New mixture models and algorithms in the **mixtools** package

Didier Chauveau

MAPMO - UMR 6628 - Université d’Orléans

1ères Rencontres BoRdeaux Juillet 2012
Outline

1. Mixture models and EM algorithm-ology

2. Quick look at some mixtools’ current algorithms
   - Univariate (semiparametric) mixtures
   - Nonparametric multivariate “EM” algorithms

3. New models and algorithms for mixtools next version
   - Nonlinear smoothed MM for nonparametric mixtures
   - Mixture models for censored lifetime data
   - And more...
Univariate mixture example 1: Old Faithful wait times

Time between Old Faithful eruptions (minutes)

- Obvious bimodality
- Normal-looking components?
- Classification of individuals?

from www.nps.gov/yell

D. Chauveau – RR Bordeaux 2012  New algorithms in mixtools
Univariate mixture example 2: lifetime data

No obvious multimodality but 2 failure sources suspected?
Univariate mixture example 2: lifetime data

No obvious multimodality but 2 failure sources suspected?

True components and mixture densities for these simulated data differences in the tails only!
Finite mixture model and missing data setup

A complete $i$th observation $Y_i = (X_i, Z_i)$ consists of:

- The missing 0/1 latent variable $Z_i = (Z_{i1}, \ldots, Z_{im})$,

$$Z_{ij} = \begin{cases} 1 & \text{if } X_i \text{ comes from component } j \\ 0 & \text{otherwise} \end{cases}, \quad \mathbb{P}(Z_{ij} = 1) = \lambda_j$$

- The observed, incomplete data of interest $X_i$ with $j$th component density $(X_i | Z_{ij} = 1) \sim f_j$

Most models in mixtools share in common a finite mixture pdf

$$g_\theta(x) = \sum_{j=1}^{m} \lambda_j f_j(x)$$

with the model parameters: $\theta = (\lambda, f) = (\lambda_1, \ldots, \lambda_m, f_1, \ldots, f_m)$
Some mixture estimation problems in **mixtools**

**Goal:** Estimate the parameter $\theta$ given an iid sample $x$ from

**Univariate Cases:** $x \in \mathbb{R}$
- some parametric families

\[
g_\theta(x) = \sum_{j=1}^{m} \lambda_j f(x|\phi_j)
\]

- semi-parametric location-shift mixture

\[
g_\theta(x) = \sum_{j=1}^{m} \lambda_j f(x - \mu_j)
\]

- mixture of regressions, …
Some mixture estimation problems in **mixtools**

**Goal:** Estimate the parameter $\theta$ given an iid sample $x$ from

**Univariate Cases:** $x \in \mathbb{R}$
- some parametric families

$$g_\theta(x) = \sum_{j=1}^{m} \lambda_j f(x | \phi_j)$$

- semi-parametric location-shift mixture

$$g_\theta(x) = \sum_{j=1}^{m} \lambda_j f(x - \mu_j)$$

- mixture of regressions, . . .

**Multivariate cases:** $x \in \mathbb{R}^r$
- parametric (Gaussian, . . .)

$$g_\theta(x) = \sum_{j=1}^{m} \lambda_j f(x | \phi_j)$$

- Fully nonparametric, e.g.

$$g_\theta(x) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_k)$$

*Conditional independence of $x_1, \ldots, x_r$ given $z$*
Multivariate example: Water-level data

Example from psychometrics Thomas Lohaus Brainerd (1993).

- Subjects are shown $r = 8$ vessels orientations, presented in this order
- They draw the water surface for each
- Measure = signed angle formed by surface with horizontal
Multivariate example: Water-level data

Example from psychometrics Thomas Lohaus Brainerd (1993).

- Subjects are shown $r = 8$ vessels orientations, presented in this order
- They draw the water surface for each
- Measure = signed angle formed by surface with horizontal

Assume that opposite clock-face orientations lead to same behavior = conditionally iid responses
(1 & 7), (2 & 8), (4 & 10), (5 & 11),
Water-level data: histograms by blocks of iid coord.

