# Clustering on Unobserved Data using Mixture of Gaussians

Lu Ye
Dept. of Mathematics
York University
4700 Keele Str
Toronto, ON M3J 1P3
lye@mathstat.yorku.ca

Minas E. Spetsakis
Dept. of Computer Science
Center for Vision Research
York University
4700 Keele Str
Toronto, ON M3J 1P3
minas@cs.yorku.ca

**ABSTRACT**

This report provides an review of Clustering using Mixture Models and the Expectation Maximization method[1] and then extends these concepts to the problem of clustering of unobserved data where we cluster a set of vectors $u_i$ for $i = 1..N$ for which we only know the probability distribution. This problem has several applications in Computer Vision where we want to cluster noisy data.

## 1. Clustering

Data clustering is an important statistical technique, closely related to unsupervised learning. Clustering is the process of grouping the samples into clusters, so that samples with similar properties belong to the same cluster. A cluster is just a set whose members are similar but are different from the members of other clusters. For example, samples in a particular space where the distances between two samples in a cluster is less than minimal.

One of the most popular clustering techniques is based on Expectation Maximization (EM). EM is a method for estimating parameters using partially observed data[2, 3]. Since the clustering problem would be trivial if we already knew the membership of every sample to a particular cluster, we treat this membership information as the unobserved part of the data and apply EM. In this report we develop a clustering method that treats not only the membership of the data as unobserved but the data as well.

Edge grouping is a potential application for clustering. Edges can be considered as the sets of many edgels (edge elements) and these edgels can be detected in an image using any of the standard edge detection techniques.

If we happen to know that our edgels belong to straight lines and we have an estimate of the slope of the lines then we can group the edgels into straight line edges using a clustering technique. One of the problems is that some of the information is not very accurate and we can only assume that we know the probability distribution of the data, but not the data itself. So we essentially have partially observed data.

## 2. Mixture Model

The underlying model of EM clustering is the Mixture Model[4]. This model assumes that each sample comes from a cluster $\omega_j$, where $j = 1, \ldots, K$. The way to generate samples form the mixture model is as follows. We select one of the clusters $\omega_j$ with probability $P(\omega_j) = \pi_j$ and then generate a sample $x$ out of a probability distribution $p(x|\omega_j, \theta)$ or $p(x|\theta_j)$ where $\theta = \left\{ \theta_j,\ j = 1, \ldots, K \right\}$ and $\theta_j$ is the vector of parameters associated with the cluster $\omega_j$, which in our case contains the mean $\mu_j$, the covariance matrix $C_j$ and the mixture probability $\pi_j$. One should notice that $p(x|\theta_j)$ and $p(x|\omega_j, \theta)$ are the same. The first one denotes the probability density of $x$ given the parameters $\theta_j$ of the cluster $\omega_j$ and the other denotes the probability density of $x$ given the parameters $\theta$ of all the clusters and the fact that we have selected cluster $\omega_j$. The density $p(x|\theta_j)$ is called component density. The probability density function of a sample x from the mixture model is then

$$p(x|\theta) = \sum_{j=1}^{K} p(x|\omega_j, \theta)\pi_j. \tag{2.1}$$

For the mixture probabilities $\pi_j$, the following is true

$$\sum_{j=1}^{K} \pi_j = 1.$$

## 3. Maximum-Likelihood Estimation

The Maximum-Likelihood (ML) method can be used for parameter estimation from a set of samples $x_i, i = 1..N$. We assume that $p(x|\theta_j)$ is of known parametric form with parameter vectors $\theta_j$ which are unknown. The mixture probabilities $\pi_j$ are also unknown. The maximum-likelihood estimate of the parameters $\theta$ related to a set of data $D = x_i, i = 1, \ldots, N$ is obtained by maximizing the log likelihood of the parameters. The likelihood of the parameters is the probability of the data given the parameters

$$L(\theta) = p(D|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

assuming the data is independent. By substituting the probability density function Eq. (2.1), we get

$$p(D|\theta) = \prod_{i=1}^{N} \sum_{j=1}^{K} p(x_i|\omega_j, \theta)\pi_j.$$

The reason to take the log-likelihood of the parameters is that $p(D|\theta)$ involves a product of $N$ terms. To simplify the calculation, we use log-likelihood to change products to sums. We will use the Gaussian distribution[5] as our parametric form for $p(x_i|\theta_j)$. For multidimensional data $x$ it takes the form

$$p(x|\theta_j) = \frac{1}{\sqrt{(2\pi)^m |C_j|}} e^{-\frac{(x-\mu_j)^T C_j^{-1} (x-\mu_j)}{2}} \tag{3.1}$$

where $m$ is the number of dimensions of vector $x$, $C_j$ is the covariance matrix and $\mu_j$ is the mean for the cluster $\omega_j$. The mixture probabilities, the covariance and the mean are bundled in the parameter vector $\theta_j = \left\{ \mu_j, C_j, \pi_j \right\}$.

Im most applications of ML the next step is to maximize the likelihood function (or more often the log-likelihood). Usually the result is a simple analytic expression which along with a host of other nice properties explains the popularity of ML. But in this case we have a product of sums and it is impossible to get a simple expression.

The problem would become very simple if we knew the membership of every sample. Then we would only have to solve the Gaussian parameter estimation problem for every cluster separately. But lacking this membership information we have to use Expectation Maximization.

## 4. Expectation Maximization Applied to Clustering

Expectation Maximization (EM) is an iterative algorithm that alternates between two steps: the Expectation step and the Maximization step. In the Expectation step, we

first derive the log-likelihood of the unknown parameters as a function of the unobserved data and then we compute the expected value of this likelihood using the probability density of the unobserved data. Since the density of the unobserved data involves the parameters we want to estimate we use a guess. In the Maximization step, we obtain the parameters that maximize the expected value of the likelihood. This new parameter will become the guess to be used in the next iteration. This iteration goes on until convergence or satisfaction of the termination conditions. The unknown parameter vector $\theta$ contains information appropriate for the mixture model we use. Since we will use the Gaussian distribution over the mixture model, normally the parameter vector contains three components which are the mean and the variance for each cluster and the mixture probability. These are usually the unknowns of our problem. After we know the mean, variance and the mixture probabilities of each cluster, we are able to group samples. The number of clusters $K$ is known, but we can augment the algorithm with BIC or AIC (see below) and have $K$ as unknown too.

The straight application of ML to our clustering problem is extremely difficult because it involves logarithms of sums and many complicated functions. Luckily, it can be easily put in a form where we can apply EM. We do this as follows[6]. Let $z_i$ be a vector of length $K$ (the number of clusters) and the $j^{th}$ element of the vector is 1, if the sample $x_i$ was generated by cluster $j$ and zero otherwise. Obviously there is exactly one "1" in the vector and the rest are zero so

$$\sum_{j=1}^{K} z_{ij} = 1 \tag{4.1}$$

where $z_{ij}$ is the $j^{th}$ element of $z_i$. Since we do not know the membership of every sample, $z_i$ is the unobserved part of the data. Let $D_y = y_i, i = 1, \ldots, N$ be the complete data set where $y_i = \{(x_i, z_i)\}$, $D_x = x_i, i = 1, \ldots, N$ is observed data set and $D_z = z_i, i = 1, \cdots, N$ be the unobserved data set.

