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* $\mu(X \,|\, Y = Y_{n+1} \,;\, \theta_n)$.

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

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

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

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

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)**

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

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 infcompacity 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.

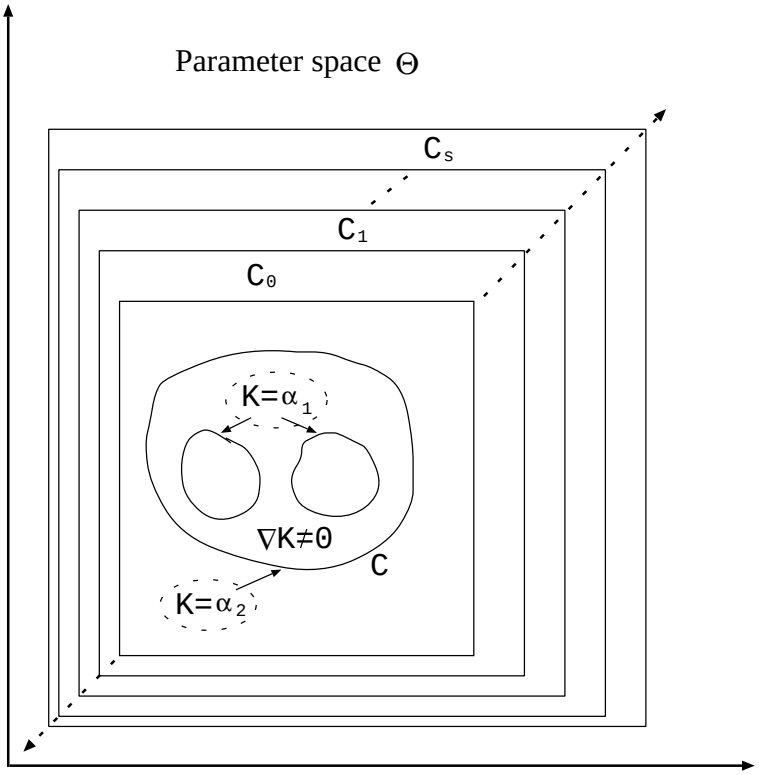

Figure 1: Barrier compacts used for truncations: case of a parameter space $\Theta = (0, \infty)^2$.

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 of $K(\theta)$, that is the RSEM algorithm avoids local maxima or saddle points. This is true, as guaranteed by the following proposition which is a straightforward application of Theorem 1 in (Brandière and Duflo, 1996).

Proposition 1 *In the framework of Theorem 1, assume in addition the variance-covariance matrix $\Gamma(\theta_\infty)$ is positive definite. Then the limit point θ_∞ is a local minima of Kullback-Leibler divergence $K(\theta)$.* □

Speed up of the RSEM algorithm by averaging: to reduce the asymptotic variance of algorithm estimate (θ_n), we adopt the averaging technique introduced by (Polyak, 1990) (see also (Polyak and Juditsky, 1992)). The idea is to use the averages $\overline{\theta}_n$ instead of (θ_n), which can be recursively computed by

$$\overline{\theta}_n = \overline{\theta}_{n-1} + \frac{1}{n}\left(\theta_n - \overline{\theta}_{n-1}\right) \ , \quad \text{with } \overline{\theta}_0 = 0 \, . \qquad \square \tag{11}$$

Explicit construction of a compact sequence (C_s) satisfying the noise control (T.c): Denote the conditional variance-covariance matrix

$$\Gamma(\theta) = \mathbb{E}\left(\varepsilon_{n+1}\varepsilon'_{n+1} \mid \mathcal{F}_n \, , \theta_n = \theta\right) = \mathbb{V}\mathrm{ar}\left(\nabla \log \pi(X_{n+1} \, ; \, \theta_n) \mid \mathcal{F}_n \, , \theta_n = \theta\right) \tag{12}$$

The p-dimensional martingale $\sum \gamma_n \varepsilon_{n+1}$ converges if and only if for all $u \in \mathbb{R}^p$, the scalar martingales $\sum \gamma_n \, u'\varepsilon_{n+1}$ converge. To this end, it will be sufficient to ensure

$$\sum \gamma_n^2 \, \lambda_{\max}\left[\Gamma(\theta_n)\right] \quad \text{converges a.s.} \tag{13}$$

where $\lambda_{\max}(\cdot)$ stands for the largest eigenvalue of a positive definite matrix.

If $\lambda_{\max}\left[\Gamma(\theta)\right]$ is bounded on whole Θ, since $\sum \gamma_n^2 < \infty$ by definition (see (8)), clearly (13) holds and $\sum \gamma_n \varepsilon_{n+1}$ converges without any additional condition on (C_s). Therefore the following construction specifically applies in the more intricate unbounded case where $\sup_\Theta \lambda_{\max}[\Gamma(\theta)] = \infty$.

