——— 22nd Annual Conference of the GfKl ———

A SEQUENTIAL MODIFICATION OF EM ALGORITHM

J. Grim

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
P.O, Box 18, CZ - 18208 Prague 8, Czech Republic

Abstract: In the framework of estimating finite mixture distributions we consider a sequential learning scheme which is equivalent to the EM algorithm in case of a repeatedly applied finite set of observations. A typical feature of the sequential version of the EM algorithm is a periodical substitution of the estimated parameters. The different computational aspects of the considered scheme are illustrated by means of artificial data randomly generated from a multivariate Bernoulli distribution.

1 Introduction

The problems of sequential estimating finite mixture distributions arise routinely in the fields of pattern recognition and signal detection, frequently in the context of unsupervised learning and neural networks. The observations are assumed to be received sequentially, one at a time and the estimates of parameters have to be updated after each observation without storing the observed data. Recently (cf. Grim (1996), Vajda, Grim (1997)) we considered a probabilistic approach to neural networks based on finite mixtures and the EM algorithm. In this case the existence of a sequential version of the EM procedure is an important condition of neurophysiological plausibility.

Sequential methods of estimating finite mixtures have been considered by many authors (cf. Titterington et al. (1985), Chapter 6 for a detailed discussion). However, in most cases the problem is not formulated in full generality and/or the solution is computationally intractable for multivariate mixtures. Also the methods are usually related to stochastic approximation techniques and therefore the important monotonic property of the EM algorithm is lost.

In the present paper we consider a sequential scheme which is equivalent to the EM algorithm in case of a repeatedly applied finite set of observations. As the equivalence does not apply for non-periodical sequences of data we use the term pseudo-sequential EM algorithm. A typical feature of this scheme is a periodical substitution of the estimated parameters. The

[0] An early version of the paper: Grim J., "A sequential modification of EM algorithm". In Studies in Classification, Data Analysis and Knowledge Organization, Gaul W., Locarek-Junge H., (Eds.), pp. 163-170, Springer, 1999.

updated parameters are not substituted into the estimated mixture immediately after each observation but only periodically after the last data vector of the training set. The considered pseudo-sequential procedure suggests some possibilities to speed up the EM algorithm and simultaneously, there is a natural possibility to extend the pseudo-sequential scheme to infinite sequences of observations. Different computational aspects of the present paper are illustrated by means of artificial multivariate binary data.

2 EM algorithm

Let $\boldsymbol{x} = (x_1, \cdots, x_N)$ be a vector of binary variables $\boldsymbol{x} \in \{0,1\}^N$ and $P(\boldsymbol{x})$ be a finite mixture of multivariate Bernoulli distributions

$$P(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} f(m) F(\boldsymbol{x}|m), \quad F(\boldsymbol{x}|m) = \prod_{n \in \mathcal{N}} f_n(x_n|m), \tag{1}$$

$$f_n(x_n|m) = \theta_{nm}^{x_n}(1-\theta_{nm})^{1-x_n}, \quad \sum_{m \in \mathcal{M}} f(m) = 1,$$

$$\mathcal{M} = \{1,2,\ldots,M\}, \quad \mathcal{N} = \{1,2,\ldots,N\}.$$

Here $f(m) \geq 0$ is the a priori weight of the m -th component and $f_n(x_n|m)$ are the related discrete distributions of the binary random variables.

The EM algorithm can be used to compute maximum-likelihood estimates of the involved parameters (cf. Dempster et al. (1977), Grim, (1982)). We assume that there is a set $\mathcal{S}$ of independent observations of a binary random vector

$$\mathcal{S} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K\}, \quad \boldsymbol{x}_k \in \{0,1\}^N \tag{2}$$

with some unknown distribution of the form (1). The corresponding log-likelihood function

$$L_{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{x \in X} \log[\sum_{m=1}^{M} f(m) F(\boldsymbol{x}|m)] \tag{3}$$

can be maximized with respect to the unknown parameters by means of the following EM iteration equations:

