Estimation de modèles de mélanges en présence de censure

Laurent Bordes - Université de Pau et des Pays de l'Adour

en collaboration avec

Didier Chauveau - Université d'Orléans

Séminaire LJK-Statistique

Plan de l'exposé

Plan

Durées de vies

Durées de vies issues d'un modèle de mélange à m composantes de densités f_z, $z = 1, \ldots, m$, où $f_z(\cdot) \in \mathcal{F}$ une famille de densités. La densité de X s'écrit

$$X \sim g(x|\boldsymbol{\theta}) = \sum_{z=1}^{m} \lambda_z f_z(x).$$

- **Cas semi-paramétrique :**
  $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{f}) = (\lambda_1, \ldots, \lambda_m, f_1, \ldots, f_m) \in \Lambda_m \times \mathcal{F}^m$ où
  $\Lambda_m = \{(\lambda_1, \ldots, \lambda_m) \in [0,1]^m; \sum_{z=1}^{m} \lambda_z = 1\}$.
- **Cas paramétrique :** $\boldsymbol{\theta}$ est réduit à $(\boldsymbol{\lambda}, \boldsymbol{\xi}) = (\lambda_1, \ldots, \lambda_m, \xi_1, \ldots, \xi_m)$ si $\mathcal{F} = \{f(\cdot|\xi); \xi \in \Xi\}$.

Durées de vies

Durées de vies issues d'un modèle de mélange à m composantes de densités f_z, $z = 1, \ldots, m$, où $f_z(\cdot) \in \mathcal{F}$ une famille de densités. La densité de X s'écrit

$$X \sim g(x|\boldsymbol{\theta}) = \sum_{z=1}^{m} \lambda_z f_z(x).$$

- **Cas semi-paramétrique :**
  $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{f}) = (\lambda_1, \ldots, \lambda_m, f_1, \ldots, f_m) \in \Lambda_m \times \mathcal{F}^m$ où
  $\Lambda_m = \{(\lambda_1, \ldots, \lambda_m) \in [0,1]^m; \sum_{z=1}^{m} \lambda_z = 1\}$.
- **Cas paramétrique :** $\boldsymbol{\theta}$ est réduit à $(\boldsymbol{\lambda}, \boldsymbol{\xi}) = (\lambda_1, \ldots, \lambda_m, \xi_1, \ldots, \xi_m)$ si $\mathcal{F} = \{f(\cdot|\xi); \xi \in \Xi\}$.
- **Données complètes :** $Y = (X, Z)$ où $Z \sim \mathcal{M}\text{ult}(1, \boldsymbol{\lambda})$ et $X|Z = z \sim f_z(\cdot)$. Une référence pour les modèles de mélange : McLachlan and Peel (2000).

Durées de vies

Durées de vies issues d'un modèle de mélange à m composantes de densités f_z, $z = 1, \ldots, m$, où $f_z(\cdot) \in \mathcal{F}$ une famille de densités. La densité de X s'écrit

$$X \sim g(x|\boldsymbol{\theta}) = \sum_{z=1}^{m} \lambda_z f_z(x).$$

- **Cas semi-paramétrique :**
  $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{f}) = (\lambda_1, \ldots, \lambda_m, f_1, \ldots, f_m) \in \Lambda_m \times \mathcal{F}^m$ où
  $\Lambda_m = \{(\lambda_1, \ldots, \lambda_m) \in [0,1]^m; \sum_{z=1}^{m} \lambda_z = 1\}$.
- **Cas paramétrique :** $\boldsymbol{\theta}$ est réduit à $(\boldsymbol{\lambda}, \boldsymbol{\xi}) = (\lambda_1, \ldots, \lambda_m, \xi_1, \ldots, \xi_m)$ si $\mathcal{F} = \{f(\cdot|\xi); \xi \in \Xi\}$.
- **Données complètes :** $Y = (X, Z)$ où $Z \sim \mathcal{M}\text{ult}(1, \boldsymbol{\lambda})$ et $X|Z = z \sim f_z(\cdot)$. Une référence pour les modèles de mélange : McLachlan and Peel (2000).

Des exemples semi-paramétriques

- **Vie accélérée :**

$$X|Z = z \sim \frac{1}{\xi_z} f\left(\frac{\cdot}{\xi_z}\right) \text{ pour } 1 \leq z \leq m$$

alors $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_m, \xi_1, \ldots, \xi_m, f) \in \Lambda_m \times (0, +\infty)^m \times \mathcal{F}$ et

$$g(x|\boldsymbol{\theta}) = \sum_{z=1}^{m} \frac{\lambda_m}{\xi_z} f\left(\frac{x}{\xi_z}\right).$$

- **Risques proportionnels :** $\mathcal{L}$ est l'ensemble des fonctions de risque

$$\alpha_{X|Z}(x|z) = \exp(\xi_z)\alpha_0(\cdot) \text{ pour } 1 \leq z \leq m.$$

alors $\theta = (\lambda_1, \ldots, \lambda_m, \xi_1, \ldots, \xi_m, \alpha_0) \in \Lambda_m \times (0, +\infty)^m \times \mathcal{L}$ et

$$\bar{G}(x|\theta) = \sum_{z=1}^{m} \lambda_z \left(\bar{F}_0(x)\right)^{\exp(\xi_z)} \text{ où } \bar{F}_0(x) = \exp\left(-\int_0^x \alpha_0(s)ds\right).$$

Des exemples semi-paramétriques

- **Vie accélérée :**

$$X|Z = z \sim \frac{1}{\xi_z} f\left(\frac{\cdot}{\xi_z}\right) \text{ pour } 1 \leq z \leq m$$

alors $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_m, \xi_1, \ldots, \xi_m, f) \in \Lambda_m \times (0, +\infty)^m \times \mathcal{F}$ et

$$g(x|\boldsymbol{\theta}) = \sum_{z=1}^{m} \frac{\lambda_m}{\xi_z} f\left(\frac{x}{\xi_z}\right).$$

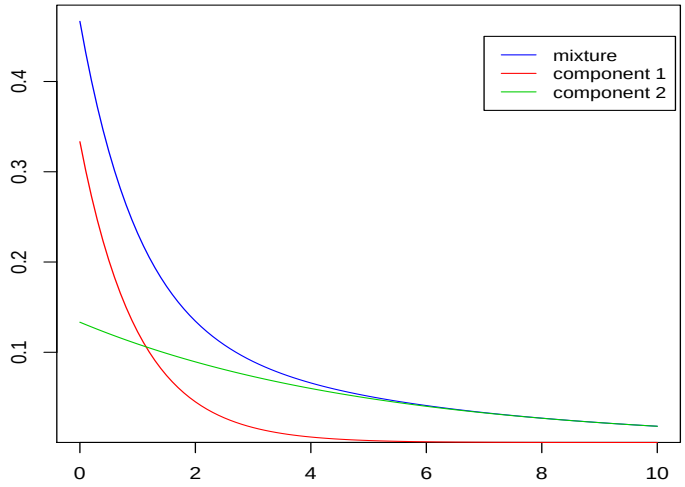
- **Risques proportionnels :** $\mathcal{L}$ est l'ensemble des fonctions de risque

$$\alpha_{X|Z}(x|z) = \exp(\xi_z)\alpha_0(\cdot) \text{ pour } 1 \leq z \leq m.$$

alors $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_m, \xi_1, \ldots, \xi_m, \alpha_0) \in \Lambda_m \times (0, +\infty)^m \times \mathcal{L}$ et

$$\bar{G}(x|\boldsymbol{\theta}) = \sum_{z=1}^{m} \lambda_z \left(\bar{F}_0(x)\right)^{\exp(\xi_z)} \text{ où } \bar{F}_0(x) = \exp\left(-\int_0^x \alpha_0(s)ds\right).$$

Exemples paramétriques



Distributions classiques en durées de vies :

$$\mathcal{F} = \{\text{Weibull}\} \cup \{\text{log-normale}\} \cup \dots$$

Objectif pratique.

Proposer des algorithmes qui permettent d'ajuster des mélanges avec des composantes issues de familles paramétriques variées...

Censure à droite

La censure est représentée par une variable aléatoire C admettant la densité q, la f.d.r. Q et la survie $\bar{Q}$. Dans le contexte usuel de censure à droite l'information disponible est

$$T = \min(X, C), \quad D = 1_{\{X \leq C\}}.$$

Les n durées sont $\boldsymbol{X} = (X_1, \ldots, X_n)$ i.i.d. $\sim g$, associées à n temps de censure $\boldsymbol{C} = (C_1, \ldots, C_n)$ i.i.d. $\sim q$. In fine les observations sont

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

où $t_i = \min(x_i, c_i)$ and $d_i = 1_{\{x_i \leq c_i\}}$.

