Social and Personality Psychology Compass 2/1 (2008): 302–317, 10.1111/j.1751-9004.2007.00054.x

An Introduction to Latent Class Growth Analysis and Growth Mixture Modeling

Tony Jung and K. A. S. Wickrama*
Iowa State University

Abstract

In recent years, there has been a growing interest among researchers in the use of latent class and growth mixture modeling techniques for applications in the social and psychological sciences, in part due to advances in and availability of computer software designed for this purpose (e.g., Mplus and SAS Proc Traj). Latent growth modeling approaches, such as latent class growth analysis (LCGA) and growth mixture modeling (GMM), have been increasingly recognized for their usefulness for identifying homogeneous subpopulations within the larger heterogeneous population and for the identification of meaningful groups or classes of individuals. The purpose of this paper is to provide an overview of LCGA and GMM, compare the different techniques of latent growth modeling, discuss current debates and issues, and provide readers with a practical guide for conducting LCGA and GMM using the Mplus software.

Researchers in the fields of social and psychological sciences are often interested in modeling the longitudinal developmental trajectories of individuals, whether for the study of personality development or for better understanding how social behaviors unfold over time (whether it be days, months, or years). This usually requires an extensive dataset consisting of longitudinal, repeated measures of variables, sometimes including multiple cohorts, and analyzing this data using various longitudinal latent variable modeling techniques such as latent growth curve models (cf. MacCallum & Austin, 2000). The objective of these approaches is to capture information about interindividual differences in intraindividual change over time (Nesselroade, 1991).

However, conventional growth modeling approaches assume that individuals come from a single population and that a single growth trajectory can adequately approximate an entire population. Also, it is assumed that covariates that affect the growth factors influence each individual in the same way. Yet, theoretical frameworks and existing studies often categorize individuals into distinct subpopulations (e.g., socioeconomic classes, age groups, at-risk populations). For example, in the field of alcohol research, theoretical literature suggests different classes

of alcohol use initiation patterns, e.g., 'early' versus 'late' onsetters (Hill, White, Chung, Hawkins, & Catalano, 2000). Using growth mixture modeling (GMM) with five different indices of alcohol use (alcohol use disorder, alcohol dependence, alcohol consequences, past year alcohol quantity and frequency, and heavy drinking), Jackson and Sher (2005) identified four distinct classes for each measure. The results of these studies confirm theoretical contentions that heterogeneity of growth trajectories exist within the larger population. In addition, these findings suggest that describing an entire population using a single growth trajectory estimate is oversimplifying the complex growth patterns that describe continuity and change among members of different groups. Instead, a latent class or growth mixture modeling approach seems to be the most appropriate method for fully capturing information about interindividual differences in intraindividual change taking into account unobserved heterogeneity (different groups) within a larger population.

Person-Centered and Variable-Centered Analyses

A useful framework for beginning to understand latent class analysis and growth mixture modeling is the distinction between person-centered and variable-centered approaches (cf. Muthén & Muthén, 2000). Variable-centered approaches such as regression, factor analysis, and structural equation modeling focus on describing the relationships among variables. The goal is to identify significant predictors of outcomes, and describe how dependent and independent variables are related. Person-centered approaches, on the other hand, include methods such as cluster analysis, latent class analysis, and finite mixture modeling. The focus is on the relationships among individuals, and the goal is to classify individuals into distinct groups or categories based on individual response patterns so that individuals within a group are more similar than individuals between groups.

Growth Mixture Modeling

Given a typical sample of individual growth trajectories (Figure 1, left), conventional growth modeling approaches give a single average growth estimate (bold line), a single estimation of variance of the growth parameters, and assumes a uniform influence of covariates on the variance and growth parameters. However, there may exist a subset of individuals (Figure 1, right) whose growth trajectories are significantly different from the overall estimate. In this example, the figure on the left-hand side represents a sample of individual adolescent mental health growth trajectories (SCL-90-R depression, anxiety, and somatic symptoms measures), with an average positive intercept and slope. The figure on the right-hand side is a subset of the entire sample, representing adolescent mental health growth

 Social and Personality Psychology Compass 2/1 (2008): 302–317, 10.1111/j.1751-9004.2007.00054.x