$$f(m|\boldsymbol{x}) = \frac{f(m)F(\boldsymbol{x}|m)}{\sum_{j \in \mathcal{M}} f(j) F(\boldsymbol{x}|j)}, \quad m \in \mathcal{M}, \quad \boldsymbol{x} \in \mathcal{S}, \tag{4}$$

$$f^{'}(m) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} f(m|\boldsymbol{x}), \quad m \in \mathcal{M}, \tag{5}$$

$$f_n^{'}(\xi|m) = \frac{1}{\sum_{x \in \mathcal{S}} f(m|\boldsymbol{x})} \sum_{x \in \mathcal{S}} \delta(\xi, x_n) f(m|\boldsymbol{x}), \quad \xi \in \{0,1\}, \quad n \in \mathcal{N}. \tag{6}$$

Here $f^{'}, f_n^{'}$ are the new values of parameters and $\delta(\xi, x_n)$ denotes the delta-function. The EM algorithm produces a non-decreasing sequence of values of the log-likelihood function converging to a local or global maximum of $L_{\mathcal{S}}$. The proof of convergence properties is largely based on the following inequality first proved by Schlesinger (1968) for successive values $L_{\mathcal{S}}, L_{\mathcal{S}}^{'}$:

$$L_{\mathcal{S}}^{'} - L_{\mathcal{S}} = \sum_{m\in\mathcal{M}} f^{'}(m) \log \frac{f^{'}(m)}{f(m)} + \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} \sum_{m\in\mathcal{M}} f(m|\boldsymbol{x}) \log \frac{f(m|\boldsymbol{x})}{f^{'}(m|\boldsymbol{x})} +$$

$$+ \sum_{m\in\mathcal{M}} f^{'}(m) \sum_{n\in\mathcal{N}} \sum_{\xi\in\{0,1\}} f_n^{'}(\xi|m) \log \frac{f_n^{'}(\xi|m)}{f_n(\xi|m)} \geq 0. \tag{7}$$

For a detailed discussion of different aspects of convergence see e.g. Dempster et al. (1977), Grim (1982), Wu (1983), Titterington et al. (1985).

In the following sections we illustrate different computational aspects of the considered procedures by means of artificial data randomly generated from a 16-dimensional Bernoulli distribution. The parameters of the source mixture having three components were chosen randomly from suitably defined intervals (cf. Grim (1983)). In order to avoid any small sample effects we used a sufficiently large data set of 10000 binary vectors. The lower five-tuple of curves on Fig. 1 shows typical convergence curves of the EM algorithm starting form five different randomly chosen points. For the sake of comparison the same sets of initial parameters have been used in all computational experiments.

3 Pseudo-sequential EM algorithm

We create an infinite data sequence by repeating the finite set (2):

$$\{\boldsymbol{x}^{(t)}\}_{t=0}^{\infty}, \qquad \boldsymbol{x}^{(t)} = \boldsymbol{x}_k \in \mathcal{S}, \quad k = (t \; mod \; K) + 1. \tag{8}$$

It is easily verified that the EM algorithm (4) - (6) can be equivalently rewritten as follows:

$$f(m|\boldsymbol{x}^{(t)}) = \frac{f(m)F(\boldsymbol{x}^{(t)}|m)}{\sum_{j\in\mathcal{M}} f(j)F(\boldsymbol{x}^{(t)}|j)}, \quad m \in \mathcal{M}, \quad t = 0, 1, 2, \ldots \tag{9}$$

$$f^{(t+1)}(m) = (1 - \frac{1}{k})f^{(t)}(m) + \frac{1}{k}f(m|\boldsymbol{x}^{(t)}), \qquad f^{(0)}(m) = f_n^{(0)}(\xi|m) = 0, \tag{10}$$

$$f_n^{(t+1)}(\xi|m) = (1 - \frac{f(m|\boldsymbol{x}^{(t)})}{kf^{(t+1)}(m)})f_n^{(t)}(\xi|m) + \delta(\xi, x_n^{(t)})\frac{f(m|\boldsymbol{x}^{(t)})}{kf^{(t+1)}(m)}, \tag{11}$$

$$f^{'}(m) = f^{(K)}(m), \quad f_n^{'}(\xi|m) = f_n^{(K)}(\xi|m), \quad \xi \in \{0,1\}. \tag{12}$$

As expressed by Eqs. (12) the updated parameters $f^{(t)}(m), f_n^{(t)}(\xi|m)$ are not substituted into $f^{'}(m), f_n^{'}(\xi|m)$ immediately after each observation but only periodically, at the end of each cycle, i.e. for $t = K$.

