RECURSIVE EM ALGORITHM WITH APPLICATIONS TO DOA ESTIMATION

Olivier Cappé, Maurice Charbit and Eric Moulines

CNRS LTCI & GET / Télécom Paris,
46 rue Barrault, 75634 Paris cedex 13, France

ABSTRACT

We propose a new recursive EM (REM) algorithm that can be used whenever the complete-data model associated to the observed data belongs to an exponential family of distributions. The main characteristic of our approach is to use a stochastic approximation algorithm to approximate the conditional expectation of the complete-data sufficient statistic rather than the unknown parameter itself. Compared to existing approaches, the new algorithm requires no analytical gradient or Hessian computation, it deals with parameter constraints straightforwardly and the resulting estimate can be shown to be Fisher-efficient in general settings. This approach is illustrated on the classic direction of arrival (DOA) model.

1. INTRODUCTION

The EM algorithm [1] is a very popular tool for maximum-likelihood (or maximum a posteriori) estimation. The common strand to problems where this approach is applicable is a notion of *incomplete-data*, which includes the conventional sense of missing data but is much broader than that. The EM algorithm demonstrates its strength in situations where some hypothetical experiment yields "complete" data that are related to the parameters more conveniently than the measurements are. The EM algorithm has several appealing properties. Because it relies on complete-data computations, it is generally simple to implement; at each iteration, the E-step only involves computing conditional expectation given the observed date; the M-step only involves complete-data maximum-likelihood estimation, which is most often in simple closed-form. Moreover, it is numerically stable, in the sense that it each iteration of the algorithm increases the likelihood (of the observed data).

For large sample sizes however, the EM algorithm becomes time and memory consuming since each iteration involves all the available observations. To overcome this limitation, it is of interest to consider recursive implementations of the EM algorithm. By "recursive" we mean that each observation is only used once and that the required computations can be carried out sequentially. Whereas the standard EM algorithm is suitable only for batch or off-line processing, recursive versions of the algorithm are suitable for on-line processing.

Although several recursive EM implementations have been proposed in the literature [2, 3, 4, 5], we feel that most of them are more related to the principle known as Fisher scoring in the statistical literature than to the EM algorithm directly – see (7) and discussion below. These algorithms involve computing the gradient of the log-likelihood –which is readily available due to Fisher formula [1]– but also some form of approximation of the observed-data Fisher information matrix. With those gradient-based algorithms, it is hard to deal with the parameter constraints in some models and setting the scale of the gradient step-size usually is a non-trivial task.

In this communication, we propose a new recursive EM algorithm that is clearly more reminiscent of the EM algorithm and, in many cases, easier to implement than previously mentioned algorithms, while alleviating some of the problems discussed above. This algorithm may be used whenever the complete-data model belongs to a (curved) exponential family. Section 2 covers the algorithm in the general setting of independent observations and we next consider consider its application to DOA estimation in Section 3.

2. RECURSIVE EM (REM)

Let $\{Y_1, \cdots, Y_N\}$ be a sequence of i.i.d. random variables whose common probability density function (pdf), with respect to some measure μ on $\mathbb{R}^{n_y}$, is denoted by $\pi(y)$ and $\{g(y;\vartheta); \vartheta \in \Theta\}$ a parametric family of pdfs. The maximum likelihood estimator (MLE) is given by

$$\hat{\vartheta}_{\mathrm{ML}}(Y_1, \cdots, Y_N) = \arg\max_{\vartheta\in\Theta} N^{-1} \sum_{n=1}^{N} \log g(Y_n;\vartheta) \ . \quad (1)$$

Under standard regularity assumptions, $\hat{\vartheta}(Y_1, \cdots, Y_N)$ converges, as N goes to infinity, to the value

$$\vartheta^\star = \arg\min_{\vartheta\in\Theta} \mathrm{K}\left(\pi \,\|g(\cdot,\vartheta)\right) \ ,$$

where $\mathrm{K}\left(p \,\|q\right) = -\int \log \frac{q(y)}{p(y)} \, p(y)\mu(dy)$ is the Kullback-Leibler divergence between p and q.

