The EM Algorithm.
An Application to Finite Mixture Distributions

Comunicación Técnica No. D-95-18 (PE/CIMAT)

Eloísa Díaz Francés
CIMAT;
Dept. of Probability and Statistics;
A.P. 402; Guanajuato, Gto. 36000;
México; email: diazfran@cimat.mx

Abstract

The goal of this work is to provide a general overview of the EM algorithm that may aid possible users to gain quick understanding of its main ideas, as well as to suggest a set of basic and didactical bibliographical references that may help to obtain a better insight of it. The EM algorithm iteratively computes maximum likelohood estimates when the observations can be viewed as incomplete data, and consists of two stages: an Expectation step followed by a Maximization step (thus the name). The application of this algorithm for finding maximum likelihood estimates in the case of finite mixture distributions is discussed. A mixture distribution sample may be seen as an incomplete data case, since there is lack of information on which component of the distribution originated each of the observations. Two examples are presented of the use of the EM algorithm in the cases of finite mixtures of Normal as well as of Gumbel distibutions.

1 The EM Algorithm

The term EM was introduced in Dempster, Laird, and Rubin (1977) to designate a very general iterative algorithm for maximum likelihood estimation in incomplete data settings. The name refers to the two stages that constitute the algorithm: an Expectation step followed by a Maximation one. In that paper, Dempster *et al.* formalized and generalized relatively old approaches for handling missing data which had been proposed in special contexts in scientific literature. Their algorithm is closely related to the intuitive idea of filling in missing values, improving the estimates of the parameters, and then iterating. Overall, it consists of the following steps.

1. The parameters are assigned initial values by any reasonable method.
2. The missing values are replaced by their expected values (conditioned to the observed values and the current values of the parameters).
3. The parameters are re-estimated using the observed and updated missing values.
4. The missing values are recalculated assuming the new parameters are correct.
5. The parameters are re-estimated once again as in the third step, and so forth until a certain stopping criterion is fulfilled.

These ideas had been proposed in special contexts by different authors since 1926. However, the relevance of Dempster *et al.* (1977) was to expose the full generality of the algorithm, to prove several of its attractive properties, and to give a wide range of examples where the EM could be applied, including problems not usually considered to arise from missing data.

Denote the complete data vector by $Y = (Y_{obs}, Y_{mis})$,where Y_{obs} represents the observed part and Y_{mis} denotes the missing values. Let the complete data density be $f(Y;\Theta)$, which is related to the observed (or incomplete) sampling density $g(Y_{obs};\Theta)$ by

$$g\left(Y_{obs};\Theta\right) = \int\limits_{X(Y_{obs})} f(Y;\Theta)dY = \int\limits_{X(Y_{obs})} f\left(Y;\Theta\right)dY_{mis}, \tag{1}$$

where X is the sampling space and $X\left(Y_{obs}\right)$ is the subset of X which consists of all posible values that could be taken by the missing values given the observed ones. Note that $g\left(Y_{obs};\Theta\right)$ is like a marginal density in the sense that it was obtained by integrating the complete data density over all possible values for the missing ones given what was observed. The objetive is then to maximize the *observed likelihood,* which is proportional to $g\left(Y_{obs};\Theta\right)$, and thus obtain ML estimators for Θ. If (1) is differentiable and unimodal, ML estimators may be found as solutions to the equation $\bigtriangledown g\left(Y_{obs};\theta\right)$. When closed-form solutions cannot be found, then an iterative approach like the EM algorithm might prove helpful in solving the problem. The two steps of the EM algorithm will be next described.

1.1 E-STEP

The E-step finds the conditional expectations of the missing data (or complete sufficient statistics) given the observed data and current estimated parameters, and then substitutes these expectations for the missing data. Actually,the missing data are not necessarily substituted by EM; rather, the functions of Y_{mis} appearing in the complete data log likelihood are. Notice that the complete data likelihood is proportional to the complete data density; that is, $L\left(\Theta;Y\right) = C\left(Y\right)f\left(Y;\Theta\right)$, where $C\left(Y\right)$ is a positive bounded function of complete data which does not depend on the parameter Θ (which may be multidiminsional).