It should be emphasized that the EM algorithm and its sequential version (9) - (12) are equivalent in the sense that they produce identical sequences of parameters for identical initial values. Obviously, the important monotonic property (7) and all the well known convergence properties of the EM algorithm remain valid for the sequential scheme (9) - (12). Nevertheless, we use the term "pseudo-sequential" because the equivalence does not apply to data sequences which are not periodical. Fig. 2 illustrates the non-monotonic behavior of the pseudo-sequential EM algorithm when it is applied to a non-periodical sequence of data. In Sec. 6 we suggest a truly-sequential version of the EM algorithm but the justification is only heuristical.

Note that the initial values $f^{(0)}(m)$, $f_n^{(0)}(\xi|m)$ are irrelevant since for t being a multiple of K the first term on the right-hand side of Eqs. (10), (11) is zero. Let us recall also that the sequential procedure is invariant with respect to the order of data vectors between substitutions (cf. (5),(6)).

Remark. The periodical substitution of parameters can be interpreted from a neurophysiological point of view. It is generally assumed that the adaptivity of neurons is based on some relatively slow biochemical processes. For this reason the functional properties of neurons cannot be expected to change continuously, as an immediate consequence of a specific activity of neurons. We can rather assume that the functioning of neural network specifically influences e.g. the concentration of some chemical stuffs or energetic balance of neurons and, in this way, some metabolical changes or growth processes responsible for adaptation can be initialized. Consequently, some delay would occur between a specific activity of a neuron and its adaptive changes. In this sense, periodical substitutions could correspond to sleep phases or to daily cycles. The invariance of adaptive changes with respect to data ordering is also a relevant argument for the present interpretation.

4 Truncated iteration cycle

Motivated by the sequential EM algorithm (9) - (12) we consider first some possibilities to speed up the convergence of the EM algorithm. As it can be assumed in case of randomly chosen starting points, the estimated parameters usually change substantially at initial phases of computation. In other words, at initial iterations the computed estimates are "handicapped" by their previous inaccurate values influencing the weights (9). For obvious reasons this handicap cannot be fully removed by using larger data set but, on the other hand, it could be possible to save computing time by using only some smaller portion of data when computing the initial "rough" estimates.

Proceeding along this line we assume a partition of the sequence $\mathcal{S}$ in two parts $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ and define the parameters $f^{[i]}(m)$, $f_n^{[i]}(\xi|m)$ in analogy with (5),(6) for $\mathcal{S} \approx \mathcal{S}_i,\ i = 1, 2$. We are interested to guarantee the monotonic condition (7) by the parameters $f^{[1]}(m),\ f_n^{[1]}(\xi|m)$ based on a subset $\mathcal{S}_1 \subset \mathcal{S}$

since, in this way, we could save computing time. We can write (cf. (7))

$$L_{\mathcal{S}}^{[1]} - L_{\mathcal{S}} \geq \frac{|\mathcal{S}_1|}{|\mathcal{S}|} \sum_{m \in \mathcal{M}} f^{[1]}(m) \log \frac{f^{[1]}(m)}{f(m)} + \frac{|\mathcal{S}_2|}{|\mathcal{S}|} \sum_{m \in \mathcal{M}} f^{[2]}(m) \log \frac{f^{[1]}(m)}{f(m)} +$$

$$+ \sum_{n \in \mathcal{N}} \sum_{m \in \mathcal{M}} \{ \frac{|\mathcal{S}_1| f^{[1]}(m)}{|\mathcal{S}| f'(m)} \sum_{\xi \in \{0,1\}} f_n^{[1]}(\xi|m) \log \frac{f_n^{[1]}(\xi|m)}{f_n(\xi|m)} + \tag{13}$$

