



**Expectation Maximization Algorithm  
Application in Gaussian Mixture Models**

BY

**PRISCILLA ANGATIA SHIROKO**  
**I56/34727/2019**

A Thesis Submitted to the Department of  
Mathematics in partial fulfilment for a Degree in  
Master of Science in Mathematical Statistics

December 6, 2022

# Abstract

Gaussian mixture models are applied in machine learning specifically unsupervised machine learning. More specifically they can be used during image segmentation and music classification just to mention a few.

In this project, it is shown how the EM Algorithm is derived and how it effectively comes into use in terms of soft clustering data sets into distributions.

EM Algorithm is used to estimate parameters within a model in a fast and stable way then fills the missing data in a sample and find the values of latent variables.

The Gaussian Mixture model looks at the distributions. It groups only data points that belong to a similar distribution. This is done through soft clustering where by the points are assigned the probability of being in a certain distribution, It goes as far as clustering data points in between different distributions accurately by showing to which extent a data point falls in a particular distribution.

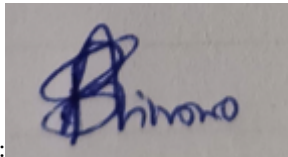
Expectation Maximum Algorithm uses the observed data to get optimum values that can be used to generate the model parameters.

Limitations anticipated within this study include;

- Expectation Maximum Algorithms have slow convergence and this convergence is made to the local optima.
- It also requires forward and backward probabilities, while numerical optimization only requires forward probability.


## Declaration and Approval

I declare that this dissertation is my original work and has not been presented for an award of a degree in any other university.



Signed: ..... Date: 06 / 12 / 2022  
Priscillah Angatia Shiroko  
I56/34727/2019

As the supervisor to this student's dissertation, I certify that this dissertation has my approval for submission.



Signed: ..... Date: **06.12.2022**  
Prof. Patrick G. O. Weke  
Department of Mathematics,  
University of Nairobi,  
Box 30197-00100, NAIROBI, KENYA.  
Email: pweke@uonbi.ac.ke

# Dedication

This project is dedicated to my daughter Ariellah Norah for being my motivation towards completing this degree, My parents and siblings for their love, encouragement and support.

# Acknowledgement

I give all thanks first and foremost to the Almighty God for giving the strength to reach this far. I also thank Prof. Patrick Weke for the guidance, correction and support given while undertaking my project. I also thank the entire Department of Mathematics and the University of Nairobi as a whole for offering all necessary support to reach this far

# Contents

<b>1</b>	<b><i>GENERAL INTRODUCTION</i></b>	<b>1</b>
1.1	Background Information . . . . .	1
1.2	Problem Statement . . . . .	1
1.3	Objectives of the Study . . . . .	2
1.4	Research Method . . . . .	2
1.5	Significance of the Study . . . . .	2
1.6	Literature Review . . . . .	3
<b>2</b>	<b><i>GAUSSIAN MIXTURE MODELS</i></b>	<b>4</b>
2.1	Gaussian Distribution (Parameter Estimation) . . . . .	4
2.2	Gaussian Mixture Models (Parameter Estimation) . . . . .	7
<b>3</b>	<b><i>DERIVATION OF THE EXPECTATION MAXIMIZATION ALGORITHM</i></b>	<b>9</b>
3.1	Log-Likelihood . . . . .	9
3.2	Convergence in the Expectation Maximization (EM) Algorithm .	12
<b>4</b>	<b><i>GENERALIZED EXPECTATION MAXIMIZATION</i></b>	<b>16</b>
4.1	Likelihood for complete data . . . . .	16
4.2	The Expectation Step . . . . .	18
<b>5</b>	<b><i>EXPECTATION MAXIMIZATION(EM) IN GAUSSIAN MIXTURE MODELS</i></b>	<b>20</b>
5.1	Expectation Maximization (EM) in Gaussian Mixture Models . .	20
5.2	Illustration of the EM Steps . . . . .	21
<b>6</b>	<b><i>DISCUSSIONS AND RECOMMENDATIONS</i></b>	<b>26</b>
6.1	Discussion . . . . .	26
6.2	Study Limitation . . . . .	26

6.3 Recommendations . . . . .	26
<b>7 REFERENCES</b>	<b>27</b>

# 1 *GENERAL INTRODUCTION*

## 1.1 Background Information

A mixture can be described as a constructed probability distribution after combining two distributions or more to get a new distribution. It can be classified as either Discrete, Finite, or Continuous.

