Muthén, B. (in press). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (ed.), *Handbook of quantitative methodology for the social sciences.* Newbury Park, CA: Sage.

Chapter 19

LATENT VARIABLE ANALYSIS

Growth Mixture Modeling and Related Techniques for Longitudinal Data

BENGT MUTHÉN

19.1. INTRODUCTION

This chapter gives an overview of recent advances in latent variable analysis. Emphasis is placed on the strength of modeling obtained by using a flexible combination of continuous and categorical latent variables. To focus the discussion and make it manageable in scope, analysis of longitudinal data using growth models will be considered. Continuous latent variables are common in growth modeling in the form of random effects that capture individual variation in development over time. The use of categorical latent variables in growth modeling is, in contrast, perhaps less familiar, and new techniques have recently emerged. The aim of this chapter is to show the usefulness of growth model extensions using categorical latent variables. The discussion also has implications for latent variable analysis of cross-sectional data.

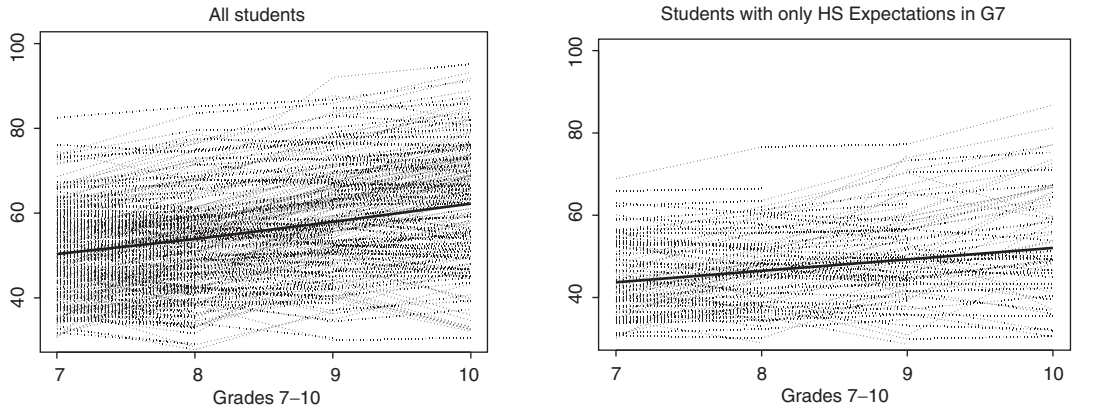
The chapter begins with two major parts corresponding to continuous outcomes versus categorical outcomes. Within each part, conventional modeling using continuous latent variables will be described first, followed by recent extensions that add categorical latent variables. This covers growth mixture modeling, latent class growth analysis, and discrete-time survival analysis. Two additional sections demonstrate further extensions. Analysis of data with strong floor effects gives rise to modeling with an outcome that is part binary and part continuous, and data obtained by cluster sampling give rise to multilevel modeling. All models fit into a general latent variable framework implemented in the Mplus program (Muthén & Muthén, 1998–2003). For overviews of this modeling framework, see Muthén (2002) and Muthén and Asparouhov (2003a, 2003b). Technical aspects are covered in Asparouhov and Muthén (2003a, 2003b).

19.2. CONTINUOUS OUTCOMES: CONVENTIONAL GROWTH MODELING

In this section, conventional growth modeling will be briefly reviewed as a background for the more general growth modeling to follow. To prepare for this

AUTHOR'S NOTE: The research was supported under grant K02 AA 00230 from NIAAA. I thank the Mplus team for software support, Karen Nylund and Frauke Kreuter for research assistance, and Tihomir Asparouhov for helpful comments. Please send correspondence to bmuthen@ucla.edu.

Figure 19.1 LSAY Math Achievement in Grades 7 to 10



transition, the multilevel and mixed linear modeling representation of conventional growth modeling will be related to representations using structural equation modeling and latent variable modeling.

To introduce ideas, consider an example from mathematics achievement research. The Longitudinal Study of Youth (LSAY) is a national sample of mathematics and science achievement of students in U.S. public schools (Miller, Kimmel, Hoffer, & Nelson, 2000). The sample contains 52 schools with an average of about 60 students per school. Achievement scores were obtained by item response theory equating. There were about 60 items per test with partial item overlap across grades. Tailored testing was used so that test results from a previous year influenced the difficulty level of the test of a subsequent year. The LSAY data used here are from Cohort 2, containing a total of 3,102 students followed from Grade 7 to Grade 12 starting in 1987. Individual math trajectories for Grades 7 through 10 are shown in Figure 19.1.

The left-hand side of Figure 19.1 shows typical trajectories from the full sample of students. Approximately linear growth over the grades is seen, with the average linear growth shown as a bold line. Conventional growth modeling is used to estimate the average growth, the amount of variation across individuals in the growth intercepts and slopes, and the influence of covariates on this variation. The right-hand side of Figure 19.1 uses a subset of students defined by one such covariate, considering students who, in seventh grade, expect to get only a high school degree. It is seen that the intercepts and slopes are considerably lower for this group of low-expectation students.

A conventional growth model is formulated as follows for the math achievement development related to educational expectations. For ease of transition between modeling traditions, the multilevel notation of Raudenbush and Bryk (2002) is chosen. For time point t and individual i, consider the variables

y_{ti} = repeated measures on the outcome (e.g., math achievement),
a_{1ti} = time-related variable (time scores) (e.g., Grades 7–10),
a_{2ti} = time-varying covariate (e.g., math course taking),
x_i = time-invariant covariate (e.g., Grade 7 expectations),

and the two-level growth model,

$$\text{Level 1: } y_{ti} = \pi_{0i} + \pi_{1i}\, a_{1ti} + \pi_{2ti}\, a_{2ti} + e_{ti}, \quad (1)$$

$$\text{Level 2: } \begin{cases} \pi_{0i} = \beta_{00} + \beta_{01} x_i + r_{0i} \\ \pi_{1i} = \beta_{10} + \beta_{11} x_i + r_{1i}. \\ \pi_{2i} = \beta_{20} + \beta_{21} x_i + r_{2i} \end{cases} \quad (2)$$

Here, π_{0i}, π_{1i}, and π_{2i} are random intercepts and slopes varying across individuals. The residuals e, r_0, r_1, and r_2 are assumed normally distributed with zero means and uncorrelated with a_1, a_2, and w. The Level 2 residuals r_0, r_1, and r_2 are possibly correlated but uncorrelated with e. The variances of e_t are typically assumed equal across time and uncorrelated across time, but both of these restrictions can be relaxed.[1]

[1]The model may alternatively be expressed as a mixed linear model relating y directly to a_1, a_2, and x by inserting (2) into (1). Analogous to a two-level regression, when either a_{ti} or π_{2ti} varies across i, there is variance heteroscedasticity for y given covariates and therefore not a single covariance matrix for model testing.

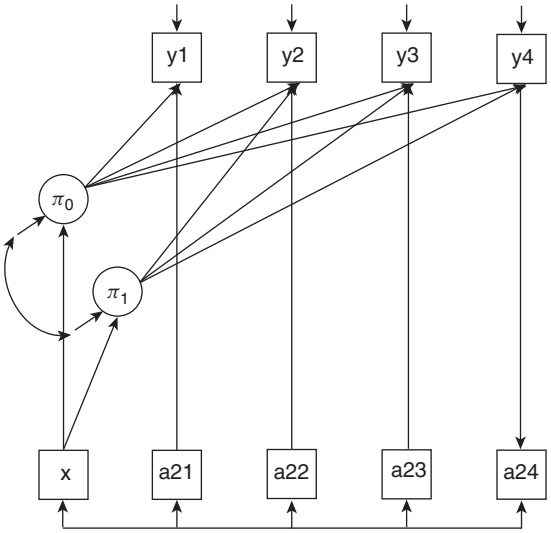
The growth model above is presented as a multilevel, random-effects model. Alternatively, the growth model can be seen as a latent variable model, where the random effects π_0, π_1, and π_2 are latent variables. The latent variables π_0, π_1 will be called growth factors and are of key interest here. As will be shown, the latent variable framework views growth modeling as a single-level analysis. A special case of latent variable modeling is obtained via mean- and covariance-structure structural equation modeling (SEM). Connections between multilevel, latent variable, and SEM growth analysis will now be briefly reviewed.

When there are individually varying times of observation, a_{1ti} in (1) varies across i for given t. In this case, a_{1ti} may be read as data. This means that in conventional multilevel modeling, π_{1i} is a (random) slope for the variable a_{1ti}. When $a_{1ti} = a_{1t}$ for all t values, a reverse view can be taken. In SEM, each a_{1t} is treated as a parameter, where a_{1t} is a slope multiplying the (latent) variable π_{1i}. For example, accelerated or decelerated growth at a third time point may be captured by $a_{1t} = (0, 1, a_3)$, where a_3 is estimated.[2]

Typically in conventional multilevel modeling, the random slope π_{2ti} (1) for the time-varying covariate a_{2t} is taken to be constant across time, $\pi_{2ti} = \pi_{2i}$. It is possible to allow variation across both t and i, although it may be difficult to find evidence for in data. In SEM, however, the slope is not random, $\pi_{2ti} = \pi_{2t}$, because conventional covariance structure modeling cannot handle products of latent and observed continuous variables.

In the latent variable modeling and SEM frameworks, the distinction between Level 1 and Level 2 is not made, but a regular (single-level) analysis is done. This is because the modeling framework considers the T-dimensional vector $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$ as a multivariate outcome, accounting for the correlation across time by the same random effects influencing each of the variables in the outcome vector. In contrast, multilevel modeling typically views the outcome as univariate, accounting for the correlation across time by the two levels of the model. From the latent variable and SEM perspective, (1) may be seen as the measurement part of the model where the growth factors π_0 and π_1 are measured by the multiple indicators y_t. In (2), the structural part of the model relates growth factors and random slopes to other variables. A growth model diagram corresponding to the SEM perspective is shown in Figure 19.2, where circles correspond to latent variables and boxes correspond to observed variables.

Figure 19.2 Growth Model Diagram

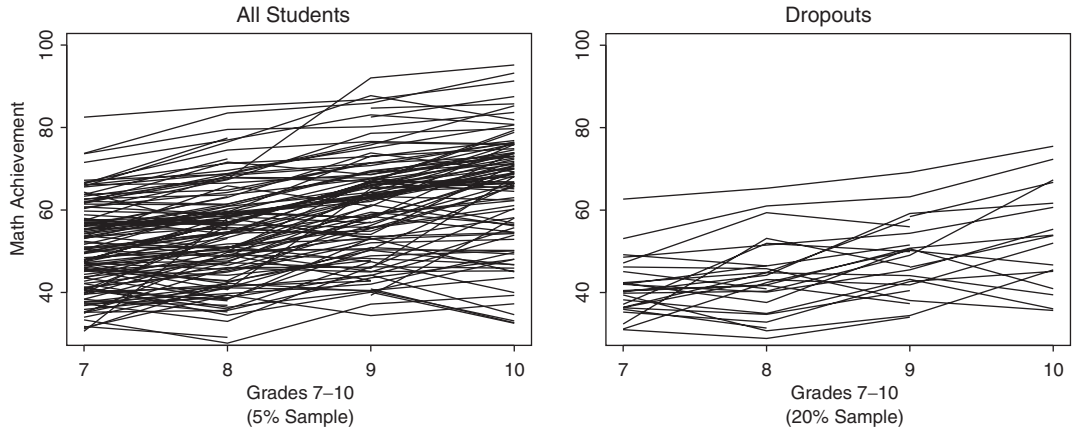


There are several advantages of placing the growth model in an SEM or latent variable context. Growth factors may be regressed on each other—for example, studying growth while controlling for not only observed covariates but also the initial status growth factor. Or, a researcher may want to study growth in a latent variable construct measured with multiple indicators. Other advantages of growth modeling in a latent variable framework include the ease with which to carry out analysis of multiple processes, both parallel in time and sequential, as well as multiple groups with different covariance structures. More generally, the growth model may be only a part of a larger model, including, for instance, a factor analysis measurement part for covariates measured with errors, a mediational path analysis part for variables influencing the growth factors, or a set of variables that are influenced by the growth process (distal outcomes).

