# EM algorithm and variants: an informal tutorial

Alexis Roche

CEA – Service Hospitalier Frédéric Joliot

4, place du Général Leclerc, 91401 Orsay, France

roche@shfj.cea.fr

## 1 Introduction

The expectation-maximization (EM) algorithm introduced by Dempster et al [12] in 1977 is a very general method to solve maximum likelihood estimation problems. In this informal report, we review the theory behind EM as well as a number of EM variants, suggesting that beyond the current state of the art is an even much wider territory still to be discovered.

## 2 EM background

Let $Y$ a random variable with probability density function (pdf) $p(y|\theta)$, where $\theta$ is an unknown parameter vector. Given an outcome $y$ of $Y$, we aim at maximizing the likelihood function $\mathcal{L}(\theta) \equiv p(y|\theta)$ wrt $\theta$ over a given search space $\Theta$. This is the very principle of maximum likelihood (ML) estimation. Unfortunately, except in not very exciting situations such as, e.g. estimating the mean and variance of a Gaussian population, a ML estimation problem has generally no closed-form solution. Numerical routines are then needed to approximate it.

### 2.1 EM as a likelihood maximizer

The EM algorithm is a class of optimizers specifically taylored to ML problems, which makes it both general and not so general. Perhaps the most salient feature of EM is that it works iteratively by maximizing successive local approximations of the likelihood function. Therefore, each iteration consists of two steps: one that performs the approximation (the E-step) and one that maximizes it (the M-step). But, let's make it clear, not any two-step iterative scheme is an EM algorithm. For instance, Newton and quasi-Newton methods [27] work in a similar iterative fashion but do not have much to do with EM. What essentially defines an EM algorithm is the philosophy underlying the local approximation scheme – which, for instance, doesn't rely on differential calculus.

The key idea underlying EM is to introduce a latent variable $Z$ whose pdf depends on $\theta$ with the property that maximizing $p(z|\theta)$ is easy or, say, easier than maximizing $p(y|\theta)$. Loosely speaking, we somewhat enhance the incomplete data by guessing some useful additional information. Technically, $Z$ can be any variable such that $\theta \to Z \to Y$ is a Markov chain[1], i.e. we assume that $p(y|z,\theta)$ is independent from $\theta$, yielding a Chapman-Kolmogorov equation:

$$p(z,y|\theta) = p(z|\theta)p(y|z) \tag{1}$$

Reasons for that definition will arise soon. Conceptually, $Z$ is a complete-data space in the sense that, if it were fully observed, then estimating $\theta$ would be an easy game. We will emphasize that the convergence speed of EM is highly dependent upon the complete-data specification, which is widely arbitrary despite some estimation problems may have seamingly "natural" hidden variables. But, for the time being, we assume that the complete-data specification step has been accomplished.

## 2.2 EM as a consequence of Jensen's inequality

Quite surprisingly, the original EM formulation stems from a very simple variational argument. Under almost no assumption regarding the complete variable $Z$, except its pdf doesn't vanish to zero, we can bound the variation of the log-likelihood function $L(\theta) \equiv \log p(y|\theta)$ as follows:

$$\begin{aligned}
L(\theta) - L(\theta') &= \log \frac{p(y|\theta)}{p(y|\theta')} \\
&= \log \int \frac{p(z,y|\theta)}{p(y|\theta')}\, dz \\
&= \log \int \frac{p(z,y|\theta)}{p(z,y|\theta')}\, p(z|y,\theta')\, dx \\
&= \log \int \frac{p(z|\theta)}{p(z|\theta')}\, p(z|y,\theta')\, dx \qquad (2) \\
&\geq \underbrace{\int \log \frac{p(z|\theta)}{p(z|\theta')}\, p(z|y,\theta')\, dx}_{\text{Call this } Q(\theta,\theta')} \qquad (3)
\end{aligned}$$

Step (2) results from the fact that $p(y|z,\theta)$ is independent from $\theta$ owing to (1). Step (3) follows from Jensen's inequality (see [9] and appendix A.2) along with the well-known concavity property of the logarithm function. Therefore, $Q(\theta,\theta')$ is an auxiliary function for the log-likelihood, in the sense that: *(i)* the likelihood variation from $\theta'$ to $\theta$ is always greater than $Q(\theta,\theta')$, and *(ii)* we have $Q(\theta',\theta') = 0$. Hence, starting from an initial guess $\theta'$, we are guaranteed to increase the likelihood value if we can find a $\theta$ such that $Q(\theta,\theta') > 0$. Iterating such a process defines an EM algorithm.

There is no general convergence theorem for EM, but thanks to the above mentioned monotonicity property, convergence results may be proved under mild regularity conditions. Typically, convergence towards a non-global likelihood maximizer, or a saddle point, is a worst-case scenario. Still, the only trick behind EM is to exploit the concavity of the logarithm function!

[1] In many presentations of EM, $Z$ is as an aggregate variable $(X,Y)$, where $X$ is some "missing" data, which corresponds to the special case where the transition $Z \to Y$ is deterministic. We believe this restriction, although important in practice, is not useful to the global understanding of EM. By the way, further generalizations will be considered later in this report.

### 2.3 EM as expecation-maximization

Let's introduce some notations. Developing the logarithm in the right-hand side of (3), we may interpret our auxiliary function as a difference: $Q(\theta, \theta') = Q(\theta|\theta') - Q(\theta'|\theta')$, with:

$$Q(\theta|\theta') \equiv \int \log p(z|\theta)\, p(z|y, \theta')\, dx \tag{4}$$

Clearly, for a fixed $\theta'$, maximizing $Q(\theta, \theta')$ wrt $\theta$ is equivalent to maximizing $Q(\theta|\theta')$. If we consider the residual function: $R(\theta|\theta') \equiv L(\theta) - Q(\theta|\theta')$, the incomplete-data log-likelihood may be written as:

$$L(\theta) = Q(\theta|\theta') + R(\theta|\theta')$$

The EM algorithm's basic principle is to replace the maximization of $L(\theta)$ with that of $Q(\theta|\theta')$, hopefully easier to deal with. We can ignore $R(\theta|\theta')$ because inequality (3) implies that $R(\theta|\theta') \geq R(\theta'|\theta')$. In other words, EM works because the auxiliary function $Q(\theta|\theta')$ always deteriorates as a likelihood approximation when $\theta$ departs from $\theta'$. In an ideal world, the approximation error would be constant; then, maximizing $Q$ would, not only increase, but truly maximize the likelihood. Unfortunately, this won't be the case in general. Therefore, unless we decide to give up maximizing the likelihood, we have to iterate – which gives rise to quite a popular statistical learning algorithm.

