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

1. Mixture models and EM algorithms
   - Motivations, examples and notation
   - Review of EM algorithm-ology

2. The semi-parametric univariate case

3. Multivariate non-parametric “EM” algorithms
   - Model and algorithms
   - Examples

4. Nonlinear smoothed Likelihood maximization
Finite mixture estimation problem

**Goal:** Estimate $\lambda_j$ 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:** $x \in \mathbb{R}^r$
  
  $$g(x) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_k)$$

*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...)

D. Chauveau – June 2010
Nonparametric multivariate mixtures
Univariate example: Old Faithful wait times (min.)

Time between Old Faithful eruptions

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

from www.nps.gov/yell
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
Notational convention

We have:

- $n =$ # of individuals in the sample
- $m =$ # of Mixture components
- $r =$ # of Repeated measurements (coordinates)

Thus, the log-likelihood given data $x_1, \ldots, x_n$ 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

$$1 \leq i \leq n, \quad 1 \leq j \leq m, \quad 1 \leq k \leq r$$
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)?
For MLE in finite mixtures, EM algorithms are standard.

A “complete” observation \((X, Z)\) consists of:

- The observed, “incomplete” data \(X\)
- The “missing” vector \(Z\), defined by

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

What does this mean?

- In simulations: We generate \(Z\) first, then \(X|Z_j = 1 \sim f_j\)
- In real data, \(Z\) 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

\[
\theta = (\lambda, \phi) = (\lambda_1, \ldots, \lambda_m, \phi_1, \ldots, \phi_m)
\]

**Example:** the Gaussian mixture model,

\[
f(x; \phi_j) = f\left(x; (\mu_j, \sigma_j^2)\right) = \text{the pdf of } \mathcal{N}(\mu_j, \sigma_j^2).
\]
Let $\theta^t$ be an "arbitrary" value of $\theta$

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

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

**M-step:** Maximize the “complete data” loglikelihood

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

**Iterate:** Let $\theta^{t+1} = \arg \max_{\theta} L_c(\theta)$ and repeat.
Typical M-step: for $j = 1, \ldots, m$

\[
\begin{align*}
\lambda_{j}^{t+1} &= \frac{\sum_{i=1}^{n} Z_{ij}^{t}}{n} \\
\mu_{j}^{t+1} &= \frac{\sum_{i=1}^{n} Z_{ij}^{t} x_{i}}{n \lambda_{j}^{t+1}} \\
\sigma_{j}^{2t+1} &= \frac{\sum_{i=1}^{n} Z_{ij}^{t} (x_{i} - \mu_{j}^{t+1})^2}{n \lambda_{j}^{t+1}}
\end{align*}
\]
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

\[ \lambda_1 = 0.361 \]

In R with **mixtools**, type

```r
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 \( \lambda_j \) 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:

$$\theta = (\lambda, \mu, f)$$
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:

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

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

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

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

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

Time between Old Faithful eruptions

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

- Semiparametric EM

  ```
  R> spEMsymmloc(waiting,
  R+ mu=c(55,80),
  R+ h=4) # bandwidth 4
  R> \hat{\mu} = (54.7, 79.8)
  ```

\( \lambda_1 = 0.361 \)
\( \lambda_1 = 0.353 \)
The blessing of dimensionality (!)

Recall the model in the \textbf{multivariate case}, \( r > 1 \):

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

\textit{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) \ldots \textit{from at least one point of view, the \textquote{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 \textit{identically distributed}. 

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

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

- 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:

\[
Z^t_{ij} \equiv E_{\theta^t}[Z_{ij}|x_i] = \frac{\lambda^t_j \prod_{k=1}^r f^t_{jb_k}(x_{ik})}{\sum_{j'} \lambda^t_{j'} \prod_{k=1}^r f^t_{j'b_k}(x_{ik})}
\]

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

\[
\lambda^t_{j+1} = \frac{1}{n} \sum_{i=1}^n Z^t_{ij}
\]

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

\[
f_{j\ell}^{t+1}(u) = \frac{1}{n h C_{\ell} \lambda^t_{j}} \sum_{k=1}^r \sum_{i=1}^n Z^t_{ij} \mathbb{I}_{\{b_k=\ell\}} K \left( \frac{u - x_{ik}}{h} \right)
\]

where \( C_{\ell} = \sum_{k=1}^r \mathbb{I}_{\{b_k=\ell\}} = \# \) of coordinates in block \( \ell \)
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 $f_{22} \equiv Beta(1, 5)$
Iterative and per component & block bandwidth

Estimated sample size for \( j \)th component and \( \ell \)th 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 \( \sigma \)'s and \( IQR \)'s have to be estimated per iteration/component/block
Augment each M-step to include

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

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

NB: these “parameters” are not in the model
Let $x^\ell$ denote the $nC^\ell$ data in block $\ell$, 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 $\alpha$-quantile estimate:

$$Q^{t+1}_{j\ell, \alpha} = x^\ell_{\tau(i_\alpha)}$$

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 $IQR^{t+1}_{j\ell} = Q_{j\ell, 0.75}^{t+1} - Q_{j\ell, 0.25}^{t+1}$
Simulated trivariate benchmark models

Comparisons with Hall et al. (2005) inversion method

\( m = 2, \ r = 3, \ b = (1, 2, 3), \ 3 \) models

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

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

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

- **Normal**
- **Double Exponential**
- **t(10)**
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

Block 1: 1:00 and 7:00 orientations

Appearance of Vessel at Orientation = 1:00

Mixing Proportion (Mean, Std Dev)
- 0.077 (-32.1, 19.4)
- 0.431 (-3.9, 23.3)
- 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)
- 0.077 (-31.4, 55.4)
- 0.431 (-11.7, 27.0)
- 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)
- 0.077 (43.6, 39.7)
- 0.431 (11.4, 27.5)
- 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)
- 0.077 (27.5, 19.3)
- 0.431 (2.0, 22.1)
- 0.492 (-0.1, 6.1)
The Water-level data, \( m = 4 \) components, 4 blocks

**Block 1: 1:00 and 7:00 orientations**

**Appearance of Vessel at Orientation = 1:00**

Mixing Proportion (Mean, Std Dev)
- 0.049 (−31.0, 10.2)
- 0.117 (−22.9, 35.2)
- 0.355 (0.5, 16.4)
- 0.478 (−1.7, 5.1)

**Block 2: 2:00 and 8:00 orientations**

**Appearance of Vessel at Orientation = 2:00**

Mixing Proportion (Mean, Std Dev)
- 0.049 (−48.2, 36.2)
- 0.117 (0.3, 51.9)
- 0.355 (−14.5, 18.0)
- 0.478 (−2.7, 4.3)

**Block 3: 4:00 and 10:00 orientations**

**Appearance of Vessel at Orientation = 4:00**

Mixing Proportion (Mean, Std Dev)
- 0.049 (58.2, 16.3)
- 0.117 (−0.5, 49.0)
- 0.355 (15.6, 16.9)
- 0.478 (0.9, 5.2)

**Block 4: 5:00 and 11:00 orientations**

**Appearance of Vessel at Orientation = 5:00**

Mixing Proportion (Mean, Std Dev)
- 0.049 (28.2, 12.0)
- 0.117 (18.0, 34.6)
- 0.355 (−1.9, 14.8)
- 0.478 (0.3, 5.3)
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**

