CHAPTER 6

Growth mixture modeling: Analysis with non-Gaussian random effects

Bengt Muthén and Tihomir Asparouhov

Contents

6.1 Introduction

This chapter gives an overview of non-Gaussian random-effects modeling in the context of finite-mixture growth modeling developed in Muthén and Shedden (1999), Muthén (2001a, 2001b, 2004), and Muthén et al. (2002), and extended to cluster samples and cluster-level mixtures in Asparouhov and Muthén (2008). Growth mixture modeling represents

unobserved heterogeneity between the subjects in their development using both random effects (e.g., Laird and Ware, 1982) and finite mixtures (e.g., McLachlan and Peel, 2000). This allows different sets of parameter values for mixture components corresponding to different unobserved subgroups of individuals, capturing latent trajectory classes with different growth curve shapes. This chapter discusses examples motivating modeling with such trajectory classes. A general latent-variable modeling framework is presented together with its maximum likelihood estimation. Examples from criminology, mental health, and education are analyzed. The choice of a normal or a non-parametric distribution for the random effects is discussed and investigated using a simulation study. The discussion will refer to growth mixture modeling techniques as implemented in the Mplus program (Muthén and Muthén, 1998–2006) and input scripts for the analyses are available at `http://www.statmodel.com`.

The outline of this chapter is as follows. Section 6.2 presents examples with substantive questions that motivate growth mixture analysis. Section 6.3 describes the general model. Section 6.4 discusses estimation and model assessment. Section 6.5 illustrates the modeling with a series of examples. Section 6.6 compares the parametric and non-parametric versions of the random-effect model. Section 6.7 concludes.

6.2 Examples

The following examples show the breadth of longitudinal studies that may be approached by growth mixture modeling.

6.2.1 Example 1: Clinical trials with placebo response

The first example concerns analysis of data from a double-blind 8-week randomized trial on depression medication (Leuchter et al., 2002). Of particular interest is how to assess medication effects in the presence of placebo response. Placebo response is an improvement in depression ratings that is unrelated to medication. The improvement is often seen as an early steep drop in depression, often followed by a later upswing. Figure 6.1 shows results for a two-class growth mixture model for the sample of 45 placebo group subjects using the Hamilton depression scale (Ham-D). The first two time points are before randomization and the next nine time points are after randomization. The responder class is shown in the left panel and the non-responder class in the right panel. The solid curve is the estimated mean curve, whereas the broken curves are observed individual trajectories for individuals classified as most likely belonging to this class.

Placebo response confounds the estimation of the true effect of medication and is an important phenomenon, given its high prevalence of 25–60%. Because placebo response is pervasive, the statistical modeling should account for this when estimating medication effects. This can be done by acknowledging the qualitative heterogeneity in trajectory shapes for responders and non-responders using growth mixture modeling. The estimation of medication effects using growth mixture modeling is described in Muthén et al. (2007). The medication effect is estimated in line with the approach of the next example.

6.2.2 Example 2: Randomized interventions with treatment effects varying across latent classes

The second example concerns a randomized preventive field trial conducted in Baltimore public schools (Dolan et al., 1993; Ialongo et al., 1999). The study applied a universal intervention aimed at reducing aggressive-disruptive behavior during first and second grade to improve reading and reduce aggression with outcomes assessed through middle school and beyond (Kellam et al., 1994). Children were followed from first to seventh grade with
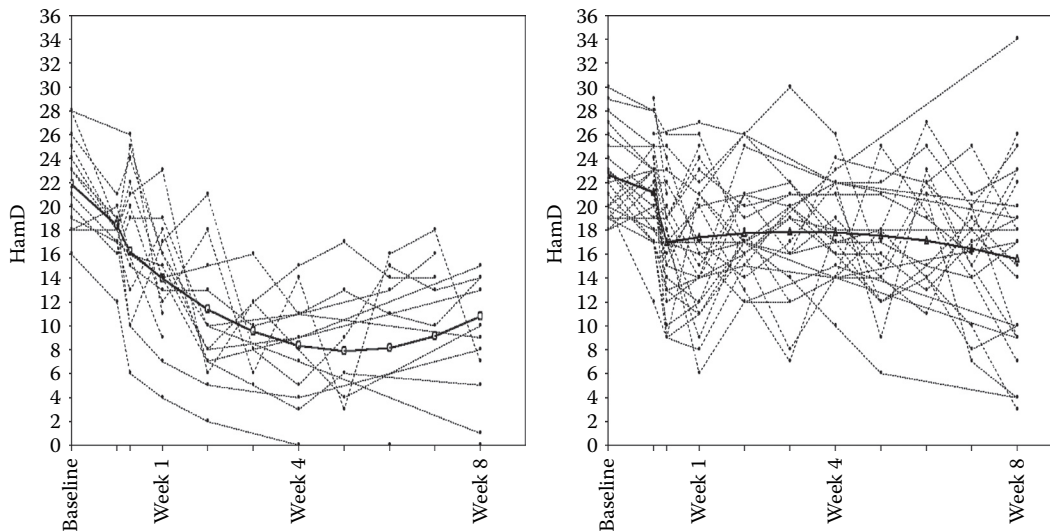
Figure 6.1 Two-class growth mixture model for depression in a placebo group.

respect to the course of aggressive behavior, and a follow-up to age 18 also allowed for the assessment of intervention impact on more distal events, such as the probability of juvenile delinquency as indicated by juvenile court records. The intervention was administered after one pre-intervention time point in fall of first grade.

Key scientific questions addressed whether the intervention reduced the slope of the aggression trajectory across the grades, whether the intervention was different in impact for children who initially display higher levels of aggression, and whether the intervention impacted distal outcomes. Analyses of these hypotheses were presented in Muthén et al. (2002). Allowing for multiple trajectory classes in the growth model gave a flexible way to assess differential effects of the intervention. The analyses focused on boys and intervention status as defined by classroom assignment in fall of first grade, resulting in a sample of 119 boys in the intervention group and 80 boys in the control group. Figure 6.2 shows results from a four-class growth mixture model for the 119 boys. For each combination of latent-trajectory class and intervention condition, the estimated mean growth curve is shown together with observed individual trajectories for individuals estimated to be most likely a member of the class. An intervention effect in terms of reducing aggressive behavior is seen for the high class and perhaps also for the low starting ("LS") class, whereas the other two classes show no effects.