A mixture created by combining numerous Gaussian distributions is known as a Gaussian Mixture Model. It is predicated on the idea that each data point is produced by a combination of a limited Gaussian distributions' number with unknown characteristics.

Application of Gaussian mixture models is applied in machine learning specifically unsupervised machine learning. More specifically they can be used during image segmentation and music classification just to mention a few.

They use a clustering format where we try to find cluster points using unsupervised learning in the datasets that shares common characteristics. In the clustering analysis, the Expectation Maximum Algorithm is used to fill in missing data in a sample, estimate model parameters quickly and steadily, and determine the values of latent variables.

Dempster (1977) introduced the Expectation Maximum Algorithm to obtain the Maximum Likelihood Estimates of incomplete data. A broad applicable algorithm for computing maximum likelihood estimates from incomplete data was presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm was derived.

## 1.2 Problem Statement

Most research works have applied the Expectation Maximum Algorithm to obtain Maximum Likelihood Estimates for missing or incomplete data sets in Gaussian mixture models. The EM approach, which is frequently used to estimate the model's parameters, mainly relies on the estimation of insufficient data. However, it doesn't make use of any data to lessen the uncertainty caused by missing data.

This project aims to do parameter estimation for the Gaussian mixture models using the Expectation Maximum Algorithm.



### **1.3 Objectives of the Study**

The study's main objective is to estimate the parameters for the Gaussian mixture model using the EM Algorithm.

The specific objectives are to;

- I. Derive the Expectation Maximization Algorithm.
- II. Estimate Gaussian Mixture models' parameters using the EM Algorithm method.
- III. Apply the EM Algorithm in the Gaussian Mixture Model.

### **1.4 Research Method**

The breakdown of the method used for the estimation of parameters using the EM Algorithm method is given below;

- I. Derive the Expectation Maximum Algorithm.
- II. The system is provided a collection of imperfect observed data with the presumption that the observed data originates from a particular model..
- III. E – Step is employed, in which values of the missing or partial data are estimated or conjectured using the observed data. updating the variables, in short.
- IV. M – Step is applied whereby complete data sets gotten in the E – Step is used to perform the updating of the values of the parameters. In summary, updating the hypothesis.
- V. Check whether if values converge or not. If yes, then stop, otherwise, repeat steps two, three, and four until the convergence occurs.

### **1.5 Significance of the Study**

The world is changing at a fast pace and embracing technology to make life easier. Machine learning is one core aspect that has been embraced by individuals

and largely by companies to manage data and algorithms and improve accuracy in terms of data analysis.

This study seeks to play a contributory role in making readers gain a further understanding of EM Algorithms for Gaussian Mixture models and broaden the reader's understanding of unsupervised learning which is an area under machine learning.

## 1.6 Literature Review

Clustering is a method used to place data points that have similar characteristics into groups. It is a type of unsupervised learning broken down into two types; soft clustering and hard clustering. Using a probability- model based approach, it is assumed that the data follows a mixture model of probability distributions in which Expectation and Maximization Algorithm is used as stated by Yang,Lai and Lin (2012).

Mixture models can be described as a combination of multiple distributions. It used probability as a tool to project presence of sub-populations in an overall population and the observation's distribution.

A Gaussian mixture model is a type of mixture model meaning that it uses probability to assign data points to a certain number of Gaussian distributions using soft clustering. The idea of Gaussian mixtures was popularized by Duda and Hart (1973).

Expectation and Maximization Algorithm for Gaussian Mixtures performs maximum likelihood estimation with missing values. The process was introduced by Dempster, Laird and Rubin (1977) It is an iteration approach that cycles between two steps that is; estimating the missing values and optimizing the model and the two steps are repeated until convergence occurs. It is a good estimation for missing variables as will be seen in this paper.

The current values of the existing parameters are used to calculate weights, then the weighted joint log-likelihood is maximized in each iteration. In short in each procedure the expectation is maximized hence the name Expectation Maximization Algorithm.

In the field of discrete choice modelling, EM algorithms have been used by Bhat (1997a) and Train (2008a,b).

