A Stochastic EM algorithm for a semiparametric mixture model

Laurent Bordes, Didier Chauveau, Pierre Vandekerkhove

An EM algorithm for a semiparametric mixture model

Laurent BORDES[1] Didier CHAUVEAU[2] Pierre VANDEKERKHOVE[3]

[1] Université de Technologie de Compiègne, [2] Université d'Orléans & CNRS, [3] Université de Marne-la-Vallée & CNRS

January 2006

Abstract. Recently several authors considered finite mixture models with semi-/non-parametric component distributions. Identifiability of such model parameters is generally not obvious, and when it occurs, inference methods are rather specific to the mixture model under consideration. In this paper we propose a generalization of the EM algorithm to semiparametric mixture models. Our approach is methodological and can be applied to a wide class of semiparametric mixture models. The behavior of the EM type estimators we propose is studied numerically through several Monte Carlo experiments but also by comparison with alternative methods existing in the literature. In addition to these numerical experiments we provide applications to real data showing that our estimation methods behaves well, that it is fast and easy to be implemented.

Keywords. EM algorithm, finite mixture model, semiparametric model

AMS 2000 subject Classification. 62G05, 62G07, 62F10.

1 Introduction

Probability density functions (pdf) of m-component mixture models are defined in a general setup by

$$g(x) = \sum_{j=1}^{m} \lambda_j \, f_j(x), \quad \sum_{j=1}^{m} \lambda_j = 1, \quad x \in \mathbb{R}^p,$$

where the unknown mixture proportions $\lambda_j \geq 0$ and the unknown pdf's f_j have to be estimated. It is commonly assumed that the f_j's belong to a parametric family $\mathcal{F} = \{f(\cdot|\xi), \xi \in \mathbb{R}^d\}$ indexed by an Euclidean parameter ξ, so that the pdf g becomes

$$g_\theta(x) = g(x|\theta) = \sum_{j=1}^{m} \lambda_j \, f(x|\xi_j) \tag{1}$$

where $\theta = (\lambda_j, \xi_j)_{j=1,\dots,m} \in \Theta$ is the Euclidean model parameter. Note that the Bayesian-like notations $g(x|\theta)$ and $f(x|\xi_j)$ are usually preferred to the classical ones in this setup, even if the context is not Bayesian (θ is not a random variable), as, e.g., in the seminal paper of Dempster *et al.* [7]. We will use the Bayesian notation when necessary, e.g. when the parameter itself needs subscripts. When the number of components m is fixed the parametric mixture model of equation (1) has been well-studied; e.g. Titterington *et al.* [22], Lindsay [14] and McLachlan and Peel [16] are general references to the broad literature on this topic.

Nonparametric approaches for mixtures are motivated by the fact that the choice of a parametric family $\mathcal{F}$ may be difficult. However the model (1) can be made more flexible assuming that the number of components m is unknown; in that case m has to be estimated, see e.g. Leroux [13], Dacunha-Castelle and Gassiat [6], Lemdani and Pons [12]. But if the number of components is specified but that little is known about subpopulations (e.g., tails) another way to make the model more flexible is to avoid parametric assumption on $\mathcal{F}$. For example, one may state the model where for $j = 1, \dots, m$ we have $f_j \in \mathcal{F} = \{\text{continuous pdf on } \mathbb{R}^p\}$. Of course, such a model is very flexible since each component distribution can be itself a mixture distribution, and obviously, without additional assumptions on $\mathcal{F}$ the resulting model parameter's are not identifiable. Nevertheless, if training data are available such models become identifiable, and then, the component distributions can be estimated nonparametrically, see for example Hall [8], Titterington [21].

Note also that in the nonparametric setup without training data, specific methods to estimate mixture weights have been developed by Hettmansperger and Thomas [10] and Cruz-Medina and Hettmansperger [5].

Recently, Hall and Zhou [9] looked at p-variate data drawn from a mixture of two distributions, each having independent nonparametric components, and proved that under mild regularity assumptions their model is identifiable for $p \geq 3$. The non-identifiability for $p \leq 2$ requires to restrain the class of pdf $\mathcal{F}$. For example, for $p = 1$, restraining $\mathcal{F}$ to the location-shifted symmetric pdf, we obtain the following semiparametric mixture model:

$$g_\varphi(x) = g(x|\varphi) = \sum_{j=1}^{m} \lambda_j \, f(x - \mu_j), \quad x \in \mathbb{R}, \tag{2}$$

where the λ_j's, the μ_j's and $f \in \mathcal{G} = \{\text{even pdf on } \mathbb{R}\}$ are unknown. Hence the model parameter is

$$\varphi = (\theta, f) = ((\lambda_j, \mu_j)_{j=1,\dots,m}, f) \in \Phi = \Theta \times \mathcal{G},$$

where

$$\Theta = \left\{ (\lambda_j, \mu_j)_{j=1,\dots,m} \in \{(0,1) \times \mathbb{R}\}^m; \sum_{j=1}^{m} \lambda_j = 1 \text{ and } \mu_i \neq \mu_j \text{ for } 1 \leq i < j \leq m \right\}.$$

See Bordes *et al.* [2] and Hunter *et al.* [11] for identifiability results. In [2], for $m = 2$, the authors propose an estimator of (θ, f) for $m = 2$. Because $g = A_\theta f$, where A_θ is an invertible operator from $L^1(\mathbb{R})$ to $L^1(\mathbb{R})$, and f is an even pdf, they propose a contrast function for θ that depends only on g. Given a sample of independent g-distributed random variables they estimate g. Then, replacing g by its estimator in the contrast function, they propose a minimum contrast estimator for θ, and then, inverting A_θ and replacing θ by its estimator they obtain an estimator of the pdf f (which generally is not a pdf because the estimator of g has no reason to be in the range of the operator A_θ). This method has several limitations. For example, for $m = 3$, even if the model is identifiable (see [11]) the operator A_θ maybe not invertible and then the estimation method fails. On the other hand, the method cannot be naturally generalized to p-variate data. Furthermore the numerical computation involved by the method is time consuming which can be a drawback for large sample size. In [11] an alternative method of estimation is proposed but it seems that it suffers from similar weakness.

In parametric setup one main problem is the computation of maximum likelihood (ML) estimates; parameter estimates cannot in general be obtained in closed form from mixture structures. Conventional algorithms, such as the Newton-Raphson, have long been known to lead to difficulties; see Lindsay (p.65, [14]). The computational issue has largely been resolved, however, with the development of the EM algorithm after its formalization by Dempster *et al.* [7]. See McLachlan and Krishnan [15] for a detailed account of the EM algorithm. Moreover, in the parametric setup, the ML method can be applied easily, which is not longer true in the semiparametric setup. That is another reason why we propose an alternative method to estimate parameters of semiparametric mixture models when the number of components is fixed.

In Section 2 we give a brief description of the EM algorithm in the missing data setup. In Section 3 we show how to extend this method to the semiparametric setup by introducing one step of nonparametric estimation of the unknown mixed pdf. Although this method is applied to model (2), it is worth to mention that it is extendable to more general semiparametric models provided that they are identifiable. Because our approach is methodological only (convergence of our estimators has to be proved) Section 4 is devoted to a Monte Carlo study of the behavior of our estimators. In the same section, several known examples with real data are addressed while concluding remarks are given in Section 5.

