CYFS nebraska center for research children youth families & schools

research methodology series

Introduction to Mixture Modeling

Kevin A. Kupzyk, MA

Methodological Consultant, CYFS SRM Unit

Outline

- Variable- vs. person-centered analyses
- Traditional methods
- Latent Class Analysis vs. Latent Profile Analysis
- Mixture modeling

- Data structure and analysis examples
- Longitudinal extensions

Person-centered analysis

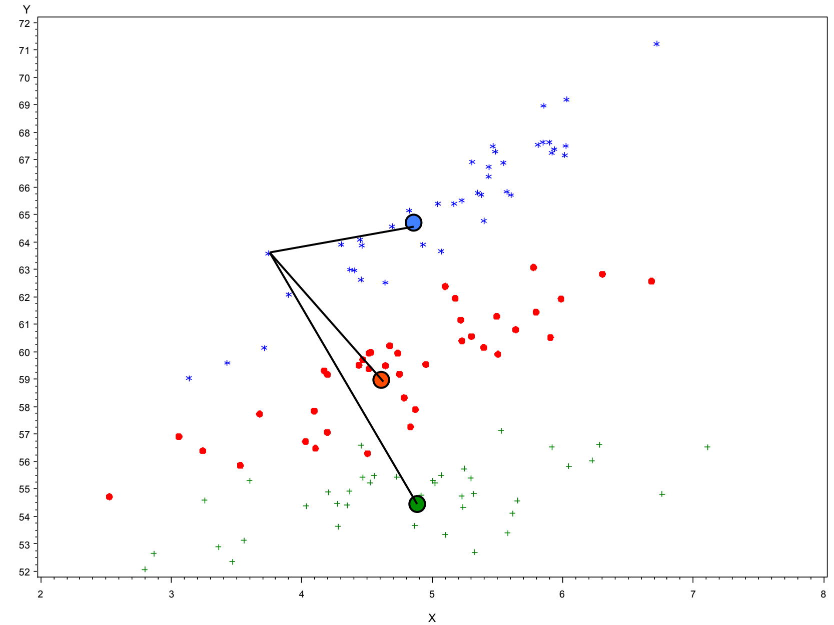
- Person*item data structure
- Variable-centered: correlations among variables are of most interest
  - Factor analysis
  - Structure among columns
  - Predicting outcomes
- Person-centered: Structure among rows is of most interest
  - Relationships among individuals
  - Grouping individuals based on shared characteristics
  - Identifying qualitatively different groups

		Factor 1			Factor 2	
	child_id	ScaleA	ScaleB	ScaleC	ScaleD	ScaleE
	101	4.00	4.57	1.75	4.50	4.40
	102	4.88	4.57	5.00	4.88	4.40
	103	3.63	4.14	2.25	4.13	3.80
	104	4.00	3.57	2.00	3.25	.60
Group 1	105	3.75	4.14	2.50	3.88	3.40
	106	4.63	4.57	2.50	5.00	4.60
	107	4.63	4.29	3.75	4.13	3.80
	108	4.25	3.71	3.50	4.75	4.20
	109	4.88	4.71	4.75	4.50	4.80
	110	4.50	4.71	3.75	4.75	5.00
Group 2	111	3.63	4.00	4.50	2.63	3.80
	112	3.88	4.00	4.00	3.75	3.60
	113	5.00	5.00	1.50	5.00	5.00
	114	3.88	5.00	2.50	3.75	3.40
	115	4.00	4.57	3.75	4.75	3.80
	116	2.75	3.57	.50	3.00	2.00
Group 3	117	5.00	5.00	4.25	5.00	4.40
	118	4.13	2.57	1.00	3.00	4.00
	119	.86	4.00	1.25	.00	2.33
	120	3.25	4.29	4.25	1.71	2.80
	121	3.75	3.29	.75	4.00	4.40

Traditional Methods

- K-means clustering
- Hierarchical clustering

  - Using Euclidean distance
    - Distance between the individual and the cluster mean
  - All variables need to be on the same scale
  - Continuous variables only
  - Dependent on start values
  - No fit statistics available
  - Sample dependent
    - Not model based
    - Not replicable



What is mixture modeling?

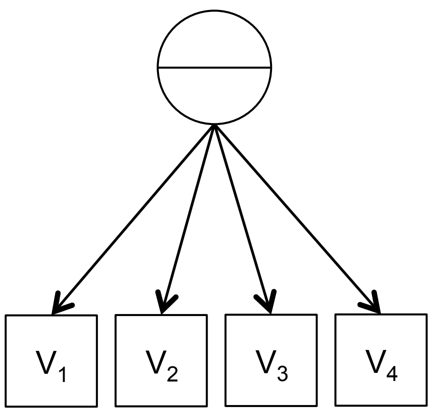
- Modeling a “mixture” of sub-groups within a population
- “Finite” number of homogeneous categories.
- Assumes the population is a mixture of qualitatively different groups of individuals
- Identified based on similarities in response patterns
- You might hypothesize that your population is made up of different types of individuals, families, etc.
  - Demographic or academic risk factors often co-occur (diagnostic comorbidity)

- Latent Class Analysis (LCA) and Latent Profile Analysis (LPA) are special cases of mixture models

Terminology

Outcome/Dependent Variable

			Continuous	**Categorical**
Predictor Variable(s)	Observed	**Continuous**	Regression	Logistic Regression
		Categorical	ANOVA/Regression	Non-Parametric (e.g. Chi-Square)
	Latent	**Continuous**	Factor Analysis	Item Response Theory
		Categorical	Latent Profile Analysis	Latent Class Analysis



"Finite Mixture Models"

Getting started

- First pick appropriate measures
  - Demographics
  - Outcome measures
  - Stuff you're interested in
- Pick a software program
  - *M*plus*
  - Latent Gold
  - SAS (LCA, LTA, TRAJ)

Evaluating model fit

- *BIC, AIC (Information Criteria)
  - To compare competing models
  - Look for lowest value
- Entropy
  - Measure of classification uncertainty
  - Ranges from 0 to ∞, lower is better
- Relative Entropy
  - Ranges from 0 to 1, higher is better
  - This is what M*plus* provides, but it's called "Entropy"

Evaluating model fit

- Likelihood ratio test
  - Problematic due to categorical latent variable
- (Vuong-)Lo-Mendel-Rubin likelihood ratio test
  - TECH11 in M*plus*
  - Compares estimated model with a model with one less class
  - $p<.05$ indicates the model with more classes fits significantly better
- Bootstrap Likelihood ratio test
  - TECH14 in M*plus*
  - Compares estimated model with a model with one less class
  - Often inconclusive

