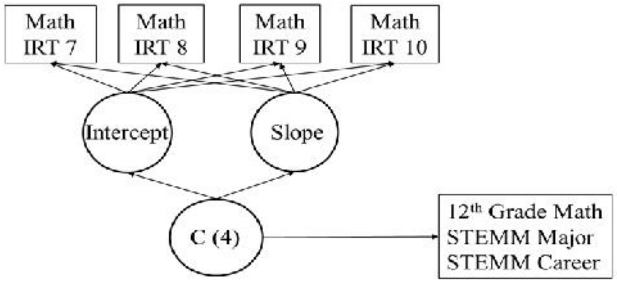
A demonstration of different methods to include distal outcomes in mixture models of classification

Ryan Grimm, Karen Nylund-Gibson, University of California, Santa Barbara

Introduction

Mixture models have increased in popularity in the social sciences over the last decade. This study focuses on how best to include auxiliary information, such as covariates or distal outcomes, into the mixture model. We examine different ways distal outcomes can be included into mixture models, describing the different methods and illustrating how they can be used in a real data analysis context.



Methods

Participants and Data

We used data from the Longitudinal Study of American Youth (LSAY; Miller, 1987-1994, 2007). We used Math IRT scores collected during the fall of grades 7-10 as indicators of latent classes. Three distal outcomes were included: (1) 12th grade Math IRT score, was continuous, (2) whether or not students anticipated pursuing a STEMM major and, (3) whether or not they later entered a STEMM career – were dichotomous.

Analyses

We fit a series of unconditional Growth Mixture Models (GMM) using M*plus* 7.11 (Muthén & Muthén, 1998-2013). We began by running a 1-class model, then iteratively increased the number of latent classes by one until adding classes no longer adequately explained the heterogeneity in the sample. We included the distal outcomes using six different methods.

1) The **distal-as-indicator approach** includes the distals in the beginning of the modeling process.
2) The **classify-analyze approach** (Clogg, 1995) assigns individuals to classes, then performs a subsequent analysis of the distals using class assignment as a grouping variable.
3) The **pseudo-class draw** approach classifies individuals and performs the subsequent analysis multiple times, then combines the results.
4) The **three-step approach** (Vermunt, 2010) accounts for the imperfect class assignment by including classification error into a subsequent model when analyzing distals. This can be conducted manually or automatically in Mplus (Muthén & Muthén, 1998-2012) and both were conducted.
5) Finally**, Lanza et al.** (2013) proposed an alternative three-step approach that avoids changes in class enumeration.

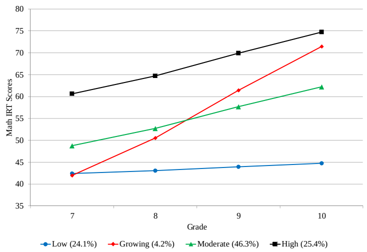


Figure 1. Growth plot of Math IRT scores for the unconditional 4-class GMM.

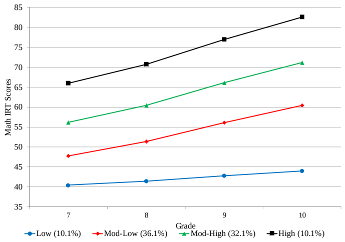


Figure 2. Growth plot of the conditional 4-class GMM using the distal-as-indicator approach.