2 Missing data setup and EM algorithm

The methodology we present in this paper involves the representation of the mixture problem as a particular case of maximum-likelihood estimation (MLE) when the observations can be viewed as incomplete data. This setup implies consideration of two sample spaces, the sample space of the (incomplete) observations, and a sample space of some “complete” observations, the characterization of which being that the ML estimation can be performed explicitly at this level, i.e. the MLE based on the complete-data is in closed form. Reference papers and monographs on this subjects are, e.g., Dempster *et al.* [7], Redner and Walker [18], McLachlan and Peel [16] and references therein. We give below a brief description of this setup in general, with some details for the parametric mixture of location-shifted pdf (2).

The (observed) data consist in n i.i.d. observations denoted by $\mathbf{x} = (x_1, \dots, x_n)$ from a pdf $g(\cdot|\theta)$. It is common to denote the pdf of the sample by $\mathbf{g}(\cdot|\theta)$ (the n-fold product of $g(\cdot|\theta)$), so that we write simply $\mathbf{x} \sim \mathbf{g}(\cdot|\theta)$. For the m-component mixture model, $g(x|\theta)$ is given by (1).

In the missing data setup, $\mathbf{g}(\cdot|\theta)$ is called the incomplete-data pdf, and the associated log-likelihood is

$$L_{\mathbf{x}}(\theta) = \sum_{i=1}^{n} \log g(x_i|\theta).$$

The (parametric) ML estimation consists in finding $\hat{\theta}_{\mathbf{x}} = \operatorname{argmax}_{\theta\in\Theta} L_{\mathbf{x}}(\theta)$. Calculating $\hat{\theta}_{\mathbf{x}}$ for the mixture model is known to be a difficult problem, and considering $\mathbf{x}$ as an incomplete data resulting from a non-observed complete-data helps.

The associated complete-data is denoted by $\mathbf{y} = (y_1, \dots, y_n)$, with associated pdf $\mathbf{h}(\mathbf{y}|\theta) = \prod_{i=1}^{n} h(y_i|\theta)$ (there exists a many-to-one mapping from $\mathbf{y}$ to $\mathbf{x}$, representing the loss of information). In the parametric mixture model (1), $y_i = (x_i, z_i)$, where $z_i \in \{1, \dots, m\}$ is the (missing) component allocation associated to the observed x_i, i.e.

$$(X_i|Z_i = j) \sim f(\,\cdot\,|\xi_j) \text{ and } \mathbb{P}(Z_i = j) = \lambda_j, \quad j = 1, \dots, m.$$

The complete-data pdf for one observation is thus

$$h(y|\theta) = h((x, z)|\theta) = \lambda_z f(x|\xi_z),$$

and the associated complete-data log-likelihood is $\log \mathbf{h}(\mathbf{y}|\theta) = \sum_{i=1}^{n} \log h(y_i|\theta)$. It is easy to check that for model (1), the complete-data MLE $\hat{\theta}_{\mathbf{y}}$ based on $\log \mathbf{h}(\mathbf{y}|\theta)$ maximization is easy to find, provided that this being the case for the parametric family $\mathcal{F}$.

2.1 The EM algorithm for the parametric mixture model

The EM algorithm iteratively maximizes, instead of the observed log-likelihood $L_{\mathbf{x}}(\theta)$, the operator

$$Q(\theta|\theta^t) = \mathbb{E}[\log \mathbf{h}(\mathbf{y}|\theta)|\mathbf{x}, \theta^t],$$

where θ^t is the current value at step t. The iteration $\theta^t \to \theta^{t+1}$ is defined in the above general setup by

1. E-step: `compute` $Q(\theta|\theta^t)$
2. M-step: `set` $\theta^{t+1} = \text{argmax}_{\theta\in\Theta} Q(\theta|\theta^t)$

The operator $Q(\,\cdot\,|\theta^t)$ is an expectation relatively to the distribution $\mathbf{k}(\mathbf{y}|\mathbf{x},\theta)$ of $\mathbf{y}$ given $\mathbf{x}$, for the value θ^t of the parameter. In the mixture model,

$$\mathbf{k}(\mathbf{y}|\mathbf{x},\theta) = \prod_{i=1}^{n} k(y_i|x_i,\theta) = \prod_{i=1}^{n} k(z_i|x_i,\theta),$$

since the $(z_i|x_i), i = 1,\ldots,n$, are independent. The $\mathbf{z}$ are discrete here, and their distribution is given through the Bayes formula by

$$k(j|x,\theta^t) = \mathbb{P}(Z = j|x,\theta^t) = \frac{\lambda_j^t f(x|\xi_j^t)}{\sum_{\ell=1}^m \lambda_\ell^t f(x|\xi_\ell^t)}. \tag{3}$$

In the case of a location-shifted mixture model with pdf (2) and known component density f, i.e. when the parametric family is $\mathcal{F} = \{f(\cdot|\mu) = f(\cdot - \mu), \mu \in \mathbb{R}^p\}$, this gives

$$k(j|x,\theta^t) = \frac{\lambda_j^t f(x-\mu_j^t)}{\sum_{\ell=1}^m \lambda_\ell^t f(x-\mu_\ell^t)}, \quad j = 1,\ldots,m. \tag{4}$$

Finally, since the component parameters are expectations (i.e. $\mathbb{E}(X|Z=j) = \mu_j$) in the location-shifted mixture model, the EM algorithm implementation for the iteration $\theta^t \to \theta^{t+1}$ is given by standard calculations (see, e.g., Redner and Walker [18]):

1. E-step: `for` $i = 1,\ldots,n$ `and` $j = 1,\ldots,m$, `compute` $k(j|x_i,\theta^t)$
2. M-step: `update` θ^{t+1} `with:`

$$\lambda_j^{t+1} = \frac{1}{n}\sum_{i=1}^n k(j|x_i,\theta^t) \tag{5}$$

$$\mu_j^{t+1} = \frac{\sum_{i=1}^n k(j|x_i,\theta^t)\, x_i}{\sum_{i=1}^n k(j|x_i,\theta^t)}, \quad j = 1,\ldots,m. \tag{6}$$

3 A semiparametric EM algorithm

We consider now on the semiparametric location-shifted mixture model (2), where the pdf f itself is an unknown, even density, considered as a parameter which has to be estimated from the data $\mathbf{x}$. One major difference with the methods in Bordes *et al.* [2] or Hunter *et al.* [11] is that our proposed methodology may be applied for any number m of mixture components, and can naturally be generalized to p-variate data $x \in \mathbb{R}^p$, $p \geq 1$ (even if this does not mean that the corresponding models are identifiable). This approach can also be generalized straightforwardly to a finite mixture of unknown symmetric densities that differ from location and scale parameters. However, we describe it for the location-shifted mixture model, since identifiability has been proved for $m = 2$ or $m = 3$ in this case.

