Fast Incremental Stochastic Version of the EM algorithm

B. Karimi[1,2], **M. Lavielle**[1,2], **E. Moulines**[2]

INRIA[1], CMAP École Polytechnique[2]

belhal.karimi@polytechnique.edu

1 Problem statement

A wide class of statistical problems involves observed and unobserved data. We can consider, for example, inverse problems concerning deconvolution, source separation, change-points detection, etc. Linear and nonlinear mixed effects models can also be considered as incomplete-data models. Estimation of the parameters of these models is a difficult challenge. In particular, the likelihood of the observations cannot usually be maximized in closed form. The EM algorithm proposed by Dempster, Laird and Rubin led to many variants when the conditional expectation of the complete log-likelihood is intractable. The MCEM (Meng, 1993) and the SAEM (Delyon, 1999) are two of them.
Following Neal, Hinton and Gunawardana efforts in justifying a variant version of the EM algorithm considering an incremental scheme, we decided to focus on the Incremental EM, MCEM and SAEM for continuous random variables.

2 Convergence Theorems

2.1 IEM algorithm

Following Gunawardana work we can show that the IEM on missing data problem ($Y = (Y_i)_{i=1}^N$ are observed and $Z = (Z_i)_{i=1}^N$ are latent) can be seen as a Alternatively minimizing problem. In that context we proposed a criteria that we will be minimizing alternatively, for all $\theta \in \Theta$ and $(\delta_i)_{i=1}^N \in \Theta^N$:

$$A\colon \Theta^{N+1} \to \mathbb{R}$$

$$(\theta, \delta_1, \dots, \delta_N) \mapsto D_{KL}(\prod_{i=1}^N P_{Z_i|Y_i,\delta_i} \parallel \prod_{i=1}^N P_{Z_i|Y_i,\theta}) - \sum_{i=1}^N \log p(y_i, \theta)$$

Two successive mappings $F_{I_k}\colon \Theta^{N+1} \to \Theta^{N+1}$, where I_k is the index picked at iteration k, and $B\colon \Theta^{N+1} \to \Theta^{N+1}$ decrease this criteria, strictly when the input is outside the solution set.

$$F_{I_k}(\theta, \delta_{I_k}, \delta_{-I_k}) = (\theta, \arg\min_{\delta_{I_k} \in \Theta} D_{KL}(P_{z_{I_k}|y_{I_k},\delta_{I_k}} \parallel P_{z_{I_k}|y_{I_k},\theta^{k-1}}), \delta_{-I_k})$$

$$B(\theta, (\delta_i)_{i=1}^N) = (\arg\min_{\theta \in \Theta} A(\theta, \delta_1^{(k)}, \dots, \delta_N^{(k)}), (\delta_i)_{i=1}^N)$$

Using Zangwill global convergence theorem allows us to write this first convergence theorem.

Let $\{(\theta^t, \delta_1^t, \dots, \delta_N^t)\}_{k=0}^\infty$ be a sequence generated from a tuple $(\theta^0, \delta_1^0, \dots, \delta_N^0)$ by the iterative application, with its associated solution set $\mathcal{L}_{IEM} = \{(\theta^*, \delta_1^*, \dots, \delta_N^*) \in \Theta^{N+1}, \theta^* \in \mathcal{L}_{EM}, \theta^* = \delta_1^* = \dots = \delta_N^*\}$, of the point-to-set map $T = T_N \circ \dots \circ T_1$, where $T_i = B \circ F_i$ at each iteration t:

$$(\theta^t, \delta_1^t, \dots, \delta_N^t) = T(\theta^{t-1}, \delta_1^{t-1}, \dots, \delta_N^{t-1})$$

Theorem 2.1: IEM convergence theorem

Let $T : \Theta^{N+1} \to \Theta^{N+1}$ be the composition of the point-to-set maps B and F_i as defined in the algorithm.
With certain assumptions of compactness and continuity:

1. All accumulation points of the sequence $\{(\theta^t, \delta_1^t, \dots, \delta_N^t)\}_{t=0}^\infty$ lie in the solution set $\mathcal{L}_{IEM}$
2. The sequence $\{\theta_t\}_{t=0}^\infty$ converges to stationary points of the incomplete data likelihood, i.e. all the accumulation points of this sequence are in $\mathcal{L}_{EM} = \{\theta \in \Theta, \partial_\theta l(\theta) = 0\}$

