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

Although the expectation-maximization (EM) algorithm [1] is a very popular and often efficient approach to *maximum* likelihood (or *maximum a posteriori*) estimation in incomplete data models, as it a simple use algorithm, it has one major issue : the computation of the expectation with respect to the conditional distribution. Indeed, in certain situations, the EM is not applicable because the expectation step cannot be performed in closed form. To overcome this restriction, many different options have been proposed. The first one is to replace the expectation step by a sampling of the unobserved data step. We refer to this EM version as the Stochastic EM (SEM) algorithm [2]. In particular, in the SEM, only one sample of the latent variable is drawn. A possible generalization of the SEM is the Monte-Carlo EM (MCEM) [3], in which a Monte-Carlo implementation of the expectation in the E-step is carried out. In an alternative way, Delyon, Lavielle and Moulines [4] proposed to replace the expectation step of the EM algorithm by one iteration of a stochastic approximation procedure, referred to as SAEM, standing for stochastic approximation EM. In addition to avoiding the computation of the expectation, introducing randomness may enable to escape local maxima. However, this is not yet theoretically proved nor numerically illustrated in the literature.

The convergence of the SAEM toward local *maxima* has been proved in [4] and its numerical efficiency has been demonstrated in several situations such as in inference in hidden Markov models [5]. 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 [6] 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 [7] 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 [8], [9], [10].

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, Picchini and Samson [11] have proposed to couple the SAEM algorithm to an approximate Bayesian computation step (ABC, see [12] 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, [11] do not provide any theoretical guarantee of its convergence. More broadly, when sampling from the posterior distribution is prohibitive, one may want to shift to variational inference

- *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

Manuscript received April 19, 2005; revised August 26, 2015.

[13], [14], [15].

Behind variational inference, the main idea is to replace the objective function by a minorant function which is a trade off between a likelihood and a Kullback-Leibler (KL) divergence between the conditional distribution and a parametric probability density function. This minimizing function is then optimized by a stochastic gradient descent. These methods are known to converge toward local minima for bounded parameters or positive objective functions.

We propose here a new stochastic approximation version of the EM algorithm where we do not sample from the exact distribution but rather from a distribution which converges to the conditional one along the algorithm iterations. This new procedure allows us to derive a wide class of SAEM-like algorithms, including the "trick" initialized SAEM of [7] and the ABC-SAEM algorithms, to cope with intractable or difficult sampling. We refer to this new algorithm as the approximated-SAEM.

This general framework allows us to build a procedure, with the thought of the simulated annealing version of the SAEM [6], to favor convergence toward the global *maxima*. We introduce a sequence of temperatures and sample from a tempered version of the conditional distribution. Therefore, the conditional 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 Lavielle and Moulines [6] as we do not modify the model but only the sampling-step.

In Section 2, we introduce our new stochastic approximation version of the EM algorithm, namely the approximated-SAEM, and prove the convergence of this algorithm toward local *maxima* under usual assumptions. The demonstration of the convergence, by its similarity with the proof of the convergence of the SAEM, highlights the unconstraining nature of the different assumptions and therefore the great applicability of our algorithm. Then, we provide a theoretical study of the convergence of the tempering-SAEM toward local *maxima*. We also give an heuristic to the convergence toward "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 [16]. 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. This averaging concerns the conditional expected log-likelihood

$$\theta \mapsto Q(\theta|\theta_k) = \int_{\mathcal{Z}} \log q(y,z|\theta) q(z|y,\theta_k)\ \mathrm{d}z\,,$$

where θ_k denotes the current optimal parameter. Starting from $Q_0(\theta) = 0$ for all $\theta \in \Theta$, we build an approximation of $\theta \mapsto Q(\theta|\theta_k)$ through stochastic approximation of the compete log-likelihood. We denote Q_k this approximation. 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 [4]: an hypothesis stated with a $(^\star)$ means that it is a direct generalization of the corresponding one in [4]; 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

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

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,

$$\text{for all } y \in \mathcal{Y}, \quad \ell\colon \theta \mapsto \int_{\mathcal{Z}} q(y,z;\theta)\, \mathrm{d}\mu(z)$$
$$\text{and} \quad L\colon (s,\theta) \mapsto -\psi(\theta) + \langle\, s \,|\, \phi(\theta)\,\rangle\,.$$

Note that if the function ℓ is defined as above, it implicitly depends on y. However, as the observations y are assumed to be i.i.d, the counterpart function that would be defined on the whole set $\mathcal{Y}$ would be a sum over y of the defined ℓ.

(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. Let s_k denote the k^{th} approximation of the sufficient statistics. Then, the k-th iteration of the approximated-SAEM summarizes in:

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

where s_k is initialized to zero: $s_0(\theta) = 0$ for all $\theta \in \Theta$.

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. Let, for all set $\mathcal{X}$, $\mathrm{clos}(\mathcal{X})$ denotes the closure of $\mathcal{X}$ and consider the following assumptions which are generalization of the ones of [4]:

(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 $\theta_0, \ldots, \theta_k$, *both* simulated latent variables $\tilde{z}_1, \ldots, \tilde{z}_k$ and $z_1, \ldots, z_k$ 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 [4].

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 $\big(\tilde{q}_k(\cdot;\theta)\big)_{k\in\mathbb{N}}$ converges 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 [17]. 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 [17] 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

Note that the statement of Theorem 2.1 is very similar to the corresponding one of [4], namely Theorem 5 which establish the convergence of the SAEM. In other words, approximate the posterior distribution in the S-step does not require supplementary considerations to still guarantee the convergence of the sequence $(\theta_k)_{k\in\mathbb{N}}$. Thus, the scope of application of the approximated-SAEM algorithm is at least as unrestrictive as the one of the SAEM.

The proof of the theorem consists in applying Theorem 2 of [4]. We recall this theorem in Appendix A. In particular, (SA0 - 4) refer to their hypothesis (SA0 - 4). Moreover, since it is used in our demonstration, we also recall Lemma 2 from the same paper. 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 (Algorithm 1).

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 [4] (see Appendix A) depends only of the meanfield of the model, it can be applied as it is. Thus, (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 $n \in \mathbb{N}$,

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

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

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

According to Theorem 2.15 of [18], 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 [4] applies and

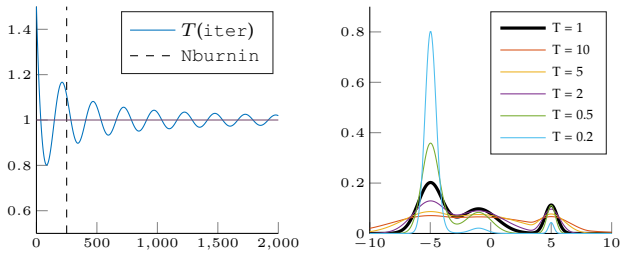
$$\limsup_{k\to\infty} d(s_k, \{s \in \mathcal{S} | \partial_s V(s) = 0\}) = \limsup_{k\to\infty} d(s_k, \{s \in \mathcal{S} | F(s) = 0\}) = 0\,.$$

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.