In the expectation step, the expression for expectation is

$$Q(\theta, \theta^t) = E_{D_z}\left\{ \ln p(D_y|\theta) \middle| D_x, \theta^t \right\}.$$

Although it looks intimidating at first, it can be tamed in a few simple steps. The first step is to derive the form of $\ln p(D_y|\theta)$ which is a function of $D_x$, $D_z$ and $\theta$. Once we simplify the $\ln p(D_y|\theta)$, we can get its expected value by applying the standard rules. The subscript $D_z$ in $E_{D_z}$ means that the random variables are the elements of $D_z$ and the probability of $D_z$ is conditioned on $D_x$ and $\theta^t$, where $\theta^t$ is just a guess, but it is usually the result of the $t^{th}$ (or previous) iteration. The result is a function of three parameters: $\theta$, the unknown mixture parameters, $\theta^t$ the guess for $\theta$ and the observed data $D_x$, but for simplicity we drop $D_x$ and simply write $Q(\theta, \theta^t)$.

Using the rule of joint probabilities and assuming that the $y_i$s, the elements of $D_y$, are independent, we can rewrite $p(D_y|\theta)$ as follows:

$$p(D_y|\theta) = \prod_{i=1}^{N} p(y_i|\theta) \tag{4.2}$$

Since we have $y_i = \{(x_i, z_i)\}$, $p(y_i|\theta)$ can be written as

$$p(y_i|\theta) = p(x_i, z_i|\theta)$$

Based on the properties of conditional probability $p(x_i, z_i|\theta)$ has the following probability density expression.

$$p(x_i, z_i|\theta) = p(x_i|z_i, \theta)p(z_i|\theta)$$

As we mentioned above, $z_i$ is the vector, whose $j^{th}$ element is 1 if the sample $x_i$ was generated by the cluster $j$. Therefore, the probability of $z_i$ given the parameter vector $\theta$, is equal to the mixture probability $\pi_j$, i.e. the probability to select cluster $j$ among the clusters of the mixture model.

The probability of $x_i$ given $z_i$ and $\theta$, $p(x_i|z_i, \theta)$ has an alternative form which is $p(x_i|\theta_j)$, where $\theta_j$ is the vector whose parameters associated with the cluster $\omega_j$. So we have

$$p(x_i, z_i|\theta) = p(x_i|\theta_j)\pi_j.$$

Thus $p(y_i|\theta)$ can be written as

$$p(y_i|\theta) = p(x_i|\theta_j)\pi_j \tag{4.3}$$

This equation only presents the correspondence of $p(y_i|\theta)$ to the parameter vector of the $j^{th}$ cluster $\theta_j$. Recalling that $z_i$ vector, $z_{ij} = 1$ and $z_{ik} = 0$ for $k \neq j$, we have the following new expression for $p(y_i|\theta)$:

$$p(y_i|\theta) = \prod_{k=1}^{K} (p(x_i|\theta_k)\pi_k)^{z_{ik}}$$

The term $(p(x_i|\theta_k)\pi_k)^{z_{ik}}$ is equal to $p(x_i|\theta_j)\pi_j$ when $z_{ij} = 1$, and it is 1 when $k \neq j$. Thus, the product terms of all K clusters is the same as Eq. (4.3). By grouping all the samples $y_i$ in $D_y$ as in Eq. (4.2), the probability density function of the $D_y$ set given $\theta$ becomes:

$$p(D_y|\theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} (p(x_i|\theta_k)\pi_k)^{z_{ik}}$$

Then we should obtain $\ln p(D_y|\theta)$ by simply taking the logarithm on the above equation:

$$\ln p(D_y|\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \ln(p(x_i|\theta_k)\pi_k)$$

and by applying the logarithm rule once more

$$\ln p(D_y|\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \ln(p(x_i|\theta_k)) + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \ln(\pi_k) \tag{4.4}$$

Eq. (4.4) is our form for $\ln p(D_y|\theta)$ and we can now calculate the expected value of it to conclude the Expectation Step. Since the form of $\ln p(D_y|\theta)$ is the sum of two terms, the expected value should be the sum of the expected values of those two terms. We also need to keep in mind that the random variables are the unobserved data $z_{ik}$ only. The

other terms of the equation should be treated as the constants in this situation. The expected values of constants are themselves.

$$E_z\left\{ \ln p(D_y|\theta)\middle|D_x, \theta^t\right\} = \sum_{i=1}^{N}\sum_{k=1}^{K} E_z\left\{ z_{ik}|D_x, \theta^t\right\} \ln(p(x_i|\theta_k)) + \sum_{i=1}^{N}\sum_{k=1}^{K} E_z\left\{ z_{ik}|D_x, \theta^t\right\} \ln(\pi_k) \tag{4.5}$$

Let's rename $E_z\left\{ z_{ik}|D_x, \theta^t\right\}$ as $\bar{z}_{ik}$, then we have

$$E_z\left\{ \ln p(D_y|\theta)\middle|D_x, \theta^t\right\} = \sum_{i=1}^{N}\sum_{k=1}^{K} \bar{z}_{ik} \ln(p(x_i|\theta_k)) + \sum_{i=1}^{N}\sum_{k=1}^{K} \bar{z}_{ik} \ln(\pi_k) \tag{4.6}$$

Based on the definition of the $z_{ik}$ vector, we notice that $z_{ik}$ takes two values only: zero and one. So

$$\bar{z}_{ik} = E_z\left\{ z_{ik}|D_x, \theta^t\right\} = 0 \cdot P(z_{ik} = 0|D_x, \theta^t) + 1 \cdot P(z_{ik} = 1|D_x, \theta^t) = P(z_{ik} = 1|D_x, \theta^t)$$

is the probability of $z_{ik}$ to be 1. This is the same as $P(\omega_k|x_i, \theta^t)$ the probability the sample $x_i$ to be generated by the $k^{th}$ cluster among all the clusters. Applying the Bayes rule, we have $\bar{z}_{ik}$ corresponding to the whole model:

$$\bar{z}_{ik} = E_z\left\{ z_{ik}|D_x, \theta^t\right\} = P(\omega_k|x_i, \theta^t) = \frac{p(x_i|\theta^t, \omega_k)P(\omega_k|\theta^t)}{\sum_{j=1}^{K} p(x_i|\theta^t, \omega_j)P(\omega_j|\theta^t)} = \tag{4.7}$$

$$\frac{p(x_i|\theta^t, \omega_k)\pi_k^t}{\sum_{j=1}^{K} p(x_i|\theta^t, \omega_j)\pi_j^t} = \frac{p(x_i|{\theta_k}^t)\pi_k^t}{\sum_{j=1}^{K} p(x_i|{\theta_j}^t)\pi_j^t}$$

Since we use a known parametric form for the probability, Gaussian in this case, we can compute $p(x_i|\theta^t)$ and so compute $\bar{z}_{ik}$ from Eq. (4.7) and the we can plug the value of $\bar{z}_{ik}$ into Eq. (4.5), to get the final result of the Expectation Step.

In the Maximization Step, we are supposed to maximize the above expectation function values in order to get the optimal solution for the unknown $\theta$ which can be used in the next round as $\theta^t$ to estimate a new expectation function and from this to get the new $\theta$. There is one constraint in our mixture model, which is

$$\sum_{j=1}^{K} \pi_j = 1. \tag{4.8}$$

Therefore, we maximize the expectation function $Q(\theta, \theta^t)$ subject to this constraint. We define a new $Q'(\theta, \theta^t)$ which takes into account the constraint of Eq. (4.8) by using the Lagrange multiplier.

$$Q'(\theta, \theta^t) = \sum_{i=1}^{N}\sum_{k=1}^{K} \bar{z}_{ik} \ln(p(x_i|\theta_k)) + \sum_{i=1}^{N}\sum_{k=1}^{K} \bar{z}_{ik} \ln(\pi_j) + \lambda(1 - \sum_{j=1}^{K} \pi_j)$$

Using standard calculus procedures, we can find the maximum value of a function subject

to the constraint by taking derivatives over the unknowns of the function and $\lambda$, set the derivatives to zero, and solve the resulting equations to find the value of the unknowns and $\lambda$ that maximize the function. In the our case the unknown is $\theta$ the vector containing all the parameters of the mixture model $\pi_j$, $\mu_j$ and $C_j$ for $j = 1..K$.