## 2 GAUSSIAN MIXTURE MODELS

### 2.1 Gaussian Distribution (Parameter Estimation)

A random variable X follows a Gaussian distribution if;

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x - \mu)^2 / 2\sigma^2}$$

(2.1)

Equation (2.1) above has 2 parameters and 1 input. The two parameters are  $\mu$  as the mean and  $\sigma^2$  as variance. The mean shifts the center of the Gaussian curve while variance measures the wideness of the Gaussian curve. When the  $\mu =$  then it means that the largest value is 1. Also, when the variance is not big but small, then the data is less spread out and is closer to the mean but when the variance is large, the data is more spread out and moves far from the mean. From the equation we are able to tell the observation probability of the input x, given a certain distribution.

For the multivariate, the probability density function is given by;

$$f(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^{|\Sigma|} |\sigma^2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

(2.2)

Where we have the input as x, mean as  $\mu$  and  $\Sigma$  as the covariance matrix. The mean is still the center of the data but this time it is vector similarly to the input. It therefore must vary similar to the input. The covariance tells each variance's dimension and the inputs relationship.

To find the maximum likelihood estimate for unknown mean  $\mu$  for a Gaussian distribution whose known variance  $\sigma^2$  is , we first;

- i) The log-likelihood.
- ii) Differentiate the log-likelihood.
- iii) Set the differentiated log-likelihood to 0.

Assuming all points are independent, then the joint likelihood which is a product of the likelihood of each point is;

$$f_X(x) = \prod_{i=1}^n f(x_i; \mu, \sigma_i^2) \tag{2.3}$$

when we refer to Equation (2.1), we get equation (2.4) below from equation (2.3) above

$$f_X(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - \mu)^2 / 2\sigma^2} \tag{2.4}$$

The log-likelihood function then becomes

$$\ln f_X(x) = \sum_{i=1}^n \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2} (x_i - \mu)^2 \tag{2.5}$$

Differentiating;

$$\frac{d}{d\mu} \ln f_X(x) = \sum_{i=1}^n \frac{1}{\sigma^2} (x_i - \mu)^2$$

(2.6)

Equating the likelihood to zero (0)

$$\sum_{i=1}^n \frac{1}{\sigma^2} (x_i - \mu)^2 = 0$$

(2.7)

We get the parameter that maximizes the likelihood to be equal to;

$$\hat{\mu} = \frac{1}{N} \sum_i x_i$$

(2.8)

The variance as seen has no effect on the estimates.

## 2.2 Gaussian Mixture Models (Parameter Estimation)

Gaussian mixture models help in modelling sets of data from several different clusters where by each cluster has different properties from the other but data points within each cluster through the Gaussian distribution can be modelled. Data points that belong in a single distribution are grouped together. Suppose from the data we have data, we want to get parameters that will maximize the likelihood of observing the data. We therefore get the probability as follows;

$$p(x) = \sum_{i=1}^n \phi_i f(x; \mu_i, \Sigma_i)$$

$$\sum_{i=1}^n \phi_i = 1$$

(2.9)

where we have the weight as  $\phi_i$ , mean as  $\mu_i$  and the covariance matrix as  $\Sigma_i$  per Gaussian. The weights sum upto 1.

We use Gaussian mixture models when we can tell which particular gaussians combination a particular point comes from and if so we can get the means and covariances.

Assume we have data from K Gaussian distributions. The likelihood then will be;

$$\ln p(x; \pi, \mu, \Sigma) = \sum_{i=1}^N \ln \left( \sum_{k=1}^K \pi_k N(x_i | \mu_k, \sigma_k) \right)$$

(2.10)

Applying the same principle as done under section 2.1, differentiation of equation (2.10) and equating the function to 0, we get optimal distribution parameter. The parameter that will maximize the likelihood is shown as;

i)

$$\pi_K = \frac{1}{N} \sum_n (Y_n = K)$$

(2.11)

From the Bayes Rule we know that;

ii)

$$\mu_K = \frac{\sum_n (Y_n = K) x_n}{\sum_n (Y_n = K)}$$

(2.12)

We apply equation (2.12) in equation (2.11) to get

iii)

$$\sigma_k^2 = \frac{\sum_n (Y_n = K) \|x_n - \mu_k\|^2}{\sum_n (Y_n = K)}$$

(2.13)

If we can not figure out which Gaussians combinations a point comes from then it makes it difficult to get the means and covariances hence we bring in the Expectation Maximization approach to estimate the weights, means and variances.

### **3 *DERIVATION OF THE EXPECTATION MAXIMIZATION ALGORITHM***

#### **3.1 Log-Likelihood**

In formulating the Expectation Maximization the log-likelihood of the complete full data set comprising of the observations made ( $X$ ) and the missing data sets ( $Z$ ) is key.