Opposite clock-face = conditionally iid responses

1:00 and 7:00 orientations

2:00 and 8:00 orientations

4:00 and 10:00 orientations

5:00 and 11:00 orientations

D. Chauveau – RR Bordeaux 2012

New algorithms in mixtools
General context of missing data: A fraction $x$ of $y$ is observed
MLE’s on observed data ($x$) and complete data ($y$):

$$
\hat{\theta}_x = \arg \max_{\theta \in \Theta} L_x(\theta), \quad \hat{\theta}_y = \arg \max_{\theta \in \Theta} L_y^c(\theta),
$$

where $L_x(\theta) = \sum_{i=1}^n \log g_\theta(x_i)$ and $L_y^c(\theta) = \ldots$
EM Algorithm (1/3)  Dempster Laird Rubin (1977)

General context of missing data: A fraction $x$ of $y$ is observed
MLE’s on observed data ($x$) and complete data ($y$):

$$
\hat{\theta}_x = \arg \max_{\theta \in \Theta} L_x(\theta), \quad \hat{\theta}_y = \arg \max_{\theta \in \Theta} L_y^c(\theta),
$$

where $L_x(\theta) = \sum_{i=1}^n \log g_\theta(x_i)$ and $L_y^c(\theta) = \ldots$

**Intuition:** often the MLE on the complete data $\hat{\theta}_y$ is available, while $\hat{\theta}_x$ is not

→ Try to maximize instead of $L_y^c(\theta)$, its expectation given $x$

$$
Q(\theta|\theta') := \mathbb{E} [L_y^c(\theta)| x; \theta']
$$

for “some” $\theta'$ ...
EM Algorithm (2/3): general principle

\[ \theta^0 = \text{some "arbitrary" initialization}, \quad \theta^t = \text{current value:} \]

\[ \text{EM } \theta^t \rightarrow \theta^{t+1} \text{ iteration} \]

**Expectation step:** compute \( \theta \mapsto Q(\theta|\theta^t) \)

**Maximization step:** set \( \theta^{t+1} = \arg \max_{\theta \in \Theta} Q(\theta|\theta^t) \).

Why does it work under mild conditions?

**Ascent property for the observed loglikelihood**

\[ L_x(\theta^{t+1}) \geq L_x(\theta^t) \]

alternatives: GEM, ECM, MM, Stochastic-EM,…
EM Algorithm (3/3): parametric mixture $f_j = f(x|\phi_j)$

**E-step:** Amounts to find the **posterior probabilities**

\[
Z_{ij}^t := \mathbb{E}_{\theta^t}[Z_{ij}|x_i] = \mathbb{P}_{\theta^t}[Z_{ij} = 1|x_i] = \frac{\lambda_j^t f(x_i|\phi_j^t)}{\sum_{j'} \lambda_{j'}^t f(x_i|\phi_{j'}^t)}
\]

**M-step:** Maximization looks like weighted MLE

\[
\lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^{n} Z_{ij}^t \quad \text{generic for EM algorithms for mixtures}
\]

and for e.g., the Gaussian $f(x|\phi_j) = \text{the pdf of } \mathcal{N}(\mu_j, \nu_j)$,

\[
\mu_j^{t+1} = \frac{\sum_{i=1}^{n} Z_{ij}^t x_i}{\sum_{i=1}^{n} Z_{ij}^t}, \quad \nu_j^{t+1} = \frac{\sum_{i=1}^{n} Z_{ij}^t (x_i - \mu_j^{t+1})^2}{\sum_{i=1}^{n} Z_{ij}^t}
\]
About Stochastic EM versions

In some (mixture) setup, it may be useful to simulate the latent data from the posterior probabilities:

\[ \hat{Z}_i^t \sim Mult \left( 1 ; Z_{i1}^t, \ldots, Z_{im}^t \right), \quad i = 1, \ldots, n \]

Then:

- The “complete” data \((x, \hat{Z}^t)\) allows direct computation of the MLE
- the sequence \((\theta^t)_{t \geq 1}\) becomes a **Markov Chain**
- Historically, parametric Stochastic EM introduced by Celeux Diebolt (1985, 1986, . . .)
- General convergence properties: Nielsen 2000
About Stochastic EM versions

In some (mixture) setup, it may be useful to simulate the latent data from the posterior probabilities:

\[ \hat{Z}_i^t \sim \text{Mult}(1; Z_{i1}^t, \ldots, Z_{im}^t), \quad i = 1, \ldots, n \]

Then:

- The “complete” data \((x, \hat{Z}^t)\) allows direct computation of the MLE
- The sequence \((\theta^t)_{t \geq 1}\) becomes a **Markov Chain**
- Historically, parametric Stochastic EM introduced by Celeux Diebolt (1985, 1986, . . .)
- General convergence properties: Nielsen 2000
- In non-parametric framework: Stochastic EM for reliability mixture models, Bordes C (2010, 2012) more on this later
Outline: Next up...