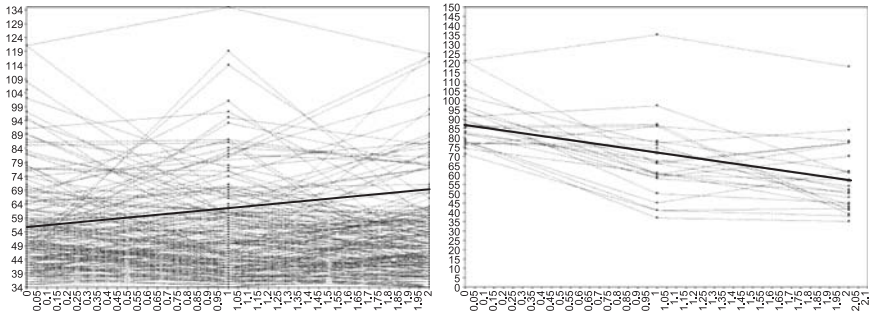

Figure 1 Individual trajectories for adolescent mental health (left) and the recovery class (right).

trajectories that are decreasing in poor mental health symptomology, that is, improving mental health. Individuals in this 'recovery' group have a higher intercept and a negative slope, characteristics of the growth parameters that are clearly different from that of the whole sample.

The conventional growth model can be described as a multilevel, random-effects model (Raudenbush & Bryk, 2002). According to this framework, intercept and slope vary across individuals and this heterogeneity is captured by random effects (i.e., continuous latent variables). However, as mentioned previously, this approach assumes that the growth trajectories of all individuals can be adequately described using a single estimate of growth parameters (both the mathematical form and the magnitude). Underlying this framework is the assumption that all individuals are drawn from a single population with common parameters. GMM, on the other hand, relaxes this assumption and allows for differences in growth parameters across unobserved subpopulations. This is accomplished using latent trajectory classes (i.e., categorical latent variables), which allow for different groups of individual growth trajectories to vary around different means (with the same or different forms). The results are separate growth models for each latent class, each with its unique estimates of variances and covariate influences. This modeling flexibility is the basis of the GMM framework (cf. Muthén & Asparaouhov, 2006).

Latent class growth analysis (LCGA) is a special type of GMM, whereby the variance and covariance estimates for the growth factors within each class are assumed to be fixed to zero. By this assumption, all individual growth trajectories within a class are homogeneous. This framework of growth modeling has been extensively developed by Nagin and colleagues (cf. Nagin & Land, 1993) and is embodied in the SAS procedure Proc Traj (Jones, Nagin, & Roeder, 2001). The benefit of this approach is the identification of distinct classes prior to conducting GMM. It serves as a starting point for conducting GMM. In terms of computation, it is easy to specify in Mplus and the zero constraints on the variance estimates allow for faster model convergence (cf. Kreuter & Muthén, 2007).

Current Issues and Debate

Much of the current issues and debate surround three main areas: (i) the determination of latent trajectory classes; (ii) which model fit index to use; and (iii) the problem of convergence. The first issue is concerned with the question whether latent classes really exist and if so, how many? For example, Bauer and Curran (2003a, b) cautioned that the existence of multiple classes may simply be due to skewed or nonnormally distributed data.

Assuming there are multiple classes, how does one determine how many there are? Currently, methods for determining the number of components in a growth mixture model consists of finding the model with the smallest Bayesian information criteria (BIC) value and a significant Lo, Mendell, and Rubin (2001) likelihood ratio test (LMR-LRT) statistic. More recently, however, further simulations have demonstrated that while the BIC performed the best among the information criteria-based indices, the bootstrap likelihood ratio test (BLRT) proved to a better indicator of classes across all of the models considered. All of these fit indices are available in Mplus (see Nylund, Asparouhov, & Muthén, 2007, for a discussion on fit indices). Analogous to determining the number of factors using exploratory factor analysis, the number of classes should ultimately be determined by a combination of factors in addition to fit indices, including one's research question, parsimony, theoretical justification, and interpretability (cf. Bauer & Curran, 2003b; Muthén, 2003; Rindskopf, 2003).

