HHS Public Access

Author manuscript

Behav Res Ther. Author manuscript; available in PMC 2018 November 01.

Published in final edited form as:

Behav Res Ther. 2017 November ; 98: 91–102. doi:10.1016/j.brat.2017.04.003.

Fitting Latent Variable Mixture Models

Gitta H Lubke[1,2,*] and **Justin Luningham**[1]

[1]Department of Psychology, University of Notre Dame, Notre Dame, Indiana [2]Department of Biological Psychiatry, VU University, Amsterdam, The Netherlands

Abstract

Latent variable mixture models (LVMMs) are models for multivariate observed data from a potentially heterogeneous population. The responses on the observed variables are thought to be driven by one or more latent continuous factors (e.g. severity of a disorder) and/or latent categorical variables (e.g., subtypes of a disorder). Decomposing the observed covariances in the data into the effects of categorical group membership and the effects of continuous trait differences is not trivial, and requires the consideration of a number of different aspects of LVMMs. The first part of this paper provides the theoretical background of LVMMs and emphasizes their exploratory character, outlines the general framework together with assumptions and necessary constraints, highlights the difference between models with and without covariates, and discusses the interrelation between the number of classes and the complexity of the within-class model as well as the relevance of measurement invariance. The second part provides a growth mixture modeling example with simulated data and covers several practical issues when fitting LVMMs.

Introduction

Latent variable mixture models (LVMMs) combine latent class analysis models and factor models or more complex structural equation models (Muthén, 2001). LVMMs are most commonly used to investigate population heterogeneity, which refers to the presence of subgroups in the population. LVMMs can serve to analyse data from heterogeneous populations without knowing beforehand which individual belongs to which of the subgroups.

The simplest types of mixture models are latent class analysis (LCA) models. These models are designed for multiple observed variables (e.g., symptom endorsements, of questionnaire items), and have a single latent class variable that groups the individuals in a sample into a user-specified number of latent groups (Lazarsfeld & Henry, 1968; McCutcheon, 1987). LCA models do not have factors within class, and the covariances between the observed variables *within class* are constrained to zero[1]. This is a very stringent assumption. Suppose we have 5 observed items measuring some disorder. Not allowing these items to covary

[*]Correspondence to: Gitta Lubke, 110 Haggar Hall, Notre Dame, Indiana 46556, (574) 631-8789, glubke@nd.edu.

[1]This is called the “local independence assumption”

within class means that there are no systematic severity differences between participants within a class in LCA models. The covariances between observed variables *in the total sample* only deviate from zero due to mean differences between the classes.

Factor models on the other hand are models for a single homogeneous population (i.e., no differences between subtypes), and observed variables in the sample are assumed to covary due to systematic differences along the underlying continuous latent factors (Bollen, 1989).

LVMMs can have one or more latent class variables, and permit the specification of factor models, growth models, or even more complex models within each class. If the within class model is a factor model, the resulting LVMM is often called factor mixture model. Covariances between observed variables in the total sample are attributed partially to mean differences between classes, and partially to continuous latent factors within each class. For example, consider data collected on several questionnaire items that measure anger. Suppose the population consists of two groups, a majority group of participants with very low levels of anger and a smaller group characterized by high scores on most of the items. The observed anger items in the total sample covary because of the mean differences between the two groups. In addition, the items can also covary if there are differences in the severity of anger *within* each group. These two sources of covariance are modeled in LVMMs by using latent categorical and latent continuous variables.

Latent class models are a special case of the LVMM where factor variances (or, alternatively, factor loadings) are zero. In the anger example this would mean that all participants within the low-scoring class do not differ in the severity of anger (i.e., zero anger factor variance within group). The same holds for the high scoring group: the assumption of the latent class model is no variability of anger within group because if there were systematic anger differences within class then the items would in fact covary. The observed covariances between the anger items in this model are modeled to be entirely due to mean differences between the groups. Factor models for a homogeneous population are also a special case: they are LVMMs with a single latent class. In the anger example this would boil down to neglecting the presence of two subgroups, and attributing all covariances to one underlying anger factor within a single homogenous population.

The LVMM framework is extremely flexible, and permits the specification of different types of mixture models. Models such as path models, factor models, survival models, growth curve models, and more general structural equation models can all be specified for multiple subgroups instead of for a single homogeneous population (see for instance Arminger et al. 1999; Dolan & vd Maas, 1998; Jedidi et al., 1997; Muthén & Shedden, 1999; Muthén & Muthén, 2000; Ram & Grimm, 2009; Varriale & Vermunt, 2012; Vermunt 2008; Yung, 1997). The flexibility comes at a price. The framework is built on a set of assumptions that should be realistic for the data. Further, in order to estimate a model, all relations between observed variables, between observed variables and latent variables, and between latent variables have to be specified. It is therefore necessary to decide whether within-class model parameters are class specific, or are the same for all classes (i.e., class invariant). As will be discussed in this paper, the interpretation of the model depends on these decisions. It is important to note that different within-class parameterization can influence how many

classes best fit the data (Lubke & Neale, 2008). However, comparing a set of carefully parameterized mixture models can provide great insight into the processes and interrelations between variables when the assumption of population homogeneity is unrealistic.

The paper is organized into two main parts. The first part provides the theoretical background. After discussing the generally exploratory character of mixture analyses, the modeling framework is presented together with some of the necessary assumptions and constraints. The first part concludes with the discussion of issues that deserve consideration prior to fitting models to data, such as the interrelation between number of classes and within-class model complexity, measurement invariance, and models with and without covariates. The second part consists of a growth mixture analysis with covariates, and illustrates some of the practical issues discussed in the first part of the paper.

Part I: Exploration of Heterogeneity using Mixture Models

Latent variable mixture models (LVMMs) afford the possibility to detect groups of subjects in a sample, and to investigate the differences between the groups. LVMMs differ from other techniques to detect groups in data, such as taxometrics and cluster analysis, in that they require the user to specify all relations between observed and latent variables in the model (Meehl, 1995; Lubke & Miller, 2014). LVMMs are therefore prone to misspecifications. However, if there is sufficient *a priori* knowledge to specify these relations, then LVMMs usually have more power to detect groups in the data (Lubke & Tueller, 2010).

LVMMs differ from multi-group models in that it is not necessary to know which subject belongs to which group. Group membership is unobserved, or latent. Mixture models are therefore especially useful if the causes of the grouping are not known *a priori*. The grouping variable is formalized as a latent categorical variable, and the groups are called latent classes. In a cross-sectional setting, classes can consist of subjects with class-specific response profiles (e.g., high scores on some questionnaire items but low on others, or high on all), and in a longitudinal setting classes are characterized by class-specific trajectories over time (e.g., an increasing risk trajectory and a low constant trajectory).

If the process that causes the grouping is not well understood, then it is unlikely that the exact number of latent classes or the within-class structure are known. Mixture analyses are therefore often rather exploratory in character. Typically, a set of models with an increasing number of latent classes and different within-class structures is fitted to the data (e.g., more vs. less constrained models, see part 2, applied example). Model selection is based on measures such as the Bayesian Information Criterion (BIC) or the bootstrapped likelihood ratio test (Schwarz, 1978; McLachlan & Peel, 2000). Of course there is nothing wrong with exploratory analyses, quite the contrary. One can learn a lot from investigating heterogeneity, and such an analysis can be much more insightful about the structure in the data than incorrectly assuming that the data were sampled from a single homogeneous population. However, the exploratory character of a mixture analysis needs to be taken into account when best-fitting models are interpreted, and results need to be validated before specific conceptual conclusions concerning the class structure and within-class parameters can be drawn.

The Modeling Framework

This section provides a brief overview of the key aspects of the LVMM framework so that the practical challenges in an empirical mixture analysis, as illustrated in part 2 of the paper, can be fully appreciated.

