David R. Hunter, Prabhani Kuruppumullage Don, and Bruce G. Lindsay

An Expansive View of EM Algorithms

Contents

0.1 Introduction

In their most basic form, EM algorithms provide a general method used to find local maxima of a likelihood function when we may conceive of some subset of the full data as having been unobserved in the experiment. These algorithms are the computational workhorses of maximum likelihood estimation in mixture models, which provide perhaps the best-known class of examples in which EM algorithms are effective. In adopting the "expansive" view in this chapter, we explain the EM mechanism without reference specifically to mixture models or even maximum likelihood. Indeed, we argue for a definition of EM that includes any algorithm based on the same basic principle that guarantees the algorithms' success in their original framework. Then, returning to the topic of finite mixtures, we describe several algorithms we consider instances of EM that extend this framework, each one of them related to mixture models. In this way, we hope to both elucidate the workings of EM and demonstrate the broad applicability of these ideas, which we feel is a testament to the genius and simplicity of the EM scheme.

The name "EM algorithm" can trace its origin to the seminal paper of Dempster et al. (1977). Yet as with so many episodes in science, the true story of the algorithm's origins are somewhat murkier. For one thing, "the EM algorithm" is not an algorithm at all: As Dempster et al. (1977) acknowledge in a footnote, "... our use of the term 'algorithm' can be criticized because we do not specify the sequence of computing steps actually required to carry out a single E- or M-step." Indeed, the iteratively applied E- and M-steps of Dempster et al. (1977) are best viewed as a recipe for creating algorithms, and for this reason we refer throughout this article not to "the EM algorithm" but rather to "EM algorithms" more broadly. For another thing, there are numerous examples of EM algorithms in the literature that predate the Dempster et al. (1977) article. Any attempt to catalog them here would be incomplete, so we instead refer interested readers to the book-length treatment of EM

by McLachlan and Krishnan (2007). Yet it was Dempster et al. (1977) who first articulated the general framework for creating EM algorithms, an ingenious innovation that we feel deservedly places their paper among the most highly-cited statistical articles in history.

While many treatments of EM algorithms begin with the missing-data framework and concepts such as "complete-data log likelihood" and "observed-data log likelihood," we take a different approach here. We survey the EM landscape from high altitude, beginning with a development of the essential *product-of-sums* form taken by many difficult-to-maximize likelihoods for which EM algorithms provide a helpful tool. We eventually return to earth in Section 0.3 and explain how the more general development relates to the specific maximum-likelihood-with-missing-data paradigm. We conclude with several illustrative examples as well as some thoughts on convergence criteria.

0.2 The Product-of-Sums Formulation

We start with a maximization problem involving a real-valued objective function $L(\theta)$ of the possibly vector-valued parameter θ. Typically, a global maximum is desired; but here, we will generally mean a local maximum, both for theoretical reasons (e.g., likelihood theory typically guarantees only that some local maximum is a consistent estimator) and for practical ones (most optimization methods, including those described here, attain at best a local maximum).

An EM algorithm is a specialized algorithm, in that it can only be used on objective functions with a particular mathematical structure one might call *product-of-sums*: To wit, let us suppose that $L(\theta)$ has the following product representation:

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta) \tag{1}$$

Note that while we have used a single index i in the product, it can be thought of as shorthand for a set of indices, say (i, j, k), over which products are taken. In addition, we assume that each $L_i(\theta)$ has the summation representation

$$L_i(\theta) = \sum_{g=1}^{G_i} L_{ig}(\theta) \quad \text{or} \quad L_i(\theta) = \int L_{ig}(\theta)\, dg, \tag{2}$$

where $L_{ig}(\theta)$ is a nonnegative function of θ for all i and g. Since we may take the logarithm of the $L(\theta)$ function without changing the θ values that optimize it, it will be convenient to define

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log L_i(\theta) = \sum_{i=1}^{n} \log \left(\sum_{g=1}^{G_i} L_{ig}(\theta) \right). \tag{3}$$

We might call the form of $\ell(\theta)$ in equation (3) *sum-of-log-of-sums*. We allow the possibility that $\ell(\theta)$ takes on the value of $-\infty$ when $L(\theta) = 0$, although that issue generally has no bearing on an EM algorithm as long as one initiates the algorithm using some $\theta^{\langle 0 \rangle}$ satisfying $L(\theta^{\langle 0 \rangle}) > 0$.

