McGill University

Faculty of Science

Mathematics and Statistics

Winter 2022: Directed Reading Program

Etienne SEBAG

Mentor: Mr. James McVittie

Abstract

The project commences by exploring the **Expectation-Maximization (E-M) algorithm**. This is a well-known and widely used method in statistics used to find maximum likelihood estimates of parameters in models where we regard the observations as incomplete data. The paper then investigates a detailed example using the multinomial distribution to further motivate and illustrate the algorithm. Another application of the algorithm, applied to a Gaussian mixture model, is also investigated.

Contents

Chapter 1

Introduction and Explanation of Algorithm

1.1 Motivation, Set-up of the Algorithm, Key Definitions and Properties

The Expectation-Maximization algorithm (short: E-M algorithm) is an **iterative procedure** used to find maximum likelihood estimates in statistical models having incomplete data. Whilst the term *incomplete data* may seem slightly enigmatic, a precise mathematical definition can be offered: [4].

Definition 1.1.1. ***Incomplete Data.***
Let f be a non-injective mapping between two sample spaces $\mathcal{X}$ and $\mathcal{Y}$. Assume that the data collected, ie: the observed data, is denoted by $\boldsymbol{y}$. We say that $\boldsymbol{y}$ is a realization from the sample space $\mathcal{Y}$ and as such $\boldsymbol{y} \in \mathcal{Y}$. Let $\boldsymbol{x} \in \mathcal{X}$ be data observations which are not observed directly. Instead, $\boldsymbol{x}$ is observed indirectly through the observed data $\boldsymbol{y}$. By considering the many-to-one mapping $f : \mathcal{X} \to \mathcal{Y}$, and recalling that we only observe $\boldsymbol{y}$, we assume that $\boldsymbol{x}$ is only known to reside in $\mathcal{X}(\boldsymbol{y})$, a subset of $\mathcal{X}$ specified by the relationship $\boldsymbol{y} = f(\boldsymbol{x}) \in \mathcal{Y}$.

Under this setup, we say that $\mathbf{x}$ is the **complete data** and that $\mathbf{y}$ is the **incomplete data**. As a trivial example, a complete data set can be thought of to be draws from a population, and an incomplete data set can be regarded as draws from a subset of that population. Another more compelling example of the distinction between complete and incomplete can be found in Chapter 2. Typically, when working with problems which lend themselves naturally to an application of the E-M algorithm, the incomplete data $\mathbf{y}$ is known

and fixed, whereas the complete data $\mathbf{x}$ can be appropriately chosen in certain circumstances to facilitate the computations of the algorithm.

The E-M Algorithm was first discussed in a 1977 paper by Dempster, Laird, and Rubin when they introduced the algorithm and guaranteed that its solution converges to the maximum likelihood estimates of the desired parameters. The authors also extended the algorithm to various levels of generality, the broadest one being that of applying the algorithm to exponential families. Under the formalism of exponential families, we will see in section 1.2 how the algorithm is greatly simplified. Indeed, their paper also pushed the algorithm towards its full generality by arguing that it indeed achieves convergence towards the MLE's even outside of the exponential family.

In the current statistical literature, the E-M algorithm has a tremendous number of applications to virtually any model with **latent variables**. Examples of where this algorithm is commonly applied includes: factor analysis, censored data, finite mixtures, hyperparameter estimation, and iteratively re-weighted least squares, among others.

Definition 1.1.2. ***Latent Variable***
A ***latent variable*** *is a variable whose value is not directly observed. Rather, its value can only be inferred from other known values in the model. Following the terminology introduced earlier, if* $\boldsymbol{x}$ *is latent then it is unobservable.*

It is also worthwhile to mention that the E-M algorithm also works under a Bayesian framework to find maximum a posteriori (MAP) estimates. However, this paper will focus to an overview of the algorithm under a parametric model where the parameter to be estimated is a fixed value (i.e. under a frequentist paradigm). In brief, the algorithm works towards finding the MLE's by iteratively applying a two-fold sequence of steps: an *E-step (expectation step)* and an *M-step (maximization step)*. In order to understand what happens at each of those steps, we need to introduce additional notation:

Let $f(\mathbf{x}|\theta)$ denote the probability density(mass) function associated with the complete-data vector $\mathbf{x}$ which is dependent on a (possibly multivariate) unknown parameter $\theta \in \Theta$ **(complete-data specification)** where Θ denotes the parameter space. Similarly, let $g(\mathbf{y}|\theta)$ the **incomplete-data specification**.