Specifically, let $\Theta^{(t)}$ be the current estimate of the parameter Θ. The E step of EM finds the conditional expectation of the complete data log likelihood, given that the correct values of Θ were $\Theta^{(t)}$, and given the observed data Y_{obs}. This conditional expectation is a function of Θ and will be denoted by $Q\left(\Theta \mid \Theta^{(t)}\right)$:

$$Q\left(\Theta \mid \Theta^{(t)}\right) = \int l\left(\Theta;Y\right) f\left(Y_{mis} \mid Y_{obs},\Theta^{(t)}\right) dY_{mis} = E\left[l\left(\Theta;Y\right) \mid Y_{obs},\Theta^{(t)}\right].$$

It is crucial to distinguish the two arguments of Q. The first one, Θ, is an argument of the full log likelihood $l\left(\Theta;Y\right)$. On the other hand, $\Theta^{(t)}$ is the parameter of the conditional distribution of the complete data Y, given the observed Y_{obs}, which is used to calculate the conditional expectation. Thus, the expectation step may be defined as follows: given the current estimate $\Theta^{(t)}$ of the parameter, calculate $Q\left(\Theta \mid \Theta^{(t)}\right)$ as a function of the dummy argument Θ (see Cox and Oakes, 1984). Also, note that the conditional expectation provides a way of filling in for the missing data (or functions of them), since it replaces the missing values with their expected counterparts given the distribution of the complete data as well as given the observed data. This will be clearly exhibited in the example of Little and Rubin (1987), presented in Section 1.5.

1.2 M-STEP

The M or Maximization step of EM determines $\Theta^{(t+1)}$ by finding the value of Θ that maximizes this expected log likelihood $Q\left(\Theta \mid \Theta^{(t)}\right)$, (viewed as a function of Θ). That is,

$$Q\left(\Theta^{(t+1)} \mid \Theta^{(t)}\right) \geq Q\left(\Theta \mid \Theta^{(t)}\right), \text{ for all } \Theta.$$

The heuristic idea here is that one would like to choose Θ to maximize $l\left(\Theta;Y\right)$, but since this loglikelihood is unknown, what can be maximized instead is its current expectation given the observed data and the current estimate $\Theta^{(t)}$ (Dempster et al., 1977).

2 Basic Properties and Advantages of the EM Algorithm

One of the most important properties of the EM is that each iteration of the algorithm increases the log-likelihood $l\left(\Theta \mid Y_{obs}\right)$, and if this function is bounded, then the sequence $\left\{l\left(\Theta^{(t)} \mid Y_{obs}\right)\right\}_{t=1}^{\infty}$ converges to a stationary value

of $l(\Theta \mid Y_{obs})$. Overall, if the sequence $\left\{\Theta^{(t)}\right\}_{t=1}^{\infty}$ converges, it will do so to a local maximum or saddle point of $l(\Theta \mid Y_{obs})$. In this sense, the EM converges reliably.

An advantage of the EM algorithm is that the two step that it involves are often very easy to calculate and to program; however, the rate of convergence can be extremely slow if a lot of data are missing (see Dempster et al., 1977). In the applications of the algorithm to finite Normal or Gumbel mixtures distributions, when there were two components in the mixtures and a sample size of 500, the EM converged in less than five seconds and ten iterations in most of the simulated samples (an 80486 personal IBM compatible computer was used and the algorithm was programmed in Gauss VM 3.6 computer language from Aptech Systems, Inc.).

In the case that the underlying complete data come from an exponential family whose ML estimates are easily computed, then each maximization step of the EM algorithm will be likewise easily calculated. Actually, in this situation, the EM algorithm may be formulated in terms of complete sufficient statistics.

