EM-like algorithms for semi- and non-parametric estimation in multivariate mixtures

Didier Chauveau

MAPMO - UMR 6628 - Université d'Orléans

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

WU – Wien, June 17 2010

Outline

Finite mixture estimation problem

Goal: Estimate λ_j and f_j (or f_{jk}) given an i.i.d. sample from

Univariate Case: $x \in \mathbb{R}$

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

Multivariate case: $\mathbf{x} \in \mathbb{R}^r$

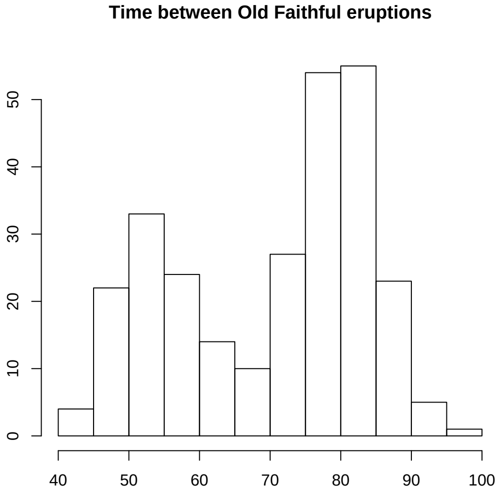
$$g(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_k)$$

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

Motivations:

Do not assume any more than necessary about the parametric form of f_j or f_{jk} (e.g., avoid assumptions on tails...)

Univariate example: Old Faithful wait times (min.)



from www.nps.gov/yell

- Obvious bimodality
- Normal-looking components ?
- More on this later!

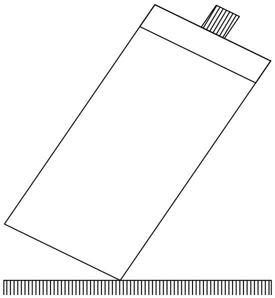
Multivariate example: Water-level data

Example from Thomas Lohaus and Brainerd (1993).

The task:

- Subjects are shown 8 vessels, pointing at 1:00, 2:00, 4:00, 5:00, 7:00, 8:00, 10:00, and 11:00
- They draw the water surface for each
- Measure: (signed) angle formed by surface with horizontal

Vessel tilted to point at 1:00



Notational convention

We have:

- $n =$ # of individuals in the sample
- $m =$ # of **M**ixture components
- $r =$ # of **R**epeated measurements (coordinates)

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

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

- Note the subscripts: Throughout, we use

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

For the examples

The Old Faithful geyser data

- Number of observations: $n = 272$
- Number of coordinates: $r = 1$ (univariate).
- Number of mixture components $m = 2$ (obviously)

The Water-level dataset

- Number of subjects: $n = 405$
- Number of coordinates (repeated measures): $r = 8$.
- What should m be (and mean for child development) ?

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 mixture model

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

The parameter of the mixture model is

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

Example: the Gaussian mixture model,

$f(x; \phi_j) = f\left(x; (\mu_j, \sigma_j^2)\right) =$ the pdf of $\mathcal{N}(\mu_j, \sigma_j^2)$.

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

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

M-step: Maximize the "complete data" loglikelihood

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

Iterate: Let $\boldsymbol{\theta}^{t+1} = \arg\max_\theta L_c(\boldsymbol{\theta})$ and repeat.

