EM-like algorithms for nonparametric estimation in multivariate mixtures

Didier Chauveau

MAPMO - UMR 6628 - Université d’Orléans

Joint work with D. Hunter & T. Benaglia (Penn State University, USA)

COMPSTAT 2010 – Paris, August 24th 2010

Outline

Outline: Next up. . .

1. Mixture models and EM algorithms
   - Motivations, examples and notation
   - Review of EM algorithm-ology
2. Multivariate non-parametric "npEM" algorithms
   - Model and algorithm
   - Examples
   - Adaptive bandwidths in the npEM algorithm
3. Further extensions

Outline: Next up. . .

Finite mixture estimation problem

Multivariate observation $\mathbf{x} = (x_1, \ldots, x_r) \in \mathbb{R}^r$ from the mixture

$$g(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j \, \mathbf{f}_j(\mathbf{x})$$

Finite mixture estimation problem

Multivariate observation $\mathbf{x} = (x_1, \ldots, x_r) \in \mathbb{R}^r$ from the mixture

$$g(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j \, \mathbf{f}_j(\mathbf{x})$$

Assume independence of $x_1, \ldots, x_r$ conditional of the component from which $\mathbf{x}$ comes (Hall and Zhou 2003,...):

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

i.e. the dependence is induced by the mixture.

Finite mixture estimation problem

Multivariate observation $\mathbf{x} = (x_1, \ldots, x_r) \in \mathbb{R}^r$ from the mixture

$$g(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j \, \mathbf{f}_j(\mathbf{x})$$

Assume independence of $x_1, \ldots, x_r$ conditional of the component from which $\mathbf{x}$ comes (Hall and Zhou 2003,...):

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

i.e. the dependence is induced by the mixture.

Goal: Estimate $\theta = (\boldsymbol{\lambda}, \mathbf{f})$ given an i.i.d. sample from g

Nonparametric mixture model

In parametric case $\mathbf{f}_j(\cdot) \equiv f(\cdot; \phi_j) \in \mathcal{F}$, a *parametric family* indexed by a parameter $\phi \in \mathbb{R}^d$

The parameter of the mixture model is

$$\boldsymbol{\theta} = (\boldsymbol{\lambda}, \phi) = (\lambda_1, \ldots, \lambda_m, \phi_1, \ldots, \phi_m)$$

Usual example: the univariate Gaussian mixture model, $f(x; \phi_j) = f\left(x; (\mu_j, \sigma_j^2)\right) =$ the pdf of $\mathcal{N}(\mu_j, \sigma_j^2)$.

Nonparametric mixture model

In parametric case $\mathbf{f}_j(\cdot) \equiv f(\cdot; \phi_j) \in \mathcal{F}$, a *parametric family* indexed by a parameter $\phi \in \mathbb{R}^d$

The parameter of the mixture model is

$$\boldsymbol{\theta} = (\boldsymbol{\lambda}, \phi) = (\lambda_1, \ldots, \lambda_m, \phi_1, \ldots, \phi_m)$$

Usual example: the univariate Gaussian mixture model, $f(x; \phi_j) = f\left(x; (\mu_j, \sigma_j^2)\right) =$ the pdf of $\mathcal{N}(\mu_j, \sigma_j^2)$.

Motivations here:

Do not assume any parametric form for the f_{jk}'s (e.g., avoid assumptions on tails...)

Notational convention

We have:

- $n =$ # of individuals in the sample
- $m =$ # of **M**ixture components
- $r =$ # of **R**epeated measurements (coordinates)
- Throughout, we use the subscripts:

$$1 \leq i \leq n, \quad 1 \leq j \leq m, \quad 1 \leq k \leq r$$

Notational convention

We have:

- $n = \#$ of individuals in the sample
- $m = \#$ of **M**ixture components
- $r = \#$ of **R**epeated measurements (coordinates)
- Throughout, we use the subscripts:

$$1 \leq i \leq n, \quad 1 \leq j \leq m, \quad 1 \leq k \leq r$$

The log-likelihood given data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

$$L(\theta) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_{ik}) \right)$$

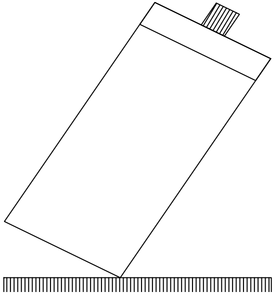
Motivating example: Water-level data

Example from Thomas Lohaus and Brainerd (1993).