In the standard EM approach, we introduce a family of pdfs $\{f(y,z;\vartheta)\, , \vartheta \in \Theta\}$,

$$g(y;\vartheta) = \int f(y,z;\vartheta)\lambda(dz) \ ,$$

where λ denotes a measure on $\mathbb{R}^{n_z}$. The pdfs $f(\cdot;\vartheta)$ and $g(\cdot;\vartheta)$ are, respectively, referred to as the complete-data and observed (or incomplete) likelihood and z is interpreted as unobservable or missing data. The EM algorithm is an iterative optimization algorithm to compute the MLE. Each iteration consists of two successive steps, known as the E-step and the M-step. In the E-step, one evaluates the conditional expectation

$$Q(\vartheta;\vartheta_p) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\left[\log f(Y_n, Z_n;\vartheta) \,|\, Y_n;\vartheta_p\right] \ , \tag{2}$$

where ϑ_p is the current fit for the parameter ϑ. In the M-step, the value of ϑ maximizing $Q(\vartheta;\vartheta_p)$ is found, yielding the new parameter estimate ϑ_{p+1}. The essence of the EM algorithm is that increasing $Q(\vartheta;\vartheta_p)$ forces an increase of the likelihood [1].

In the sequel, we assume that the complete-data model belongs to a curved exponential family:

$$\log f(y,z;\vartheta) = h(y,z) - \psi(\vartheta) + \langle S(y,z), \phi(\vartheta)\rangle \ , \tag{3}$$

where the symbol $\langle\cdot,\cdot\rangle$ denotes the scalar product. The EM re-estimation functional $Q(\vartheta;\vartheta')$ may then be expressed as

$$Q(\vartheta;\vartheta') = L\left(N^{-1}\sum_{i=1}^{N}\bar{s}(Y_i;\vartheta');\vartheta\right) \ , \tag{4}$$

where $L(s;\vartheta) = -\psi(\vartheta) + \langle s, \phi(\vartheta)\rangle$ and

$$\bar{s}(y;\vartheta) \overset{\text{def}}{=} \mathrm{E}\left[S(Y,Z)\,|\,Y=y;\vartheta\right] \ . \tag{5}$$

The k-th iteration of the EM algorithm updates ϑ_k according to

$$\vartheta_{k+1} = \bar{\theta}\left(N^{-1}\sum_{n=1}^{N}\bar{s}(Y_n;\vartheta_k)\right) \ , \tag{6}$$

where $\bar{\theta}(s) = \arg\max_{\vartheta\in\Theta} L(s;\vartheta)$.

In a recursive framework, the data are run through once sequentially and the parameter update must be computable from ϑ_{n-1} and Y_n, where ϑ_n denotes the current value of the parameter estimate after n observations. To our best knowledge, the first recursive parameter estimation procedure for incomplete data model has been proposed by [2]. It is given by:

$$\hat{\vartheta}_n = \hat{\vartheta}_{n-1} + \gamma_n I_f^{-1}(\hat{\vartheta}_{n-1}) U(Y_n;\hat{\vartheta}_{n-1}) \ , \tag{7}$$

where $\{\gamma_n\}$ is a non-increasing sequence of positive numbers, $U(y;\vartheta) = \nabla_\vartheta \log g(y;\vartheta)$ is the score function and $I_f(\vartheta)$ is the Fisher information matrix (FIM) associated to a complete observation. This recursion is recognized as a stochastic approximation procedure on ϑ. It is often referred to in the literature under the name of recursive EM, but we find that this term is somewhat misleading because, contrary to the EM, it is a gradient algorithm. This algorithm may be seen as a recursive implementation of the gradient EM algorithm presented in [6]. The works presented in [4, 5] build on the idea of [2], whereas [3] is actually closer to the present contribution, although limited to a specific model.

