Wiley StatsRef:
Statistics Reference Online

EM Algorithm

By *Xiao-Li Meng*

The EM algorithm is an iterative procedure for computing **maximum–likelihood estimates** or **posterior** modes in problems with **incomplete data** or problems that can be formulated as such (e.g., with **latent structures**). The name EM, proposed by Dempster et al. [5], highlights the two steps performed at each iteration. In its general form, the E-step, an abbreviation for the **expectation** (not estimation) step, computes the conditional expectation of the complete-data log likelihood function given the observed data and the parameter estimate from the previous iteration. The M-step, or maximization step, then maximizes this expected log **likelihood** function to determine the next iterate of the parameter estimate. The algorithm iterates between the E-step and M-step until convergence. An easily readable summary of the basic theoretical properties of EM can be found in the entry on the **Missing Information Principle**, which also contains a simple yet informative numerical illustration. The current entry supplements and updates these entries with emphasis on extensions and variations of the EM algorithm.

A mathematical description is as follows. Let $L(\theta|Y_{obs})$ be the log likelihood function (or more generally the log of a posterior density) to be maximized over $\theta \in \Theta$, where Y_{obs} are the observed data. Suppose we have a way of augmenting Y_{obs} to obtain $Y = (Y_{obs}, Y_{mis})$, where we shall call the unobserved part Y_{mis} the *missing data* and Y the *complete data* (more precisely, the augmented data). The log likelihood function given Y will be denoted by $L(\theta|Y)$. Starting from an initial guess $\theta^{(0)} \in \Theta$, the EM algorithm forms a sequence of iterates $\theta^{(t)}$, $t = 0, 1, \ldots$ by repeating the following two steps:

E-Step: Compute the conditional expectation of $L(\theta|Y)$ assuming $\theta = \theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) \equiv \int L(\theta|Y) f(Y_{\text{mis}}|Y_{\text{obs}}, \theta = \theta^{(t)}) d Y_{\text{mis}}; \tag{1}$$

M-Step: Determine $\theta^{(t+1)}$ by maximizing $Q(\theta|\theta^{(t)})$, that is, determine $\theta^{(t+1)}$ such that

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}) \text{ for all} \theta \in \Theta. \tag{2}$$

The simplicity of EM depends on the functional form of $L(\theta|Y) = \log f(Y_{obs}, Y_{mis}|\theta)$. It is often possible in practice to construct Y_{mis}, typically corresponding to real unobserved variables or **latent variables** assumed for the model, such that $f(Y_{obs}, Y_{mis}|\theta)$ is from an exponential family and $L(\theta|Y)$ is easy to maximize as a function of θ. When $f(Y|\theta)$ is from an exponential family, $L(\theta|Y)$ is linear in the complete-data **sufficient**

This article was originally published online in 2006 in Encyclopedia of Statistical Sciences, © John Wiley & Sons, Inc. and republished in Wiley StatsRef: Statistics Reference Online, 2014.

statistics $S(Y)$, and thus the E-step reduces to calculating the expected value of $S(Y)$:

$$S^{(t)}(Y) = \int S(Y) f(Y_{\text{mis}} | Y_{\text{obs}}, \theta = \theta^{(t)}) d Y_{\text{mis}},$$

which often only involves standard manipulation of conditional expectations. The M-step then maximizes $Q(\theta|\theta^{(t)}) = L(\theta|S^{(t)}(Y))$, which can be maximized by the same simple method that maximizes the complete-data log likelihood, $L(\theta|S(Y))$. The fact that one can take advantage of the simplicity of complete-data maximum-likelihood estimation is one of the principal reasons for the popularity of EM in practice.

In fact the EM algorithm was motivated by this simplicity. For many years prior to the publication of ref. 5, practitioners had been interested in the following strategy for handling missing observations. If the values of the missing observations were known, then standard complete-data techniques could be applied to fit the posited model. On the other hand, if the parameters of the posited model were known, then the missing observations could be imputed (or predicted) using the observed data and the model. The EM algorithm provides a corrected formulation of this ad hoc "chicken-and-egg" approach (e.g., Beale [3]), namely, the correct procedure is to impute not the individual missing observations, but rather the complete-data sufficient statistics, or more generally the log likelihood function itself, by the conditional expectation defined in (1).

Because the idea behind EM is so intuitive, the algorithm or mathematical formulations of it had been documented in various forms by a handful of authors before Dempster et al. [5], who traced back the relevant literature to as early as McKendrik's paper [22] in 1926. Among them, Hartley [8], Orchard and Woodbury [32], Sundberg [38], and Baum et al. [2] are particularly significant, as discussed in Meng and van Dyk [31], who noted that it is very difficult, if not impossible, to trace the origin of EM.

