STATISTICAL HORIZONS

Finite Mixture Modeling

Jeffrey Harring, Ph.D.

Upcoming Seminar:

May 3-4, 2019, Philadelphia, Pennsylvania

May 3-4, 2019
Temple University
Philadelphia, PA

Finite Mixture Modeling

JEFFREY R. HARRING
University of Maryland

General Mixture Specification

- **General Mixture Specification**

$$f(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \varphi_k f_k(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_k)$$

Composite distribution with parameters $\boldsymbol{\varphi}$ **and** $\boldsymbol{\theta}$ is the **sum** of *K* **component distributions** each with component-specific parameters $\boldsymbol{\theta}_k$ and weights or **mixing proportions**, φ_k, with

$$0 \le \varphi_k \le 1 \qquad \varphi_k = 1 - \sum_{j=1}^{K-1} \varphi_j$$

General Mixture Specification

$$f(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\varphi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \varphi_k f_k(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_k)$$

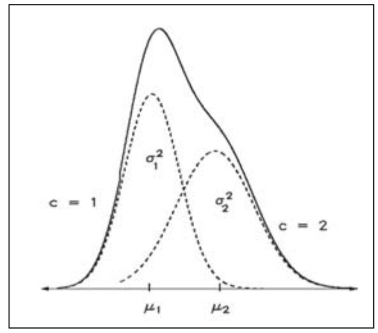
- **The class (component)-specific probability distributions can take on a variety of forms to represent different types of models and modeling for different purposes: distribution estimation, statistical and measurement-based modeling such as regression, IRT, SEM, LGM, etc.**

Class-component probability distributions typically come from the same family, but do not have to

Mixtures of Univariate Distributions

Univariate Mixtures

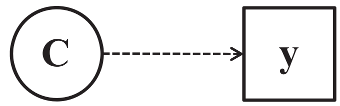
- For example, consider a composite distribution to be a mixture of two normal distributions



Class Proportion

$$f_k(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_k) = \sum_{k=1}^{2} \varphi_k f_k(\mathbf{y} \mid \boldsymbol{\theta}_k)$$

$$= \varphi_1 f_1(\mathbf{y} \mid \boldsymbol{\theta}_1) + (1-\varphi_1) f_2(\mathbf{y} \mid \boldsymbol{\theta}_2)$$



$$f_k(y_i \mid \boldsymbol{\theta}_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(y_i - \boxed{\mu_k})^2}{2\sigma_k^2}\right\}$$

Class-specific variances

Class-specific means

Systematic Measurement Error

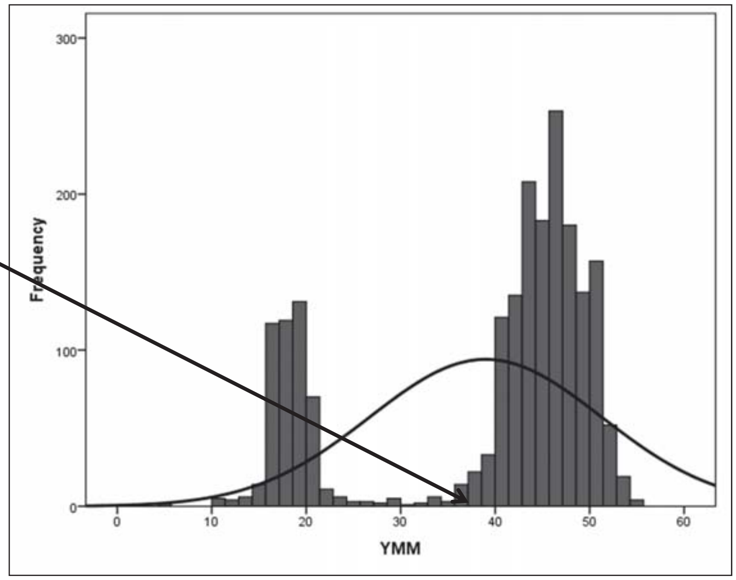
- Liu, Hancock, & Harring (2010) discussed how mixtures of univariate normal distributions could be used to examine potential systematic measurement problems
- Issues with the existing data from the National Survey of Child and Adolescent Well-being from initial exploration
- Survey collected longitudinal data from children who were subject to child abuse and/or neglect
- Key variable: head circumference of children up to 4 years old in centimeters