3 The EM Algorithm when $f(Y;\Theta)$ Belongs to the Exponential Family.

When the complete data density has the regular exponential-family form:

$$f(Y;\Theta) = \frac{b(Y)}{a(\Theta)} \exp\left(\Theta\, t(Y)'\right),$$

the steps of the EM algorithm may be expressed in terms of a vector of complete sufficient statistics $t(Y)$, as follows (if Θ is k-dimensional, then $t(Y)$ is also k-dimensional):

E-step: Obtain an approximation to the complete-data vector of sufficient statistics $t(Y)$ by calculating the conditional expectation of the sufficient statistics given the observed data and the current estimate of the parameters,

$$E\left[t(Y) \mid Y_{obs}, \Theta^{(t)}\right].$$

M-step: Obtain $\Theta^{(t+1)}$ as the solution for Θ to the set of equations formed by equating the expectation of the complete-data sufficient statistic to the one calculated in the E-step,

$$E\left[t(Y);\Theta\right] = E\left[t(Y) \mid Y_{obs}, \Theta^{(t)}\right]. \tag{2}$$

These equations (2) are the familiar form of the likelihood equations for maximum-likelihood estimation given data from a regular exponential family.

Note that in this special case, the M-step involves the solution of a system of equations. However, this is equivalent to maximizing the likelihood funtion since these equations arise after differentiating $Q\left(\Theta \mid \Theta^{(t)}\right)$ and equating the corresponding partial derivatives to zero. (See Cox & Oakes, 1984, Section 11.4 for a full proof of this result).

A very illuminating and simple example, presented in Little and Rubin (1987), will hopefully serve to better understand the application of the EM algorithm to obtain ML estimates under an incomplete data setting, where the complete data density pertains to the exponential family.

3.1 Example 1. Univariate Normal Data, Little & Rubin (1987).

Suppose y_i are $iid\ N\left(\mu, \sigma^2\right)$, where y_i for $i = 1, ...m$ are observed, and y_i for $i = (m+1), ..., n$ are missing $(n > m)$. The expectation of each missing y_i given Y_{obs} and $\Theta = \left(\mu, \sigma^2\right)$ is μ. Since the normal density belongs to the regular exponential family, the log likelihood based on the complete data is linear in the sufficient statistics and the EM may be applied as in Section 1.4. The vector of sufficient statistics is

$$t(Y)' = \left(\textstyle\sum_1^n y_i, \sum_1^n y_i^2\right).$$

The E-step then calculates,

$$E\left[\textstyle\sum_1^n y_i \mid Y_{obs}, \Theta^{(t)}\right] = \sum_1^m y_i + (n-m)\,\mu^{(t)} \tag{3}$$

and

$$E\left[\sum_1^n y_i^2 \mid Y_{obs}, \Theta^{(t)}\right] = \sum_1^m y_i^2 + (n-m)\left[\left(\mu^{(t)}\right)^2 + \sigma^{2(t)}\right], \tag{4}$$

for current estimates $\mu^{(t)}$and $\sigma^{2(t)}$. (Initial values have to be provided for the first iteration). Notice in (3) that there are $(n-m)$ missing observations which are replaced by their expected value $\mu^{(t)}$. In (4), the sum of the squares of the $(n-m)$ missing observations is replaced by their expected value,

$$E\left(y_i^2\right) = \left(\mu^{(t)}\right)^2 + \sigma^{2(t)}.$$

For the M-step the expectation of the complete-data sufficient statistics $E\left[t\left(Y\right);\Theta\right]$ has to be calculated

$$E\left[\sum_1^n y_i; \Theta\right] = \sum_1^n E\left(y_i\right) = n\mu \tag{5}$$

and

$$E\left[\sum_1^n y_i^2; \Theta\right] = \sum_1^n E\left(y_i^2\right) = \sum_1^n \left(\sigma^2 + \mu^2\right) = n\sigma^2 + n\mu^2. \tag{6}$$

The updated version of the parameters $\mu^{(t+1)}$ and $\sigma^{2(t+1)}$ is obtained by equating (3) to (5) and solving for μ, and by equating (4) to (6) and solving for σ^2. Afterwards, substituting $\mu^{(t+1)}$ for μ, as well as $\sigma^{2(t+1)}$ for yields:

$$\mu^{(t+1)} = E\left[\sum_1^n y_i \mid Y_{obs}, \Theta^{(t)}\right]/n, \tag{7}$$

and

$$\sigma^{2(t+1)} = E\left[\sum_1^n y_i^2 \mid Y_{obs}, \Theta^{(t)}\right]/n - \left(\mu^{(t+1)}\right)^2 \tag{8}$$

Therefore to proceed with subsequent iterations, (3) and (4) may be updated with $\mu^{(t+1)}$ and $\sigma^{2(t+1)}$,and then $\mu^{(t+2)}$ and $\sigma^{2(t+2)}$ may be obtained by calculating (7) and (8) with the updated values. The iterations will end when a stopping criterium is fulfilled.