If f is unknown the probabilities $k(j|x_i, \theta^t)$'s of the missing data conditionally to the observations, given by (4), are unknown. Hence the operator $Q(\theta|\theta^t)$ of the parametric EM itself is unknown. This is not surprising since the Euclidean parameter θ alone does not completely characterized the distribution of the data.

The parameter of the semiparametric model is $\varphi = (\theta, f) \in \Phi = \Theta \times \mathcal{F}$, where $\mathcal{F}$ is the set of continuous even pdf's over $\mathbb{R}$. In this framework, we still have that the pdf of the observed and complete data are

$$\begin{aligned} g_\varphi(x) = g(x|\varphi) &= \sum_{j=1}^{m} \lambda_j f(x - \mu_j) \\ h(y|\varphi) &= h((x,z)|\varphi) = \lambda_z f(x - \mu_z), \end{aligned}$$

and, formally, the log-likelihood associated to $\mathbf{x}$ for the parameter φ is

$$L_{\mathbf{x}}(\varphi) = \sum_{i=1}^{n} \log g(x_i|\varphi).$$

To design an EM-like algorithm which "mimic" the parametric version, we have to define, for a current value $\varphi^t = (\theta^t, f^t)$ of the parameter at iteration t, the operator

$$Q(\varphi|\varphi^t) = \mathbb{E}[\log \mathbf{h}(\mathbf{y}|\varphi)|\mathbf{x}, \varphi^t].$$

As in the parametric case, the expectation is taken with respect to the distribution of the $\mathbf{y}$ given $\mathbf{x}$, for the value φ^t of the parameter:

$$\mathbf{k}(\mathbf{y}|\mathbf{x}, \varphi^t) = \prod_{i=1}^{n} k(y_i|x_i, \varphi^t) = \prod_{i=1}^{n} k(z_i|x_i, \varphi^t),$$

where

$$k(j|x, \varphi^t) = \mathbb{P}(Z = j|x, \varphi^t) = \frac{\lambda_j^t f^t(x - \mu_j^t)}{\sum_{\ell=1}^{m} \lambda_\ell^t f^t(x - \mu_j^t)}, \quad j = 1, \ldots, m. \tag{7}$$

Hence $Q(\varphi|\varphi^t)$ is given by

$$Q(\varphi|\varphi^t) = \sum_{i=1}^{n} \sum_{j=1}^{m} k(j|x_i, \varphi^t)[\log(\lambda_j) + \log f(x_i - \mu_j)]. \tag{8}$$

For a given initialization $\varphi^0 = (\theta^0, f^0)$, a formal EM algorithm for estimating φ is thus

1. E-step: `compute` $Q(\varphi|\varphi^t)$ `using (7) and (8);`
2. M-step: `choose` φ^{t+1} `which maximizes` $Q(\varphi|\varphi^t)$.

The main difficulty to apply the above algorithm is to determine an estimate f^{t+1} of f such that $\varphi^{t+1} = (\theta^{t+1}, f^{t+1})$ maximizes $Q(\cdot|\varphi^t)$, since standard nonparametric density estimates do not insure this property. In the next section, we propose instead an heuristic approach based on the model (location-shifted mixture) and the parametric EM maximization step.

3.1 Methodology for the semiparametric EM

We focus first on the maximization for the Euclidean part of the parameter. Consider the parametric mixture model (1), and assume that the complete-data $(\mathbf{x}, \mathbf{z})$ are observed. Denote by $\{x_i : z_i = j, i = 1, \dots, n\}$ the sub-sample of the observations belonging to the jth component. Without any assumption on the common pdf f, the MLE of the proportions of the mixture are

$$\hat{\lambda}_j = \frac{\sum_{i=1}^n \mathbb{I}_{z_i=j}}{n}, \quad j = 1, \dots, m, \tag{9}$$

(where $\mathbb{I}_{z_i=j}$ equals 1 when $z_i = j$). Note that in (9), actually only $m-1$ weights have to be estimated. Consider further the particular case where the ξ_j's are expectation parameters, i.e. $\mathbb{E}(X|Z = j) = \xi_j$ (note that this does not require the parametric pdf $f(\cdot|\xi)$ to be even). Then, the unbiased and consistent estimates of the ξ_j's are the sub-sample empirical averages

$$\hat{\xi}_j = \frac{\sum_{i=1}^n x_i \mathbb{I}_{z_i=j}}{\sum_{i=1}^n \mathbb{I}_{z_i=j}}, \quad j = 1, \dots, m. \tag{10}$$

In addition, the $\hat{\xi}_j$'s given by (10) are the MLEs of the ξ_j's when, e.g., f belongs to an exponential family with associated sufficient statistic $T(x) = x$ (see, e.g., Sundberg [20] and Redner and Walker [18]).

When the $\mathbf{z}$ are missing, the MLE on the complete-data has to be replaced by the parametric EM. Its M-step given by equations (5) and (6), which comes from direct maximization of $Q(\,\cdot\,|\theta^t)$, can be viewed as equations (9) and (10), where each unknown $\mathbb{I}_{z_i=j}$ has been replaced by its expectation counterpart, conditionally to its associated observations x_i, and for the current value of the parameter; that is precisely $\mathbb{E}(\mathbb{I}_{Z_i=j}|x_i, \theta^t) = k(j|x_i, \theta^t)$ given by (3). This is a well-known property of EM, which comes from the fact that the formula involving the missing data in the expectation is linear.

The heuristic approach we suggest to implement the semiparametric EM algorithm is based on this property, since the component parameters are expectations in the location-shifted semiparametric mixture. The idea is to iteratively:

1. compute an estimate f^{t+1} of f, possibly using θ^t
2. substitute f^{t+1} in the M-step of the parametric EM (5)–(6) to compute θ^{t+1}.

We turn now to the determination of an estimate of f, given the Euclidean parameter θ or an estimate θ^t, and using the model assumption, i.e. the fact that the mixture has an effect only on the location parameter. Then it seems reasonable to estimate f using a nonparametric density estimate based on *all* the data $\mathbf{x}$ appropriately "centered back", by removing the shift of localization for each observation. In the sequel, we denote by $\tilde{x}_i$ the ith observation "centered back", and by $\tilde{\mathbf{x}} = (\tilde{x}_1, \dots, \tilde{x}_n)$ the corresponding vector.