6.2.3 Example 3: High school dropout predicted by failing math achievement development

The third example concerns growth mixture modeling of mathematics achievement development in US schools. Muthén (2004) analyzed longitudinal math scores for students in grades 7–10 from the Longitudinal Study of American Youth (LSAY) and found a problematic trajectory class with an exceptionally low starting point in grade 7 as well as a low growth rate; see Figure 6.3. The class membership was strongly related to covariates such as grade 7 measures of having low schooling expectations and dropout thoughts. Taken together with the poor math development, this suggests that the class consists of students who are disengaged from school. Class membership was also highly predictive of dropping out by grade 12, a binary "distal" outcome. In a further analysis, Muthén (2004) carried out a growth mixture analysis where the clustering of students within schools was taken
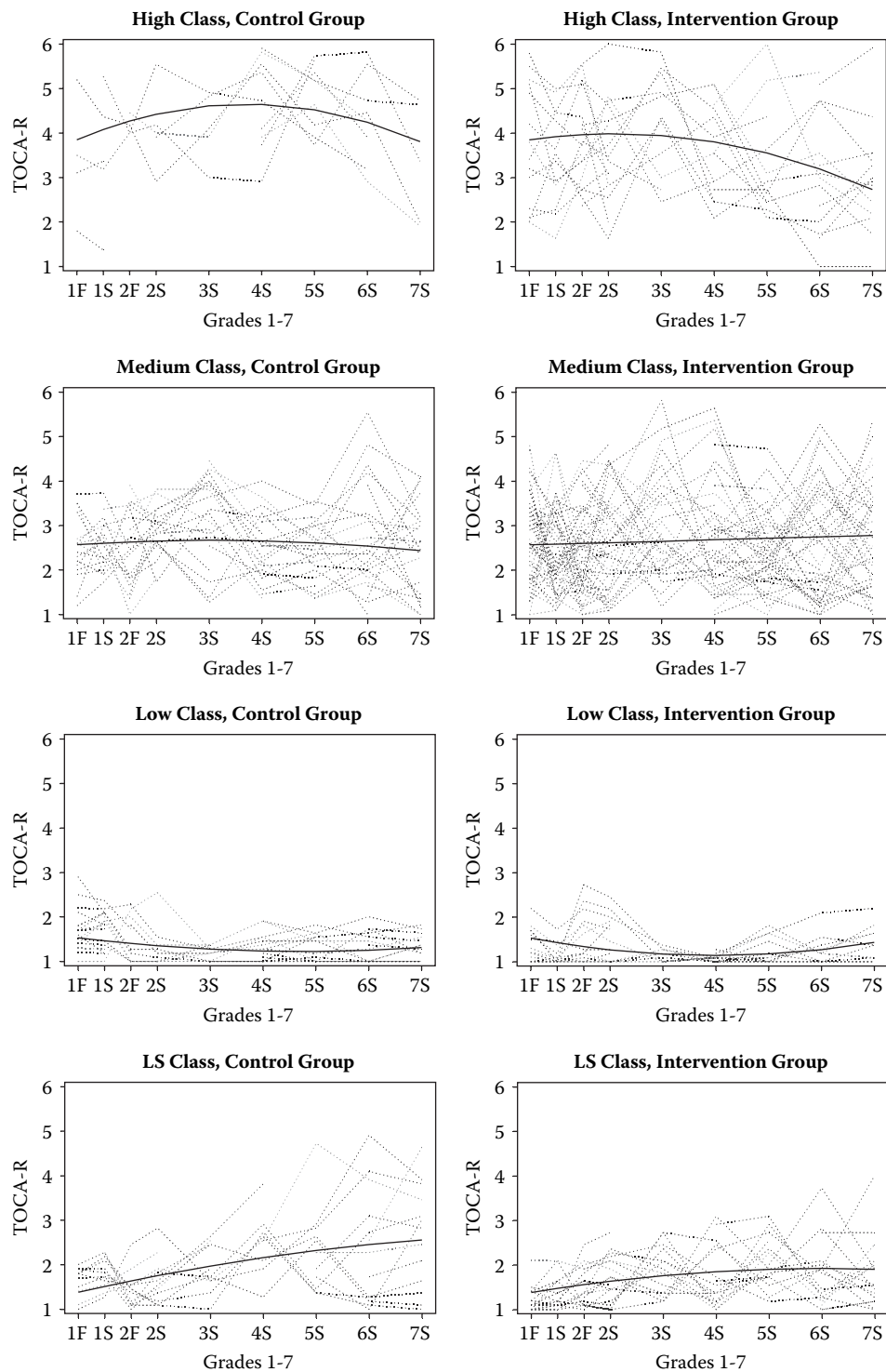
Figure 6.2 Four-class growth mixture model for aggressive behavior in control and intervention groups.
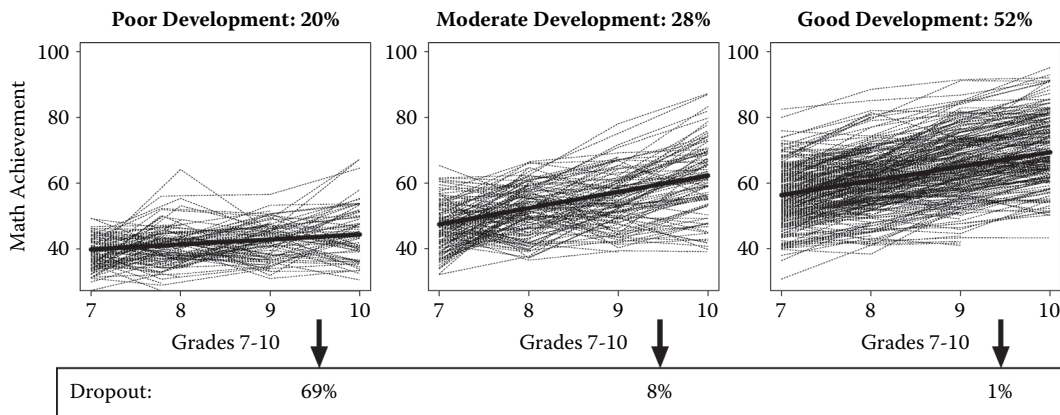
Figure 6.3 Three-class growth mixture model for math achievement related to high school dropout.

into account by allowing random-effect variation across schools. The school variation was represented in the random effects for the growth, the random intercept in the logistic regression for dropping out, and the random intercept in the multinomial regression predicting latent-class membership as a function of student-level covariates. Furthermore, school-level covariates corresponding to poverty of the school neighborhood and teaching quality in the school were used to predict across-school variation in the random coefficients.

6.2.4 Example 4: Age–crime curves

The fourth example concerns criminal activity of 13,160 males born in Philadelphia, Pennsylvania in 1958 (D'Unger et al., 1998; D'Unger, Land, and McCall, 2002; Loughran and Nagin, 2006). Annual counts of police contacts are available from age 4 to 26 of this birth cohort. The aggregate age–crime curve follows the well-known pattern of increasing annual convictions throughout the subjects' teenage years and decreasing annual convictions thereafter. The criminology literature has focused extensively on identifying groups of individuals with similar patterns or careers of delinquent and criminal offending. To quote D'Unger et al. (1998, p. 1595): "This question of how many latent classes of criminal careers are optimal, and why the number of categories itself is important, has gained salience for criminological theory in light of recent theoretical debates". The authors go on to mention Moffit (1993) as a key contributor to the notion of different trajectory classes, proposing a distinction between the trajectory of "life-course persistents" versus "adolescence limiteds" depending on the behaviors persisting over the life course or seen only during adolescence. The debate continues, as seen in Sampson and Laub (2005) discussing the "group-based" analysis approach of Nagin (1999, 2005), Nagin and Land (1993), and Roeder, Lynch, and Nagin (1999). The Philadelphia crime data will be analyzed in a new way in this chapter. The analyses to be presented have two special features. First, the outcome variable is a count variable that is very skewed, with a large number of zeros at each point in time. Second, it is of interest to contrast the "group-based" approach with random-effects models.

6.2.5 Example 5: Classification of schools based on achievement development

The fifth example extends the achievement analyses discussed in Example 3 by using a school-level latent-class variable, enabling a classification of schools as more or less successful. The LSAY data discussed in Example 3 are from a limited number of schools and

analyses are instead performed on data from grades 8, 10, and 12 of the National Education Longitudinal Study (NELS). NELS surveyed 913 schools and a total of 14,217 students. In the analyses to be presented, student growth rate is regressed on the growth intercept in grade 8 and allows this relationship to vary across the school-level latent classes. It has been argued in the education literature that a weak relationship is an indicator of a school being egalitarian (e.g., Choi and Seltzer, 2006). The means of the random intercept and the intercept of the random growth rate are also allowed to vary across the school-level latent classes. Both types of school-level latent-class features are useful for determining school quality.

6.2.6 Other applications

Other applications of growth mixture modeling found in the literature include Verbeke and Lesaffre (1996), see also Pearson et al. (1994), who considered different groups of males with linear or exponential growth in prostate-specific antigen (PSA); Muthén and Shedden (1999) and Muthén and Muthén (2000), with application to the development of heavy drinking and alcohol dependence; Lin et al. (2002), with application to PSA and prostate cancer, combining growth mixture modeling with survival analysis; Croudace et al. (2003), with application to bladder control; Muthén et al. (2003), with application to reading failure, including the modeling of a kindergarten process for phonemic awareness linked to a later process of word recognition; and Muthén and Masyn (2005), with application to aggressive behavior and juvenile delinquency, combining growth mixture modeling and discrete time survival analysis. Related applications to latent-class membership representing non-participation (non-compliance) and complier-average causal effect estimation in intervention studies (Angrist, Imbens, and Rubin, 1996) are given in Jo (2002), Jo and Muthén (2003), and are also generalizable to longitudinal studies (see Yau and Little, 2001; Dunn et al., 2003; Muthén, Jo, and Brown, 2003), including time-varying compliance (Lin, Ten Have, and Elliott, 2006).

6.3 Growth mixture modeling

This section describes the general growth mixture modeling framework (see also Asparouhov and Muthén, 2008). The description is closely related to the implementation in the Mplus software version 4.2 and higher (Muthén and Muthén, 1998–2006). To familiarize readers with the general Mplus modeling framework, the section starts with a simple growth example put into the conventional linear mixed-effects model as well as the Mplus modeling framework.

6.3.1 Specification of a simple growth model

Consider a single growth process with no latent-trajectory classes, no clustering, linear growth for a continuous outcome y, a time-invariant covariate x and a time-varying covariate w,

$$Y_{ij} = \eta_{0i} + \eta_{1i}a_{ij} + \kappa_i w_{ij} + \epsilon_{ij}, \tag{6.1}$$

where a_{ij} are time scores ($j = 1, 2, \ldots, T$), the random intercept η_{0i} and the random slope η_{1i} represent the growth process, κ_i is a random slope, and ϵ is a normally distributed residual. The random intercepts and slopes are expressed as

$$\eta_{0i} = \alpha_0 + \gamma_0 x_i + \zeta_{0i}, \tag{6.2}$$

$$\eta_{1i} = \alpha_1 + \gamma_1 x_i + \zeta_{1i}, \tag{6.3}$$

$$\kappa_i = \alpha_2 + \gamma_2 w_{ij} + \zeta_{2i}, \tag{6.4}$$

where the αs and γs are parameters and the ζs are normally distributed residuals. In multilevel terms, equation (6.1) represents level-1 variation across time and (6.2)–(6.4) represent level-2 variation across individuals. Consider the mixed linear model formulation for $Y_i = (Y_{i1}, \ldots, Y_{iT})'$,

$$\boldsymbol{Y}_i = X_i \boldsymbol{\beta} + Z_i \boldsymbol{b}_i + \boldsymbol{e}_i, \tag{6.5}$$

where some individuals may not be observed at all occasions T, leading to missing data. In this example, let

$$\Lambda_i = \begin{pmatrix} 1 & a_{i1} \\ 1 & a_{i2} \\ 1 & a_{i3} \\ \vdots & \vdots \\ 1 & a_{iT} \end{pmatrix},$$

so that in (6.5) we have

$$X_i = \left(\Lambda_i \, \boldsymbol{w}_i \, \Lambda_i x_i \, \boldsymbol{w}_i x_i \right),$$
$$\beta = (\alpha_0, \alpha_1, \ \alpha_2, \ \gamma_0, \gamma_1, \ \gamma_2)',$$
$$Z_i = \left(\Lambda_i \, \boldsymbol{w}_i \right),$$
$$\boldsymbol{b}_i = (\zeta_{0i}, \zeta_{1i}, \ \zeta_{2i})',$$
$$\boldsymbol{e}_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})'.$$