Deux niveaux de données complètes ($T \perp C$)

- **Données complètes :** (X, Z)

$$f^c(x, z|\boldsymbol{\theta}) = f^c_{\boldsymbol{\theta}}(X = x, Z = z) = \lambda_z f_z(x).$$

- **Données complètes :** (T, D, Z)

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

et de manière similaire $f^c_{\boldsymbol{\theta}}(t, 0, z) = \lambda_z \bar{F}_z(t) q(t)$, la densité pour données complètes se résume en

$$f^c(t, d, z|\boldsymbol{\theta}) = \left[\lambda_z f_z(t) \bar{Q}(t)\right]^d \left[\lambda_z \bar{F}_z(t) q(t)\right]^{1-d}.$$

Deux niveaux de données complètes ($T \perp C$)

- **Données complètes :** (X, Z)

$$f^c(x, z|\boldsymbol{\theta}) = f^c_{\boldsymbol{\theta}}(X = x, Z = z) = \lambda_z f_z(x).$$

- **Données complètes :** (T, D, Z)

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

et de manière similaire $f^c_{\boldsymbol{\theta}}(t, 0, z) = \lambda_z \bar{F}_z(t) q(t)$, la densité pour données complètes se résume en

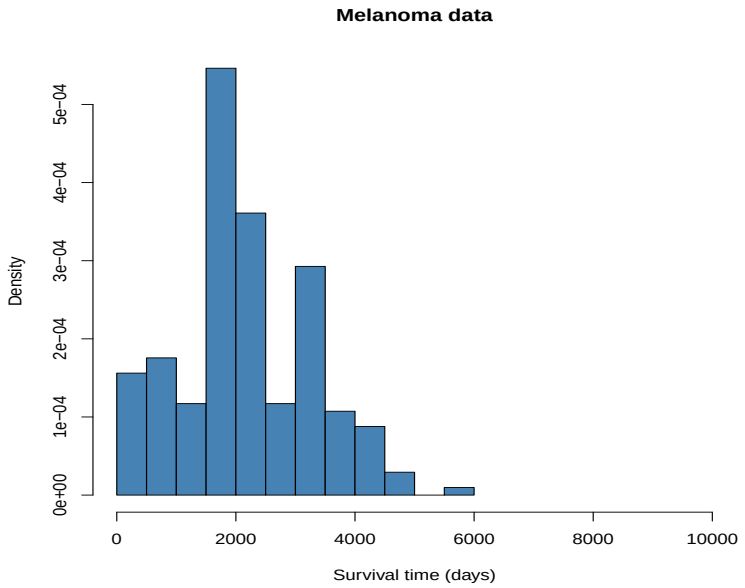
$$f^c(t, d, z|\boldsymbol{\theta}) = \left[\lambda_z f_z(t) \bar{Q}(t)\right]^d \left[\lambda_z \bar{F}_z(t) q(t)\right]^{1-d}.$$

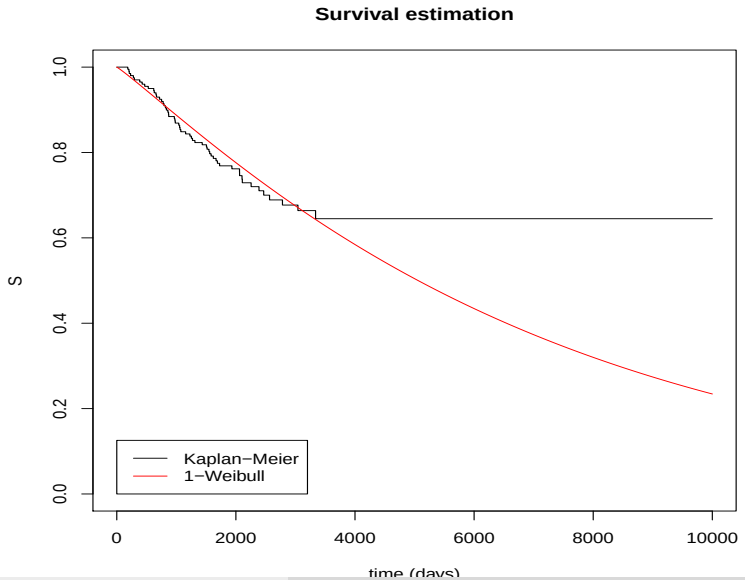
Plan

Données mélanome (Andersen et al., 1993)

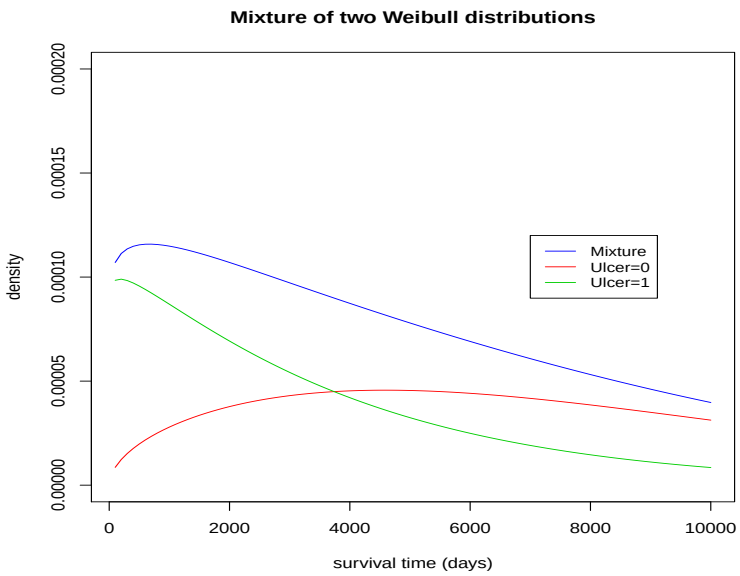
- Effet de groupe : ulcer=0/1
- Taille échantillon : 205
- Censure : 148 (72%)

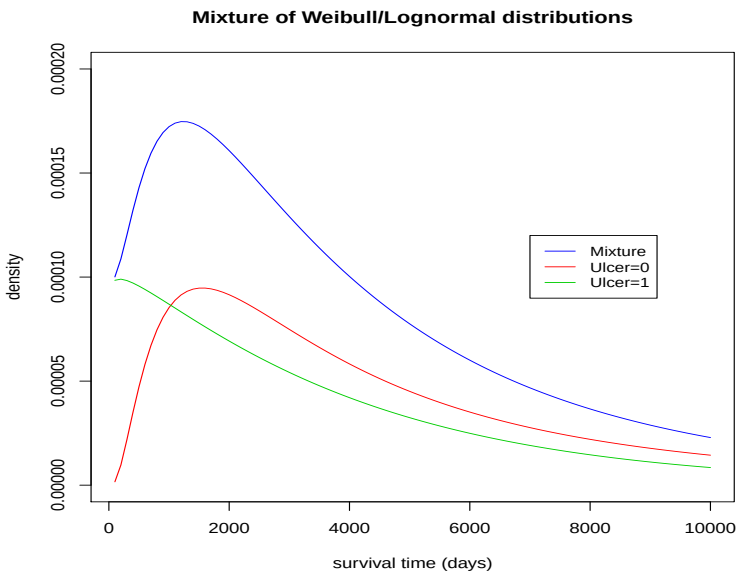
Variables	Description
`time`	survie ou censure
`status`	indicateur censure
`ulcer`	0 or 1