1. Mixture models and EM algorithm-ology

2. Quick look at some mixtools' current algorithms
   - Univariate (semiparametric) mixtures
   - Nonparametric multivariate “EM” algorithms

3. New models and algorithms for mixtools next version
   - Nonlinear smoothed MM for nonparametric mixtures
   - Mixture models for censored lifetime data
   - And more...
In R with `mixtools`, type

```R
R> data(faithful)
R> attach(faithful)
R> normalmixEM(waiting, mu=c(55,80), sigma=5)
```

number of iterations = 24
Old Faithful data with parametric Gaussian EM

In R with `mixtools`, type

```r
R> data(faithful)
R> attach(faithful)
R> normalmixEM(waiting, mu=c(55,80), sigma=5)
```

number of iterations = 24

- Gaussian EM result: \( \hat{\mu} = (54.6, 80.1) \)
One originality of **mixtools**: Its capability to handle various nonparametric components $f_j$'s in $g_\theta(x) = \sum_{j=1}^{m} \lambda_j f_j(x)$, for several models

Univariate identifiable semiparametric mixtures

One originality of **mixtools**: 
Its capability to handle various nonparametric components $f_j$’s in $g_\theta(x) = \sum_{j=1}^{m} \lambda_j f_j(x)$, for several models

**Location-shift semi-parametric mixture model:**

$$g_\theta(x) = \lambda_1 f(x - \mu_1) + (1 - \lambda_1) f(x - \mu_2)$$

This model is **identifiable** when $f(\cdot)$ is **even**. Bordes Mottelet Vandekerkhove (2006), Hunter Wang Hettmansperger (2007)

mixtools’ Semi-parametric “EM” algorithm

**E-step:** Same as usual:

\[ Z_{ij}^t \equiv \mathbb{E}_{\theta^t}[Z_{ij}|x_i] = \frac{\lambda_j^t f^t(x_i - \mu_j^t)}{\lambda_1^t f^t(x_i - \mu_1^t) + \lambda_2^t f^t(x_i - \mu_2^t)} \]

**M-step:** Maximize complete data “loglikelihood” for \( \lambda \) and \( \mu \):

\[
\lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^{n} Z_{ij}^t \\
\mu_j^{t+1} = (n\lambda_j^{t+1})^{-1} \sum_{i=1}^{n} Z_{ij}^t x_i
\]

**Weighted KDE-step:** Update \( f^t \) (for some bandwidth \( h \)) by

\[
f^{t+1}(u) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j=1}^{2} Z_{ij}^t K\left( \frac{u - (x_i - \mu_j^{t+1})}{h} \right), \quad \text{then symmetrize.}
\]
Location-shift semi-parametric mixture model

Time between Old Faithful eruptions (minutes)

\[ \lambda_1 = 0.361 \]

Gaussian EM:
\[ \hat{\mu} = (54.6, 80.1) \]
Location-shift semi-parametric mixture model

\[ g_\theta(x) = \lambda_1 f(x - \mu_1) + (1 - \lambda_1) f(x - \mu_2) \]

Time between Old Faithful eruptions (minutes)

- Gaussian EM:
  \[ \hat{\mu} = (54.6, 80.1) \]

- Semiparametric EM
  \[ R> \text{spEMsymloc(waiting, } \mu=c(55,80), \ h=4) \ # \text{ bandwidth} \]
  \[ \hat{\mu} = (54.7, 79.8) \]

\[ \lambda_1 = 0.361 \]
\[ \lambda_1 = 0.353 \]
Outline: Next up...

1. Mixture models and EM algorithm-ology
2. Quick look at some mixtools’ current algorithms
   - Univariate (semiparametric) mixtures
   - Nonparametric multivariate “EM” algorithms
3. New models and algorithms for mixtools next version
   - Nonlinear smoothed MM for nonparametric mixtures
   - Mixture models for censored lifetime data
   - And more...
Identifiability: the blessing of dimensionality (!)

Recall the model in the **multivariate case**, \( r > 1 \):

\[
g_{\theta}(x) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_k)
\]

**N.B.:** Assume conditional independence of \( x_1, \ldots, x_r \)

- **Hall Zhou (2003)** show that when \( m = 2 \) and \( r \geq 3 \), the model is identifiable under mild restrictions on the \( f_{jk}(\cdot) \)
- **Hall et al. (2005)** ... *from at least one point of view, the ‘curse of dimensionality’ works in reverse.*
- **Allman et al. (2008)** give mild sufficient conditions for identifiability whenever \( r \geq 3 \)
The notation gets even worse. . .

