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

Laurent Bordes, Didier Chauveau

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

Laurent Bordes[1] Didier Chauveau[2]

[1] LMA - UMR CNRS 5142
Université de Pau et des Pays de l'Adour
laurent.bordes@univ-pau.fr
[2]MAPMO - CNRS UMR 7349 - Fédération Denis Poisson
Université d'Orléans
didier.chauveau@univ-orleans.fr

June 6, 2013

Abstract

Mixture models in reliability bring a useful compromise between parametric and nonparametric models, when several failure modes are suspected. The classical methods for estimation in mixture models rarely handle the additional difficulty coming from the fact that lifetime data are often censored, in a deterministic or random way. We present in this paper several iterative methods based on EM and Stochastic EM methodology, that allow us 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 are demonstrated in simulation studies and an actual dataset example.

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

1 Introduction

Statistical analysis of reliability or lifetime data is usually based on parametric or nonparametric models. On the nonparametric side, when the classical parametric distributions (e.g., exponential, Weibull, lognormal) fail to properly fit the data, the nonparametric Kaplan-Meier survival estimate is a standard approach. However on

several occasions both the simple parametric and fully nonparametric models fail to capture a phenomenon quite common in this context, which is the heterogeneity of the underlying population. As an example, consider satellite reliability models as in Castet and Saleh (2010) and Dubos et al. (2010), where two non-observed failure causes, infant mortality and wear out, are suspected. In such situations a (usually parametric) mixture model (see, e.g., McLachlan and Peel, 2000) is appropriate and can significantly improve the quality of the fit. In mixture models, estimates of the parameters of the distribution comprise estimates of the proportions of each (non-observed) subpopulation corresponding to each failure mode, that may be of great interest for the end user. Moreover, when the statistical inference is done by using Expectation-Maximization (EM) algorithms (see Section 1.1), estimates of the individual probabilities that each observation come from each component are provided, allowing for unsupervised clustering of the data. We propose an example of such situation in Section 4.5, where a new semiparametric mixture model provides an alternative to the nonparametric Kaplan-Meier approach, taking into account the two suspected failure modes.

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 Cramèr-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 (MLE's), mostly in the well-known class of EM algorithms (see Section 1.1 below), but few of them handle the additional problem of censoring. Atkinson (1992) derived an EM algorithm for a finite mixture of two univariate normal distributions for deterministically right-censored data. Chauveau (1995) proposed extentions of EM and of Stochastic EM algorithms (introduced by Celeux and Diebolt, 1986) to handle Type-I deterministic right censoring or truncation, for exponential and Weibull mixtures. More recently, Balakrishnan and Mitra (2011) fitted right censored and left truncated data by adapting the EM algorithm to a finite mixture of lognormal distributions; Lee and Scott (2012) proposed EM solutions for deterministically censored or truncated multivariate Gaussian mixtures.

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 (iid) lifetime observations taking values in $\mathbb{R}^+$, from a lifetime probability density function (pdf)

$$\boldsymbol{X} = (X_1, \ldots, X_n) \text{ iid} \sim g(x|\boldsymbol{\theta}),$$

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.

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

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ are the component weights satisfying $\sum_{j=1}^{m} \lambda_j = 1$, and $\boldsymbol{f} = (f_1, \ldots, f_m)$ 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 $\boldsymbol{x} = (x_1, \ldots, x_n)$ iid $\sim g$ are associated to n censoring times $\boldsymbol{c} = (c_1, \ldots, c_n)$ iid $\sim h$. The observations are finally in the censoring case

$$(\boldsymbol{t}, \boldsymbol{d}) = ((t_1, d_1), \ldots, (t_n, d_n)),$$

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

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 $\boldsymbol{g}(\cdot|\boldsymbol{\theta})$, corresponds to the *incomplete* data pdf, associated to the log-likelihood $\ell_{\boldsymbol{x}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log g(x_i|\boldsymbol{\theta})$. When maximizing $\ell_{\boldsymbol{x}}(\boldsymbol{\theta})$ leads to a difficult problem (such as in, e.g., the mixture model situation), considering $\boldsymbol{x}$ as an incomplete data resulting from a non-observed and suitable *complete* data $\boldsymbol{y}$ often helps. Assuming $\boldsymbol{y}$ comes from a complete data pdf $\boldsymbol{g}^c$, the EM algorithm iteratively maximizes the operator

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

the expectation being taken relatively to the conditional distribution of $(\boldsymbol{y}|\boldsymbol{x})$, 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 $(\boldsymbol{\theta}^k, k = 1, 2, \ldots)$:

1. E-step: `compute` $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$

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

In the mixture framework, the complete data associated to $\boldsymbol{x}$ correspond to the situation where the component of origin $z \in \{1,\ldots,m\}$ of each individual data x is observed. The complete data distribution of (X,Z) is given by $\mathbb{P}_{\boldsymbol{\theta}}(Z=z)=\lambda_z$ and $(X|Z=z)\sim f_z$. Computing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$ reduces then to the computation of the essential ingredient of any EM algorithm for finite mixture, the *posterior* probabilities

$$p_{ij}^k = \mathbb{P}_{\boldsymbol{\theta}^k}\left(Z_i=j \mid X_i=x_i\right) = \frac{\lambda_j^k f_j^k(x_i)}{\sum_{j'=1}^m \lambda_{j'}^k f_{j'}^k(x_i)}, \quad i=1,\ldots,n; \quad j=1,\ldots,m. \quad (2)$$

p_{ij}^k is the probability that the ith observation belongs to component j, conditional on the data and the current value of the parameter at iteration k.

In the present setup the observed data $(\boldsymbol{t},\boldsymbol{d})$ depends on $\boldsymbol{x}$ which comes from a finite mixture with pdf g, hence missing data are naturally involved. When censored lifetimes $(\boldsymbol{t},\boldsymbol{d})$ are observed, we may think of two stages for the associated complete data: the simplest one is to consider the component indicators $\boldsymbol{z}=(z_1,\ldots,z_n)$ as missing like in the usual mixture situation, so that $(\boldsymbol{t},\boldsymbol{d},\boldsymbol{z})$ is the complete data. But we can also consider in addition the censored observations $(x_i, i\in\{1,\ldots,n\} : d_i=0)$ as missing (this is the case in Chauveau (1995) for deterministic censoring), so that the complete data is $(\boldsymbol{x},\boldsymbol{z})$. 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. 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).