A third issue that is often raised is the problem of nonconvergence and local solutions (cf. Hipp & Bauer, 2006). Trying to mathematically model a sample distribution that consists of a mixture of many different kinds of subdistributions (i.e., a finite mixture model) is extremely difficult. Such attempts are notorious for convergence issues due to likelihood estimation problems (e.g., local minima and maxima and singularities). Like other methods such as cluster analysis, latent class analysis, and finite fixture modeling, growth mixture models are also susceptible to local solutions. The problem of local solutions is where during curve estimation a largest value (maximum) or smallest value (minimum) that a function takes is identified for only a given area on that curve, but that is not necessarily the largest or smallest value for the entire curve (i.e., the global minimum or maximum). The problem with local solutions in latent class analysis has long been known (Goodman, 1974). In mixture modeling, parameters are estimated by the method of maximum likelihood and are iterative in nature (e.g., EM algorithm). Ideally, the iteration will result in successful convergence on the global maximum solution, that is, the parameter estimates associated with the largest loglikelihood. However, the algorithm cannot distinguish between a global maximum and a local maximum. As long as it reaches some maximum, the algorithm will

 Social and Personality Psychology Compass 2/1 (2008): 302–317, 10.1111/j.1751-9004.2007.00054.x

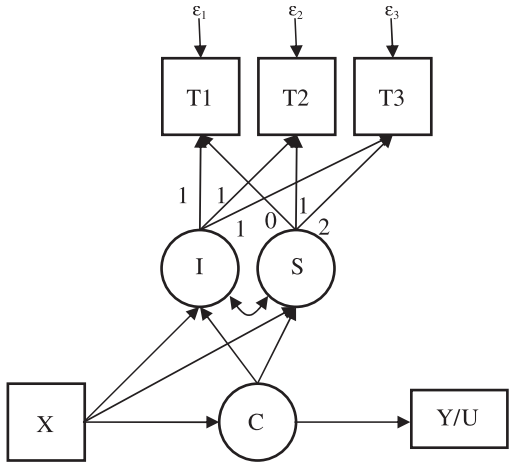

Figure 2 Representation of a growth mixture model with covariates.

terminate. Fortunately, the Mplus software incorporates the use of random starting values, with sufficient user flexibility, to avoid local solutions in GMM.

GMM and LCGA in Mplus

This section outlines the basic steps for specifying a simple LCGA and GMM model in Mplus Version 4.1, briefly explains the different user-modifiable options, and highlights specific parts of the output that the beginning user needs to be aware of. Readers are recommended to refer to Chapter 8 in the Mplus User's Manual available at www.statmodel.com for a complete treatment of longitudinal mixture modeling. Examples of input and output for more complex analyses, with more detailed instructions are available at www.statmodel.com/examples/penn.shtml. The general latent variable growth mixture model can be represented as follows:

The growth mixture model in Figure 2 consists of the following components: (i) a univariate latent growth curve of observed variable T with an intercept (I) and slope (S), (ii) a categorical variable for class (C), and (iii) covariates or predictor variables (X). A distal continuous outcome variable (Y) or a dichotomous outcome variable (U) can be also added to the model by regressing Y onto C, but is not shown here. The simple univariate latent growth curve with latent growth factors, intercept (I) and slope (S), are formed by the observed variables T1, T2, and T3 that represent repeated measures across three time points. A fourth repeated measure (T4) could also be added to the model to estimate a quadratic growth factor (Q), but for sake of simplicity only the slope factor is

considered here. As an aside, estimating additional growth factors, for example, a quadratic term, will add computational burden, so it is not unusual to see the variance of the quadratic term and other growth factors in select classes fixed to zero to aid in convergence during GMM.

The categorical latent class variable (C) is related to the covariates (X) by way of multinomial logistical regression. The Mplus multinomial regression assigns each individual fractionally to all classes using the posterior probabilities, obtained through the EM iterations. The first set of fractional assignments is based on the starting values, and they are then iteratively improved on until convergence. In the case of a dichotomous covariate (e.g., 0 = females, 1 = males), the coefficient is the increase in the log-odds of being in the disengaged versus the normative class for a one-unit increase in X, for example, when comparing males to females in this example. Hence, a coefficient of 1.0 implies that the odds of being in the disengaged class versus the normative class is exp(1) = 2.72 times higher for males than females. The same odds ratio interpretation applies to each class when including a dichotomous distal outcome variable (U). See Muthén (2004, 349) for further detail on multinomial logistic regression in Mplus.