Parametric Gaussian EM

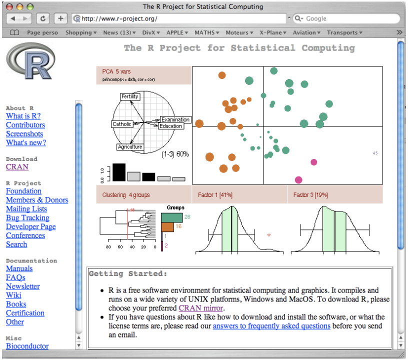
Typical M-step: for $j = 1, \ldots, m$

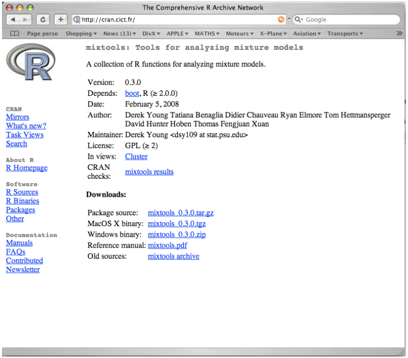
$$
\begin{aligned}
\lambda_j^{t+1} &= \frac{\sum_{i=1}^n Z_{ij}^t}{n} \\
\mu_j^{t+1} &= \frac{\sum_{i=1}^n Z_{ij}^t \, x_i}{n\lambda_j^{t+1}} \\
\sigma_j^{2\,t+1} &= \frac{\sum_{i=1}^n Z_{ij}^t \, (x_i - \mu_j^{t+1})^2}{n\lambda_j^{t+1}}
\end{aligned}
$$

Advertising!

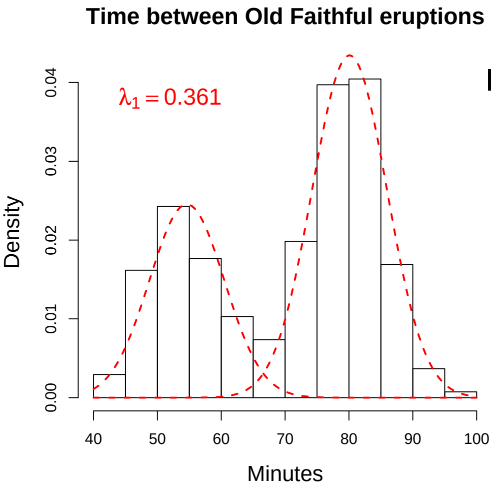
All computational techniques in this talk are implemented in the **mixtools** package for the **R** Statistical Software

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





Old Faithful data with parametric Gaussian EM



In R with **mixtools**, type

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

  number of iterations= 24
  ```
- Gaussian EM result:
  $\hat{\mu} = (54.6, 80.1)$
  ```

Identifiability

Univariate Case

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

Identifiability means: $g(x)$ uniquely determines all λ_j and f_j (up to permuting the subscripts).

- **Parametric case:** When $f_j(x) = f(x; \phi_j)$, generally no problem
- **Nonparametric case:** We need *some* restrictions on f_j

How to restrict f_j in the univariate ($r = 1$) case?

Bordes Mottelet and Vandekerkhove (2006) and Hunter Wang and Hettmansperger (2007) both showed that,
For $m = 2$, g is identifiable, at least when $\lambda_1 \neq 1/2$, if

$$f_j(x) \equiv f(x - \mu_j)$$

for some density $f(\cdot)$ that is symmetric about the origin.

Location-shift semiparametric mixture model with parameter:

$$\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\mu}, f)$$

A semi-parametric "EM" algorithm

Assume that

$$g(x) = \sum_{j=1}^{2} \lambda_j f(x - \mu_j),$$

where $f(\cdot)$ is a symmetric density.

Bordes Chauveau and Vandekerkhove (2007) introduce an EM-like algorithm that includes a kernel density estimation step.

- It is *much* simpler than the algorithms of Bordes et al. (2006) or Hunter et al. (2007).

An "EM" algorithm for $m = 2$, $r = 1$:

E-step: Same as usual:

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

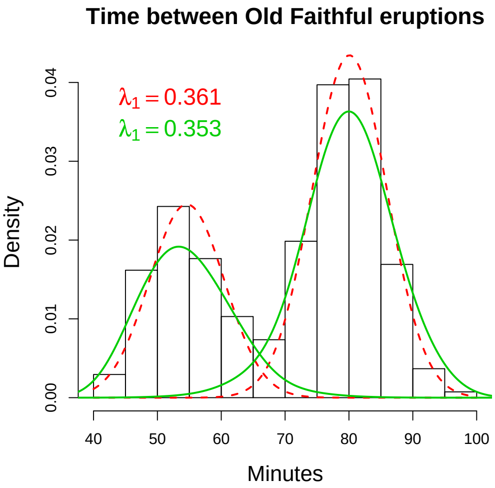
M-step: Maximize complete data "loglikelihood" for $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$:

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

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

$$f^{t+1}(u) = \frac{1}{nh}\sum_{i=1}^n \sum_{j=1}^2 Z_{ij}^t K\left(\frac{u - x_i + \mu_j^{t+1}}{h}\right), \quad \text{then symmetrize.}$$

Old Faithful data again (in mixtools)



- Gaussian EM:
  $\hat{\mu} = (54.6, 80.1)$
- Semiparametric EM

```
R> spEMsymmloc(waiting,
R+     mu=c(55,80),
R+     h=4) # bandwidth 4
```

  $\hat{\mu} = (54.7, 79.8)$

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 $b_k \in \{1, \ldots, B\}$, $k = 1, \ldots, r$.
- The model becomes

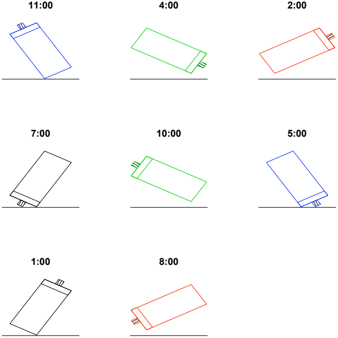
$$g(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{j b_k}(x_k)$$

- 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

- 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) generalized

E-step: Same as usual:

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

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

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

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

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): $f_{11} \equiv t(2)$ and $f_{22} \equiv \mathcal{B}eta(1,5)$

Iterative and per component & block bandwidth

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

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

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

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

Define the weighted α-quantile estimate:

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

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

Simulated trivariate benchmark models

Comparisons with Hall et al. (2005) inversion method
$m = 2$, $r = 3$, $\mathbf{b} = (1, 2, 3)$, 3 models

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

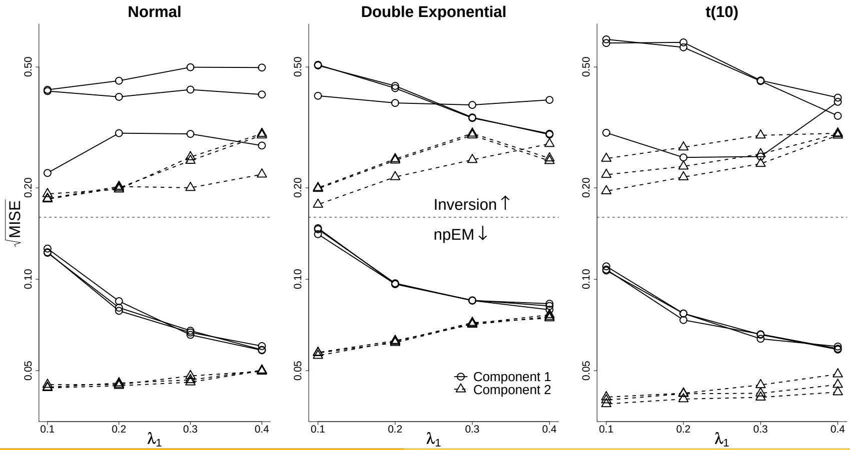
$$\mathrm{MISE}_{jk} = \frac{1}{S} \sum_{s=1}^{S} \int \left(\hat{f}_{jk}^{(s)}(u) - f_{jk}(u) \right)^2 du$$