First, we take the partial derivative with respect to $\pi_j$. In $Q'(\theta, \theta^t)$, only the last two terms involve the variable $\pi_j$ so the first term will give zero. Since we only take derivatives of the whole $Q'$ term with respect to one particular $\pi_j$ term, only the $j^{th}$ element of the whole $Q'$ term will have non zero value. So we can get rid of the summation over $j = 1, \cdots, K$, and we have

$$\frac{\partial Q'}{\partial \pi_j} = \sum_{i=1}^{N} \bar{z}_{ik} \frac{1}{\pi_j} - \lambda = 0$$

$$\lambda = \sum_{i=1}^{N} \bar{z}_{ik} \frac{1}{\pi_j}$$

$$\pi_j = \sum_{i=1}^{N} \frac{\bar{z}_{ik}}{\lambda} \tag{4.9}$$

If we take the summation from 1 to $K$ on both sides we can use Eq. (4.8) to eliminate $\pi_j$ and have

$$\sum_{j=1}^{K} \sum_{i=1}^{N} \frac{\bar{z}_{ik}}{\lambda} = 1$$

from which we get

$$\lambda = \sum_{j=1}^{K} \sum_{i=1}^{N} \bar{z}_{ik}$$

So we plug the $\lambda$ into Eq. (4.9), and get

$$\pi_j = \frac{\sum_{i=1}^{N} \bar{z}_{ik}}{\sum_{k=1}^{K} \sum_{i=1}^{N} \bar{z}_{ik}}$$

We notice that the denominator in the above expression is

$$\sum_{j=1}^{K} \sum_{i=1}^{N} \bar{z}_{ik} = \sum_{i=1}^{N} \sum_{j=1}^{K} \bar{z}_{ik} = \sum_{i=1}^{N} 1 = N$$

after taking into account Eq. (4.1). So finally

$$\pi_j = \frac{1}{N} \sum_{i=1}^{N} \bar{z}_{ik}. \tag{4.10}$$

Second, we take the partial derivative with respect to the unknown $\mu_j$, the mean of the $j^{th}$ cluster. It is only involved in the Gaussian distribution for our mixture model, so

we only need to be concerned with the first term of the expectation function when taking derivatives. In the first term of the function, $\bar{z}_{ik}$ uses the parameters $\theta^t$ which is the guess and we consider it independent from $\theta$ and constant. By the same reason as above, we eliminate one summation from $j = 1, \ldots, K$ by taking derivative with respect to the particular $\mu_j$. The definition of the Gaussian density from Eq. (3.1) is

$$p(x|\theta_j) = \frac{1}{\sqrt{(2\pi)^m |C_j|}} e^{-\frac{(x-\mu_j)^T C_j^{-1}(x-\mu_j)}{2}}$$

and the log is

$$\ln \frac{1}{\sqrt{(2\pi)^m |C_j|}} - \frac{(x_i - \mu_j)^T C_j^{-1}(x_i - \mu_j)}{2}$$

The derivative then is as follows:

$$\frac{\partial Q'}{\partial \mu_j} = \frac{\partial}{\partial \mu_j} \left( \sum_{i=1}^{N} \bar{z}_{ik} \ln \frac{1}{\sqrt{(2\pi)^m |C_j|}} - \sum_{i=1}^{N} \bar{z}_{ik} \frac{(x_i - \mu_j)^T C_j^{-1}(x_i - \mu_j)}{2} \right) = 0$$

The derivative of the first term equals to zero. The derivative of $\frac{(x_i - \mu_j)^T C_j^{-1}(x_i - \mu_j)}{2}$ with respect to $\mu_j$ is $C_j^{-1}(x_i - \mu_j)$. The derivative of the second part is

$$\frac{\partial Q'}{\partial \mu_j} = \sum_{i=1}^{N} \bar{z}_{ik} C_j^{-1}(x_i - \mu_j) = C_j^{-1} \sum_{i=1}^{N} \bar{z}_{ik}(x_i - \mu_j) = 0$$

which results to

$$\sum_{i=1}^{N} \bar{z}_{ik} x_i = \sum_{i=1}^{N} \bar{z}_{ik} \mu_j$$

or

$$\mu_j = \frac{\sum_{i=1}^{N} \bar{z}_{ik} x_i}{\sum_{i=1}^{N} \bar{z}_{ik}}$$

and using Eq. (4.10) we get

$$\mu_j = \frac{1}{N\pi_j} \sum_{i=1}^{N} \bar{z}_{ik} x_i.$$

It's time to derive the last unknown variable $C_j$ of our model in the Maximization step. The steps of derivation are similar to the derivation of $\mu_j$. We only work with the first term of expectation function. The difference begins with the following step:

$$\frac{\partial Q'}{\partial C_j} = \frac{\partial}{\partial C_j}\left(\sum_{i=1}^{N} \bar{z}_{ik} \ln \frac{1}{\sqrt{(2\pi)^m |C_j|}} - \sum_{i=1}^{N} \bar{z}_{ik} \frac{(x_i - \mu_j)^T C_j^{-1}(x_i - \mu_j)}{2}\right) = 0$$

The above differentiation involves derivatives of a determinant and a quadratic product which we can get from the Matrix Reference Manual (http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/index.html):

$$\frac{\partial}{\partial C_j}(x_i - \mu_j)^T C_j^{-1}(x_i - \mu_j) = -C_j^{-1}(x_i - \mu_j)(x_i - \mu_j)^T C_j^{-1}$$

$$\frac{\partial}{\partial C_j}|C_j| = |C_j| C_j^{-1}$$

and using the chain rule

$$\frac{\partial}{\partial C_j} \ln \frac{1}{\sqrt{(2\pi)^m |C_j|}} = \frac{\partial}{\partial C_j}\left(-\frac{1}{2}\ln|C_j|\right) = -\frac{1}{2} C_j^{-1}$$

and we get

$$\frac{\partial Q'}{\partial C_j} = -\frac{1}{2}\sum_{i=1}^{N} \bar{z}_{ik}\left(C_j^{-1} + C_j^{-1}(x_i - \mu_j)(x_i - \mu_j)^T C_j^{-1}\right) = 0$$

After that, we multiply $C_j$ twice on both sides of above equation to eliminate $C_j^{-1}$, and we get the following:

$$\sum_{i=1}^{N} \bar{z}_{ik}(-C_j + (x_i - \mu_j)(x_i - \mu_j)^T) = 0$$

$$\sum_{i=1}^{N} \bar{z}_{ik} C_j = \sum_{i=1}^{N} \bar{z}_{ik}(x_i - \mu_j)(x_i - \mu_j)^T$$

which gives

$$C_j = \frac{\sum_{i=1}^{N} \bar{z}_{ik}(x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{N} \bar{z}_{ik}}$$

By using Eq. (4.10), we have the $C_j$ for the next iteration.

$$C_j = \frac{1}{N\pi_j}\sum_{i=1}^{N} \bar{z}_{ik}(x_i - \mu_j)(x_i - \mu_j)^T$$

In conclusion, the above derivation explains the internal detail steps of EM algorithm. When we run the EM algorithm each time, we can directly use the obtained equations to calculate the covariance, means and mixture parameters.