LPA example

- 220 Preschool Children
- 51 outcome variables
  - La Familia – Family Literacy Activities
  - Parental Stress Index
  - Maternal Depression
  - Parent-Teacher Relationship Scale
  - Bracken Basic School Concepts and School Readiness
  - Teacher and parent-reported social/emotional scales

LPA example

Mplus - time1

File Edit View Mplus Graph Window Help

time1

4.3750	4.7143	1.500		
4.5000	4.8571	3.000		
4.5000	5.0000	5.000		
4.3750	4.8571	4.500		
4.3750	4.2857	3.000		
-999.0000	-999.0000	-999.000		
3.0000	3.4286	2.000		
4.7500	4.7143	3.250		
4.1250	4.2857	3.000		
5.0000	4.4286	2.500		
2.7500	2.8571	2.250		
4.6250	4.5714	3.000		
4.3750	5.0000	5.000		
5.0000	4.7143	1.750		
4.8750	4.7143	4.500		
4.3750	4.7143	3.000		
5.0000	4.8571	2.250		
3.8750	5.0000	2.500		
4.1250	4.0000	3.250		
4.8750	4.8571	4.500		
3.6250	4.1429	2.250		
5.0000	4.7143	4.000		
3.5000	4.1429	4.500		
3.5000	4.1429	4.500		
4.5000	4.7143	3.750		
4.8750	4.5714	3.000		
4.6000	5.0000	4.000		
4.7500	5.0000	3.500		
5.0000	5.0000	3.000		
4.2500	4.0000	2.0000	3.8750	4.0000
3.8750	4.2857	1.5000	4.3750	4.0000
4.3750	4.4286	1.5000	3.0000	3.0000
4.5000	4.0000	2.5000	4.0000	3.2000
5.0000	5.0000	3.0000	5.0000	5.0000
4.3750	3.8571	1.0000	2.8750	2.0000

Mplus - Mptext1

File Edit View Mplus Graph Window Help

Mptext1

```
TITLE:          Mixture Modeling - LPA Example

DATA:           File is time1.dat;
                FORMAT is 51f13.4;

VARIABLE:       NAMES are V1-V51;
                USEVARIABLES are V1-V51;
                MISSING = all(-999);
                CLASSES=c(2);

ANALYSIS:       TYPE=MIXTURE;

OUTPUT:         TECH1 TECH11 TECH14;
```

LPA example

Mplus - time1

File Edit View Mplus Graph Window Help

time1

```
THE MODEL ESTIMATION TERMINATED NORMALLY



MODEL FIT INFORMATION

Number of Free Parameters                        53

Loglikelihood

          H0 Value                        -1715.420
          H0 Scaling Correction Factor        1.393
            for MLR

Information Criteria

          Akaike (AIC)                     3536.840
          Bayesian (BIC)                   3714.988
          Sample-Size Adjusted BIC         3547.047
            (n* = (n + 2) / 24)
```