```r
> 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**

```r
> 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 \( f_j, f_\ell, f \) remain fully unspecified

For all these situations special cases of the npEM algorithm can easily be designed (some are already in \texttt{mixtools}).
Further extensions: Stochastic npEM versions

In some 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 sequence $\left( \theta^t \right)_{t \geq 1}$ 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)
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 \( \Omega \subset \mathbb{R}^r \)

\[
S f(x) = \int_\Omega K_h(x - u)f(u) \, du,
\]

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

- Nonlinear smoothing

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

\( \mathcal{N} \) is multiplicative: \( \mathcal{N} f_j = \prod_k \mathcal{N} f_{jk} \)
Smoothing the mixture

For \( f = (f_1, \ldots, f_m) \), define

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

Goal: minimizing the objective function

\[
\ell(\theta) = \ell(f, \lambda) := \int_{\Omega} g(x) \log \frac{g(x)}{[M_\lambda \mathcal{N} f](x)} \, dx.
\]

with \( f_{jk} \)'s univariate pdf and \( \sum_{j=1}^{m} \lambda_j = 1 \).
Majorization-Minimization (MM) trick

MM trick: instead of $\ell$, minimize a majorizing function:

$$b^0(\theta) + \text{constant} \geq \ell(\theta),$$

with $b^0(\theta^0) + \text{constant} = \ell(\theta^0), \quad \theta^0 = \text{current value}$

Set

$$w^0_j(x) := \frac{\lambda_j^0 N f^0_j(x)}{\mathcal{M} \lambda_0^0 N f^0(x)}, \quad \sum_{j=1}^m w^0_j(x) = 1$$

$$b^0(f, \lambda) := - \int g(x) \sum_{j=1}^m w^0_j(x) \log [\lambda_j N f_j(x)] \, dx$$

Then

$$b^0(f, \lambda) - b^0(f^0, \lambda^0) \geq \ell(f, \lambda) - \ell(f^0, \lambda^0)$$
MM (Majorization-Minimization) “algorithm”

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

$$\hat{\lambda}_j = \int g(x) w^0_j(x) \, dx$$
$$\hat{f}_{jk}(u) \propto \int K_h(x_k - u) g(x) w^0_j(x) \, dx, \quad u \in \mathbb{R}$$

Theorem: Descent property (like a true EM)

$$\ell(\hat{f}, \hat{\lambda}) \leq \ell(f^0, \lambda^0).$$
Discrete version: given the sample $x_1, \ldots, x_n$ iid $\sim g$

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

The corresponding 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:**

\[ w_{ij}^t = \frac{\lambda_j^t \mathcal{N}(x_i) \mathcal{M}}{\sum_{j'=1}^{m} \lambda_{j'}^t \mathcal{N}(x_i)} = \frac{\lambda_j^t \mathcal{N}(x_i)}{\sum_{j'=1}^{m} \lambda_{j'}^t \mathcal{N}(x_i)} . \]

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

\[ \lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^{n} w_{ij}^t \quad (1) \]

**WKDE-step:** For each \( j \) and \( k \), let

\[ f_{jk}^{t+1}(u) = \frac{1}{nh\lambda_j^{t+1}} \sum_{i=1}^{n} w_{ij}^t K\left(\frac{u - x_{ik}}{h}\right) . \quad (2) \]
Mixture models and EM algorithms
The semi-parametric univariate case
Multivariate non-parametric “EM” algorithms
Nonlinear smoothed Likelihood maximization

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.
Allman, E. S., Matias, C. and Rhodes, J. A. (2008), Identifiability of latent class models with many observed variables, preprint.


References, part 2 of 2