0.2.1 Iterative algorithms and the ascent property

To start an EM algorithm requires an initial parameter value, which we denote $\theta^{\langle 0 \rangle}$. Starting with $s = 0$, the algorithm applies a rule that maps $\theta^{\langle s \rangle}$ to $\theta^{\langle s+1 \rangle}$, then increases s by 1,

and then repeats the process until some stopping criterion is achieved (see Section 0.5 for some thoughts about stopping). The key to any EM algorithm is the fact that, in the product-of-sums case, it is possible to create the rule mapping $\theta^{\langle s \rangle}$ to $\theta^{\langle s+1 \rangle}$ in such a way that

$$L(\theta^{\langle s+1 \rangle}) \geq L(\theta^{\langle s \rangle}) \tag{4}$$

is guaranteed. Inequality (4) is called the ascent property of an EM algorithm.

The EM approach is specific to the product-of-sums situation and involves the creation and maximization, for each iteration s, of a surrogate function of θ that depends on $\theta^{\langle s \rangle}$. We denote this surrogate function by $Q(\theta \mid \theta^{\langle s \rangle})$. One can think of $Q(\theta \mid \theta^{\langle s \rangle})$ as a local approximation to $L(\theta)$ in a neighborhood of $\theta^{\langle s \rangle}$, where this approximation has an important property called *minorization* that we describe in the next section.

0.2.2 Creating a Minorizing Surrogate Function

Referring once again to equation (3) defining $\ell(\theta)$, let us ignore the summation over i for the present and focus on the expression

$$\log L_i(\theta) = \log \left(\sum_{g=1}^{G_i} L_{ig}(\theta) \right)$$

for a particular i. As a special case of Jensen's inequality, if we define

$$w_{ig}^{\langle s \rangle} = \frac{L_{ig}(\theta^{\langle s \rangle})}{\sum_{h=1}^{G_i} L_{ih}(\theta^{\langle s \rangle})} \tag{5}$$

for $g = 1, \ldots, G_i$, then

$$\log L_i(\theta) - \log L_i(\theta^{\langle s \rangle}) \geq \sum_{g=1}^{G_i} w_{ig}^{\langle s \rangle} \log \left[\frac{L_{ig}(\theta)}{L_{ig}(\theta^{\langle s \rangle})} \right]. \tag{6}$$

Inequality (6), which is the key to constructing the surrogate function for an EM algorithm, may also be derived directly from the concavity of the logarithm function since, by definition, any concave function $\varphi(\cdot)$ must satisfy

$$\varphi \left(\sum_{g=1}^{G_i} w_{ig}^{\langle s \rangle} \left[\frac{L_{ig}(\theta)}{L_{ig}(\theta^{\langle s \rangle})} \right] \right) \geq \sum_{g=1}^{G_i} w_{ig}^{\langle s \rangle} \varphi \left[\frac{L_{ig}(\theta)}{L_{ig}(\theta^{\langle s \rangle})} \right] \tag{7}$$

because $w_{i1}^{\langle s \rangle}, \ldots, w_{iG_i}^{\langle s \rangle}$ are nonnegative constants that sum to unity. If we now sum over i, the left side of inequality (6) becomes $\ell(\theta) - \ell(\theta^{\langle s \rangle})$.

Therefore, if we define

$$\begin{aligned} Q(\theta \mid \theta^{\langle s \rangle}) &= \sum_{i=1}^{n} \sum_{g=1}^{G_i} w_{ig}^{\langle s \rangle} \log L_{ig}(\theta) \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G_i} \left[\frac{L_{ig}(\theta^{\langle s \rangle})}{\sum_{h=1}^{G_i} L_{ih}(\theta^{\langle s \rangle})} \right] \log L_{ig}(\theta), \end{aligned} \tag{8}$$

then we may conclude that

$$\ell(\theta) - \ell(\theta^{\langle s \rangle}) \geq Q(\theta \mid \theta^{\langle s \rangle}) - Q(\theta^{\langle s \rangle} \mid \theta^{\langle s \rangle}). \tag{9}$$

In other words, at iteration s, the increase in $\ell(\theta)$ must always be at least as large as the increase in $Q(\theta \mid \theta^{\langle s \rangle})$. In particular, if we find θ that makes the latter increase as large as possible by maximizing with respect to θ, we will guarantee some increase in the value of $\ell(\theta)$ above $\ell(\theta^{\langle s \rangle})$. Thus, definition (8) together with maximization of $Q(\cdot \mid \theta^{\langle s \rangle})$ guarantees the characteristic ascent property (4) of all EM algorithms. We will see in Section 0.3 that creating the $Q(\theta \mid \theta^{\langle s \rangle})$ function of Equation (8) is called the "E-step" of an EM algorithm. The key to the success of the EM paradigm is the fact that $Q(\theta \mid \theta^{\langle s \rangle})$ is generally much easier to maximize directly than the original objective function $\ell(\theta)$ of Equation (3).