Given a current parameter estimate $\theta_n$:

- E-step: form the auxiliary function $Q(\theta|\theta_n)$ as defined in (4), which involves computing the posterior distribution of the unobserved variable, $p(z|y, \theta_n)$. The "E" in E-step stands for "expectation" for reasons that will arise in section 2.4.

- M-step: update the parameter estimate by maximizing the auxiliary function:

  $$\theta_{n+1} = \arg\max_{\theta} Q(\theta|\theta_n)$$

  An obvious but important generalization of the M-step is to replace the maximization with a mere increase of $Q(\theta|\theta_n)$. Since, anyway, the likelihood won't be maximized in one iteration, increasing the auxiliary function is enough to ensure that the likelihood will increase in turn, thus preserving the monotonicity property of EM. This defines generalized EM (GEM) algorithms. More on this later.

### 2.4 Some probabilistic interpretations here...

For those familiar with probability theory, $Q(\theta|\theta')$ as defined in (4) is nothing but the conditional expectation of the complete-data log-likelihood in terms of the observed variable, taken at $Y = y$, and assuming the true parameter value is $\theta'$:

$$Q(\theta|\theta') \equiv \mathrm{E}\big[\log p(Z|\theta)|y, \theta'\big] \tag{5}$$

This remark explains the "E" in E-step, but also yields some probabilistic insight on the auxiliary function. For all $\theta$, $Q(\theta|\theta')$ is an estimate of the the complete-data log-likelihood that is built upon the knowledge of the incomplete data and under the naive assumption (but what else can we do?) that the true parameter value is known. In some way, it is not far from being the "best" estimate that we can possibly make without knowing $Z$, because conditional expectation is, by definition, the estimator that minimizes the conditional mean squared error[2].

Having said that, we might still be a bit suspiscious. While we can grant that $Q(\theta|\theta')$ is a reasonable estimate of the complete-data log-likelihood, recall that our initial problem is to maximize the *incomplete-data* (log) likelihood. How good a fit is $Q(\theta|\theta')$ for $L(\theta)$? To answer that, let's see a bit more how the residual $R(\theta|\theta')$ may be interpreted. We have:

$$\begin{aligned} R(\theta|\theta') &= \log p(y|\theta) - \int \log p(z|\theta)\, p(z|y,\theta')\, dz \\ &= \int \log \frac{p(y|\theta)}{p(z|\theta)}\, p(z|y,\theta')\, dz \\ &= \int \log \frac{p(y|z,\theta)}{p(z|y,\theta)}\, p(z|y,\theta')\, dz, \end{aligned} \tag{6}$$

where the last step relies on Bayes' law. Now, $p(y|z,\theta) = p(y|z)$ is independent from $\theta$ by the Markov property (1). Therefore, using the simplified notations $q_\theta(z) \equiv p(z|y,\theta)$ and $q_{\theta'}(z) \equiv p(z|y,\theta')$, we get:

$$R(\theta|\theta') - R(\theta'|\theta') = \underbrace{\int \log \frac{q_{\theta'}(z)}{q_\theta(z)}\, q_{\theta'}(z)\, dz}_{\text{Call this } D(q_{\theta'}\|q_\theta)} \tag{7}$$

In the language of information theory, this quantity $D(q_{\theta'}\|q_\theta)$ is known as the Kullback-Leibler distance, a general tool to assess the deviation between two pdfs [9]. Although it is not, strictly speaking, a genuine mathematical distance, it is always positive and vanishes iff the pdfs are equal which, again and not surprinsingly, comes as a direct consequence of Jensen's inequality.

What does that mean in our case? We noticed earlier that the likelihood approximation $Q(\theta|\theta')$ cannot get any better as $\theta$ deviates from $\theta'$. We now realize from equation (7) that this property reflects an implicit strategy of ignoring the variations of $p(z|y,\theta)$ wrt $\theta$. Hence, a perfect approximation would be one for which $p(z|y,\theta)$ is independent from $\theta$. In other words, we would like $\theta \to Y \to Z$ to define a Markov chain... But, look, we already assumed that $\theta \to Z \to Y$ is a Markov chain. Does the Markov property still hold when permuting the roles of $Y$ and $Z$?

From the fundamental data processing inequality [9], the answer is: of course not. But these details are unnecessary here. Just remember that the validity domain of $Q(\theta|\theta')$ as a local likelihood approximation is controlled by the amount of information that both $y$ and $\theta$ contain about the complete data. We are now going to study this aspect more carefully.

## 2.5 EM as a fix point algorithm and local convergence

Well, you may not give a damn about the previous section, but you'll be glad to know that EM is a fix point algorithm:

$$\theta_{n+1} = \Phi(\theta_n) \qquad \text{with} \quad \Phi(\theta') = \arg\max_{\theta\in\Theta} Q(\theta,\theta')$$

[2] For all $\theta$, we have: $Q(\theta|\theta') = \arg\min_{\mu} \int \Big[\log p(z|\theta) - \mu\Big]^2 p(z|y,\theta')\, dz$.

Assume the sequence $\theta_n$ converges towards some value $\hat{\theta}$ – for instance, the maximum likelihood estimate in a beautiful world. Under the assumption that $\Phi$ is continuous, $\hat{\theta}$ must be a fix point for $\Phi$, i.e. $\hat{\theta} = \Phi(\hat{\theta})$. Furthermore, we can approximate the sequence's asymptotic behavior using a first-order Taylor expansion of $\Phi$ around $\hat{\theta}$, which leads to:

$$\theta_{n+1} \approx S\hat{\theta} + (I - S)\theta_n \qquad \text{with} \quad S = I - \frac{\partial \Phi}{\partial \theta}\Big|_{\hat{\theta}}$$