Now, the overarching premise of the E-M algorithm lies on the fact that traditional maximization of the log-likelihood of the observed data $\mathbf{y}$ (yielding the MLE) might not be analytically solvable. Hence, in order to find the MLE of our observed data $\mathbf{y}$, the E-M algorithm actually finds a way to incorporate the complete-data specification $f(\mathbf{x}|\theta)$. In other words, via the E-M algorithm, we find a way to link the observed incomplete-data model to a complete-data model in order to enable easier maximum-likelihood computations. To this end, for continuous densities, we have that the complete-data specification is related to the incomplete data specification via:

$$g(\mathbf{y}|\theta) = \int_{\mathbf{x}\in\mathcal{X}(\mathbf{y})} f(\mathbf{x}|\theta)dx \tag{1.1}$$

Similarly, when dealing with discrete random variables, we simply replace the integral sign in Equation (1.1) with a summation sign. Furthermore, Equation (1.1) illustrates the essence of the algorithm: we are resorting to using the complete-data specification and averaging it over all possible latent/missing variable settings [5].

Procedure:
The first step of the algorithm is an initialization step, whereby we let $\theta^0 \in \Theta$ be an initial reasonable estimate of the true parameter θ. At the first iteration, we begin with the E-step, in which we are interested in calculating an expectation quantity denoted $Q(\theta|\theta^0)$, and whose formula is given by:

$$Q(\theta|\theta^0) = \mathbb{E}\left(\ln f(X|\theta)|\mathbf{y}, \theta^0\right) \tag{1.2}$$

In other words, at the first E-step, we are computing the expected value of the log-likelihood of the complete-data specification, given the observed data $\mathbf{y}$ AND the first estimate of the parameter θ^0.

In general, at the k-th iteration of the algorithm, the quantity $Q(\theta|\theta^{k-1}) = \mathbb{E}\left(\ln f(X|\theta)|\mathbf{y}, \theta^{k-1}\right)$ is computed, where we condition on the last iterative estimate of the parameter, θ^{k-1}, which was found at the k-th iteration.

The M-step, at the k-th iteration, simply works to find the new, updated parameter θ^k which maximizes $Q(\theta|\theta^{k-1})$. In other words, the M-step can be summarized as follows:

$$\theta^k = argmax\ Q(\theta, \theta^{k-1}) \tag{1.3}$$

The recursive procedure continues by alternating between the E-step and an M-step. It is important to mention that each step of the E-M algorithm **monotonically increases the log-likelihood of the complete-data specification**. In this sense, the algorithm can only be improved on at each step. Furthermore, convergence of the E-M algorithm to a critical point of the likelihood function happens only when we assume that the likelihood function of θ is uniformly bounded from above. Finally, it is worthwhile to state that the rate of convergence of the algorithm can be increased whenever there is a reduction in the amount of missing data.

1.2 Connection to Exponential Families

Recall that the broadest level of generalization of the E-M algorithm introduced in the 1977 paper was when the complete-data specification $f(\mathbf{x}|\theta)$ is in an exponential family form. Let X be a random variable distributed according to an **exponential family** and let S_X denote the support of X, $\theta \in \Theta \subset \mathbb{R}^d$ ($d \geq 1$ is the dimension of the parameter space). The pdf/pmf for a member of the parametric **exponential family** is given by:

$$f(\mathbf{x}|\theta) = h(x)c(\theta)\ exp\left\{\omega^T(\theta) \cdot T(x)\right\}, \forall x \in S_X \tag{1.4}$$

with $h(x) > 0$, $c(\theta) > 0$ and also:

- $T(x) = \Big(t_1(x), t_2(x), ..., t_k(x)\Big)^T$ such that the $t_j(x)$'s are real-valued functions depending only on x.
- $\omega(\theta) = \Big(\omega_1(\theta), \omega_2(\theta), ..., \omega_k(\theta)\Big)^T$ such that the $\omega_j(\theta)$'s are real-valued functions depending only on θ.

Suppose that θ^{k-1} denotes the current estimate of θ at the k-th iteration of the algorithm. Observe that, when $f(\mathbf{x}|\theta)$ is an exponential family, the computation of $\ln f(\mathbf{x}|\theta)$ simplifies to:

$$\ln\big(h(x)\big) + \ln\big(c(\theta)\big) + \omega^T(\theta) \cdot T(x)$$

In the E-step, we compute $Q(\theta, \theta^{k-1})$, given by:

$$Q(\theta, \theta^{k-1}) = \mathbb{E}\Bigg(\ln(h(X)) + \ln\left(c(\theta)\right) + \omega^T(\theta) \cdot T(X) \Bigg| \mathbf{y}, \theta^{k-1} \Bigg)$$

(Remember that, at the k-th iteration, we are working to find an updated estimate for θ, denoted θ^k).

