



**HAL**  
open science

# EM and Stochastic EM algorithms for reliability mixture models under random censoring

Laurent Bordes, Didier Chauveau

► **To cite this version:**

Laurent Bordes, Didier Chauveau. EM and Stochastic EM algorithms for reliability mixture models under random censoring. 2012. hal-00685823v1

**HAL Id: hal-00685823**

**<https://hal.science/hal-00685823v1>**

Preprint submitted on 6 Apr 2012 (v1), last revised 10 Jun 2013 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# EM and Stochastic EM algorithms for reliability mixture models under random censoring

Laurent BORDES<sup>1</sup>

Didier CHAUVEAU<sup>2</sup>

<sup>1</sup> LMA - UMR CNRS 5142

Université de Pau et des Pays de l'Adour

laurent.bordes@univ-pau.fr

<sup>2</sup>MAPMO - CNRS UMR 7349 - Fédération Denis Poisson

Université d'Orléans

didier.chauveau@univ-orleans.fr

April 6, 2012

## Abstract

We present in this paper several iterative methods based on EM and Stochastic EM methodology, that allow to estimate parametric or semiparametric mixture models for randomly right censored lifetime data, provided they are identifiable. We consider different levels of completion for the (incomplete) observed data, and provide genuine or EM-like algorithms for several situations. In particular, we show that in censored semiparametric situations, a stochastic step is the only practical solution allowing computation of nonparametric estimates of the unknown survival function. The effectiveness of the new proposed algorithms is demonstrated in simulation studies.

**Keywords.** Censored data; EM algorithm; finite mixture; semiparametric mixtures; reliability; survival data.

## 1 Introduction

Estimating unknown parameters of a reliability mixture model may be a more or less intricate problem, especially if durations are censored. In the parametric framework one approach consists in minimizing the distance between a parametric distribution and its nonparametric estimate. Several distances may be chosen: e.g. Hellinger in Karunamuni and Wu (2009) or Cramér-von Mises in Beutner and Bordes (2011). These methods fail to account semiparametric mixture models without training data. There are many iterative algorithms to reach mixture models maximum likelihood estimates, mostly in the well-known class of EM algorithms (see section 1.1 below),

but few of them handle the additional problem of censoring. Chauveau (1995) proposed extensions of EM and of the Stochastic EM algorithm (Celeux and Diebolt, 1986) to handle Type-I deterministic right censoring. One advantage of the Stochastic EM algorithm is that it can be extended easily to some semiparametric mixture models provided they are identifiable (see e.g. Bordes et al. (2007)).

We present in this paper several iterative methods based on Monte Carlo simulation and Stochastic EM-like algorithms for estimation of identifiable (semi-)parametric right censored reliability mixture models. We first detail in this section the general model and notations that will be used throughout the paper. The objective is to fit  $n$  independent and identically distributed (i.i.d.) lifetime observations taking values in  $\mathbb{R}^+$ , from a lifetime density

$$\mathbf{X} = (X_1, \dots, X_n) \text{ i.i.d. } \sim g(x|\boldsymbol{\theta}),$$

where  $\boldsymbol{\theta}$  denotes the model parameter. It will always be assumed that these lifetime data come from a finite mixture of  $m$  components, i.e.

$$g(x|\boldsymbol{\theta}) = \sum_{j=1}^m \lambda_j f_j(x), \quad \boldsymbol{\theta} = (\boldsymbol{\lambda}, \mathbf{f}), \quad (1)$$

where  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$  are the component weights satisfying  $\sum_{j=1}^m \lambda_j = 1$ , and  $\mathbf{f} = (f_1, \dots, f_m)$  are the component densities. We define the cumulative density function (cdf) of component  $j$  by  $F_j(\cdot)$ , the mixture cdf by  $G(x|\boldsymbol{\theta}) = \sum_{j=1}^m \lambda_j F_j(x)$ , with corresponding survival (reliability) functions  $\bar{F}_j(\cdot) = 1 - F_j(\cdot)$  and  $\bar{G}(\cdot) = 1 - G(\cdot)$ .

We will in addition often allow the models to handle random right-censored data. This random censoring process is described by a random variable  $C$  with density function  $h$ , cdf  $H$  and survival function  $\bar{H}(\cdot) = 1 - H(\cdot)$ . In the right censoring setup the available information is

$$T = \min(X, C), \quad D = \mathbb{I}(X \leq C).$$

The  $n$  lifetime data  $\mathbf{x} = (x_1, \dots, x_n)$  iid  $\sim g$  are associated to  $n$  censoring times  $\mathbf{c} = (c_1, \dots, c_n)$  iid  $\sim h$ . The observations are finally in the censoring case

$$(\mathbf{t}, \mathbf{d}) = ((t_1, d_1), \dots, (t_n, d_n)),$$

where  $t_i = \min(x_i, c_i)$  and  $d_i = \mathbb{I}(x_i \leq c_i)$ ,  $i = 1, \dots, n$ .

## 1.1 Missing data and EM algorithms

The association of EM algorithms with mixture models has a long history since the seminal paper of Dempster et al. (1977) in which the initials ‘‘EM’’ were coined, and a finite mixture model was presented as a missing data example. A recent account of EM principle, properties and generalizations can be found in McLachlan and

Krishnan (2008), and mixture models are deeply detailed in McLachlan and Peel (2000). In the missing data setup, the  $n$ -fold product of  $g$ , say  $\mathbf{g}(\cdot|\boldsymbol{\theta})$ , corresponds to the *incomplete* data pdf, associated to the log-likelihood  $\ell_{\mathbf{x}}(\boldsymbol{\theta}) = \sum_{i=1}^n \log g(x_i|\boldsymbol{\theta})$ . When maximizing  $\ell_{\mathbf{x}}(\boldsymbol{\theta})$  leads to a difficult problem (such as in, e.g., the mixture model situation), considering  $\mathbf{x}$  as an incomplete data resulting from a non-observed and suitable *complete* data  $\mathbf{y}$  often helps. Assuming  $\mathbf{y}$  comes from a complete data pdf  $\mathbf{g}^c$ , the EM algorithm iteratively maximizes the operator

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k) = \mathbb{E}[\log \mathbf{g}^c(\mathbf{y}|\boldsymbol{\theta})|\mathbf{x}, \boldsymbol{\theta}^k],$$

the expectation being taken relatively to the conditional distribution of  $(\mathbf{y}|\mathbf{x})$ , for the value  $\boldsymbol{\theta}^k$  of the parameter at iteration  $k$ . Given an arbitrary starting value  $\boldsymbol{\theta}^0$ , the EM algorithm generates a (deterministic) sequence  $(\boldsymbol{\theta}^k, k = 1, 2, \dots)$  :

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

In the present setup the observed data  $(\mathbf{t}, \mathbf{d})$  depends on  $\mathbf{x}$  which comes from a finite mixture with pdf  $g$ , hence missing data are naturally involved (McLachlan and Peel, 2000). In the mixture framework, the complete data associated to  $\mathbf{x}$  correspond to the situation where the component of origin  $z \in \{1, \dots, m\}$  of each individual lifetime  $x$  is observed. The complete data distribution of  $(X, Z)$  is given by  $\mathbb{P}_{\boldsymbol{\theta}}(Z = z) = \lambda_z$  and  $(X|Z = z) \sim f_z$ .

When censored lifetimes  $(\mathbf{t}, \mathbf{d})$  are observed, we may think of two stages for the associated complete data: the simplest one is to consider the component indicators  $\mathbf{z} = (z_1, \dots, z_n)$  as missing like in the usual mixture situation, so that  $(\mathbf{t}, \mathbf{d}, \mathbf{z})$  is the complete data. But we can also consider in addition the censored observations  $(x_i, i \in \{1, \dots, n\} : d_i = 0)$  as missing (this is the case in Chauveau (1995) for deterministic censoring), so that the complete data is  $(\mathbf{x}, \mathbf{z})$ . This latter model allows in the stochastic EM machinery (introduced in Section 2.3) to plug standard MLE of the parameters for simple random sample from each population, whereas the former gives a simpler algorithmic implementation but requires MLE on censored data, that may be more complex numerically, as it will be illustrated in further Sections.

