Centre de Recherche

SAMOS - MATISSE

CNRS, UMR 8595

Statistique Appliquée &
Modélisation Stochastique

On recursive estimation in incomplete data models

J. YAO

Prépublication du SAMOS N° 81 ***(Version modifiée)***

Février 1999

On recursive estimation in incomplete data models

Jian-feng YAO SAMOS, Université Paris I

Running title : Recursive estimation from incomplete data

Contact address:

J.YAO, SAMOS, Université Paris I, 90 rue de Tolbiac , 75634 Paris Cedex 13, FRANCE
e-mail: yao@univ-paris1.fr Fax. (+33 1) 40 77 19 22

Abstract

We consider a new recursive algorithm for parameter estimation from an independent incomplete data sequence. The algorithm can be viewed as a recursive version of the well-known EM algorithm, augmented with a Monte-Carlo step which restores the missing data. Based on recent results on stochastic algorithms, we give conditions for the a.s. convergence of the algorithm. Moreover, asymptotical variance of this estimator is reduced by a simple averaging. Application to finite mixtures is given with a simulation experiment.

AMS Classification Code : 62 F 12, 62 L 20

Keywords: Incomplete data; EM algorithm; recursive estimation; mixtures; stochastic algorithm.

1 Introduction

In many statistical models data are observed only upon some deterministic distorsion. Such examples include censored data, mixture models and the deconvolution problem. The most popular parameter estimator for these models is probably the EM estimator found in the seminar paper (Dempster et al., 1977). An up to date report on its various extensions can be found in (Meng and van Dyk, 1997) with an extensive discussion.

Nevertheless on-line parameter estimation for theses models has not been fully addressed. Standard recursive estimation based on the observed likelihood (Fabian, 1978) do not directly apply, mainly because this likelihood is ill-defined for most of incomplete data sequence. Recently, Ryden (Rydén, 1994) develops a projection-based recursive likelihood estimator for the mixture problem. However it is not clear how to extend this approach to other incompletely observed models.

We propose in this paper a new recursive algorithm, called RSEM, which can be viewed as a recursive version of the EM algorithm, augmented with a Monte-Carlo data imputation step which restores the missing data. To state the problem, let us call, following (Dempster et al., 1977), *complete data* some *unobservable* random variable X with domain $\mathcal{X}$. We indeed observe a transformation of this, say $Y = \Phi(X)$, where Φ is a known non-random map from $\mathcal{X}$ into some observation space $\mathcal{Y}$. Typically Φ is a "many-to-one" map, we thus call the observations $\mathcal{Y}$ *incomplete data*, since many information is lost by Φ.

Assume that X (resp. Y) has a density $\pi(x\,;\theta)$ (resp. $g(y\,;\theta)$) with respect to some σ-finite measure dx (resp. dy). These densities are linked by the following

$$g(y\,;\theta) = \int_{\mathcal{X}(y)} \pi(x\,;\theta)\, dx \ ,$$

where $\mathcal{X}(y)$ denotes the set $\{x : \Phi(x) = y\}$. It is worth noting that the conditional distribution of X given $Y = y$, denoted $\mu(X \mid Y = y\,;\,\theta)$, has the density

$$k(x \mid y\,;\,\theta) = \mathbf{1}_{\mathcal{X}(y)}(x)\pi(x\,;\theta)/g(y\,;\,\theta) \ .$$

Let an i.i.d. sequence $Y_1, \ldots, Y_n, \ldots$ of Y be successively observed. For the on-line parameter estimation purpose, we consider the following recursive algorithm, named as RSEM

(for recursive SEM-like algorithm). This algorithm is first introduced in (Duflo, 1996). The sequence (γ_n) below (*gains sequence*) is any sequence of positive numbers which decreases to 0. The parameter space is some open subset Θ of $\mathbb{R}^p$.

Algorithm RSEM :

(i) *Pick an arbitrary initial value* $\theta_0 \in \Theta$.

(ii) *At each time* $n \geq 1$, *perform the following steps with the new observation* Y_{n+1} *:*

R-step (restoration) : *restore the corresponding unobserved complete data* X *by drawing an independent sample* X_{n+1} *from the conditional distribution* $\mu(X \mid Y = Y_{n+1} \; ; \; \theta_n)$.