Our proposal consists in replacing the E-step by a recursive stochastic approximation step, while keeping the maximization step unchanged, that is

$$\hat{s}_n = \hat{s}_{n-1} + \gamma_n(\bar{s}(Y_n;\hat{\vartheta}_{n-1}) - \hat{s}_{n-1}) \ , \quad \hat{\vartheta}_n = \bar{\theta}\left(\hat{s}_n\right) \ , \tag{8}$$

where γ_n is a sequence of decreasing step-sizes. This new algorithm is fully analyzed in [7] where it is shown that it is Fisher-efficient in rather general settings (not assuming in particular that $\pi = g(\cdot;\vartheta^\star)$ for some parameter value $\vartheta^\star$) for choices of the step-sizes such that $\sum_{n=1}^{\infty}\gamma_n = \infty$ and $\sum_{n=1}^{\infty}\gamma_n^2 < \infty$ (typically, take $\gamma_n = n^{-\alpha}$ with $0.5 < \alpha \leq 1$).

More precisely – see Theorem 5 of [7], we may show that under suitable assumptions and with step-sizes $\gamma_n \equiv n^{-\alpha}$, with $0.5 < \alpha < 1$, $\gamma_n^{-1/2}(\hat{\vartheta}_n - \vartheta^\star)$ converges in distribution to a zero mean Gaussian distribution with covariance matrix $\Sigma(\vartheta^\star)$ solution of the Lyapunov equation involving the FIM of a complete observation and the covariance of the observed data Fisher score. When π belongs to the parametric family of distributions under consideration; that is $\pi = g(\cdot,\vartheta^\star)$, the solution of this Lyapunov equation is the FIM associated to the observations.

$$I_\pi(\vartheta^\star) \overset{\text{def}}{=} -\nabla_\theta^2\,\mathrm{K}\left(g_{\vartheta^\star}\,\|g_\vartheta\right)|_{\vartheta=\vartheta^\star} \ , \tag{9}$$

otherwise it has a more complicated expression (see [7] for details).

This first result is not entirely satisfying as it shows that the rate of convergence is $\gamma_n^{-1/2}$ rather than $n^{-1/2}$, that is, $n^{-\alpha/2}$ with the choice discussed above, which can be much slower. In addition, this result also suggests that using α close to 1, *i.e.*, fast decreasing step-sizes, is the best option to maximize the convergence rate. In practice, one should however remember that this result pertains to the large sample behavior of the estimates and, in most models, taking α close to 1 results in the algorithm converging too slowly. A better solution proposed by [8] and further generalized in [9] consists in using step-sizes with slower decay (typically α closer to 0.5 than to 1) and to perform *averaging*; ϑ is then estimated by

$$\tilde{\vartheta}_n = (n-n_0)^{-1}\sum_{k=n_0+1}^{n}\hat{\vartheta}_n \ ,$$

where n_0 is a lag after which averaging effectively starts. It is proved in [7] that the averaged estimator $\tilde{\vartheta}_n$ is indeed Fisher-efficient, that is, it converges at rate $n^{-1/2}$ to a centered Gaussian distribution with covariance matrix equal to the inverse of the Fisher information matrix defined in (9). The practical implications of these results will be further illustrated below (in Section 4) for the DOA model.

3. APPLICATION TO DOA ESTIMATION

Consider an array with M sensors receiving signals from K far-field narrow-band sources with $M > K$. The measured array output is a linear combination of the incoming waveforms, corrupted by additive Gaussian noise. The vector of array outputs at time index n is represented as