The main contribution of ref. 5, besides the general recognition of the E and M steps, is a list of potential applications, many of which had not been previously viewed as incomplete-data problems (e.g., mixture model, variance components, factor analysis). Since the publication of ref. 5 in 1977, EM has seen a remarkable array of applications. A 1992 preliminary bibliographic review by Meng and Pedlow [25] found over 1000 EM-related articles in nearly 300 journals, approximately 85% of which are in fields other than statistics. A citation study by Stigler [37] lists [5] as one of the six "high-count" papers (the other five are Cox's model, the Kaplan–Meier estimator, the Box–Cox transformation, Efron's bootstrap, and Duncan's multiple range tests).

A second principal reason for the popularity of EM in practice is its stability: starting from any initial value inside the parameter space Θ, the sequences of EM iterates $\theta^{(t)}$, $t = 0, 1, \ldots$, will always "climb uphill" along the log likelihood being maximized:

$$L(\theta^{(t+1)}|Y_{\text{obs}}) \geq L(\theta^{(t)}|Y_{\text{obs}}) \quad \text{for all } t = 0, 1, \ldots.$$

This property is shared by any *generalized EM* (GEM), defined in [5] as any algorithm that satisfies a weaker version of (2): $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})$ for all t. In addition, many practical implementations indicate that the "uphill step size" in log likelihood, $L(\theta^{(t+1)}|Y_{obs}) - L(\theta^{(t)}|Y_{obs})$, is particularly large l at the first few iterations (e.g., $t \leq 5$), especially when the initial value is far from the convergent value; an example is given in Meng and van Dyk [31]. These properties suggest that EM often moves the iterates quickly into a neighborhood (defined by the likelihood value) of a local mode of the likelihood, albeit the final convergence to the local mode can be quite slow. In theory, EM can also converge to a saddlepoint (see Wu [49] and Boyles [4]), but this problem can be easily avoided if we run EM several times with several "overdispersed" initial values. Running EM with several starting values is a recommended practice because it helps to detect the problem of multiple modes of the likelihood, or more generally of a posterior density, a problem of great importance for statistical inferences because a point summary can be very misleading when the posterior has several (important) modes. Simple implementation, stable

convergence, and the possibility of straightforward detection of multiple modes make EM an attractive algorithm for statisticians, to whom computation is a tool.

1 Illustration with Multivariate *T** Models

The multivariate (including univariate) t is a common model for statistical analysis, especially for robust estimation (cf. Lange et al. [16] and the ESS entry on **Iteratively Reweighted Least Squares**). It also provides a simple yet informative illustration of EM as well as its various extensions. Let $\boldsymbol{t}_p$ denote a p-dimensional t-variable with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and degrees of freedom ν; the density for t_p then is

$$f(\mathbf{x}|\mu, \Sigma, \nu) \propto |\Sigma|^{-1/2} \times [\nu + (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)]^{-(\nu+p)/2}, \quad \mathbf{x} \in R^p.$$

Let us first consider the case where the number of degrees of freedom, ν, is known. Fitting this model to a data set $Y_{obs} = (y_1, \ldots, y_n)$ amounts to maximizing the likelihood function $L(\boldsymbol{\mu}, \Sigma|Y_{\text{obs}}) = \prod_i f(\mathbf{y}_i|\nu, \boldsymbol{\mu}, \Sigma)$, which is known to have no general closed-form solution. At the first sight, EM seems irrelevant here because there are no "missing data." However, if we consider how the t-distribution is derived from the normal distribution and why it is useful for **robust estimation**, we will see an obvious way of constructing a **data-augmentation** scheme for implementing EM. It is well known that the t-density is derived from the following representation of $\boldsymbol{t}_p$:

$$\mathbf{t}_p = \mu + \Sigma^{1/2}\mathbf{Z}/\sqrt{q}, \tag{3}$$

where $\boldsymbol{Z} \sim N_p(\mathbf{0}, \mathbf{I}_p)$ is a standard p-dimensional normal random variable, $q \sim \chi^2_\nu/\nu$ is a mean chi-squared variable with ν degrees of freedom, and $\boldsymbol{Z}$ and q are independent.