E-step (Estimation) : *update the estimation by*

$$\theta_{n+1} = \theta_n + \gamma_n \nabla \log \pi(X_{n+1}; \theta_n). \tag{1}$$

In a non recursive framework, the above restoration R-step is the key novelty found in a class of so-called SEM, MCEM or SRE algorithms as introduced in (Celeux and Diebolt, 1985), (Wei and Tanner, 1990) and (Qian and Titterington, 1991) respectively. For a recent account on SEM algorithms, we refer to (Celeux et al., 1996) or (Lavielle and Moulines, 1995). This R-step plays a similar role as the E-step of the EM algorithm : it enables the use of the likelihood π from the complete data for parameter updating.

This paper is devoted to an accurate study of the RSEM algorithm including convergence and asymptotic efficiency. We shall first show in §2 that the RSEM algorithm is a stochastic gradient algorithm which minimizes Kullback-Leibler divergence from the true model. Unfortunately, for most of incomplete data models of interest, it is not clear from the current state of the stochastic algorithms theory (up to our knowledge), whether or not the RSEM algorithm should converge without transformation (see §2). Therefore, we shall (§3) truncate this algorithm at randomly varying bounds in spirit of (Chen et al., 1988; Chen, 1993). This technique is surprisingly effective and yields the a.s. convergence of the truncated algorithm.

Finally in §4, we implement the RSEM algorithm for finite mixtures and report a simulation experiment.

2 Does the RSEM algorithm converge ?

From now on, the true parameter value is denoted θ_*. Let us define the Kullback-Leibler divergence

$$K(\theta) := \int_{\mathcal{Y}} g(y \; ; \; \theta_*) \log \frac{g(y \; ; \; \theta_*)}{g(y \; ; \; \theta)} dy \;\; , \tag{2}$$

and some assumptions on the smoothness of the likelihood function.

Assumptions (S). Θ is an open convex set of $\mathbb{R}^p$. The likelihood $\theta \mapsto \log \pi(x \; ; \; \theta)$ is twice differentiable on Θ such that :

(a). K is twice continuously differentiable on Θ and the identity (2) can be twice differentiated under the integral sign.

(b). For all θ and for all y, we have $\nabla g(y\,;\theta) = \int_{\mathcal{X}(y)} \nabla\pi(x\,;\theta)dx$

These conditions are merely what is necessary for a statistical model to be regular. In particular, the assumption **(S)-(b)** implies the following identity which says that the *incomplete score function* is equal to the average of the unobserved complete score function:

$$\nabla \log g(y\,;\theta) = \int_{\mathcal{X}(y)} \nabla \log \pi(x\,;\theta) k(x \mid y\,;\theta) dx \ . \tag{3}$$

We shall also use the filtration $\mathcal{F} = (\mathcal{F}_n)$, with $\mathcal{F}_0 = \{\Omega, \emptyset\}$, $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n\,; X_1, \ldots, X_n)$ for $n \geq 1$, which is the natural filtration associated to the RSEM algorithm.

Lemma 1 *Under the assumptions* ***(S)-(a)-(b),*** *the RSEM algorithm is a stochastic gradient algorithm as follows*

$$\theta_{n+1} = \theta_n - \gamma_n \left[\nabla K(\theta_n) + \varepsilon_{n+1}\right] \ ,$$

where

$$\varepsilon_{n+1} = -\left\{\nabla \log \pi(X_{n+1}\,;\theta_n) - \mathbb{E}\left[\nabla \log \pi(X_{n+1}\,;\theta_n) \mid \mathcal{F}_n\right]\right\},$$

is a $\mathcal{F}$-adapted noise sequence, that is $\mathbb{E}(\varepsilon_{n+1}|\mathcal{F}_n) = 0$ for all n.