Let us define for large enough $\rho > 0$,

$$D_\rho := \left\{\theta \in \Theta : \, d(\theta, \partial\Theta) \geq \frac{1}{\rho} \quad \text{and} \quad ||\theta|| \leq \rho\right\} . \tag{14}$$

When $\rho \to \infty$, $D_\rho \to \Theta$. Assume that on D_ρ, $\lambda_{\max}[\Gamma(\theta)]$ can be bounded by an increasing map $\varphi(\rho)$

$$\sup_{D_\rho} \lambda_{\max}[\Gamma(\theta)] \leq \varphi(\rho)\,. \tag{15}$$

Necessarily, $\lim \varphi(\rho) = \infty$ when $\rho \to \infty$. Let be some constants $b > 1$ and $c > 0$. We may take, by inversion of φ, a strictly increasing sequence (ρ_s) such that

$$\varphi(\rho_s) \;\leq\; c\frac{s^{2\alpha-1}}{(\log s)^b}\,, \quad \text{for } s \geq 2. \tag{16}$$

For example, if φ is continuous in addition, we can take $\rho_s = \varphi^{-1}\left(cs^{2\alpha-1}(\log s)^{-b}\right)$. Finally let us take for large enough s, say $s \geq s_0$,

$$C_s := D_{\rho_s}\,. \tag{17}$$

Thus for $n \geq s_0$, since $\theta_n \in C_{\sigma(n)} \subset C_n = D_{\rho_n}$, we have

$$\lambda_{\max}[\Gamma(\theta_n)] \leq \varphi(\rho_n) \leq c\frac{n^{2\alpha-1}}{(\log n)^b}\,.$$

Hence, (13) as well as **(T.c)** holds.

Summarizing, we have

Proposition 2 *In the case where $\lambda_{\max}[\Gamma(\theta)]$ is unbounded on Θ, assume there exists an upper bound $\varphi(\rho)$ satisfying (15) for large ρ. With the compact sets C_s defined in (16)-(17), it holds that $\sum \gamma_n \varepsilon_{n+1}$ converges a.s. (condition **(T.c)**).*

Such a construction will be explicited in §4 for finite mixtures. □

4 Application to finite mixtures

This section is devoted to show the effectiveness of the truncated RSEM algorithm for mixture model. Since the moment estimators introduced in (Pearson, 1894), finite mixture models has been and continue to be widely analysed by statisticians. We refer to (Redner and Walker, 1984) and (Titterington et al., 1985) for classical backgrounds. Recent advances including Bayesian approach, stochastic estimators or dimension testing can be found in (Celeux and Diebolt, 1985; Dacunha-Castelle and Gassiat, 1997; Robert, 1996; Richardson and Green, 1997; Celeux et al., 1996). For recursive estimation, (Rydén, 1994) proposed a recursive likelihood estimator using projection on some suitable compact subset of Θ. Despite of some similarity, several differences arise between Rydén's algorithm and the RSEM algorithm. First the RSEM algorithm is based on the complete data likelihood up to a restoration step; second, instead of projections on a compact subset of Θ as proposed in Rydén's procedure, the RSEM algorithm achieves stabilisation by truncations at randomly varying bounds.

Let be m a positive integer, Γ an open convex set of $\mathbb{R}^q$ and $\mathcal{D} = \{f(\cdot\,;\phi)\ ;\ \phi \in \Gamma\}$ a parametric family of univariate densities. A finite mixture with m components is a variable Y with the following density

$$g(y\,;\,\theta) = \sum_{k=1}^{m} \alpha_k\, f(y\,;\phi_k) \tag{18}$$

where $\boldsymbol{\alpha} := (\alpha_k)$ is a probability distribution (*mixing distribution*) on the set $\{1, \ldots, m\}$, and $\boldsymbol{\phi} := (\phi_1, \ldots, \phi_m)$ are component parameters of the mixture. Therefore, the whole parameter vector is $\theta = (\boldsymbol{\alpha}, \boldsymbol{\phi})$ which belongs to $\Theta = \Delta \times \Gamma^m$ where Δ denotes the open simplex $\{\alpha_1 > 0, \ldots, \alpha_{m-1} > 0,\ \alpha_1 + \cdots + \alpha_{m-1} < 1\}$.

A classical way to view such a mixture model as an incomplete data model is to think about an unobservable location variable Z on the integers $\{1, \ldots, m\}$, and that the observation Y is drawn conditionally to this location (see (Dempster et al., 1977)). That is to consider the vector $X = (Y, Z)$ where

$$Z \in \{1, \ldots, m\}, \quad \text{with } \mathbb{P}(Z = k) = \alpha_k\,,$$

and

$$\mu(Y \mid Z = k) \sim f(y\,;\ \phi_k)dy.$$

It is easy to see that the marginal distribution of Y is exactly the mixture (18). The missing data here is the location variable Z.