**Motivation:**
the Water-level data with 4 “blocks” of 2 similar responses

- Let the \( r \) coordinates be grouped into \( B \leq r \) blocks of conditionally iid coordinates.
  \( b_k \in \{1, \ldots, B\} \) is the block index of the \( k \)th coordinate

- The model becomes

\[
g(x) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{j b_k}(x_k)
\]

**Special cases:**
- \( b_k = k \) for \( k = 1, \ldots, r \): general model of conditional independence as in (Hall et al. 2005...)
- \( b_k = 1 \) for all \( k \): Conditionally i.i.d. assumption (Elmore et al. 2004)

E-step: Same as for a genuine parametric EM,

\[ Z_{ij}^t = \frac{\lambda_j^t \prod_{k=1}^r f_{j,b_k}^t(x_{ik})}{\sum_{j'} \lambda_{j'}^t \prod_{k=1}^r f_{j',b_k}^t(x_{ik})} \]

M-step: Maximize complete data “loglikelihood” for \( \lambda \):

\[ \lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^n Z_{ij}^t \]

WKDE-step: Update estimate of \( f_{j,\ell} \) (component \( j \), block \( \ell \)) by

\[ f_{j,\ell}^{t+1}(u) = \frac{1}{nh_{j,\ell}^{t+1} C_\ell \lambda_j^{t+1}} \sum_{k=1}^r \sum_{i=1}^n Z_{ij}^t \mathbb{1}_{\{b_k=\ell\}} \mathcal{K} \left( \frac{u - x_{ik}}{h_{j,\ell}^{t+1}} \right) \]

where \( C_\ell = \# \) of coordinates in block \( \ell \),
\( h_{j,\ell}^{t+1} = \) Iterative and per component & block adaptive bandwidth
The Water-level data

Dataset previously analysed with conditional i.i.d. assumption. Hettmansperger Thomas (2000), Elmore et al. (2004)
The non appropriate conditional i.i.d. assumption masks interesting features that our model reveals

In \texttt{mixtools}: npEM algorithm

R> library(mixtools)
R> data(Waterdata)
R> a <- npEM(Waterdata, 3,
   blockid=c(1,2,3,4,2,1,4,3),
   h=4) # user-fixed bandwidth here
R> par(mfrow=c(2,2))
R> plot(a)
R> summary(a) # some statistics per components
...
The Water-level data, $m = 3$ components, 4 blocks

Block 1: 1:00 and 7:00 orientations

Appearance of Vessel at Orientation = 1:00

Mixing Proportion (Mean, Std Dev)
- Blue: 0.077 (−32.1, 19.4)
- Green: 0.431 (−3.9, 23.3)
- Red: 0.492 (−1.4, 6.0)

Block 2: 2:00 and 8:00 orientations

Appearance of Vessel at Orientation = 2:00

Mixing Proportion (Mean, Std Dev)
- Blue: 0.077 (−31.4, 55.4)
- Green: 0.431 (−11.7, 27.0)
- Red: 0.492 (−2.7, 4.6)

Block 3: 4:00 and 10:00 orientations

Appearance of Vessel at Orientation = 4:00

Mixing Proportion (Mean, Std Dev)
- Blue: 0.077 (43.6, 39.7)
- Green: 0.431 (11.4, 27.5)
- Red: 0.492 (1.0, 5.3)

Block 4: 5:00 and 11:00 orientations

Appearance of Vessel at Orientation = 5:00

Mixing Proportion (Mean, Std Dev)
- Blue: 0.077 (27.5, 19.3)
- Green: 0.431 (2.0, 22.1)
- Red: 0.492 (−0.1, 6.1)
Outline: Next up...

1. Mixture models and EM algorithm-ology

2. Quick look at some mixtools’ current algorithms
   - Univariate (semiparametric) mixtures
   - Nonparametric multivariate “EM” algorithms

3. New models and algorithms for mixtools next version
   - Nonlinear smoothed MM for nonparametric mixtures
   - Mixture models for censored lifetime data
   - And more...
From npEM to NEMS for nonparametric mixtures

**Motivation:** Our npEM algorithms are not true EM

**Idea:** combine regularization and npEM approach to define an algorithm with a **provable ascent property**

Levine Hunter C *(Biometrika 2011)*

“Nonparametric” in smooth EM literature refers to a continuous mixing distribution

- true EM but ill-posed difficulties, Vardi et al. (1985)
- Smoothed EM (EMS), Silverman et al. (1990)
- Nonlinear EMS (NEMS) regularization approach Eggermont, LaRiccia (1995); Eggermont (1992, 1999)
Smoothing the mixture