(a) Tempering scheme (b) Tempering distributions

Fig. 2. *Construction of the temperature scheme.* Fig. 2a: Evolution of the temperature over iteration for the tempering-SAEM. Fig. 2b: Influence of the temperature over the pattern of the distribution.

Moreover, some conditions upon which the convergence toward local *maxima* is guaranteed are given in Section 7 of [4]. 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 $(T_k)_{k\in\mathbb{N}}$ 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 the function $\theta \mapsto q(z|y;\theta)$, it exists $M \in \mathbb{R}$ such that

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

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

Note that our tempering-SAEM differs from the simulated annealing version of [6] 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 $(T_k)_{k\in\mathbb{N}}$ 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 2b). Thus, finding an appropriate sequence $(T_k)_{k\in\mathbb{N}}$ to keep a balance between both behaviors is a great methodological challenge.

We propose here an oscillatory tempering pattern which oscillate around one with decreasing amplitude. In other words given the decreasing and amplitude rate a and b, the scaling-parameter r and the delay c, we define our sequence of temperatures by: for all $k \in \mathbb{N}$,

$$T_k = 1 + a^\kappa + b\,\frac{\sin(\kappa)}{\kappa}\,, \quad \text{where} \quad \kappa = \frac{k + c\times r}{r}\,.$$

We design this scheme to decrease, with an exponential rate toward 1, with dampened oscillations. In this form, the tempering scheme includes the tempering scheme used by MONOLIX. Even so, the experiments conducted in section 3 tend to show that the exponential decrease is not necessary. In particular, a is set to zero for all the experiments.

In order the tempering scheme to converge toward 1, we just need to require that the exponential rate $a \in [0,1[$. In particular, the parameter b can be chosen independently negative or positive. A positive b will flattened the distribution at the beginning of the optimization procedure. On the contrary, a negative b will make the profile of the distribution look more prickly. It can be interesting to enforce the distinction of two close modes, as in Section 3.1.2.

Due to the oscillations of the temperature, the latent variable z_k will explore and gather in turns. Thus, in case of multimodal density, the latent variable will be able to switch from one mode to an other during the heating steps and to explore these same modes during the cooling phases. In particular, during the optimization, the tempering-SAEM may escape from local minima in which the SAEM would get stuck. Figures 13 and 14 effectively illustrate this phenomenon. In this way, the local *maxima* of the likelihood can be avoided. Moreover, as the approximated distributions regularly gather around the modes of the posterior 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) 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 [16], [19], [20].

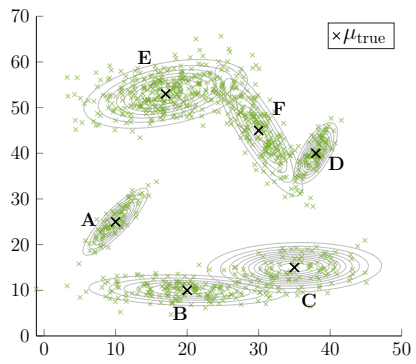
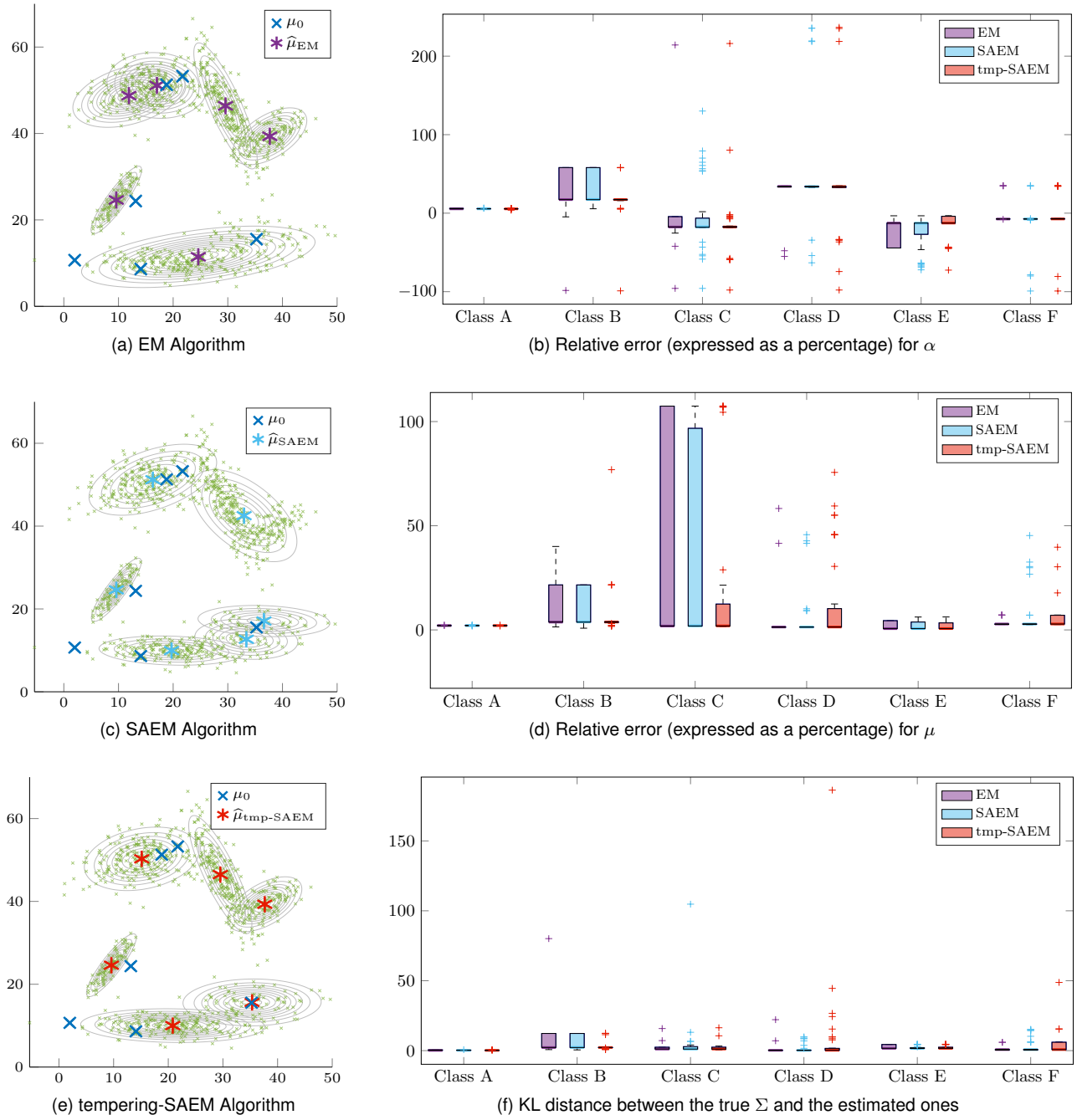


Fig. 3. : *Learning dataset* used to perform the experience regarding Section 3.1.1.



