A New Class of EM Algorithms. Escaping Local Minima and Handling Intractable Sampling

Stéphanie Allassonnière, Juliette Chevallier

A New Class of EM Algorithms. Escaping Local Minima and Handling Intractable Sampling

Stéphanie Allassonnière and Juliette Chevallier

Abstract—

The expectation-maximization (EM) algorithm is a powerful computational technique for maximum likelihood estimation in incomplete data models. When the expectation step cannot be performed in closed form, a stochastic approximation of EM (SAEM) can be used. The convergence of the SAEM toward local maxima of the observed likelihood has been proved and its numerical efficiency has been demonstrated. However, despite appealing features, the limit position of this algorithm can strongly depend on its starting position. Moreover, sampling from the posterior distribution may be intractable or have a high computational cost. To cope with this two issues, we propose here a new stochastic approximation version of the EM in which we do not sample from the exact distribution in the expectation phase of the procedure. We first prove the convergence of this algorithm toward local maxima of the observed likelihood. Then, we propose an instantiation of this general procedure to favor convergence toward global maxima. Experiments on synthetic and real data highlight the performance of this algorithm in comparison to the SAEM.

Index Terms—EM-like algorithm, stochastic approximation, stochastic optimization, tempered distribution, theoretical convergence.

1 Introduction

The expectation-maximization (EM) algorithm [1] is a popular and often efficient approach to *maximum* likelihood (or *maximum a posteriori*) estimation in incomplete data models. In certain situations, however, the EM is not applicable because the expectation step cannot be performed in closed form. To deal with these problems, [2] proposed to replace the expectation step of the EM by one iteration of a stochastic approximation procedure, referred to as SAEM, standing for stochastic approximation EM.

The convergence of the SAEM toward local *maxima* has been proved in [2] and its numerical efficiency has been demonstrated in several situations such as in inference in hidden Markov models [3]. However, despite appealing features, the limit position of this algorithm can strongly depend on its initialization. In order to avoid convergence toward local *maxima*, Lavielle and Moulines [4] have proposed a simulated annealing version of the SAEM. The main idea was to allow the procedure to better explore the state-space by considering a tempered version of the model. More precisely, assuming that the data are corrupted by an additive Gaussian noise with variance σ^2, at each iteration k of the SAEM algorithm, they consider the "false" model in which the noise variance is equal to $((1+T_k)\sigma)^2$, where (T_k) is a positive sequence ot temperatures that decreases slowly toward 0. Therefore, the bigger T_k is, the more the likelihood of the model is flattened and the optimizing sequence can escape easily from local *maxima*. The simulations gave good results but there were no theoretical guarantee for this procedure. Based on the same idea, Lavielle [5] has proposed to use the simulated-annealing process as a "trick" to better initialize the SAEM algorithm. This initialization scheme is implemented in the MONOLIX software and gives impressive results on real data [6], [7], [8].

All theoretical results regarding the convergence of the SAEM algorithm assume that we are able to sample from the posterior distribution, but in practice it may be intractable or have a high computational cost. To overcome this issue, Umberto and Samson [9] have proposed to couple the SAEM algorithm to an approximate Bayesian computation step (ABC, see [10] for a review), leading to the ABC-SAEM method in which ABC is used to sample from an approximation to the posterior distribution. Simulations show that this algorithm can be calibrated to return accurate inference, and in some situations it can outperform a version of the SAEM incorporating the bootstrap filter. However, [9] do not provide any theoretical guarantee of its convergence.

We propose here a new stochastic approximation version of the EM in which we do not sample from the exact distribution in the expectation phase. This new procedure allows us to derive a wide class of SAEM-like algorithms, including the "trick" initialized SAEM of [5] and the ABC-SAEM algorithms to cope with intractable or difficult sampling.

This general framework allows us to build a procedure, with the thought of the simulated annealing version of the SAEM [4], to prevent convergence toward local *maxima*. We introduce a sequence of temperatures and sample from a tempered version of the posterior distribution. Therefore, the posterior-likelihood of the model is "flattened" and the optimizing sequence can escape more easily from local *maxima*. We refer to this particular instantiation as the tempering-SAEM. Note that our tempering-SAEM differs to the ones of [4] as we do not modify the model but only the sampling-step.

In Section 2, we introduce our new stochastic approxima-

- *Stéphanie Allassonnière is with the Centre de Recherche des Cordeliers, Université Paris-Descartes, Paris.*
- *Juliette Chevallier is with Centre de Mathématiques Appliquées, Écoles polytechniqye, Palaiseau.*
  E-mail: juliette.chevallier@polytechnique.edu

tion version of the EM algorithm, namely the approximated-SAEM, and prove the convergence of this algorithm toward local *maxima* under usual assumptions. Thus, we provide a theoretical study of the convergence of the tempering-SAEM toward local *maxima*. We also give a heuristic to the convergence to "less local" *maxima*. Section 3 is dedicated to experiments. The first application we take into account is the *maximum* likelihood estimation of the parameters of a multivariate Gaussian mixture models. This example supports the previous heuristic discussion and gives intuitions into the behavior of the tempering-SAEM algorithm. The second application consists in independent factor analysis [11]. In both applications, we focus on the contribution of the tempering-SAEM in comparison to the SAEM.

2 Maximum Likelihood Estimation through an EM-Like Algorithm

We use in the sequel the classical terminology of the missing data problem, even though the approaches developed here apply to a more general context.

Let $\mathcal{Y} \subset \mathbb{R}^{n_y}$ denote the set of observations, $\mathcal{Z} \subset \mathbb{R}^{n_z}$ the set of latent variables and $\Theta \subset \mathbb{R}^{n_\theta}$ the set of admissible parameters. Let μ be a σ-finite positive Borel measure on $\mathcal{Z}$. For sake of simplicity, we will use the notation q for different likelihoods, specifying their variables in brackets. In particular, for all $(y;\theta) \in \mathcal{Y} \times \Theta$, $q(y,\,\cdot\,;\theta)$ is the complete likelihood given the observation y and parameter θ and we assume it is integrable with respect to the measure μ. As for, we note $q(y;\theta) = \int_{\mathcal{Z}} q(y,z;\theta)\,\mathrm{d}\mu(z)$ the observed likelihood and $q(z|y;\theta) = \frac{q(y,z;\theta)}{q(y;\theta)}$ the posterior distribution of the missing data z given the observed data y. Our goal is to estimate the parameters that maximize the likelihood of the observations of n independent samples of a random variable Y, *i.e.* that maximize the observed data likelihood.

2.1 A New Stochastic Approximation Version of the EM Algorithm

We propose in this contribution a generalization of the SAEM algorithm, referred to as approximated-SAEM. Similar to the SAEM, the basic idea is to split the E-step into a simulation step and a stochastic averaging procedure. In the original SAEM, the S-step consists in generating realizations of the missing data vector under the posterior distribution $q(\cdot|y;\theta)$. Here, we propose to sample under *approximation* of the posterior distribution. The following paragraph describes this new algorithm.