K-means clustering is a simple and classical example by applying EM algorithm. It can be viewed as the problem of estimating the means of K Gaussians Mixture Model. The special assumption of K-means is that the mixing probabilities $\pi_j$ are equal and each

Gaussian distribution has the same variance.

## 5. Bayesian Information Criterion Applied to Clustering

The EM clustering algorithm that we outlined above assumes a Mixture of Gaussians as the underlying model. After the particular mixture model is defined, the EM algorithm is be applied to find all the parameters except the number of clusters. So we have to provide the number of clusters through some other method. The Bayesian Information Criterion (BIC) is a widely used method for this purpose[7, 8], and can be applied in our case to find the number of clusters.

Since we use EM to find the maximum mixture likelihood after each iteration, we can get reliable approximation of BIC. The expression of BIC is

$$BIC = -2L_M(x, \theta) + m_M log(N) \tag{5.1}$$

where $L_M(x, \theta)$ is the maximized mixture loglikelihood for the model $M$, $m_M$ is the number of independent parameters to be estimated in the model and $N$ is the number of samples. The maximized mixture loglikelihood function is

$$L_M(x, \theta) = \log \prod_{i=1}^{N} p(x_i|\theta) = \log \prod_{i=1}^{N} \sum_{j=1}^{K} p(x_i|\theta_j)\pi_j$$

Recalling the material discussed in the previous sections, $p(x_i|\theta_j)$ is the probability of $x_i$ given that it belongs to the $j^{th}$ cluster, which is a Gaussian distribution and $\pi_j$ is the the mixture probability of the model $M$. The smaller the value of BIC, the stronger the evidence for the model.

The BIC can be used not only to determine the number of clusters by comparing models, but also help avoid local minima in our clustering problem. Since we use random starting points, usually by selecting random initial values for parameters $\theta_j$, $\theta_j = (\mu_j, C_j, \pi_j)$ where $j = 1, \cdots, K$, we get distinct clustering results when we run the EM algorithm several times. If it is our lucky day, we get reasonably good initial values for $\theta_j$, and we get desirable results in the first attempt. However, we can't control our luck. To solve this problem, we apply the EM clustering on the same data several times using random starting points and calculate the BIC to examine the maximum likelihood for each complete run by applying Eq. (5.1). Then we select the run that has minimum BIC.

## 6. Expectation Maximization Clustering Applied on Unobserved Data

In the previous sections, we discussed the EM method and the clustering problem and applied the EM algorithm on the clustering of $N$ samples into $K$ clusters. We assumed that the underlying model is a Mixture of Gaussians and the coordinates of the samples are given. Next we study the problem where the coordinates of samples are not given directly and we only know the probability distributions of the samples $p(u_i|D_i)$, where $u_i$ are the samples and $D_i$ is the data where these probabilities are based on. The union of all $D_i$ is $D$. We will again use EM to cluster these samples. The general procedure of EM algorithm remain the same but we have different assumptions and preconditions.

As before, we have both a set of observed data and a set of unobserved data. We are given the information $D$, and we have $N$ unobserved samples $u_i$, $i = 1, \ldots, N$. As before $\psi_i$ is the membership vector of that sample where as $\psi_{ij} = 1$ if sample $u_i$ was generated by cluster $\omega_j$ and is equal to zero otherwise. The complete the data $y_i$ for each sample is $y_i = (D_i, \psi_i, u_i)$. The complete data set is $D_y = (y_i, i = 1, \ldots, N)$ where $y_i = (D_i, \psi_i, u_i)$ . The unobserved data is $D_z = (z_i, i = 1, \ldots, N)$ where $z_i = (\psi_i, u_i)$. Since we assume Mixture of Gaussians, we have to compute the parameter vector $\theta = (\theta_j, j = 1..K)$ where $\theta_j = (\mu_j, C_j, \pi_j)$ is the mean, variance and mixture probabilities of each cluster.

### 6.1. Expectation

The first step in EM is to derive the expression for expectation. In this problem, $D$ is the only given data, $D_y$ is the complete data and $D_z$ is the unobserved data. The general expression for the expectation for this problem is

$$Q(\theta, \theta^t) = E_{D_z}\left\{ \ln p(D_y|\theta) \middle| D, \theta^t \right\}$$

but now $D_y$ and $D_z$ contain different data sets and we need to derive $\ln p(D_y|\theta)$ under different assumptions. As before, we write the $p(D_y|\theta)$ as:

$$p(D_y|\theta) = \prod_{i=1}^{N} p(y_i|\theta) \tag{6.1}$$

Since $y_i = (D_i, \psi_i, u_i)$, we can write

$$p(y_i|\theta) = p(D_i, \psi_i, u_i|\theta)$$

and we rewrite $p(D_i, \psi_i, u_i|\theta)$ as

$$p(D_i, \psi_i, u_i|\theta) = p(\psi_i, u_i|\theta)\, p(D_i|\theta, \psi_i, u_i)$$

and $p(\psi_i, u_i|\theta)$ as

$$p(\psi_i, u_i|\theta) = p(\psi_i|\theta)\, p(u_i|\theta, \psi_i).$$

so

$$p(y_i|\theta) = p(\psi_i|\theta)\, p(u_i|\theta, \psi_i)\, p(D_i|\theta, \psi_i, u_i)$$

We now introduce the “quazi-dependence” assumption that $D$ (or its components) are not directly related to $\theta$ (or its components) but only through the corresponding $u$. So $p(D_i|\theta, \psi_i, u_i) = p(D_i|u_i)$. So the $p(y_i|\theta)$ becomes

$$p(y_i|\theta) = p(\psi_i|\theta)\, p(u_i|\theta, \psi_i)\, p(D_i|u_i) \tag{6.2}$$

The first term in Eq. (6.2) is $p(\psi_i|\theta)$ and $\psi_i$ is the membership vector for the $i^{th}$ sample. Then

$$\psi_{ij} = \begin{cases} 1 & \text{if } u_i \text{ generated } by\ \omega_j \\ 0 & \text{otherwise} \end{cases}$$

and

$$\sum_{j=1}^{K} \psi_{ij} = 1 \tag{6.3}$$

and we can write

$$p(\psi_i|\theta) = \prod_{k=1}^{K} \pi_k{}^{\psi_{ik}}$$

The value of the above expression is $\pi_j$ if the $i^{th}$ sample is the member of $j^{th}$ cluster, or $\psi_{ij} = 1$. We can apply the same technique to $p(u_i|\theta, \psi_i)$ and have

$$p(u_i|\theta, \psi_i) = \prod_{k=1}^{K} p(u_i|\theta_k)^{\psi_{ik}}.$$

Then the Eq. (6.2) can be rewritten as

$$p(y_i|\theta) = \left(\prod_{k=1}^{K} \pi_k{}^{\psi_{ik}}\right)\left(\prod_{k=1}^{K} p(u_i|\theta_k)^{\psi_{ik}}\right) p(D_i|u_i)$$