Now suppose the observed data $\boldsymbol{Y}_{obs} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ are n independent realizations of $\boldsymbol{t}_p$. If we also had the values of the corresponding q, namely $Y_{mis} = (q_1, \ldots, q_n)$, then the maximum-likelihood estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ would be the same as that from the normal regression problem, $\mathbf{y}_i|q_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}/q_i)$, independently for $i = 1, \ldots, n$. Thus, given the augmented data, $\boldsymbol{Y} = (\mathbf{y}_i, q_i)$, $i = 1, \ldots, n$, maximizing $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\boldsymbol{Y})$ (i.e., the M-step) has a closed-form solution. The E-step takes care of the problem that $Y_{mis} = (q_1, \ldots, q_n)$ are in fact unknown. Since in this case $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\boldsymbol{Y})$ is linear in $Y_{mis} = (q_1, \ldots, q_n)$, the E-step at the $(t + 1)$st iteration only requires evaluating

$$w_i^{(t+1)} \equiv E(q_i|\mathbf{y}_i, \mu^{(t)}, \Sigma^{(t)}) = \frac{\nu + p}{\nu + d_i^{(t)}}, \quad i = 1, \ldots, n, \tag{4}$$

where $d_i^{(t)} = (\mathbf{y}_i - \mu^{(t)})^\top [\Sigma^{(t)}]^{-1} (\mathbf{y}_i - \mu^{(t)})$. The M-step then finds the maximum-likelihood estimate of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pretending $q_i = w_i^{(t+1)}$:

$$\mu^{(t+1)} = \frac{\sum_i w_i^{(t+1)} \mathbf{y}_i}{\sum_i w_i^{(t+1)}}, \tag{5}$$

$$\Sigma^{(t+1)} = \frac{1}{n}\sum_i w_i^{(t+1)}(\mathbf{y}_i - \boldsymbol{\mu}^{(t+1)}) \times (\mathbf{y}_i - \boldsymbol{\mu}^{(t+1)})^\top. \quad (6)$$

Thus the algorithm is trivial to program. The same idea also applies when q of (3) follows other types of distributions, as considered in Chapter 10 of Little and Rubin [18] and in Lange and Sinsheimer [17]. The resulting algorithms are special cases of **iteratively reweighted least squares**, a term which is easy to understand in view of the iterative forms given in (5) and (6). There are many examples of using the EM algorithm for problems that at first appear unrelated to incomplete data; a particularly interesting example is given by Rubin and Szatrowski [33] for the problem of estimating patterned covariance matrices.

2 Statistically Motivated Extensions and Variations

Extending and improving the EM algorithm has resulted in two kinds of generalizations. Here we use the *t*-model to illustrate the first kind (the second kind will be discussed later)—statistically motivated generalizations, by which we mean generalizations that are motivated by statistical considerations, in terms of both techniques and purposes, and that follow closely the original theme of EM, that is, simplicity and stability. The central theme of these generalizations is to enhance the simplicity of EM as well as to improve its speed of convergence in cases where the conventional implementations can be slow, as compared to other algorithms such as Newton–Raphson (when human time for implementing or monitoring the algorithms is not taken into account).

To illustrate, suppose the number of degrees of freedom, ν, in the *t*-model is also unknown. Even with the augmented data $\boldsymbol{Y} = (\boldsymbol{y}_i, q_i)$, $i = 1, \ldots, n$, the maximum-likelihood estimate of ν is not in closed form and thus requires iterations. To implement the original EM, we need to find the joint maximum-likelihood estimate of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. In this case $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu|\boldsymbol{Y}) = L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{Y}) + L(\nu|Y_{mis})$, and thus we can maximize the two terms separately (e.g., see Liu and Rubin [20]. In general, however, this may not be the case, and numerical procedures may be needed to find the joint maximum-likelihood estimate. In such cases it is sometimes possible to "break" the joint maximization problem into several conditional maximization ones, each of which is an easier computational task (e.g., analytic solutions, maximizing in lower dimension) than that required by the joint maximization. For example, we can first maximize $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu|\boldsymbol{Y})$ over $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with ν fixed at the previous estimate, and then in turn maximize $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu|\boldsymbol{Y})$ over ν with $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ set to the values we just obtained (for this particular problem, these two steps of conditional maximizations are the same as maximizing the joint likelihood). Meng and Rubin [28] gave three examples where such conditional maximizations (CMs) are advantageous because they eliminate the nested iterations required to implement the original M-step; in one of their examples involving **missing data** with a **log-linear model**, they use CM steps that are more sophisticated than simple partitions of the parameter vector. They show that such an *expectation/conditional maximization* (ECM) algorithm not only enhances the computational simplicity of EM in some cases but also maintains its convergence properties (e.g., the monotonic convergence of the likelihood along any ECM sequence).