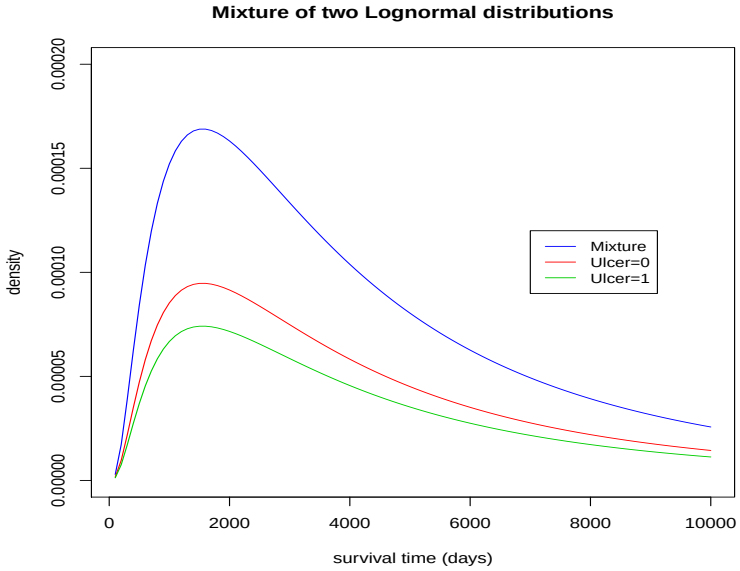


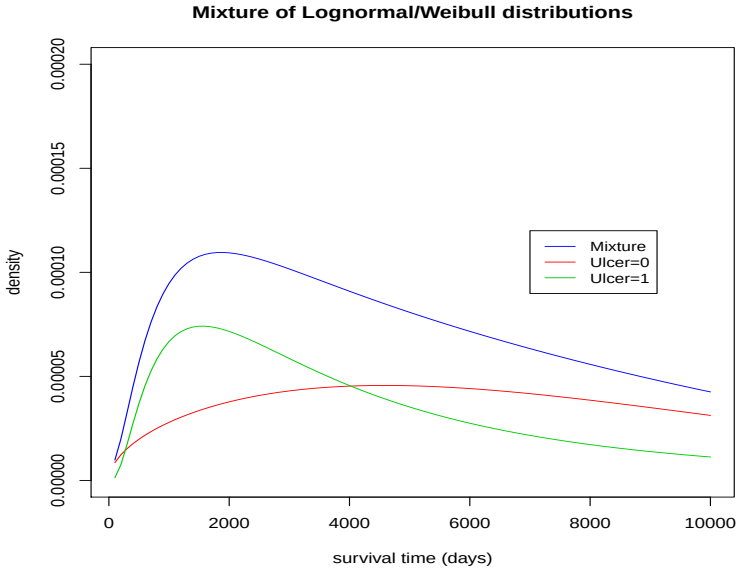


Données mélanome



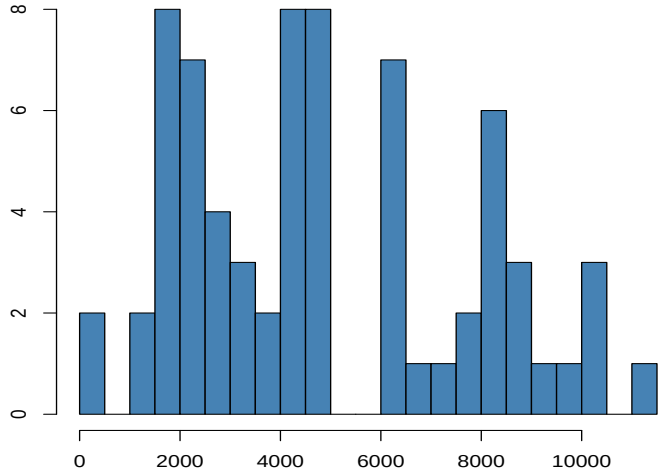




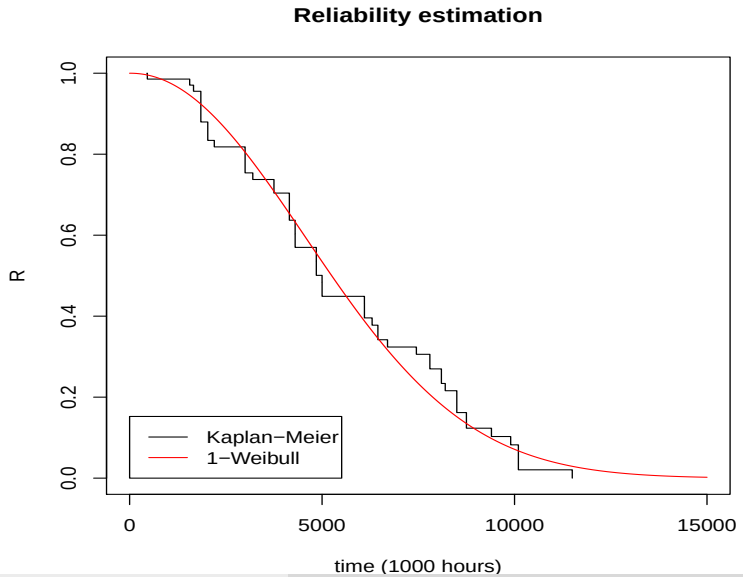


Durées de fonctionnement de ventilateurs (Nelson, 1982)

- Temps (1000 heures)
- Taille échantillon : 70
- Censure : 12



Variables	Description
`time`	survie ou censure
`status`	indicateur de censure

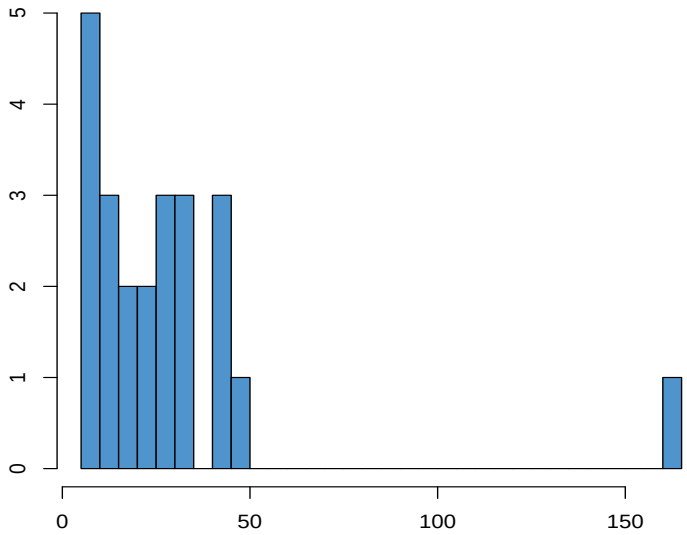


Leucémie myeloïde (Miller, 1997)

- Effet de groupe
- Taille échantillon : 23
- Censures : 5

groupe	échelle
Maintained	63.3
Nonmaintained	25.1

Variables	Description
time	survie ou censure
status	indicateur de censure
x	chimiothérapie d'appoint



Plan

Données manquantes

Nous avons :

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ données complètes avec $Y_i = (X_i, Z_i) \sim f(\cdot|\boldsymbol{\theta})$
- $\boldsymbol{X} = (X_1, \ldots, X_n)$ données observées
- $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ données manquantes

alors

$$\boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{Z}) \sim f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) = f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \times f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})$$

La log-vraisemblance lorsque les composantes de $\boldsymbol{Y}$ sont indépendantes

$$\ell^c(\boldsymbol{y}, \boldsymbol{\theta}) = \log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{Y_i|X_i}(y_i|x_i, \boldsymbol{\theta}) + \ell(\boldsymbol{x}, \boldsymbol{\theta})$$

où $\ell(\boldsymbol{x}, \boldsymbol{\theta})$ est la log-vraisemblance des données observées et $\boldsymbol{\theta} \in \Theta$.

Dempster, Laird et Rubin (1977) Algorithme EM (1/2)

Log-vraisemblance observée

$$\ell(\boldsymbol{x}|\boldsymbol{\theta}) = \mathbb{E}\left[\ell(\boldsymbol{x}|\boldsymbol{\theta})|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{\theta}'\right] = \underbrace{\mathbb{E}\left[\ell^c(\boldsymbol{Y}|\boldsymbol{\theta})|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{\theta}'\right]}_{Q(\boldsymbol{\theta}|\boldsymbol{\theta}')}$$
$$\underbrace{-\sum_{i=1}^{n}\mathbb{E}\left[\log f_{Y_i|X_i}((x_i,Z_i)|X_i=x_i,\boldsymbol{\theta})|X_i=x_i,\boldsymbol{\theta}'\right]}_{H(\boldsymbol{\theta},\boldsymbol{\theta}')},$$

où

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}') &= \mathbb{E}\left[\text{log-vraisemblance complète en } \boldsymbol{\theta}|\, \text{données observées}, \boldsymbol{\theta}'\right] \\ &= \sum_{i=1}^{n}\mathbb{E}\left[\log f(x_i,Z_i|\boldsymbol{\theta})|X_i=x_i,\boldsymbol{\theta}'\right] \end{aligned}$$

Dempster, Laird et Rubin (1977) Algorithme EM (2/2)

Principe de l'algorithme EM

Initialisation : valeur initiale $\boldsymbol{\theta}^{(0)}$ pour $\boldsymbol{\theta}$,
soit $\boldsymbol{\theta}^{(k)}$ la valeur courante de $\boldsymbol{\theta}$ obtenue à l'itération k :

Espérance : calculer $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$,

Maximisation : chercher $\boldsymbol{\theta}^{(k+1)} = \arg\max_{\theta\in\Theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$.

Pourquoi ça marche ?

$$\begin{aligned}
&\ell(\boldsymbol{x}|\theta^{(k+1)}) - \ell(\boldsymbol{x}|\theta^{(k)}) \\
= &\underbrace{\left[Q(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)}) - Q(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)})\right]}_{\geq 0 \text{ via étape } M} - \underbrace{\left[H(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)}) - H(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)})\right]}_{\leq 0 \text{ via l'inégalité de Jensen}}.
\end{aligned}$$

Celeux et Diebolt (1986) Algorithme St-EM

On a $Y = (X, Z) \sim f_Y(\cdot|\boldsymbol{\theta}) \Rightarrow Z|X = x \sim f_{Z|X}(\cdot|x, \boldsymbol{\theta})$.
On observe $x_1, \ldots, x_n$ i.i.d. issues de la distribution de X.

Principe de l'algorithme St-EM

Initialisation : valeur initiale $\boldsymbol{\theta}^{(0)}$ requise pour $\boldsymbol{\theta}$,
soit $\boldsymbol{\theta}^{(k)}$la valeur courante de $\boldsymbol{\theta}$ obtenue à l'itération k :

Espérance Stochastique : pour $i = 1, \ldots, n$ générer $z_i^{(k)}$ selon
$Z_i^{(k)} \sim k_{Z|X}(\cdot|x_i, \boldsymbol{\theta}^{(k)})$.

Maximisation : poser $y_i^{(k)} = (x_i, z_i^{(k)})$ et déterminer

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\theta \in \Theta} \log \prod_{i=1}^{n} f_Y(y_i^{(k)}|\boldsymbol{\theta}).$$

Remarque.

Pour des résultats asymptotiques voir Nielsen (2000).