4.1 Kullback-Leibler divergence between finite mixtures

Let us show that in general, the inf-compacity (5) does not hold for mixtures. First consider the case with one of the mixing probability α_k, say α_1, tends to 0. The density $g(\cdot\,;\theta)$ degenerates to the density of a smaller mixture with $m - 1$ components. Therefore, we may have

$$\lim_{\alpha_1 \to 0} K(\theta) < \infty,$$

when for example, the probability distributions in $\mathcal{D}$ are all equivalents. This finite limiting behaviour may still happen if the component parameters $\boldsymbol{\phi}$ go to the boundary. Consider indeed the family of exponential distributions $f(y\ ;\phi_k) = \phi_k e^{-\phi_k y}$, $\phi_k \in (0, \infty)$ and take $m = 2$. The parameter vector is $\theta = (\alpha, \phi_1, \phi_2)$ with $\Theta = (0,1) \times (0,\infty) \times (0,\infty)$. It is easy to see that

$$\text{with fixed } (\alpha, \phi_1), \quad \lim_{\phi_2 \to \infty} K(\theta) < \infty.$$

4.2 Set up of the RSEM algorithm for mixtures

Let $\theta_n = (\alpha_{1,n}, \ldots, \alpha_{m-1,n}, \phi_{1,n}, \ldots \phi_{m,n})$ be the current estimate at each step n. The restoration R-step draws a sample Z_{n+1} from the following conditional distribution $\mu(Z \mid Y = Y_{n+1}\,;\,\theta_n)$

$$\mathbb{P}(Z = k \,|\, Y = Y_{n+1}\,;\,\theta_n) = \frac{\alpha_{k,n} f(Y_{n+1}\ ;\ \phi_{k,n})}{\sum_{\ell=1}^{m} \alpha_{\ell,n} f(Y_{n+1}\ ;\ \phi_{\ell,n})}, \qquad 1 \le k \le m. \tag{19}$$

The p.d.f of the complete data $X = (Y, Z)$ is $\pi(x\,;\theta) = \pi(y, z\,;\,\theta) = \alpha_z f(y\ ;\ \phi_z)$. However, the mixing probabilities $\boldsymbol{\alpha} := (\alpha_k)$ have to be kept within the simplex Δ and this constraint is

particularly hard to satisfy for a recursive computation procedure. For instance, the updating rule (1) applied to α_1 gives at step n, $\alpha_{1,n+1} = \alpha_{1,n} + \gamma_n \frac{\partial}{\partial \alpha_1} \log \pi(X_{n+1}; \theta_n)$, and we see that even if $\alpha_{1,n} \in (0,1)$, $\alpha_{1,n+1}$ may escape from $(0,1)$ with a positive probability. Therefore to overcome such numerical instability, we propose a new parametrisation based on the Logit transformation. That is instead of $\boldsymbol{\alpha} := (\alpha_k)$, we shall use $\boldsymbol{\omega} := (\omega_k)$ defined as

$$\omega_k = \log\left(\frac{\alpha_k}{\alpha_m}\right), \quad \text{for } 1 \leq k < m\,, \quad \text{and} \quad \omega_m = 0 \tag{20}$$

Note that $\boldsymbol{\omega} := (\omega_k)$ belongs to $\mathbb{R}^{m-1}$ and this transformation has an easy inversion formula. Hence the new parameters are $\theta = (\omega_1, \ldots, \omega_{m-1}, \phi_1, \ldots \phi_m)$.

The estimation E-step writes as

$$\left\{ \begin{array}{rcll} \omega_{k,n+1} & = & \omega_{k,n} + \gamma_n \left[\mathbf{1}_{Z_{n+1}=k} - \alpha_{k,n} \right] , & 1 \leq k < m \\ \phi_{k,n+1} & = & \phi_{k,n} + \gamma_n \mathbf{1}_{Z_{n+1}=k} \nabla \log f(Y_{n+1}\,;\, \phi_{k,n}) \,, & 1 \leq k \leq m \end{array} \right. \tag{21}$$

Recall that the final estimates are the averages $(\bar{\boldsymbol{\omega}}_n)$ and $(\bar{\boldsymbol{\phi}}_n)$, as defined in (11).

4.3 A simulation experiment

We shall consider two univariate Gaussian mixtures with two components each, taken in (Redner and Walker, 1984). Thus $\theta = (\omega_1, m_1, \sigma_1^2, m_2, \sigma_2^2)$ and $g(.;\theta) = \alpha_1 \mathcal{N}(m_1; \sigma_1^2) + (1 - \alpha_1)\mathcal{N}(m_2; \sigma_2^2)$. The two "true" mixtures are $\mathbf{M_1}$ = 0.3*$\mathcal{N}$(3 ; 1) + 0.7*$\mathcal{N}$(-3 ; 1) and $\mathbf{M_2}$ = 0.3*$\mathcal{N}$(1 ; 1) + 0.7*$\mathcal{N}$(-1 ; 1). The first is a well-separated bimodal distribution, while the second is an unimodal one. Note that the smoothness conditions **(S)** are met in this Gaussian mixture case.