The task:

- $n = 405$ subjects are shown $r = 8$ vessels, pointing at 1, 2, 4, 5, 7, 8, 10 and 11 o'clock
- They draw the water surface for each
- Measure: (signed) angle formed by surface with horizontal

Vessel tilted to point at 1:00



Outline: Next up. . .

Review of standard EM for mixtures

For MLE in finite mixtures, EM algorithms are standard.

A “complete” observation $(X, \mathbf{Z})$ consists of:

- The observed, “incomplete” data X
- The “missing” vector $\mathbf{Z}$, defined by

$$\text{for } 1 \leq j \leq m, \; Z_j = \begin{cases} 1 & \text{if } X \text{ comes from component } j \\ 0 & \text{otherwise} \end{cases}$$

Review of standard EM for mixtures

For MLE in finite mixtures, EM algorithms are standard.

A "complete" observation $(X, \mathbf{Z})$ consists of:

- The observed, "incomplete" data X
- The "missing" vector $\mathbf{Z}$, defined by

$$\text{for } 1 \leq j \leq m,\ Z_j = \begin{cases} 1 & \text{if } X \text{ comes from component } j \\ 0 & \text{otherwise} \end{cases}$$

What does this mean?

- In simulations: We generate $\mathbf{Z}$ first, then $X|\mathbf{Z}_j = 1 \sim f_j$
- In real data, $\mathbf{Z}$ is a latent variable whose interpretation depends on context.

Parametric (univariate) EM algorithm for mixtures

Let θ^t be an “arbitrary” value of θ

Parametric (univariate) EM algorithm for mixtures

Let $\boldsymbol{\theta}^t$ be an "arbitrary" value of $\boldsymbol{\theta}$

E-step: Amounts to find the conditional expectation of each Z

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

Parametric (univariate) EM algorithm for mixtures

Let $\boldsymbol{\theta}^t$ be an “arbitrary” value of $\boldsymbol{\theta}$

E-step: Amounts to find the conditional expectation of each Z

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

M-step: Maximize the “complete data” loglikelihood

$$\boldsymbol{\theta}^{t+1} = \arg\max_{\theta} \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij}^t \log \left[\lambda_j f(x_i; \phi_j)\right]$$

Parametric (univariate) EM algorithm for mixtures

Let $\boldsymbol{\theta}^t$ be an “arbitrary” value of $\boldsymbol{\theta}$

E-step: Amounts to find the conditional expectation of each Z

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

M-step: Maximize the “complete data” loglikelihood

$$\boldsymbol{\theta}^{t+1} = \arg\max_{\theta} \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij}^t \log\left[\lambda_j f(x_i; \phi_j)\right]$$

Typically: $\lambda_j^{t+1} = \dfrac{\sum_{i=1}^n Z_{ij}^t}{n}$, $\mu_j^{t+1} = \dfrac{\sum_{i=1}^n Z_{ij}^t x_i}{\sum_{i=1}^n Z_{ij}^t}$, ...

Semiparametric univariate mixture & EM-like algorithm

Identifiability: $g(x)$ uniquely determines all λ_j and f_j's

- **Parametric case:** When $f_j(x) = f(x; \phi_j)$, generally OK
- **Nonparametric case:** Some restrictions on f_j are needed

Semiparametric univariate mixture & EM-like algorithm

Identifiability: $g(x)$ uniquely determines all λ_j and f_j's

- **Parametric case:** When $f_j(x) = f(x; \phi_j)$, generally OK
- **Nonparametric case:** Some restrictions on f_j are needed

Bordes Mottelet and Vandekerkhove (2006) and Hunter Wang and Hettmansperger (2007) both showed that:
for f symmetric about the origin and $\lambda_1 \neq 1/2$,

$$g_{\boldsymbol{\theta}}(x) = \sum_{j=1}^{2} \lambda_j f(x - \mu_j)$$

is identifiable for the parameter $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\mu}, f)$.