Proof. We have by **(S)-(a)-(b)**

$$\begin{aligned} h(\theta) \ &: = \mathbb{E}\left[\nabla \log \pi(X_{n+1}\,;\theta_n = \theta) \mid \mathcal{F}_n\right] = \mathbb{E}\left[\nabla \log \pi(X_1\,;\theta_0 = \theta)\right] \\ &= \int_{\mathcal{Y}} \int_{\mathcal{X}(y)} \nabla \log \pi(x\,;\theta)\, k(x \mid y\,;\theta) g(y\,;\theta_*)\, dx\, dy \\ &= \int_{\mathcal{Y}} \left[\nabla \log g(y\,;\theta)\right]\ g(y\,;\theta_*)\, dy = -\nabla K(\theta) \ . \end{aligned}$$

■

Therefore, the RSEM algorithm is a stochastic gradient algorithm obtained by perturbation of the following gradient system

$$\dot{\theta} = -\nabla K(\theta) \ . \tag{4}$$

The convergence analysis of such an algorithm is usually set up in two steps (see e.g. (Duflo, 1996)). First we attempt to *stabilize* the algorithm, that is to check whether a.s. the sequence (θ_n) would lives in a compact subset of Θ or not. After the algorithm is stabilized, there are many methods to ensure the a.s. convergence of (θ_n) to an point (equilibrium) of $\{\nabla K = 0\}$. Recent results based on the so-called ODE method (Kushner and Clark, 1978) is reported in (Fort and Pagès, 1996).

The gradient system (4) has a natural Lyapounov function $V = K$: for any solution $(\theta_t)_{t\geq 0}$ of the system, the function $t \mapsto K(\theta_t)$ is decreasing. If in addition K is *inf-compact*, i.e.

$$\text{for all } a \in \mathbb{R},\ \{K \leq a\} \text{ is a compact set of } \Theta, \quad (\text{i.e. } \lim_{\theta \to \partial\Theta} K(\theta) = +\infty\,) \tag{5}$$

the whole trajectory $(\theta_t)_{t\geq 0}$ would stay in the compact set $\{K \leq K(\theta_0)\}$. Therefore this inf-compacity of K is a basic tool to stabilize such an algorithm. Another widely used requirement for stabilization of this system is the following Lipshitz condition

$$\nabla K \text{ is Lipshitz on } \Theta. \tag{6}$$

Unfortunately, Kullback-Leibler divergence K from incomplete data models often does not meet both the conditions (5) and (6). In the example below with censorated data, (6) is not fulfilled, while for a finite mixture (cf. §4), (5) is no longer satisfied. Consequently, we do not know whether or not the RSEM algorithm could be stabilized without transformation (actually our simulation experiment in §4 in the mixture case seems to show it would be *not*). Therefore we shall develop in the next section a truncated version of the RSEM algorithm which will converge.

Example with censored data. Let X be a real-valued variable and $Y = \min(X, c)$ with a known constant $c \in \mathbb{R}$. Let also $a_\theta = \mathbb{P}(X < c\,;\,\theta)$ and the reference measures be $dx = 1_{x \neq c} dm(x) + d\delta_c(x)$, $dy = 1_{y<c} dm(y) + d\delta_c(y)$ with the Lebesgue measure m and the Dirac mass δ_c at c. Then

$$g(y\,;\,\theta) = \pi(y\,;\,\theta) 1_{y<c} + (1 - a_\theta) 1_{y=c}\ ,$$

and

$$K(\theta) = \int_{-\infty}^{c} \pi(y\,;\,\theta_*) \log \frac{\pi(y\,;\,\theta_*)}{\pi(y\,;\,\theta)} dm(y) + (1 - a_{\theta_*}) \log \frac{(1 - a_{\theta_*})}{(1 - a_\theta)}\ .$$

Assume the exponential law $\pi(x\,;\,\theta) = \theta e^{-\theta x}$, $x \geq 0$ for X. Thus we find $\Theta = (0, \infty)$, $\partial\Theta = \{0, \infty\}$, $a_{\theta_*} = 1 - e^{-c\theta_*}$ and in final

$$K(\theta) = a_{\theta_*} \left[\frac{\theta}{\theta_*} - 1 - \log \frac{\theta}{\theta_*} \right]\ ,\quad K'(\theta) = a_{\theta_*} \left[\frac{1}{\theta_*} - \frac{1}{\theta} \right]\ ,\quad K''(\theta) = \frac{a_{\theta_*}}{\theta^2}.$$

Therefore for this simple model, the condition (5) is fulfilled, while (6) is not.