By taking the log of $p(y_i|\theta)$ we get:

$$\log p(y_i|\theta) = \sum_{k=1}^{K} \psi_{ik} \log \pi_k + \sum_{k=1}^{K} \psi_{ik} \log p(u_i|\theta_k) + \log p(D_i|u_i) \tag{6.4}$$

and from Eq. (6.1)

$$\log p(D_y|\theta) = \sum_{i=1}^{N} \log p(y_i|\theta)$$

We substitute Eq. (6.4) into the above equation, we get the function of $\log p(D_y|\theta)$

$$\log p(D_y|\theta) = \sum_{i=1}^{N}\sum_{k=1}^{K} \psi_{ik} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} \psi_{ik} \log p(u_i|\theta_k) + \sum_{i=1}^{N} \log p(D_i|u_i) \tag{6.5}$$

The final step to derive the expectation function is to take expected value of Eq. (6.5) with respect to the unobserved data $D_z$ given the data $D$ and a guess of the clustering parameters $\theta^t$. The expectation equation is:

$$\begin{aligned} E_{D_z}\left\{ \log p(D_y|\theta) \middle| D, \theta^t \right\} = \sum_{i=1}^{N} \Bigg( & \sum_{k=1}^{K} E\left\{ \psi_{ik} \middle| \theta^t, D \right\} \log \pi_k + \\ & \sum_{k=1}^{K} E\left\{ \psi_{ik} \log p(u_i|\theta_k) \middle| \theta^t, D \right\} + \\ & E\left\{ \log p(D_i|u_i) \middle| \theta^t, D \right\} \Bigg) \end{aligned} \tag{6.6}$$

where we assume that all the expected values are taken over $D_z$. We first derive $E\left\{ \psi_{ik}|\theta^t, D \right\}$, which we name $\bar{\psi}_{ik}$.

$$\bar{\psi}_{ik} = E\left\{ \psi_{ik} \middle| \theta^t, D \right\} = 0 \cdot P(\psi_{ik} = 0|\theta^t, D) + 1 \cdot P(\psi_{ik} = 1|\theta^t, D)$$

since $\psi_{ik}$ takes only the values 0 and 1. So

$$\bar{\psi}^{ik} = P(\psi_{ik} = 1|\theta^t, D)$$

Since we need explicit dependence on $u_i$ we write

$$P(\psi_{ik} = 1|\theta^t, D) =$$
$$\int P(\psi_{ik} = 1, u_i|\theta^t, D)du_i =$$
$$\int p(u_i|\theta^t, D)P(\psi_{ik} = 1|\theta^t, D, u_i)du_i$$

By invoking again the quazi-dependence assumption we notice that $\psi_{ik}$ doesn't depend on $D$, since $u_i$ is given, so we drop $D$. We also notice that $p(u_i|\theta^t, D) = p(u_i|\theta^t, D_i)$.

$$\int p(u_i|\theta^t, D_i)P(\psi_{ik} = 1|\theta^t, u_i)du_i = \int p(u_i|\theta^t, D_i)P(\omega_k|\theta^t, u_i)du_i =$$
$$\int \frac{p(D_i|\theta^t, u_i)p(u_i|\theta^t)}{p(D_i|\theta^t)} \frac{p(u_i|{\theta_k}^t)P(\omega_k|\theta^t)}{p(u_i|\theta^t)} du_i =$$
$$\int \frac{p(D_i|\theta^t, u_i)p(u_i|{\theta_k}^t){\pi_k}^t}{p(D_i|\theta^t)} du_i$$

In $p(D_i|\theta^t, u_i)$, $\theta^t$ has no direct relationship with $D_i$ but only through $u_i$, so we invoke the quazi-dependence assumption and drop $\theta^t$ so $p(D_i|\theta^t, u_i) = p(D_i|u_i)$ which is the likelihood of $u_i$ given the data $D$ and is assumed known. Since ${\pi_k}^t$ and $p(D_i|\theta^t)$ don't contain the $u_i$ variable, we have

$$\int \frac{p(D_i|u_i)p(u_i|{\theta_k}^t){\pi_k}^t}{p(D_i|\theta^t)} du_i =$$
$$\frac{{\pi_k}^t}{p(D_i|\theta^t)} \int p(D_i|\theta^t, u_i)p(u_i|{\theta_k}^t)du_i$$

Then we get

$$\begin{aligned}
\bar{\psi}_{ik} &= \frac{{\pi_k}^t}{p(D_i|\theta^t)} \int p(D_i|u_i)p(u_i|{\theta_k}^t)du_i = \\
&\frac{{\pi_k}^t}{\int p(D_i, u_i|\theta^t)du_i} \int p(D_i|u_i)p(u_i|{\theta_k}^t)du_i = \\
&\frac{{\pi_k}^t}{\int p(D_i|u_i)p(u_i|\theta^t)du_i} \int p(D_i|u_i)p(u_i|{\theta_k}^t)du_i = \\
&\frac{{\pi_k}^t}{\int p(D_i|u_i)\sum_{j=1}^{K} {\pi_j}^t p(u_i|{\theta_j}^t)du_i} \int p(D_i|u_i)p(u_i|{\theta_k}^t)du_i
\end{aligned} \tag{6.7}$$

and if we exchange the summation and the integral we get

$$\frac{{\pi_k}^t}{\sum_{j=1}^{K}\int p(D_i|u_i)\pi_j{}^t\, p(u_i|\theta_j{}^t)du_i}\int p(D_i|u_i)p(u_i|\theta_k{}^t)du_i =$$

$$\frac{{\pi_k}^t\int p(D_i|u_i)p(u_i|\theta_k{}^t)du_i}{\sum_{j=1}^{K}\pi_j{}^t\int p(D_i|u_i)p(u_i|\theta_j{}^t)du_i}$$

We define $g_{ik}$

$$g_{ik} = p(D_i|\theta_k{}^t) = \int p(D_i|u_i)p(u_i|\theta_k{}^t)du_i =$$

and we rewrite the expression of $\bar{\psi}_{ik}$ as

$$\bar{\psi}_{ik} = \frac{{\pi_k}^t g_{ik}}{\sum_{j=1}^{K}\pi_j{}^t g_{ij}} \tag{6.8}$$

We can compute $g_{ik}$ by evaluating the integral using one of the methods described later and we can the compute $\bar{\psi}_{ik}$. The first term of Eq. (6.6) can be written as follows

$$\sum_{k=1}^{K} E\left\{\psi_{ik}\middle|\theta^t, D\right\}\log\pi_k = \sum_{k=1}^{K}\bar{\psi}_{ik}\log\pi_k \tag{6.9}$$

We now simplify the second term of Eq.(6.6).

$$\sum_{k=1}^{K} E_{D_z}\left\{\bar{\psi}_{ik}\log p(u_i|\theta_k)\middle|\theta^t, D\right\} =$$

$$0 + \sum_{k=1}^{K} E_{u_i}\left\{\log p(u_i|\theta_k)\middle|D, \theta^t, \psi_{ik}=1\right\}P(\psi_{ik}=1|D,\theta^t)$$

The second factor $P(\psi_{ik}=1|D,\theta^t)$ is the same as the $\bar{\psi}_{ik}$ that we have derived. We use the result of Eq. (6.7) directly, and have

$$E_{u_i}\left\{\log p(u_i|\theta_k)\middle|D,\theta^t,\psi_{ik}=1\right\}\bar{\psi}_{ik} = E_{u_i}\left\{\log p(u_i|\theta_k)\middle|D,\theta_k{}^t\right\}\bar{\psi}_{ik} =$$

$$\bar{\psi}_{ik}\int\log p(u_i|\theta_k)p(u_i|D_i,\theta_k{}^t)du_i$$

All the $\theta^t$ parameters are given and only $\log p(u_i|\theta_k)$ contains $\mu_k$ and $C_k$ which will be estimated in the Maximization step.