The more general latent variable approach to growth goes beyond the SEM approach by handling (1) as stated (i.e., allowing individually varying times of observation and random slopes for time-varying covariates). Here, $a_{1ti} = a_{1t}$ and $\pi_{2ti} = \pi_{2t}$ are allowed as special cases. The latent variable approach thereby combines the strength of conventional multilevel modeling and SEM. An overview showing the advantages of this combined type of modeling is given in Muthén

[2] When choosing $a_{11} = 0$, π_{0i} is defined as the initial status of the growth process. In multilevel analysis, a_{1ti} is often centered at the mean (e.g., to avoid collinearity when using quadratic growth), whereas in SEM, parameters may get highly correlated.

Figure 19.3 LSAY Math Achievement in Grades 7 to 10 and High School Dropout



and Asparouhov (2003a), and a technical background is given in Asparouhov and Muthén (2003a). In addition, general latent variable modeling allows modeling with a combination of continuous and categorical latent variables to more realistically represent longitudinal data. This aspect is the focus of the current chapter.

19.3. Continuous Outcomes: Growth Mixture Modeling

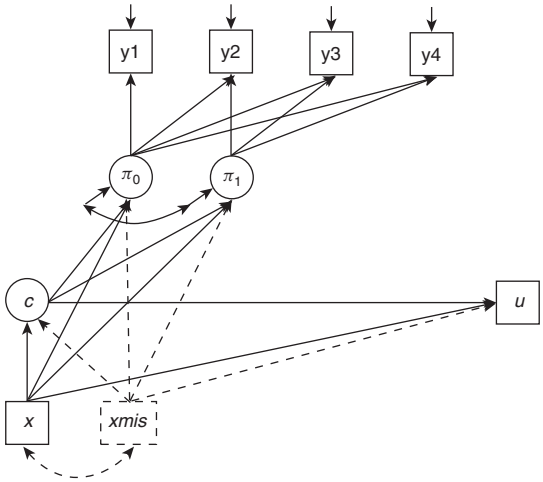
The model in (1) and (2) has two key features. On one hand, it allows individual differences in development over time because the growth intercept π_{0i} and growth slope π_{1i} vary across individuals, resulting in individually varying trajectories for y_{ti} over time. This heterogeneity is captured by random effects (i.e., continuous latent variables). On the other hand, it assumes that all individuals are drawn from a single population with common population parameters. Growth mixture modeling relaxes the single population assumption to allow for parameter differences across unobserved subpopulations. This is accomplished using latent trajectory classes (i.e., categorical latent variables). This implies that instead of considering individual variation around a single mean growth curve, the growth mixture model allows different classes of individuals to vary around different mean growth curves. The combined use of continuous and categorical latent variables provides a very flexible analysis framework. Growth mixture modeling was introduced in Muthén and Shedden (1999) with extensions and overviews in Muthén and Muthén (1998–2003) and Muthén (2001a, 2001b, 2002).

Consider again the math achievement example and the math development shown in the right-hand part of Figure 19.3. This is the development for individuals who are later classified as having dropped out by Grade 12. Note that although Figure 19.1 considers an antecedent of development, Grade 7 expectations, Figure 19.3 considers a consequence of development, high school dropout. It is seen that, with a few exceptions, the high school dropouts typically have a lower starting point in Grade 7 and grow slower than the average students in the left-hand part of the figure. This suggests that there might be an unobserved subpopulation of students who, in Grades 7 through 10, show poor math development and who have a high risk for dropout. In educational dropout research, such a subpopulation is often referred to as "disengaged," where disengagement has many hypothesized predictors. The subpopulation membership is not known during Grades 7 through 10 but is revealed when students drop out of high school. The subpopulation membership can, however, be inferred from the Grade 7 through 10 math achievement development.

19.3.1. Growth Mixture Model Specification

To introduce growth mixture modeling (GMM), consider a latent categorical variable c_i representing the unobserved subpopulation membership for student i, $c_i = 1, 2, \ldots, K$. Here, c will be referred to as a latent class variable or, more specifically, a trajectory class variable. Assume tentatively that in the math achievement example, $K = 2$, representing a disengaged class ($c = 1$) and a normative class ($c = 2$). An example of the different parts of the model is shown

Figure 19.4 GGMM Diagram



in the model diagram in Figure 19.4. The model has covariates x and *xmis*, a latent class variable c, repeated continuous outcomes y, and a distal dichotomous outcome u. For simplicity, time-varying covariates are not included in this example. The covariate x influences c and has direct effects on the growth factors π_0 and π_1, as well as a direct effect on u. In this section, the *xmis* covariate will be assumed to have no role in the model. Its effects will be studied in later sections.

Consider first the prediction of the latent class variable by the covariate x using a multinomial logistic regression model for K classes,

$$P(c_i = k|x_i) = \frac{e^{\gamma_{0k}+\gamma_{1k}x_i}}{\sum_{s=1}^{K} e^{\gamma_{0s}+\gamma_{1s}x_i}}, \qquad (3)$$

with the standardization $\gamma_{0K} = 0$, $\gamma_{1K} = 0$. With a binary $c(c = 1, 2)$, this gives

$$P(c_i = 1|x_i) = \frac{1}{1 + e^{-l_i}}, \qquad (4)$$

where l is the logit (i.e., the log odds),

$$\log[P(c_i = 1|x_i)/P(c_i = 2|x_i)] = \gamma_{01} + \gamma_{11}\, x_i, \qquad (5)$$

so that γ_{11} is the increase in the log odds of being in the disengaged versus the normative class for a unit increase in x. For example, assume that x is dichotomous and scored 0, 1 for females versus males. From (4), it follows that $e^{\gamma_{11}}$ is the odds ratio for being in the disengaged class versus the normative class when comparing males to females. For example, $\gamma_{11} = 1$ implies that the odds of being in the disengaged class versus the normative class is $e^1 = 2.72$ times higher for males than females.

Generalizing (1) and (2), GMM considers a separate growth model for each of the two latent classes. Key differences across classes is typically found in the fixed effects β_{00}, β_{10}, and β_{20} in (2). For example, the disengaged class would have lower β_{00} and β_{10} values (i.e., lower means) than the normative class. Class differences may also be found in the covariate influence, with class-varying β_{01}, β_{11}, and β_{21}. In addition, class-varying variances and covariances for the r residuals may be found. In (1), the type of growth function for Level 1 is perhaps different across class as well. For example, although the disengaged class may be well represented by linear growth, the normative class may show accelerated growth over some of the grades (e.g., calling for a quadratic growth curve). Here, the variance for the e residual may also be class varying.

The basic GMM can be extended in many ways. One important extension is to include an outcome that is predicted from the growth. Such an outcome is often referred to as a *distal outcome*, whereas in this context, the growth outcomes are referred to as *proximal outcomes.* Dropping out of high school is an example of such a distal outcome in the math achievement context. Given that the growth is succinctly summarized by the latent trajectory class variable, it is natural to let the latent trajectory class variable predict the distal outcome. With the example of a dichotomous distal outcome u scored 0, 1, this model part is given as a logistic regression with covariates c and x,

$$P(u_i = 1|c_i = k, x_i) = \frac{1}{1 + e^{\tau_k - \kappa_k x_i}}, \qquad (6)$$

where the main effect of c is captured by the class-varying thresholds τ_k (an intercept with its sign reversed), and κ_k is a class-varying slope for x. For each class, the same odds ratio interpretation given above can be applied also here. Model extensions of this type will be referred to as general growth mixture modeling (GGMM).

19.3.1.1. Latent Class Growth Analysis

A special type of growth mixture model has been studied by Nagin and colleagues (see, e.g., Nagin, 1999; Nagin & Land, 1993; Roeder, Lynch, & Nagin, 1999) using the SAS procedure PROC TRAJ (Jones, Nagin, & Roeder, 2001). See also the 2001 special issue of *Sociological Methods & Research* (Land, 2001). The models studied by Nagin are characterized by having zero variances and covariances for r in (2); that is, individuals within a class are treated as

homogeneous with respect to their development.[3] Analysis with zero growth factor variances and covariances will be referred to as latent class growth analysis (LCGA) in this chapter. As will be discussed in the context of categorical outcomes, the term LCGA is motivated by it being more similar to latent class analysis than growth modeling.

LCGA may be useful in two major ways. First, LCGA may be used to find cut points on the GMM growth factors. A k-class GMM that has within-class variation may have a model fit similar to that of a $k + m$-class LCGA for some $m > 0$. The extra m classes may be a way to objectively find cut points in the within-class variation of a GMM to the extent that such further grouping is substantively useful. This situation is similar to the relationship between factor analysis and latent class analysis, as discussed in Muthén (2001a), where latent classes of individuals were identified along factor dimensions. From a substantive point of view, however, this poses the challenge of how to determine which latent classes represent fundamentally different trajectories and which represent only minor variations. Second, as pointed out in Nagin's work, the latent classes of LCGA may be viewed as producing a nonparametric representation of the distribution of the growth factors, resulting in a semi-parametric model. This view will be further discussed in the next section.

LCGA is straightforward to specify within the general Mplus framework. The zero variance restriction makes LCGA easy to work with, giving relatively fast convergence. If the model fits the data, the simplicity can be a practically useful feature. Also, LCGA can be used in conjunction with GMM as a starting point for analyses. Section 19.3.4.1 discusses the use of LCGA on data that have been generated by a GMM in which covariates have direct influence on the growth factors. This misapplication leads to serious distortions in the formation of the latent classes.

19.3.1.2. Nonparametric Estimation of Latent Variable Distributions

In the GMM described earlier, the normality assumption for the residuals on Level 1 and Level 2 is applied to each class. Within class, the latent variables of π_0, π_1, and π_2 of (2) may have a nonnormal distribution due to the influence of a possibly nonnormal x covariate, and the distribution of y in (1) is further influenced by possibly nonnormal Level 1 covariates. This implies that the distribution of the outcomes y can be nonnormal within class. Strong nonnormality for y is obtained when latent classes with different means and variances are mixed together.

The normality assumption for the residuals is not innocuous in mixture modeling. Alternative distributions would result in somewhat different latent class formations. The literature on nonparametric estimation of random-effect distributions reflects such a concern, especially with categorical and count outcomes already in nonmixture models. Maximum likelihood estimation for logistic models with random effects typically uses Gauss-Hermite quadrature to integrate out the normal random effects. The quadrature uses fixed nodes and weights for a set of quadrature points. As pointed out by Aitkin (1999), a more flexible distributional form is obtained if both the nodes and the weights are estimated, and this approach is an example of mixture modeling. The mixture modeling approximation to a continuous random-effect distribution, such as a random intercept growth factor, is illustrated in Figure 19.5 using an approximately normal distribution as well as a skewed distribution. In both cases, five nodes and weights are used, corresponding to a mixture with five latent classes. Aitkin argues that the mixture approach may be particularly suitable with categorical outcomes in which the usual normality assumption for the random effects has scarce empirical support. For an overview of related work, see also Heinen (1996); for a more recent discussion in the context of logistic growth analysis, see Hedeker (2000).