1.2 Nonparametric estimation under censoring

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

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

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

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

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

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

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 $\hat{f}(t) = \hat{\alpha}(t)\hat{\bar{F}}(t)$. We will use this estimate in the present paper, even though f could also be estimated by smoothing the Kaplan-Meier estimator.

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 $t_{(1)} < \cdots < t_{(n)}$ 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

$$\hat{A}(t) = \sum_{\{i:t_{(i)} \leq t\}} \frac{d_{(i)}}{n-i+1}, \tag{3}$$

$$\hat{\bar{F}}(t) = \prod_{\{i:t_{(i)} \leq t\}} \left(1 - \frac{d_{(i)}}{n-i+1}\right), \tag{4}$$

and

$$\hat{\alpha}(t) = \sum_{i=1}^{n} \frac{1}{b_n} \mathcal{K}\left(\frac{t-t_{(i)}}{b_n}\right) \frac{d_{(i)}}{n-i+1}. \tag{5}$$

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 jth 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

$$X \sim g(x|\boldsymbol{\theta}) = \sum_{j=1}^{m} \lambda_j f(x|\xi_j), \tag{6}$$

where $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\xi}) = (\lambda_1, \ldots, \lambda_m, \xi_1, \ldots, \xi_m)$ is the (Euclidean) model parameter.

Note that in the present model all components share densities from the same parametric family (shape) $\mathcal{F}$. This is the description commonly used in finite mixture

models for ease of notation and simplicity in the maximization (M-step) calculation. However, it is usually straightforward to extend this model to mixtures with parametric densities from different families, i.e. $f_j(\cdot) = f_j(\cdot|\xi_j) \in \mathcal{F}_j$ for $j = 1, \ldots, m$. The only consequence is that each jth M-step for ξ_j then depends on the parametric family $\mathcal{F}_j$. In usual mixture models (usual meaning without censoring), the jth M-step looks like a "weighted MLE" for ξ_j in the parametric family $\mathcal{F}_j$, where the weights are the posterior probabilities given by (2) (see e.g. McLachlan and Krishnan, 2008, for the Gaussian mixture case).

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 $(\boldsymbol{t}, \boldsymbol{d}, \boldsymbol{z})$

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 $f(\cdot|\boldsymbol{\theta})$) is given by

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

and similarly $f^c(t, 0, z|\boldsymbol{\theta}) = \lambda_z \bar{F}(t|\xi_z)h(t)$, where $F_j(\cdot) = F(\cdot|\xi_j)$ denotes the cdf of the jth component. This can be summarized by

$$f^c(t, d, z|\boldsymbol{\theta}) = \left[\lambda_z f(t|\xi_z)\bar{H}(t)\right]^d \left[\lambda_z \bar{F}(t|\xi_z)h(t)\right]^{1-d}. \tag{7}$$

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

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

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

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

where $\boldsymbol{f}^c$ 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) amounts to

compute the *posterior* probability that the ith observation (an observed or censored lifetime) belongs to component j, conditional on the data and the current value of the parameter at iteration k:

$$\begin{aligned} p_{ij}^k &:= \mathbb{P}(Z_i = j | t_i, d_i, \boldsymbol{\theta}^k) \\ &= \left(\frac{\lambda_j^k f(t_i|\xi_j^k)\bar{H}(t_i)}{\sum_{\ell=1}^p \lambda_\ell^k f(t_i|\xi_\ell^k)\bar{H}(t_i)} \right)^{d_i} \left(\frac{\lambda_j^k \bar{F}(t_i|\xi_j^k)h(t_i)}{\sum_{\ell=1}^p \lambda_\ell^k \bar{F}(t_i|\xi_\ell^k)h(t_i)} \right)^{1-d_i} \\ &= \left(\frac{\lambda_j^k f(t_i|\xi_j^k)}{\sum_{\ell=1}^p \lambda_\ell^k f(t_i|\xi_\ell^k)} \right)^{d_i} \left(\frac{\lambda_j^k \bar{F}(t_i|\xi_j^k)}{\sum_{\ell=1}^p \lambda_\ell^k \bar{F}(t_i|\xi_\ell^k)} \right)^{1-d_i} \qquad (8) \\ &= \lambda_j^k \bar{F}(t_i|\xi_j^k) \left(\frac{\alpha(t_i|\xi_j^k)}{\sum_{\ell=1}^p \lambda_\ell^k \alpha(t_i|\xi_\ell^k)\bar{F}(t_i|\xi_\ell^k)} \right)^{d_i} \left(\sum_{\ell=1}^p \lambda_\ell^k \bar{F}(t_i|\xi_\ell^k) \right)^{d_i - 1}, \qquad (9) \end{aligned}$$

where equation (9) is a rewriting of equation (8) 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

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k) &= \sum_{i=1}^n \sum_{j=1}^m p_{ij}^k \left[\log(\lambda_j) + d_i \log f(t_i|\xi_j) + (1-d_i) \log \bar{F}(t_i|\xi_j) \right] \\ &\quad + R(\boldsymbol{t}, \boldsymbol{d}, \boldsymbol{\theta}^k, h), \qquad (10) \end{aligned}$$

where the remaining term $R(\boldsymbol{t}, \boldsymbol{d}, \boldsymbol{\theta}^k, h)$ does not depends on $\boldsymbol{\theta}$ but only on the data, current parameter and censoring distribution. The maximization for $\boldsymbol{\lambda}$ is direct and does not depends 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 $\mathcal{F}$.

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

1. **E-step:** *Calculate the posterior probabilities p_{ij}^k using Equation (8), for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$.*
2. **M-step:** *Set*

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

This algorithm behavior is illustrated in Section 4.