Within the LVMM framework the population can consist of *k=1, …, K* latent classes. If *K=1*, then there is only a single class (i.e. a single homogeneous population). The *K=1* case therefore includes factor models, structural equation models, and growth models for a single homogeneous population. In case *K>1*, then a model needs to be specified for each of the classes. These within-class models are estimated jointly using a mixture distribution. A mixture distribution is a weighted sum of *K* component distributions, and is denoted as

$$f(Y) = \sum_{k=1}^{K} \pi_k f_k(Y;\theta_k) \quad \text{Eq(1)}$$

where *Y* is a vector of observed random variables, π_k is a weight that quantifies the relative size of the k^{th} component, and θ_k is a vector of model parameters for the k^{th} component (see McLachlan & Peel, 2000, for more detail on mixture distributions). The most common choice for the component distributions $f_k$is the multivariate normal distribution, although other distributions such as the Poisson distribution can be chosen to accommodate non-normal observed data (e.g., counts of cigarettes, etc.). In case each set of observed variables within class, Y_k, is multivariate normally distributed, we have $Y_k \sim MVN\left(\mu_k, \sum_k\right)$, where the parameter vector $\theta_k$contains the parameters that structure the component specific means, μ_k, , and covariance matrices, $\sum_k$

$$\mu_k = \nu_k + \Lambda_k(\mathrm{I} - \mathrm{B}_k^{-1}\alpha_k, \quad \text{Eq(2)}$$

$$\sum_k = \Lambda_k(\mathrm{I} - \mathrm{B}_k^{-1}\Psi_k[(\mathrm{I} - \mathrm{B}_k^{-1t}\Lambda_k^t + \Theta_k, \quad \text{Eq(3)}$$

where ν_k are the intercepts, Λ_k is the factor loading matrix, I is an identity matrix of appropriate dimensions, B_k contains regression coefficients of the regressions between factors, α_k are the factor means, $\Psi_k$the factor covariance matrix, and $\Theta_k$the covariance matrix of residuals. Basically, a full structural equation model can be specified within class, that is, Eq(2) and Eq(3) are the standard equations for structural equation models, and the reader is referred to textbooks such as Bollen (1989) for more detail on possible submodels.

At least some of these parameters can be constrained to be equal across classes. Parameter constraints and their interpretation are discussed later.

Mixture models with different numbers of classes cannot be compared with standard likelihood ratio tests since the test statistic does not have a known distribution (McLachlan

& Peel, 2000). Models have to be compared using information criteria such as the Bayesian Information Criterion (BIC) or bootstrapped likelihood ratio tests (Schwarz, 1978; McLachlan & Peel, 2000; Nylund et al., 2007).

Assumptions, necessary constraints, and sets of random starts

The LVMM framework inherits some of the assumptions of LCA models and structural equation models.

As in LCA, one assumes that each mixture component corresponds to a group or cluster of subjects. This is called the *direct interpretation* of mixture distributions (Titterington et al. 1985). However, it is important to realize that mixture distributions are not only useful to model clustered data, they can also be used to approximate distributions with an unknown functional form. For instance, skewed distributions can be approximated by a mixture of normal component distributions (Pearson, 1894, 1895). Of course, observing skewed data does not mean that there are necessarily distinct groups of subjects in the thin tail of the distribution. This is called the *indirect interpretation* of mixtures because the components do not necessarily correspond to meaningful clusters of subjects in the data (Titterington et al. 1985). Deciding whether the latent classes of a selected model represent meaningful groups of subjects is an important task of the researcher. Inclusion of covariates (e.g., age, sex, etc.), and class-predicted outcomes (e.g., adverse or beneficial outcomes such as dropout or improved health) can support the interpretation of the classes in terms of distinct groups. These and other means of validation are discussed in the section on validation methods.

As in structural equation modeling, the estimation of LVMMs is based on the assumption that the observed variables, Y, are linearly related to the factors. In addition, suitable distributions have to be chosen for the factors and the measurement errors, most commonly the multivariate normal distribution. Categorical observed data can be modeled either using a threshold model that treats Y as an unobserved continuous response variable that is partitioned into response categories using thresholds, or by replacing the linear regressions of observed variables on the factors by logistic or multinomial regression (Agresti, 2002; Muthén & Asparouhov, 2002). As mentioned above, other distributions for the observed variables can also be chosen (e.g., Poisson distribution for count data such as numbers of cigarettes per time interval).

Several constraints are necessary to identify the model. First, the weights, or proportions, of the component distributions have to sum to one to ensure the mixture distribution is a probability distribution, $\sum_{k=1}^{K}\pi_k=1$. Second, the latent factors within each class need to be scaled, that is, the mean and variance of each factor have to be identified. The constraints to identify the class-specific factor means are the same as in multi-group analysis, and most commonly consist of fixing one factor mean to zero, and estimating factor mean differences in the other classes (Sorbom, 1974). Constraints regarding the factor variance are the same as in confirmatory factor analysis, and consist of either fixing the variance to one, or fixing the factor loading of one of the items on the factor to one. In mixture analyses, that last option is preferable since it is often unrealistic that the factor variance is the same across classes. For instance, if a set of depression symptoms is observed in a sample from the

general population, the majority class will likely answer zero on most symptoms and the depression factor variance in that class will be close to zero. Conversely, the depression factor variance can be much larger in the affected group. Hence specifying equality of factor variance across classes may reflect an incorrect assumption about the data. Minimal constraints necessary to fit a model with categorical outcomes are discussed in detail in Millsap & Tein (2004).

When fitting mixture models it is necessary to use different sets of random starting values. If multiple sets of starting values lead to the same best fitting model the likelihood is said to have been replicated. However, replicating the likelihood is not sufficient to ensure a proper solution has been found. The number of random starts depends on model complexity and the quality of the data, and should be increased if the likelihood has not been replicated.

The number of classes and within class model complexity

The goals of fitting mixture models to empirical behavioral data are most commonly to establish the number classes and their relative size, and to evaluate the difference between the classes with respect to the means, variances, and other model parameters of interest. The following three paragraphs cover three issues that are important when designing an analysis that consists of comparing models with an increasing number of classes. These relate to the fact that the number of classes of the best-fitting models is affected by (1) the interdependence of the number of classes and the leniency of the within class model, (2) the statistical power to detect classes, and (3) the measurement properties of the items.

Interdependence of the number of classes and within class parameterization

Models with very constrained within-class parameterizations are likely to require more classes to fit the data than models that are more lenient within class (Lubke & Neale, 2006, 2008). Model leniency (or, reversely, model complexity) refers to the number of freely estimated parameters within class. An example of a constrained model is a latent class model, which constrains the within-class covariances between observed variables to zero. When fitting latent class models, the observed covariances in the total sample are attributed entirely to mean differences between the classes. If there are in fact considerable covariances *within class* in the data (e.g., there are severity differences within class), then these covariances will result in the need of additional classes to fit the data. This is called “overextraction of classes” because it would be more appropriate to fit models with fewer classes that permit observed variables to covary within class.

In factor and structural equation mixture models, observed variables are not modeled to be independent within class. Permitting within-class covariation takes care of part of the total observed covariances in the total sample, and as a result a smaller number of classes might suffice to explain the observed covariances in the total sample.

Another example of the interrelation between within-class complexity and number of classes concerns factor variances within class, for instance the variance of growth factors in a growth mixture model. Constraining the factor variances to be equal across classes might necessitate additional classes compared to models with class-specific factor variances. The

reason is again that part of the overall variance is not adequately accounted for by the within-class parameters. For instance, a majority class that stays at low levels of substance use over time might have very little variance in intercept and slope whereas subjects in a risk group might vary considerably. Constraining factor variances to be equal would lead to models with additional risk classes to capture the larger variance in the risk group. When exploring heterogeneity it is therefore useful to compare models with different levels of complexity within class as well as models with an increasing number of classes.

The trade-off between the number of classes and the within-class model complexity directly affects the results of model comparisons, because adding a class with only a few class-specific parameters can result in a more parsimonious model than estimating many class-specific parameters in a model with one class less. Model comparisons are evaluated with criteria that stress parsimony. Criteria such as the BIC are not infallible, however. Models with more classes and fewer class-specific parameters often have a quite different conceptual interpretation compared to models with fewer classes and more class specific parameters. Therefore the interpretation of model comparisons has to be cast in the context of this trade-off. This issue plays an especially important role when analyzing ordered categorical data, and when investigating measurement invariance (see next two paragraphs, and next section).

Power to detect classes

The power to detect classes depends to a large extent on the distance between the classes and on the sample size, especially the size of the smaller classes. This has been shown for relatively simple mixture models with simulated data (Lubke & Muthén, 2007; Tueller & Lubke, 2010). Low power due to smaller class separation can be compensated by a larger sample size. Mild misspecifications such as omitting cross-loadings do not seem to have a dramatic effect on the power to detect the true number of classes. When fitting more complex models it is important to realize that adding a class can require the estimation of a substantially larger number of parameters. Fit indices such as the BIC have a penalty for the number of parameters, and decisions based on the BIC might incorrectly be in favor of a model with fewer classes, especially when sample size is small. This issue is also relevant when analyzing ordinal items with class-specific thresholds, and measurement non-invariant models (see below). The sample size needed to detect classes in any given analysis depends on the characteristics of the data and the sample. Therefore, unfortunately, there is no good rule of thumb regarding sample size that is valid for different types of mixture analyses.