Head Circumference

- **Overall, treating as 1 normal distribution**

$$\bar{y} = 38.99$$
$$s = 150.97$$

Clearly these statistics do not adequately summarize head circumference



Research Questions

Univ-ME-2-Class.inp ME-Data-Mplus.dat

- **Were head circumference measurements sampled from a homogeneous normally distributed population with a single mean and variance?**
- **OR were head circumference measurements sampled from a heterogeneous population consisting of multiple latent classes (i.e., unknown groups), each characterized by a normal distribution with a unique mean and variance?**

Fit successively greater number of classes allowing for class-specific means and variances

Use fit indices and substantive considerations to choose among models

Mplus

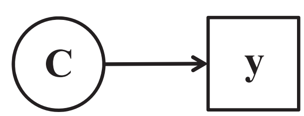
- Mplus commands are written in scripts with sections known as commands. These commands include:
  - TITLE:
  - DATA:
  - VARIABLE:
  - ANALYSIS:
  - MODEL:
  - OUTPUT:

The comment character is !

- Other commands that we'll discuss
  - DEFINE:
  - SAVEDATA:
  - PLOT:

Mplus Input for Head Circumference

```
TITLE:       Univariate normal mixture model

DATA:        FILE IS ME-Data-Mplus.dat;

VARIABLE:    NAMES = ID sex age y;
             MISSING ARE ALL (-99);
             USEVARIABLES ARE y;
             CLASSES = c(2);

ANALYSIS:    TYPE = MIXTURE;
!Default estimator is an accelerated EM algorithm;
             STARTS = 50 10; STSCALE = 5;
             STITERATIONS = 50; ITERATIONS = 1000;
             SDITERATIONS = 250; MITERATIONS = 500;
             MCONVERGENCE = 1E-5;
```



Mplus Input Cont.

```
MODEL:
%overall%
y*; [y*];
%c#1%
y*15; [y*45];
%c#2%
y*5.6; [y*18];
```

Must accompany mixture models

Class-specific models can be specified

Variance of a variable is specified by its name while the mean (or intercept) is specified with brackets [ ]. An asterisk allows the user to specify an initial starting value for the parameter

Mplus Input Cont.

```
OUTPUT: TECH1 TECH4;

SAVEDATA:

FILE IS ME-Data-Mplus-2.dat;

SAVE = CPROB;
```

The <u>CPROB</u> argument tells Mplus to save posterior probabilities and most likely class membership to file…

ME-Data-Mplus-2.dat

Mplus Output Excerpts

Test of Model Fit:

```
THE MODEL ESTIMATION TERMINATED NORMALLY

Number of Free Parameters                              5

Loglikelihood

          H0 Value                             -6501.699

Information Criteria

          Akaike (AIC)                         13013.398
          Bayesian (BIC)                       13041.472
          Sample-Size Adjusted BIC             13025.587
            (n* = (n + 2) / 24)
         Entropy                                   0.998
```


Mplus Output Excerpts

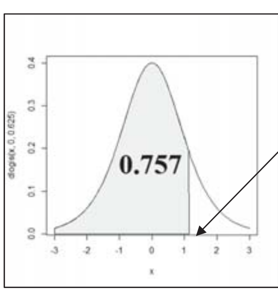
```
FINAL CLASS COUNTS AND PROPORTIONS FOR THE
LATENT CLASSES BASED ON THE ESTIMATED MODEL
    Latent
   Classes
       1       1534.55873          0.75669
       2        493.44127          0.24331
```

$$2028 \times .75669 = 1534.5587$$

```
Categorical Latent Variables

 Means
    C#1                        1.135      0
```