The Mplus framework uses the general model expression for observed vectors $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$,

$$\boldsymbol{Y}_i = \boldsymbol{\nu} + \Lambda \boldsymbol{\eta}_i + K \boldsymbol{X}_i + \boldsymbol{\epsilon}_i, \tag{6.6}$$

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + B \boldsymbol{\eta}_i + \Gamma \boldsymbol{X}_i + \boldsymbol{\zeta}_i, \tag{6.7}$$

implying

$$\boldsymbol{Y}_i = \boldsymbol{\nu} + \Lambda \left(I - B \right)^{-1} \boldsymbol{\alpha} + \Lambda \left(I - B \right)^{-1} \Gamma \, \boldsymbol{X}_i + K \, \boldsymbol{X}_i + \Lambda \left(I - B \right)^{-1} \boldsymbol{\zeta}_i + \boldsymbol{\epsilon}_i,$$

where the first row refers to fixed effects and the second row to random effects. The regression parameter arrays Λ, K, B, and Γ are allowed to vary across i as a function of observed variables or they can be unobserved random slopes. The model equations (6.6) and (6.7) capture the level-1 and level-2 expressions for the linear growth example in (6.1) and (6.2)–(6.4). The notation of (6.6) and (6.7) follows that of the linear growth example with $B = 0$ and with the vector $\boldsymbol{X}_i$ containing both the time-varying covariate w_{ij} and the time-invariant covariate x_i. The model of (6.6) and (6.7) includes the mixed linear model of (6.5) as a special case. In latent-variable modeling terms, (6.6) is referred to as the measurement part of the model, where the latent-variable vector $\boldsymbol{\eta}_i$ is measured by the indicators $\boldsymbol{Y}_i$. Here, Λ may contain parameters. A frequent example is when $a_{it} = a_t$, so that a_t can be treated as parameters, for example capturing deviations from linear growth shape (fixing two a_t values for identification, typically $a_1 = 0, a_2 = 1$). Another example is where multiple indicators of a factor are available at each time point, where different indicators have different factor loadings λ. With $\Lambda_i = \Lambda$, (6.6) also covers factor analysis with covariates. Furthermore, (6.7) is referred to as the structural part, containing regressions among the latent variables. The regression matrix B has zero diagonal elements, but the off-diagonal elements may be used to regress random effects on each other. For example, the growth slope (growth rate) η_{1i}, or the random slope κ_i may be expressed as a function of the intercept (initial status) η_{0i}. More generally, (6.7) also covers structural equation modeling. In this way, the extensions of (6.6) and (6.7) to finite mixtures and cluster samples presented in this chapter pertain to not only growth models but also factor analysis and structural equation models, as well as combinations of such models and growth models (Muthén, 2002).

6.3.2 A general multilevel mixture model

Let Y_{kij} be the jth observed dependent variable for individual i in cluster k. Three types of variables are considered in the analyses to be presented: binary and ordered categorical variables, continuous normally distributed variables, and counts following the Poisson or zero-inflated Poisson distribution. Let C_{ki} be a latent categorical variable for individual i in cluster k which takes values $1, \ldots, L$. Let D_k be a cluster-level latent categorical variable for cluster k which takes values $1, \ldots, M$. The choice of L and M will be discussed in Section 6.4.1. To construct a model for observed binary and ordered categorical variables we proceed in line with Muthén (1984) by defining an underlying continuous, normally distributed latent variable Y_{kij}^* such that, for a set of threshold parameters τ_{cdsj},

$$Y_{kij} = s|_{C_{ki}=c, D_k=d} \Leftrightarrow \tau_{cdsj} < Y_{kij}^* < \tau_{cd,s+1,j}.$$

For continuous normally distributed variables we define $Y_{kij}^* = Y_{kij}$. For counts $Y_{kij}^* = \log(\lambda_{kij})$, where λ_{kij} is the rate of the Poisson distribution. Let $\boldsymbol{Y}_{ki}^*$ be the J-dimensional vector of all dependent variables and $\boldsymbol{X}_{ki}$ be the Q-dimensional vector of all individual-level covariates. Using latent-variable terms, the "measurement" part of the model is defined by

$$\boldsymbol{Y}_{ki}^*|_{C_{ki}=c, D_k=d} = \boldsymbol{\nu}_{cdk} + \Lambda_{cdk}\,\boldsymbol{\eta}_{ki} + K_{cdk}\,\boldsymbol{X}_{ki} + \boldsymbol{\epsilon}_{ki}, \tag{6.8}$$

where $\boldsymbol{\nu}_{cdk}$ is a J-dimensional vector of intercepts, Λ_{cdk} is a $J \times m$ slope matrix for the m-dimensional random-effect vector $\boldsymbol{\eta}_{ki}$, K_{cdk} is a $J \times Q$ slope matrix for the covariates, and $\boldsymbol{\epsilon}_{ki}$ is a J-dimensional vector of residuals with mean zero and covariance matrix Θ_{cd}. For a categorical variable Y_{kij} a normality assumption for ϵ_{kij} is thus equivalent to a probit regression for Y_{kij} on η_{kij} and X_{kij}. Alternatively, ϵ_{kij} can have a logistic distribution, resulting in a logistic regression. For a count variable Y_{kij} the residual ϵ_{kij} is assumed to be zero. For normally distributed continuous variables Y_{kij} the residual variable ϵ_{kij} is assumed normally distributed.

The "structural" part of the model is defined by

$$\boldsymbol{\eta}_{ki}|_{C_{ki}=c, D_k=d} = \boldsymbol{\alpha}_{cdk} + B_{cdk}\boldsymbol{\eta}_{ki} + \Gamma_{cdk}\boldsymbol{X}_{ki} + \boldsymbol{\zeta}_{ki}, \tag{6.9}$$

where $\boldsymbol{\alpha}_{cdk}$ is an m-dimensional vector of intercepts, B_{cdk} is an $m \times m$ structural regression parameter matrix, Γ_{cdk} is a $m \times Q$ slope parameter matrix, and $\boldsymbol{\zeta}_{ki}$ is an m-dimensional vector of normally distributed residuals with covariance matrix Ψ_{cd}. The model for the latent categorical variable C_{ki} is a multinomial logit model

$$\Pr(C_{ki} = c|D_k = d) = \frac{\exp(a_{cdk} + \boldsymbol{b}_{cdk}'\boldsymbol{X}_{ki})}{\sum_s \exp(a_{sdk} + \boldsymbol{b}_{sdk}'\boldsymbol{X}_{ki})}. \tag{6.10}$$

Some parameters have to be restricted for identification purposes. For example, the variance of ϵ_{kij} should be 1 for categorical variables Y_{kij} under probit and $\pi^2/\sqrt{3}$ under logit. Also $a_{Ldk} = b_{Ldk} = 0$.

The multilevel part of the model is introduced as follows. Each of the intercepts, slopes or loading parameters in equations (6.8)–(6.10) can be either a fixed coefficient or a random effect that varies across clusters k. Let $\boldsymbol{\eta}_k$ be the vector of all such random effects and let X_k be the vector of all cluster-level covariates. The between-level model for $\boldsymbol{\eta}_k$ is then

$$\boldsymbol{\eta}_k|_{D_k=d} = \boldsymbol{\mu}_d + B_d\boldsymbol{\eta}_k + \Gamma_d\boldsymbol{X}_k + \boldsymbol{\zeta}_k, \tag{6.11}$$

where $\boldsymbol{\mu}_d$, B_d and Γ_d are fixed parameters and $\boldsymbol{\zeta}_k$ is a normally distributed residual with covariance Ψ_d. The model for the between level categorical variable D is also a multinomial logit regression

$$\Pr(D_k = d) = \frac{\exp(a_d + \boldsymbol{b}_d'\boldsymbol{X}_k)}{\sum_s \exp(a_s + \boldsymbol{b}_s'\boldsymbol{X}_k)}. \tag{6.12}$$

Equations (6.8)–(6.12) comprise the definition of a multilevel latent-variable mixture model. There are many extensions of this model that are possible in the Mplus framework. For example, observed dependent variables can be incorporated on the between level. Other extensions arise from the fact that a regression equation can be constructed between any two variables in the model. Such equations can be fixed- or random-effect regressions. The model can also accommodate multiple latent-class variables on the within and the between level. Other types of dependent variables can also be incorporated in this model such as censored, nominal, semi-continuous, and time-to-event survival variables; see Olsen and Schafer (2001) and Asparouhov, Masyn, and Muthén (2006).