3 Stabilization and convergence of the truncated RSEM algorithm

We investigate in this section a truncated version of the RSEM algorithm. This method, called *truncations at randomly varying bounds*, is introduced in (Chen et al., 1988) for the Robbins-Monro algorithm. Although the initial purpose of these authors seems to weak conditions on both the regression function and the error process, such a truncation method proves to be powerful for the stabilization task of a wider class of stochastic algorithms.

An extension of this method for our RSEM algorithm is as follows. Assume we can find in Θ a sequence of increasing compact subsets $C_0 \subset \cdots \subset C_s \subset \cdots$ which tends to its boundary (see Fig. 1)

$$\Theta \subseteq \lim_{s \to \infty} \uparrow C_s \subseteq \overline{\Theta}. \qquad \textbf{(T.a)}$$

Let us fix some point in the first compact $w \in C_0$. The basic idea of truncations is to bring the algorithm (θ_n) back to this fixed point w, every time θ_n will leave some compact C_s. The index s of such a barrier compact C_s is randomly selected, according to the own truncation history of each algorithm path. Moreover, s increases if truncations repeat, so that next barrier compacts go farther.

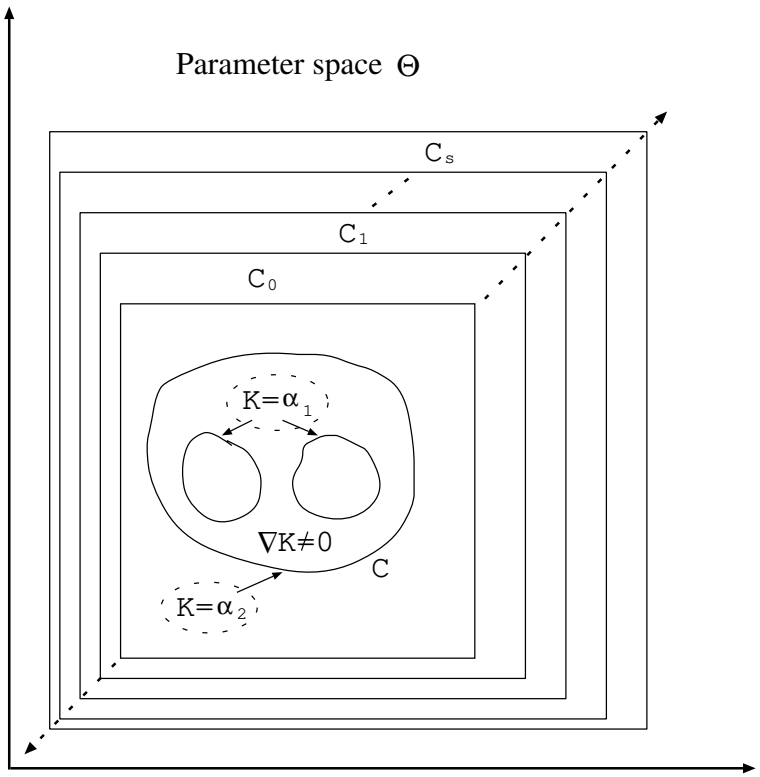

Figure 1: Barrier compacts used for truncations: case of a parameter space $\Theta = (0, \infty)^2$.

More precisely, to truncate the RSEM algorithm, we replace the updating rule (1) by the following

$$\theta_{n+1} = \begin{cases} \theta_n + \gamma_n \nabla \log \pi(X_{n+1};\, \theta_n)\,, & \text{if } \theta_n + \gamma_n \nabla \log \pi(X_{n+1};\, \theta_n) \in C_{\sigma(n)} \\ w\,, & \text{elsewise}\,. \end{cases} \tag{7}$$

Here $\sigma(n)$ stands for the number of truncations up to time $n-1$

$$\sigma(n) = \#\{i : 1 \le i \le n-1 \text{ and } \theta_i + \gamma_i \nabla \log \pi(X_{i+1}; \theta_i) \notin C_{\sigma(i)}\} \text{ for } n \ge 2,$$

with $\sigma(1) = 0$.

Furthermore, the compact sequence (C_s) should be carefully determined from a theoretic point of view. First, we shall assume (see Fig. 1)

C_0 is convex and there are constants $0 < \alpha_1 < \alpha_2$ such that **(T.b)**

(i). $C := \{K \le \alpha_2\}$ is compact and $C \subseteq C_0$.