Semiparametric univariate mixture & EM-like algorithm

Identifiability: $g(x)$ uniquely determines all λ_j and f_j's

- **Parametric case:** When $f_j(x) = f(x; \phi_j)$, generally OK
- **Nonparametric case:** Some restrictions on f_j are needed

Bordes Mottelet and Vandekerkhove (2006) and Hunter Wang and Hettmansperger (2007) both showed that:
for f symmetric about the origin and $\lambda_1 \neq 1/2$,

$$g_{\boldsymbol{\theta}}(x) = \sum_{j=1}^{2} \lambda_j f(x - \mu_j)$$

is identifiable for the parameter $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\mu}, f)$.

Bordes Chauveau and Vandekerkhove (2007) introduced a stochastic EM-like algorithm that includes a Kernel Density Estimation (KDE) step.

Outline: Next up. . .

Outline: Next up. . .

The blessing of dimensionality (!)

Recall the model in the multivariate case, $r > 1$:

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

N.B.: Assume conditional independence of $x_1, \ldots, x_r$

- Hall and 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.*

The blessing of dimensionality (!)

Recall the model in the multivariate case, $r > 1$:

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

N.B.: Assume conditional independence of $x_1, \ldots, x_r$

- Hall and 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. . .

Suppose some of the r coordinates are *identically distributed*.

- Let the r coordinates be grouped into B blocks of iid coordinates.
  Denote the block index of the kth coordinate by $b_k \in \{1, \dots, B\}$, $k = 1, \dots, r$.
- The model becomes

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

The notation gets even worse...

Suppose some of the r coordinates are *identically distributed*.

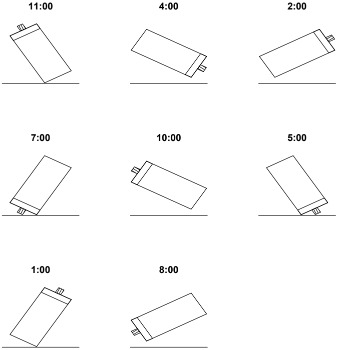
- Let the r coordinates be grouped into B blocks of iid coordinates.
  Denote the block index of the kth coordinate by $b_k \in \{1, \ldots, B\}$, $k = 1, \ldots, r$.
- The model becomes

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

- Special cases:
  - $b_k = k$ for each k: Fully general model, seen earlier (Hall et al. 2005; Qin and Leung 2006)
  - $b_k = 1$ for each k: Conditionally i.i.d. assumption (Elmore et al. 2004)

Motivation: The water-level data example again

8 vessels, presented in order 11, 4, 2, 7, 10, 5, 1, 8   o'clock

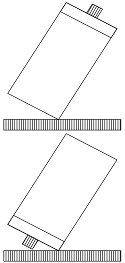


Motivation: The water-level data example again

8 vessels, presented in order 11, 4, 2, 7, 10, 5, 1, 8 o'clock

- Assume that opposite clock-face orientations lead to conditionally iid responses (same behavior)

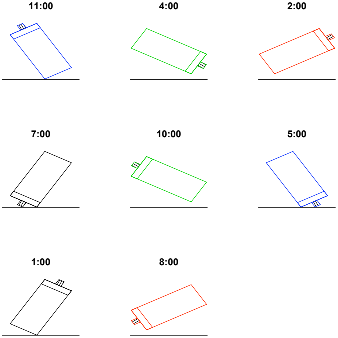
Vessel tilted to point at 1:00 and 7:00



Motivation: The water-level data example again

8 vessels, presented in order 11, 4, 2, 7, 10, 5, 1, 8 o'clock

- Assume that opposite clock-face orientations lead to conditionally iid responses (same behavior)
- $B = 4$ blocks defined by $\mathbf{b} = (4, 3, 2, 1, 3, 4, 1, 2)$
- e.g., $b_4 = b_7 = 1$, i.e., block 1 relates to coordinates 4 and 7, corresponding to clock orientations 1:00 and 7:00