$$Y_n = \mathbf{A}(\boldsymbol{\theta})X_n + B_n = \sum_{k=1}^{K} \mathbf{a}(\theta_k)X_{n,k} + B_n \ ,$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ represents the unknown DOA parameters, $\mathbf{a}(\theta)$ is the complex array response to a unit waveform with incoming angle θ, $X_n = (X_{n,1}, \ldots, X_{n,K})$ is the emitted signal, and B_n the additive noise. We assume that the array is linear and uniform, which implies that, for any angle θ, $\mathbf{a}^H(\theta)\mathbf{a}(\theta) = C$. We further assume that the vector of signal waveforms X_n is a stationary white complex Gaussian noise and that the K sources are independent with powers $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. The additive noise B_n is also a stationary white complex Gaussian noise assumed for simplicity to be spatially white with power υ. The parameter ϑ thus encompasses both the DOA parameters $\boldsymbol{\theta}$, the powers $\boldsymbol{\alpha}$, and the noise variance υ. The likelihood of one observation is $g(y;\vartheta) = \mathcal{N}(0, \boldsymbol{\Gamma}(\vartheta))$, with

$$\boldsymbol{\Gamma}(\vartheta) = \mathbf{A}(\boldsymbol{\theta})\mathbf{P}(\boldsymbol{\alpha})\mathbf{A}^H(\boldsymbol{\theta}) + \upsilon\mathbf{I_M} \ , \tag{10}$$

where $\mathcal{N}$ denotes the complex multivariate Gaussian distribution, $\mathbf{P}(\boldsymbol{\alpha}) = \operatorname{diag}(\alpha_1, \cdots, \alpha_K)$, the superscript H denotes the conjugate-transpose, and $\mathbf{I_M}$ is the M-dimensional identity matrix.

We may represent the array response Y_n as a superposition of K independent complex Gaussian vectors $Z_{n,k}$ with zero mean and covariance

$$\boldsymbol{\Gamma}_{\boldsymbol{k}}(\vartheta) = \alpha_k \mathbf{a}(\theta_k)\mathbf{a}^H(\theta_k) + \upsilon_k \mathbf{I_M} \ , \tag{11}$$

where $\sum_{k=1}^{K} \upsilon_k = \upsilon$; taking, for instance, $\upsilon_k = \upsilon/K$. In the EM terminology, the vectors $Z_{n,1}, \ldots, Z_{n,K}$ form the complete data. The joint pdf of the complete data is given by

$$\begin{aligned}\log(f(z_{n,1}, \cdots, z_{n,K}; \vartheta)) &= -M \log \pi \\ &- \psi(\vartheta) + \sum_{k=1}^{K} \operatorname{trace}\left[S(z_{n,k})\phi_k(\vartheta)\right] \ , \end{aligned} \tag{12}$$

where

$$\begin{aligned}\psi(\vartheta) &= M \sum_{k=1}^{K} \log(\upsilon/K) + \sum_{k=1}^{K} \log(1 + CK\upsilon^{-1}\alpha_k) \ , \\ S(z) &= zz^H \ , \\ \phi_k(\vartheta) &= -K\upsilon^{-1}\mathbf{I_M} + \frac{K^2\upsilon^{-2}\alpha_k}{1 + CK\upsilon^{-1}\alpha_k} \ \mathbf{a}(\theta_k)\mathbf{a}^H(\theta_k) \ . \end{aligned}$$

Thus, in the DOA model the complete-data sufficient statistics correspond to the K empirical covariance matrices $S(z_{n,1})$, $\ldots, S(z_{n,K})$.

We first turn to the E-step of the EM algorithm. Following (5), we need to compute conditional expectation of $S(Z_{n,k})$

$$\bar{s}_k(Y_n;\vartheta) \stackrel{\text{def}}{=} \mathrm{E}\left[Z_{n,k}Z_{n,k}^H \,\middle|\, Y_n; \vartheta \right] \ , \tag{13}$$