Step 1: Specify a single-class latent growth curve model

The initial step prior to specifying latent classes is to specify a single-class growth model. For example, a univariate growth curve model without covariates may be an initial starting point for beginning users. In Mplus, a univariate latent growth curve can be specified as follows:

```
Title:      UNIVARIATE GROWTH CURVE
Data:       file is 'C:\My Folder\filename.dat';
Variable:   names are id sex t1 t2 t3 x;
            usevar = t1–t3;
            missing = all (999);
Analysis:   type = missing H1;
Model:      i s | t1@0 t2@1 t3@2;
Output:     sampstat standardized tech1;
```

In this dataset, there are six variables and missing values are coded as '999'. For the present analysis, only the three observed variables in the dataset are needed (USEVAR = T1 – T3). The TYPE = MISSING syntax invokes the full-information maximum likelihood algorithm for handling missing data. The MODEL syntax specifies the appropriate factor loadings corresponding to the equidistant time intervals, 0, 1, and 2, for the intercept (I) and slope (S). The zero time score for the slope growth factor at time 1 defines the intercept growth factor as an initial status factor. The coefficients of

the intercept growth factor are fixed at one as the default. The user has the option of freely estimating a factor loading by simply deleting it (e.g., delete '@1' to freely estimate the second time point, T2). Note that the | syntax is a new option introduced with Mplus Version 3, which replaces the BY syntax option. The BY option is a general approach to defining latent variables in Mplus. The new | syntax is used to name and define random effect variables and is used for specifying growth models. It calls for a different start value algorithm whereby parameters are perturbed differently as a function of their standard deviations. This syntax specifies the following latent univariate growth curve model:

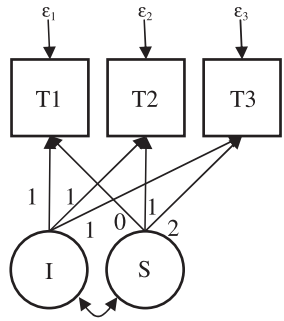

Figure 3 A latent variable representation of a univariate growth curve model.

Step 2: Specify an unconditional latent class model without covariates (X) or distal outcomes (Y or U)

However, if the covariates have significant direct effects on the growth factors (I and S) and on class (C), then it is important to keep in mind that the unconditional model will lead to distorted results since the observed variables (T1–T3) will be incorrectly related to class (C). This is analogous to a misspecified regression model, where the slope estimate will be distorted when important predictors are left out of the equation. Therefore, it is important to follow-up the unconditional model with the appropriate conditional model and to compare the results (Muthén, 2004).

At this step, specifying a LCGA model with no within-class variance is also recommended as an initial exploratory option. As discussed previously, LCGA is a useful initial modeling step prior to specifying a GMM model. The main difference is that LCGA assumes no within class variance on the growth factors, whereas GMM freely estimates the within class variances. The benefits of fixing the within-class variances to zero are the clearer identification of classes and also the less computational burden – factors that are important at this stage in model building. It is also recommended

that LCGA be used with the conditioned model before proceeding to GMM. In Mplus, an unconditional LCGA model can be specified as follows:

```
Title:      LATENT CLASS GROWTH ANALYSIS
Data:       file is 'C:\My Folder\filename.dat';
Variable:   names are id sex t1 t2 t3 x;
            usevar = t1–t3;
            missing = all (999);
            CLASSES = c(3);
Analysis:   type = MIXTURE missing;
            STARTS = 10 2;
            STITERATIONS = 10;
Model:      % OVERALL%
            i s | t1@0 t2@1 t3@2;
            i-s@0;
Output:     sampstat standardized tech1
              TECH8 TECH11 TECH14;
PLOT:       SERIES = t1–t3 (s);
            TYPE = PLOT3;
```

The bold portions of the syntax are the parts that add the LCGA model (C) to the existing univariate growth curve model. The TYPE = MIXTURE syntax invokes the mixture model algorithm. Here, a three-class model is being initially examined with the syntax CLASSES = c(3). After successful convergence, note the model fit using the BIC value and then proceed to check models with more classes by replacing the '3' with the appropriate number in the syntax. The best fitting model will have the smallest BIC value. However, BIC is only one of the options for determining model fit. In addition to the BIC, the LMR-LRT and the BLRT tests discussed earlier can also be examined by including TECH11 or TECH14 syntax, respectively, in the OUTPUT line. In this example, TECH8 shows the iteration process so that the user can be aware of iteration progress.