This expression shows that the rate of convergence is controlled by $S$, a square matrix that is constant across iterations. Hence, $S$ is called the speed matrix, and its spectral radius[3] defines the global speed. Unless the global speed is one, the local convergence of EM is only linear. We may relate $S$ to the likelihood function by exploiting the fact that, under sufficient smoothness assumptions, the maximization of $Q$ is characterized by:

$$\frac{\partial Q}{\partial \theta^t}(\theta_{n+1}, \theta_n) = 0$$

From the implicit function theorem, we get the gradient of $\Phi$:

$$\frac{\partial \Phi}{\partial \theta} = -\Big(\frac{\partial^2 Q}{\partial \theta \partial \theta^t}\Big)^{-1} \frac{\partial^2 Q}{\partial \theta' \partial \theta^t} \quad \Rightarrow \quad S = \Big(\frac{\partial^2 Q}{\partial \theta \partial \theta^t}\Big)^{-1} \Big[\frac{\partial^2 Q}{\partial \theta \partial \theta^t} + \frac{\partial^2 Q}{\partial \theta' \partial \theta^t}\Big]$$

where, after some manipulations:

$$\begin{aligned}
\frac{\partial^2 Q}{\partial \theta \partial \theta^t}\Big|_{(\hat{\theta},\hat{\theta})} &= \int p(z|y,\hat{\theta}) \underbrace{\frac{\partial^2 \log p(z|\theta)}{\partial \theta \partial \theta^t}\Big|_{\hat{\theta}}}_{\text{Call this } -J_z(\hat{\theta})} dz \\
\frac{\partial^2 Q}{\partial \theta' \partial \theta^t}\Big|_{(\hat{\theta},\hat{\theta})} &= \underbrace{\frac{\partial^2 \log p(y|\theta)}{\partial \theta \partial \theta^t}\Big|_{\hat{\theta}}}_{\text{Call this } -J_y(\hat{\theta})} - \frac{\partial^2 Q}{\partial \theta \partial \theta^t}\Big|_{(\hat{\theta},\hat{\theta})}
\end{aligned}$$

The two quantities $J_y(\hat{\theta})$ and $J_z(\hat{\theta})$ turn out to be respectively the observed-data information matrix and the complete-data information matrix. The speed matrix is thus given by:

$$S = \mathcal{J}_z(\hat{\theta})^{-1} J_y(\hat{\theta}) \qquad \text{with} \quad \mathcal{J}_z(\hat{\theta}) \equiv \mathrm{E}\big[J_z(\hat{\theta})|y,\hat{\theta}\big] \tag{8}$$

We easily check that: $\mathcal{J}_z(\hat{\theta}) = J_y(\hat{\theta}) + \mathcal{F}_{z|y}(\hat{\theta})$, where $\mathcal{F}_{z|y}(\hat{\theta})$ is the Fisher information matrix corresponding to the posterior pdf $p(z|y,\hat{\theta})$, which is always symmetric positive. Therefore, we have the alternative expression:

$$S = \big[J_y(\hat{\theta}) + \mathcal{F}_{z|y}(\hat{\theta})\big]^{-1} J_y(\hat{\theta})$$

For fast convergence, we want $S$ close to identity, so we better have the posterior Fisher matrix as "small" as possible. To interpret this result, let's imagine that $Z$ is drawn from $p(z|y,\hat{\theta})$, which is not exactly true since $\hat{\theta}$ may be at least slightly different from the actual parameter value. The Fisher information matrix represents the average information that the complete data contains about $\hat{\theta}$ conditional to the observed data. In this context, $\mathcal{F}_{z|y}(\hat{\theta})$ is a measure of missing information, and the speed matrix is the fraction of missing data. The conclusion is that the rate of convergence of EM is governed by the fraction of missing data.

[3] Let $(\lambda_1, \lambda_2, \dots, \lambda_m)$ be the complex eigenvalues of $S$. The spectral radius is $\rho(S) = \min_i |\lambda_i|$.

## 2.6 EM as a proximal point algorithm

Chrétien & Hero [7] note that EM may also be interpreted as a proximal point algorithm, i.e. an iterative scheme of the form:

$$\theta_{n+1} = \arg\max_{\theta} \big[ L(\theta) - \lambda_n \Psi(\theta, \theta_n) \big], \tag{9}$$

where $\Psi$ is some positive regularization function and $\lambda_n$ is a sequence of positive numbers.

Let us see where this result comes from. In section 2.4, we have established the fundamental log-likelihood decomposition underlying EM, $L(\theta) = Q(\theta|\theta') + R(\theta|\theta')$, and related the variation of $R(\theta|\theta')$ to a Kullback distance (7). Thus, for some current estimate $\theta_n$, we can write:

$$Q(\theta|\theta_n) = L(\theta) - D(q_{\theta_n} \| q_\theta) - R(\theta_n|\theta_n),$$

where $q_\theta(z) \equiv p(z|y,\theta)$ and $q_{\theta_n}(z) \equiv p(x|y,\theta_n)$ are the posterior pdfs of the complete data, under $\theta$ and $\theta_n$, respectively. From this equation, it becomes clear that maximizing $Q(\theta|\theta_n)$ is equivalent to an update rule of the form (9) with:

$$\Psi(\theta, \theta_n) = D(q_{\theta_n} \| q_\theta), \qquad \lambda_n \equiv 1$$

The proximal interpretation of EM is very useful to derive general convergence results [7]. In particular, the convergence rate may be superlinear if the sequence $\lambda_n$ is chosen so as to converge towards zero. Unfortunately, such generalizations are usually intractable because the objective function may no longer simplify as soon as $\lambda_n \neq 1$.

## 2.7 EM as maximization-maximization

Another powerful way of conceptualizing EM is to reinterpret the E-step as another maximization. This idea, which was formalized only recently by Neal & Hinton [26], appears as a breakthrough in the general understanding of EM-type procedures. Let us consider the following function:

$$L(\theta, q) \equiv \mathrm{E}_q\big[\log p(Z, y|\theta)\big] + H(q) \tag{10}$$

where $q(z)$ is some pdf (yes, any pdf), and $H(q)$ is its entropy [9], i.e. $H(q) \equiv -\int \log q(z)\, q(z)\, dz$. We easily obtain an equivalent expression that involves a Kullback-Leiber distance:

$$L(\theta, q) = L(\theta) - D(q \| q_\theta),$$

where we still define $q_\theta(z) \equiv p(z|y,\theta)$ for notational convenience. The last equation reminds us immediately of the proximal interpretation of EM which was briefly discussed in section 2.6. The main difference here is that we don't impose $q(z) = p(z|y,\theta)$ for some $\theta$. Equality holds for *any* distribution!