4 Application of EM to Mixture Distributions

The crucial problem when dealing with a mixture distribution is that one does not know which distribution in the mixture was the one that produced each observation in the sample. If this is known, then there is no mixture problem at all, since one can apply the usual statistical procedures to each separate distribution after classifying the observations in the sample. Therefore, the case of mixture distributions may be seen as an incomplete data problem.

The complete data may be defined to be then, a sample of size n of random *iid* elements $Y= (X, \varepsilon)$, where X is an observed random variable and ε is an unobserved random vector of dimension c (in the case of a mixture of c distributions). Note that each observation x_i is associated with an unobserved indicator vector $\varepsilon_{i=}(\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{ic})$, whose components are indicator variables which are all zero except for one, say $\varepsilon_{ij} = 1$, and indicates that the observation x_i pertains to the j-th distribution. That is, the observed part of Y is $Y_{obs} = X$ and the unobserved part is $Y_{mis} = \varepsilon$.

A "complete data" finite mixture density of c components may be represented in the following way:

$$f\left(x_i;\theta\right) = \prod_{k=1}^{c} f_k\left(x_i;\theta_k\right)^{\varepsilon_{ki}}, \tag{9}$$

where $f_k\left(x_i;\theta_k\right)$ is the density of the k-th component and the $\varepsilon_{ki},\ k = 1, ..., c$, are unobserved indicator variables which indicate which component generated the i-th observation.

An observed finite mixture density of c components may be expressed with the aid of additional parameters $p_k,\ k = 1, ..., c$ which indicate the proportion in which the k-th component contributes to the global density and which add up to one, $\sum_{k=1}^{c} p_k = 1$,

$$f\left(x_i;\theta\right) = \sum_{k=1}^{c} p_k\ f_k\left(x_i;\theta_k\right). \tag{10}$$

The relationship between these two expressions is

$$E[\varepsilon_{ki}] = P[\varepsilon_{ki} = 1] = p_k, \text{ for } i = 1, ..., n.$$

The expression (9) leads directly to the application of the EM algorithm since it makes evident the missing part of the observations (the indicator variables $\varepsilon_{ki},\ i = 1, ...n.$ and $k = 1, ..., c$). This form will simplify calculations when working with the corresponding log likelihood.

The complete data log likelihood from (9) is given by

$$l(\theta; Y) = \sum_{i=1}^{n}\sum_{k=1}^{c} \varepsilon_{ik} \ln[f_k(\theta_k; x_i)].$$

Calculating the E-step with this expression yields:

$$\begin{aligned} Q\left(\Theta \mid \Theta^{(t)}\right) &= \int l(\Theta; Y) f\left(Y_{mis} \mid Y_{obs}, \Theta^{(t)}\right) dY_{mis} = E\left[l(\Theta; Y) \mid Y_{obs}, \Theta^{(t)}\right] = \\ &= \sum_{k=1}^{c} E\left[\left(\sum_{i=1}^{n} \varepsilon_{ik} \ln\left(f_k(\theta_k; x_i)\right)\right) \mid x_i, \Theta^{(t)}\right] \\ &= \sum_{k=1}^{c}\sum_{i=1}^{n} \ln[f_k(\theta_k; x_i)]\, E\left[\varepsilon_{ik} \mid x_i, \Theta^{(t)}\right]. \end{aligned} \tag{11}$$

Note that here, the indicator variables which were unobserved, will be replaced with a conditional expected value. Using Bayes Theorem, the "replacing value" is equal to

$$\begin{aligned} E\left[\varepsilon_{ik} \mid \Theta^{(t)}\right] &= P\left[\varepsilon_{ik} = 1 \mid x_i, \Theta^{(t)}\right] = \frac{P[x_i \mid \varepsilon_{ik} = 1] P[\varepsilon_{ik} = 1]}{\sum_{j=1}^{c} P[x_i \mid \varepsilon_{ij} = 1] P[\varepsilon_{ij} = 1]} = \\ &= \frac{f_k\left(\theta_k^{(t)}; x_i\right) p_k^{(t)}}{f\left(\theta^{(t)}; x_i\right)} = P[k \mid x_i]^{(t)}. \end{aligned} \tag{12}$$