Simplifying further, the term $\ln(c(\theta))$ is treated as a constant and hence pulled out of the expectation yielding:

$$Q(\theta, \theta^{k-1}) = \ln\left(c(\theta)\right) + \mathbb{E}\bigg(\ln(h(X)) + \sum_{i=1}^{n} \omega_i(\theta)\ t_i(X) \mid \mathbf{y}, \theta^{k-1} \bigg)$$

$$Q(\theta, \theta^{k-1}) = \ln\left(c(\theta)\right) + \mathbb{E}\bigg(\ln(h(X) \mid \mathbf{y}, \theta^{k-1} \bigg) + \sum_{i=1}^{n} \omega_i(\theta) \mathbb{E}\bigg(t_i(X) \mid \mathbf{y}, \theta^{k-1} \bigg) \tag{1.5}$$

When working with exponential families, we have the well known result that when X has a distribution belonging to the exponential family, $\forall j \in \{1, ..., d\}$:

$$\mathbb{E}\left\{ \frac{\partial}{\partial \theta_j} \bigg(\sum_{i=1}^{n} \omega_i(\theta) t_i(X) \bigg) \right\} = -\frac{\partial}{\partial \theta_j} \bigg(ln\ c(\theta) \bigg) \tag{1.6}$$

As part of the M-step, we can maximize Equation (1.5) with respect to θ by taking the derivative with respect to θ and then setting it equal to 0. Note that the expectation of the term $\ln(h(X))$ gets dropped at this stage as it will depend only on θ^{k-1} and $\boldsymbol{y}$.

$$\frac{\partial}{\partial \theta}\ ln(c(\theta))\ +\ \frac{\partial}{\partial \theta} \left\{ \sum_{i=1}^{n} \omega_i(\theta) \mathbb{E}\bigg(t_i(X) \mid \mathbf{y}, \theta^{k-1} \bigg) \right\} = 0$$

The first term on the left side of the above equality can be expressed alternatively via Equation (1.6):

$$-\mathbb{E}\left\{ \frac{\partial}{\partial \theta_j} \bigg(\sum_{i=1}^{n} \omega_i(\theta) t_i(X) \bigg) \right\}\ +\ \frac{\partial}{\partial \theta} \left\{ \sum_{i=1}^{n} \omega_i(\theta) \mathbb{E}\bigg(t_i(X) \mid \mathbf{y}, \theta^{k-1} \bigg) \right\} = 0$$

From which we obtain:

$$\frac{\partial}{\partial \theta} \left\{ \sum_{i=1}^{n} \omega_i(\theta) \mathbb{E}\bigg(t_i(X) \mid \mathbf{y}, \theta^{k-1} \bigg) \right\}\ -\ \mathbb{E}\left\{ \frac{\partial}{\partial \theta_j} \bigg(\sum_{i=1}^{n} \omega_i(\theta) t_i(X) \bigg) \right\} = 0 \tag{1.7}$$

Working with Equation (1.7), we can treat each term one at a time. The first term may be expressed as:

$$\begin{aligned}
&\frac{\partial}{\partial \theta} \left\{ \sum_{i=1}^{n} \omega_i(\theta) \mathbb{E}\bigg(t_i(X) \mid \mathbf{y}, \theta^{k-1} \bigg) \right\} \\
&= \frac{\partial}{\partial \theta} \bigg(\omega_1(\theta) \mathbb{E}\big(t_1(X) \mid \mathbf{y}, \theta^{k-1} \big) + ... + \omega_n(\theta) \mathbb{E}\big(t_n(X) \mid \mathbf{y}, \theta^{k-1} \big) \bigg) \\
&= \frac{\partial}{\partial \theta} \bigg(\omega_1(\theta) \mathbb{E}\big(t_1(X) \mid \mathbf{y}, \theta^{k-1} \big) \bigg) + ... + \frac{\partial}{\partial \theta} \bigg(\omega_n(\theta) \mathbb{E}\big(t_n(X) \mid \mathbf{y}, \theta^{k-1} \big) \bigg) \\
&= \mathbb{E}\bigg(t_1(X) \mid \mathbf{y}, \theta^{k-1} \bigg) \frac{\partial}{\partial \theta} \left\{ \omega_1(\theta) \right\} + ... + \mathbb{E}\bigg(t_n(X) \mid \mathbf{y}, \theta^{k-1} \bigg) \frac{\partial}{\partial \theta} \left\{ \omega_n(\theta) \right\}
\end{aligned}$$

$$= \sum_{i=1}^{n} \left\{ \mathbb{E}\big(t_i(X) \mid \mathbf{y}, \theta^{k-1}\big) \frac{\partial}{\partial \theta} \omega_i(\theta) \right\}$$

Working with the second term from Equation (1.7), and following a similar argument as above (using the sum rule for derivatives, rules of expectations), we have that:

$$\mathbb{E}\left\{ \frac{\partial}{\partial \theta_j} \left(\sum_{i=1}^{n} \omega_i(\theta) t_i(X) \right) \right\} = \sum_{i=1}^{n} \left\{ \mathbb{E}\big(t_i(X)\big) \frac{\partial}{\partial \theta} \omega_i(\theta) \right\}$$