Minorization is a key characteristic guaranteed by Inequality (9). A function is said to minorize another function at the point θ_0 if it provides a uniform lower bound and if equality is attained at θ_0. In other words, Inequality (9) implies that the function

$$Q(\theta \mid \theta^{\langle s \rangle}) - Q(\theta^{\langle s \rangle} \mid \theta^{\langle s \rangle}) + \ell(\theta^{\langle s \rangle}) \tag{10}$$

is a minorizer of $\ell(\theta)$ at $\theta^{\langle s \rangle}$. The general theory of minorization-maximization (MM) algorithms (Hunter and Lange, 2004), of which any EM algorithm is a special case, shows that iteratively minorizing at the current parameter value and then maximizing the resulting minimizer will guarantee the ascent property (4).

The same techniques explained above may be applied to the case in which $\ell(\theta)$ is a sum-of-log-of-integrals instead of a sum-of-log-of-sums. In this case, Equation (8) is replaced by

$$Q(\theta \mid \theta^{\langle s \rangle}) = \sum_{i=1}^{n} \int \left[\frac{L_{ig}(\theta^{\langle s \rangle})}{\int L_{ih}(\theta^{\langle s \rangle})\, dh} \right] \log L_{ig}(\theta)\, dg. \tag{11}$$

Technical notes: Earlier, we insisted only that $L_{ig}(\theta)$ be nonnegative, not strictly positive; yet it appears at first glance that zero values may create problems in (6). Fortunately, such problems are easily avoided. First, there is no loss of generality in restricting attention to the subset of Θ consisting of those θ for which $\ell(\theta)$ is finite, and the left side of (9) is always well-defined in this set Θ. Second, we are justified in ignoring any summands on the right side of (6) for which $L_{ig}(\theta^{\langle s \rangle}) = 0$ because the limit of $t \log(1/t)$ as $t \to 0$ is zero. This convention works even when $L_{ig}(\theta)$ is also zero, for omitting such g from both sides of (6) changes nothing. Finally, in case $L_{ig}(\theta)$ is ever zero while $L_{ig}(\theta^{\langle s \rangle})$ is not, inequality (6) remains true because if we allow the right side to take the value $-\infty$.

0.3 Likelihood As a Product of Sums

The product representation over variable i in (1) typically arises in likelihood methods because independent observations $Y_1, Y_2, ..., Y_n$ lead to such a product representation. In the statistics framework, the inner summation representation (2) arises in a more specialized way. When one has a known parametric joint density $f(y, z; \theta)$ for two variables Y and Z, then the marginal density for Y typically has the representation

$$f(y; \theta) = \sum_{z} f(y, z; \theta) \quad \text{or} \quad \int f(y, z; \theta)\, dz,$$

where the choice of sum or integral depends on whether the density function $f(y, z; \theta)$ is a probability mass function or a Lebesgue density on the Z space. Thus, in this context, the variable z plays the role of summation index g.

Therefore, if we are given a hypothetical data set $(Y_1, Z_1), (Y_2, Z_2), ..., (Y_n, Z_n)$, each

pair of which is independent of the other pairs, then the objective (log likelihood) function could be written as

$$\ell(\theta) = \log\left(\prod_{i=1}^{n} f_i(y_i;\theta)\right) \tag{12}$$

$$= \sum_{i=1}^{n} \log\left(\sum_{z} f_i(y_i, z;\theta)\right). \tag{13}$$

Moreover, the density for the ith pair (Y_i, Z_i) is $f_i(y, z)$, but the statistician only observes the data $(Y_1, ..., Y_n)$, and so must use the log likelihood given in (12). For this reason, we sometimes refer to $\ell(\theta)$ as the observed-data log likelihood function; other chapters in this book use the notation $\mathcal{L}(\theta)$ or $\mathcal{L}_O(\theta)$ instead of $\ell(\theta)$. Just as we call $(Y_1, \ldots, Y_n)$ the observed data, we call $(Z_1, \ldots, Z_n)$ the missing data.