The categorical latent variable means are on the logit scale…

Converting to the probability scale requires…



$$\text{Prob} = \frac{\exp(1.135)}{1 + \exp(1.135)} = 0.75669$$

Mplus Output Excerpts

```
FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASS
PATTERNS BASED ON ESTIMATED POSTERIOR PROBABILITIES

   Class 1     1534.55873          0.75669
   Class 2      493.44127          0.24331
```

	Posterior Probabilities		
Subject	1	2	Modal Class Assignment
1	1.000	0.000	1
2	1.000	0.000	1
3	1.000	0.000	1
2028	0.000	1.000	2
	Avg. Class 1 0.75669	Avg. Class 2 0.24331	

Mplus Output Excerpts

```
CLASSIFICATION OF INDIVIDUALS BASED ON THEIR MOST LIKELY
LATENT CLASS MEMBERSHIP

Class Counts and Proportions

   Class 1      1535           0.75690
   Class 2       493           0.24310
```

Subject	Modal Class Assignment
1	1
2	1
3	1
⋮	⋮
2028	2

How many subjects have a modal class assignment of 1? Of 2?

Mplus Output Excerpts

```
Average Latent Class Probabilities for
Most Likely Latent Class Membership
(Row) by Latent Class (Column)

                        1         2

   Class 1          0.999     0.001
   Class 2          0.001     0.999
```

Latent Class Model Assignment	*n*	Latent Class Average Posterior Probabilities 1	2	3
1	4	.936	.064	.000
2	5	.083	.841	.076
3	4	.031	.083	.886

Average posterior probability for Class 1 for cases assigned to Class 1 (want high)

Average posterior probability for Class 1 for cases not assigned to Class 1 (want low)

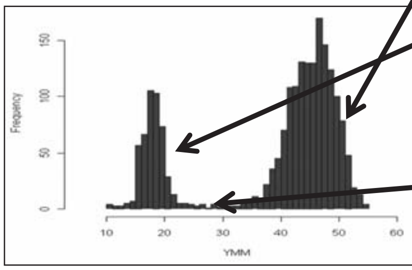
Subject	Posterior Probabilities 1	2	3
1	.999	.001	.000
2	.743	.255	.002
3	1.000	.000	.000
4	1.000	.000	.000
5	.295	.557	.158
6	.015	.800	.185
7	.000	1.000	.000
8	.105	.856	.039
9	.001	.999	.000
10	.001	.001	.998
11	.100	.300	.600
12	.020	.030	.950
13	.001	.000	.999

Mplus Output Excerpts

```
                        Estimate      S.E.   Est./S.E.    P-Value
Latent Class 1

 Means
    Y                     45.660     0.104     437.994      0.000

 Variances
    Y                     14.925     0.792      18.836      0.000

Latent Class 2

 Means
    Y                     18.252     0.113     161.483      0.000

 Variances
    Y                      5.666     1.113       5.089      0.000

Categorical Latent Variables

 Means
    C#1                    1.135     0.052      21.768      0.000
```

Posterior Probabilities

- **Using model parameters and a person's data, we can calculate their probability of membership in each class (more on this when we talk through estimation)**



Subject	HC	Posterior Probabilities 1	Posterior Probabilities 2	Modal Class Assignment
1	53.3	**1.000**	0.000	1
2	47.0	**1.000**	0.000	1
3	39.0	**1.000**	0.000	1
14	21.0	0.000	**1.000**	2
15	50.0	**1.000**	0.000	1
68	29.0	**0.824**	0.176	1
69	20.0	0.000	**1.000**	2
89	28.0	0.196	**0.804**	2
124	45.3	**1.000**	0.000	1
659	49.0	**1.000**	0.000	1
670	17	0.000	**1.000**	2
900	26.5	**0.565**	0.435	1
1000	16	0.000	**1.000**	2