for $k = 1, \ldots, K$. It is easy to derive that

$$\begin{aligned}\bar{s}_k(Y;\vartheta) &= \boldsymbol{\Gamma}_{\boldsymbol{k}}(\vartheta) - \boldsymbol{\Gamma}_{\boldsymbol{k}}(\vartheta)\boldsymbol{\Gamma}^{-1}(\vartheta){\boldsymbol{\Gamma}_{\boldsymbol{k}}}^H(\vartheta) \\ &+ \boldsymbol{\Gamma}_{\boldsymbol{k}}(\vartheta)\boldsymbol{\Gamma}^{-1}(\vartheta)\left(YY^H\right)\boldsymbol{\Gamma}^{-1}(\vartheta){\boldsymbol{\Gamma}_{\boldsymbol{k}}}^H(\vartheta) \ , \end{aligned} \tag{14}$$

where $\boldsymbol{\Gamma}(\vartheta)$ and $\boldsymbol{\Gamma}_{\boldsymbol{k}}(\vartheta)$ are respectively given by (10) and (11).

The M-step then consists in maximizing $L(s_1, \ldots, s_K; \vartheta)$ defined by (4). To do so, we note that it is first necessary to maximize separately with respect to θ_k only the function $\mathbf{a}^H(\theta_k)s_k\mathbf{a}(\theta_k)$ to obtain

$$\begin{aligned} m_k &\stackrel{\text{def}}{=} \max_{\theta} \mathbf{a}^H(\theta)s_k\mathbf{a}(\theta) \ , \\ \bar{\theta}_k(s_k) &= \arg\max_{\theta} \mathbf{a}^H(\theta)s_k\mathbf{a}(\theta) \ , \end{aligned} \tag{15}$$

using one dimensional line searches. Now, the maximization with respect to the $(K+1)$ positive parameters $\alpha_1, \cdots, \alpha_K$, and υ yields

$$\bar{\upsilon}(s_1, \ldots, s_K) = \frac{1}{(M-1)} \sum_{k=1}^{K} (\operatorname{trace}(s_k) - m_k/C) \tag{16}$$

and

$$\bar{\alpha}_k(s_1, \ldots, s_K) = \frac{m_k - C\bar{\upsilon}(s_1, \ldots, s_K)/K}{C^2} \tag{17}$$

Following (8), the REM algorithm then consists in approximating the K statistics $\hat{s}_{n,k}$ by

$$\hat{s}_{n,k} = \hat{s}_{n-1,k} + \gamma_n \left(\bar{s}_k(Y_n; \hat{\vartheta}_{n-1}) - \hat{s}_{n-1,k} \right) \ ,$$

where $\bar{s}_k(Y_n; \hat{\vartheta}_{n-1})$ is computed according to (14). Then, $\hat{\vartheta}_n$ is obtained applying (15), (16), and (17), as in the (non-recursive) EM algorithm.

4. NUMERICAL RESULTS

In this section, we study the performance of the proposed algorithm in the scenario considered in [10]: three sources with equal power are located at $\boldsymbol{\theta}^\star = [24°, 28°, 45°]$; the array consists of $M = 15$ sensors with equal inter-spacing of half wavelength; the signal-to-noise ratio for each path is kept at 0 dB. In the results below, we use the same initial parameter guess $\hat{\vartheta}_1$ as in [10].

The sequence of step-sizes is chosen as $\gamma_n = n^{-0.6}$. Note that although, we could obviously select different sequences that behaves similarly for large n, one of the merit of the proposed algorithm is that the absolute scale of γ_n is in some sense fixed by the fact that $\gamma_n = 1$ amounts to taking $\hat{s}_n = \bar{s}(Y_n; \hat{\vartheta}_{n-1})$, i.e, not performing any smoothing on the sufficient statistic.

The performance is estimated by averaging the squared error $\|\hat{\theta}_N - \theta_{\mathrm{true}}\|^2$ over one hundred independent simulated trajectories of array outputs. For reference, we also plot the Cramer-Rao lower bound (CRB) for the DOA model [11].