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

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 $\boldsymbol{z}$ and $\{x_i : d_i = 0, 1 \leq i \leq n\}$. The EM methodology aims to iteratively maximize

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

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 δ_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

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

Note that as far as $\boldsymbol{\lambda}$ is concerned, this expression is exactly Equation (10) for the case where $(\boldsymbol{t}, \boldsymbol{d}, \boldsymbol{z})$ 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.

Exemple 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 $\boldsymbol{\theta}^k \to \boldsymbol{\theta}^{k+1}$ for the parametric case and complete data $(\boldsymbol{x}, \boldsymbol{z})$ is given by:*

1. **E-step:** *Calculate the posterior probabilities p_{ij}^k as in Equation (8), for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$.*
2. **M-step:** *Set*

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

Note that the update for ξ_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 ξ_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 expand 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 $\boldsymbol{z}$ 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 $\boldsymbol{z}$ are missing; but due to the censoring process, some of the x_i's are also missing (replaced by the incomplete observations $\{t_i, i = 1, \ldots, n : d_i = 0\}$. 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 $\boldsymbol{z}$, which has been the preferred solution here (i.e. the complete data are $(\boldsymbol{t}, \boldsymbol{d}, \boldsymbol{z})$), 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). Morevover simulation of $\boldsymbol{z}$ alone seems to be up to know the only practical solution for the semiparametric extensions we propose later on. An example of this St-EM approach for censored mixtures of Weibull distributions is given in Section 4.3.

Let $\boldsymbol{p}_i^k = (p_{i1}^k, \ldots, p_{im}^k)$ denote the posterior probability vector associated to observation i, and $Z_i \sim \mathcal{M}ult(1, \boldsymbol{p}_i^k)$ a multinomial distributed random variable with

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

1. **E-step:** Calculate p_{ij}^k from Equation (8), for all $i = 1, \dots, n$ and $j = 1, \dots, m$.
2. **Stochastic step:** Simulate $Z_i^k \sim \mathcal{M}ult(1, \boldsymbol{p}_i^k)$, $i = 1, \dots, n$, and define the subsets
$$\chi_j^k = \{i \in \{1, \dots, n\} : Z_i^k = j\}, \quad j = 1, \dots, m. \tag{11}$$
3. **M-step:** For each component $j \in \{1, \dots, m\}$
$$\lambda_j^{k+1} = \text{Card}(\chi_j^k)/n,$$
and
$$\xi_j^{k+1} = \arg\max_{\xi \in \Xi} L_j(\xi), \tag{12}$$
where
$$L_j(\xi) = \prod_{i \in \chi_j^k} (f(t_i|\xi))^{d_i} (\bar{F}(t_i|\xi))^{1-d_i}. \tag{13}$$

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 $I^{-1}(\boldsymbol{\theta}_0)\left[2I - \{I + F(\boldsymbol{\theta}_0)\}^{-1}\right]$ 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 $I^{-1}(\boldsymbol{\theta}_0)\left[I - \{I + F(\boldsymbol{\theta}_0)\}^{-1}\right]$. 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 $(\boldsymbol{\theta}^k)_{k \geq 0}$ 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)). The Louis method is applied to a mixture of lognormal distributions in Balakrishnan and Mitra (2011).

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:

$$g(x|\lambda, \xi, f) = \lambda f(x) + (1-\lambda)\xi f(\xi x), \quad x > 0. \tag{14}$$

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/ξ 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 semiparametric models like (14) 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 (14) 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 (14) 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 (14) are uncensored then only a n-sample $\boldsymbol{t}$ is observed (note that $\boldsymbol{t}$ is in this case nothing but $\boldsymbol{x}$ since there is no censored data).

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

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

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

Step 2. Simulate $Z_i^k \sim \mathcal{M}ult(1, \boldsymbol{p}_i^k))$ and define subsets

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

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 $\boldsymbol{t}^k = (t_1^k, \ldots, t_n^k)$ be the "unscaled sample" $\{t_i; i \in \chi_1^k\} \cup \{\xi^k t_i; i \in \chi_2^k\}$. Set:

$$f^{k+1}(x) = \sum_{i=1}^{n} \frac{1}{nb} \mathcal{K}\left(\frac{x - t_i^k}{b}\right), \tag{15}$$

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

Remark: Note that at the third step ξ^k is updated using a moment estimation method instead of a ML 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 (14) are randomly censored like in Section 1, so that only a n-sample $(\boldsymbol{t}, \boldsymbol{d})$ 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 (8) or (9), 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 (4), and $\alpha(\cdot)$ by the kernel estimate (5). 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 ith 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 (4). In this case, stochastic versions of the EM algorithm provide the workable solutions that we propose here: simulation of $\boldsymbol{z}$ 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 semiparametric scale mixture and censored data

Step 1. E-step: For each item $i \in \{1, \ldots, n\}$,
if $d_i = 0$

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

else

$$p_{i1}^k = \frac{\lambda^k \alpha^k(t_i)\bar{F}^k(t_i)}{\lambda^k \alpha^k(t_i)\bar{F}^k(t_i) + (1-\lambda^k)\xi^k \alpha^k(\xi^k t_i)\bar{F}^k(\xi^k t_i)}, \tag{17}$$

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

Step 2. Stochastic step: Simulate $Z_i^k \sim \mathcal{M}ult(1, \boldsymbol{p}_i^k)$, $i = 1, \ldots, n$, and define the subsets

$$\chi_j^k = \{i \in \{1, \ldots, n\} : Z_i^k = j\}, \quad j = 1, 2. \tag{18}$$

Step 3. Update the Euclidean parameters:

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

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

$$\xi^{k+1} = \frac{\int_0^{M_1^k} \bar{S}_1^k(s)\, ds}{\int_0^{M_2^k} \bar{S}_2^k(s)\, ds} \tag{19}$$