Substituting our simplified expressions back into Equation (1.7), we obtain:

$$\sum_{i=1}^{n} \left\{ \mathbb{E}\big(t_i(X)) \mid \mathbf{y}, \theta^{k-1}\big) \frac{\partial}{\partial \theta} \omega_i(\theta) \right\} \; - \; \sum_{i=1}^{n} \left\{ \mathbb{E}\big(t_i(X)\big) \frac{\partial}{\partial \theta} \omega_i(\theta) \right\} = 0$$

By combining the summations:

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \omega_i(\theta) \left(\mathbb{E}(t_i(X)) \mid \mathbf{y}, \theta^{k-1}) - \mathbb{E}(t_i(X)) \right) = 0$$

Which implies that: $\mathbb{E}\left(t_i(X)) \mid \mathbf{y}, \theta^{k-1}\right) - \mathbb{E}\left(t_i(X)\right) = 0$; if and only if:

$$\frac{\partial}{\partial \theta} \omega_i(\theta) \neq 0 \; \forall \; i \in \{1, ..., n\} \;\; \text{and} \; sgn\left(\frac{\partial}{\partial \theta} \omega_i(\theta) \right) = sgn\left(\mathbb{E}(t_i(X)) \mid \mathbf{y}, \theta^{k-1}) - \mathbb{E}(t_i(X)) \right)$$

Hence, when the complete-data specification is an exponential family, the M-step amounts to having:

$$\boxed{\mathbb{E}\left(t_i(X)) \mid \mathbf{y}, \theta^{k-1}\right) = \mathbb{E}\left(t_i(X)\right)}$$

Chapter 2

Worked Multinomial Genetic Model Example

In order to better understand the E-M algorithm in action, we can proceed with a detailed mathematical treatment applied to an illustrating example mentioned in Dempster, Laird, and Rubin's original paper.

In this example, our sample size is $n = 197$ animals, which are distributed following a multinomial distribution into $k = 4$ categories. We designate our **observed data vector** as $\mathbf{y}$ and have that:

$$\mathbf{y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

Further assume that the model specifies the following probabilities:

$$p = (p_1, p_2, p_3, p_4) = \left(\frac{1}{2} + \frac{1}{4}\theta,\ \frac{1}{4}(1-\theta),\ \frac{1}{4}(1-\theta),\ \frac{1}{4}\theta\right)$$

Note that $\theta \in \Theta$ is an unknown population parameter which needs to be estimated and that we have $0 \leq \theta \leq 1$.

In order to be in a framework where we can apply the E-M algorithm, we will now regard $\mathbf{y}$ as being **incomplete data** and actually denote by $\mathbf{x}$ the complete data vector coming from a multinomial population with $k = 5$ categories.

Hence, we have: $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ and with corresponding probabilities:

$$q = (q_1, q_2, q_3, q_4, q_5) = \left(\frac{1}{2},\ \frac{1}{4}\theta,\ \frac{1}{4}(1-\theta),\ \frac{1}{4}(1-\theta),\ \frac{1}{4}\theta\right)$$

Comparing the complete data vector $\mathbf{x}$ with the incomplete data vector $\mathbf{y}$, it is easy to observe the following one-to-one correspondences between their respective probabilities. Indeed: $q_3 = p_2$, $q_4 = p_3$, and $q_5 = p_4$. Also notice that $q_1 + q_2 = p_1$. Hence, we can say that the complete data $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ is related to the incomplete data whereby $y_1 = x_1 + x_2,\ y_2 = x_3,\ y_3 = x_4,\ y_4 = x_5$.

In this example, from the fact that $y_1 = x_1 + x_2$, we can say that x_1 and x_2 are latent variables because we never directly observe their individual values. In other words, even though we have observed what y_1 is, just knowing its value will not tell us specifically the values of x_1 and x_2. From our complete data vector $\mathbf{x}$, we also have its **complete-data specification**, which is simply the probability mass function (or probability density function) of the complete data. It is designated by $f(x|\theta)$. Here, recall that $f(x|\theta)$ has a multinomial pmf given by:

$$f(x|\theta) = \frac{n!}{x_1!x_2!x_3!x_4!x_5!}\left(\frac{1}{2}\right)^{x_1}\left(\frac{1}{4}\theta\right)^{x_2}\left(\frac{1}{4}(1-\theta)\right)^{x_3}\left(\frac{1}{4}(1-\theta)\right)^{x_4}\left(\frac{1}{4}\theta\right)^{x_5} \tag{2.1}$$

We can take the logarithm of the complete data specification to obtain the **complete-data log-likelihood**, which is what we will use throughout the remainder of the algorithm. Let $\mathcal{L}(x|\theta)$ be the complete-data likelihood. We now calculate $\ln(\mathcal{L}(x|\theta))$, the complete-data log-likelihood, as a function of θ. This means that any terms independent of θ have been discarded. We hence obtain:

$$\ln(\mathcal{L}(x|\theta)) = x_2 \ln\left(\frac{1}{4}\theta\right) + x_3 \ln\left(\frac{1}{4}(1-\theta)\right) + x_4 \ln\left(\frac{1}{4}(1-\theta)\right) + x_5 \ln\left(\frac{1}{4}\theta\right)$$

By further simplifying using logarithmic rules, and only retaining the terms containing θ, we have, for the complete-data log-likelihood:

$$\ln(\mathcal{L}(x|\theta)) = (x_2 + x_5)\ln(\theta) + (x_3 + x_4)\ln(1-\theta) \tag{2.2}$$

The main issue with Equation (2.2) is that this log-likelihood equation cannot be maximized as x_2 is a latent variable and as such is unobservable. However, the E-step of the E-M algorithm provides a solution.

