Maximum Smoothed Likelihood for Multivariate Nonparametric Mixtures

David Hunter

Pennsylvania State University, USA

Joint work with: Tom Hettmansperger, Hoben Thomas, Didier Chauveau, Pierre Vandekerkhove, Laurent Bordes, Tatiana Benaglia, Michael Levine, Xiaotian Zhu

Research partially supported by NSF Grants SES–0518772, DMS–1209007

Outline

1. Nonparametric mixtures and parameter identifiability

2. Motivating example; extension to multivariate case

3. Smoothed maximum likelihood
1. Nonparametric mixtures and parameter identifiability

2. Motivating example; extension to multivariate case

3. Smoothed maximum likelihood
We first introduce nonparametric finite mixtures

\[ X \sim g(x) = \int f_{\phi}(x) \, dQ(\phi) \]  

Sometimes, \( Q(\cdot) \) is the “nonparametric” part; e.g., work by Bruce Lindsay assumes \( Q(\cdot) \) is unrestricted. However, in this talk we assume that

- \( f_{\phi}(\cdot) \) is (mostly) unrestricted
- \( Q(\cdot) \) has finite support

So (1) becomes

\[ g(x) = \sum_{j=1}^{m} \lambda_j f_j(x) \quad \text{... and we assume } m \text{ is known.} \]
Can we learn about male / female *subpopulations* (i.e., parameters)?

What can we say about individuals?
Old Faithful Geyser waiting times provide another simple univariate example

Let $m = 2$, so assume we have a sample from

$$\lambda_1 f_1(x) + \lambda_2 f_2(x).$$

Why do we need any assumptions on $f_j$?
With no assumptions, parameters are not identifiable

\[ \lambda_1 f_1 \quad \lambda_2 f_2 \]

Multiple different parameter combinations

\[(\lambda_1, \lambda_2, f_1, f_2)\]

give the same mixture density.

Thus, some constraints on \( f_j \) are necessary.

NB: Sometimes, there is no obvious multi-modality.
The univariate case is identifiable under some assumptions

It is possible to show\(^1\) that if

\[ g(x) = \sum_{j=1}^{2} \lambda_j f_j(x), \]

the \(\lambda_j\) and \(f_j\) are uniquely identifiable from \(g\) if \(\lambda_1 \neq 1/2\) and

\[ f_j(x) \equiv f(x - \mu_j) \]

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

\(^1\)cf. Bordes, Mottelet, and Vandekerkhove (2006); Hunter, Wang, and Hettmansperger (2007)
A modified EM algorithm may be used for estimation

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

- The “observed” data \(X\)
- The “unobserved” 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: Generate \(Z\) first, then \(X \mid Z \sim \prod_j [f_j(\cdot)]^{Z_j}\)
- In real data, \(Z\) is a *latent variable* whose interpretation depends on context.
Standard EM for finite mixtures looks like this:

**E-step:** Amounts to finding the conditional expectation of each $Z_i$:

\[
\hat{Z}_{ij} \overset{\text{def}}{=} E\theta Z_{ij} = \frac{\lambda_j f_j(x_i)}{\hat{\lambda} \cdot \hat{f}(x_i)}
\]

**M-step:** Amounts to maximizing the “expected complete data loglikelihood”

\[
L_c(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{Z}_{ij} \log [\lambda_j f_j(x_i)] \quad \Rightarrow \quad \hat{\lambda}_{\text{next}} = \frac{1}{n} \sum_{i} \hat{Z}_i
\]

**Iterate:** Let $\hat{\theta}_{\text{next}} = \arg\max_{\theta} L_c(\theta)$ and repeat.

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

May 2013  |  Multivariate Nonparametric Mixtures
A modified EM algorithm may be used for estimation

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

\[
\hat{Z}_{ij} \overset{\text{def}}{=} E_{\hat{\theta}} Z_{ij} = \frac{\hat{\lambda}_j \hat{f}_j(x_i)}{\hat{\lambda} \cdot \hat{f}(x_i)} \quad \text{and} \quad \hat{\lambda}^{\text{next}} = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_i
\]

**KDE-step:** Update \( \hat{f}_j \) using a weighted kernel density estimate. Weight each \( x_i \) by the corresponding \( \hat{Z}_{ij} \).