Since the third term of Eq. (6.6) doesn't contain any variables that need to be estimated in the maximization step, it is irrelevant and we can ignore it.

### 6.2. Maximization

In the Maximization step, we use the same method to get the optimal solution for parameters in $\theta_j = \left\{\mu_j, C_j, \pi_j\right\}$. We also need to use the constraint $\sum_{j=1}^{K}\pi_j = 1$. I use the defined symbols to substitute some complex terms in Expectation step. By using the

Lagrange multiplier, we get the maximization function.

$$Q'(\theta, \theta^t) = \sum_{i=1}^{N}\left(\sum_{k=1}^{K} \bar{\psi}_{ik} \log \pi_k + \sum_{k=1}^{K} \bar{\psi}_{ik} \int \log p(u_i|\theta_k) p(u_i|D_i, {\theta_k}^t) du_i\right) + \lambda\left(1 - \sum_{k=1}^{K} \pi_k\right) \tag{6.10}$$

We want to take partial derivatives over $Q'(\theta, \theta^t)$ with respect to the components of $\theta_j = \{\mu_j, C_j, \pi_j\}$ and $\lambda$.

First, we take the partial derivative with respect to $\pi_j$. Only the first term and the lagrange term contain the variable $\pi_j$ in the Eq. (6.10), so the other terms will vanish. Since the integral operator and partial derivative are the linear operators, we can interchange them

$$\frac{\partial Q'}{\partial \pi_j} = \sum_{i=1}^{N} \bar{\psi}_{ik} \frac{1}{\pi_j} - \lambda = 0$$

$$\lambda = \sum_{i=1}^{N} \bar{\psi}_{ik} \frac{1}{\pi_j}$$

and after solving for $\pi_j$

$$\pi_j = \frac{1}{\lambda} \sum_{i=1}^{N} \bar{\psi}_{ik} \tag{6.11}$$

Using the constraint of the mixture probabilities (by taking the derivative with respect to $\lambda$)

$$\sum_{k=1}^{K} \pi_k = 1$$

we get

$$1 = \frac{1}{\lambda} \sum_{k=1}^{K} \sum_{i=1}^{N} \bar{\psi}_{ik}$$

$$\lambda = \sum_{k=1}^{K} \sum_{i=1}^{N} \bar{\psi}_{ik}$$

So we plug the expression of $\lambda$ into the Eq. (6.11) and get

$$\pi_j = \frac{\sum_{i=1}^{N} \bar{\psi}_{ik}}{\sum_{k=1}^{K} \sum_{i=1}^{N} \bar{\psi}_{ik}} \tag{6.12}$$

Next we take the partial derivatives with respect to $\mu_j$ and set it to 0 to get the optimal value of $\mu_j$ which only appears in the log-Gaussian $\log p(u_i|\theta_k)$. The other terms will vanish.

$$\frac{\partial Q'}{\partial \mu_j} = \frac{\partial \sum_{i=1}^{N}\sum_{k=1}^{K} \bar{\psi}_{ik} \int \log p(u_i|\theta_k) p(u_i|D_i, {\theta_k}^t) du_i}{\partial \mu_k} = 0 \tag{6.13}$$

The distribution of $p(u_i|\theta_k)$ is a Gaussian with probability density function

$$p(u_i|\theta_k) = \frac{1}{\sqrt{(2\pi)^m |C_k|}} e^{-\frac{(u_i-\mu_k)^T {C_k}^{-1}(u_i-\mu_k)}{2}}$$

and

$$\log p(u_i|\theta_k) = \log \frac{1}{\sqrt{(2\pi)^m |C_k|}} - \frac{(u_i-\mu_k)^T {C_k}^{-1}(u_i-\mu_k)}{2}$$

Then we plug the expression of $\log p(u_i|\theta_k)$ into the Eq. (6.13) and we get

$$\frac{\partial Q'}{\partial \mu_j} = \frac{\partial \sum_{i=1}^{N}\sum_{k=1}^{K} \bar{\psi}^{ik} \int \log\left(\frac{1}{\sqrt{(2\pi)^m |C_k|}} - \frac{(u_i-\mu_k)^T {C_k}^{-1}(u_i-\mu_k)}{2}\right) p(u_i|D_i, {\theta_k}^t) du_i}{\partial \mu_k} = 0$$

After changing the order of the summation, the integral and the derivative, and ignoring the term that vanishes we get

$$\frac{\partial Q'}{\partial \mu_j} = \sum_{i=1}^{N}\sum_{k=1}^{K} \bar{\psi}_{ik} \int \frac{1}{2}\left(\partial \frac{(u_i-\mu_k)^T {C_k}^{-1}(u_i-\mu_k) p(u_i|D_i, {\theta_k}^t)}{\partial \mu_j}\right) du_i = 0$$

As mentioned in the last problem, the derivative of $\frac{(u_i-\mu_k)^T {C_k}^{-1}(u_i-\mu_k)}{2}$ with respect to $\mu_k$ is ${C_k}^{-1}(u_i-\mu_k)$. Thus we get

$$\frac{\partial Q'}{\partial \mu_j} = \sum_{i=1}^{N} \bar{\psi}_{ik} \int {C_j}^{-1}(u_i-\mu_j) p(u_i|D_i, {\theta_k}^t) du_i =$$

$$\sum_{i=1}^{N} \bar{\psi}_{ik} \int p(u_i|D_i, {\theta_k}^t) {C_j}^{-1} u_i du_i - \sum_{i=1}^{N} \bar{\psi}_{ik} \int p(u_i|D_i, {\theta_k}^t) {C_j}^{-1} \mu_j du_i = 0.$$

Since $\mu_j$ can be moved out of the integral and since ${C_j}^{-1}$ is not singular and can be eliminated, we get

$$\sum_{i=1}^{N} \bar{\psi}_{ik} \int p(u_i|D_i, {\theta_k}^t) u_i du_i = \mu_j \sum_{i=1}^{N} \bar{\psi}_{ik} \int p(u_i|D_i, {\theta_k}^t) du_i$$

and

$$\mu_j = \frac{\sum_{i=1}^{N} \bar{\psi}_{ik} \int p(u_i|D_i, {\theta_k}^t) u_i du_i}{\sum_{i=1}^{N} \bar{\psi}_{ik}}$$

since $\int p(u_i|D_i, {\theta_k}^t) du_i = 1$. Furthermore, we can write

$$p(u_i|D_i, \theta_k{}^t) = \frac{p(D_i|u_i, \theta_k{}^t)p(u_i|\theta_k{}^t)}{p(D_i|\theta_k{}^t)}$$

and from the quazi-dependence assumption, we have $p(D_i|u_i, \theta_k{}^t) = p(D_i|u_i)$. Thus we rewrite above expression as

$$\mu_j = \frac{\sum_{i=1}^{N} \frac{\bar{\psi}_{ik}}{p(D_i|\theta_k{}^t)} \int p(D_i|u_i)p(u_i|\theta_k{}^t)u_i du_i}{\sum_{i=1}^{N} \bar{\psi}_{ik}}$$

Recall that we denoted $p(D_i|\theta_k{}^t)$ as $g_{ik}$, and similarly we denote $g'_{ik} = \int p(D_i|u_i)p(u_i|\theta_k{}^t)u_i du_i$. Then we get

$$\mu_j = \frac{\sum_{i=1}^{N} \bar{\psi}_{ik} \frac{g'_{ik}}{g_{ik}}}{\sum_{i=1}^{N} \bar{\psi}_{ik}}$$