(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 ones

Fig. 4. *Performance of the estimation for the Multivariate Gaussian mixture model.* Figs. 4a, 4c and 4e: Qualitative comparison of the *maximum* likelihood estimation of the parameters. The estimation is performed with the same initial points (the blue crosses). Figs. 4b, 4d and 4f: Relative error (expressed as a percentage) for the weights α and the centroïds μ. Kullback-Leibler distance between the true covariance matrices Σ and the estimated ones, for 500 runs and $n = 1000$.

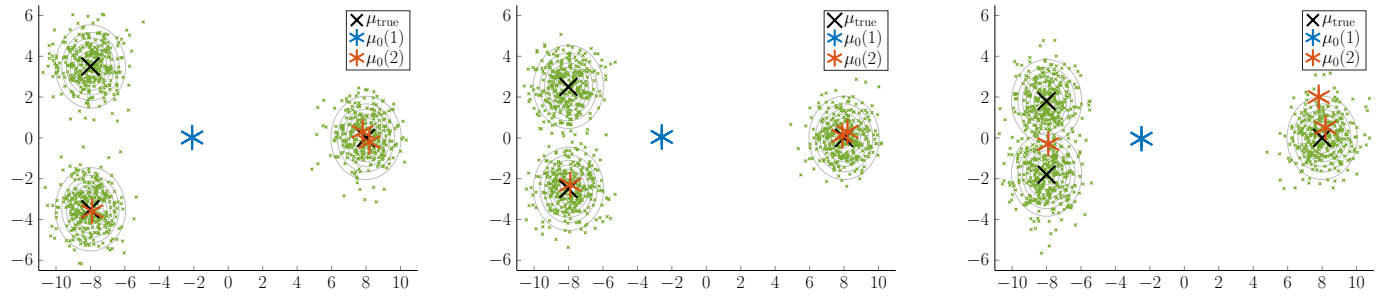
3.1 Multivariate Gaussian Mixture Models

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 the weights $\alpha = (\alpha_j)_{j\in[\![1,m]\!]}[0,1]^m$ such that $\sum_{j=1}^m \alpha_j = 1$, the centroids $\mu = (\mu_j)_{j\in[\![1,m]\!]} \in \mathbb{R}^{md}$ and the covariance matrices $\Sigma = (\Sigma_j)_{j\in[\![1,m]\!]} \in (\mathcal{S}_d\mathbb{R})^m$, we assume that

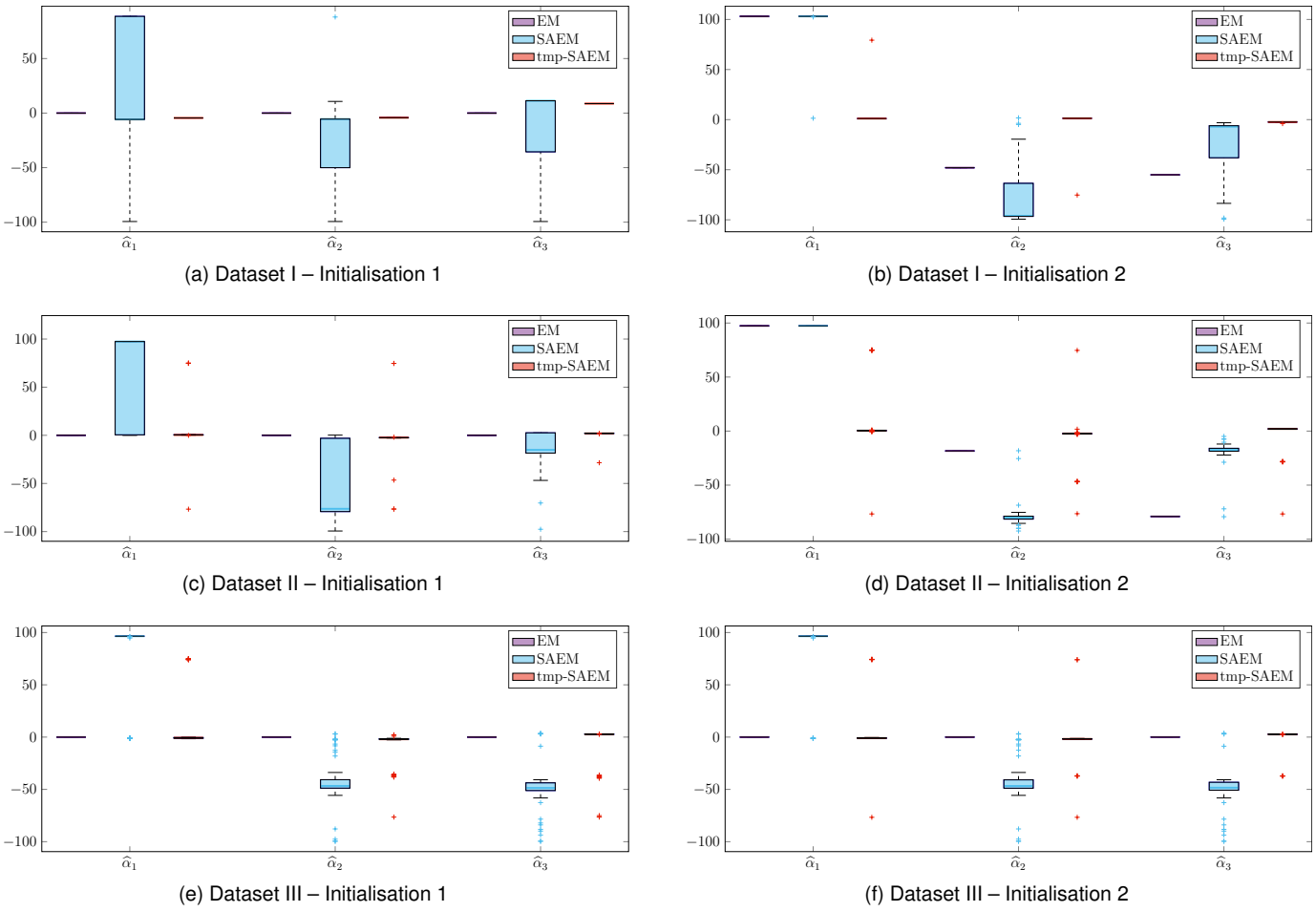
$$y|z,\theta \sim \bigotimes_{i=1}^{n} \mathcal{N}(\mu_{z_i}, \Sigma_{z_i}) \quad \text{and} \quad z|\theta \sim \sum_{j=1}^{m} \alpha_j \delta_j \,,$$

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.



(a) Dataset I (b) Dataset II (c) Dataset III

Fig. 5. *The three scattered clusters' datasets* used to perform the experiences regarding Section 3.1.2. For each dataset, we consider two possible initial positions for the means μ: either all at the barycenter of the dataset (the blue asterisk) or two of them in the single right cluster and the last mean on the left side (the orange asterisks).



(a) Dataset I – Initialisation 1

(b) Dataset I – Initialisation 2

(c) Dataset II – Initialisation 1

(d) Dataset II – Initialisation 2

(e) Dataset III – Initialisation 1

(f) Dataset III – Initialisation 2

Fig. 6. Relative error (expressed as a percentage) for the weights α, for 100 runs and $n = 1000$, according to the dataset and the type of initialization.

