Semi- and Non-Parametric Mixture Models: A Progress Report

Tom Hettmansperger
Tatiana Benaglia
Tracey Wrobel
Penn State University

Dave Hunter
Didier Chauveau
Hoben Thomas

Current and Future Trends in Nonparametrics
October 11-12, 2007
University of South Carolina
Water Level Task
405 Children
Ages 11-16

Measurement = angular error
Clock settings: 1,2,4,5,7,8,10,11
Piaget:

Age 4: no understanding
Ages 5-7: confused but learning
Age 9: should understand

Data: 405 vectors of 8 measurements

Problem: Fit a 3 component multivariate mixture without assuming a parametric form for the underlying model

\[ f(x) = \lambda_1 f_1(x) + \lambda_2 f_2(x) + (1 - \lambda_1 - \lambda_2) f_3(x) \]

\( x \) is an 8 \( \times \) 1 vector of measurements
We will focus on 2-component mixtures

**The Model:** \[ f(x) = \lambda f_1(x) + (1 - \lambda) f_2(x) \]

**The Data:** \( x_1, \ldots, x_n \) \( m \times 1 \) vectors of measurements

**The Problem:** Fit the model to the data, making minimal assumptions on \( f_1 \) and \( f_2 \).

**The Issues:** Identifiability and computability

**Want List:** Estimates of \( f_1, f_2 \) and marginal estimates of means, standard deviations,...
Identifiability

Suppose

\[
f(x) = \lambda \prod_{j=1}^{m} f_{1j}(x_j) + (1 - \lambda) \prod_{j=1}^{m} f_{2j}(x_j)
\]

conditionally independent measurements but not necessarily identically distributed. No assumptions on the marginal distributions.

Result: A k-component mixture is identifiable provided \( m \geq m_k \) where \( m_k \geq m_k^* \) and \( 2^{m_k^*} - 1 \geq km_k^* + 1. \)

\[
(k, m_k^*) : (2, 3), (3, 4), (4, 5), (5, 5)\ldots
\]

Hall, Neeman, Pakyari, and Elmore (2005)
Conditional Independence

\[ f(x_1, x_2) = \lambda f_{11}(x_1)f_{12}(x_2) + (1 - \lambda)f_{21}(x_1)f_{22}(x_2) \]

\[ EX_1 = \lambda \mu_{11} + (1 - \lambda)\mu_{21} \]
\[ EX_2 = \lambda \mu_{12} + (1 - \lambda)\mu_{22} \]

\[ VarX_1 = \lambda_1 \sigma_{11}^2 + \lambda_2 \sigma_{21}^2 + \lambda_1 \lambda_2 (\mu_{11} - \mu_{21})^2 \]
\[ VarX_2 = \lambda_1 \sigma_{12}^2 + \lambda_2 \sigma_{22}^2 + \lambda_1 \lambda_2 (\mu_{12} - \mu_{22})^2 \]

\[ Cov(X_1, X_2) = \lambda_1 \lambda_2 (\mu_{11} - \mu_{21})(\mu_{12} - \mu_{22}) \]

Note: \( Cov(X_1, X_2) = 0 \) for scale mixtures.
Let $S_0$ denote the estimate of the covariance matrix assuming conditional independence and let $S$ denote the usual sample covariance matrix.

**Hope:** $S_0$ and $S$ are close.

A check on conditional independence:

Bootstrap 95% confidence interval for $\frac{\lambda_{(m)}}{\lambda^0_{(m)}}$, the ratio of maximum eigen values for $S$ and $S_0$.

**Want:** 95% confidence interval to contain 1.
If the 95% confidence interval contains 1 then we proceed to fit the model assuming conditional independence. Otherwise, we may have identifiability problems.

**A possibility:** Transform $Y = S_0^{1/2} S^{-1/2} X$

Then $Y$ has the covariance structure roughly corresponding to conditional independence (at least conditionally uncorrelated).

Fit the conditionally independent model to the $Y$ data. The $Y$ data are like vectors of scores made up of linear combinations of the original measurements.
Water Level Data
Two analyses: first using all $m = 8$ measurements and secondly using $m = 4$ measurements corresponding to clock settings 1, 2, 4, 5 on the right side of the clock.