Finally, we need to derive the last unknown parameter $C_j$. As before, only $\log p(u_i|\theta_k)$ contains $C_j$. We start the derivation from the following:

$$\frac{\partial Q'}{\partial C_j} =$$
$$\partial \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} \bar{\psi}_{ik} \int p(u_i|D_i, \theta_k{}^t) \log \frac{1}{\sqrt{(2\pi)^m|C_k|}} du_i}{\partial C_j} -$$
$$\frac{1}{2} \partial \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} \bar{\psi}_{ik} \int (u_i - \mu_k)^T C_k{}^{-1}(u_i - \mu_k) p(u_i|D_i, \theta_k{}^t) du_i}{\partial C_j} = 0$$

which is

$$\frac{\partial Q'}{\partial C_j} =$$
$$-\frac{1}{2} \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} \bar{\psi}_{ik} \partial \int (\log|C_k|) p(u_i|D_i, \theta_k{}^t) du_i}{\partial C_j} -$$
$$\frac{1}{2} \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} \bar{\psi}_{ik} \partial \int (u_i - \mu_k)^T C_k{}^{-1}(u_i - \mu_k) p(u_i|D_i, \theta_k{}^t) du_i}{\partial C_j} = 0$$

Since

$$\frac{\partial \log|C_k|}{\partial|C_k|} = \frac{1}{|C_k|}$$

and the derivative of a determinant is

$$\frac{\partial|C_k|}{\partial C_k} = |C_k| C_k^{-1}$$

the derivative of the logarithm of the determinant is

$$\frac{\partial \log|C_k|}{\partial C_k} = C_k^{-1}.$$

The differentiation of a quadratic product is

$$\frac{\partial (u_i - \mu_k)^T C_k^{-1} (u_i - \mu_k)}{\partial C_k} = -C_k^{-1}(u_i - \mu_k)(u_i - \mu_k)^T C_k^{-1}$$

We combine the above two equations and get

$$\frac{\partial Q'}{\partial C_j} = -\frac{1}{2}\sum_{i=1}^{N} \bar{\psi}_{ik} \int C_j^{-1} p(u_i|D_i, \theta_k{}^t) du_i + \frac{1}{2}\sum_{i=1}^{N} \bar{\psi}_{ik} \int p(u_i|D_i, \theta_k{}^t) C_j^{-1}(u_i - \mu_j)(u_i - \mu_j)^T C_j^{-1} du_i = 0$$

After left and right multiplying by $C_j$ both sides and omitting non-zero constants, we get

$$\sum_{i=1}^{N} \bar{\psi}_{ik} C_j \int p(u_i|D_i, \theta_k{}^t) du_i - \sum_{i=1}^{N} \bar{\psi}_{ik} \int p(u_i|D_i, \theta_k{}^t)(u_i - \mu_j)(u_i - \mu_j)^T du_i = 0$$

and get finally

$$C_j = \frac{\sum_{i=1}^{N} \bar{\psi}_{ik} \int p(u_i|D_i, \theta_k{}^t)(u_i - \mu_j)(u_i - \mu_j)^T du_i}{\sum_{i=1}^{N} \bar{\psi}_{ik} \int p(u_i|D_i, \theta_k{}^t) du_i}$$

The integral in the denominator is equal to one so we get

$$C_j = \frac{\sum_{i=1}^{N} \bar{\psi}_{ik} \int \frac{p(D_i|u_i) p(u_i|\theta_k{}^t)(u_i - \mu_j)(u_i - \mu_j)^T}{p(D_i|\theta_k{}^t)} du_i}{\sum_{i=1}^{N} \bar{\psi}_{ik}}$$

As before, we denote $g''_{ik} = \int p(D_i|u_i) p(u_i|\theta_k{}^t)(u_i - \mu_j)(u_i - \mu_j)^T du_i$. Then we have

$$C_j = \frac{\sum_{i=1}^{N} \bar{\psi}_{ik} \frac{g''_{ik}}{g_{ik}}}{\sum_{i=1}^{N} \bar{\psi}_{ik}}$$

### 6.3. Computing $g_{ik}$, $g'_{ik}$ and $g''_{ik}$

All the computations described above are simple summations and averages of the $g_{ik}$'s, $g'_{ik}$'s and $g''_{ik}$'s. The only non trivial computation involves $g_{ik}$, $g'_{ik}$ and $g''_{ik}$ themselves. The reason is that they contain integration.

The actual method used to compute them will depend on the form in which $p(D_i|u_i)$ is given. If we are given samples on a regular grid (which can be efficient for low dimensionality $u_i$) we can compute $p(u_i|{\theta_k}^t)$ on the same grid and the integral becomes a discrete summation. Let ${}^m u_i$ be the discrete samples of $u_i$ for $m = 1..M$. We can write then

$$p(D_i|u_i) = \sum_{m=1}^{M} p(D_i|{}^m u_i) Sa({}^m u_i - u_i) =$$

so the integral

$$g_{ik} = \int_{u_i} p(D_i|u_i) p(u_i|{\theta_k}^t) du_i =$$

$$\int_{u_i} \sum_{m=1}^{M} p(D_i|{}^m u_i) Sa({}^m u_i - u_i) p(u_i|{\theta_k}^t) du_i =$$

$$\sum_{m=1}^{M} p(D_i|{}^m u_i) \int_{u_i} Sa({}^m u_i - u_i) p(u_i|{\theta_k}^t) du_i$$

is transformed into a sum of several definite integrals that can be computed analytically. If we approximate the sampling function with the zero mean Gaussian with variance $C_s$ the integral becomes much simpler

$$g_{ik} = \sum_{m=1}^{M} p(D_i|{}^m u_i) \int_{u_i} N({}^m u_i - u_i; 0, C_s) p(u_i|{\theta_k}^t) du_i =$$

$$\sum_{m=1}^{M} p(D_i|{}^m u_i) \int_{u_i} N({}^m u_i - u_i; 0, C_s) N(u_i; {\mu_k}^t, {C_k}^t) du_i =$$

$$\sum_{m=1}^{M} p(D_i|{}^m u_i) N({}^m u_i; {\mu_k}^t, {C_k}^t + C_s)$$

where $C_s$ is a matrix chosen so that the Gaussian approximates the sampling function. A diagonal matrix with every member of the diagonal equal to 0.4 is adequate for our purposes if the samples are on integer values of $u_i$. A similar procedure can be followed to compute the other $g$'s

$$g'_{ik} = \sum_{m=1}^{M} p(D_i|{}^m u_i) N({}^m u_i; {\mu_k}^t, {C_k}^t + C_s)\, {}^m u_i$$

$$g''_{ik} = \sum_{m=1}^{M} p(D_i|{}^m u_i) N({}^m u_i; {\mu_k}^t, {C_k}^t + C_s)({}^m u_i - \mu_k)^T({}^m u_i - \mu_k)$$

It is obvious that the above procedure is not suitable for high dimensionality $u_i$'s. An alternative is that $p(D_i|u_i)$ is given in a standard analytic form which would allow the analytic computation of the $g$'s. There are many good candidate standard forms but one of the most adaptable is Weighted Sum of Gaussians (similar to Mixture of Gaussians but without the requirement that the mixture probabilities sum up to one). Let

$$p(D_i|u_i) = \sum_{m=1}^{m_i} q_{im} N(u_i, \mu_{im}, C_{im})$$

where $q_{im}$ is the mixture weight for Gaussian $m$ for data point $u_i$, and $\mu_{im}$ and $C_{im}$ are the $m^{th}$ mean and covariance for point $u_i$. As before $p(u_i|{\theta_k}^t)$ is a Gaussian, which we denote as $N(u_i; \mu_k^t, C_k^t)$. So we can write

$$\begin{aligned} g_{ik} &= \int p(D_i|u_i)\, p(u_i|{\theta_k}^t) du_i = \\ &\int \left( \sum_{m=1}^{m_i} q_{im} N(u_i, \mu_{im}, C_{im}) \right) N(u_i, \mu_k^t, C_k^t) du_i \\ &\sum_{m=1}^{m_i} q_{im} \int N(u_i, \mu_{im}, C_{im}) N(u_i, \mu_k^t, C_k^t) du_i \end{aligned}$$