A small parametric bootstrap study can help to provide insight into the power to distinguish between alternative models (Muthén & Muthén, 2002). The recommendation to assess power is the same as in any non-mixture analysis, and it should therefore be regarded as an integral part of the analysis. It consists of four steps, namely (1) fitting a model of interest to the empirical data, (2) saving the parameter estimates, (3) generating multiple simulation data sets using the saved model parameters, and (4) fitting the alternative models to each set of simulated data to obtain the proportion of simulations that correctly select the true data-generating model. This can be repeated for multiple models of interest, and can aid the interpretation of the empirical results in the context of power. Bootstrap options are conveniently integrated in modeling software such as Mplus (Muthén & Muthén, 1998–

2012). Note however that in an empirical analysis the processes that lead to the observed data are often much more complex than the fitted models, and the power to discriminate between a set of fitted models that do not include the true data-generating model may be underestimated in such bootstrap simulations (Lubke et al., 2016).

Measurement properties of the observed items

The measurement properties of the observed items can have a substantial effect on whether or not a model with a correct number of classes is accepted as a best-fitting model in a model comparison (Lubke & Miller, 2014). Location, scale, and response format of the items need to be considered preferably prior to an analysis.

The item means of the observed variables (or location parameters in case of binary or ordinal items) should cover the whole range of the construct that is driving the classification. For instance, if the goal of an analysis is to detect different classes of a personality disorder, then the items need to cover the full range of the severity of the disorder. Limiting the set of observed items to those at the high end of the range will not permit distinguishing between groups of individuals at the lower end of the range (Lubke & Miller, 2014; Lubke & Spies, 2008). Item means are a function of the item's content; for instance, a question regarding the frequency of suicidal thought will discriminate mainly at the higher end of depression, and does not help to distinguish between unaffected individuals since these will most likely all answer zero. Therefore the item content of the observed items should be evaluated prior to an analysis to ensure that the items included in the analysis differ gradually with respect to the probability of being endorsed.

Simulation studies also suggest that the response format of the items has an impact on the number of classes in the best-fitting model (Lubke & Neale, 2008). Compared to normally distributed outcomes, analyses done with binary items show a decrease in power to detect classes. Using Likert items with multiple response categories can lead to accepting models with too few classes in case class-specific thresholds are estimated. In that case adding an additional class results in estimating a substantial number of additional parameters, which can lead to a higher BIC than a model with one class less. Although in practice the selection of items might be limited because a standard questionnaire is used, it is necessary to consider the potential impact of item format and item content when designing the analysis as well as when contextualizing the results.

Relevance of measurement invariance

Measurement invariance (MI) refers to the equality across groups of the parameters of the measurement model that relates observed variables to underlying latent variables (Mellenbergh, 1989; Meredith, 1993). The definition of MI is

$$f(Y;\eta) = f(Y;\eta, s), \quad \text{Eq(4)}$$

where Y is again a vector of observed random variables, η is a vector of latent variables, and s is a grouping variable. In case of mixture modeling the grouping variable, s, is latent. The

definition in Eq(4) means that the distribution of observed variables conditional on the latent variables η equals the distribution of observed variables conditional on η *and* the grouping variable. In other words, the grouping does not affect the measurement model that relates the observed variables Y to the latent variables η. Simply put, in case MI holds, the measurement model is the same in all groups. A common misunderstanding is that testing equality of the variance of the latent variables η is part of an investigation of MI. This is not the case since MI involves the distribution of observed variables conditional on η.

More specifically, in case Y is related to η in a factor model, then the parameters of the intercepts, loadings, and residual variances have to be the same across the groups. Since the definition in Eq(4) concerns the distribution of Y conditional on latent variables η the distribution of the latent variables η may differ across groups. MI can be established by comparing a set of models with different invariance constraints. In case of normally distributed outcomes within group, the set consists of (1) a model that does not constrain intercepts, loadings, and residual variances to be equal across groups but has factor means set to zero in all groups, (2) the same model but now with class-invariant loadings, (3) a model with class-invariant loadings and intercepts, and estimated factor means in $k-1$ of K groups, and (4) a full measurement invariant model which adds equal residual variances to model (3). The four models are fitted and compared using appropriate fit indices. With ordered categorical data, threshold and loading invariance is tested jointly (Muthén & Asparouhov, 2002; Millsap & Tein, 2004). A model with class-invariant loadings and thresholds and class-specific factor means and covariances is compared to a model with class-specific loadings and thresholds and zero factor means.

If model (4) fits the data well, one can conclude that the data do not provide evidence against MI. MI implies that the factors are measured in the same way in all groups, so the groups can be compared with respect to the factor means and factor covariances. Since the interpretation of latent variables hinges on how they are measured by the observed items, MI supports the same interpretation of the factors in all groups. If MI does not hold up in a model comparison, then the interpretation is more cumbersome because apparently the factors have a somewhat different interpretation across groups. In that case groups can only be compared with respect to the observed items. Models with class-specific factor loadings and/or intercepts imply that the factor structure differs across classes, and classes therefore differ *qualitatively*.

Investigating MI is a bit more complicated in mixture analyses than in multi-group analyses because groups are unobserved (i.e., the grouping variable s is latent). Testing MI in mixture models involves in principle the comparison of models (1)-(4). The main difference is that all models have to be compared for different numbers of classes. As a result, there are a number of potential complications.

First, as explained above, constraints on within-class parameters can often necessitate additional classes to account for the observed variability in the total sample. In the context of MI, this means that MI constraints can result in models with more classes. For example, a 4-class measurement invariant model might have a similar fit to a 2-class model without MI.

This is the reason why models with different MI constraints have to be compared with different numbers of classes.

Second, also mentioned above, in the case of ordered categorical outcomes (e.g., 5-point Likert items) an additional class with class-specific thresholds and loadings implies a quite large increase in the number of estimated parameters. Simulation studies have shown that the BIC can incorrectly favor MI models with an additional class over a much less parsimonious non-MI model with fewer classes (Lubke & Neale, 2008). The trade-off between adding a class and adding additional within-class parameters is not always easy to solve, and several alternative fitted models can have a similar BIC. In such a case validation can provide useful information to support model selection.

To summarize, the conceptual interpretation of non-MI models is different from that of MI-models because, in case MI holds, classes can be compared quantitatively with respect to the factors. When MI does not hold, classes should be compared with respect to the response profiles of the observed variables or other class-specific parameters of the distribution of Y. When fitting mixture models one has to carefully choose which parameters are constrained to be class-specific, and which are class-invariant, rather than relying on defaults in model fitting software. Choosing a set of alternative models for a model comparison and other practical issues are discussed in more detail in the next section.

Models with and without covariates

There is now a considerable body of literature concerning whether or not to include covariates when deciding on the number of classes (Asparouhov & Muthen, 2014; Kim, Vermunt, Bakk, Jaki, & Van Horn, 2016; Li & Hser, 2011; Nylund-Gibson & Masyn, 2016; Vermunt, 2010). Results of simulation studies seem to converge to the conclusion that is not necessary to include covariates to detect the correct number of classes. Nevertheless, parameters of models with and without covariates (including class proportions) can differ dramatically, thus leading to quite different conceptual interpretations. This occurs if covariates have direct effects on the observed variables within class over and above the effect that is mediated by class membership. The situation is similar to a scenario where a variable X (e.g., gender) has effects on two other variables (e.g., anger and depression). In that case the effect size of the association between anger and depression would depend on whether or not effects of gender on depression and anger are included in the analysis. In the context of mixture models, if covariates have an effect on the class variable and on observed variables within class, then the unconditional model (i.e., the model without the covariates) can have different class proportions compared to the conditional model (i.e., the model with covariates). This is due to the fact that the regression of the observed Y variables on the class variable C depends on whether or not covariates are included: the covariates partially “explain” the association between the Y and the class variable C. Stated otherwise, the grouping of subjects into classes depends on whether or not information contained in covariates is included in the analysis. Likewise, the conceptual interpretation of a conditional model is different from the interpretation of the unconditional model.

It is for the researcher to decide which model more appropriately reflects the research question at hand. As the example above illustrates, fitting a model with covariates addresses a different research question than a model without covariates. The question is whether a clustering of subjects is sought conditional on covariates (e.g., conditional on sex), or not.

From a statistical perspective, deciding on the number of classes can be done without covariates, and covariate effects can also be tested post-hoc, after a mixture model has been fitted to the data. For instance, in a mixture analysis of anger, one can test post-hoc whether the prevalence of males in a high scoring class is higher than the prevalence of females. Different ways to conduct such tests in a 3-step procedure are described in detail in (Asparouhov & Muthen, 2014; Masyn, 2016; Vermunt, 2010).

Model selection uncertainty

It is a very common strategy to fit additional models based on results of the previously fitted models. This clearly illustrates the exploratory character of mixture modeling. Even if the majority of the fitted models are planned, it is possible that additional models adapt more and more to sample-specific idiosyncrasies. Importantly, due to sampling fluctuation, this increases the uncertainty that the same model would be selected in a new sample.