The nonparametric "EM" (npEM) algorithm

E-step: Same as usual, but now f_{jb_k} is part of the parameter:

$$Z_{ij}^t \equiv \mathbb{E}_{\theta^t}[Z_{ij}|\mathbf{x}_i] = \frac{\lambda_j^t \prod_{k=1}^r f_{jb_k}^t(x_{ik})}{\sum_{j'} \lambda_{j'}^t \prod_{k=1}^r f_{j'b_k}^t(x_{ik})}$$

The nonparametric "EM" (npEM) algorithm

E-step: Same as usual, but now f_{jb_k} is part of the parameter:

$$Z_{ij}^t \equiv \mathbb{E}_{\theta^t}[Z_{ij}|\mathbf{x}_i] = \frac{\lambda_j^t \prod_{k=1}^r f_{jb_k}^t(x_{ik})}{\sum_{j'} \lambda_{j'}^t \prod_{k=1}^r f_{j'b_k}^t(x_{ik})}$$

M-step: Maximize "complete data loglikelihood" for $\boldsymbol{\lambda}$:

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

The nonparametric "EM" (npEM) algorithm

E-step: Same as usual, but now f_{jb_k} is part of the parameter:

$$Z_{ij}^t \equiv \mathbb{E}_{\boldsymbol{\theta}^t}[Z_{ij}|\mathbf{x}_i] = \frac{\lambda_j^t \prod_{k=1}^r f_{jb_k}^t(x_{ik})}{\sum_{j'} \lambda_{j'}^t \prod_{k=1}^r f_{j'b_k}^t(x_{ik})}$$

M-step: Maximize "complete data loglikelihood" for $\boldsymbol{\lambda}$:

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

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

$$f_{j\ell}^{t+1}(u) = \frac{1}{nhC_\ell \lambda_j^{t+1}} \sum_{k=1}^r \sum_{i=1}^n Z_{ij}^t \mathbb{I}_{\{b_k=\ell\}} K\left(\frac{u - x_{ik}}{h}\right)$$

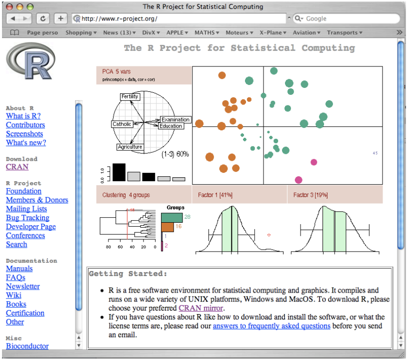
where $C_\ell = \sum_{k=1}^r \mathbb{I}_{\{b_k=\ell\}} = \#$ of coordinates in block ℓ

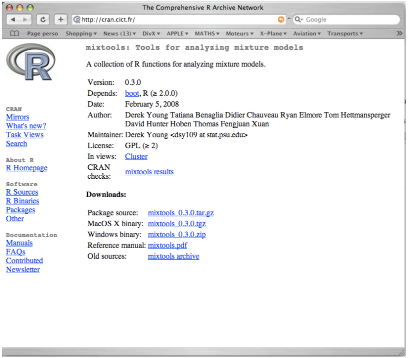
Outline: Next up. . .

Advertising!

All computational techniques in this talk are implemented in the **mixtools** package for the **R** Statistical Software

`www.r-project.org` `cran.cict.fr/web/packages/mixtools`





Simulated trivariate benchmark models

Comparisons with Hall et al. (2005) inversion method
$m = 2$, $r = 3$, conditional independence (no blocks)

Simulated trivariate benchmark models

Comparisons with Hall et al. (2005) inversion method
$m = 2$, $r = 3$, conditional independence (no blocks)

For $j = 1, 2$ and $k = 1, 2, 3$, we compute as in Hall et al.