In terms of the number of iterations, ECM typically is slower than EM. This is expected, although not always true (see the counterexample provided in Meng [24]), but the tradeoff is typically worthwhile, since ECM saves programming time. Nevertheless, Liu and Rubin [19] noticed that sometimes a simple modification can speed up ECM substantially, without sacrificing its simplicity or stability. The *t*-model with unknown ν was the motivating example for Liu and Rubin's algorithm. They noticed that when $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are fixed, maximizing the actual (conditional) log likelihood $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu|\boldsymbol{Y}_{obs})$ over ν and maximizing the complete-data conditional log likelihood $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu|\boldsymbol{Y})$ over ν require a similar one-dimensional optimization routine. Thus they suggest maximizing the actual conditional log likelihood of ν instead of

its (expected) complete-data conditional log likelihood when implementing the CM step that updates ν. In general, depending on the nature of each CM step, they propose to maximize either the actual conditional log likelihood or the (expected) complete-data conditional log likelihood. They term this "hybrid" algorithm the *expectation/conditional maximization either* (ECME) algorithm, and provide theoretical and empirical evidence for its stability and fast convergence. (Meng and van Dyk [31] noted that with ECME, in order to guarantee monotone convergence of the likelihood values, within each iteration those CM steps that act on the (expected) complete-data conditional log likelihood must be performed before those CM steps that act on the actual conditional log likelihood.)

To understand why ECME can substantially reduce the number of iterations of ECM without sacrificing the stability, we need to review the relationship between the rate of convergence of EM and the data-augmentation scheme underlying EM. As shown in ref. 5, under mild regularity conditions, the (matrix) rate of convergence of EM equals the fractions of missing information, that is,

$$\mathbf{DM}^{\mathrm{EM}} = \mathbf{I}_{\mathrm{mis}}\mathbf{I}_{\mathrm{com}}^{-1}. \tag{7}$$

Here, $\boldsymbol{DM}^{EM}$ is the Jacobian matrix of the EM mapping [i.e., the mapping defined by the EM iteration: $\theta^{(t+1)} = \boldsymbol{M}^{EM}(\theta^{(t)}]$ evaluated at the convergent value of $\theta^{(t)}$, $t \geq 0$, and $\boldsymbol{I}_{mis}$ and $\boldsymbol{I}_{com}$ are, respectively, the (expected) missing and complete information, both in terms of the observed **Fisher information** (see refs. 26 and 29 for detail). Intuitively, (7) says that the less we augment the data, the faster the corresponding EM will converge. Thus, if we have two augmentation schemes yielding two EM implementations, we should use the one with less augmentation *if* its computation is no harder than that required by the other. The ECM formulation provides a more flexible framework for data augmentation, because we can use different augmentation schemes for implementing different CM steps. The ECME algorithm takes advantage of such flexibility by allowing no augmentation in some CM steps and thus can speed up the algorithm. Fessler and Hero [7] proposed a more general *space-alternating generalized EM algorithm* (they abbreviate it as SAGE) and used it for **image reconstruction**, a problem that was handled with the EM algorithm originally by Shepp and Vardi [34]; also see Vardi et al. [46], Lange and Carson [15], and Silverman et al. [35]. Summarizing these methods, Meng and van Dyk [31] describe a general framework, the *alternating expectation/conditional maximization* (AECM) algorithm, which allows different amounts of augmentation at the different CM steps, and thus provides more flexibility in constructing faster algorithms without sacrificing the stability or the simplicity of the original EM.

Meng and van Dyk [31] also introduce another way of speeding up EM by using a working parameter to search for efficient data augmentation schemes for constructing EM-type algorithms. The *t*-model was also their motivating example. They noticed that by multiplying both the numerator and denominator in (3) by $|\sum|^{-a/2}$ with a an arbitrary constant, one obtains

$$\mathbf{t}_p = \boldsymbol{\mu} + \frac{|\sum|^{-a/2}\sum^{1/2}\mathbf{Z}}{\sqrt{q(a)}}, \tag{8}$$

where $\mathbf{Z}$ and q are as before, and $q(a) = q|\sum|^{-a}$. Although (8) is mathematically equivalent to (3), it provides a different data-augmentation scheme when $a \neq 0$, because when $q(a)$ is *assumed* known it also contributes to the estimation of $\sum$. By treating $\boldsymbol{Y}(a) = (\boldsymbol{y}_i, q_i(a))$, $i = 1, \ldots, n$ as the augmented data, we can derive the corresponding EM algorithm, whose speed of convergence depends on the choice of a. Meng and van Dyk [31] show that the optimal a, which maximizes the speed of the algorithm, is $a_{opt} = 1/(\nu + p)$. (Here ν is assumed to be known; if it is unknown, we will replace it with the estimate of ν from the previous iteration, and thus provide a good illustration of AECM, since the data-augmentation scheme varies with iteration.) The corresponding optimal algorithm only differs from the original EM, (5)–(6), by a trivial modification that is accomplished by replacing the denominator n in (6) with the sum of the

weights:

$$\sum_{\text{opt}}^{(t+1)} = \frac{\sum_i w_i^{(t+1)} (\mathbf{y}_i - \boldsymbol{\mu}^{(t+1)})(\mathbf{y}_i - \boldsymbol{\mu}^{(t+1)})^{\mathrm{T}}}{\sum_i w_i^{(t+1)}}. \tag{9}$$

