USING MACHINE LEARNING TO UNCOVER POPULATION HETEROGENEITY IN LONGITUDINAL STUDY

By

Youngjun Lee

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods—Doctor of Philosophy

2022

ABSTRACT

USING MACHINE LEARNING TO UNCOVER POPULATION HETEROGENEITY IN LONGITUDINAL STUDY

By

Youngjun Lee

Machine learning has been an emerging data analytic tool in the fields of quantitative social and behavioral sciences. Among others, model-based recursive partitioning (MOB) is one of the popular comprehensive approaches incorporating parametric model into tree-based algorithm. It has gained growing interests as a complementary data analytic tool to address population heterogeneity by detecting parameter instability over candidate covariates. Structural equation models using tree algorithm (SEM Trees) has particularly shown its benefits for discovering informative covariates and their complex interactions that predict differences in structural parameters with interpretable results, which in turn produces distinct homogeneous subgroups. While all previous studies make important contributions to use this approach, it has been less examined to investigate the performance of SEM Trees where there exist interaction effects of various types of covariates (i.e., categorical, ordinal, and continuous), which is the key motivation of this study.

This study has three main purposes. First, it aims to introduce a framework of MOB for educational researchers and guide them when it can be beneficial with an illustrative example using nationally representative longitudinal data (High School Longitudinal Study of 2009). A parametric latent growth curve model (LGCM) is used as a template model along with MOB. Second, a simulation study for a given LGCM is conduced to investigate the performance of MOB, which provides researchers with statistical evidence of how well MOB recovers true subgroups. Simulation conditions include a) effect size (0.2, 0.4, 0.6, 0.8, and 1.0), b) sample size (1,000, 2,000, 5,000, 10,000, and 20,000), c) three different test statistic for ordinal

covariate (chi-square, adapted maximum Lagrange multiplier, and a weighted double maximum), d) pre pruning option of limiting the minimum sample size per subgroup (250 vs. none), and e) post pruning option (BIC vs. none). The main evaluation criteria are a) statistical power to recover true subgroups, b) overall classification accuracy and precision, c) accuracy of cut points of ordinal/continuous covariates and labels of categorical covariates, and d) bias and root mean squared error (RMSE) of the parameter estimates per subgroup. Third, the simulation is parallelly conducted with GMM, and the results of it are compared with the ones of MOB.

The key findings suggest that medium effect size (0.4 - 0.6) with relatively large sample sizes (5,000, 10,000, and 20,000) and large effect size (0.8 - 1.0) with adequate sample size (1,000 or 2,000) are enough to distinguish the difference in focal parameters, recovering the true number of subgroups. In addition, treating ordinal variables as either ordinal or categorical is not different in terms of recovering the true subgroups. However, the empirical study suggests that using test statistic for the ordinal covariates is desired when there exist association between the outcome and ordinal covariate. Post pruning using BIC and limiting the minimum size per subgroup simultaneously are also desired options. Without the post pruning with BIC, MOB tends to over-extract the subgroups across conditions. With the same simulated datasets, GMM produced neither accurate subgroups nor reliable parameter estimates.

This study sheds light on how to uncover subpopulations using MOB algorithm with a popular parametric model for longitudinal study. This approach is beneficial for large-scale data such as more than 10,000 sizes with large number of potential covariates. Limitations and future directions are also discussed. The findings play a critical role to lay the groundwork of extending the application of MOB into various statistical models by investigating its performance regarding complex covariate effects.

This dissertation is dedicated to my parents, wife, and friends.
Thank you for always believing in me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 Background and rationale

Machine learning has been an emerging data analytic tool in the fields of methodological statistics as well as social and behavioral sciences since the first appearance of automated interaction detection (AID) for tree-structured regression analysis introduced by Morgan and Sonquist (1963). Emergence of new academic communities such as *Educational Data Mining and Learning Analytics* reflect the growing trend in application of machine learning within educational contexts (e.g., Baker et al., 2016; Paquette et al., 2020). Machine learning refers to either supervised or unsupervised process of modeling that automatically reveals patterns of variation in large-scale datasets or so-called big data. The main goal of machine learning is to build reliable and accurate predictive models, and it has several advantageous features over traditional statistical regression models. First, it can handle high-dimensional predictors even when the number of the predictors (i.e., the number of columns) is larger than the sample size (i.e., the number of rows). Second, there are no statistical assumptions with a model. That is, it has merely set of optimal tuning parameters and several best performing classifiers (i.e., algorithms) to enhance the performance of prediction. Third, no prior knowledge is required to select predictors (covariates or features) for constructing a model, which allows a screening of informative predictors in exploratory research. Fourth, machine learning automatically detects nonlinearity and complex interaction effects of covariates with an iterative algorithmic approach.

Among many machine learning approaches, tree-based methods, also known as recursive partitioning or decision-tree, have been extensively and increasingly employed in educational and psychological research (e.g., Grimm & Jacobucci, 2020; Jacobucci & Grimm, 2020; Strobl et al., 2011; Strobl et al., 2015). A well-known algorithm within the realm of tree-based methods is the

classification and regression trees (CART; Breiman et al., 2017). To briefly explain the tree algorithm, a structured tree is grown by recursively partitioning the samples using available variables, starting from the entire sample (i.e., root node) through subgroups (i.e., child nodes or inner nodes) to final subgroups (i.e., terminal nodes or final nodes) of subjects based on the values of variables selected by algorithms. Thus, the subgroups of subjects are determined by a set of variables that are related to the outcome, in which the values of the outcome are different across subgroups. Two key strengths of using tree-based methods are *interpretability* and *predictability*. Although the predictive power of decision-tree is relatively weaker than other modern approaches (Fernandez-Delgado et al., 2014), such as random forests introduced by Breiman (2001), a single tree is still significantly valuable due to its simple, intuitive, and clear interpretability because it produces a visualization of the tree with the results (e.g., Eo & Cho, 2013; Le & Moore, 2020). Consequently, one can easily understand the structure, composition, and characteristics of the subgroups depicted by the tree.

In contrast to the machine learning detecting complex interactions automatically, traditional statistical models should specify all the interaction terms. In the realm of social sciences, "the testing of interactions is at the very heart of theory testing" (Cohen et al., 2014). Interactions can be a form of person-person, context-context, or person-context (Bauer & Shanahan, 2007). If the interaction effects are well-established and correctly specified in a model, the results of the statistical models are clearly interpretable, unbiased, and efficient (De Gonzalez & Cox, 2007). However, when given many variables in datasets, specifying all possible combinations of interactions of the variables produces complex higher-order interactions. This not only makes ones hard to understand the terms in a model, but also is not common in practice. Even if higher-order interactions are modeled correctly, the interpretation is greatly limited as the effects of each covariate are estimated controlling for both effects of other key covariates and interaction effects.

Moreover, for categorical variables, multiple dummy variables have to be created and the quantity of their interactions with other covariates would increase dramatically. For instance, if there are three covariates including gender (e.g., female, male) and race/ethnicity (e.g., White, Black, Hispanic, Asian, and others), and categorized socio-economic status (e.g., low, medium, and high), the possible number of interaction terms is twenty-two. As researchers also tend to include and interpret interaction terms based on the statistical significance of each term, the model specification is likely to be exploratory and subject to a great degree of modifications, which may lead to spurious results.

To partially compensate for specifying complex interaction effects of the statistical models, there has been substantially growing interests in combining parametric models and tree-based methods over the last two decades (see Loh, 2014 for more details). This analytic approach is beneficial for finding informative covariates as their higher-order interactions and nonlinearity effects can be automatically detected. Among many others, the model-based recursive partitioning (MOB; Zeileis et al., 2008) provides a unified framework that fits a parametric model *locally*, in which the heterogeneous subsamples are determined by testing overall parameter instability. In many cases of social science research, it may be unrealistic to assume that a global model fits the whole sample at a satisfactory. Instead, it would be more reasonable to assume that data for varying subsamples of subjects well fits diverse models (Zeileis et al, 2008). As regards, MOB attempts to find such subsamples given available covariates using a huge recursive searching method. Within this framework, the definition of a covariate is a candidate variable that could be potentially related to the interested outcome(s) in the available datasets. This concept of covariate is broader than the one used in statistics. That is, all the variables that could be related to the outcome(s) that researchers believe can be the candidates of covariates, making algorithm(s) detect and find those relationship automatically. The basic idea of the generic MOB is that a particular parametric model

is fitted to each subsample, which can be in a form of OLS regressions, generalized linear models, item response models, or structural equation models. The estimated parameters are then tested whether they are statistically different depending on the values of covariates. It has been shown that these new approaches adopting machine learning offer valuable opportunities to answer novel research questions for social and behavioral research that is notably different from what traditional parametric statistical models can answer; for example, how the subgroups (formed based on the combinations of covariates) differ in their characteristics, raising the issues of population heterogeneity (e.g., Serang, 2021).

Over the past decade, researchers have been actively adopting MOB particularly for structural equation models, so called (SEM Trees), which was introduced by Brandmaier et al. (2013). SEM Trees integrates the comprehensive and flexible SEM framework with tree algorithm. The main goal of SEM Trees is analogous to MOB, which is to identify subgroups having similar covariance structures or item response patterns using a data-driven, but theory-constrained search as SEM Trees utilizes a template structural parametric model derived from an existing theory. SEM Trees are currently implemented via specific software packages, such as `semtree` (Brandmaier et al., 2013) by either connecting it with `OpenMx` (Neale et al., 2015) or `lavaan` (Rosseel, 2012) to estimate SEM models. Stegmann et al. (2018) further proposed an approach called nonlinear longitudinal recursive partitioning with an associated package of `longRpart2`, which is useful to model inherent nonlinearity of changes. Recently, Serang and his colleagues (2020) extended SEM Trees to make it easily available with popular commercial software `Mplus` by connecting it with `MplusTrees` in R package. They claim that it has some advantages that can cover broader range of SEMs estimated from `Mplus` over `OpenMx` or `lavaan`. Currently, there are a few simulation studies employing SEM Trees for longitudinal data (Usami et al., 2017; Usami et al., 2019). The results indicate that the informativeness of a dichotomous covariate related to the true subgroups,

which is measured by their correlation, was the most critical factor to recover the true number of subgroups.

SEM Trees originally use likelihood-ratio test (LRT; testing change of deviance) by default for evaluating heterogeneity of parameters to search for optimal split points in covariates. The LRT-guided split evaluation appears to be powerful and efficient for dichotomous covariates, although it requires an additional step of locating the optimal cut point for categorical, ordinal, or continuous covariates that have more than two unique values. The likelihood ratio for each possible splitting values of each covariate should be calculated, which in turn produces computationally demanding processes. To complement this, Arnold et al. (2021) recently added options of various score-based testing methods for ordinal and continuous covariates into existing `semtree`. They determined that not only it was more computationally efficient than LRT, but also it showed having higher enough power to detect group differences and unbiased estimates in the selected covariates to grow a tree. A key difference between SEM Trees and MOB is how they choose the split points of a covariate, which is one of the key elements to determine the characteristics of subgroups. While SEM Trees locates the cut points by means of score-based testing, MOB locates the cut points by comparing likelihood ratios. That is, MOB first selects the covariate using test statistic. Then, it determines the cut point by optimizing the sum of the loss function between the two resulting subgroups. This makes MOB computationally more efficient than the original SEM Trees. However, no prior studies have explored the performance of MOB for SEMs.

More importantly, while all previous studies make important contributions, most of the simulation studies merely examined single informative covariate that is directly related to the terminal subgroups. A few empirical studies using SEM Trees did explore various types of covariates showing interactions (e.g., Stegmann et al., 2018) and yet, their applications in terms of adopting options for finding optimal and generalizable results are largely insufficient. Finding

empirical literature using SEM Trees or MOB is also limited in social science literature. Furthermore, until today, there is no simulation study that investigates the performance of MOB for varying types of covariates and their interaction effects within the context of longitudinal SEM, which is the key motivation for this study. To the best of my knowledge, this is the first simulation study to employ the MOB algorithm from `partykit` for the SEM trees rather than using `semtree` to scrutinize how well it performs.

1.2 Research purposes and questions

This study has three main purposes to address the needs of using MOB approach: a) Demonstrate how to use MOB with longitudinal data using two R packages, `partykit` and `lavaan` with an empirical data (High School Longitudinal Study of 2009). b) Investigate the performance of MOB with latent growth curve model (LGCM) having interactions of multiple types of covariates (categorical, ordinal, and continuous) via a simulation study. C) Compare the results of MOB with the ones of growth mixture model (GMM).

To accomplish the above three research purposes, there are specific nine research questions to be answered:

1) How can the approach of MOB with longitudinal data be used to find heterogeneous subgroups?
2) For a given population model of LGCM and the certain number of covariates, how well MOB correctly determine the true number of subgroups?
3) For a given population model of LGCM and the certain number of covariates, how accurately and precisely MOB classify the true subgroups?
4) For a given population model of LGCM and the certain number of covariates, how well MOB recover the splitting points of the covariates?
5) For a given population model of LGCM and the certain number of covariates, how

accurately and precisely MOB recover the parameter estimates (mean intercept)?

6) What is the best option for test statistic of the ordinal covariates?
7) When is the post pruning option of BIC more desirable than without it?
8) When is limiting the minimum sample size per a subgroup more desirable than without it?
9) For a given population model of LGCM, how well GMM correctly determine the true number of subgroups compared to MOB?

The first research question is answered by an illustrative example with detailed procedures. This study employs an empirical data from a nationally representative longitudinal study. Using the results from the empirical data, a population model is specified and datasets for simulations are generated. The second to the eighth research questions are answered by a Monte Carlo simulation study. The performance of MOB was evaluated under various conditions including a) effect size (0.2, 0.4, 0.6, 0.8, and 1.0), b) sample size (1,000, 2,000, 5,000, 10,000, and 20,000), c) treatment of ordinal covariate with different test statistic (chi-square, adapted maximum Lagrange multiplier, and a weighted double maximum), d) pre-pruning option limiting minimum sample size per subgroup (250 vs. none), and e) post-pruning option (BIC vs. none). To answer the nineth research question, the simulated datasets were simultaneously fitted to a growth mixture model (GMM) to see how the results are different from each other.

The next chapter reviews traditional statistical approaches dealing with population heterogeneity depending on the types of variables (outcomes), assumptions, and contexts, followed by reviewing the detailed procedures of MOB and previous literature. Chapter 3 describes how to use MOB with the empirical longitudinal data to find heterogeneous subgroups, followed by the interpretation of the resulting tree and figures. Chapter 4 presents a population model along with simulation designs and evaluation criteria for the simulation study. Chapter 5 describes the results in

the order of the research questions, a) statistical power to determine the true number of subgroups, b) overall classification accuracy and precision of the subgroups, c) accuracy of the splitting points of the covariates, d) bias and root mean squared error (RMSE) of the parameter estimates, and e) desirable options for test statistic of the ordinal covariates, post pruning method using BIC, and limiting minimum sample size per a subgroup. Primary findings, implications, limitations of this study are discussed in Chapter 6.

CHAPTER 2. LITERATURE REVIEW

2.1 Overview of statistical models explaining population heterogeneity

Population heterogeneity has gained attention in social and behavioral science literature (Lubke & Muthén, 2005). Sources of the heterogeneity can be either *observed* or *unobserved*. For the former, heterogeneity is often explained by fixed effects of covariates such as demographics, contextual backgrounds, or behavioral characteristics and psychological traits. These fixed effects reveal the difference of the outcome between the observed covariates specified by researchers in a statistical model. The unexplained individual differences can be captured by residual (error term) in the regression model. If the data has a hierarchically nested structure (i.e., students are nested within schools, and the schools are nested in counties, and so forth), multilevel models (also known as hierarchical linear models; Raudenbush & Bryk, 2002) can be used to specify the associated random effects in a model not only to capture the remained heterogeneity of the individuals, but also to explain the differences of parameters of regression coefficients by regressing them on observed covariates at each level. This approach has been widely used for nested data structure. Another popular approach to investigating heterogeneity is to employ multi-group structural equation models (MGSEM; Jöreskog, 1971) or differential item functioning in item response theory (DIF; Mellenbergh, 1989), which tests separate structural factor models or item response functions with two or more pre-defined grouping covariates such as gender or race. This is especially useful when there is a small number of groups to be tested for comparisons. However, it becomes tiresome and infeasible with numerous groups because multiple estimation and testing should be done to compare parameters across every pair of groups, requiring adjustments of multiple statistical testing. Eventually, this reduces statistical power.

If the source(s) of the heterogeneity is unobserved or latent rather than observed, other

approaches can be used to specify and estimate them in a statistical model. Within a comprehensive and general statistical framework of latent variable modeling (LVM), a popular flexible modeling approach to investigating heterogeneity is called finite mixture model (FMM). FMM can be defined as a parametric statistical model assuming the presence of unobserved distinct groups, also known as latent classes (McLachlan & Basford, 1988; McLachlan et al., 2019). If it is reasonable to assume that the sample consists of several latent groups showing different characteristics or distributions in terms of measured outcome variable(s), a variant of FMM can be a decent choice for researchers to adopt depending on the types of observed indicators and latent variables as well as research questions. Under the umbrella of the extended FMM framework, representative examples of sub-models that explore unobserved heterogeneity include (a) latent class analysis for both categorical observed and latent variables (LCA; McCutcheon, 1987), (b) latent profile analysis for continuous observed indicators and categorical latent variables (LPA; Gibson, 1959), (c) factor mixture model/analysis (FMA; Lubke & Muthén, 2005) and mixture item response theory (Mixture IRT; Rost, 1990) for continuous and categorical/ordinal indicators, respectively, specifying simultaneously both categorical and continuous latent variables for cross-sectional outcomes of measurement models, and (d) growth mixture model for longitudinal outcomes (GMM; Muthén & Asparouhov, 2007; Muthén & Shedden, 1999).

Once the number of the latent classes and their structures are determined via a series of process of selecting the best fitting model based on some criteria such as AIC (Akaike Information Criterion; Akaike, 1974) or BIC (Bayesian Information Criterion; Schwarz, 1978) or relevant other statistical testing and theoretical consideration, a natural interest is to identify the characteristics of the latent classes using available covariates and previous knowledge from established theory. Since the latent class itself does not sufficiently inform meticulous implications, informative covariates should be added to explain the latent class memberships. However, there exist mixed

recommendations for guiding and determining the number of classes whether one should include covariates during the modeling process or after class enumeration. While some argue that simultaneous estimation of a correct model specification with the added covariates produces reliable class enumeration (e.g., Lubke & Muthén, 2007), others, for example, Vermunt (2010) and Nylund-Gibson and Masyn (2016) suggest that the number of latent classes should be determined without covariates first, then covariates or distal outcomes would be added to examine their associations with latent class memberships. The latter approach is called as three-step approach in contrast to one-step approach for the former (Asparouhov & Muthén, 2013). Following the three-step approach, it provides estimated posterior probabilities for each observation to belong to certain classes accounting for uncertainty in class membership. The class membership, which is a nominal variable generated from the posterior probabilities, is then explained by the added covariates using multinomial logistic regression model (Nylund-Gibson et al., 2014). However, the relationship between the covariates and the class membership is typically presumed to be linear, and their interactions are manually formed by one's own choice/decision. If there are many available potential informative covariates that are likely to interact with others, and less established prior knowledge about the relationship, it challenges researchers to correctly specify those complex interactions that might be associated with the class membership. The model-based recursive partitioning method adopting machine learning technique of the decision tree algorithm can handle this issue of complex interactions of the covariates, which helps to understand and interpret the subgroups that are distinct in terms of the specified statistical models.

2.2 Model-based recursive partitioning

The model-based recursive partition is introduced by Zeileis et al (2008) to not only uncover subgroups but also investigate different treatment effects depending on the groups. MOB employs an *empirical score function* for detecting the parameter instability (Zeileis & Hornik,

2007). The score is a case-wise derivative of the estimation function at the estimated parameters. It is used to inspect if the parameter estimates fluctuate *randomly* around their mean of zero or exhibit systematic deviations from zero over the values of covariates, which leading to construct relevant test statistic for any types of covariates. It has been widely used in various social and behavioral research including but not limited to psychometric research, such as measurement invariance (e.g., Merkle et al., 2014). MOB is also adopted to beta regression for limited responses (Grün et al., 2012) and Rasch item response theory (Rasch Trees; Strobl et al., 2015). Recently, it was extended to a linear mixed model that handles multilevel data structure (Fokkema et al., 2018). They conducted a simulation study to see if it recovers treatment-subgroup interactions under nested data structure. The results showed higher accuracy and predictive power for recovering the interaction of fixed effects, while the random effects were set to constant across subgroups.

The general procedures of MOB algorithm have four steps (Zeileis et al., 2008):

(1) A parametric model is chosen by a researcher and fitted to all samples via a selected estimation method that should have a form that either maximizes or minimizes an objective function.

(2) Stability (or volatility) of parameter estimates is assessed for every covariate considered ($Z_p = Z_1, Z_2, \dots, Z_P$). If any overall instability is detected with respect to particular covariate(s), a covariate showing the highest instability are chosen based on the p-value from the test statistics. If there is no significant volatility detected across all the values of the covariates, the process stops.

(3) The fitted model is divided into a set of segmented models according to the split points (values) of the covariates that are searched and computed to locally fit the model better. The number of splits can be either fixed or adaptively chosen. The split points are determined through optimizing the sum of the log-likelihoods of two partitioned models. That is, with

the selected covariate, a split point that improves the highest model fit is determined and the samples are divided into another subsample.

(4) The split is done with the selected covariates and the steps (1) - (3) are repeated until there is no more significant instability. Some stopping rules also can be used by researchers depending on research questions and sample sizes, and if the criteria are met, splitting does not proceed anymore.

The details of how the above steps are carried out are described in the following.