First, define the log-likelihood of the complete data set as;

$$L(X|\theta_i) = \log p(X|\theta_i) \tag{3.1}$$

The complete data pdf;

$$p(X|\theta_i) \tag{3.2}$$

can be factored as;

$$p(X|\theta_i) = \frac{p(Z, X|\theta_i)}{p(Z|X, \theta_i)} \tag{3.3}$$



Third, we use equation 3.1 and equation 3.3 to get the log-likelihood of the incomplete data set.

$$\mathbb{L}(X|\theta_i) = \log p(X|\theta_i) \sum_Z P(Z|X, \theta_i) \tag{3.4}$$

Remember that;

$$\sum_Z P(Z|X, \theta_i) = 1 \tag{3.5}$$

is differentiable in  $\theta$

$$\mathbb{L}(X|\theta_i) = \sum_Z P(Z|X, \theta_i) \log p(X|\theta_i) \tag{3.6}$$

is finite for all estimated  $\theta$

Using equation (3.3) above,

$$\mathbb{L}(X|\theta_i) = \sum_Z P(Z|X, \theta_i) \log \frac{p(Z, X|\theta_i)}{p(Z|X, \theta_i)}$$

(3.7)

is differentiable with respect to estimated  $\theta$  for fixed  $\theta$ .

$$\mathbb{L}(X|\theta_i) = \sum_Z P(Z|X, \theta_i) \log p(Z, X|\theta_i) - \sum_Z P(Z|X, \theta_i) \log p(Z|X, \theta_i)$$

(3.8)

$$\mathbb{L}(X|\theta_i) = E_Z(\log p(Z, X|\theta_i)|X, \theta_i) - E_Z(\log p(Z|X, \theta_i)|X, \theta_i)$$

(3.9)

by definition of expectation,

$$\mathbb{L}(X|\theta_i) = Q(\theta_i|\theta_i) + R(\theta_i|\theta_i)$$

(3.10)

$E_Z$  is the Expectation for the missing data sets.

$-E_Z(\log p(Z|X, \theta_i)|X, \theta_i)$  represents the incomplete-data likelihood and the expectation of the completed-data likelihood difference.

### 3.2 Convergence in the Expectation Maximization (EM) Algorithm

The Jensen's inequality mentions that if  $f_x$  is a function on  $T_X$ , and  $E(f(X))$  and  $f(E(X))$  are finite, then

$$E(f(X)) \geq f(E(X))$$

(3.11)

With the Jensen's inequality in mind and using the equation  $E_Z(\log p(Z, X|\theta_i)|X, \theta_i)$ , we show that it can be improved with each iteration that is the M-Step and so will the likelihood function.

The proof of convergence begins with the observation of the following relationship;

$$L(X|\theta) = \log p(X|\theta) = \log\left(\sum_Z P(Z, X|\theta_i)\right) = \log\left(\sum_Z P(Z|X, \theta_i) \frac{p(Z, X|\theta_i)}{p(Z|X, \theta_i)}\right)$$

(3.12)

With the equation 3.12 and the Jensen's inequality we obtain;

$$L(X|\theta) = \log p(X|\theta)$$

(3.13)

Using equation (3.3)

$$L(X|\theta) = \log \left( \sum_Z P(Z|X, \theta_i) \frac{p(Z, X|\theta)}{p(Z|X, \theta_i)} \right)$$

(3.14)

is differentiable with respect to estimated  $\theta$  for fixed  $\theta$ .

$$L(X|\theta) = \log \left( E_Z \frac{p(Z, X|\theta)}{p(Z|X, \theta_i)} | X, \theta_i \right) \geq E_Z \left( \log \frac{p(Z, X|\theta)}{p(Z|X, \theta_i)} | X, \theta_i \right)$$

(3.15)

Using equation (3.3)

$$L(X|\theta) = \sum_Z P(Z|X, \theta_i) \log \frac{p(Z, X|\theta)}{p(Z|X, \theta_i)}$$

(3.16)

is differentiable with respect to estimated  $\theta$  for fixed  $\theta$ .

$$L(X|\theta) = \sum_Z P(Z|X, \theta_i) \log p(Z, X|\theta) - \sum_Z P(Z|X, \theta_i) \log p(Z|X, \theta_i)$$

(3.17)

is continuous in  $\theta$  for fixed  $\theta$ .

