Stochastic EM-like Algorithms for Fitting Finite Mixture of Lifetime Regression Models Under Right Censoring
Laurent Bordes, Didier Chauveau

To cite this version:

HAL Id: hal-01478523
https://hal.science/hal-01478523
Submitted on 12 Mar 2017

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.
Stochastic EM-like Algorithms for Fitting Finite Mixture of Lifetime Regression Models Under Right Censoring

Laurent Bordes∗ Didier Chauveau†

Abstract
Finite mixture of models based on the proportional hazards or the accelerated failure time assumption lead to a large variety of lifetime regression models. We present several iterative methods based on EM and Stochastic EM methodologies, that allow fitting parametric or semiparametric mixture of lifetime regression models for randomly right censored lifetime data including covariates. Their identifiability is briefly discussed and in the semiparametric case we show that simulating the missing data coming from the mixture allows to use the ordinary partial likelihood inference method in an EM algorithm’s M-step. The effectiveness of the new proposed algorithms is illustrated through simulation studies.

Key Words: Right censoring, EM algorithm, proportional hazards model, semiparametric mixture models

1. Introduction
In survival analysis it is frequent that the duration of interest is observed with covariates influencing its probability distribution. The semiparametric proportional hazards model (PHM) is probably the most famous lifetime regression model since Cox (1972) introduced the partial likelihood function that allows estimating the Euclidean regression parameter, considering that the baseline hazard rate function is a nuisance parameter. When the duration of interest depends on several explanatory variables and that quantitative or ordinal explanatory variables are missing, then the associated survival function is simply a finite mixture of survival functions potentially dependent of the observed covariates. In the parametric case there is a huge number of papers dealing with inference methods for finite mixture models taking into account the fact that often the lifetime is incompletely observed due to censoring or truncation. See e.g. Chauveau (1995), Beutner and Bordes (2011), Balkrishnan and Mitra (2011, 2014), Bordes and Chauveau (2014) for contributions. However very few papers deal with semiparametric finite mixture of lifetime models. Recently, Bordes and Chauveau (2016) proposed to fit a semiparametric two-component mixture model under right censoring using a stochastic EM-like algorithm. Nevertheless it is worth noting that there are very special kinds of two-component semiparametric mixture models that are common in lifetime data analysis, that is the mixture of a nonparametric lifetime model and a mass at 0 (zero-inflated model) or at infinity (cure model). The later model has motivated important developments with and without explanatory variables (see for instance Yin and Ibrahim, 2005).

During several decades, mixture models have considerably expanded from both theoretical and applied point of view (see for example McLachlan and Peel, 2000) as well as specific estimation methods, especially those based on EM algorithm (see McLachlan and Krishnan, 2008) or their stochastic versions (see e.g., Celeux and Diebolt, 1986; Celeux et al., 1996) for which there are few theoretical results (see Nielsen, 2000). Some of the estimation methods which have

∗Univ. Pau & Pays de l’Adour, CNRS, IPRA-LMAP, UMR 5142, Pau, France; laurent.bordes@univ-pau.fr
†Univ. Orléans, CNRS, MAPMO, UMR 7349, Orléans, France; didier.chauveau@univ-orleans.fr
been developed to fit the proportional hazards model with missing covariates are also close
to estimation methods required to fit mixture models (see for instance Chen and Little,
1999).

In this paper first we briefly introduce the general framework on finite mixture of life-
time regression models, right censored data and the semiparametric estimation method for
the PHM. Then, in Section 2, we introduce several classes of parametric and semipara-
tmetric finite mixture models based on the PHM. Section 3 is devoted to a genuine EM
algorithm in the parametric setup while Section 4 deals with an adaptation of the stochastic
EM-like algorithm for semiparametric models. Several numerical illustrations are given in
Section 5 and a discussion ends the paper in Section 6.

