EM-like algorithms for nonparametric multivariate mixture models

Didier Chauveau

MAPMO - UMR 6628 - Université d'Orléans

Joint work with:
David Hunter (Le Studium, CNRS Orléans and Penn State University),
and Tatiana Benaglia (Penn State University)

Workshop on mixture models and their applications
Pau, June 23–24, 2008
Outline

1. Motivations, examples and notation
2. Review of EM algorithm-ology
3. The semi-parametric univariate case
4. The multivariate non-parametric “EM” algorithm
Outline: Next up...

1. Motivations, examples and notation
2. Review of EM algorithm-ology
3. The semi-parametric univariate case
4. The multivariate non-parametric “EM” algorithm
Finite mixture estimation problem

**Goal:** Estimate $\lambda_j$ and $f_j$ (or $f_{jk}$) given an i.i.d. sample from:

**Univariate Case**

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

**Goal:** Estimate $\lambda_j$ and $f_j$ (or $f_{jk}$) given an i.i.d. sample from:

### Univariate Case

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

### Multivariate case

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

**N.B.:** Assume conditional independence of $x_1, \ldots, x_r$. 
Finite mixture estimation problem

**Goal:** Estimate $\lambda_j$ and $f_j$ (or $f_{jk}$) given an i.i.d. sample from:

Univariate Case

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

Multivariate case

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

**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...)

June 2008  Nonparametric 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

June 2008    Nonparametric mixtures
Multivariate example: Water-level angles

This example is due to 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 nightmare

We have:

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

We have:

- \( n = \# \text{ of individuals in the sample} \)
- \( m = \# \text{ of Mixture components} \)
- \( r = \# \text{ of Repeated measurements (coordinates)} \)

Thus, the log-likelihood given data \( x_1, \ldots, x_n \) is

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

- 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

For the Old Faithful geyser data
- Number of observations: $n = 272$
- Number of coordinates: $r = 1$ (univariate).
- Number of mixture components $m = 2$ (obviously)

For the Water-level dataset:
- Number of subjects: $n = 405$
- Number of coordinates (repeated measures): $r = 8$.
- What should $m$ (number of mixture components) be?
Outline: Next up...

1. Motivations, examples and notation
2. Review of EM algorithm-ology
3. The semi-parametric univariate case
4. The multivariate non-parametric “EM” algorithm
Review of standard EM for mixtures

For MLE in finite mixtures, EM algorithms are standard. A “complete” observation \((X, Z)\) consists of:

- The “observed” 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.
Review of standard EM for mixtures

For MLE in finite mixtures, EM algorithms are standard.

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

- The “observed” 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 \(X \mid Z_j = 1 \sim f_j\)
- In real data, \(Z\) is a \textit{latent variable} whose interpretation depends on context.
**Motivations, examples and notation**

Review of EM algorithm-ology

The semi-parametric univariate case

The multivariate non-parametric “EM” algorithm

---

**Parametric (univariate) EM algorithm for mixtures**

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

\[
\hat{Z}_{ij} \equiv \mathbb{E}_{\hat{\theta}}[Z_{ij}|x_i] = \mathbb{P}[Z_{ij} = 1|x_i] = \frac{\hat{\lambda}_j \hat{f}_j(x_i)}{\sum_{j'} \hat{\lambda}_{j'} \hat{f}_{j'}(x_i)}
\]

---

June 2008  
Nonparametric mixtures
**Parametric (univariate) EM algorithm for mixtures**

**E-step:** Amounts to find the conditional expectation of each \( Z \)

\[
\hat{Z}_{ij} \equiv \mathbb{E}_\theta[Z_{ij}|x_i] = \mathbb{P}[Z_{ij} = 1|x_i] = \frac{\hat{\lambda}_j \hat{f}_j(x_i)}{\sum_{j'} \hat{\lambda}_{j'} \hat{f}_{j'}(x_i)}
\]

*N.B.: In parametric case \( f_j(x) \equiv f(x; \phi_j) \). We let \( \theta \) denote \((\lambda, \phi)\)*
Parametric (univariate) EM algorithm for mixtures

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

$$
\hat{Z}_{ij} \equiv \mathbb{E}_{\hat{\theta}}[Z_{ij}|x_i] = P[Z_{ij} = 1|x_i] = \frac{\hat{\lambda}_j \hat{f}_j(x_i)}{\sum_{j'} \hat{\lambda}_{j'} \hat{f}_{j'}(x_i)}
$$

*N.B.: In parametric case $f_j(x) \equiv f(x; \phi_j)$. We let $\theta$ denote $(\lambda, \phi)$*

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

$$
L_c(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{Z}_{ij} \log [\lambda_j f_j(x_i)]
$$
Parametric (univariate) EM algorithm for mixtures