Step 1: Provide an initial value/estimate for the parameter θ at the 0-th iteration. Denote this by θ^0.

Step 2: E-step We are interested now in the computation of $Q(\theta|\theta^0)$:

$$Q(\theta \mid \theta^0) = \mathbb{E}\left(\ln(\mathcal{L}(X|\theta)) \mid \mathbf{y}, \theta^0\right)$$

$$Q(\theta \mid \theta^0) = \mathbb{E}\bigg((X_2 + X_5)\ln(\theta) + (X_3 + X_4)\ln(1-\theta) \mid y_1, y_2, y_3, y_4, \theta^0 \bigg)$$

Now, remark that $y_2 = x_3, y_3 = x_4, y_4 = x_5$ have all been revealed, and, as such, X_3, X_4 and X_5 are deterministic and no longer random. Hence, we can simplify as follows:

$$Q(\theta \mid \theta^0) = \mathbb{E}\bigg((X_2 \mid y_1, \theta^0) + X_5 \bigg) \ln(\theta) + (X_3 + X_4)\ \ ln(1-\theta)$$

Let Y_1 be the random variable corresponding to y_1. At this point, we need to compute the conditional expectation of X_2 given $Y_1 = X_1 + X_2$ given $\Theta = \theta^0$: Notice how in the E-step, we are, in essence, replacing X_2 - the latent random variable, by its conditional expectation given the observed data vector $\mathbf{y}$. The question becomes how we can find the expectation $\mathbb{E}\big(X_2 \mid Y_1\ ,\ \theta^0\big)$.

Note that the conditional distribution of $X_2 \mid Y_1,\ \theta^0$ is binomial since the counts in a single cell in a multinomial experiment precisely follow a binomial distribution. Precisely, we obtain:

$$X_2 \mid Y_1,\ \theta^0 \sim Binom\bigg(y_1, \frac{\frac{1}{4}\theta^0}{\frac{1}{2} + \frac{1}{4}\theta^0} \bigg)$$

Since there are $n = y_1 = 125$ trials, the probability of success of $X_2 \mid Y_1$ is indeed given by $\{\mathbb{P}(X_2 = x_2)/\mathbb{P}(Y_1 = x_1 + x_2)\} = (\theta^0/4)\ /\ (1/2 + \theta^0/4)$.

Standard results about the expectation of a binomial random variable imply that:

$$\mathbb{E}\big(X_2 \mid Y_1\ ,\ \theta^0\big) \ = \ \frac{y_1 \times \frac{1}{4}\theta^0}{\frac{1}{2} + \frac{1}{4}\theta^0} \ = \ \frac{\frac{125}{4}\theta^0}{\frac{1}{2} + \frac{1}{4}\theta^0} \ = \ \frac{125\theta^0}{2 + \theta^0} \ = \ \frac{y_1\theta^0}{2 + \theta^0} \ = \ x_2^{(0)}$$

Hence, we ultimately obtain:

$$Q(\theta \mid \theta^0) = \big(x_2^{(0)} + x_5\big)\ln(\theta) + \big(x_3 + x_4\big)\ln(1-\theta)$$

Step 3: M-step

In this step, we choose an update $\theta^{(1)}$ so that the quantity $Q(\theta \mid \theta^0)$ is maximized. Hence, let us maximize $Q(\theta \mid \theta^0)$ with respect to θ. We take the derivative, set it equal to zero, and solve for our new parameter update θ^1.

$$\frac{\partial}{\partial\theta}\Big\{\big(x_2^{(0)}+x_5\big)\ln(\theta)+\big(x_3+x_4\big)\ln(1-\theta)\Big\}=0$$

$$\frac{1}{\theta}\big(x_2^{(0)}+x_5\big)-\frac{1}{1-\theta}\big(x_3+x_4\big)=0$$

$$\big(x_2^{(0)}+x_5\big)\big(1-\theta\big)=\theta\big(x_3+x_4\big)$$

$$x_2^{(0)}-x_2^{(0)}\theta+x_5-x_5\theta=x_3\theta+x_4\theta$$

Collecting the θ terms on one side:

$$x_2^{(0)}+x_5=x_3\theta+x_4\theta+x_5\theta+x_2^{(0)}\theta$$

$$x_2^{(0)}+x_5=\theta\big(x_3+x_4+x_5+x_2^{(0)}\big)$$

And finally, solving for θ will give us our new updated parameter estimate θ^0 and complete the first iteration.