2.2.1 Parameter estimation for a template model

The first step is to fit a parametric model to a whole sample using M-estimators such as ordinary least squares or maximum likelihood estimation. The specification of a template parametric model is determined at this stage according to research questions and established theory. Focusing on maximum likelihood estimation under multivariate normality assumption, the likelihood function given n independent and identically distributed observations y_i $(i = 1, 2, \dots, n)$ and a set of parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_K\}$ is obtained by the product of the individual densities as

$$L(\boldsymbol{\theta}; y_i) = \prod_{i=1}^{n} (2\pi)^{-\frac{k}{2}} |\Sigma(\boldsymbol{\theta})|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\big(y_i - \mu(\boldsymbol{\theta})\big)'\Sigma(\boldsymbol{\theta})^{-1}\big(y_i - \mu(\boldsymbol{\theta})\big)\right\}, \qquad (2\text{-}1)$$

where $\mu(\theta)$ is the $k \times 1$ mean vector and $\Sigma(\theta)$ is the $k \times k$ covariance matrix. A set of parameters is estimated consistently and efficiently by maximizing the above likelihood function or minimizing the negative log-likelihood. The y_i individuals equally contribute to the whole log-likelihood as

$$lnL(\boldsymbol{\theta}; y_i) = -\frac{1}{2}\left\{nkln(2\pi) + nln|\Sigma(\boldsymbol{\theta})| + \sum_{i=1}^{n}\big(y_i - \mu(\boldsymbol{\theta})\big)'\Sigma(\boldsymbol{\theta})^{-1}\big(y_i - \mu(\boldsymbol{\theta})\big)\right\}. \qquad (2\text{-}2)$$

The above function is used typically via well-established iterative ways to find the parameter estimates, $\hat{\theta}$. It is also more widely employed to assess the goodness of model fit between two competing models to select a better fitting model. Multiplying the log-likelihood by -2, the

difference between two log-likelihoods of competing models asymptotically follows a χ^2 distribution with q (the difference in the number of parameters between two models) degrees of freedom. This χ^2 is used as a test statistic for likelihood ratio testing (LRT). LRT is initially suggested by Brandmaier et al. (2013) to determine whether or not the samples are to be divided into sub-samples according to the values of covariates as indicated earlier. MOB uses LRT for locating the split points, but selecting a covariate is completed by other test statistic constructed using a score function.

The score is defined as the gradient of the log-likelihood function with respect to the vector of k parameters. The individual scores are obtained from the individual likelihoods

$$lnL(\boldsymbol{\theta}; y_i) = -\frac{1}{2}\Big\{kln(2\pi) + ln|\Sigma(\boldsymbol{\theta})| + \big(y_i - \mu(\boldsymbol{\theta})\big)'\Sigma(\boldsymbol{\theta})^{-1}\big(y_i - \mu(\boldsymbol{\theta})\big)\Big\}, \tag{2-3}$$

by taking the partial derivatives of them with respect to each parameter where the expected gradient of the function is zero. Parameters estimates $\hat{\theta}$ can be computed by a summation of the partial derivatives of the individual log-likelihood function with respect to a set of $\hat{\theta}$ under mild regularity conditions (White, 1994) as

$$\sum_{i=1}^{n} -\frac{\partial lnL\big(\widehat{\boldsymbol{\theta}}; y_i\big)}{\partial \widehat{\boldsymbol{\theta}}} = 0. \tag{2-4}$$

Then, the individual scores are calculated by solving the first partial derivatives of the individual log-likelihood function with respect to each $\widehat{\boldsymbol{\theta}}$. The score function is represented as a matrix form:

$$s\big(\widehat{\boldsymbol{\theta}}; y_i\big) = \begin{bmatrix} \frac{\partial lnL\big(\hat{\theta}_1; y_1\big)}{\partial \hat{\theta}_1} & \cdots & \frac{\partial lnL\big(\hat{\theta}_K; y_1\big)}{\partial \hat{\theta}_K} \\ \vdots & \ddots & \vdots \\ \frac{\partial lnL\big(\hat{\theta}_1; y_n\big)}{\partial \hat{\theta}_1} & \cdots & \frac{\partial lnL\big(\hat{\theta}_K; y_n\big)}{\partial \hat{\theta}_K} \end{bmatrix}. \tag{2-5}$$

These scores represent the extent to which an individual's log-likelihood is maximized by

each *k* parameters. The closer the individual score is to zero, the better the individual fits the model. On the other hand, large values of scores imply misfit between the individual and model. Thus, it is possible to inspect if the scores deviate from zero systematically according to particular covariates.

2.2.2 Testing instability of parameter estimates

The second step is to test if all parameter estimates are stable, or they fluctuate over a set of partitioning covariates (Z_p) using the empirical score function. Based on the matrix of the empirical contributions to the gradient, the instability of the parameter estimates is tested if splitting the sample with respect to one of covariates (Z_p) improves the model fit. One of the methods is to check if the scores fluctuate randomly around zero or deviate systematically from zero. Under parameter stability, the empirical score function fluctuates randomly around its expected value of zero. If there are some instabilities over parameters, systematic departures from zero for sub-subsamples related to certain covariates can be detected. Intuitively, the idea is similar to examining the randomness of residuals in linear regression. The deviations are monitored by the empirical fluctuation process, which is defined as the *K*-dimensional cumulative score process,

$$B(\widehat{\boldsymbol{\theta}}; j) = \frac{1}{\sqrt{n}} I(\hat{\theta})^{-1/2} \sum_{i=1}^{\lfloor nj \rfloor} s(\hat{\theta}; y_i, Z_p) \quad (0 \le j \le 1), \tag{2-6}$$

where *n* is the total sample size within a subgroup (node), *j* is the number of sorted samples by a candidate covariate Z_p that is being examined. $\lfloor nj \rfloor$ is a floor operator producing an integer part of *nj*. $B_p(\widehat{\boldsymbol{\theta}}; j)$ is the partial sum process of the scores to *nj*th samples ordered by Z_p, scaled by the inverse of the square root of both *n* and the estimated covariance matrix of Fisher's information, $I(\hat{\theta})$, evaluated at the parameter estimates. This produces an $n \times K$ matrix for a *p*th covariate accounting for the ordering of individuals simultaneously.

$B(\widehat{\boldsymbol{\theta}}; j)$ converges to a univariate distribution of Brownian bridge by a functional central

limit theorem under the null hypothesis that the parameter estimates are stable (see Merkle et al., 2014; Zeileis & Hornik, 2007 for details). Formally, the functional central limit theorem holds as

$$B(\hat{\boldsymbol{\theta}}; \cdot) \xrightarrow{d} B^0(\cdot), \tag{2-7}$$

where $\xrightarrow{d}$ denotes convergence in distribution and $B^0(\cdot)$ is a *k*-dimensional Brownian bridge. The Brownian bridge is a stochastic process that is pinned at the start (*i=0*) and the end (*i=n*). The expected value of the bridge is zero with variance $j(1-j)$, implying that most volatility occurs in the middle of the bridge. Thus, an empirical cumulative score can be represented within an $n \times K$ matrix with elements $B_p(\hat{\boldsymbol{\theta}}; i/n)$. This will be denoted as $B(\hat{\boldsymbol{\theta}})_{ik}$. Test statistic of a single value of scalar can be derived by aggregating $B(\hat{\boldsymbol{\theta}})_{ik}$ over *i* individuals and *k* parameters. Each row of the matrix represents a cumulative sum of scores of individuals who belong to i/n percentile of the covariate, Z_p or below. Different ways of aggregating them produce different test statistic depending on the types of covariates (Merckle & Zeileis, 2013; Merckle et al., 2014). Then, the null hypothesis of parameter homogeneity can be tested by comparing the test statistic obtained by aggregating $B(\hat{\boldsymbol{\theta}})_{ik}$ with the corresponding analogous statistic from a Brownian bridge. Currently, MOB provides different test statistic for continuous and categorical covariates as a default option, and users can optionally utilize other two test statistic for ordinal covariates (Hothorn & Zeileis, 2015).

a) Test statistic for categorical covariates (declared as 'factor' in R)

Assessing the parameter instabilities for categorical covariates, Z_p having *M* levels of categories, is achieved by constructing a test statistic, through summing the squared differences in the sum of scores corresponding to the associated category *m = 1, 2, ..., M* of the covariate over *K* parameters,

$$LM = \sum_{m=1}^{M} \sum_{k=1}^{K} \left(B(\hat{\boldsymbol{\theta}})_{i_m k} - B(\hat{\boldsymbol{\theta}})_{i_{(m-1)} k} \right)^2, \tag{2-8}$$

where $i_{(m-1)}$ is the size of individuals within *m-1* category. LM follows χ^2 distribution asymptotically with $K(M-1)$ degrees of freedom. This captures the fluctuation within each of the categories of the partitioning covariate, Z_p. This is a Lagrange multiplier (LM) type test statistic, which is asymptotically equivalent to the corresponding LRT. Using LM is especially beneficial compared to LRT approach in terms of reducing computational burden because the model is fitted once in the current subgroup (node) to estimate parameters and corresponding score functions are computed per node. Then, the scores are simply reordered and aggregated, producing a test statistic each time (Zelleis et al., 2008). The corresponding vector of p-values for the Z_p covariates can also be obtained (Hjort & Koning, 2002). This test statistic can also be used if the ordinal variable is treated as unordered/categorical in the analysis.

b) Test statistic for ordinal covariates (declared as 'ordered' in R)

Although the above LM statistic for categorical variables can be used by default for ordinal variables, two statistics for the ordinal variables were proposed by Merkle et al. (2014). The first one is a weighted double maximum (WDM_O) and the second one is an adapted $maxLM_O$. Their associated test statistic can be also obtained. The former employs multivariate normal probability to calculate the p-values, and the latter gets p-values by means of simulating the critical values on the fly (Kleiber et al., 2002) requiring some computation time. Formally, these statistics can be represented as

$$WDM_O = \max_{m=1,\ldots,M-1} \left\{ \frac{i_m}{n}\left(1-\frac{i_m}{n}\right) \right\}^{-1/2} \max_{k=1,\ldots,K} \left| B\left(\widehat{\boldsymbol{\theta}}\right)_{ik} \right|, \tag{2-9}$$

$$maxLM_O = \max_{m=1,\ldots,M-1} \left\{ \frac{i_m}{n}\left(1-\frac{i_m}{n}\right) \right\}^{-1} \sum_{k=1}^{K} B\left(\widehat{\boldsymbol{\theta}}\right)_{ik}^2. \tag{2-10}$$

c*) Test statistic for continuous covariates (declared as 'numeric' or 'integer' in R)*

Among the proposed three different test statistic for continuous covariates (Merkle & Zeileis, 2013), a currently available statistic is the supremum of Lagrange multiplier that can be presented as

$$maxLM_C = \max_{i=\underline{i},\ldots,\bar{i}} \left\{ \frac{i}{n}\left(1 - \frac{i}{n}\right) \right\}^{-1} \sum_{k=1}^{K} B\left(\widehat{\boldsymbol{\theta}}\right)_{ik}^{2}. \quad (2\text{-}11)$$

This is the maximum of the sum of the squares of the $B_{ik}\left(\widehat{\boldsymbol{\theta}}\right)$ in each k-dimensional vector of parameters over all single-split j samples, scaled by its variance component. Since $maxLM_C$ considers the values of all parameters at a single point, it is suitable when there exists a noticeable single change point in several parameters. The lower and upper bounds $[\underline{i}, \bar{i}]$ are typically specified for the continuous covariates since few individuals belonging to the extreme values could have an effect on instability of the test statistic. $maxLM_C$ is asymptotically equivalent to the supremum of likelihood-ratio statistics (Chow, 1960; Zelleis et al., 2008), and the asymptotic p-values can be obtained from a table proposed by Hansen (1997). If there are many ties in the partitioning covariates, the maximum value is not unique, and the results may be affected by the ordering of the individuals. In this case, it is suggested to either investigate the results by breaking ties randomly or treat the continuous variable as an ordinal variable.

2.2.3 Partitioning sample into subgroups along with selected covariates

In the third step, sample in the current node is divided into child nodes along with the partitioning covariate, Z_{p*}. The covariate showing the strongest association with the parameter instability producing the highest p-value is firstly chosen. Then, a split value (breakpoint) that optimizes the estimating function with the largest improvement of the model fit is computed. This is achieved by an exhaustive iterative search procedure, during which the two models consisting of the divided subsamples within each node are fit and the split point is determined for each noticeable breakpoint in the chosen Z_{p*} covariate based on likelihood ratio testing.

2.2.4 Repeating steps until stopping rules are met

Steps 1) - 3) are repeated until there is no significant instability of parameters is detected in the current nodes. The node showing stable/homogeneous set of parameters is called a terminal node, which is the uncovered distinct subgroup. Optionally, there are pre- and post- pruning strategies to determine the optimal tree size. The options for the former include a setting of a minimal number of sizes for a node, for example, setting up a node size of 100, and/or to use a Bonferroni-adjusted p-value. However, when the sample size is very large that the traditional significance level is not useful, the resulting trees typically produce excessive number of terminal nodes because small parameter instabilities can be detected, which is not concise to interpret the trees. In this case, a grown large tree is pruned back if the splits did not improve the model based on AIC or BIC (Su et al., 2004). That is, the large tree is pruned/evaluated based on whether or not the sum of model fit of the divided subgroups improves the model fit of the previous group including the divided subgroups statistically. This option is currently available in `partykit` (Hothorn & Zeileis, 2015). Among others, limiting the sample size per a subgroup (node) and post pruning method using BIC are investigated in a simulation study.

CHAPTER 3. AN ILLUSTRATIVE EXAMPLE

To achieve the first research purpose, this chapter demonstrates how to use MOB with a parametric template model of latent growth curve model (LGCM) with an educational empirical data. Before describing the procedure of MOB, it is necessary to understand how MOB works to uncover statistically distinct and interpretable groups creating a visualized tree. I used two categorical covariates, gender, and race. They are considered to be related to the outcome, academic achievement. The outcome variables are overall GPA scores from 9th to 12th grades, which produces four repeated measurements. A quadratic latent growth curve model was chosen as a template parametric model rather than linear growth model, to estimate the fixed effects of the intercept, linear slope and quadratic slope plus their covariance-variance and the residuals. The result shows how the changes in GPAs over four years differ across demographics.

An example of LGCM tree using MOB is presented in Figure 1. The composition of the subgroups can be found in this visualized tree directly. Following the terminology of the machine learning community, the top of the tree is called *root node*, which indicates the whole sample that is used for the study. The subsamples (subgroups) in the middle of the tree are called *inner nodes* or *child nodes*. These subgroups were determined by a race covariate first in this example because the parameters of some ethnic groups are most significantly different from other ethnic groups. Specifically, the left side of the tree consists of three race groups, Black (BLK), Hispanic (HIP), and Other ethnic groups (OTS), and the right side of the tree consists of two race categories, Asian (ASA) and White (WHT). Then, the machine learning algorithm built in MOB keeps continuing to find heterogeneity (instability) of the model parameters. The second covariate dividing the subgroups is the gender, and the subsamples are divided into subsamples again and again until either some stopping rules are met or there is no more significant instability

of the parameters across the values of the covariates. The final distinct subgroups are called *terminal nodes*. Each final distinct subgroups have their sample size and own model parameters including the intercept, linear slope, and quadratic slope, to distinguish their change trajectories based on the latent growth curve model.

In this example, there are nine distinct subgroups showing significantly different initial GPA score and their changes over time. For example, the third subgroup from the left at the bottom of the tree is the group of Black male students (Node 8), which shows the lowest initial GPA score. In contrast, the sixth subgroup from the left at the bottom of the tree is the group of Asian female students (Node 13) showing the highest initial GPA score. This tree easily informs us the composition of the subgroups, which enables us to understand complex interaction effects between the race and gender. The information in the bottom of the tree can be customized by researchers. For instance, all the parameter estimates including the random effects and their standard errors can be also presented while this study presents three coefficients only for the illustrative purpose.

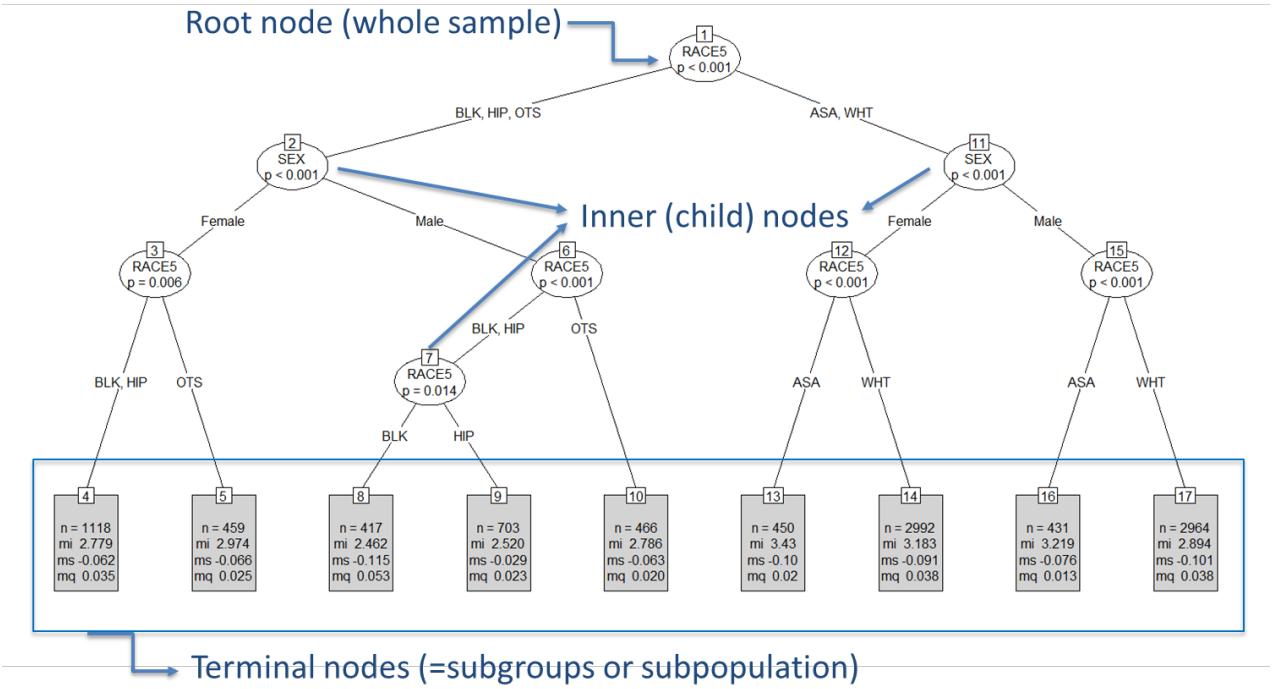


Figure 1. An example of LGCM tree using MOB.

Using the parameter estimates of each terminal node, the expected GPA change trajectories across four grades can be visualized as Figure 2. Interestingly, most of the subgroups show similar change trajectories that slightly decrease from the first year to the third year then increase at the four years. This result is expected and can be interpreted as common in population because most of the students and schools would make their great efforts to manage the GPA scores of the 12th grade. However, this analysis reveals that two other subgroups (Node 13 and Node 16) do not follow the same trend. Node 13 and Node 16 in the Figure 2 show that their overall GPA scores *constantly* decreased over the four years despite of their highest initial GPA score at the first year.

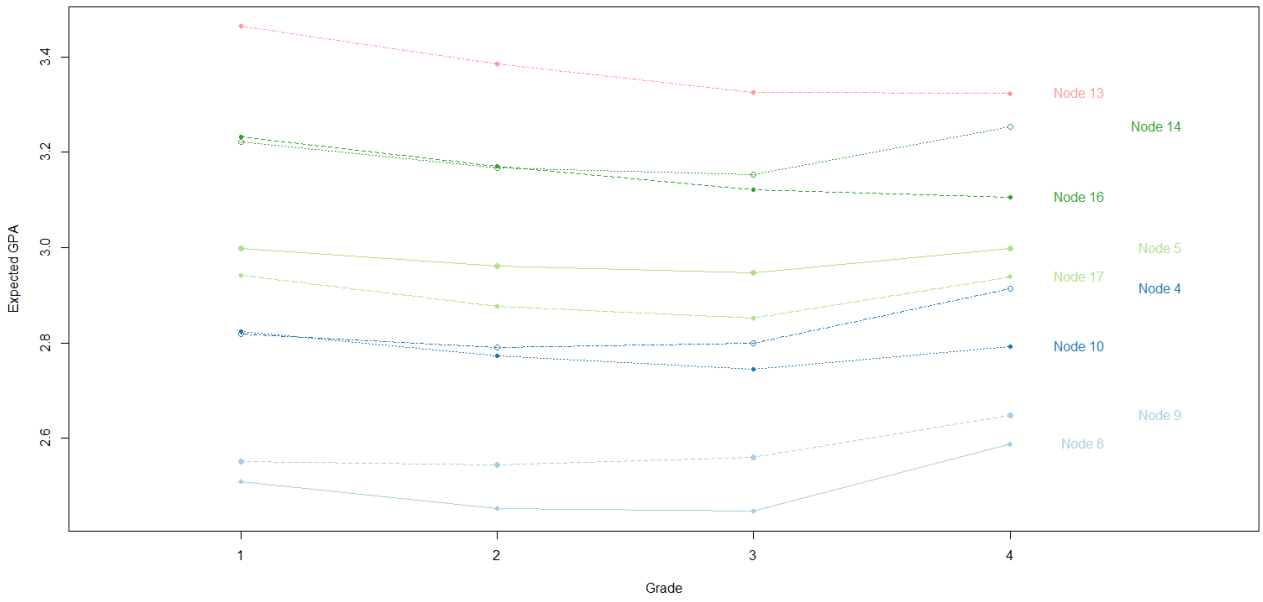


Figure 2. An example of expected GPA changes across four grades (Nodes are subgroups).

