Identifying the Correct Number of Classes in a Growth Mixture Model

Davood Tofighi
Craig Enders
Arizona State University

Mixture Modeling

• Heterogeneity exists such that the data are comprised of two or more latent classes with different distributions

\[ f(y) = \sum_{k=1}^{K} \pi_k \phi_k (y; \mu_k, \Sigma_k) \]
Growth Mixture Modeling (GMM)

- $k$ latent classes with different growth trajectories and variance components
- Class membership may be related to covariates and distal outcomes
- GMM is analogous to a multiple-group growth model, but group membership is unobserved

Religiousness Example

- Three classes of religious development
Enumerating Latent Classes

- How many trajectory classes are there?
- Information-based criteria
  - BIC, AIC, etc.
- Likelihood ratio tests
  - Lo, Mendell, Rubin (2001)
- Goodness of fit tests
  - Tests based on model-implied skewness and kurtosis (Muthén, 2003)

Bayesian Information Criterion

- Based on the log likelihood and penalty terms related to model complexity
  \[ \text{BIC} = -2LL + p \ln(N) \]
- The sample-size adjusted BIC (SABIC) replaces \( N \) with \( (N + 2) / 24 \)
Akaike Information Criterion

- Similar idea as the BIC ...
  \[ AIC = -2LL + 2p \]

- The consistent AIC (CAIC) is
  \[ CAIC = -2LL + p(\ln[N] + 1) \]

- A sample-size adjusted CAIC (SACAIC) replaces \( N \) with \( (N + 2) / 24 \)

Likelihood Ratio Tests

- Likelihood ratio tests can be used to compare a \( k \) versus \( k - 1 \) class model
  \[ LRT = -2(LL_{k-1} - LL_k) \]

- The LRT is not chi-square distributed

- Class probabilities for the nested \( k - 1 \) class model are at the boundary (zero)
Lo, Mendell, and Rubin (2001)

- Derived an appropriate reference distribution for the LRT, and an ad hoc adjustment to the test statistic
- LMR and adjusted LMR (ALMR)
- A small $p$ value suggests that the $k$ class model is favored over $k - 1$ classes

Bootstrapping The LRT

- The $k$ and $k - 1$ class models are fit to a number of bootstrap samples
- A $p$ value for the LRT is obtained from the empirical reference distribution of bootstrapped LRT values
- See Nylund, Asparouhov, and Muthén (2006) for detailed simulation results
Fit Tests Based On Multivariate Skewness And Kurtosis

- Model-implied skewness and kurtosis from the $k$ class model are compared to the sample moments (Muthén, 2003)
- Analogous to the GOF test in SEM
- A large $p$ value indicates that the $k$ class model accurately reproduces the higher-order moments
- Herein referred to as MST and MKT

An Artificial Data Example

- An artificial data set ($N = 1000$) was generated from a population with three trajectory classes
- A sequence of models was fit ($k = 1$ to $4$) to illustrate the class extraction process
Analysis Results

<table>
<thead>
<tr>
<th></th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 3$</th>
<th>$k = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC</td>
<td>27333.77</td>
<td>26680.68</td>
<td><strong>26623.59</strong></td>
<td>26649.66</td>
</tr>
<tr>
<td>SABIC</td>
<td>27311.53</td>
<td>26633.03</td>
<td><strong>26550.54</strong></td>
<td>26560.73</td>
</tr>
<tr>
<td>LMR</td>
<td>N/A</td>
<td>$p &lt; .001$</td>
<td>$p &lt; .001$</td>
<td>$p = .09$</td>
</tr>
<tr>
<td>BLRT</td>
<td>N/A</td>
<td>$p &lt; .001$</td>
<td>$p = 1.00$</td>
<td>$p = 1.00$</td>
</tr>
<tr>
<td>MST</td>
<td>$p &lt; .001$</td>
<td>$p = .26$</td>
<td>$p = .82$</td>
<td>$p = .86$</td>
</tr>
<tr>
<td>MKT</td>
<td>$p &lt; .001$</td>
<td>$p = .92$</td>
<td>$p = .49$</td>
<td>$p = .55$</td>
</tr>
</tbody>
</table>

More On The LRT

- Multiple $k - 1$ class models are possible
- For example, when testing a 3 class model, three different 2 class models could result
- *Mplus* discards the first class when fitting the $k - 1$ class model
- Starting values must be used to ensure that classes are ordered properly
### Example

<table>
<thead>
<tr>
<th>Correct Ordering</th>
<th>Incorrect Ordering</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 1-class model</td>
<td>• 1-class model</td>
</tr>
<tr>
<td>– $LL_1 = -13642.71$</td>
<td>– $LL_1 = -13642.71$</td>
</tr>
<tr>
<td>• 2-class model</td>
<td>• 2-class model</td>
</tr>
<tr>
<td>– $LL_1 = -13642.71$</td>
<td>– $LL_1 = -13642.71$</td>
</tr>
<tr>
<td></td>
<td>– $LL_2 = -13288.53$</td>
</tr>
<tr>
<td>• 3-class model</td>
<td>• 3-class model</td>
</tr>
<tr>
<td>– $LL_2 = -13288.53$</td>
<td>– $LL_2 = -13265.72$</td>
</tr>
<tr>
<td></td>
<td>– $LL_3 = -13232.36$</td>
</tr>
</tbody>
</table>