Applying Eggermont (1992, 1999) idea to component pdf’s

- Nonlinear smoothing

\[ \mathcal{N}f(x) = \exp \int_{\Omega} K_h(x - u) \log f(u) \, du, \]

where \( K_h(u) = h^{-r} \prod_{k=1}^{r} K(h^{-1}u_k) \) is a product kernel

D. Chauveau – RR Bordeaux 2012
Smoothing the mixture

Applying Eggermont (1992, 1999) idea to component pdf’s

- Nonlinear smoothing

\[
\mathcal{N} f(x) = \exp \int_{\Omega} K_h(x - u) \log f(u) \, du,
\]

where \( K_h(u) = h^{-r} \prod_{k=1}^{r} K(h^{-1}u_k) \) is a product kernel

- For \( \mathbf{f} = (f_1, \ldots, f_m) \), define

\[
\mathcal{M}_\lambda \mathcal{N} \mathbf{f}(x) := \sum_{j=1}^{m} \lambda_j \mathcal{N} f_j(x)
\]

Goal: minimizing the \( \infty \)-sample objective function (Kullback)

\[
\ell(\theta) = \ell(\mathbf{f}, \lambda) := \int_{\Omega} g(x) \log \frac{g(x)}{[\mathcal{M}_\lambda \mathcal{N} \mathbf{f}](x)} \, dx.
\]
Majorization-Minimization (MM) trick

**Principle:** Let $\theta^t$ be the current value of $\theta$ in an algorithm.

A function $B(\theta|\theta^t)$ is said to majorize $\ell(\theta)$ at $\theta^t$ provided

$$B(\theta|\theta^t) \geq \ell(\theta) \text{ for all } \theta$$

$$B(\theta^t|\theta^t) = \ell(\theta^t)$$

MM minimization algorithm: set $\theta^{t+1} = \arg \min_{\theta} B(\theta|\theta^t)$

The MM algorithm satisfies the descent property

$$\ell(\theta^{t+1}) \leq \ell(\theta^t)$$

and we can define such a majorizing function $B(\theta|\theta^t)$ here!
MM algorithm with a descent property

Smoothed “log-likelihood” version for finite sample-size given the sample $x_1, \ldots, x_n$ iid $\sim g$

$$\ell_n(f, \lambda) := - \sum_{i=1}^{n} \log[M_X N_f](x_i)$$

The following MM algorithm satisfies a descent property

$$\ell_n(f^{t+1}, \lambda^{t+1}) \leq \ell_n(f^t, \lambda^t)$$
nonparametric Maximum Smoothed Likelihood (npMSL) algorithm

**E-step:** (requires additional univariate numerical integrations)

\[ w_{ij}^t = \frac{\lambda_j^t N f_j^t(x_i)}{M \lambda_t N f_t^t(x_i)} = \frac{\lambda_j^t \prod_{k=1}^r N f_{jk}^t(x_{ik})}{\sum_{j'=1}^m \lambda_{j'} \prod_{k=1}^r N f_{j'k}^t(x_{ik})} \]

**M-step:** for \( j = 1, \ldots, m \)

\[ \lambda_{j}^{t+1} = \frac{1}{n} \sum_{i=1}^n w_{ij}^t \]

**WKDE-step:** For each component \( j \) and coord. \( k \) (or blocks),

\[ f_{jk}^{t+1}(u) = \frac{1}{nh \lambda_j^{t+1}} \sum_{i=1}^n w_{ij}^t K \left( \frac{u - x_{ik}}{h} \right) \]
The Water-level data again, npEM vs. npMSL (dotted)

npMSL in *mixtools*:

R> b <- npMSL(Waterdata, 3, blockid=c(1,2,3,4,2,1,4,3), h=4) # bandwidth

R> plot(b)
Further extensions: Semiparametric models

Component or block density may differ only in location and/or scale parameters, e.g.

\[ f_{j\ell}(x) = \frac{1}{\sigma_{j\ell}} f_j \left( \frac{x - \mu_{j\ell}}{\sigma_{j\ell}} \right) \]

or

\[ f_{j\ell}(x) = \frac{1}{\sigma_{j\ell}} f_{\ell} \left( \frac{x - \mu_{j\ell}}{\sigma_{j\ell}} \right) \]

or

\[ f_{j\ell}(x) = \frac{1}{\sigma_{j\ell}} f \left( \frac{x - \mu_{j\ell}}{\sigma_{j\ell}} \right) \]

where \( f_j, f_{\ell}, f \) remain fully unspecified

For all these situations special cases of the npEM/npMSL algorithm can be designed (some are already in \texttt{mixtools}).
Outline: Next up...