Alternatively, select randomly

\[
Z_i \sim \text{Mult}(1, \hat{Z}_i)
\]

and use a standard KDE for the \( x_i \) selected into each component. (cf. Bordes, Chauveau, & Vandekerkhove, 2007).
For the symmetric location family assumption, the modified EM looks like this

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

\[
\hat{Z}_{ij} \equiv E_{\hat{\theta}} Z_{ij} = \frac{\hat{\lambda}_j \hat{f}(x_i - \mu_j)}{\hat{\lambda}_1 \hat{f}(x_i - \mu_1) + \hat{\lambda}_2 \hat{f}(x_i - \mu_2)}
\]

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

\[
\hat{\lambda}_j^{\text{next}} = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_{ij}, \quad \hat{\mu}_j^{\text{next}} = (n\hat{\lambda}_j)^{-1} \sum_{i=1}^{n} \hat{Z}_{ij} x_i
\]

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

\[
\hat{f}^{\text{next}}(u) = (nh)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} \hat{Z}_{ij} K \left( \frac{u - x_i + \hat{\mu}_j}{h} \right), \text{then symmetrize.}
\]
Compare two solutions for Old Faithful data

Time between Old Faithful eruptions

Minutes

Density

0.00 0.01 0.02 0.03 0.04

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

Semiparametric EM with bandwidth $= 4$:
$
\hat{\mu} = (54.7, 79.8)
$

Both algorithms are implemented in the `mixtools` package for R (Benaglia et al., 2009).
Compare two solutions for Old Faithful data

Time between Old Faithful eruptions

$\lambda_1 = 0.361$

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

Both algorithms are implemented in the `mixtools` package for R (Benaglia et al., 2009).
Compare two solutions for Old Faithful data

Time between Old Faithful eruptions

- Gaussian EM: \( \hat{\mu} = (54.6, 80.1) \)
- Semiparametric EM with bandwidth = 4: \( \hat{\mu} = (54.7, 79.8) \)
- Both algorithms are implemented in \texttt{mixtools} package for R (Benaglia et al., 2009).
Outline

1. Nonparametric mixtures and parameter identifiability
2. Motivating example; extension to multivariate case
3. Smoothed maximum likelihood
Multivariate example: Water-level angles

This example is due to Thomas, Lohaus, & 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 on each vessel.
- Measure: (signed) angle with horizontal formed by line drawn.

Vessel tilted to point at 1:00.
After loading `mixtools`,

```r
R> data(Waterdata)
R> Waterdata
Trial.1  Trial.2  Trial.3  Trial.4  Trial.5  Trial.6  Trial.7  Trial.8
1   -16     30      -5       8       30      1       0      -26
2   -12     13     -15     -64      21      17      -46     -66
3     7     -85     -27     -81      39      -2      -10     -12
4     6      55      -5       1      36      -3       5      -14
5    32      58     -61     -30      60      28     -30     -60
...
```

For these data,

- Number of subjects: \( n = 405 \)
- Number of coordinates (repeated measures): \( r = 8 \).
- What should \( m \) (number of mixture components) be?
We will assume conditional independence

Each \( f_j(x) \) density on \( \mathbb{R}^r \) is assumed to be the product of its marginals:

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

- We call this assumption *conditional independence* (cf. Hall and Zhou, 2003; Qin and Leung, 2006)
- Very similar to repeated measures models:
  - In RM models, we often assume measurements are independent conditional on the individual.
  - Here, we have component-specific effects instead of individual-specific effects.
There exists a nice identifiability result for conditional independence when $r \geq 3$

Recall the conditional independence finite mixture model:

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

Allman, Matias, & Rhodes (2009) use a theorem by Kruskal (1976) to show that if:

- $f_{1k}, \ldots, f_{mk}$ are linearly independent for each $k$;
- $r \geq 3$

... then $g(x)$ uniquely determines all the $\lambda_j$ and $f_{jk}$ (up to label-switching).
Some of the marginals may be assumed identical

Let the \( r \) coordinates be grouped into \( B \) i.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, Neeman, Pakyari, & Elmore 2005; Qin & Leung 2006)
- \( b_k = 1 \) for each \( k \): Conditionally i.i.d. assumption (Elmore, Hettmansperger, & Thomas 2004)
The water-level data may be blocked

8 vessels, presented in the 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 \( b = (4, 3, 2, 1, 3, 4, 1, 2) \)
- 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” algorithm is easily extended to the multivariate conditional independence case.

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

\[
\hat{Z}_{ij} = \frac{\hat{\lambda}_j \hat{f}_j(x_i)}{\hat{\lambda} \cdot \hat{f}(x_i)} = \frac{\hat{\lambda}_j \prod_{k=1}^{r} \hat{f}_{jk}(x_{ik})}{\hat{\lambda} \cdot \hat{f}(x_i)}
\]

and

\[
\hat{\lambda}_{\text{next}} = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_i
\]

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

\[
\hat{f}_{jk}(u) = \frac{1}{nh\hat{\lambda}_{j\text{next}}} \sum_{i=1}^{n} \hat{Z}_{ij} K \left( \frac{u - x_{ik}}{h} \right).
\]

(Benaglia, Chauveau, Hunter, 2009)
The Water-level data, three components

Block 1: 1:00 and 7:00 orientations

- 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 = 1:00

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)

Appearance of Vessel at Orientation = 2:00

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)

Appearance of Vessel at Orientation = 4:00

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)

Appearance of Vessel at Orientation = 5:00

May 2013  Multivariate Nonparametric Mixtures
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)

Appearance of Vessel at Orientation = 1:00

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)

Appearance of Vessel at Orientation = 2:00

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)

Appearance of Vessel at Orientation = 4:00

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)

Appearance of Vessel at Orientation = 5:00

May 2013 Multivariate Nonparametric Mixtures
1. Nonparametric mixtures and parameter identifiability
2. Motivating example; extension to multivariate case
3. Smoothed maximum likelihood
The previous algorithm is not truly EM

- Does this algorithm maximize any sort of log-likelihood? Does it have an EM-like ascent property?
- Are the estimators consistent and, if so, at what rate?

*Empirical evidence: Rates of convergence similar to those in non-mixture setting.*

We might hope that the algorithm has an ascent property for

\[
\ell(\lambda, f) = \sum_{i=1}^{n} \log \sum_{j=1}^{m} \lambda_j f_j(x_i)
\]

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

Unfortunately, \( \ell(\lambda^{\text{next}}, f^{\text{next}}) \not\geq \ell(\lambda, f) \)
Smoothing the likelihood

- No ascent property for

\[ \ell(\lambda, f) = \sum_{i=1}^{n} \log \sum_{j=1}^{m} \lambda_j f_j(x_i). \]

- However, we borrow an idea from Eggermont and LaRiccia (1995) and introduce a nonlinearly smoothed version:

\[ \ell^n_S(\lambda, f) = \sum_{i=1}^{n} \log \sum_{j=1}^{m} \lambda_j [N f_j](x_i), \]

where

\[ [N f_j](x) = \exp \int \frac{1}{h^r} K_r \left( \frac{x - u}{h^r} \right) \log f_j(u) \, du. \]
The “infinite sample” case: Minimum K-L divergence

Whereas \( \ell^n_S(\lambda, f) = \sum_{i=1}^n \log \sum_{j=1}^m \lambda_j \mathcal{N}(f_j(x_i)), \)
we may define \( e_j = \lambda_j f_j \) and write

\[
\ell^\infty(e) = \int g(x) \log \left( \frac{g(x)}{\sum_j \mathcal{N}(e_j(x))} \right) \, dx + \sum_j \int e_j(x) \, dx.
\]