Looking at the previous Figure 1, the Nodes 13 and 16 are the groups of Asian female and Asian male students, respectively. Moreover, regarding the change between the third and fourth year, the amount of increase for the nodes 14 (White female), 4 (Black / Hispanic female), and 8 (Black male) is large compared to other groups. The lowest initial GPA group is Black and Hispanic male student groups (nodes 8 and 9). The combination of two covariates produces ten intersectional groups (interaction effect) though, nine groups are statistically different in terms of the change in GPAs over time. This section would not attempt to interpret the substantive meaning of the results because this analysis used only two covariates of the race and gender, to provide an overview of the tree and what information MOB can produce for the interpretation. Next section describes how to use MOB in detail.

3.1 Data

The data used for this study is High School Longitudinal Study (HSLSL:09). This is a representative sample of the U.S. high school students collecting a variety of information on

students' academic outcomes, experiences, environments, and backgrounds. It used a stratified two-stage sampling design, which samples schools first followed by students from the selected schools as the second step. It started to collect the data from the academic year of 2009, and follow-up data were collected in 2012, 2013, and 2016. In 2013, the transcripts were collected to add more detailed information on students' academic outcomes containing students' course taking and GPA scores. In addition, the HSLS:09 contains numerous covariates on students' experiences, demographics, and their noncognitive scale scores related to the educational outcomes. To producing replicable results, this study uses publicly available datasets.

The interested educational outcomes are the overall GPA scores from 9^{th} to 12^{th} grades, which is the same as the above exemplary analysis. These measures are non-cumulative GPA scores. Measurement invariance across years is assumed for the purpose of this study that GPA score measures the same construct across years. The outcomes are used to fit unconditional latent growth curve model. One of the most important reasons to choose GPA score as an outcome for this study is that GPA scores have been known as strong indicators to predict academic success in college and career (Allensworth & Clark, 2020). According to their recent study, high school GPA is the most critical indicator of the academic readiness and performance for students and institutions of higher education. In that sense, the high school GPA scores of the different students would follow different trajectories having different initial points and shapes. Some of groups of students could share their backgrounds, demographics, and experiences in schools in terms of the GPA scores. The purpose of this approach, MOB, is to find those subgroups using available predictors/covariates.

The covariates found to be related to GPA scores are selected based on previous literature to avoid any irrelevant covariates and to reduce the estimation time (e.g., Bowers &

Sprott, 2012). A few available student- and school-level covariates are considered for this study though, there is no limit to the number of covariates. That is, if there is no existing prior literature matching to the research topic or if the research purpose is to explore any potentially relevant covariates, researchers can add any kinds of covariates without any limitations while this would increase the estimation time. One of the benefits of using MOB is to use covariates as they are. It is not required to make multiple dummy variables for making groups with categorical variables, which is commonly done in regression modeling. In addition, it is unnecessary to assume that the ordinal responses are either continuous or categorical. Researchers can declare the ordinal variables as ordinal, and MOB can get the relevant test statistic for the hypothesis testing of the ordinal variables as I described in Chapter 2. A list of the used four outcomes and ten covariates are presented in Table 1. The covariates of the first year (9th grade) were only chosen to be related to the initial GPA score and its change over time.

Table 1. A list of variables of High School Longitudinal Study of 2009 data

Variables	Descriptions	Types
Outcomes		
GPA9 GPA10 GPA11 GPA12	Overall GPA scores from 9th to 12th grades (four repeated measures and non-cumulative GPA scores)	Continuous
Covariates		
SEX	Female or male	Categorical
RACE	Black, Hispanic, Asian, White, and others	Categorical
LOCA	School location: city, suburban, town, and rural	Categorical
FLUNCH*	Categorized percentage of students enrolled in the school who receive free or reduced-price lunch; 0 = 0%, 1= more than 0% but less than 10%, 2 = at least 10% but less than 20%, …, 11 = 100%. This has 11 ordered categories	Ordinal
HACT	Hours spent on extracurricular activities on typical school day; 1 = less than 1 hour, 2 = 1 to 2 hours, …, 6 = 5 or more hours. This has six ordered categories	Ordinal
MISBEHAV*	Frequency of student in-class misbehavior at this school; 1 = Daily, 2 = At least once a week, 3 = At least once a month, 4 = On occasion, and 5 = Never happens. This has five ordered categories	Ordinal
SES	Socio-economic status scale	Continuous
MATEFF	Standardized scale of student's math self-efficacy; higher values represent higher math self-efficacy	Continuous
BEHAVSCH	Standardized scale of student's answer about in-school behavior within last 6 months. Higher values represent more positive assessments of the school's problems, i.e., fewer problems are indicated	Continuous
SCHCLI	Standardized scale of administrator's assessment of school climate; higher values represent more positive assessments of the school's climate, i.e., fewer problems are indicated	Continuous

Note. All continuous variables are rounded to one decimal place for the analysis. FLUNCH and MISBEHAV are restricted-use variables. The covariates are chosen from the first year (9th grade) only.

First, the categorical covariates are 1) gender (female and male), 2) race (Black, Hispanic, Asian, White, and Others), and 3) school location (city, suburban, town, and rural). Second, the ordinal covariates are 1) hours spent on extracurricular activities on typical school day (HACT; 1 = less than 1 hour, 2 = 1 to 2 hours, …, 6 = 5 or more hours), which has six ordered categories, 2) the categorized percentage of students enrolled in the school who receive free or reduced-price lunch (FLUNCH; 0 = 0%, 1 = more than 0% but less than 10%, 2 = at least 10% but less than 20%, …, 11 = 100%), 3) the frequency of student in-class misbehavior at this school (MISBEHAV; 1 = daily, 2 = at least once a week, 3 = at least once a month, 4 = on occasion, and 5 = never happens). Third, the continuous covariates are 1) socio-economic status (SES; standardized scale score), 2) student's math self-efficacy scale score (MATEFF; higher values represent higher math self-efficacy), 3) school's motivation scale score (BEHAVSCH; higher values represent more positive assessments of the school's problem), and 4) scale score of the administrator's assessment of school climate (SCHCLI; higher values represent more positive assessments of the school's climate, i.e., fewer problems are indicated). All these scale scores are standardized composite scores and rounded to one decimal place to reduce the estimation time.

3.2 Template model: LGCM

To use a latent growth curve model (LGCM) as a template model of the MOB, the first step is to scrutinize descriptive statistics of the variables that are intended to be used based on previous literature or theories. After excluding all non-responses for each variable, 9,275 samples were chosen to be analyzed for the illustrative purpose assuming the missingness of missing completely at random. The descriptive statistics of the continuous/ordinal variables and frequency and proportions of the categorical variables are presented in Table 2 and Table 3,

respectively. Although the descriptive statistics of the ordinal variables are presented with the continuous variables together, the analysis treats the ordinal variables as ordinal rather than categorical or continuous. This feature was discussed in detail Chapter 2.

Table 2. Descriptive statistics of continuous and ordinal variables without sampling weights (n=9,275)

	Mean	*SD*	*Min*	*Max*	*Skewness*	*Kurtosis*
GPA9	2.982	0.741	0.130	4.000	3.870	-0.667
GPA10	2.937	0.750	0.000	4.000	4.000	-0.632
GPA11	2.938	0.741	0.000	4.000	4.000	-0.686
GPA12	3.027	0.727	0.000	4.000	4.000	-0.971
FLUNCH*	4.494	2.468	1.000	12.000	11.000	0.296
HACT*	2.600	1.434	1.000	6.000	5.000	0.805
MISBEHAV*	2.116	1.151	1.000	5.000	4.000	0.652
SES	0.178	0.780	-1.800	2.900	4.700	0.286
MATEFF	0.059	0.997	-2.500	1.700	4.200	-0.356
BEHAVSCH	0.124	0.868	-5.600	1.200	6.800	-1.640
SCHCLI	0.253	0.995	-3.200	2.600	5.800	-0.411

Note. FLUNCH, HACT, and MISBEHAV are ordinal variables. FLUNCH = categorized percentage of students enrolled in the school who receive free or reduced-price lunch, HACT = hours spent on extracurricular activities on typical school day. MISBEHAV = frequency of student in-class misbehavior at this school. SES = socio-economic status scale score. MATEFF = student's math self-efficacy scale score. BEHAVSCH = student's school motivation scale score. SCHCLI = scale score of the administrator's assessment of school climate.

Table 3. Frequency and proportions of categorical variables without sampling weights

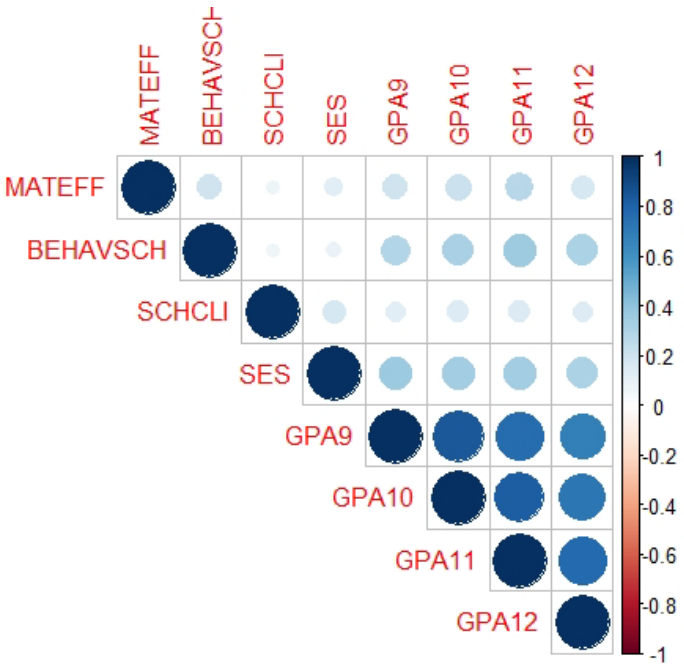
Gender	Female	Male	Total	Female	Male	Total
Race		*n*			*%*	
Others	410	458	868	4.4	4.9	9.4
Asian	373	397	770	4.0	4.3	8.3
Black	336	333	669	3.6	3.6	7.2
Hispanic	640	638	1,278	6.9	6.9	13.8
White	2,856	2,834	5,690	30.8	30.6	61.3
Total	4,615	4,660	9,275	49.8	50.2	100.0
	City	Suburb	Town	Rural	Total	
School location	2,518 (0.27%)	3,287 (0.35%)	1,207 (0.13%)	2,263 (0.24%)	9,275 (100.0%)	

Correlations between continuous variables are estimated and visually presented with the coefficients in Table 4. Color intensity and the size of the circle are proportional to the correlation coefficients. In the right side of the correlogram, the legend color shows the correlation coefficients and the corresponding colors. The SES and school motivation (BEHAVSCH) have stronger associations (around 0.3s) with GPA scores than two other covariates (around 0.1 - 0.2).

Table 4. Correlation matrix between continuous variables with a visualized figure

	GPA9	GPA10	GPA11	GPA12	SES	MAT-EFF	BEHA-VSCH	SCH-CLI
GPA9	1.000	0.843	0.761	0.684	0.354	0.216	0.290	0.132
GPA10	0.843	1.000	0.821	0.721	0.343	0.219	0.323	0.154
GPA11	0.761	0.821	1.000	0.776	0.341	0.272	0.361	0.160
GPA12	0.684	0.721	0.776	1.000	0.307	0.176	0.320	0.152
SES	0.354	0.343	0.341	0.307	1.000	0.133	0.080	0.177
MATEFF	0.216	0.219	0.272	0.176	0.133	1.000	0.200	0.066
BEHAVSCH	0.290	0.323	0.361	0.320	0.080	0.200	1.000	0.079
SCHCLI	0.132	0.154	0.160	0.152	0.177	0.066	0.079	1.000

Note. GPA9 - GPA12 = Overall GPA scores from the 9th to the 12th grade. SES = socio-economic status scale score. MATEFF = student's math self-efficacy scale score. BEHAVSCH = student's school motivation scale score. SCHCLI = scale score of the administrator's assessment of school climate.



The general procedure is the same as the traditional approach. One of the most important parts of this is to examine whether the growth pattern shows linear or quadratic or other nonlinear forms. Then, researchers can compare the model fit indices to find the best fitting model. Among them, ones can also determine if the random effects are either constrained to be the same across times or freely estimated by comparing all the candidate models. In this study, the average trend of GPA scores across four grades (9th to 12th) was examined by drawing a spaghetti plot of 50 randomly chosen students. In Figure 3, it is not easy to find a noticeable general trend pattern of the GPA score across years. However, it is hard to find dramatic changes of GPA score from the year to year. Keeping this in mind, both linear and quadratic latent growth curve models are considered to be fitted to this data.

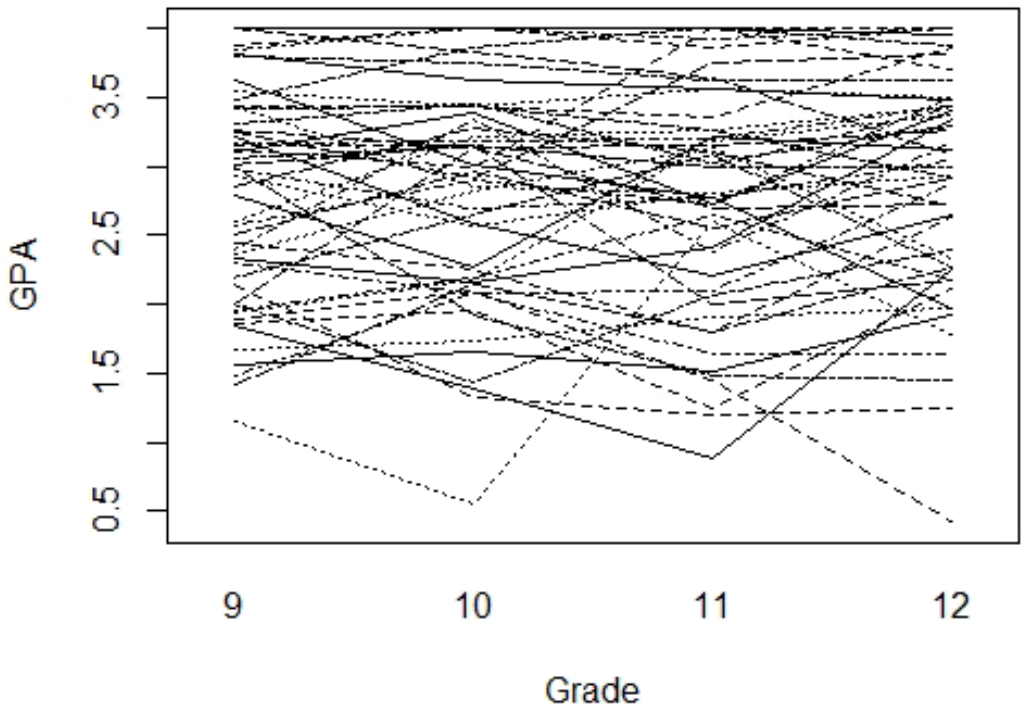


Figure 3. Spaghetti plot of overall GPA for 50 randomly chosen students.

Following the classical notations from Bollen and Curran (2006), a latent growth curve model can be represented as

$$\begin{aligned} Y_i &= \Lambda \cdot \eta_i + \varepsilon_i, \\ \eta_i &= \mu_{\eta_i} + \zeta_i, \end{aligned} \tag{3-1}$$

where Y_i is a $T \times 1$ vector of T repeated measures for ith individual ($i = 1, 2, \ldots, n$), Λ is a $T \times k$ matrix of factor loadings (intercept, linear slope, and quadratic slope: these are fixed based on the coding of time points), η_i is a $k \times 1$ vector of k latent factors, and ε_i is a $T \times 1$ vector of residuals. The vector of latent factors is decomposed into $\mu_{\eta_i} + \zeta_i$, which are the mean and deviance, respectively.

A combined form can be expressed as $Y_i = \Lambda \cdot (\mu_{\eta_i} + \zeta_i) + \varepsilon_i$. Then, the model-implied covariance matrix is written as

$$\Sigma = \Lambda \Psi \Lambda' + \Theta_{\varepsilon_i}, \tag{3-2}$$

where Σ is the covariance matrix of the responses Y_i, Ψ is the covariance structure of latent factors of ζ_i, and Θ_{ε_i} is the covariance structure of the residuals, which is a diagonal matrix consisting of all the variance components. With four time points, the factor loadings of intercept, linear slope, and quadratic slope were fixed and coded as [1, 1, 1, 1], [0, 1, 2, 3], and [0, 1, 4, 9], respectively. The variance of residuals (ε_i) are fixed to be the same across time points to avoid negative variances and to enhance model fit in this study. Depending on distribution of data, the residuals can be freely estimated. They are assumed to be normally and equivalently distributed as $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ following a common practice though, one can specify different distributional forms.

Denoting θ to be all parameters to be estimated, $\mu(\theta)$ and $\Sigma(\theta)$ are model-implied mean and covariance structures, respectively. Without covariates/predictors, the parameters

estimated from this model specification are a) means of three latent factors (fixed effects; intercept (μ_I), linear slope (μ_S), and quadratic slope (μ_Q)), b) their variances (σ_I^2, σ_S^2, and σ_Q^2) and covariances (σ_{IS}, σ_{IQ}, and σ_{SQ}), and c) a fixed residual (σ_ε^2). In total, there are ten parameters. Parameters are estimated by employing the robust maximum likelihood (MLR) estimation under the mild multivariate normality assumption.

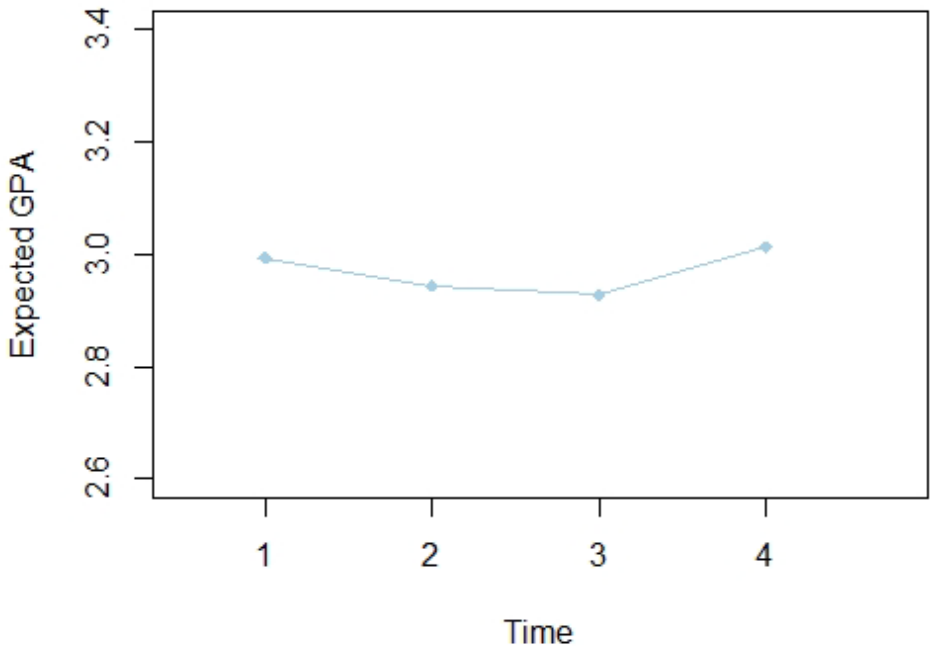


Figure 4. The expected GPA score over four years from quadratic latent growth curve model.

This study examined a variety of models including the linear latent growth curve model and quadratic LGCM with freely estimated residuals across times. However, the model fit indices of the linear model were poorer than the ones of the quadratic model. In addition, there were negative variance estimates with the freely estimated residuals. Based on this result, this study decided to use the quadratic LGCM as a template model. Table 5 shows the parameter

estimates and the model fit indices of the unconditional quadratic LGCM. As displayed in Figure 4, the expected overall GPA score decreases from the first year (9th grade) to the third year (11th grade) and then, increases in the fourth year (12th grade). The variances and covariances of the latent growth factors are significant except the covariance between intercept and linear slope. The intercept variance is fairly large compared to the slopes, indicating that there are significant individual differences in the GPA score at 9th grade (i.e., Time point 1), however, the overall patterns of the changes over time is alike among individuals. To explore what demographic, environmental, and behavioral factors are associated with the initial status (i.e., GPA in 9th grade) and changes over time, informative covariates can be used to explain the differences.