Assume we have an initial guess $\theta_n$ and try to find $q$ that maximizes $L(\theta_n, q)$. From the above discussed properties of the Kullback-Leibler distance, the answer is $q(z) = q_{\theta_n}(z)$. Now, substitute $q_{\theta_n}$ in (10), and maximize over $\theta$: this is the same as performing a standard M-step[4]! Hence, the conventional EM algorithm boils down to an alternate maximization of $L(\theta, q)$ over a search space $\Theta \times \mathcal{Q}$, where $\mathcal{Q}$ is a suitable set of

[4] To see that, just remember that $p(z,y|\theta) = p(z|\theta)p(y|z)$ where $p(y|z)$ is independent from $\theta$ due to the Markov property (1).

pdfs, i.e. $\mathcal{Q}$ must include all pdfs from the set $\{q(z) = p(z|y, \theta),\ \theta \in \Theta\}$. It is easy to check that any global maximizer $(\hat{\theta}, \hat{q})$ of $L(\hat{\theta}, \hat{q})$ is such that $\hat{\theta}$ is also a global maximizer of $L(\theta)$. By the way, this is also true for local maximizers under weak assumptions [26].

The key observation of Neal & Hinton is that the alternate scheme underlying EM may be replaced with other maximization strategies without hampering the simplicity of EM. In the conventional EM setting, the auxiliary pdf $q_n(z)$ is always constrained to a specific form. This is to say that EM authorizes only specific pathways in the expanded search space $\Theta \times \mathcal{Q}$, yielding some kind of "labyrinth" motion. Easy techniques exist to find its way in a labyrinth, such as breaking the walls or escaping through the roof. Similarly, one may consider relaxing the maximization constraint in the E-step. This leads for instance to incremental and sparse EM variants (see section 3).

# 3 Deterministic EM variants

We first present deterministic EM variants as opposed to stochastic variants. Most of these deterministic variants attempt at speeding up the algorithm, either by simplifying computations, or by increasing the rate of convergence (see section 2.5).

## 3.1 CEM

Classification EM [6]. The whole EM story is about introducing a latent variable $Z$ and performing some inference about its posterior pdf. We might wonder: why not simply estimate $Z$? This is actually the idea underlying the CEM algorithm, which is a simple alternate maximization of the functional $p(z, y|\theta)$ wrt both $\theta$ and $z$. Given a current parameter estimate $\theta_n$, this leads to:

- Classification step: find $z_n = \arg\max_z p(z|y, \theta_n)$.
- Maximization step: find $\theta_{n+1} = \arg\max_\theta p(z_n|\theta)$.

Notice that a special instanciation of CEM is the well-known $k$-means algorithm. In practice, CEM has several advantages over EM, like being easier to implement and typically faster to converge. However, CEM doesn't maximize the incomplete-data likelihood and, therefore, the monotonicity property of EM is lost. While CEM estimates the complete data explicitely, EM estimates only sufficient statistics for the complete data. In this regard, EM may be understood as a fuzzy classifier that avoids the statistical efficiency problems inherent to the CEM approach. Yet, CEM is often useful in practice.

## 3.2 Aitken's acceleration

An early EM extension [12, 21, 22]. Aitken's acceleration is a general purpose technique to speed up the convergence of a fixed point recursion with asymptotic linear behavior. Section 2.5 established that, under appropriate smoothness assumptions, EM may be approximated by a recursion of the form:

$$\theta_{n+1} \approx S\hat{\theta} + (I - S)\theta_n,$$

where $\hat{\theta}$ is the unknown limit and $S$ is the speed matrix given by (8) which depends on this limit. Aitken's acceleration stems from the remark that, if $S$ was known, then the limit could be computed explicitely in a single iteration, namely: $\hat{\theta} \approx \theta_0 + S^{-1}(\theta_1 - \theta_0)$ for some starting value $\theta_0$. Despite that $S$ is unknown and the sequence is not strictly linear, we are tempted to consider the following modified EM scheme. Given a current parameter estimate $\theta_n$,

- E-step: compute $Q(\theta|\theta_n)$ and approximate the inverse speed matrix: $S_n^{-1} = J_y(\theta_n)^{-1}\,\mathcal{J}_z(\theta_n)$.
- M-step: unchanged, get an intermediate value $\theta^* = \arg\max_{\theta} Q(\theta|\theta_n)$.
- Acceleration step: update the parameter using $\theta_{n+1} = \theta_n + S_n^{-1}(\theta^* - \theta_n)$.

It turns out that this scheme is nothing but the Newton-Raphson method to find a zero of $\theta \mapsto \Phi(\theta) - \theta$, where $\Phi$ is the map defined by the EM sequence, i.e. $\Phi(\theta') = \arg\max_\theta Q(\theta|\theta')$. Since the standard EM sets $S_n = I$ on each iteration, it may be viewed as a first-order approach to the same zero-crossing problem, hence avoiding the expense of computing $S_n$. Beside this important implementational issue, convergence is problematic using Aitken's acceleration as the monotonicity property of EM is generally lost.

### 3.3 AEM

Accelerated EM [15]. To trade off between EM and its Aitken's accelerated version (see section 3.2), Jamshidian and Jennrich propose a conjugate gradient approach. Don't be messed up: this is not a traditional gradient-based method (otherwise there would be no point to talk about it in this report). No gradient computation is actually involved in here. The "gradient" is the function $\Phi(\theta) - \theta$, which may be viewed as a generalized gradient for the incomplete-data log-likelihood, hence justifying the use of the generalized conjugate gradient method (see e.g. [27]). Compared to the Aitken's accelerated EM, the resulting AEM algorithm doesn't require computing the speed matrix. Instead, the parameter update rule is of the form:

$$\theta_{n+1} = \theta_n + \lambda_n d_n,$$

where $d_n$ is a direction composed from the current direction $\Phi(\theta_n) - \theta_n$ and the previous directions (the essence of conjugate gradient), and $\lambda_n$ is a step size typically computed from a line maximization of the complete-data likelihood (which may or may not be cumbersome). As an advantage of line maximizations, the monotonicity property of EM is safe. Also, from this generalized gradient perspective, it is straightforward to devise EM extensions that make use of other gradient descent techniques such as the steepest descent or quasi-Newton methods [27].

### 3.4 ECM

Expectation Conditional Maximization [23]. This variant (not to be confused with CEM, see above) was introduced to cope with situations where the standard M-step is intractable. It is the first on a list of coordinate ascent-based EM extensions.