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 [21]. However, the computational cost can be prohibitive. A faster procedure is to use the SAEM algorithm. Nevertheless, both algorithms are very sensitive to its initial position: solutions can highly depend on their starting point and consequently produce sub-optimal *maximum* likelihood estimates [22]. The tempering-SAEM appears as a way to escape from local *maxima* and reach global *maxima* more often.

3.1.1 *Insensitivity of the tempering-SAEM to Initialization*

To estimate the sensitivity of the tempering-SAEM algorithm to its initial position, we generate a synthetic dataset (Figure 3) and perform the estimation 500 times for the three algorithms, with the same sequences of points chosen at random within the dataset.

The relative errors for α and μ and the Kullback-Leibler divergence between the true covariance matrices Σ and the

TABLE 1
Mean (standard deviation) relative errors (expressed as a percentage) for the estimated parameters of the GMM within the *dataset I*, according to the initial positions of the centroïdes and the type of algorithm. Over 100 runs.

	EM - 1	**EM** - 2	**SAEM** - 1	**SAEM** - 2	**tmp-SAEM** - 1	**tmp-SAEM** - 2
$\widehat{\alpha}_1$	0.00	103.10	24.83 (46.24)	99.46 (18.99)	−4.46 (0.00)	2.01 (8.54)
$\widehat{\alpha}_2$	0.00	−48.02	−19.41 (25.89)	−77.72 (28.72)	−4.23 (0.00)	0.39 (8.35)
$\widehat{\alpha}_3$	0.00	−55.08	−5.42 (26.22)	−21.87 (24.94)	8.69 (0.00)	−2.40 (0.18)
$\widehat{\mu}_2$	78.46	39.44	14.28 (18.28)	38.07 (7.15)	1.24 (0.00)	1.62 (4.18)
$\widehat{\mu}_2$	78.46	185.93	58.54 (84.14)	168.49 (37.28)	0.17 (0.00)	2.56 (17.03)
$\widehat{\mu}_3$	126.23	0.73	2.86 (4.06)	2.94 (4.40)	0.34 (0.00)	1.03 (0.01)
$\widehat{\Sigma}_1$	1503.22	306.00	104.94 (216.86)	295.19 (31.90)	0.99 (0.00)	7.08 (33.01)
$\widehat{\Sigma}_2$	1503.22	7.26	19.04 (98.85)	18.85 (5.96)	4.78 (0.00)	2.16 (2.41)
$\widehat{\Sigma}_3$	1503.22	8.90	5.19 (0.20)	6.07 (0.21)	2.35 (0.00)	1.52 (0.27)

TABLE 2
Mean (standard deviation) relative errors (expressed as a percentage) for the estimated parameters of the GMM within the *dataset II*, according to the initial positions of the centroïdes and the type of algorithm. Over 100 runs.

	EM - 1	**EM** - 2	**SAEM** - 1	**SAEM** - 2	**tmp-SAEM** - 1	**tmp-SAEM** - 2
$\widehat{\alpha}_1$	0.00	97.4	55.16 (48.27)	97.4 (0.00)	0.34 (19.25)	3.81 (18.33)
$\widehat{\alpha}_2$	0.00	−18.36	−44.85 (38.38)	−79.44 (7.78)	−2.01 (18.19)	−3.67 (12.17)
$\widehat{\alpha}_3$	0.00	−79.04	−10.31 (15.68)	−17.96 (7.78)	1.67 (3.15)	−0.14 (10.95)
$\widehat{\mu}_2$	70.97	29.12	17.05 (13.80)	29.12 (0.00)	3.34 (7.31)	3.21 (6.94)
$\widehat{\mu}_2$	70.97	192.96	104.28 (92.43)	187.28 (1.11)	5.31 (25.61)	9.47 (39.60)
$\widehat{\mu}_3$	132.29	13.28	2.15 (1.82)	2.94 (1.09)	0.79 (0.48)	1.45 (5.57)
$\widehat{\Sigma}_1$	1438.21	154.34	88.00 (57.50)	154.34 (0.00)	7.81 (27.00)	10.60 (34.47)
$\widehat{\Sigma}_2$	1438.21	10.58	44.17 (608.01)	13.68 (0.01)	7.28 (27.00)	3.48 (13.66)
$\widehat{\Sigma}_3$	1438.21	13.60	7.64 (0.13)	10.12 (0.00)	4.14 (0.90)	4.63 (2.97)

TABLE 3
Mean (standard deviation) relative errors (expressed as a percentage) for the estimated parameters of the GMM within the *dataset III*, according to the initial positions of the centroïdes and the type of algorithm. Over 100 runs.

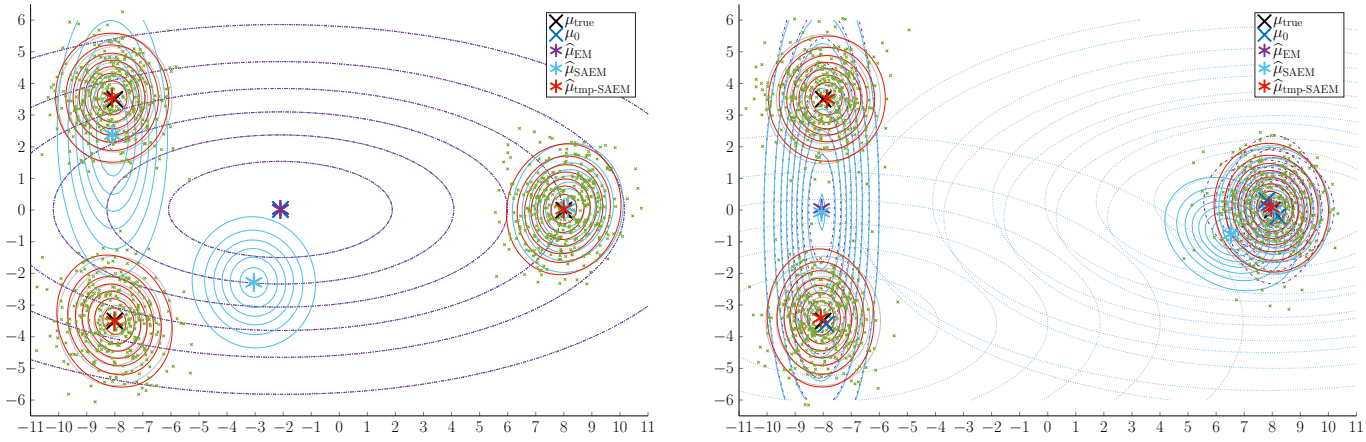
	EM - 1	**EM** - 2	**SAEM** - 1	**SAEM** - 2	**tmp-SAEM** - 1	**tmp-SAEM** - 2
$\widehat{\alpha}_1$	0.00	96.5	82.77 (34.01)	94.91 (12.40)	2.99 (22.04)	68.43 (20.77)
$\widehat{\alpha}_2$	0.00	−24.85	−41.62 (23.97)	−44.71 (18.13)	−4.64 (19.36)	−33.88 (9.76)
$\widehat{\alpha}_3$	0.00	−71.65	−41.15 (25.54)	−50.20 (18.62)	1.65 (6.42)	−34.55 (11.24)
$\widehat{\mu}_2$	70.56	20.82	18.11 (6.74)	20.51 (2.46)	3.58 (7.16)	19.26 (5.28)
$\widehat{\mu}_2$	70.56	196.07	158.34 (69.90)	187.25 (24.20)	9.84 (38.15)	174.04 (51.80)
$\widehat{\mu}_3$	131.10	5.64	6.46 (4.50)	7.43 (3.48)	0.95 (1.04)	7.10 (1.93)
$\widehat{\Sigma}_1$	1451.58	87.14	75.49 (8.35)	85.79 (1.10)	10.40 (21.34)	80.38 (22.82)
$\widehat{\Sigma}_2$	1451.58	5.51	29.18 (194.51)	12.84 (1.76)	6.42 (11.62)	11.61 (2.48)
$\widehat{\Sigma}_3$	1451.58	6.99	7.07 (0.21)	7.92 (0.21)	3.06 (0.67)	7.49 (1.49)