Table 5. Results of unconditional quadratic latent growth curve model

Parameters		Estimates	S.E.	p-value
Means	Intercept (μ_I)	2.984	0.008	< 0.001
	Linear slope (μ_S)	-0.087	0.005	< 0.001
	Quadratic slope (μ_Q)	0.034	0.002	< 0.001
Variances	Intercept (σ_I^2)	0.471	0.008	< 0.001
	Linear slope (σ_S^2)	0.057	0.005	< 0.001
	Quadratic slope (σ_Q^2)	0.007	0.001	< 0.001
Covariances	I ~ S (σ_{IS})	0.001	0.004	0.791
	I ~ Q (σ_{IQ})	-0.012	0.001	< 0.001
	S ~ Q (σ_{SQ})	-0.016	0.002	< 0.001
Variances	Residuals (σ_ε^2)	0.084	0.001	< 0.001
Model fit indices				
	AIC	51542.013		
	BIC	51613.364		
	CFI	0.994		
	TLI	0.991		
	RMSEA	0.064		

Note. AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), CFI (Comparative Fit Index), TLI (Tucker Lewis Index), RMSEA (Root Mean Square Error of Approximation)

3.3 Using MOB to grow a tree

The above unconditional model can be used as a template model to grow trees with available potentially informative covariates by partitioning recursively. The idea of this study was introduced by Zeileis (2020). Thus, most of the details of how to connect two R packages of `partykit` and `lavaan` is referenced by his paper. The detailed R codes for this empirical analysis was presented in Appendix. There are two ways to determine the optimal size of the node. The first option is to set up the minimal size per subgroup in advance. The second option is to use post-pruning strategy that prunes splits back using information criteria such as BIC as described earlier. The partitioning algorithm stops when there is no significant parameter instability based on pre-specified Type-I error rate (e.g., a = 0.05 with Bonferroni correction controlling for familywise error rate), however, small differences can be identified as they are significant with large size dataset like this study. To avoid such large number of terminal nodes with small sizes, a post-pruning method adopting BIC is employed. Additionally, the minimum sample size per node is set to 250 to get stable parameter estimates for each subgroup. Previous literature suggests to have minimum sample size of 250 to correctly detect nonlinear changes with four time points (Diallo et al., 2014). Indeed, when the minimum number of nodes was set as less than 250 (e.g., 100, 150, or 200), there were negatively estimated variances of the linear slope for a specific terminal node. The test statistic for the ordinal covariate was set to use either the weighted double maximum (WDM_O) or the adapted $maxLM_O$ because the results were the same. However, using the test statistic of the categorical covariate for the ordinal covariate, which is the default option in `partykit,` was not the same as the above two. This different feature was investigated in the simulation study. The following is the specific procedure how MOB is used to grow a tree.

Firstly, the model was fitted with default options without both limiting the number of subgroups and post pruning of using BIC. The default option for the ordinal covariate was to use the test statistic of the categorical covariate. The resulting tree produced *57* terminal nodes, which is too large to interpret the resulting tree. Also, the number of sample size of the smallest subgroup was 36. This would not be acceptable to say that the parameter estimates are stable. Furthermore, there are a lot of subgroups in a tree. It is neither meaningful to interpret the tree nor concise to have better understanding about the composition of the subgroups.

Secondly, the model was fitted with the same options of the above with adding the minimum sample size of 250. The resulting tree produced *25* terminal nodes, which was smaller than the above. The first choice of splitting covariate was the SES as well. Still, there were a lot of overlapped subgroups in terms of trajectories. In addition, the tree structure was very complicated. Next, the same model was fitted with *both* limiting the number of subgroups and post pruning of using BIC. The default option for the ordinal covariate was used again. The resulting tree produced *9* terminal nodes, which is concise and meaningful to interpret the resulting tree. The first covariate used for splitting was SES again. Interestingly, the ordinal covariates were not used to split the subgroups in a tree so far. Only the continuous and categorical covariates were used for splitting.

Finally, the option using test statistic of WDM_O which is used for the ordinal covariates, was added to the existing other options of the above. The resulting tree produced *13* terminal nodes, which was larger than the above. More importantly, the ordinal covariates, such as HACT and FLUNCH covariates, were used to split the groups. The parameter estimates of MOB with the quadratic LGCM for each subgroup are presented in Table 6. The first column represents the composition of the subgroups with the node number. The substantive

interpretation of the results can be accomplished by using the Table 6 along with Figures 5 and 6. First, the subgroup that shows the lowest mean intercept was the group of students who 1) spent more than one hour on extracurricular activities on typical school day, 2) have less than or equal to -1.2 scale score of the school motivation, 3) have less than or equal to 0.6 scale score of socio-economic status, and 4) attending a school in which the percentage of students who receive free or reduced-price lunch is more than 20% (Node 18: HACT > 1H & FLUNCH > 20% & BEHAVSCH ≤ -1.2 & SES ≤ 0.6). This group shows dramatic decrease of GPA score from the first to the second year, and then it increases from the third to the fourth year looking at the Figure 6. Second, the subgroup that shows the highest mean intercept was the group of *female* students who 1) spent more than one hour on extracurricular activities on typical school day, 2) have larger than 0.5 scale score of socio-economic status, and 3) attending a school in which the percentage of students who receive free or reduced-price lunch is less than or equal to 20% (Node 14: HACT > 1H & FLUNCH ≤ 20% & SES > 0.5 & Female). This group shows slight decrease of GPA score from the first to the second year, and then it slightly increases from the third to the fourth year looking at the Figure 6.

There are distinct subgroups showing notable different trajectories compared to other subgroups. Looking at the Figure 6, purple and pink lines show gradual increase of GPA score across the school years. Using the information from the Table 6, these groups are node 6 (purple) and node 5 (pink). The Node 6 is the group of students who 1) spent less than or equal to one hour on extracurricular activities on typical school day, 2) have larger than -0.1 scale score of the school motivation, and 3) have larger than 0.1 scale score of socio-economic status (HACT ≤ 1H & BEHAVSCH > -0.1 & SES > 0.1). The Node 5 is the group of students who 1) spent less than or equal to one hour on extracurricular activities on typical school day, 2) have larger than -0.1

scale score of the school motivation, and 3) have smaller than 0.1 scale score of socio-economic status (HACT ≤ 1H & BEHAVSCH > -0.1 & SES ≤ 0.1). The only difference between these two groups is the socio-economic status, making the difference of the mean intercept of GPA score in the first year. These groups spent less than one hour on extracurricular activities though, they had higher behavioral motivation in school on average, leading to gradual increase of GPA score across years. The interesting finding is that the overall GPA scores of these groups are higher than the ones of the node 3 group (HACT ≤ 1H & BEHAVSCH ≤ -0.1) across years. The node 3 group had lower behavioral motivation in school on average, leading to the second lowest overall GPA score across years regardless of their socio-economic status.

Table 6. Parameter estimates of the LGCM tree of MOB using all covariates

Subgroup composition (Node #)	μ_I	μ_S	μ_Q	σ_I^2	σ_S^2	σ_Q^2	σ_{IS}	σ_{IQ}	σ_{SQ}	σ_ε^2
HACT > 1H & FLUNCH > 20% & BEHAVSCH ≤ -1.2 & SES ≤ 0.6 (18)	2.450	-0.261	0.083	0.456	0.067	0.007	-0.042	0.005	-0.018	0.189
HACT ≤ 1H & BEHAVSCH ≤ -0.1 (3)	2.474	-0.154	0.057	0.494	0.106	0.013	-0.018	-0.010	-0.031	0.128
HACT ≤ 1H & BEHAVSCH > -0.1 & SES ≤ 0.1 (5)	2.748	0.000	0.014	0.484	0.095	0.009	-0.039	-0.004	-0.024	0.113
HACT > 1H & FLUNCH > 20% & BEHAVSCH > -1.2 & SES ≤ 0.6 (19)	2.771	-0.146	0.050	0.424	0.044	0.007	0.007	-0.015	-0.014	0.108
HACT > 1H & FLUNCH ≤ 20% & BEHAVSCH > 0.0 & SES ≤ 0.5 (10)	2.780	-0.153	0.053	0.383	0.040	0.006	0.014	-0.014	-0.012	0.091
HACT > 1H & FLUNCH > 20% & BEHAVSCH > 0.3 & Others or Black or Hispanic (22)	2.866	-0.033	0.021	0.450	0.047	0.004	-0.032	-0.005	-0.010	0.090
HACT > 1H & FLUNCH > 20% & BEHAVSCH > 0.3 & SES < -0.3 & Asian or White (24)	3.001	-0.068	0.032	0.439	0.084	0.008	-0.040	-0.003	-0.021	0.073
HACT ≤ 1H & BEHAVSCH > -0.1 & SES > 0.1 (6)	3.138	-0.039	0.023	0.384	0.048	0.006	0.002	-0.012	-0.015	0.063
HACT > 1H & FLUNCH ≤ 20% & BEHAVSCH > 0.0 & SES ≤ 0.5 (11)	3.177	-0.055	0.020	0.320	0.023	0.003	0.006	-0.010	-0.007	0.057
HACT > 1H & FLUNCH > 20% & BEHAVSCH ≤ 0.3 & SES > 0.6 (20)	3.198	-0.163	0.052	0.341	0.039	0.005	-0.004	-0.005	-0.011	0.072
HACT > 1H & FLUNCH ≤ 20% & SES > 0.5 & Male (13)	3.240	-0.039	0.012	0.261	0.026	0.004	0.005	-0.007	-0.008	0.045
HACT > 1H & FLUNCH > 20% & BEHAVSCH > 0.3 & SES > -0.3 & Asian or White (25)	3.349	-0.088	0.033	0.260	0.033	0.004	0.017	-0.013	-0.010	0.063
HACT > 1H & FLUNCH ≤ 20% & SES > 0.5 & Female (14)	3.479	-0.063	0.022	0.206	0.041	0.004	-0.008	-0.006	-0.010	0.032

Note. FLUNCH, HACT, and MISBEHAV are ordinal variables. FLUNCH = categorized percentage of students enrolled in the school who receive free or reduced-price lunch, HACT = hours spent on extracurricular activities on typical school day. SES = socio-economic status scale score. BEHAVSCH = student's school motivation scale score.

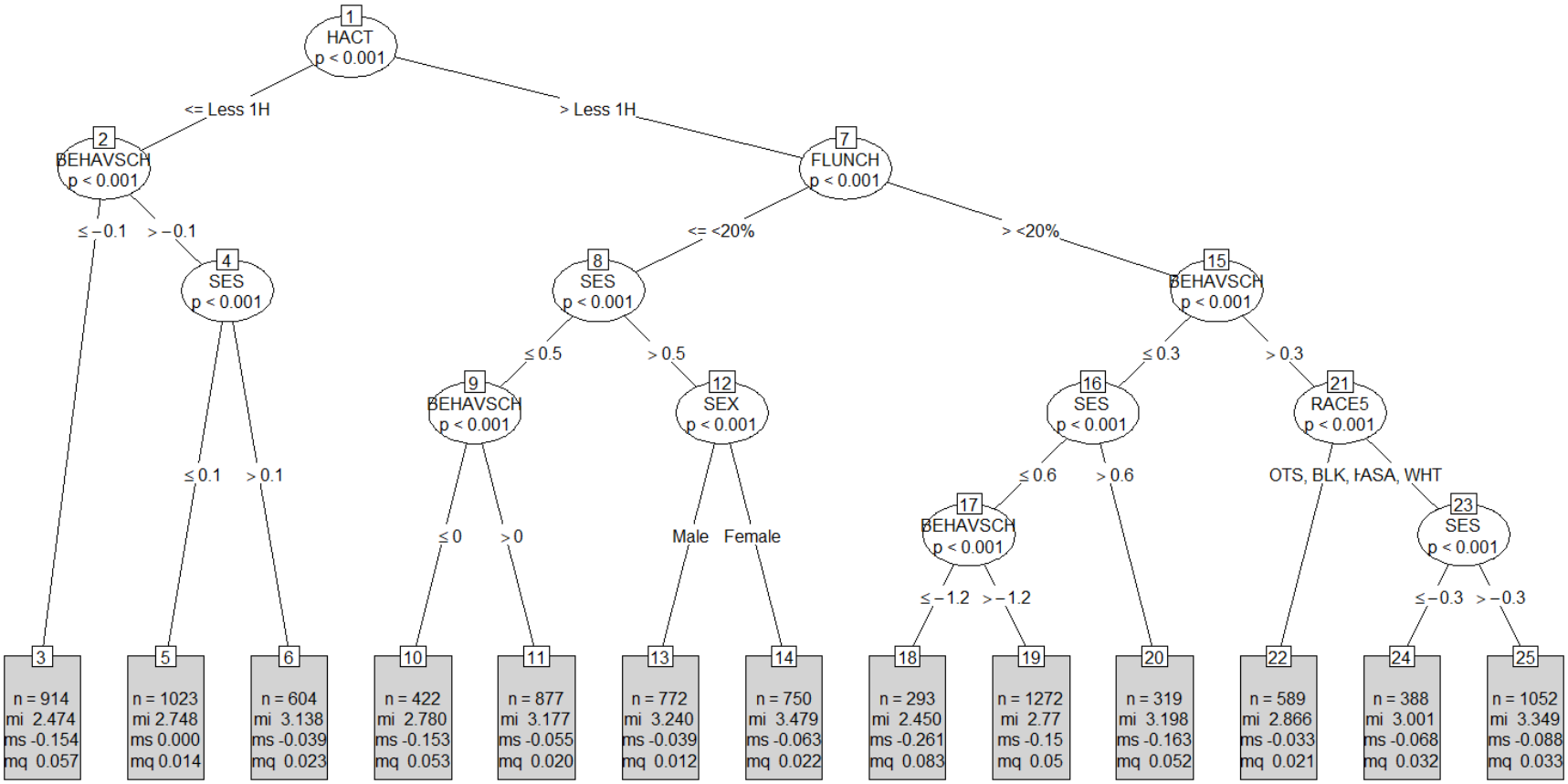


Figure 5. A LGCM tree of MOB with minimum size of 250 per node and pruning of BIC using WDM_O statistic (BLK = Black, HIP = Hispanic, OTS = Others, ASA = Asian, WHT = White, mi = μ_I; mean intercept, ms = μ_S; mean linear slope, mq = μ_Q; mean quadratic slope).

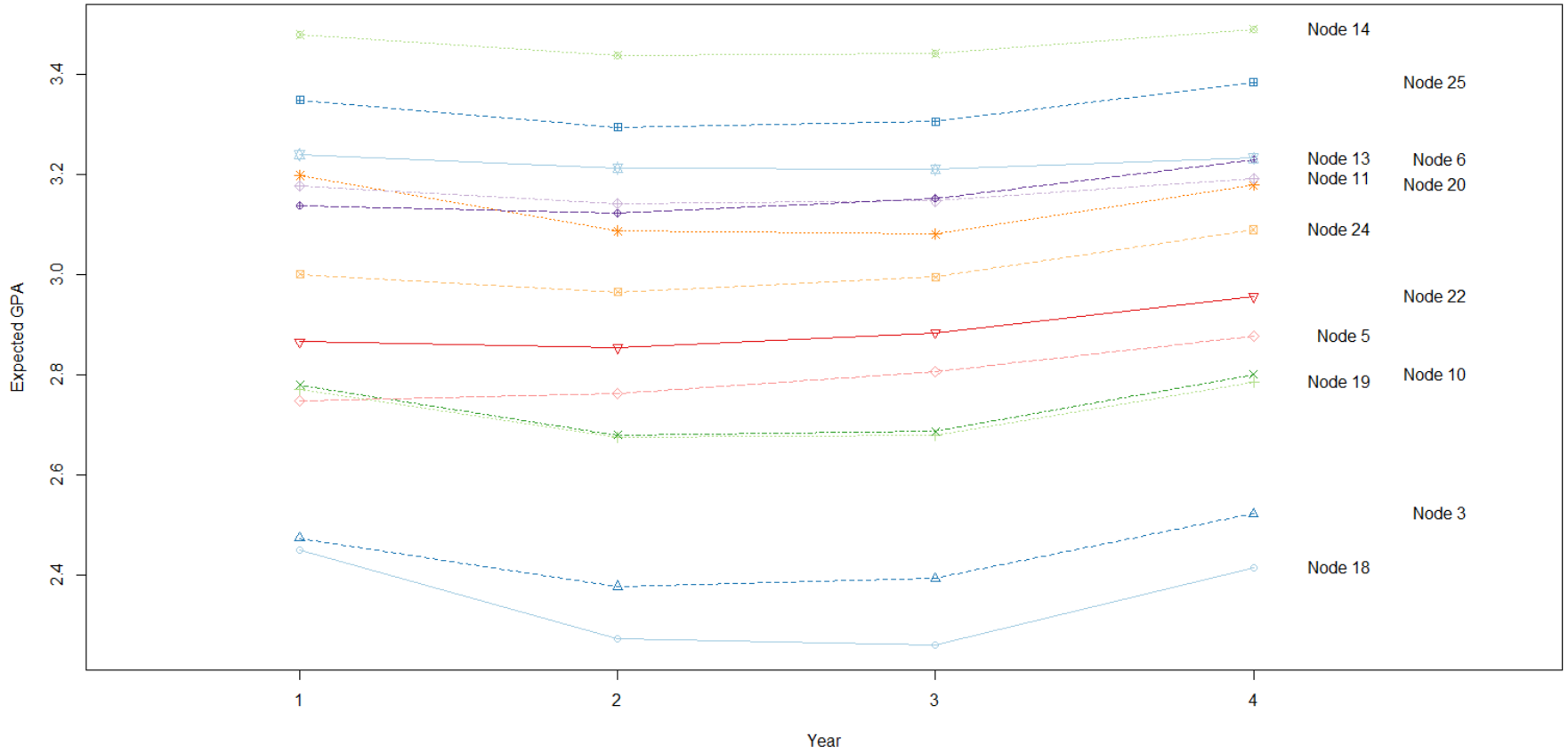


Figure 6. Expected GPA changes over four years of the 13 distinct subgroups (nodes).

CHAPTER 4. METHODS

This chapter describes how to evaluate the performance of model-based recursive partitioning (MOB) with latent growth curve model (LGCM) under a population model. The simulated datasets are also analyzed using growth mixture model (GMM) to compare the results each other. The data generation and analysis of MOB with LGCM were conducted using R 4.0.5 software (R Core Team, 2021) with the two suggested packages of `partykit` (Hothorn & Zeileis, 2015) and `lavaan` (Rosseel, 2012). Also, `MplusAutomation` (Hallquist & Wiley, 2018) and Mplus version 8 (Muthén & Muthén, 1998-2017) were used together in R software to fit GMM.

4.1 Population model for data generation

A plausible population model was chosen to simulate datasets based on the result of the empirical analysis described in Chapter 3. The values of the parameters are presented in Table 7. This study only considers the differences of the mean intercept because the results of the empirical analysis showed that most of the subgroups had similar change trajectories across time. Moreover, the random effects of the subgroups were very similar to each other (see Table 6). Thus, this study tries to mimic the empirical results.

The number of true subgroups is set to four. The four subgroups are determined by three covariates having interaction effect. This study presumes that the four subgroups have only different intercepts depending on the effect size with little variance. There are two reasons for this scenario. The first reason of considering one population model is that the purpose of this study is to investigate the performance of MOB with LGCM under a few different conditions

and other available options for researcher need to choose. This helps ones to not only interpret the results more concisely, but also focus on the practical options that needed to be investigated. The second reason is that the mean intercept difference dramatically increased when the amount of the random effect of the intercept was large. This is unrealistic in practice. Therefore, the random effect is set to small imitating the empirical results as well. All the remaining parameters including the linear and quadratic slopes as well as their variance-covariance components were assumed to be the same across four groups for the purpose of study.

Table 7. Parameters of a population model

Parameters		Coefficients
Means	Intercept (μ_I)	2.800
	Linear slope (μ_S)	-0.100
	Quadratic slope (μ_Q)	0.060
Variances	Intercept (σ_I^2)	0.471
	Linear slope (σ_S^2)	0.057
	Quadratic slope (σ_Q^2)	0.007
Covariances	I ~ S (σ_{IS})	0.001
	I ~ Q (σ_{IQ})	-0.012
	S ~ Q (σ_{SQ})	-0.016
Variances	Residuals (σ_ε^2)	0.084

The mean intercept was varied across subgroups systematically using the Cohen's effect size to enhance interpretation. Based on the values of the effect size, the mean intercept values were calculated to have different values. The Cohen's effect size is obtained by the mean difference between two groups ($\mu_{1I} - \mu_{2I}$) divided by the pooled standard deviation (σ_I) as the equation (4-1) where *I* indicates the intercept.

$$Cohen's\ d = \frac{\mu_{1I} - \mu_{2I}}{\sigma_I}. \quad (4\text{-}1)$$

Since the effect size is ranged from 0.2 to 1.0, their corresponding mean intercept differences are 0.14, 0.27, 0.41, 0.55, and 0.69. Thus, the expected GPA score changes of the four subpopulations over four years can be represented as Figure 7. While the mean intercept is different across four subgroups, the change trajectory is not different from each other. The specific population mode with the parameters to simulate datasets for each group across the effect size is presented in Table 8.

Table 8. Parameters of fixed effects for each subgroup depending on effect size

Effect size	Subgroup	Intercept (μ_I)	Linear slope (μ_S)	Quadratic slope (μ_Q)
0.2	G1	2.660	-0.100	0.060
	G2	2.800	-0.100	0.060
	G3	2.940	-0.100	0.060
	G4	3.080	-0.100	0.060
0.4	G1	2.530	-0.100	0.060
	G2	2.800	-0.100	0.060
	G3	3.070	-0.100	0.060
	G4	3.340	-0.100	0.060
0.6	G1	2.390	-0.100	0.060
	G2	2.800	-0.100	0.060
	G3	3.210	-0.100	0.060
	G4	3.620	-0.100	0.060
0.8	G1	2.250	-0.100	0.060
	G2	2.800	-0.100	0.060
	G3	3.350	-0.100	0.060
	G4	3.900	-0.100	0.060
1	G1	2.110	-0.100	0.060
	G2	2.800	-0.100	0.060
	G3	3.490	-0.100	0.060
	G4	4.180	-0.100	0.060

Note. G1-G4 are names of subgroups. The variance-covariance components and residuals are the same across subgroups as the population model.