Let us first describe the flavor of the method in what can be considered as an "ideal situation". Assume that the complete-data $\mathbf{y} = (\mathbf{x}, \mathbf{z})$ is available, and that θ is known. Then a consistent estimate of f would be given by the following steps:

1. compute $\tilde{\mathbf{x}} = (\tilde{x}_1, \dots, \tilde{x}_n)$, where $\tilde{x}_i = x_i - \mu_{z_i}$, $i = 1, \dots, n$
2. compute a kernel density estimate using some kernel K and bandwidth h_n,

$$\hat{f}_{\tilde{\mathbf{x}}}(u) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{u - \tilde{x}_i}{h_n}\right).$$

Assume now that the $\mathbf{z}$ are missing, but that the true parameter φ is known. The difficulty then is to recover a sample from f be given a sample from g_φ. We may think of several strategies, which consist intuitively in allocating each observed x_i to a component j, and given this allocation to "recenter" x_i by substracting μ_j to it. The allocation can only be deduced from the posterior probabilities $k(j|x_i, \varphi)$ given by (7). Then we may think of an "expectation strategy" following the EM principle:

$$\tilde{x}_i = x_i - \sum_{j=1}^{m} k(j|x_i, \varphi)\, \mu_j, \quad i = 1, \dots, n.$$

We may also use the maximum of the posterior probabilities, as it is usually done in classification algorithms based on EM:

$$\tilde{x}_i = x_i - \mu_{j_i^*}, \quad j_i^* = \underset{j \in \{1, \dots, m\}}{\operatorname{argmax}}\ k(j|x_i, \varphi), \quad i = 1, \dots, n.$$

Unfortunately, even with φ known, none of these strategies return a sample f-distributed, as it can be checked on simple explicit situations. To recover a sample from f, we *need* to simulate the ith allocation according to the posterior probabilities $(k(j|x_i, \varphi), j = 1, \dots, m)$, i.e. from a multinomial distribution of order 1:

```
S-1: for $i = 1, \dots, n$, simulate $Z(x_i, \varphi) \sim \mathcal{M}(1; k(j|x_i, \varphi), j = 1, \dots, m)$;
S-2: set $\tilde{x}_i = x_i - \mu_{Z(x_i, \varphi)}$,
```

where $Z(x,\varphi) \in \{1,\dots,m\}$ and $\mu_{Z(x,\varphi)} = \mu_j$ when $Z(x,\varphi) = j$. The result below states that this procedure returns a sample f-distributed, in the multidimensional situation.

Lemma 1 *If* $\mathbf{X}$ *is a sample from the pdf* g_φ *of the* m*-components location-shifted mixture model, then* $\tilde{\mathbf{X}}$ *given by the Stochastic step (S-1 and S-2) above, where* φ *is known, is a sample from* f*.*

Proof. Since $\mathbf{X} = (X_1,\dots,X_n)$ is i.i.d. from g_φ, it is enough to check the property for $X \in \mathbb{R}^p$. Let $X \sim g_\varphi$, and $\tilde{X} = X - \mu_{Z(X,\varphi)}$. For $y = (y_1,\dots,y_p) \in \mathbb{R}^p$, and $\mu_\ell = (\mu_{\ell,1},\dots,\mu_{\ell,p}) \in \mathbb{R}^p$, we denote by

$$\mathbb{P}_\varphi(\tilde{X} < y) = \mathbb{P}_\varphi(\tilde{X}_1 < y_1,\dots,\tilde{X}_p < y_p)$$

the multidimensional cdf of the random vector $\tilde{X}$. Then

$$\begin{aligned}
\mathbb{P}_\varphi(\tilde{X} < y) &= \int \mathbb{P}\left(X - \mu_{Z(X,\varphi)} < y | X = x\right) g_\varphi(x)\,dx \\
&= \int \sum_{\ell=1}^{m} \mathbb{P}\left(x - \mu_{Z(x,\varphi)} < y | Z(x,\varphi) = \ell\right) k(\ell|x,\varphi) g_\varphi(x)\,dx \\
&= \sum_{\ell=1}^{m} \lambda_\ell \int \mathbb{I}_{\{x_1-\mu_{\ell,1}<y_1\}} \times \cdots \times \mathbb{I}_{\{x_p-\mu_{\ell,p}<y_p\}} f(x-\mu_\ell)\,dx_1 \dots dx_p \\
&= \sum_{\ell=1}^{m} \lambda_\ell \mathbb{P}_\varphi(X_1 < y_1,\dots,X_p < y_p) = F(y),
\end{aligned}$$

where F is the cdf of X. □

It appears that this simulation step is analog to the "Stochastic EM" (SEM) simulation step for the missing data in parametric mixture situations (see, e.g., Celeux and Diebolt [3]). But the stochastic step was introduced there in an attempt to accelerate EM's convergence or to avoid stabilization on saddle points in parametric mixture situations. This stochastic step has also proved to be useful in situations where the integral in $Q(\,\cdot\,|\theta^t)$ is not in closed form due to specific missing data situations, preventing the maximization to be worked out explicitly, as in Chauveau [4]. It is also present in the Monte-Carlo EM (MCEM) algorithm of Wei and Tanner [23], where $Q(\,\cdot\,|\theta^t)$ is replaced by its Monte-Carlo approximation based on several simulated realizations of the missing data at each EM step. The interesting point is that in the present semiparametric situation, this stochastic step is required to recover f from g_φ.

We may thus make use of this "asymptotic result" (in the sense that, when φ^t is close to φ, the sample $\tilde{\mathbf{X}}$ should be approximately f-distributed) to design an estimate of f at iteration $t+1$, when only the current value of the parameter φ^t is available. The S-step becomes:

S-1 `for` $i = 1, \ldots, n$, `simulate` $Z^{t+1}(x_i, \varphi^t) \sim \mathcal{M}(1; k(j|x_i, \varphi^t), j = 1, \ldots, m)$;

S-2 `set` $\tilde{x}_i^{t+1} = x_i - \mu^t_{Z^{t+1}(x_i,\varphi^t)}$.

It is then possible to compute a kernel density estimate of f based on the centered data $\tilde{\mathbf{x}}^{t+1}$, and, using the symmetric assumption in the model, to symmetrize this kernel density estimate to obtain an estimate f^{t+1} of f.

Finally, the step $\varphi^t \to \varphi^{t+1}$ of the semiparametric EM algorithm (SP-EM) is defined by:

1. **E-step:** `compute` $k(j|x_i, \varphi^t)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$ using (7).

2. **S-step:**

   - `for` $i = 1, \ldots, n$, `draw` $Z^{t+1}(x_i, \varphi^t) \sim \mathcal{M}(1; k(j|x_i, \varphi^t), j = 1, \ldots, m)$;
   - `set` $\tilde{x}_i^{t+1} = x_i - \mu^t_{Z^{t+1}(x_i,\varphi^t)}$.