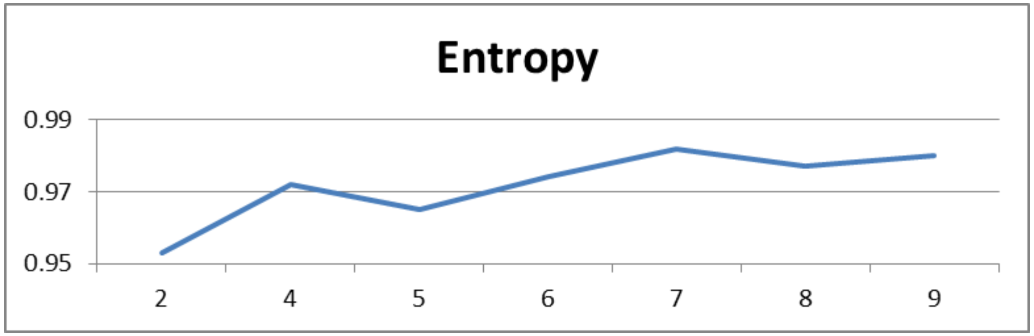
Mplus - time1

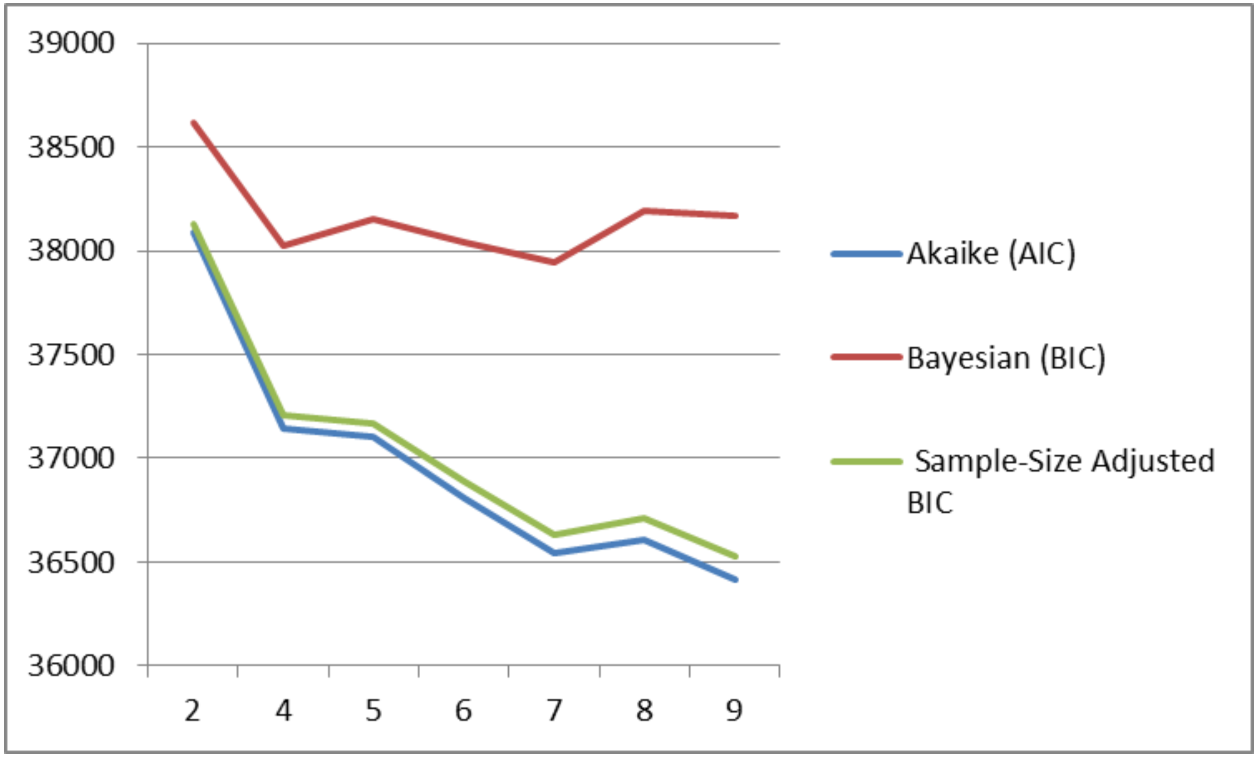
File Edit View Mplus Graph Window Help

time1

```
MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

Latent Class 1

 Means
    V1                 1.704      0.310      5.498      0.000
    V2                 2.714      0.281      9.648      0.000
    V3                 1.597      0.218      7.323      0.000
    V4                 1.730      0.306      5.655      0.000
    V5                 1.201      0.345      3.483      0.000
    V6                 2.421      0.191     12.679      0.000
    V7                 2.327      0.150     15.524      0.000
    V8                 2.305      0.080     28.888      0.000
    V9                 1.693      0.150     11.308      0.000
    V10                2.233      0.081     27.462      0.000

 Variances
    V1                 0.411      0.061      6.691      0.000
    V2                 0.285      0.035      8.151      0.000
    V3                 1.137      0.086     13.153      0.000
    V4                 0.660      0.086      7.646      0.000
    V5                 0.841      0.118      7.135      0.000
    V6                 0.168      0.025      6.611      0.000
    V7                 0.115      0.017      6.754      0.000
    V8                 0.049      0.006      7.944      0.000
    V9                 0.147      0.017      8.738      0.000
    V10                0.050      0.005      9.371      0.000

Latent Class 2

 Means
    V1                 4.360      0.059     73.979      0.000
    V2                 4.440      0.057     78.435      0.000
    V3                 3.169      0.127     24.942      0.000
    V4                 4.153      0.101     41.310      0.000
    V5                 4.138      0.084     49.058      0.000
    V6                 1.387      0.051     27.062      0.000
    V7                 1.574      0.046     34.252      0.000
    V8                 1.556      0.033     46.977      0.000
    V9                 1.159      0.027     42.173      0.000
```

LPA example


Entropy
0.99
0.97
0.95
2
4
5
6
7
8
9

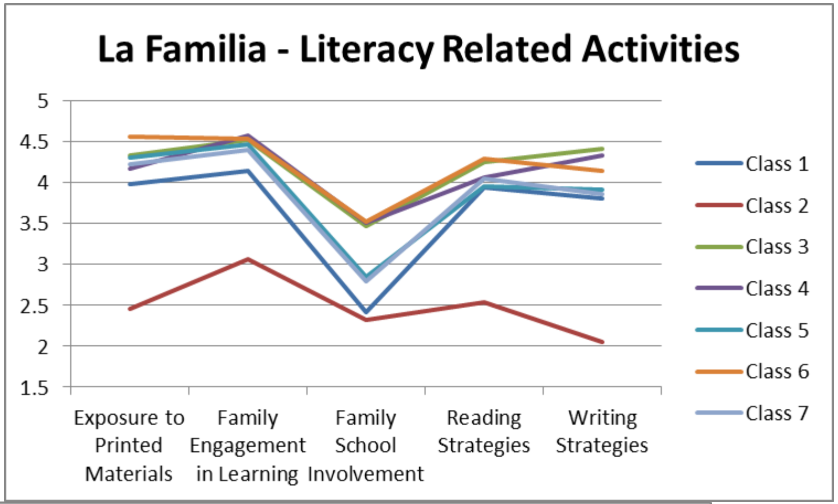

39000
38500
38000
37500
37000
36500
36000
2
4
5
6
7
8
9
Akaike (AIC)
Bayesian (BIC)
Sample-Size Adjusted BIC

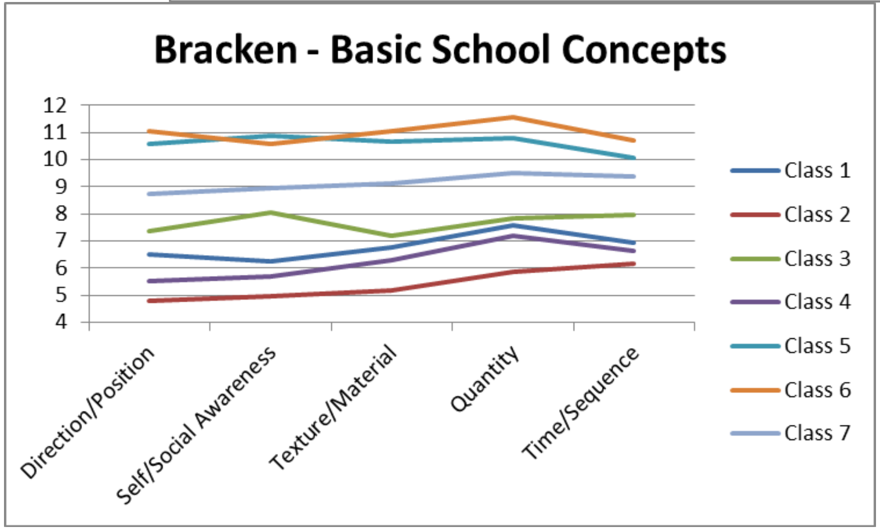
Model Estimates

FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES BASED ON THE ESTIMATED MODEL

LatentClasses

1	25.01285	0.11369
2	16.84910	0.07659
3	23.84887	0.10840
4	30.96302	0.14074
5	33.33118	0.15151
6	38.91238	0.17687
7	51.08259	0.23219





LCA Example

- 220 Preschool children and families
- 42 dichotomous demographic variables (yes/no)
  - Does your child speak English?
  - Does the child have an identified disability?
  - Speech-Language Impairment
  - Is there a father figure living in the home?
  - Unemployed
  - School lunch/ breakfast program
  - Is your child on any medications?
  - Parent's clinical depression

Syntax

cprobs

1.	0.112	0.888	2.000
1.	1.000	0.000	1.000
0.	1.000	0.000	1.000
1.	0.973	0.027	1.000
1.	1.000	0.000	1.000
1.	0.724	0.276	1.000
1.	1.000	0.000	1.000
1.	1.000	0.000	1.000
1.	1.000	0.000	1.000
1.	1.000	0.000	1.000
1.	1.000	0.000	1.000
1.	1.000	0.000	1.000
1.	1.000	0.000	1.000
1.	0.001	0.999	2.000
1.	0.999	0.001	1.000
1.	1.000	0.000	1.000
1.	0.018	0.982	2.000
1.	0.027	0.973	2.000
1.	0.000	1.000	2.000
0.	0.000	1.000	2.000
1.	0.000	1.000	2.000
0.	0.143	0.857	2.000
0.	0.000	1.000	2.000
0.	0.000	1.000	2.000
1.	0.003	0.997	2.000
0.	0.993	0.007	1.000
0.	0.551	0.449	1.000
1.	0.958	0.042	1.000
0.	0.066	0.934	2.000
1.	0.000	1.000	2.000
0.	0.000	1.000	2.000