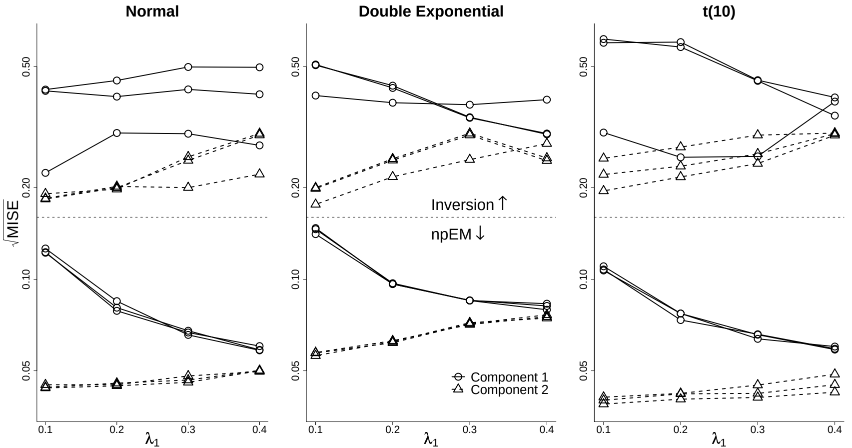
$$\mathrm{MISE}_{jk} = \frac{1}{S} \sum_{s=1}^{S} \int \left(\hat{f}_{jk}^{(s)}(u) - f_{jk}(u) \right)^2 du$$

over S replications, where $\hat{Z}_{ij}$'s are the final posterior, and

$$\hat{f}_{jk}(u) = \frac{1}{nh\hat{\lambda}_j} \sum_{i=1}^{n} \hat{Z}_{ij} K\left(\frac{u - x_{ik}}{h} \right)$$

MISE comparisons with Hall et al (2005) benchmarks

$n = 500$, $S = 300$ replications, 3 models, log scale



The Water-level data

Previously analysed using mixtures by Hettmansperger and Thomas (2000), and Elmore et al. (2004), using Assumptions and model:

- $r = 8$ coordinates assumed conditionally i.i.d.
- *Cutpoint approach* $=$ binning data in p-dim vectors
- mixture of multinomial identifiable whenever $r \geq 2m - 1$ (Elmore and Wang 2003)

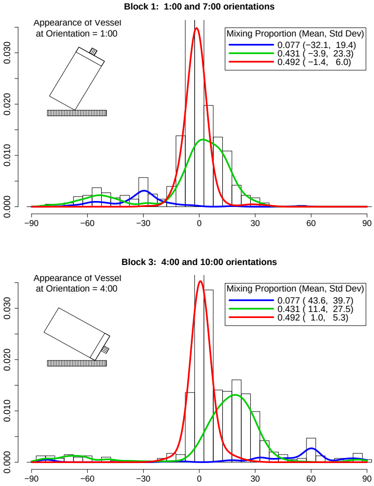
The Water-level data

Previously analysed using mixtures by Hettmansperger and Thomas (2000), and Elmore et al. (2004), using Assumptions and model:

- $r = 8$ coordinates assumed conditionally i.i.d.
- *Cutpoint approach* $=$ binning data in p-dim vectors
- mixture of multinomial identifiable whenever $r \geq 2m - 1$ (Elmore and Wang 2003)

The non appropriate i.i.d. assumption masks interesting features that our model reveals

The Water-level data, $m = 3$ components, 4 blocks



Outline: Next up. . .