6.4 Estimation

The above model is estimated by the maximum likelihood estimator using the EM algorithm where the latent variables C_{ki}, $\boldsymbol{\eta}_{ki}$, D_k and $\boldsymbol{\eta}_k$ are treated as missing data. The observed-data likelihood is given by

$$\prod_k \sum_d \Pr(D_k = d) \int \psi_k(\boldsymbol{\eta}_k) \prod_i \Big(\sum_c \Pr(C_{ki} = c) \int f_{ki}(\boldsymbol{Y}_{ki}) \psi_{ki}(\boldsymbol{\eta}_{ki}) d\boldsymbol{\eta}_{ki}\Big) d\boldsymbol{\eta}_k, \quad (6.13)$$

where f_{ki}, ψ_{ki} and ψ_k are the likelihood functions for $\boldsymbol{Y}_{ki}$, $\boldsymbol{\eta}_{ki}$ and $\boldsymbol{\eta}_k$, respectively. Numerical integration is utilized in the evaluation of the above likelihood using both adaptive and non-adaptive quadrature (see Schilling and Bock, 2005). The method can be described as follows. Suppose that η is a continuously distributed random-effect variable with density function ψ. Then

$$\int f(\eta)\psi(\eta)d\eta \approx \sum_{q=1}^{Q} w_q f(n_q), \quad (6.14)$$

where n_q are the nodes of the numerical integration and w_q are the weights. The weights are computed as $w_q = \psi(n_q)/\sum_{i=1}^{Q} \psi(n_i)$. The numerical integration method approximates the continuous distribution for η with a categorical distribution, that is, we can assume that the variable η takes the values n_q with probabilities w_q. Using this method the likelihood (6.13) is approximated by

$$\prod_k \sum_d \Pr(D_k = d) \sum_q \Pr(\eta_k = n_{qk}) \prod_i \Big(\sum_c \Pr(C_{ki} = c) \sum_r \Pr(\eta_{ki} = n_{rki}) f_{ki}(\boldsymbol{Y}_{ki})\Big)$$

$$= \prod_k \sum_{d,q} \Pr(D_k = d, \eta_k = n_{qk}) \prod_i \Big(\sum_{c,r} \Pr(C_{ki} = c, \eta_{ki} = n_{rki}) f_{ki}(\boldsymbol{Y}_{ki})\Big), \quad (6.15)$$

where n_{qk} and n_{rki} are the nodes of the numerical integration.

The EM algorithm is as follows. First compute the posterior distribution for the latent variables. The posterior joint distribution for D_k and $\boldsymbol{\eta}_k$ is computed as follows:

$$p_{dqk} = \Pr(D_k = d, \boldsymbol{\eta}_k = n_{qk}|*)$$

$$= \frac{\Pr(D_k = d, \boldsymbol{\eta}_k = n_{qk}) \prod_i \Big(\sum_{c,r} \Pr(C_{ki} = c, \boldsymbol{\eta}_{ki} = n_{rki}) f_{ki}(\boldsymbol{Y}_{ki})\Big)}{\sum_{d,q} \Pr(D_k = d, \boldsymbol{\eta}_k = n_{qk}) \prod_i \Big(\sum_{c,r} \Pr(C_{ki} = c, \boldsymbol{\eta}_{ki} = n_{rki}) f_{ki}(\boldsymbol{Y}_{ki})\Big)}.$$

The posterior conditional joint distribution for C_{ki} and $\boldsymbol{\eta}_{ki}$ is computed as follows:

$$p_{crki|dq} = \Pr(C_{ki} = c, \boldsymbol{\eta}_{ki} = n_{rki}|*, D_k = d, \boldsymbol{\eta}_k = n_{qk})$$

$$= \frac{\Pr(C_{ki} = c, \boldsymbol{\eta}_{ki} = n_{rki}) f_{ki}(Y_{ki})}{\sum_{c,r} \Pr(C_{ki} = c, \boldsymbol{\eta}_{ki} = n_{rki}) f_{ki}(Y_{ki})}.$$

The expected complete-data log-likelihood is now given by

$$\sum_{dqk} p_{dqk} \, \log(\Pr(D_k = d, \boldsymbol{\eta}_k = n_{qk})) + \sum_{dcqrki} p_{dqk} \, p_{crki|dq} \, \log(\Pr(C_{ki} = c, \boldsymbol{\eta}_{ki} = n_{rki}))$$

$$+ p_{dqk} \, p_{crki|dq} \sum_{dcqrki} \log(f_{ki}(Y_{ki})),$$

which is maximized with respect to the model parameters.

An alternative algorithm for obtaining the maximum likelihood estimates can be constructed by directly optimizing (6.15) with a standard maximization algorithms such as the Fisher scoring and the quasi-Newton algorithms. Such alternative algorithms can be used in combination with the EM algorithm to achieve faster convergence, an approach known as the accelerated EM algorithm (AEM). The AEM algorithm is implemented in Mplus.

A number of different integration methods can be used in (6.14). Mplus implements three different integration methods: rectangular, Gauss–Hermite and Monte Carlo integration. In addition, adaptive integration can be used. With this method, the integration nodes are concentrated in the area where the posterior distribution of the random effects is non-zero.

The estimation implemented in Mplus allows missing at random data for all dependent variables (Little and Rubin, 2002). Non-ignorable missing data is discussed in Muthén et al. (2003). It should be noted that mixture models in general are prone to have multiple local maxima of the likelihood and the use of many different sets of starting values in the interactive maximization procedure is strongly recommended. An automatic random starts procedure is implemented in the Mplus program, where starting values given by the user or produced automatically by the program are randomly perturbed.

6.4.1 Model assessment

For comparison of fit of models that have the same number of classes and are nested, the usual likelihood ratio chi-square difference test can be used, as long as the requirement is fulfilled of not having parameters on the border of the admissible parameter space in the more restricted model. Comparison of models with different numbers of classes violates this requirement with zero probability parameters. Deciding on the number of classes is instead typically accomplished by a Bayesian information criterion (BIC: Schwartz, 1978; Kass and Raftery, 1993),

$$BIC = -2 \log L + r \log n,$$

where r is the number of free parameters in the model and n is the sample size. The lower the BIC value, the better the model. The number of classes is increased until a BIC minimum is found. Although not chi-square distributed, the usual likelihood ratio statistic for comparing models with different number of classes can still be used, assessing the distribution of the statistic by bootstrap techniques. McLachlan and Peel (2000, Chapter 6) discuss a parametric bootstrapped likelihood ratio approach proposed by Aitkin, Anderson, and Hinde (1981). Although computationally intensive, it has been found to perform well in simulation studies using latent-class and growth mixture models, outperforming BIC in some instances (Nylund, Asparouhov, and Muthén, 2007).

The fit of the model to data for continuous variables can be studied by comparing for each class estimated moments with moments created by weighting the individual data by the estimated conditional probabilities (Roeder, Lynch and Nagin, 1999). To check how closely the estimated average curve within each class matches the data, it is also useful to randomly assign individuals to classes based on individual estimated conditional class

probabilities. Plots of the observed individual trajectories together with the model-estimated average trajectory can be used to check assumptions using class membership determined by "pseudo-class" draws (Bandeen-Roche et al., 1997). Wang, Brown, and Bandeen-Roche (2005) present methods for residual checking based on these ideas. With categorical and count outcomes, model fit may be investigated with respect to univariate and bivariate frequency tables, as well as frequencies for response patterns that do not have too small expected counts. Finally, it is important to note that the need for latent classes may be due to non-normality of the outcomes rather than substantively meaningful subgroups (see McLachlan and Peel, 2000, pp. 14–17; Bauer and Curran, 2003). To support a substantive interpretation of the latent classes, the researcher should consider not only the outcome variable in question, but also antecedents (covariates predicting latent-class membership), concurrent outcomes, and distal outcomes (predictive validity); see also related arguments in Muthén (2004).

6.5 Examples

This section presents analyses of the crime data of Example 4, the aggressive behavior data of Example 2, and the math achievement data of Example 5. The Example 4 analysis uses a growth mixture model for crime counts. Examples 2 and 5 consider multilevel growth mixture modeling of cluster data. Example 2 examines intervention effects that vary across both student-level and classroom-level latent-class variables. Example 5 considers students within school where student growth characteristics vary across a school-level latent-class variable.

6.5.1 Analysis of Example 4: Age–crime curves

The analysis of the Philadelphia data with counts of criminal activity for 13,160 males aged 4–26 will compare two different approaches, a "group-based" approach and growth mixture analysis (for more extended comparisons, see Kreuter and Muthén, 2007, 2008). The group-based analysis is associated with the work of of Nagin and Land (1993), Nagin (1999, 2005), Roeder, Lynch, and Nagin (1999), and Jones, Nagin, and Roeder (2001). This approach is commonly seen in the criminology literature and was used by D'Unger et al. (1998), D'Unger, Land, and McCall (2002), and Loughran and Nagin (2006) for these data. The group-based analysis does not cover cluster sampling and has the further restrictions of zero within-class variances $\Psi_c = 0$, as well as $\Theta_c = \theta I$. The group-based approach is further discussed in Muthén (2004) where it is referred to as latent-class growth analysis (LCGA), given its similarity to latent-class analysis (LCA). Both LCGA and LCA search for classes of individuals defined by conditional independence of the repeated measures given class. In contrast, a growth mixture model (GMM) allows for within-class correlations between repeated measures. Such correlation may, for example, be due to omitted time-varying covariates. If within-class correlation is ignored, a distorted class formation is obtained. Within-class correlation is obtained in GMMs by allowing for random effects with non-zero within-class variances. Both LCGA and GMMs use a zero-inflated Poisson model in line with Roeder, Lynch, and Nagin (1999). For time point j, individual i, and cluster k,

$$Y_{kij}|_{C_{ki}=c} = \begin{cases} 0 & \text{with probability } \pi_{kij}, \\ \text{Poisson}(\lambda_{ckij}) & \text{with probability } 1 - \pi_{kij} \end{cases}$$

where λ is the Poisson rate. In line with previous modeling of the Philadelphia data, a quadratic growth curve is used. Drawing on (6.8) and (6.9) of the general model in Section 6.3.2, the growth mixture zero-inflated Poisson model for these data is expressed in

terms of the log rate as

$$\log \lambda_{ij|C_i=c} = \eta_{0i} + \eta_{1i}\, a_{ij} + \eta_{2i}\, a_{ij}^2,$$
$$\eta_{0i|C_i=c} = \alpha_{0c} + \zeta_{0i},$$
$$\eta_{1i|C_i=c} = \alpha_{1c} + \zeta_{1i},$$
$$\eta_{2i|C_i=c} = \alpha_{2c} + \zeta_{2i}.$$

To make analysis results comparable to the LCGA of Loughran and Nagin (2006), a minority of individuals with more than 10 criminal offenses in any given year are deleted, reducing the sample size only from 13,160 to 13,126, and combining the data into two-year intervals. Loughran and Nagin (2006) settled on a four-class solution: non-offenders, adolescent-limited, and high and low chronic (persisting criminal activity at age 26). D'Unger et al. (1998) and D'Unger, Land, and McCall (2002) used a random subset ($n = 1000$) of the data and concluded based on BIC that a five-class LCGA solution was preferred. Their five classes were labeled: non-offenders, high and low adolescent-peaked, and high and low chronic.