Let $\gamma = (\gamma_k)_{k\in\mathbb{N}}$ be a sequence of positive step-size for the stochastic approximation, and $\tilde{q} = (\tilde{q}_k)_{k\in\mathbb{N}}$ be a sequence of *approximated* distributions on $\mathcal{Z}\times\Theta$ such that for all $k \in \mathbb{N}$ and all $\theta \in \Theta$, $\tilde{q}_k(\cdot;\theta)$ is integrable on $\mathcal{Z}$ with respect to the measure μ. As in the SAEM, once the step size γ_k decreases, we can consider a constant number of simulations. In practice (and from now on to avoid cumbersome notations), as the S-step is generally the most computationally costly, we set this number to one. Then, the approximated-SAEM iterates the following three steps:

S-step: Sample the latent variable $\tilde{z}_k$ under the approximated density $\tilde{q}_k(\cdot;\theta_k)$;

SA-step: Update $Q_k(\theta)$ as

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k\big(\log q(y,\tilde{z}_k;\theta) - Q_{k-1}(\theta)\big)\,;$$

M-step: Maximize $Q_k(\theta)$ in the feasible set Θ, *i.e.* find $\theta_{k+1} \in \Theta$ such that

$$\forall\theta\in\Theta,\quad Q_k(\theta_{k+1}) \geqslant Q_k(\theta)\,.$$

Note that without approximation, *i.e.* if the approximated densities $\tilde{q}_k$ match with the correct posterior distribution, we feature the classical SAEM. Moreover, the approximated densities $\tilde{q}_k$ may not depend on the observations y, as in variational Bayesian methods or may be done by ABC samplers as in ABC-SAEM. In Section 2.2, we propose a way to build a sequence $\tilde{q}$ leading to good properties in practice and theoretical guarantees are given in the following section.

2.1.1 Curved Exponential Family

Before establishing the convergence of this procedure, we briefly recall the hypothesis required to prove the convergence of the EM. More precisely, we restrict our attention to models for which the complete data likelihood belongs to the curved exponential family. In this paragraph and the following, we keep the notations of [2]: an hypothesis stated with a $(^\star)$ means that it is a direct generalization of the corresponding one in [2]; on the contrary, hypothesis stated without are unchanged compared to the original one.

(M1*) The parameter space Θ is an open subset of $\mathbb{R}^{n_\theta}$. For all $y \in \mathcal{Y}$, $z \in \mathcal{Z}$ and $\theta \in \Theta$, the complete data likelihood function can be expressed as

$$q(y,z;\theta) = \exp\big(-\psi(\theta) + \langle\, S(y,z)\,|\,\phi(\theta)\,\rangle\big)$$

where $S : \mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_s}$ is a Borel function and $\mathcal{S}$ is an open subset of $\mathbb{R}^{n_s}$. The convex hull of $S(\mathbb{R}^{n_z})$ is included in $\mathcal{S}$. For all $\theta \in \Theta$, all $y \in \mathcal{Y}$ and all $k \in \mathbb{N}$, we have

$$\int_{\mathcal{Z}} \|S(y,z)\|\,\tilde{q}_k(z;\theta)\,\mathrm{d}\mu(z) < +\infty$$

$$\text{and}\quad \int_{\mathcal{Z}} \|S(y,z)\|\,q(z|y;\theta)\,\mathrm{d}\mu(z) < +\infty\,.$$

Let $\ell\colon \Theta \to \mathbb{R}$ and $L\colon \mathcal{S}\times\Theta \to \mathbb{R}$ defined as,

$$\text{for all } y\in\mathcal{Y},\quad \ell\colon \theta \mapsto \int_{\mathcal{Z}} q(y,z;\theta)\,\mathrm{d}\mu(z)$$

$$\text{and}\quad L\colon (s,\theta) \mapsto -\psi(\theta) + \langle\, s\,|\,\phi(\theta)\,\rangle\,.$$

(M2) The functions $\psi\colon \Theta\to\mathbb{R}$ and $\phi\colon\Theta\to\mathcal{S}$ are twice continuously differentiable on Θ;

(M3) The function $\bar{s} : \Theta \to \mathcal{S}$ is continuously differentiable on Θ, where $\bar{s}$ is defined as: $\forall y \in \mathcal{Y}$,

$$\bar{s}\colon \theta \mapsto \int_{\mathcal{Z}} S(y,z) q(z|y;\theta)\,\mathrm{d}\mu(z) = \mathbb{E}_\theta\left[S(Z)\right]\,;$$

(M4) The function $\ell\colon \Theta\to\mathbb{R}$ is continuously differentiable and for all $y\in\mathcal{Y}$ and $\theta\in\Theta$

$$\partial_\theta \int_{\mathcal{Z}} q(y,z;\theta)\,\mathrm{d}\mu(z) = \int_{\mathcal{Z}} \partial_\theta\, q(y,z;\theta)\,\mathrm{d}\mu(z)\,;$$

(M5) There exists a continuously differentiable function $\hat{\theta}\colon \mathcal{S}\to\Theta$ such that

$$\forall\theta\in\Theta,\quad \forall s\in\mathcal{S},\quad L(s,\hat{\theta}(s)) \geqslant L(s,\theta)\,.$$

Hypothesis (M1*) differs from (M1) as we do not only require the function $z \mapsto \|S(z;\theta)\|$ to be integrable with respect to the posterior measure $q(\cdot|y;\theta)\,\mathrm{d}\mu$, but also with respect to all approximated distributions $\tilde{q}_k(\cdot;\theta)\,\mathrm{d}\mu$, for all parameters $\theta \in \Theta$, all observations $y \in \mathcal{Y}$ and all iterations $k \in \mathbb{N}$. For most models of practical interest (see for instance Section 3.2), the function $L(s;\cdot)$ has a unique global *maximum* and the existence and the differentiability of $\hat{\theta}$ is a direct consequence of the implicit function theorem.

For exponential families, the SA-step is more conveniently (and equivalently) replaced by an update of the estimation of the conditional expectation of the sufficient statistics. Namely, the k-th iteration of the approximated-SAEM summarizes in:

$$s_k = s_{k-1} + \gamma_k\big(S(y,\tilde{z}_k) - s_{k-1}\big) \qquad (1)$$
$$\text{and} \quad \theta_k = \hat{\theta}(s_k) \quad \text{where} \quad \tilde{z}_k \sim \tilde{q}_k(\cdot;\theta_{k-1})\,.$$

2.1.2 Convergence Toward Local Maxima

Let $\widetilde{\mathcal{F}} = \{\widetilde{\mathcal{F}}_k\}_{k\in\mathbb{N}}$ the natural filtration with respect to the process $(\tilde{z}_k)_{k\in\mathbb{N}}$ and $\mathcal{F} = \{\mathcal{F}_k\}_{k\in\mathbb{N}}$ the natural filtration with respect to the process $(z_k)_{k\in\mathbb{N}}$ where $z_k \sim q(\cdot|y;\theta_{k-1})$ for all k. Consider the following assumptions which are generalization of the ones of [2]:

- (SAEM1) For all $k \in \mathbb{N}$, $\gamma_k \in [0,1]$, $\sum_{k=1}^{\infty}\gamma_k = \infty$ and $\sum_{k=1}^{\infty}\gamma_k^2 < \infty$;
- (SAEM2) The functions $\psi\colon \Theta \to \mathbb{R}$ and $\phi\colon \Theta \to \mathcal{S}$ are m times differentiable;
- (SAEM3*) For all positive Borel functions ϕ, for all $k \in \mathbb{N}$ and all $y \in \mathcal{Y}$,

$$\mathbb{E}\big[\phi(Z_{k+1})\big|\widetilde{\mathcal{F}}_k\big] = \int_{\mathcal{Z}} \phi(z)\tilde{q}_k(z;\theta_k)\,\mathrm{d}\mu(z)$$

and

$$\mathbb{E}\big[\phi(Z_{k+1})\big|\mathcal{F}_k\big] = \int_{\mathcal{Z}} \phi(z)q_k(z|y;\theta_k)\,\mathrm{d}\mu(z)\,;$$

- (SAEM4*) For all $\theta \in \Theta$, all $y \in \mathcal{Y}$ and all $k \in \mathbb{N}$,

$$\int_{\mathcal{Z}} \|S(y,z)\|^2\,\tilde{q}_k(z;\theta)\,\mathrm{d}\mu(z) < +\infty\,.$$

Assumption (SAEM1) is characteristic of stochastic approximation procedures in which the step-size have to decrease not too fast. Like Assumption (M1*), (SAME3*) is similar to (SAEM3), except that we assume that, given $\theta_0,\dots,\theta_k$, *both* simulated latent variables $\tilde{z}_1,\dots,\tilde{z}_k$ and $z_1,\dots,z_k$ are conditionally independent, given their respective natural filtration. In Assumption (SAEM4*), we demand the integrability of $z \mapsto \|S(y,z)\|^2$ with respect to the measures $\tilde{q}_k(z;\theta)\,\mathrm{d}\mu$.