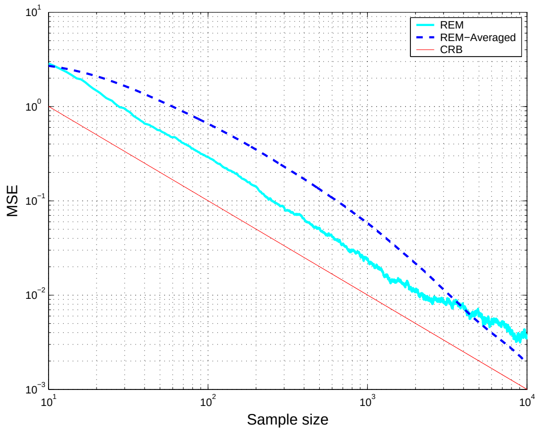

Fig. 1. MSE of REM with and without averaging as a function of the number of samples, compared to the CRB (MSE estimated from 100 independent runs); $n_{\min} = 0$.

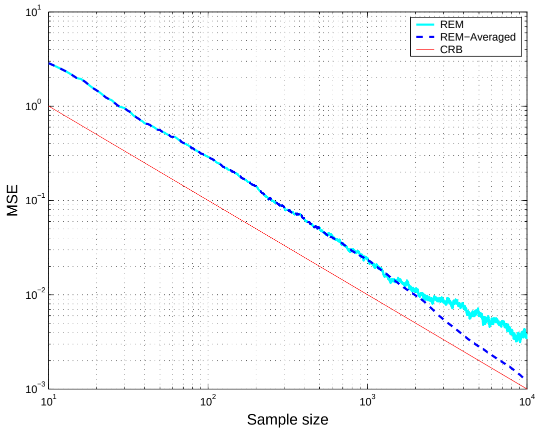

Fig. 2. Same figure with $n_{\min} = 500$.

On figures 1 and 2, we first observe that in the asymptotic regime (sample sizes larger than 2,000), the curve that pertains to the REM algorithm without averaging (solid bold curve) doesn't have the same slope as the CRB, confirming that the estimate converges at a rate which is lower than $n^{-1/2}$. On figure 1, averaging (dotted curve) appears to yield better results only for larger sample sizes but is much worse for small to intermediate sample sizes. This is due to the fact that for small sample sizes, the estimation error is dominated by the bias caused by the mismatch between the initial value and $\vartheta^\star$. In this regime, averaging only worsens the problem and it is recommend to start the averaging process only when the estimate gets reasonably close to the true value. Figure 2 (with $n_0 = 500$) shows that in this case, averaging is very beneficial and that it does reach the CRB for larger sample sizes.

5. REFERENCES

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[2] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. Roy. Statist. Soc. Ser. B*, vol. 46, no. 2, pp. 257–267, 1984.

[3] M. Sato and S. Ishii, "On-line EM algorithm for the normalized gaussian network," *Neural Computation*, 2000.

[4] P-J. Chung and F. B. Böhme, "Recursive EM and SAGE-inspired algorithms with application to DOA estimation," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2664–2677, 2005.

[5] Z. Liu, J. Almhana, V. Choulakian, and R. McGorman, "Online EM algorithm for mixture with application to internet traffic modeling," *Comput. Statist. Data Anal.*, 2000.

[6] K. Lange, "A gradient algorithm locally equivalent to the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, vol. 57, no. 2, pp. 425–437, 1995.

[7] C. Andrieu, O. Cappé, M. Charbit, and E. Moulines, "Recursive EM algorithm for parameter estimation in incomplete data model," preprint, 2005.

[8] B. T. Polyak, "New method of stochastic approximation type," *Automation Remote Contr.*, vol. 7, pp. 937–946, 1991.

[9] H. J. Kushner and J. Yang, "Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes," *SIAM J. Contr. Optim.*, vol. 31, pp. 1045–1062, 1993.

[10] P-J. Chung and F. B. Böhme, "Recursive EM algorithm for stochastic ML DOA estimation," *IEEE ICASSP*, vol. III, pp. 3029–3032, 2002.

[11] P. Stoica and A. Nehorai, "Performance study of conditional and unconditional direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 10, pp. 1783–1795, 1990.