The STARTS and STITERATIONS lines are not required for conducting LCGA since Mplus automatically sets these parameters at default values. However, adjusting these values may aid in obtaining successful convergence. The STARTS syntax specifies the number of random sets of starting values (default = 10) followed by the number of final optimizations (default = 2), which optimizes the two best sets identified by the highest loglikelihood values after the initial round of optimizations given by the syntax STITERATIONS (default = 10). This feature, which can be user defined, is rare in other software programs and is one of the main approaches to addressing problems relating to nonconvergence and local maxima. For a more thorough investigation of multiple solutions, it is recommended that the user change the default values for the number of random sets and start iterations to:

 Social and Personality Psychology Compass 2/1 (2008): 302–317, 10.1111/j.1751-9004.2007.00054.x

```
        STARTS = 100 10;
        STITERATIONS = 10;
OR:
        STARTS = 500 20;
        STITERATIONS = 20;
```

Certainly, the user may change the start values to even higher numbers to ensure successful convergence. However, changing to higher values will increase computational burden and increase computation time.

The syntax %OVERALL% specifies the same model and free estimates across all classes. The intercepts and residual variances of the growth factors are estimated as the default, and the growth factor residual covariance is estimated as the default because the growth factors do not influence any variable in the model except their own indicators. The intercepts of the growth factors are not held equal across classes as the default. However, the residual variances and residual covariance of the growth factors are held equal across classes as the default. In this example, the syntax i-s@0 fixes all within-class variances to zero, consistent with the LCGA approach. Removing this line will set the variances of I and S as equal across all classes and estimate the variances of the growth parameters. If separate estimates of the within-class variances are desired for each class, it is necessary to add the following changes to the MODEL line in the syntax:

```
Model:      %OVERALL%
            i s | t1@0 t2@1 t3@2;
            %c#1%
            i s ;
            %c#2%
            i s;
            %c#3%
            i s;
```

These added lines in bold tell Mplus to estimate the unique variances of intercept and slope for each class. This would be equivalent to a GMM model where the within-class variances are allowed to be freely estimated instead of fixed to zero as in LCGA. However, because freely estimating the variances for all growth factors in each class separately adds considerable computational burden, the user must choose growth parameter variances to freely estimate carefully. In general, increasing model complexity by adding classes, adding covariates, allowing across-class variation in covariance matrices can add to computation time, convergence problems, improper solutions, and overall model instability. Hence, it is not unusual to decide to only estimate intercept and not slope, or limit changes to a particular class rather than across all classes. These decisions should be made after successfully running a LCGA, then looking at the graphics using the

estimated means and observed individual values plot given by the PLOT syntax [Alt-V or from pull-down menu: Graph → View Graphs] and determine if any class needs its own class-specific variance. After examining the variances of the growth factors for each class using the charts, it is necessary to reanalyze the model with the class-specific variance in line with the results of the initial LCGA findings.

Step 3: Determine the number of classes

It is important to reiterate the point made earlier that determining the number of classes depends on a combination of factors in addition to fit indices, including one's research question, parsimony, theoretical justification, and interpretability. Keeping this in mind, fit indices and tests of model fit should not be the final word in deciding on the number of classes. However, they are useful in the initial exploratory stages of analyses. At this point, studies are ongoing and results are not conclusive regarding the best fit indices. Using simulations, Nylund et al. (2007) has determined that of all the fit indices and tests available in Mplus, the BLRT performed the best, followed by BIC and then ABIC. However, Nylund et al. (2007, 33–34) recommends: 'due to the increased amount of computing time of the BLRT, it may be better to not request the BLRT in the initial steps of model exploration. Instead, one could use the BIC and the LMR *P*-values as guides to get close to possible solutions and then once a few plausible models have been identified, reanalyze these model requesting the BLRT'. Following their recommendation, the model with a low BIC value and a significant LMR *P*-value comparing the k and the $k-1$ class model should initially guide our analysis. In other words, comparing the current model against the model with 1 less class than the current model of choice should give a LMR *P*-value less than 0.05.