$$\boxed{\theta^{(1)}=\frac{x_2^{(0)}+x_5}{x_3+x_4+x_5+x_2^{(0)}}}$$

Now, the E-steps and M-steps alternate as the algorithm progresses. In general, we obtain, at the k-th iteration, that:

$$\theta^{(k)}=\frac{x_2^{(k-1)}+x_5}{x_3+x_4+x_5+x_2^{(k-1)}}\quad\text{where}\quad x_2^{(k-1)}=\frac{y_1\theta^{(k-1)}}{2+\theta^{(k-1)}}$$

Chapter 3

The E-M Algorithm and Gaussian Mixture Models

3.1 Background Information, Definitions

We start this section by introducing what is known as a **mixture of Gaussians**. Simply put, a Gaussian mixture is a model in which there is a linear superposition of a finite number of Gaussian densities. We say that each Gaussian density is a **component** of the mixture model, and further, that each Gaussian density contributes its own mean and variance-covariance matrix.

From this point forward, we shall denote by K the number of Gaussian distributions in our model.

Figure 3.1 beneath is a visual representation of a simple Gaussian mixture model in the one-dimensional case. [3].

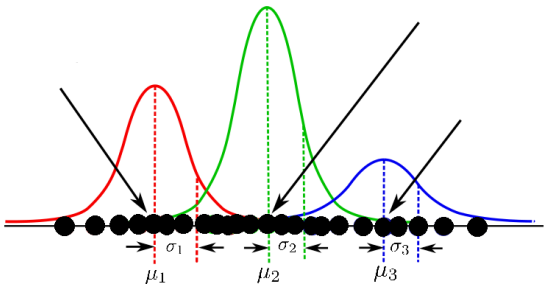

Figure 3.1: Gaussian Mixture with $K = 3$)

Gaussian mixture models are closely tied to the concept of **unsupervised classification**, as implied by figure 3.1. Indeed, when we are concerned with the issue of clustering, we want to group/classify observations according to data points that are "similar" via a suitable notion of distance. To quantify the similarity between two data points, we could, as the most general example, choose the Euclidean distance if we are operating under the assumption that the observations can be represented as vectors in $\mathbb{R}^p$.

Needless to say, when we create a cluster, the observations within that cluster are close in space and hence "similar". The connection with Gaussian mixture models lies within the fact that, **each of these K Gaussian distributions actually represents one cluster**. Therefore, a Gaussian mixture model groups together observations that come from a specific Gaussian density.

In figure 3.1, we have an underlying population model represented by the black data points along $\mathbb{R}$. The aim of classification studies is to figure out which data point belongs to which cluster/Gaussian component.

An important aspect of a Gaussian mixture model is the so-called **mixing probability**.

Definition 3.1.1. ***Mixing Probability (Mixing Coefficients)***
In a Gaussian mixture model, the ***mixing probability****, denoted by π_k, represents the probability, (or proportion), that the population is described by the k^{th} component (or, equivalently, the k^{th} Gaussian density). We can think of this mixing coefficient as the "weight" that the k^{th} component contributes to the overall Gaussian mixture model. As a general rule, we have that:*

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

Furthermore, since these mixing coefficients are probabilities themselves, we also have:

$$0 \leq \pi_k \leq 1$$

.

Now, the equation describing Gaussian mixture models is given below:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{3.1}$$

Equation (3.1) again reminds us that each of the K components of a mixture model has its own mean $\boldsymbol{\mu}_k$ and its own covariance matrix $\boldsymbol{\Sigma}_k$. [2].

3.2 Application of the E-M Algorithm

In turns out that the Gaussian mixture model is a very well-suited model in order to investigate the E-M algorithm, provided we can introduce latent variables somehow into the set-up: [1]

Let $\mathbf{x} = (x_1, ..., x_n)$ be a vector of n independent observations sampled from a Gaussian mixture where we will assume that each component of the mixture (ie: each Gaussian/normal density) is univariate. We will also further assume that the Gaussian mixture has $K = 2$ components. Hence, let $\mathbf{z} = (z_1, z_2)$ be the latent (unobservable) variables that specify from which of the two univariate normal densities an observation is coming from.

We further have that $\mathbb{P}(Z_i = 1) = \tau_1$ and that $\mathbb{P}(Z_i = 2) = \tau_2 = 1 - \tau_1$, where τ_1 and τ_2 represent the mixing coefficients as outlined in Definition (3.1.1). We are thus assuming that observations originate either from a first (1) or second (2) component. It is straightforward to realize that, for every one of the n observations, each of the Z_i's is a **Bernoulli** random variable. It is also important to remark that knowing the value of τ_1 immediately provides us with the value of τ_2.