3. **Nonparametric step:** (update of the functional parameter)

   - `kernel density estimate`

   $$\hat{f}_{\tilde{\mathbf{x}}^{t+1}}(u) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{u - \tilde{x}_i^{t+1}}{h_n}\right); \tag{11}$$

   - `symmetrization`

   $$f^{t+1}(u) = \frac{\hat{f}_{\tilde{\mathbf{x}}^{t+1}}(u) + \hat{f}_{\tilde{\mathbf{x}}^{t+1}}(-u)}{2}. \tag{12}$$

4. **M-step:** (parametric EM strategy to update the Euclidean parameter)

$$\begin{aligned}
\lambda_j^{t+1} &= \frac{1}{n} \sum_{i=1}^{n} k(j|x_i, \varphi^t); \\
\mu_j^{t+1} &= \frac{\sum_{i=1}^{n} k(j|x_i, \varphi^t)\, x_i}{\sum_{i=1}^{n} k(j|x_i, \varphi^t)}, \quad j = 1, \ldots, m.
\end{aligned}$$

An alternative algorithm may be used, in the flavor of the SEM algorithm, i.e. taking advantage of the simulated complete-data to compute the MLE of the Euclidean parameter :

Replace M-step 4 above with:

$$\begin{aligned}
\lambda_j^{t+1} &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{Z_i^{t+1}(x_i,\varphi^t)=j\}}; \\
\mu_j^{t+1} &= \frac{\sum_{i=1}^{n} x_i\, \mathbb{I}_{\{Z_i^{t+1}(x_i,\varphi^t)=j\}}}{\sum_{i=1}^{n} \mathbb{I}_{\{Z_i^{t+1}(x_i,\varphi^t)=j\}}}, \quad j = 1, \ldots, m.
\end{aligned}$$

Note that this algorithm is well-defined and easy to implement also for multidimensional models ($x_i \in \mathbb{R}^p$, $p > 1$). However, the symmetry condition on f, necessary for identifiability in the univariate case, has then to be replaced by other conditions, as in Hall and Zhou [9]. The corresponding symmetrization (12) in step 3 of the SP-EM algorithm has also to be changed accordingly.

The remaining implementation issue is the definition of the initialization for the nonparametric part of the parameter. Indeed as usual, EM needs an initial value θ^0 which may be chosen arbitrarily or data-driven (see Section 4.3). But in this semiparametric case, it also needs a starting value f^0, to compute the E-step of the very first iteration using (7). Since f is assumed symmetric with zero location, we have to re-center the data once without the availability of the posterior probabilities (which need an estimate of f to be computed). We suggest to use a nearest-neighbour (or K-means) approach based on the initial location parameters. For example in the case of $m = 2$ components,

$$\tilde{x}_i^0 = x_i - \mu_1^0 \mathbb{I}_{\{|x_i - \mu_1^0| \leq |x_i - \mu_2^0|\}} - \mu_2^0 \mathbb{I}_{\{|x_i - \mu_1^0| > |x_i - \mu_2^0|\}},$$

and to compute f^0 using (11) and (12) as in the general case.

From the theoretical point of view, the sequence of Euclidean parameter $(\theta^t)_{t\geq 0}$ is a marginal of a Markov chain, the definition of which depends on the strategy selected for the M-Step. The asymptotic behavior of this Markov chain is an ongoing work.

4 Simulation and examples

We applied this SP-EM algorithm on synthetic simulated examples and on real data cases corresponding to model (2). To compare with competing methods, we choose to simulate the synthetic models proposed in Bordes *et al.* [2]. We then apply the semiparametric EM to the well-known Old Faithful geyser data used in Hunter *et al.* [11], and to the rainfall data already used in [2]. All these applications are location-shifted semiparametric mixtures with two components and univariate observations.

Computations have been performed with `Matlab`, using the EM strategy for the M-step of the semiparametric EM. The needed kernel density estimates (11) have been computed using the appropriate function (`ksdensity`) from the `Matlab` statistics toolbox, with different but non adaptive settings for the bandwidth h_n (see below).

Standard EM algorithms use for the stopping criterion a distance between two consecutive iterations, e.g. running EM until $||\theta^{t+1} - \theta^t|| \leq \varepsilon$. But in our case, the sequence of Euclidean parameters $(\theta^t)_{t\geq 0}$ is a marginal of a Markov chain, so that pointwise convergence cannot be obtained theoretically (see Celeux and Diebolt [3]). Instead, we used for the stopping criterion the minimum between the EM criterion above, and a fixed, predetermined number of iterations large enough so that the sequence of Euclidean parameter stabilizes. As it can be seen on the figures, stabilization occurs rather quickly in all the examples we have tried.

All the figures for the detailed runs include six panels. The three top panels show the sequence of each coordinate $(\lambda^t, \mu_1^t, \mu_2^t)$ of the Euclidean parameter in solid line, together with the sequence of empirical means (e.g., $s \to \sum_{t=1}^{s} \lambda^t / s$ in top left panels) in dashed line. For the simulated cases, the true value is also depicted as a constant line. The bottom left panel is an histogram of the actual or simulated data, together with the true pdf in Monte Carlo studies. The bottom middle panel is a plot of the pdf f estimated via the SP-EM algorithm (in dashed line), against the true pdf in simulated situations, or against the parametric f estimated by ML method in the Gaussian mixture model for real data cases (solid line).

4.1 Monte Carlo study from a Gaussian mixture

We have simulated first the two components Gaussian mixture already used for illustration purpose in [2]. The model is

$$g(x|\varphi) = \lambda\phi_{\mu_1}(x) + (1-\lambda)\phi_{\mu_2}(x),$$

where ϕ_μ is the pdf of the Gaussian $\mathcal{N}(\mu, 1)$, i.e. f is the pdf of $\mathcal{N}(0, 1)$.

The results presented consist first in two detailed sample runs for $n = 100$ and $n = 300$ observations, and true Euclidean parameter $\theta = (\lambda = 0.15, \mu_1 = -1, \mu_2 = 2)$. This is the most difficult situation of the three choices in [2], in the sense that λ is small, so that the pdf $g(\cdot|\varphi)$ is weakly bumped (see Figure 1, bottom right). It is well known that mixtures are more difficult to estimate when the component densities overlap, or when the weight of one component is small. For this model, the bandwidth has been set here close to the minimizer of the mean integrated square error $h_n = (4/3n)^{1/5}$ (see [2]). The initial Euclidean parameter has been set arbitrarily to $\theta^0 = (0.5, -1.5, 2.5)$. Note that only about 7 seconds were necessary to run using `Matlab` 50 SP-EM iterations for the $n = 300$ case on an average computer.

Figure 1 top panels show that SP-EM stabilizes after about 30 iterations for the three Euclidean parameters. Bottom panels show that the moderate bump of g_φ has been recovered by the algorithm, even for this rather small sample size, in comparison with the difficulty of the problem and the fact that both θ and f are unknown. Figure 2 is provided to show on this particular run the improvement obtained by increasing the sample size to $n = 300$. The effect is visible on the noise of the sequence of the Euclidean parameters (marginals of a markov chain), and on the recovering of f and g_φ.