$m = 8$ measurements:

$\frac{\lambda_{(m)}}{\lambda_{0(m)}} = 1.55$ and

95% confidence interval: $(1.15, 1.98)$
$m = 4$ measurements

\[
\frac{\lambda_{(m)}}{\lambda^0_{(m)}} = 1.03 \text{ and }
\]

95% confidence interval: $(0.96, 1.28)$
\[
S = \begin{pmatrix}
233 & 63 & -4 & -37 \\
63 & 714 & -96 & -20 \\
-4 & -96 & 581 & 15 \\
-37 & -20 & 15 & 354 \\
\end{pmatrix}
\]

eigen values: 772, 533, 361, 216

\[
S_0 = \begin{pmatrix}
233 & 49 & -18 & -52 \\
49 & 714 & -44 & -72 \\
-18 & -44 & 581 & 41 \\
-52 & -72 & 141 & 354 \\
\end{pmatrix}
\]

eigen values: 752, 569, 349, 213

Proceed with the analysis of the 4 measurement data.
**Computability**

Again, the discussion will be confined to 2 components.

\[
L = \prod_{i=1}^{n} (\lambda \prod_{j=1}^{m} f_{1j}(x_{ij}) + (1 - \lambda) \prod_{j=1}^{m} f_{2j}(x_{ij}))
\]

If we know which component \( x_{ij} \) belongs to then letting \( z_i = 1 \) if the first component and 0 otherwise, the complete likelihood is:

\[
L_c = \prod_{i=1}^{n} \prod_{j=1}^{m} \{ f_{1j}(x_{ij})^{z_i} f_{2j}(x_{ij})^{(1-z_i)} \lambda^{z_i} (1 - \lambda)^{(1-z_i)} \}
\]

"EM algorithm" next
Initial Values:

a. Use a 2-means clustering algorithm and let $z_i^{(0)} = 1$ if the vector of measurements $x_i$ is in the first cluster and 0 otherwise.

Then compute $\lambda^{(0)} = \text{ave}(z_i^{(0)})$

b. Using $z_i^{(0)} \; i = 1, \ldots, n$ and $\lambda^{(0)}$ compute:

$$f_{1j}^{(0)}(u) = \frac{1}{\lambda^{(0)} nh} \sum_{i=1}^{n} z_i^{(0)} K\left(\frac{u - x_{ij}}{h}\right)$$

Similarly for $f_{2j}^{(0)}(u)$. 
Updating and Iterations:

E step:

\[ z_i^{(t+1)} = \frac{\lambda^{(t)} \prod_{j=1}^{m} f_{1j}^{(t)}(x_{ij})}{\lambda^{(t)} \prod_{j=1}^{m} f_{1j}^{(t)}(x_{ij}) + (1 - \lambda^{(t)}) \prod_{j=1}^{m} f_{2j}^{(t)}(x_{ij})} \]

"M step"

\[ \lambda^{(t+1)} = \text{ave}(z_i^{(t+1)}) \]

\[ \mu_{1j}^{(t+1)} = \frac{\sum_{i=1}^{n} z_i^{(t+1)} x_{ij}}{\sum_{i=1}^{n} z_i^{(t+1)}} \]

\[ f_{1j}^{(t+1)}(u) = \frac{1}{\lambda^{(t+1)} nh} \sum_{i=1}^{n} z_i^{(t+1)} K\left(\frac{u - x_{ij}}{h}\right) \]

Stopping: When change in \( \lambda^{(t)} \), and \( \mu_{kj}^{(t)} \) for \( k = 1, 2 \) and \( j = 1, \ldots, m \) is sufficiently small.
Attractive since:

Fast to compute for general $m$, the number of measurements, and $k$, the number of components.

Performed well in simulation studies

Easy to determine features of the component marginal distributions, eg. means, medians, stdevs, pdfs, and cdfs.

Motivated by work of Bordes, Chauveau, Vandekerkhove (2007)
Water Level Data, 4 measurements

Friedman Test:
$S = 237.40 \text{ DF } = 3 \text{ P } = 0.000$

<table>
<thead>
<tr>
<th>meas</th>
<th>n</th>
<th>Est med</th>
<th>Sum of Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>405</td>
<td>-0.125</td>
<td>1079</td>
</tr>
<tr>
<td>2</td>
<td>405</td>
<td>4.125</td>
<td>1297</td>
</tr>
<tr>
<td>3</td>
<td>405</td>
<td>-3.625</td>
<td>755.5</td>
</tr>
<tr>
<td>4</td>
<td>405</td>
<td>-1.875</td>
<td>918.5</td>
</tr>
</tbody>
</table>

Suggesting that the 4 measures are not identically distributed.