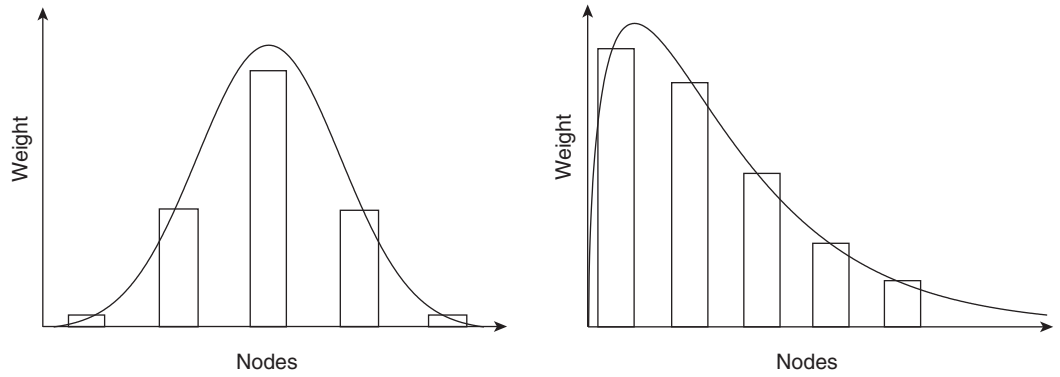
The Mplus latent variable framework can be used for this type of nonparametric approach. Corresponding to Figure 19.5, a random intercept growth factor distribution can be represented by a five-class mixture. Here, the estimation of the nodes is obtained by estimating the growth factor means in the different classes, and the estimation of the weights is obtained by estimating the class probabilities (the growth factor variance parameter is held fixed at zero). If a single-class model is considered, the other parameters of the model are held equal across classes; otherwise, they are not.

19.3.1.3. Growth Mixture Modeling Estimation

The growth mixture model can be estimated by maximum likelihood using an EM algorithm. For a given solution, each individual's probability of

[3] Nagin's work focuses on count data using Poisson distributions. As discussed in later sections, modeling with count outcomes and categorical outcomes can also use nonzero variance for r.

Figure 19.5 Random-Effects Distributions Represented by Mixtures



membership in each class can be estimated, as well as the individual's score on the growth factors π_{0i} and π_{1i}. Measures of classification quality can be considered based on the individual class probabilities, such as entropy. This has been implemented in the Mplus program (Muthén & Muthén, 1998–2003). Technical aspects of the modeling, estimation, and testing are given in Technical Appendix 8 of the *Mplus User's Guide* (Muthén & Muthén, 1998–2003), Muthén and Shedden (1999), and Asparouhov and Muthén (2003a, 2003b). Missing data on y are handled using MAR. Muthén, Jo, and Brown (2003) discuss nonignorable missing data modeling using missing data indicators. As with mixture modeling in general, local optima are often encountered in the likelihood. This phenomenon is well known, for example, in latent class analysis, particularly in models with many classes and data that carry limited information about the class membership. Because of this, the use of several different sets of starting values is recommended, and this is automated in Mplus.

19.3.1.4. The LSAY Example

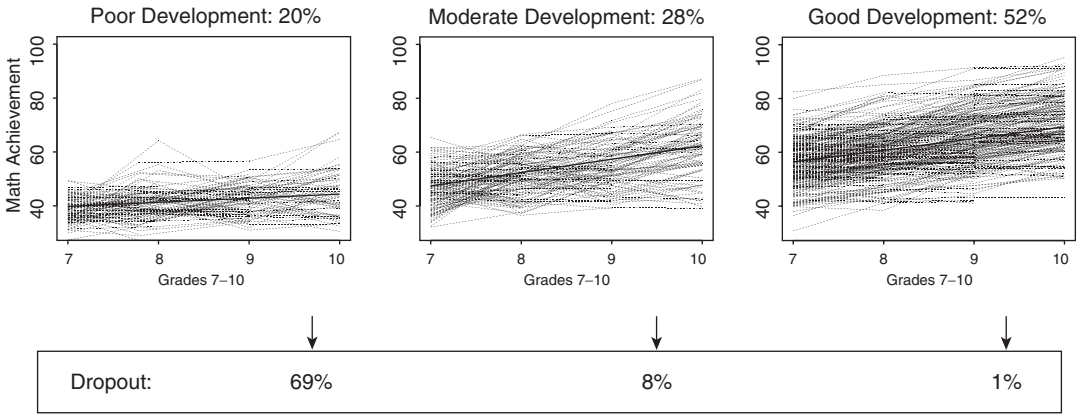
To conclude this section in a concrete way using the LSAY math achievement data, a brief preview of the analyses in Section 3.5 is of interest. Figure 19.6 shows that three latent trajectory classes are found, including their class probabilities, the mean trajectory and individual variation for each class, and the probability of dropping out of high school for each class. Of the students, 20% are found to belong to a disengaged class with poor math development. Membership in the disengaged class dramatically enhances the risk of dropping out of high school, raising the dropout percentage from 1% and 8% to 69%. Section 3.5 presents the covariates predicting latent trajectory class membership, and it is found that having low educational expectations and dropout thoughts already by Grade 7 are key predictors.

Before going through the analysis steps for the LSAY math achievement example, model interpretation, estimation, and model selection procedures will be discussed. Latent variable modeling requires good analysis strategies, and this is even more true in the framework of growth mixture modeling, where both continuous and categorical latent variables are used. Many statistical procedures have been suggested within the related statistical area of finite mixture modeling (see, e.g., McLachlan & Peel, 2000), and some key ideas and new extensions will be briefly reviewed. Both substantive and statistical considerations are critical and will be discussed. Early prediction of class membership is also of interest in growth mixture modeling and will be briefly covered. In the LSAY math achievement example, it is clearly of interest to make such early predictions of risk for high school dropout to make interventions possible.

19.3.2. Substantive Theory and Auxiliary Information for Predicting and Understanding Model Results

GGMM should be investigated using substantively based theory and evidence. Auxiliary information can be used to more fully understand model results even at an exploratory stage, when little theory exists. Once substantive theory has been formulated, it can be used to predict a related set of events that can then be tested.

Figure 19.6 LSAY Math Achievement in Grades 7 to 10 and High School Dropout



Substantive theory building typically does not rely on only a single outcome measured repeatedly, accumulating evidence for a theory only by sorting into classes observed trajectories on a single outcome variable. Instead, many different sources of auxiliary information are used to check the theory's plausibility. Mental health research may find that a pattern of a high level of deviant behavior at ages when this is not typical is often accompanied with a variety of negative social consequences, so that there is a distinct subtype. A good education study of failure in school also considers what else is happening in the student's life, involving predictions of accompanying problems. Gene-environment interaction theories may predict the emergence of problems as a response to adverse life events at certain ages. These are the situations when GGMM is particularly useful. GGMM can include the auxiliary information in the model and test if the classes formed have the characteristics on the auxiliary variables that are predicted by theory. Auxiliary information may take the form of antecedents, concurrent events, or consequences. These are briefly discussed in turn below.

19.3.2.1. Antecedents

Auxiliary information in the form of antecedents (covariates) of class membership and growth factors should be included in the set of covariates to correctly specify the model, find the proper number of classes, and correctly estimate class proportions and class membership. The fact that the "unconditional model" without covariates is not necessarily the most suitable for finding the number of classes has not been fully appreciated and will be discussed below.

An important part of GGMM is the prediction of class membership probabilities from covariates. This gives the profiles of the individuals in the classes. The estimated prediction of class membership is a key feature in examining predictions of theory. If classes are not statistically different with respect to covariates that, according to theory, should distinguish classes, crucial support for the model is absent.

Class variation in the influence of antecedents (covariates) on growth factors or outcomes also provides a better understanding of the data. As a caveat, one should note that if a single-class model has generated the data with significant positive influence of covariates on growth factors, GGMM that incorrectly divides up the trajectories in, say, low, medium, and high classes might find that covariates have lower and insignificant influence in the low class due to selection on the dependent variable. If a GGMM has generated the data, however, the selected subpopulation is the relevant one to which to draw the inference. In either case, GGMM provides considerably more flexibility than what can be achieved with conventional growth modeling. As an example, consider Muthén and Curran's (1997) analysis of a preventive intervention with a strong treatment-baseline interaction. The intervention aimed at changing the trajectory slope of aggressive-disruptive behavior of children in Grades 1 through 7. No main effect was found, but Muthén and Curran used multiple-group latent growth curve modeling to show that the initially more aggressive children benefited from the intervention

in terms of lowering their trajectory slope. The Muthén-Curran technique is not, however, able to capture a nonmonotonic intervention effect that exists for children of medium-range aggression and is absent for the most or least aggressive children. In contrast, such a nonmonotonic intervention effect can be handled using GGMM with the treatment/control dummy variable as a covariate having class-varying slopes (see Muthén et al., 2002). There are probably many cases in which the effect of a covariate is not strong or even present, except in a limited range of the growth factor or outcome.

19.3.2.2. Concurrent Events and Consequences (Distal Outcomes)

Modeling with concurrent events and consequences speaks directly to standard considerations of concurrent and predictive validity. In GGMM, concurrent events can be handled as time-varying covariates that have class-varying effects, as time-varying outcomes predicted by the latent classes, or as parallel growth processes. Consequences can be handled as distal outcomes predicted by the latent classes or as sequential growth processes. Examples of distal outcomes in GGMM include alcohol dependence predicted by heavy drinking trajectory classes (Muthén & Shedden, 1999) and prostate cancer predicted by prostate-specific antigen trajectory classes (Lin, Turnbull, McCulloch, & Slate, 2002).

A very useful feature of GMM, even if a single-class nonnormal growth model cannot be rejected, is that cut points for classification are provided. For instance, individuals in the high class, giving the higher probability for the distal outcome, are identified, whereas this information is not provided by the conventional single-class growth analysis. It is true that this classification is done under a certain set of model assumptions (e.g., within-class conditional normality of outcomes given covariates), but even if the classification is not indisputable, it is nevertheless likely to be useful in practice. In single-class analysis, one may estimate individuals' values on the growth factors and attempt a classification, but it can be very difficult to identify cut points, and the classification is inefficient. The added classification information in GMM versus conventional single-class growth modeling is analogous to the earlier discussion of latent class and latent profile analysis adding complementary information to factor analysis. In addition, GMM classification is an important tool for early detection of likely membership in a problematic class, as will be discussed in the example below.

19.3.3. Statistical Aspects of Growth Mixture Modeling: Studying Model Estimation Quality and Power by Monte Carlo Simulation Studies

Because growth mixture modeling is a relatively new technique, rather little is known about requirements in terms of sample size and the number of time points needed for good estimation and strong power. Monte Carlo studies are useful for gaining understanding about this. Figure 19.4 shows a prototypical growth mixture model with a distal outcome. The following is a brief description of how a Monte Carlo study can be carried out based on this model using Mplus. For background about Monte Carlo studies of latent variable models using Mplus, see Muthén and Muthén (2002). As argued in that article, general rules of thumb are not likely to be dependable, but Monte Carlo studies can be done in settings similar to those of the study at hand.

A total of 100 data sets were generated according to the Figure 19.4 model without the *xmis* covariates, using a sample size of 3,000, similar to that of LSAY. Here, the class percentages are 27% and 73%. Maximum likelihood estimation was carried out and results summarized over the 100 replications. The Mplus output contains average parameter estimates, parameter estimate standard deviations, average standard errors, 95% coverages, and power estimates. Here, *power* refers to the proportion of replications in which the hypothesis that the parameter value is zero is rejected.[4]

The results indicate very good estimation of parameters and standard errors as well as good coverage. The quality is a function of the sample size, the number of time points, the separation between the classes, and the within-class variation. Here, the intercept growth factor means in the two classes are one standard deviation apart. As examples of the power estimates, the regression coefficient for the slope growth factor on the covariate is 0.43 for the smaller class, which has a smaller coefficient, and 1.00 for the larger class, which has a larger coefficient. Changing the sample size to 300, the results are still acceptable, although the power estimates for the slope growth factor regression coefficients are now reduced to 0.11 and 0.83.