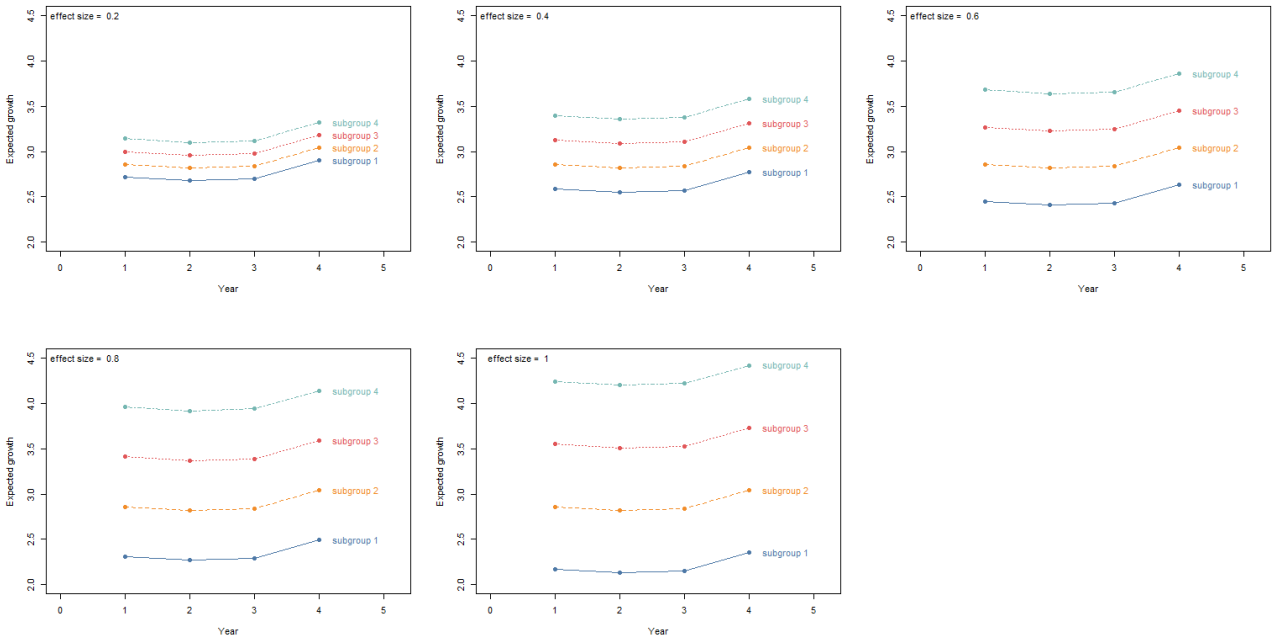


Figure 7. The expected GPA changes over four years of the four subpopulations depending on effect size.

The four repeated outcomes are generated by the quadratic latent growth curve model for each subgroup. Three informative covariates are used to differentiate the four subgroups showing interaction effects of the covariates. Three types of covariates, which are categorical, ordinal, and continuous covariates, are used. In addition, four noise variables (two continuous, one ordinal and one categorical variables) are generated regardless of the subgroups. A true tree structure of the four subgroups with the cut-points is visualized in Figure 8. Within the square boxes at the bottom of the tree, all the parameter values of the specific subgroup are presented. Looking at the tree, the first splitting covariate is the ordinal covariate with the cut point of 2. The first subgroup from the left has the lowest mean intercept of 2.660, and this group has an interaction between the ordinal and continuous covariates. That is, if the informative ordinal covariate is less than or equal to 2, the whole samples are divided into subsamples first, then the

subsamples are again divided into two subsamples if the informative continuous covariate is less than or equal to -0.7. If the informative continuous covariate is more than -0.7, the second subgroup from the left has higher mean intercept than the first subgroup. Back to the top of the tree, if the informative ordinal covariate is larger than 2, the subsamples are divided into the right part of the tree. Then, the subsamples are divided into two subgroups based on the values of the informative categorical covariate. The splitting points of these informative covariates are the same across simulation conditions.

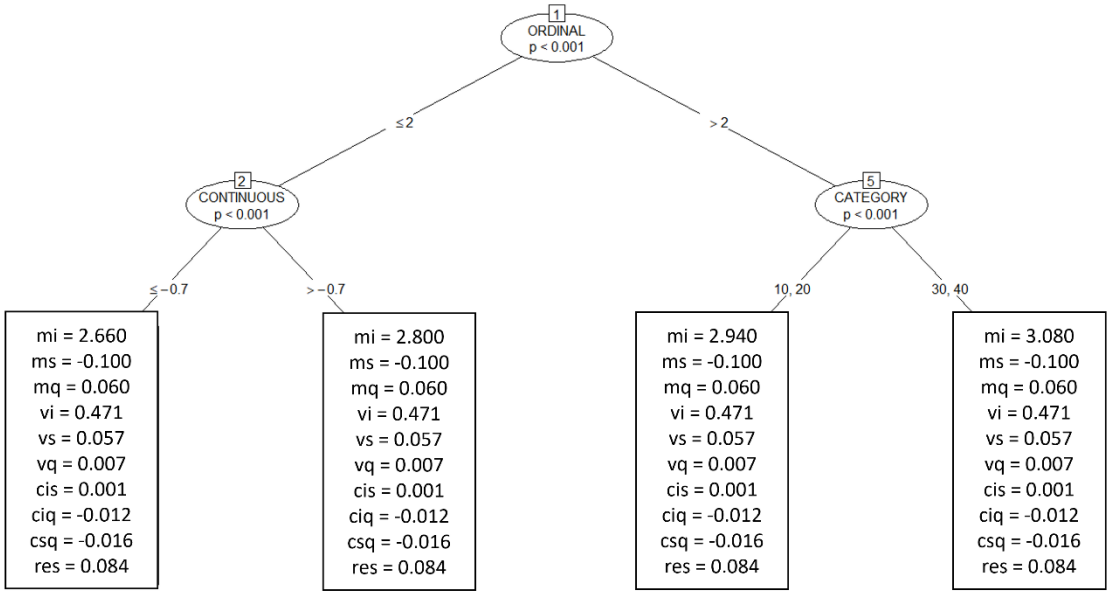


Figure 8. A true tree structure of four subgroups using the effect size of 0.2.

4.2 Simulation design

This study considered five simulation conditions; a) effect size (0.2, 0.4, 0.6, 0.8, and 1.0), b) sample size (1,000, 2,000, 5,000, 10,000, and 20,000), c) treatment of ordinal covariate with different test statistic (chi-square test statistic, adapted maximum Lagrange multiplier, and a weighted double maximum), d) minimum sample size per subgroup (250 vs. none), and e) post-pruning option (BIC vs. none). Crossing the conditions fully (5 x 5 x 3 x 2 x 2), there are 300 conditions. The number of replications was 100. Since the conditions of c), d), and e) are the options of MOB, 2,500 (5 x 5 x 100) datasets in total were simulated and analyzed by MOB with LGCM and GMM, respectively.

The third purpose of this study is to investigate how well GMM extracts the true number of subgroups (latent class in GMM) with the same datasets generated from the population model. Unconditional GMM without covariates was fitted to the data to enumerate the number of latent classes (subgroups). The covariates were not used to predict the latent classes because it is infeasible to make all possible interaction effects among the three informative and four noninformative covariates, which was considered for the population model. Moreover, GMM requires several more steps to find the best fitting model among multiple candidate models, which has one, two, three, four, or five classes. In addition, it is suggested to increase the number of initial random starts and final stage optimization to avoid local maxima solutions, ensuring that the best likelihood was replicated. Thus, the number of 1,000 and the number of 100 were chosen for the initial random starts and the final stage optimization, respectively. In sum, five candidate models were fitted to the same data and the best fitting model was chosen with the lowest value of BIC and significant statistical testing of Lo-Mendell-Rubin likelihood ratio test (Chen et al., 2017; Lo et al., 2001; Nylund et al., 2007). That is, if a

model shows the lowest BIC value as well as the significant LRT result which the p-value is less than 0.05, the model was chosen to be the best fitting model and the relevant estimates were stored to be evaluated. Entropy was also examined, but it was not considered to determine the number of classes.

4.3 Evaluation criteria

For each condition, a) recovery of the true number of subgroups, b) overall classification accuracy and precision of the subgroups, c) accuracy of the splitting points of the covariates, d) average bias and root mean squared error (RMSE) of focal fixed effects, which is the mean intercept estimate, and e) desirable options for test statistic of the ordinal covariates, post pruning method using BIC, and limiting minimum sample size per a subgroup, were evaluated.

First, the recovery of the true number of subgroups was evaluated using two statistics, the mean number of estimated subgroups (*MNS*) and the statistical power (*SP*) to correctly recover the true number of subgroups (C_T) among the number of the estimated subgroups ($\hat{C}_r$). SP can be calculated as the sum of the number of estimated subgroups that is equal to the true number of subgroups (P), where *r* is the *r*th replication and *R* is the total number of replications. While the former statistic informs whether the number of subgroups is overestimated or underestimated, the latter one informs the power to recover the true number of subgroups. These statistics, however, do not tell whether the composition of the subgroups is correctly recovered or not.

$$MNS = \frac{1}{R}\sum_{r=1}^{R=100} \hat{C}_r \tag{4-2}$$

$$SP = \sum_{r=1}^{R=100} |\hat{C}_r \in C_T| \quad (4\text{-}3)$$

Second, overall classification accuracy and precision of the subgroups were calculated using the information from a confusion table presented in Table 4.3. The confusion matrix provides a tabular summary of the actual subgroup labels versus the predicted ones. Since this study has the four number of subgroups, the confusion table consists of 4 x 4 matrix. Let N be the total number of samples in the confusion table. $Diag$ is defined as the number of correctly classified samples per subgroup. This produces a vector of $Diag = [N11, N22, N33, N44]$. The number of samples per subgroup is a vector of $TN = [TN1, TN2, TN3, TN4]$ and the number of predictions per subgroup is a vector of $PN = [PN1, PN2, PN3, PN4]$.

Table 9. Confusion table

		Predicted subgroup				
		G1	G2	G3	G4	Total
True subgroup	G1	N11	N12	N13	N14	TN1
	G2	N21	N22	N23	N24	TN2
	G3	N31	N32	N33	N34	TN3
	G4	N41	N42	N43	N44	TN4
	Total	PN1	PN2	PN3	PN4	*N*

Note. G1-G4 are names of subgroups. N11-N44 are the number of samples in each cell.

The overall classification accuracy is regarded as a metric to evaluate overall performance of the model. This can be calculated as the total number of correct predictions divided by the total number of predictions made for the dataset, which represented in the (4-4) equation. The prediction is the number of positive subgroup predictions that actually belong to the positive subgroup. This is obtained by the equation (4-5), which is the $Diag$ divided by PN.

The recall is the number of positive subgroup predictions made out of all positive individuals in the dataset. This is obtained by the equation (4-6), which is the $Diag$ divided by TN. The F1 is a single score that balances both the concerns of precision and recall in one number. This is the harmonic average of recall and precision. This can be obtained by two times precision multiplied by recall divided by the sum of the precision and recall as presented in the equation (4-7). All metrics are ranged between 0 and 1.

The precision, recall, and F1 score are the subgroup-specific metrics, which produces vectors having four values, respectively. Since this study aims to evaluate the overall performance of the classification, those three metrics (precision, recall, and F1) are *averaged* over four subgroups rather than using subgroup-specific metrics, resulting in macro-averaged precision, recall, and F1 score from the equations (4-8) - (4-10).

$$Accuracy = \frac{\sum Diag}{N} = \frac{N11 + N22 + N33 + N44}{N}, \tag{4-4}$$

$$Precision = \frac{Diag}{PN}, \tag{4-5}$$

$$Recall = \frac{Diag}{TN}, \tag{4-6}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + recall}, \tag{4-7}$$

$$Macro\ Averaged\ Precision\ (AP) = \sum Precision * \frac{1}{4} \tag{4-8}$$

$$Macro\ Averaged\ Recall\ (AR) = \sum Recall * \frac{1}{4}, \tag{4-9}$$

$$Macro\ Averaged\ F1\ (MAF1) = \sum F1 * \frac{1}{4}, \tag{4-10}$$

Then, the evaluation was conducted by calculating the average of each metric over the replication for the simulation study, which are the mean accuracy, mean macro averaged precision, mean macro averaged recall, and mean macro averaged F1 represented by the equations (4-11) - (4-14).

$$Mean\ Accuracy\ (MA) = \frac{1}{R}\sum_{r=1}^{R=100} Accuracy_r \tag{4-11}$$

$$Mean\ Macro\ Averaged\ Precision\ (MAP) = \frac{1}{R}\sum_{r=1}^{R=100} AP_r \tag{4-12}$$

$$Mean\ Macro\ Averaged\ Recall\ (MAR) = \frac{1}{R}\sum_{r=1}^{R=100} AR_r \tag{4-13}$$

$$Mean\ Macro\ Averaged\ F1\ (MAF1) = \frac{1}{R}\sum_{r=1}^{R=100} AF1_r \tag{4-14}$$

Third, it is also important to make sure if the composition of the estimated subgroups is correct. This is evaluated by checking the splitting points of the covariates. The accuracy of the splitting (cut) points can be calculated using an average of split point where *S* stands for splitting point.

$$MS = \frac{1}{R}\sum_{r=1}^{R=100} \hat{S}_r \tag{4-8}$$

Fourth, it is necessary to examine the parameter estimates of each subgroup if the

estimates are unbiased and efficient. This also can be evaluated using bias and RMSE of the estimates. Since the population model focuses on the differences of the mean intercept, only the mean intercept was evaluated using below two equations.

$$Bias_E = \frac{1}{R}\sum_{r=1}^{R=100} \hat{\theta}_r - \theta_T \tag{4-9}$$

$$RMSE_E = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \theta_T\right)^2} \tag{4-10}$$

Fifth, a simulation is parallelly conducted with unconditional GMM using the same datasets, and the results of MOB are compared with the ones of GMM. The insights from the comparative study reveal how different approaches can be useful.

CHAPTER 5. RESULTS

This chapter presents the results of the simulation study organized into six sections corresponding to the research questions described in Chapter 1. The first four sections show the results of the comprehensive simulation studies that investigate the performance of MOB with LGCM from a certain population model under considered conditions. Statistical power to detect the true number of subgroups, classification accuracy and precision, accuracy of splitting point of the covariates, and bias and root mean squared error of parameter estimates are presented. The fifth section reveals desirable several options of test statistic for ordinal covariates, post pruning option using BIC, and setting the minimum sample size per a subgroup. The last section briefly presents the results of comparison between MOB and GMM primarily focusing on the number of subgroups.

5.1 Statistical power to recover the true number of subgroups

This section is to answer the second research question that how well MOB correctly determine the true number of subgroups for a given population model. Table 10 shows the results of the average number of the estimated subgroups and the statistical power to recover the true number of subgroups. The analysis was conducted using WDM_O test statistic that is the weighted double maximum for ordinal covariate. The results show that as the magnitude of effect size and the sample size increase, the statistical power increases with the post pruning method using BIC. The highlighted cells in the table with green color indicate 95-100% of recovery rate of the true number of subgroups. Without the pruning option using BIC, there was a tendency to over extract the number of subgroups. Even without setting up the minimum sample size per subgroup, using BIC as the pruning method worked to recover the true number of subgroups

with conditions that the effect size is larger than or equal to 0.4 and the sample size is larger than or equal to 10,000.

With the small effect size of 0.2, the sample sizes of 1,000 and 2,000 were not sufficient to recover the true number of subgroups. The average number of estimated subgroups is ranged from 1.00 to 2.15 in this condition. As the sample sizes increase to 5,000, 10,000, and 20,000, the average number of the estimated subgroup and the power increase as well. However, the power was ranged from 75% to 84%, which was not enough to correctly detect the true number of subgroups. When the effect size is small, post pruning option makes less the number of subgroups, which the power is 0%.

With the conditions of medium effect size of 0.4 and 0.6 using BIC pruning option, the sample sizes of 10,000 and 5,000, respectively, were sufficient to recover the true number of subgroups. The mean estimated number of subgroups with the sample size of 2,000 and the minimum sample size of 250 without the BIC pruning option was 4.0, which is the same with the true number of subgroups, and its power to recover it was 94%. With the large effect sizes of 0.8 and 1.0, the results with sample sizes of 1,000 and 2,000 show that the true number of subgroups without pruning and with the minimum sample size of 250 was perfectly recovered. This feature of options will be described in the fifth section in detail.

Table 10. Mean of estimated number of subgroups (MNS) and statistical power (P) to recover true number of subgroups using WDM_O statistics

Effect Size	Min #	Pruning	N=1,000		N=2,000		N=5,000		N=10,000		N=20,000	
			MNS	P	MNS	P	MNS	P	MNS	P	MNS	P
0.2	No	No	2.13	1	2.15	2	3.47	35	4.27	78	4.24	80
		BIC	1.00	0	1.00	0	2.00	0	2.00	0	2.00	0
	250	No	2.02	0	2.09	0	3.50	41	4.18	84	4.33	75
		BIC	1.00	0	1.03	0	2.00	0	2.00	0	2.00	0
0.4	No	No	2.48	8	4.13	84	4.17	86	4.16	85	4.19	84
		BIC	2.00	0	2.00	0	2.04	0	4.00	100	4.00	100
	250	No	2.88	17	4.00	94	4.21	81	4.11	90	4.14	87
		BIC	2.00	0	2.00	0	2.05	0	4.00	100	4.00	100
0.6	No	No	4.16	84	4.22	83	4.23	80	4.27	79	4.27	78
		BIC	2.00	0	2.11	1	4.00	100	4.00	100	4.00	100
	250	No	3.98	98	4.03	97	4.15	87	4.23	84	4.14	87
		BIC	2.00	0	2.12	3	4.00	100	4.00	100	4.00	100
0.8	No	No	4.16	84	4.32	74	4.21	82	4.23	78	4.18	84
		BIC	2.17	2	4.00	100	4.00	100	4.00	100	4.00	100
	250	No	4.00	100	4.02	98	4.19	85	4.12	89	4.21	83
		BIC	2.15	2	4.00	100	4.00	100	4.00	100	4.00	100
1.0	No	No	4.12	88	4.26	79	4.25	77	4.22	84	4.19	85
		BIC	3.99	99	4.00	100	4.00	100	4.00	100	4.00	100
	250	No	4.00	100	4.00	100	4.17	84	4.27	78	4.17	83
		BIC	4.00	100	4.00	100	4.00	100	4.00	100	4.00	100

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.
MNS = mean number of estimated subgroups. P = statistical power.

5.2 Overall classification accuracy and precision

This section is to answer the third research question that how well MOB accurately and precisely classify the true subgroups for a given population model. Tables 11 - 15 show the results of the classification accuracy, macro-averaged precision, recall, and F1 score using WDM_O test statistic which is the weighted double maximum for ordinal covariate. The tables are presented in the order of effect size. Since the values of recall and F1 score were very similar to the accuracy and precision, this section focuses on describing the accuracy and precision.

First, looking at the Table 11, most of the means of classification accuracy is below 0.90. The maximum of the classification accuracy was 0.90 with the sample size of 20,000 without both minimum sample size and BIC pruning option. Some cells with the sample sizes of 1,000 and 2,000 empty because those conditions did not even recover the four true subgroups once. As sample size increases, the classification accuracy, precision, recall, and F1 score tend to increase. When the effect size was 0.2, the classification accuracy and precision were around 0.90 and 0.98, respectively without BIC pruning option. Even though the power to recover the true number of subgroups was ranged from 75 - 84% without the BIC pruning option for the sample sizes of 10,000 and 20,000 in this condition, their classification accuracy and precision were ranged from 0.89 - 0.99.

Second, Table 12 shows the results of the effect size of 0.4. For the cells of perfectly recovering the true number of subgroups, their classification accuracy and precision were also 100%. These conditions are the sample sizes of 10,000 and 20,000 with the BIC pruning option. Interestingly, classification accuracy and precision were 0.97 and 0.98, respectively, for the sample size of 2,000 using minimum sample size option without pruning method. As the effect

size increases from 0.4 to 0.6, the classification accuracy and precision tend to increase as well under relatively small sample sizes, such as 1,000 and 2,000 with the option of setting minimum sample size of 250. As presented in Table 13, the performance of MOB in terms of the classification accuracy and precision was almost perfect under smaller sample sizes (1,000 and 2,000) with the option of setting the minimum sample size of 250 rather than with BIC pruning option. However, the performance was reversed under larger sample sizes (5,000 - 20,000) with the BIC pruning option regardless of the minimum sample size of 250.

Third, the results of large effect sizes (0.8 and 1.0) are presented in Tables 14 and 15, respectively. There were conditions that perfectly classified the subgroups with smaller sample sizes, such as 1,000 and 2,000. For the sample size of 1,000, MOB performed perfectly with the option of minimum sample size of 250 without BIC pruning when the effect sizes were 0.8 and 1.0. As the sample size increases to 2,000, MOB performs perfectly except one condition that there are no options of setting minimum sample size and BIC pruning. As the sample size increases more than 5,000, MOB requires pruning option using BIC to perfectly classify the subgroups. This is because MOB tends to over-estimate the number of subgroups without pruning with the larger sample size as presented in Table 10.