This replacement does not change the limit, because $\sum_i w_i^{(t+1)} \to n$ as $t \to \infty$, as proved by Kent et al. [11], who proposed this modification based on a "curious likelihood identity" regarding the t-density. Although this replacement does not require any new computation, the resulting gain in speed of convergence is quite dramatic, especially for small ν and large p. Meng and van Dyk [31] provide numerical examples showing substantial reductions (e.g., by 90%) of the number of iterations required for convergence. They also illustrate that the "working parameter" approach can be applied to other problems (e.g., Poisson models for image reconstruction or [43] random-effects models) for speeding up EM (or ECM) without sacrificing its stability or simplicity.

3 Numerically Motivated Variations and Other Supplements

A second kind of variation is to combine EM with various numerical acceleration techniques, in the hope that the resulting algorithm will converge more rapidly than the original EM. These variations include combining EM with Aitken's acceleration (e.g., Louis [21]), with various Newton-type accelerations (e.g., Lange [13,14]), and with conjugate-gradient acceleration (e.g., Jamshidian and Jennrich [9]). Since these accelerated variations typically require extra programming as well as careful monitoring for convergence (e.g., the monotonic convergence of the likelihood is no longer automatic), they are best suited for numerically (as well as statistically) sophisticated users, in order to maintain reliable results with the added advantage of being computationally faster. Careful monitoring is also needed when implementing the E-step via Monte Carlo simulation or numerical integration (e.g., Wei and Tanner [47], Meng and Schilling [30]), as the simulation or numerical errors from the E-step can destroy the stability of EM.

The main idea underlying EM has also been applied to settings that go beyond the general EM theory developed in ref. 5 and in ref. 49. For example, the use of EM in nonparametric estimation, in the form of Efron's [6] **self-consistent estimators**, which includes the **Kaplan–Meier estimator** [10] and Turnbull's algorithms [40,41] for grouped, censored, and truncated data; see Laird [12] and Vardi and Lee [45]. However, a satisfactory general theory for EM beyond the parametric family is still lacking, and it is unclear whether such a theory exists in the sense of providing constructive insight (e.g., as provided by [5]), not merely an omnibus mathematical summary.

Finally, several supplemental methods and algorithms have been developed for computing the observed Fisher information $\boldsymbol{I}_{obs}$, and thus the asymptotic variance–covariance, associated with the maximum-likelihood estimates (or posterior modes), when implementing EM or its extensions (e.g., ECM). Louis [21] gives an algebraic formula expressing $\boldsymbol{I}_{obs}$ via conditional expectations (at the final E-step) of the complete-data observed Fisher information and the cross product of the complete-data score function. Meilijson [23] discusses an approximation method that is applicable when the observed data can be formulated as an independently and identically distributed sample. Baker [1] develops a method of computing $\boldsymbol{I}_{obs}$ when using EM for **categorical data** by taking advantage of a matrix link between the complete and incomplete data. Expanding upon an idea presented by Smith [36] in his discussion of ref. 5, Meng and Rubin [26] formulate a *supplemented EM* (SEM) algorithm for computing the asymptotic

variance $\mathbf{I}_{\text{obs}}^{-1}$ via

$$\begin{aligned}\mathbf{V}_{\text{obs}} &\equiv \mathbf{I}_{\text{obs}}^{-1} = \mathbf{I}_{\text{com}}^{-1}(\mathbf{I} - \mathbf{DM}^{\text{EM}})^{-1} \\ &\equiv \mathbf{V}_{\text{com}}(\mathbf{I} - \mathbf{DM}^{\text{EM}})^{-1},\end{aligned} \tag{10}$$

which is a consequence of (7) and of the information identity $\boldsymbol{I}_{obs} = \boldsymbol{I}_{com} - \boldsymbol{I}_{mis}$. The usefulness of SEM comes from the facts that the computation for the rate of convergence matrix $\boldsymbol{DM}^{EM}$ only requires the computer code for implementing E and M steps, and that the calculation for $\boldsymbol{I}_{com}$ (or $\mathbf{V}_{\text{com}} = \mathbf{I}_{\text{com}}^{-1}$) is straight-forward whenever the complete-data observed Fisher information is a linear function of the unobserved complete-data sufficient statistics, which is the case when EM is most useful.