## 1.2 Nonparametric estimation under censoring

We recall here some classical results concerning estimation in nonparametric situations, when the available data are  $(\mathbf{t}, \mathbf{d})$  from a single (i.e., non mixture) lifetime distribution  $F$ . Let us introduce the two counting processes  $N$  and  $Y$  defined by

$$N(t) = \sum_{i=1}^n \mathbb{I}(t_i \leq t, d_i = 1) \quad \text{and} \quad Y(t) = \sum_{i=1}^n \mathbb{I}(t_i \geq t) \quad t \geq 0,$$

counting respectively the number of failures in  $[0, t]$  and the number of items at risk at time  $t^-$ . The Nelson-Aalen estimator of the cumulative hazard rate function  $A$  is defined by

$$\hat{A}(t) = \int_0^t \frac{dN(s)}{Y(s)} = \sum_{\{i:t_i \leq t\}} \frac{\Delta N(t_i)}{Y(t_i)} \quad t \geq 0,$$

where  $\Delta N(s) = N(s) - N(s^-)$ . The Kaplan-Meier estimator of the survival function  $\bar{F}$  is defined by

$$\hat{\bar{F}}(t) = \prod_{s \leq t} (1 - \Delta \hat{A}(s)) = \prod_{s \leq t} \left(1 - \frac{\Delta N(s)}{Y(s)}\right) \quad t \geq 0.$$

Let  $\mathcal{K}$  be a kernel function and  $b_n$  a bandwidth satisfying  $b_n \searrow 0$  and  $nb_n \nearrow +\infty$ , it is well known that the hazard rate function  $\alpha(\cdot) = f(\cdot)/\bar{F}(\cdot)$  can be estimated nonparametrically by

$$\hat{\alpha}(t) = \int_0^{+\infty} \mathcal{K}_{b_n}(t-s) d\hat{A}(s) = \sum_{i=1}^n \mathcal{K}_{b_n}(t-t_i) \frac{\Delta N(t_i)}{Y(t_i)},$$

where  $\mathcal{K}_{b_n}(\cdot) = b_n^{-1} \mathcal{K}(\cdot/b_n)$ . Then  $f = \alpha \times \bar{F}$  can be estimated by  $\hat{f}(t) = \hat{\alpha}(t) \hat{\bar{F}}(t)$ .

Since we consider that the unknown distribution is absolutely continuous with respect to the Lebesgue measure we have  $t_i \neq t_j$  for  $i \neq j$  with probability 1. Let us denote by  $t_{(1)} < \dots < t_{(n)}$  the ordered durations, and write  $d_{(i)}$  the corresponding censoring indicators ( $d_{(i)} = d_j$  if  $t_{(i)} = t_j$ ). The estimates can be written

$$\hat{A}(t) = \sum_{\{i:t_{(i)} \leq t\}} \frac{d_{(i)}}{n-i+1}, \quad (2)$$

$$\hat{\bar{F}}(t) = \prod_{\{i:t_{(i)} \leq t\}} \left(1 - \frac{d_{(i)}}{n-i+1}\right), \quad (3)$$

and

$$\hat{\alpha}(t) = \sum_{i=1}^n \frac{1}{b_n} \mathcal{K}\left(\frac{t-t_{(i)}}{b_n}\right) \frac{d_{(i)}}{n-i+1}. \quad (4)$$

For more properties about these estimators see, e.g., Andersen et al. (1993).

## 2 Parametric mixture model with censored data

If we assume that the  $j$ th component density is restricted to  $f_j(\cdot) = f(\cdot|\xi_j) \in \mathcal{F}$ , where  $\mathcal{F}$  is a parametric family indexed by a Euclidean parameter  $\xi \in \mathbb{R}^d$ . Model (1) becomes

$$X \sim g(x|\boldsymbol{\theta}) = \sum_{j=1}^m \lambda_j f(x|\xi_j), \quad (5)$$

where  $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\xi}) = (\lambda_1, \dots, \lambda_m, \xi_1, \dots, \xi_m)$  is the (Euclidean) model parameter. The cdf of the  $j$ th component reduces to  $F_j(\cdot) = F(\cdot|\xi_j)$ . As explained in Section 1.1, two EM algorithms can be defined in this case, depending on the desired level for the complete data.

## 2.1 EM algorithm for complete data $(\mathbf{t}, \mathbf{d}, \mathbf{z})$

We consider here that the missing information is only the component of origin of the  $n$  lifetimes. The complete data pdf (where informally densities and probabilities are denoted  $f_{\boldsymbol{\theta}}$ ) is given by

$$\begin{aligned} f_{\boldsymbol{\theta}}^c(T = t, D = 1, Z = z) &= \mathbb{P}_{\boldsymbol{\theta}}(z) f_{\boldsymbol{\theta}}(D = 1, T = t|Z = z) \\ &= \lambda_z f_{\boldsymbol{\theta}}(C \geq X, X = t|z) \\ &= \lambda_z \mathbb{P}_{\boldsymbol{\theta}}(C \geq t) f_{\boldsymbol{\theta}}(X = t|z) \\ &= \lambda_z f(t|\xi_z) \bar{H}(t), \end{aligned}$$

and similarly  $f_{\boldsymbol{\theta}}^c(t, 0, z) = \lambda_z \bar{F}(t|\xi_z) h(t)$ . This can be summarized by

$$f^c(t, d, z|\boldsymbol{\theta}) = [\lambda_z f(t|\xi_z) \bar{H}(t)]^d [\lambda_z \bar{F}(t|\xi_z) h(t)]^{1-d}. \quad (6)$$

The observed data log-likelihood is then given by taking the marginal of the complete-data pdf w.r.t.  $z$ ,

$$\begin{aligned} \ell_{\mathbf{t}, \mathbf{d}}(\boldsymbol{\theta}) &= \log \left( \prod_{i=1}^n f(t_i, d_i|\boldsymbol{\theta}) \right) \\ &= \sum_{i=1}^n \log \left( \bar{H}(t_i)^{d_i} h(t_i)^{1-d_i} \right) + \sum_{i=1}^n \log \left( \sum_{j=1}^m \lambda_j f(t_i|\xi_j)^{d_i} \bar{F}(t_i|\xi_j)^{1-d_i} \right), \end{aligned}$$

where the first sum does not depend on  $\boldsymbol{\theta}$ . The EM methodology amounts here to iteratively maximize

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k) = \mathbb{E} \left[ \log \mathbf{f}^c(\mathbf{t}, \mathbf{d}, \mathbf{Z}|\boldsymbol{\theta}) \mid \mathbf{t}, \mathbf{d}, \boldsymbol{\theta}^k \right] = \sum_{i=1}^n \mathbb{E} \left[ \log f^c(t_i, d_i, Z_i|\boldsymbol{\theta}) \mid t_i, d_i, \boldsymbol{\theta}^k \right],$$

where  $\mathbf{f}^c$  denotes the  $n$ -fold product of  $f^c$ , and the rightmost term comes from the iid assumption on the complete data. Computing this expectation (in  $Z$ ) requires the essential ingredient of any EM algorithm for finite mixture, i.e. the definition of  $p_{ij}^k$ , the *posterior* probability that the  $i$ th observation (an observed or censored lifetime) belongs to component  $j$ , conditional on the data and the current value of

the parameter at iteration  $k$  :

$$\begin{aligned} p_{ij}^k &:= \mathbb{P}(Z_i = j | t_i, d_i, \boldsymbol{\theta}^k) \\ &= \left( \frac{\lambda_j^k f(t_i | \xi_j^k) \bar{H}(t_i)}{\sum_{\ell=1}^p \lambda_\ell^k f(t_i | \xi_\ell^k) \bar{H}(t_i)} \right)^{d_i} \left( \frac{\lambda_j^k \bar{F}(t_i | \xi_j^k) h(t_i)}{\sum_{\ell=1}^p \lambda_\ell^k \bar{F}(t_i | \xi_\ell^k) h(t_i)} \right)^{1-d_i} \\ &= \left( \frac{\lambda_j^k f(t_i | \xi_j^k)}{\sum_{\ell=1}^p \lambda_\ell^k f(t_i | \xi_\ell^k)} \right)^{d_i} \left( \frac{\lambda_j^k \bar{F}(t_i | \xi_j^k)}{\sum_{\ell=1}^p \lambda_\ell^k \bar{F}(t_i | \xi_\ell^k)} \right)^{1-d_i} \end{aligned} \quad (7)$$