The following theorem ensures the convergence of our new stochastic approximation version of the EM algorithm. This theorem is the approximated counterpart of Theorem 5 of [2]. Let $\ell\colon \Theta \to \mathbb{R}$ defined as, for all $y \in \mathcal{Y}$,

$$\ell\colon \theta \mapsto \int_{\mathcal{Z}} q(y,z;\theta)\,\mathrm{d}\mu(z)\,.$$

Theorem 2.1 (Convergence of the approximated-SAEM)**.** *Assume that (M1*), (M2-5), (SAEM1), (SAEM2), (SAME3*) and (SAEM4*) hold. Assume in addition that:*

(A) *For all $y \in \mathcal{Y}$, the sequence $\big(\tilde{q}_k(\cdot;\theta)\big)_{k\in\mathbb{N}}$ converge in mean on every compact subset of Θ for the measure $S.\mu$ to $q(\cdot|y;\theta)$, that is to say for all observations $y \in \mathcal{Y}$ and all compact $\mathcal{K} \subset \Theta$,*

$$\lim_{k\to\infty}\Big\{\sup_{\theta\in\mathcal{K}}\int_{\mathcal{Z}} S(y,z)\big(\tilde{q}_k(z;\theta) - q(z|y;\theta)\big)\,\mathrm{d}\mu(z)\Big\} = 0\,;$$

(B) *With probability 1, clos $(\{s_k\}_{k\in\mathbb{N}^*})$ is a compact subset of $\mathcal{S}$.*

Let $\mathcal{L} = \{\theta \in \Theta | \partial_\theta \ell(\theta) = 0\}$. Then, with probability 1,

$$\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0\,.$$

Hypothesis (A) makes explicit what we mean by sequence of approximated densities. In particular, it allows a wide variety of numerical schemes; we propose an example of practical interest in Section 2.2. Note that (SAEM4*) and (A) ensure the function $z \mapsto \|S(y,z)\|^2$ to be integrable with respect to the measure $q(y,z\theta)\,\mathrm{d}\mu$.

In practice, checking the compactness condition (B) may be intractable. In that case, we have to recourse to a stabilization procedure. We proceed as in [12]. Let $(\mathcal{K}_n)_{n\in\mathbb{N}}$ be an exhaustion by compact sets of the space $\mathcal{S}$, *i.e.* be a sequence of compact subsets of $\mathcal{S}$ such that

$$\bigcup_{n\in\mathbb{N}} \mathcal{K}_n = \mathcal{S} \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \mathcal{K}_n \subset \mathrm{int}(\mathcal{K}_{n+1})$$

where $\mathrm{int}(A)$ denotes the interior of the set A. The main idea is to reset the sequence s_k to an arbitrary point every time s_k wanders out of the compact subset $\mathcal{K}_{n_k}$, where n_k is the number of projections up to the k-th iteration. Let $\varepsilon = (\varepsilon_k)_{k\in\mathbb{N}}$ be a monotone non-increasing sequence of positive numbers and let K be a subset of $\mathcal{Z}$. Last, let $\Pi\colon \mathcal{Z}\times\mathcal{S} \to K \times \mathcal{K}_0$ be a measurable function (See [12] for details about the way to choose Π). The stochastic approximation with truncation on random boundaries summarizes as:

Fig. 1. Stochastic approximation with truncation on random boundaries

Set $n_0 = 0$, $s_0 \in \mathcal{K}_0$ and $\tilde{z}_0 \in K$
for all $k \in \mathbb{N}$ **do**
  Sample $\tilde{z}^* \sim \tilde{q}_k(\cdot;\theta_{k-1})$
  Compute $s^* = s_{k-1} + \gamma_k\big(S(y,\tilde{z}^*) - s_{k-1}\big)$
  if $s^* \in \mathcal{K}_{n_{k-1}}$ **then**
    Set $(\tilde{z}_k, s_k) = (\tilde{z}^*, s^*)$ [
  else
    Set $(\tilde{z}_k, s_k) = \Pi(\tilde{z}_{k-1}, s_{k-1})$ and $n_k = n_{k-1} + 1$
    Set $\theta_k = \hat{\theta}(s_k)$]
  end if
end for

The proof of the theorem consists in applying the (recalled in Appendix A) theorem 2 of [2]. In particular, (SA0-4) refer to their hypothesis (SA0-4). For sake of simplicity, we prove the convergence of the approximated-SAEM under the compactness condition (B). However, the result remains true even if (B) is not satisfied, on condition of having recourse to this truncation on random boundaries procedure.

Proof. As for all $k \in \mathbb{N}$, $\gamma_k \in [0,1]$, (SA0) is verified under (M1⋆) and (SAEM1). Moreover, (SA1) is implied by (SAEM1) and (SA3) by (B). Note that under Assumption (B), there exists, with probability 1, a compact set K such that for all $k \in \mathbb{N}$, $s_k \in K$.

Let, for all $s \in \mathcal{S}$ and $k \in \mathbb{N}$, $h(s) = \bar{s}(\hat{\theta}(s)) - s$,

$$e_k = S(y, \tilde{z}_k) - \mathbb{E}\left[S(y, \tilde{z}_k)|\widetilde{\mathcal{F}}_{k-1}\right]$$
$$\text{and} \quad r_k = \mathbb{E}\left[S(y, \tilde{z}_k)|\widetilde{\mathcal{F}}_{k-1}\right] - \bar{s}(\hat{\theta}(s_{k-1}))$$

such that Equation (1) writes on Robbins-Monro type approximation procedure.

As Lemma 2 of [2] depends only of the meanfield of the model, it can be applied as it is. More precisely, (SA2.i) is satisfied with the Lyapunov function $V = -\ell \circ \hat{\theta}$ and

$$\{s \in \mathcal{S}|F(s) = 0\} = \{s \in \mathcal{S}|\partial_s V(s) = 0\}\,,$$
$$\hat{\theta}\left(\{s \in \mathcal{S}|F(s) = 0\}\right) = \{\theta \in \Theta|\partial_\theta \ell(\theta) = 0\} = \mathcal{L}\,.$$

Moreover, (SA2.ii) is satisfied due to the Sard theorem and (SAEM2). We only need to focus on (SA4).