We now describe in details the set up of the algorithm and the constants used throughout all the simulations.

Initialisation: The algorithm starts with

$$\theta_0 = (\,0,\, \frac{3}{2} m_{1,*}\,,\, \frac{1}{2}\,\sigma_{1*}^2\,, \frac{3}{2}\, m_{2,*},\, \frac{1}{2}\,\sigma_{2*}^2\,).$$

Again the same initial value is used in (Redner and Walker, 1984) for their experiments on the EM algorithm. Note that $\omega_1 = 0$ iff $\alpha_1 = 1/2$. □

Gains sequence: The gains sequence is a piecewise constant version of the general rule (8) with exponent $\alpha = 3/4$: we set

$$\text{for each } n \in [100(k-1),\, 100k) \text{ with integer } k \geq 1, \quad \gamma_n = \frac{10}{(100k+10)^{3/4}}. \quad \square$$

Compacts C_s for truncation: For sake of simplicity, the description is given by means of the proportion parameter α_1: the actually used bounds for ω_1 could be easily derived. Thus C_s will take a product form $A_s \times M_s \times V_s \times M_s \times V_s$, for $\alpha_1 \in A_s$, $m_k \in M_s$ and $\sigma_k^2 \in V_s$. We use a two-staged set-up for (C_s).

(i) For 100 first compacts with $0 \leq s < 100$, a constant step size is used

$$\begin{aligned} A_s &= \left[10^{-1} - s10^{-4},\ 1 - \left(10^{-1} - s10^{-4}\right)\right], \\ M_s &= \left[-100 - s,\ 100 + s\right], \\ V_s &= \left[10^{-3} - s10^{-6},\ 100 + s\right]. \end{aligned}$$

Note that the starting compact is

$$C_0 = \left\{0.1 \leq \alpha_1 \leq 0.9,\ |m_k| \leq 100,\ 10^{-3} \leq \sigma_k^2 \leq 100\right\} \tag{22}$$

In particular, the true parameter θ_* is believed to belong to C_0. In real data case, C_0 should be set with a rough preliminary estimate of θ_*.

(ii) For next compacts with $s \geq 100$, we use the construction defined in Proposition 2. Taking into account the believed fact $\theta_* \in C_0$, this procedure yields the following

$$C_s = \left\{ \frac{1}{\rho_s} \leq \alpha_1 \leq 1 - \frac{1}{\rho_s},\ |m_k| \leq \rho_s,\ \frac{1}{\rho_s} \leq \sigma_k^2 \leq \rho_s \right\} \tag{23}$$

with

$$\rho_s = 10^3 \left\{ \frac{\left(\frac{s}{100}\right)^{\frac{1}{2}}}{\left(\frac{\log s}{\log 100}\right)^2} \right\}^{\frac{1}{8}}. \tag{24}$$

Detailed derivations of these formula are postponed to Lemma 2 at the end of the section. It is worth noting that $\rho_{100} = 10^3$, so that $C_{99} \subset C_{100}$ which links up two stages well.

Finally, the restart value w used after each truncation is set to be the same as the starting value : $w = \theta_0$. $\square$

For each of the two mixtures $\mathbf{M_1}$ and $\mathbf{M_2}$, 100 independent runs RSEM over $N = 1000$ iterations are generated and then averaged. Tab.1 and Tab.2 show respectively statistics about finally minimised Kullback distance $K(\overline{\theta}_N)$, number of truncations $\sigma(N)$ and the last truncation time T defined as

$$T = \max\left\{n : \theta_n + \gamma_n \nabla \log(X_{n+1}; \theta_n) \notin C_{\sigma(n)} \text{ and } n \leq N\right\}.$$

	Mean of $K(\overline{\theta}_N)$	Starting value $K(\theta_0)$	ratio $K(\overline{\theta}_N)/K(\theta_0)$
Mixture $\mathbf{M_1}$	0.0538	2.4819	2.2%
Mixture $\mathbf{M_2}$	0.0152	0.2386	6.3%

Table 1: Mean of the minimised Kullback distances from 100 independent runs of the truncated RSEM algorithm, with their ratio to the starting value.

	Number of runs actually truncated	Sum of $\sigma(N)$	Mean of T
Mixture $\mathbf{M_1}$	37	47	119.1
Mixture $\mathbf{M_2}$	91	265	243.5

Table 2: Truncation statistics from 1000 independent runs of the truncated RSEM algorithm: number of runs with at least one truncation, total number of truncations $\sum \sigma(N)$ over 100 independent runs and mean of the last truncation time T from those runs with truncations.