$$= \lambda_j^k \bar{F}(t_i | \xi_j^k) \left( \frac{\alpha(t_i | \xi_j^k)}{\sum_{\ell=1}^p \lambda_\ell^k \alpha(t_i | \xi_\ell^k) \bar{F}(t_i | \xi_\ell^k)} \right)^{d_i} \left( \sum_{\ell=1}^p \lambda_\ell^k \bar{F}(t_i | \xi_\ell^k) \right)^{d_i-1}, \quad (8)$$

where equation (8) is a rewriting of equation (7) using only survival and hazard rate function for component  $j$ ,  $\alpha(\cdot | \xi_j) = f(\cdot | \xi_j) / \bar{F}(\cdot | \xi_j)$ , that will be used later in Section 3.2. Note that these posterior probabilities do not depend on the censoring distribution. Then

$$\begin{aligned} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^k) &= \sum_{i=1}^n \sum_{j=1}^m p_{ij}^k [\log(\lambda_j) + d_i \log f(t_i | \xi_j) + (1 - d_i) \log \bar{F}(t_i | \xi_j)] \\ &\quad + R(\mathbf{t}, \mathbf{d}, \boldsymbol{\theta}^k, h), \end{aligned} \quad (9)$$

where the remaining term  $R(\mathbf{t}, \mathbf{d}, \boldsymbol{\theta}^k, h)$  does not depend on  $\boldsymbol{\theta}$  but only on the data, current parameter and censoring distribution. The maximization for  $\boldsymbol{\lambda}$  is straightforward and does not depend on the parametric family considered. Hence the EM implementation is straightforward if the maximization of  $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^k)$  in  $\boldsymbol{\xi}$  is feasible for the parametric family  $\mathcal{F}$ .

**Example 1** If  $\mathcal{F}$  is the family of exponential distributions with rate parameter  $\xi > 0$ ,  $f(x | \xi) = \xi \exp(-\xi x)$ , the iteration  $\boldsymbol{\theta}^k \rightarrow \boldsymbol{\theta}^{k+1}$  for the parametric case and complete data  $(\mathbf{t}, \mathbf{d}, \mathbf{z})$  is given by:

1. **E-step:** Calculate the posterior probabilities  $p_{ij}^k$  using Equation (7), for all  $i = 1, \dots, n$  and  $j = 1, \dots, m$ .
2. **M-step:** Set

$$\begin{aligned} \lambda_j^{k+1} &= \frac{1}{n} \sum_{i=1}^n p_{ij}^k \quad \text{for } j = 1, \dots, m \\ \xi_j^{k+1} &= \frac{\sum_{i=1}^n p_{ij}^k d_i}{\sum_{i=1}^n p_{ij}^k t_i} \quad \text{for } j = 1, \dots, m. \end{aligned}$$

This algorithm behavior is illustrated in Section 4.

## 2.2 EM algorithm for complete data $(\mathbf{x}, \mathbf{z})$

In this case the complete data pdf is given by  $f^c(X = x, Z = z|\boldsymbol{\theta}) = \lambda_z f(x|\xi_z)$ , and the missing information is  $(\mathbf{z}, (x_i > t_i, i \in \{1, \dots, n\} : d_i = 0))$ . The EM methodology aims to iteratively maximize

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k) = \mathbb{E} \left[ \log \mathbf{f}^c(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \mid \mathbf{t}, \mathbf{d}, \boldsymbol{\theta}^k \right] = \sum_{i=1}^n \mathbb{E} \left[ \log f^c(X_i, Z_i|\boldsymbol{\theta}) \mid t_i, d_i, \boldsymbol{\theta}^k \right],$$

where each expectation is in this case w.r.t.  $(X_i, Z_i)$ . Computing this expectation requires the (posterior) probability of  $(X_i, Z_i)$  given  $T_i = t_i$ ,  $D_i = d_i$  and for the parameter value  $\boldsymbol{\theta}^k$ . Since the distribution of  $(X|t, d, Z = j)$  is a Dirac measure  $\delta_t$  when  $d = 1$ , and the pdf of  $(X|X > t, Z = j)$  when  $d = 0$ , we get

$$\begin{aligned} f(x, j|t_i, d_i, \boldsymbol{\theta}^k) &= p_{ij}^k f(x|t_i, d_i, Z = j, \boldsymbol{\theta}^k) \\ &= \lambda_j^k \left( \frac{\mathbb{I}(x = t_i) f(t_i|\xi_j^k)}{\sum_{\ell=1}^m \lambda_\ell^k f(t_i|\xi_\ell^k)} \right)^{d_i} \left( \frac{\mathbb{I}(x > t_i) f(x|\xi_j^k)}{\sum_{\ell=1}^m \lambda_\ell^k \bar{F}(t_i|\xi_\ell^k)} \right)^{1-d_i}. \end{aligned}$$

Again these posterior probabilities do not depend on the censoring distribution which cancels out. Then

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k) &= \sum_{i=1}^n \sum_{j=1}^m p_{ij}^k \log(\lambda_j) \\ &\quad + \sum_{i=1}^n \sum_{j=1}^m p_{ij}^k \left[ d_i \log f(t_i|\xi_j) + (1 - d_i) \int_{t_i}^{+\infty} \log f(x|\xi_j) \frac{f(x|\xi_j^k)}{\bar{F}(t_i|\xi_j^k)} dx \right]. \end{aligned}$$

Note that as far as  $\boldsymbol{\lambda}$  is concerned, this expression is exactly Equation (9) for the case where  $(\mathbf{t}, \mathbf{d}, \mathbf{z})$  is the complete data, so that the M-step for  $\boldsymbol{\lambda}$  is identical in both situations. This EM implementation is not straightforward in general since, except for very specific parametric families, calculation of  $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$  is not obtained in closed form and has to be calculated and maximized by numerical methods.

**Example 2** If  $\mathcal{F}$  is the family of exponential distributions with rate parameter  $\xi > 0$ ,  $f(x|\xi) = \xi \exp(-\xi x)$  and  $\bar{F}(x|\xi) = \exp(-\xi x)$ . The iteration  $\boldsymbol{\theta}^k \rightarrow \boldsymbol{\theta}^{k+1}$  for the parametric case and complete data  $(\mathbf{x}, \mathbf{z})$  is given by:

1. **E-step:** Calculate the posterior probabilities  $p_{ij}^k$  as in Equation (7), for all  $i = 1, \dots, n$  and  $j = 1, \dots, m$ .
2. **M-step:** Set

$$\begin{aligned} \lambda_j^{k+1} &= \frac{1}{n} \sum_{i=1}^n p_{ij}^k \quad \text{for } j = 1, \dots, m \\ \xi_j^{k+1} &= \frac{\sum_{i=1}^n p_{ij}^k}{\sum_{i=1}^n p_{ij}^k \left( t_i + (1 - d_i)/\xi_j^k \right)} \quad \text{for } j = 1, \dots, m. \end{aligned}$$



Note that the update for  $\xi_j$  can also be written as the weighted average

$$\xi_j^{k+1} = \frac{\sum_{i=1}^n p_{ij}^k}{\sum_{i=1}^n p_{ij}^k \left( d_i t_i + (1 - d_i)(t_i + 1/\xi_j^k) \right)}$$

which means that each observed failure ( $d_i = 1$ ) contributes with  $t_i$ , and each censored lifetime ( $d_i = 0$ ) contributes with  $t_i + 1/\xi_j^k$ , as in simple censored sample case. This algorithm behavior is illustrated in Section 4.