estimated one are compiled in Figures 4b, 4d and 4f. The class refer to the ones of Figure 3. 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, the tempering-SAEM is always competitive with the EM and the SAEM and most of the time greater. In other words, 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. Figures 4a, 4c and 4e displays the result of a type run for each of the three algorithms, with the same initial points (the blue crosses). 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.

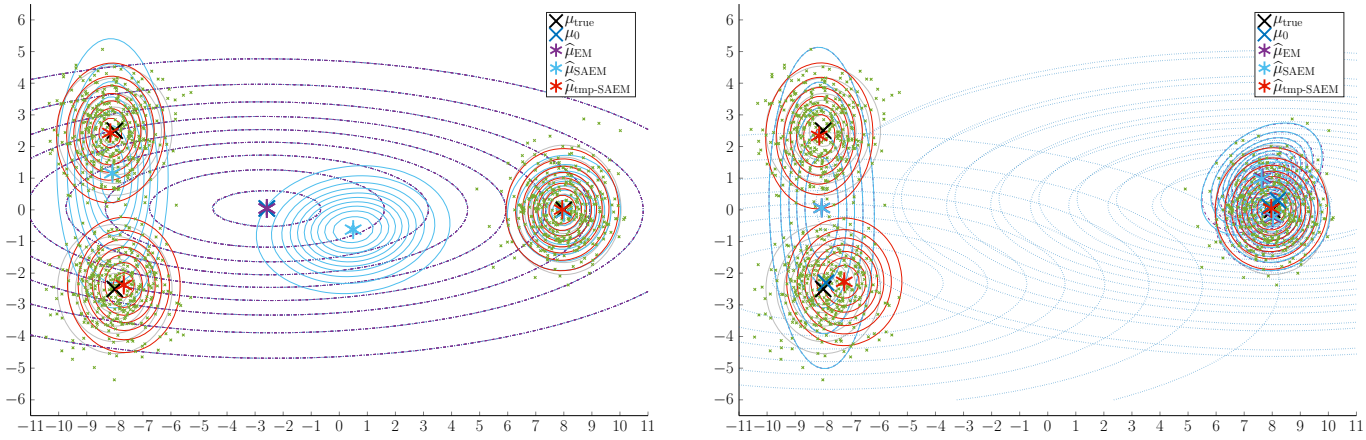
3.1.2 *Escaping Local Minima*

We then consider a situation known to be badly managed by the EM algorithm. Namely, we consider a three clusters' dataset. One cluster is on the far right side, two are on the far left side and all the three clusters are equiprobable. Moreover, we want to study the influence of the distance between the two left clusters. So we build three datasets: one where the two left clusters are properly distinct, one where they are close and a last one where they are almost merged. The three datasets are displayed at Figure 5. For each dataset, we perform the optimization for two different initial positions, referred as initialization 1 and 2 in the following. In the first case, the three centroïdes μ are initialized at the barycenter



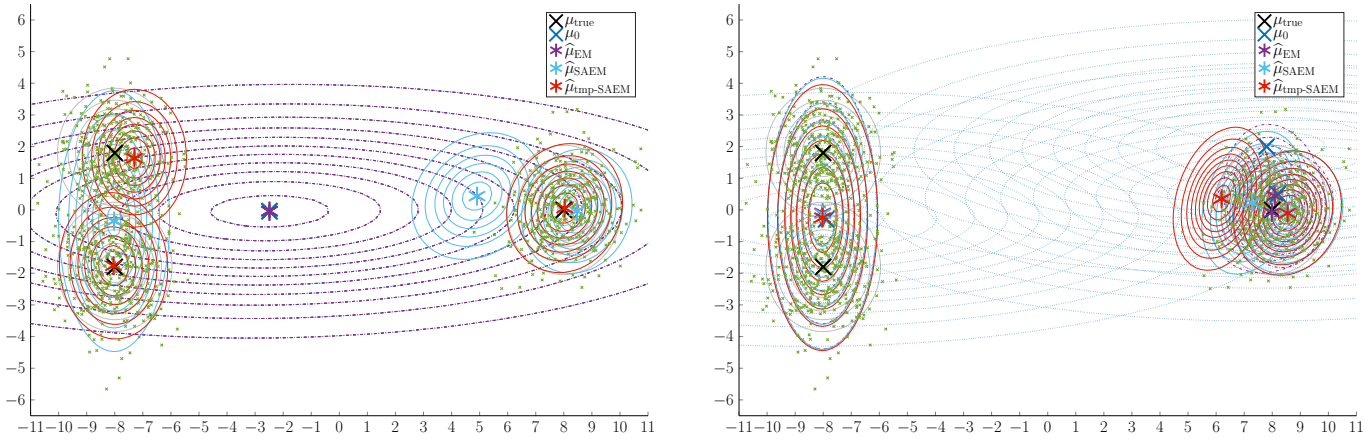
(a) Dataset I – Initialisation 1 (b) Dataset I – Initialisation 2

Fig. 7. *Dataset I*. Mean centroïdes and covariance matrices estimated by the EM, the SAEM and the tempering-SAEM algorithms within the dataset I, according to the initial position of the means. In purple dashed lines, the covariance matrix estimated by the EM; in blue plain lines, the one estimated by the SAEM and in bold red lines the one estimated by the tempering-SAEM. In dotted blue lines the initial covariance matrices associated to the two different initial means (the blue crosses).



(a) Dataset II – Initialisation 1 (b) Dataset II – Initialisation 2

Fig. 8. *Dataset II*. Mean centroïdes and covariance matrices estimated by the EM, the SAEM and the tempering-SAEM algorithms within the dataset II, according to the initial position of the means. Same conventions as for the previous figure.



(a) Dataset III – Initialisation 1 (b) Dataset III – Initialisation 2

Fig. 9. *Dataset III*. Mean centroïdes and covariance matrices estimated by the EM, the SAEM and the tempering-SAEM algorithms within the dataset III, according to the initial position of the means. Same conventions as for the previous figure.

of the observed data. In the second one, we initialize two means on the right side and one on the left side.