Bandwidth issues in the kernel density estimates

Crude method :

- use R default (Silverman’s rule) based on *sd* (standard deviation) and *IQR* (InterQuartileRange) computed by pooling the $n \times r$ data points,

$$h = 0.9 \min \left\{ sd, \frac{IQR}{1.34} \right\} (nr)^{-1/5}$$

Bandwidth issues in the kernel density estimates

Crude method :

- use R default (Silverman's rule) based on *sd* (standard deviation) and *IQR* (InterQuartileRange) computed by pooling the $n \times r$ data points,

$$h = 0.9 \min \left\{ sd, \frac{IQR}{1.34} \right\} (nr)^{-1/5}$$

- Inappropriate for mixtures, e.g. for components with supports of different locations and/or scales
Example (see later): $f_{11} \equiv$ Student and $f_{22} \equiv \mathcal{B}eta$

Iterative and per component & block bandwidths

Estimated sample size for jth component and ℓth block

$$\sum_{i=1}^{n} \sum_{k=1}^{r} \mathbb{I}_{\{b_k=\ell\}} Z_{ij}^{t} = nC_{\ell}\lambda_j^{t}$$

Iterative and per component & block bandwidths

Estimated sample size for jth component and ℓth block

$$\sum_{i=1}^{n} \sum_{k=1}^{r} \mathbb{I}_{\{b_k=\ell\}} Z_{ij}^{t} = nC_\ell \lambda_j^t$$

Iterative bandwidth $h_{j\ell}^{t+1}$ applying (e.g.) Silverman's rule

$$h_{j\ell}^{t+1} = 0.9 \min \left\{ \sigma_{j\ell}^{t+1}, \frac{IQR_{j\ell}^{t+1}}{1.34} \right\} (nC_\ell \lambda_j^{t+1})^{-1/5}$$

where σ's and IQR's have to be estimated per iteration/component/block

Iterative and per component/block sd’s

Augment each M-step to include

$$\mu_{j\ell}^{t+1} = \frac{\displaystyle\sum_{i=1}^{n}\sum_{k=1}^{r} Z_{ij}^{t}\mathbb{I}_{\{b_k=\ell\}}x_{ik}}{nC_\ell\lambda_j^{t+1}},$$

$$\sigma_{j\ell}^{t+1} = \left[\frac{\displaystyle\sum_{i=1}^{n}\sum_{k=1}^{r} Z_{ij}^{t}\mathbb{I}_{\{b_k=\ell\}}(x_{ik}-\mu_{j\ell}^{t+1})^2}{nC_\ell\lambda_j^{t+1}}\right]^{1/2}$$

NB: these “parameters” are not in the model

Iterative and per component/block quantiles

Let $\mathbf{x}^\ell$ denote the nC_ℓ data in block ℓ, and $\tau(\cdot)$ be a permutation on $\{1, \ldots, nC_\ell\}$ such that

$$x^\ell_{\tau(1)} \leq x^\ell_{\tau(2)} \leq \cdots \leq x^\ell_{\tau(nC_\ell)}$$

Iterative and per component/block quantiles

Let $\mathbf{x}^\ell$ denote the nC_ℓ data in block ℓ, and $\tau(\cdot)$ be a permutation on $\{1, \dots, nC_\ell\}$ such that

$$x^\ell_{\tau(1)} \leq x^\ell_{\tau(2)} \leq \cdots \leq x^\ell_{\tau(nC_\ell)}$$

Define the weighted α-quantile estimate:

$$Q^{t+1}_{j\ell,\alpha} = x^\ell_{\tau(i_\alpha)}, \quad \text{where } i_\alpha = \min\left\{ s : \sum_{u=1}^{s} Z^t_{\tau(u)j} \geq \alpha n C_\ell \lambda_j^{t+1} \right\}$$

Iterative and per component/block quantiles

Let $\mathbf{x}^\ell$ denote the nC_ℓ data in block ℓ, and $\tau(\cdot)$ be a permutation on $\{1, \ldots, nC_\ell\}$ such that

$$x^\ell_{\tau(1)} \le x^\ell_{\tau(2)} \le \cdots \le x^\ell_{\tau(nC_\ell)}$$

Define the weighted α-quantile estimate:

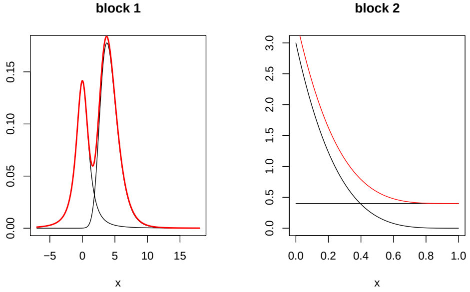
$$Q^{t+1}_{j\ell,\alpha} = x^\ell_{\tau(i_\alpha)}, \quad \text{where } i_\alpha = \min\left\{ s : \sum_{u=1}^{s} Z^t_{\tau(u)j} \ge \alpha n C_\ell \lambda^{t+1}_j \right\}$$

Set $IQR^{t+1}_{j\ell} = Q^{t+1}_{j\ell,0.75} - Q^{t+1}_{j\ell,0.25}$

Iterative & adaptive bandwidth illustration

Multivariate example with $m = 2$, $r = 5$, $B = 2$ blocks

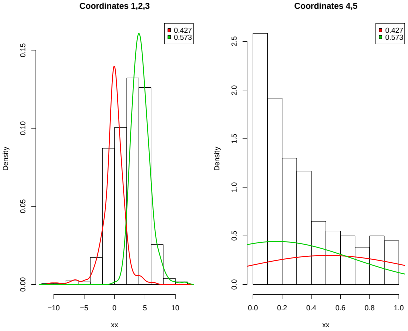
- Block 1 $= (x_1, x_2, x_3)$, components $f_{11} = t(2,0)$, $f_{21} = t(10,4)$
- Block 2 $= (x_4, x_5)$, components $f_{12} = \mathcal{U}_{[0,1]}$, $f_{22} = \mathcal{B}eta(1,5)$



Simulated data, $n = 300$ individuals

Default bandwidth

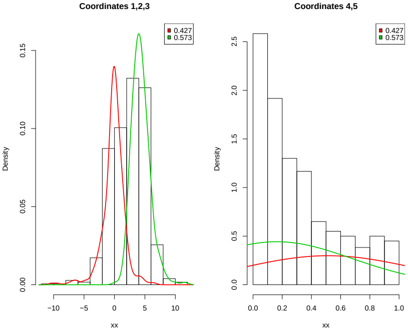
```
> blockid = c(1,1,1,2,2)
> a = npEM(x, 2, blockid)
> plot(a, breaks = 18)
> a$bandwidth
[1] 0.5238855
```



Simulated data, $n = 300$ individuals

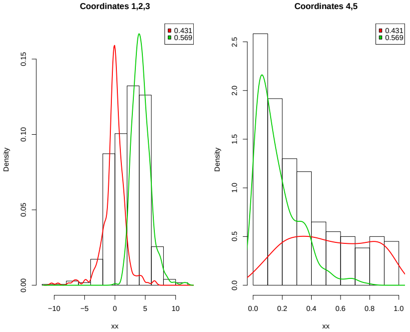
Default bandwidth

```
> blockid = c(1,1,1,2,2)
> a = npEM(x, 2, blockid)
> plot(a, breaks = 18)
> a$bandwidth
[1] 0.5238855
```

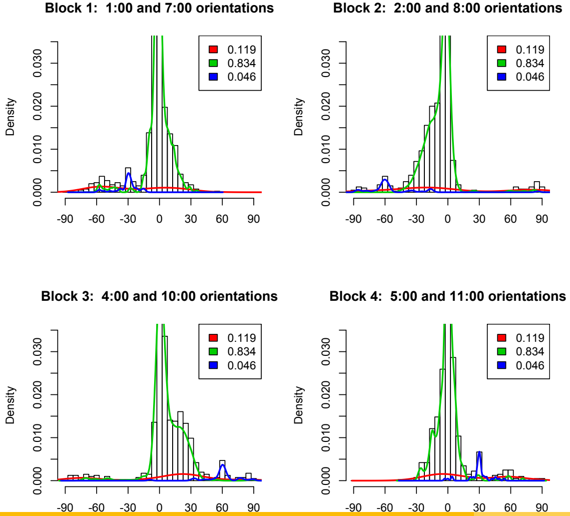


Bandwidth per block & component

```
> b = npEM(x, 2, blockid, samebw=FALSE)
> plot(b, breaks = 18)
> b$bandwidth
        component 1 component 2
block 1 0.38573749 0.35232409
block 2 0.08441747 0.04388618
```