Table 11. Means of classification accuracy (MA), macro-averaged precision (MAP), recall (MAR), and F1 (MAF1) using $\mathrm{WDM_O}$ statistics (Effect size=0.2)

Sample Size	Min #	Pruning	Effect Size = 0.2			
			MA	MAP	MAR	MAF1
1,000	No	No	0.27	0.27	0.51	0.51
		BIC	0.25	0.25	-	-
	250	No	0.25	0.25	0.50	0.50
		BIC	0.25	0.25	-	-
2,000	No	No	0.27	0.27	0.51	0.51
		BIC	0.25	0.25	-	-
	250	No	0.27	0.27	0.52	0.52
		BIC	0.25	0.25	-	-
5,000	No	No	0.66	0.72	0.79	0.79
		BIC	0.25	0.25	0.50	0.50
	250	No	0.69	0.74	0.81	0.81
		BIC	0.25	0.25	0.50	0.50
10,000	No	No	0.89	0.97	0.89	0.89
		BIC	0.25	0.25	0.50	0.50
	250	No	0.90	0.95	0.90	0.90
		BIC	0.25	0.25	0.50	0.50
20,000	No	No	0.90	0.99	0.90	0.90
		BIC	0.25	0.25	0.50	0.50
	250	No	0.87	0.95	0.87	0.87
		BIC	0.25	0.25	0.50	0.50

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

Table 12. Means of classification accuracy (MA), macro-averaged precision (MAP), recall (MAR), and F1 (MAF1) using WDM_O statistics (Effect size=0.4)

Sample Size	Min #	Pruning	Effect size = 0.4			
			MA	MAP	MAR	MAF1
1,000	No	No	0.35	0.35	0.58	0.58
		BIC	0.25	0.25	0.50	0.50
	250	No	0.61	0.64	0.82	0.82
		BIC	0.25	0.25	0.50	0.50
2,000	No	No	0.91	0.95	0.92	0.92
		BIC	0.25	0.25	0.50	0.50
	250	No	0.97	0.98	0.97	0.97
		BIC	0.25	0.25	0.50	0.50
5,000	No	No	0.92	0.97	0.92	0.92
		BIC	0.27	0.27	0.52	0.52
	250	No	0.92	0.97	0.92	0.92
		BIC	0.26	0.25	0.51	0.51
10,000	No	No	0.93	0.96	0.93	0.93
		BIC	1.00	1.00	1.00	1.00
	250	No	0.94	0.96	0.94	0.94
		BIC	1.00	1.00	1.00	1.00
20,000	No	No	0.91	0.95	0.91	0.91
		BIC	1.00	1.00	1.00	1.00
	250	No	0.92	0.96	0.92	0.92
		BIC	1.00	1.00	1.00	1.00

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

Table 13. Means of classification accuracy (MA), macro-averaged precision (MAP), recall (MAR), and F1 (MAF1) using WDM_O statistics (Effect size=0.6)

Sample Size	Min #	Pruning	Effect size=0.6			
			MA	MAP	MAR	MAF1
1,000	No	No	0.88	0.92	0.89	0.89
		BIC	0.25	0.25	0.50	0.50
	250	No	1.00	1.00	1.00	1.00
		BIC	0.25	0.25	0.50	0.50
2,000	No	No	0.90	0.95	0.90	0.90
		BIC	0.29	0.29	0.54	0.54
	250	No	0.98	0.99	0.98	0.98
		BIC	0.29	0.29	0.53	0.53
5,000	No	No	0.90	0.94	0.90	0.90
		BIC	1.00	1.00	1.00	1.00
	250	No	0.93	0.96	0.93	0.93
		BIC	1.00	1.00	1.00	1.00
10,000	No	No	0.87	0.93	0.87	0.87
		BIC	1.00	1.00	1.00	1.00
	250	No	0.91	0.95	0.91	0.91
		BIC	1.00	1.00	1.00	1.00
20,000	No	No	0.89	0.94	0.89	0.89
		BIC	1.00	1.00	1.00	1.00
	250	No	0.92	0.99	0.92	0.92
		BIC	1.00	1.00	1.00	1.00

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

Table 14. Means of classification accuracy (MA), macro-averaged precision (MAP), recall (MAR), and F1 (MAF1) using $\mathrm{WDM_O}$ statistics (Effect size=0.8)

Sample Size	Min #	Pruning	Effect size=0.8			
			MA	MAP	MAR	MAF1
1,000	No	No	0.91	0.95	0.91	0.91
		BIC	0.29	0.29	0.54	0.54
	250	No	1.00	1.00	1.00	1.00
		BIC	0.30	0.30	0.55	0.55
2,000	No	No	0.89	1.00	0.89	0.89
		BIC	1.00	1.00	1.00	1.00
	250	No	0.98	0.99	0.98	0.98
		BIC	1.00	1.00	1.00	1.00
5,000	No	No	0.92	0.98	0.92	0.92
		BIC	1.00	1.00	1.00	1.00
	250	No	0.89	0.92	0.89	0.89
		BIC	1.00	1.00	1.00	1.00
10,000	No	No	0.90	0.96	0.90	0.90
		BIC	1.00	1.00	1.00	1.00
	250	No	0.95	0.97	0.95	0.95
		BIC	1.00	1.00	1.00	1.00
20,000	No	No	0.92	0.97	0.92	0.92
		BIC	1.00	1.00	1.00	1.00
	250	No	0.91	0.99	0.91	0.91
		BIC	1.00	1.00	1.00	1.00

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

Table 15. Means of classification accuracy (MA), macro-averaged precision (MAP), recall (MAR), and F1 (MAF1) using WDM_O statistics (Effect size=1.0)

Sample Size	Min #	Pruning	Effect size=1.0			
			MA	MAP	MAR	MAF1
1,000	No	No	0.96	0.99	0.96	0.96
		BIC	1.00	1.00	1.00	1.00
	250	No	1.00	1.00	1.00	1.00
		BIC	1.00	1.00	1.00	1.00
2,000	No	No	0.87	0.93	0.87	0.87
		BIC	1.00	1.00	1.00	1.00
	250	No	1.00	1.00	1.00	1.00
		BIC	1.00	1.00	1.00	1.00
5,000	No	No	0.90	0.98	0.90	0.90
		BIC	1.00	1.00	1.00	1.00
	250	No	0.91	0.94	0.91	0.91
		BIC	1.00	1.00	1.00	1.00
10,000	No	No	0.92	0.96	0.92	0.92
		BIC	1.00	1.00	1.00	1.00
	250	No	0.87	0.92	0.87	0.87
		BIC	1.00	1.00	1.00	1.00
20,000	No	No	0.93	0.97	0.93	0.93
		BIC	1.00	1.00	1.00	1.00
	250	No	0.90	0.96	0.90	0.90
		BIC	1.00	1.00	1.00	1.00

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

5.3 Accuracy of splitting points of covariates

This section investigates how accurate the composition of the subgroups is by evaluating the mean of the splitting points of covariates. Firstly, it was investigated if the noise covariates were used for splitting at any stages. The results show that the noise covariates were not selected as the splitting covariates under all conditions. Thus, those results are not presented in this section. The averages of the splitting points of the informative covariates are presented in Table 16. If the number of the detected subgroups is not equal to four, it is not possible to calculate the means of all splitting points of the informative covariates. This is because not only is the structure of the tree different from the true tree, but also the number of replications is not sufficient to calculate the mean. Since the first splitting point is 2 for the informative (true) ordinal covariate, the means of them are calculated across replications if the number of the detected subgroups is larger than 1. If MOB did not differentiate the true subgroups, leading to produce just one subgroup, there was not the splitting point. The splitting point of the ordinal covariate is presented in the column of *O* in Table 16. The results show that the true splitting point of 2 for the true informative ordinal covariate was correctly detected for the first splitting under most of the conditions except two conditions of the effect size of 0.2 and the sample sizes of 1,000 and 2 000 with BIC pruning option.

The second true splitting is done by both a continuous covariate and a categorical covariate as described in Figure 8. The true splitting point of the continuous covariate is -0.7. The composition of the categorical covariate was also investigated though, the splitting composition was perfect if the average estimated number of subgroups was four and correctly recovered. If the number of the determined subgroups under some conditions is larger than 4, such as 5 or 6 or 7, the subgroup was divided into additional subgroups according to the values

of the continuous covariate.

Table 16. Mean of splitting points of covariates (MS) using WDM_O statistics

Effect Size	Min #	Pruning	N=1,000		N=2,000		N=5,000		N=10,000		N=20,000	
			O	C	O	C	O	C	O	C	O	C
0.2	No	No	2	-	2	-	2	-	2	-	2	-
		BIC	-	-	-	-	2	-	2	-	2	-
	250	No	2	-	2	-	2	-	2	-	2	-
		BIC	-	-	-	-	2	-	2	-	2	-
0.4	No	No	2	-	2	-	2	-	2	-	2	-
		BIC	2	-	2	-	2	-	2	-0.69	2	-0.70
	250	No	2	-	2	-	2	-	2	-	2	-
		BIC	2	-	2	-	2	-	2	-0.69	2	-0.70
0.6	No	No	2	-	2	-	2	-	2	-	2	-
		BIC	2	-	2	-	2	-0.69	2	-0.70	2	-0.70
	250	No	2	-0.70	2	-0.64	2	-	2	-	2	-
		BIC	2	-	2	-	2	-0.70	2	-0.70	2	-0.70
0.8	No	No	2	-	2	-	2	-	2	-	2	-
		BIC	2	-	2	-0.68	2	-0.69	2	-0.70	2	-0.70
	250	No	2	-0.70	2	-0.65	2	-	2	-	2	-
		BIC	2	-	2	-0.67	2	-0.70	2	-0.70	2	-0.70
1.0	No	No	2	-	2	-	2	-	2	-	2	-
		BIC	2	-0.74	2	-0.67	2	-0.69	2	-0.70	2	-0.70
	250	No	2	-0.70	2	-0.68	2	-	2	-	2	-
		BIC	2	-0.70	2	-0.70	2	-0.70	2	-0.70	2	-0.70

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.
O = splitting point of ordinal covariate. C = splitting point of continuous covariate.

5.4 Bias and RMSE of parameter estimates

To further check the unbiasedness and efficiency of the parameter estimates, bias and root mean squared error (RMSE) were calculated. Since it is infeasible to calculate the bias and RMSE if the number of estimated subgroups is not the same with the true number of subgroups, only the results of the conditions that perfectly recovered the true number of subgroups are presented in Tables 17 - 20 for bias and 21 - 24 for RMSE, respectively. Also, means of intercept, linear slope, and quadratic slope with their variances are presented in this section because the biases of covariances and residuals are close to zero across all conditions.

When the effect size is 0.4, the conditions that perfectly recovered the number of true subgroups are when the sample sizes are 10,000 and 20,000 with BIC post pruning option. Looking at the bias of these conditions in Table 17, the range of all bias in this condition is from -0.003 to 0.003. Focusing on the values of the mean intercept, the range of bias for the mean intercept is from -0.003 to 0.002. Since the values of the mean intercept parameters with the effect size of 0.4 are 2.53 (G1), 2.80 (G2), 3.07 (G3), and 3.34 (G4) for each true subgroup, the relative magnitude of bias (which is calculated by the bias divided by the corresponding parameter value times 100) is ranging from -0.09% to 0.09%. Table 21 shows the RMSE of these conditions, and the values are close to zeros.

When the effect size is 0.6, the conditions that perfectly recovered the number of true subgroups are when the sample sizes are 5,000, 10,000, and 20,000 with BIC post pruning option. Looking at the bias of these conditions in Table 18, the range of all bias in this condition is from -0.005 to 0.004. Focusing on the values of the mean intercept, the range of bias for the mean intercept is from -0.002 to 0.002. Since the values of the mean intercept parameters with the effect size of 0.6 are 2.39 (G1), 2.80 (G2), 3.21 (G3), and 3.62 (G4) for each true subgroup,

the relative magnitude of bias is ranging from -0.21% to 0.17%. Table 22 shows the RMSE in these conditions, and the values are very close to zeros.

When the effect size is 0.8, the conditions that perfectly recovered the number of true subgroups additionally included the sample sizes of 1,000 and 2,000 in certain pruning methods. Specifically, when the sample size is 1,000, setting the minimum sample size of 250 *without* the post pruning option using BIC recovered the true number of subgroups. Other conditions recovered the true number of subgroups include the post pruning option using BIC regardless of the minimum sample size of 250 across different sample sizes (2,000, 5,000, 10,000, and 20,000). Looking at the bias of these conditions in Table 19, the range of all bias in this condition is from -0.011 to 0.01. Focusing on the values of the mean intercept, the range of bias for the mean intercept is from -0.011 to 0.008. Since the values of the mean intercept parameters with the effect size of 0.8 are 2.25 (G1), 2.80 (G2), 3.35 (G3), and 3.90 (G4) for each true subgroup, the relative magnitude of bias is ranging from -0.29% to 0.24%. Table 23 shows the RMSE in these conditions, and the values are very close to zeros.

When the effect size is 1.0, the conditions that perfectly recovered the number of true subgroups included two more conditions than the ones of the effect size of 0.8. These conditions are when 1) sample size of 1,000 setting the minimum size of 250 per subgroup *with* the post pruning method using BIC, 2) sample size of 2,000 setting the minimum size of 250 per subgroup without the pruning method using BIC. Likewise, other conditions recovered the true number of subgroups include the post pruning option using BIC regardless of the minimum sample size of 250 across all sample sizes (1,000, 2,000, 5,000, 10,000, and 20,000). Looking at the bias of these conditions in Table 20, the range of all bias in these conditions is from -0.012 to 0.013. Focusing on the values of the mean intercept, the range of bias for the mean intercept is

from -0.009 to 0.008. Since the values of the mean intercept parameters with the effect size of 1.0 are 2.11 (G1), 2.80 (G2), 3.49 (G3), and 4.18 (G4) for each subgroup, the relative magnitude of bias is ranging from -0.38% to 0.24%. Table 24 shows the RMSE in these conditions, and the values are very close to zeros.

Table 17. Bias of parameter estimates of conditions perfectly recovered true subgroups using $\mathrm{WDM_O}$ statistics (Effect size=0.4)

N	Min #	Pruning	Subgroups	mi	ms	mq	vi	vs	vq
10,000	No	BIC	G1	0.002	-0.001	0.000	0.003	-0.002	0.000
			G2	0.002	0.001	0.000	0.002	0.001	0.000
			G3	0.002	-0.001	0.000	0.003	0.001	0.000
			G4	0.001	0.001	0.000	0.000	0.002	0.000
	250	BIC	G1	-0.002	0.000	0.000	0.002	-0.002	0.000
			G2	-0.001	-0.001	0.000	0.001	-0.002	0.000
			G3	-0.002	0.000	0.000	0.002	-0.002	0.000
			G4	-0.003	0.002	0.000	0.003	0.001	0.000
20,000	No	BIC	G1	0.001	0.000	0.000	0.002	0.000	0.000
			G2	0.002	-0.002	0.000	0.001	0.000	0.000
			G3	0.002	0.000	0.000	0.000	0.001	0.000
			G4	0.001	0.001	0.000	0.000	-0.001	0.000
	250	BIC	G1	0.001	0.000	0.000	-0.001	0.002	0.000
			G2	0.001	0.001	0.000	0.000	0.000	0.000
			G3	0.001	0.001	0.000	-0.001	0.000	0.000
			G4	0.001	-0.001	0.000	-0.001	0.000	0.000

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
mi = mean intercept. ms = mean linear slope. mq = mean quadratic slope.
vi = variance of intercept. vs = variance of linear slope. vq = variance of quadratic slope.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

Table 18. Bias of parameter estimates of conditions perfectly recovered true subgroups using $\mathrm{WDM_O}$ statistics (Effect size=0.6)

N	Min #	Pruning	Subgroups	mi	ms	mq	vi	vs	vq
5,000	No	BIC	G1	0.004	0.001	-0.001	0.003	0.001	0.000
			G2	0.003	0.000	0.000	0.004	-0.001	0.000
			G3	0.001	0.000	0.000	0.003	-0.001	0.000
			G4	0.001	0.002	-0.001	0.002	-0.004	0.000
	250	BIC	G1	-0.005	0.001	0.000	0.001	-0.001	0.000
			G2	-0.001	-0.001	0.000	0.004	0.000	0.000
			G3	-0.002	-0.002	0.001	0.003	-0.004	0.000
			G4	-0.004	0.001	0.000	-0.001	0.003	0.000
10,000	No	BIC	G1	0.001	-0.002	0.001	-0.001	0.000	0.000
			G2	0.000	0.000	0.000	-0.001	0.000	0.000
			G3	0.002	-0.002	0.001	-0.002	0.000	0.000
			G4	-0.001	0.003	-0.001	-0.004	0.000	0.000
	250	BIC	G1	-0.001	-0.001	0.001	-0.003	0.002	0.000
			G2	-0.001	0.000	0.000	-0.002	-0.001	0.000
			G3	-0.002	0.001	0.000	-0.005	0.001	0.000
			G4	0.000	-0.001	0.000	-0.004	0.000	0.000
20,000	No	BIC	G1	0.002	0.001	0.000	-0.002	0.000	0.000
			G2	0.003	0.001	0.000	-0.001	0.000	0.000
			G3	0.004	0.001	0.000	-0.001	0.000	0.000
			G4	0.003	0.001	-0.001	0.000	0.000	0.000
	250	BIC	G1	0.000	0.000	0.000	-0.001	-0.002	0.000
			G2	-0.001	-0.001	0.000	0.000	0.001	0.000
			G3	-0.001	0.001	0.000	0.000	0.000	0.000
			G4	-0.001	-0.001	0.000	-0.002	-0.001	0.000

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
mi = mean intercept. ms = mean linear slope. mq = mean quadratic slope.
vi = variance of intercept. vs = variance of linear slope. vq = variance of quadratic slope.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

Table 19. Bias of parameter estimates of conditions perfectly recovered true subgroups using WDM_O statistics (Effect size=0.8)

N	Min #	Pruning	Subgroups	mi	ms	mq	vi	vs	vq
1000	250	No	G1	0.004	0.005	-0.001	-0.004	-0.002	0.000
			G2	0.005	0.006	-0.002	0.010	0.001	0.000
			G3	0.008	0.002	0.000	-0.003	-0.002	0.000
			G4	0.007	0.005	-0.002	0.003	-0.005	0.000
2000	No	BIC	G1	-0.006	-0.002	0.001	0.000	-0.003	0.000
			G2	-0.008	0.000	0.001	0.000	0.000	0.000
			G3	-0.005	0.002	-0.001	0.001	-0.004	0.000
			G4	-0.011	0.006	-0.002	0.002	-0.005	-0.001
	250	BIC	G1	-0.002	-0.005	0.002	-0.007	-0.006	-0.001
			G2	-0.004	-0.003	0.002	-0.006	-0.001	0.000
			G3	0.001	-0.010	0.003	-0.008	-0.004	0.000
			G4	-0.003	-0.004	0.001	-0.004	-0.001	0.000
5000	No	BIC	G1	-0.003	0.004	-0.001	0.002	-0.001	0.000
			G2	0.000	0.002	-0.001	0.001	-0.001	0.000
			G3	0.000	0.002	-0.001	0.003	-0.002	0.000
			G4	0.000	-0.003	0.001	0.003	0.001	0.000
	250	BIC	G1	0.004	0.002	-0.001	-0.004	-0.002	0.000
			G2	0.005	-0.001	-0.001	-0.003	0.000	0.000
			G3	0.003	0.001	0.000	-0.002	0.000	0.000
			G4	0.004	0.000	0.000	-0.004	-0.001	0.000
10000	No	BIC	G1	0.001	0.000	0.000	0.001	0.000	0.000
			G2	0.000	0.003	-0.001	0.002	-0.002	0.000
			G3	0.001	0.000	0.000	0.000	0.000	0.000
			G4	-0.001	0.003	0.000	0.002	-0.003	0.000
	250	BIC	G1	-0.001	-0.002	0.001	-0.002	-0.001	0.000
			G2	0.001	-0.002	0.001	-0.001	0.001	0.000
			G3	0.000	-0.001	0.000	-0.003	0.002	0.000
			G4	0.000	-0.002	0.001	-0.004	0.001	0.000
20000	No	BIC	G1	-0.001	-0.001	0.000	0.001	0.000	0.000
			G2	-0.001	-0.001	0.000	0.001	0.000	0.000
			G3	-0.001	-0.001	0.000	0.002	0.000	0.000
			G4	-0.001	0.000	0.000	0.002	0.000	0.000
	250	BIC	G1	-0.001	-0.001	0.000	0.001	0.001	0.000
			G2	-0.001	0.000	0.000	-0.001	-0.001	0.000
			G3	0.000	-0.001	0.000	-0.001	0.000	0.000
			G4	-0.001	0.000	0.000	-0.001	0.001	0.000

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
mi = mean intercept. ms = mean linear slope. mq = mean quadratic slope.
vi = variance of intercept. vs = variance of linear slope. vq = variance of quadratic slope.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

Table 20. Bias of parameter estimates of conditions perfectly recovered true subgroups using WDM_O statistics (Effect size=1.0)

N	Min #	Pruning	Subgroups	mi	ms	mq	vi	vs	vq
1000	250	No	G1	0.004	0.001	-0.001	0.002	-0.004	0.000
		No	G2	-0.001	0.002	0.000	0.007	0.004	0.000
		No	G3	0.004	-0.007	0.003	0.000	-0.002	0.000
		No	G4	0.006	-0.012	0.004	0.004	0.001	0.000
	250	BIC	G1	-0.005	0.005	0.000	0.007	-0.002	0.000
		BIC	G2	0.001	-0.004	0.002	0.009	-0.001	0.000
		BIC	G3	-0.002	0.001	0.000	0.013	0.002	0.000
		BIC	G4	0.000	-0.004	0.002	0.003	0.002	0.000
2000	No	BIC	G1	-0.008	0.003	0.000	0.004	0.003	0.000
		BIC	G2	-0.009	0.003	-0.001	0.001	-0.002	0.000
		BIC	G3	-0.007	-0.001	0.001	0.001	-0.001	0.000
		BIC	G4	-0.007	-0.001	0.000	-0.002	0.001	0.000
	250	No	G1	-0.005	0.006	-0.002	0.004	0.000	0.000
		No	G2	-0.004	0.004	-0.001	0.000	-0.003	0.000
		No	G3	-0.002	-0.001	0.001	0.007	-0.003	0.000
		No	G4	-0.001	0.000	0.000	0.003	0.001	0.000
	250	BIC	G1	0.001	-0.005	0.002	-0.004	-0.007	0.000
		BIC	G2	-0.001	0.003	-0.001	-0.007	-0.007	-0.001
		BIC	G3	0.000	-0.003	0.001	-0.003	-0.002	0.000
		BIC	G4	-0.002	0.001	0.000	0.001	-0.003	0.000
5000	No	BIC	G1	-0.001	-0.001	0.000	-0.004	-0.002	0.000
		BIC	G2	-0.001	-0.001	0.000	-0.005	0.001	0.000
		BIC	G3	-0.001	-0.002	0.001	-0.003	-0.002	0.000
		BIC	G4	0.001	-0.005	0.001	-0.004	0.002	0.000
	250	BIC	G1	-0.001	0.001	0.000	0.003	-0.003	0.000
		BIC	G2	0.000	0.000	0.000	0.000	-0.006	0.000
		BIC	G3	0.002	-0.001	0.000	0.004	-0.002	0.000
		BIC	G4	0.000	0.001	0.000	0.001	-0.001	0.000
10000	No	BIC	G1	0.002	0.000	0.000	-0.003	0.001	0.000
		BIC	G2	0.004	-0.005	0.001	0.000	0.000	0.000
		BIC	G3	0.002	-0.002	0.000	0.000	0.002	0.000
		BIC	G4	0.002	0.000	0.000	-0.002	0.002	0.000
	250	BIC	G1	0.001	-0.002	0.001	-0.001	-0.001	0.000
		BIC	G2	0.000	-0.001	0.001	-0.002	-0.003	0.000
		BIC	G3	0.003	-0.002	0.001	0.000	-0.001	0.000
		BIC	G4	0.000	-0.001	0.000	-0.001	-0.002	0.000