$$+ \frac{|\mathcal{S}_2| f^{[2]}(m)}{|\mathcal{S}| f'(m)} \sum_{\xi \in \{0,1\}} f_n^{[2]}(\xi|m) \log \frac{f_n^{[1]}(\xi|m)}{f_n(\xi|m)} \}$$

and further, using notation

$$I(f^{'}(\cdot), f(\cdot)) = \sum_{m \in \mathcal{M}} f^{'}(m) \log \frac{f^{'}(m)}{f(m)}, \tag{14}$$

we can write the inequality

$$L_{\mathcal{S}}^{[1]} - L_{\mathcal{S}} \geq \frac{|\mathcal{S}_1|}{|\mathcal{S}|} \; I(f^{[1]}(\cdot), f(\cdot)) + \frac{|\mathcal{S}_2|}{|\mathcal{S}|} \min_{m \in \mathcal{M}} \{ \log \frac{f^{[1]}(m)}{f(m)} \} +$$

$$+ \sum_{n \in \mathcal{N}} \sum_{m \in \mathcal{M}} \{ \frac{|\mathcal{S}_1|}{|\mathcal{S}_1| + |\mathcal{S}_2| / f^{[1]}(m)} \; I(f_n^{[1]}(\cdot|m), f_n(\cdot|m)) + \tag{15}$$

$$+ \frac{|\mathcal{S}_2|}{|\mathcal{S}_2| + |\mathcal{S}_1| * f^{[1]}(m)} \min_{\xi \in \{0,1\}} \{ \log \frac{f_n^{[1]}(\xi|m)}{f_n(\xi|m)} \} \}.$$

If the right hand side of the last inequality is positive then the increase of the criterion L_S is guaranteed without including the remaining data $\mathcal{S}_2$ into computation. This condition can be used to choose the minimum necessary size of the sequence $\mathcal{S}_1$ since all the involved quantities are available at each step of the sequential process. As a result we would obtain so called generalized EM algorithm (cf. Dempster et al. (1977)) with similar properties. Unfortunately, the lower bound obtained in (15) is probably to rough, since in our numerical experiments we achieved only small savings in initial phases of computation.

5 Intermediate updating of parameters

Another way to speed up the convergence of the EM algorithm is to utilize the information accumulating sequentially in the parameters (10), (11) before the end of the substitution period. In particular, we assume a finite partition

$$\mathcal{S} = \bigcup_{i=1}^{I} \mathcal{S}_i, \quad N_i = \sum_{l=1}^{i} |\mathcal{S}_l|, \quad (N_I = |\mathcal{S}|) \tag{16}$$

and define the intermediate updates of the estimated parameters

$$f^{[l-1]}(m|\boldsymbol{x}) = \frac{f^{[l-1]}(m)F^{[l-1]}(\boldsymbol{x}|m)}{\sum_{j\in\mathcal{M}} f^{[l-1]}(j)F^{[l-1]}(\boldsymbol{x}|j)}, \quad \boldsymbol{x}\in\mathcal{S}, \quad m\in\mathcal{M}, \tag{17}$$

$$f^{[i]}(m) = \frac{1}{N_i}\sum_{l=1}^{i}\sum_{x\in\mathcal{S}_l} f^{[l-1]}(m|\boldsymbol{x}), \quad i=1,\ldots,I, \quad f^{[0]}(\cdot)\approx f(\cdot), \tag{18}$$

$$f_n^{[i]}(\xi|m) = \frac{1}{N_i f^{[i]}(m)}\sum_{l=1}^{i}\sum_{x\in\mathcal{S}_l}\delta(\xi,x_n)f^{[l-1]}(m|\boldsymbol{x}), \quad \xi\in\{0,1\}. \tag{19}$$

At the end of the substitution period ($i = I$) the estimated parameters are given by

$$\tilde{f}(m) = f^{[I]}(m), \;\; \tilde{f}_n(\xi|m) = f_n^{[I]}(\xi|m), \;\; \xi\in\{0,1\}, \;\; m\in\mathcal{M}, \;\; n\in\mathcal{N}. \tag{20}$$