Step 3'. Update the functional parameters α and $\bar{F}$:
Let $\boldsymbol{t}^k = (t_1^k, \ldots, t_n^k)$ be the order statistic of $\{t_i; i \in \chi_1^k\} \cup \{\xi^k t_i; i \in \chi_2^k\}$, so that $t_1^k \leq \cdots \leq t_n^k$, and $\boldsymbol{d}^k = (d_1^k, \ldots, d_n^k)$ be the corresponding indicators. Set:

$$\alpha^{k+1}(x) = \sum_{i=1}^{n} \frac{1}{b} \mathcal{K}\left(\frac{x - t_i^k}{b}\right) \frac{d_i^k}{n - i + 1}, \tag{20}$$

and

$$\bar{F}^{k+1}(x) = \prod_{\{i: t_i^k \leq x\}} \left(1 - \frac{d_i^k}{n - i + 1}\right), \tag{21}$$

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

Note that here also ξ^k is updated using a moment estimation method instead of a ML 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 semiparametric, genuine and stochastic EM or EM-like algorithms that have been proposed in this paper. All the algorithms shown here are implemented — and 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). Several models are tested on various simulated data, and this section ends with a study on a real dataset from aeronautical industry.

4.1 Initialization of EM-type algorithms

As it is typically the case for EM algorithms, the choice of the starting parameter value $\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}$ is important. A common initialization procedure when experimenting on synthetic data consists in simply starting the algorithms from the true parameter. The argument supporting that approach is that on real data, a usual practice consists in starting the algorithm from several values either taken from a grid, or randomly drawn from a uniform distribution on the parameter space, and retaining the EM estimate achieving the maximum of the observed likelihood among all the trials. If this *exhaustive exploration of* $\boldsymbol{\Theta}$ if done with enough precision (number of trials), then this estimate corresponds to the location of the global maximum closest to the true parameter value. We first tried initialization from the true values in all our experiments on synthetic data, but choose instead to present here only practical, data-driven or realistic initialization procedures. For instance, we applied the initialization by this exploration of $\boldsymbol{\Theta}$ in Section 4.3 where it is detailed.

The disadvantage of this approach is its high computing requirement, so that for the particular case of univariate data where $m = 2$ components are suspected, and assuming that these components have an effect at least in localization (as it is often the case for lifetime data $\boldsymbol{t}$), a faster approach, that we call here *splitting the data* amounts to:

1. Split the data at a threshold s into two subsamples $\{t_i : t_i \leq s\}$ and $\{t_i : t_i > s\}$. Ideally, s corresponds to a visible separation between two modes from the histogram of $\boldsymbol{t}$, or a threshold between two types of lifetimes obtained by prior information on the data.

2. Fit an appropriate model and compute a MLE from each subsample: in parametric mixture case, the ideal model is simply the parametric family $\mathcal{F}_j$ and the MLE is the corresponding $\hat{\boldsymbol{\xi}}_j$. Set $\lambda_1^0 = \#\{t_i : t_i \leq s\}/n$, to obtain an initial parameter $\boldsymbol{\theta}^0 = \boldsymbol{\theta}^0(s)$.

3. Run the St-EM algorithm from that $\boldsymbol{\theta}^0(s)$ to get an estimate $\hat{\boldsymbol{\theta}}(s)$.

4. If there is no obvious unique choice for s, apply steps 1-2-3 above for a set of plausible $s \in \mathcal{S}$ (typically a grid within the observations range) and retain the estimate achieving the largest log-likelihood: $\hat{\boldsymbol{\theta}} = \text{argmax}_{s\in\mathcal{S}}\, \ell(\hat{\boldsymbol{\theta}}(s))$.

When appropriate, this one-dimensional approach is obviously simpler than exploring the multidimensional parameter space $\boldsymbol{\Theta}$. It can be viewed also as a crude way of completing the data by setting the missing data to $z_i = j$ for i's in subsample j. We applied initialization based on splitting the data in Sections 4.4 and 4.5.

Finally, when performing Monte-Carlo experiments, care should be taken to prevent or handle a possible "label-switching" as much as possible. This label-switching issue arises because the particular ordering of the subscripts $j = 1, \ldots, m$ 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.

4.2 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 ($(\boldsymbol{t}, \boldsymbol{d}, \boldsymbol{z})$ and $(\boldsymbol{x}, \boldsymbol{z})$ (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%. Even if this true model have rather different rates, the component densities are severely overlapping because of the shape of the exponential distribution.

We applied on this model the two EM strategies introduced in Sections 2.1 and 2.2, i.e. for the two possible levels of complete data, over 300 Monte-Carlo replications and sample size $n = 500$. For the initialization, we define a very crude data-driven procedure to compute initial rates $\boldsymbol{\xi}$ using a "uniform binning" of the data (even simpler than the splitting of the data presented in Section 4.1). We split $\boldsymbol{t}$ in $m = 3$ bins of equal empirical probabilities 1/3, and from each bin j we compute ξ_j^0 as the inverse of the average of the data within the bins. This implies that the ξ_j^0's are decreasing like the true $\boldsymbol{\xi}$. Note that this is not a constraint since a mixture is define up to a permutation of the labels (label-switching). For the weights,

the initialization has been completely non informative by simply set $\lambda_j = 1/m$ for all $j = 1, \ldots, m$.

Results of this Monte-Carlo experiment are given in Fig. 1, which shows the good behavior of these algorithms and no clear winner between both strategies, even for this rather coarse initialization procedure. In this experiment no label switching occured between the rate parameters. In less than 1% of the replications, λ_3 have been wrongly estimated to very small weight. A more careful examination of these cases show that the corresponding rates ξ_3 are then also very small ($< 10^{-3}$), so that these cases correspond to a degenerate estimation of the third component.