1. Mixture models and EM algorithm-ology

2. Quick look at some mixtools’ current algorithms
   - Univariate (semiparametric) mixtures
   - Nonparametric multivariate “EM” algorithms

3. New models and algorithms for mixtools next version
   - Nonlinear smoothed MM for nonparametric mixtures
   - Mixture models for censored lifetime data
   - And more...
Bordes C (2010, 2012)
Typical data from reliability and life testing + mixture model:

- lifetime data on $\mathbb{R}^+$ $(X_1, \ldots, X_n)$ iid $\sim g_\theta = \sum_j \lambda_j f_j(x)$
- censoring times $(C_1, \ldots, C_n)$ iid $\sim q$ (unknown)
- Random (right) censoring $T_i = \min(X_i, C_i), D_i = \mathbb{I}\{X_i \leq C_i\}$
- Observed (incomplete) data $(t, d) = ((t_i, d_i), i = 1, \ldots n)$
Mixture models for censored lifetime data

Bordes C (2010, 2012)  
Typical data from reliability and life testing + mixture model:

- lifetime data on $\mathbb{R}^+$ $(X_1, \ldots, X_n)$ iid $\sim g_\theta = \sum_j \lambda_j f_j(x)$
- censoring times $(C_1, \ldots, C_n)$ iid $\sim q$ (unknown)
- Random (right) censoring $T_i = \min(X_i, C_i)$, $D_i = I\{X_i \leq C_i\}$
- Observed (incomplete) data $(t, d) = (((t_i, d_i), i = 1, \ldots n)$

A semiparametric example:

**Semiparametric accelerated lifetime mixture model**

$$g_\theta(x) = \lambda f(x) + (1 - \lambda) \xi f(\xi x)$$

Scaling model: $(X|Z_1 = 1)$ has distribution $U \sim f$, and $(X|Z_2 = 1) \sim U/\xi$, identifiable under some restrictions
Semi-parametric Stochastic-EM for censored data (1)

Denote $F$ the cdf of $f$, $S = 1 - F$ the Survival function and $\alpha = f / S$ the hazard rate

Building blocks:

- For a single censored sample $(t, d)$, $S$ and $\alpha$ can be estimated nonparametrically using Kaplan-Meier and Nelson-Aalen estimators.
Denote $F$ the cdf of $f$, $S = 1 - F$ the Survival function and $\alpha = f/S$ the hazard rate.

**Building blocks:**

- For a single censored sample $(t, d)$, $S$ and $\alpha$ can be estimated nonparametrically using Kaplan-Meier and Nelson-Aalen estimators.
- $E(X|Z_j = 1) = \int_0^{+\infty} S_j(s) ds$, where $S_j(s) = P(X > s|Z_j = 1)$ is the $j$th component survival.
Semi-parametric Stochastic-EM for censored data (1)

Denote $F$ the cdf of $f$, $S = 1 - F$ the Survival function and $\alpha = f/S$ the hazard rate

**Building blocks:**

- For a single censored sample $(t, d)$, $S$ and $\alpha$ can be estimated nonparametrically using Kaplan-Meier and Nelson-Aalen estimators.

\[ E(X|Z_j = 1) = \int_{0}^{\infty} S_j(s) ds, \]

where $S_j(s) = P(X > s | Z_j = 1)$ is the $j$th component survival

- In this scaling model $\xi = \frac{E(X|Z_1=1)}{E(X|Z_2=1)}$
Semi-parametric Stochastic-EM for censored data (1)

Denote $F$ the cdf of $f$, $S = 1 - F$ the Survival function and $\alpha = f/S$ the hazard rate

Building blocks:

- For a single censored sample $(t, d)$, $S$ and $\alpha$ can be estimated nonparametrically using Kaplan-Meier and Nelson-Aalen estimators.
- $E(X | Z_j = 1) = \int_0^{+\infty} S_j(s)ds$, where $S_j(s) = P(X > s | Z_j = 1)$ is the $j$th component survival.
- In this scaling model $\xi = \frac{E(X | Z_1 = 1)}{E(X | Z_2 = 1)}$.
- If $X$ comes from component 2, then $\xi X \sim f$, so that $\{X_i : Z_{i1} = 1\} \cup \{\xi X_i : Z_{i2} = 1\}$ iid $\sim f$. 
Semi-parametric Stochastic-EM for censored data (1)