Mixture models provide a wide variety of observed- and missing-data examples well suited to EM algorithms. In the finite mixture case, we posit a population composed of a certain number, say G, of distinct subgroups. Let η_g denote the proportion of the population consisting of subgroup g, $1 \leq g \leq G$. We assume that variable(s) Y is measured for each individual in a sample from the population, and that Y is distributed according to a density $f_g(y)$ for individuals in the gth subgroup. If the data $y_1, \ldots, y_n$ are observed without their corresponding subgroup labels $z_1, \ldots, z_n$, the "missing data" paradigm fits perfectly: There is a clear sense in which each observed y_i is merely a function of the complete (y_i, z_i) and each z_i is missing. In this example, we may rewrite equation (13) as

$$\ell(\theta) = \sum_{i=1}^{n} \log\left[\sum_{g=1}^{G} \eta_g f_g(y_i)\right]. \tag{14}$$

Although a growing body of literature considers the model of Equation (14) without specifying the parametric form of $f_g(y)$—see Chauveau et al. (2015) for a survey of some of this work—we often assume a parametric form for the density function, so we may replace $f_g(\cdot)$ by $f(\cdot;\theta_g)$. In this context, the $L_{ig}(\theta)$ function of Equation (3) becomes $\eta_g f(y_i;\theta_g)$. The weights defined in equation (5) are given by

$$w_{ig}^{\langle s\rangle} = \frac{\eta_g^{\langle s\rangle} f(y_i;\theta_g^{\langle s\rangle})}{\sum_{h=1}^{G} \eta_h^{\langle s\rangle} f(y_i;\theta_h^{\langle s\rangle})},$$

and the surrogate function for the mixture model is

$$Q^{\langle s\rangle}(\theta) = \sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}^{\langle s\rangle} \log \eta_g + \sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}^{\langle s\rangle} \log f(y_i;\theta_g). \tag{15}$$

Notice that in (15), each element of $\theta = (\eta_1, \ldots, \eta_G, \theta_1, \ldots, \theta_G)$ appears alone in its own separate sum. This means that $Q^{\langle s\rangle}(\theta)$ is much easier to maximize than $\ell(\theta)$. This is typical of EM algorithms, which often replace a single difficult maximization with a series of much easier ones. The surrogate function $Q^{\langle s\rangle}(\theta)$ is sometimes called the complete-data log likelihood and denoted by $\mathcal{L}_C^{\langle s\rangle}(\theta)$ or simply $\mathcal{L}_C(\theta)$, though it should always be remembered that this complete-data log likelihood function depends on the current iterate $\theta^{\langle s\rangle}$ even when this fact is not reflected explicitly by the notation.

There is no generic maximizer of (15) with respect to θ_g, though often when we know the functional form of $f(\cdot;\theta_g)$ an explicit maximizer can be derived. On other other hand,

the maximizer with respect to the η_g parameters, which must sum to unity, is always the same for any finite mixture model:

$$\eta_g^{\langle s+1\rangle} = \frac{1}{n}\sum_{i=1}^{n} w_{ig}^{\langle s\rangle}. \tag{16}$$

Figure 1 illustrates the $\ell(\theta)$ function and several of its its minorizing functions defined by equation (10) for a simple finite mixture model example: Consider a mixture problem with $G = 2$ in which both subgroup densities are completely known: $g_1(\cdot)$ and $g_2(\cdot)$ are both normal with unit variance and means 0 and 1, respectively. Here, the only unknown parameter is $\eta \equiv \eta_1$ (since η_2 is just $1 - \eta_1$). Starting from $\eta^{\langle 0\rangle} = 0.4$, the figure depicts the first three iterations of an EM algorithm.

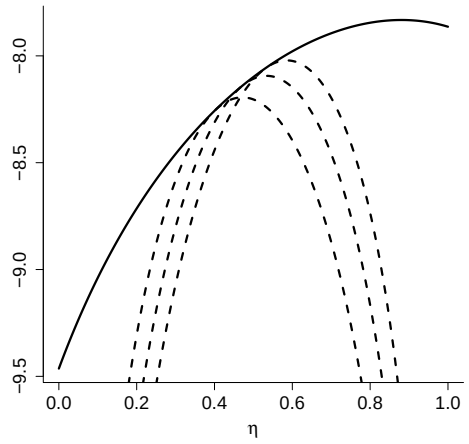

FIGURE 1
The solid curve is the observed-data log likelihood function $\ell(\eta)$ for the two-component mixture model $\eta N(0,1)+(1-\eta)N(1,1)$ with $n = 6$. The data are $-1.0, -0.5, 0.0, 0.5, 0.8, 1.6$. The dotted curves are the functions that minorize $\ell(\eta)$ at the points $\eta^{\langle 0\rangle} = 0.4$, $\eta^{\langle 1\rangle} = 0.472$, and $\eta^{\langle 2\rangle} = 0.534$.