As a simple illustration, consider the following example detailed in refs. 5 and 26. Suppose the complete data $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5)$ have a multinomial distribution with cell probabilities

$$\left(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right), \qquad 0 \le \theta \le 1.$$

The observed counts are $\boldsymbol{Y}_{obs} = (Y_1 + Y_2, Y_3, Y_4, Y_5) = (125, 18, 20, 34)$ and we want to compute the maximum-likelihood estimate for θ based on $\boldsymbol{Y}_{obs}$. Notice that if $\mathbf{Y}$ were observed, the MLE of θ would be immediate:

$$\theta^* = \frac{Y_2 + Y_5}{Y_2 + Y_3 + Y_4 + Y_5}. \tag{11}$$

Also note that the log likelihood $L(\theta|\mathbf{Y})$ is linear in $\boldsymbol{Y}$, so in the E-step we only need to replace the missing Y_2 by its conditional expectation. Thus at the $(t+1)$st iteration, we calculate for the E-step

$$Y_2^{(t+1)} = 125\frac{\theta^{(t)}/4}{1/2 + \theta^{(t)}/4}; \tag{12}$$

and for the M-step, following (11),

$$\theta^{(t+1)} = \frac{Y_2^{(t+1)} + 34}{Y_2^{(t+1)} + 72}. \tag{13}$$

Substituting (12) into (13) yields the EM iteration $\theta^{(t+1)} = M^{EM}(\theta^{(t)})$ for this problem. We note that for this problem iteration is unneeded, because we can analytically solve $\theta^* = M^{EM}(\theta^*)$, which is a quadratic equation, to obtain the MLE of θ:

$$\theta^* = (15 + \sqrt{53809})/394 \cong 0.626821498.$$

Table **1** displays the convergence of EM to this solution from the initial value $\theta^{(0)} = 0.5$. The second column in the table gives the corresponding values of $\phi = \arcsin\sqrt{\theta}$, which is the well-known variance-stabilizing transformation for the binomial proportion θ. The third and fourth columns give their corresponding deviations $d_\theta^{(t)} = \theta^{(t)} - \theta^*$ and $d_\phi^{(t)} = \phi^{(t)} - \phi^*$, respectively. The fifth and sixth columns are the corresponding ratios of successive deviations. The ratios are essentially constant for $t \ge 3$, which implies that the rate of convergence for EM is $DM^{EM} = 0.1328$. This rate of convergence is invariant under any one-to-one differentiable transformation of the parameter, and it is recommended in ref. 26 to use transformations (e.g., a variance-stabilizing transformation) that can improve both the accuracy of large-sample normal approximation and the numerical stability of the SEM implementation.

Since the complete-data density is

$$f(\mathbf{Y}|\theta) \propto \theta^{Y_2+Y_5}(1-\theta)^{Y_3+Y_4},$$

Table 1. The EM Iterations for the Multinomial Example

t	$\theta^{(t)}$	$\phi^{(t)}$	$d_\theta^{(t)}$	$d_\phi^{(t)}$	$d_\theta^{(t+1)}/d_\theta^{(t)}$	$d_\phi^{(t+1)}/d_\phi^{(t)}$
0	0.50000000	0.78539816	−0.12682150	−0.12822228	0.1465	0.1490
1	0.60824742	0.89450953	−0.01857408	−0.01911092	0.1346	0.1352
2	0.62432105	0.91103720	−0.00250045	−0.00258324	0.1330	0.1331
3	0.62648888	0.91327661	−0.00033262	−0.00034383	0.1328	0.1328
4	0.62677732	0.91357478	−0.00004418	−0.00004567	0.1328	0.1328
5	0.62681563	0.91361438	−0.00000587	−0.00000606	0.1328	0.1328
6	0.62682072	0.91361964	−0.00000078	−0.00000081	0.1328	0.1328
7	0.62682140	0.91362034	−0.00000010	−0.00000011	·	·
8	0.62682149	0.91362043	−0.00000001	−0.00000001	·	·
9	0.62682150	0.91362044	−0.00000000	−0.00000000	·	·

the complete-data variance for $\theta - \theta^*$ is simply the ordinary binomial variance $\theta(1-\theta)/n$, where $n = Y_2 + Y_3 + Y_4 + Y_5$. Thus, from (10), the asymptotic variance of $\theta - \theta^*$ based on the observed counts $\mathbf{Y}_{obs}$ is

$$\frac{\theta^*(1-\theta^*)}{n^*} \times \frac{1}{1 - DM^{\mathrm{EM}}} = \frac{0.0023}{1 - 0.1328} = 0.0026,$$

where $n^* = Y_2^* + 72 = 101.83$ is the conditional expectation of n given $\mathbf{Y}_{obs}$ and $\theta = \theta^*$. Similarly, since the complete-data variance of $\phi - \phi^*$ is $1/(4n)$, the asymptotic variance of $\phi - \phi^*$ based on $\mathbf{Y}_{obs}$ is

$$\frac{1}{4n^*(1 - DM^{\mathrm{EM}})} = 0.0028.$$

Further work by van Dyk et al. [44] illustrate how to implement SEM when using ECM, based on the relationship between the rates of convergence of EM and of ECM established in ref. 24. Another interesting use of the rate of the convergence of EM is for estimating the number of components with mixture models, as proposed by Windham and Cutler [48] and further explored in van Dyk [42].