$$L(X|\theta) = E_Z(\log p(Z, X|\theta)|X, \theta_i) - E_Z(\log p(Z|X, \theta_i)|X, \theta_i)$$

(3.18)

Express as;

$$L(X|\theta) = Q(\theta|\theta_i) + R(\theta_i|\theta_i)$$

(3.19)

In the Maximization step of the i-th iteration,  $\theta^*$  is chosen according to;

$$\theta^* = \operatorname{argmax}_\theta Q(\theta|\theta_i)$$

(3.20)

a  $\theta^*$  at iteration  $n$  can be chosen such that;

$$Q(\theta^*|\theta_i) \geq Q(\theta_i|\theta_i)$$

(3.21)

This proves the existence of a sufficient condition to prove the convergence property of the EM algorithm.

## 4 *GENERALIZED EXPECTATION MAXIMIZATION*

The generalized EM convergence is slow but offers a more flexible and general framework for dividing into the EM Steps, the optimization process. It is applied in cases where it is difficult to get  $\theta^*$ .

### 4.1 Likelihood for complete data

With the assumption that a set of missing data is known, then optimization of a likelihood function is the first step. But to compute the complete-data likelihood's expectation in the E-step, we begin by getting the missing data sets's expectation.

If  $X$  is equal to  $x_i$ ;  $s = 1, \dots, S$  contains  $S$  vectors that are statistically independent and  $Z = z_s \in C$ ;  $s = 1, \dots, S$ , where  $z_s = C^{(j)}$  shows that  $x_s$  is generated by the  $s$ -th mixture, then;

$p(Z, X|\theta)$  can be written as;

$$p(Z, X|\theta) = \prod_{s=1}^S p(z_s, x_s|\theta) \tag{4.1}$$

Indicator variables come in to show the status of the missing data sets

$$\Delta = (\delta_s^{(j)}; j = 1, \dots, J \text{ and } s = 1, \dots, S) \tag{4.2}$$

Since only one of the terms in  $(\delta_s^{(j)}; j = 1, \dots, J)$  is equal to one for each and all others equal to 0, we express  $p(Z, X|\theta)$  as:

$$p(Z, X|\theta) = \prod_{s=i}^S \sum_{j=i}^J \delta_s^{(j)} p(x_s, z_s|\theta)$$

(4.3)

Find the likelihood of the first set

$$p(Z, X|\theta) = \prod_{s=i}^S \sum_{j=i}^J \delta_s^{(j)} p(x_s, z_s=c^{(j)}|\theta)$$

(4.4)

Find the likelihood of the next set

$$p(Z, X|\theta) = \prod_{s=i}^S \sum_{j=i}^J \delta_s^{(j)} p(x_s, \delta_s^{(j)} = 1|\theta)$$

(4.5)

The likelihood of the completed-data together is therefore obtain as;



$$\begin{aligned}
\log p(Z, X|\theta) &= \sum_{s=i}^S \log \left( \sum_{j=1}^J \delta_s^{(j)} p(x_s, \delta_s^{(j)} = 1|\theta) \right) \\
\log p(Z, X|\theta) &= \sum_{s=i}^S \log \left( \sum_{j=1}^J \delta_s^{(j)} p(x_s|\delta_s^{(j)} = 1, \theta) P(\delta_s^{(j)} = 1, \theta) \right) \\
\log p(Z, X|\theta) &= \sum_{s=i}^S \log \left( \sum_{j=1}^J \delta_s^{(j)} p(x_s|\delta_s^{(j)} = 1, \phi^{(j)}) P(\delta_s^{(j)} = 1) \right) \\
\log p(Z, X|\theta) &= \sum_{s=i}^S \sum_{j=1}^J \delta_s^{(j)} \log(p(x_s|z_s = 1, \phi^{(j)}) \pi^{(j)}) \\
\log p(Z, X|\theta) &= \sum_{s=i}^S \sum_{j=1}^J \delta_s^{(j)} \log(p(x_s|z_s = C^{(j)}, \phi^{(j)}) \pi^{(j)})
\end{aligned} \tag{4.6}$$

With only one non-zero term in the summation  $\sum_{j=1}^J$ , we can extract  $\delta_s^{(j)}$  without affecting the result from the log function.