Set for all $n \in \mathbb{N}^*$, $E_n = \sum_{k=1}^n \gamma_k e_k$. The sequence $(E_n)_{n\in\mathbb{N}^*}$ is a $\widetilde{\mathcal{F}}$-martingale: for all $m > n$, $\mathbb{E}\big[E_m|\widetilde{\mathcal{F}}_n\big] = E_n$ as for all $k > n$, $\widetilde{\mathcal{F}}_n \subset \widetilde{\mathcal{F}}_{k-1}$. Moreover, for all $n \in \mathbb{N}$,

$$\mathbb{E}\left[\left\|S(y, \tilde{z}_{n+1}) - \mathbb{E}\left[S(y, \tilde{z}_{n+1})|\widetilde{\mathcal{F}}_{n+1}\right]\right\|^2 \Big| \widetilde{\mathcal{F}}_n\right] \leqslant \mathbb{E}\left[\left\|S(y, \tilde{z}_{n+1})\right\|^2 \Big| \widetilde{\mathcal{F}}_n\right] < \infty \text{ a.s.}$$

since by (B) and (M5), with probability 1, $\hat{\theta}(s_n)$ is in the compact set $\hat{\theta}(K) \subset \Theta$. So,

$$\sum_{n=1}^{\infty} \mathbb{E}\left[\left\|E_{n+1} - E_n\right\|^2 \Big| \widetilde{\mathcal{F}}_n\right] \leqslant \sum_{n=1}^{\infty} \gamma_{n+1}^2\, \mathbb{E}\left[\left\|S(\tilde{z}_{n+1})\right\|^2 \Big| \widetilde{\mathcal{F}}_n\right] < \infty \text{ a.s.}\,.$$

According to Theorem 2.15 of [13], with probability 1, $\lim_{n\to\infty} E_n$ exists. Moreover,

$$r_n = \int_Z S(y, z)\left(q(z|y, \hat{\theta}(s_{n-1})) - \tilde{q}_n(z, \hat{\theta}(s_{n-1}))\right) \mathrm{d}\mu(z)$$

for all $n \in \mathbb{N}$, which converge to 0 according to hypothesis (A), proving (SA4).

Thus, Theorem 2 of [2] applies and

$$\limsup_{k\to\infty} d(s_k, \{s \in \mathcal{S}|\partial_s V(s) = 0\}) = \limsup_{k\to\infty} d(s_k, \{s \in \mathcal{S}|F(s) = 0\}) = 0\,.$$

Lastly, by continuity of $\hat{\theta}\colon \mathcal{S} \to \Theta$,

$$\limsup_{k\to\infty} d\left(\hat{\theta}(s_k), \hat{\theta}\big(\{s \in \mathcal{S}|F(s) = 0\}\big)\right) = \limsup_{k\to\infty} d(\theta_k, \mathcal{L}) = 0\,.$$

□

The obtained results demonstrate that, under appropriate conditions, the sequence $(\theta_k)_{k\in\mathbb{N}}$ converges to a connected component of the set $\mathcal{L}$ of stationary points of ℓ. Moreover, some conditions upon which the convergence toward local *maxima* is guaranteed are given in Section 7 of [2]. As this conditions only depend on the design of the model and not on the definition of the optimizing sequence $(\theta_k)_{k\in\mathbb{N}}$, the corresponding theorems remain exact in our context leading to classical hypothesis ensuring convergence toward local *maxima*.

2.2 A Tempering Version of the SAEM

We focus in the following on an instantiation of the approximated-SAEM, leading to the *tempering-SAEM*. Let $T = (T_k)_{k\in\mathbb{N}}$ be a sequence of positive numbers such that $\lim_{k\to\infty} T_k = 1$. We set, for all $y \in \mathcal{Y}$, all $z \in \mathcal{Z}$, all $\theta \in \Theta$ and all $k \in \mathbb{N}$,

$$\tilde{q}_k(z; \theta) = \frac{1}{c_\theta(T_k)}\, q(z|y; \theta)^{1/T_k}$$

where $c_\theta(T_k)$ is a scaling constant.

Let $y \in \mathcal{Y}$ and $\mathcal{K} \subset \Theta$ compact. Then, by continuity of $q(z|y;\cdot)$, it exists $M \in \mathbb{R}$ such that

$$\sup_{\theta\in\mathcal{K}} \left|S(y, z)\left(\tilde{q}_k(z; \theta) - q(z|y; \theta)\right)\right| \leqslant \sup_{\theta\in\mathcal{K}} M\left|1 - \frac{1}{c_\theta(T_k)} \exp\left(-\Big(1 - \frac{1}{T_k}\Big) q(z|y, \theta_k)\right)\right|\,.$$

Thus, as $\mathcal{K}$ is compact, (A) is satisfied.

Note that our tempering-SAEM differs from the simulated annealing version of [4] as we do not modify the model but only the sampling-step of the estimation algorithm.

2.2.1 Escape Local Maxima

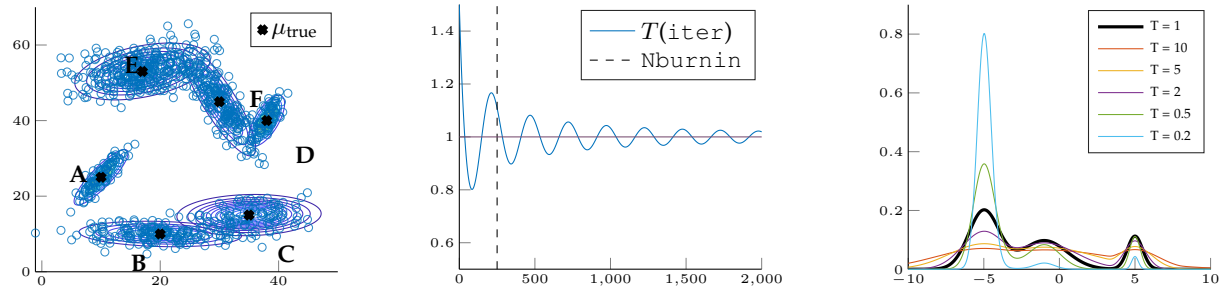
This scheme has been built with the intuition of the simulated annealing: the sequence T has to be interpreted as a sequence of temperatures. The higher T_k, the more the corresponding distribution $\tilde{q}_k$ lies flat and the (approximated) hidden variable z_k is able explore all the set $\mathcal{Z}$. On the contrary, a low temperature will freeze the exploration of z_k (see Figure 2c). Thus, finding an appropriate sequence T to keep a balance between both behaviors is a great methodological challenge.

We propose here an oscillatory tempering pattern: we start from a high temperature and then we oscillate around one with decreasing amplitude. In other words, given an (high) initial temperature T_0, the decreasing and amplitude rate a, b and the delay r, we define our sequence of temperatures by for all $k \in \mathbb{N}$,

$$T_k = \tanh\left(\frac{k}{2r}\right) + \left(T_0 - \frac{2\sqrt{2}}{3\pi}\, b\right) \times a^{k/r} + b\,\frac{\sin(\kappa)}{\kappa}$$

where $\kappa = \frac{k}{r} + \frac{3\pi}{4}$. We design this scheme to decrease, with an exponential rate, from T_0 to 1, with dampened oscillations.

Due to the oscillations of the temperature, the latent variable z_k will explore and gather in turns, leading to the possibility to switch from one mode to the other in a multimodal density during heating and explore these modes during cooling steps. In this way, the local *maxima* of the likelihood can be avoided, especially during the firsts iterations. Moreover, as the approximated distributions regularly gather around the modes of the posterior



(a) Dataset (b) Tempering scheme (c) Tempering distributions

Fig. 2. *Applying the tempering-SAEM to Gaussian mixture model.* Fig. 2a: Learning dataset for the multivariate GMM (see Section 3.1). Fig. 2b: Evolution of the temperature over iteration for the tempering-SAEM. Fig. 2c: Influence of the temperature over the pattern of the distribution.

distribution $q(\cdot|y;\theta_k)$, the exploration of z will stabilize and the algorithm will converge.