We then conduct a Monte Carlo study to compute mean and standard errors on Monte-carlo replications, on the same settings as those used in [2]. The initial

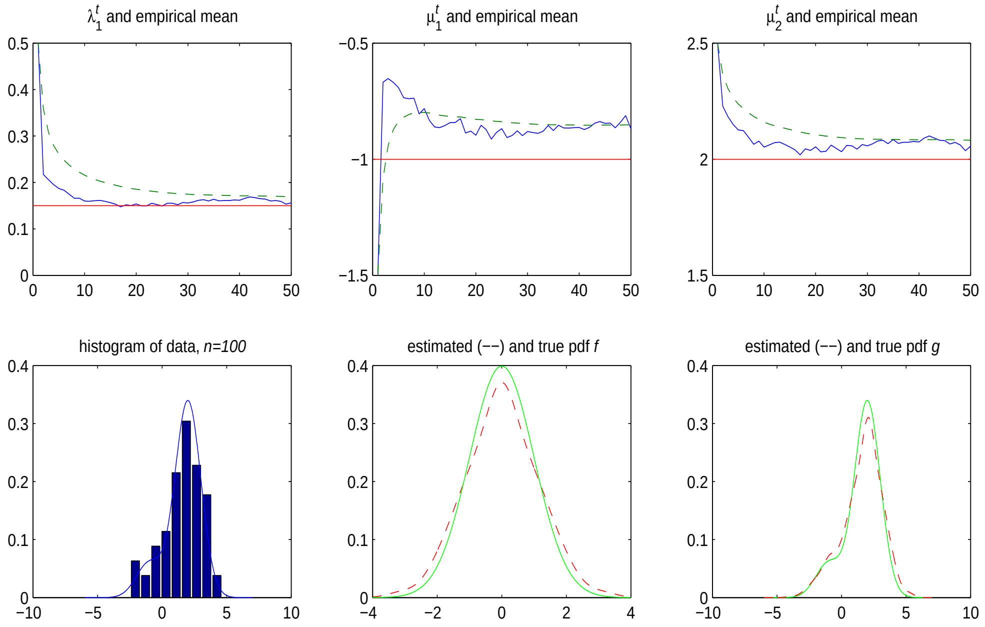

Figure 1: Semiparametric EM for the Gaussian mixture, $n = 100$, $\theta^0 = (0.5, -1.5, 2.5)$.

Euclidean parameter θ^0 has been set here to the true value, to avoid a possible bias introduced by different starting values among replications, convergence to saddle points or to estimates corresponding to one empty component (λ close to 0 or 1), or label switching difficulties, as this is often the case in EM studies (see, e.g., Redner and Walker [18]). The results are displayed in Table 1. This allows us to compare our results with [2] and with the MLE of the parametric Gaussian mixture model given therein. The standard errors are comparable to those obtained by the method of [2], while the estimates given by the SP-EM are slightly more biased, particularly the weight of component one (λ) which tends to be under-estimated, even for the highly bumped model. It is interesting to point out that our estimates are also in the range of the parametric MLE given in [2].

We finally did a Monte Carlo study to estimate the decreasing behavior of the standard error of the parameters when n increases, in order to estimate the rate of convergence. Standard errors computed over replications are then fitted using a standard Least-Square method to different rates, from $\log(n)^{-1}$ to $n^{-\gamma}$ for selected values of $\gamma \in (0, 1)$. Results for the best fitted curves, which correspond clearly to the rate $\mathcal{O}(n^{-1/2})$, are displayed in Figure 3.

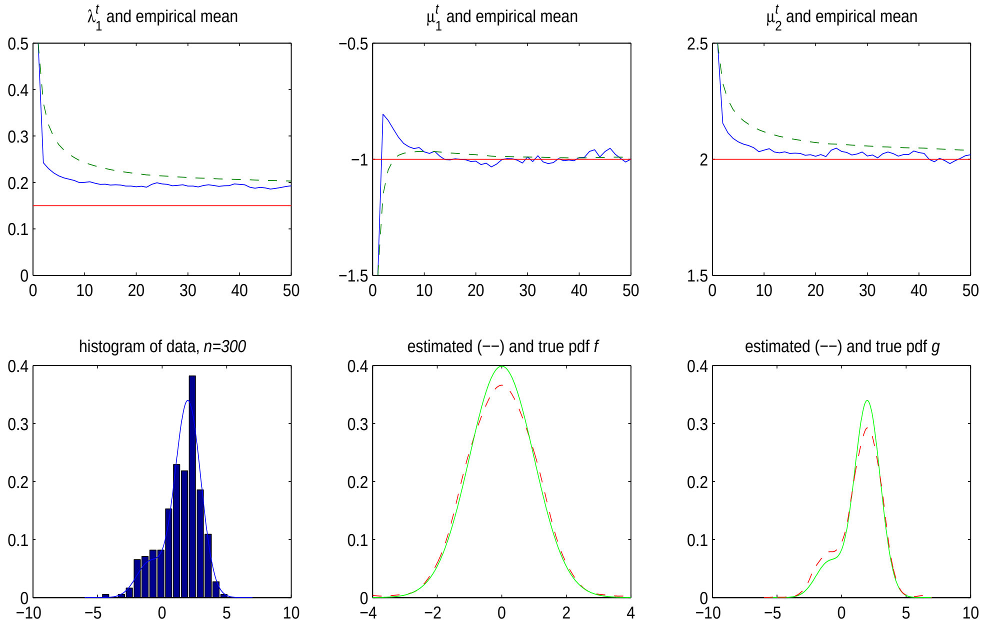

Figure 2: Semiparametric EM for the Gaussian mixture, $n = 300$, $\theta^0 = (0.5, -1.5, 2.5)$.

4.2 Monte Carlo study from a mixture of a trimodal density

We also simulate the 2 components location-shifted mixture of a trimodal pdf, also proposed as an example in Bordes *et al.* [2]. The model is

$$g(x|\varphi) = \lambda f(x - \mu_1) + (1 - \lambda) f(x - \mu_2)$$

with true parameter $\theta = (\lambda = 0.25, \mu_1 = 0, \mu_2 = 4)$, and f being itself a 3 components Gaussian mixture depicted in the bottom-middle panel of Figure 4. We just provide a sample run for a small sample size of $n = 100$ here, for brevity. Here again, the SP-EM algorithm stabilizes after a small number of iterations (Figure 4, top panels), and the estimates are computed over 50 iterations. The reconstruction of the mixing pdf f (Figure 4, bottom middle panel) is quite good, as well as the reconstruction of the data pdf g_φ.

4.3 Actual data examples

We choose to apply the SP-EM algorithm on two motivating and known datasets. In each of these actual situations, the initial Euclidean parameter θ^0 has been determined from the data, by choosing empirically a threshold c between the two bumps looking at the histogram of the data (bottom left panel of Figures 5 and 6), and

Table 1: Empirical means and standard error of (λ, μ_1, μ_2), based on 200 Monte-Carlo replications of 50 SP-EM iterations; Initial $\theta =$ true value $= (\lambda, -1, 2)$.

n	λ	$\bar{\lambda}$	$\bar{\mu}_1$	$\bar{\mu}_2$	$\sigma(\bar{\lambda})$	$\sigma(\bar{\mu}_1)$	$\sigma(\bar{\mu}_2)$
100	0.15	0.123	-1.069	1.924	0.049	0.540	0.145
200	0.15	0.133	-1.027	1.958	0.035	0.289	0.095
100	0.25	0.226	-0.980	1.905	0.060	0.414	0.172
200	0.25	0.237	-1.009	1.946	0.041	0.194	0.104
100	0.35	0.343	-0.893	1.906	0.062	0.337	0.218
200	0.35	0.344	-0.955	1.960	0.039	0.182	0.111

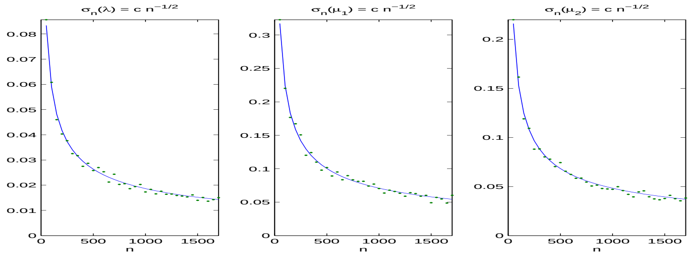

Figure 3: plots and $\mathcal{O}(n^{-1/2})$ Least-Square fitting of standard errors for the Gaussian mixture model, $\lambda = 0.35$.

considering that observations $x_i \leq c$ belong to the first component, while observations $x_i > c$ belong to the second component. Then relative component weight and component mean are computed.