The Mplus Monte Carlo facility is quite flexible. For example, to study model misspecification, one could analyze a different model than the one that generated the data. In latent class models, the misspecification may concern the number of classes. For Monte Carlo

[4]The Mplus input and output for this analysis are given in Example 1 at www.statmodel.com/mplus/examples/penn.html.

designs that are not offered in Mplus, externally generated data can be analyzed using the RUNALL utility.[5] An extensive Monte Carlo study of growth mixture and related factor mixture models is given in Lubke and Muthén (2003).

19.3.4. Statistical Aspects of Growth Mixture Modeling: Model Selection Procedures

This section gives an overview of strategies and methods for model selection and testing. An emphasis is placed on practical analysis steps and recent testing developments.

19.3.4.1. Analysis Steps

In conventional growth modeling, a common analysis strategy is to first consider an "unconditional model" (i.e., not introducing covariates for the growth factors). This strategy can lead to confusion with growth mixture modeling. Consider the growth mixture model diagram shown earlier in Figure 19.4. Here the model has covariates x and *xmis*, a latent class variable c, repeated continuous outcomes y, and a distal dichotomous outcome u. The covariate x influences c, has direct effects on the growth factors π_0 and π_1, and also has direct effects on u.

Consider first an analysis of this model without u and without the xs. Here, the class formation is based on information from the observed variables y, channeled through the growth factors. A distorted analysis is obtained if the xs are excluded because they have direct effects on the growth factors. This is because the only observed variables, y, are incorrectly related to c if the xs are excluded. The distortion can be understood based on the analogy of a misspecified regression analysis. Leaving out an important predictor, the slope for the other predictor is distorted. In Figure 19.4, the other predictor is the latent class variable c, and the distortion of its effect on the growth factors causes incorrect evaluation of the posterior probabilities in the E step and therefore incorrect class probability estimates and incorrect individual classification. If, on the other hand, the x covariates do not have a direct influence on the growth factors (and no direct influence on y), the "unconditional model" without the xs would be correct, giving correct class probabilities and growth curves for y.

To further explicate the reasoning above, consider a data set generated by the model in Figure 19.4 without the *xmis* covariate, using the Monte Carlo feature of Mplus discussed earlier.[6] Analysis of the generated data by the correct model recovers the population parameters well, as expected. The estimated Class 1 probability of 0.26 is close to the true value of 0.27. The entropy is not large, despite the correctness of the model, 0.57, but this is a function of the degree of separation between the classes and the within-class variation. In line with the discussion above, the influence of the covariate x is of special interest. The model that generated the data has a positive slope for the influence of x on being in the smaller Class 1, positive slopes for the influence on the growth factors, and a positive slope for the influence on u. The estimated class-specific means and variances of the x covariate are 0.63 and 0.79 for Class 1 and -0.20 and 0.82 for Class 2. The higher mean for Class 1 is expected, given the positive slope for the influence on the Class 1 membership. Being in Class 1, in turn, implies higher means for the growth factors. Within class, the growth factor means are higher due to the direct positive influence of x on the growth factors. With x left out of the model, the latent class variable alone needs to account for the differences in growth factor values across individuals. As a result, the class probabilities are misestimated. In the generated data example, the Class 1 probability is now misestimated as 0.35.[7]

Analyzing the Figure 19.4 model excluding u but correctly including x gives the correct answer in terms of class membership probabilities for c and growth curves for y. This is because excluding u does not imply that the observed variables (y or x) are incorrectly related to c. Excluding u simply makes the standard errors larger and worsens the classification precision (entropy). In the generated data example, the Class 1 probability is well estimated as 0.26, whereas the entropy is now lowered to 0.50.[8]

In practice, model estimation with and without a distal outcome u may give different results for the class probabilities and growth curves for two reasons. First, if you include u but misspecify the model by not allowing direct effects from the xs to u, you get distorted parameter estimates (e.g., incorrect class probabilities) by the same regression misspecification analogy given above. In the generated data example,

[5] See http://www.statmodel.com/runutil.html.

[6] The Mplus input and output for this analysis is given in Example 2 at www.statmodel.com/mplus/examples/penn.html.

[7] The Mplus input and output for this analysis is given in Example 3 at www.statmodel.com/mplus/examples/penn.html.

[8] The Mplus input and output for this analysis are given in Example 4 at www.statmodel.com/mplus/examples/penn.html.

this misspecification gave the strongly distorted Class 1 probability estimate as 0.40. Second, key covariates may have been left out of the model (i.e., may not have been measured or are missing), causing a model misspecification. The notation *xmis* in Figure 19.4 refers to such a covariate. Consider two cases, both assuming that *xmis* is not available. First, if *xmis* influences only u and not the growth factors, the analysis excluding u gives correct results, but the analysis including u gives incorrect and hence different results. Second, if *xmis* influences both the growth factors and u, the analyses with and without u give incorrect results and are different.

In conclusion, the proper choice of covariates is important in growth mixture modeling. Substantive theory and previous analyses are needed to make a choice that is sufficiently inclusive. The covariates should be allowed to influence not only class membership but also the growth factors directly, unless there are well-motivated reasons not to. An analysis without covariates can be useful to study different growth in different trajectory classes. However, it should not be expected that the class distribution or individual classification remains the same when adding covariates. It is the model with covariates properly included that gives the better answer.

It should also be noted that choosing the correct within-class variance structure is important. The data above were generated from a model with class-varying variances for the residuals of e in (1). Misspecifying the model by holding these variances equal across class leads to an estimated Class 1 probability of 0.23. Larger distortions would be obtained if the growth factor variances differ across classes.

It is instructive to consider model misspecification results if data generated by the growth mixture model are analyzed by a latent class growth analysis. In the generated data example above, LCGA leads to a misspecified model. The misspecification can be studied in two steps, first by restricting the residual (co)variances and second by also not allowing the direct influence from x to the growth factors. In both cases, the distal outcome is u. In the first step, the estimated Class 1 probability is found to be 0.42, a value far off from the true probability of 0.27. In the second step, the estimated Class 1 probability is even more strongly distorted, 0.51. It is noteworthy that the misspecification of not letting x have a direct effect on the growth factors cannot be discovered using LCGA. Note that in the last two analyses, the entropy values are strongly overestimated, 0.80 and 0.85. It is also likely that more than two classes are needed to account for the within-class variation. This implies that some of the classes are merely slight variations on a theme and do not have a substantial meaning.

19.3.4.2. Equivalent Models

With latent variable models in general and mixture models in particular, the phenomenon of equivalent models may be encountered. Here, *equivalent models* means that two or more models fit the same data approximately the same so that there is no statistical basis on which to base a model choice. Consider two psychometric examples. First, in exploratory factor analysis, a rotated solution using uncorrelated factors gives the same estimated correlation matrix as a rotated solution with correlated factors. Second, Bartholomew and Knott (1999, pp. 154–155) point out a well-known psychometric fact that a covariance matrix generated by a latent profile model (a latent class model with continuous outcomes) can be perfectly fitted by a factor analysis model. A covariance matrix from a k-class model can be fitted by a factor analysis model with $k - 1$ factors. Molenaar and von Eye (1994) show that a covariance matrix generated by a factor model can be fitted by a latent class model. This should not be seen as a problem but merely as two ways of looking at the same reality. The factor analysis informs about underlying dimensions and how they are measured by the items, whereas the latent profile analysis sorts individuals into clusters of individuals who are homogeneous with respect to the item responses. The two analyses are not competing but are complementary.

The issue of alternative explanations is classic in finite mixture statistics. Mixtures have two separate uses. One is to simply fit a nonnormal distribution without a particular interest in the mixture components. The other is to capture substantively meaningful subgroups. For a historical overview, see, for instance, McLachlan and Peel (2000, pp. 14–17), who refer to a debate about blood pressure. A classic example concerns data from a univariate (single-class) lognormal distribution that are fitted well by a two-class model that assumes within-class normality and has different means. Bauer and Curran (2003) consider the analogous multivariate case arising with growth mixture modeling.[9] The authors use a Monte Carlo simulation study to show that a multiclass growth mixture model can be arrived at using conventional Bayesian information criterion (BIC) approaches (see below) to determine the number of classes when data, in fact, have been generated by a nonnormal multivariate

[9] Multivariate formulas that show equivalence are not given.

distribution that is skewed and kurtotic. Although the authors only consider GMM, the resulting overextraction of classes would be more pronounced for LCGA. Bauer and Curran's study serves as a caution to researchers to not automatically assume that the latent trajectory classes of a growth mixture model have substantive meaning. Their article is followed by three commentaries and a rejoinder that place the discussion in a larger context. Two of the commentaries, including one by Muthén (2003), point out that BIC does not address model fit to data but is a relative fit measure comparing competing models. Muthén discusses new mixture tests that aim to address data fit, which are mentioned below. The use of these alternative models ultimately has to be guided by arguments related to substantive theory, auxiliary information, predictive validity, and practical usefulness.

19.3.4.3. Conventional Mixture Tests

The selection of the number of latent classes has been discussed extensively in the statistical literature on finite mixture modeling (see, e.g., McLachlan & Peel, 2000). The likelihood ratio comparing a $k - 1$ and a k-class model does not have the usual large-sample chi-square distribution due to the class probability parameter being at the border (zero) of its admissible space. A commonly used alternative procedure is the BIC (Schwartz, 1978), defined as

$$\text{BIC} = -2 \log L + p \ln n, \tag{7}$$

where p is the number of parameters and n is the sample size. Here, BIC is scaled so that a small value corresponds to a good model with a large log-likelihood value and not too many parameters.

Consider as an example the generated data example of the previous section. Here, the analysis without the x covariate or the u distal outcome gave the following BIC values for one, two, and three classes: 39,676.166, 39,603.274, and 39,610.785. This points correctly to two classes, despite the fact that the model is misspecified due to not including x and its direct effect on the growth factors. This fortunate outcome cannot be relied on, however.

19.3.4.4. New Mixture Tests

This section briefly describes two new mixture test approaches. A key notion is the need for checking how well the mixture model fits the data, not merely basing a model choice on k classes fitting better than $k - 1$ classes. It should be emphasized that there are many possibilities for checking model fit against data in mixture settings, and the methodology for this is likely to expand considerably in the future. One promising approach is the residual diagnostics based on pseudo-classes, proposed in Wang, Brown, and Bandeen-Roche (2002).

Lo, Mendell, and Rubin (2001) proposed a likelihood ratio–based method for testing $k - 1$ classes against k classes. The Lo-Mendell-Rubin approach has been criticized (Jeffries, 2003), although it is unclear to which extent the critique affects its use in practice. The Lo-Mendell-Rubin likelihood ratio test (LMR LRT) avoids a classic problem of chi-square testing based on likelihood ratios. This concerns models that are nested, but the more restricted model is obtained from the less restricted model by a parameter assuming a value on the border of the admissible parameter space—in the present case, a latent class probability being zero. It is well known that such likelihood ratios do not follow a chi-square distribution. Lo, Mendell, and Rubin consider the same likelihood ratio but derive its correct distribution. A low p-value indicates that the $k - 1$-class model has to be rejected in favor of a model with at least k classes. The Mplus implementation uses the usual Mplus mixture modeling assumption of within-class conditional normality of the outcomes given the covariates. When nonnormal covariates are present, this allows a certain degree of within-class nonnormality of the outcomes. The LMR LRT procedure has been studied for GMMs by Monte Carlo simulations (Masyn, 2002). More investigations of performance in practice are, however, of interest, and readers can easily conduct studies using the Mplus Monte Carlo facility for mixtures.