Mplus - Mptext1

File Edit View Mplus Graph Window Help

Mptext1

```
TITLE:          Mixture Modeling - LCA Example

DATA:           File is time1demo.dat;
                FORMAT is 42f6.0;

VARIABLE:       NAMES are D1-D42;
                USEVARIABLES are D1-D42;
                MISSING = all(-999);
                CATEGORICAL = D1-D42;
                CLASSES=c(2);

ANALYSIS:       TYPE=MIXTURE;

OUTPUT:         TECH1 TECH11 TECH14;

SAVEDATA:       SAVE=CPROBABILITIES;
                FILE=cprobs.dat;
```

Results

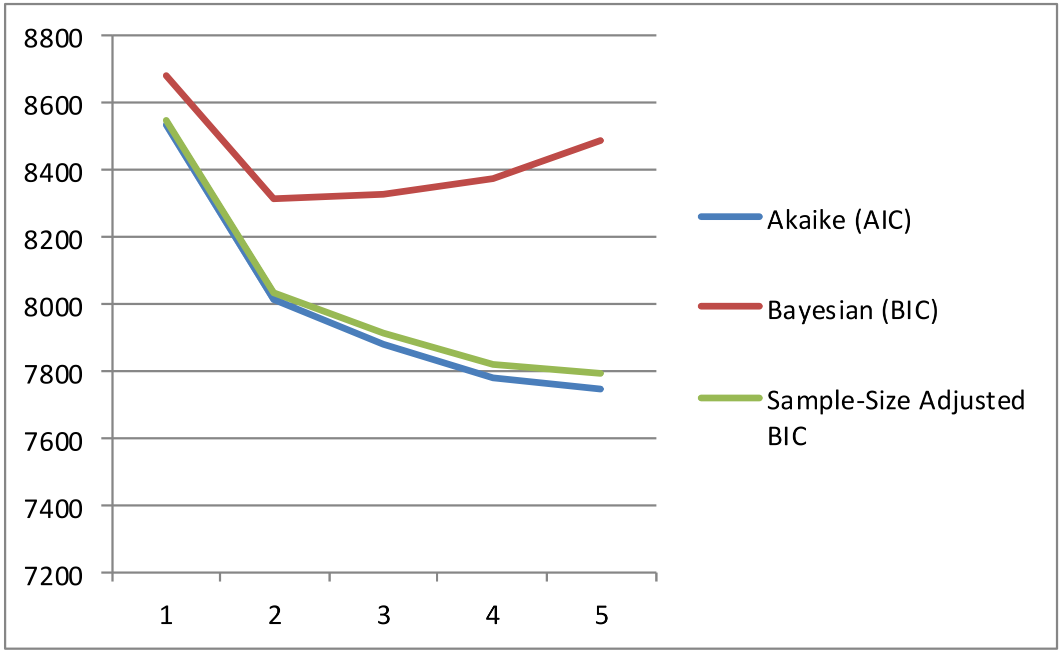
$$\frac{\exp(-.942)}{1+\exp(-.942} = .280$$

time1demo

```
MODEL RESULTS

                                                  Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

Latent Class 1

 Thresholds
    D1$1              -0.942      0.416     -2.267      0.023
    D2$1               2.429      0.644      3.772      0.000
    D3$1               1.980      0.704      2.811      0.005
    D4$1               2.197      0.530      4.145      0.000
    D5$1               2.990      0.948      3.153      0.002
    D6$1               1.747      0.872      2.003      0.045

Latent Class 2

 Thresholds
    D1$1             -15.000      0.000    999.000    999.000
    D2$1              -0.661      0.262     -2.523      0.012
    D3$1              -1.965      0.499     -3.939      0.000
    D4$1               1.917      0.275      6.970      0.000
    D5$1               1.975      0.292      6.770      0.000
    D6$1              -0.342      0.431     -0.792      0.428

Categorical Latent Variables

 Means
    C#1               -0.829      0.291     -2.853      0.004


RESULTS IN PROBABILITY SCALE

Latent Class 1

 D1
    Category 1         0.280      0.084      3.345      0.001
    Category 2         0.720      0.084      8.580      0.000
 D2
    Category 1         0.919      0.048     19.179      0.000
    Category 2         0.081      0.048      1.690      0.091
 D3
    Category 1         0.879      0.075     11.700      0.000
    Category 2         0.121      0.075      1.616      0.106
 D4
    Category 1         0.900      0.048     18.862      0.000
    Category 2         0.100      0.048      2.097      0.036
 D5
    Category 1         0.952      0.043     22.029      0.000
    Category 2         0.048      0.043      1.107      0.268
 D6
    Category 1         0.852      0.110      7.725      0.000
    Category 2         0.148      0.110      1.347      0.178
 D7
```

Results

	1	**2**	3	4	5
Akaike (AIC)	8537.02	8016.994	7882.698	7783.848	7744.665
Bayesian (BIC)	8682.946	8312.24	8327.263	8377.733	8487.87
Sample-Size Adjusted BIC	8546.679	8036.536	7912.124	7823.157	7793.858
VLMR-LRT		0.0001	0.0652	0.5232	0.2954
LMR ADJUSTED LRT		0.0001	0.0668	0.525	0.2974
BOOTSTRAPPED LRT		0.0000	0.0000	0.0000	0.0000



Results

UNIVARIATE PROPORTIONS ANDCOUNTS FOR CATEGORICAL VARIABLES

D1: Does your child speak English?

Category 1: No	0.085	18
Category 2: Yes	0.915	195

D2: Is your child enrolled in child care or cared for outside of the home on a regular

Category 1: No	0.516	110
Category 2: Yes	0.484	103

D3: Has your child ever been in a child care arrangement?

Category 1: No	0.339	61
Category 2: Yes	0.661	119

D4: Does the child have an identified disability?

Category 1: No	0.88	184
Category 2: Yes	0.12	25

D5: Has the child been referred for evaluation for development delays through

Category 1: No	0.897	156
Category 2: Yes	0.103	18

D6: Does the child have an indvidualize Educational Plan?