Table 6.1 gives results for 1–4 classes of GMM and 4–8 classes for LCGA. In addition to log-likelihood values, number of parameters, and BIC, the table shows fit to the data in terms of the number of standardized residuals that are significant at the 5% level for the 10 most frequent response patterns across time (comprising 78% of the data and eliminating only patterns with observed frequency less than 100). The one-class GMM is the conventional random-effects model. Here, 5 of the 10 residuals show significant misfit, illustrating the need for a more flexible model. The two- and the three-class GMMs obtain considerably improved BIC values. The three-class GMM reduces the number of significant residuals from 5 to 1, indicating the appropriateness of the mixture modeling. The four-class GMM adds relatively little improvement. The three-class GMM displays the three themes of non-offenders, adolescent-limited, and chronic. Figure 6.4 shows the mean trajectories for the three-class GMM. The four-class GMM splits the adolescent-limited class into two, where the total percentage for those two classes is about the same as for the adolescent-limited class of the three-class GMM.

The four-class LCGA is the same as presented in Loughran and Nagin (2006) and the five-class LCGA shows the same types of trajectory classes as in D'Unger et al.'s analysis. Neither of these two models fit the data well. An eight-class LCGA is needed to get a reduction to one significant residual. In contrast, the three-class GMM has only one significant residual and the four-class GMM has none. With three classes the GMM gives a better BIC value than any of the LCGA models shown in Table 6.1. The BIC values for the four-class LCGA

Table 6.1 Age–crime curves: Log-likelihood and BIC comparisons for GMM and LCGA

Model	Log-Likelihood	# Parameters	BIC	# Significant Residuals
1-class GMM	−40,606	17	81,373	5
2-class GMM	−40,422	21	81,044	4
3-class GMM	−40,283	25	80,803	1
4-class GMM	−40,237	29	80,748	0
4-class LCGA	−40,643	23	81,503	4
5-class LCGA	−40,483	27	81,222	3
6-class LCGA	−40,410	31	81,114	3
7-class LCGA	−40,335	35	81,003	2
8-class LCGA	−40,263	39	80,896	1

EXAMPLES   155

used in Loughran and Nagin (2006) and the five-class model used in D'Unger et al. (1998) and D'Unger, Land, and McCall (2002) are considerably worse than the BIC value for the three-class GMM. Furthermore, the three-class GMM uses two parameters less than the five-class LCGA, but has a better log-likelihood by 200 points. This illustrates the importance of using random effects to allow for variations on the themes of the trajectory shapes of the classes. The LCGA approach leads to a proliferation of classes, all of which may not have substantive salience.

6.5.2 Analysis of Example 2: Varying intervention effects on classroom aggressive behavior

The Baltimore randomized field trial discussed in Section 6.2.2 was repeated for several cohorts of students. The Section 6.2.2 analysis considered cohort 1 data, whereas data from cohort 3 (Ialongo et al., 1999) are analyzed here. A total of 362 boys in 27 classrooms are considered over four time points: fall of first grade, spring of first grade, spring of second grade, and spring of third grade. The average number of boys per classroom is 13.4. It is of interest to study if teachers in classrooms with higher aggressiveness levels have a more difficult time successfully implementing the intervention aimed at reducing aggressive-disruptive behavior. For the first grade, there is substantial variation across classrooms in the aggressiveness scores as evidenced by the intraclass correlations at the four time points: $0.11, 0.16, 0.04, 0.00$. In addition to student-level trajectory classes, the use of latent classes on the classroom level makes it possible to more fully explore variation in intervention effects.

Drawing on the Section 6.3.2 general model, the two-level GMM is expressed as follows using a quadratic curve shape,

$$Y_{kij|C_{ki}=c,D_k=d} = \eta_{0ki} + \eta_{1ki}\, a_{ij} + \eta_{2ki}\, a_{ij}^2 + \epsilon_{kij},$$
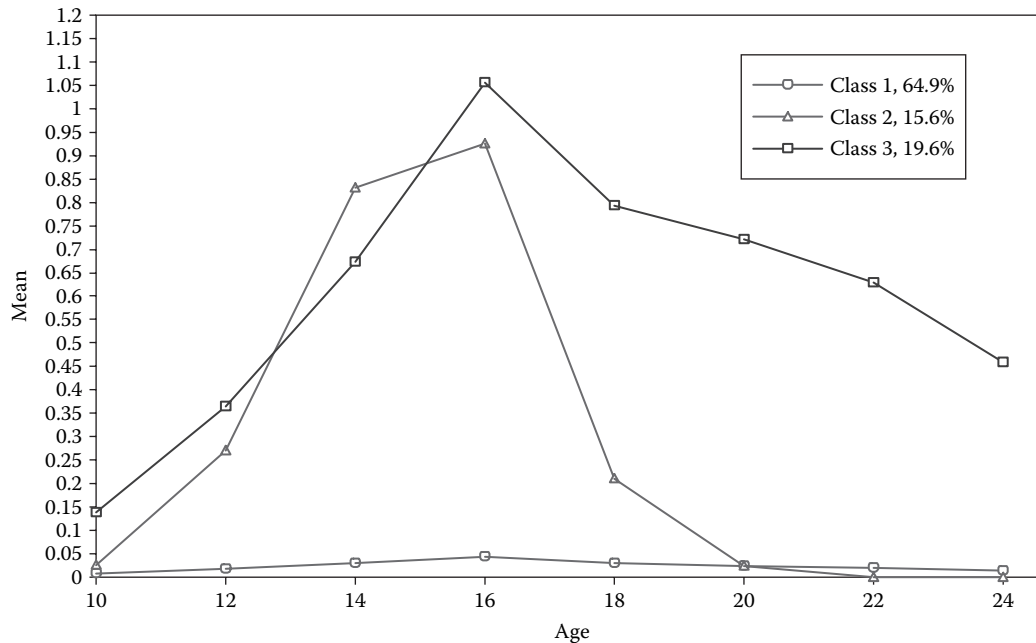

Figure 6.4 Estimated mean trajectories from a three-class growth mixture model for criminal activities.

for $j = 1, 2, 3, 4$, with variation across students within classrooms expressed as

$$\eta_{0ki}|_{C_{ki}=c, D_k=d} = \alpha_{cdk0} + \zeta_{0ki},$$
$$\eta_{1ki}|_{C_{ki}=c, D_k=d} = \alpha_{cdk1} + \zeta_{1ki},$$
$$\eta_{2ki}|_{C_{ki}=c, D_k=d} = \alpha_{cdk2} + \zeta_{2ki},$$

and variation across classrooms expressed as

$$\alpha_{cdk0}|_{C_{ki}=c, D_k=d} = \alpha_{cd0} + \zeta_{30k},$$
$$\alpha_{cdk1}|_{C_{ki}=c, D_k=d} = \alpha_{cd1} + \gamma_{cd1} Z_k + \zeta_{31k},$$
$$\alpha_{cdk2}|_{C_{ki}=c, D_k=d} = \alpha_{cd2} + \gamma_{cd2} Z_k + \zeta_{32k}.$$

Here, $a_{i1} = 0$ to center the intercept η_0 at the pre-intervention time point. Z is a treatment-control dummy variable on the classroom level. For reasons of parsimony, the student-level latent-class variable C and the classroom-level latent-class variable D are taken to have an additive effect on the means α_{cd0}, α_{cd1}, and α_{cd2}. The γ intervention effects are, however, allowed to vary across combinations of C and D classes. The linear and quadratic slopes were found to have zero variance across classrooms. The intraclass correlation is captured by the classroom variation in the random intercept of the growth model, α_{cdk0}.

The latent categorical variable C_{ki} follows the multinomial logistic regression

$$\Pr(C_{ki} = c | D_k = d) = \frac{\exp(a_{cdk})}{\sum_s \exp(a_{sdk})},$$

where in this application

$$a_{cdk}|_{D_k=d} = a_c + \zeta_{ck}. \tag{6.16}$$

The analyses indicate that $V(\zeta_{ck}) = 0$, that is, the random intercepts for the latent-class variable C do not vary across classrooms. In other applications, however, this variance can be substantial.