### 2.3 Stochastic EM algorithms

The advantage of using a genuine EM algorithm as in Sections 2.1 and 2.2 is that it has a provable ascent property for the observed log-likelihood, as any EM does. On the other hand, using an EM algorithm requires the implementation of the M-step for the component parameters (the  $\xi_j$ 's), which is specific of the parametric family  $\mathcal{F}$ , and may often be tedious, particularly here where expression of survival functions are needed (e.g. for deterministic censored mixtures of Weibull distributions, see Chauveau (1995)).

When this is the case, Stochastic versions of EM like the one initially introduced by Celeux and Diebolt (1986) may overcome this difficulty at the expense of the loss of the ascent property, and more complicated convergence properties (general results from Nielsen (2000) give conditions for convergence of the sequence of estimates, which is generally a Markov Chain). In the mixture problem without censoring, the Stochastic EM (St-EM) principle consists in simulating the missing data  $\mathbf{z}$  from the posterior probabilities, so that the estimation gets back to standard MLE procedures applied to  $m$  simple random samples, allowing in particular usage of standard MLE software packages.

In the present parametric setup with censoring, we have considered two kinds of missing data: because of the mixture the  $\mathbf{z}$  are missing; but due to the censoring process, some of the  $x_i$ 's are also missing (replaced by the incomplete observations  $\{t_i, i = 1, \dots, n : d_i = 0\}$ ). We may think about simulating all the missing data as in Chauveau (1995) in the deterministic type-I censoring case, or simulating just the indicator  $\mathbf{z}$ , which has been the preferred solution here (i.e. the complete data are  $(\mathbf{t}, \mathbf{d}, \mathbf{z})$ ), since it allows straightforward M-step implementation by calling standard MLE packages for right censored data from standard distributions, as e.g. the `survival` package (Therneau and Lumley, 2009) for the R statistical software (R Development Core Team, 2010). Moreover simulation of  $\mathbf{z}$  alone seems to be up to now the only practical solution for the semi-parametric extensions we propose later on. An example of this St-EM approach for censored mixtures of Weibull distributions is given in Section 4.2.

Let  $\mathbf{p}_i^k = (p_{i1}^k, \dots, p_{im}^k)$  denote the posterior probability vector associated to observation  $i$ , and  $Z_i \sim \text{Mult}(1, \mathbf{p}_i^k)$  a multinomial distributed random variable with

parameters 1 and  $\mathbf{p}_i^k$  (i.e.,  $Z_i \in \{1, \dots, m\}$  with probabilities  $\mathbb{P}(Z_i = j) = p_{ij}^k$ ). The iteration  $\boldsymbol{\theta}^k \rightarrow \boldsymbol{\theta}^{k+1}$  of the St-EM algorithm is given by:

1. **E-step:** Calculate  $p_{ij}^k$  from Equation (7), for all  $i = 1, \dots, n$  and  $j = 1, \dots, m$ .
2. **Stochastic step:** Simulate  $Z_i^k \sim \text{Mult}(1, \mathbf{p}_i^k)$ ,  $i = 1, \dots, n$ , and define the subsets

$$\chi_j^k = \{i \in \{1, \dots, n\} : Z_i^k = j\}, \quad j = 1, \dots, m. \quad (10)$$

3. **M-step:** For each component  $j \in \{1, \dots, m\}$

$$\lambda_j^{k+1} = \text{Card}(\chi_j^k)/n,$$

and

$$\xi_j^{k+1} = \arg \max_{\xi \in \Xi} L_j(\xi), \quad (11)$$

where

$$L_j(\xi) = \prod_{i \in \chi_j^k} (f(t_i|\xi))^{d_i} (\bar{F}(t_i|\xi))^{1-d_i}. \quad (12)$$

**Asymptotic results of ergodic averages from a St-EM algorithm** From Nielsen (2000), several asymptotic results may be derived under regularity assumptions. Let us temporarily add the subscript  $n$  to  $\boldsymbol{\theta}^k$  to remember that our estimates depend on both  $k$  and  $n$ ) The main important result deals with the asymptotic behavior (in  $n$ ) of the weak limit (in  $k$ )  $\boldsymbol{\theta}_n$  of  $(\boldsymbol{\theta}_n^k)_{k \geq 0}$ . Nielsen (2000) shows that if  $\boldsymbol{\theta}_0$  is the true value of the unknown parameter  $\boldsymbol{\theta}$ , then  $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$  converges in distribution to a centered Gaussian random vector with variance-covariance matrix  $I^{-1}(\boldsymbol{\theta}_0) [2I - \{I + F(\boldsymbol{\theta}_0)\}^{-1}]$  where  $I(\boldsymbol{\theta}_0)$  denotes the observed data information and  $F(\boldsymbol{\theta}_0)$  is the expected fraction of missing information. As a consequence the basic Stochastic EM algorithm produces estimators whose the asymptotic variance can be divided into a model part  $I^{-1}(\boldsymbol{\theta}_0)$  and a simulation part  $I^{-1}(\boldsymbol{\theta}_0) [I - \{I + F(\boldsymbol{\theta}_0)\}^{-1}]$ . Various strategies can be used to reduce the simulation part of the variance. For example Nielsen shows that averaging the last  $J$  iterations of the Markov chain, i.e. taking the weak limit (in  $k$ ) of  $(\boldsymbol{\theta}_n^{k-J+1} + \dots + \boldsymbol{\theta}_n^k)/J$ , reduces the simulation part of the asymptotic variance to a  $\mathcal{O}(1/J)$ . Similar asymptotic results are obtained by increasing the number of simulation per iteration as for example in the MCEM algorithm by Wei and Tanner (1990).

In this paper two different strategies are proposed. Either we use the standard St-EM algorithm that produces a Markov chain  $(\boldsymbol{\theta}_n^k)_{k \geq 0}$  and our final estimate is the ergodic mean  $\bar{\boldsymbol{\theta}}^K$  of the first  $K$  iterates, or at iteration  $k+1$  missing data are simulated by fixing  $\boldsymbol{\theta}$  to the mean of the first  $k$  iterates, i.e.  $\boldsymbol{\theta} = (\boldsymbol{\theta}^1 + \dots + \boldsymbol{\theta}^k)/k$ . Again the final estimate of  $\boldsymbol{\theta}$  is nothing but the ergodic mean  $\bar{\boldsymbol{\theta}}^K$  of the first  $K$  iterates. In the latter case the sequence  $(\boldsymbol{\theta}_n^k)_{k \geq 0}$  is no longer a Markov chain but it

generally results in more stable algorithm and in both cases we can expect that the variance part due to simulation is almost deleted. Then the asymptotic variance of the estimator can be derived following the method of Louis (1982) (see also formula (42) in Nielsen (2000)).

### 3 Semiparametric two-components mixture models

We consider now a mixture of accelerated lifetime model, where two lifetime populations are mixed with lifetime distributions equal up to a scale parameter:

$$g(x|\lambda, \xi, f) = \lambda f(x) + (1 - \lambda)\xi f(\xi x), \quad x > 0. \quad (13)$$

This model means that the lifetime has the distribution of a r.v. (say)  $U \sim f$  when belonging to component 1, and of  $U/\xi$  when belonging to component 2. The unknown parameter is  $\boldsymbol{\theta} = (\lambda, \xi, f) \in (0, 1) \times \mathbb{R}_*^+ \times \mathcal{F}$  where  $\mathcal{F}$  is a set of density functions. Let us define by  $\boldsymbol{\theta}^0 = (\lambda^0, \xi^0, f^0)$  and  $\boldsymbol{\theta}^k = (\lambda^k, \xi^k, f^k)$  the initial and current values of  $\boldsymbol{\theta}$ .