Table 20 (cont'd)

20000	No	BIC	G1	0.003	0.000	0.000	0.000	0.000	0.000
		BIC	G2	0.002	0.000	0.000	0.000	0.000	0.000
		BIC	G3	0.002	-0.001	0.000	0.000	-0.001	0.000
		BIC	G4	0.003	-0.001	0.000	0.001	0.000	0.000
	250	BIC	G1	0.001	0.001	0.000	0.003	0.000	0.000
		BIC	G2	0.002	0.001	0.000	0.001	-0.001	0.000
		BIC	G3	0.003	0.000	0.000	0.003	-0.001	0.000
		BIC	G4	0.001	0.001	0.000	0.001	0.002	0.000

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
mi = mean intercept. ms = mean linear slope. mq = mean quadratic slope.
vi = variance of intercept. vs = variance of linear slope. vq = variance of quadratic slope.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

Table 21. RMSE of parameter estimates of conditions perfectly recovered true subgroups using WDM_O statistics (Effect size=0.4)

N	Min #	Pruning	Subgroups	mi	ms	mq	vi	vs	vq
10,000	No	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000
	250	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000
20,000	No	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000
	250	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
mi = mean intercept. ms = mean linear slope. mq = mean quadratic slope.
vi = variance of intercept. vs = variance of linear slope. vq = variance of quadratic slope.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

Table 22. RMSE of parameter estimates of conditions perfectly recovered true subgroups using WDM_O statistics (Effect size=0.6)

N	Min #	Pruning	Subgroups	mi	ms	mq	vi	vs	vq
5,000	No	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.001	0.000	0.000
	250	BIC	G1	0.001	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.001	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000
10,000	No	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000
	250	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000
20,000	No	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000
	250	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
mi = mean intercept. ms = mean linear slope. mq = mean quadratic slope.
vi = variance of intercept. vs = variance of linear slope. vq = variance of quadratic slope.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

Table 23. RMSE of parameter estimates of conditions perfectly recovered true subgroups using WDM_O statistics (Effect size=0.8)

N	Min #	Pruning	Subgroups	mi	ms	mq	vi	vs	vq
1000	250	No	G1	0.002	0.001	0.000	0.004	0.001	0.000
			G2	0.002	0.001	0.000	0.004	0.001	0.000
			G3	0.002	0.001	0.000	0.003	0.001	0.000
			G4	0.002	0.001	0.000	0.003	0.001	0.000
2000	No	BIC	G1	0.001	0.000	0.000	0.001	0.001	0.000
			G2	0.002	0.000	0.000	0.001	0.001	0.000
			G3	0.001	0.001	0.000	0.001	0.000	0.000
			G4	0.002	0.001	0.000	0.001	0.000	0.000
	250	BIC	G1	0.001	0.001	0.000	0.001	0.000	0.000
			G2	0.001	0.000	0.000	0.002	0.000	0.000
			G3	0.001	0.001	0.000	0.001	0.001	0.000
			G4	0.001	0.000	0.000	0.001	0.001	0.000
5000	No	BIC	G1	0.001	0.000	0.000	0.000	0.000	0.000
			G2	0.001	0.000	0.000	0.000	0.000	0.000
			G3	0.001	0.000	0.000	0.000	0.000	0.000
			G4	0.001	0.000	0.000	0.000	0.000	0.000
	250	BIC	G1	0.001	0.000	0.000	0.000	0.000	0.000
			G2	0.001	0.000	0.000	0.001	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.001	0.000	0.000	0.001	0.000	0.000
10000	No	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000
	250	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000
20000	No	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000
	250	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
			G2	0.000	0.000	0.000	0.000	0.000	0.000
			G3	0.000	0.000	0.000	0.000	0.000	0.000
			G4	0.000	0.000	0.000	0.000	0.000	0.000

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
mi = mean intercept. ms = mean linear slope. mq = mean quadratic slope.
vi = variance of intercept. vs = variance of linear slope. vq = variance of quadratic slope.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

Table 24. RMSE of parameter estimates of conditions perfectly recovered true subgroups using WDM_O statistics (Effect size=1.0)

N	Min #	Pruning	Subgroups	mi	ms	mq	vi	vs	vq
1000	250	No	G1	0.002	0.001	0.000	0.003	0.001	0.000
		No	G2	0.002	0.001	0.000	0.003	0.001	0.000
		No	G3	0.002	0.001	0.000	0.002	0.001	0.000
		No	G4	0.002	0.001	0.000	0.003	0.001	0.000
	250	BIC	G1	0.002	0.001	0.000	0.002	0.001	0.000
		BIC	G2	0.002	0.001	0.000	0.002	0.001	0.000
		BIC	G3	0.002	0.001	0.000	0.002	0.001	0.000
		BIC	G4	0.002	0.001	0.000	0.003	0.001	0.000
2000	No	BIC	G1	0.001	0.001	0.000	0.001	0.000	0.000
		BIC	G2	0.001	0.001	0.000	0.001	0.000	0.000
		BIC	G3	0.001	0.001	0.000	0.001	0.000	0.000
		BIC	G4	0.001	0.000	0.000	0.002	0.000	0.000
	250	No	G1	0.001	0.001	0.000	0.001	0.000	0.000
		No	G2	0.001	0.001	0.000	0.001	0.001	0.000
		No	G3	0.001	0.000	0.000	0.002	0.001	0.000
		No	G4	0.001	0.001	0.000	0.001	0.000	0.000
	250	BIC	G1	0.001	0.001	0.000	0.001	0.001	0.000
		BIC	G2	0.001	0.000	0.000	0.001	0.000	0.000
		BIC	G3	0.001	0.001	0.000	0.001	0.000	0.000
		BIC	G4	0.001	0.000	0.000	0.001	0.001	0.000
5000	No	BIC	G1	0.001	0.000	0.000	0.000	0.000	0.000
		BIC	G2	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G3	0.001	0.000	0.000	0.001	0.000	0.000
		BIC	G4	0.001	0.000	0.000	0.001	0.000	0.000
	250	BIC	G1	0.000	0.000	0.000	0.001	0.000	0.000
		BIC	G2	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G3	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G4	0.000	0.000	0.000	0.000	0.000	0.000
10000	No	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G2	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G3	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G4	0.000	0.000	0.000	0.000	0.000	0.000
	250	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G2	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G3	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G4	0.000	0.000	0.000	0.000	0.000	0.000

Table 24 (Cont'd)

20000	No	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G2	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G3	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G4	0.000	0.000	0.000	0.000	0.000	0.000
	250	BIC	G1	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G2	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G3	0.000	0.000	0.000	0.000	0.000	0.000
		BIC	G4	0.000	0.000	0.000	0.000	0.000	0.000

Note. WDM_O = a test statistic of weighted double maximum for ordinal covariate.
mi = mean intercept. ms = mean linear slope. mq = mean quadratic slope.
vi = variance of intercept. vs = variance of linear slope. vq = variance of quadratic slope.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.

5.5 Several desirable options

This section focuses on differences of the results stem from several available options. As I described before, three options were considered in this study: 1) three different test statistic for the ordinal covariate (LM, $maxLM_O$, and WDM_O), 2) an option of pre-pruning option whether or not to limit the minimum sample size per subgroup, and 3) an option whether or not to use post pruning using BIC. The mean of estimated number of subgroups and the statistical power to recover the true number of subgroups were presented as an order of effect size in Tables 25 - 29.

5.5.1 Test statistics of ordinal covariates

Regardless of treating the ordinal covariates as categorical or ordinal, there was not noticeable significant difference in terms of the mean of estimated number of subgroups and the statistical power to recover the true number of subgroups between three different test statistic across most of conditions except a few conditions. There were minor discrepancies between the different test statistic when the sample sizes were relatively small (1,000 and 2,000) with the

medium or large effect sizes (0.4, 0.6, 0.8, 1.0).

First, Table 26 shows the results of the effect size of 0.4. When the sample size was 2,000 with limiting the minimum size per subgroup without post pruning using BIC, the means of the estimated subgroups were 3.96 for the LM (test statistic for categorical covariate), 4.02 for the $maxLM_O$(LM test statistic for ordinal covariate), and 4.00 for the WDM_O (double maximum test statistic for ordinal covariate), respectively. Their statistical powers were 94%, 96%, and 94%, respectively. However, this discrepancy was due to the informative continuous covariate, which means that the split point of the ordinal covariate was 2 for all conditions. Thus, the condition of the effect size of 0.4 with the sample size of 2,000 sufficiently works well to recover the true number of subgroups with the pre-pruning option of setting the minimum sample size of 250 in this study even though it was not perfectly recovered. However, with the same conditions of the options, the number of estimated subgroups increased as the sample size increases regardless of the different test statistic for the ordinal covariate.

Second, Table 27 shows the results of the effect size of 0.6. When the sample size was 1,000 with limiting the minimum size per subgroup without post pruning using BIC, the means of the estimated subgroups across three test statistic of LM, $maxLM_O$, and WDM_O were 3.98, 3.99, and 3.98, respectively, and their statistical powers were also 98%, 99%, and 98%, respectively. With the same condition, as the sample size was 2,000, the means of the estimated subgroups across three options were 4.01, 4.00, and 4.03 with their corresponding statistical powers of 99%, 100%, and 97%, respectively. Likewise, this was due to the informative continuous covariate not by the informative ordinal covariate. Thus, the condition of the effect size of 0.6 with the sample size of 1,000 or more sufficiently works well to recover the true number of subgroups with the options of pre-pruning of setting the minimum sample size of 250

without post-pruning method using BIC. However, with the same conditions of the options, the number of estimated subgroups tend to increase as the sample size increases regardless of the different test statistic for the ordinal covariate like the results of the effect size of 0.4.

Third, Table 28 shows the results of the effect size of 0.8. When the sample size was 1,000 with limiting the minimum size per subgroup without post pruning using BIC, the means of the estimated subgroups across three test statistic of LM, $maxLM_O$, and WDM_O were the same as 4.00 showing 100% of statistical powers. With the same condition, as the sample size was 2,000, the means of the estimated subgroups across three options were 4.00, 4.02, and 4.02 with their corresponding statistical powers of 100%, 98%, and 98%, respectively. Likewise, this was due to the informative continuous covariate not by the informative ordinal covariate. However, with the same conditions of the options, the number of estimated subgroups tend to increase as the sample size increases regardless of the different test statistic for the ordinal covariate like the results of the effect sizes of 0.4 and 0.6. The powers were less than 90% with this condition. Thus, when there is mean difference with the effect size of 0.8 and the sample sizes of 1,000 or 2,000, MOB sufficiently works well to recover the true number of subgroups with the options of pre-pruning of setting the minimum sample size of 250 without post-pruning method using BIC. Looking at Table 29, the results of the effect size of 1.0 show a similar pattern of the results of the effect size of 0.8 except one condition. This condition will be discussed next section. Like the other results, there were not discrepancies of the results between three test statistic.

5.5.2 Post pruning method using BIC

The post pruning option with BIC plays a key role to determine the final number of subgroups. With or without it produces different number of subgroups under most of the conditions. Across most of the conditions, the mean of estimated number of subgroups tends to

increase without the post pruning option, which means that MOB tends to over-extract the number of subgroups. When the effect size was 0.2, there were no cells which correctly recovered the true number of subgroups even if the sample size was large. Looking at the Table 25, the means of estimated number of subgroups increase as the sample size increase without the post pruning option. Even if MOB was used with setting the minimum size of 250 without post pruning of BIC, the mean of estimated number of subgroups increased. This is because tiny parameter instabilities can be detected with large sample sizes. Using post pruning of BIC consistently reduced the number of subgroups compared to the one without it. When the sample sizes were 1,000 or 2,000, the averages of determined number of subgroups were around 1, which MOB failed to differentiate the group differences that improve the model fit statistically. Their statistical power was also zero across conditions. In addition, as the sample size increases to 5,000 or more, the averages of determined number of subgroups were around 2 with the post pruning using BIC. That is, even though MOB might differentiate distinct subgroups with around 80% statistical powers at first, the model fit of the subgroups did not improve compared to the model without the subgroups, determining the number of subgroups as two instead of four.

Next, when the effect size was 0.4, using post pruning with BIC works well to recover the true number of subgroups under the sample sizes of 10,000 and 20,000. Looking at Table 26, with the increased effect size of 0.4 compared to 0.2, the statistical powers to recover the true number of subgroups nearly 99-100% where there are 10,000 samples or more regardless of setting the minimum sample size per subgroup. As the effect size increases, required sample size for recovering the true number of subgroups decreases. Looking at Table 27, when the effect size was 0.6, the average number of estimated subgroups was nearly 4.00 and the statistical power was nearly 100% with the 5,000 of sample size using post pruning method. When the effect size

was 0.8, the 2,000 of sample size was sufficient to recover the true number of subgroups as presented in Table 28. When the effect size was 1.0, the 1,000 of sample size was sufficient to recover the true number of subgroups as presented in Table 29. Cleary, post pruning using BIC helps to avoid over-fitting issues (growing large tree) under certain conditions.

5.5.3 Limiting minimum sample size per subgroup

Limiting the minimum sample size per subgroup as 250 was used to correctly detect the nonlinear changes with four time points and get stable parameter estimates for each subgroup. Without this, smaller sample size per subgroup can be used to fit the quadratic latent growth curve model. Two notable findings how limiting the minimum size works are when the effect sizes are 0.6 and 0.8 with the sample size of 1,000. Looking at Table 27, the average of estimated number of subgroups was ranged from 3.98 to 3.99, and their statistical power was ranged from 99% to 100% without post pruning using BIC. However, when the post pruning of BIC was used with the same condition, the average of estimated number of subgroup and the statistical power were 2.00 and 0%, respectively. Also, when the sample size increases to 2,000, the average of estimated number of subgroups and statistical power were ranged from 4.00 to 4.03 and from 97% to 100%, respectively. When the effect size was 0.8 in Table 28, the average of estimated number of subgroups and the statistical power were 4.00 and 100% with 1,000 samples, respectively, and the average of estimated number of subgroups was ranged from 4.00 to 4.02 and the statistical power was ranged from 98% to 100% with 2,000 samples. However, as the sample size increases more than 2,000, MOB without the post pruning using BIC over-extracted the number of subgroups. This means that when the effect sizes are medium (0.4 - 0.6) with relatively small sample size, such as 2,000, limiting the minimum sample size per subgroup works better than post pruning using BIC to recover the true number of subgroups.

Table 25. Mean of estimated number of subgroups (MNS) and statistical power (P) to recover true number of subgroups (Effect size = 0.2)

Min #	Pruning	Ordinal	N=1,000		N=2,000		N=5,000		N=10,000		N=20,000	
			MNS	P	MNS	P	MNS	P	MNS	P	MNS	P
No	No	LM	2.06	0	2.15	2	3.13	31	4.19	83	4.19	82
		$maxLM_O$	2.06	1	2.09	1	3.41	34	4.25	80	4.24	80
		WDM_O	2.13	1	2.15	2	3.47	35	4.27	78	4.24	80
	BIC	LM	1.00	0	1.00	0	2.00	0	2.00	0	2.00	0
		$maxLM_O$	1.00	0	1.01	0	2.00	0	2.00	0	2.00	0
		WDM_O	1.00	0	1.00	0	2.00	0	2.00	0	2.00	0
	No	LM	1.96	0	2.11	0	3.35	31	4.19	82	4.29	78
		$maxLM_O$	2.00	0	2.10	0	3.48	40	4.20	83	4.23	78
		WDM_O	2.02	0	2.09	0	3.50	41	4.18	84	4.33	75
	BIC	LM	1.00	0	1.04	0	2.00	0	2.00	0	2.00	0
		$maxLM_O$	1.00	0	1.03	0	2.00	0	2.00	0	2.00	0
		WDM_O	1.00	0	1.03	0	2.00	0	2.00	0	2.00	0
250	No	LM	2.06	0	2.15	2	3.13	31	4.19	83	4.19	82
		$maxLM_O$	2.06	1	2.09	1	3.41	34	4.25	80	4.24	80
		WDM_O	2.13	1	2.15	2	3.47	35	4.27	78	4.24	80
	BIC	LM	1.00	0	1.00	0	2.00	0	2.00	0	2.00	0
		$maxLM_O$	1.00	0	1.01	0	2.00	0	2.00	0	2.00	0
		WDM_O	1.00	0	1.00	0	2.00	0	2.00	0	2.00	0
	No	LM	1.96	0	2.11	0	3.35	31	4.19	82	4.29	78
		$maxLM_O$	2.00	0	2.10	0	3.48	40	4.20	83	4.23	78
		WDM_O	2.02	0	2.09	0	3.50	41	4.18	84	4.33	75
	BIC	LM	1.00	0	1.04	0	2.00	0	2.00	0	2.00	0
		$maxLM_O$	1.00	0	1.03	0	2.00	0	2.00	0	2.00	0
		WDM_O	1.00	0	1.03	0	2.00	0	2.00	0	2.00	0

Note. LM = Lagrange multiplier (LM) type test statistic for categorical covariate.
$maxLM_O$= Test statistic of adapted maximum of LM for ordinal covariate.
WDM_O = Test statistic of weighted double maximum for ordinal covariate.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.
MNS = mean number of estimated subgroups. P = statistical power.

Table 26. Mean of estimated number of subgroups (MNS) and statistical power (P) to recover true number of subgroups (Effect size = 0.4)

Min #	Pruning	Ordinal	N=1,000		N=2,000		N=5,000		N=10,000		N=20,000	
			MNS	P	MNS	P	MNS	P	MNS	P	MNS	P
No	No	LM	2.47	5	4.11	79	4.27	78	4.17	83	4.22	86
		$maxLM_O$	2.39	5	4.12	86	4.15	87	4.27	78	4.28	78
		WDM_O	2.48	8	4.13	84	4.17	86	4.16	85	4.19	84
	BIC	LM	2.00	0	2.00	0	2.03	0	4.00	100	4.00	100
		$maxLM_O$	2.00	0	2.00	0	2.03	0	3.99	99	4.00	100
		WDM_O	2.00	0	2.00	0	2.04	0	4.00	100	4.00	100
	No	LM	2.73	12	3.96	94	4.21	82	4.23	78	4.28	78
		$maxLM_O$	2.86	12	4.02	96	4.22	79	4.17	84	4.19	84
		WDM_O	2.88	17	4.00	94	4.21	81	4.11	90	4.14	87
	BIC	LM	2.00	0	2.00	0	2.06	0	4.00	100	4.00	100
		$maxLM_O$	2.00	0	2.00	0	2.07	1	4.00	100	4.00	100
		WDM_O	2.00	0	2.00	0	2.05	0	4.00	100	4.00	100
250	No	LM	2.47	5	4.11	79	4.27	78	4.17	83	4.22	86
		$maxLM_O$	2.39	5	4.12	86	4.15	87	4.27	78	4.28	78
		WDM_O	2.48	8	4.13	84	4.17	86	4.16	85	4.19	84
	BIC	LM	2.00	0	2.00	0	2.03	0	4.00	100	4.00	100
		$maxLM_O$	2.00	0	2.00	0	2.03	0	3.99	99	4.00	100
		WDM_O	2.00	0	2.00	0	2.04	0	4.00	100	4.00	100
	No	LM	2.73	12	3.96	94	4.21	82	4.23	78	4.28	78
		$maxLM_O$	2.86	12	4.02	96	4.22	79	4.17	84	4.19	84
		WDM_O	2.88	17	4.00	94	4.21	81	4.11	90	4.14	87
	BIC	LM	2.00	0	2.00	0	2.06	0	4.00	100	4.00	100
		$maxLM_O$	2.00	0	2.00	0	2.07	1	4.00	100	4.00	100
		WDM_O	2.00	0	2.00	0	2.05	0	4.00	100	4.00	100

Note. LM = Lagrange multiplier (LM) type test statistic for categorical covariate.
$maxLM_O$= Test statistic of adapted maximum of LM for ordinal covariate.
WDM_O = Test statistic of weighted double maximum for ordinal covariate.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.
MNS = mean number of estimated subgroups. P = statistical power.