Muthén and Asparouhov (2002) proposed a new approach for testing the fit of a k-class mixture model for continuous outcomes. As opposed to the LMR LRT, this procedure concerns a test of a specific model's fit against data. The procedure relies on testing if the multivariate skewness and kurtosis (SK) estimated by the model fit the corresponding sample quantities. The sampling distributions of the SK tests are assessed by computing these values over a number of replications in data generated from the estimated mixture model. Obtaining low p-values for skewness and kurtosis indicates that the k-class model does not fit the data. Univariate and bivariate test results are also provided for each variable and pair of variables. These tests may provide a useful complement to the LMR LRT. Currently, the SK tests are not available with missing data. Given the inherent sensitivity to outliers, the SK testing should be preceded by outlier investigations.

The SK procedure needs further investigation but is offered here as an example of the many possibilities of testing a mixture model against data (see also Wang et al., 2002).

19.3.5. The LSAY Math Achievement Example

This section returns to the analysis of the mathematics achievement data from the LSAY data mentioned earlier. Based on the educational literature, the following covariates are included: female; Hispanic; Black; mother's education; home resources; the student's educational expectations, measured in seventh grade (1 = high school only, 2 = vocational training, 3 = some college, 4 = bachelor's degree, 5 = master's degree, 6 = doctorate); the student's thoughts of dropping out, measured in seventh grade; whether the student has ever been arrested, measured by seventh grade; and whether the student has ever been expelled by seventh grade. Corresponding to individuals with complete data on the covariates, the analyses consider a subsample of 2,757 of the total 3,116 individuals. The analyses were carried out by maximum likelihood estimation using Mplus Version 2.13.

19.3.5.1. Statistical Checking

The univariate skewness and kurtosis sample values in the LSAY data are as follows:

$$\text{Skewness} = (0.168 \;\; 0.030 \;\; 0.063 \;\; -0.077), \quad (8)$$

$$\text{Kurtosis} = (-0.551 \;\; -0.338 \;\; -0.602 \;\; -0.559). \quad (9)$$

In line with the earlier discussion of the LMR LRT, due to the low nonnormality in the outcomes, it is plausible that this test is applicable in the LSAY analysis for testing a one-class model versus more than one class. In the LSAY analysis, this test points to at least two classes with a strong rejection ($p = .0000$) of the one-class model. The SK tests carried out on the listwise present subsample of 1,538 reject the one-class model ($p = .0000$ for both multivariate skewness and multivariate kurtosis) but do not reject two classes ($p = .4300$ and .5800). The LMR LRT for two versus three or more classes obtained a high p-value (.6143) in support of two classes. Taken together, the statistical evidence points to at least two classes. Given that the skewness and kurtosis tests found that two- and three-class GMMs fit the data, the LMR LRT is useful for testing the multiclass alternatives against each other.

19.3.5.2. Substantive Checking and Further Statistical Analysis

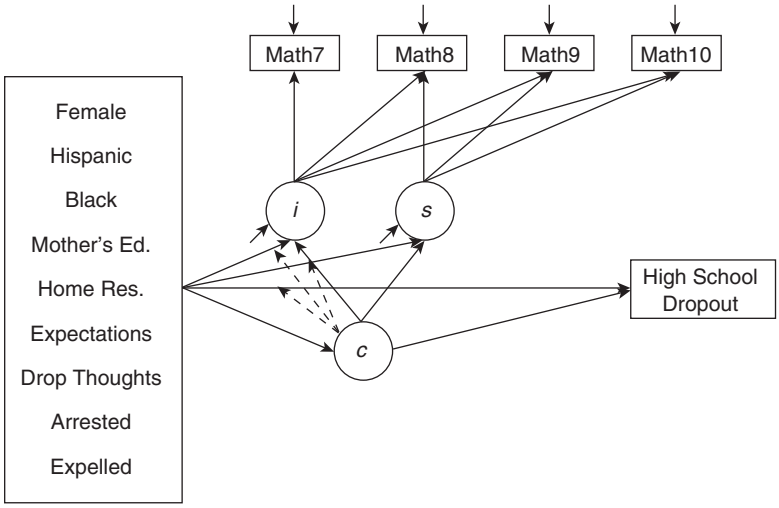
This section compares analysis results using a conventional one-class growth model and different forms of GMMs and discusses substantive meaningfulness based on educational theory, auxiliary information, and practical usefulness. Figure 19.7 shows a diagram of the general model.

19.3.5.2.1. Conventional one-class growth modeling. As a first step, the conventional one-class growth model results are considered. Briefly stated, a linear growth model fits reasonably well and has a positive growth rate mean of about 1 standard deviation across the four grades. The covariates with significant influence (sign in parentheses) on the initial status are as follows: female (+), Hispanic (−), Black (−), mother's education (+), home resources (+), expectations (+), dropout thoughts (−), arrest (−), and expelled (−). The covariates with significant influence (sign in parentheses) on the growth rate are as follows: female (−), Hispanic (−), home resources (+), expectations (+), and expelled (−).

19.3.5.2.2. Two-class GMM. The two-class solution is characterized by a low class of 41%, which, in comparison to the high class, has a lower initial status mean and variance, a lower growth rate mean, and a higher growth rate variance. It is interesting to consider what characterizes these students apart from their poor mathematics achievement development. The multinomial logistic regression for class membership indicates that, relative to the high class, the odds of membership in the low class are significantly increased by being male, being Hispanic, having a mother with a low level of education, having low seventh-grade educational expectations, having had seventh-grade thoughts of dropping out, having been arrested, and having been expelled. The low class appears to be a class of students with problems both in and out of school. The profile of the low class is reminiscent of individuals at risk for dropping out of high school (see, e.g., Rumberger & Larson, 1998, and references therein). Many of these students are "disengaged," to use language from high school dropout theories.

The within-class influence of the covariates on the initial status and growth rate factors varies significantly across class. The low class has no significant predictors of growth rate, whereas the growth rates of the two higher classes are significantly enhanced in well-known ways by being male, having a mother with a high level of education, having high home resources,

Figure 19.7 GGMM Diagram for LSAY Data



and having high expectations. To the extent that the low class has substantive meaning, the findings suggest that different processes are in play for students in the low class.

19.3.5.2.3. Three-class GMM including a distal outcome. To more specifically investigate the data from the high school dropout perspective and further characterize the low class, the distal binary outcome of dropping out of high school, as recorded in Grade 12, was added. The overall dropout rate in the sample is 14.7%, or 458 individuals. Here, class membership in the GMM is, to some extent, also determined by the Grade 12 dropout indicator and not only by the covariates and math achievement development. Adding the distal outcome, the LMR LRT rejected the two-class model in favor of at least three classes ($p = .0060$). The three-class solution produces a more distinct low class of 19%, a middle class of 28%, and a high class of 52%. Here, the low class (estimated as 536 students) has a lower growth rate mean and lower growth rate variance than in the two-class solution without the distal outcome.[10]

[10]The Akaike information criterion (AIC) points to at least three classes, whereas the Bayesian information criterion (BIC) points to two classes. The one-class log-likelihood, number of parameters, AIC, and BIC values are as follows: −30,021.955, 27, 60,097.909, and 60,257.791. The two-class log-likelihood, number of parameters, AIC, BIC, and entropy values are as follows: −29,676.457, 63, 59,478.914, 59,851.971, and 0.552. The three-class log-likelihood, number of parameters, AIC, BIC, and entropy values are as follows: −29,566.679, 99, 59,331.359, 59,917.591, and 0.620.

The class membership regression part of the model indicates that for the low class relative to the highest class, the same covariates as in the two-class solution are significant, except that Hispanic and mother's education are insignificant, whereas Black and home resources are significant. Interestingly, comparing the middle class to the high class, the disengagement covariates of low educational expectations, seventh-grade dropout thoughts, having been arrested, and having been expelled are no longer significant. This suggests that the low class is now a more distinct class that is more specifically characterized as disengaged and at risk for high school dropout. The two higher classes may or may not make a substantively meaningful distinction among students, but their presence helps to isolate the low class. In a two-class solution including the distal outcome, the low class is not very different from the more unspecific low class of the initial two-class solution without the distal outcome. It is interesting to note that although the LMR LRT does not point to three classes without the distal outcome, the three-class solution without the distal outcome shows a similar low class as in the three-class solution with the distal outcome. As will be shown next, the three-class solution with the distal outcome gets not only statistical support from the LMR LRT but also substantive support from predicting dropout.

Further bolstering the notion that the low class is prone to high school dropout, the probability of dropping out, as estimated from the three-class model, is distinctly different in the low class. The probabilities are .692 for the low class, .076 for the middle class,

and .006 for the high class. Other concurrent and distal events were also added to the three-class model to further understand the context of the low class, including responses to the following 10th grade question: "How many of your friends will drop out before graduating from high school?" (1 = *none*, 2 = *a few*, 3 = *some*, 4 = *most*). Treating this as an ordered polytomous outcome influenced by class and the covariates resulted in estimated probabilities for response in either of the three highest categories (few, some, most): .259 for the low class, .117 for the middle class, and .030 for the high class. Considerably more students in the low class have friends who are also thinking of dropping out. In contrast, heavy alcohol involvement in Grade 10 was not distinctly different in the low class. The estimated growth curves and individual trajectories can be seen in Figure 19.6.

19.3.5.2.4. Practical usefulness. An educational researcher is likely to find it interesting that the analyses suggest that dropout by Grade 12 can be predicted already by the end of Grade 10 with the help of information on problematic math achievement development. Whether the division into growth mixture classes is meaningful is largely a substantive question. An argument in favor of there being a distinct "failing class" is obtained from the distal outcome of high school dropout. The fact that the dropout percentage is dramatically higher for the low class than for the other two, 69% versus 8% and 1%, suggests that the three classes are not merely gradations on an achievement development scale but that the low class represents a distinct group of students.

From the point of view of intervention, it is valuable to explore whether a dependable classification into the low class can be achieved earlier than Grade 10. GGMM can help answer this question. For example, by Grade 7, the covariates and the first math achievement outcome are available, and given the estimated three-class model, new students can be classified based on the model and their Grade 7 data. GGMM allows the investigation of whether this information is sufficient or if math achievement trend information provided by adding Grade 8 information (or Grades 8 and 9 information) is needed before a useful classification can be made.

19.4. Categorical Outcomes: Conventional Growth Modeling

With categorical outcomes, the Level 1 model part (1) has to be replaced with a model that describes the probability of the outcome at different time points for different individuals. This model has been studied by Hedeker and Gibbons (1994). Here, logistic regression will be used, so that with the example of a binary outcome u scored 0 and 1,

$$P(u_{ti} = 1|a_{1ti}, a_{2ti}, x_i) = \frac{1}{1 + e^{\tau - logit(u_{ti})}}, \qquad (10)$$

$$\text{Level 1 (Within): } \text{logit}(u_{ti}) = \pi_{0i} + \pi_{1i}\, a_{1ti} + \pi_{2ti} a_{2ti} + e_{ti}, \qquad (11)$$

$$\text{Level 2 (Within): } \begin{cases} \pi_{0i} = \beta_{00} + \beta_{01} x_i + r_{0i} \\ \pi_{1i} = \beta_{10} + \beta_{11} x_i + r_{1i}. \\ \pi_{2i} = \beta_{20} + \beta_{21} x_i + r_{2i} \end{cases} \qquad (12)$$

A perhaps more common parameterization is to fix the threshold parameter τ in (10) at zero, which enables the identification of β_{00}.[11] The variance of e is not a free parameter but is fixed in line with logistic regression. With ordered polytomous outcomes, Mplus uses the proportional odds logistic regression model (see, e.g., Agresti, 1990, pp. 322–324). This may be thought of as a threshold model for a latent response variable, so that with C categories, there is a series of $C-1$ ordered thresholds. The thresholds are held equal across time. As a standardization, $\beta_{00} = 0$ may be chosen, or alternatively, the first threshold may be set at zero. Hedeker and Gibbons (1994) describe maximum likelihood estimation and show that this requires heavier computations than with continuous outcomes, calling on numerical integration using quadrature methods. The computational burden is directly related to the number of random effects (i.e., the number of coefficients π for which the variance of r is not fixed at zero).