As a first step, a model without the classroom-level latent-class variable D was explored. As judged by BIC, the conventional single-class random-effects growth model is clearly outperformed by growth mixture modeling, with a three-class model giving the lowest BIC. The log-likelihood for the conventional model is -4157.98 with 14 parameters and a BIC of 8398, while the three-class GMM has a log-likelihood of -4048.84 with 26 parameters and a BIC of 8251. The three-class model has a significant classroom variance for the random intercept. Second, two latent classes for D were added to the model resulting in latent classes with low versus high classroom-level aggression (51% versus 49%). The log-likelihood is -4041.42 with 34 parameters and a BIC of 8283. This BIC is not as good as for the previous model with no classroom-level latent-class variable, but it is not known how BIC performs in settings with multilevel latent-class variables. The three student-level latent-trajectory classes show a low-increasing class of 68%, a medium-increasing class of 19%, and a high-decreasing class of 12%. The mean curves for these three latent classes are shown in Figure 6.5 as pairs of control and intervention curves. Results for the latent class consisting of classrooms with low aggression level are given in the left plot and results for the latent class consisting of classrooms with high aggression level are given in the right plot. The plots suggest that in classrooms with a low level of aggression, students who are in the two highest trajectory classes benefit from the intervention. In classrooms with a high level of aggression, however, only students who are in the lowest trajectory class benefit from the intervention. This suggests that the intervention may be harder for teachers to implement well in high-aggressive classrooms. The results should be interpreted with caution, however, given the sample of only 27 classrooms and other competing models. An alternative model lets the C and D latent-class variables have an interactive effect on the
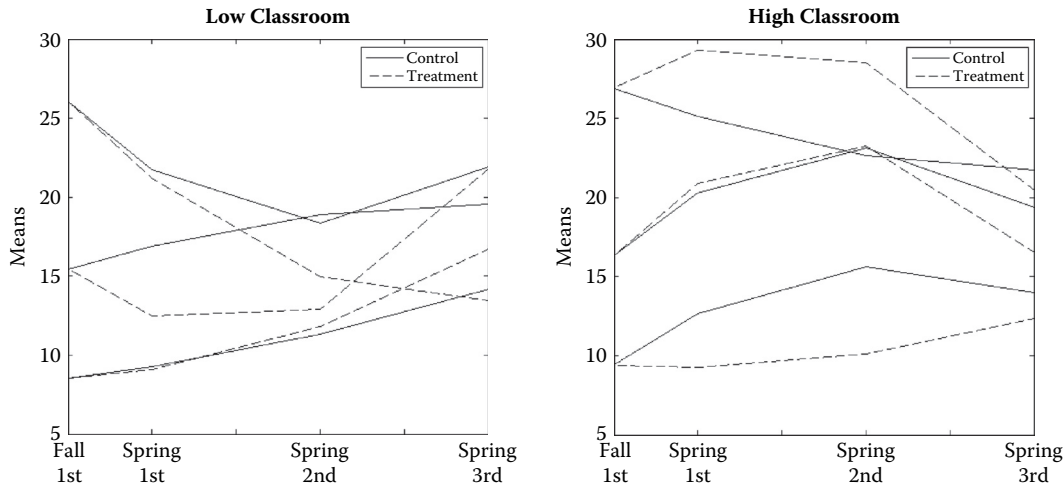
Figure 6.5 Estimated mean trajectories from a growth mixture model of classroom aggressive behavior.

random-effect growth means and lets the random-effect means of a_{cdk} in (6.16) be influenced by the latent classes of D. This significantly improves the log-likelihood, but the increased number of parameters on the classroom level results in a less stable solution. The resulting split into 7 and 20 classrooms for the latent classes of D causes estimated outcome mean differences with high variability.

6.5.3 Analysis of Example 5: Classification of schools based on achievement development

The NELS math achievement data from grades 8, 10, and 12 discussed in Section 6.2.5 are analyzed here. NELS surveyed 913 schools and a total of 14,217 students. The NELS analysis illustrates two features of the Section 6.3.2 model, taking into account the school clusters and using a school-level latent-class variable.

In the NELS analysis, student growth rate is regressed on the growth intercept in grade 8 using a random slope that varies across schools. This random slope and the means of the random intercept and the intercept of the random growth rate are allowed to vary across the school-level latent classes. Letting school-level latent classes influence student-level relations helps identify the school-level latent classes. Extending the example of Section 6.2.3 to clusters k and a cluster-level latent-class variable D_k, variation across grades is expressed as

$$Y_{kij|D_k=d} = \eta_{0ki} + \eta_{1ki}\, a_{ij} + \epsilon_{kij},$$

for $j = 1, 2, 3$, with variation across students expressed as

$$\eta_{0ki|D_k=d} = \alpha_{d0} + \zeta_{0ki},$$
$$\eta_{1ki|D_k=d} = \alpha_{d1} + \beta_{dk}\, \eta_{0ki} + \zeta_{1ki},$$

where the variation across schools is accomplished by the variation of α_{d0}, α_{d1}, and β_{dk} across the classes of D. A single-class growth model, that is, a conventional three-level analysis, obtains a log-likelihood of $-31{,}791$ with 10 parameters and a BIC of 63,678. A two-class GMM obtains a log-likelihood of $-31{,}545$ with 16 parameters and a BIC of 63,243. A three-class GMM obtains a log-likelihood of $-31{,}434$ with 22 parameters and a BIC of 63,079. A four-class model does not improve the log-likelihood further. The three-class model shows that the growth rate is significantly positively related to the growth intercept defined at grade 8 only for a class of 52% of the schools who have average growth over grades 8–12. A higher developing class of 25% and a lower developing class of 23% have small and

insignificant relationships. This illustrates the possibility of finding clusters of schools with different achievement profiles. School-level covariates predicting school class membership can give further understanding of the school classes.

6.6 Parametric versus non-parametric random-effect models

Titterington, Smith, and Makov (1985) make a distinction in the use of finite-mixture modeling in terms of direct and indirect applications. A direct application uses mixtures to represent "the underlying physical phenomenon," whereas with the indirect application the mixture components "do not necessarily have a direct physical interpretation." The examples discussed so far can be seen as attempts at direct application, where trajectory classes are given substantive interpretation and results are presented for each mixture component rather than mixing over the classes. Examples of indirect applications include outlier detection and representation of non-normal distributions. Mixture modeling of non-normal distributions is the focus of this section. A growth model with a non-parametric representation of the random-effects distribution is presented and a simulation study compares the use of such a model to the conventional random-effect growth model assuming normality.

It has been argued that with categorical and count outcomes, the typical normality assumption for random effects in repeated-measurement modeling may be less well supported by data (see also Aitkin, 1999). Deviations from normality may strongly affect the results. With categorical and count outcomes, maximum likelihood leads to the use of numerical integration which is computationally heavy and intractable when the number of random effects is large. Numerical integration uses fixed quadrature points and weights according to a normal distribution. A non-parametric approach instead considers a discretized distribution, estimating the points and the weights using a finite-mixture model. The latent class means are the points and the class probabilities are the weights. In this way, the non-parametric approach both avoids the normality specification and is computationally less demanding.

In this section we describe and compare the general parametric and non-parametric random-effect models. Both of these models are special cases of the general model described in equations (6.8)–(6.12). Both of these modeling alternatives attempt to capture cluster-specific effects. The difference between the two models is the underlying assumption for the distributions of the cluster-specific random effects. In the parametric model the random effects are assumed to a have conditionally normal distribution, that is, the conditional distribution of the random effects, given all covariates, is assumed to be normal. In the non-parametric model the random effect are assumed to have a non-parametric conditional distribution.

The parametric random-effect model is well established and frequently used in practice. Butler and Louis (1992) show that the normality assumption in the parametric model does not affect the fixed slopes in the model. Verbeke and Lesaffre (1996) show that more accurate estimates can be obtained for the random effects if a non-normal distribution is estimated. Aitkin (1999) gives the general modeling approach to the non-parametric random-effect models that we follow here. First we give the complete description of the two modeling alternatives and show how they fit in the general modeling framework (6.8)–(6.12).

6.6.1 Parametric random-effect model

This model is a special case of model (6.8)–(6.12) for the case of no categorical latent variables. The within-level model is given by

$$\boldsymbol{Y}_{ki}^* = \boldsymbol{\nu}_k + \Lambda_k \boldsymbol{\eta}_{ki} + K_k \boldsymbol{X}_{ki} + \boldsymbol{\epsilon}_{ki}, \tag{6.17}$$

$$\boldsymbol{\eta}_{ki} = \boldsymbol{\alpha} + B_k \boldsymbol{\eta}_{ki} + \Gamma_k \boldsymbol{X}_{ki} + \boldsymbol{\zeta}_{ki}. \tag{6.18}$$

The coefficients $\boldsymbol{\nu}_k$, Λ_k, K_k, B_k, and Γ_k can be either fixed coefficients that are the same across cluster or random effects that vary across cluster. Let $\boldsymbol{\eta}_k$ represent the vector of all such random effects. The between-level model is described by

$$\boldsymbol{\eta}_k = \boldsymbol{\mu} + B\boldsymbol{\eta}_k + \Gamma\,\boldsymbol{X}_k + \boldsymbol{\zeta}_k. \tag{6.19}$$

The random-effect residuals $\boldsymbol{\zeta}_{ki}$ and $\boldsymbol{\zeta}_k$ are assumed normally distributed. This assumption is the difference between the parametric and the non-parametric model. Note that the distributional assumption for $\boldsymbol{\epsilon}_{ki}$ is determined by the type of observed variable we model.