TECHNICAL 11 OUTPUT	
VUONG–LO–MENDELL–RUBIN LIKELIHOOD	
RATIO TEST FOR 2 (H0) VERSUS 3 CLASSES	
H0 Loglikelihood Value	–256.426
Two Times the Loglikelihood Difference	20.216
Difference in the Number of Parameters	3
Mean	–10.943
Standard Deviation	29.068
***P*-value**	**0.0476**
LO–MENDELL–RUBIN ADJUSTED LRT TEST	
Value	18.483
P-value	0.0560
TECHNICAL 14 OUTPUT	
Random Starts Specification for the *k*–1 Class Model	
Number of initial stage random starts	0
Number of final stage optimizations for the initial stage random starts	0
Random Starts Specification for the *k* Class Model	

Number of initial stage random starts	20
Number of final stage optimizations	5
Number of bootstrap draws requested	Varies
PARAMETRIC BOOTSTRAPPED LIKELIHOOD RATIO TEST FOR 2 (H0) VERSUS 3 CLASSES	
H0 Loglikelihood Value	–256.426
Two times the Loglikelihood Difference	20.216
Difference in the Number of Parameters	3
Approximate *P*-value	**0.0000**
Successful Bootstrap Draws	20

The results of the LMR-LRT and the BLRT can be seen in the output under the TECHNICAL 11 and TECHNICAL 14 sections, respectively. Here, the LMR-LRT and the BLRT both show a statistically significant difference between the two-class versus three-class models. This suggests that the three-class model gives significant improvement in model fit over the three-class model. The next step would be to compare the three-class model to the four-class model, and so on until the tests result in nonsignificance.

Other considerations include successful convergence, high entropy value (near 1.0), no less than 1% of total count in a class, and high posterior probabilities (near 1.0). In the output window, one can find the following information under the TESTS OF MODEL FIT section:

TESTS OF MODEL FIT

Information Criteria

Number of Free Parameters	14
Akaike (AIC)	4752.782
Bayesian (BIC)	**4799.641**
Sample-Size Adjusted BIC (n* = (n + 2)/24)	4755.281
Entropy	**0.923**

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

Latent Classes

1	171.37719	**0.81608**
2	27.44010	**0.13067**
3	11.18271	**0.05325**

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

	1	2	3
1	**0.975**	0.012	0.014
2	0.055	**0.941**	0.003
3	0.030	0.000	**0.970**

There are no set cut-off criteria for deciding whether the entropy is reasonably high, however, in the output above, note that the entropy is near 1.0. Also the proportions for the latent classes are all above .01 or 1%. Finally, the posterior probabilities (diagonal values) are all reasonably high, near 1.0. Together, these results suggest good model fit for the

three-class model. Once the number of classes has been decided, one can look at the graphs and see the observed and estimated means and trajectories for each class. In Mplus, selecting Alt-V or from pull-down menu, Graph → View Graphs → Estimated Means and Observed Individual Values to obtain the following graphs:

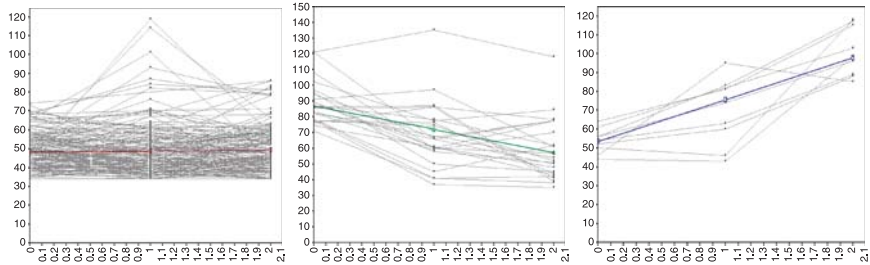

Figure 4 Estimated means and observed individual growth trajectories for each latent class.

Step 4: Address convergence issues

As discussed previously, The STARTS syntax, which can be user defined, is one of the main approaches to addressing problems relating to nonconvergence and local maxima. The user can change the default values for the number of random sets and start iterations to higher values. However, even with successful convergence, it is necessary to check whether the solutions are local solutions. To do this, check the estimates in the output using the OPTSEED syntax on the seed values from the best loglikelihood values. If the estimates are replicated, then most likely you did not run into local solutions. The best loglikelihood values are ordered from best to worst for you by Mplus. Alongside these loglikelihood values are the SEED values in the output:

RANDOM STARTS RESULTS RANKED FROM THE BEST TO THE WORST LOGLIKELIHOOD VALUES		
Initial stage loglikelihood values, seeds, and initial stage start numbers:		
–2362.540	**939021**	8
–2362.677	**373505**	88
–2363.424	**436460**	89
–2363.447	**76974**	16
–2363.507	**402224**	91
–2363.513	**467339**	66
–2363.767	**364676**	27
–2365.631	**902278**	21
–2366.060	**170954**	86
–2367.730	**830392**	35