In ECM, the M-step is replaced with a number of lower dimensional maximization problems called CM-steps. This implies decomposing the parameter space as a sum of subspaces, which, up to some possible reparameterization, is the same as splitting the parameter vector into several blocks, $\theta = (t_1, t_2, \dots, t_s)$. Starting from a current estimate $\theta_n$, the CM-steps update one coordinate block after another by partially maximizing the auxiliary $Q$-function, yielding a scheme similar in essence to Powell's multidimensional optimization method [27]. This produces a sequence $\theta_n = \theta_{n,0} \to \theta_{n,1} \to \theta_{n,2} \to \dots \to \theta_{n,s-1} \to \theta_{n,s} = \theta_{n+1}$, such that:

$$Q(\theta_n|\theta_n) \leq Q(\theta_{n,1}|\theta_n) \leq Q(\theta_{n,2}|\theta_n) \leq \dots \leq Q(\theta_{n,s-1}|\theta_n) \leq Q(\theta_{n+1}|\theta_n)$$

Therefore, the auxiliary function is guaranteed to increase on each CM-step, hence globally in the M-step, and so does the incomplete-data likelihood. Hence, ECM is a special case of GEM (see section 2.3).

### 3.5 ECME

ECM either [18]. This is an extension of ECM where some CM-steps are replaced with steps that maximize, or increase, the incomplete-data log-likelihood $L(\theta)$ rather than the auxiliary $Q$-function. To make sure that the likelihood function increases globally in the M-step, the only requirement is that the CM-steps that act on the actual log-likelihood be performed after the usual $Q$-maximizations. This is because increasing the $Q$-function only increases likelihood from the starting point, namely $\theta_n$, which is held fixed during the M-step (at least, this is what we assume)[5].

Starting with $Q$-maximizations is guaranteed to increase the likelihood, and of course subsequent likelihood maximizations can only improve the situation. With the correct setting, ECME is even more general than GEM as defined in section 2.3 while inheriting its fundamental monotonicity property. An example application of ECME is in mixture models, where typically mixing proportions are updated using a one-step Newton-Raphson gradient descent on the incomplete-data likelihood, leading to a simple additive correction to the usual EM update rule [18]. At least in this case, ECME has proved to converge faster than standard EM.

### 3.6 SAGE

Space-Alternating Generalized EM [13, 14]. In the continuity of ECM and ECME (see sections 3.4 and 3.5), one can imagine defining an auxiliary function specific to each coordinate block of the parameter vector. More technically, using a block decomposition $\theta = (t_1, t_2, \dots, t_s)$, we assume that, for each block $i = 1, \dots, s$, there exists a function $Q_i(\theta|\theta')$ such that, for all $\theta$ and $\theta'$ with identical block coordinates except (maybe) for the $i$-th block, we have: $L(\theta) - L(\theta') \geq Q_i(\theta|\theta') - Q_i(\theta'|\theta')$.

[5] For example, if one chooses $\theta^*$ such that $L(\theta^*) \geq L(\theta_n)$ and, then, $\theta_{n+1}$ such that $Q(\theta_{n+1}|\theta_n) \geq Q(\theta^*|\theta_n)$, the only conlusion is that the likelihood increases from $\theta_n$ to $\theta^*$, but may actually decrease from $\theta^*$ to $\theta_{n+1}$ because $\theta^*$ is not the starting point of $Q$. Permuting the $L$-maximization and the $Q$-maximization, we have $Q(\theta^*|\theta_n) \geq Q(\theta_n|\theta_n)$, thus $L(\theta^*) \geq L(\theta_n)$, and therefore $L(\theta_{n+1}) \geq L(\theta_n)$ since we have assumed $L(\theta_{n+1}) \geq L(\theta^*)$. This argument generalizes easily to any intermediate sequence using the same cascade inequalities as in the derivation of ECM (see section 3.4).

This idea has two important implications. First, the usual ECM scheme needs to be rephrased, because changing the auxiliary function across CM-steps may well result in decreasing the likelihood, a problem worked around in ECME with an appropriate ordering of the CM-steps. In this more general framework, though, there may be no such fix to save the day. In order to ensure monotonicity, at least some CM-steps should start with "reinitializing" their corresponding auxiliary function, which means... performing an E-step. It is important to realize that, because the auxiliary function is coordinate-specific, so is the E-step. Hence, each "CM-step" becomes an EM algorithm in itself which is sometimes called a "cycle". We end up with a nested algorithm where cycles are embedded in a higher-level iterative scheme.

Furthermore, how to define the $Q_i$'s? From section 2, we know that the standard EM auxiliary function $Q(\theta|\theta')$ is built from the complete-data space $Z$; see in particular equation (5). Fessler & Herro introduce *hidden-data spaces*, a concept that generalizes complete-data spaces in the sense that hidden-data spaces may be coordinate-specific, i.e. there is a hidden variable $Z_i$ for each block $t_i$. Formally, given a block decomposition $\theta = (t_1, t_2, \dots, t_s)$, $Z_i$ is a hidden-data space for $t_i$ if:

$$p(z_i, y|\theta) = p(y|z_i, \{t_{j\neq i}\})\, p(z_i|\theta)$$

This definition's main feature is that the conditional probability of $Y$ knowing $Z_i$ is allowed to be dependent on every parameter block but $t_i$. Let us check that the resulting auxiliary function fulfils the monotonicity condition. We define:

$$Q_i(\theta|\theta') \equiv E\big[\log p(Z_i|\theta)|y, \theta'\big]$$

Then, applying Jensen's inequality (3) like in section 2.2, we get:

$$L(\theta) - L(\theta') \geq Q_i(\theta|\theta') - Q_i(\theta'|\theta') + \int \log \frac{p(y|z_i,\theta)}{p(y|z_i,\theta')}\, p(z_i|y,\theta')\, dz_i$$

When $\theta$ and $\theta'$ differ only by $t_i$, the integral vanishes because the conditional pdf $p(y|z_i,\theta)$ is independent from $t_i$ by the above definition. Consequently, maximizing $Q_i(\theta|\theta')$ with respect to $t_i$ only (the other parameters being held fixed) is guaranteed to increase the incomplete-data likelihood. Specific applications of SAGE to the Poisson imaging model or penalized least-squares regression were reported to converge much faster than standard EM.

### 3.7 CEMM

Component-wise EM for Mixtures [4]. Celeux et al extend the SAGE methodology to the case of constrained likelihood maximization, which arises typically in mixture problems where the sum of mixing proportions should equate to one. Using a Lagrangian dualization approach, they recast the initial problem into unconstrained maximization by defining an appropriate penalized log-likelihood function. The CEMM algorithm they derive is a natural coordinatewise variant of EM whose convergence to a stationary point of the likelihood is established under mild regularity conditions.