1.1 Data and Cox regression under right censoring

Let \((X_1, Z_1), \ldots, (X_n, Z_n)\) be \(n\) i.i.d. copies of \((X, Z) \in [0, +\infty) \times \mathbb{R}^p\) where
the conditional pdf of the lifetime \(X\) given \(Z = z\) is \(g(x|z, \theta)\). We assume that these lifetimes
data come from a finite mixture of \(m\) components

\[
g(x|z, \theta) = \sum_{j=1}^{m} \alpha_j f_j(x|z),
\]

where \(\theta = (\alpha, f)\) with \(\alpha = (\alpha_1, \ldots, \alpha_m) \in [0, 1]^m\) and \(\sum_{j=1}^{m} \alpha_j = 1\) are the component
weights and \(f = (f_1, \ldots, f_m)\) are the component conditional pdf. Considering the cdf we
can write

\[
G(x|z, \theta) = \sum_{j=1}^{m} \alpha_j F_j(x|z),
\]

or

\[
\bar{G}(x|z, \theta) = \sum_{j=1}^{m} \alpha_j \bar{F}_j(x|z),
\]

where \(\bar{G} = 1 - G\) and \(\bar{F}_j = 1 - F_j\) are conditional survival functions. In addition we
assume that lifetimes \(X_i\) are possibly right censored by censoring times \(C_i\) such that instead
of observing \(X_i\) we observe

\[T_i = X_i \wedge C_i\quad \text{and} \quad D_i = \mathbb{I}(X_i \leq C_i)\]

for \(i \in \{1, \ldots, n\}\). We assume that \(\{(X_1, C_1, Z_1), \ldots, (X_n, C_n, Z_n)\}\) are i.i.d. copies of
\((X, C, Z)\), that conditionally on \(Z = z\), \(X\) and \(C\) are independent, the conditional pdf of
\(C\) is written \(h(c|z)\), \(H(c|z)\) and \(\bar{H}(c|z)\) are the corresponding conditional cdf and survival
functions. We write \(v(z)\) the pdf of \(Z\) on \(\mathbb{R}^p\).

Finally we observe \(\{t, d, z\} = \{(t_1, d_1, z_1), \ldots, (t_n, d_n, z_n)\}\) where \(t_i = x_i \wedge c_i\) and
d_i = \(\mathbb{I}(x_i \leq c_i)\) for \(1 \leq i \leq n\). The observed covariates are not time-dependent here but
considering time-dependent covariates should be possible.

We assume that for all \(z\), \(G(\cdot|z)\) and \(\bar{H}(\cdot|z)\) are absolutely continuous with respect
to the Lebesgue measure, thus with probability one we have \(T_i \neq T_j\) for all \(i \neq j\). Let
\((i_1, \ldots, i_n)\) be the permutation of \((1, \ldots, n)\) such that \(t_{i_1} < t_{i_2} < \cdots < t_{i_n}\). For simplicity,
from now on we rewrite \((t_k, d_k, z_k) = ((t_{i_k}, d_{i_k}, z_{i_k}))\) for \(1 \leq k \leq n\).

Let us recall that a duration \(X\) follows a proportional hazards rate model if conditionally
on \(Z = z\) its hazard rate function (or risk function) is defined by

\[
\lambda_{X|Z}(x|z) = e^{\beta^T z} \lambda_0(x),
\]
where $\beta \in \mathbb{R}^p$ is an unknown regression parameter and $\lambda_0$ is an unknown baseline hazard rate function. By the Cox partial likelihood principle $\beta$ can be estimated by

$$\hat{\beta} = \arg \max_{\beta \in \mathbb{R}^p} L_n(\beta)$$

where

$$L_n(\beta) = \prod_{i=1}^n \left( \frac{e^{\beta^T z_i}}{\sum_{j \geq i} e^{\beta^T z_j}} \right)^{d_i}.$$

The cumulative hazard rate function $\Lambda_0(x) = \int_0^x \lambda_0(s) ds$ is estimated by

$$\hat{\Lambda}_0(x) = \sum_{i: t_i \leq x} \frac{d_i}{\sum_{j \geq i} e^{\beta^T z_j}},$$

the conditional survival function $S_{X|Z}(x|z)$ is estimated by

$$\hat{S}_{X|Z}(x|z) = \exp \left( -e^{\hat{\beta}^T z} \hat{\Lambda}_0(x) \right).$$