From Tab.1, we see that the mixture $\mathbf{M_2}$ is more difficult to handle with since the minimisation ratio (6.3%) is about 3 times bigger than for the mixture $\mathbf{M_1}$ (2.2%), even absolute Kullback distances are smaller. A histogram of minimised Kullback distances is given on the top row in Fig. 2. These distances are highly grouped near 0.

For the truncation behaviour shown in Tab.2, the mixture $\mathbf{M_2}$ need more truncations (91 truncated runs against 37 ones, and 265 truncations in total against 47). Also the last truncation can occur much later (243.5 in average against 119.1). The middle row in Fig. 2 show the corresponding histograms. Note that the maximum of the truncation number observed over 100 runs is respectively 9 for $\mathbf{M_2}$ and 4 for $\mathbf{M_1}$. Again these histograms are concentrated near 0. Thus it seems true that few truncations are enough to stabilize the algorithm.

Fig.3 shows the evolution of a density estimate $g(.;\overline{\theta}_n)$ for time $n = 0, 250, 500$ and 1000, taken from one of the 100 generated samples for the mixture $\mathbf{M_1}$. This sample run has the following characteristics: $\sigma(N) = 1$;

Proof. Let us denote respectively by $\mathbb{E}_{n,\theta}$ and $\mathbb{V}_{n,\theta}$, the conditional expectation and variance with respect to $[\mathcal{F}_n\,, \theta_n = \theta]$. Recall that $\Gamma(\theta) = \mathbb{V}_{n,\theta}[\nabla \log \pi(X_{n+1}; \theta_n)]$. We have

$$\lambda_{\max}[\Gamma(\theta)] \leq \text{trace } [\Gamma(\theta)] = \sum_j \mathbb{V}_{n,\theta}\left[\frac{\partial}{\partial\theta_j}\log\pi(X_{n+1};\theta_n)\right] .$$

where θ_j stands for ω_1, m_k and σ_k^2. For mixture of two univariate normal variables, we have by (21)

$$\left\{\begin{array}{lcl} \frac{\partial}{\partial\omega_1}\log\pi(X_{n+1};\theta_n) & = & \mathbf{1}_{Z_{n+1}=1} - \alpha_{1,n} \\ \frac{\partial}{\partial m_k}\log\pi(X_{n+1};\theta_n) & = & \mathbf{1}_{Z_{n+1}=k}\,\frac{Y_{n+1}-m_{k,n}}{\sigma^2_{k,n}} \\ \frac{\partial}{\partial\sigma_k^2}\log\pi(X_{n+1};\theta_n) & = & \mathbf{1}_{Z_{n+1}=k}\left[-\frac{1}{2\sigma^2_{k,n}} + \frac{1}{2\sigma^4_{k,n}}\,(Y_{n+1}-m_{k,n})^2\right] \end{array}\right. \tag{26}$$

Straightforward computation shows that for $\theta \in D_\rho$

$$\begin{aligned} \mathbb{V}_{n,\theta}\left[\mathbf{1}_{Z_{n+1}=1} - \alpha_{1,n}\right] &\leq 1\,, \\ \mathbb{E}_{n,\theta}\left[\mathbf{1}_{Z_{n+1}=k}\,\frac{Y_{n+1}-m_{k,n}}{\sigma^2_{k,n}}\right]^2 &\leq 2\rho^2\left(\rho^2+\mu_{2,*}\right)\,, \\ \mathbb{E}_{n,\theta}\left\{\mathbf{1}_{Z_{n+1}=k}\left[-\frac{1}{2\sigma^2_{k,n}} + \frac{1}{2\sigma^4_{k,n}}\,(Y_{n+1}-m_{k,n})^2\right]\right\}^2 &\leq \rho^4\left(\rho^4+2\mu_{2,*}\rho^2+\mu_{4,*}\right)\,. \end{aligned}$$

where $\mu_{j,*} = \mathbb{E}_{\theta_*}|Y|^j$. Hence for $\rho \geq 1$,

$$\text{trace } [\Gamma(\theta)] \leq 2\rho^4\left(\rho^4 + 4\mu_{2,*}\rho^2 + \mu_{4,*} + 2\right) + 1\,.$$

It remains to bound the expectations $\mu_{j,*}$. But if W is a normal variable $\mathcal{N}(m,\sigma^2)$, we have

$$\mathbb{E}|W|^j \leq 2^{j-1}\left(|m|^j + 4\sigma^j\right)\,.$$

Now we have assumed $\theta_* \in C_0$, so that by (22), $|m_{k,*}| \leq 100$ and $10^{-3} \leq \sigma^2_{k,*} \leq 100$. It follows

$$\mu_{2,*} \leq 10^5, \quad \mu_{4,*} \leq 4\cdot 10^9.$$

Hence for $\rho \geq 10^3$,

$$2\rho^4\left(\rho^4 + 4\mu_{2,*}\rho^2 + \mu_{4,*} + 2\right) + 1 \leq 3\rho^8\,.$$

and (25) follows.