Mplus rank orders the best loglikelihood values to the worst. The middle column of numbers consists of the seed values. We need to check the model parameter estimates using OPTSEED on the seed values from the best loglikelihood values. If the estimates are replicated, then most likelihood you did not run into local solutions. A successfully converged model will have the best loglikelihood values repeated at least twice. In this example, the first two values are the best loglikelihoods, with seed values 939021 and 373505 respectively. Next, we need to back to our original syntax and add a syntax line for OPTSEED = 939021; under the ANALYSIS line as follows:

```
Analysis:    type = MIXTURE missing;
             OPTSEED = 939021;
```

After running this model, run another model using OPTSEED = 373505, then compare the results of both outputs. We should see our estimates replicated. If the estimates are replicated then we can trust that we did not find local solutions.

Step 5: Specify a conditional latent class model with covariates (X)

Next, Steps 2–4 as outlined above need to be repeated using a conditioned model. As mentioned previously, if the covariates have significant direct effects on the growth factors (I and S) and on class (C), then the unconditional model will lead to distorted results since only the observed variables (T1–T3) will be incorrectly related to class (C). Therefore, it is important to follow-up the unconditional model with the appropriate conditional model and to compare the results. The conditioned model can be specified by adding the following bolded lines to the existing syntax:

```
Title:       LATENT CLASS GROWTH ANALYSIS
Data:        file is 'C:\My Folder\filename.dat';
Variable:    names are id sex t1 t2 t3 x;
             usevar = t1 – t3;
             missing = all (999);
             CLASSES = c(3);
Analysis:    type = MIXTURE missing;
             STARTS = 10 2;
             STITERATIONS = 10;
Model:       %OVERALL%
             i s|t1@0 t2@1 t3@2;
             i-s@0;
             i s ON x;
             c#1 ON x;
             c#2 ON x;
Output:      sampstat standardized tech1 TECH8 TECH11;
PLOT:        SERIES = t1 – t3 (s);
             TYPE = PLOT3;
```


Up to this point, only the unconditional model was considered for pedagogical purposes. Furthermore, one disadvantage of conducting class analysis with covariates is that Mplus does not give graphs of the estimated and observed values for each class. Hence, it is wise to conduct initial exploratory analyses with the unconditioned model to at least see how the individual growth trajectories and classes are distributed.

The first ON statement regresses the intercept and slope growth factors onto the time-invariant covariate (X). The second ON statement describes the multinomial logistic regression of the categorical latent variable (C) on the time-invariant covariate (X) when comparing class 1 to classes 2 and 3.

One final recommendation is to obtain the predicted class for each individual by saving the posterior probabilities and class assignments into an output text file. This option can be specified as follows:

```
Title:      LATENT CLASS GROWTH ANALYSIS
Data:       file is 'C:\My Folder\filename.dat';
Variable:   names are id sex t1 t2 t3 x;
            usevar = t1 – t3;
            IDVARIABLE = id;
            missing = all (999);
            CLASSES = c(3);
SAVEDATA:   FILE IS C:\MY FOLDER\CLASSoutput;
            save = cprobabilities;
Analysis:   type = MIXTURE missing;
```

The IDVARIABLE syntax must correspond to the subject identifier variable in the dataset, here it is simply id. The SAVEDATA line specifies the location of the folder and filename 'CLASSoutput' by which the output is saved. The output text file can then be simply opened with MS WordPad and will look like the following:

```
45.000    49.000    53.000    501.000 . . . . . 1.000
50.000    43.000    68.000    502.000 . . . . . 1.000
87.000    78.000    62.000    503.000 . . . . . 2.000
38.000*             38.000    504.000 . . . . . 1.000
54.000    48.000    48.000    505.000 . . . . . 1.000
68.000    87.000    78.000    506.000 . . . . . 1.000
95.000    50.000    45.000    507.000 . . . . . 2.000
52.000*   *                   508.000 . . . . . 1.000
44.000    43.000    98.000    509.000 . . . . . 3.000
```