4.3.1 Old Faithful geyser data

The Old Faithful geyser dataset gives measurements in minutes of the eruptions lengths, and of the time between eruptions. This dataset is included in the standard R distribution, and has already been used by Hunter *et al.* [11] as a benchmark for the location shifted mixture model.

The initial Euclidean parameter θ^0 has been determined from the data, by choosing thresholds $c \in [60, 70]$. Several trials show that the result is not sensitive to the threshold. For the run detailed in Figure 5, the choice $c = 65$ gives $(\lambda^0 = 0.35, \mu_1^0 = 54.05, \mu_2^0 = 79.79)$. The semiparametric EM stabilizes rather quickly, since the model is highly bumped, and n is large. The parameter estimates are computed after 60 iterations. Table 2 shows that our estimates are comparable to those obtained by [11], and also close to the parametric MLE of the Gaussian mixture model with equal variances assumption, which is reasonable for these data.

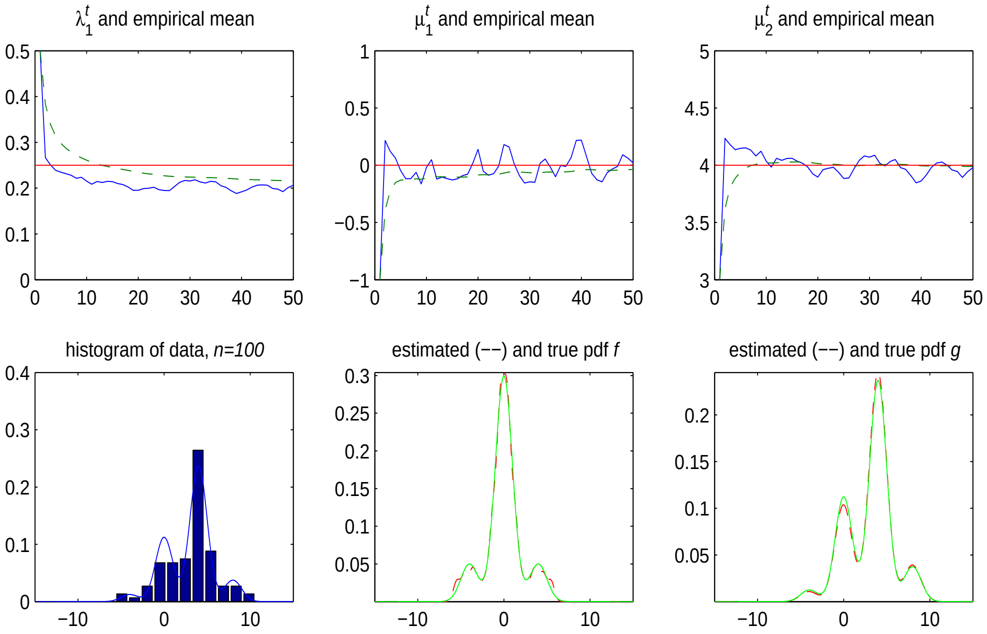

Figure 4: Semiparametric EM for trimodal pdf f, $n = 100$, initial value $\theta^0 = (0.5, -1, 3)$; estimates are $\hat{\lambda} = 0.216, \hat{\mu}_1 = -0.037, \hat{\mu}_2 = 3.989$.

4.3.2 Precipitation data

To compare again with the method in Bordes *et al.* [2], we apply the SP-EM algorithm to the rainfall data, which give the amount of precipitation in inches for cities in the US (see McNeil [17]). These data are interesting here since the Gaussian mixture model seems not reasonable in the tails. The threshold for computing the initial θ^0 has been set to $c = 26$ from the data, and gave $(\lambda^0 = 0.243, \mu_1^0 = 15.182, \mu_2^0 = 41.206)$. The bandwidth has been set to 2.5 by trial-and-error, since too large values

Table 2: Parameter estimates for the Old Faithful geyser waiting data, using the normal homoscedastic mixture approach (NMLE), the semiparametric estimation from [11] (SP), and the semiparametric EM algorithm (SP-EM).

	λ	μ_1	μ_2	σ^2
NMLE	0.361	54.61	80.09	34.45
SP	0.352	54.0	80.0	
SP-EM	0.359	54.592	80.046	

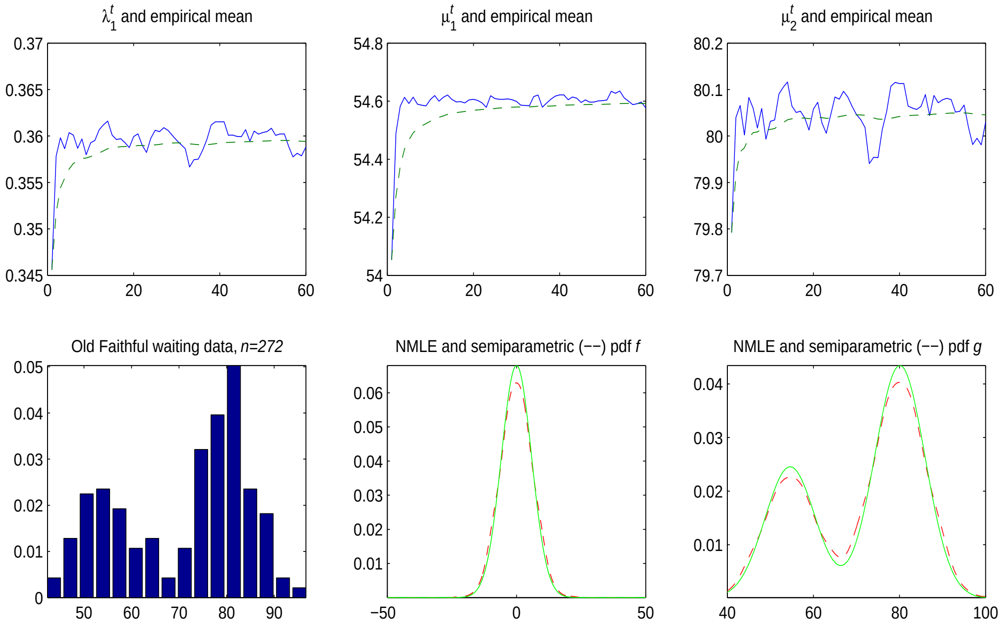

Figure 5: Semiparametric EM and NMLE for the Old Faithful Geyser waiting time data; SP-EM estimates are $(\hat{\lambda} = 0.359, \hat{\mu}_1 = 54.592, \hat{\mu}_2 = 80.046)$.

result in the algorithm emptying one component. The reason for this is that $n = 70$ is small here, and few observations come from the leftmost component (e.g., 17 observations for $c = 26$). The solution founded is rather unstable due to this small sample size. Figure 6 compares the results obtained with the parametric MLE of the homoscedastic Gaussian model (NMLE), as given in [2], and the estimates given by the SP-EM algorithm after 60 iterations. It is interesting to note that, as with the estimates from [2], the difference with the parametric Gaussian model consist essentially in the two bumps in the tails of f.