### Purpose Of Study

• In the previous example, the fit indices did not agree on the number of classes.
• Which index should be used to determine the number of classes?
• We designed a Monte Carlo simulation study to address this question.
Simulation Procedure

- Data were generated from a population with three trajectory classes
- A sequence of GMMs was fit ($k = 2$ to $4$)
- Extraction was performed with and without covariates
- In what proportion of replications was the $k = 3$ class model recovered?

Data Generation Model
Population Trajectory Classes

Manipulated Variables

- **Number of repeated measures**
  - $t = 4, 7$
- **Sample size**
  - $N = 400, 700, 1000, 2000$
- **Mixing proportions**
  - 20%, 33%, 47% and 7%, 36%, 57%
- **Within-class normality**
  - $S = 0, K = 0$ and $S = 1, K = 1$
- **Class separation**
  - “High” and “Low”
Manipulating Class Separation

- Definition of separation is subjective, but based on previous experience (e.g., McCullough, Enders, & Brion, 2005)
- Within-class variance components were increased in magnitude to create the low separation condition
- Mean growth trajectories did not change

Class 1: High Versus Low Separation

High Separation

Low Separation
Class 2: High Versus Low Separation

High Separation

Low Separation

Class 3: High Versus Low Separation

High Separation

Low Separation
Average Class Probabilities

<table>
<thead>
<tr>
<th>Low Separation</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.85</td>
<td>.03</td>
<td>.12</td>
</tr>
<tr>
<td>2</td>
<td>.02</td>
<td>.81</td>
<td>.18</td>
</tr>
<tr>
<td>3</td>
<td>.07</td>
<td>.13</td>
<td>.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High Separation</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.91</td>
<td>.08</td>
<td>.01</td>
</tr>
<tr>
<td>2</td>
<td>.05</td>
<td>.89</td>
<td>.08</td>
</tr>
<tr>
<td>3</td>
<td>.01</td>
<td>.11</td>
<td>.89</td>
</tr>
</tbody>
</table>

Sample Size (No Covariates)

Normality, High Separation, $t = 4$, Proportions = 20:33:47

- BIC
- SABIC
- AIC
- CAIC
- SACAIC
- LMR
- MST
- MKT

- $N = 400$
- $N = 700$
- $N = 1000$
- $N = 2000$
Sample Size (With Covariates)
Normality, High Separation, $t = 4$, Proportions = 20:33:47

Class Separation (No Covariates)
Normality, $N = 1000$, $t = 4$, Proportions = 20:33:47
Class Separation (Covariates)

Normality, $N = 1000$, $t = 4$, Proportions = 20:33:47

Mixing Proportion (No Covariates)

Normality, $N = 1000$, $t = 4$, High Separation
Mixing Proportion (Covariates)
Normality, \( N = 1000, t = 4, \) High Separation

Normality (No Covariates)
\( N = 1000, t = 4, \) Proportions = 20:33:47, High Separation
Normality (Covariates)

$N = 1000, t = 4$, Proportions $= 20:33:47$, High Separation

- The SABIC very accurately detected the number of latent classes
- At small $N$s, it had a slight tendency to extract too few classes
- Note that the $k$ class model was retained if the SABIC decreased by any amount
- Other information-based criteria performed poorly
Likelihood Ratio Tests

- The LMR and ALMR performed well, but were somewhat less powerful than the SABIC
- LMR tended to extract too few classes at small $Ns$, and too many classes at large $Ns$

MSK and MKT

- MSK and MKT uniformly extracted too few classes
- The performance of these measures may be model-dependent
- MSK was slightly more accurate than the LMR in a pilot study with a slightly different set of mixture distributions
The Use Of Covariates

- The inclusion of covariates dramatically decreased power, and resulted in the extraction of too few classes
- e.g., At N = 400, the SABIC extracted two classes 39% of the time, compared to 14% when covariates were excluded
- The use of covariates should be avoided unless N is very large

Nonnormal Data

- Mild violations of within-class normality led to the extraction of too many classes
- None of the tests we studied was immune to this problem
- Is bootstrapping the LRT a solution?
Bootstrapping The LRT

- Preliminary results suggest that the bootstrap is less powerful than the LMR
- e.g., The $k = 2$ class model was correctly rejected about 75% and 10% of the time in the high and low separation conditions, respectively
- See Nylund, Asparouhov, and Muthén (2006) for detailed simulation results

How Likely Is Over-Extraction?

- Extracting too many classes frequently produced problematic solutions (e.g., negative variances, unstable solutions)
- Estimating class-specific variances probably prevents over-extraction
- Invoking constraints to attain convergence (e.g., fixing variances to zero) is likely a sign of over-extraction