19.5. Categorical Outcomes: Growth Mixture Modeling

The conventional growth modeling for categorical outcomes given in (11) and (12) can be extended to growth mixture modeling with latent trajectory classes. This is a new technique introduced in Asparouhov and Muthén (2003b), using maximum likelihood estimation based on an EM algorithm with numerical integration. In line with the latent variable approach to

[11] The Mplus input and output for these analyses are given in Example 5 at www.statmodal.com/mplus/examples/penn.html.

growth modeling with continuous outcomes discussed in Section 19.2, the Asparouhov-Muthén approach allows a_{1ti} in (11) to be handled as data or as parameters to be estimated. Furthermore, the π_{2ti} slopes can be random for the time-varying covariates a_{2ti}.[12] The Hedeker-Gibbons model is obtained as a special case with a single latent class.

As in (3), the covariate effect on class membership is a multinomial logistic regression,

$$P(c_i = k|x_i) = \frac{e^{\gamma_{0k}+\gamma_{1k}x_i}}{\sum_{s=1}^{K} e^{\gamma_{0s}+\gamma_{1s}x_i}}. \quad (13)$$

The growth mixture extension of (10) is

$$P(u_{ti} = 1|a_{1ti}, a_{2ti}, x_i, c_i = k) = \frac{1}{1 + e^{\tau - logit(u_{tik})}}, \quad (14)$$

where the added conditioning on c and the subscript k emphasize that the growth model for u, as expressed by the logits, varies across classes. In line with the extension for continuous outcomes, the different latent classes have different growth models (11) and (12), with key differences typically found in the β coefficients but also in the (co)variances of the Level 2 residuals r. Typically, the thresholds τ would be time and class invariant to represent measurement invariance, although class invariance is not necessary. Generalizations to including distal outcomes u_d, as in (15), is of interest also here:

$$P(u_{di} = 1|c_i = k, x_i) = \frac{1}{1 + e^{\tau_k - \kappa_k x_i}}, \quad (15)$$

with coefficients varying across classes k.

Model building and testing strategies for categorical outcomes are in line with those discussed earlier for continuous outcomes.

19.5.1. Categorical Outcomes: Latent Class Growth Analysis

Latent class growth analysis (LCGA) for categorical outcomes considers the model in (11) through (13) with the restriction of zero variances and covariances for the residuals r. Background references for LCGA include Nagin (1999), Nagin and Land (1993), and Nagin and Tremblay (2001).

[12] Threshold parameters are useful with ordered polytomous outcomes, in which case β_{00} can be fixed at zero, or, alternatively, the first threshold is fixed at zero.

It is instructive to relate LCGA to latent class analysis (LCA). As in LCGA, LCA considers multiple u variables seen as indicators of c and assumed conditionally independent given c. As in LCGA, there are no continuous latent variables to explain further within-class correlation among the u variables. Typically, all outcomes are categorical. Continuous outcomes are, however, possible, giving rise to latent profile analysis. In LCA, the multiple indicators are cross-sectional measures, not longitudinal. When the multiple indicators correspond to repeated measures over time, latent classes may correspond to different trends, and trend structures can be imposed across the indicators' probabilities. To clarify this, consider again (14):

$$P(u_{ti} = 1|a_{1ti}, a_{2ti}, x_i, c_i = k) = \frac{1}{1 + e^{\tau - logit(u_{tik})}}. \quad (16)$$

This means that with, for example, linear growth over T time points, the probabilities of the T u variables are structured according to a logit-linear trend, where the intercept and slope factors have different means across the classes. Note here that τ is held equal across time points. In contrast, LCA considers

$$P(u_{ti} = 1|x_i, c_i = k) = \frac{1}{1 + e^{\tau_{tk}}}, \quad (17)$$

where the τ_{tk} thresholds vary in an unrestricted fashion across the u variables and across the classes. In this way, LCGA gives a more parsimonious description of longitudinal data than LCA.

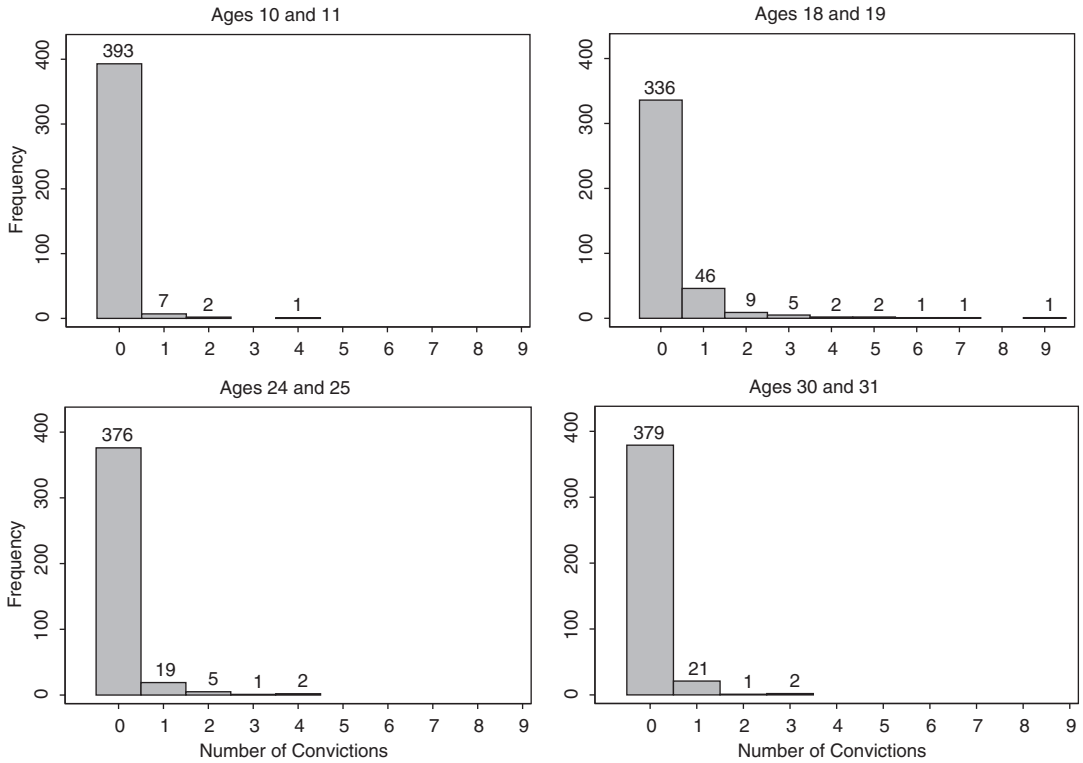
Models with more than one latent class variable are also of interest. Examples of LCGA with multiple-class variables are given in Muthén and Muthén (2000), Muthén (2001a), and Nagin and Tremblay (2001). In this connection, it is useful to consider another important class of growth models, latent transition analysis (LTA). LTA uses time-specific latent class variables measured by multiple indicators at each time point to study class membership change over time.

Both LCA and LTA can be generalized to include random effects as in growth mixture modeling (Asparouhov & Muthén, 2003b). All of these model variations can be captured in a general latent variable modeling framework and are included in Mplus.

19.5.2. Categorical Outcomes: Comparing LCGA and GMM on Delinquency Data

Nagin and Land (1993), Nagin (1999), Roeder et al. (1999), and Jones et al. (2001) used PROC TRAJ

Figure 19.8 Frequency Distributions for Cambridge Data



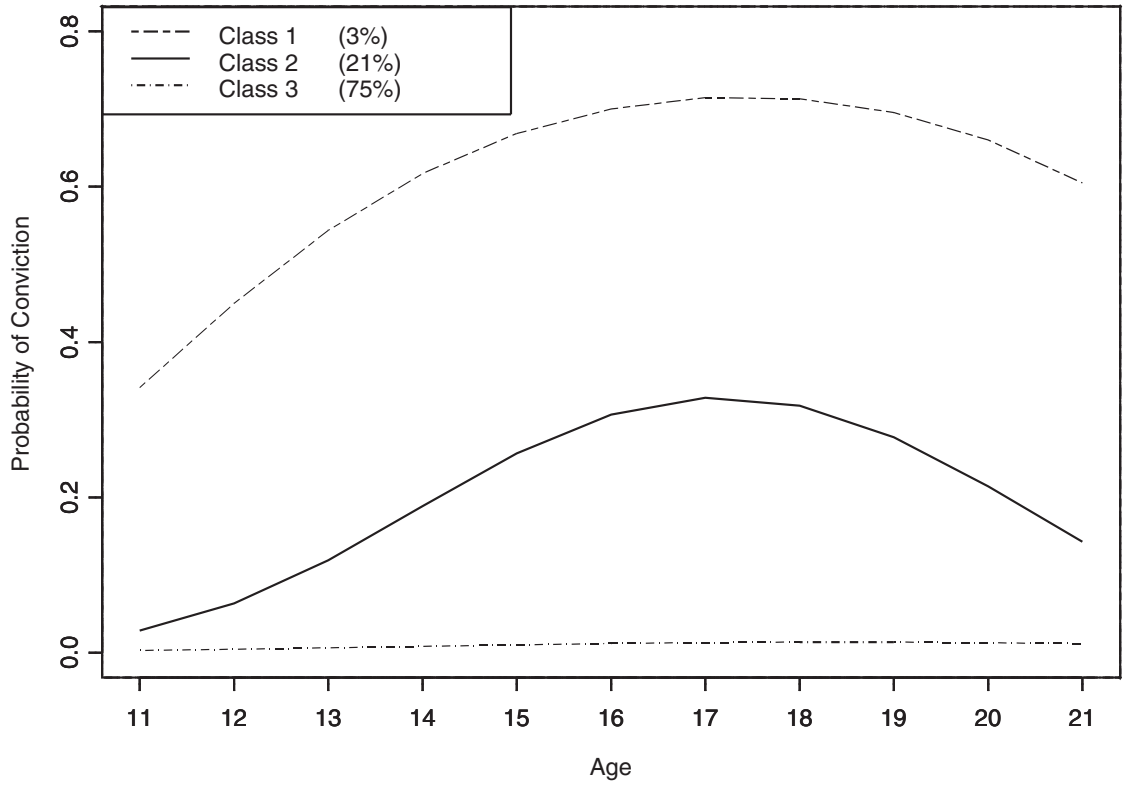
LCGA to study the development of delinquency over ages 10 to 32 in a sample of 411 boys in a working-class section of London (Farrington & West, 1990). These "Cambridge data" were studied from the substantive perspective of the Moffitt (1993) theory of adolescent-limited versus life course–persistent antisocial behavior. This theory suggests two major trajectory classes. Using different ways to aggregate and model the outcomes, Nagin and Land found four classes, Nagin three classes, Roeder et al. four classes, and Jones et al. three classes. The Nagin (1999) approach of considering 2-year intervals and excluding the 8 boys who died during the study will be used here, resulting in 11 time points and $n = 403$. The frequency distributions are shown in Figure 19.8. Only ages 11 to 21 will be used here.

Given that few individuals have more than two convictions in the 2-year interval, data will be coded as 0, 1, and 2 for zero, one, or more convictions in the last 2 years; 65% have 0 value at all 11 time points. A logistic ordered polytomous response model will be used, and three types of analyses will be illustrated: latent class growth analysis, conventional growth modeling, and growth mixture modeling. The analyses draw on Muthén, Kreuter, and Asparouhov (2003).