Although the analysis of this algorithm is heuristic, the simulations (see the following section and Figure 3) confirm the intuition and give good results. A theoretical analysis is an ongoing problem.

3 Application and Experiments

As explained in the previous paragraph, the tempering-SAEM allows us to escape from local *maxima*. To illustrate this phenomenon, we propose two applications: cluster analysis through Gaussian mixture model and independent factor analysis which can lead to blind source separation [11], [14], [15].

3.1 Multivariate Gaussian Mixture Models

Before considering a more realistic application, we first present an application of the tempering-SAEM to multivariate Gaussian mixture model (GMM). Actually, in spite of an apparent simplicity, this model illustrates well the main features of our algorithm.

Let $y = (y_i)_{i\in[\![1,n]\!]} \in \mathbb{R}^{nd}$ be a n-sample of $\mathbb{R}^d$. We assume that y is distributed under a a weighted sum of m d-dimensional Gaussians: Given $\alpha = (\alpha_j)_{j\in[\![1,m]\!]}[0,1]^m$ such that $\sum_{j=1}^m \alpha_j = 1$, $\mu = (\mu_j)_{j\in[\![1,m]\!]} \in \mathbb{R}^{md}$ and $\Sigma = (\Sigma_j)_{j\in[\![1,m]\!]} \in (\mathcal{S}_d\mathbb{R})^m$, we assume that

$$y|z,\theta \sim \bigotimes_{i=1}^{n} \mathcal{N}(\mu_{z_i}, \Sigma_{z_i}) \quad \text{and} \quad z|\theta \sim \sum_{j=1}^{m} \alpha_j \delta_j$$

where $\theta = (\alpha, \mu, \Sigma)$ and $z = (z_i)_{i\in[\![1,n]\!]}$ is the latent variable specifying the identity of the mixture component of each observation. In the following, we compare the efficiency of the EM, the SAEM and the tempering-SAEM algorithms to produce a *maximum* likelihood estimate of the parameters with the *a priori* given exact number of components m.

Classically, as closed-form expressions are possible for finite GMM, the EM algorithm is a very popular technique used to produce the *maximum* likelihood estimation of the parameters [16]. However, the computational cost can be prohibitive. A faster procedure is to use the SAEM algorithm. Nevertheless, both algorithms are very sensitive to the initial position: solutions can highly depend on their starting point and consequently produce sub-optimal *maximum* likelihood estimates [17]. The tempering-SAEM appears as a way to escape from local *maxima* and reach global *maxima* more often.

To quantify this assertion, we have generated a synthetic dataset (Figure 2a) and performed the estimation 500 times for the three algorithms. The relative errors for α and μ and the KullbackLeibler divergence between the true covariance matrices Σ and the estimated one are compiled in Figures 3b, 3d and 3f. The class refer to the ones of Figure 2a. We consider the algebraic relative error for α so that we can deduce if the studied algorithm tend to empty (class E) or overfill (class B) the classes. First thing to remark is that the tempering-SAEM is always competitive with the EM and the SAEM and most of the time greater. That is to say that the global *maximum* is more often reached while tempering the posterior distribution. Moreover, while EM and SAEM achieve fairly identical results, the tempering-SAEM is able to discriminate overlapped classes. Class A, which is the only isolated class, is seemingly the best learned. The EM and SAEM seem to empty the class C for the benefit of the class B and merge them together on a "super-class" as if there were only 5 components in the Gaussian mixture.

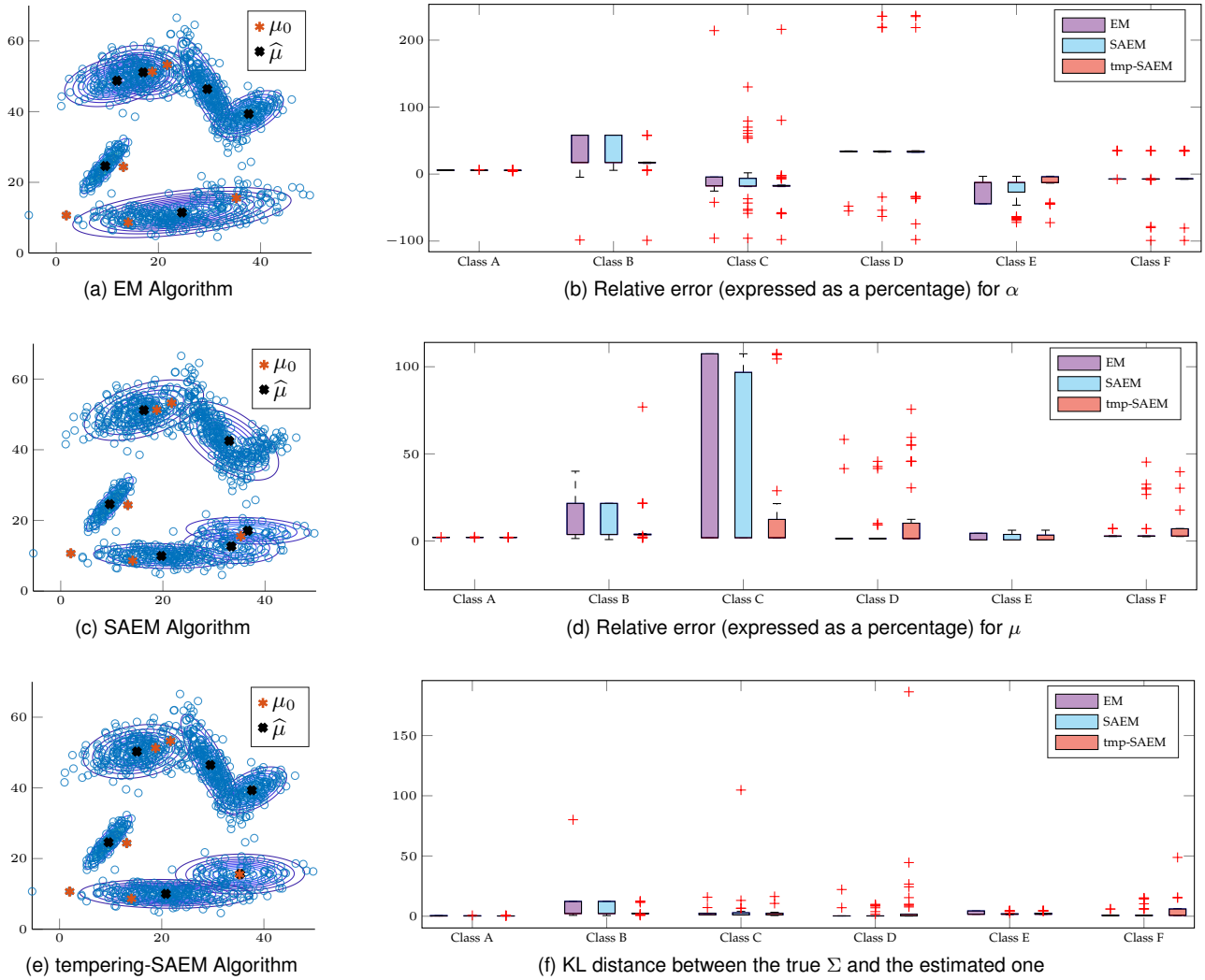
The three procedures are detailed in Appendix B.