This is the probability that a given x_i comes from the density f_k. After substituting (12) in (11) the following is obtained:

$$\begin{aligned} Q\left(\theta \mid \theta^{(t)}\right) &= \sum_{k=1}^{c}\sum_{i=1}^{n} \ln[f_k(\theta_k; x_i)] \frac{f_k\left(\theta^{(t)}; x_i\right) p_k^{(t)}}{f\left(\theta^{(t)}; x_i\right)} \\ &= \sum_{k=1}^{c}\sum_{i=1}^{n} \ln[f_k(\theta_k; x_i)] P[k \mid x_i]^{(t)}. \end{aligned} \tag{13}$$

Now, for the M-step, $Q\left(\theta \mid \theta^{(t)}\right)$has to be maximized with respect to Θ, which is equivalent to finding the solution of the following equations that arise after differentiating with respect to each parameter and equating to zero:

$$\begin{aligned} \frac{\partial Q\left(\theta \mid \theta^{(t)}\right)}{\partial \theta_{kj}} &= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_{kj}} \left(\ln[f_k(\theta_k; x_i)]\, P[k \mid x_i]^{(t)}\right) \\ &= \sum_{i=1}^{n} P[k \mid x_i]^{(t)} \frac{\partial}{\partial \theta_{kj}} \left(\ln[f_k(\theta_k; x_i)]\right) = 0 \end{aligned} \tag{14}$$

This is done for all components of the mixture ($k = 1, ..., c$) as well as for each of the parameters $[j = 1, ..., \dim(\Theta_k)]$ belonging to each component.

Note that these maximum likelihood equations for estimating the parameters Θ are weighted averages (where $P[k \mid x_i]^{(t)}$are the weights) of the usual maximum likelihood equations for each component of the mixture (see Everitt and Hand, 1981).

In order to iterate, rename the solution of equations (14) as $\theta^{(t+1)}$, replace and calculate again (13) and solve again equations (14), etc. This iterative procedure will take place until a stopping criterion is satisfied. To complete an iteration, an estimation of $p_k,\ k = 1, ..., c$, must be performed since the following posterior probabilities must be obtained in each step,

$$P\left[k \mid x_i\right]^{(t)} = \frac{f_k\left(\theta^{(t)}; x_i\right) p_k}{f\left(\theta^{(t)}; x_i\right)}.$$

To estimate $p_k,\ k = 1, ..., c$ one can maximize the observed loglikelihood which is obtained from expression (10) and proceed in the usual way. In this case a maximization of the "observed" log likelihood function subject to the restriction that $\sum p_k = 1$, has to be performed:

$$l_{obs} \ln L\left(X, \theta\right) - \lambda\left(\sum_{k=1}^{c} p_k - 1\right),$$

where λ is a Lagrange multiplier.

Differentiating with respect to p_k yields

$$\begin{aligned} \frac{\partial l_{obs}}{\partial p_k} &= \sum_{i=1}^{n} \frac{\partial}{\partial p_k}\left(\ln \sum_{k=1}^{c} p_k f_k\left(x_i; \theta_k\right)\right) - \lambda = \\ &= \sum_{i=1}^{n} \frac{f_k\left(x_i; \theta_k\right)}{f\left(x_i; \theta\right)} - \lambda = 0 \end{aligned} \tag{15}$$

If both sides of this equation are multiplied by p_k and one adds up over all k, the value for the Lagrange multiplier λ can be obtained:

$$\sum_{k=1}^{c} p_k \sum_{i=1}^{n} \frac{f_k\left(x_i; \theta_k\right)}{f\left(x_i; \theta\right)} = \sum_{i=1}^{n} \lambda p_k = \lambda \sum_{k=1}^{c} p_k = \lambda.$$

The left side of this equation may be simplified to

$$\sum_{k=1}^{c} \sum_{i=1}^{n} p_k \frac{f_k\left(x_i; \theta_k\right)}{f\left(x_i; \theta\right)} = \sum_{i=1}^{n} \frac{\sum_{k=1}^{c} p_k f_k\left(x_i; \theta_k\right)}{f\left(x_i; \theta\right)} = \sum_{i=1}^{n} \frac{f\left(x_i; \theta\right)}{f\left(x_i; \theta\right)} = n.$$

So therefore, $\lambda = n$.