- We wish to minimize \( \ell^\infty(e) \) over vectors \( e \) of functions not necessarily integrating to unity.
- The added term ensures that the solution satisfies the usual constraints, i.e., \( \sum_j \lambda_j = 1 \) and \( \int f_j(x) \, dx = 1 \).
- This may be viewed as minimizing a (penalized) K-L divergence between \( g(\cdot) \) and \( \sum_j \mathcal{N}e_j(\cdot) \).
The finite- and infinite-sample problems lead to EM-like (actually MM) algorithms

**E-step:**

Old: \( \hat{Z}_{ij} = \frac{\hat{\lambda}_j \hat{f}_j(x_i)}{\sum_{j'} \hat{\lambda}_{j'} \hat{f}_{j'}(x_i)} \)

**M-step:**

\[ \hat{\lambda}^{\text{next}} = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_i \]

**KDE-step:** (Part 2 of M-step)

\[ \hat{f}_{jk}^{\text{next}}(u) = \frac{1}{nh\hat{\lambda}_j^{\text{next}}} \sum_{i=1}^{n} \hat{Z}_{ij} K \left( \frac{u - x_{ik}}{h} \right). \]
The finite- and infinite-sample problems lead to EM-like (actually MM) algorithms

**E-step:** (Technically an M-step, for “minorization”)

Old: \( \hat{Z}_{ij} = \frac{\hat{\lambda}_j \hat{f}_j(x_i)}{\sum_{j'} \hat{\lambda}_{j'} \hat{f}_{j'}(x_i)} \)

New: \( \hat{Z}_{ij} = \frac{\lambda_j \hat{f}_j(x_i)}{\sum_{j'} \lambda_{j'} \hat{f}_{j'}(x_i)} \)

**M-step:**

\[
\hat{\lambda}^{\text{next}} = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_i
\]

**KDE-step:** (Part 2 of M-step)

\[
\hat{f}_{jk}^{\text{next}}(u) = \frac{1}{nh\hat{\lambda}_j^{\text{next}}} \sum_{i=1}^{n} \hat{Z}_{ij} K \left( \frac{u - x_{ik}}{h} \right)
\]
First “M” in MM is for “minorization”

- Recall the smoothed loglikelihood:

\[
\ell_n^S(\lambda, f) = \sum_{i=1}^n \log \sum_{j=1}^m \lambda_j \mathcal{N}(f_j(x_i))
\]

- Define

\[
Q(\lambda, f | \hat{\lambda}, \hat{f}) = \sum_{i=1}^n \sum_{j=1}^m \hat{Z}_{ij} \log \{ \lambda_j \mathcal{N}(f_j(x_i)) \}
\]

- We can prove

\[
\ell_n^S(\lambda, f) - \ell_n^S(\hat{\lambda}, \hat{f}) \geq Q(\lambda, f | \hat{\lambda}, \hat{f}) - Q(\hat{\lambda}, \hat{f} | \hat{\lambda}, \hat{f}).
\]

- We say that \(Q(\cdot | \hat{\lambda}, \hat{f})\) is a \textit{minorizer} of \(\ell_n^S(\cdot)\) at \((\lambda, f)\).
MM is a generalization of EM

- The minorizing equation is
  \[ \ell^n_S(\lambda, f) - \ell^n_S(\hat{\lambda}, \hat{f}) \geq Q(\lambda, f | \hat{\lambda}, \hat{f}) - Q(\hat{\lambda}, \hat{f} | \hat{\lambda}, \hat{f}). \]  
  (2)

- By (2), increasing \( Q(\cdot | \hat{\lambda}, \hat{f}) \) leads to an increase in \( \ell^n_S(\cdot) \).

- \( Q(\lambda, f | \hat{\lambda}, \hat{f}) \) is maximized at \( (\hat{\lambda}^{\text{next}}, \hat{f}^{\text{next}}) \).