0.4 Non-standard Examples of EM Algorithms

This section presents several examples, each related in some way to mixture models, that illustrate the main theme of this chapter. That is, although an EM algorithm often follows a precise recipe once the observed and missing data are defined, this standard paradigm is not necessary to define an EM algorithm. All that one needs is the product-of-sums, or sum-of-log-of-sums, representation.

0.4.1 Modes of a Density

Li et al. (2007) introduced a Modal EM (MEM) algorithm that finds the local maxima, or modes, of a given density. Consider a 2-component normal mixture density $f(y) =$

$\eta_1 f_1(y) + \eta_2 f_2(y)$, where y is real-valued, η_1 and η_2 are the prior probabilities of the two mixture components, and $f_g(y) = \mathcal{N}(y|\mu_g, \sigma^2)$ is the density of component g. Given an initial value $y^{\langle 0 \rangle}$, MEM solves a local maximum of the mixture density, $f(y)$. We see that

$$\log f(y) = \log \{\eta_1 f_1(y) + \eta_2 f_2(y)\}$$

is in the log-of-sums format. Taking $n = 1$ and following the steps explained in Section 0.2.2, we can obtain the surrogate $Q\left(y|y^{\langle 0 \rangle}\right)$ as

$$Q\left(y|y^{\langle 0 \rangle}\right) = w_1(y^{\langle 0 \rangle}) \log \mathcal{N}(y; \mu_1, \sigma^2) + w_2(y^{\langle 0 \rangle}) \log \mathcal{N}(y; \mu_2, \sigma^2),$$

where

$$w_1(y^{\langle 0 \rangle}) = 1 - w_2(y^{\langle 0 \rangle}) = \frac{\eta_1 \mathcal{N}(y^{\langle 0 \rangle}; \mu_1, \sigma^2)}{\eta_1 \mathcal{N}(y^{\langle 0 \rangle}; \mu_1, \sigma^2) + \eta_2 \mathcal{N}(y^{\langle 0 \rangle}; \mu_2, \sigma^2)}.$$

Solving for y, we have

$$y = w_1(y^{\langle 0 \rangle})\mu_1 + w_2(y^{\langle 0 \rangle})\mu_2.$$

Thus, the MEM algorithm in this case iterates between calculating $w_1(y^{\langle s-1 \rangle})$ and $w_2(y^{\langle s-1 \rangle})$ at the E-step, and calculating $y^{\langle s \rangle}$ at the M-step, until convergence.

0.4.2 Gradient Maxima

The gradient function in its original form (Lindsay, 1995) can be used to test whether a latent distribution, say $\mathcal{H}_0$, is the nonparametric maximum likelihood estimator in an infinite-dimensional space of mixing distributions. In the space of distribution functions, define a path from $\mathcal{H}_0$ to any other distribution $\mathcal{H}_1$ as $\mathcal{H}_\alpha = (1 - \alpha)\mathcal{H}_0 + \alpha\mathcal{H}_1$. Notice that for every α, this construction generates an intermediate distribution. Let $L^*(\alpha) = L(\mathcal{H}_\alpha)$ be the likelihood along the above path. Then the derivative of $\log L^*(\alpha)$ at $\alpha = 0$ is the *directional derivative* corresponding to the path from $\mathcal{H}_0$ to $\mathcal{H}_1$ and it has the form

$$D_{\mathcal{H}_0}(\mathcal{H}_1) = \sum_{i=1}^{d} n(i) \left(\frac{L_i(\mathcal{H}_1)}{L_i(\mathcal{H}_0)} - 1 \right),$$

where d is the number of distinct observations and $n(i)$ is the multiplicity of the ith observation. The *gradient function* is defined as a special case of the directional derivative with degenerate $\mathcal{H}_1$ at ϕ and it has the form

$$D_{\mathcal{H}_0}(\phi) := D_{\mathcal{H}_0}(\Delta_\phi) = \sum_{i=1}^{d} n(i) \left(\frac{L_i(\phi)}{L_i(\mathcal{H}_0)} - 1 \right).$$

Lindsay (1995) also shows that any $\mathcal{H}$ is a maximum likelihood estimator if and only if $D_{\mathcal{H}}(\phi) \geq 0$ for all ϕ.