Model selection uncertainty refers to the probability that different models would be selected as the “best-fitting model” in different samples from the same population. Especially in mixture analyses, model selection uncertainty can be substantial, and it is advisable not to treat the best-fitting model as if it were the only possible model for the data. Furthermore, it is important to realize that the p-values for parameter estimates in modeling software output do not take this uncertainty into account, and therefore do not reflect a 0.05 Type I error (Hurvich & Tsai, 1990; Lubke & Campbell, 2016). In fact, Type I error can be grossly inflated when model comparisons and tests of parameter significance are conducted in the same data. It is therefore advisable to split the sample into an exploratory and confirmatory subsample, and fit all models in the exploratory set. The selected model or models can then be fitted to the confirmatory subsample to obtain parameter estimates and their significance. If the original sample size is too small to split, it is preferable to limit the number of planned models. In that case the study needs to be reported as an exploratory analysis, and parameter significance cannot be considered.

In sum, it is good to plan the analysis beforehand, design the planned models according to pre-existing knowledge and the research questions at hand, and include models in the analysis plan that can refute hypotheses about the structure in the data. Most importantly, one should stick to the analysis plan without adding additional models that are based on intermediate results to avoid capitalization on chance.

Practical issues to consider before fitting mixture models

Exploratory data analysis prior to fitting models

Given the interdependence of the number of classes needed to fit the data and the within-class model constraints, it is advisable to gain as much information as possible about the

data before specifying and fitting models. Exploratory data analysis (Behrens, 1997; Tukey, 1977) is a very important step before conducting a quality mixture analysis. The sample data should be split into an exploratory and a confirmatory set such that all exploration and initial model fitting can be done in the exploratory set. The confirmatory set serves to fit the final models. Data splitting avoids compromising Type I error when assessing parameter significance in the selected final models (see also section on Validation). Summary statistics of the items, bivariate and trivariate plots, and other exploratory tools can all help to inform the subsequent model specifications. Although mixture distributions are not necessarily bi- or multimodal (McLachlan & Peel, 2000), it is useful to examine the observed item distributions for skewness, kurtosis, outliers, and in case of ordinal outcomes, a preponderance of zero's. A thorough data exploration provides an indispensable basis to understand the results of the comparison of a set of mixture models. It provides the basis to contextualize and interpret model-fitting results, and can also greatly facilitate the understanding of problems such as non-convergence of fitted models.

Designing increasingly lenient models

Similar to forward and backward stepwise approaches in regression, model comparisons can be done by starting with the most constrained model or by starting with the most lenient model. When comparing mixture models it is more practical to start with constrained models because estimation is much faster and the risk of non-convergence is usually smaller. A possible strategy is to first fit latent class models that constrain the within-class covariance matrix to be diagonal in accordance with the assumption of local independence (i.e., zero covariance within class). Comparing latent class models with an increasing number of classes will provide an approximate upper bound for the required number of classes. Gradually permitting more complexity within class such as including factors within class will usually result in the need for fewer classes. For instance, when fitting growth mixture models one can start with fixed effects models where the growth factors have zero variance and all within class variability is considered to be error (Nagin, 1999). Next, models with intercept random effects can be fitted, followed by models with both intercept and slope variance. This approach is illustrated for longitudinal data in Muthén & Muthén (2000).

A strategy of fitting gradually more lenient models also reveals at which point the specified models overfit the data, which can lead to large standard errors for some parameter estimates, improbable parameter estimates, or non-convergence. Although non-convergence can have different causes, knowing which additional estimated parameters resulted in non-convergence can often substantially narrow the potential sources of the problem. Non-converged models should always be reported along with the converged ones. For instance, non-convergence of models with class-specific loadings is not necessarily due to overfitting but might be due to insufficient sample size. When accepting a more constrained, converged model, the interpretation of the model comparison needs to take this possibility into account.

Before fitting a set of alternative models, it is necessary to consider which parameters need to be specified as class specific or class invariant in each of the models. As mentioned above, the substantial interpretation of a model hinges on which parameters are specified to be class

specific or class invariant. Therefore, all a priori knowledge about the structure in the data should be translated into an appropriate parameterization of the within class models.

Empirical underidentification

Careful consideration of the within-class parameterization is especially important when the expected size of one or more classes is small, because insufficient within class sample size can lead to empirical underidentification. This refers to a situation in which the fitted model is identified, but the sample data do not support the estimation of all model parameters. A classic example is a 2-factor model with two observed indicators for each of the factors. Although the model is identified, it is possible that in a given sample the correlation between the factors approaches zero. Fitting the 2-factor model to these data really means that two separate single factor models are estimated simultaneously. However, since single factor models with two indicators are not identified, this will result in problems. Empirical underidentification is not uncommon when the size of one or more classes is small relative to the number of class specific parameters. Constraints such as fixing loadings to be class invariant can help to stabilize the model estimation. Since different constraints affect model interpretation, the chosen constraints need to be explicitly reported.

Missing Data

Assumptions concerning missingness and appropriate ways of estimating models in the light of missingness are similar to non-mixture analysis. Specific to mixtures is the consideration whether missingness is expected to be the same across classes. This issue has been investigated in the context of growth mixture modeling where attrition over time a commonly observed pattern (Lu et al., 2011).

Validation of best-fitting model(s)

The validation of best fitting models can take different forms, and can include (1) comparison to previous results, (2) including additional variables in the analysis, and (3) validation in new data.

When comparing results of a model comparison to the results of previous studies, it is necessary to consider potential differences between the populations from which the samples were drawn, differences in sample size, and differences in the measurement instrument (if any). To illustrate the effect of differences between populations, consider measuring a psychiatric disorder in a general population sample vs. a clinical sample. Class solutions will not be easily comparable because the unaffected major cluster in the general population is missing in the clinical sample. Differences in sample size can affect the number of classes between the studies due to differences in power. Different measurement instruments can be differentially sensitive to detection of classes in the higher or lower ranges of the construct on which the sample is clustered (see paragraph on measurement properties). However, when differences between studies are appropriately taken into account, a comparison to previous results can be a very useful contribution to the validation of results.

Inclusion of covariates and class-predicted outcomes can validate the interpretation of the latent classes if expectations considering the impact of the covariates or class predicted

outcomes are formulated a priori. For instance, high school dropout can be an expected outcome of a smaller risk trajectory class but might not be expected for the majority class in a growth mixture analysis. Validating the class solution with secondary variables such as covariates and class predicted outcomes can help characterize the classes in more detail and can support a direct interpretation of the latent classes (see previous paragraph on covariates).

Validation in a new sample is usually the strongest type of validation. Having access to an additional sample from the same population provides the means to compare the class structure and the model parameter estimates. Obviously, collecting an additional sample might not be feasible due to monetary or other constraints. However, in light of increasing collaborations between research groups and efforts to make data publicly available, this type of validation will hopefully become more frequent. Splitting the available sample into an exploratory and a confirmatory set is a viable option with a sufficient sample size. Model selection can then be done in the exploratory set. The advantage of sample splitting is that fitting the selected model to the confirmatory set can be also be used for statistical testing of parameter significance. It avoids the inflated Type I errors that occur when carrying out model selection and statistical testing of parameters in a single sample (Lubke & Campbell, 2016; Lubke et al., 2016). Grimm et al. (2016) propose to use *k*-fold cross validation which also leads to correct Type-I error in the selected model. However, splitting the total sample into for instance 5 folds severely decreases the power to detect classes. Therefore this option is especially useful when a very large sample is available.

Summary of necessary steps

The following list summarizes the issues covered in Part I, and can be used as a guideline.

1. exploratory analysis (preferably in a partition of the data that is not used for significance testing): this includes considering which distribution to use for the observed data (counts, or normally distributed, or categorical?), assessing response probabilities on the observed items (especially for likert-type items), checking item correlations, selecting observed items for the analysis, fitting initial single class and mixture models to assess the need for multiple classes.
2. prepare and stick to an analysis plan: based on the exploratory analysis, design models that are likely supported by the data and that represent the research question(s) at hand. Then carry out the analysis plan. Any additional models that are fitted based on results of the planned models are post-hoc analyses, and should be reported as such. Note that the exploratory part usually takes considerably longer than designing and carrying out the main analysis!
3. report and interpret the results: most commonly the resolution in the data limit the complexity of the mixture models that can be fitted to the data. Non-convergence is therefor not uncommon for the more complex models. This should be reported, and included in the discussion. It is also important to consider that in a new sample results might look different. This so-called

sampling fluctuation is especially important to recognize in mixture modeling (for an example see Lubke & Campbell, 2016).

Part II: Illustration of growth mixture modeling using simulated data

This illustration concerns a longitudinal analysis using growth mixture models, and assumes the reader is familiar with linear and quadratic growth curve models for a single homogeneous population. For this illustration data were generated for an intermediate sample size (N = 1200) with 5 measurement occasions.