- Thus, the MM algorithm guarantees that
  \[ \ell^n_S(\hat{\lambda}^{\text{next}}, \hat{f}^{\text{next}}) \geq \ell^n_S(\hat{\lambda}, \hat{f}). \]
The Water-level data revisited

- “NEMS” = nonlinearly smoothed EM-like algorithm (Levine, Chauveau, Hunter 2011)
- Colored lines = original (non-smoothed) algorithm
- Very little difference in any example we’ve seen
Next step: Use ascent property to establish theory

- Empirical results suggest consistency and convergence rate results are possible.
- Eggermont and LaRiccia (1995) prove asymptotic results for a related problem; however, it appears that their methods do not apply directly here.

\[ f_{21} \text{: slope } = -0.365 \]

\[ \log(n) \]

\[ \text{log}(\text{MISE}) \]

\[ 4.5 \quad 5.0 \quad 5.5 \quad 6.0 \quad 6.5 \quad 7.0 \quad 7.5 \quad 8.0 \]

\[ -3.8 \quad -3.4 \quad -3.0 \quad -2.6 \]

\[ \lambda^2 \text{: slope } = -0.488 \]

\[ \log(n) \]

\[ \text{log}(\text{MSE}) \]
References

EXTRA SLIDES
Consider the simplest case: Univariate $x_i$ and $J \in \{1, 2\}$:

- $Y = X\beta_J + \epsilon$ where $\epsilon \sim f$
- Fix $X = x_0$. 

May 2013

Multivariate Nonparametric Mixtures
Identifiability for mixtures of regressions: Intuition

Consider the simplest case: Univariate $x_i$ and $J \in \{1, 2\}$:

- $Y = X\beta_J + \epsilon$ where $\epsilon \sim f$
- Fix $X = x_0$.
- Conditional distribution of $Y$ when $X = x_0$ not necessarily identifiable as mixture of shifted versions of $f$, even if $f$ is assumed (say) symmetric.
Consider the simplest case: Univariate $x_i$ and $J \in \{1, 2\}$:

- $Y = X\beta_J + \epsilon$ where $\epsilon \sim f$
- Fix $X = x_0$.
- Conditional distribution of $Y$ when $X = x_0$ not necessarily identifiable as mixture of shifted versions of $f$, even if $f$ is assumed (say) symmetric.
- Identifiability depends on using additional $X$ values that change the relative locations of the mixture components.
Next allow an intercept: \( Y = \beta J_1 + X\beta J_2 + \epsilon, \) with \( \epsilon \sim f \).

- Even if \( f \) is assumed (say) symmetric about zero, identifiability can fail:
- Additional \( X \) values give no new information if the regression lines are parallel.
Mixtures of simple linear regressions

Next allow an intercept: \( Y = \beta J_1 + X\beta J_2 + \epsilon \), with \( \epsilon \sim f \).

- Even if \( f \) is assumed (say) symmetric about zero, identifiability can fail:
- Additional \( X \) values give no new information if the regression lines are parallel.
Theorem (Hunter and Young 2012)

Theorem: If the support of the density $h(x)$ contains an open set in $\mathbb{R}^p$, then the parameters in

$$
\psi(x, y) = h(x) \sum_{j=1}^{m} \lambda_j f(y - x^t \beta_j),
$$

are uniquely identifiable from $\psi(x, y)$.

- When $X$ is one-dimensional, there is a nice way to see this based on an idea of Laurent Bordes and Pierre Vandekerkhove.
Single predictor case

Let

\[ g_x(y) = \sum_{j=1}^{m} \lambda_j f(y - \mu_j + \beta_j x) \]

denote the conditional density of \( y \) given \( x \).

Next, define

\[ R_k(a, b) = \int x^k g_x(a + bx) \, dx. \]

A bit of algebra gives:

\[ R_0(a, b) = \sum_{j=1}^{m} \frac{\lambda_j}{|b - \beta_j|} \]

This identifies the \( \beta_j, \lambda_j \);

\[ R_1(a, b) = \sum_{j=1}^{m} \frac{\lambda_j(\mu_j - a)}{(b - \beta_j)^2} \]

This identifies the \( \mu_j \);