Category 1: No	0.587	27
Category 2: Yes	0.413	19

FINAL CLASS COUNTS AND PROPORTIONS FOR
THE LATENT CLASSES BASED ON THE ESTIMATED MODEL

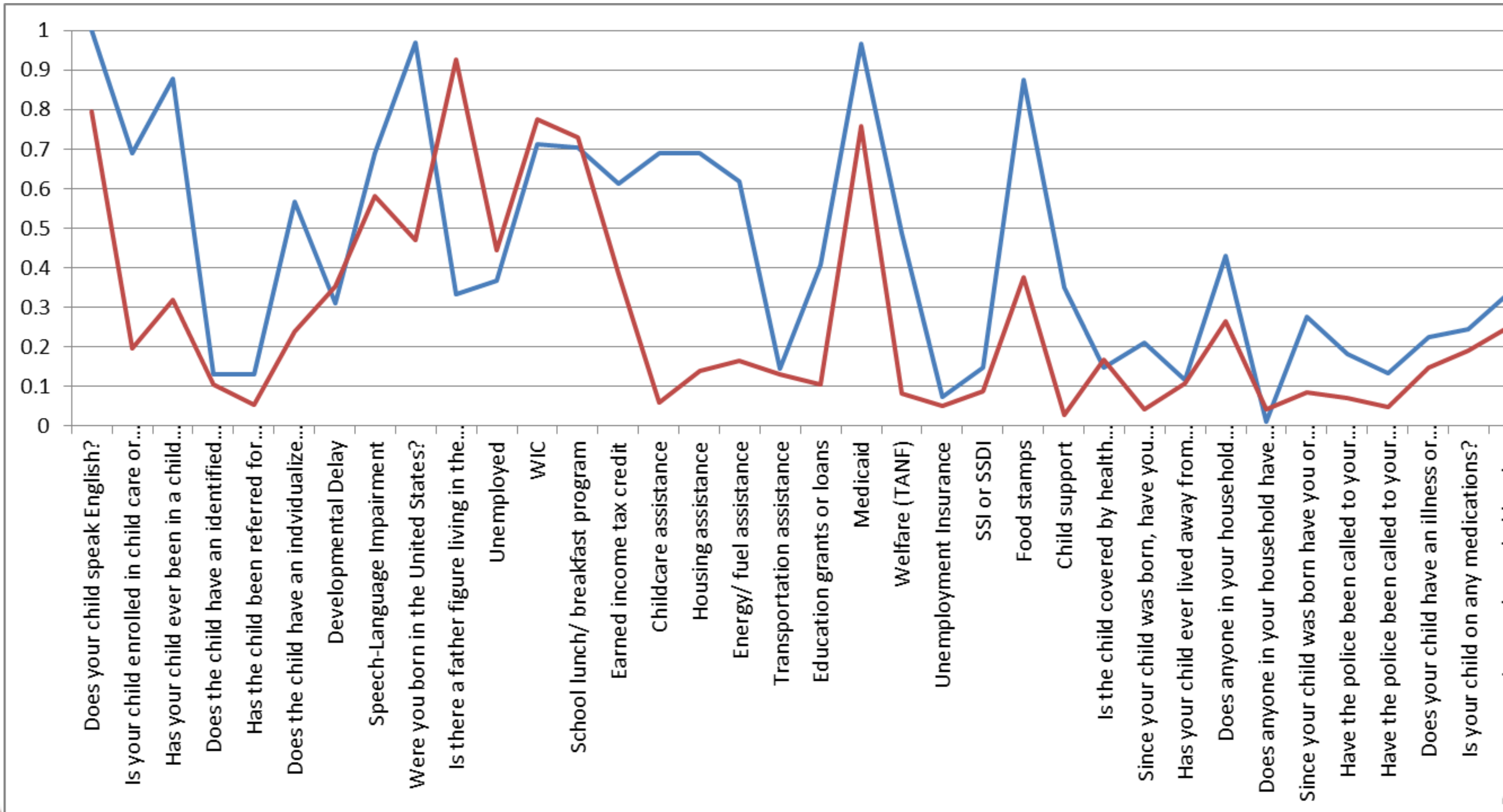
Latent Classes		
1	128.54808	0.58431
2	91.45192	0.41569

CLASSIFICATION OF INDIVIDUALS BASED ON THEIR MOST LIKELY LATENT CLASS MEMBERSHIP

Class Counts and Proportions

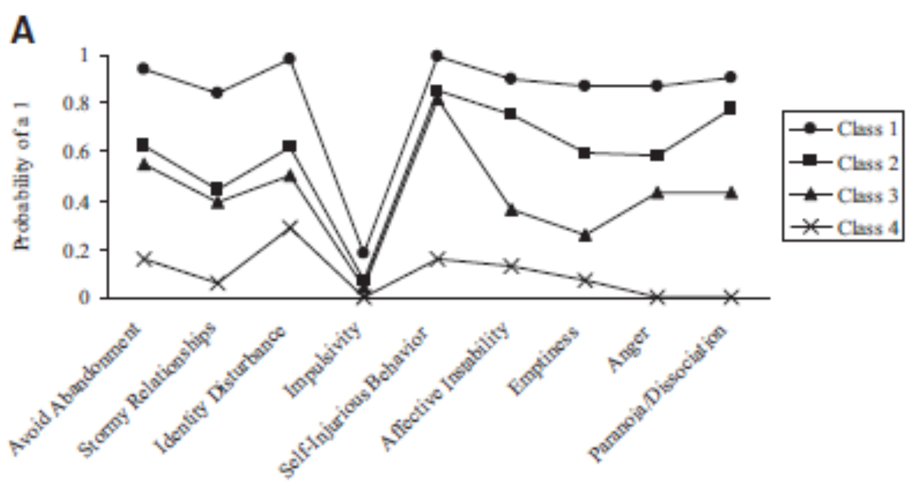
Latent Classes		
1	131	0.59545
2	89	0.40455

Results in Probability Scale



Profile Interpretability

- Sometimes profiles will be fairly similar
- Profiles with few participants may be difficult to interpret or validate
- Describe the subgroups identified using line graphs or proportions
- Which items or scales are most useful for differentiating classes?
  - Conditional probabilities of responses
  - Cabell et al. 2011
  - Bornovalova et al. 2010