With this set-up, we have:

$$X_i \mid (Z_i = 1) \sim \mathcal{N}(\mu_1, \sigma_1^2) \text{ and } X_i \mid (Z_i = 2) \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

The overarching goal of applying the E-M algorithm to this setup is to estimate the vector of unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\tau}, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

The incomplete-data likelihood is denoted by $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$ and it is equal to:

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^{n} \left\{ \left(\tau_1 f(x_i; \mu_1, \sigma_1^2)\right) + \left(\tau_2 f(x_i; \mu_2, \sigma_2^2)\right) \right\}$$

where $f(\cdot)$ represents the density of the univariate normal distribution.

In the E-step, we work with the complete-data likelihood in this case, which we specify if we know the values of the latent variables $\mathbf{z}$.

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} \left[\tau_1 f(x_i; \mu_1, \sigma_1^2)\right]^{\mathbb{1}(z_i=1)} \left[(1-\tau_1) f(x_i; \mu_2, \sigma_2^2)\right]^{\mathbb{1}(z_i=2)}$$

where, for each $i \in \{1, ..., n\}$, the indicator function $\mathbb{1}(z_i = 1)$ is equal to one if $z_i = 1$ and is equal to zero otherwise. Similarly, the indicator function $\mathbb{1}(z_i = 2)$ is equal to one if $z_i = 2$ and equal to zero otherwise. Conclusively, for each i-th observation, the complete-data likelihood consists of just one term; either $\tau_1 f(x_i; \mu_1, \sigma_1^2)$ or $(1-\tau_1) f(x_i; \mu_2, \sigma_2^2)$.

As a preliminary to the E-step of the algorithm, we must proceed by taking the logarithm of the complete-data likelihood, yielding:

$$\ln(\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})) = \sum_{i=1}^{n} \ln \left\{ \left[\tau_1 f(x_i; \mu_1, \sigma_1^2)\right]^{\mathbb{1}(z_i=1)} \left[(1-\tau_1) f(x_i; \mu_2, \sigma_2^2)\right]^{\mathbb{1}(z_i=2)} \right\}$$

$$\ln(\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})) = \sum_{i=1}^{n} \mathbb{1}(z_i = 1) \ln\left(\tau_1 f(x_i; \mu_1, \sigma_1^2)\right) + \mathbb{1}(z_i = 2) \ln\left((1-\tau_1) f(x_i; \mu_2, \sigma_2^2)\right)$$

$$\ln(\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})) = \ln(\tau_1) \sum_{i=1}^{n} \mathbb{1}(z_i = 1) \; + \; \sum_{i=1}^{n} \mathbb{1}(z_i = 1) \ln\left(f(x_i; \mu_1, \sigma_1^2)\right) \; +$$

$$\ln(1-\tau_1) \sum_{i=1}^{n} \mathbb{1}(z_i = 2) \; + \; \sum_{i=1}^{n} \mathbb{1}(z_i = 2) \ln\left(f(x_i; \mu_2, \sigma_2^2)\right)$$

Analogously to the multinomial example, the problem with the above likelihood is that it cannot be maximized directly because the z_i's are latent variables. Hence, this is where we invoke the E-step of the algorithm. We first provide an initial value/estimate for the parameter θ at the 0-th iteration denoted by θ^0 and then we compute the conditional expectation $Q(\theta, \theta^0)$.

$$Q(\theta, \theta^0) = \mathbb{E}\Big(\ln(\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})) \mid X_i = \boldsymbol{x}_i \; ; \; \theta^0 \Big)$$

$$Q(\theta, \theta^0) = \ln(\tau_1) \sum_{i=1}^{n} \mathbb{E}\Big(\mathbb{1}(Z_i = 1) \mid x_i, \theta^0 \Big) + \sum_{i=1}^{n} \mathbb{E}\Big(\mathbb{1}(Z_i = 1) \mid x_i, \theta^0 \Big) \ln\left(f(x_i; \mu_1, \sigma_1^2)\right) \; +$$

$$\ln(1-\tau_1) \sum_{i=1}^{n} \mathbb{E}\Big(\mathbb{1}(Z_i = 2) \mid x_i, \theta^0 \Big) + \sum_{i=1}^{n} \mathbb{E}\Big(\mathbb{1}(Z_i = 2) \mid x_i, \theta^0 \Big) \ln\left(f(x_i; \mu_2, \sigma_2^2\right)$$

Now, we need to remember that the expected value of an indicator random variable is simply the probability of the event defined by the indicator function. Hence:

$$\mathbb{E}\Big(\mathbb{1}(Z_i = 1) \mid x_i, \theta^0\Big) = \mathbb{P}\Big(\mathbb{1}(Z_i = 1) \mid x_i, \theta^0\Big)$$