Finally, to define ρ_s and $C_s = D_{\rho_s}$ for $s \geq 100$, let us take $b = 2$ in the rule (16). Since φ is continuous, and we have chosen the exponent $\alpha = 3/4$ for (γ_n), we are looking for ρ_s satisfying

$$\varphi(\rho_s) = 3\,(\rho_s)^8 = c\frac{s^{1/2}}{(\log s)^2}\,, \quad \text{for large } s\,.$$

We fix the constant c by the initial condition

$$\text{for } s = 100,\ \rho_s = 10^3.$$

Hence (24) follows. ■

Acknowledgement. *I thank Marie Duflo for several helpful discussions on this work.*

References

Brandière, O. and Duflo, M. (1996). Les algorithmes stochastiques contournent-ils les pièges? *Ann. Inst. Henri Poincaré*, 32:395–427.

Celeux, G., Chauveau, D., and Diebolt, J. (1996). Stochastic versions of the EM algorithm: an experimental study in the mixture case. *J. Statist. Computation and Simulation*, 55:287–314.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quaterly*, 2:73–82.

Chen, H. (1993). Asymptotically efficient stochastic approximation. *Stochastics and Stochastics Reports*, 45:1–16.

Chen, H., Guo, L., and Gao, A. (1988). Convergence and robustess of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stochastic Processes and their Applications*, 27:217–231.

Dacunha-Castelle, D. and Gassiat, E. (1997). Testing in locally conic models and applications to mixture models. *ESAIM Probab. Statist.*, 1:285–317.

Dempster, A., Liard, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Soc. Stat. B*, 39:1–38.

Duflo, M. (1996). *Algorithmes Stochastiques*. Springer-Verlag, Berlin.

Fabian, V. (1978). On asymptotic efficient recursive estimation. *Ann. of Stat.*, 6:854–856.

Fort, J. and Pagès, G. (1996). Convergence of stochastic algotithms : from the Kushner-Clark theorem to the Lyapounov functional method. *Adv. Appl. Probab.*, 28:1072–1094.

Kushner, H. and Clark, D. (1978). *Stochastic Approximation for Constrained and Unconstrained Systems*. Springer-Verlag, Berlin.

Lavielle, M. and Moulines, E. (1995). On a stochastic approximation version of the EM algorithm. Technical Report 95.08, Université Paris-Sud.

Meng, X. and van Dyk, D. (1997). The EM algorithm - an old folk-song sung to a fast new tune. *J. Royal Statist. Soc.*, 59:511–567.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Phil. Trans. Royal Soc. A*, 185:71–110.

Polyak, B. T. (1990). New stochastic approximation type procedures. *Avtomat. i Telemekh.*, 7:98–107. English translation: Automation and remote control **51**, 1991.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. control and optimization*, 30:838–855.

Qian, W. and Titterington, D. (1991). Estimation of parameters in hidden Markov models. *Phil. Trans. R. Soc. Lond.*, 337:407–428.

Redner, R. and Walker, H. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26:195–239.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. B*, 59:731–792.

Robert, C. (1996). Mixtures of distributions: inference and estimation. In W. R. Gilks, S. R. and Spiegelhater, D. J., editors, *Markov Chain Monte Carlo in Pratice*, chapter 24, pages 441–464. Chapman and Hall, London.

Rydén, T. (1994). Asymptoticallly efficient recursive estimation for incomplete data models using the observed information. Technical report, Dept. of Mathematical Statistics, University of Lund.

Titterington, D., Smith, A., and Markov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Wei, G. and Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. American Statist. Association*, 85:699–704.