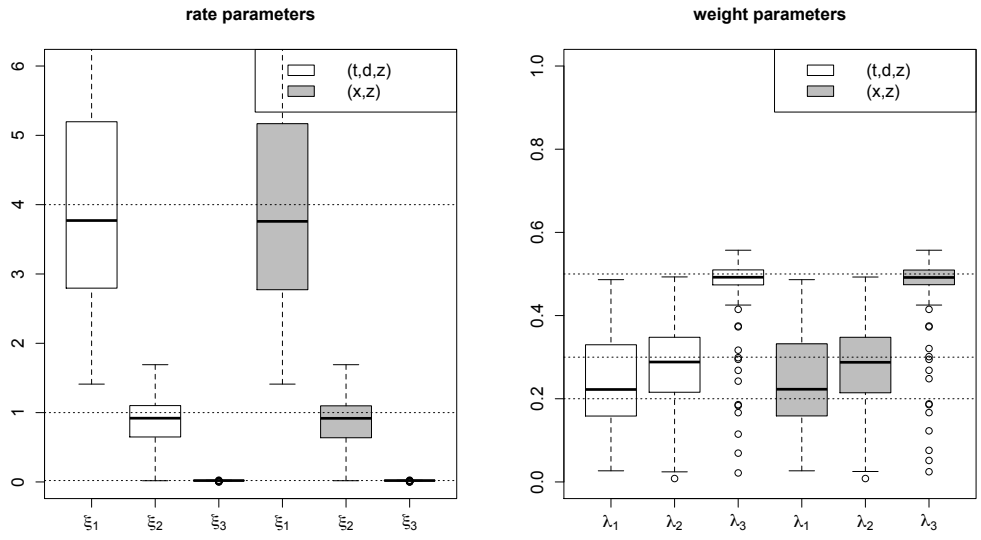

Figure 1: *Boxplots of estimates for 300 replications of two EM algorithms started from a* $\boldsymbol{\theta}^0$ *given by the "uniform binning" described above, for* $n = 500$ *sample size and average censoring rate of 16%. EM algorithms using complete data* $(\boldsymbol{t}, \boldsymbol{d}, \boldsymbol{z})$ *are in white, and complete data* $(\boldsymbol{x}, \boldsymbol{z})$ *in grey. Horizontal dotted lines are true values.*

4.3 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

$$\bar{G}(x) = \sum_{j=1}^{m} \lambda_j \exp\left[-\left(\frac{x}{\eta_j} \right)^{\beta_j} \right] \tag{22}$$

with shape parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$ and scale parameters $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)$. The Weibull distribution is commonly used in reliability analysis because it can captures 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 the real data are not publicly available, with similar but rounded parameters (see the true values in Table 1). 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 (11), (12) and (13), 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 $n = 1400$, to which we applied a somehow strong random, exponentially-distributed censoring process, achieving in average 31% of censored observations.

The initialization of the St-EM algorithm has revealed more tricky for this model. We first tried the *splitting the data* approach explained in Section 4.1, computing the MLE from each subsample using survreg() and optimizing the cutting point based on the log-likelihood. This approach reveals itself non appropriate here, for a reason due to the model for the satellite data: component 1 called "infant mortality" by Dubos et al. (2010) since $\beta_1 < 1$ is actually associated to a long life duration of several thousands years, whereas component 2 called "wearout" ($\beta_2 > 1$) is associated to a short life of about 9 years. Hence the region of interest of component 2 is imbedded in the region of interest of component 1, so that, as suggested by an histogram, no splitting can lead to reasonable MLE's and starting parameters. We have thus implemented the more demanding initialization by exhaustive exploration of the 5-dimensional parameter space $\boldsymbol{\Theta}$. This approach is not data-driven but we added some "prior information" using what is reasonably expected for these satellite data: assuming that a good starting parameter could be in the subset

$$\tilde{\Theta} = \{\lambda_1 > 0.6, \quad \beta_1 < 1 < \beta_2 < 10, \quad 1000 < \eta_1 < 10,000, \quad \eta_2 < 50\},$$

we applied the following algorithm for each replication:

1. simulate B starting points $\boldsymbol{\theta}^0(1), \ldots, \boldsymbol{\theta}^0(B)$ iid$\sim \mathcal{U}_{\tilde{\Theta}}$;

2. run B St-EM algorithms started from the $\boldsymbol{\theta}^0(u)$'s, for 100 iterations, giving St-EM estimates $\hat{\boldsymbol{\theta}}(u)$, $u = 1, \ldots, B$;
3. set $\tilde{\boldsymbol{\theta}} = \text{argmax}_{1 \leq u \leq B} \ell_{\boldsymbol{t},\boldsymbol{d}}(\hat{\boldsymbol{\theta}}(u))$ the estimate achieving the best observed log-likelihood;
4. run 1000 iterations of the St-EM algorithm started from $\tilde{\boldsymbol{\theta}}$ to get a final estimate.

Since this approach involves repetitive tasks for each replication, the code has been written taking advantage of recent advances in High Performance Computing in R using the Rmpi package (Yu, 2012). Our code run St-EM's on multicore computers or actual clusters, and we ran it on the regional cluster CCSC[1].

Each St-EM algorithm was ran using the "averaged" strategy where missing data at iteration k are simulated from a conditional distribution with parameter based on 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	
	$j=1$	$j=2$	$j=1$	$j=2$	$j=1$	$j=2$
$\boldsymbol{\lambda}$	0.95	0.05	0.9504	0.0496	0.0073	0.0073
$\boldsymbol{\beta}$	0.45	7.00	0.452	7.461	0.0168	1.3814
$\boldsymbol{\eta}$	4100	9	4190.1	9.004	503.298	0.228

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 31% are censored in average. Initialization done as described above with $B = 480$ starting points iid$\sim \mathcal{U}_{\tilde{\Theta}}$.

These results show the good behavior of this St-EM strategy in this situation where one component is characterized by a rather small wheight ($\lambda_2 = 5\%$), compensated by a large dataset ($n = 1400$), and even with 31% of censored data. The somehow large standard deviation for η_1 is due to about 1% (3 cases out of 300) of too large estimates $\hat{\eta}_3 > 6000$ years.