Eqs. (17) - (20) correspond to one iteration of the EM algorithm but they are not equivalent to the original Eqs. (4) - (6). The increment of the log-likelihood function corresponding to the Eqs. (17) - (20) can be expressed in the form

$$\tilde{L}_{\mathcal{S}} - L_{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{i=1}^{I}\sum_{x\in\mathcal{S}_i}\sum_{m\in\mathcal{M}} f^{[i-1]}(m|\boldsymbol{x})\log\frac{f(m|\boldsymbol{x})}{\tilde{f}(m|\boldsymbol{x})}+$$

$$+\sum_{m\in\mathcal{M}}\tilde{f}(m)\log\frac{\tilde{f}(m)}{f(m)} + \sum_{m\in\mathcal{M}}\tilde{f}(m)\sum_{n\in\mathcal{N}}\sum_{\xi\in\{0,1\}}\tilde{f}_n(\xi|m)\log\frac{\tilde{f}_n(\xi|m)}{f_n(\xi|m)}. \tag{21}$$

Generally, expression (21) may be negative because of the first sum, but we had to use a very small data set ($|\mathcal{S}| \approx 10^2$) to demonstrate the non-monotonic behavior of the sequential procedure (17) - (20). In computational experiments the intermediately updated parameters (20) essentially improved the initial iterations.

It appears that a fixed partition (16) increases the increments of initial iterations but disturbs the final convergence. In accordance with this idea we obtained the best results by making the partition (16) coarser after each substitution and by using non-partitioned set $\mathcal{S}$ in the final stages of computation. The convergence curves obtained for $|\mathcal{S}_{i+1}| = |\mathcal{S}_i| + 500i$ are shown on Fig.1 (upper five-tuple of curves).

Let us recall that by intermediate updating of parameters we obtain a procedure which is not more equivalent to the EM algorithm and therefore the basic convergence properties are not guaranteed. Nevertheless, we can treat the initial computation as a heuristical improvement of starting values, as long as it holds $|\mathcal{S}_i| \leq |\mathcal{S}|$. Further iterations using the non-partitioned set $\mathcal{S}$ correspond to the standard EM algorithm again.

6 Concluding remarks

It can be seen that Eqs. (17) - (20) represent a truly-sequential procedure for an infinitely large set $\mathcal{S}$ and for $I \to \infty$. However, the corresponding computational experiments have shown a relatively slow convergence (cf. Fig.3). The value of the criterion is given by the formula

$$L^{[i]} = \frac{1}{N_i} \sum_{l=1}^{i} \sum_{x \in \mathcal{S}_l} \log[\sum_{m=1}^{M} f^{[l-1]}(m) F^{[l-1]}(\boldsymbol{x}|m)], \tag{22}$$

$$N_i = \sum_{l=1}^{i} |\mathcal{S}_l|, \quad i = 1, \dots, I.$$

and the iteration steps are recomputed to the multiples of 10000 in analogy with Fig. 2, though an exact comparison with periodical sequences is not possible.

Let us recall also (cf. Sec. 5) that, in general, the convergence properties of the truly-sequential procedure are not guaranteed.

References

DEMPSTER, A.P., LAIRD, N.M. and D.B. Rubin (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statist. Society,* **B 39***, 1-38.*

GRIM, J. (1982): On numerical evaluation of maximum–likelihood estimates for finite mixtures of distributions. *Kybernetika, (18), 3, 173–190*

GRIM, J. (1983): Application of finite mixtures to multivariate statistical pattern recognition. *Proc. DIANA Conference on Discriminant Analysis*, pp. 153–166, MÚČSAV Praha 1983

GRIM, J. (1996): Maximum likelihood design of layered neural networks. In: Proceedings of the 13th International Conference on Pattern Recognition **IV** (pp. 85-89). IEEE Computer Society Press, Los Alamitos.

