# Theory and Practice of Expectation Maximization (EM) Algorithm

**Chandan K. Reddy**
*Wayne State University, USA*

**Bala Rajaratnam**
*Stanford University, USA*

## INTRODUCTION

In the field of statistical data mining, the Expectation Maximization (EM) algorithm is one of the most popular methods used for solving parameter estimation problems in the maximum likelihood (ML) framework. Compared to traditional methods such as steepest descent, conjugate gradient, or Newton-Raphson, which are often too complicated to use in solving these problems, EM has become a popular method because it takes advantage of some problem specific properties (Xu et al., 1996). The EM algorithm converges to the local maximum of the log-likelihood function under very general conditions (Demspter et al., 1977; Redner et al., 1984). Efficiently maximizing the likelihood by augmenting it with latent variables and guarantees of convergence are some of the important hallmarks of the EM algorithm.

EM based methods have been applied successfully to solve a wide range of problems that arise in fields of pattern recognition, clustering, information retrieval, computer vision, bioinformatics (Reddy et al., 2006; Carson et al., 2002; Nigam et al., 2000), etc. Given an initial set of parameters, the EM algorithm can be implemented to compute parameter estimates that locally maximize the likelihood function of the data. In spite of its strong theoretical foundations, its wide applicability and important usage in solving some real-world problems, the standard EM algorithm suffers from certain fundamental drawbacks when used in practical settings. Some of the main difficulties of using the EM algorithm on a general log-likelihood surface are as follows (Reddy et al., 2008):

- EM algorithm for mixture modeling converges to a local maximum of the log-likelihood function very quickly.
- There are many other promising local optimal solutions in the close vicinity of the solutions obtained from the methods that provide good initial guesses of the solution.
- Model selection criterion usually assumes that the global optimal solution of the log-likelihood function can be obtained. However, achieving this is computationally intractable.
- Some regions in the search space do not contain any promising solutions. The promising and non-promising regions co-exist and it becomes challenging to avoid wasting computational resources to search in non-promising regions.

Of all the concerns mentioned above, the fact that most of the local maxima are not distributed uniformly makes it important to develop algorithms that not only help in avoiding some inefficient search over the low-likelihood regions but also emphasize the importance of exploring promising subspaces more thoroughly (Zhang et al, 2004). This subspace search will also be useful for making the solution less sensitive to the initial set of parameters. In this chapter, we will discuss the theoretical aspects of the EM algorithm and demonstrate its use in obtaining the optimal estimates of the parameters for mixture models. We will also discuss some of the practical concerns of using the EM algorithm and present a few results on the performance of various algorithms that try to address these problems.

## BACKGROUND

Because of its greedy nature, the EM algorithm converges to a local maximum on the log-likelihood surface. Hence, the final solution will be very sensitive to the given initial set of parameters. This local maxima problem (popularly known as the initializa-

tion problem) is one of the well studied issues in the context of the EM algorithm. Several algorithms have been proposed in the literature to try and solve this issue (Reddy, 2007).

Although EM and its variants have been extensively used in the literature, several researchers have approached the problem by identifying new techniques that give good initialization. More generic techniques like deterministic annealing (Ueda et al., 1998), genetic algorithms (Pernkopf et al., 2005) have been successfully applied to obtain good parameter estimates. Though, these techniques have asymptotic guarantees, they are very time consuming and hence cannot be used in most practical applications. Some problem specific algorithms like split and merge EM (Ueda et al., 2000), component-wise EM (Figueiredo et al., 2002), greedy learning (Verbeek et al., 2003), parameter space grid (Li, 1999) have also been proposed in the literature. Some of these algorithms are either computationally very expensive or infeasible when learning mixture models in high dimensional spaces (Li, 1999). In spite of the high computational cost associated with these methods, very little effort has been taken to explore promising subspaces within the larger parameter space. Most of the above mentioned algorithms eventually apply the EM algorithm to move to a locally maximal set of parameters on the log-likelihood surface. Simpler practical approaches like running EM from several random initializations, and then choosing the final estimate that leads to the local maximum with the highest log-likelihood value to a certain extent have also been successful.

For a problem with a non-uniform distribution of local maxima, it is difficult for most methods to search neighboring subspaces (Zhang et al, 2004). Though some of these methods apply other additional mechanisms (like perturbations) to escape out of local optimal solutions, systematic methods for searching the subspace have not been thoroughly studied. More recently, TRUST-TECH based Expectation Maximization (TRUST-TECH-EM) algorithm has been developed by Reddy et al (2008), which applies some properties of the dynamical system of the log-likelihood surface to identify promising initial starts for the EM algorithm. This dynamical system approach will reveal more information about the neighborhood regions and helps in moving to different basins of attraction in the neighborhood of the current local maximum.

## MAIN FOCUS

In this section, we will first discuss the theoretical aspects of the EM algorithm and prove some of its basic properties. We will then demonstrate the use of the EM algorithm in the context of mixture models and give some comparative results on multiple datasets.

## THEORY OF THE EM ALGORITHM

Formally consider the problem of maximizing the likelihood function $L(\theta;x)$ arising from a density $f(x;\theta)$, with $x$ denoting the data or sample, and $\theta$ the parameter of interest. As noted above, in both theoretical and applied problems maximizing $L(\theta;x)$ can often be a difficult task. Let us assume that we can identify another random variable y such that

$$f(x;\theta) = \int f(x,y;\theta)dy \qquad (1)$$

and where the likelihood function arising from $f(x,y;\theta)$ is relatively easier to maximize. The variable y is often called the "hidden", "latent" or "missing" data and together $(x,y)$ is often referred to as the "complete" data.

The EM algorithm maximizes the original likelihood function by working with the complete or augmented likelihood. The expectation or E-step takes the expected value of the complete likelihood over the missing data given the original data $y$ and a starting parameter value. This process gives rise to an expected (rather conditional expectation) version of the complete likelihood which is easier to maximize. The E-step essentially has the effect of "substituting" values for the hidden variable $y$. The maximization or M-step optimizes the resulting conditional expectation of the complete likelihood leading to a new parameter estimate. Based on the new parameter estimate, the E-step and the M-step are repeated back and forth in an iterative manner (McLachlan et al., 1997).

We shall prove below that every EM-step gives an improvement in the likelihood in the original problem but let us first formally state the EM algorithm.