4 A Concluding Note on Misuse

The EM algorithm has been one of the most popular computational methods in applied statistics. The central idea underlying it, namely constructing convenient statistical algorithms via data augmentation, has also helped in formulating other widely applicable computational methods, most notably Tanner and Wong's [39] algorithm for simulating a posterior distribution. Like many popular methods in statistics, however, it has also been misused or misperceived. For example, one misperception is to regard EM as an estimation procedure (though the idea underlying EM can be useful for constructing estimators in settings beyond parametric estimations) and to study the properties of "EM estimators," forgetting that it is simply a computational method intended for computing maximum-likelihood estimates, or more generally posterior modes (e.g., see the rejoinder of Meng and Rubin in ref. 27 for examples of such misperception). There is also sometimes a problem of overuse. For example, if in an application of an EM-type algorithm one finds that the E-step needs numerical implementation and there is not much simplification in the resulting M or CM steps, then it is wise at least to consider alternative methods.

These problems can be avoided if we keep in mind that computational methods are only a tool for **statistical inferences**, and it is typically more effective in practice to use computational methods that

deliver reliable results with as little human effort as possible. The key idea underlying EM, enhanced by the very general AECM formulation, provides an encompassing approach for constructing algorithms that are ideal for statisticians. In particular, as demonstrated in refs. [7, 19], and 31, it allows us to construct algorithms that are *simple, stable,* and *fast,* the three features that are appreciated by every user.

5 Related Articles

See also **Missing Data, Types of**; **Turnbull estimator**; **Optimization in Statistics**; **Statistics and Computing**; **Computer-Intensive Statistical Methods**.

References

[1] Baker, S. G. (**1992**). A simple method for computing the observed information matrix when using the EM algorithm with categorical data. *J. Comput. Graph. Statist.*, *1*, 63–76. Correction, *1*, 180.

[2] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (**1970**). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, *41*, 164–171.

[3] Beale, E. M. L. (**1977**). Discussion of the paper by Dempster, Laird, and Rubin. *J. R. Statist. Soc. B*, *39*, 22–24.

[4] Boyles, R. A. (**1983**). On the convergence of the EM algorithm. *J. R. Statist. Soc. B*, *45*, 47–50.

[5] Dempster, A. P., Laird, N. M., and Rubin, D. B. (**1977**). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, *39*, 1–38.

[6] Efron, B. (**1967**). The two sample problem with censored data. *Proc. Fifth Berkeley Symp. Math. Statist. Probab., Vol. 4*, pp. 831–853.

[7] Fessler, J. A. and Hero, A. O. (**1994**). Space alternating generalized EM algorithm. *IEEE Trans. Signal Process.*, *42*, 2664–2677.

[8] Hartley, H. O. (**1958**). Maximum likelihood estimation from incomplete data. *Biometrics*, *14*, 174–194.

[9] Jamshidian, M. and Jennrich, R. I. (**1993**). Conjugate gradient acceleration of the EM algorithm. *J. Amer. Statist. Ass.*, *88*, 221–228.

[10] Kaplan, E. L. and Meier, P. (**1958**). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Ass.*, *53*, 457–481.

[11] Kent, J. T., Tyler, D. E., and Vardi, Y. (**1994**). A curious likelihood identity for the multivariate *t*-distribution. *Commun. Statist. Simulation*, *23*, 441–453.

[12] Laird, N. (**1978**). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Ass.*, *73*, 805–815.

[13] Lange, K. (**1995**). A gradient algorithm locally equivalent to the EM algorithm. *J. R. Statist. Soc. B*, *56*, 425–437.

[14] Lange, K. (**1995**). A quasi-Newton acceleration of the EM algorithm. *Statist. Sinica*, *5*, 1–18.

[15] Lange, K. and Carson, R. (**1984**). EM reconstruction algorithms for emission and transmission tomography. *J. Comput. Assisted Tomography*, *8*, 302–316.

[16] Lange, K., Little, R. J. A., and Taylor, J. M. G. (**1989**). Robust statistical modeling using the *t* distribution. *J. Amer. Statist. Ass.*, *84*, 881–896.

[17] Lange, K. and Sinsheimer, J. S. (**1993**). Normal/independent distributions and their applications in robust regression. *J. Comput. Graph. Statist.*, *2*, 175–198.

[18] Little, R. J. A. and Rubin, D. B. (**1987**). *Statistical Analysis with Missing Data*. Wiley, New York.

[19] Liu, C. and Rubin, D. B. (**1994**). The ECME algorithm: a simple extension of EM and ECM with fast monotone convergence. *Biometrika*, *81*, 633–648.