Plan

EM paramétrique : données complètes $= (\boldsymbol{t}, \boldsymbol{d}, \boldsymbol{z})$

Densité des données complètes

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

Alors

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

Le calcul de $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$ requiert le calcul des probabilités *a posteriori* suivantes

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

Durées exponentielles : données complètes $= (\boldsymbol{t}, \boldsymbol{d}, \boldsymbol{z})$

Algorithme EM : $\boldsymbol{\theta}^k \to \boldsymbol{\theta}^{k+1}$

1. **Étape E :** Calculer les probabilités a posteriori p_{ij}^k selon l'équation (1), pour $i = 1, \ldots, n$ et $j = 1, \ldots, m$.
2. **Étape M :** Itérer selon

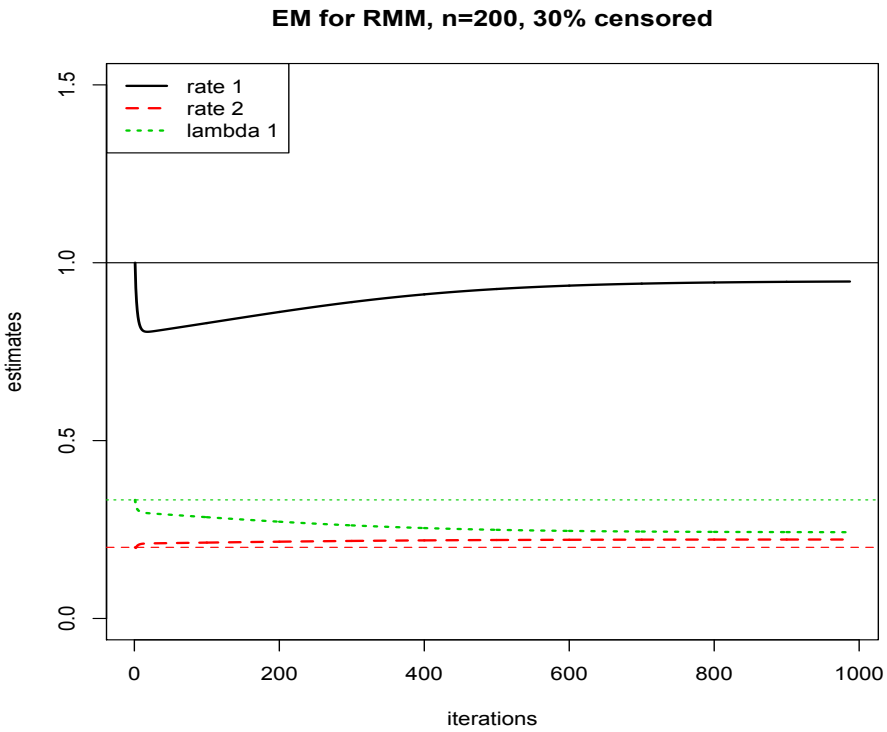
$$\lambda_j^{k+1} = \frac{1}{n} \sum_{i=1}^{n} p_{ij}^k \quad \text{pour } j = 1, \ldots, m$$

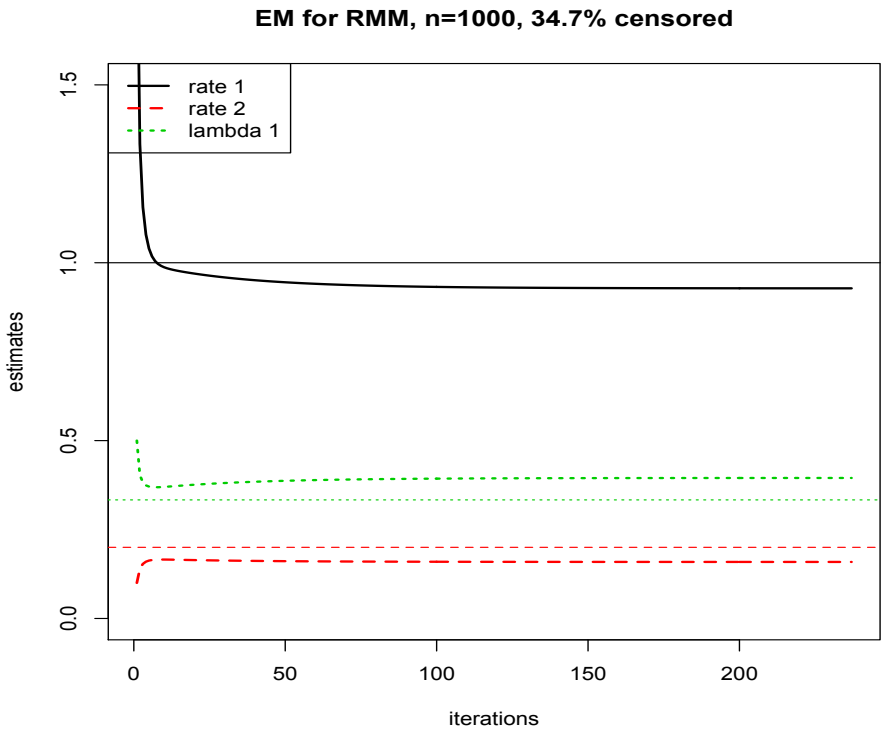
$$\xi_j^{k+1} = \frac{\sum_{i=1}^{n} p_{ij}^k \, d_i}{\sum_{i=1}^{n} p_{ij}^k \, t_i} \quad \text{pour } j = 1, \ldots, m.$$

Exemple simulé

$$g(x) = \lambda_1 \xi_1 \exp(-\xi_1 x) + \lambda_2 \xi_2 \exp(-\xi_2 x) \quad x > 0,$$

avec $\xi_1 = 1$ – – –, $\xi_2 = 0.2$ – – – et $\lambda_1 = 1/3$ – – –.





EM paramétrique : données complètes $= (\boldsymbol{x}, \boldsymbol{z})$

La densité des données complètes est

$$f^c(x, z) = \lambda_z f_z(x).$$

Alors

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

Le calcul de $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$ requiert le calcul des probabilités *a posteriori* suivantes

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

Durées exponentielles : données complètes $= (\boldsymbol{x}, \boldsymbol{z})$

Algorithme EM : $\boldsymbol{\theta}^k \to \boldsymbol{\theta}^{k+1}$

1. **Étape E :** Calculer les probabilités *a posteriori* : p_{ij}^k comme dans l'équation (1), pour $i = 1, \ldots, n$ et $j = 1, \ldots, m$.
2. **Étape M :** Pour $j = 1, \ldots, m$

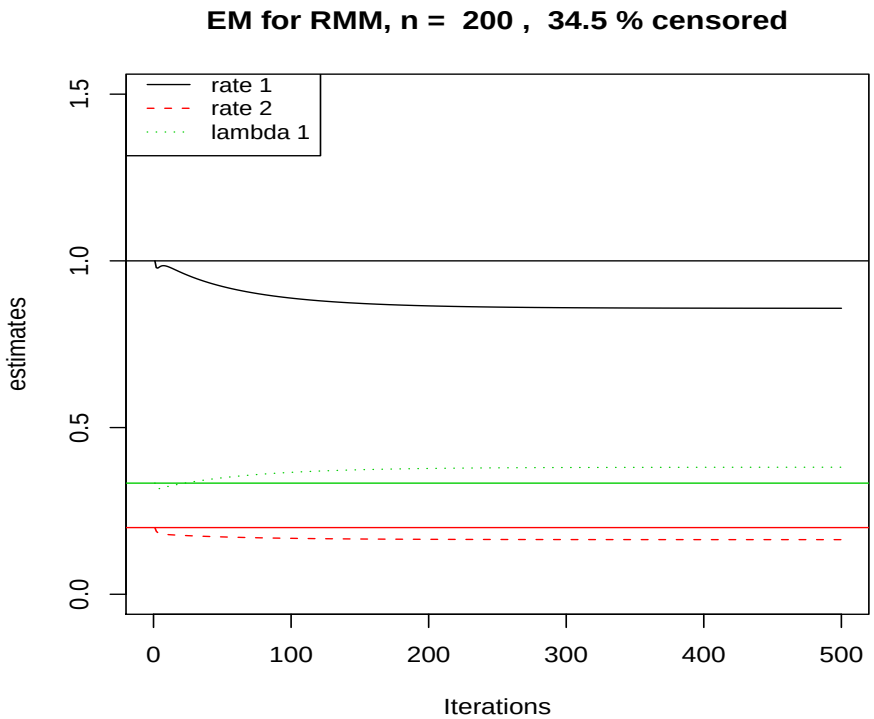
$$\lambda_j^{k+1} = \frac{1}{n} \sum_{i=1}^{n} p_{ij}^k,$$

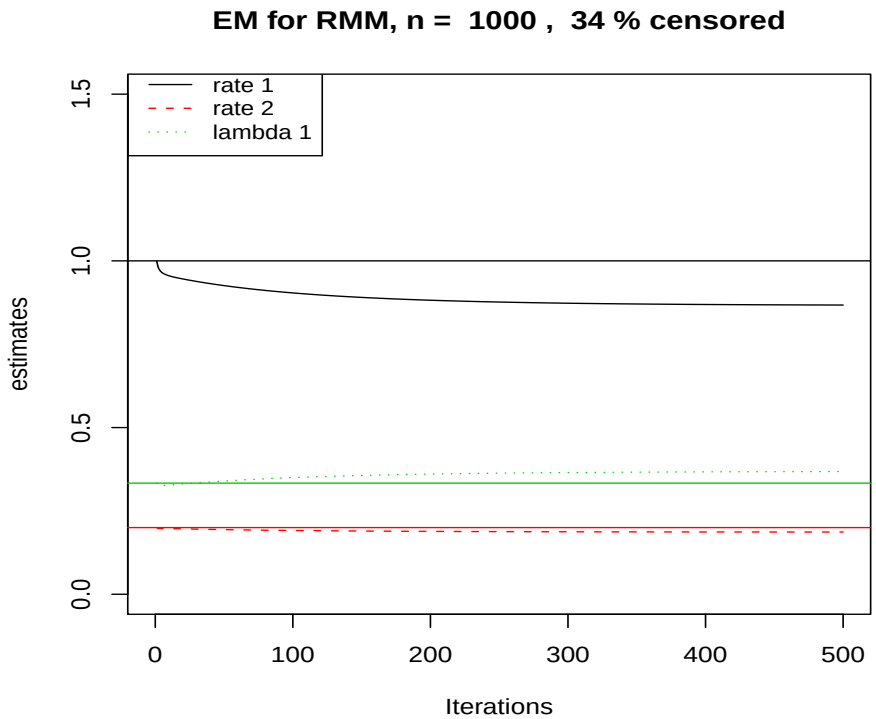
$$\xi_j^{k+1} = \frac{\sum_{i=1}^{n} p_{ij}^k}{\sum_{i=1}^{n} \left(d_i t_i p_{ij}^k + (1 - d_i) \frac{\lambda_j^k (1 + \xi_j^k t_i) \exp(-\xi_j^k t_i)}{\xi_j^k \sum_{\ell=1}^{P} \lambda_\ell^k \exp(-\xi_\ell^k t_i)} \right)}.$$