In addition, if $K$ is a kernel function and $b = b_n$ a bandwidth such that $(b_n)_{n \geq 1} \not\to 0$ and $(nb_n)_{n \geq 1} \not\to +\infty$, then $\lambda_0(x)$ is estimated by

$$\hat{\lambda}_0(x) = \frac{1}{b} \sum_{i=1}^n K \left( \frac{x - t_i}{b} \right) \frac{d_i}{\sum_{j \geq i} e^{\beta^T z_j}}.$$

Note that:

1. Maximizing $L_n(\beta)$ with respect to $\beta$ is generally done using differential optimization method since $\beta \rightarrow L_n(\beta)$ belongs to $C^\infty(\mathbb{R}^p)$ and is convex.
2. $S_{X|Z}(x|z)$ can also be estimated using a product-limit type estimator.
3. The package `survival` (Therneau and Lumley, 2009) for the R statistical software (R Core Team, 2013) gives all the previous quantities except $\hat{\lambda}_0$.

### 2. Some finite mixtures of the proportional hazards model

We describe in this section four possible models, denoted M1–M4, for which each component in (3) follows a semiparametric Proportional Hazards Model (PHM).

**M1: Common covariate effect with dependent baseline risk functions.**

For $1 \leq j \leq m$ we have $\tilde{F}_j(x|z, \theta) = \{S_0(x)\}^{\exp(\beta^T z + \gamma_j)}$, then $\theta = (S_0(\cdot), \alpha, \beta, \gamma)$ where $\gamma = (\gamma_2, \ldots, \gamma_m)$ ($\gamma_1 = 0$ for identifiability reasons), hence $\theta \in \mathcal{S} \times \mathbb{R}^{2m + p - 2}$ where $\mathcal{S}$ denotes the set of survival functions.

**M2: Common baseline risk function with independent covariate effects.**

For $1 \leq j \leq m$ we have $\tilde{F}_j(x|z, \theta) = \{S_0(x)\}^{\exp(\beta^T z)}$ then $\theta = (S_0(\cdot), \alpha, \beta)$ where $\beta = (\beta_1, \ldots, \beta_m)$, hence $\theta \in \mathcal{S} \times \mathbb{R}^{m(p+1)-1}$.

**M3: Common covariate effect with independent baselines (NP).**

For $1 \leq j \leq m$ we have $\tilde{F}_j(x|z, \theta) = \{S_{0j}(x)\}^{\exp(\beta^T z)}$, then $\theta = (S_{01}(\cdot), \ldots, S_{0m}(\cdot), \alpha, \beta)$, hence $\theta \in \mathcal{S}^m \times \mathbb{R}^{m(p+1)}$.

**M4: Independent covariate effects and baselines.**

For $1 \leq j \leq m$ we have $\tilde{F}_j(x|z, \theta) = \{S_{0j}(x)\}^{\exp(\beta^T z)}$, then $\theta = (S_{01}(\cdot), \ldots, S_{0m}(\cdot), \alpha, \beta)$ where $\beta = (\beta_1, \ldots, \beta_m)$, hence $\theta \in \mathcal{S}^m \times \mathbb{R}^{m(p+1)-1}$.

Note that we have some hierarchy for these models: Model 1 $\subset$ Model 3 $\subset$ Model 4, and Model 2 $\subset$ Model 4.
3. Genuine EM–algorithm in the parametric set-up

In the parametric situation the complete data pdf \( f^c \) is defined by

\[
f^c_{T,D,Z,J}(t,d,z,j|\theta) = \alpha_j \left[ f(t|\gamma_j, \beta, z)H(t|z) \right]^d \left[ \tilde{F}(t|\gamma_j, \beta, z)h(t|z) \right]^{1-d} v(z)
\]

where \( v \) does not depend on \( \theta = (\alpha, \gamma, \beta) \) and \( J \sim \text{Mult}(1, \alpha) \) is the missing data, independent of \((T, D, Z)\). In the sequel we write \( f^c \) for \( f^c_{T,D,Z,J} \). The complete data likelihood function \( \ell^c \) is defined by

\[
\ell_{t,d,z,j}(\theta) = \log \left( \prod_{i=1}^{n} f^c(t_i, d_i, z_i, j_i|\theta) \right)
= \sum_{i=1}^{n} \log \left( (\tilde{H}(t_i|z_i))^d_i(h(t_i|z_i))^{1-d_i} v(z_i) \right)
+ \sum_{i=1}^{n} \log \left( \alpha_j \left( f(t_i|\gamma_j, \beta, z_i) \right)^{d_i} \tilde{F}(t_i|\gamma_j, \beta, z_i)^{1-d_i} \right)
\]

where in the right hand side of the last equality the first term does not depend on \( \theta \) and \( j = (j_1, \ldots, j_n) \) is the unobserved realization of \((J_1, \ldots, J_n)\).