## 4.2 The Expectation Step

Using the expectation of equation 44 we get,

$$\begin{aligned}
Q(\theta|\theta_i) &= E_Z(\log p(Z, X, \theta_i)) \\
Q(\theta|\theta_i) &= \sum_{s=i}^S \sum_{j=1}^J E \delta_s^{(j)} | x_s, \theta_i \log(p(x_s|\delta_s^{(j)} = 1, \phi^{(j)}) \pi^{(j)})
\end{aligned} \tag{4.7}$$

We then define

$$v_i^{(j)}(x_s) = E(\delta_s^{(j)} | x_s, \theta_i) = P(\delta_s^{(j)} = 1 | x_s, \theta_i)$$

(4.8)

and denote  $\pi_i^{(j)}$  at iteration  $i$  as the  $j$ -th mixture coefficient. With the Bayes theorem, express  $v_i^{(j)}(x_s)$  as

$$\begin{aligned} v_i^{(j)}(x_s) &= P(\delta_s^{(j)} = 1 | x_s, \theta_i) \\ &= \frac{p(x_s | \delta_s^{(j)} = 1, \theta_i) P(\delta_s^{(j)} = 1 | x_s, \theta_i)}{p(x_s | \theta_i)} \\ &= \frac{p(x_s | \delta_s^{(j)} = 1, \phi_i^{(j)}) P(\delta_s^{(j)} = 1 | x_s, \theta_i)}{p(x_s | \theta_i)} \\ &= \frac{p(x_s | \delta_s^{(j)} = 1, \phi_i^{(j)}) \pi_i^{(j)}}{\sum_{k=1}^J p(x_s | \delta_s^{(j)} = 1, \phi_i^{(k)}) \pi_i^{(k)}} \end{aligned}$$

(4.9)

## 5 *EXPECTATION MAXIMIZATION(EM) IN GAUSSIAN MIXTURE MODELS*

### 5.1 Expectation Maximization (EM) in Gaussian Mixture Models

Expectation Maximization is an algorithm that is used when there are latent variables also known as missing values or incomplete data. From the existing, we find the optimum values for the missing variables and then find the parameters of the model after which step back and update the values for the latent variables. Simply put, the algorithm begins with some initial estimates which it iteratively updates through an E-step and M-step until convergence occurs.

The following is a breakdown on the steps followed in EM;

- i) An incomplete existing data set is uploaded to the system putting in mind the assumption that the existing data is from a specified model.
- ii) Expectation step also known as the E-Step: In this step we calculate the probability of a particular gaussian generating a specific point. The existing data is what we estimate or guess the value of the missing variables with. We compute as follows;

$$W_i = \frac{\phi_i f(x; \mu_i, \Sigma_i)}{\sum_{j=1}^n \phi_j f(x; \mu_j, \Sigma_j)} \tag{5.1}$$

We are looking at the probability that  $x^k$  was is gaussian i generated for an element in row k and column i with the rows as the points and the columns as the Gaussians.

- iii) Maximization step also known as the M-Step: Using the expectations generated in the E-Step, we update the parameter's values that is the means, weights and covariances.

To do a weight  $\phi_i$  update, we add the probability each point is generated by Gaussian i then perform division by the number of points.

$$\phi_i = \frac{1}{N} \sum_{i=1}^N W_i$$

(5.2)

For a mean  $\mu_i$  update, we calculate all points weighted's mean by the probability being generated by Gaussian i of that point.

$$\mu_i = \frac{\sum_{i=1}^N W_i x}{\sum_{i=1}^N W_i}$$

(5.3)

For covariance  $\Sigma_i$ , we sum of all points weighted's covariance by the probability being generated by Gaussian i of that point. Perform the same for all the Gaussian i.

$$\sum_i = \frac{\sum_{i=1}^N W_i (x - \mu_i)(x - \mu_i)^T}{\sum_{i=1}^N W_i}$$

(5.4)

iv) Check whether the values are converging or not. If yes, then stop otherwise repeat step ii and step iii until the convergence occurs.

Convergence is attained by calculating the log-likelihood value after each and every iteration and stopping when it stops making a significant change from one iteration to the other.

## 5.2 Illustration of the EM Steps

Let  $p(x_s | \delta_s^{(j)} = 1, \phi^{(j)})$  be a Gaussian distribution, from which a the j-th cluster's model parameter  $\varphi^{(j)} = (\mu^{(j)}, \tau^{(j)})$  consisting of a vector mean and a covariance matrix of full rank.