Denote $F$ the cdf of $f$, $S = 1 - F$ the Survival function and $\alpha = f / S$ the hazard rate

**Building blocks:**

- For a single censored sample $(t, d)$, $S$ and $\alpha$ can be estimated nonparametrically using Kaplan-Meier and Nelson-Aalen estimators.
- $E(X|Z_j = 1) = \int_0^{+\infty} S_j(s)ds$, where $S_j(s) = \mathbb{P}(X > s|Z_j = 1)$ is the $j$th component survival.
- In this scaling model $\xi = \frac{E(X|Z_1 = 1)}{E(X|Z_2 = 1)}$
- If $X$ comes from component 2, then $\xi X \sim f$, so that $\{X_i : Z_{i1} = 1\} \cup \{\xi X_i : Z_{i2} = 1\}$ iid $\sim f$

→ Estimates $\hat{\xi}$ and $\hat{f} = \hat{\alpha} \hat{S}$ available in mixture setup if $Z$ is observed

D. Chauveau – RR Bordeaux 2012

New algorithms in mixtools
Semi-parametric “St-EM” for censored data (2)

Stochastic version required here!

**E-step:** if \( d_i = 0 \) (censored lifetime)

\[
Z_{i1}^t = \frac{\lambda^t S^t(t_i)}{\lambda^t S^t(t_i) + (1 - \lambda^t)S^t(\xi^t t_i)}
\]

else \((d_i = 1\) observed lifetime)

\[
Z_{i1}^t = \frac{\lambda^t \alpha^t(t_i) S^t(t_i)}{\lambda^t \alpha^t(t_i) S^t(t_i) + (1 - \lambda^t)\xi^t \alpha^t(\xi^t t_i) S^k(\xi^k t_i)}
\]

**S-step:** Simulate \( \widehat{Z}_{i1}^t \sim B(p_{i1}^t) \), \( i = 1, \ldots, n \), and set

\[
\chi_j^t = \{ i \in \{1, \ldots, n\} : \widehat{Z}_{i1}^t = 1 \}, \quad j = 1, 2
\]
Semi-parametric “St-EM” for censored data (3)

M-step for $\lambda, \xi$:

$$
\lambda^{t+1} = \frac{\text{Card}(\chi_1^t)}{n}
, \quad 
\xi^{t+1} = \frac{\int_0^{\tau_1^t} \hat{S}_1^t(s) ds}{\int_0^{\tau_2^t} \hat{S}_2^t(s) ds}
, \quad 
\tau_j^t = \max_{i \in \chi_j^t} t_i
$$

where $\hat{S}_j^t$ is the Kaplan-Meier estimate based on $\{(t_i, d_i); i \in \chi_j^t\}$

Nonparametric-step for $\alpha, S$: Let $t^t$ be the order statistic of 
$\{t_i; i \in \chi_1^t\} \cup \{\xi^t t_i; i \in \chi_2^t\}$ and $d^t$ the associated censoring indicators

$$
\alpha^{t+1}(s) = \sum_{i=1}^n \frac{1}{n} \frac{d_i^t}{h} \left( \frac{s - t_i^t}{h} \right) 
$$

$$
S^{t+1}(s) = \prod_{i: t_i^t \leq s} \left( 1 - \frac{d_i^t}{n - i + 1} \right)
$$
Simulated example: scale mixture of Lognormals

\( n = 300, \) 15% censored

\[
\begin{align*}
\text{Scaling} & \\
\text{Weight of component 1} & \\
\text{Survival function} & \\
\text{Density} &
\end{align*}
\]

```
R> library(mixtools)
R> library(survival) # for KM estimation
# simulating data...
R> s <- spRMMSM(t, d, scaling=0.1)
R> plot(s)
```
Outline: Next up . . .

1. Mixture models and EM algorithm-ology

2. Quick look at some mixtools’ current algorithms
   - Univariate (semiparametric) mixtures
   - Nonparametric multivariate “EM” algorithms

3. New models and algorithms for mixtools next version
   - Nonlinear smoothed MM for nonparametric mixtures
   - Mixture models for censored lifetime data
   - And more . . .
Other new models/algorithms available (1)

- Parametric (Exponential, Weibull, lognormal...) reliability mixture models for censored data
Other new models/algorithms available (1)