The data

The data file contains N = 1200 for a single outcome variable observed at 5 equally spaced time points. At each time point the observed variable is a categorical item scored 0, 1, or 2, which could for instance indicate alcohol consumption (0=never, 1=1–2 glasses per week, 2=more than 2 glasses per week). The data-generating model is a 2-class quadratic growth mixture model with 3 covariates predicting class membership. The 3 covariates were generated as binary variables, representing for example gender or minority ethnic group membership, and have differential effects predicting class membership as well as the random intercept in each class (i.e., class –specific covariate effects). Individuals scoring a 1 on the first covariate are more likely to be in class 1, whereas individuals who score a 1 on the second and third covariates are less likely to be in class 1. The covariates have a larger impact on the intercept factor in class 1. The data-generating model specifies expected class proportions of about 75% and 25%. Class 1 is the majority class, which starts at a low level and remains basically flat during the interval of the study due to a small positive linear slope and a small negative quadratic slope. Class 2 is the smaller class, which starts at a bit higher level and increases initially before declining again, with a large positive linear slope and moderate negative quadratic slope. This scenario could represent data concerning alcohol problems or substance use during adolescence, where a majority of the sample does not display these behaviors at any given time. However, there is a smaller risk group characterized by a quadratic average trajectory over the course of the study.

Research Questions

The three main goals of the mixture analysis are to explore (1) whether there are multiple subgroups in the population that differ with respect to their developmental trajectories, (2) whether it is worthwhile to allow for random effects within each group or whether variability around an average class trajectory is largely measurement error, and (3) whether the inclusion of covariates changes interpretation of the structure in the data.

Analysis plan

The crucial first step of the analysis is to randomly split the data into an exploratory set and confirmatory set before any data exploration takes place. For this analysis, the two sets will be of equal size, with N=600 each. Prior to fitting models, exploratory data analysis (EDA) is conducted in the exploratory set to observe response frequencies of the categorical dependent variables, the average response trajectory over time, and the frequencies of covariate responses.

To address the research questions, a set of increasingly complex models is designed, and then fitted to the exploratory part of the data. This forms the basis to select the most appropriate model or models, which are then fitted in the confirmatory data part in order to get parameter estimates with correct standard errors (Hurvich & Tsai, 1990; Lubke & Campbell, 2016). It is possible that at some level of complexity models become unstable, for instance, variance parameters can have large standard errors, or models do not converge. This can occur for different reasons, including that the model is overfitting the data (i.e., the model is more complex than the growth process in the population), or that the data do not provide sufficient resolution to estimate the parameters (e.g., small sample size, categorical data, missing data, etc.).

It is common practice to fit models without covariates first to establish the number of classes (Asparouhov & Muthen, 2014; Masyn, 2016). In this analysis we fit the same models with class membership predicted by the covariates to illustrate how inclusion of covariates can change class sizes and therefore interpretation of results. We also fit a set of models that include class-specific covariate effects on the latent intercept factor in addition to covariate prediction of class membership. Model comparison will only be carried out on these models fitted to the exploratory set of the data. A very small subset of the best-fitting models from the exploration (1–3 models) is fitted in the confirmatory set for model inference and interpretation.

Models without covariates—Models without covariates are unconditional models (i.e., they are estimated without conditioning on covariates). The following models without covariates will be fitted:

> Models 1–4: quadratic growth mixture model with 2, 3, 4, and 5 classes, respectively; no variances or covariances for the growth factors.
>
> Models 5–7: quadratic growth mixture model with 2, 3, and 4 classes; random intercept factor within class.
>
> Models 8–10: quadratic growth mixture model with 2, 3, and 4 classes; random intercept and linear slope factors within class, and intercept-slope covariance.
>
> Models 11–12: quadratic growth mixture model with 2, and 3 classes; random intercept, linear slope, and quadratic slope factors, and their covariances, within class.

Note that the number of classes decreases as the model parameterizations become more complex. As explained previously, it is expected that fewer classes are needed to explain to the total covariance in the data as within-class models increase in complexity.

Models with covariates—To determine if covariates impact model selection and/or class membership estimates, the following models with covariates (i.e., conditional models) will be fitted:

> Conditional models 1–12: as in models 1–12 without covariates, with class membership predicted by the covariates.

Conditional models 13–15: quadratic growth mixture models with 2, 3, and 4 classes; class membership predicted by the covariates and class-specific covariate effects on the intercept factor; no linear or quadratic slope variances (only the residual variance of the intercept regressed on the covariates estimated).

Conditional models 16–18: as in models 13–15, but with random linear slope factor and intercept-slope factor covariance.

Conditional models 19–20: as in models 16–17 (2 and 3 classes only), but with random linear and quadratic slopes and growth factor covariances.

The conditional models 13–20 with the additional effect of the covariates on the intercept factor could represent a case where the covariate value is expected to impact class membership as well as the baseline level in each class. In the example of adolescent substance use, a researcher may expect that gender could significantly differentiate the classes, and within each class, males may have a higher baseline level than females.

Model fitting

Models were fitted using *Mplus 7.4* (Muthén & Muthén, 1998–2015). Models were initially fitted with 500 initial random starts and the best 50 carried out to the default convergence. If the best log-likelihood value was not replicated, the random starts were increased to 2000 with the best 200 iterated to convergence, respectively. The relative fits of the models were compared using BIC, and models were also checked for proper convergence. We do not advise using the Lo-Mendell-Rubin test since it has been criticized to rely on incorrect assumptions (Jeffries, 2003). The bootstrapped likelihood ratio test might also not be an ideal choice since classes resulting from fitting say Model A do not contain exactly the same subjects (i.e., the same grouping) ad those resulting from fitting Model B. Likelihood ratio testing, however, relies on identical groups when comparing multi-group (or multiclass) models. Of course no index for model comparisons is flawless, and BIC for instance only selects the correct model asymptotically (i.e. when sample size approaches infinity). Models were deemed properly converged if estimation converged properly, parameters estimates were acceptable (e.g., no variance estimates were negative), and standard errors were reasonably small.

Results

Exploratory Data Analysis

Descriptive statistics of the observed variables and covariates in the exploratory data are presented in Table 1. The responses are seen to have a preponderance of zeros, although responses on ‘1’ and ‘2’ tend to increase from time 1 to time 5. The covariates have 0/1 response proportions of 48/52%, 64/36%, and 68/32%. Plotting the mean responses at each time point reveals what appears to be a quadratic trajectory. However, due to the zero-inflated responses, we would not expect all individuals to experience initial growth, thus informing our choice to fit quadratic growth mixture models.

Models without covariates

The fit statistics for the 12 models without covariates are presented in Table 2. Though AIC and sample-size adjusted BIC are also presented, the model comparison is based on BIC as the fit criteria of choice. The three best-fitting unconditional models are models 5, 6 and 8. Model 5 is the two-class model with random intercepts, model 6 is the three-class model with random intercepts, and model 8 is the two-class model with random intercepts and linear slopes. Overall, model 5 fits the best, with models 6 and 8 having nearly identical BICs. The pattern of models indicate that model fit greatly improves when random intercepts are introduced, but then gets worse as random linear and quadratic slopes are introduced, and as more classes are estimated with these random slopes.

Additionally, model 7 (four-class model with random intercepts) and model 11 (two classes with random intercepts, linear slopes, and quadratic slopes) do not converge due to a non-positive definite Fisher Information matrix. Therefore, fit statistics are not obtained. Models 10 and 12 report fit statistics, but register an error with the first-order derivative product matrix. This error can be due to starting values, scaling issues such as inflated thresholds, or unidentified models. This error also registers the parameter at the source of the problem. In this case, random slopes are the problem parameters for both models 10 and 12, indicating that slope variance cannot be reliably estimated in these data.

Models with covariates

Next, the models conditional on the covariates were fitted in the exploratory data. The fit statistics and model summaries for these 20 models are presented in Table 3. The best-fitting model according to BIC in these conditional models is model 13, which is the two-class model with class membership predicted by the covariates and within-class effects of the covariates on the intercept factor. As this is a regression of the intercept factor on the covariates, a residual variance for the intercept is estimated. The second-best-fitting model is model 16, which is the same as model 13 but with random linear slopes. Models 12 and 20, the 3-class models including variances and covariances for all growth factors, had a non-positive definite Fisher Information matrix, giving an indication that random effects for all of the growth factors in 3 classes may be overfitting the data. Models 18 and 19 have non-positive derivative matrices, and identify random slopes as the problem parameter. Random linear slopes in 4 classes (model 18) and random quadratic slopes (model 19) also appear to overfit these data.

Additionally, several models registered an error with the first-order derivative product matrix. Models 5, 6, 9, and 11 identified the regression of a covariate on the class variable as a potential source of the problem. Further inspection revealed that these estimates were quite inflated, with logistic regression coefficients in the range of 12–14 (these were generated around values of 1 and 2). These estimates correspond to improbable odds ratios of class membership. Models 13–17, however, account for the additional class-specific effects of the covariates on the intercept term, and these models generally display better fit to the data. The unaccounted conditional effects of the covariates on the intercept could potentially inflate the prediction of class membership and therefore contribute to the problems in models 5, 6, 9, and 11.