Nonparametric or semi-parametric models like (13) are generally not identifiable without additional assumptions on the underlying nonparametric densities. Indeed identifiability, whereby the distribution of the data uniquely determines the parameter values, is a difficult question. In the particular case of model (13) where  $X$  is distributed according to a mixture of scaled random variables  $f$ -distributed, transforming  $X$  to  $Y = \log(X)$  gives a mixture of the common density  $\varphi(y) = e^y f(e^y)$  differing only by a shift parameter. It has been proved that if  $\varphi(\cdot)$  is even, then such shift-location semiparametric mixtures are identifiable (see Bordes et al. (2006) for the two-component case, and Hunter et al. (2007) for the two and three-component cases). Therefore, identifiability of model (13) holds if the density function  $f$  satisfies the constraint that  $y \mapsto e^y f(e^y)$  is an even function. This family of density functions includes for example log-normal distributions.

#### 3.1 Semiparametric St-EM algorithm without censoring

If observations from the scale mixture model (13) are uncensored then only a  $n$ -sample  $\mathbf{t}$  is observed (note that  $\mathbf{t}$  is in this case nothing but  $\mathbf{x}$  since there is no censored data).

**Step 1.** For each item  $i \in \{1, \dots, n\}$ :

$$p_{i1}^k = \frac{\lambda^k f^k(t_i)}{\lambda^k f^k(t_i) + (1 - \lambda^k)\xi^k f^k(\xi^k t_i)},$$

then set  $\mathbf{p}_i^k = (p_{i1}^k, 1 - p_{i1}^k)$ .

**Step 2.** Simulate  $Z_i^k \sim \text{Mult}(1, \mathbf{p}_i^k)$  and define subsets

$$\chi_j^k = \{i \in \{1, \dots, n\} : Z_i^k = j\}, \quad j = 1, 2.$$

**Step 3.** Update Euclidean parameters:

$$\begin{aligned} \lambda^{k+1} &= n_1/n \quad \text{where } n_1 = \text{Card}(\chi_1^k), \\ \xi^{k+1} &= \frac{n - n_1}{n_1} \frac{\sum_{i \in \chi_1^k} t_i}{\sum_{i \in \chi_2^k} t_i} \end{aligned}$$

**Step 3'.** Update the functional parameters  $f$ : Let  $\mathbf{t}^k = (t_1^k, \dots, t_n^k)$  be the “unscaled sample”  $\{t_i; i \in \chi_1^k\} \cup \{\xi^k t_i; i \in \chi_2^k\}$ . Set:

$$f^{k+1}(x) = \sum_{i=1}^n \frac{1}{nb} \mathcal{K} \left( \frac{x - t_i^k}{b} \right), \quad (14)$$

where  $\mathcal{K}$  is a kernel function and  $b$  a bandwidth.

**Remark:** Note that at the third step  $\xi^k$  is updated using a moment estimation method instead of a Maximum Likelihood principle. This latter method is hard to use here since it requires to estimate nonparametrically the first derivative of  $f$  which generally leads to unstable estimates. This and the additional nonparametric step 3' may precludes application of general results from Nielsen (2000) on St-EM convergence. Hence convergence of this algorithm is yet only based on empirical numerical evidence.

### 3.2 Semiparametric St-EM algorithm for right censored data

We consider now that lifetime data from the scale mixture model (13) are randomly censored like in Section 1, so that only a  $n$ -sample  $(\mathbf{t}, \mathbf{d})$  is observed, so that the nonparametric estimation techniques for censored data (recalled in Section 1.2) have to be used to estimate the functional parameter  $f$ . In particular, for computing the posterior probabilities using (7) or (8), estimates of the survival function  $\bar{F}$  and the hazard rate  $\alpha(\cdot)$  associated to  $f$  are needed.  $\bar{F}$  is naturally estimated by the Kaplan-Meier estimator (3), and  $\alpha(\cdot)$  by the kernel estimate (4). It appears that modified versions of kernel density estimates can be imbedded in “EM-like” algorithms for semi- or non-parametric mixtures. For example, Benaglia et al. (2009a) define a weighted kernel density estimate for the density  $f_j$  of component  $j$ , in which the  $i$ th observation is weighted according to its posterior probability  $p_{ij}^k$  at step  $k$ . Unfortunately, there is no direct way to use these posterior probabilities to define a similar weighted version of the Kaplan-Meier estimate (3). In this case, stochastic versions of the EM algorithm provide the workable solutions that we propose here: simulation of  $\mathbf{z}$  as in Section 2.3 allows to define at each iteration sub-samples corresponding to each component, from which Kaplan-Meier estimates can be directly computed.

**St-EM for semi-parametric scale mixture and censored data**

**Step 1.** E-step: For each item  $i \in \{1, \dots, n\}$ ,

$$\text{if } d_i = 0, \quad p_{i1}^k = \frac{\lambda^k \bar{F}^k(t_i)}{\lambda^k \bar{F}^k(t_i) + (1 - \lambda^k) \bar{F}^k(\xi^k t_i)}, \quad (15)$$

else

$$p_{i1}^k = \frac{\lambda^k \alpha^k(t_i) \bar{F}^k(t_i)}{\lambda^k \alpha^k(t_i) \bar{F}^k(t_i) + (1 - \lambda^k) \xi^k \alpha^k(\xi^k t_i) \bar{F}^k(\xi^k t_i)}, \quad (16)$$

then set  $\mathbf{p}_i^k = (p_{i1}^k, 1 - p_{i1}^k)$ .

**Step 2.** Stochastic step: Simulate  $Z_i^k \sim \text{Mult}(1, \mathbf{p}_i^k)$ ,  $i = 1, \dots, n$ , and define the subsets

$$\chi_j^k = \{i \in \{1, \dots, n\} : Z_i^k = j\}, \quad j = 1, 2. \quad (17)$$

**Step 3.** Update the Euclidean parameters:

$$\lambda^{k+1} = n_1/n \quad \text{where } n_1 = \text{Card}(\chi_1^k).$$

Let  $\bar{S}_j^k$  denotes the Kaplan-Meier survival estimate associated to  $\{t_i : i \in \chi_j^k\}$ , for  $j = 1, 2$ , and  $M_j^k = \max_{i \in \chi_j^k}(t_i)$ , then set

$$\xi^{k+1} = \frac{\int_0^{M_1^k} \bar{S}_1^k(s) ds}{\int_0^{M_2^k} \bar{S}_2^k(s) ds} \quad (18)$$

**Step 3'.** Update the functional parameters  $\alpha$  and  $\bar{F}$ :

Let  $\mathbf{t}^k = (t_1^k, \dots, t_n^k)$  be the order statistic of  $\{t_i; i \in \chi_1^k\} \cup \{\xi^k t_i; i \in \chi_2^k\}$ , so that  $t_1^k \leq \dots \leq t_n^k$ , and  $\mathbf{d}^k = (d_1^k, \dots, d_n^k)$  be the corresponding indicators. Set:

$$\alpha^{k+1}(x) = \sum_{i=1}^n \frac{1}{b} \mathcal{K} \left( \frac{x - t_i^k}{b} \right) \frac{d_i^k}{n - i + 1}, \quad (19)$$

and

$$\bar{F}^{k+1}(x) = \prod_{\{i: t_i^k \leq x\}} \left( 1 - \frac{d_i^k}{n - i + 1} \right), \quad (20)$$

where  $\mathcal{K}$  is a kernel function and  $b$  a bandwidth.

Note that here also  $\xi^k$  is updated using a moment estimation method instead of a Maximum Likelihood principle. Implementation considerations such as starting values, choices for kernels and bandwidths are discussed in Section 4.

## 4 Implementation and Examples

We propose in this Section some examples illustrating most of the parametric and semi-parametric, genuine and stochastic EM or EM-like algorithms that have been introduced in this paper. Note that all the algorithms shown here will be publicly available in an upcoming version of the `mixtools` package (Benaglia et al., 2009b) for the R statistical software (R Development Core Team, 2010).

As is typically the case for EM-type algorithms, the choice of the initial starting parameter values is important. When performing Monte-Carlo experiments, we always started the algorithms from the true parameter values to prevent a possible “label-switching” as much as possible. This label-switching issue arises because the particular ordering of the subscripts  $j = 1, \dots, m$  in equation such as (1) is arbitrary: A permutation of these subscripts gives exactly the same density function, so that the best we can do is to estimate the parameters up to a permutation of the labels. However, in a simulation study based on Monte-Carlo replications, this issue can lead to flawed average estimates because there is no guarantee that only estimates from the “same” component are averaged together. For a fuller account of the label-switching issue, see McLachlan and Peel (2000), and Celeux et al. (1996) for label switching issues in Stochastic-EM algorithms.