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

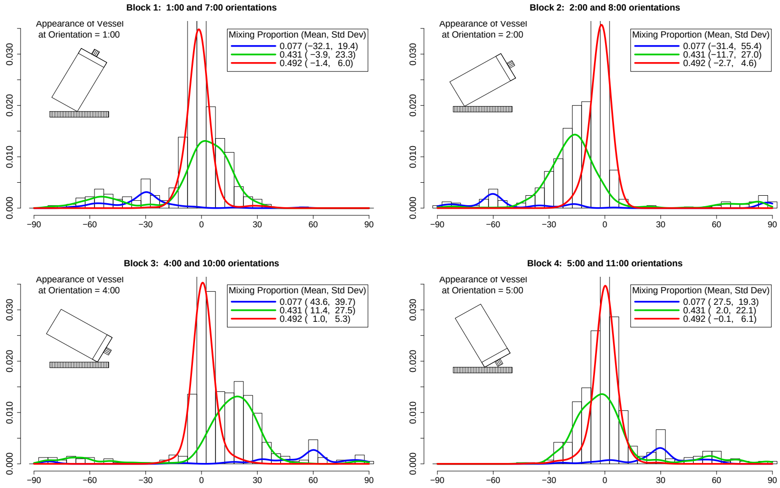
Dataset previously analysed by Hettmansperger and Thomas (2000), and Elmore et al. (2004)

Assumptions and model:

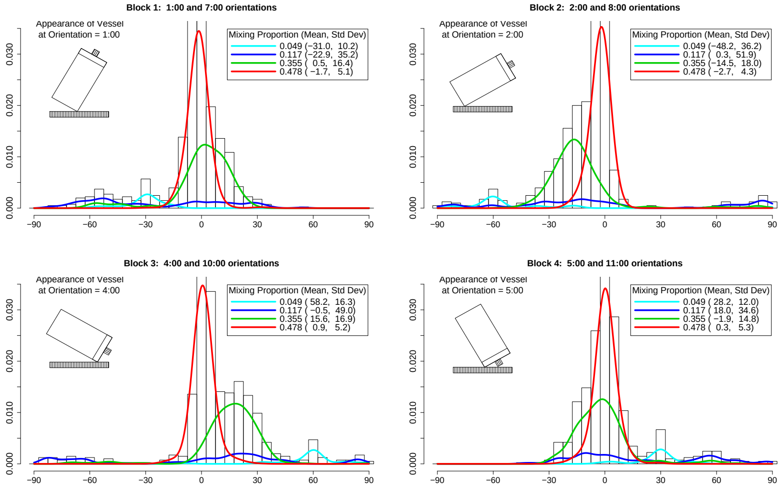
- $r = 8$ coordinates assumed conditionally i.i.d.
- *Cutpoint approach* = binning data in p-dim vectors
- mixture of multinomial identifiable whenever $r \geq 2m - 1$ (Elmore and Wang 2003)

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

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



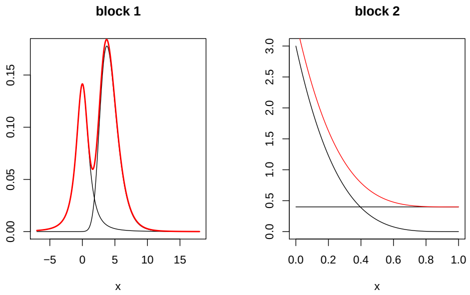
The Water-level data, $m = 4$ components, 4 blocks



Iterative bandwidth $h_{j\ell}^t$ illustration

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

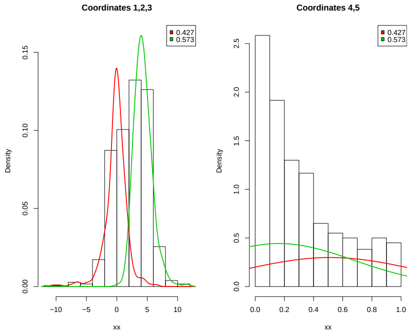
- Block 1: coordinates $k = 1, 2, 3$,
  components $f_{11} = t(2, 0)$, $f_{21} = t(10, 4)$
- Block 2: coordinates $k = 4, 5$,
  components $f_{12} = \mathcal{B}(1, 1) = \mathcal{U}_{[0,1]}$, $f_{22} = \mathcal{B}(1, 5)$