Let us return again to the finite mixture model example of Equation (14). Equation (15) gives the surrogate function for the EM algorithm and Equation (16) gives the formula for the M-step of the mixing parameter update $\eta_g^{\langle s+1 \rangle}$. In this setting, the gradient function $D(\phi^\star)$ for a single parameter value $\phi^\star$ is

$$D(\phi^\star) = \sum_{i=1}^{n} \left(\frac{f(y_i; \phi^\star)}{\sum_{h=1}^{G} \eta_h f(y_i; \phi_h)} - 1 \right) = -n + \sum_{i=1}^{n} \frac{f(y_i; \phi^\star)}{\sum_{h=1}^{G} \eta_h f(y_i; \phi_h)}$$

As Lindsay (1995) explained, if for some $\phi^\star$, $D(\phi^\star) > 0$, then a higher likelihood can be

obtained, improving the model fit. The $\phi^\star$ that maximize $D(\phi^\star)$ will be the best candidates for improvement, and adding the constant n does not change the maximizer. Therefore, we use an EM algorithm on $D(\phi^\star) + n$. For notational simplicity, let

$$\alpha_i = \frac{1}{\sum_{h=1}^{G} \eta_h f(y_i; \phi_h)}.$$

Then

$$\log[D(\phi^\star) + n] = \log \sum_{i=1}^{n} \alpha_i f(y_i; \phi^\star),$$

which is in the log-of-sums format. The surrogate function for maximizing $D(\phi^\star)$ is

$$Q^{\langle s \rangle}(\phi^\star) = \sum_{i=1}^{n} w_i^{\langle s \rangle} \log f(y_i; \phi^\star),$$

where

$$w_i^{\langle s \rangle} = \frac{\alpha_i f\left(y_i; \phi^{\star\langle s \rangle}\right)}{\sum_{i=1}^{n} \alpha_i f\left(y_i; \phi^{\star\langle s \rangle}\right)}.$$

Once we know the form of $f(y_i; \phi)$, we can maximize $Q^{\langle s \rangle}(\phi^\star)$ for $\phi^\star$ to find candidates, if they exist, that would improve our current solution.

0.4.3 Two-step EM

Govaert and Nadif (2003) first proposed the block mixture model for clustering rows and columns simultaneously. Let $Y = \{y_{ij} : i \in I \text{ and } j \in J\}$ be the data matrix, where I is the set of R row indices and J is the set of C column indices. Suppose that the row labels $A_i = a_i$ are drawn independently from the density $p(a)$ on the set $\{1, 2, \ldots, G_1\}$, and the column labels $B_j = b_j$ are drawn independently from the density $q(b)$ on $\{1, 2, \ldots, G_2\}$. Let $f(y_{ij}; \theta_{a_i, b_j})$ be the density of the ijth observation conditioned on the true labels **a** and **b**. The unconditional density of Y is

$$f(X) = \sum_{a_1}^{G_1} \cdots \sum_{a_R}^{G_1} \sum_{b_1}^{G_2} \cdots \sum_{b_C}^{G_2} \left(\prod_{i=1}^{R} \prod_{j=1}^{C} f(y_{ij}; \theta_{a_i, b_j}) \right) \left(\prod_{i=1}^{R} p(a_i) \right) \left(\prod_{j=1}^{C} q(b_j) \right). \tag{17}$$

It can be shown (Govaert and Nadif, 2003, 2005; Wyse and Friel, 2012) that a standard EM algorithm cannot be directly applied in this setting due to the large number of operations involved. As an alternative to the full likelihood of Equation (17), Kuruppumullage Don (2014) proposed estimating the parameters using the composite likelihood

$$\begin{aligned} \log L_{RC}(\theta) = & \underbrace{\sum_{i=1}^{R} \log \left\{ \sum_{a=1}^{G_1} p_a \left(\prod_{j=1}^{C} \sum_{b=1}^{G_2} q_b f(y_{ij}; \theta_{ab}) \right) \right\}}_{\mathrm{cl}_1(\theta)} \\ & + \underbrace{\sum_{j=1}^{C} \log \left\{ \sum_{b=1}^{G_2} q_b \left(\prod_{i=1}^{R} \sum_{a=1}^{G_1} p_a f(y_{ij}; \theta_{ab}) \right) \right\}}_{\mathrm{cl}_2(\theta)}. \end{aligned} \tag{18}$$