### 3.8 AECM

Alternating ECM [24, 25]. In an attempt to summarize earlier contributions, Meng & van Dyk propose to cast a number of EM extensions into a unified framework, the so-called AECM algorithm. Essentially, AECM is a SAGE algorithm (itself a generalization of both ECM and ECME) that includes another data augmentation trick. The idea is to consider a family of complete-data spaces indexed by a working parameter $\alpha$. More formally, we define a joint pdf $q(z, y|\theta, \alpha)$ as depending on both $\theta$ and $\alpha$, yet imposing the constraint that the corresponding marginal incomplete-data pdf be preserved:

$$p(y|\theta) = \int q(z, y|\theta, \alpha)\, dz,$$

and thus independent from $\alpha$. In other words, $\alpha$ is identifiable only given the complete data. A simple way of achieving such data augmentation is to define $Z = f_{\theta,\alpha}(Z_0)$, where $Z_0$ is some reference complete-data space and $f_{\theta,\alpha}$ is a one-to-one mapping for any $(\theta, \alpha)$. Interestingly, it can be seen that $\alpha$ has no effect if $f_{\theta,\alpha}$ is insensitive to $\theta$. In AECM, $\alpha$ is tuned beforehand so as that to minimize the fraction of missing data (8), thereby maximizing the algorithm's global speed. In general, however, this initial mimimization cannot be performed exactly since the global speed may depend on the unknown maximum likelihood parameter.

### 3.9 PX-EM

Parameter-Expanded EM [19, 17]. Liu et al revisit the working parameter method suggested by Meng and van Dyk [24] (see section 3.8) from a slighlty different angle. In their strategy, the joint pdf $q(z, y|\theta, \alpha)$ is defined so as to meet the two following requirements. First, the baseline model is embedded in the expanded model in the sense that $q(z, y|\theta, \alpha_0) = p(z, y|\theta)$ for some null value $\alpha_0$. Second, which is the main difference with AECM, the observed-data marginals are consistent up to a many-to-one reduction function $r(\theta, \alpha)$,

$$p(y|r(\theta, \alpha)) = \int q(z, y|\theta, \alpha)\, dz,$$

for all $(\theta, \alpha)$. From there, the trick is to to "pretend" estimating $\alpha$ iteratively rather than pre-processing its value.

The PX-EM algorithm is simply an EM on the expanded model with additional instructions after the M-step to apply the reduction function and reset $\alpha$ to its null value. Thus, given a current estimate $\theta_n$, the E-step forms the auxiliary function corresponding to the expanded model from $(\theta_n, \alpha_0)$, which amounts to the standard E-step because $\alpha = \alpha_0$. The M-step then provides $(\theta^*, \alpha^*)$ such that $q(y|\theta^*, \alpha^*) \geq q(y|\theta_n, \alpha_0)$, and the additional reduction step updates $\theta_n$ according to $\theta_{n+1} = r(\theta^*, \alpha^*)$, implying $p(y|\theta_{n+1}) = q(y|\theta^*, \alpha^*)$. Because $q(y|\theta_n, \alpha_0) = p(y|\theta_n)$ by construction of the expanded model, we conclude that $p(y|\theta_{n+1}) \geq p(y|\theta_n)$, which shows that PX-EM preserves the monotonicity property of EM.

In some way, PX-EM capitalizes on the fact that a large deviation between the estimate of $\alpha$ and its known value $\alpha_0$ is an indication that the parameter of interest $\theta$ is poorly estimated. Hence, PX-EM adjusts the M-step for this deviation via the reduction function. A variety of examples where PX-EM converges much faster than EM is reported in [19]. Possible variants of PX-EM include the coordinatewise extensions underlying SAGE.

### 3.10 Incremental EM

Following the maximization-maximization approach discussed in section 2.7, Neal & Hinton [26] address the common case where observations are i.i.d. Then, we have $p(y|\theta) = \prod_i p(y_i|\theta)$ and, similarly, the global EM objective function (10) reduces to:

$$L(\theta, q) = \sum_i \left\{ \mathrm{E}_{q_i} \left[ \log p(Z_i, y_i|\theta) \right] + H(q_i) \right\},$$

where we can search for $q$ under the factored form $q(z) = \prod_i q_i(z)$. Therefore, for a given $\theta$, maximizing $L(\theta, q)$ wrt $q$ is equivalent to maximizing the contribution of each data item wrt $q_i$, hence splitting the global maximization problem into a number of simpler maximizations. Incremental EM variants are justified from this remark, the general idea being to update $\theta$ by visiting the data items sequencially rather than from a global E-step. Neal & Hinton demonstrate an incremental EM variant for mixtures that converges twice as fast as standard EM.

### 3.11 Sparse EM

Another suggestion of Neal & Hinton [26] is to track the auxiliary distribution $q(z)$ in a subspace of the original search space $\mathcal{Q}$ (at least for a certain number of iterations). This general strategy includes sparse EM variants where $q$ is updated only at pre-defined plausible unobserved values. Alternatively, "winner-take-all" EM variants such as the CEM algorithm [6] (see section 3.1) may be seen in this light. Such procedures may have strong computational advantages but, in counterpart, are prone to estimation bias. In the maximization-maximization interpretation of EM, this comes as no surprise since these approaches "project" the estimate on a reduced search space that may not contain the maximum likelihood solution.

## 4 Stochastic EM variants

While deterministic EM variants were mainly motivated by convergence speed considerations, stochastic variants are more concerned with other limitations of standard EM. One is that the EM auxiliary function (4) involves computing an integral that may not be tractable in some situations. The idea is then to replace this tedious computation with a stochastic simulation. As a typical side effect of such an approach, the modified algorithm inherits a lesser tendancy to getting trapped in local maxima, yielding improved global convergence properties.

### 4.1 SEM

Stochastic EM [5]. As noted in section 2.4, the standard EM auxiliary function is the best estimate of the complete-data log-likelihood in the sense of the conditional mean squared error. The idea underlying SEM, like other stochastic EM variants, is that there might be no need to ask for such a "good" estimate. Therefore, SEM replaces the standard auxiliary function with:

$$\hat{Q}(\theta|\theta') = \log p(z'|\theta'),$$

where $z'$ is a random sample drawn from the posterior distribution of the unobserved variable[6], $p(z|y,\theta')$. This leads to the following modified iteration; given a current estimate $\theta_n$:

- Simulation step: compute $p(z|y,\theta_n)$ and draw an unobserved sample $z_n$ from $p(z|y,\theta_n)$.
- Maximization step: find $\theta_{n+1} = \arg\max_\theta p(z_n|\theta)$.