Simulated data, $n = 300$ individuals

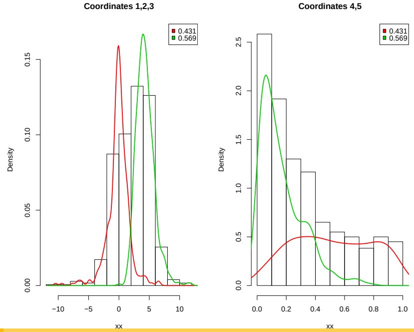
Default bandwidth

```
> id = c(1,1,1,2,2)
> a = npEM(x, centers, id, eps=1e-8)
> plot(a, breaks = 18)
> a$bandwidth
[1] 0.5238855
```



Bandwidth per block & component

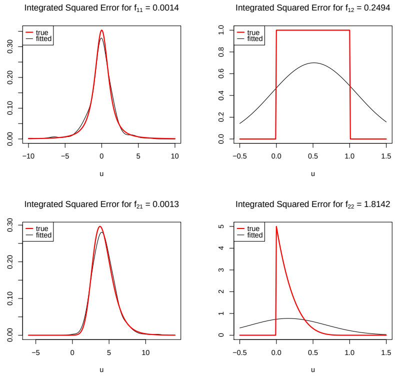
```
> b = npEM(x, centers, id, eps=1e-8, samebw=FALSE)
> plot(b, breaks = 18)
> b$bandwidth
        component 1 component 2
block 1 0.38573749 0.35232409
block 2 0.08441747 0.04388618
```



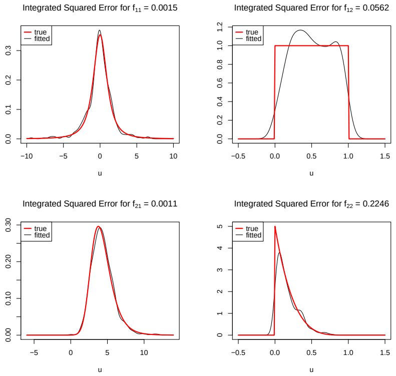
Integrated Squared Error for densities $f_{j\ell}$'s

Using `ise.npEM()` in mixtools:

Default bandwidth



Bandwidth per block & component



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_ℓ, 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:

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

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 (2010)

Pros and cons of npEM

- **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.
- **Pro:** No need to assume conditionally i.i.d. (not the case for cutpoint approach)
- **Pro:** No loss of information from categorizing data.
- **Con:** Not a true EM algorithm (no monotonicity property)

From EM to NEMS for “nonparametric” mixtures

Nonparametric in this literature relates to the mixing distribution

- true EM but ill-posed difficulties , Vardi et al. (1985)
- Smoothed EM (EMS), Silverman et al. (1990)
- regularization approach from Eggermont and LaRiccia (1995) and Eggermont (1999): Nonlinear EMS (NEMS)

Goal: combining regularization and npEM approach
Joint work with M. Levine and D. Hunter (2010)

Smoothing the log-density

Following Eggermont (1992, 1999):

- Smoothing, for $f \in L_1(\Omega)$ and $\Omega \subset \mathbb{R}^r$

$$\mathcal{S}f(\mathbf{x}) = \int_\Omega K_h(\mathbf{x}-\mathbf{u})f(\mathbf{u})\,d\mathbf{u},$$

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

$$\mathcal{N}f(\mathbf{x}) \;=\; \exp\{(\mathcal{S}\log f)(\mathbf{x})\} \;=\; \exp\int_\Omega K_h(\mathbf{x}-\mathbf{u})\log f(\mathbf{u})\,d\mathbf{u}.$$

  $\mathcal{N}$ is multiplicative: $\mathcal{N}f_j = \prod_k \mathcal{N}f_{jk}$

Smoothing the mixture

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

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

Goal: minimizing the objective function

$$\ell(\boldsymbol{\theta}) = \ell(\mathbf{f}, \boldsymbol{\lambda}) := \int_\Omega g(\mathbf{x}) \log \frac{g(\mathbf{x})}{[\mathcal{M}_{\boldsymbol{\lambda}} \mathcal{N} \mathbf{f}](\mathbf{x})} \, d\mathbf{x}.$$

with f_{jk}'s univariate pdf and $\sum_{j=1}^{m} \lambda_j = 1$.