3.2 Independent Factor Analysis

The decomposition of a sample of multi-variable data on a relevant subspace is a recurrent problem in many different fields from source separation problem in acoustic signals to computer vision and medical image analysis. Independent component analysis has become one of the standard approaches. This technique relies upon a data augmentation scheme, where the (unobserved) input are viewed as the missing data. We observe multivariable data y which are measured by n sensors and supposed to arise from m source signals x, that are linearly mixed together by some linear transformation H, and corrupted by an additive Gaussian noise ε. Simply put, we observe $y = (y^{(t)})_{t\in[\![1,T]\!]}$ where each measurement is a point of $\mathbb{R}^n$ and assumed to be given by $y^{(t)} = Hx^{(t)} + \varepsilon^{(t)}$ where $H \in \mathcal{M}_{n,m}\mathbb{R}$, $x^{(t)} \in \mathbb{R}^m$ and $\varepsilon^{(t)} \overset{i.i.d}{\sim} \mathcal{N}(0, \lambda I_n)$, $\lambda \in \mathbb{R}$. The suitability of the SAEM algorithm in this context has been demonstrated in [14] and [15]. We propose here to modify the learning principle to make the procedure less susceptible to trapping states.

As in [14] and [11], we assume that:

1) $(x^{(t)})_{t\in[\![1,T]\!]}$ and $(\varepsilon^{(t)})_{t\in[\![1,T]\!]}$ are independent;



(a) EM Algorithm

(b) Relative error (expressed as a percentage) for α

(c) SAEM Algorithm

(d) Relative error (expressed as a percentage) for μ

(e) tempering-SAEM Algorithm

(f) KL distance between the true Σ and the estimated one

Fig. 3. *Multivariate Gaussian mixture model.* Figs. 3a, 3c and 3e: Qualitative comparison of the *maximum* likelihood estimation of the parameters. The estimation is performed with the same initial points (in orange). Figs. 3b, 3d and 3f: Relative error (expressed as a percentage) for the weights α and the centrods μ. Kullback-Leibler distance between the true covariance matrices Σ and the estimated ones, for 500 runs and $n = 1000$.

2) $(x^{(t)})_{t\in[\![1,T]\!]}$ is an i.i.d sequence of random vectors, with independent component. Each component $x_i^{(t)}$ is given by a mixture of k Gaussians indexed by $z_i^{(t)} \in [\![1,k]\!]$ with means $\mu_{z_i(t)}$, variances $\sigma^2_{z_i(t)}$ and mixing proportions $\alpha_{z_i(t)}$:

$$q(x_i^{(t)};\theta_i^{(t)}) = \sum_{z_i^{(t)}=1}^{k} \alpha_{z_i(t)}\, \mathcal{G}\left(x_i^{(t)} - \mu_{z_i(t)}; \sigma^2_{z_i(t)}\right)$$

$$\theta_i^{(t)} = \left(\alpha_{z_i(t)}, \mu_{z_i(t)}, \sigma^2_{z_i(t)}\right)$$

where for all vectors x and μ and all symmetric matrix Σ, $\mathcal{G}(x-\mu,\Sigma)$ refers to the (multivariate) Gaussian distribution.

This model is called independent factor analysis (IFA). The problem is to find the value of the parameter $W = (H, \lambda, \theta)$ given y. Identifiability in this model is discussed in [18]. Basically, the sources are defined only to within an order permutation and scaling. To avoid trivialities, we fix the variances $(\sigma_j^2)_{j\in[\![1,k]\!]}$ to one [15]. Note that this definition of the IFA model is somewhat less general that the one introduced by Attias [11] in which the components are supposed to be independent but not necessarily identically distributed. Nevertheless, it has been shown that restrictive IFA models can perform well in practice [15].

The likelihood of the IFA can be put in exponential form using the sufficient statistics, for all $j \in [\![1,k]\!]$,

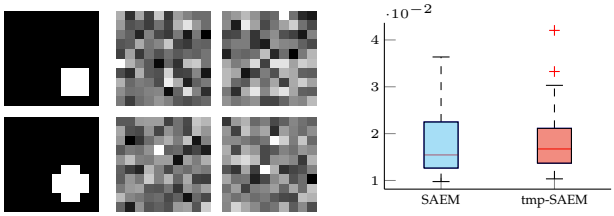
$$S_{1,j}(x,y,z) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{z_i=j\}}\,; \qquad S_4(x,y,z) = y^{\,t}y\,;$$

$$S_{2,j}(x,y,z) = \frac{1}{m}\sum_{i=1}^{m} x_i\,\mathbb{1}_{\{z_i=j\}}\,; \qquad S_5(x,y,z) = y^{\,t}x\,;$$

$$S_{3,j}(x,y,z) = \frac{1}{m}\sum_{i=1}^{m} x_i^2\,\mathbb{1}_{\{z_i=j\}}\,; \qquad S_6(x,y,z) = x^{\,t}x\,.$$

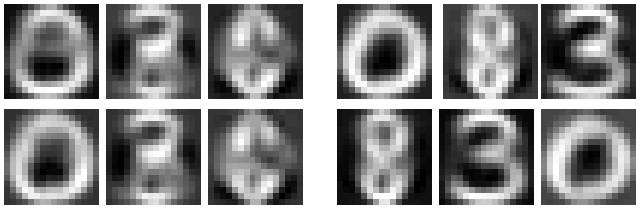
The M-step is then given by

$$H = [S_5]\,([S_6])^{-1}\;;\quad \alpha = [S_1]\;;\quad \mu = \frac{[S_2]}{[S_1]}\;;\quad \sigma^2 = \mathbf{1}_k\,;$$
$$\lambda = \|\,[S_6]\,\|_2^2 - 2\langle\, H \mid [S_5]\,\rangle + \langle\, {}^tHH \mid [S_6]\,\rangle$$



(a) Decomposition images (left) and samples of the dataset (b) KL distances

Fig. 4. *Independent factor analysis – BG-ICA.* Kullback-Leibler distance between the source matrix H used to build the dataset and the estimated one. The dataset consists of 100 images distributed in accordance with the two-components Bernouilli-Gaussian model build from the square and the cross binary images.



(a) SAEM (b) tempering-SAEM

Fig. 5. *Independent factor analysis – USPS dataset.* Results of the independent factor estimation on a balanced mix of digits 0, 3 and 8 from the USPS database. The dataset is composed of 50 samples of each digits.

where $\mathbf{1}_k$ stands for the k-vector off all 1 and the brackets denote the empirical-average. Moreover, it is possible to compute the conditional distribution of the hidden variable (x, z) given observed values of y and the E-step can be computed exactly [11]: For all $\zeta \in [\![1, k]\!]^m$,

$$\mathbb{P}(z = \zeta | y; W) = \frac{\alpha_\zeta \, \mathcal{G}\left(y - H\mu_\zeta; H\Delta_\zeta \, {}^tH + \lambda I_n\right)}{\sum_z \alpha_z \, \mathcal{G}\left(y - H\mu_z; H\Delta_z \, {}^tH + \lambda I_n\right)}$$

$$\text{and} \quad q(x|y, z; W) = \mathcal{G}\left(x - \nu_{y,z}; \Sigma_z\right)$$

where

$$\alpha_z = \prod_{i=1}^{m} \alpha_{z_i} \; ; \quad \mu_z = (\mu_{z_i})_i \; ; \quad \Delta_z = \mathcal{D}iag\left((\sigma^2_{z_i})_i\right) \; ;$$

$$\Sigma_z = \left(\frac{1}{\lambda} \, {}^tHH + \Delta_z^{-1}\right)^{-1} \; ; \; \nu_{y,z} = \Sigma_z \left(\frac{1}{\lambda} \, {}^tHy + \Delta_z^{-1}\mu_z\right).$$

Thus, as well as for the GMM, we can compare the efficiency of SAEM vs tempering-SAEM algorithms in this context.