The EM genuine algorithm consists in providing iterates \((\theta^k)_{k\geq 0}\) by iteratively maximizing \(Q(\theta|\theta^k)\) where

\[
Q(\theta|\theta^k) = \sum_{i=1}^{n} \mathbb{E} \left[ \log(f^c(T_i, D_i, Z_i, J_i|\theta)|t_i, d_i, z_i, \theta^k) \right].
\]

Calculating the above conditional expectation requires to calculate the posterior probabilities

\[
\alpha_{ij}^k = \Pr(J_i = j|t_i, d_i, z_i, \theta^k) = \frac{f^c(t_i, d_i, z_i, j_i|\theta^k)}{f^c(t_i, d_i, z_i|\theta^k)}
= \frac{\alpha_j^k \left( f(t_i|\gamma_j^k, \beta^k, z_i) \right)^{d_i} \tilde{F}(t_i|\gamma_j^k, \beta^k, z_i)^{1-d_i}}{\sum_{j=1}^{m} \alpha_j^k \left( f(t_i|\gamma_j^k, \beta^k, z_i) \right)^{d_i} \tilde{F}(t_i|\gamma_j^k, \beta^k, z_i)^{1-d_i}}.
\]

The important point here is that the posterior probabilities in (4) neither depend on the censoring distribution nor on the covariate distribution. Thus we obtain

\[
Q(\theta|\theta^k) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij}^k \left[ \log \alpha_j + d_i \log f(t_i|\gamma_j, \beta, z_i) + (1-d_i) \log \tilde{F}(t_i|\gamma_j, \beta, z_i) \right]
+ R(t, d, z, h, v, \theta^k),
\]

where \( R \) does not depend on \( \theta \). Thus we delete \( R \) in the definition of \( Q(\theta|\theta^k) \).

**Example 1** We consider a finite mixture model where the j-th component survival function is defined by \( \tilde{F}(t|\gamma_j, z) = \exp(-\gamma_j t e^{\beta^T z}) \), corresponding to a parametric proportional hazard rate model with exponential (thus constant) baseline hazard rate function. Setting \( \gamma_j = e^{\xi_j - \xi_1} \) we remark that this model belongs to the M1 family of lifetime regression models with \( S_0(t) = \exp(-e^{\xi_1} t) \). The identifiability of this model parameters can be proved using Teicher (1967) and assuming that the covariates vectors \( z \) generate \( \mathbb{R}^p \). The j-th component pdf is therefore defined by

\[
f(t|\gamma_j, \beta, z) = \gamma_j \exp(\beta^T z - \gamma_j t e^{\beta^T z}),
\]
leading to
\[
Q(\theta | \theta^k) \propto \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij}^k \left[ \log \alpha_j + d_i \left\{ \log \gamma_j + \beta^T z_i - \gamma_j t_i e^{\beta^T z_i} \right\} - (1 - d_i) \gamma_j t_i e^{\beta^T z_i} \right]
\]
\[
\propto \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij}^k \left[ \log \alpha_j + d_i \left\{ \log \gamma_j + \beta^T z_i \right\} - \gamma_j t_i e^{\beta^T z_i} \right],
\]
where \(\propto\) means "equal to, up to a term that does not depend of the parameter of interest".