We make the assumption that the Gaussian MM:

$$\theta = (\pi^{(j)}, \mu^{(j)}, \tau^{(j)};$$

$j=1, \dots, J$

(5.5)

$\tau^{(j)}$ ,  $\pi^{(j)}$  and  $\mu^{(j)}$  and respectively denote, the covariance matrix, mixture coefficient and mean vector, of the  $j$ -th component density. It's then given by;

$$P(x_s, \theta) = \sum_{j=1}^J \pi^{(j)} p(x_s | \delta_s^{(j)} = 1, \phi^{(j)})$$

(5.6)

where

$$p(x_s | \delta_s^{(j)} = 1, \phi^{(j)}) = (2\pi)^{-\frac{i}{2}} |\tau^{(j)}|^{-\frac{1}{2}} \exp\left(\frac{-1}{2} (x_s - \mu^{(j)})^T \tau^{(j)} (x_s - \mu^{(j)})\right)$$

(5.7)

The EM iteration is as follows after the initialization of  $\theta_0$ :

i) E-step. In  $i$ -th iteration, calculate  $h_n^{(j)}(x_s)$  for  $j$  and  $s$  using Equations 5.8 and 5.6, M-step follows.

ii) M-step. Maximization of  $Q(\theta|\theta_i)$  w.r.t  $\theta$  to find  $\theta^*$  should be done. Exchange  $\theta_i$  with  $\theta^*$ . After which we increase  $i$  by 1 and redo the E-step to convergence.

To get  $\mu^{(k)*}$  we set;

$$\frac{\sigma Q(\theta|\theta_i)}{\sigma \mu^{(k)}} = 0 \tag{5.8}$$

which gives

$$\mu^{(k)*} = \frac{\sum_{s=1}^S h_i^{(k)}(x_s) x_s}{\sum_{s=1}^S h_i^{(k)}(x_s)} \tag{5.9}$$

To determine  $\tau^{(k)*}$  we set;

$$\frac{\sigma Q(\theta|\theta_i)}{\sigma \mu^{(k)}} = 0 \tag{5.10}$$

which gives

$$\tau^{(k)*} = \frac{\sum_{s=1}^S h_i^{(k)}(x_s)(x_s - \mu^{(k)*})(x_s - \mu^{(k)*})^S}{\sum_{s=1}^S h_i^{(k)}(x_s)}$$

(5.11)

We maximize  $Q(\theta|\theta_i)$  to determine  $\tau^{(k)*}$  w.r.t  $\tau^{(k)}$  subject to  $\sum_{j=1}^J \pi^{(j)} = 1$ , which brings;

$$\pi^{(k)*} = \frac{1}{S} \sum_{s=1}^S h_i^{(k)}(x_s)$$

(5.12)



Figure 1: EM Algorithm cycle



## **6 DISCUSSIONS AND RECOMMENDATIONS**

### **6.1 Discussion**

The Gaussian Mixture model looks at the distributions. It groups only data points that belong to a similar distribution. This is done through soft clustering where by the points are assigned the probability of being in a certain distribution, It goes as far as clustering data points in between different distributions accurately by showing to which extent a data point falls in a particular distribution.

Expectation Maximum Algorithm uses the observed data to get optimum values that can be used to generate the model parameters.

### **6.2 Study Limitation**

Limitations anticipated within this study include;

- Expectation Maximum Algorithms have slow convergence and this convergence is made to the local optima.
- It also requires forward and backward probabilities, while numerical optimization only requires forward probability.

### **6.3 Recommendations**

Future studies can look at estimation methods that can reduce the uncertainty of missing or incomplete data.

## **7 REFERENCES**

Kung, Mak Wang, Lin. (2005). Biometric Authentication: A machine learning approach.

Bhat, C. (1997a), 'An endogenous segmentation mode choice model with an application to intercity travel', *Transportation Science* 31, 34-48.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1-22

Fu, Zhaoxia Wang, Liming. (2012). Color Image Segmentation Using Gaussian Mixture Model and EM Algorithm.

Duda, RO Hart, PE (1973). *Pattern Classification and Scene Analysis*.

Miin-Shen Yang, Chion-Yo Lai Chih-Ying Lin (2012). A robust EM Clustering Algorithm for Gaussian Mixture Models.

Train, K. (2008a), 'EM algorithms for nonparametric estimation of mixing distributions', *Journal of Choice Modelling* 1, 40-69.

Train, K. (2008b), 'A recursive estimator for random coefficient models', working paper, Department of Economics, U. of California, Berkeley.