19.5.2.1. Latent Class Growth Analysis of the Cambridge Data

Latent class growth analysis was performed with two, three, and four classes applying a quadratic growth curve for all classes. The corresponding BIC values were 2,230.014, 2,215.251, and 2,227.976. This points to the three-class model as being the best. This model has a log-likelihood value of –1,071.632, 12 parameters, and an entropy of 0.821. The estimated class percentages are 3%, 21%, and 75%, arranging the curves from high to low. The LMR LRT also points to three classes in that the test of the two-class model against the three-class model has a p-value of .0030, suggesting rejection, whereas the three-class model tested against the four-class model has a p-value of .1554. The estimated three-class growth curves for the

Figure 19.9 Three-Class LCGA for Cambridge Data



probability of having at least one conviction are shown in Figure 19.9.[13]

19.5.2.2. Growth and Growth Mixture Analysis of the Cambridge Data

Conventional one-class growth modeling of the ordered polytomous outcome used a centering of the time scale at age 17 and let the intercept and linear slope growth factors be random, and the quadratic slope factor variance was fixed at zero. The intercept and linear slope were allowed to correlate. This one-class growth model resulted in a log-likelihood value of –1,072.396 with seven parameters and a BIC value of 2,186.785.[14] The linear slope variance is not significant and will, for simplicity, be set to zero in subsequent analyses. In the growth mixture analyses to follow, this intercept variance was allowed to vary across the classes.

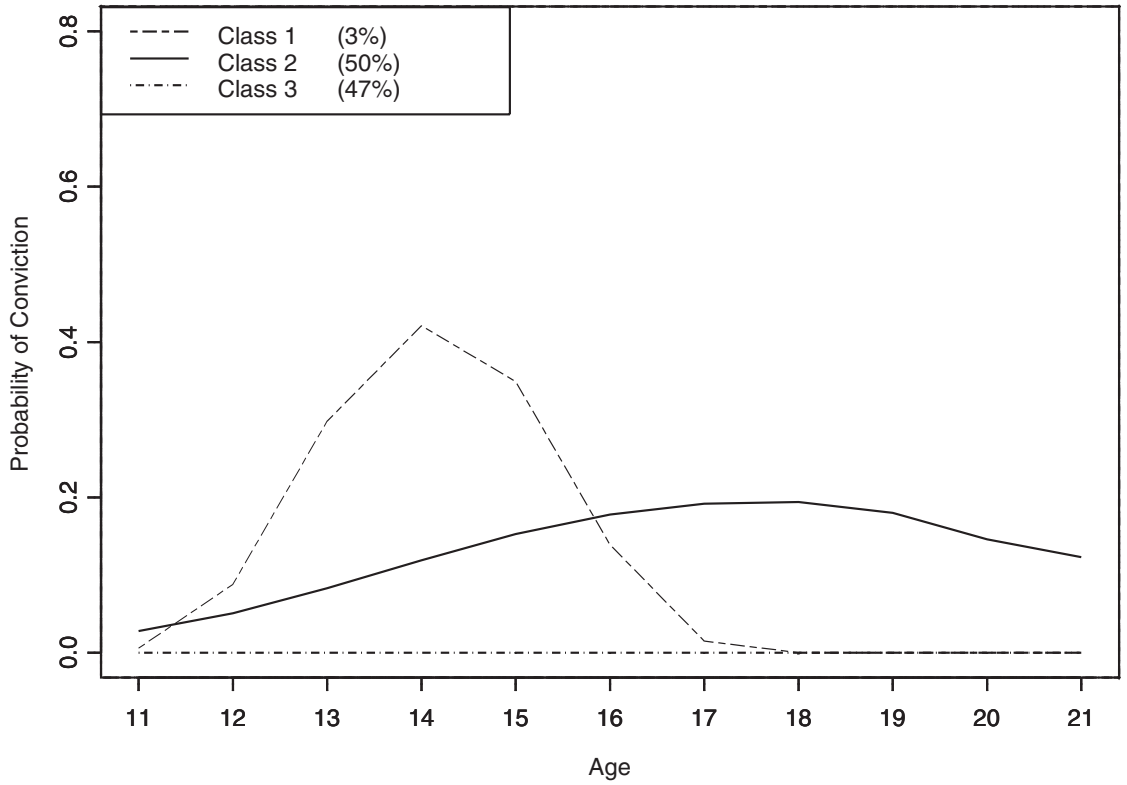
The two-class growth mixture modeling resulted in a log-likelihood value of –1,070.898, a BIC of 2,201.785, 10 parameters, and an entropy of 0.414. The estimated class percentages are 46% and 54%, arranging the classes from high to low. The intercept variance is significant in both classes and lower in the low class. The LMR LRT p-value for one class tested against two classes is .0362, pointing to the need for at least two classes.

A specific three-class growth mixture model was considered next, in which one class was specified to have zero probability of conviction throughout the time period. This zero class corresponds to the notion that some individuals do not get involved in delinquency activities at all. In the other two classes, the intercept variance was allowed to be free to be estimated and different across those classes. This model resulted in a log-likelihood value of –1,066.767, a BIC of 2,199.523, 11 parameters, and an entropy of 0.535. The estimated class percentages are 3%, 50%, and 47%, arranging the classes from high to

[13]The Mplus input and output for these analyses are given in Example 6 at www.statmodel.com/mplus/examples/penn.html.

[14]The Mplus input and output for these analyses are given in Example 7 at www.statmodel.com/mplus/examples/penn.html. This analysis was carried out by Mplus Version 3.

Figure 19.10 Three-Class LCGA for Cambridge Data



low. The intercept variance is nonsignificant for the highest class but significant for the middle class.[15] An interesting finding is that this three-class GMM, which allows within-class variation, has 1 parameter less than the three-class LCGA but a better fit in terms of log-likelihood and BIC values. The zero class is smaller in the GMM than in the LCGA, 47% versus 75%. The fact that 64% of the individuals have observed values at zero throughout, whereas the GMM zero class has only 47% prevalence, is due to the fact that the individuals who are most likely to be in the low class according to the posterior probabilities have a sizable probability of being in the middle class. The estimated three-class growth curves for the probability of having at least one conviction are shown in Figure 19.10. These curves are clearly different from the LCGA curves in Figure 19.9, with Class 1 and Class 2 peaking at different ages for GMM but not for LCGA. This may lead to different substantive interpretations in the context of Moffitt's (1993) theory.

[15]The Mplus input and output for these analyses are given in Example 8 at www.statmodel.com/mplus/examples/penn.html. This analysis was carried out by Mplus Version 3.

19.5.3. Categorical Outcomes: Discrete-Time Survival Analysis

Discrete-time survival analysis (DTSA) uses the categorical variables u to represent events modeled by a logistic hazard function (cf. Muthén & Masyn, in press). For an overview of conventional DTSA, see, for example, Singer and Willet (1993). Consider a set of binary 0/1 variables u_j, $j = 1, 2, \ldots, r$, where $u_{ij} = 1$ if individual i experiences the nonrepeatable event in time period j, and define j_i as the last time period in which data were collected for individual i. The hazard is the probability of experiencing the event in time period j given that it was not experienced prior to j. The hazard is written as

$$h_{ij} = \frac{1}{1 + e^{-(-\tau_j + \kappa_j x_i)}}, \tag{18}$$

where a proportional-odds assumption is obtained by dropping the j subscript for κ_j. Discrete-time survival analysis is fitted into the general mixture model above by noting that the likelihood is the same as for u related to c and x in a single-class model.

The fact that individual i does not have observations on u after time period j_i is handled as missing data. For example, with five time periods ($r = 5$), an individual who experiences the event in Period 4 has the data vector $\mathbf{u}_i'$

$$(0 \quad 0 \quad 0 \quad 1 \quad 999),$$

with 999 representing missing data. An individual who is censored in Period 5 has the data vector $\mathbf{u}_i'$

$$(0 \quad 0 \quad 0 \quad 0 \quad 0),$$

whereas an individual who is censored in Period 4 has the data vector $\mathbf{u}_i'$

$$(0 \quad 0 \quad 0 \quad 999 \quad 999).$$

Muthén and Masyn (in press) also propose general discrete-time survival mixture analysis (DTSMA) models, in which different latent classes have different hazard and survival functions. For example, a growth mixture model for y can be combined with a survival model for u.

19.6. Combination of Categorical and Continuous Outcomes: Modeling With Zeros

In the previous section, it was seen that the u variables need not represent conventional categorical outcomes but can be used as indicators of events. In this section, this idea is taken further by using the u variables as indicators of zero values on a continuous and on a count outcome variable.

Growth mixture modeling is useful for describing growth in outcomes that can be seen as continuous but nonnormally distributed. A type of nonnormality that cannot be well captured by mixtures of normal distributions arises in studies in which a significant number of individuals are at the lowest value of an outcome, for example, representing absence of a behavior. Applications include alcohol, drug, and tobacco use among adolescents. Censored-normal models are often used for outcomes of this kind, including classic Tobit regression analysis (Amemiya, 1985; Tobin, 1958) and LCGA in the PROC TRAJ program (Jones et al., 2001). A recent article by Olsen and Schafer (2001) gives an excellent overview of several related modeling efforts. Censored-normal models have been criticized (see, e.g., Duan, Manning, Morris, & Newhouse, 1983) because of the limitation of assuming that the same set of covariates influences both the decision to engage in the behavior and the amount observed. A two-part modeling approach proposed in Olsen and Schafer avoids this limitation.

To simplify the discussion, the lowest value will be taken to be zero. It is useful to distinguish between two kinds of zero outcomes. First, individuals may have zero values at a given time point because their behavioral activity is low and is zero during certain periods ("random zeros"). Second, individuals may not engage in the activity at all and therefore have zeros throughout all time points of the study ("structural zeros"). Olsen and Schafer (2001) proposed a two-part model for the case of random zeros, whereas Carlin, Wolfe, Brown, and Gelman (2001) considered the case of structural zeros. In both articles, a random-effects logistic regression was used to express the probabilities of nonzeros versus zeros.

Olsen and Schafer (2001) studied alcohol use in Grades 7 through 11. To capture the changing zero status across time, they expressed the logistic regressions for each time point as a random-effects growth model. The term *two-part model* refers to having both a logistic model part to model the probability of nonzero versus zero outcomes (Part 1) and a continuous-normal or lognormal model part for the values of the nonzero outcomes (Part 2). In Olsen and Schafer, the two parts have correlated random effects. The two parts are also allowed to have different covariates, avoiding the limitation of censored-normal modeling.

Carlin et al. (2001) studied cigarette smoking among adolescents. A two-class model was used with a "zero class" (structural zeros) representing individuals not susceptible to regular smoking (also referred to as "immunes"). As pointed out in Carlin et al., an individual with zeros throughout the study does not necessarily belong to the zero class but may show zeros by chance. In their analysis, the estimated proportion of immunes was 69%, whereas the empirical proportion with all zeros was 77%. Because of this, an ad hoc analysis based on deleting individuals with all zeros may lead to distorted results.

Inspired by Olsen and Schafer (2001) and Carlin et al. (2001), Muthén (2001b) proposed a generalization of growth mixture modeling to handle both random and structural zeros in a two-part model. Multiple latent classes are used to represent the growth in the probability of nonzero values in Part 1 as well as

the growth in the nonzero outcomes in Part 2. For the Part 1 modeling of the probability of nonzero values, Muthén considered a latent class growth alternative to the random-effects modeling of Olsen and Schafer (2001) and Carlin et al. (2001)—that is, a model in line with Nagin (1999). The use of latent classes for the Part 1 modeling of the probability of nonzero values may be seen as a semi-parametric alternative to a random-effects model in line with Aitkin (1999). In addition to accounting for random zeros as in Olsen and Schafer, Muthén's Part 1 approach incorporates Carlin et al.'s concept of a zero class that has zero probability of nonzero values throughout the study. A further advantage of the proposed approach is that covariates are allowed to have a different influence in different classes. For the Part 2 modeling of the nonzero outcomes, the proposed modeling extends the Olsen-Schafer growth model to a growth mixture model. The Olsen-Schafer model, the mixture version of Olsen-Schafer, the Carlin et al. model, and the Muthén two-part growth mixture model can all be fitted into the general latent variable modeling framework of Mplus.