- Parametric (Exponential, Weibull, lognormal...) reliability mixture models for censored data
- Gaussian mixtures with linear constraints on the parameters, e.g. for $p \leq m$, known $M, C$,

$$
\mu = \begin{pmatrix} 
\mu_1 \\
\vdots \\
\mu_m 
\end{pmatrix} = M\beta + C = M \begin{pmatrix} 
\beta_1 \\
\vdots \\
\beta_p 
\end{pmatrix} + \begin{pmatrix} 
C_1 \\
\vdots \\
C_m 
\end{pmatrix}
$$

and similar constraints on variances

\[ \rightarrow \text{requires ECM and/or MM algorithms C Hunter (2011)} \]
Other new models/algorithms available (1)

- Parametric (Exponential, Weibull, lognormal...) reliability mixture models for censored data
- Gaussian mixtures with linear constraints on the parameters, e.g. for $p \leq m$, known $M, C$,

\[
\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix} = M\beta + C = M \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} C_1 \\ \vdots \\ C_m \end{pmatrix}
\]

and similar constraints on variances

→ requires ECM and/or MM algorithms C Hunter (2011)

- Advances in mixtures of regression: predictor-dependent mixing proportions Young, Hunter (CSDA 2010)
Other new models (2): Mixtures in FDR estimation

In multiple testing (e.g., micro-arrays) we consider

- $n$ multiple tests $\rightarrow p$-values $p = (p_1, \ldots, p_n)$
- question: False Discovery Rate (FDR) $= \text{expected proportion of falsely rejected } H_0$’s
Other new models (2): Mixtures in FDR estimation

In multiple testing (e.g., micro-arrays) we consider

- $n$ multiple tests $\rightarrow p$-values $\mathbf{p} = (p_1, \ldots, p_n)$
- question: False Discovery Rate (FDR) = expected proportion of falsely rejected $H_0$’s
- mixture modelling: latent variable $Z_i = 1$ if $H_0$ is true and $Z_i = 0$ if $H_0$ is rejected (= interesting case)

$$g_\theta(p) = \lambda_0 f_0(p) + (1 - \lambda_0)f_1(p)$$

- theoretically $(p_i|H_0$ true) $\sim \mathcal{U}_{[0,1]} \equiv f_0$ known!
In multiple testing (e.g., micro-arrays) we consider

- $n$ multiple tests $\rightarrow p$-values $\mathbf{p} = (p_1, \ldots, p_n)$
- question: False Discovery Rate (FDR) $= \text{expected proportion of falsely rejected } H_0$’s
- mixture modelling: latent variable $Z_i = 1$ if $H_0$ is true and $Z_i = 0$ if $H_0$ is rejected (= interesting case)

$$g_\theta(p) = \lambda_0 f_0(p) + (1 - \lambda_0) f_1(p)$$

- theoretically ($p_i | H_0 \text{ true} \sim U_{[0,1]} \equiv f_0$ known!
- nonparametric assumption for $f_1$ as e.g. in Robin et. al. 2007 or the fdrtool package Strimmer 2008
Other new models (2): Mixtures in FDR estimation

In multiple testing (e.g., micro-arrays) we consider

- *n* multiple tests $\rightarrow$ *p*-values $p = (p_1, \ldots, p_n)$
- question: False Discovery Rate (FDR) = expected proportion of falsely rejected $H_0$'s
- **mixture modelling**: latent variable $Z_i = 1$ if $H_0$ is true and $Z_i = 0$ if $H_0$ is rejected (= interesting case)

$$g_\theta(p) = \lambda_0 f_0(p) + (1 - \lambda_0) f_1(p)$$

- theoretically $(p_i | H_0 \text{ true}) \sim \mathcal{U}_{[0,1]} \equiv f_0 \text{ known!}$
- **nonparametric assumption** for $f_1$ as e.g. in Robin et. al. 2007 or the **fdrtool** package Strimmer 2008

In **mixtools**: a semiparametric EM with one component known, for FDR estimation C Saby (2011, 2012)
MERCI DE VOTRE ATTENTION!

ET

useR!
For Further Reading (1/3)


Bordes, L. and Chauveau D.
EM and Stochastic EM algorithms for reliability mixture models under random censoring.
*Preprint HAL, 2012*

Levine, M., Hunter, D. R., Chauveau, D.
Maximum Smoothed Likelihood for Multivariate Mixtures.

E. S. Allman, C. Matias, and J. A. Rhodes.
Identifiability of parameters in latent structure models with many observed variables.
S. F. Nielsen. 
The stochastic EM algorithm: Estimation and asymptotic results. 

D. Young and D. Hunter. 
Mixtures of regressions with predictor-dependent mixing proportions. 