In Section 3.1, we were interested in the performance of our algorithm for data generated according to the true model. We relax here this assumption and observe $T = 100$ images distributed in accordance with the Bernoulli-Gaussian model (BG-ICA [15]), with two components. The components are represented as two-dimensional binary images. The first one is a black image with a white cross in the top left corner. The second one has a white square in the bottom right corner. At Figure 4, we present the two decomposition images, 4 typical observations and the Kullback-Leibler distance between the true H (in the BG-ICA model) and the estimated one for 50 runs.

This experience confirms the robustness of the tempering-SAEM. Moreover, one could have feared that the augmentation of the number of hyper-parameters due to the choice of the temperature scheme would increase the variance. Figure 4 eliminates this assumption. However, the context is very favorable to the SAEM algorithm which obtain very good and hard to outperformed results. To measure the efficiency of the tempering-SAEM, we test it on the USPS database, which contains gray-level images of handwritten digits.

We consider a balanced mix of the digits 0, 3 and 8, which consists of 50 samples for each of the three digits. We then run both the SAEM and the tempering-SAEM. We present at Figure 5 two typical runs (in line). If the two of them succeed in discriminate 0 against 3 and 8, the tempering-SAEM outperform the SAEM algorithm concerning 3 versus 8. Thus, the tempering-SAEM produces meaningful sources, which could be the result of a clustering procedure, while the SAEM runs into difficulties. Hence, this experience suggests that the tempering-SAEM can indeed escape from local *maxima* in which the SAEM can be trapped.

Finally, applying the tempering-SAEM for independent factor analysis aims to check that the advantages of the tempering-SAEM over the SAEM can improve significantly the results of maximum likelihood estimation in complex hierarchical models.

3.3 Discussion and Perspective

We propose here a new stochastic approximation version of the EM algorithm. The benefit of this general procedure is twofold: we can deal with the problem of intractable or difficult sampling in one hand and favor convergence toward global *maxima* in the other hand.

Our first contribution is theoretical with the proof of the convergence of the approximated-SAEM toward local *maxima*. This result gives an *a posteriori* justification for some existent schemes like the ABC-SAEM or MONOLIX. Moreover, our general framework is versatile enough to encompass a wide range of algorithms. Our second contribution goes this way by proposing an instantiation of this general procedure to prevent convergence toward local *maxima*, referred to as tempering-SAEM. This tempering-SAEM method is the one used in the MONOLIX software. We have applied this algorithm in both synthetic and real data frameworks and obtained improved results with respect to the state of the art algorithms in both cases.

APPENDIX A
THEOREM 2 AND LEMMA 2 OF [2]

In order our article to be more self-contained, we recall Theorem 2 and Lemma 2 of [2]. Actually, the proof of Theorem 2.1 is based on this theorem which establish the convergence of Robin-Monroe type approximation procedure, *i.e.* the convergence of sequences defined recursively as

$$\forall k \in \mathbb{N}, \quad s_k = s_{k-1} + \gamma_k \big(h(s_k) + r_k + e_k\big) \, .$$

Theorem A.2 (Delyon, Lavielle, Moulines)**.** *Assume that*

(SA0) *With probability 1, for all $k \in \mathbb{N}$, $s_k \in \mathcal{S}$.*

(SA1) $(\gamma_k)_{k\in\mathbb{N}^*}$ *is a decreasing sequence of positive numbers such that* $\sum_{k=1}^{\infty}\gamma_k=\infty$.

(SA2) *The vector field* h *si continuous on* $\mathcal{S}$ *and there exists a continuously differentiable function* $V:\mathcal{S}\to\mathbb{R}$ *such that :*

(i) *for all* $s\in\mathcal{S}$, $F(s)=\langle \mathrm{d}_s V(z)\,|\,h(s)\rangle\leqslant 0$,
(ii) *int* $(V(\mathcal{L}))=\varnothing$ *where* $\mathcal{L}=\{s\in\mathcal{S}|F(s=0)\}$.

(SA3) *With probability 1, clos* $(\{s_k\}_{k\in\mathbb{N}})$ *is a compact subset of* $\mathcal{S}$.

(SA4) *With probability 1,* $\sum\gamma_k e_k$ *exists and is finite,* $\lim r_k=0$.

Then, with probability 1, $\overline{\lim}\, d(s_k,\mathcal{L})=0$.

Lemma A.2. *Assume (M1-M5) and (SAEM2). Then (SA2) is satisfied with* $V=-\ell\circ\hat{\theta}$. *Moreover,*

$$\{s\in\mathcal{S}|F(s)=0\}=\{s\in\mathcal{S}|\,\mathrm{d}_s V(s)=0\}$$
$$\textit{and}\quad \hat{\theta}\left(\{s\in\mathcal{S}|F(s)=0\}\right)=\{\theta\in\Theta|\,\mathrm{d}_\theta\ell(\theta)=0\}$$

where $F\colon s\mapsto\langle \mathrm{d}_s V(s)\,|\,h(s)\rangle$.

APPENDIX B
MULTIVARIATE GAUSSIAN MIXTURE MODEL

We give here some details about the estimation procedure in the multivariate Gaussian mixture model. The complete log-likelihood of the GMM model is

$$\log q(y,z;\theta)=-n\log 2\pi - \sum_{j=1}^{m}\sum_{i=1}^{n}\left(\frac{1}{2}\log|\Sigma_j|-\log\alpha_j + {}^t(y_i-\mu_j)\,\Sigma_j^{-1}\,(y_i-\mu_j)\right)\mathbb{1}_{\{z_i=j\}}\,.$$

B.1 Estimation through the EM Algorithm

Let t index the current iteration. The general EM algorithm iterates the following two steps:

E-step: Compute $Q(\theta|\theta^t)=\mathbb{E}\left[\log q(y,z;\theta)|y,\theta^t\right]$;
M-step: Set $\theta^{t+1}=\mathrm{argmax}_{\theta\in\Theta}\,Q(\theta|\theta^t)$.

For all $(i,j)\in[\![1,n]\!]\times[\![1,m]\!]$, set $\tau_{i,j}=\mathbb{P}\left[z_i=j|y_i,\theta^t\right]$. Then,

$$Q(\theta|\theta^t)=-n\log 2\pi - \sum_{j=1}^{m}\sum_{i=1}^{n}\left(\frac{1}{2}\log|\Sigma_j|-\log\alpha_j + {}^t(y_i-\mu_j)\,\Sigma_j^{-1}\,(y_i-\mu_j)\right)\tau_{i,j}\,.$$

According to Bayes' rule,

$$\tau_{i,j}=\frac{\alpha_j\,\mathcal{G}(y_i-\mu_j;\Sigma_j)}{\sum_{j=1}^{m}\alpha_j\,\mathcal{G}(y_i-\mu_j;\Sigma_j)}$$

where $\mathcal{G}(y-\mu;\Sigma)$ refers to the Gaussian distribution with mean μ and covariance matrix Σ. Lastly, a straightforward computation gives

$$\alpha_j^{t+1}=\frac{1}{n}\sum_{i=1}^{n}\tau_{i,j}\quad,\quad \mu_j^{t+1}=\frac{\sum_{i=1}^{n}\tau_{i,j}y_i}{\sum_{i=1}^{n}\tau_{i,j}}$$
$$\text{and}\quad \Sigma_j^{t+1}=\frac{\sum_{i=1}^{n}\tau_{i,j}(y_i-\mu_j^{t+1})\,{}^t(y_i-\mu_j^{t+1})}{\sum_{i=1}^{n}\tau_{i,j}}\,.$$

B.2 Estimation through the SAEM Algorithm

Given a sequence of positive step-size for the stochastic approximation $\gamma=(\gamma_t)_{t\in\mathbb{N}}$, the general SAEM algorithm iterates the following two steps:

SAE-step: Sample *a* new hidden variable z^{t+1} according to the conditional distribution $q(z|y,\theta^t)$ and compute

$$Q_{t+1}(\theta)=Q_t(\theta)+\gamma_t\left(\log q(y,z;\theta^t)-Q_t(\theta)\right)\,;$$

M-step: Set $\theta^{t+1}=\mathrm{argmax}_{\theta\in\Theta}\,Q_{t+1}(\theta)$.