The product of two Gaussians is a simple function of the two sets of parameters $C_{im}$, $C_j$, $\mu_{im}$ and $\mu_j$

$$\begin{aligned} &N(u_i; \mu_{im}, C_{im}) N(u_i; \mu_k^t, C_k^t) = \\ &N(u_i; \mu_{imk}, C_{imk})\, \frac{N(0; \mu_{im}, C_{im}) N(0; \mu_k^t, C_k^t)}{N(0; \mu_{imk}, C_{imk})} = \\ &N(u_i; \mu_{imk}, C_{imk})\, f_{imk} \end{aligned}$$

where

$$\begin{aligned} C_{imk} &= \left( C_{im}^{-1} + (C_k^t)^{-1} \right)^{-1} \\ \mu_{imk} &= C_{imk}(C_{im}^{-1}\mu_{im} + (C_k^t)^{-1}\mu_k^t) \\ f_{imk} &= \frac{N(0; \mu_{im}, C_{im}) N(0; \mu_k^t, C_k^t)}{N(0; \mu_{imk}, C_{imk})} \end{aligned}$$

and since we know that the integral of the Gaussian is the unity

$$g_{ik} = \sum_{m=1}^{m_i} q_{im} f_{imk}$$

In a very similar fashion we can derive the expression for $g'_{ik}$ and $g''_{ik}$

$$g'_{ik} = \int p(D_i|u_i)p(u_i|{\theta_k}^t)u_i du_i =$$
$$\int \left( \sum_{m=1}^{m_i} q_{im} N(u_i, \mu_{im}, C_{im}) \right) N(u_i, \mu_k^t, C_k^t) u_i du_i$$
$$\sum_{m=1}^{m_i} q_{im} \int N(u_i; \mu_{imk}, C_{imk}) f_{imk} u_i du_i$$
$$\sum_{m=1}^{m_i} q_{im} f_{imk} \int N(u_i; \mu_{imk}, C_{imk}) u_i du_i =$$
$$\sum_{m=1}^{m_i} q_{im} f_{imk} \mu_{imk}$$

and

$$g''_{ik} = \sum_{m=1}^{m_i} q_{im} f_{imk} C_{imk}$$

### 6.4. Summary of the Algorithm

While the derivation is rather complicated the steps one needs to follow to execute one iteration of the algorithm are straightforward. We will follow the established practice and name the two distinct groups of steps *Expectation* and *Maximization* respectively although this association is by no means direct. The Expectation step is

$$\bar{\psi}_{ik} = \frac{{\pi_k}^t g_{ik}}{\sum_{j=1}^{K} {\pi_j}^t g_{ij}}$$

and the Maximization step is:

$$\pi_j = \frac{\sum_{i=1}^{N} \bar{\psi}_{ik}}{\sum_{k=1}^{K} \sum_{i=1}^{N} \bar{\psi}_{ik}}$$
$$\mu_j = \frac{\sum_{i=1}^{N} \bar{\psi}_{ik} \frac{g'_{ik}}{g_{ik}}}{\sum_{i=1}^{N} \bar{\psi}_{ik}}$$
$$C_j = \frac{\sum_{i=1}^{N} \bar{\psi}_{ik} \frac{g''_{ik}}{g_{ik}}}{\sum_{i=1}^{N} \bar{\psi}_{ik}}$$

After we compute the parameters of all the Gaussians we can compute the $g_{ik}$'s to be used in the next step. If the $p(D_i|u_i)$ is given in discrete form as a set of $M$ samples on a regular grid in the space spanned by $u_i$ then

$$g_{ik} = \sum_{m=1}^{M} p(D_i|{}^m u_i) N({}^m u_i; {\mu_k}^t, {C_k}^t + C_s)$$

$$g'_{ik} = \sum_{m=1}^{M} p(D_i|{}^m u_i)\, {}^m u_i N({}^m u_i; {\mu_k}^t, {C_k}^t + C_s)$$

$$g''_{ik} = \sum_{m=1}^{M} p(D_i|{}^m u_i)({}^m u_i - \mu_k)^T ({}^m u_i - \mu_k) N({}^m u_i; {\mu_k}^t, {C_k}^t + C_s)$$

If on the other hand $p(D_i|u_i)$ is given as a weighted sum of Gaussians then

$$C_{imk} = \left( C_{im}^{-1} + (C_k^t)^{-1} \right)_{-1}$$

$$\mu_{imk} = C(C_{im}^{-1}\mu_{im} + (C_k^t)^{-1}\mu_k^t)$$

$$f_{imk} = \frac{N(0; \mu_{im}, C_{im}) N(0; \mu_k^t, C_k^t)}{N(0; \mu_{imk}, C_{imk})}$$

$$g_{ik} = \sum_{m=1}^{m_i} q_{im} f_{imk}$$

$$g'_{ik} = \sum_{m=1}^{m_i} q_{im} f_{imk_{imk}}$$

$$g''_{ik} = \sum_{m=1}^{m_i} q_{im} f(C_{im}, C_k, \mu_{im}, \mu_k) C_{imk}$$

## 7. Conclusion

In this report, we reviewed briefly the concepts of *Clustering*, *Expectation Maximization* and *Mixture of Gaussians* and then developed the background for clustering of unobserved data under the Mixture of Gaussians model. We used *Maximum Likelihood* estimation to do this and since this problem involves hidden variables (the unobserved data and the membership) we used the *Expectation Maximization* (EM) method.

Our fundamental assumption is that the data is not directly observed but we assume that its probability distribution is known. Each datum $u_i$ is conditioned on a set of known parameters $D_i$ and that datum $u_{i'}$ is independent of $D_i$ if $i = i'$. We also assume that the $D_i$s are not directly dependent on the clustering parameters which gave arise to the quazi-dependence assumption we used throughout the report. Armed with these assumptions we developed the equations for the iterative estimation of the gaussian clusters. If the probability distribution of the data is assumed to be a parametric form the formulas could be further further simplified and we plan to do this for future research.

## References

1. Frank Dellaert, *The Expectation Maximization Algorithm*, http://www.cc.gatech.edu/~dellaert/html/publications.htm (Feb. 2002).
2. Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum likelihood from incomplete data via the EM Algorithm," *Journal of the Royal Statistical Society* **B 39** pp. 1-38 (1977).
3. Geoffrey J. McLachlan and Thriyambakam Krishnan, *The EM Algorithm and Extensions*, Wiley-Interscience (1997).
4. Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, Wiley Interscience (2001).
5. Athanasios Papoulis, *Probability, Random Variables and Stochastic Processes*, McGrow-Hill (1984).
6. George Bebis, *Mathematical Methods Computer Vision*, http://www.cs.unr.edu/~bebis/MathMethods/EM/lecture.pdf ().
7. C. Fraley and A. E. Raftery, "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis," *Tech. Report*, ().
8. Gigeon Schwarz, "Estimating the Dimension of a Model," *The annals of Statistics* **6**(2) pp. 461-464 (March. 1978).