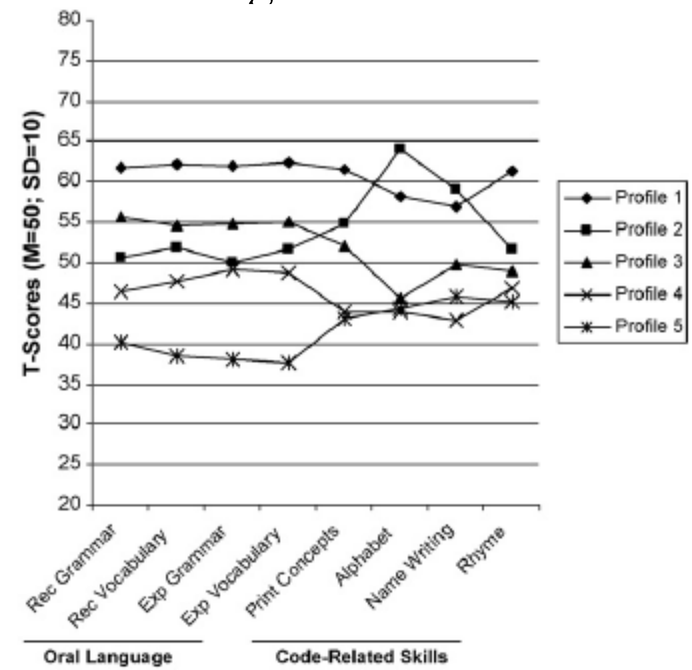
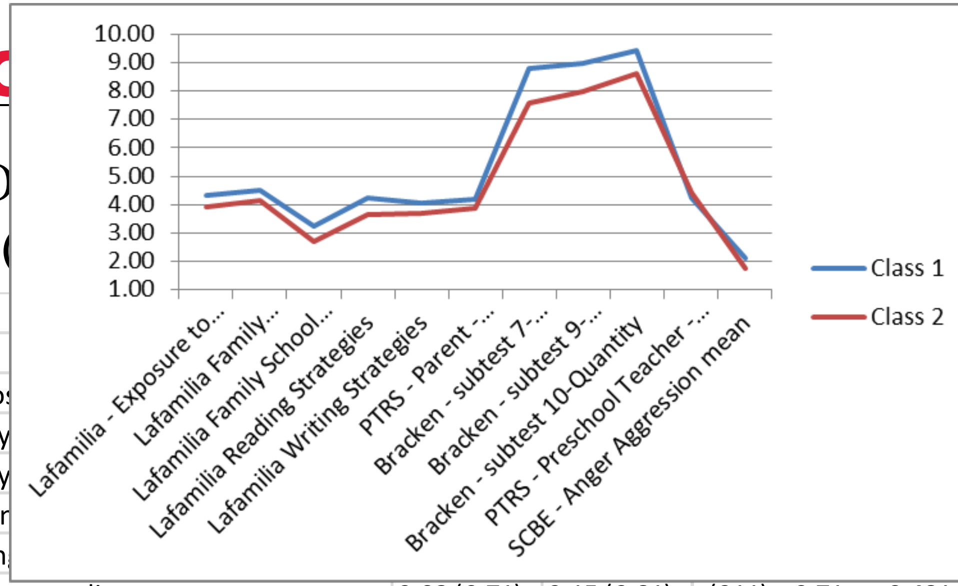
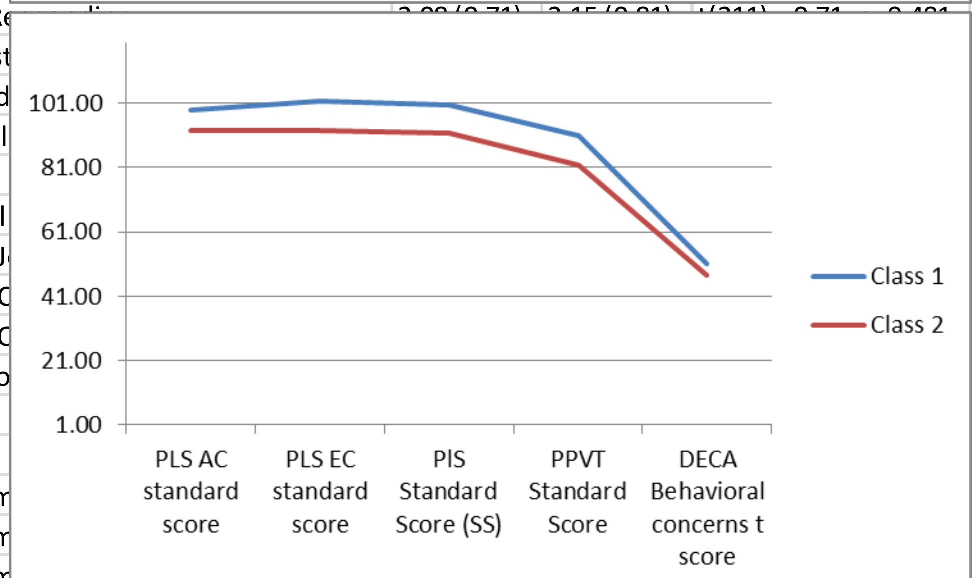


Fig. 1. Profiles of emergent literacy skills.
Profile 1: Highest emergent literacy (14%).
Profile 2: Average oral language, strength in alphabet knowledge (16.3%).
Profile 3: High average oral language, weakness in alphabet knowledge (24.2%).
Profile 4: Low average oral language, broad code-related weaknesses (22.5%).
Profile 5: Lowest oral language, broad code-related weaknesses (22.9%).