5 Perspectives

The proposed algorithm is fast, computationally simple, and it seems as efficient as the competing method of Bordes *et al.* [2]. It potentially works for $m > 2$ components, multidimensional situations $p > 1$, i.e. more general mixtures, provided the model being identifiable. It may also be used in applications using mixtures with one component known (as this is the case, e.g., in microarray data, see Robin *et al.* [19], and Bordes *et al.* [1]).

Theoretically, we are currently studying the convergence of the simulated Markov

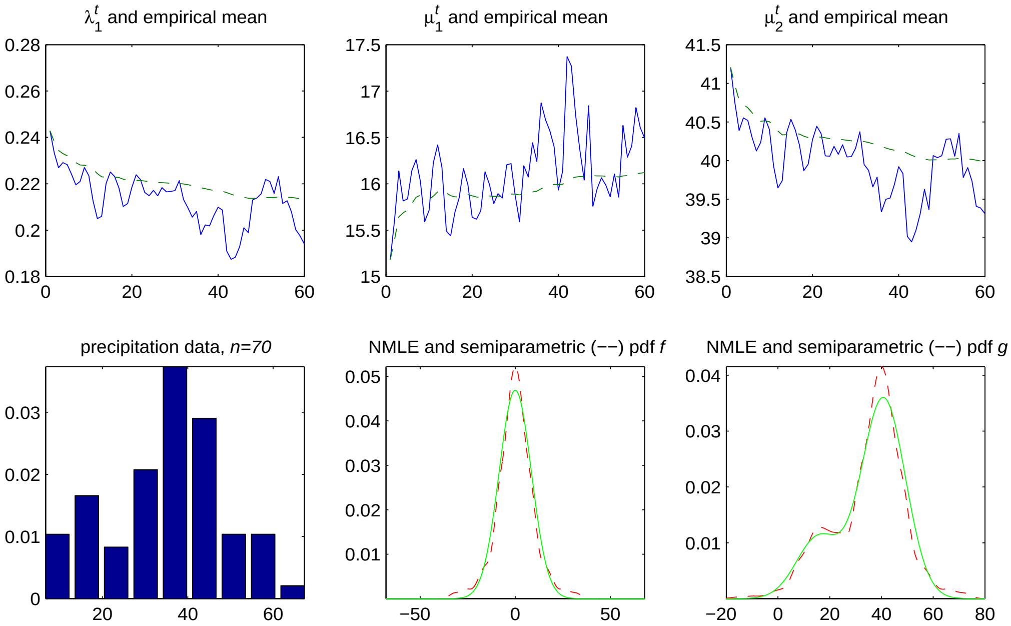

Figure 6: Semiparametric EM and NMLE for precipitation data; SP-EM estimates are $(\hat{\lambda} = 0.213, \hat{\mu}_1 = 16.12, \hat{\mu}_2 = 39.983)$.

chains and their limiting distribution, for the two semiparametric EM algorithms proposed.

References

[1] Bordes, L., Delmas, C. and Vandekerkhove, P. (2005). Semiparametric estimation of a two-component mixture model when a component is known. *Submitted for publication.*

[2] Bordes, L., Mottelet, S. and Vandekerkhove, P. (2005). Semiparametric estimation of a two-component mixture model, *Ann. Statist.* (to appear).

[3] Celeux, G. and Diebolt, J. (1992). A Stochastic Approximation Type EM Algorithm for the Mixture Problem, *Stoch. Stoch. Rep.* **41**, 119–134.

[4] Chauveau, D. (1995). A Stochastic EM Algorithm for Mixtures with Censored Data, *J. Statist. Plann. Inference*, **46**, 1–25.

[5] Cruz-Medina, I. R. and Hettmansperger, T. P. (2004). Nonparametric estimation in semi-parametric univariate mixture models. *J. Stat. Comput. Simul.*, **74**, 513–524.

[6] Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist.*, **27**, 1178–1209.

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

[8] Hall, P. (1981). On the nonparametric estimation of mixture proportions. *J. Roy. Statist. Soc. Ser. B*, **43**, 147–156.

[9] Hall, P. and Zhou, X-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.*, **31**, 201–224.

[10] Hettmansperger, T. P. and Thomas, H. (2000). Almost nonparametric inference for repeated measures in mixture models. *J. Roy. Statist. Soc. Ser. B*, **62**, 811–825.

[11] Hunter, D.R., Wang, S. and Hettmansperger, T.P. (2004). *Inference for mixtures of symmetric distributions.* Tech. report 04-01, Penn State University.

[12] Lemdani, M. and Pons, O. (1999). Likelihood ratio tests in contamination models. *Bernoulli*, **5**, 705–719.

[13] Leroux, B.G. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.*, **20**, 1350–1360.

[14] Lindsay, B.G. (1995). *Mixture Models: Theory, Geometry and Applications.* NSFCBMS Regional Conference Series in Probability and Statistics, **5**, IMS and ASA.

[15] McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* Wiley, New York.

[16] Mclachlan, G. and Peel, D.A. (2000). *Finite Mixture Models.* Wiley, New York.

[17] McNeil D.R. (1977). *Interactive Data Analysis*, Wiley, New York.

[18] Redner, R. A. and Walker, H. F. (1984). Mixtures densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**, 195–249.

[19] Robin, S., Bar-Hen, A. and Daudin, J.J. (2005). *A semiparametric approach for mixture models: Application to local FDR estimation.* Preprint INA/INRIA, France.

[20] Sundberg, R. (1974). Maximum Likelihood Theory for Incomplete Data from an Exponential Family, *Scand. J. Statist.*, **1**, 49–58.

[21] Titterington, D. M. (1983). Minimum-distance non-parametric estimation of mixture proportions. *J. Roy. Statist. Soc. Ser. B*, **45**, 37–46.

[22] Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, Chichester.

[23] Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms, *J. Amer. Statist. Assoc.*, **85** , 699–704.

[24] Wu, C.F. (1983). On the convergence properties of the EM algorithm, *Ann. Statist.*, **11**, 95–103.

Corresponding author
Didier Chauveau
Laboratoire MAPMO - UMR 6628 - Fédération Denis Poisson
Université d'Orléans
BP 6759, 45067 Orléans cedex 2, FRANCE.
Email: `didier.chauveau@univ-orleans.fr`