The GMM belongs to the curved exponential family. Actually, for all y, z and θ,

$$\log q(y,z;\theta)=-n\log(2\pi) + \sum_{j=1}^{m}\left(\log\alpha_j-\frac{1}{2}\log|\Sigma_j|+\langle\,\mu_j\,{}^t\mu_j\,|\,\Sigma_j^{-1}\,\rangle_{\mathcal{F}}\right)S_{1,j}(y,z) + \sum_{j=1}^{m}\left[\,\langle\,\Sigma_j^{-1}\,|\,S_{3,j}(y,z)\,\rangle_{\mathcal{F}}-2\langle\,\Sigma_j^{-1}\mu_j\,|\,S_{2,j}(y,z)\,\rangle\,\right]$$

where, for all $j\in[\![1,m]\!]$,

$$S_{1,j}(y,z)=\sum_{i=1}^{n}\mathbb{1}_{z_i=j}\quad;\quad S_{2,j}(y,z)=\sum_{i=1}^{n}y_i\,\mathbb{1}_{z_i=j}$$
$$\text{and}\quad S_{3,j}(y,z)=\sum_{i=1}^{n}y_i\,{}^t y_i\,\mathbb{1}_{z_i=j}\,.$$

So, the SAE-step is replaced by an update of the estimation of the conditional expectation of the sufficient statistics, namely, for all $\ell\in\{1,2,3\}$, and all j,

$$S_{\ell,j}^{t+1}=S_{\ell,j}^{t}+\gamma_t\left(S_{\ell,j}(y,z^{t+1})-S_{\ell,j}^{t}\right)$$

where, for all i, z_i^{t+1} is sampled from the discrete law $\sum_{j=1}^{m}\tau_{i,j}\delta_j$ where $\tau_{i,j}=\mathbb{P}\left[z_i=j|y_i,\theta^t\right]$ as in the EM-case.

The M-step can also be computed in close-form:

$$\alpha_j^{t+1}=\frac{1}{n}S_{1,j}\quad,\qquad \mu_j^{t+1}=\frac{S_{2,j}}{S_{1,j}}$$
$$\text{and}\qquad \Sigma_j^{t+1}=\frac{S_{3,j}-S_{2,j}\,{}^t\mu_j^{t+1}}{S_{1,j}}\,.$$

B.3 Estimation through the tmp-SAEM Algorithm

The previous computation remain true except that the hidden variables z_i^{t+1} are now sampled from the tempered conditional distribution $\frac{1}{c(T_t)}\sum_{j=1}^{m}\tau_{i,j}^{1/T_t}\delta_j$ where $c(T_t)=\sum_{j=1}^{m}\tau_{i,j}^{1/T_t}$ and T_t is defined in Section 2.2.

ACKNOWLEDGMENTS

Ce travail bénéficie d'un financement public Investissement d'avenir, réference ANR-11-LABX-0056-LMH. This work was supported by a public grant as part of the Investissement d'avenir, project reference ANR-11-LABX-0056-LMH.

REFERENCES

[1] A. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[2] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a stochastic approximation version of the EM algorithm," *The Annals of Statistics*, vol. 27, no. 1, pp. 94–128, 1999.

[3] O. Cappé, É. Moulines, and T. Rydén, *Inference in Hidden Markov Models*, ser. Springer Series in Statistics. Springer, 2005.

[4] M. Lavielle and E. Moulines, "A simulated annealing version of the em algorithm for non-gaussian deconvolution," *Statistics and Computing*, vol. 7, no. 4, pp. 229–236, 1997.

[5] M. Lavielle, *Mixed effects models for the population approach: models, tasks, methods and tools*. CRC press, 2014.

[6] M. Lavielle and F. Mentré, "Estimation of population pharmacokinetic parameters of saquinavir in hiv patients with the monolix software," *Journal of pharmacokinetics and pharmacodynamics*, vol. 34, no. 2, pp. 229–249, 2007.

[7] A. Samson, M. Lavielle, and F. Mentré, "Extension of the saem algorithm to left-censored data in nonlinear mixed-effects model: Application to hiv dynamics model," *Computational Statistics & Data Analysis*, vol. 51, no. 3, pp. 1562–1574, 2006.

[8] P. L. Chan, P. Jacqmin, M. Lavielle, L. McFadyen, and B. Weatherley, "The use of the saem algorithm in monolix software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics parameters of maraviroc in asymptomatic hiv subjects," *Journal of pharmacokinetics and pharmacodynamics*, vol. 38, no. 1, pp. 41–61, 2011.

[9] U. Picchini and A. Samson, "Coupling stochastic em and approximate bayesian computation for parameter inference in state-space models," *Computational Statistics*, vol. 33, no. 1, pp. 179–212, 2018.

[10] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder, "Approximate bayesian computational methods," *Statistics and Computing*, vol. 22, no. 6, pp. 1167–1180, 2012.

[11] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.

[12] C. Andrieu, É. Moulines, and P. Priouret, "Stability of stochastic approximation under verifiable conditions," *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 283–312, 2006.

[13] P. Hall and C. C. Heyde, *Martingale limit theory and its application*, ser. Probability and mathematical statistics. Academic Press, 1980.

[14] E. Moulines, J.-F. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *Acoustics, Speech, and Signal Processing*, vol. 5. IEEE, 1997, pp. 3617–3620.

[15] S. Allassonnière and L. Younes, "A stochastic algorithm for probabilistic independent component analysis," *The Annals of Applied Statistics*, vol. 6, no. 1, pp. 125–160, 03 2012.

[16] G. McLachlan and D. Peel, *Finite Mixture Models*, ser. Wiley Series in Probability and Statistics. Wile, 2000.

[17] C. Biernacki, G. Celeux, and G. Govaert, "Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models," *Computational Statistics & Data Analysis*, vol. 41, no. 3, pp. 561 – 575, 2003.

[18] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287 – 314, 1994, higher Order Statistics.

Stéphanie Allassonnière received her PhD degree in Applied Mathematics (2007), studies one year as postdoctoral fellow in the Center for Imaging Science, JHU, Baltimore. She joined the Applied Mathematics department of Ecole Polytechnique in 2008 as assistant professor and moved to Paris Descartes school of medicine in 2016 as Professor. Her researches focus on statistical analysis of medical databases in order to: understanding the common features of populations, designing classification, early prediction and decision support systems.

Juliette Chevallier received the graduate degree from University Paris-Sud, Orsay. She is carrying her PhD thesis in applied mathematics at the École polytechnique, Palaiseau. Her interests are ranging from fundamental subjects such as Riemannian geometry or stochastic optimization to high-dimensional statistics and application to medicine. In particular, she has worked on the statistical analysis of longitudinal manifold-valued data with application to chemotherapy monitoring.