Choosing confirmatory models

The fit criteria of the conditional and unconditional models cannot be compared directly because they use different data, but the choice between the two modeling approaches for making inference is very important. For example, observing the estimated class probabilities of the unconditional model 5 and the conditional model 13 (for comparison, not for parameter inference), it is seen that the unconditional model 5 estimated class proportions are roughly 52% for class 1 and 48% for class 2, whereas the conditional model 13 has estimated class proportions of 76% and 24%, respectively. The unconditional model 5 is characterized by a slightly larger class that starts at a low level, declines with a negative linear slope, and accelerates in its decline due to a negative quadratic slope. The slightly smaller has the quadratic growth trajectory that was observed in the exploratory data analysis. Both classes have very large intercept factor variances. Conditional model 13 contains a majority class with a low initial level, small positive linear slope and small negative quadratic slope. The much smaller class in conditional model 13 starts a bit higher, and has a larger positive linear slope and larger negative quadratic slope compared to the majority class.

In the exploratory model comparison, the two-class models with random intercepts were generally the best-fitting types of models, with random linear slopes being potentially important for the data. In the conditional modeling, it appeared that adding the covariates predicting both class membership and the within-class intercept factor stabilized the models, as models 13, 14, and 16 fit better than any models without the conditional intercepts. Additionally, a work-around for comparing the conditional and unconditional models is including the covariates in the unconditional models with coefficients fixed at zero; this resulted in a BIC of 4022.13 for model 5, compared to the BIC of 3614.37 and 3636.26 for the conditional models 13 and 16, respectively. Therefore, it was decided to fit models 13 and 16 in the confirmatory data.

Confirmatory models

Models 13 and 16 with covariates included were fitted in the confirmatory data. Model 13 (BIC = 3535.49, AIC = 3447.55) demonstrated better fit to the data than model 16 (BIC = 3558.13, AIC = 3452.60), although their parameter estimates were generally fairly similar. The only difference between model 13 and model 16 is the inclusion of within-class linear slope variances and intercept-slope covariances. These estimates are not significant in model 16, which is not unexpected given that model 13 displays better fit.

Model 13 estimated a majority class of 72.5% and smaller class of 27.5%, and model 16 estimated a similar 73/27% split. Selected parameter estimates from each model are presented in Table 4. Both models estimate a majority class that starts at low levels and has a significant but small linear increase, followed by a decline towards the initial level due to the negative quadratic slope. The baseline level, or intercept factor, varies significantly depending on X1 and X3, but not X2. The smaller class starts a bit higher and is characterized by a much more pronounced quadratic curve, with a larger positive linear slope and a negative quadratic slope. The intercept factor is significantly predicted by all 3 covariates in this class. Membership in class 1 is negatively predicted by X2 and X3.

Drawing inference on these final models, it appears these data are composed of 2 latent classes, a majority class of about 72.5% and smaller class of 27.5%. The average trajectories of the two classes, computed from class probabilities and class-specific parameter estimates, are presented in Figure 1. The majority class stays within a very small range over time, on average, even though the slope factor means are significantly different from zero. For a study of adolescent alcohol use, this class represents the majority who remain at relatively lows levels of consumption throughout adolescence. Within this class, those who score a 1 on X1 and X3 (say, males and adolescents who started drinking before the study) would have a higher baseline level of drinking over time compared to, say, females and adolescents who did not drink at a young age. However, early-onset drinkers remain less likely to be in this class, which is why the average trajectory remains lower than class 2.

Class 2, the smaller class, starts a bit higher and experiences much more increase over the course of the study. This class represents a smaller group of adolescents at risk for higher levels of substance use, with a large up-tick in use in early adolescence, followed by a deceleration and slight decline. However, these adolescents remain at relatively high risk over the 5 measurement occasions of this hypothetical study. In class 2, the baseline level is also higher for males and early drinkers, but it is significantly reduced for those with X2 (e.g., adolescents in a minority ethnic group). Within each class, the average linear and quadratic slopes best describe the change process over time, as seen by the best model fixing the slope variances to zero.

Part II Summary

The overall results address the main research questions highlighted at the outset of the data analysis. There are two subgroups in the population that have meaningfully different trajectories. The intercept factor, or baseline level, varies across individuals, but fixing the linear and quadratic slopes is adequate for these data. Including the covariates in the mixture model drastically changes the interpretation of the data structure. Similar conditional and unconditional models differed in the estimated class proportions and the trajectories of the classes. Furthermore, the covariates explained some of the variability in the intercept factors within each class. Including the covariates was crucial in this analysis.

Part II of this paper demonstrated how to systematically fit GMMs, taking into account the theoretical and practical issues discussed in part I. First, research questions were specified and an analysis plan was created to answer these questions. We followed the analysis plan, detailed the exploratory nature of our model comparisons in the data, and justified choices for subsequent steps of model fitting. We randomly split our data into exploratory and confirmatory subsets, which allows researchers to make inferences after exploratory model comparisons with controlled type I errors and valid standard errors for parameter estimates (Lubke & Campbell, 2016; Lubke et al., 2016). Following this procedure led to best-fitting models that were indeed congruent with the data-generating model.

Conclusion

The mixture modeling framework is largely an exploratory device. A number of assumptions and constraints are necessary to fit mixture models to data. These assumptions need to be realistic for a given data set, and should correspond to existing knowledge about the data. Apart from assumptions that are inherited from structural equation modeling (e.g. distributional assumptions, linear relations), the assumption that each mixture component corresponds to a meaningful group in the sample (i.e., the direct interpretation of mixtures) needs to be considered when interpreting results of an analysis.

A mixture modeling analysis usually consists of comparing models with an increasing number of classes and different within-class parameterizations. Importantly, each model consists of a system of equations relating the observed variables to continuous and categorical latent variables and their interrelations, thus representing a combined hypothesis that is evaluated by fitting the model to the data. Comparing multiple models therefore equates to comparing numerous alternative combined hypotheses. Fitting a sequence of models in which each additional model depends on the result of the previous models is not advisable because it is likely that the models can adapt more and more to sample specific aspects, thereby reducing the likelihood of replicating the results in a new sample. To avoid this potential capitalization on chance, it is good practice to split the sample into an exploratory and confirmatory part, and design an analysis plan before fitting models to the data. Clarity regarding the number of fitted models and the model-fitting strategy is essential when reporting results of a mixture analysis. Non-converged (but potentially more interesting) models should also be reported such that the reader can evaluate the results of the full set of models that were fitted to the data.

Due to the potential that a different model can get selected in a different sample from the same population (i.e., model selection uncertainty), it is also good practice to avoid focusing on a single best-fitting model (Lubke & Campbell, 2016, Lubke et al., 2016). As illustrated in Part 2, considering a small number of best-fitting models can be highly informative.

In sum, LVMMs are a complex and sophisticated device to investigate population heterogeneity. They require careful consideration of the within-class parameterization, and should include some form of validation. Though LVMMs are complex tools that can at times be tedious to fit, carefully evaluating a set of LVMMs can provide much more insight into the structure of the data than models that ignore potential population heterogeneity.

References

Agresti, A. Categorical data analysis. 2nd. New York, NY: Wiley-Interscience; 2002.

Arminger G, Stein P, Wittenberg J. Mixtures of conditional mean- and covariance-structure models. Psychometrika. 1999; 64(4):475–494. DOI: 10.1007/bf02294568

Asparouhov, T., Muthén, BO. Multilevel mixture models. In: Hancock, GR., Samuelsen, KM., editors. Advances in latent variable mixture models. Charlotte, NC: Information Age Publishing; 2008. p. 27-52.

Asparouhov T, Muthén BO. Auxiliary variables in mixture modeling: Three-Step approaches using M plus. Structural Equation Modeling: A Multidisciplinary Journal. 2014; 21(3):329–341. DOI: 10.1080/10705511.2014.915181

Behrens JT. Principles and procedures of exploratory data analysis. Psychological Methods. 1997; 2(2):131–160. DOI: 10.1037//1082-989x.2.2.131

Bollen, KA. Structural equations with latent variables. New York: Wiley; 1989.

Dolan CV, van der Maas HLJ. Fitting multivariage normal finite mixtures subject to structural equation modeling. Psychometrika. 1998; 63(3):227–253. DOI: 10.1007/bf02294853

Grimm KJ, Mazza GL, Davoudzadeh P. Model selection in finite mixture models: A k-fold cross-validation approach. Structural Equation Modeling: A Multidisciplinary Journal. 2017; 24(2):246–256.

Jedidi K, Jagpal HS, DeSarbo WS. STEMM: A general finite mixture structural equation model. Journal of Classification. 1997; 14(1):23–50. DOI: 10.1007/s003579900002

Jeffries NO. A note on 'Testing the number of components in a normal mixture'. Biometrika. 2003; 90(4):991–994.