6.6.2 Non-parametric random-effect model

This model is a special case of model (6.8)–(6.12) where the random effects $\boldsymbol{\eta}_{ki}$ and $\boldsymbol{\eta}_k$ do not have normally distributed residuals $\boldsymbol{\zeta}_{ki}$ and $\boldsymbol{\zeta}_k$. The within-level model is given by

$$\boldsymbol{Y}_{ki}^* = \boldsymbol{\nu}_k + \Lambda_k\boldsymbol{\eta}_{ki} + K_k\boldsymbol{X}_{ki} + \boldsymbol{\epsilon}_{ki},$$

$$\boldsymbol{\eta}_{ki}|_{C_{ki}=c} = \boldsymbol{\alpha}_c + B_k\boldsymbol{\eta}_{ki} + \Gamma_k\boldsymbol{X}_{ki}, \tag{6.20}$$

$$\Pr(C_{ki} = c) = p_c, \tag{6.21}$$

where p_c are parameters to be estimated. The coefficients $\boldsymbol{\nu}_k$, Λ_k, K_k, B_k, and Γ_k can again be either fixed coefficients or random effects. Let $\boldsymbol{\eta}_k$ represent the vector of all random effects. The between-level model is given by

$$\boldsymbol{\eta}_k|_{D_k=d} = \boldsymbol{\mu}_d + B\boldsymbol{\eta}_k + \Gamma\boldsymbol{X}_k, \tag{6.22}$$

$$\Pr(D_k = d) = q_d, \tag{6.23}$$

where q_d are parameters to be estimated. The random-effect model (6.20)–(6.23) can alternatively be presented as in equations (6.18)–(6.19), considering the mixture across classes

$$\boldsymbol{\eta}_{ki} = \boldsymbol{\alpha} + B_k\boldsymbol{\eta}_{ki} + \Gamma_k\boldsymbol{X}_{ki} + \boldsymbol{\zeta}_{ki},$$

$$\boldsymbol{\eta}_k = \boldsymbol{\mu} + B\boldsymbol{\eta}_k + \Gamma\boldsymbol{X}_k + \boldsymbol{\zeta}_k,$$

where $\boldsymbol{\alpha} = \sum_c \boldsymbol{\alpha}_c p_c$, $\boldsymbol{\mu} = \sum_d \boldsymbol{\mu}_d q_d$, and $\boldsymbol{\zeta}_{ki}$ and $\boldsymbol{\zeta}_k$ are non-parametric zero-mean residuals that are freely estimated. The residual $\boldsymbol{\zeta}_{ki}$ takes the values $\boldsymbol{\alpha}_c - \boldsymbol{\alpha}$ with probability p_c and the residual $\boldsymbol{\zeta}_k$ takes values $\boldsymbol{\mu}_d - \boldsymbol{\mu}$ with probability q_d. The variance and covariance for the non-parametric effects can also be computed; for example, the variance of $\boldsymbol{\zeta}_{ki}$ is $\sum_c p_c(\boldsymbol{\alpha}_c - \boldsymbol{\alpha})(\boldsymbol{\alpha}_c - \boldsymbol{\alpha})'$.

6.6.3 Simulation study

A simulation study is conducted to compare the performance of the parametric and non-parametric random-effect models for data generated with non-normal random effects. Consider a logistic growth model with 10 binary items $U_1, \ldots, U_{10}$,

$$\log\left[\frac{\Pr(U_{ij} = 1)}{\Pr(U_{ij} = 0)}\right] = \eta_{0i} + \eta_{1i}\,a_{ij}, \tag{6.24}$$

where the time scores $a_{ij} = (j-1)/2$, and η_0 and η_1 are non-normal random effects. Generation of η_0 and η_1 used the following finite mixture of normal distributions:

$$0.67 \cdot N(\mu_1, \sigma^2) + 0.09 \cdot N(\mu_2, \sigma^2) + 0.24 \cdot N(\mu_3, \sigma^3).$$

To generate η_0, the following parameters were used: $\mu_1 = 2$, $\mu_2 = 1$, $\mu_3 = 0$, and $\sigma = 0.4$. To generate η_1, the following parameters were used: $\mu_1 = -0.3$, $\mu_2 = -0.4$, $\mu_3 = -1$, and $\sigma = 0.1$. From these values, 100 samples of size 2000 were generated according to the model (6.24). The data were analyzed using the parametric linear model (PM) and

Table 6.2 Comparing the parametric (PM) and non-parametric (NPM) random-effect models

Parameter	True value	PM bias	NPM bias	PM MSE	NPM MSE
m_0	1.421	0.004	-0.010	0.0018	0.0018
m_1	-0.480	0.073	0.004	0.0056	0.0003
v_0	0.733	0.563	0.080	0.3302	0.0122
v_1	0.088	-0.033	-0.004	0.0012	0.0001
ρ	0.247	-0.006	0.003	0.0008	0.0002

the non-parametric linear model (NPM). PM is a conventional single-class growth model with random normal effects as in (6.17)–(6.18). Drawing on (6.20) and (6.22), the NPM is expressed as

$$Y_{ij}^* = \eta_{0i} + \eta_{1i}\, a_{ij} + \epsilon_{ij},$$

$$\eta_{0i|C_i=c} = \alpha_{0c},$$
$$\eta_{1i|C_i=c} = \alpha_{1c},$$

so that the random effects are represented by a mixture distribution. A more general form would allow within-class variation for residuals ζ as in (6.9).

The parameter estimates are summarized in Table 6.2. The means of η_0 and η_1 are denoted by m_0 and m_1 and the variances by v_0 and v_1. The covariance of η_0 and η_1 is denoted by ρ. The results are presented for the non-parametric model with three nodes, since three nodes were determined to be sufficient for most replications using the McLachlan and Peel (2000) parametric likelihood ratio test. The estimates on which Table 6.2 is based are computed for the mixture over the three classes in line with (6.20) and (6.22). The results in Table 6.2 clearly indicate the advantages of the NPM method. The NPM parameter estimates have substantially smaller bias and smaller mean squared error (MSE) for several parameters.

In general it is difficult to evaluate model fit for random-effect models. There is no general unrestricted model which can be used for comparison. In this simulated example, however, there is such a model, namely, the completely unrestricted contingency table for the binary items. In addition, the Pearson chi-square test can be used to test the fit of the model. The data were generated according to a linear growth model with non-normal random effects; that is, the true model is a linear random-effect growth model. Both the PM and NPM models are linear random-effect growth models but are based on different assumptions on the distribution of the random effects; neither assumption specifies the true random-effect distribution. This situation is typical in practical applications, where the true random-effect distribution is unknown and the modeling assumptions are likely to deviate from the true distribution to some extent. It is assumed that the distributional misspecification will not interfere with the basic structure of the model and that the estimated model will provide a good fit for the data despite the distributional misspecifications. The Pearson test of fit can be used to directly compare the sensitivity of the PM and NPM models. If the Pearson test rejects the model, one concludes that the model fit is poor. In practical applications the lack of model fit could incorrectly be interpreted as evidence for deficiency in the linear growth structure of the model rather than as possible misspecification in the random-effect distribution.

In the current simulated example one wants the Pearson test to reject the model no more than the nominal 5% of the time. Table 6.3 contains the Pearson test of fit results for the PM linear growth model as well as the NPM linear growth model using 3, 4, 5, and 6 nodes. The rejection rate in Table 6.3 is the percentage of times the linear growth model was rejected incorrectly. Also presented are the average test statistic value and the degrees of freedom. These two values should generally be close because the expected value of the chi-square distribution is equal to the degrees of freedom. The parametric approach

Table 6.3 Pearson test of fit for the binary linear growth model

Model	PM	NPM(3)	NPM(4)	NPM(5)	NPM(6)
Rejection rate	100%	22%	15%	8%	4%
Average test statistic	1252	1038	1025	1007	997
Degrees of freedom	1018	1015	1012	1009	1006

rejected the correct model 100% of the time. There is a large difference between the degrees of freedom and the average test value for the parametric model. On the other hand, the non-parametric approach leads to the correct conclusion. The Pearson rejection rate converges to the nominal 5% value as the number of nodes is increased from 3 to 6. The average test statistic value also converges towards its expected value.

As yet another test of model fit, the estimated parametric and non-parametric models are used to compute response pattern frequencies. Table 6.4 shows the top five most frequent patterns in one of the generated data sets as well as the predicted frequency for these patterns, using both the parametric and the non-parametric random-effect models. For the non-parametric model three-node estimation is used. In this comparison as well one can clearly see the advantage of the non-parametric approach. The non-parametric pattern frequencies are more closely matching the observed values than the parametric pattern frequencies.

The simulation study clearly demonstrates the fact that model estimation with normal random effects is not robust to the distributional assumptions. If the random effects are truly non-normal the parametric model can actually have a poor fit to the data. This phenomenon is exacerbated by the fact that in general random-effect models have no test of model fit. Thus researchers can easily be mislead into believing that a random-effect model is a useful model. Violations of the normality assumptions can lead to poor estimates and inference. In addition, a comparison between the non-parametric model and the parametric model that is limited to the fixed-effect parameters can be misleading, because even when the fixed parameter estimates as well as the random-effect parameter estimates are identical the models implied by the two approaches are quite different. Even when the parametric model fits the first- and second-order moments of the data, the model could have a much worse fit to the data than the non-parametric model, simply because it will not fit higher-order moments. The non-parametric model described in this section overcomes these shortcomings and should be used more frequently in practice.