Exemple simulé

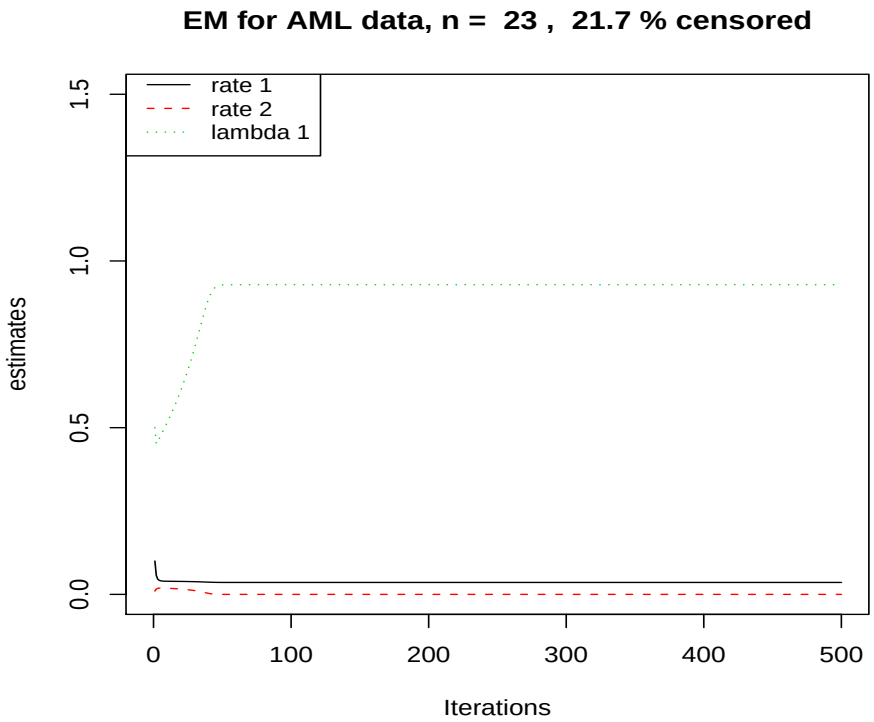
$$g(x) = \lambda_1 \xi_1 \exp(-\xi_1 x) + \lambda_2 \xi_2 \exp(-\xi_2 x) \quad x > 0,$$

avec $\xi_1 = 1$ – – –, $\xi_2 = 0.2$ – – – et $\lambda_1 = 1/3$ – – –.

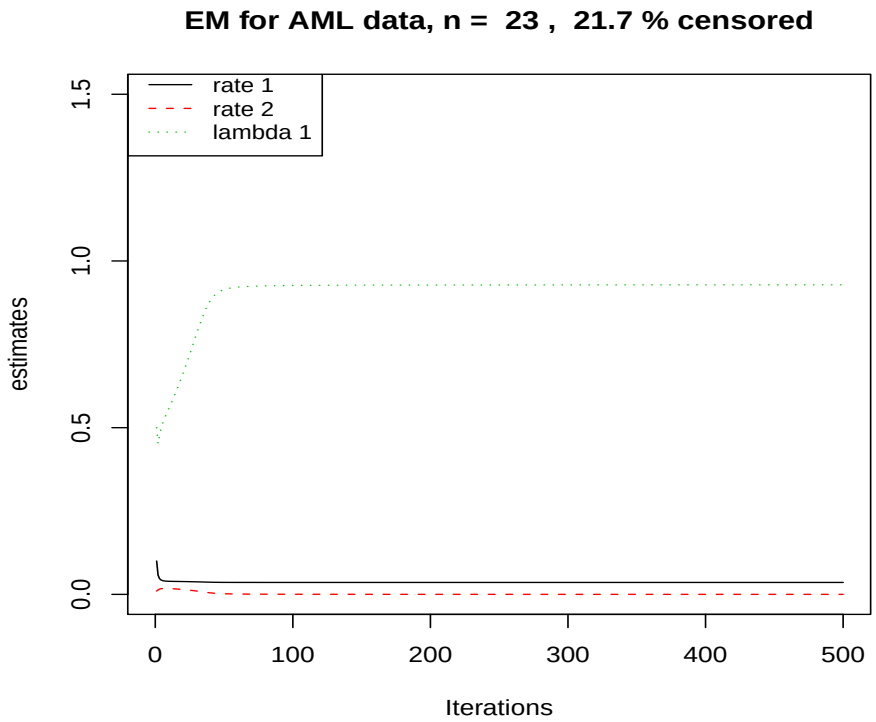




Application aux données AML



Données complètes $= (T, D, Z)$



Données complètes $= (X, Z)$

Remarques sur l'algorithme EM paramétrique

- Quel que soit le choix des données complètes l'étape M pour les λ_j est toujours explicite.
- Le calcul de $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$ dépend fortement du choix de la famille paramétrique $\mathcal{F}$.
- Hormis le cas exponentiel peu de chance de localiser explicitement les maxima de $\boldsymbol{\theta} \mapsto Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$.
- Maximiser $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$ peut être aussi *compliqué* que maximiser la vraisemblance des données observées.

Plan

L'approche St-EM [1/2]

- Idée de Celeux et Diebolt (1985, 1986) : à chaque itération ajouter une étape **stochastique** où les données manquantes sont simulées selon leur distribution *a posteriori* pour la valeur courante $\boldsymbol{\theta}^k$ du paramètre inconnu $\boldsymbol{\theta}$.
- Quel choix pour les *données complètes* ? Il est suffisant de choisir $(\boldsymbol{t}, \boldsymbol{d}, \boldsymbol{z})$.
- Notons $\boldsymbol{p}_i^k = (p_{i1}^k, \ldots, p_{im}^k)$ le vecteur des probabilités *a posteriori* de l'observation i. Considérons $Z \sim \mathcal{M}\mathrm{ult}(1, \boldsymbol{p}_i^k)$ une v.a. de loi multinomiale de paramètres 1 et $\boldsymbol{p}_i^k$ (i.e. $Z \in \{1, \ldots m\}$ avec les probabilités $\mathbb{P}(Z = j) = p_{ij}^k$)).

L'approche St-EM [2/2]

Algorithme St-EM : $\boldsymbol{\theta}^k \rightarrow \boldsymbol{\theta}^{k+1}$

1. **Étape E :** Calcul des probabilités *a posteriori* p_{ij}^k comme dans l'équation (1), pour $i = 1, \ldots, n$ et $j = 1, \ldots, m$.
2. **Étape Stochastique :** Simuler $Z_i^k \sim \mathcal{M}\mathrm{ult}(1, \boldsymbol{p}_i^k)$, $i = 1, \ldots, n$, et definir les sous-ensembles