Note that starting the algorithms from the true parameter is not an “unfair” procedure, since on real data the common practice consists in starting the algorithm from several values uniformly spreaded on the parameter space (or on the region of interest of it). Then the retained EM estimate is the estimate achieving the maximum of the observed likelihood among all the trials, which is typically the location of the global maximum closest to the true parameter value.

### 4.1 EM algorithm for parametric model with censored data

The algorithms of Section 2 for mixture of lifetime distributions with censored data has been implemented for exponential densities,

$$g(x) = \sum_{j=1}^m \lambda_j \xi_j \exp(-\xi_j x) \quad x > 0,$$

and the two types of complete data  $((\mathbf{t}, \mathbf{d}, \mathbf{z})$  and  $(\mathbf{x}, \mathbf{z})$  (see Examples 1 and 2). We choose here a  $m = 3$  components model with true parameters  $\boldsymbol{\lambda} = (0.2, 0.3, 0.5)$  and  $\boldsymbol{\xi} = (4, 1, 0.02)$ . We assume that  $C$  is uniformly distributed on  $[0, c]$ , where we have chosen  $c = 150$  in order to achieve an average censoring rate of about 16%. Fig. 1 shows the good behavior of a particular EM sequence for  $n = 500$  observations among which 13.8% are censored, and even using initial values that are not the true values:  $\boldsymbol{\lambda}^0 = (1/3, 1/3, 1/3)$  is “non-informative”, and  $\boldsymbol{\xi}^0 = (5, 0.5, 0.1)$  is just ordered like the true  $\boldsymbol{\xi}$ . Estimates are accurate except for  $\xi_1$  in this particular case.

We also apply on this model the two EM strategies, i.e. for the two levels of complete data, over 300 Monte-Carlo replications started with the true  $\theta$ . As explained previously, this initialization together with a true model with somehow different rates has been chosen to avoid label switching across replications in Monte-Carlo simulations (even if the component densities are severely overlapping because of the shape of the exponential component densities). In this experiment no label switching occurred between rate parameters. Results of this Monte-Carlo experiment are given in Fig. 2, which shows the good behavior of these algorithms, similar behaviour and no clear winner between both strategies.

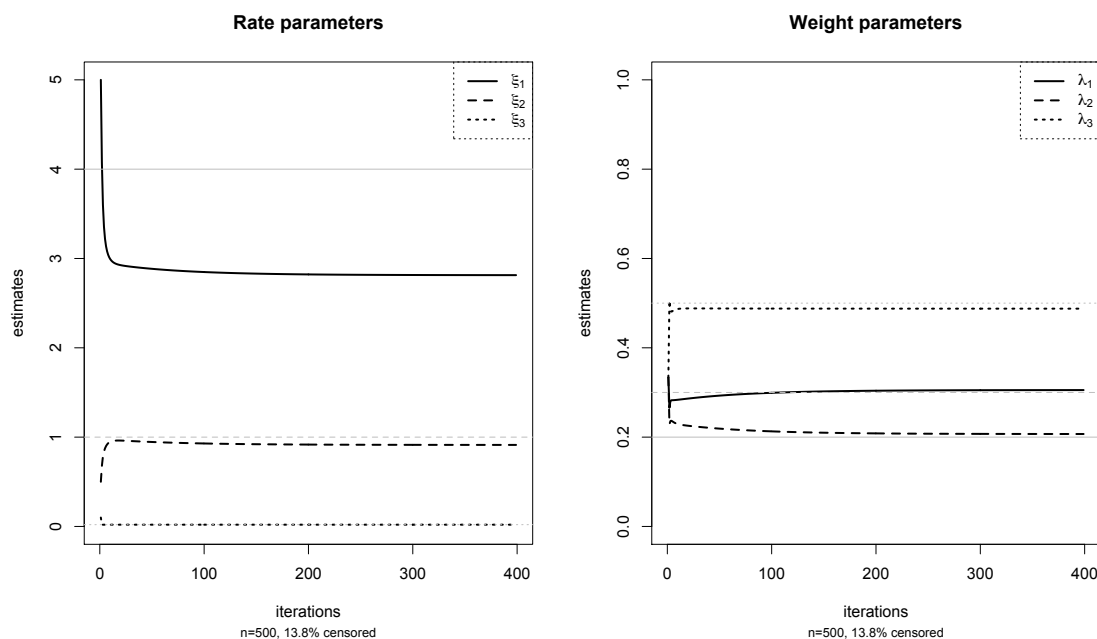


Figure 1: Sequence of output of the EM algorithm started from “arbitrary” parameters  $\lambda^0 = (1/3, 1/3, 1/3)$  and  $\xi^0 = (5, 0.5, 0.1)$ , for  $n = 500$  observations with 13.8% censored; horizontal grey lines are true values.

## 4.2 Stochastic EM for parametric model with censored data

As explained in Section 2.3, using stochastic versions of EM only makes sense when dealing with parametric families for which the M-steps of the genuine EM algorithms for censored mixture (Sections 2.1 and 2.2) are not in closed form. This is the case, e.g., for a mixture of Weibull distributions which can be expressed in terms of its survival function

$$\bar{G}(x) = \sum_{j=1}^m \lambda_j \exp \left[ - \left( \frac{x}{\eta_j} \right)^{\beta_j} \right]$$

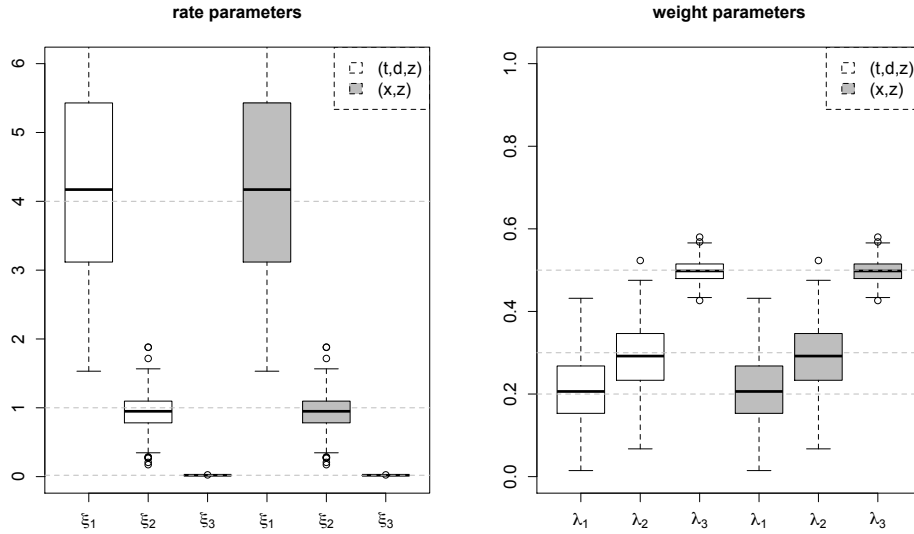


Figure 2: *Boxplots of estimates for 300 replications of two EM algorithms started from the true  $\theta$ , for  $n = 500$  sample size and average censoring rate of 16%; white boxplots are EM algorithms using complete data  $(\mathbf{t}, \mathbf{d}, \mathbf{z})$ , and grey boxplots are EM algorithms using complete data  $(\mathbf{x}, \mathbf{z})$ ; horizontal dotted lines are true values.*

with shape parameters  $\beta = (\beta_1, \dots, \beta_m)$  and scale parameters  $\eta = (\eta_1, \dots, \eta_m)$ . Weibull distributions are commonly used in reliability analysis because this distribution can capture different kinds of failure behavior through its shape parameter: infant mortality when  $0 < \beta < 1$ , constant failure rate (exponential distribution) when  $\beta = 1$ , wear-out when  $\beta > 1$ .