For each situation, we perform the estimation through all the three algorithms. We present at Tables 1, 2 and 3 the relative errors for the different parameters. As previously, we consider the algebraic relative error for the weights α. To better understand the behavior of the different algorithms, we also provide a box plot of these relative errors at Figure 6. The SAEM algorithm tends to empty classes for the benefit of other(s). It seems to be less the case for the tempering-SAEM. Note that, whatever the dataset, if the mean parameters are initialized to the mean of the dataset, the EM algorithm does not move. Thus, the error concerning the mixture proportion α seems to be very small, but this is only due to the initialization of the parameters α in favor to an equiprobable mixture.

Figures 7, 8 and 9 display the mean of the estimated $\widehat{\mu}$ and $\widehat{\Sigma}$ by the three algorithms (EM vs SAEM vs tempering-SAEM), for each dataset and each initial position.

The tempering-SAEM succeed to accurately estimate all the parameters related to the first and second datasets (Figure 7, Figure 8, Table 1 and Table 2), even when two of the mean parameters μ are initialized in the right cluster (Initialization 2). When the parameters μ are initialized to the barycenter of the dataset, the tempering-SAEM still accurately estimates the different parameters, including for the dataset III where the left clusters are merged (Figure 9a). However, when two of the mean parameters are initialized within the single right cluster, the tempering-SAEM does not succeed to capture the two left classes if the left clusters are to close (Figure 9b), but this can be easily explained by the distribution of the observations (Figure 5c). Still concerning initialization 2 and dataset 3, the tempering-SAEM is nevertheless at least competitive with the SAEM algorithm. Most interesting behavior: even though the tempering-SAEM does not always explain the whole distribution, the relative error may fall to zero with the tempering-SAEM, whereas it is never the case for the SAEM algorithm. In other words, the tempering-SAEM favor the convergence toward global maxima, and may succeed to almost surely reach them as in the first dataset, that is exactly the expected behavior of this algorithm.

Last, we present at Figures 13 and 14, in Appendix B, the evolution of the means and their associated covariance matrices. The lines 6 and 7 of Figure 13 illustrate the capacity of the tempering-SAEM to distinguish two close classes. On the contrary, the SAEM algorithm does not seem to be able to do so and remains trapped in a local minimum.

This experiment also highlights the benefits of an oscillating temperature scheme over a simple warming up phase at the beginning of the optimization. Indeed, initializing the centroïds to the mean of the dataset may be interpreted as the limit case of heating the conditional distribution for the first iterations. However, our experiments show that the SAEM initialized at the mean (Initialization 1) behave less well than the tempering-SAEM, whatever the initialization.

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 [19] and [20]. We propose here to modify the learning principle to make the procedure less susceptible to trapping states.

As in [16] and [19], we assume that:

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

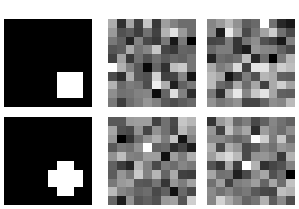
This model is called independent factor analysis (IFA). The problem is to find the value of the parameter $W = (H, \lambda, \theta)$ given y. Identifiability in this model is discussed in [23]. Basically, the sources are defined only to within an order permutation and scaling. To avoid trivialities, we fix the variances $(\sigma_j^2)_{j\in[\![1,k]\!]}$ to one [20]. Note that this definition of the IFA model is somewhat less general that the one introduced by Attias [16] 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 [20].

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

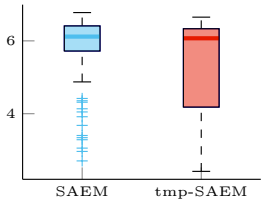
$$S_{1,j}(x,y,z) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{z_i=j\}}\,; \qquad S_4(x,y,z) = y^{\,t}y\,;$$
$$S_{2,j}(x,y,z) = \frac{1}{m}\sum_{i=1}^{m} x_i\,\mathbb{1}_{\{z_i=j\}}\,; \quad S_5(x,y,z) = y^{\,t}x\,;$$
$$S_{3,j}(x,y,z) = \frac{1}{m}\sum_{i=1}^{m} x_i^2\,\mathbb{1}_{\{z_i=j\}}\,; \quad S_6(x,y,z) = x^{\,t}x\,.$$

The M-step is then given by

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



(a) Decomposition images (left) and samples of the dataset



(b) $\mathbb{L}^2$ norm

Fig. 10. *Independent factor analysis – BG-ICA* renormalized $\mathbb{L}^2$ norm 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.

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 [16]: 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

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

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

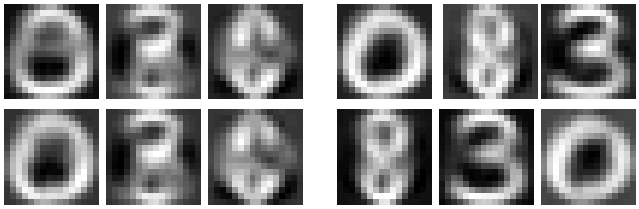
Thus, as well as for the GMM, we can compare the efficiency of SAEM vs tempering-SAEM algorithms in this context.

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, [20]), 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 10, we present the two decomposition images, 4 typical observations and the renormalized $\mathbb{L}^2$ norm between the true H (in the BG-ICA model) and the estimated one for 100 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 10 eliminates this assumption. However,

TABLE 4
Mean (standard deviation) of the p-values for the five decomposition vectors presented at Figure 12, over 50 runs.

		SAEM	tmp-SAEM
Groups 1 vs 2&3	$10^{-3}\times$	0.43 (0.31)	0.37 (0.27)
Groups 1 vs 2	$10^{-3}\times$	11.76 (7.43)	11.33 (7.18)



(a) SAEM (b) tempering-SAEM

Fig. 11. *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 digit.

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 11 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.

Last, we consider a dataset consisting of 101 hippocampi surfaces. The subjects of the dataset can be split in three groups of size 57, 32 and 12 respectively. The first group corresponds to healthy patients; the next two groups correspond to patients with Alzheimer's disease, at two stages of advancement (mild and advanced). Over each hippocampus, a scalar field represent the deformation of the considered hippocampus regarding a template one. Thus we can study the diversity of atrophy patterns, depending on the patient's state of health. We have computed $m = 5$ decomposition vectors based on the complete data set. Figure 12 presents these decomposition vectors mapped on the meshed hippocampus for both SAEM and tempering-SAEM algorithms. For comparison purpose, we enforce the same colorbar for both experiments and all hippocampi. Then, it seems that the two algorithms behave in much the same way, at least visually. This experiment attests to the reliability of the tempering-SAEM.

At Table 4, we provide the p-values obtained from the comparison of the five columns of H among the three subgroups. The test is based on a Hoteling T-statistic evaluated on the coefficients, the p-value being computed using permutation sampling. Following [20], we compute the p-value for two different comparisons: the first one compare the healthy patients with respect to Alzheimer's and dementia patients (the two last groups). The second test compare the healthy patient with respect to the mild Alzheimer's patients (the second group). Due to the stochasticity of the SAEM algorithm, we computed an average and a standard deviation of the p-values over 50 runs, with the same initial conditions. Thus, the tempering-SAEM algorithm always

behaves at least as well as the SAEM algorithm.