4.4 Stochastic EM for semiparametric model with censored data

We consider the mixture of accelerated lifetime model of Equation (14), in which the unknown parameter is $\boldsymbol{\theta} = (\lambda, \xi, f)$ 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

[1] *Centre de Calcul Scientifique en région Centre*, `http://cascimodot.fdpoisson.fr/?q=ccsc`

this case the M-step for the functional parameter uses a kernel density estimate on the unscaled sample (15) 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 (much harder) 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 initialization of its first E-step (Equations (16) and (17)) initial values for both the Euclidean parameters (λ^0, ξ^0) and the nonparametric survival and hazard rate functions $\bar{F}^0$ and α^0, evaluated at the t_i's and the $\xi^0 t_i$'s. These require an initial "unscaled" sample $\boldsymbol{t}^0$ which in turn requires initial component indicators $(Z_1^0, \ldots, Z_n^0)$. Since a typical histogram of data issued from this model reveals two bumps, we first applied a splitting of the data as in Section 4.1, with a cutting point set at $s = 60$ from the typical histogram, from which we compute an initial proportion λ^0 and, taking into account only the t_i's associated to $d_i = 1$, mean times for each subsample and their ratio that gives an initial scaling ξ^0. The means of each subsample have been also used as initial centers of a k-means clustering algorithm on $\boldsymbol{t}$ from which the Z_i^0's are assigned, and initial $\boldsymbol{t}^0$, $\bar{F}^0$ and α^0 can be built using an initial Step 3', Equations (20) and (21). Actually, these computations are done inside the function running the semiparametric St-EM algorithm in the mixtools package.

We then ran the semiparametric St-EM of Section 3.2 for $K = 300$ iterations, where the bandwidth b involved in the nonparametric kernel estimate of the hazard rate $\alpha(\cdot)$ has been set by calling the R default function `bw.nrd0` (Silverman's "rule of thumb"), applied to the first subsample only, since most of the data from this subsample are f-distributed, the other subsample, mostly issued from the scaled f, would have lead to an over-estimation of b.

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,

$$\hat{\lambda} = \frac{1}{K}\sum_{k=1}^{K} \lambda^k; \quad \hat{\xi} = \frac{1}{K}\sum_{k=1}^{K} \xi^k. \tag{23}$$

For the nonparametric $\bar{F}$ and α, 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 $\boldsymbol{t}^K$. Then, since we are in the censored data situation, the density f itself is estimated by plug-in $f^K(\cdot) = \bar{F}^K(\cdot)\alpha^K(\cdot)$. The other approach we tested tries to mimic the average strategy used for the scalar parameters, and is for that denoted "average" later. It amounts to use $\hat{\xi}$ from Equation (23) instead of ξ^K to compute an "averaged" unscaled ordered sample $\hat{\boldsymbol{t}}$ from which $\hat{\bar{F}}(\cdot)$ and $\hat{\alpha}(\cdot)$ are estimated using an additional Step 3' (as in (20) and (21)). 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 initialized by the above procedure. 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, standard deviations and mse's 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:

$$\text{MISE} = \frac{1}{R}\sum_{r=1}^{R}\int_0^\infty \left(\hat{f}^{(r)}(u) - f(u)\right)^2 du \approx 0.00079$$

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 rth 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 λ is rather good. The estimation of ξ is also good, even though it is slightly biased. Figure 2 shows a typical result on a simulated sample of size $n = 500$ from this model. Note that this plot is the default 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). Fig. 2 illustrates the good estimation of the scalar parameters (despite the tendency to over estimate ξ), and of the survival function through Kaplan-Meier estimates on the unscaled samples $\boldsymbol{t}^K$ or $\hat{\boldsymbol{t}}$. The typical estimate of the density f is, not surprisingly, not so precise since it includes a kernel density estimate of the hazard rate (20), 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
λ	0.4	0.398	0.0207	0.00043
ξ	0.1	0.112	0.00793	0.0002

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

4.5 An example on actual data

We study in this section an actual dataset consisting of $n = 2057$ observations $(\boldsymbol{t}, \boldsymbol{d})$ among which 26% are censored. These lifetimes correspond to times of wearout of

mechanical parts collected by the Turbomeca (TM) company[2]. These mechanical parts (inspected on a regular basis) are used in two "configurations", and depending on that are replaced typically after 3000 or 6000 hours. The configurations are not observed here, so that it is reasonable to assume that a two-components mixture model corresponding to two different average lifetimes is appropriate, even though the histogram of $\boldsymbol{t}$, depicted in Fig. 3, does not show by itself a clear bimodal picture.

Parametric models. We first choose to fit parametric two-components mixtures with Stochastic EM algorithms from Section 2.3, for two parametric families for which the MLE on single sample of censored data is implemented in the survreg() function of the survival package, so that the M-step (equations(12) and (13)) does not require much additional coding. We applied mixtures of Weibull distributions as in Section 4.3 model (22), and mixtures of lognormal distributions.

For actual data the initialization of EM or St-EM algorithms is a crucial question, since poor starting points can lead to stabilization near unmeaningful local maxima of the likelihood. This remains an issue even for stochastic versions because of the averaging procedure detailed in Section 2.3. We choose first to apply the splitting procedure for univariate data described in Section 4.1, where for each subsample a parametric (Weibull or lognormal) model has been fit using the survival package, to obtain an initial $\boldsymbol{\theta}^0$ after log-likelihood optimization in the cutting point s since the histogram of the TM data (Fig. 3) does not exhibit a bimodal shape. We applied the splitting procedure for a sequence of $s \in \mathcal{S} = \{1000, 1200, \ldots, 5000\}$. Surprisingly, the estimates were heavily dependent from the starting positions (even after about $K = 10,000$ St-EM iterations), and the largest values of the log-likelihood were obtained for degenerate models corresponding more or less to a single Weibull or lognormal distribution. This suggest that these parametric families may not be appropriate for the TM data. To confirm this, we even ran the same procedure on simulated but similar data from true Weibull mixtures. The procedure above with optimization of $s \mapsto \ell(\hat{\boldsymbol{\theta}}(s))$ resulted in good St-EM estimates and less dependence from the initial $\boldsymbol{\theta}^0(s)$'s. These results are not detailed here for brevity. However, from the histogram (Fig. 3) and the prior knowledge we have for the TM data, a binning procedure with $s \approx 3000$ seemed to be a reasonable choice, so we illustrate the approach with that initialization here. The estimated Weibull and lognormal components from each mixture models are provided in Fig. 3, and the corresponding mixture survival functions are plotted for comparison in Fig. 4.