Post Hoc

- Conduct ANO … *icantly* on any variables (

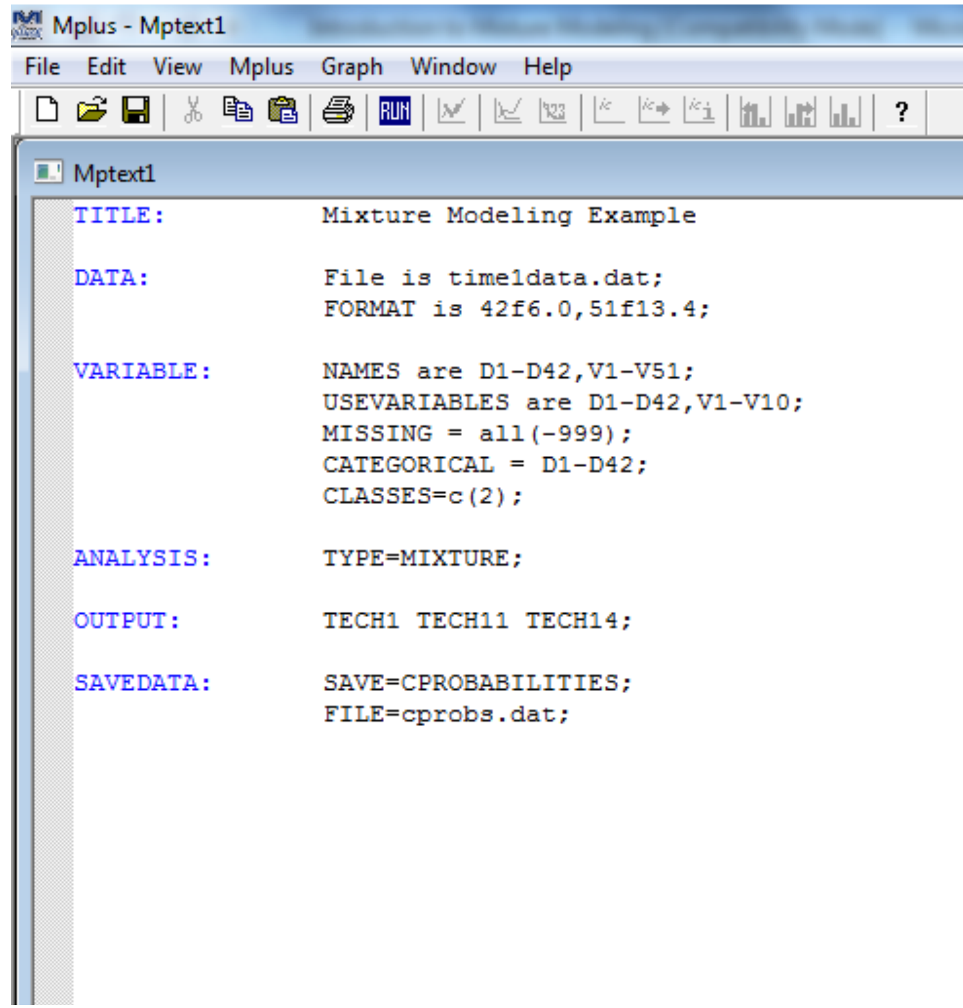




Lafamilia - Expos			
Lafamilia Family			
Lafamilia Family			
Lafamilia Readir			
Lafamilia Writin			
PSI Defensive Re			
PSI Parental Dist			
PSI Parent-Child			
PSI Difficult Chil			
PSI Total Stress			
CESD - Maternal			
PTRS - Parent - J			
PTRS - Parent - C			
PTRS - Parent - C			
PSOC Satisfactio			
PSOC Efficacy			
PSOC Total			
Family Involvem			
Family Involvem			
Family Involvem			
SCBE - Parent - social competence	39.86 (9.28)	41.13 (9.5)	t(199)=-0.95, p=0.343
SCBE - Parent - anxiety withdrawal	15.84 (5.1)	16.86 (6.08)	t(199)=-1.28, p=0.201
SCBE - Parent - anger agression	20.76 (7.92)	20.81 (7.82)	t(199)=-0.04, p=0.969

Finite mixture model – LCA and LPA

- Same syntax as before
- Added 10 continuous variables to USEVARIABLES list
- CATEGORICAL list does not change
- Will get both means and probabilities
- Everything is interpreted the same



Longitudinal Analyses

- Assuming everyone follows the same trajectory may be wrong
- Two options
  - Perform mixture model at baseline and see if trajectories differ across groups
  - Perform a growth mixture model to see if there are classes of trajectories

Mixture Model with longitudinal data

Sturge-Apple et al. (2010). Typologies of family functioning and children's adjustment during the early school years. *Child Development*, 81, 1320–1335.

- Cohesive families have kids with better adjustment
- First, a latent class analysis/latent profile analysis was used to identify groups/types at wave 1.

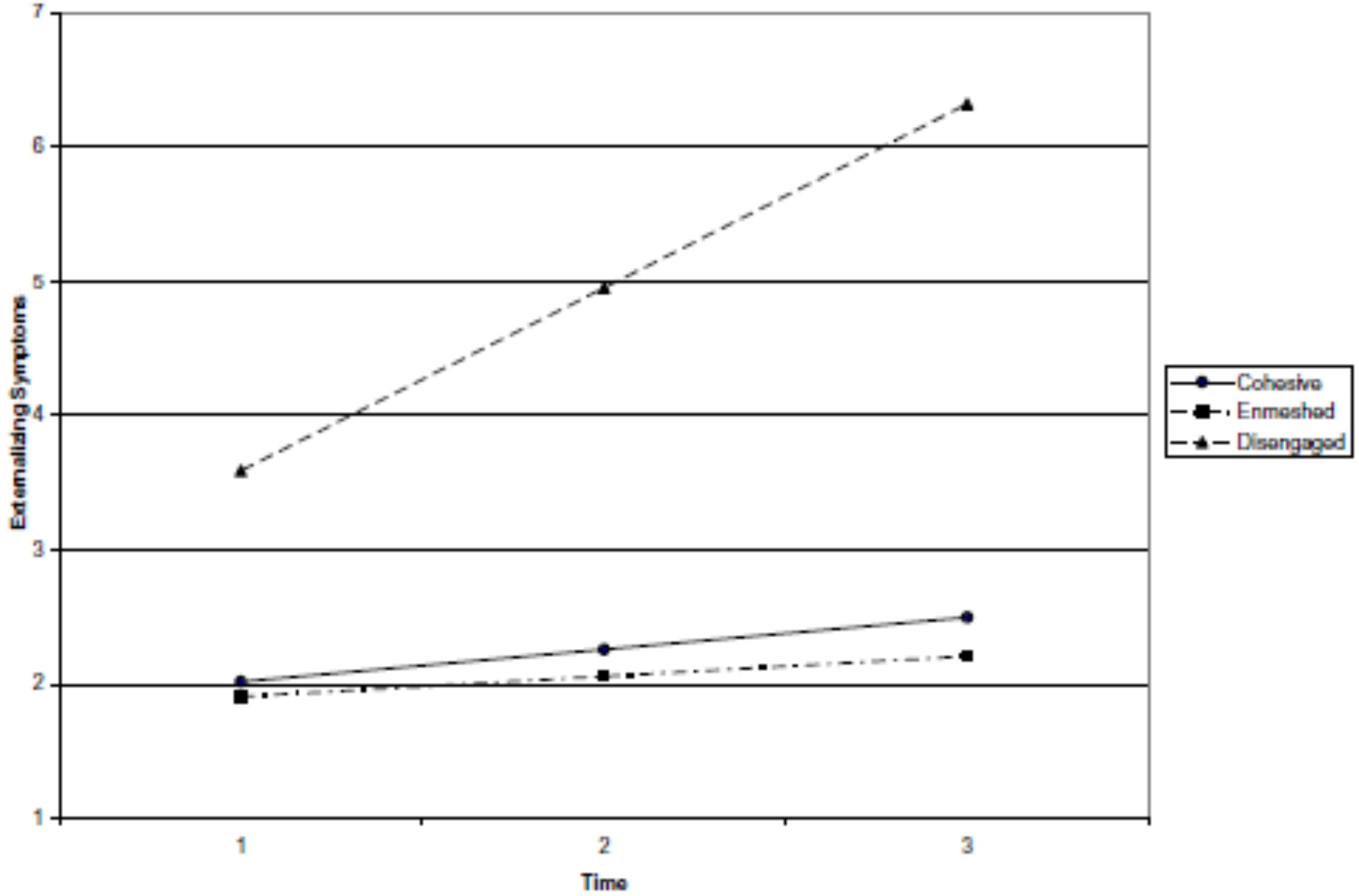
Table 2
Means, Standard Deviations, and ANOVA Comparisons of the Three Family Typologies on Seven Defining Variables