Table 1

Class-specific Means and All Pairwise Class Comparisons

	Approach				
Pairwise Class Mean Comparisons	Distal-as-Indicator	Classify-Analyze	PC draw	Manual 3-step	Lanza et al.
Math Gr. 12					
Low vs Growing (Mod-Low)	**48.97 v 64.38**	**50.88 v 73.78**	**52.54 v 74.58**	**55.63 v 71.89**	**48.97 v 74.47**
Low vs Moderate (Mod-High)	**48.97 v 76.98**	**50.88 v 66.64**	**52.54 v 67.20**	**55.63 v 67.72**	**48.97 v 64.14**
Low vs High	**48.97 v 89.01**	**50.88 v 82.56**	**52.54 v 79.46**	**55.63 v 77.80**	**48.97 v 88.70**
Growing (Mod-Low) vs Moderate	**64.38 v 76.98**	**73.78 v 66.64**	**74.58 v 67.20**	**71.89 v 67.72**	**74.47 v 64.14**
Growing (Mod-Low) vs High	**64.38 v 89.01**	**73.78 v 82.56**	**74.58 v 79.46**	**71.89 v 77.80**	**74.47 v 88.70**
Moderate (Mod-High) vs High	**76.98 v 89.01**	**66.64 v 82.56**	**67.20 v 79.46**	**67.72 v 77.80**	**64.14 v 88.70**
STEMM Major					
Low vs Growing (Mod-Low)	**.05 v .09**	.06 v .19	**.07 v .20**	**.08 v .19**	**.05 v .20**
Low vs Moderate (Mod-High)	**.05 v .22**	**.06 v .13**	**.07 v .14**	**.08 v .16**	**.05 v .10**
Low vs High	**.05 v .46**	**.06 v .31**	**.07 v .28**	**.08 v .26**	**.05 v .50**
Growing (Mod-Low) vs Moderate	**.09 v .22**	.19 v .13	.20 v .14	.19 v .16	.20 v .10
Growing (Mod-Low) vs High	**.09 v .46**	.19 v .31	.20 v .28	.19 v .26	**.20 v .50**
Moderate (Mod-High) vs High	**.22 v .46**	**.13 v .31**	**.14 v .28**	**.16 v .26**	**.10 v .50**
STEMM Career					
Low vs Growing (Mod-Low)	.01 v .01	.01 v .09	**.01 v .09**	**.02 v .09**	.01 v .09
Low vs Moderate (Mod-High)	**.01 v .09**	**.01 v .04**	**.01 v .05**	**.02 v .06**	.01 v .01
Low vs High	**.01 v .30**	**.01 v .17**	**.01 v .15**	**.02 v .13**	**.01 v .33**
Growing (Mod-Low) vs Moderate	**.01 v .09**	.09 v .04	.09 v .05	.09 v .06	.09 v .01
Growing (Mod-Low) vs High	**.01 v .30**	.09 v .17	.09 v .15	.09 v .13	**.09 v .33**
Moderate (Mod-High) vs High	**.09 v .30**	**.04 v .17**	**.05 v .15**	**.06 v .13**	**.01 v .33**

Note. Boldface indicates significant mean differences at p < .05. Mod-Low = Moderate-Low; Mod-High = Moderate-High. [a]Class labels in parentheses refer to the labels used only in the distal-as-indicator approach. [b]Estimates could not be provided due to classification error in the third step.

References

Clogg, C.C. (1995). Latent class models. In G. Arminger, C.C. Clogg, M.E. Sobel (Eds.), *Handbook of statistical modeling for the social sciences* (pp. 311-359). New York, NY: Plenum.

Lanza, S.T., Tan X., & Bray B.C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling, 20*, 1-26.

Miller, J.D. (1987-1994, 2007). *Longitudinal Study of American Youth* [Data file and code book]. Ann Arbor, MI: Inter-university Consortium for Political and Social Science Research.

Muthén, L.K. & Muthén, B.O. (1998-2013). *Mplus user's guide* (6th ed.). Los Angeles, CA

Vermunt, J. K. (2010). Latent Class Modeling with Covariates: Two improved Three-Step Approaches. *Political Analysis*, *18*, 450–469. doi:10.1093/pan/mpq025

Results

GMM Results

We chose the 4-class model as the preferred model. The unconditional growth plot can be seen in Figure 1. We labeled the emergent classes *Low*, *Growing*, *Moderate*, and *High*.

Comparison of Distal Outcome Approaches

Figure 2 presents the growth plot produced by the distal-as-indicator approach. This plot differed from the other approaches, which all produced plots identical to the unconditional model. Due to the shift in classes, the *Growing* class was renamed *Moderate-Low* and the *Moderate* class was renamed *Moderate-High*.

Class-specific means of the distal outcomes were compared across approaches and are presented in Table 1. The means could not be computed using the automatic 3-step approach due to classification error being greater than 20% at the third step. The distal-as-indicator approach estimated a lower mean for the *Mod-Low* class than the *Mod-High* class. This differed from all other approaches. The means estimated by the classify-analyze, PC draw, and manual 3-step approaches tended to be fairly similar compared to the means estimated by the distal-as-indicator and Lanza approaches. Generally, the greatest differences occurred in mean estimates for the *High* class.

Table 1 also presents statistically significant mean differences of all pairwise class comparisons. There were two discrepancies between the classify-analyze and PC draw and manual 3-step approaches with the categorical variables. Five discrepancies occurred between the distal-as-indicator approach and classify-analyze, PC draw, and manual 3-step approaches. All involved the *Moderate-Low* class and the categorical distals. Four discrepancies occurred between the Lanza approach and the PC draw and manual 3-step approaches.

Discussion

In this study, we illustrate the influence of auxiliary variables on class enumeration and differences among the methods. We demonstrate substantive interpretation of both the emergent classes and distal outcomes may be considerably impacted by the method chosen to include distal outcomes. While drawbacks associated with some of these methods have been documented (e.g. Clogg, 1995), others have yet to be studied in-depth.

A real-data example was provided and thus, the obtained results may not generalize to other datasets. Further work should be conducted to examine each of the methods' capacities to accurately model data. Simulation studies should be undertaken to compare known and estimated parameters with each method. This study emphasizes the importance of this work to applied social science research.