In order to calculate this probability, we need to use Bayes' rule:

$$\mathbb{P}\big(\mathbb{1}(Z_i = 1) \mid x_i, \theta^0\big) = \frac{\mathbb{P}\big(X_i = \boldsymbol{x}_i; \theta^0 \mid z_i = 1\big)\mathbb{P}\big(Z_i = 1\big)}{\mathbb{P}\big(X_i = \boldsymbol{x}_i; \theta^0\big)}$$

Denote this above probability of interest, which amounts to being the conditional expectation we are after, by $(T_i)_1$.Therefore, we have, at the 0-th iteration:

$$\mathbb{E}\Big(\mathbb{1}(Z_i = 1) \mid x_i, \theta^0\Big) = {(T_i)_1}^{(0)} = \frac{{\tau_1}^{(0)} f(x_i; \mu_1^{(0)}, \sigma_1^{2(0)})}{{\tau_1}^{(0)} f(x_i; \mu_1^{(0)}, \sigma_1^{2(0)}) \ + \ (1 - {\tau_1}^{(0)}) f(x_i; \mu_2^{(0)}, \sigma_2^{2(0)})}$$

Similarly, we also have an expression for $(T_i)_2$:

$$\mathbb{E}\Big(\mathbb{1}(Z_i = 2) \mid x_i, \theta^0\Big) = {(T_i)_2}^{(0)} = \frac{(1 - {\tau_1}^{(0)}) f(x_i; \mu_2^{(0)}, \sigma_2^{2(0)})}{{\tau_1}^{(0)} f(x_i; \mu_1^{(0)}, \sigma_1^{2(0)}) \ + \ (1 - {\tau_1}^{(0)}) f(x_i; \mu_2^{(0)}, \sigma_2^{2(0)})}$$

Note: The T_i terms are called **membership probabilities**.

We can now replace the respective T_i terms into the expression for $Q(\theta, \theta_0)$ we had on the previous page:

$$Q(\theta, \theta^0) = \ln(\tau_1) \sum_{i=1}^{n} (T_i)_1 + \sum_{i=1}^{n} (T_i)_1 \ln\big(f(x_i; \mu_1, \sigma_1^2\big) + \ln(1-\tau_1) \sum_{i=1}^{n} (T_i)_2 + \sum_{i=1}^{n} (T_i)_2 \ln\big(f(x_i; \mu_2, \sigma_2^2\big)$$

Recall that, for a normally distributed random variable with mean μ and variance σ^2, its density function $f(\cdot)$ is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Taking the natural logarithm of the above:

$$\ln f(x) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

Hence, we can now fully write our expression for $Q(\theta, \theta^0)$ before moving on to the M-step of the algorithm.

$$Q(\theta,\theta^0) = \ln(\tau_1)\sum_{i=1}^{n}(T_i)_1 \; + \; \sum_{i=1}^{n}(T_i)_1\bigg(-\frac{1}{2}\ln(2\pi\sigma_1^2) - \frac{(x_i-\mu_1)^2}{2\sigma_1^2}\bigg) \; + \; \ln(1-\tau_1)\sum_{i=1}^{n}(T_i)_2 \; +$$

$$\sum_{i=1}^{n}(T_i)_2\bigg(-\frac{1}{2}\ln(2\pi\sigma_2^2) - \frac{(x_i-\mu_2)^2}{2\sigma_2^2}\bigg)$$

Now, in the M-step, determining the values that maximize the parameter vector $\boldsymbol{\theta}$ is relatively simple. As an example, we will show the M-step to obtain the MLE of μ_1. In order to do this, it suffices to differentiate the expression $Q(\theta,\theta^0)$ with respect to μ_1, setting it equal to 0, and solving accordingly.

$$\frac{\partial Q(\theta,\theta^0)}{\partial\mu_1} = \sum_{i=1}^{n}(T_i)_1\left[\frac{x_i-\mu_1}{\sigma_1^2}\right] = 0$$

$$\implies \sum_{i=1}^{n}(T_i)_1 x_i - \sum_{i=1}^{n}(T_i)_1\mu_1 = 0$$

$$\implies \hat{\mu}_1^{(1)} = \frac{\sum_{i=1}^{n}(T_i)_1^{(0)}x_i}{\sum_{i=1}^{n}(T_i)_1^{(0)}}$$

Acknowledgments:

I would like to thank my mentor, Mr. James McVittie, for his consistent support throughout this semester. I would also like to thank the DRP committee members for organizing this program and setting up the networks between the mentors and mentees. It has been a great fulfilling experience!

Bibliography

[1] Expectation–maximization algorithm, March 2022. Page Version ID: 1078223536.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[3] Oscar Contreras Carrasco. Gaussian Mixture Models Explained, February 2020.

[4] George Casella and Roger L. Berger. *Statistical Inference Vol. 70*. Duxbury Press Belmont, Ca, 1990.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. Publisher: [Royal Statistical Society, Wiley].