6.7 Conclusions

The range of growth mixture analyses described in this chapter shows the usefulness of extending the conventional random-effects repeated-measures model (mixed linear model) to include different latent classes of development. Growth mixture modeling is a powerful analytic tool when applied to randomized trials as well as to non-experimental research. The idea of detecting different treatment effects for individuals belonging to different trajectory classes has important implications for designing future intervention studies. It is possible to

Table 6.4 Observed and estimated pattern frequencies for the top five most frequent patterns

Pattern	Observed	PM-Estimated	NPM-Estimated
1111111111	141	246	135
0000000000	83	66	74
1000000000	61	39	69
1111111110	55	36	54
1111111101	49	33	47

select different interventions for individuals belonging to different trajectory classes using longitudinal screening procedures. One may attempt to classify individuals into their most likely trajectory class based on a set of initial repeated measurements before the intervention starts. Alternatively, one may administer a universal intervention and follow up with a targeted intervention for individuals who show little or no intervention effect.

In non-experimental research, the exploration of different trajectory classes provides a more nuanced picture of development, where different classes can have different antecedents and consequences. As a form of cluster analysis, both the use of individual-level and cluster-level latent-class variables defined from longitudinal data have great promise.

This chapter also studied non-parametric representation of random-effect distributions in growth models. It was shown that mixture modeling with latent classes can provide a good representation of non-normal effects in cases where the conventional approach drawing on normality gives misleading results. In this connection, it is interesting to consider the age–crime analysis of Example 4 in Section 6.5.1. Here, three substantively meaningful trajectory classes were found using growth mixture analysis, each showing within-class variation on the theme of the trajectory shape. The variation was captured by normally distributed random effects. The latent-class growth analysis that was presented may be seen as representing the random effects non-parametrically. In this example, the non-parametric approach leads to a proliferation of latent classes. An alternative is to postulate three substantive classes and allow for within-class variation by a non-parametric approach. The estimates would be mixed together across the non-parametric classes, but not across the substantive classes.

The statistical developments behind growth mixture modeling are still relatively recent and much statistical research remains to be done. For example, the quality of parameter recovery under different conditions needs to be more fully understood. The Monte Carlo simulation facility in Mplus is useful in this regard. Tests need to be further developed to compare models differing not only in the number of latent classes, but also in their random-effect specification. Residual checking of model fit needs to be further developed. Furthermore, much more practical experience is needed with these analyses to create guidelines for how to build models. Hopefully this chapter will stimulate such statistical work.

Acknowledgments

References

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

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society B* **144**, 419–461.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**, 444–455.

Asparouhov, T. and Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock and K. M. Samuelson (eds.), *Advances in Latent Variable Mixture Models*, pp. 27–51. Charlotte, NC: Information Age Publishing.

Asparouhov, T., Masyn, K., and Muthén, B. (2006). Continuous time survival in latent variable models. In *2006 Proceedings of the American Statistical Association, Biometrics Section*, pp. 180–187. Alexandria, VA: American Statistical Association.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* **92**, 1375–1386.

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

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46**, 443–459.

Butler, S. M. and Louis, T. A. (1992). Random effects models with non-parametric priors. *Statistics in Medicine* **11**, 19812000.

Choi, K., and Seltzer, M. (2006). Modeling heterogeneity in relationships between initial status and rates of change: Treating latent variable regression coefficients as random coefficients in a three-level hierarchical model. Accepted for publication in *Journal of Educational and Behavioral Statistics.*

Croudace, T. J., Jarvelin, M. R., Wadsworth, M. E., and Jones, P. B. (2003). Developmental typology of trajectories to nighttime bladder control: Epidemiologic application of longitudinal latent class analysis. *American Journal of Epidemiology* **157**, 834–42.

Dolan, L., Kellam, S. G., Brown, C. H., Werthamer-Larsson, L., Rebok, G. W., Mayer, L. S., Laudolff, J., Turkkan, J. S., Ford, C., and Wheeler, L. (1993). The short-term impact of two classroom based preventive intervention trials on aggressive and shy behaviors and poor achievement. *Journal of Applied Developmental Psychology* **14**, 317–345.

D'Unger, A. V., Land, K. C., and McCall, P. L. (2002). Sex differences in age patterns of delinquent/criminal careers: Results from Poisson latent class analyses of the Philadelphia cohort study. *Journal of Quantitative Criminology* **18**, 349–375.

D'Unger, A. V., Land, K. C., McCall, P. L., and Nagin, D. S. (1998). How many latent classes of delinquent/criminal careers? Results from mixed Poisson regression analysis. *American Journal of Sociology* **103**, 1593–1630.

Dunn, G., Maracy, M., Dowrick, C., Ayuso-Mateos, J. L., Dalgard, O. S., Page, H., Lehtinen, V., Casey, P., Wilkinson, C., Vasquez-Barquero, J. L., and Wilkinson, G. (2003). Estimating psychological treatment effects from a randomized controlled trial with both non-compliance and loss to follow-up. *British Journal of Psychiatry* **183**, 323–331.

Ialongo, L. N. , Werthamer, S., Kellam, S. K., Brown, C. H., Wang, S., and Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression and antisocial behavior. *American Journal of Community Psychology* **27**, 599–641.

Jo, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics* **27**, 385–409.

Jo, B. and Muthén, B. (2003). Longitudinal studies with intervention and noncompliance: Estimation of causal effects in growth mixture modeling. In S. P. Reise and N. Duan (eds.), *Multilevel Modeling: Methodological Advances, Issues, and Applications*, pp. 112–139. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Kass R. E. and Raftery, A. E. (1993). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kellam, S. G., Rebok, G. W., Ialongo, N., and Mayer, L. S. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry* **35**, 359–382.

Kreuter, F. and Muthén, B. (2007). Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology.* To appear.

Kreuter, F. and Muthén, B. (2008). Longitudinal modeling of population heterogeneity: Methodological challenges to the analysis of empirically derived criminal trajectory profiles. In G. R.

Hancock and K. M. Samuelson (eds.), *Advances in Latent Variable Mixture Models*. Charlotte, NC: Information Age Publishing.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Leuchter, A. F., Cook, I. A., Witte, E. A., Morgan, M., and Abrams, M. (2002). Changes in brain function of depressed subjects during treatment with placebo. *American Journal of Psychiatry* **159**, 122–129.

Lin, J. Y., Ten Have, T. R., and Elliott, M. R. (2006). Longitudinal nested compliance class model in the presence of time-varying noncompliance. *Journal of the American Statistical Association*. To appear.

Lin, H., Turnbull, B. W., McCulloch, C. E., and 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.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York: Wiley.

Loughran, T. and Nagin, D. S. (2006). Finite sample effects in group-based trajectory models. *Sociological Methods & Research* **35**, 250–278.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

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

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **49**, 115–132.

Muthén, B. (2001a). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In L. M. Collins and A. Sayer (eds.), *New Methods for the Analysis of Change*, pp. 291–322. Washington, DC: APA.

Muthén, B. (2001b). Latent variable mixture modeling. In G. A. Marcoulides and R. E. Schumacker (eds.), *New Developments and Techniques in Structural Equation Modeling*, pp. 1–33. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Muthén, B., Jo, B., and Brown, C. H. (2003). Comment on the Barnard, Frangakis, Hill and 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. and Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics* **30**, 27–58.

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

Muthén, B. and Muthén, L. (1998–2006). *Mplus User's Guide*. Los Angeles: Muthén and Muthén.

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

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C., Wang, C. P., Kellam, S., Carlin, J., and Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics* **3**, 459–475.

Muthén, B., Khoo, S. T., Francis, D., and Boscardin, K. (2003). Analysis of reading skills development from Kindergarten through first grade: An application of growth mixture modeling to

sequential processes. In S. P. Reise and N. Duan (eds.), *Multilevel Modeling: Methodological Advances, Issues, and Applications* (2003). Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, B., Brown, C. H., Leuchter, A., and Hunter, A. (2007). General approaches to analysis of course: Applying growth mixture modeling to randomized trials of depression medication. In P. E. Shrout (ed.), *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures.* Washington, DC: American Psychiatric Publishing. To appear.

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

Nagin, D. S. (2005). *Group-Based Modeling of Development.* Cambridge, MA: Harvard University Press.

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

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

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

Pearson, J. D., Morrell, C. H., Landis, P. K., Carter, H. B., and Brant, L. J. (1994). Mixed-effect regression models for studying the natural history of prostate disease. *Statistics in Medicine* **13**, 587–601.

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

Sampson, R. J. and Laub, J. H. (2005). Seductions of methods: Rejoinder to Nagin and Tremblay's "Developmental trajectory groups: Fact or fiction?". *Criminology* **43**, 905–913.

Schilling, S. and Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika* **70**, 533–555.

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

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217–221.

Werthamer-Larsson, L., Kellam, S. G., and Wheeler, L. (1991). Effect of first-grade classroom environment on child shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology* **19**, 585–602.

Wang, C. P., Brown, C. H., and Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association* **100**, 1054–1076.

Yau, L. H. Y. and Little, R.J. (2001). Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association* **96**, 1232–1244.