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

$$\hat{Z}_{ij} \equiv \mathbb{E}_\theta[Z_{ij}|x_i] = \mathbb{P}[Z_{ij} = 1|x_i] = \frac{\hat{\lambda}_j \hat{f}_j(x_i)}{\sum_{j'} \hat{\lambda}_{j'} \hat{f}_{j'}(x_i)}$$

N.B.: In parametric case $f_j(x) \equiv f(x; \phi_j)$. We let $\theta$ denote $(\lambda, \phi)$

M-step: Maximize the complete data loglikelihood

$$L_c(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{Z}_{ij} \log [\lambda_j f_j(x_i)]$$

Iterate: Let $\hat{\theta} = \arg \max_\theta L_c(\theta)$ and repeat.
All computational techniques in this talk are implemented in a package called **mixtools** for R Statistical Software

www.r-project.org  
cran.cict.fr/web/packages/mixtools
Old Faithful data with parametric normal EM

In R with `mixtools`, type

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

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

In R with **mixtools**, type

- \( \text{R> data(faithful)} \)
- \( \text{R> attach(faithful)} \)
- \( \text{R> ans=normalmixEM(} \)
  - \( \text{R+ waiting,} \)
  - \( \text{R+ mu=c(55,80),} \)
  - \( \text{R+ sigma=5,} \)
  - \( \text{R+ fast=T)} \)

number of iterations = 24

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

\( \lambda_1 = 0.361 \)
Outline: Next up...
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 $m = 2$

$$g(x) = \sum_{j=1}^{2} \lambda_j f_j(x)$$

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.
Exploiting identifiability: An “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:

\[
\hat{Z}_{ij} \equiv \mathbb{E}_\hat{\theta}[Z_{ij}|x_i] = \frac{\hat{\lambda}_j \hat{f}(x_i - \hat{\mu}_j)}{\hat{\lambda}_1 \hat{f}(x_i - \hat{\mu}_1) + \hat{\lambda}_2 \hat{f}(x_i - \hat{\mu}_2)}
\]
An “EM” algorithm for \( m = 2, r = 1 \):

E-step: Same as usual:

\[
\hat{Z}_{ij} \equiv \mathbb{E}_{\hat{\theta}}[Z_{ij}|x_i] = \frac{\hat{\lambda}_j \hat{f}(x_i - \hat{\mu}_j)}{\hat{\lambda}_1 \hat{f}(x_i - \hat{\mu}_1) + \hat{\lambda}_2 \hat{f}(x_i - \hat{\mu}_2)}
\]

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

\[
\tilde{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_{ij} \quad \tilde{\mu}_j = (n\tilde{\lambda}_j)^{-1} \sum_{i=1}^{n} \hat{Z}_{ij}x_i
\]
An “EM” algorithm for $m = 2, r = 1$:

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

$$
\hat{Z}_{ij} \equiv \mathbb{E}_\theta[Z_{ij} | x_i] = \frac{\hat{\lambda}_j \hat{f}(x_i - \hat{\mu}_j)}{\hat{\lambda}_1 \hat{f}(x_i - \hat{\mu}_1) + \hat{\lambda}_2 \hat{f}(x_i - \hat{\mu}_2)}
$$

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

$$
\tilde{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_{ij} \quad \tilde{\mu}_j = (n\tilde{\lambda}_j)^{-1} \sum_{i=1}^{n} \hat{Z}_{ij} x_i
$$

**KDE-step:** Update estimate of $f$ (for some bandwidth $h$) by

$$
\tilde{f}(u) = (nh)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} \hat{Z}_{ij} K \left( \frac{u - x_i + \hat{\mu}_j}{h} \right), \text{then symmetrize.}
$$
Old Faithful data again

Time between Old Faithful eruptions

Minutes

Density

0.00
0.01
0.02
0.03
0.04

40 50 60 70 80 90 100
Old Faithful data again

Time between Old Faithful eruptions

\[ \lambda_1 = 0.361 \]

Gaussian EM:
\[ \hat{\mu} = (54.6, 80.1) \]
Old Faithful data again

Time between Old Faithful eruptions

- Gaussian EM: $\hat{\mu} = (54.6, 80.1)$
- Semiparametric EM with bandwidth $= 4$: $\hat{\mu} = (54.7, 79.8)$
Outline: Next up...

1. Motivations, examples and notation
2. Review of EM algorithm-ology
3. The semi-parametric univariate case
4. The multivariate non-parametric “EM” algorithm
The blessing of dimensionality (!)

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

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

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

- Hall and Zhou (2003) show that when $m = 2$ and $r \geq 3$, the model is identifiable without restrictions on the $f_{jk}(\cdot)$!
The blessing of dimensionality (!)

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

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

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

- Hall and Zhou (2003) show that when $m = 2$ and $r \geq 3$, the model is identifiable without restrictions on the $f_{jk}(\cdot)$!
- Hall et al (2005) remark that this is a case in which ... from at least one point of view, the ‘curse of dimensionality’ works in reverse.
The notation gets even worse. . .