Semiparametric model. The previous study argue for trying to fit the semiparametric mixture model (14) to these data. We thus splitted the data in two bins at $s = 3000$ as previously, and computed mean times and their ratio that gives an initial scaling $\xi^0 \approx 0.4$ (the censoring/observed indicator was not taken into account

[2]The authors thank the Turbomeca Company `http://www.turbomeca.com` that allowed us to use these data.

here). We then ran the semiparametric St-EM of Section 3.2 for $K = 500$ iterations, with several settings for the bandwidth b involved in the nonparametric kernel estimate of the hazard rate $\alpha(\cdot)$. The R standards settings (`bw.nrd0`, biaised cross validation, ...) return bandwidths between $b = 223$ and $b = 287$, that result in slightly too jagged density estimate, so that we finally used $b = 300$. This gave estimates of the Euclidean parameters $\hat{\lambda} = 0.421$ and $\hat{\xi} = 0.643$. Note that all the bandwidth trials gave approximately similar estimates. The fitting of the mixture survival function,

$$\hat{\lambda}\hat{S}(t) + (1 - \hat{\lambda})\hat{S}(\hat{\xi}t),$$

where $\hat{S}$ is the "average" estimate of the survival function (i.e. computed with an additional M-step based on the average of the posterior probabilities over iterations, as explained before) is compared to the plain Kaplan-Meier estimate of $(\boldsymbol{t}, \boldsymbol{d})$, and to the Weibull and lognormal mixtures fit in Fig.4. The slightly better fitting of the semiparametric mixture model (particularly near the decay at about 3000 hours), and the estimated scaling $\hat{\xi}$ in accordance with the prior information that the ratio of lifetimes between the two possible subpopulations could be approximately 1/2 show that, if there is a good reason to account for the existence of two sub-populations, then this model is a preferable solution. This illustrates the better flexibility provided by the nonparametric assumption.

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 provides practical solutions taking advantage of the simulated complete data. Several of these algorithms have been compared on a case study for actual 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 often 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 computational work in this paper has been done using an upcoming next 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.

Atkinson, S. E. (1992). The performance of standard and hybrid EM algorithms for ML estimates of the normal mixture model with censoring. *Journal of Statistical Computation and Simulation*, 44(1-2):105–115.

Balakrishnan, N. and Mitra, D. (2011). Likelihood inference for lognormal data with left truncation and right censoring with illustration. *Journal of Statistical Planning and Inference*, 144(11):3536–3553.

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. *Annals of Statistics*, 35(1):224–251.

Karunamuni, R. and Wu, J. (2009). Minimum hellinger distance estimation in a nonparametric mixture model. *Journal of Statistical Planning and Inference*, 3:1118–1133.

Lee, G. and Scott, C. (2012). EM algorithms for multivariate gaussian mixture models with truncated and censored data. *Computational Statistics and Data Analysis*, 56:2816–2829.

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*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

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.