Eigen values for $S_0$ and $S$ suggest that we need not reject the assumption of conditionally uncorrelated measures.
Proceed with fitting the model to the data...  
4 measurements, 3 components

Means:

<table>
<thead>
<tr>
<th>Meas</th>
<th>Comp 1</th>
<th>Comp 2</th>
<th>Comp 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.9</td>
<td>0.2</td>
<td>21.7</td>
</tr>
<tr>
<td>2</td>
<td>8.5</td>
<td>2.0</td>
<td>32.0</td>
</tr>
<tr>
<td>3</td>
<td>-13.2</td>
<td>-1.8</td>
<td>-19.1</td>
</tr>
<tr>
<td>4</td>
<td>-6.2</td>
<td>-1.5</td>
<td>-31.9</td>
</tr>
</tbody>
</table>

Standard Deviations:

<table>
<thead>
<tr>
<th>Meas</th>
<th>Comp 1</th>
<th>Comp 2</th>
<th>Comp 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.9</td>
<td>6.4</td>
<td>25.8</td>
</tr>
<tr>
<td>2</td>
<td>28.5</td>
<td>6.4</td>
<td>54.5</td>
</tr>
<tr>
<td>3</td>
<td>23.0</td>
<td>6.5</td>
<td>56.0</td>
</tr>
<tr>
<td>4</td>
<td>23.4</td>
<td>6.9</td>
<td>17.3</td>
</tr>
</tbody>
</table>

Lambdas: .42, .49, .09
CDF plots for the first measurement
1 o’clock
CDF plots for the second measurement
2 o’clock
CDF plots for the third measurement
4 o’clock
CDF plots for the fourth measurement
5 o’clock
If we assume iid measures, then the 4 plots are combined:

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Stdev</th>
<th>Lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.8</td>
<td>-1.0</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td>-2.6</td>
<td>-2.6</td>
<td>.44</td>
</tr>
<tr>
<td></td>
<td>45.8</td>
<td>4.9</td>
<td>.41</td>
</tr>
<tr>
<td></td>
<td>20.7</td>
<td>20.7</td>
<td>.41</td>
</tr>
</tbody>
</table>
Other work: Qin and Leung (2006)
2 components and 3 measurements

Conditionally independent model:
\[ f(x) = \lambda \prod_{j=1}^{3} f_j(x_j) + (1 - \lambda) \prod_{j=1}^{3} g_j(x_j) \]

Exponential tilt:
\[ g(x_j) = f_j(x_j) \exp(\beta_{0j} + \beta_{1j}x_j + \beta_{2j}x_j^2) \]

The algorithm:
1. determine initial values for \( \lambda, \beta_{0j}, \beta_{1j}, \beta_{2j} \)
   \( j = 1, 2, 3 \)
2. use empirical likelihood to estimate \( F_j \)
3. use EM to estimate \( \lambda, \beta_{0j}, \beta_{1j}, \beta_{2j} \)
   \( j = 1, 2, 3 \)

\[ L = \prod_{i=1}^{n} \prod_{j=1}^{3} \left( \lambda + (1 - \lambda)e^{\beta_{0j} + \beta_{1j}x_j + \beta_{2j}x_j^2} \right) dF_j(x_{ij}) \]
The Univariate Case

Identifiability:
Model:
\[ f(x) = \lambda g(x - \mu_1) + (1 - \lambda) g(x - \mu_2) \]

where \( g \) is symmetric about 0.

Bordes, Mottelet, Vandekerkhove (2006)

Computatatability:
Very expensive. Algorithms only for 2 component case.

Two possibilities:
1. "EM" algorithm (Bordes, Chauveau, Vandekerkhove (2007))

Suppose we have initial values for $\mu_1, \mu_2,$ and $g(.)$.

**E step**

$$z_i^{(t+1)} = \frac{\lambda^{(t)} g^{(t)}(x_i - \mu_1^{(t)})}{\lambda^{(t)} g^{(t)}(x_i - \mu_1^{(t)}) + (1 - \lambda^{(t)}) g^{(t)}(x_i - \mu_2^{(t)})}$$

"M step"

$$\lambda^{(t+1)} = \text{ave}(z_i^{(t+1)}) \quad \text{and} \quad \mu_1^{(t+1)} = \text{ave}(z_i^{(t+1)} x_i)$$

$$g^{(t+1)}(u) = \frac{1}{2nh} \sum_{i=1}^{n} \sum_{j=1}^{2} z_{ij}^{(t+1)} \left\{ K\left(\frac{u - x_i - \mu_j^{(t+1)}}{h}\right) + K\left(\frac{-u - x_i - \mu_j^{(t+1)}}{h}\right) \right\}$$
2. Exponential Tilt Model

\[ f(x) = \lambda g_0(x)e^{\beta_0 + \beta_{11}x + \beta_{21}x_2} + \\
(1 - \lambda)g_0(x)e^{\beta_{02} + \beta_{12}x + \beta_{22}x_2} \]

\( g_0(x) \) is called the carrier density.


**Computation**: discretize the data, use a kernel density estimator for \( g_0(x) \), and an EM algorithm to estimate \( \beta_s \). Computation is carried out via a mixture of Poisson regressions.

Computation is fast and works for \( k \) components.
Example: Time between eruptions of Old Faithful Geyser

Issues: identifiability, estimation of the carrier...
The End