Substituting λ in (15) and multiplying by p_k:

$$\sum_{i=1}^{n} \frac{p_k f_k\left(x_i; \theta_k\right)}{f\left(x_i; \theta\right)} = n p_k. \tag{16}$$

Using Bayes Theorem as in (12), thus noting that

$$p_k = P\left[\varepsilon_{ik} = 1\right], \quad \text{and} \quad f_k\left(x_i; \theta_k\right) = P\left[x_i \mid \varepsilon_{ik} = 1\right],$$

substituting the elements of the above sum (16) yields

$$\sum_{i=1}^{n} P\left[k \mid x_i\right] = n p_k.$$

Therefore a natural estimator for p_k in the iteration $(t + 1)$ is

$$p_k^{(t+1)} = \frac{\sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)}}{n}.$$

It has been shown, so far, that if one obtains by any reasonable method initial estimates for the parameters, $\theta_k^{(t)}$ and of the mixing proportions $p_k^{(t)}, k = 1, ..., c$ (the superindex indicates the number of the iteration), the EM algorithm may be reduced to the following two steps:

- **E-step.** Calculate (starting with $t = 1$)

$$P\left[k \mid x_i\right]^{(t)} = \frac{f_k\left(\theta_k^{(t)}; x_i\right) p_k^{(t)}}{f\left(\theta^{(t)}; x_i\right)}.$$

  These posterior probabilities provide useful information that may help to identify the distribution component that generated the i-th observation. This is a very useful and informative by-product of the EM algorithm. There are some clustering techniques which use this information for allocation of the observations to different clusters. McLachlan and Basford (1988) discuss these ideas in great detail.

- **M-step.** Use $P\left[k \mid x_i\right]^{(t)}$ and solve equations (14) updating the parameters $\theta^{(t+1)}$ (or numerically maximize $Q\left(\theta \mid \theta^{(t)}\right)$), and estimate $p_k^{(t+1)}$ as

$$p_k^{(t+1)} = \frac{\sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)}}{n}.$$

One must return to the E-step and iterate once again until a certain stopping criterion is fulfilled. The idea is to stop iterating when the updated parameters are not changing significantly with respect to the ones calculated in the previous step.

In order to clarify the application of the EM algorithm to the case of finite mixtures of distributions, two examples will be given in the following final sections.

4.1 Example 2. Finite Mixtures of Normal Densities

In this case, there is a location parameter μ_k, and a scale parameter σ_k to be estimated for each of the c densities f_k in the mixture. Additionally there are c parameters p_k to be estimated that indicate the amount of participation in the mixture of each density. The k-th density is given by

$$f_k\left(x_i; \mu_k; \sigma_k^2\right) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{-\left(x_i - \mu_k\right)^2}{2\sigma_k^2}\right).$$

The logarithm of this density is:

$$\ln[f_k\left(x_i; \sigma_k^2\right)] = -\ln\sqrt{2\pi} - \frac{1}{2}\ln\sigma_k^2 - \frac{\left(x_i - \mu_k\right)^2}{2\sigma_k^2}.$$

Replacing in (14), and equating to zero, the following expression is obtained,

$$\frac{\partial Q}{\partial_{\mu_k}} = \sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)} \frac{\partial}{\partial_{\mu_k}}\left(\ln[f_k\left(x_i; \mu_k, \sigma_k^2\right)]\right) = \sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)} \frac{x_i - \mu_k}{\sigma_k^2} = 0.$$

Solving this equation for μ_k and renaming it $\mu_k^{(t+1)}$,

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)} x_i}{\sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)}}, \tag{17}$$

which is clearly a weighted average of the observations, and the weight depends on the probability of each observation belonging to the k-th density.