[20] Liu, C. and Rubin, D. B. (**1995**). ML estimation of the *t* distribution using EM and its extensions, ECM and ECME. *Statist. Sinica*, *5*, 19–40.

[21] Louis, T. A. (**1982**). Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. B*, *44*, 226–233.

[22] McKendrik, A. G. (**1926**). Applications of mathematics to medical problems. *Proc. Edinburgh Math. Soc.*, *44*, 98–130.

[23] Meilijson, I. (**1989**). A fast improvement to the EM algorithm on its own terms. *J. R. Statist. Soc. B*, *51*, 127–138.

[24] Meng, X. L. (**1994**). On the rate of convergence of the ECM algorithm. *Ann. Statist.*, *22*, 326–339.

[25] Meng, X. L. and Pedlow, S. (**1992**). EM: A bibliographic review with missing articles. *Proc. Statist. Comput. Sect*. American Statistical Association, Washington, pp. 24–27.

[26] Meng, X. L. and Rubin, D. B. (**1991**). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Amer. Statist. Ass.*, *86*, 899–909.

[27] Meng, X. L. and Rubin, D. B. (**1992**). Recent extensions to the EM algorithm (with discussion). In *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds. Oxford University Press, pp. 307–320.

[28] Meng, X. L. and Rubin, D. B. (**1993**). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, *80*, 267–278.

[29] Meng, X. L. and Rubin, D. B. (**1994**). On the global and componentwise rates of convergence of the EM algorithm. *Linear Algebra Appl.*, *199*, 413–425.

[30] Meng, X. L. and Schilling, S. (**1996**). Fitting full-information item factor models and an empirical investigation of bridge sampling. *J. Amer. Statist. Ass.*, *91*, 1254–1267.

[31] Meng, X. L. and van Dyk, D. (**1997**). The EM algorithm (with discussion). *J. R. Statist. Soc. B*, *59*, to appear.

[32] Orchard, T. and Woodbury, M. A. (**1972**). A missing information principle: theory and application. *Proc. 6th Berkeley Symp. Math. Statist. Probab., Vol. 1*, pp. 697–715.

[33] Rubin, D. B. and Szatrowski, T. H. (**1982**). Finding maximum likelihood estimates of patterned covariance matrices by the EM algorithm. *Biometrika*, *69*, 657–660.

[34] Shepp, L. A. and Vardi, Y. (**1982**). Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Med. Imaging*, *1*, 113–121.

[35] Silverman, B. W., Jones, M. C., Wilson, J. D., and Nychka, D. W. (**1990**). A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion). *J. R. Statist. Soc. B*, *52*, 271–324.

[36] Smith, C. A. B. (**1977**). Discussion of the paper by Dempster, Laird, and Rubin. *J. R. Statist. Soc. B*, *39*, 24–25.

[37] Stigler, S. (**1994**). Citation patterns in the journals of statistics and probability. *Statist. Sci.*, *9*, 94–108.

[38] Sundberg, R. (**1976**). An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Comm. Statist. Simulation Comput. B*, *5*(1), 55–64.

[39] Tanner, M. A. and Wong, W. H. (**1987**). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Ass.*, *82*, 528–550.

[40] Turnbull, B. W. (**1974**). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Ass.*, *69*, 169–173.

[41] Turnbull, B. W. (**1976**). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Statist. Soc. B*, *38*, 290–295.

[42] van Dyk, D. A. (**1995**). Construction, Implementation, and Theory of Algorithms Based on Data Augmentation and Model Reduction. Ph.D. thesis, Department of Statistics, University of Chicago.

[43] van Dyk, D. A. and Meng, X. L. (**1997**). On the orderings and groupings of conditional maximizations within ECM-type algorithms. *J. Comput. Graph. Statist.*, to appear.

[44] van Dyk, D. A., Meng, X. L., and Rubin, D. B. (**1995**). Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance. *Statist. Sinica*, *5*, 55–76.

[45] Vardi, Y. and Lee, D. (**1993**). From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems (with discussion). *J. R. Statist. Soc. B*, *55*, 569–612.

[46] Vardi, Y., Shepp, L. A., and Kaufman, L. (**1985**). A statistical model for positron emission tomography. *J. Amer. Statist. Ass.*, *80*, 8–37.

[47] Wei, C. G. and Tanner, M. A. (**1990**). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Ass.*, *85*, 699–704.

[48] Windham, M. P. and Cutler, A. (**1992**). Information ratios for validating mixture analyses. *J. Amer. Statist. Ass.*, *87*, 1188–1192.

[49] Wu, C. F. J. (**1983**). On the convergence properties of the EM algorithm. *Ann. Statist.*, *11*, 95–103.