The question of the proper treatment of zeros also arises with count variables. Roeder et al. (1999) considered zero-inflated Poisson modeling (ZIP) (Lambert, 1992) in the context of LCGA. When a count outcome is modeled by ZIP, it is assumed that a zero value can be observed for two reasons. The ZIP model is a two-class mixture model, similar in spirit to that of Carlin et al. (2001). First, if an individual is in the zero class, a zero count has probability 1. Second, if an individual is in the nonzero class, the probability of a zero count is expressed by the Poisson distribution. The probability of being in the zero class can be modeled by covariates that are different from those that predict the counts for the nonzero class. In longitudinal data, this probability can be modeled to vary across time. The model by Roeder et al. considered an LCGA for the nonzero part.

19.7. Multilevel Growth Mixture Modeling

This final section returns to the analysis of the LSAY math achievement example. Longitudinal data are often collected through cluster sampling. This was the case in the LSAY study, in which students were observed within randomly sampled schools. This gives rise to three-level data with variation across time on Level 1, variation across individuals on Level 2, and variation across clusters on Level 3. This section discusses three-level growth modeling and its new extension to three-level growth mixture modeling. Due to lack of space, details of the modeling will not be discussed here, but an analysis of the LSAY example will instead be discussed in general terms. The reader is referred to Asparouhov and Muthén (2003b) for technical details.

The model diagram of Figure 19.11 is useful for understanding the general ideas of the multilevel growth mixture modeling. This is the LSAY math achievement example discussed in Section 19.3.5. In Figure 19.11, the observed math variable rectangles at the top of the figure represent the Level 1 variation across time. The latent variable circles, labeled *i* and *s*, represent the Level 2 variation in the intercept and slope growth factors across students. The *ib*, *cb*, *sb*, and *hb* latent variable circles represent the Level 3 variation across schools. Here, *b* refers to between-school variation. One aim of three-level growth modeling is the decomposition of the intercept variance into *i* and *ib* variation and the decomposition of the slope variance into *s* and *sb* variation. Furthermore, it is of interest to describe part of this variation by school-level covariates, as shown at the bottom of the diagram.

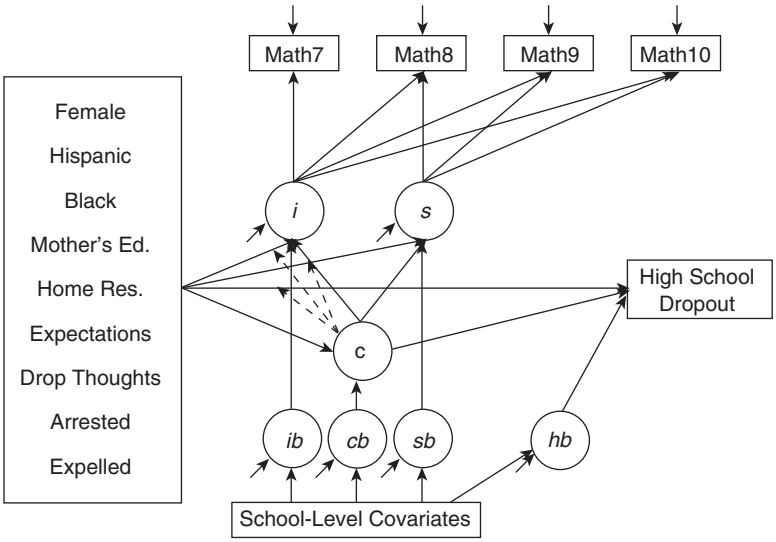
Figure 19.11 also includes a distal outcome of high school dropout and considers across-school variation in its intercept *hb* (there may also be across-school variation in some of the slopes). The intercept variation is again described by school-level covariates. This model part is analogous to two-level logistic regression (see, e.g., Hedeker & Gibbons, 1994). In Figure 19.11, a new feature is that the two-level logistic regression has as one of its predictors a latent categorical variable *c*, the latent trajectory class variable.

A key new feature in Figure 19.11 is the across-school variation *cb* in the individual-level latent class variable *c*. This part of the model makes it possible to study the influence of school-level variables on the class member probability for the students. This corresponds to multinomial logistic regression with random effects, except that the dependent variable is latent.

The model in Figure 19.11 was analyzed using maximum likelihood estimation in Mplus.[16] A key school-level variable used in the modeling was a school poverty index, measured as the percentage of the student body receiving full school lunch support. It was found that this school poverty index did not have a significant effect on the probability of dropping out

[16]The Mplus input and output for these analyses are given in Example 9 at www.statmodel.com/mplus/examples/penn.html. This analysis was carried out by Mplus Version 3.

Figure 19.11 Multilevel GGMM for LSAY Data



of high school. It did, however, have a significant influence on *c* in the sense that a high index value resulted in a higher probability of being a member of the class with a poor math achievement trajectory in Grades 7 through 10. The growth mixture analyses reported on earlier showed that membership in the failing class gave a very high risk of dropping out of high school. In this way, the multilevel growth mixture modeling implies that school poverty does not influence dropout directly but indirectly, in that it influences achievement trajectory class, which in turn influences dropout. This is an interesting new type of mediational process, whereby the mediator is not only categorical but also latent.

The general latent variable modeling framework considered here allows multilevel modeling, such as three-level growth modeling, not only for continuous outcomes but also for categorical outcomes. In this way, multilevel modeling is available in Mplus for GGMM, LCGA, LCA, LTA, and DTSMA.

19.8. Conclusions

This chapter has shown how modeling, using a combination of continuous and categorical latent variables, provides an extremely flexible analysis framework. Different traditions such as growth modeling, latent class analysis, and survival analysis are brought together using the unifying theme of latent variable modeling. New developments in these areas have been presented. Not only does this create more interesting analysis options in each area, but the combination of model parts that is possible leads to even further opportunities for investigating data. Several such combinations were not discussed but include the following (see also Muthén, 2001a, 2002; Muthén & Asparouhov, 2003a, 2003b):

- Multiple-process growth mixture modeling
  Parallel (dual) processes: studying relations between concurrent outcomes
  Sequential processes: predicting later growth from earlier growth
- Multiple-group growth mixture modeling: studying similarities and differences across known groups
- Multiple indicator growth mixture modeling: studying growth in a latent variable construct
- Embedded growth mixture modeling: combining the growth model with LCA, factor analysis, path analysis, and SEM components
- Combined growth mixture and discrete-time survival modeling: predicting survival from trajectory classes and vice versa

Mplus covers these models for outcomes that are continuous, binary, ordered polytomous, two-part, zero-inflated Poisson, or combinations thereof, allowing both missing data and cluster data.

REFERENCES

Agresti, A. (1990). *Categorical data analysis.* New York: John Wiley.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics, 55,* 117–128.

Amemiya, T. (1985). *Advanced econometrics.* Cambridge, MA: Harvard University Press.

Asparouhov, T., & Muthén, B. (2003a). *Full-information maximum-likelihood estimation of general two-level latent variable models.* Manuscript in preparation.

Asparouhov, T., & Muthén, B. (2003b). *Maximum-likelihood estimation in general latent variable modeling.* Manuscript in preparation.

Carlin, J. B., Wolfe R., Brown, C. H., & Gelman, A. (2001). A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics, 2,* 397–416.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Arnold.

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

Duan, N., Manning, W. G., Morris, C. N., & Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics, 1,* 115–126.

Farrington, D. P., & West, D. J. (1990). The Cambridge study in delinquent development: A prospective longitudinal study of 411 males. In H.-J. Kernere & G. Kaiser (Eds.), *Criminality: Personality, behavior, and life history.* New York: Springer-Verlag.

Hedeker, D. (2000). *A fully semi-parametric mixed-effects regression model for categorical outcomes.* Paper presented at the Joint Statistical Meetings, Indianapolis, IN.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics, 50,* 933–944.

Heinen, T. (1996). Latent class and discrete latent trait models: Similarities and differences. Thousand Oaks, CA: Sage.

Jeffries, N. O. (2003). A note on "Testing the number of components in a normal mixture." *Biometrika, 90,* 991–994.

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.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34,* 1–13.

Land, K. C. (2001). Introduction to the special issue on finite mixture models. *Sociological Methods & Research, 29,* 275–281.

Lin, H., Turnbull, B. W., McCulloch, C. E., & Slate, E. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association, 97,* 53–65.

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

Lubke, G., & Muthén, B. (2003). *Performance of factor mixture models.* Manuscript submitted for publication.

Masyn, K. (2002, June). *Latent class enumeration revisited: Application of Lo, Mendell, and Rubin to growth mixture models.* Paper presented at the meeting of the Society for Prevention Research, Seattle, WA.

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

Miller, J. D., Kimmel, L., Hoffer, T. B., & Nelson, C. (2000). *Longitudinal study of American youth: User's manual.* Evanston, IL: Northwestern University, International Center of the Advancement of Scientific Literacy.

Moffitt, T. E. (1993). Adolescence-limited and life-course persistent antisocial behavior. *Psychological Review, 100,* 674–701.

Molenaar, P. C., & von Eye, A. (1994). On the arbitrary nature of latent variables. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis* (pp. 226–242). Thousand Oaks, CA: Sage.

Muthén, B. (2001a). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Mahwah, NJ: Lawrence Erlbaum.

Muthén, B. (2001b). *Two-part growth mixture modeling.* Draft.

Muthén, B. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81–117.

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

Muthén, B., & Asparouhov, T. (2002). *Mixture testing using multivariate skewness and kurtosis.* Manuscript in preparation.

Muthén, B., & Asparouhov, T. (2003a). *Advances in latent variable modeling, part I: Integrating multilevel and structural equation modeling using Mplus.* Manuscript in preparation.

Muthén, B., & Asparouhov, T. (2003b). *Advances in latent variable modeling, part II: Integrating continuous and categorical latent variable modeling using Mplus.* Manuscript in preparation.

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C., et al. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics, 3,* 459–475.

Muthén, B., & Curran, P. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2,* 371–402.

Muthén, B., Jo, B., & Brown, H. (2003). Comment on the Barnard, Frangakis, Hill & Rubin article, Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association, 98,* 311–314.

Muthén, B., Kreuter, F., & Asparouhov, T. (2003). *Applications of growth mixture modeling to non-normal outcomes.* Manuscript in preparation.

Muthén, B., & Masyn, K. (in press). Mixture discrete-time survival analysis. *Journal of Educational and Behavioral Statistics.*

Muthén, B., & Muthén, L. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling

with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24,* 882–891.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55,* 463–469.

Muthén, L., & Muthén, B. (1998–2003). *Mplus user's guide.* Los Angeles: Author.

Muthén, L. K., & Muthén, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4,* 599–620.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods, 4,* 139–157.

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.

Nagin, D. S., & Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods, 6,* 18–34.

Olsen, M. K, & Schafer, J., L. (2001). A two-part random effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association, 96,* 730–745.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Roeder, K., Lynch, K. G., & Nagin, D. S. (1999). Modeling uncertainty in latent class membership: A case study in criminology. *Journal of the American Statistical Association, 94,* 766–776.

Rumberger, R. W., & Larson, K. A. (1998). Student mobility and the increased risk of high school dropout. *American Journal of Education, 107,* 1–35.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics, 18,* 155–195.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica, 26,* 24–36.

Wang, C. P., Brown, C. H., & Bandeen-Roche, K. (2002). *Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior.* Manuscript submitted for publication.