	Cohesive (C; n = 137)		Enmeshed (E; n = 51)		Disengaged (D; n = 43)			
	M	*SD*	*M*	*SD*	*M*	*SD*	$F(2, 230)$	Post hoc
Wave 1								
Interparental hostility	−.46	.53	1.47	.79	−.27	.64	187.50***	E > C, D
Interparental withdrawal	−.36	.67	−.18	.74	1.38	.97	90.50***	D > E, C
Parental emotional availability	.31	.82	.01	1.02	−.99	.86	36.17***	E, C > D
Parental intrusiveness	−.14	.95	.09	1.07	.34	.99	4.20***	D > C
Child relatedness	.18	.96	−.12	1.01	−.44	.98	7.23***	E, D > C
Triadic competition	−.08	.90	.40	1.23	−.28	.88	6.20***	E > C, D
Triadic cooperation	.18	.91	−.16	.98	−.37	1.18	5.97***	C > D, E
Triadic cohesiveness	.27	.95	−.20	.92	−.61	.91	15.59***	C > E, D

Note. Post hoc comparisons used Tukey's HSD to control for alpha level, ">" refers to significantly larger whereas "," refers to not significantly different at alpha = .05 level. ANOVA = analysis of variance.
***$p \leq .001$.

Mixture Model with longitudinal data

- The second analysis links types with trajectories (Latent Growth Curve; LGC)

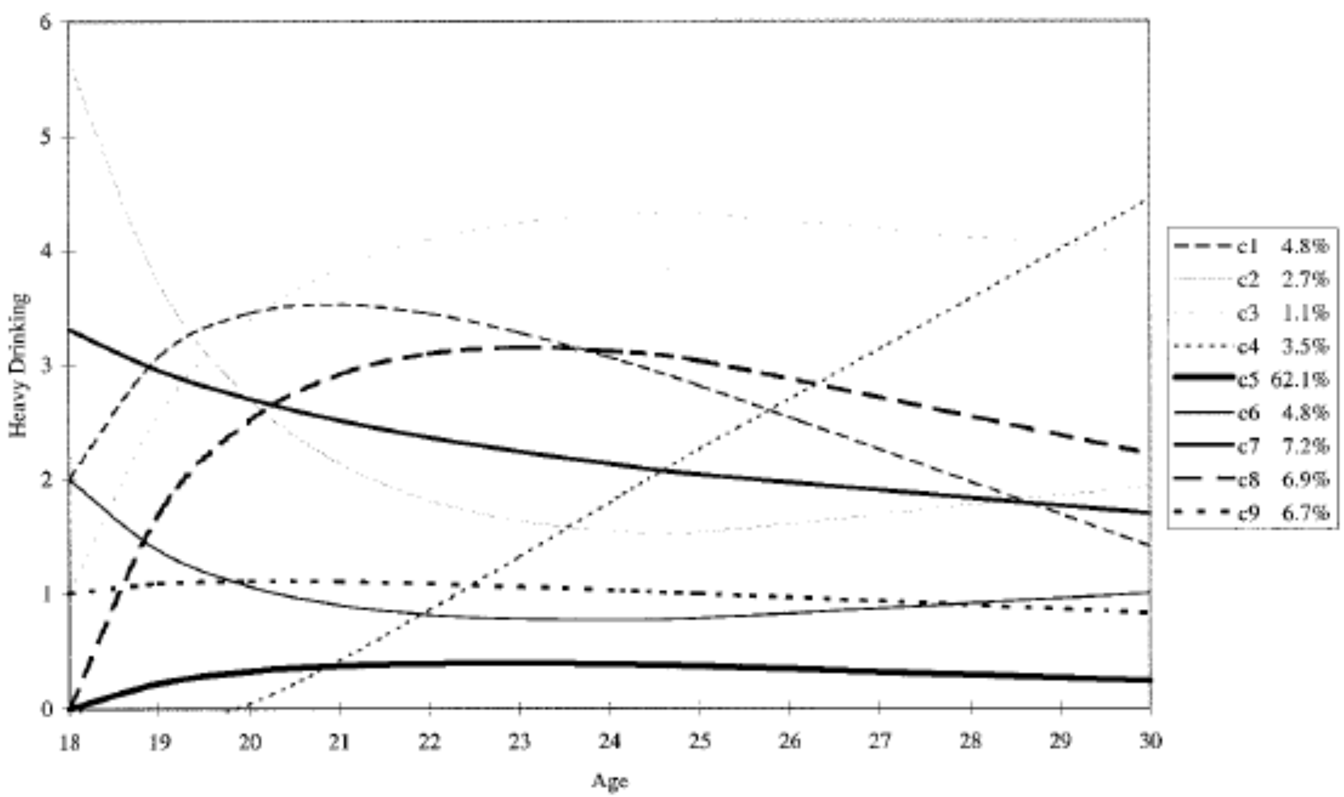


Growth Mixture Modeling

Muthen & Muthen (2000) Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alchoholism: Clinical and Experimental Research*, 24, 882-891.

- Looking for heterogeneity in developmental trajectories

NLSY Mean Curves for 9 Classes with Zero Factor Variance
(BIC=16597.712)



Limitations

- May need to use multiple starts
- Can take a long time to estimate
- Solutions may change depending on the set of predictors
- Exploratory in nature
- Not guaranteed to produce interpretable profiles

Conclusions

- Can help identify at-risk individuals
  - May want to target them for intervention
- Flexible (can use categorical or continuous outcome and predictor variables; model cross-sectional or longitudinal data)
- Useful for condensing a large amount of information in order to see patterns in your data
- Useful for when groups are unknown
- Avoids some of the problems of traditional clustering methods
- Profile interpretability is key

References

Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling*, 14, 671-694.

Lazarsfeld, P. F., & Henry, N. W. (1968). Latent structure analysis. New York: Houghton Mifflin.

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park: Sage.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Nylund, K. L., Asparouhov, T., & Muthen, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural Equation Modeling, 14, 535–569.

kkupzyk2@unlnotes.unl.edu

Thank You