By solving normal equations with respect to \(\alpha_j\) for \(j = 1, \ldots, m\) we obtain
\[
\alpha_{j}^{k+1} = \frac{\sum_{i=1}^{n} \alpha_{ij}^k d_i}{\sum_{i=1}^{n} \alpha_{ij}^k t_i e^{\beta^T z_i}}.
\]

Then we write the normal equations for \(\gamma_j\):
\[
\frac{\partial Q(\theta | \theta^k)}{\partial \gamma_j} = \sum_{i=1}^{n} \alpha_{ij}^k \left[ \frac{d_i}{\gamma_j} - t_i e^{\beta^T z_i} \right] = 0
\]
for \(1 \leq j \leq m\). Considering \(\beta\) as known we solve the above equations by setting
\[
\gamma_{j}^{k+1}(\beta) = \frac{\sum_{i=1}^{n} \alpha_{ij}^k d_i}{\sum_{i=1}^{n} \alpha_{ij}^k t_i e^{\beta^T z_i}}.
\]
for \(1 \leq j \leq m\). Thus profiling the remaining part of \(Q(\cdot | \theta^k)\) as a function of \(\beta\) we estimate \(\beta\) by
\[
\beta^{k+1} = \arg \max_{\beta \in \mathbb{R}^p} Q(\beta | \theta^k)
\]
where
\[
Q(\beta | \theta^k) \propto \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij}^k \left\{ d_i \beta^T z_i - d_i \log \left( \sum_{i=1}^{n} \alpha_{ij}^k t_i e^{\beta^T z_i} \right) \right\}.
\]
Finally \(\gamma^{k+1} = (\gamma_{1}^{k+1}(\beta^{k+1}), \ldots, \gamma_{m}^{k+1}(\beta^{k+1}))\).

4. Stochastic EM–like algorithms for semiparametric models

Hereafter we consider semiparametric models 1–4. If parameter identifiability is generally well studied for parametric finite mixture models (see e.g. Teicher, 1967), the identifiability of semi- or non-parametric finite mixture model’s parameters is generally a difficult task for which there are few general tools at the exception of Allman et al. (2009). Even if this point is not discussed in details here we can say briefly that we obtained a partial identifiability results for \(\theta\) in the model \(\{(x, z) \mapsto G(x|z; \theta) = \alpha(F(x))^{\gamma + e^{\beta^T z}}, \theta = (\alpha, \gamma, \beta, F(\cdot)) \in \Theta = [0, 1] \times (1, +\infty) \times \mathbb{R}^p \times \mathcal{S}\} where \(\mathcal{S}\) is the class of continuous survival functions.

4.1 Stochastic EM–like principle

The missing data are the component numbers \(J_1, \ldots, J_n\), the common distribution of which is defined by \text{Mult}(1, \alpha). Conditionally on \(Z = z\), the survival function of \(X\) is defined by
\[
S(x|z) = \sum_{j=1}^{m} \alpha_j \tilde{F}_j(t|z, \theta)
\]
where survival functions $\tilde{F}_j(t|z, \theta)$ are defined by one of the formula for the models M1–M4. In the parametric set-up the general principle of the Stochastic EM (St–EM) algorithm is to produce a sequence of iterates $\theta^k$ (a Markov chain) such that its ergodic mean converges to the unknown value of the Euclidean parameter $\theta$ (see Nielsen, 2000). In the semiparametric set-up there is only empirical evidence that the St–EM algorithm performs well (see, e.g. Bordes and Chauveau, 2016). Given the value of the parameter $\theta^k$ at the $k$th iteration, the general St–EM algorithm follows the following steps.

Step 1. For each item $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$ calculate
\[
\alpha^k_{ij} = \frac{\alpha^k_j \left( \lambda^k_j(t_i|z_i) \right) d_i \tilde{F}^k_j(t_i|z_i)}{\sum_{l=1}^m \alpha^k_l \left( \lambda^k_l(t_i|z_i) \right) d_i \tilde{F}^k_l(t_i|z_i)}.
\]

Step 2. For each item $i \in \{1, \ldots, n\}$ simulate a realization $j^k_i$ of $\text{Mult}(1, (\alpha^k_{1i}, \ldots, \alpha^k_{im}))$ and for $1 \leq j \leq m$ define the $m$ sets
\[
X^k_l = \{ i \in \{1, \ldots, n\}; j^k_i = l \} \text{ for } 1 \leq l \leq m.
\]

We have $\bigcup_{l=1}^m X^k_l = \{1, \ldots, n\}$.

Step 3. Update the Euclidean parameters:
For $j \in \{1, \ldots, m\}$ $\alpha^{k+1}_j = \text{Card}(X^k_j)/n$.