2.2 IMCEM algorithm

One important assumption is needed in order to use the deterministic mapping T to prove the convergence of the stochastic one P_t. For all $(\theta_t, \delta_1^t, \dots, \delta_N^t) \in \Theta^{N+1}$, $w.p.1$:

$$\lim_t |A \circ P_t(\theta_t, \delta_1^t, \dots, \delta_N^t) - A \circ T(\theta_t, \delta_1^t, \dots, \delta_N^t)| = 0$$

Theorem 2.2: IMCEM convergence theorem

With certain assumptions of compactness and continuity and a sequence of parameter estimates $\{\theta_t\}$ satisfying the IMCEM algorithm:

1. if $A(\mathcal{L} \cap \Theta, \dots, \mathcal{L} \cap \Theta)$ has an empty interior, then the accumulation points of $\{\theta_t\}_{t \geq 0}$ are in $\mathcal{L}_{EM}$

3 Incremental Algorithms in Practice

3.1 IMCEM on a simple Gaussian case

Let's consider the case when all the variables of interest are Gaussian.

$$Y_i = Z_i + \epsilon_i$$

Where $Z_i \sim \mathcal{N}(\theta, \omega^2)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Since the Z_i and ϵ_i are i.i.d we have that $Y_i \sim \mathcal{N}(\theta, \sigma^2 + \omega^2)$ and $Y_i|Z_i \sim \mathcal{N}(Z_i, \sigma^2)$.
The goal is to find an estimate of the mean θ that maximizes the likelihood $p(y, \theta)$ considering that σ^2 and ϵ^2 are known.
We can now apply our maximization step:

$$\theta_{N+j} = \hat{\Theta}(S) = \frac{\alpha}{N} \sum_{i=1}^N \theta_{N+j-i} + (1-\alpha)\bar{y} + \bar{e}_{N+j}$$

Where $\bar{e} \sim \mathcal{N}(0, \frac{\gamma^2}{M_{N+j}N})$

If we define the vector of parameter as follow (with $k = N + j$):

$$\theta_k = \begin{pmatrix} \theta_k \\ .. \\ \theta_{k-N+1} \end{pmatrix} = \rho\theta_{k-1} + (1-\alpha)\bar{y}e_1 + \bar{e}_k e_1 \text{ where } \rho = \begin{pmatrix} \frac{\alpha}{N} & .. & .. & \frac{\alpha}{N} \\ 1 & 0 & .. & 0 \\ 0 & 1 & .. & 0 \\ .. & .. & .. & .. \\ .. & .. & .. & 0 \end{pmatrix}$$

Now if we consider a scheme where not only one individual is picked at each iteration but a batch pN (where p is a percentage). In that case we can write in scalar (to facilitate the notation we'll consider M=1 and $\bar{y} = 0$):

$$\theta_k = \rho^{1/p}\theta_{k-1/p} + \frac{1-\rho^{1/p}}{1-\rho}\bar{e}_k$$

In that case we can calculate the expectation and the variance of our estimator θ_k in the stationary regime:

$$\mathbb{E}[\theta_k] = \rho^{k/p}\theta_0 \tag{1}$$

$$\text{Var } \theta_k = \frac{\gamma^2}{N(1-\rho)^2} \frac{1-\rho^{1/p}}{1+\rho^{1/p}} \tag{2}$$

With these two expressions we understand what strategy is best for the choice of the batch size at each iteration. Indeed the bias is small when p is small so one should start with picking one individual first to kill the bias and the variance is decreasing when p is increasing. So once the bias is killed one should increase the size of the batch to kill the variance of the estimator.

This result implies as well that the Online EM algorithm introduced by (Cappe and Moulines, 2007) is the best strategy to follow even when all the data is initially available. In other words, even though one has access to the whole observed dataset, one should consider increasing batch of individuals at each iteration.