In this file, the first three columns correspond to the actual observed data for t1, t2, and t3, respectively. The asterisks correspond to missing values. The fourth column corresponds to the id variable. There will be many columns following the id column that contains the posterior probabilities

and class probabilities (not shown here). The column of interest is the very last column at the far right that consists of the latent class number assigned by Mplus. Here, the values range from one to three since a three-class model was estimated. One can easily export this class assignment information back into the original dataset to be used for further analyses, such as conducting a test of mean differences across the classes on the covariates using ANOVA, or using class membership as a predictor for distal outcome. Since Mplus does not give the graphs for each class from a conditioned latent class model, the user may opt to use these individual class assignments as the grouping variable for plotting each class separately using a different software package such as SAS, SPSS, or Excel. These class information can also be used for other analysis using these software packages.

Summary

Conventional growth model approaches such as the multilevel, random-effects model, assumes that the growth trajectories of all individuals can be adequately described using a single estimate of growth parameters. GMM and LCGA relax this assumption and allow for differences in growth parameters across unobserved subpopulations using latent trajectory classes. LCGA estimates a mean growth curve for each class, but no individual variation around the mean growth curve is allowed. GMM, on the other hand, combines the features of the random effects model and LCGA by estimating both mean growth curves for each class and individual variation around these growth curves by estimating growth factor variances for each class.

Acknowledgment

The dataset used in this paper comes from the Iowa Single Parent Project (Ron Simons, P.I.). The authors would like to thank Dr. Simons for letting them use his dataset for their paper.

Short Biographies

Tony Jung's research interests center on longitudinal data analysis and applying statistical methods to family research. He holds a PhD in Human Development and Family Studies from Iowa State University, and is currently pursuing an MS in Statistics.

K. A. S. Wickrama's research interests include social determinants of health and health inequality across the life course, racial/ethnical inequalities in mental and physical health of children and adults, international development and health, and application of advanced statistical methods to social epidemiology. He holds a PhD in Human Development and Family Studies from Iowa State University.

Endnote

⋆ Correspondence address: Institute for Social and Behavioral Research, 2625 North Loop Drive, Suite 500, Ames, Iowa 50010, USA. Email: s2kas@iastate.edu.

References

Bauer, D. J., & Curran, P. J. (2003a). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, **8**, 338–363.

Bauer, D. J., & Curran, P. J. (2003b). Overextraction of latent trajectory classes: Much ado about nothing? Reply to Rindskopf (2003), Muthén (2003), and Cudeck and Henly (2003). *Psychological Methods*, **8**, 384–393.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.

Hill, K. G., White, H. R., Chung, I., Hawkins, J. D., & Catalano, R. F. (2000). Early adult outcomes of adolescent binge drinking: Person- and variable-centered analyses of binge drinking trajectories. *Alcoholism: Clinical & Experimental Research*, **24**, 892–901.

Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods*, **11**, 36–53.

Jackson, K. M., & Sher, K. J. (2005). Similarities and differences of longitudinal phenotypes across alternate indices of alcohol involvement: A methodologic comparison of trajectory approaches. *Psychology of Addictive Behaviors*, **19**, 339–351.

Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research*, **29**, 374–393.

Kreuter, F., & Muthén, B. (2007). Longitudinal modeling of population heterogeneity: Methodological challenges to the analysis of empirically derived criminal trajectory profiles. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in Latent Variable Mixture Models*. Charlotte, NC: Information Age Publishing.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, **88**, 767–778.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, **51**, 201–226.

Muthén, B. (2003). Statistical and substantive checking in growth mixture modeling: Comment on Bauer and Curran (2003). *Psychological Methods*, **8**, 369–377.

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of Quantitative Methodology for the Social Sciences* (pp. 345–368). Newbury Park, CA: Sage Publications.

Muthén, B., & Asparouhov, T. (2006). Growth mixture analysis: Models with non-Gaussian random effects. Forthcoming In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Advances in Longitudinal Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, **24**, 882–891.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology*, **31**, 327–362.

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In L. A. Collins & J. L. Horn (Eds.), *Best Methods for the Analysis of Change* (pp. 92–106). Washington, DC: American Psychological Association.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rindskopf, D. (2003). Mixture or homogeneous? Comment on Bauer and Curran (2003). *Psychological Methods*, **8**, 364–368.

 Social and Personality Psychology Compass 2/1 (2008): 302–317, 10.1111/j.1751-9004.2007.00054.x