VAJDA, I., GRIM, J. (1997): About the maximum information and maximum likelihood principles. (to appear in) *Kybernetika.*

SCHLESINGER, M.I. (1968): Relation between learning and self-learning in pattern recognition. (in Russian) *Kibernetika, (Kiev), 2, 81-88.*

TITTERINGTON, D.M., SMITH, A.F.M., & MAKOV, U.E. (1985): Statistical analysis of finite mixture distributions. John Wiley & Sons, New York.

WU, C.F.J. (1983): On the convergence properties of the EM algorithm. *Annals of Statistics, (11), 95 - 103.*

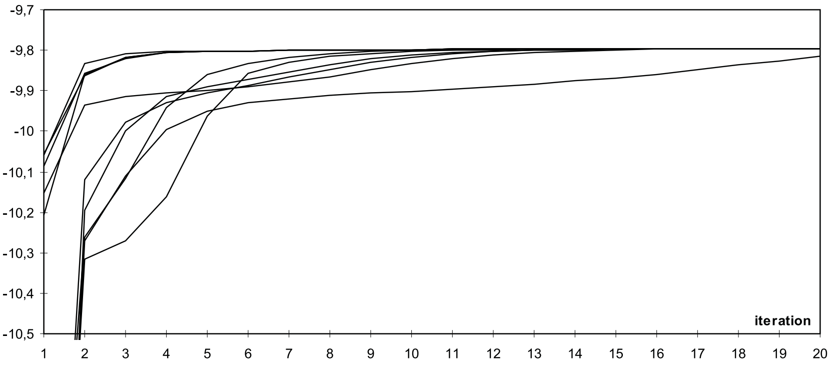

Fig 1. Convergence of EM algorithm from five different starting points compared with an accelerated modification (upper five-tuple of curves, cf. Sec.5).

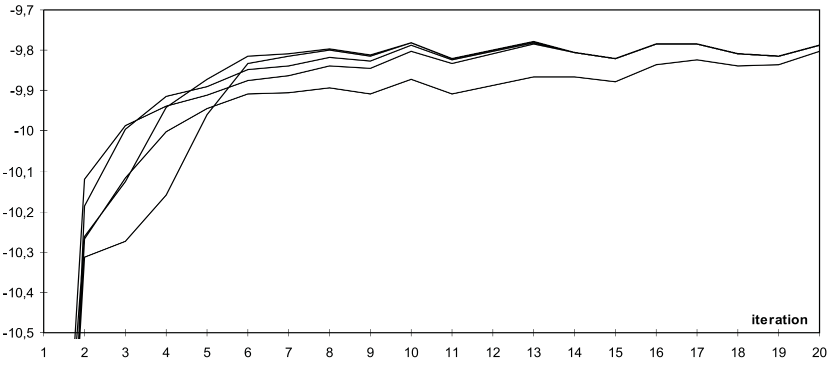

Fig 2. EM algorithm applied to infinite sequence of data sets ($|\mathcal{S}_i| = 10000$).

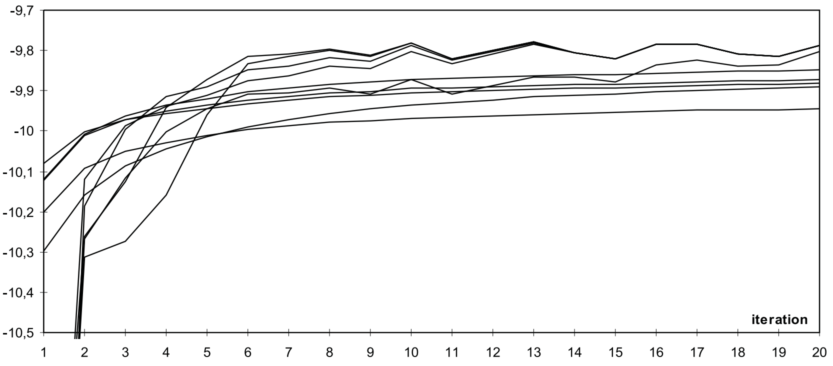

Fig 3. A sequential modification of EM algorithm (cf. Sec.6) applied to infinite sequence of data sets (comparison with Fig.2).