The update of the regression parameter depends on the model under consideration.
We just detail here the situation for the first two models M1 and M2, but other models can be derived similarly:

(3.1) for Model 1: Calculate
\[
(\beta^{k+1}, \gamma^{k+1}) = \arg \max_{\beta \in \mathbb{R}^p, \gamma \in \mathbb{R}^{m-1}} L^{(1,k)}(\beta, \gamma),
\]

where
\[
L^{(1,k)}(\beta, \gamma) = \prod_{i=1}^n \left( \frac{\exp(\beta^T z_i + \gamma_j^k)}{\sum_{l=1}^m \exp(\beta^T z_l + \gamma_j^k)} \right)^{d_i}.
\]

(3.2) for Model 2:
First method, for $j \in \{1, \ldots, m\}$
\[
\beta^{k+1} = \arg \max_{\beta \in \mathbb{R}^p} L^{(2,k)}_j(\beta)
\]
where
\[
L^{(2,k)}_j(\beta) = \prod_{i \in X^k_j} \left( \frac{\exp(\beta^T z_i)}{\sum_{l \geq i, l \in X^k_j} \exp(\beta^T z_l)} \right)^{d_i}.
\]

Second method
\[
(\beta^{1,k+1}, \ldots, \beta^{m,k+1}) = \arg \max_{(\beta_1, \ldots, \beta_m) \in \mathbb{R}^{pm}} L^{(2,k)}(\beta_1, \ldots, \beta_m),
\]

where
\[
L^{(2,k)}(\beta_1, \ldots, \beta_m) = \prod_{i=1}^n \left( \frac{\exp(\beta^T z_i)}{\sum_{l=1}^m \exp(\beta^T z_l)} \right)^{d_i}.
\]

This second approach is based on a profile likelihood approach.
Step 4. Update the functional parameters: here as well we just detail the situation for M1 and M2.

(4.1) for Model 1:
\[ \Lambda_{k+1}^0(t) = \sum_{i:t_i \leq t} \frac{d_i}{\sum_{l=1}^n \exp(z_i^T \beta_{k+1}^l + \gamma_{jk}^l)}. \]

(4.2) for Model 2:
\[ \Lambda_{k+1}^0(t) = \sum_{i:t_i \leq t} \frac{d_i}{\sum_{l=1}^n \exp(z_i^T \beta_{k+1+1}^l)}. \]

Step 5. Kernel estimators: for \( j \in \{0, 1, \ldots, m\} \)
\[ \lambda_{j+1}^k(t) = \sum_{i=1}^n \frac{1}{b} K \left( \frac{t - t_i}{b} \right) \Delta \Lambda_{j+1}^k(t_i), \]
where the bandwidth has to be tuned following, e.g., rules proposed in Bordes and Chauveau (2016) and \( \Delta \Lambda_{j+1}^k(t_i) = \Lambda_{j+1}^k(t_i) - \Lambda_{j+1}^k(t_i^-). \)

Remark. It is easy to check that for models 1 and 2, since the baseline hazard rate is shared by all components, in \( \alpha_{ij}^k \) the baseline hazard rate can be factorized in the numerator and then it disappears. The consequence is that for these two models the above step 5 can be skipped.

5. Numerical study and real data analysis

5.1 M1 in a parametric case, genuine EM algorithm

We propose here an experiment in the situation of Example 1, i.e. when the \( j \)-th component survival function is defined by \( \tilde{F}(t|\gamma_j, z) = \exp(-\gamma_j t e^{\beta^T z}) \), corresponding to a parametric proportional hazard rate model with exponential (thus constant) baseline hazard rate function. In other words, the \( j \)-th component given the covariate \( z \) comes from the exponential distribution \( \mathcal{E}(\gamma_j e^{\beta^T z}) \). We choose here \( p = 2 \) independent and binary covariates \( Z = (Z_1, Z_2) \), each of which being Bernoulli \( B(0.5) \) distributed. We simulate a \( m = 2 \)-component mixture with parameters \( \alpha_1 = 30\% \), \( \gamma = (0.5, 0.1) \), \( \beta = (0.5, -0.5) \). The corresponding conditional survival functions are displayed in Fig. 1.

