smith, A.F.M., Makov, U.E. (1985). Statistical Analysis of Finite Mixture Distributions. John Wiley and sons, N.Y: Stanford Press.

[22] William C. Parr, William R. Schucany (1982). Minimum Distance Estimation of Goodness-of-fit Statistics. Journal of the Royal Statistical Society, Series B (Methodological). Vol.44, No.2, PP. 178-189, Published by Wiley.

[23] William H.P. (2008). Computational statistics with application to Bio-informatics – unit II: Gaussian mixture models and EM methods using MATLAB.

[24] Wirjanto, T.S., Xu Dinghai (2009). The Applications of Mixtures of Normal Distributions in Empirical Finance: A Selected Survey. Chinese Economic Society (CES).

[25] Withers, C.S. (1996). Moment Estimates for mixtures of several distributions and its different means and scales. Communication in Statistics, Theory and methods 25: 1799-1824.

[26] Wolfowitz, J. (1957).The Minimum Distance Method. The annals of Mathematical Statistics 28(1): 75-88.

[27] Wu, C.F. (1985). On the convergence properties of the EM algorithm. Annals of Mathematics and Statistics 11: 95-103.