Since the $\mathrm{cl}_1(\theta)$ term in equation (18) is in the form of log-of-sums, we use the surrogate construction explained in Section 0.2.2. Letting

$$L_i(\theta) = \sum_{a=1}^{G_1} p_a \left(\prod_{j=1}^{C} \sum_{b=1}^{G_2} q_b f(y_{ij}; \theta_{ab}) \right),$$

the minorizing surrogate of $\mathrm{cl}_1(\theta) = \sum_{i=1}^{R} \log L_i(\theta)$ is

$$Q_1(\theta|\theta^{\langle s \rangle}) = \sum_{i=1}^{R} \sum_{a=1}^{G_1} w_{ia}^{\langle s \rangle} \log \left\{ p_a \left(\prod_{j=1}^{C} \sum_{b=1}^{G_2} q_b f(y_{ij}; \theta_{ab}) \right) \right\}, \tag{19}$$

where

$$w_{ia}^{\langle s \rangle} = \frac{p_a^{\langle s \rangle} \left(\prod_{j=1}^{C} \sum_{b=1}^{G_2} q_b^{\langle s \rangle} f(y_{ij}; \theta_{ab}^{\langle s \rangle}) \right)}{\sum_{a'=1}^{G_1} p_{a'}^{\langle s \rangle} \left(\prod_{j=1}^{C} \sum_{b=1}^{G_2} q_b^{\langle s \rangle} f(y_{ij}; \theta_{a'b}^{\langle s \rangle}) \right)}.$$

Equation (19) can be rewritten as

$$Q_1(\theta|\theta^{\langle s \rangle}) = \sum_{i=1}^{R} \sum_{a=1}^{G_1} w_{ia}^{\langle s \rangle} \log p_a + \sum_{i=1}^{R} \sum_{a=1}^{G_1} w_{ia}^{\langle s \rangle} \sum_{j=1}^{C} \underbrace{\log \left(\sum_{b=1}^{G_2} q_b f(y_{ij}; \theta_{ab}) \right)}_{\mathrm{cl}_1^*}.$$

We see that although $Q_1(\theta|\theta^{\langle s \rangle})$ can be directly solved for p_a, solving it for q_b and θ_{ab} is still problematic due to the log-of-sums form of cl_1^*. We therefore minorize a second time, following the same construction, and define

$$\begin{aligned} Q_2(\theta|\theta^{\langle s \rangle}) = &\sum_{i=1}^{R} \sum_{a=1}^{G_1} w_{ia}^{\langle s \rangle} \log p_a + \sum_{i=1}^{R} \sum_{a=1}^{G_1} w_{ia}^{\langle s \rangle} \sum_{j=1}^{C} \sum_{b=1}^{G_2} w_{iab}^{\langle s \rangle} \log q_b \\ &+ \sum_{i=1}^{R} \sum_{a=1}^{G_1} w_{ia}^{\langle s \rangle} \sum_{j=1}^{C} \sum_{b=1}^{G_2} w_{iab}^{\langle s \rangle} \log f(y_{ij}; \theta_{ab}), \end{aligned}$$

where

$$w_{iab}^{\langle s \rangle} = \frac{q_b^{\langle s \rangle} f(y_{ij}; \theta_{ab}^{\langle s \rangle})}{\sum_{b'=1}^{G_2} q_{b'}^{\langle s \rangle} f(y_{ij}; \theta_{ab'}^{\langle s \rangle})}.$$

Notice that minorization is a transitive operation: Since $Q_1(\cdot|\theta^{\langle s \rangle})$ minorizes $\mathrm{cl}_1(\cdot)$ at $\theta^{\langle s \rangle}$ and $Q_2(\cdot|\theta^{\langle s \rangle})$ minorizes $Q_1(\cdot|\theta^{\langle s \rangle})$ at $\theta^{\langle s \rangle}$, we conclude that $Q_2(\cdot|\theta^{\langle s \rangle})$ minorizes $\mathrm{cl}_1(\cdot)$ at $\theta^{\langle s \rangle}$. Thus, we could define an EM algorithm by repeatedly creating $Q_2(\theta|\theta^{\langle s \rangle})$ in the E-step, then maximizing it in the M-step, and at each iteration we would guarantee the ascent property in the $\mathrm{cl}_1(\theta)$ function.

However, recall that the composite log likelihood function $\log L_{RC}(\theta)$ actually consists of $\mathrm{cl}_1(\theta) + \mathrm{cl}_2(\theta)$. Therefore, to search for a composite maximum likelihood estimator, we should repeat the steps above to obtain a computationally tractable minorizing function for $\mathrm{cl}_2(\theta)$, then add this minimizer to $Q_2(\theta|\theta^{\langle s \rangle})$. The resulting sum gives a minorizer of $\log L_{RC}(\theta)$ whose maximizer provides the next parameter estimate $\theta^{\langle s+1 \rangle}$.