Majorization-Minimization (MM) trick

MM trick: instead of ℓ, minimize a majorizing function:

$$\begin{array}{rrcl} & b^0(\boldsymbol{\theta}) + \text{constant} & \geq & \ell(\boldsymbol{\theta}), \\ \text{with} & b^0(\boldsymbol{\theta}^0) + \text{constant} & = & \ell(\boldsymbol{\theta}^0), \quad \boldsymbol{\theta}^0 = \text{ current value} \end{array}$$

Set

$$\begin{aligned} w_j^0(\mathbf{x}) &:= \frac{\lambda_j^0 \mathcal{N} f_j^0(\mathbf{x})}{\mathcal{M}_{\boldsymbol{\lambda}^0} \mathcal{N} \mathbf{f}^0(\mathbf{x})}, \quad \sum_{j=1}^m w_j^0(\mathbf{x}) = 1 \\ b^0(\mathbf{f}, \boldsymbol{\lambda}) &:= -\int g(\mathbf{x}) \sum_{j=1}^m w_j^0(\mathbf{x}) \log \left[\lambda_j \mathcal{N} f_j(\mathbf{x}) \right] \, d\mathbf{x} \end{aligned}$$

Then $\quad b^0(\mathbf{f}, \boldsymbol{\lambda}) - b^0(\mathbf{f}^0, \boldsymbol{\lambda}^0) \geq \ell(\mathbf{f}, \boldsymbol{\lambda}) - \ell(\mathbf{f}^0, \boldsymbol{\lambda}^0)$

MM (Majorization-Minimization) “algorithm”

Minimization of $b^0(\mathbf{f}, \boldsymbol{\lambda})$ for $j = 1, \ldots, m$ and $k = 1, \ldots, r$

$$\begin{aligned} \hat{\lambda}_j &= \int g(\mathbf{x}) w_j^0(\mathbf{x})\, d\mathbf{x} \\ \hat{f}_{jk}(u) &\propto \int K_h(x_k - u) g(\mathbf{x}) w_j^0(\mathbf{x})\, d\mathbf{x}, \quad u \in \mathbb{R} \end{aligned}$$

Theorem: Descent property (like a true EM)

$$\ell(\hat{\mathbf{f}}, \hat{\boldsymbol{\lambda}}) \leq \ell(\mathbf{f}^0, \boldsymbol{\lambda}^0).$$

MM algorithm with a descent property

Discrete version: given the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ iid $\sim g$

$$\ell_n(\mathbf{f}, \boldsymbol{\lambda}) := \int \log \frac{1}{[\mathcal{M}_{\boldsymbol{\lambda}} N\mathbf{f}](\mathbf{x})} \, dG_n(\mathbf{x}) \;=\; -\sum_{i=1}^{n} \log[\mathcal{M}_{\boldsymbol{\lambda}} N\mathbf{f}](\mathbf{x}_i)$$

The corresponding MM algorithm satisfies a descent property

$$\ell_n(\mathbf{f}^{t+1}, \boldsymbol{\lambda}^{t+1}) \leq \ell_n(\mathbf{f}^{t}, \boldsymbol{\lambda}^{t})$$

nonparametric Maximum Smoothed Likelihood (npMSL) algorithm

E-step:

$$w_{ij}^t = \frac{\lambda_j^t \mathcal{N} f_j^t(\mathbf{x}_i)}{\mathcal{M}_{\boldsymbol{\lambda}^t} \mathcal{N} \mathbf{f}^t(\mathbf{x}_i)} = \frac{\lambda_j^t \mathcal{N} f_j^t(\mathbf{x}_i)}{\sum_{j'=1}^m \lambda_{j'} \mathcal{N} f_{j'}^t(\mathbf{x}_i)}.$$