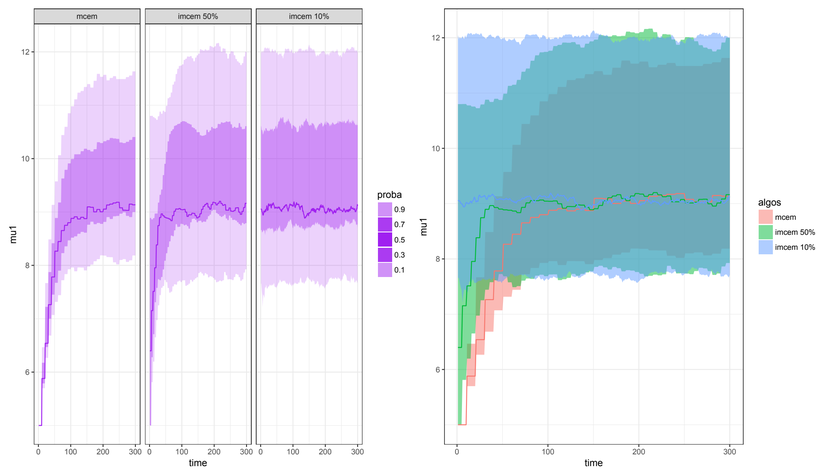

Figure 1: Incremental MCEM for 50 replicates

3.2 ISAEM on a PK-PD dataset

We use an example used by P. Girard and F. Mentre for the symposium dedicated to Comparison of Algorithms Using Simulated Data Sets and Blind Analysis, that took place in Lyon, France, September 2004.
The dataset contains 100 individuals, each receiving 3 different doses:$(0, 10, 90)$, $(5, 25, 65)$ or $(0, 20, 30)$. It was assumed that doses were given in a cross-over study with sufficient wash out period to avoid carry over. Responses y_{ij} were simulated with the following pharmacodynamic model:

$$y_{ij} = E0_i + D_{ij}Emax_i/(D_{ij} + EC50_i) + \epsilon_{ij} \tag{3}$$

Where D_{ij} is the dose given to individual i at time t_j and the individual parameters $(E0_i, Emax_i, EC50_i)$ follow log normal distribution.

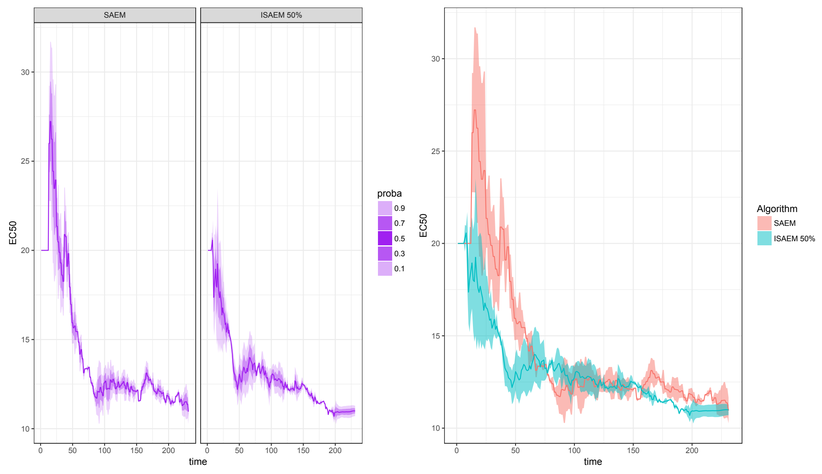

Figure 2: Incremental SAEM for 50 replicates

4 Conclusion

We've shown convergence of incremental versions of the EM and stochastic EM algorithms but incremental variants also imply choice of the individuals at each iteration and the size of the number of individuals considered.
Even though Neal and Hinton showed that incremental EM was optimally fast when only one observation was considered at each iteration, stochastic versions behave differently. Current work is focusing on this aspect.

References

(A. Gunawardana(2005)) W. Byrne A. Gunawardana. Convergence theorems for Generalized Alternating Minimization procedures. (English). 2005.

(B. Delyon and Moulines(1999)) M. Lavielle B. Delyon and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. (English). 1999.

(E.Moulines(2007)) O. Cappe E.Moulines. Online EM Algorithm for Latent Data Models. (English). 2007.

(I. Csiszar(1983)) G. Tusnady I. Csiszar. Information Geometry and Alternating Minimization Procedures. (English). 1983.

(Meng and Rubin(1993)) X.L. Meng and D.B. Rubin. Maximum likelihood estimation via the ECM algorithm: a general framework. (English). 1993.

(M.I Jordan(1995)) L. Xu M.I Jordan. On Convergence Properties of the EM Algorithm for Gaussian Mixtures. (English). 1995.

(N. Roux and Bach(2015)) M. Schmidt N. Roux and F. Bach. A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets. (English). 2015.

(R. Neal(2007)) G.Hinton R. Neal. A view of the EM algorithm that justifies incremental, sparse, and other variants. (English). 2007.

(WU(1983)) C. WU. On the convergence properties of the EM algorithm. (English). 1983.