We have chosen to apply the parametric St-EM algorithm to a synthetic model similar to those used to model satellite reliability, as in Castet and Saleh (2010) and Dubos et al. (2010). The model we precisely focused on is a  $m = 2$  components mixture of Weibull distributions fitted on  $n = 1394$  actual lifetime data, where component 1 with  $\beta_1 = 0.4477$  and  $\eta_1 = 4102$  years captures infant mortality (representing a proportion  $\lambda_1 = 0.9466$  of the population), and component 2 with  $\beta_2 = 7.163$  and  $\eta_2 = 9.2$  years corresponds to an increasing failure rate and wear-out behavior for large satellites (see Table 5 in Dubos et al. (2010)).

We have simulated artificial data based on this actual reliability model since these data are not publicly available. Within each M-step of the St-EM algorithm, the MLE on complete data from each component  $j$ ,  $\{t_i, d_i : i \in \chi_j^k\}$ , as defined in equations (10), (11) and (12), has been implemented by calling the `survreg()` function from the `survival` package (Therneau and Lumley, 2009). It itself requires an iterative optimization method within each St-EM iteration.



The Monte-Carlo experiment consists in 300 replications of censored samples of size  $n = 1400$ , from the true parameter given above, to which we applied a somehow strong random censoring process, achieving in average 30.6% of censored observations. Each St-EM algorithm was ran for  $K = 500$  iterations using the “averaged” strategy where missing data at iteration  $k$  are simulated from the mean of the first  $k$  iterates, as discussed in Section 2.3. A careful examination of the estimates reveals no label switching issue here. Results are given in Table 1 in terms of means and standard deviations over replications, for each scalar parameter.

	true		mean		stdev	
	$j = 1$	$j = 2$	$j = 1$	$j = 2$	$j = 1$	$j = 2$
$\lambda$	0.9466	0.0534	0.9466	0.0534	0.00708	0.00708
$\beta$	0.4477	7.1630	0.449	7.457	0.0172	1.0145
$\eta$	4102.000	9.201	4178.20	9.19	473.923	0.211

Table 1: Estimated means and standard deviations from 300 replications of the parametric St-EM algorithm for a Weibull mixture with  $n = 1400$  lifetime data among which 30.6% are censored in average. All the replications were started from the true parameter value.

These results show the good behavior of this St-EM strategy in this situation where one component is characterized by a rather small weight ( $\lambda_2 \approx 5\%$ ), compensated by a large dataset ( $n = 1400$ ), and even with 30% of censored data.

### 4.3 Stochastic EM for semi-parametric model with censored data

We consider the mixture of accelerated lifetime model of Equation (13), in which the unknown parameter is  $\theta = (\lambda, \xi, f)$  chosen here to be  $\lambda = 0.4$ ,  $\xi = 0.1$  and for the nonparametric part  $f$  the density of  $\mathcal{LN}(3, 0.5)$ , a lognormal distribution with mean 3 and standard deviation 0.5 on the log scale. For brevity, we do not show results on simulated or actual data for this model without censored data, because in this case the M-step for the functional parameter uses a kernel density estimate on the unscaled sample (14) which is thus very similar in its principle to the St-EM for a location-shift mixture, as in Bordes et al. (2007). We thus simulated the censored data case, where we choose for the censoring distribution a Uniform over the interval  $[20; 1300]$ , which results in a censoring rate of about 10% of the data.

The algorithm needs for the first E-step (Equations (15) and (16)) initial values for both the Euclidean parameters ( $\lambda^0, \xi^0$ ) and the nonparametric survival and hazard rate functions  $\bar{F}^0$  and  $\alpha^0$ , evaluated at the  $t_i$ 's and the  $\xi^0 t_i$ 's. These require an initial “unscaled” sample  $\mathbf{t}^0$  which in turn requires initial component indicators ( $Z_1^0, \dots, Z_n^0$ ). We choose here to first apply a kmeans clustering algorithm on the lifetime data  $\mathbf{t}$  with initial centers chosen in a data-driven manner from the histogram of these data, from which the  $Z_i^0$ 's are assigned, and initial  $\mathbf{t}^0$ ,  $\bar{F}^0$  and  $\alpha^0$  can be built using an initial Step 3' (Equations (19) and (20)). The initial scalar parameters

have been set to the true scale parameter  $\xi^0 = 0.1$  and non-informative arbitrary  $\lambda^0 = (1/2, 1/2)$ ,

Since the algorithm is stochastic, there is no pointwise convergence to expect. The algorithm is ran up to some fixed number  $K$  of iterations, and then estimates are computed. For the Euclidean parameters we proceed as usual, by taking the empirical mean over iterations sequences,

$$\hat{\lambda} = \frac{1}{K} \sum_{k=1}^K \lambda^k; \quad \hat{\xi} = \frac{1}{K} \sum_{k=1}^K \xi^k. \quad (21)$$

For the nonparametric  $\bar{F}$  and  $\alpha$ , things are not so clear. We have suggested and tried two strategies in this experiment. The simplest one, so-called “final”, consists in plotting the  $\bar{F}^K(\cdot)$  and  $\alpha^K(\cdot)$  obtained at the last iteration and evaluated at the last unscaled sample  $\mathbf{t}^K$ . Then, since we are in the censored data situation, the density  $f$  itself can only be estimated using the plug-in estimator  $f^K(\cdot) = \bar{F}^K(\cdot)\alpha^K(\cdot)$ . The other approach we tried tries to mimic the average strategy used for the scalar parameters, and is for that denoted “average” below. It amounts to use  $\hat{\xi}$  from Equation (21) instead of  $\xi^K$  to compute an “averaged” unscaled ordered sample  $\hat{\mathbf{t}}$  from which  $\hat{\bar{F}}(\cdot)$  and  $\hat{\alpha}(\cdot)$  are estimated using an additional Step 3’ (as in (19) and (20)).x Both strategies delivered very similar estimates in our experiments.

We ran  $R = 300$  replications of  $n = 500$  observations each, and for each replication the algorithm was ran for  $K = 200$  iterations. As explained before, with stochastic EM sequences, care should be taken with possible label switching issues. A careful examination of the estimates reveals no obvious label switching issue here. In particular we get  $\hat{\lambda} < 1/2$  for all the  $B$  replications. Results are given in Table 2 in terms of means and standard deviations over replications, for each scalar parameter. In addition, we have computed the error for the unknown density estimation in terms of the Mean Integrated Squared Error (MISE) over replications:

$$\text{MISE} = \frac{1}{R} \sum_{r=1}^S \int \left( \hat{f}^{(r)}(u) - f(u) \right)^2 du \approx 0.000831$$

where  $f$  is the pdf of the lognormal distribution  $\mathcal{LN}(3, 0.5)$  and the integral is computed numerically. Each estimated density  $\hat{f}^{(r)}$  at  $r$ th replication is computed, as discussed in Section (3.2), from the product  $\hat{\bar{F}}(\cdot)\hat{\alpha}(\cdot)$ , where these estimates can be “final” or “average” versions (see above). Hence this error includes the error on the estimation of the survival function.

Table 2 provides the results over replications for the scalar parameters, where we can see that the estimation of  $\lambda$  is rather good. The estimation of  $\xi$  is also good, even though it is slightly biased. Figures 3 shows a typical result on a simulated sample of size  $n = 500$  from this model. Note that this plot is the standard output of a generic `plot()` command within R, applied to a result returned by the semiparametric St-EM

algorithm implemented in the next version of the `mixtools` package (Benaglia et al., 2009b). Figure 3 illustrates the good estimation of the scalar parameters (despite the tendency to over estimate  $\xi$ , and of the survival function through Kaplan-Meier estimates on the unscaled samples  $\mathbf{t}^K$  or  $\hat{\mathbf{t}}$ ). The typical estimate of the density  $f$  is, not surprisingly, not so precise since it includes a kernel density estimate of the hazard rate (19), which itself depends on the choices for the underlying kernel and bandwidth. Some hints about these issues are given in Section 5.

	true	mean	stdev	mse
$\lambda$	0.4	0.3995	0.0211	0.00045
$\xi$	0.1	0.1118	0.00686	0.00019

Table 2: Estimated means and standard deviations from 300 replications of the semiparametric St-EM algorithm with  $n = 500$  lifetime data among which 10% are censored in average.