Finally, applying the tempering-SAEM for independent factor analysis aims to check that the advantages of the tempering-SAEM over the SAEM can improve or at least does not deteriorate 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.

This opens up new perspectives. Typically, now that we have ensured of the convergence of the approximated-SAEM, a natural opening concerns the study of the convergence of the approximated-MCMC-SAEM. Indeed, although the convergence of this algorithm has not yet been demonstrated, the tempering-MCMC-SAEM has already shown its numerical efficiency, especially in the case of medical applications [24].

APPENDIX A
THEOREM 2 AND LEMMA 2 OF [4]

In order our article to be more self-contained, we recall Theorem 2 and Lemma 2 of [4]. 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.1 (Delyon, Lavielle, Moulines [4]). *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 $V : \mathcal{S} \to \mathbb{R}$ continuously differentiable 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 (Delyon, Lavielle, Moulines [4]). *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

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

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 $\tau_{i,j} = \mathbb{P}\left[z_i = j | y_i, \theta^t\right]$.

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

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

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

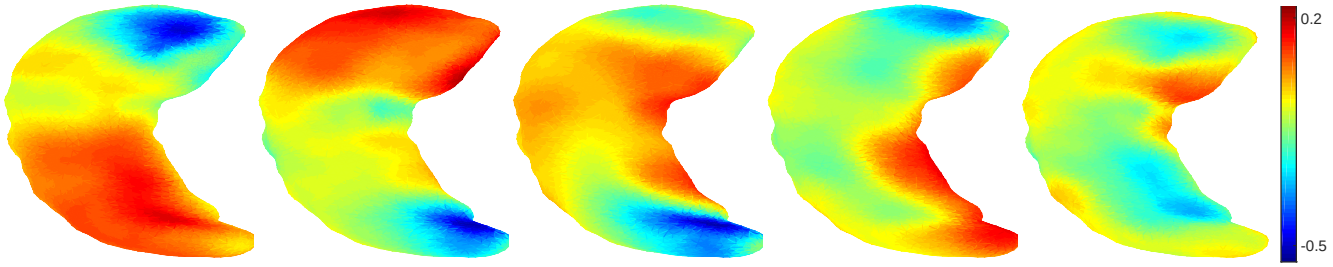
B.2 Estimation through the SAEM Algorithm

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:

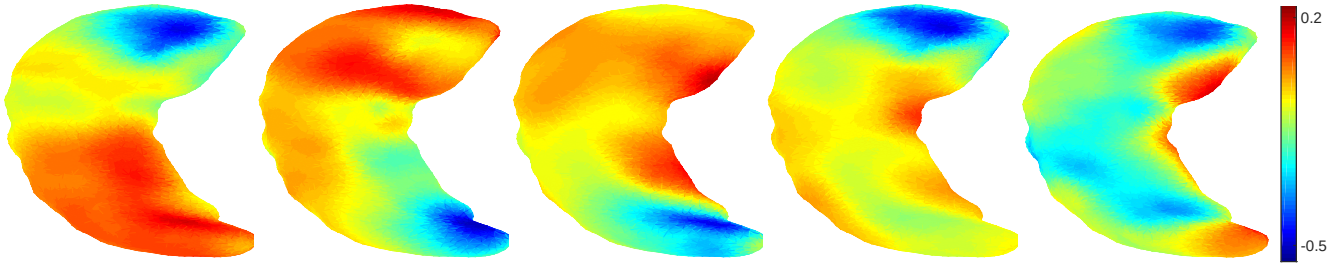
SAE-step: Sample *a* new hidden variable z^{t+1} according to the conditional distribution $q(z|y, \theta^t)$ and compute

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

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



(a) SAEM



(b) tempering-SAEM

Fig. 12. *Independent factor analysis – Hippocampi dataset.* Results of the independent factor estimation on a corpus of 101 hippocampi. Atrophy patterns of the hippocampi in the context of Alzheimers disease.

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

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

where, for all $j\in[\![1,m]\!]$, $S_{1,j}(y,z)=\sum_{i=1}^{n}\mathbb{1}_{z_i=j}$

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

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.

To stabilize the convergence of both SAEM and tempering-SAEM, we may use inverse Wishart priors for the variances and Gaussian priors for the weights.

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 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*, vol. 39, no. 1, pp. 1–38, 1977.

[2] G. Celeux and J. Diebolt, "The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem," *Computational statistics quarterly*, vol. 2, pp. 73–82, 1985.

[3] G. C. Wei and M. A. Tanner, "A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms," *Journal of the American statistical Association*, vol. 85, no. 411, pp. 699–704, 1990.

[4] 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.

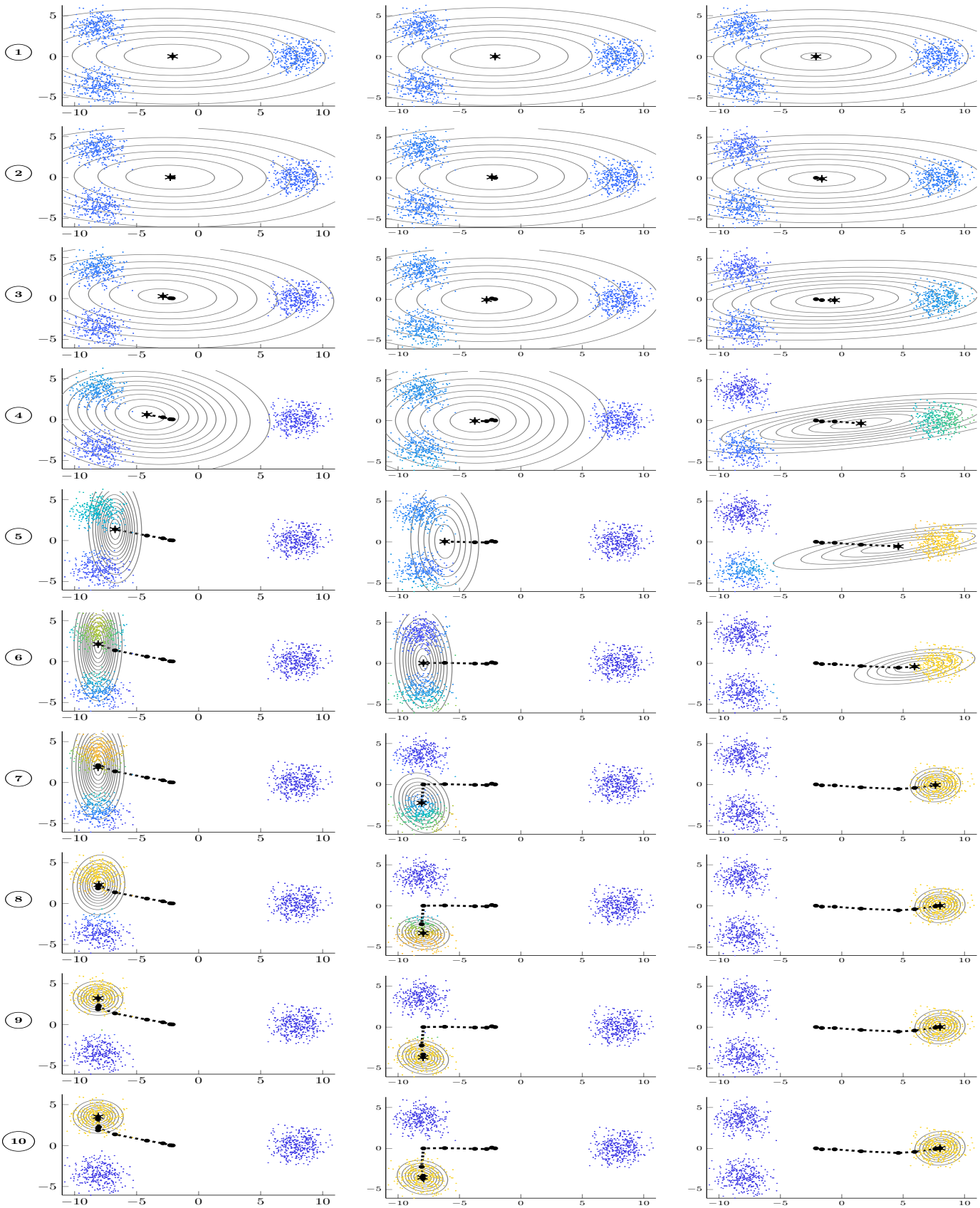
[5] O. Cappé, É. Moulines, and T. Rydén, *Inference in Hidden Markov Models*, ser. Springer Series in Statistics. Springer, 2005.