The Water-level data with adaptive bandwidth




```
> b$band
        comp 1 comp 2  comp 3
block 1 12.172 1.4597 0.97535
block 2 13.996 2.7370 2.27581
block 3 19.190 2.5545 2.27582

block 4 12.363 1.2772 1.62558
```

Pros and cons of the npEM algorithm

- **Pro:** Easily generalizes beyond $m = 2$, $r = 3$ (not the case for inversion methods)
- **Pro:** Much lower MISE for similar test problems.
- **Pro:** Computationally simple (in the mixtools package).
- **Pro:** No need to assume conditionally i.i.d., and no loss of information from categorizing data (as for for the cutpoint approach)
- **Con:** Not a true EM algorithm (no monotonicity property) $\longrightarrow$ Nonlinear Smoothed Likelihood MM algorithms Levine, Hunter and Chauveau (2010, ...)

Outline: Next up. . .

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's, f_ℓ's, or the single f remain fully unspecified

For all these situations special cases of the npEM algorithm can easily be designed (some are already in **mixtools**).

Further extensions: Stochastic npEM versions

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

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

Then the sequence $(\theta^t)_{t\geq 1}$ becomes a Markov Chain

- Historically, parametric Stochastic EM introduced by Celeux Diebolt (1985, 1986,...), see also MCMC sampling (Diebolt Robert 1994)
- In non-parametric framework: Stochastic npEM for reliability mixture models, Bordes Chauveau (COMPSTAT 2010...)

Selected references

- Allman, E.S., Matias, C. and Rhodes, J.A. (2008), Identifiability of latent class models with many observed variables, *Annals of Statistics*, **37**: 3099–3132.
- Benaglia, T., Chauveau, D., and Hunter, D. R. (2009), An EM-like algorithm for semi- and non-parametric estimation in mutlivariate mixtures, J. Comput. Graph. Statist. **18**, no. 2, 505Ð-526.
- Benaglia T., Chauveau D., Hunter D. R., Young D. S., mixtools: An R Package for Analyzing Mixture Models, Journal of Statistical Software 32 (2009), 1–29.
- Bordes, L., Mottelet, S., and Vandekerkhove, P. (2006), Semiparametric estimation of a two-component mixture model, *Annals of Statistics*, **34**, 1204–1232.
- Bordes, L., Chauveau, D., and Vandekerkhove, P. (2007), An EM algorithm for a semiparametric mixture model, *Computational Statistics and Data Analysis*, **51**: 5429–5443.
- Elmore, R. T., Hettmansperger, T. P., and Thomas, H. (2004), Estimating component cumulative distribution functions in finite mixture models, *Communications in Statistics: Theory and Methods*, **33**: 2075–2086.
- Elmore, R. T., Hall, P. and Neeman, A. (2005), An application of classical invariant theory to identifiability in nonparametric mixtures, *Annales de l'Institut Fourier*, **55**, 1: 1–28.
- Hall, P. and Zhou, X. H. (2003) Nonparametric estimation of component distributions in a multivariate mixture, *Annals of Statistics*, **31**: 201–224.
- Hall, P., Neeman, A., Pakyari, R., and Elmore, R. (2005), Nonparametric inference in multivariate mixtures, *Biometrika*, **92**: 667–678.
- Hunter, D. R., Wang, S., and Hettmansperger, T. P. (2007), Inference for mixtures of symmetric distributions, *Annals of Statistics*, **35**: 224–251.
- Thomas, H., Lohaus, A., and Brainerd, C.J. (1993). Modeling Growth and Individual Differences in Spatial Tasks, *Monographs of the Society for Research in Child Development*, **58**, 9: 1–190.