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

3. **Étape M :** Pour chaque composante $j \in \{1, \ldots, m\}$

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

et

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

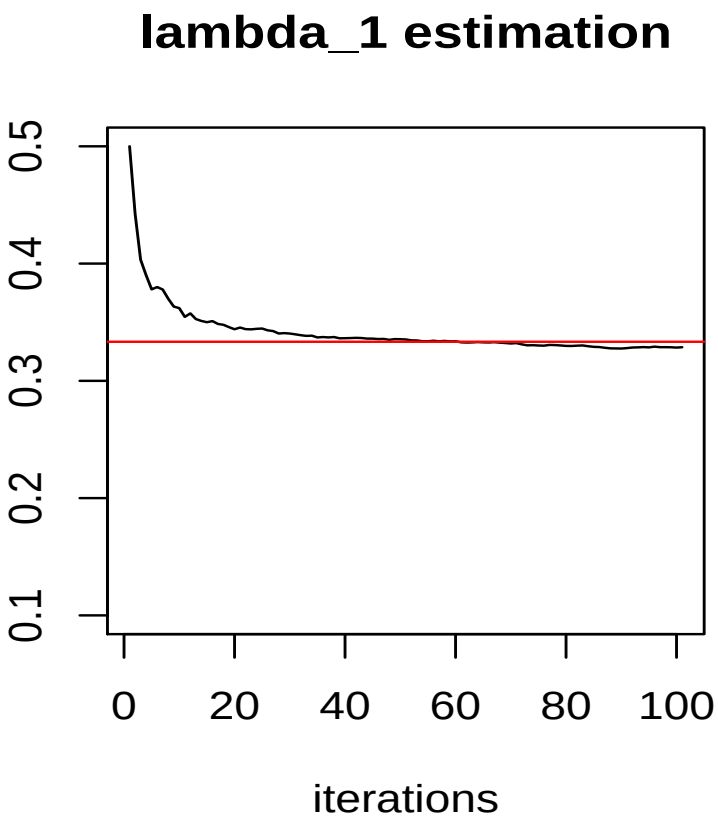
où

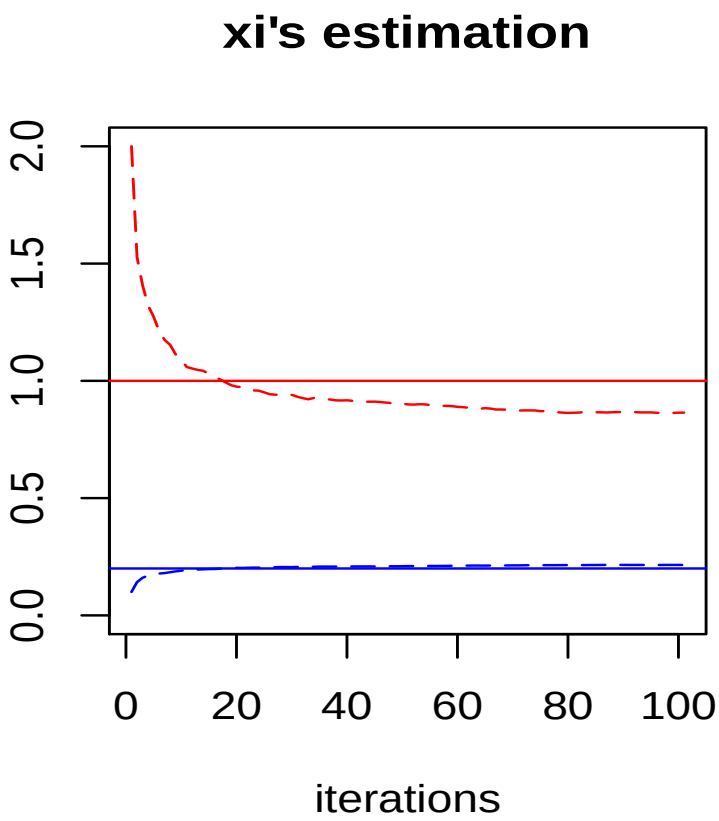
$$L_j(\xi) = \prod_{i \in \chi_j^k} (f(t_i|\xi))^{d_i} (\bar{F}(t_i|\xi))^{1-d_i}.$$

Exemple d'un mélange d'exponentielles ($n = 200$)

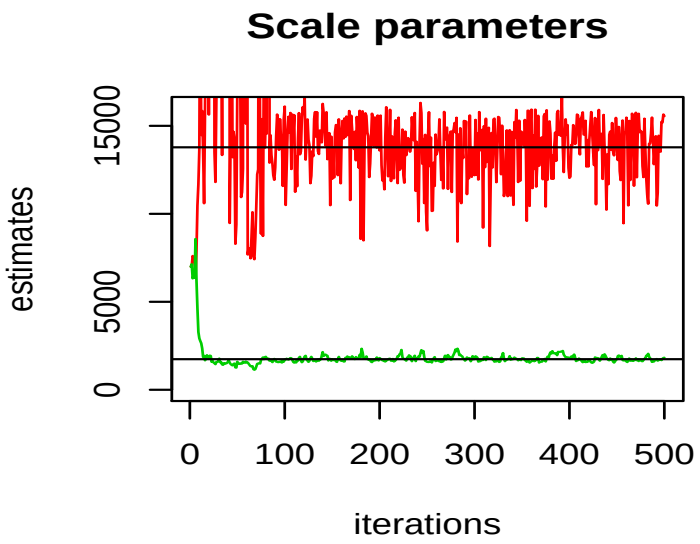
$$g(x) = \lambda\xi_1 \exp(-\xi_1 x) + (1-\lambda)\xi_2 \exp(-\xi_2 x) \quad x > 0,$$

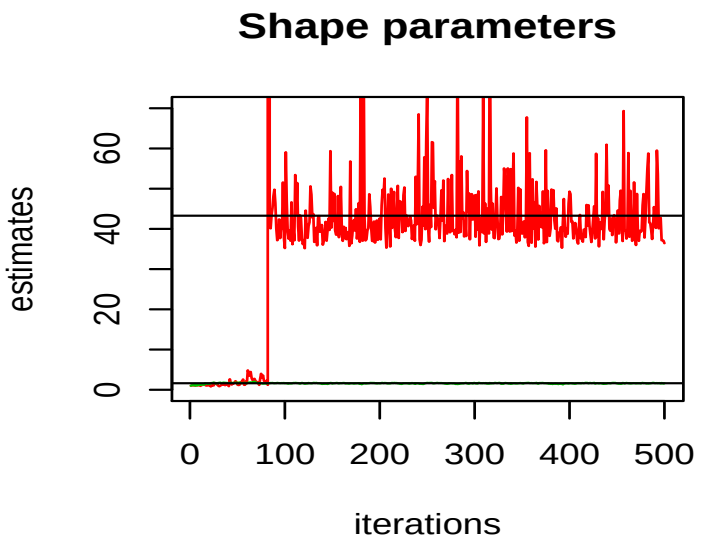
avec $\lambda = 1/3$, $\xi_1 = 1$ et $\xi_2 = 1/5$.

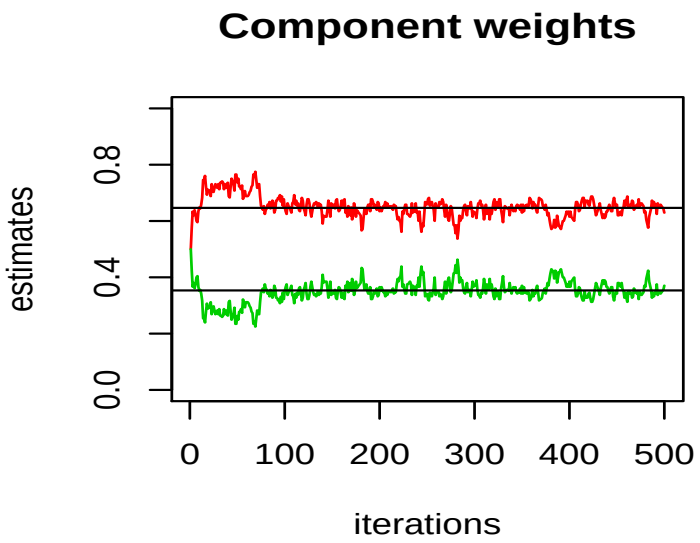


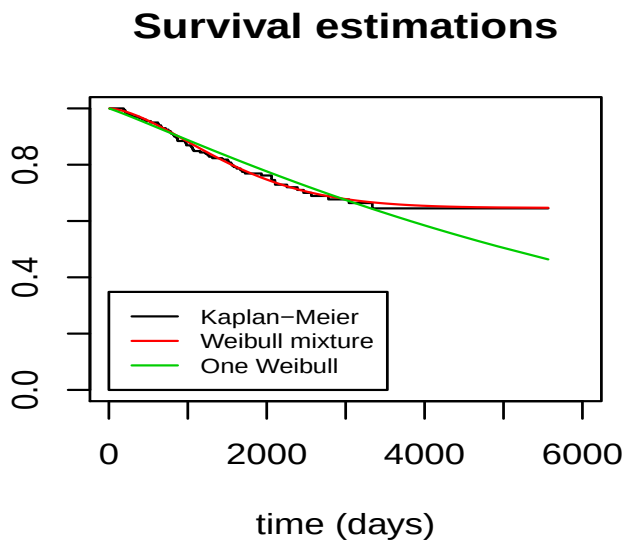


Données mélanome (deux lois de Weibull)