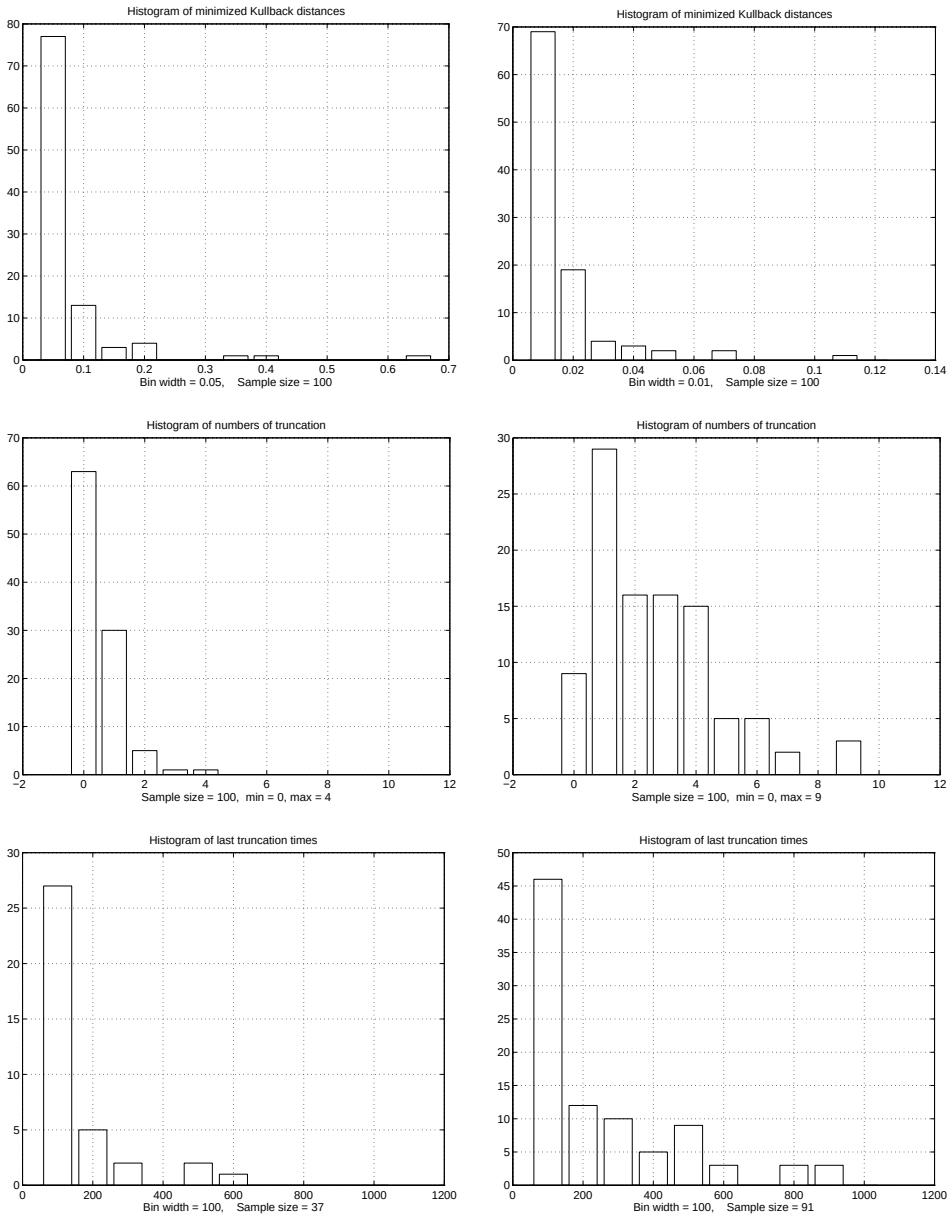

Figure 2: Histograms from 100 independent runs for the mixture $\mathbf{M_1}$ (left column) and $\mathbf{M_2}$ (right column). Top: minimised Kullback distances $K(\bar{\theta}_n)$. Middle: number of truncations $\sigma(N)$. Bottom: the last truncation time T.