[6] 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.

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

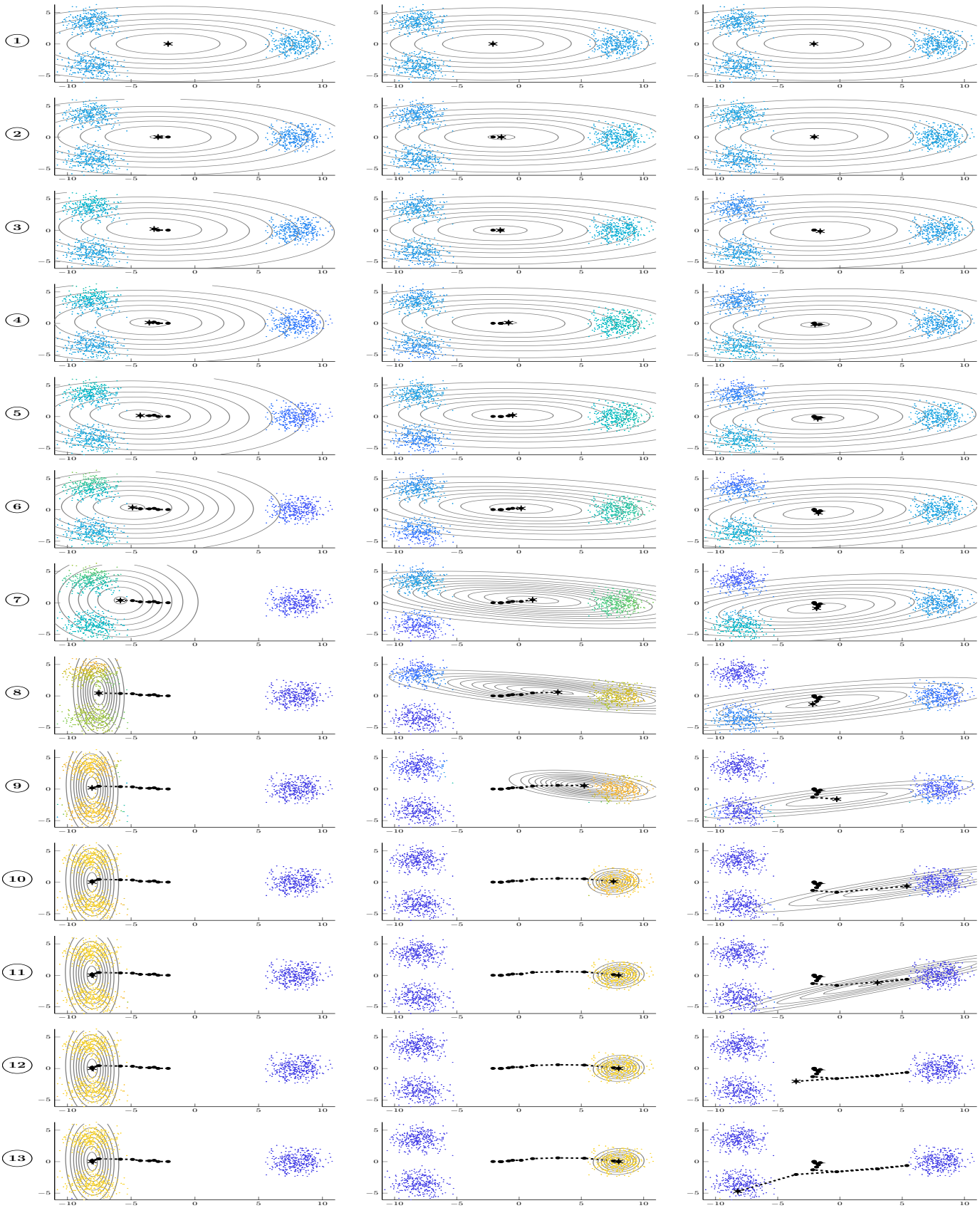
[8] M. Lavielle and F. Mentré, "Estimation of population pharmacokinetic parameters of saquinavir in hiv patients with the monolix software," *Journal of pharmacokinetics and pharmacodynamics*, vol. 34, no. 2, pp. 229–249, 2007.

[9] 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.



(a) Classe A (b) Classe B (c) Classe C

Fig. 13. *Evolution of the parameters throughout the estimation by the tempering-SAEM algorithm*, within the dataset I, with initial means at the barycenter of the dataset, for a typical run. For each class, we plot the trajectories of the mean μ and the evolution of the associated covariance. The observed data are colored according to their probability to belonging to the classes: from 0 in dark blue to 1 in yellow.



(a) Classes A & B (b) Classe C (c) Empty classe

Fig. 14. *Evolution of the parameters throughout the estimation by the SAEM algorithm*, within the dataset I, with initial means at the barycenter of the dataset, for a typical run. We keep the conventions of the previous figure.

[10] 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.
[11] 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.
[12] 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.
[13] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.
[14] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.
[15] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.
[16] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.
[17] 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.
[18] P. Hall and C. C. Heyde, *Martingale limit theory and its application*, ser. Probability and mathematical statistics. Academic Press, 1980.
[19] 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.
[20] 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.
[21] G. McLachlan and D. Peel, *Finite Mixture Models*, ser. Wiley Series in Probability and Statistics. Wile, 2000.
[22] 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.
[23] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287 – 314, 1994, higher Order Statistics.
[24] V. Debavelaere, A. Bône, S. Durrleman, and S. Allassonnière, "Clustering of longitudinal shape data sets using mixture of separate or branching trajectories," 2019, to appear in MICAI 2019.

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 École 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. 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.