The EM algorithm requires, as always, initialization values for the parameter, \( \theta^{(0)} \). For this \( m = 2 \) rather simple case, we defined a data-driven initialization: From Fig. 1 we can notice that the two component are somehow separated, whatever the values of the covariates. It is possible from an histogram of the data or prior expert opinion, to define a cutpoint \( \tau \) and two sub-samples \( t^1 \) and \( t^2 \) defined by the non censored \( t_k \)'s that are below or above \( \tau \). Then we set \( \alpha_1^{(0)} \) as the proportion of non-censored observations belonging to \( t^1, \gamma_j^{(0)} = 1/\text{mean}\{t^j\} \), and a non informative initialization \( \beta^{(0)} = (1, 1) \). For \( m > 2 \), we suggest the common procedure consisting in exploring the parameter space, running EM algorithms from several (random) initializations, and optimizing in the maximum of the log-likelihood.

Fig. 2 shows a typical result in terms of the empirical distribution of the estimates, for 300 Monte-Carlo replications of samples of size \( n = 500 \), with a censoring distribution achieving an average censoring rate of 27\%. The stopping criterion here is based on the numerical stabilization of the log-likelihood, as for any genuine EM. In the present case the EM’s required an average of 330 iterations.
survival functions
\( z=(0,0) \) or \((1,1)\)
\( z=(0,1) \)
\( z=(1,0) \)
component \( j=1 \)
component \( j=2 \)

Figure 1: Example 1, true survival functions
\[
\bar{F}(t | \gamma_j, z) = \exp(-\gamma_j t e^{\beta^T z})
\]
for each covariate possible value and each component.

5.2 Model M1 with nonparametric baseline, stochastic EM-like algorithm

This semiparametric example involves a fully unknown baseline survival function \( \bar{F}_0(\cdot) \) and conditional survival for component \( j \) from M1, \( \bar{F}_j(x | z, \theta) = [\bar{F}_0(x)]^{\exp(\beta^T z + \gamma_j)} \).

The simulated model uses for \( \bar{F}_0(x) \) a Weibull distribution with shape \( a_0 \) and scale \( b_0 \), \( \bar{F}_0(x) := \exp[-(x/b_0)^{a_0}] \). It involves \( m = 2 \) components with weight \( \alpha_1 = 30\% \), and \( p = 2 \) independent covariates uniformly distributed on the interval \([0, 2]\). The regression parameters are \( \beta = (0.5, -0.5) \) and \( \gamma_2 = 3 \) (\( \gamma_1 = 0 \) for identifiability). Hence the model parameters are \( (\alpha, \gamma_2, \beta, \bar{F}_0(\cdot)) \). Fig. 3 shows the corresponding conditional densities over the range of the possible values for the covariates, together with a typical sample distribution of non-censored data from this model. The simulation of sample data is done by simulating the covariates, computing the conditional scales given each \( i \) and component \( j \) as \( s_j(i) = b_0 \exp[-(\beta^T z_i + \gamma_j)/a_0] \) and simulating each duration \((x_i | J = j, Z = z_i) \sim \mathcal{W}(a_0, s_j(i))\), a Weibull distribution with shape \( a_0 \) and the conditional scale. Then a censoring is applied, for an average 10\% of censored observations.

As in Example 1 Section 5.1, the algorithm requires an initialization and this case is more tricky than the previous one; in particular a “non informative” initialization for the parameters \( \beta \) as in Example 1 does not work well for this more complex model. We experiment here a new data-driven initialization procedure. First, from Fig. 3 we can notice that the two components are somehow separated, whatever the values of the covariates, so that we start by defining a cutpoint \( \tau \) and two sub-samples \( t^1 \) and \( t^2 \) from an histogram of the data or prior expert opinion as in Section 5.1. A cutpoint \( \tau = 2 \) has been chosen here. Then the procedure involves the following steps:

1. Fit a single weibull distribution on the sample \((t, d)\) to get initial values for step 2.

This fit can be done by calling standard MLE packages for right censored data from standard distributions. We use the \texttt{survreg()} function for the \texttt{survival} package (Therneau and Lumley, 2009) for the \texttt{R} statistical software (R Core Team, 2013).
2. Fit separately two Weibull distributions to each subsample \((t^1, d^1)\) and \((t^2, d^2)\), where \(d^\ell\) are the censoring indicators corresponding to the lifetimes \(t^\ell\) for the \(\ell\)th subsample. This is done by applying again `survreg()` but with initial values provided by step 1.