Kim M, Vermunt J, Bakk Z, Jaki T, Van Horn ML. Modeling predictors of latent classes in regression mixture models. Structural Equation Modeling: A Multidisciplinary Journal. 2016; 23(4):601–614.

Lanza, ST., Flaherty, BP., Collins, LM. Latent class and latent transition analysis. In: Schinka, JA., Velicer, WF., editors. Handbook of psychology: Research methods in psychology. Hoboken, NJ: Wiley; 2003. p. 663-685.

Lazarsfeld, PF., Henry, NW. Latent structure analysis. New York: Houghton-Mifflin; 1968.

Li L, Hser YI. On inclusion of covariates for class enumeration of growth mixture models. Multivariate Behavioral Research. 2011; 46:266–302. DOI: 10.1080/00273171.2011.556549 [PubMed: 23904664]

Lu ZL, Zhang Z, Lubke GH. Bayesian inference for growth mixture models with latent class dependent missing data. Multivariate behavioral research. 2011; 46(4):567–597. [PubMed: 24790248]

Lubke GH, Miller PJ. Does nature have joints worth carving? A discussion of taxometrics, model-based clustering and latent variable mixture modeling. Psychological Medicine. 2014; 45(04):705–715. DOI: 10.1017/s003329171400169x [PubMed: 25137654]

Lubke GH, Campbell I, Luningham J, McArtor DB, Miller PJ, Van den Berg SM. Assessing model selection uncertainty using a bootstrap approach: An update. Structural Equation Modeling: A Multidisciplinary Journal. (In press).

Lubke GH, Dolan CV, Kelderman H, Mellenbergh GJ. On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. Intelligence. 2003; 31(6):543–566. DOI: 10.1016/s0160-2896(03)00051-5

Lubke GH, Neale MC. Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood. Multivariate Behavioral Research. 2006; 41(4):499–532. DOI: 10.1207/s15327906mbr4104_4 [PubMed: 26794916]

Lubke GH, Neale MC. Distinguishing between latent classes and continuous factors with categorical outcomes: Class Invariance of parameters of factor mixture models. Multivariate Behavioral Research. 2008; 43(4):592–620. DOI: 10.1080/00273170802490673 [PubMed: 20165736]

Lubke GH, Muthén BO. Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. Structural Equation Modeling: A Multidisciplinary Journal. 2007; 14(1):26–47. DOI: 10.1080/10705510709336735

Lubke, GH., Spies, JR. Choosing a "correct" factor mixture model: Power, limitations, and graphical data exploration. In: Hancock, GR., Samuelson, KM., editors. Advances in latent variable mixture models. Charlotte, NC: Information Age Publishing; 2008. p. 343-362.

Lubke GH, Tueller S. Latent class detection and class assignment: A comparison of the MAXEIG Taxometric procedure and factor mixture modeling approaches. Structural Equation Modeling: A Multidisciplinary Journal. 2010; 17(4):605–628. DOI: 10.1080/10705511.2010.510050 [PubMed: 24648712]

Masyn KE. Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. Structural Equation Modeling: A Multidisciplinary Journal. (In press).

McCutcheon, AL. Quantitative Applications in the Social Sciences Series. Vol. 64. Thousand Oaks, CA: Sage; 1987. Latent class analysis.

McLachlan, GJ., Peel, D. Finite mixture models. New York: Wiley; 2000.

Meehl PE. Bootstraps taxometrics: Solving the classification problem in psychopathology. American Psychologist. 1995; 50(4):266–275. DOI: 10.1037//0003-066x.50.4.266 [PubMed: 7733538]

Mellenbergh GJ. Item bias and item response theory. International Journal of Educational Research. 1989; 13(2):127–143. DOI: 10.1016/0883-0355(89)90002-5

Meredith W. Measurement invariance, factor analysis and factorial invariance. Psychometrika. 1993; 58(4):525–543. DOI: 10.1007/bf02294825

Millsap RE, Yun-Tein J. Assessing Factorial Invariance in ordered-categorical measures. Multivariate Behavioral Research. 2004; 39(3):479–515. DOI: 10.1207/s15327906mbr3903_4

Muthén, BO. Latent variable mixture modeling. In: Marcoulides, GA., Schumacker, RE., editors. New developments and techniques in structural equation modeling. Mahwah, NJ: Lawrence Erlbaum Associates; 2001. p. 1-33.

Muthén BO, Muthén LK. Integrating person-centered and variable-centered analysis: Growth mixture modeling with latent trajectory classes. Alcoholism: Clinical and Experimental Research. 2000; 24(6):882–891.

Muthén LK, Muthén BO. How to use a Monte Carlo Study to decide on sample size and determine power. Structural Equation Modeling: A Multidisciplinary Journal. 2002; 9(4):599–620. DOI: 10.1207/s15328007sem0904_8

Muthén B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics. 1999; 55(2):463–469. DOI: 10.1111/j.0006-341x.1999.00463.x [PubMed: 11318201]

Muthén, LK., Muthén, BO. Mplus user's guide. 7th. Los Angeles, CA: Muthén & Muthén; 1998–2015.

Muthén, BO., Asparouhov, T. Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. 2002. Retrieved from https://www.statmodel.com/download/webnotes/CatMGLong.pdf

Nagin DS. Analyzing developmental trajectories: A semiparametric, group-based approach. Psychological Methods. 1999; 4(2):139–157. DOI: 10.1037//1082-989x.4.2.139

Nylund KL, Asparouhov T, Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural Equation Modeling: A Multidisciplinary Journal. 2007; 14(4):535–569. DOI: 10.1080/10705510701575396

Nylund-Gibson K, Masyn K. Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. Structural Equation Modeling: A Multidisciplinary Journal. 2016; 23(6):782–797. http://dx.doi.org.proxy.library.nd.edu/10.1080/10705511.2016.1221313.

Pearson K. Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London (A). 1894; 185:71–110. DOI: 10.1098/rsta.1894.0003

Pearson K. Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. Philosophical Transactions of the Royal Society of London (A). 1895; 186:343–414. DOI: 10.1098/rsta.1895.0010

Ram N, Grimm KJ. Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. International Journal of Behavioral Development. 2009; 33(6):565–576. DOI: 10.1177/0165025409343765 [PubMed: 23885133]

Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978; 6(2):461–464. DOI: 10.1214/aos/1176344136

Sörbom D. A general method for studying differences in factor means and factor structure between groups. British Journal of Mathematical and Statistical Psychology. 1974; 27(2):229–239. DOI: 10.1111/j.2044-8317.1974.tb00543.x

Titterington, DM., Smith, AFM., Makov, UE. Statistical analysis of finite mixture distributions. New York: Wiley; 1985.

Tueller S, Lubke GH. Evaluation of structural equation mixture models: Parameter estimates and correct class assignment. Structural Equation Modeling: A Multidisciplinary Journal. 2010; 17(2):165–192. DOI: 10.1080/10705511003659318 [PubMed: 20582328]

Tukey, JW. Exploratory Data Analysis. Phillipines: Addison-Wesley Pub. Co; 1977.

Vermunt JK. Latent class and finite mixture models for multilevel data sets. Statistical Methods in Medical Research. 2008; 17:33–51. [PubMed: 17855746]

Vermunt JK. Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. Political Analysis. 2010; 18:450–469.

Varriale R, Y Vermunt JK. Multilevel mixture factor models. Multivariate Behavioral Research. 2012; 47(2):247–275. [PubMed: 26734850]

Yung YF. Finite mixtures in confirmatory factor-analysis models. Psychometrika. 1997; 62(3):297–330. DOI: 10.1007/bf02294554