(ii). $\nabla K \neq 0$ on $\{K > \alpha_1\}$.

It is worth noting that unlike the inf-compacity condition (5) where all level sets $\{K \le a\}$ must be compact, the condition **(T.b)** requires such a compacity only for small values

of a. It also happens by **(T.b),** the set of equilibrium points $\{\nabla K = 0\}$ is a subset of the level set $\{K \leq \alpha_1\}$. Therefore, the first compact C_0 should be large enough to include all target equilibrium points. On the other hand, the existence of the constants α_1 and α_2 is of theoretical importance only: practical implementation is always free from these constants (see §4.3).

Second, we should also select (C_s) in such a way to get the following control on the noise sequence (ε_n):

$$\sum_n \gamma_n \varepsilon_{n+1} \quad \text{converges almost surely .} \tag{**T.c**}$$

We shall provide a general construction of (C_s) fulfilling this requirement in Proposition 1 below, which is postponed at the end of this section.

To state our main theorem, let us fix the gains sequence (γ_n). Although many suitable general gains sequences can be used, here we shall only use the following slow gains sequence

$$\gamma_n = \frac{A}{(n+B)^\alpha}, \quad \text{with} \quad \frac{1}{2} < \alpha < 1 \,, \tag{8}$$

where A, B are some positive constants.

Theorem 1 *Assume that*

(i) the smoothness condition ***(S)*** *holds.*

(ii) the compact sequence (C_s) *fulfills Conditions* ***(T.a-b-c)***.

(iii) the equilibrium points $\{\theta : \nabla K(\theta) = 0\}$ *are isolated.*

Then, with the gains sequence (8), the truncated RSEM algorithm (θ_n)*, defined (7), converges a.s. to a point* θ_∞ *of* $\{\nabla K = 0\}$.

Proof. We proceed in two steps.

Step (1). First we show that the number of truncations is a.s. finite

$$a.s. \text{ there is a time } T \text{ such that for all } n \geq T,\ \sigma(n) \equiv \sigma(T). \tag{9}$$

It follows in particular that the algorithm lives a.s. in some (path-dependent) compact subset of Θ (stabilization). This essentially relies on several extensions of Theorem 3 of (Chen et al., 1988). As stated before, the authors considered the Robbins-Monro algorithm. Also since the state space for this algorithm is the whole Euclidean space $\mathbb{R}^p$, they used for truncations a sequence of discs $C_s = B(0, M_s)$ with radii M_s increasing to infinity. Indeed, their Theorem 3 can be extended in two directions. First we may consider a wider class of recursive algorithm of type

$$x_{n+1} = x_n + \gamma_n \left[h(x_{n+1}) + \varepsilon_{n+1} \right],$$

provided h is continuous and the following condition on the noise process (their condition A, see also the condition (2.5) in (Chen, 1993)) is fulfilled

$$\text{a.s.} \quad \lim_{n\to\infty} \gamma_n \sum_{i=0}^{n} \varepsilon_{i+1} = 0 \,. \tag{10}$$

Secondly, the original state space $\mathbb{R}^p$, as well as the disc sequence for truncation can be respectively replaced by any open convex space and any sequence of compact sets tending to its boundary. Therefore, to conclude this step we need just check (10) for the RSEM algorithm. Since $\gamma_n \downarrow 0$, (10) follows from **(T.c)** and Kronecker's lemma.

Step (2). From the first step, we know that the sequence (θ_n) lives a.s. in a compact set of Θ. To conclude the convergence of the truncated RSEM algorithm, we just apply a global Kushner-Clark theorem ((Kushner and Clark, 1978), see also for an improved version, Theorem 2 in (Fort and Pagès, 1996)): since the function K is twice continuously differentiable and the equilibria set $\{\nabla K = 0\}$ contains isolated points only, the gradient system (4) has no quasi-cycles solutions. ■

Nature of the limiting points: an important question from Theorem 1 concerns the nature of the limiting point θ_∞ of the algorithm. It is expected that θ_∞ should be (at least) a local minimum fropoi32 y FF(hendximactsush FFadde)f(ch(includ1 yw())3 y(see7 617 6ru6 145(R)84gu