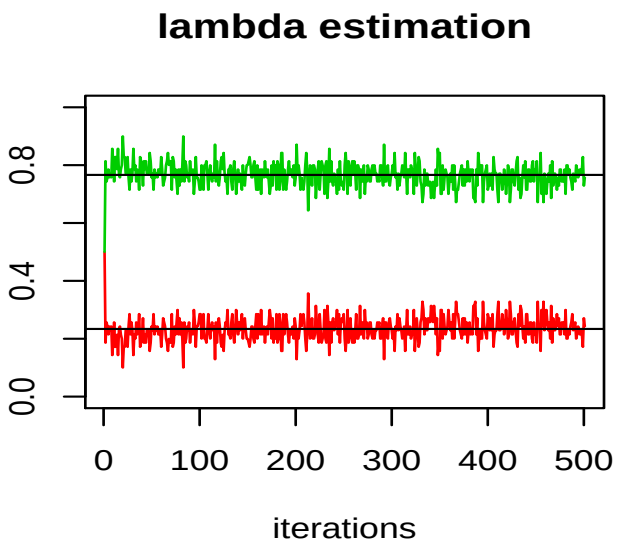


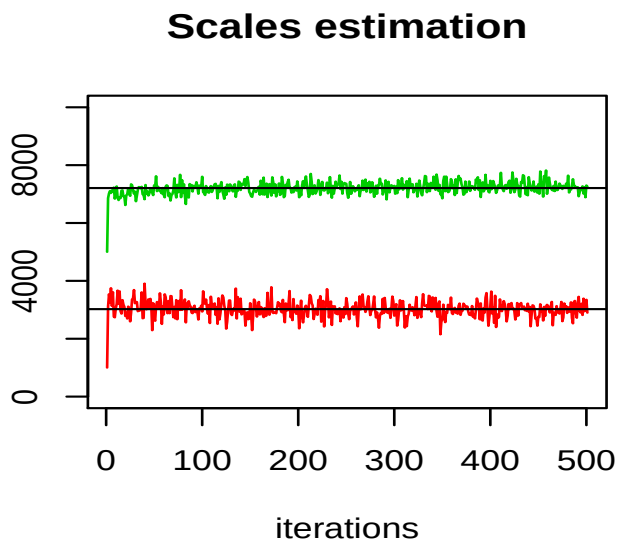


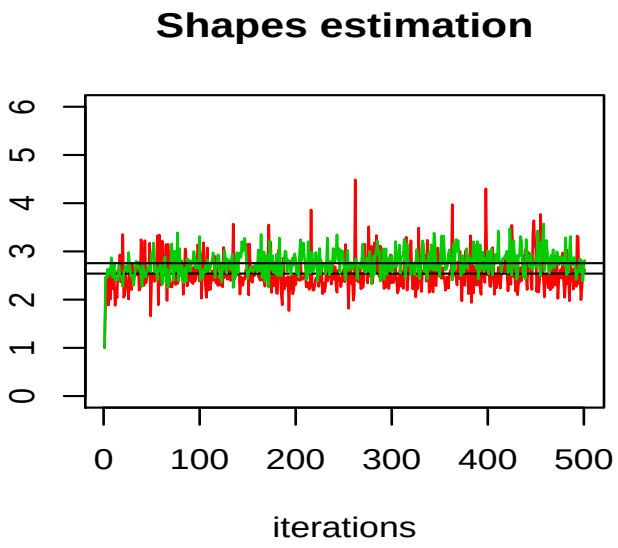


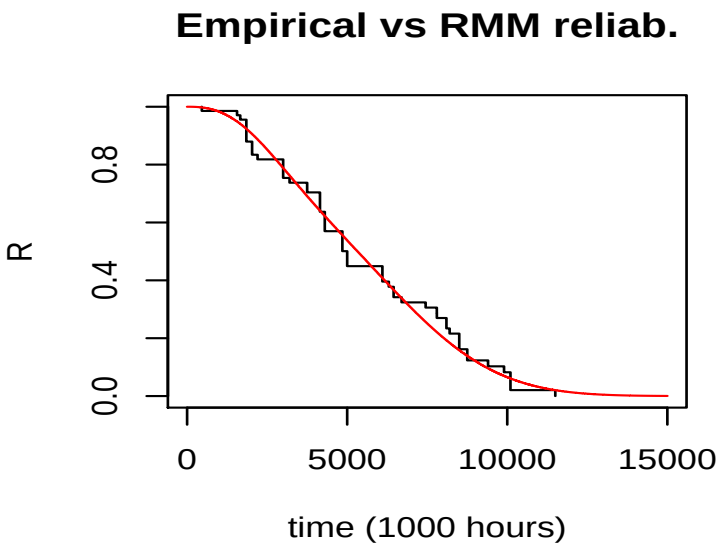


Données ventilateurs (deux lois de Weibull)









Plan

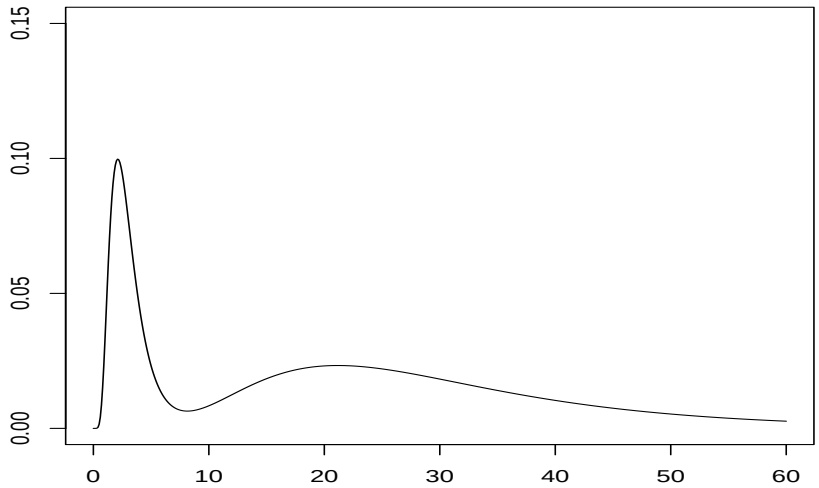
Un modèle de Mélange Semi-Paramétrique (MSP)

$$g(x|\boldsymbol{\theta}) = \lambda_1 f(x) + \lambda_2 \xi f(\xi x) \quad x > 0,$$

où $\boldsymbol{\theta} = (\lambda, \xi, f) \in (0,1) \times \mathbb{R}_*^+ \times \mathcal{F}$; $\mathcal{F}$ est une famille de densités.

Interprétation : modèle de vie accélérée pour données groupées avec deux groupes et indicateur de groupe inconnu.

Exemple : $\lambda_1 = 0.3$, $\xi = 0.1$ et $f \sim \mathcal{LN}(1, 0.5)$.



Identifiabilité de $\boldsymbol{\theta}$

Question : comment choisir $\mathcal{F}$ pour obtenir

$$\left[\forall x > 0 \quad g(x|\boldsymbol{\theta}) = g(x|\boldsymbol{\theta}')\right] \quad \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}'?$$

Question plutôt difficile. . . des réponses partielles dans Bordes, Mottelet et Vandekerkhove (2006) et dans Hunter, Wang and Hettmansperger (2007) : si $\mathcal{F}$ est une famille de densités f vérifiant $x \mapsto e^x f(e^x)$ est paire alors l'identifiabilité a lieu !

Algorithme St-EM pour le MSP [1/4]

Notations : f est la densité inconnue, notons $\bar{F}$ la fonction de fiabilité et $\alpha = f/\bar{F}$ le taux de défaillance.
À partir de $(\boldsymbol{t}, \boldsymbol{d})$:

- $\bar{F}$ est estimée non paramétriquement par Kaplan-Meier,
- α est estimée non paramétriquement en régularisant l'estimateur de Nelson-Aalen.

Comme λ^k, ξ^k, $\bar{F}^k$ et α^k sont des estimations de λ, ξ, $\bar{F}$ et α à l'étape k on a :

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

où la densité f est estimée par $f^k = \alpha^k \bar{F}^k$.

Algorithme St-EM pour le MSP [2/4]

1. **Calcul des probabilités *a posteriori* :** pour chaque individu $i \in \{1, \ldots, n\}$:
   si $d_i = 0$ alors

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

   sinon

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

   Poser $\boldsymbol{p}_i^k = (p_{i1}^k, 1 - p_{i1}^k)$.
2. **Étape Stochastique :** pour chaque individu $i \in \{1, \ldots, n\}$ simuler $Z_i^k \sim \mathcal{M}\mathrm{ult}(1, \boldsymbol{p}_i^k)$. Definir les sous-ensembles

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

Algorithme St-EM pour le MSP [3/4]

Constat : $\xi = \frac{E(X|Z=1)}{E(X|Z=2)}$ et si $S_j(s) = \mathbb{P}(X > s|Z = j)$ alors $E(X|Z = j) = \int_0^{+\infty} S_j(s)ds$.

3. **Mise à jour des paramètres λ et ξ:**

$$\begin{aligned} \lambda^{k+1} &= \frac{\mathrm{Card}(\chi_1^k)}{n}, \\ \xi^{k+1} &= \frac{\int_0^{\tau_1^k} \hat{S}_1^k(s)ds}{\int_0^{\tau_2^k} \hat{S}_2^k(s)ds}, \end{aligned}$$

où $\hat{S}_j^k$ est l'estimateur de Kaplan-Meier pour la sous-population $\{(t_\ell, d_\ell); \ell \in \chi_j^k\}$ et $\tau_j^k = \max_{\ell \in \chi_j^k} t_\ell$.

Algorithme St-EM pour le MSP [4/4]

Constat : si X provient de la composante 2 (i.e. si $Z = 2$), alors $\xi X \sim f$.

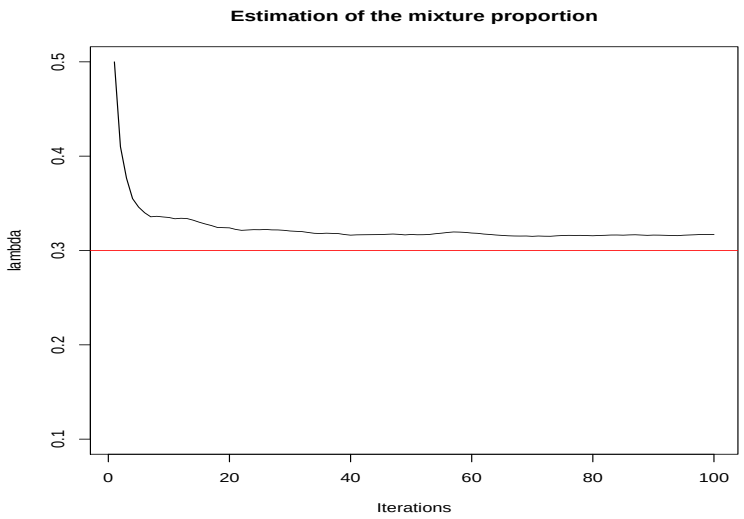
4. **Mise à jour des paramètres fonctionnels α et $\bar{F}$:** noter $\boldsymbol{t}^k = (t_1^k, \ldots, t_n^k)$ la statistique d'ordre associée à $\{t_i; i \in \chi_1^k\} \cup \{\xi^k t_i; i \in \chi_2^k\}$; écrire $\boldsymbol{d}^k = (d_1^k, \ldots, d_n^k)$ les indicateurs de censure correspondants.