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

- Let the $r$ coordinates be grouped into $B$ i.d. blocks. Denote the block of the $k$th coordinate by $b_k$, $1 \leq b_k \leq B$.
- The model becomes

$$g(x) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jb_k}(x_k)$$
The notation gets even worse…

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

- Let the $r$ coordinates be grouped into $B$ i.d. blocks. Denote the block of the $k$th coordinate by $b_k$, $1 \leq b_k \leq B$.
- 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)
The nonparametric “EM,” generalized

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

$$
\hat{Z}_{ij} \equiv \mathbb{E}_{\hat{\theta}}[Z_{ij}|x_i] = \frac{\hat{\lambda}_j \prod_{k=1}^{r} \hat{f}_{jb_k}(x_{ik})}{\sum_{j'} \hat{\lambda}_{j'} \prod_{k=1}^{r} \hat{f}_{j'b_k}(x_{ik})}
$$
The nonparametric “EM,” generalized

E-step: Same as usual:

\[ \hat{Z}_{ij} \equiv \mathbb{E}_{\hat{\theta}}[Z_{ij}|x_i] = \frac{\hat{\lambda}_j \prod_{k=1}^{r} \hat{f}_{jb_k}(x_{ik})}{\sum_{j'} \hat{\lambda}_{j'} \prod_{k=1}^{r} \hat{f}_{j'b_k}(x_{ik})} \]

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

\[ \tilde{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_{ij} \]
The nonparametric “EM,” generalized

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

\[ \hat{Z}_{ij} \equiv E_{\hat{\theta}} [Z_{ij}|x_i] = \frac{\hat{\lambda}_j \prod_{k=1}^{r} \hat{f}_{jb_k}(x_{ik})}{\sum_{j'} \hat{\lambda}_{j'} \prod_{k=1}^{r} \hat{f}_{j'k}(x_{ik})} \]

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

\[ \tilde{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_{ij} \]

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

\[ \tilde{f}_{j\ell}(u) = \frac{1}{nhC_{\ell}} \sum_{k=1}^{r} \sum_{i=1}^{n} \hat{Z}_{ij} \mathbb{I}_{\{b_k = \ell\}} K \left( \frac{u - x_{ik}}{h} \right), \quad C_{\ell} = \sum_{k=1}^{r} \mathbb{I}_{\{b_k = \ell\}} \]
The Water-level data, three components

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)

Appearance of Vessel at Orientation = 5:00

Block 2: 2:00 and 8:00 orientations

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

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

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, four components**

### Block 1: 1:00 and 7:00 orientations

- **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

- **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

- **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

- **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)

**But is the model Identifiable?**

**June 2008**
Motivations, examples and notation
Review of EM algorithm-ology
The semi-parametric univariate case
The multivariate non-parametric “EM” algorithm

The Water-level data, four components

But is the model Identifiable?

June 2008
Nonparametric mixtures
Pros and cons of np-EM compared with:

- Mixture model inversion method (Hall et al. 2005)
  - Pro: Easily generalizes beyond $m = 2, r = 3$.  
  - “Con”: Easily generalizes beyond $m = 2, r = 3$.  
  - Pro: Much lower MISE for similar test problems.  
  - Pro: Computationally simple.
Pros and cons of np-EM compared with:

- Mixture model inversion method (Hall et al. 2005)
  - **Pro:** Easily generalizes beyond \( m = 2, r = 3 \).
  - **“Con”:** Easily generalizes beyond \( m = 2, r = 3 \).
  - **Pro:** Much lower MISE for similar test problems.
  - **Pro:** Computationally simple.

- Cutpoint method (Elmore et al. 2004)
  - **Pro:** No need to assume conditionally i.i.d.
  - **Pro:** No loss of information from categorizing data.
  - **Con:** Identifiability of model not settled
Open questions

What about identifiability for $m > 2$?

*Elmore et al. (2005) apply classical invariant theory (!) and still can’t give a complete answer.*

*see C. Matias previous talk?*
Open questions

- What about identifiability for $m > 2$?
  
  Elmore et al. (2005) apply classical invariant theory (!) and still can’t give a complete answer.
  
  see C. Matias previous talk ?

- Can we have different block structure in each component?
  
  Yes, but in this case label-switching becomes an issue. An iterative approach could work here too.
Open questions

- What about identifiability for $m > 2$?
  *Elmore et al. (2005) apply classical invariant theory (!) and still can’t give a complete answer.*
  *see C. Matias previous talk ?*

- Can we have different block structure in each component?
  *Yes, but in this case label-switching becomes an issue. An iterative approach could work here too.*

- Are the estimators consistent, and if so at what rate?
  *Empirical evidence: Rates of convergence similar to those in non-mixture setting.*
References, part 1 of 2

- Benaglia, T., Chauveau, D., and Hunter, D. R. (2008), An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures, hal.archives-ouvertes.fr/hal-00193730/fr
References, part 2 of 2