3. Fit a two-component mixture of Weibull distributions with censored data to the whole sample \((t, d)\) using the specific St-EM algorithm from Bordes and Chauveau (2016). This St-EM itself requires initial parameters for weight, shape and scale per components. The initial weight \(\alpha_0^\ell\) is defined as the proportion of observations belonging to \(t^1\), and shape and scale per components are the estimates obtained in step 2.

4. Using the posterior probabilities obtained by the St-EM algorithm in step 3, simulate a starting vector \(J^0\) of component origin for each individual. This is similar to Step 2 of the Stochastic EM-like algorithm described in 4.1.

5. Fit a Cox PHM applying the function `coxph()` for a model with covariates \((z, J^0)\), i.e. \((t, d) \sim z_1 + z_2 + J^0\). This gives initial values \(\beta^0, \gamma_2^0\) and \(\bar{F}_0(\cdot)\).

We applied the above procedure to Monte-Carlo replications and several sample sizes from \(n = 500\) to \(n = 5000\), with good results suggesting “empirical convergence”. An example is displayed in Fig. 4, which shows the empirical distribution of the estimates for the scalar parameters, and the estimates of the baseline \(\bar{F}_0\) over replications, in the case of a sample of size \(n = 1000\). Table 1 below gives numerical results for two sample sizes.
Figure 3: Semiparametric example: empirical distribution of a sample of size $n = 1000$ of non-censored data and the true density functions for each component over the range of covariates $z$.

<table>
<thead>
<tr>
<th>sample size</th>
<th>stat.</th>
<th>$\alpha_1$</th>
<th>$\gamma_2$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td></td>
<td>0.3</td>
<td>3</td>
<td>0.5</td>
<td>-0.5</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>mean</td>
<td>0.29</td>
<td>3.15</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>0.018</td>
<td>0.220</td>
<td>0.077</td>
<td>0.078</td>
</tr>
<tr>
<td></td>
<td>mse</td>
<td>0.0004</td>
<td>0.0700</td>
<td>0.0060</td>
<td>0.0061</td>
</tr>
<tr>
<td>$n = 2000$</td>
<td>mean</td>
<td>0.30</td>
<td>3.10</td>
<td>0.50</td>
<td>-0.48</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>0.014</td>
<td>0.166</td>
<td>0.058</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>mse</td>
<td>0.00023</td>
<td>0.03828</td>
<td>0.00332</td>
<td>0.00290</td>
</tr>
</tbody>
</table>

Table 1: Estimated means, standard deviations and MSE’s from 100 replications of the semiparametric St-EM algorithm.

6. Discussion

We have proposed several iterative methods based on EM and Stochastic EM methodologies, for parametric and semiparametric PHM’s designed for randomly right censored lifetime data. In particular, we have illustrated the behavior of these algorithms for a parametric model allowing for a genuine EM, and a more complex semiparametric model requiring a St-EM algorithm.

For both strategies, we defined data-driven automated initialization procedures that perform in a satisfactory manner. This question of initialization can indeed be delicate, as illustrated by the semiparametric model and St-EM algorithm, for which a multiple stage procedure involving itself several simpler models and algorithms has been designed.

Asymptotic variance of the St-EM estimates is only available for parametric models (Nielsen, 2000), but in the situations experimented through Monte-Carlo simulations, our algorithms provide good estimates and decreasing MSE’s when the sample size increases, suggesting numerical evidence of convergence of these algorithms.
Figure 4: Exemple 2 semiparametric model: empirical distributions of St-EM estimates based on 100 replications of a sample of size $n = 1000$. Green (dotted) lines or curves are true values, red lines are estimates averaged over replications.

All the algorithms shown here are implemented — and will be publicly available — in an upcoming version of the mixtools package (Benaglia et al., 2009) for the R statistical software (R Core Team, 2013).

References