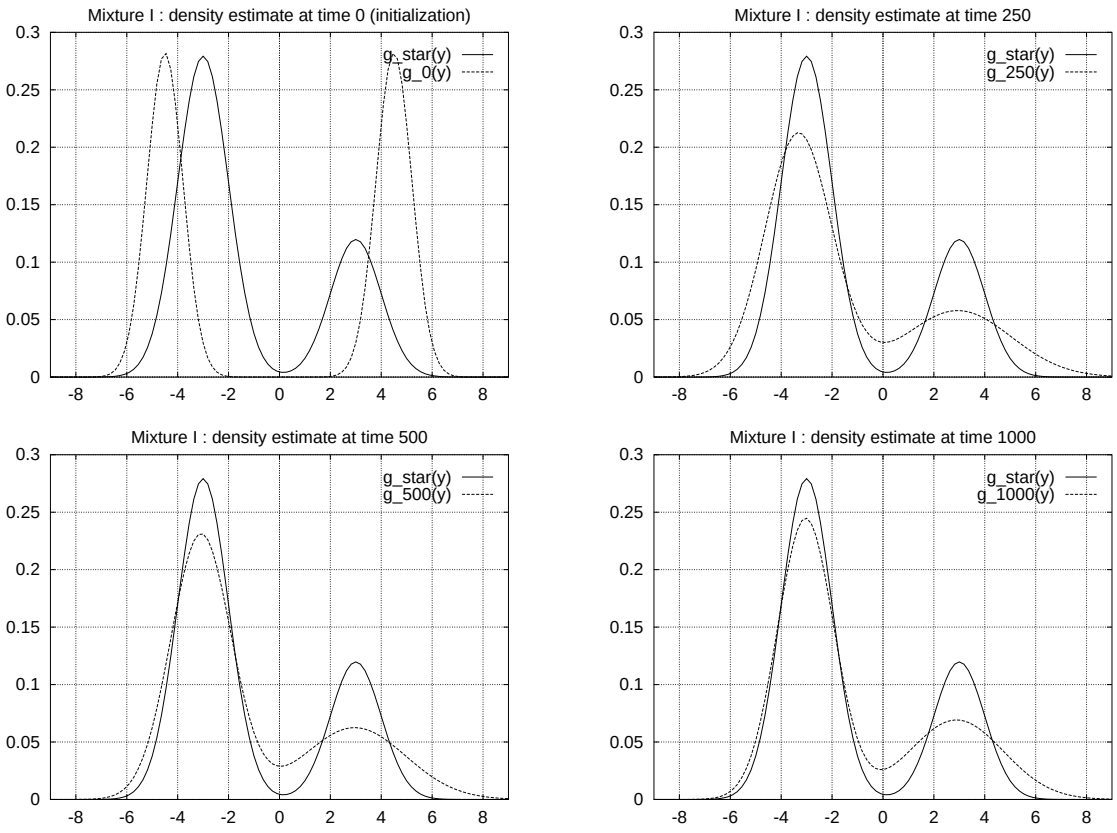

Figure 3: Evolution of density estimates for the $\mathbf{M}_1$ (dashed lines) : for time 0,250,500 and 1000 from top to bottom and left to right. The corresponding Kullback distances are 2.482, 0.155, 0.104 and 0.076. The true density is shown with solid lines.

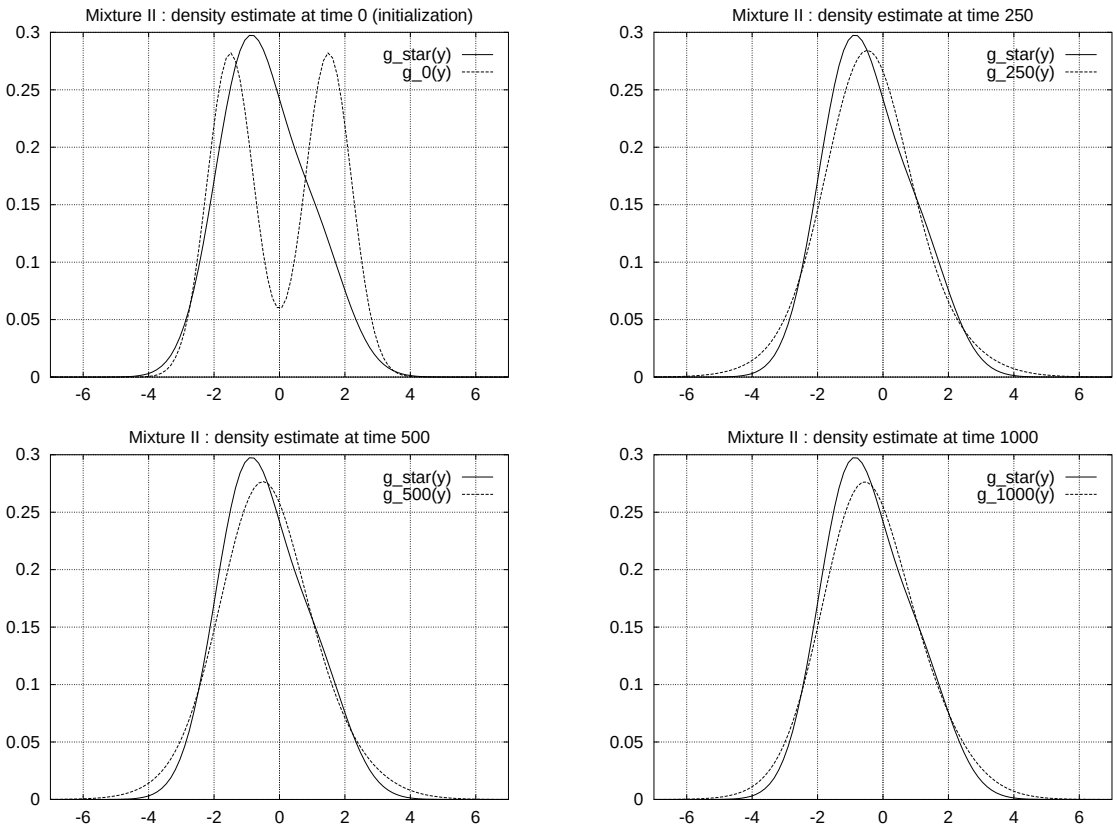

Figure 4: Evolution of density estimates for $\mathbf{M_2}$ (dashed lines): for time 0,250,500 and 1000 from top to bottom and left to right. The corresponding Kullback distances are 0.2386, 0.0254, 0.0255 and 0.0188. The true density is shown with solid lines.