Now, if one uses again (14) and equates to zero,

$$\frac{\partial Q}{\partial \sigma_k^2} = \sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)} \left(\frac{-1}{2\sigma_k^2} + \frac{\left(x_i - \mu_k\right)^2}{2\left(\sigma_k^2\right)^2}\right) = 0.$$

Thus,

$$\left(\sigma_k^2\right)^{(t+1)} = \frac{\sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)} \left(x_i - \mu_k\right)^2}{\sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)}}. \tag{18}$$

In this specific case, the EM algorithm may be reduced to the following two steps (initial estimates for the parameters must be obtained by any reasonable way).

- E-step. Calculate

$$P\left[k \mid x_i\right]^{(t)} = \frac{f_k\left(\theta_k^{(t)}; x_i\right) p_k^{(t)}}{f\left(\theta^{(t)}; x_i\right)}.$$

- M-step. Use $P\left[k \mid x_i\right]^{(t)}$ in (17) and (18) to update the parameters. That is, calculate $\mu_k^{(t+1)}$, $\left(\sigma_k^2\right)^{(t+1)}$, and $p_k^{(t+1)}$,

$$p_k^{(t+1)} = \frac{\sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)}}{n}.$$

With these updated parameters, repeat the E-step, and then the M-step, iteratively until certain stopping criterion is fulfilled.

Note that this example can be easily extended to the estimation of mixtures of multivariate normal mixtures.

4.2 Example 3. Finite Mixtures of Gumbel Densities

A finite mixture of standard Gumbel densities will be considered. That is, there is one location parameter φ_k to be estimated for each of the c densities f_k in the mixture. All of the c scale parameters are equal to unity. Additionally there are c parameters p_k to be estimated and $\sum_{k=1}^{c} p_k = 1$ holds. The density of the k-th component is:

$$f_k\left(x_i; \varphi_k\right) = \exp\left[\left(x_i - \varphi_k\right) - \exp\left(x_i - \varphi_k\right)\right].$$

The logarithm of this density is:

$$\ln[f_k\left(x_i; \varphi_k\right)] = \left(x_i - \varphi_k\right) - \exp\left(x_i - \varphi_k\right).$$

Replacing in (14) and equating to zero,

$$\frac{\partial Q}{\partial \varphi_k} = \sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)} \frac{\partial}{\partial \varphi_k}\left(\ln f_k\left(x_i; \varphi_k\right)\right) = \sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)} \left(\exp\left(x_{i-}\varphi_k\right) - 1\right) = 0.$$

The solution is thus,

$$\varphi_k^{(t+1)} = \ln\left(\frac{\sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)} \exp\left(x_i\right)}{\sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)}}\right). \tag{19}$$

Once again, the EM algorithm may be reduced to the following two steps.

- E-step. Calculate

$$P\left[k \mid x_i\right]^{(t)} = \frac{f_k\left(\varphi_k^{(t)}; x_i\right) p_k^{(t)}}{f\left(\varphi^{(t)}; x_i\right)},\ k = 1, ..., c.$$

- M-step. Use $P\left[k \mid x_i\right]^{(t)}$ in (19) to update the parameters. tah is, calculate $\varphi_k^{(t+1)}$, as well as $p_k^{(t+1)}$ by

$$p_k^{(t+1)} = \frac{\sum_{i=1}^{n} P\left[k \mid x_i\right]^{(t)}}{n},\ k = 1, ..., c.$$

With these updated parameters, repeat E-step, and then M-step, iteratively until a certain stopping criterion is fulfilled.

References

[1] Cox, D. and Oakes D. (1984). *The Analysis of Survival Data.* London: Chapman and Hall.

[2] Dempster, A., Laird, N., and Rubin D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *JRSS, Series B*, V.39, pp. $1 - 38$.

[3] Everitt, B. and Hand, D. (1981). *Finite Mixture Distributions.* London: Chapman and Hall.

[4] Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

[5] McLachlan G. J. and Basford, K. E. (1988). *Mixture Models. Inference and Applications to Clustering.* New York: Marcel Dekker, Inc.

[6] Tittering D. M., Smith, A. F. M., and Makov, U. E. (1983). *Statistical Analysis of Finite Mixture Distributions.* Chichester: John Wiley & Sons.