## 5 Discussion

We have proposed several iterative methods based on EM and Stochastic EM methodologies, for parametric and semiparametric identifiable mixture models designed for randomly right censored lifetime data. For some simple parametric situations, it was possible to define genuine EM algorithms. For more intricate models we have shown that the introduction of stochastic steps in EM-like algorithms provide practical solutions taking advantage of the simulated complete data.

For censored semiparametric mixtures, the stochastic step is an even more attractive tool since it allows direct and per-component computation of nonparametric, Kaplan-Meier estimates of the survival function. This semiparametric St-EM algorithm also requires hazard rate kernel density estimates, that raise kernel and bandwidth issues. We have for now implemented and tested three basic kernels: the Gaussian, a triangular and an adaptive triangular kernel preventing the “mass leaking” near 0. Indeed, it is known that for nonparametric estimation from survival data on the positive real line, a Gaussian kernel is not a good choice because of this mass leaking for observations too close to 0. After some preliminary experiments, we finally apply here an adaptive triangular kernel, where “adaptive” means that the shape of the triangle is adapted for observations too close to 0, for which usage of the regular triangle at the chosen bandwidth would result in a positive mass below 0. The bandwidth has simply been set here to R default, which is probably not the best method for this model. The choices of kernel and bandwidth definitely require more investigations, that are beyond the scope of the present paper.

Finally, we reiterate that the simulations in this paper have been done with an upcoming new version of the `mixtools` package for the R statistical software (Benaglia

et al., 2009b; R Development Core Team, 2010).

## References

- Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.
- Benaglia, T., Chauveau, D., and Hunter, D. R. (2009a). An EM-like algorithm for semi-and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526.
- Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009b). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.
- Beutner, E. and Bordes, L. (2011). Estimators based on data-driven generalized weighted cramer-von mises distances under censoring - with applications to mixture models. *Scandinavian Journal of Statistics*, 38(1):108–129.
- Bordes, L., Chauveau, D., and Vandekerkhove, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics and Data Analysis*, 51(11):5429–5443.
- Bordes, L., Mottelet, S., and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Annals of Statistics*, 34(3):1204–1232.
- Castet, J.-F. and Saleh, J. H. (2010). Single versus mixture weibull distributions for nonparametric satellite reliability. *Reliability Engineering and System Safety*, 95:295–300.
- Celeux, G., Chauveau, D., and Diebolt, J. (1996). Stochastic versions of the EM algorithm: An experimental study in the mixture case. *J. Statist. Comput. Simul.*, 55:287–314.
- Celeux, G. and Diebolt, J. (1986). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2:73–82.
- Chauveau, D. (1995). A stochastic EM algorithm for mixtures with censored data. *Journal of Statistical Planning and Inference*, 46(1):1–25.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.
- Dubos, G. F., Castet, J.-F., and Saleh, J. H. (2010). Statistical reliability analysis of satellites by mass category: Does spacecraft size matter? *Acta Astronautica*, 67:584–595.

- Hunter, D. R., Wang, S., and Hettmansperger, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.*, 35(1):224–251.
- Karunamuni, R. and Wu, J. (2009). Minimum hellinger distance estimation in a non-parametric mixture model. *Journal of Statistical Planning and Inference*, 3:1118–1133.
- Louis, T. (1982). Finding the observed information matrix when using the em algorithm. *J. R. Statist. Soc. Ser. B*, 44:226–233.
- McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.
- McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. New York: Wiley.
- Nielsen, S. F. (2000). The stochastic em algorithm: Estimation and asymptotic results. *Bernoulli*, 6(3):457–489.
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Therneau, T. and Lumley, T. (2009). *survival: Survival analysis, including penalised likelihood*. R package version 2.35-8.
- Wei, G. and Tanner, M. (1990). A monte carlo implementation of the em algorithm and the poor man’s data augmentation algorithm. *J. Amer. Statist. Assoc.*, 85:699–704.

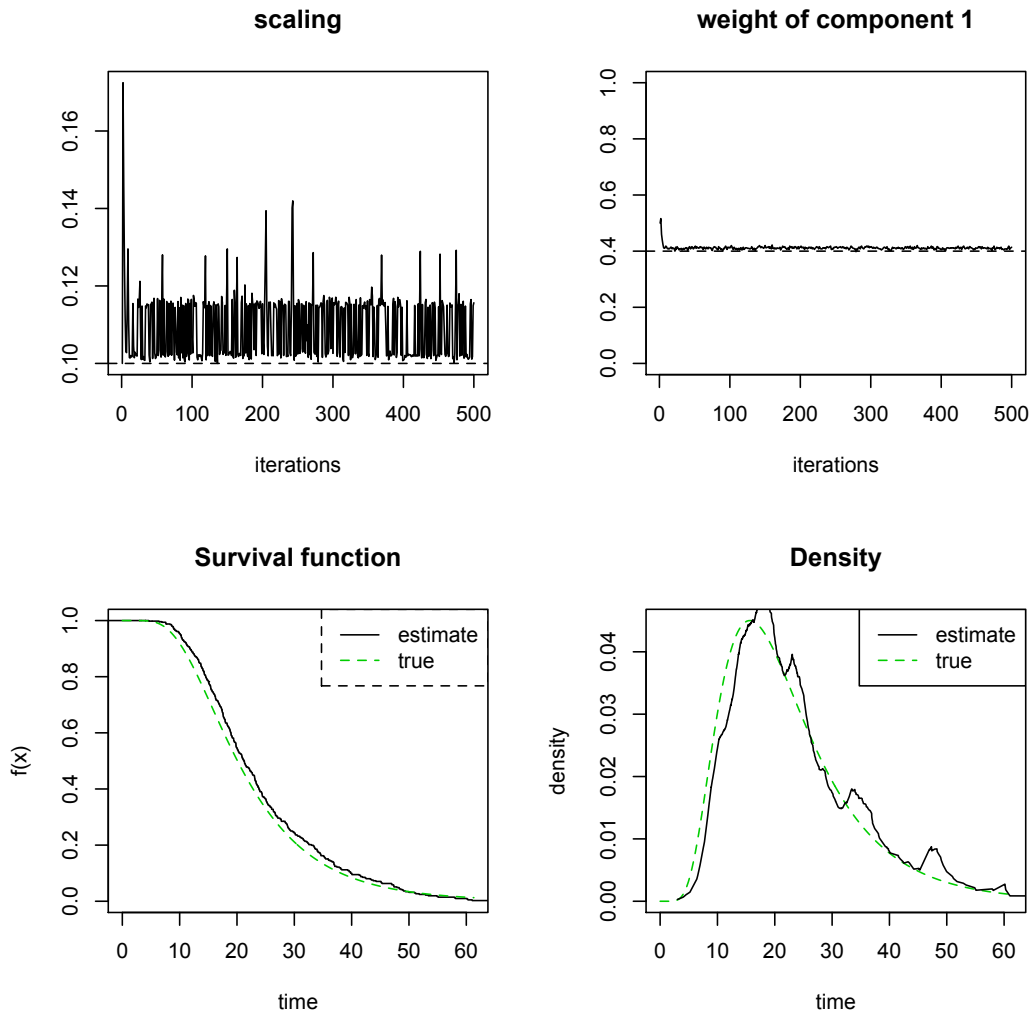


Figure 3: Sample output of the semiparametric St-EM algorithm “final” estimate. The algorithm have been started from the true scale parameter and arbitrary  $\lambda^0 = (1/2, 1/2)$ , for  $n = 500$  censored observations with 10% censored. True values are horizontal lines in top panels (scalar parameters), and dotted lines from  $\mathcal{LN}(3, 0.5)$  in bottom panels (functional parameter).