By construction, the resulting sequence $\theta_n$ is an homogeneous Markov chain[7] which, under mild regularity conditions, converges to a stationary pdf. This means in particular that $\theta_n$ doesn't converge to a unique value! Various schemes can be used to derive a pointwise limit, such as averaging the estimates over iterations once stationarity has been reached (see also SAEM regarding this issue). It was established in some specific cases that the stationary pdf concentrates around the likelihood maximizer with a variance inversly proportional to the sample size. However, in cases where several local maximizers exist, one may expect a multimodal behavior.

## 4.2 DA

Data Augmentation algorithm [28]. Going further into the world of random samples, one may consider replacing the M-step in SEM with yet another random draw. In a Bayesian context, maximizing $p(z_n|\theta)$ wrt $\theta$ may be thought of as computing the mode of the posterior distribution $p(\theta|z_n)$, given by:

$$p(\theta|z_n) = \frac{p(z_n|\theta)p(\theta)}{\int p(z_n|\theta')p(\theta')\,d\theta'}$$

where we can assume a flat (or non-informative) prior distribution for $\theta$. In DA, this maximization is replaced with a random draw $\theta_{n+1} \sim p(\theta|z_n)$. From equation (1), we easily check that $p(\theta|z_n) = p(\theta|z_n, y)$. Therefore, DA alternates conditional draws $z_n|(\theta_n, y)$ and $\theta_{n+1}|(z_n, y)$, which is the very principle of a Gibbs sampler. Results from Gibbs sampling theory apply, and it is shown under general conditions that the sequence $\theta_n$ is a Markov chain that converges in distribution towards $p(\theta|y)$. Once the sequence has reached stationarity, averaging $\theta_n$ over iterations yields a random variable that converges to the conditional mean $\mathrm{E}(\theta|y)$, which is an estimator of $\theta$ generally different from the maximum likelihood but not necessarily worse.

Interesting enough, several variants of DA have been proposed recently [20, 17] following the parameter expansion strategy underlying the PX-EM algorithm described in section 3.9.

## 4.3 SAEM

Stochastic Approximation type EM [3]. The SAEM algorithm is a simple hybridation of EM and SEM that provides a pointwise convergence as opposed to the erratic behavior of SEM. Given a current estimate $\theta_n$, SAEM performs a standard EM iteration in addition to the SEM iteration. The parameter is then updated as a weighted mean of both contributions, yielding:

$$\theta_{n+1} = (1-\gamma_{n+1})\theta_{n+1}^{EM} + \gamma_{n+1}\theta_{n+1}^{SEM},$$

[6] Notice that when $Z$ is defined as $Z = (X, Y)$, this simulation reduces to a random draw of the missing data $X$.

[7] The draws need to be mutually independent conditional to $(\theta_1, \theta_2, \dots, \theta_n)$, i.e. $p(z_1, z_2, \dots, z_n|\theta_1, \theta_2, \dots, \theta_n) = \prod_i p(z_i|\theta_i)$.

where $0 \leq \gamma_n \leq 1$. Of course, to apply SAEM, the standard EM needs to be tractable.

The sequence $\gamma_n$ is typically chosen so as to decrease from 1 to 0, in such a way that the algorithm is equivalent to SEM in the early iterations, and then becomes more similar to EM. It is established that SAEM converges almost surely towards a local likelihood maximizer (thus avoiding saddle points) under the assumption that $\gamma_n$ decreases to 0 with $\lim_{n\to\infty}(\gamma_n/\gamma_{n+1}) = 1$ and $\sum_n \gamma_n \to \infty$.

## 4.4 MCEM

Monte Carlo EM [30]. At least formally, MCEM turns out to be a generalization of SEM. In the SEM simulation step, draw $m$ independent samples $z_n^{(1)}, z_n^{(2)}, \dots, z_n^{(m)}$ instead of just one, and then maximize the following function:

$$\hat{Q}(\theta|\theta_n) = \frac{1}{m}\sum_{j=1}^{m} \log p(z_n^{(j)}|\theta),$$

which, in general, converges almost surely to the standard EM auxiliary function thanks to the law of large numbers.

Choosing a large value for $m$ justifies calling this Monte Carlo something. In this case, $\hat{Q}$ may be seen as an empirical approximation of the standard EM auxiliary function, and the algorithm is expected to behave similarly to EM. On the other hand, choosing a small value for $m$ is not forbidden, if not advised (in particular, for computational reasons). We notice that, for $m = 1$, MCEM reduces to SEM. A possible strategy consists of increasing progressively the parameter $m$, yielding a “simulated annealing” MCEM which is close in spirit to SAEM.

## 4.5 SAEM2

Stochastic Approximation EM [11]. Delyon et al propose a generalization of MCEM called SAEM, not to be confused with the earlier SAEM algorithm presented in section 4.3, although both algorithms promote a similar simulated annealing philosophy. In this version, the auxiliary function is defined recursively by averaging a Monte Carlo approximation with the auxiliary function computed in the previous step:

$$\hat{Q}_n(\theta) = (1-\gamma_n)\hat{Q}_{n-1}(\theta) + \frac{\gamma_n}{m_n}\sum_{j=1}^{m_n} \log p(z_n^{(j)}|\theta),$$

where $z_n^{(1)}, z_n^{(2)}, \dots, z_n^{(m_n)}$ are drawn independently from $p(z|y,\theta_n)$. The weights $\gamma_n$ are typically decreased across iterations in such a way that $\hat{Q}_n(\theta)$ eventually stabilizes at some point. One may either increase the number of random draws $m_n$, or set a constant value $m_n \equiv 1$ when simulations have heavy computanional cost compared to the maximization step. The convergence of SAEM2 towards a local likelihood maximizer is proved in [11] under quite general conditions.

Kuhn et al [16] further extend the technique to make it possible to perform the simulation under a distribution $\Pi_{\theta_n}(z)$ simpler to deal with than the posterior pdf $p(z|y,\theta_n)$. Such a distribution may be defined as the transition probability of a Markov chain generated by a Metropolis-Hastings algorithm. If $\Pi_\theta(z)$ is such that its associated Markov chain converges to $p(z|y,\theta)$, then the convergence properties of SAEM2 generalize under mild additional assumptions.