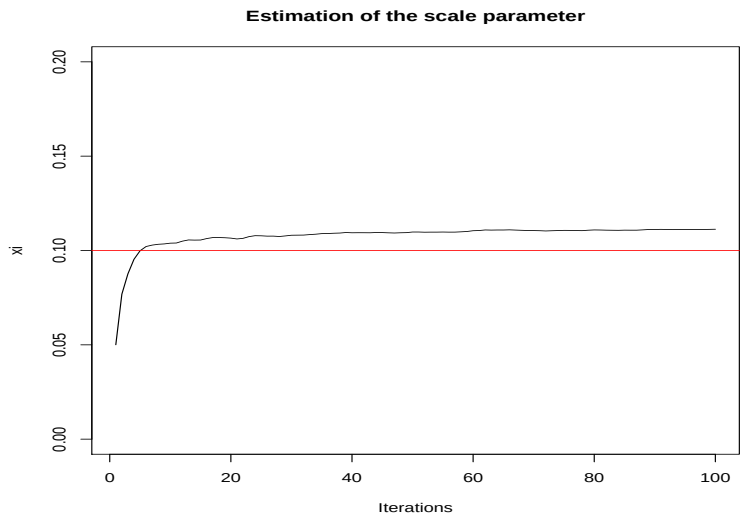
$$\begin{aligned} \alpha^{k+1}(s) &= \sum_{i=1}^{n} \frac{1}{h} \mathcal{K}\left(\frac{s - t_i^k}{h}\right) \frac{d_i^k}{n - i + 1}, \\ \bar{F}^{k+1}(s) &= \prod_{i : t_i^k \leq s} \left(1 - \frac{d_i^k}{n - i + 1}\right), \end{aligned}$$

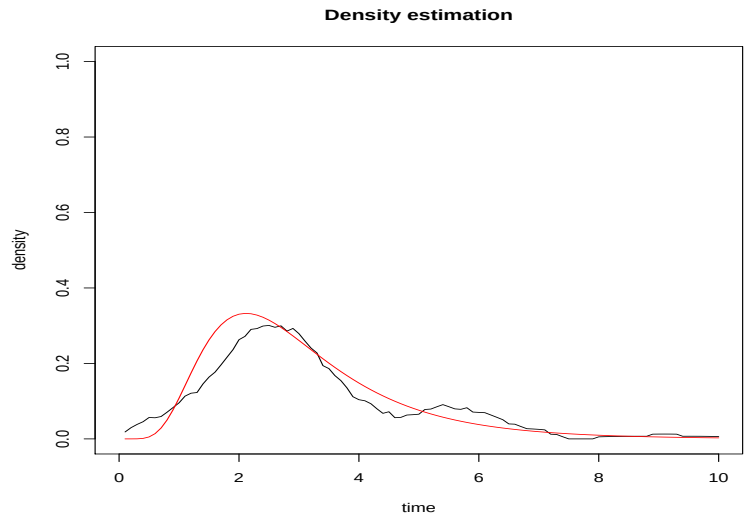
où $\mathcal{K}$ est un noyau et h une largeur de fenêtre.

Remarque : en pratique les choix de $\mathcal{K}$ et h sont importants !

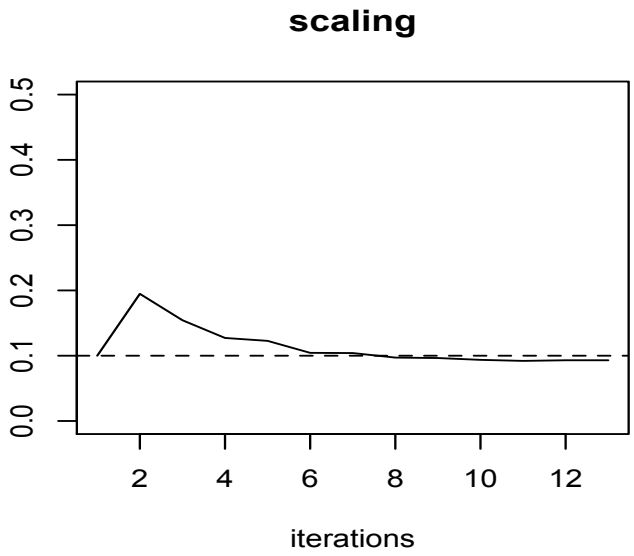
Exemple : $g(x) = 0.3f(x) + 0.7\xi f(\xi x)$, $f \sim \mathcal{LN}(1, 0.5)$.
Échantillon simulé : $n = 100$ avec 0% de censure.

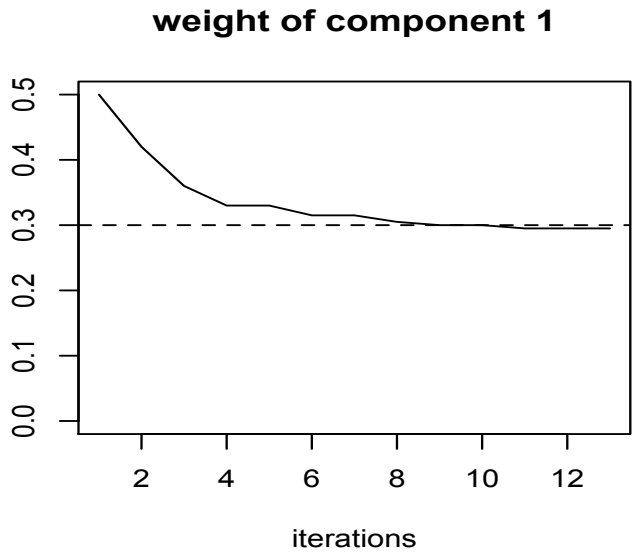


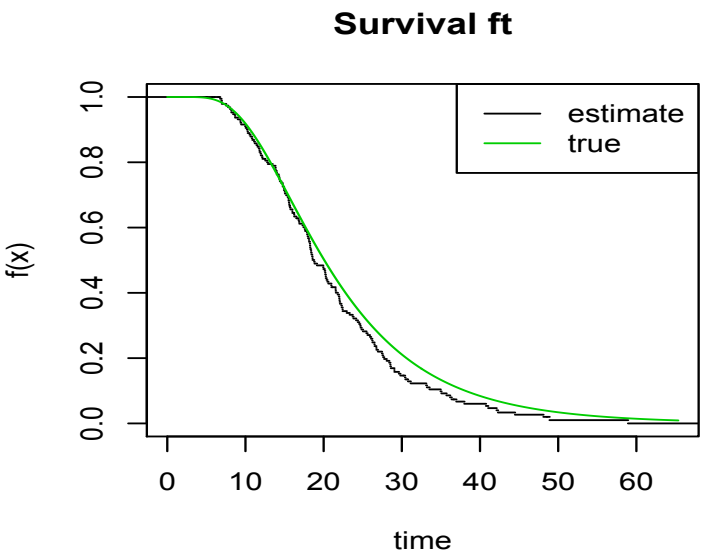


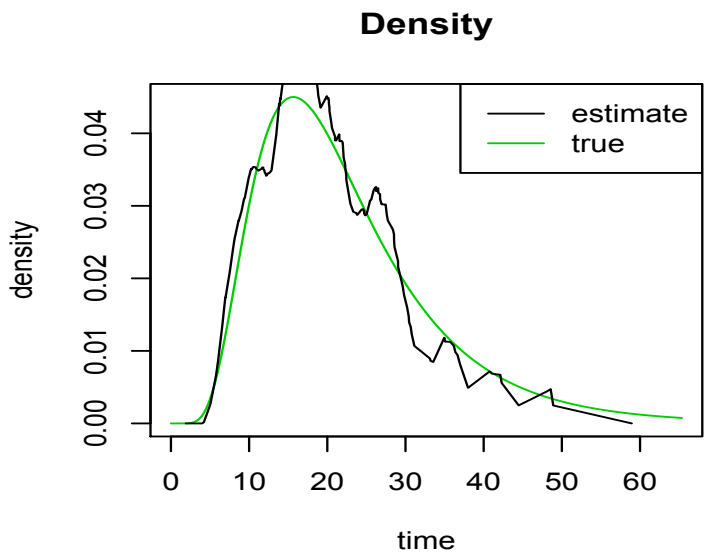


Exemple (suite) : $n = 200$ et 10% de censure.









Conclusions

- Tous les algorithmes qui ont été introduits sont (ou seront) implémentés dans la bibliothèque `mixtools` de Benaglia, Chauveau, Hunter et Young (2009) pour le logiciel statistique `R` (`R` Development Core Team, 2009).
- Les variances asymptotiques issues des estimateurs issus de l'algorithme St-EM peuvent être calculées (voir Nielsen, 2000).
- Le problème de la sélection automatique de la fenêtre en semi-paramétrique n'a pas encore été abordé.
- La méthodologie peut être étendue...

MERCI DE VOTRE ATTENTION !