M-step: for $j = 1, \ldots, m$

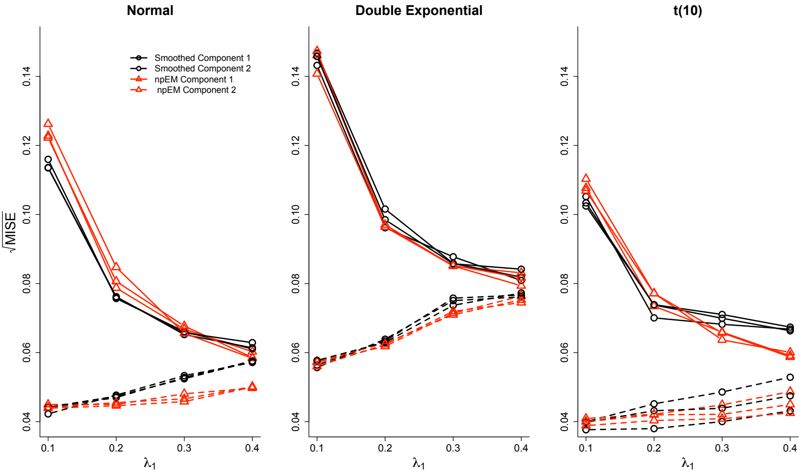
$$\lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^n w_{ij}^t \tag{1}$$

WKDE-step: For each j and k, let

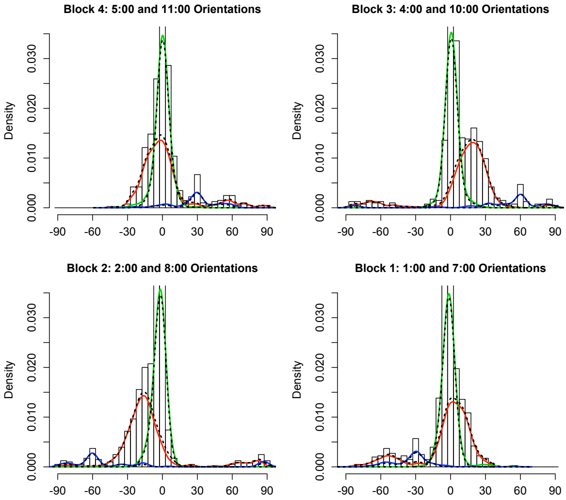
$$f_{jk}^{t+1}(u) = \frac{1}{nh\lambda_j^{t+1}} \sum_{i=1}^n w_{ij}^t K\left(\frac{u - x_{ik}}{h}\right). \tag{2}$$

npEM vs. npMSL for Hall et al benchmarks

$m = 2$, $r = 3$, $n = 500$, $S = 300$ replications, 3 models



npEM vs. npMSL for the Water-level data



$m = 3$ components
4 blocks of 2 coord. each
colored lines: npEM
dotted lines: npMSL

Conclusion. . .

Possible generalizations of the npMSL

- to block structure (see the Water-level data)
- to semiparametric (location/scale) models
- to adaptive bandwidth issue

Open questions for npEM and npMSL

- Can we have different block structure in each component?
  Yes, but in this case label-switching becomes an issue.
- Are the estimators consistent, and if so at what rate?
  Emperical evidence: Rates of convergence similar to those in non-mixture setting.

References, part 1 of 2

- Allman, E.S., Matias, C. and Rhodes, J.A. (2008), Identifiability of latent class models with many observed variables, preprint.
- 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.
- Benaglia, T., Chauveau, D., and Hunter, D. R. (2010), Bandwidth Selection in an EM-like algorithm for nonparametric multivariate mixtures, IMS Lecture Notes – Monograph Series (to appear).
- 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.

References, part 2 of 2

- 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.
- Qin, J. and Leung, D. H.-Y. (2006), Semiparametric analysis in conditionally independent multivariate mixture models, unpublished manuscript.