0.5 Stopping Rules for EM Algorithms

As popular as EM algorithms are for their computational simplicity and guaranteed monotone ascent property, they have the weakness of a linear rate of convergence, which in practice can require many iterations to estimate parameters with reasonable accuracy (Böhning et al., 1994). Commonly used stopping rules to decide whether an algorithm has reached its maximum are based on assessing the relative change of log likelihood or parameter values. That is, the algorithm is stopped if $\left|\ell^{\langle s+1\rangle} - \ell^{\langle s\rangle}\right| \leq \epsilon$ or $\left\|\theta^{\langle s+1\rangle} - \theta^{\langle s+1\rangle}\right\| \leq \epsilon$, where $\ell^{\langle s\rangle} = \ell(\theta^{\langle s\rangle})$ is the log likelihood calculated at the sth iteration and ϵ is a small positive constant.

It is worth noting that such stopping rules capture the idea of lack of progress rather than numerical accuracy. Böhning et al. (1994) developed a technique to predict the value of the log likelihood at the maximum likelihood solution, say $\hat{\ell}$, by using Aitken acceleration on log likelihood estimates. This acceleration device is applicable to any log likelihood sequence with linear convergence. The Aitken-accelerated estimate of $\hat{\ell}$ at the sth iteration is

$$\hat{\ell}^{\langle s\rangle} = \ell^{\langle s-1\rangle} + \frac{1}{1 - c^{\langle s\rangle}}\left(\ell^{\langle s\rangle} - \ell^{\langle s-1\rangle}\right),$$

where

$$c^{\langle s\rangle} = \frac{\left(\ell^{\langle s+1\rangle} - \ell^{\langle s\rangle}\right)}{\left(\ell^{\langle s\rangle} - \ell^{\langle s-1\rangle}\right)}$$

is the estimated linear rate of convergence at the tth iteration. That is, $c^{\langle s\rangle}$ is an estimate of

$$\lim_{s\to\infty} \frac{\hat{\ell} - \ell^{\langle s\rangle}}{\hat{\ell} - \ell^{\langle s-1\rangle}}, \tag{20}$$

which can be shown to exist as a positive constant for most EM algorithms. Indeed, the existence of the positive limit (20) is the defining characteristic of a linear rate of convergence.

Using the convergence properties of sequences $\ell^{\langle s\rangle}$ and $\hat{\ell}^{\langle s\rangle}$, Böhning et al. (1994) developed a useful stopping rule that actually reflects the numerical accuracy of the estimates. This rule is

$$\text{Stop EM if } 0 < \left(\hat{\ell}^{\langle s\rangle} - \ell^{\langle s\rangle}\right) < \epsilon.$$

Lindsay (1995) discussed this stopping rule, noting that that if the tolerance value is small, say, $\epsilon < 0.005$, then we are pursuing a numerical accuracy in the parameter estimates that is minor relative to the magnitude of their likelihood confidence intervals.

0.6 Conclusion

EM algorithms are the main computational workhorses for maximum likelihood calculations in mixture model contexts, and this fact does not change if we expand the notion of what defines an EM algorithm. In this chapter, we have described such an expansion, presenting several examples of algorithms that we consider EM since they are built from functions having the sum-of-log-of-sums form. In this expansive view, the E-step still relies on Jensen's inequality—an inequality based on expectation—to construct a minorizing function to maximize in the M-step; thus, these algorithms are also expectation-maximization

algorithms even though they do not fit the standard pattern first introduced in the seminal paper by Dempster et al. (1977). Each of the examples we consider here is rooted in mixture models, even though none of them uses a standard EM algorithm. Furthermore, each enjoys all of the theoretical properties—both good and bad—of more traditionally defined EM algorithms. In particular, the linear rate of convergence that characterizes EM algorithms may be exploited to engineer a stopping criterion that strives to ensure numerical accuracy in parameter estimates rather than reacting to slow progress of the algorithm.

Bibliography

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46(2):373–388.

Chauveau, D., Hunter, D. R., and Levine, M. (2015). Semi-parametric estimation for conditional independence multivariate finite mixture models. *Statistics Surveys*, 9:1–31.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, pages 1–38.

Govaert, G. and Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473.

Govaert, G. and Nadif, M. (2005). An EM algorithm for the block mixture model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):643–647.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37.

Kuruppumullage Don, P. (2014). *Estimation and Model Selection for Block Clustering with Mixtures: A Composite Likelihood Approach*. PhD thesis, The Pennsylvania State University, USA.

Li, J., Ray, S., and Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8(8):1687–1723.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applicaitons*. Institute of Mathematical Statistics, Hayward, CA.

McLachlan, G. J. and Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons, New York.

Wyse, J. and Friel, N. (2012). Block clustering with collapsed latent block models. *Statistics and Computing*, 22(2):415–428.