# 5 Conclusion

This report's primary goal is to give a flavor of the current state of the art on EM-type statistical learning procedures. We also hope it will help researchers and developers in finding the literature relevant to their current existential questions. For a more comprehensive overview, we advise some good tutorials that are found on the internet [8, 2, 1, 10, 29, 17].

# A Appendix

## A.1 Maximum likelihood quickie

Let $Y$ a random variable with pdf $p(y|\theta)$, where $\theta$ is an unknown parameter vector. Given an outcome $y$ of $Y$, maximum likelihood estimation consists of finding the value of $\theta$ that maximizes the probability $p(y|\theta)$ over a given search space $\Theta$. In this context, $p(y|\theta)$ is seen as a function of $\theta$ and called the likelihood function. Since it is often more convenient to manipulate the logarithm of this expression, we will focus on the equivalent problem of maximizing the log-likelihood function:

$$\hat{\theta}(y) = \arg\max_{\theta\in\Theta} L(y, \theta)$$

where the log-likelihood $L(y, \theta) \equiv \log p(y|\theta)$ is denoted $L(y, \theta)$ to emphasize the dependance in $y$, contrary to the notation $L(\theta)$ usually employed throghout this report. Whenever the log-likelihood is differentiable wrt $\theta$, we also define the score function as the log-likelihood gradient:

$$S(y, \theta) = \frac{\partial L}{\partial \theta}(y, \theta)$$

In this case, a fundamental result is that, for all vector $U(y, \theta)$, we have:

$$\mathrm{E}(SU^t) = \frac{\partial}{\partial \theta}\mathrm{E}(U^t) - \mathrm{E}\Big(\frac{\partial U^t}{\partial \theta}\Big)$$

where the expectation is taken wrt the distribution $p(y|\theta)$. This equality is easily obtained from the logarithm differentiation formula and some additional manipulations. Assigning the "true" value of $\theta$ in this expression leads to the following:

- $\mathrm{E}(S) = 0$
- $\mathrm{Cov}(S, S) = -\mathrm{E}\Big(\frac{\partial S^t}{\partial \theta}\Big)$ (Fisher information)
- If $U(y)$ is an unbiased estimator of $\theta$, then $\mathrm{Cov}(S, U) = \mathrm{Id}$.
- In the case of a single parameter, the above result implies $\mathrm{Var}(U) \geq \frac{1}{\mathrm{Var}(S)}$ from the Cauchy-Schwartz inequality, i.e. the Fisher information is a lower bound for the variance of $U$. Equality occurs iff $U$ is an affine function of $S$, which imposes a specific form to $p(y|\theta)$ (Darmois theorem).

### A.2 Jensen's inequality

For any random variable $X$ and any real continuous concave function $f$, we have:

$$f\big[\mathrm{E}(X)\big] \geq \mathrm{E}\big[f(X)\big],$$

If $f$ is strictly concave, equality occurs iff $X$ is deterministic.

## References

[1] A. Berger. Convexity, Maximum Likelihood and All That. Tutorial published on the web, 1998.

[2] J. Bilmes. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report ICSI-TR-97-021, University of Berkeley, Apr. 1998.

[3] G. Celeux, D. Chauveau, and J. Diebolt. On Stochastic Versions of the EM Algorithm. Technical Report 2514, INRIA, Mar. 1995.

[4] G. Celeux, S. Chrétien, F. Forbes, and A. Mkhadri. A Component-wise EM Algorithm for Mixtures. *Journal of Computational and Graphical Statistics*, 10:699–712, 2001.

[5] G. Celeux and J. Diebolt. The SEM Algorithm: a Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem. *Computational Statistics Quaterly*, 2:73–82, 1985.

[6] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14:315–332, 1992.

[7] S. Chrétien and A. O. Hero. Kullback Proximal Algorithms for Maximum Likelihood Estimation. *IEEE Transactions on Information Theory*, 46(5):1800–1810, 2000.

[8] C. Couvreur. The EM Algorithm: A Guided Tour. In *Proc. 2d IEEE European Workshop on Computationaly Intensive Methods in Control and Signal Processing*, Prague, Czech Republik, Aug. 1996.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, inc., 1991.

[10] F. Dellaert. The Expectation Maximization Algorithm. Tutorial published on the web, 2002.

[11] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, 27(1):94–128, 1999.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[13] J. A. Fessler and A. O. Hero. Space-Alternating Generalized Expectation-Maximization Algorithm. *IEEE Transactions on Signal Processing*, 42(10):2664–2677, 1994.

[14] J. A. Fessler and A. O. Hero. Penalized Maximum-Likelihood Image Reconstruction Using Space-Alternating Generalized EM Algorithms. *IEEE Transactions on Image Processing*, 4(10):1417–1429, 1995.

[15] M. Jamshidian and R. I. Jennrich. Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association*, 88:221–228, 1993.

[16] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with a MCMC procedure. Submitted, 2002.

[17] C. Liu. An Example of Algorithm Mining: Covariance Adjustment to Accelerate EM and Gibbs. To appear in Development of Modern Statistics and Related Topics, 2003.

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

[19] C. Liu, D. B. Rubin, and Y. N. Wu. Parameter Expansion to Accelerate EM: The PX-EM Algorithm. *Biometrika*, 85:755–770, 1998.

[20] J. S. Liu and Y. N. Wu. Parameter Expansion for Data Augmentation. *Journal of the American Statistical Association*, 94:1264–1274, 1999.

[21] T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, 44:226–233, 1982.

[22] I. Meilijson. A fast improvement of the EM algorithm on its own terms. *Journal of the Royal Statistical Society*, 51:127–138, 1989.

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

[24] X. L. Meng and D. A. van Dyk. The EM Algorithm - An Old Folk Song Sung To A Fast New Tune. *Journal of the Royal Statistical Society*, 59:511–567, 1997.

[25] X. L. Meng and D. A. van Dyk. Seeking efficient data augmentation schemes via conditional and marginal data augmentation. *Biometrika*, 86(2):301–320, 1999.

[26] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

[27] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*. Cambrige University Press, 2nd edition, 1992.

[28] M. A. Tanner and W. H. Wong. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82:528–550, 1987.

[29] D. A. van Dyk and X. L. Meng. Algorithms based on data augmentation: A graphical representation and comparison. In K. Berk and M. Pourahmadi, editors, *Computing Science and Statistics: Proceedings of the 31st Symposium on the Interface*, pages 230–239. Interface Foundation of North America, 2000.

[30] G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, 85:699–704, 1990.