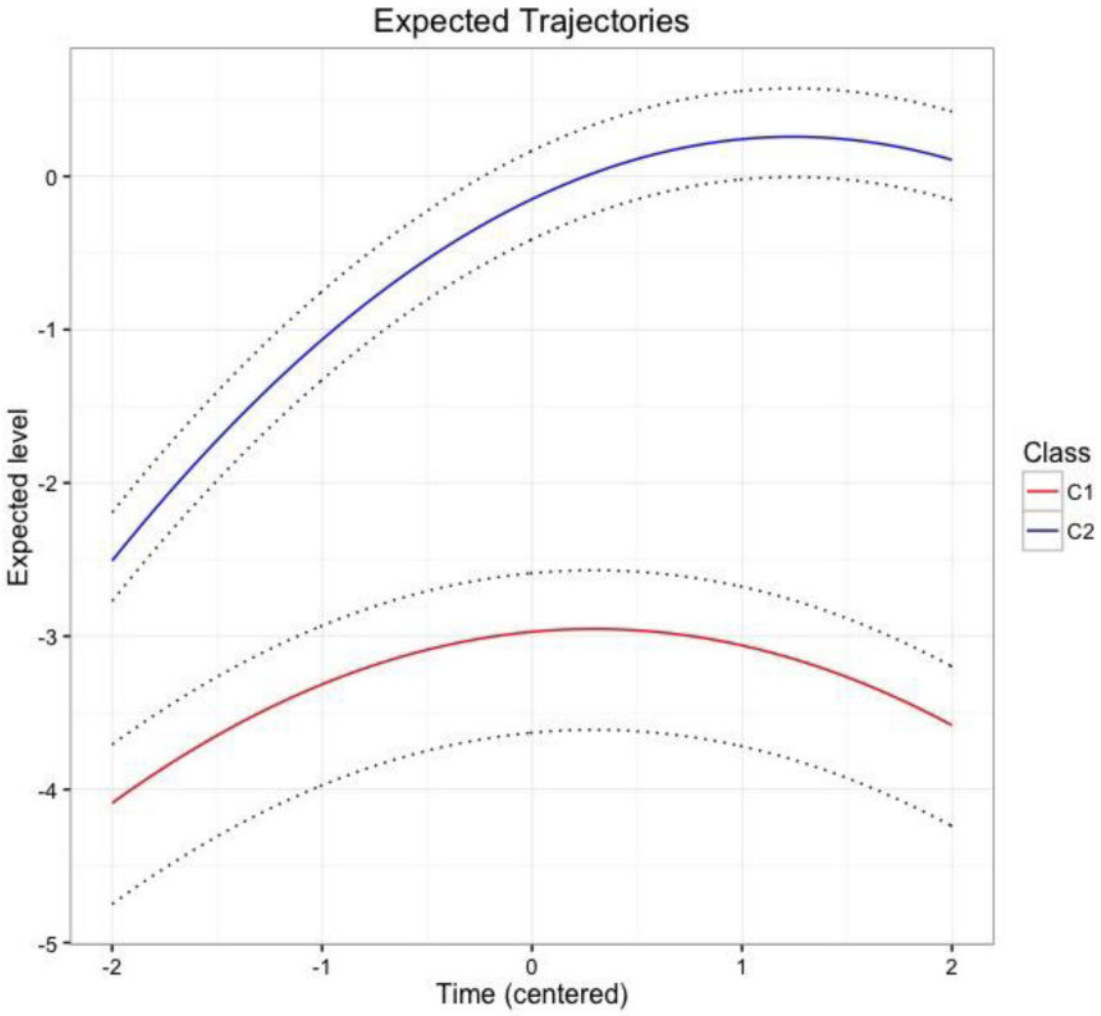

Figure 1.

Model-implied trajectories from fitting conditional model 13 in the confirmatory data. The red line represents class 1 and the blue line represents class 2. The dotted lines capture variability in the intercept factor due to the covariates, and are the expected trajectories at the extreme values of all covariates (i.e., all zeros or all ones on the three covariates).

Table 1

Summary statistics for simulated data categorical data.

Response	T1	T2	T3	T4	T5	X1	X2	X3
0	0.772	0.640	0.580	0.583	0.655	0.478	0.637	0.682
1	0.140	0.220	0.220	0.230	0.158	0.522	0.363	0.318
2	0.088	0.140	0.200	0.190	0.187	*NA*	*NA*	*NA*

T1–T5 indicates the 5 longitudinal time points; X1–X3 indicates the 3 binary predictor variables.

Table 2

Model summary and fit statistics for unconditional models (not including covariates) in the exploratory data. Italicized fit statistics indicate models that obtained information criteria, but may have untrustworthy standard errors due to non-positive definite 1st-order product derivative matrix. The problem parameter is indicated. Blank entries indicate models with non-positive definite Fisher Information matrix, for which information criteria cannot be computed. The three best BIC values are bolded.

Model	*n* class	*n* param.	Log−likelihood	AIC	BIC	aBIC	Fisher Info. Error?	1st-order Deriv. Error? (Bad param.)
m1 (unconditional)	2	9	−2082.16	4182.33	4221.90	4193.33	No	No
m2 (unconditional)	3	13	−1998.73	4023.45	4080.61	4039.34	No	No
m3 (unconditional)	4	17	−1975.34	3984.68	4059.43	4005.46	No	No
m4 (unconditional)	5	21	−1966.58	3975.17	4067.49	4000.83	No	No
m5 (unconditional)	2	11	−1972.54	3967.08	**4015.45**	3980.52	No	No
m6 (unconditional)	3	16	−1963.98	3959.95	**4030.30**	3979.51	No	No
m7 (unconditional)	4	21					Yes	
m8 (unconditional)	2	15	−1967.75	3965.5	**4031.45**	3983.83	No	No
m9 (unconditional)	3	22	−1963.02	3970.03	4066.76	3996.92	No	No
m10 (unconditional)	4	29	*−1960.47*	*3978.95*	*4106.46*	*4014.39*	No	Yes (slope variance)
m11 (unconditional)	2						Yes	
m12 (unconditional)	3	31	*−1959.74*	*3981.47*	*4117.78*	*4019.36*	No	Yes (quadratic variance)

Note. param. = parameter; AIC = Akaike information criterion; BIC = Bayesian information criterion; aBIC = sample-size adjusted BIC.

Table 3

Model summary and fit statistics for conditional models (including covariates) in the exploratory data. Italicized fit statistics indicate models that obtained information criteria, but may have untrustworthy standard errors due to non-positive definite 1st-order product derivative matrix. The problem parameter is indicated. Blank entries indicate models with non-positive definite Fisher Information matrix, for which information criteria cannot be computed. The two best BIC values are bolded.

Model	*n* class	Params	Log-likelihood	AIC	BIC	aBIC	Fisher Info. Error?	1st-order Deriv. Error? (Bad param.)
m1 (conditional)	2	12	−1917.08	3858.15	3910.91	3872.82	No	No
m2 (conditional)	3	19	−1799.51	3637.01	3720.55	3660.23	No	No
m3 (conditional)	4	26	−1763.91	3579.82	3694.14	3611.60	No	No
m4 (conditional)	5	33	−1748.29	3562.58	3707.68	3602.91	No	No
m5 (conditional)	2	14	*−1815.94*	*3659.88*	*3721.43*	*3676.99*	No	Yes (class 1 on X3)
m6 (conditional)	3	22	*−1763.59*	*3571.18*	*3667.91*	*3598.06*	No	Yes (class 1 on X3)
m7 (conditional)	4	30	−1743.59	3547.18	3679.09	3583.84	No	No
m8 (conditional)	2	18	−1804.74	3645.48	3724.63	3667.48	No	No
m9 (conditional)	3	28	*−1752.94*	*3561.89*	*3685.00*	*3596.11*	No	Yes (class 2 on X3)
m10 (conditional)	4	38	−1738.58	3553.16	3720.24	3599.61	No	No
m11 (conditional)	2	24	*−1801.65*	*3651.30*	*3756.82*	*3680.63*	No	Yes (class 1 on X3)
m12 (conditional)	3						Yes	
m13 (conditional)	2	20	−1743.22	3526.43	**3614.37**	3550.88	No	No
m14 (conditional)	3	31	−1731.19	3524.37	**3660.68**	3562.26	No	No
m15 (conditional)	4	42	−1721.62	3527.24	3711.91	3578.57	No	No
m16 (conditional)	2	24	−1741.37	3530.74	**3636.26**	3560.07	No	No
m17 (conditional)	3	37	−1727.67	3529.35	3692.03	3574.57	No	No
m18 (conditional)	4	50	*−1716.44*	*3532.89*	*3752.73*	*3593.99*	No	Yes (slope variance)
m19 (conditional)	2	30	*−1740.44*	*3540.88*	*3672.79*	*3577.54*	No	Yes (quadratic variance)
m20 (conditional)	3	46					Yes	

Note. param. = parameter; AIC = Akaike information criterion; BIC = Bayesian information criterion; aBIC = sample-size adjusted BIC.

Table 4

Key parameters of the models fitted to the confirmatory data.

Model	Class	Intercept on			Means/β_0			Class 1 on	
		X1	X2	X3	β_0	Linear	Quadratic	X1	X2	X3
13	C1	2.618***	−3.788^{+}	2.756***	−5.160***	0.127*	−0.216***	−0.533	−3.080***	−3.043***
	C2	2.746***	−3.448***	3.163***	−1.619^{+}	0.654***	−0.263***			
16	C1	2.638***	−3.917^{+}	2.810***	−2.850***	0.143***	−0.215***	−0.544	−3.056***	−2.914***
	C2	2.755***	−3.740***	3.279***	0.945	0.701***	−0.288***			

β_0 is the intercept term from the regression of the intercept factor on the covariates, i.e., the mean of the intercept factor when the covariates are all zero.

*** = $p < 0.001$,

* = $p < .05$,

$^{+}$ = $p < 0.1$.