Yu, H. (2012). *Rmpi: Interface (Wrapper) to MPI (Message-Passing Interface).*

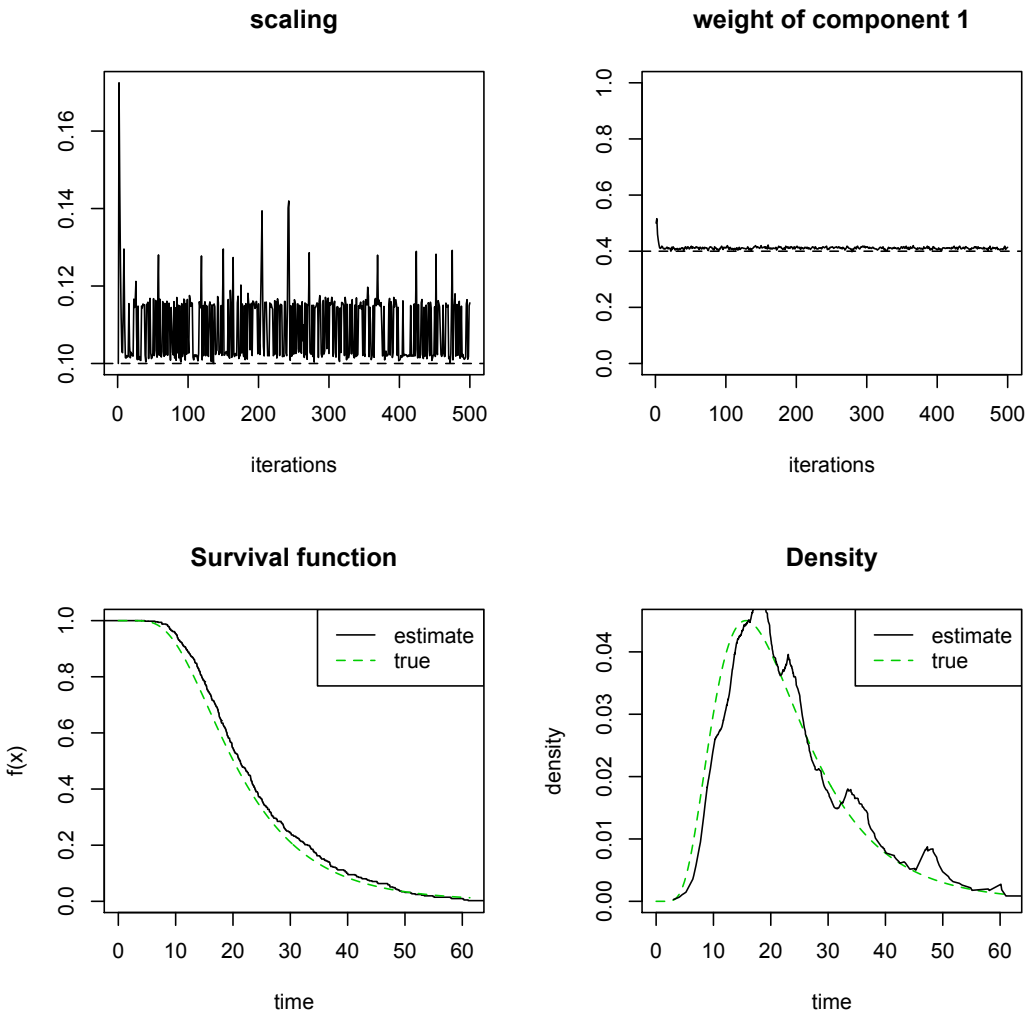

Figure 2: *Sample output of the semiparametric St-EM algorithm "final" estimate, for* $n = 500$ *censored observations with 10% censored. True values are horizontal dotted lines in top panels (scalar parameters), and dotted lines from* $\mathcal{LN}(3,\, 0.5)$ *in bottom panels (functional parameter).*

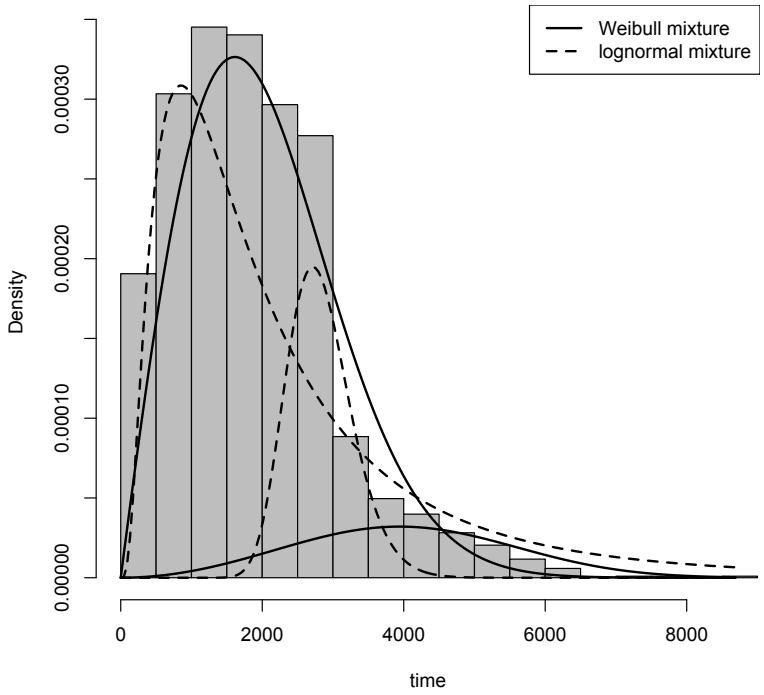

Figure 3: *The Turbomeca (TM) data together with the components* $\hat{\lambda}_j \hat{f}_j$*'s from St-EM fits for* $m = 2$ *components Weibull mixture (solid lines), and lognormal mixture (dashed lines); initialization based on a binning with* $s = 3000$.

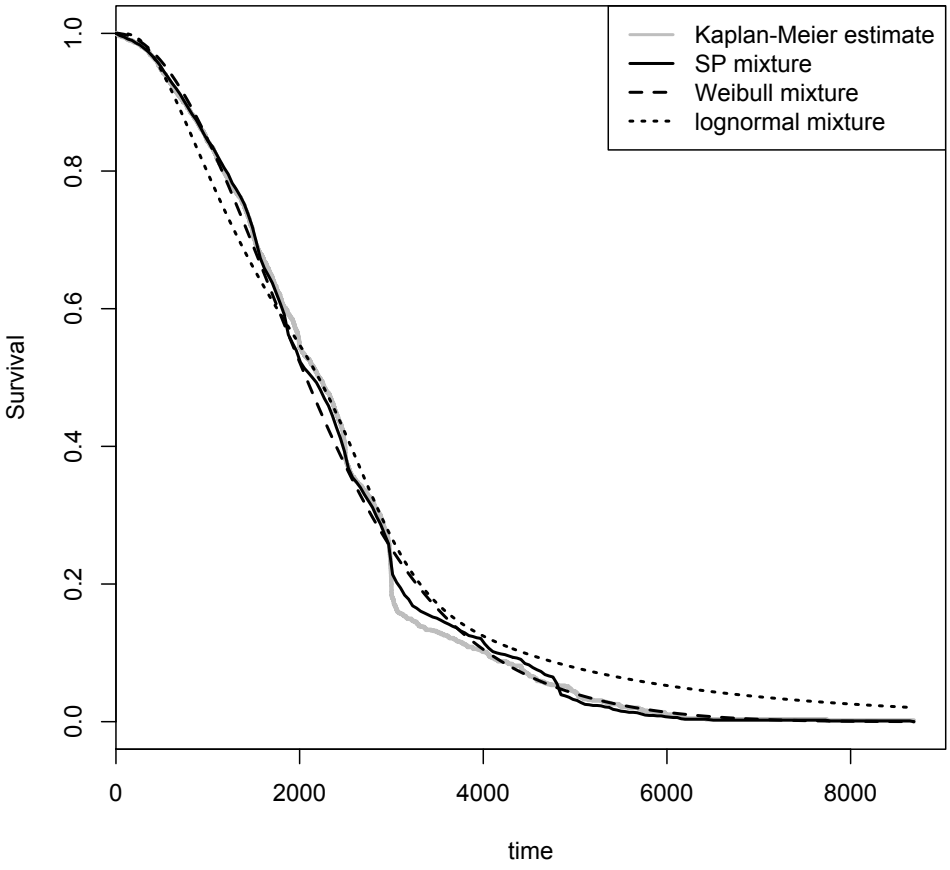

Figure 4: *Kaplan-Meier, Weibull mixture, lognormal mixture (from parametric St-EM algorithms) and semiparametric scaling mixture (from semiparametric St-EM algorithm) survival estimates for the TM data.*