Table 27. Mean of estimated number of subgroups (MNS) and statistical power (P) to recover true number of subgroups (Effect size = 0.6)

Min #	Pruning	Ordinal	N=1,000		N=2,000		N=5,000		N=10,000		N=20,000	
			MNS	P	MNS	P	MNS	P	MNS	P	MNS	P
No	No	LM	4.13	85	4.14	88	4.21	84	4.23	79	4.23	79
		$maxLM_O$	4.15	87	4.21	82	4.26	78	4.31	73	4.20	82
		WDM_O	4.16	84	4.22	83	4.23	80	4.27	79	4.27	78
	BIC	LM	2.00	0	2.21	3	4.00	100	4.00	100	4.00	100
		$maxLM_O$	2.00	0	2.15	4	4.00	100	4.00	100	4.00	100
		WDM_O	2.00	0	2.11	1	4.00	100	4.00	100	4.00	100
	No	LM	3.98	98	4.01	99	4.20	81	4.31	77	4.26	77
		$maxLM_O$	3.99	99	4.00	100	4.29	73	4.18	84	4.21	81
		WDM_O	3.98	98	4.03	97	4.15	87	4.23	84	4.14	87
	BIC	LM	2.00	0	2.11	3	4.00	100	4.00	100	4.00	100
		$maxLM_O$	2.00	0	2.15	2	4.00	100	4.00	100	4.00	100
		WDM_O	2.00	0	2.12	3	4.00	100	4.00	100	4.00	100
250	No	LM	4.13	85	4.14	88	4.21	84	4.23	79	4.23	79
		$maxLM_O$	4.15	87	4.21	82	4.26	78	4.31	73	4.20	82
		WDM_O	4.16	84	4.22	83	4.23	80	4.27	79	4.27	78
	BIC	LM	2.00	0	2.21	3	4.00	100	4.00	100	4.00	100
		$maxLM_O$	2.00	0	2.15	4	4.00	100	4.00	100	4.00	100
		WDM_O	2.00	0	2.11	1	4.00	100	4.00	100	4.00	100
	No	LM	3.98	98	4.01	99	4.20	81	4.31	77	4.26	77
		$maxLM_O$	3.99	99	4.00	100	4.29	73	4.18	84	4.21	81
		WDM_O	3.98	98	4.03	97	4.15	87	4.23	84	4.14	87
	BIC	LM	2.00	0	2.11	3	4.00	100	4.00	100	4.00	100
		$maxLM_O$	2.00	0	2.15	2	4.00	100	4.00	100	4.00	100
		WDM_O	2.00	0	2.12	3	4.00	100	4.00	100	4.00	100

Note. LM = Lagrange multiplier (LM) type test statistic for categorical covariate.
$maxLM_O$= Test statistic of adapted maximum of LM for ordinal covariate.
WDM_O = Test statistic of weighted double maximum for ordinal covariate.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.
MNS = mean number of estimated subgroups. P = statistical power.

Table 28. Mean of estimated number of subgroups (MNS) and statistical power (P) to recover true number of subgroups (Effect size = 0.8)

Min #	Pruning	Ordinal	N=1,000		N=2,000		N=5,000		N=10,000		N=20,000	
			MNS	P	MNS	P	MNS	P	MNS	P	MNS	P
No	No	LM	4.19	82	4.28	75	4.16	84	4.23	83	4.27	76
		$maxLM_O$	4.24	79	4.19	83	4.15	86	4.25	81	4.28	78
		WDM_O	4.16	84	4.32	74	4.21	82	4.23	78	4.18	84
	BIC	LM	2.20	3	4.00	100	4.00	100	4.00	100	4.00	100
		$maxLM_O$	2.19	5	4.01	99	4.00	100	4.00	100	4.00	100
		WDM_O	2.17	2	4.00	100	4.00	100	4.00	100	4.00	100
	No	LM	4.00	100	4.00	100	4.21	82	4.20	84	4.24	82
		$maxLM_O$	4.00	100	4.02	98	4.20	81	4.22	81	4.30	74
		WDM_O	4.00	100	4.02	98	4.19	85	4.12	89	4.21	83
	BIC	LM	2.15	3	4.00	100	4.00	100	4.00	100	4.00	100
		$maxLM_O$	2.11	3	4.00	100	4.02	99	4.00	100	4.00	100
		WDM_O	2.15	2	4.00	100	4.00	100	4.00	100	4.00	100
250	No	LM	4.19	82	4.28	75	4.16	84	4.23	83	4.27	76
		$maxLM_O$	4.24	79	4.19	83	4.15	86	4.25	81	4.28	78
		WDM_O	4.16	84	4.32	74	4.21	82	4.23	78	4.18	84
	BIC	LM	2.20	3	4.00	100	4.00	100	4.00	100	4.00	100
		$maxLM_O$	2.19	5	4.01	99	4.00	100	4.00	100	4.00	100
		WDM_O	2.17	2	4.00	100	4.00	100	4.00	100	4.00	100
	No	LM	4.00	100	4.00	100	4.21	82	4.20	84	4.24	82
		$maxLM_O$	4.00	100	4.02	98	4.20	81	4.22	81	4.30	74
		WDM_O	4.00	100	4.02	98	4.19	85	4.12	89	4.21	83
	BIC	LM	2.15	3	4.00	100	4.00	100	4.00	100	4.00	100
		$maxLM_O$	2.11	3	4.00	100	4.02	99	4.00	100	4.00	100
		WDM_O	2.15	2	4.00	100	4.00	100	4.00	100	4.00	100

Note. LM = Lagrange multiplier (LM) type test statistic for categorical covariate.
$maxLM_O$= Test statistic of adapted maximum of LM for ordinal covariate.
WDM_O = Test statistic of weighted double maximum for ordinal covariate.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.
MNS = mean number of estimated subgroups. P = statistical power.

Table 29. Mean of estimated number of subgroups (MNS) and statistical power (P) to recover true number of subgroups (Effect size = 1.0)

Min #	Pruning	Ordinal	N=1,000		N=2,000		N=5,000		N=10,000		N=20,000	
			MNS	P	MNS	P	MNS	P	MNS	P	MNS	P
No	No	LM	4.13	89	4.22	80	4.29	77	4.14	88	4.24	80
		$maxLM_O$	4.22	84	4.15	86	4.28	76	4.34	74	4.23	82
		WDM_O	4.12	88	4.26	79	4.25	77	4.22	84	4.19	85
	BIC	LM	3.99	99	4.00	100	4.00	100	4.00	100	4.00	100
		$maxLM_O$	4.00	100	4.00	100	4.00	100	4.00	100	4.00	100
		WDM_O	3.99	99	4.00	100	4.00	100	4.00	100	4.00	100
	No	LM	4.00	100	4.01	99	4.17	84	4.27	76	4.29	77
		$maxLM_O$	4.00	100	4.01	99	4.11	89	4.17	86	4.25	80
		WDM_O	4.00	100	4.00	100	4.17	84	4.27	78	4.17	83
	BIC	LM	4.00	100	4.00	100	4.00	100	4.00	100	4.00	100
		$maxLM_O$	4.00	100	4.00	100	4.00	100	4.00	100	4.00	100
		WDM_O	4.00	100	4.00	100	4.00	100	4.00	100	4.00	100
250	No	LM	4.13	89	4.22	80	4.29	77	4.14	88	4.24	80
		$maxLM_O$	4.22	84	4.15	86	4.28	76	4.34	74	4.23	82
		WDM_O	4.12	88	4.26	79	4.25	77	4.22	84	4.19	85
	BIC	LM	3.99	99	4.00	100	4.00	100	4.00	100	4.00	100
		$maxLM_O$	4.00	100	4.00	100	4.00	100	4.00	100	4.00	100
		WDM_O	3.99	99	4.00	100	4.00	100	4.00	100	4.00	100
	No	LM	4.00	100	4.01	99	4.17	84	4.27	76	4.29	77
		$maxLM_O$	4.00	100	4.01	99	4.11	89	4.17	86	4.25	80
		WDM_O	4.00	100	4.00	100	4.17	84	4.27	78	4.17	83
	BIC	LM	4.00	100	4.00	100	4.00	100	4.00	100	4.00	100
		$maxLM_O$	4.00	100	4.00	100	4.00	100	4.00	100	4.00	100
		WDM_O	4.00	100	4.00	100	4.00	100	4.00	100	4.00	100

Note. LM = Lagrange multiplier (LM) type test statistic for categorical covariate.
$maxLM_O$= Test statistic of adapted maximum of LM for ordinal covariate.
WDM_O = Test statistic of weighted double maximum for ordinal covariate.
Min # = minimum sample size per a subgroup. Pruning = post pruning method using BIC.
MNS = mean number of estimated subgroups. P = statistical power.

5.6 Results of growth mixture model

The same simulated datasets were fitted to growth mixture model parallelly. The results show that the number of classes (subgroups) was only *two* across all the conditions. That is, GMM failed to recover the true number of subgroups as the best fitting model across all the conditions. Specifically, the models showing the lowest BIC values were the two-class solutions. In addition, the Lo-Mendell-Rubin likelihood ratio test was not significant for three-, four-, and five- class solutions across more than 97 replications and all the conditions. When the effect size was 0.8 or 1.0 and the sample size was 10,000 or 20,000, GMMs produced significant LRT results for the four-class models even though the values of BIC of these models were higher than the ones of the two-class model. More importantly, when I look at the composition of the classified four classes, the proportions of the sample size of four-class model were approximately 49.8%, 0.01%, 50%, and 0.01% for each class, which is most of the subjects were classified into two classes. Furthermore, the parameter estimates were not consistent over the replication. Since the best solutions were the two-class model, it was not possible to calculate the classification accuracy and precision as well as the bias and RMSE of the parameter estimates. However, except means intercept, linear slope, and quadratic slope, their variances and covariances as well as the residuals were close to the parameter values because they were fixed across classes as the population model.

CHAPTER 6. CONCLUSION AND DISCUSSION

6.1 Summary of findings

This study had three main research purposes. First, it aimed to introduce and demonstrate how to use model-based recursive partitioning (MOB) approach combined with latent growth curve model (LGCM) for longitudinal study to uncover heterogeneous subpopulation. Since this approach was not introduced in the field of education and psychology, I used an empirical representative longitudinal data in education as an illustrative purpose. The procedures, findings and interpretations were presented in Chapter 3. The second purpose of this study was to investigate the performance of MOB with the quadratic latent growth curve model having two interactions among three types of covariates, one is between an ordinal and continuous covariate and another is between the ordinal and categorical covariate. Based on the results from the Chapter 3, a population model was chosen for data generation. Effect size and sample size were varied to simulate datasets, and three options in the estimation were considered as simulation conditions. A simulation study was conducted to answer seven research questions under 300 conditions from fully crossed five factors. Lastly, it aimed to compare the results of MOB with the ones of unconditional growth mixture model (GMM). There are six key findings from both illustrative analysis and simulation study to address the research questions.

First, the result of empirical study using HSLS:09 to find distinct subgroups showing different trajectories and initial status of GPA score suggests that using both options of post pruning with BIC and limiting the number of sample size per subgroup makes the resulting tree more *concise* and *easier* to interpret the composition of subgroup. In addition, using the test statistic of WDM_O for the ordinal covariate is more likely to makes MOB to use the ordinal

covariates for splitting the groups as it produces more accurate p-values than the test statistic of the categorical covariate. If there are available ordinal covariates that are strongly related to the outcomes in a parametric model and the order of the values are important to be considered in a study, it is highly suggested to declare them as ordinal and to use corresponding test statistic in software. In addition, connecting two packages should be accomplished by user's own function because commercial software is not available yet.

Second, a simulation study was conducted to investigate the performance of this approach for a given population model under five factors. The results show that the true number of subgroups was perfectly recovered when the effect size was equal to 1.0 with both sample size of 1,000 and BIC post pruning option. With the effect size of 0.8, the required sample size was 2,000 to recover the true number of subgroups perfectly. As the effect sizes of the mean intercept decrease to medium sizes, including 0.6 and 0.4, the required sample sizes also increase to 5,000 and 10,000, respectively. When the effect size was small (0.2), the maximal statistical power to recover the true number of subgroups was 84% with the sample size of 10,000.

Third, if the number of subgroups was perfectly recovered, the overall accuracy and precision were also 1.00. Even if the number of subgroups was not perfectly recovered, when the effect size and the sample size were 0.4 and 2,000, respectively, the overall accuracy and macro-averaged precision were 0.97 and 0.98, respectively. Furthermore, four noises covariates were not chosen for splitting subgroups at any conditions indicating that MOB works well to find splitting points as designed. Only the informative continuous covariate was additionally used to split the subgroups if the number of estimated subgroups was larger than 4. The splitting points of the informative ordinal and categorical covariates were nearly close

to the true point of 2 for the former and (10,20) vs. (30,40) for the latter if the number of subgroups was perfectly uncovered. The true splitting point of 2 for the informative ordinal covariate was correctly detected for the first splitting under most of the conditions except two conditions that the effect size is 0.2 and the sample sizes are 1,000 and 2 000 with BIC pruning option. This is because the uncovered number of subgroups was 1 under these conditions. Besides, the parameter estimates were also unbiased and efficient for those conditions that the number of estimated subgroups was nearly equal to four, and their statistical power was close to 100%.

Fourth, the simulation study shows that there is no evident difference among the test statistic for the ordinal covariates. This result is partly because the population model includes the main effect of the ordinal covariate only. The remained continuous and categorical covariates have the interaction effects only with the ordinal covariate. In addition, the population model is straightforward to differentiate the subgroups using one cut point only. This would make no difference among the three test statistic. However, the procedure and result of the empirical study show that if the test statistic for the ordinal covariate is used, the MOB obviously selects the ordinal covariate first to split the samples. Even though the relationship between the outcome and the ordinal covariate was strong, using the default option for the test statistic of categorical covariate as the ordinal covariate did not select the ordinal covariate at both the first stage and any other splitting stage. Thus, it is highly desirable for researchers to use all available options for the test statistic of the ordinal covariate. If there are strong relationships between the outcome and the splitting candidate ordinal covariates, it would be preferred for them to use either $maxLM_O$ or WDM_O rather than LM.

Fifth, based on the results of both empirical and simulation study, using post pruning

option of BIC helps to avoid over-fitting issues resulting in growing large size of tree generally. Moreover, it works better than limiting the number of subgroups. With smaller sample size of 1,000 and medium to large effect sizes (0.6 and 0.8), however, using the post pruning option of BIC reduced the number of subgroups unnecessarily. Limiting the number of subgroups did not impact to determine the true number of subgroups. However, it is suggested to know the adequate or required number of sample size for a certain parametric model to be used as a template model in advance. As described in the results of the empirical study, there were huge number of subgroups in a tree without limiting the sample size per subgroup. Since a parametric model is fitted to the samples of each subgroup, the adequate sample size is required to get stable and correct parameter estimates. Having a rationale to determine the adequate sample size for the specific parametric model is a task for each researcher as this study did.

Last but not least, GMM did not differentiate the true subgroups under all conditions considered in this study for a given population model of LGCM. The model fit of the four-class solution was not the best based on the value of BIC. Likelihood ratio test also did not produce significant result supporting that there are four classes in a population in terms of growth trajectories. Even if GMM finds the four latent classes as the best fitting model at a few times under certain conditions, the classification accuracy and precision were very poor because most of the samples were divided into two subgroups only and there were a few samples within the remained two subgroups. This would be partly due to the fact that the data generation was not based on the mixture model. Since the purpose of this study is not to compare the performance of these two approaches directly, it is not true that MOB performs better than GMM. Each method has its own research purposes and the approach, assumption, and conceptual framework of GMM to enumerate the number of subgroups is totally different from MOB. To better find the

best fitting model for GMM, more rigorous steps and approaches should be required. The focus of this study was to examine if MOB not only finds complex interactions between informative covariates including their cut points, but also determines the true number of subgroups correctly. Thus, the above results support the claim that the research goals are met.

6.2 Discussions

The findings of this study have several significant implications for advancements of quantitative methodology in education research. As the machine learning techniques and their algorithms are getting popular and widely used by social scientist in these days, incorporating the machine learning and a variety of statistical models can be one of great analytic tools for handling big data in the study of education.

First, it introduces a general framework of MOB and detailed procedures how to use it for social and behavioral researchers. Through the illustrative example, they can be guided when MOB can be useful and how it can be applied to different research questions and settings. Within the framework of MOB, the definition of covariate is broader than the one commonly used in statistics. A covariate can be regarded as a variable that is potentially related to the interested outcome(s) in the available datasets. Thus, this methodological approach is especially beneficial when researchers handle very large samples, such as more than 10,000, to explore unknown composition of subpopulations and uncover them with many variables that are potentially related to the interested outcome(s) and have interaction effects with other variables. Besides, the results of simulation study provide quantitative methodology community with statistical evidence of how well MOB recovers true subgroups, making them to be equipped with this analytic tool in hand. Yet, this approach would not be helpful for relatively smaller sample sizes, such as 1,000 or 2,000, because it would be unrealistic to have such a huge strong effect sizes (more than 0.8)

in social science. In fact, if there is a good theoretically established statistical model that fits data well with adequate sample size, then there is no need to consider using MOB. However, if there is a less established theoretical model and you have huge amount of variables with large sample sizes, then MOB would be a natural candidate as one of useful analytic tools. Thus, it is suggested to start with fitting the theoretical model to the whole sample of data, and then to consider the further suspected potential covariates for splitting the data to determine the subgroups. Practically, those subgroups can be either directly interpreted to explain the discrepancy of the interested outcome or a source to be analyzed as separate subsamples for any other further steps if necessary. Therefore, the number of subgroups totally depends on the purpose of study.

Second, novel but important research questions can be postulated and answered. Disparities in educational opportunities and achievements exist among students of different gender, race/ethnicity, SES, and other demographics including environmental factors. When these characteristics are considered interdependently, for example, White female student from high-SES background and Black male student from low-SES background, the educational inequalities may be found to be worse than simply examining subgroup differences by gender or race/ethnicity. In this regard, an emerging framework of *intersectionality* in education research attends to these layered marginalization of student populations. Yet, such subpopulations are mostly defined by researchers' own decisions or based on an established theory, often resulting in arbitrary groups, not empirically identified. Methodologically, MOB showed its strengths to detect and find meaningful subgroups and advance our understanding of the complex social mechanisms through examining those interaction effects. Furthermore, the resulting tree of MOB can be visualized to describe the composition of the subgroups in a tree or a trend of change,

which is very interpretable, intuitive, and straightforward. Moreover, as big data is becoming more available in education research, it is critical to begin investigating other unveiled contextual factors (e.g., average housing prices in the neighborhood) that are contributing to the increasing gap in educational opportunities. If there exist some informative covariates that are already found to be related to the outcomes, those can be included in the population model. In this case, the former covariates are not used for splitting the samples because they are already in the statistical template model. Other unveiled potential covariates then would be used for splitting the subgroups. Thus, this study ultimately contributes to informing various education policy stakeholders to make *data-driven decisions* holding statistical properties.

Third, this study has numerous potentials to be extended to other various statistical models, such as causal inference methods that aim to test possible heterogeneous treatment effects of candidate covariates. Although the simulation study was conducted for a given population having specific parameter values within a context of longitudinal study, the findings of this study can also be similarly applied to other popular statistical models. In fact, the existing approach so called SEM Trees is tailed for the SEM models specifically. This study will thus play a critical role to lay the groundwork of extending the application of MOB into various statistical models by investigating its performance regarding complex covariate effects to find subgroups.

Fourth, MOB has several options to optimize the size of subgroups with pre- and post-pruning options. This study examined their desired options under different conditions though, pruning itself totally depends on the purpose of the study. If the purpose of study is to get insights on different complex effects of the covariates on the outcome, larger number of subgroups would give ones all the details and direct interpretation of them like the traditional

regression model. Also, the interpretation of the results of MOB with large number of subgroups would not be different from the regression model if the interaction effects are not complex. On the other hands, if the purpose of study is to find a few meaningful subpopulations showing distinct different distribution and pattern of the outcome(s), optimizing the size of subgroups with pre- and post- pruning options in MOB would help reduce the number of subgroups, leading to more concise larger subgroups. In this case, a few certain targeted demographics and other background variables would be beneficial.

6.3 Limitation and future research

This study has certain limitations as followings. First, it is assumed that individuals are independent for the purpose of study even if it is not true. That is, the empirical data used in this study has multilevel (nested) data structure. However, this analytic approach using machine learning is an exploratory data analytic tool and has a complementary nature to existing traditional statistical models. Thus, it is not necessary to consider the multilevel data structure at this stage. If the interaction effects of covariates are detected through this analysis, the terms can be reflected in a statistical multilevel modeling and estimated for statistical inference as the second step.

Second, this study considered a specific population model. Different population model could result in different results in terms of performance because MOB is basically data-driven method, meaning that it depends on the available data and the parametric model used. In addition, the parametric model used in this study does not include any established covariates even if it is possible. In other words, the covariates are only used for splitting the samples. The future step would consider this population model that includes covariates that would not be used

for splitting. It also considered two types of interactions, which is the one between the ordinal and continuous covariates and another between the ordinal and categorical covariates. In this case, only the main effect of the ordinal covariate is strongly associated with the outcome. However, it is not fully addressed in literature if MOB can detect the higher-order interaction effects of the covariates when the main effects of those are not associated with the outcomes. Thus, this study is limited to the cases having interaction effects when the main effects are also associated.

Based on the limitations, this study has several next steps. The first possible extension of this study can be adopting the multilevel structural equation models accounting for a nested data structure. Distinct subgroups may be uncovered according to the higher-level covariates such as school type, districts, geographical information, or even states. The current study does not consider the multilevel structure for the simplicity. As stated earlier, any parametric models can be utilized as a template model to examine the associations between the focal model parameters and the covariates.

Secondly, even though it is not common in social science, there would be cases where there is no main effect, but it may interact with other covariates to have effects on the outcomes, producing higher order interactions. If there exist empirical data showing this particular interaction effect in educational and psychological study, a simulation condition with a different population model needs to be added to examine the performance of MOB.

Thirdly, the effects of post-pruning method are not fully examined yet in the previous literature. For large-scale datasets, post-pruning is strongly suggested to reduce the number of terminal nodes (subgroups), improving interpretability and stability of parameter estimates depending on the number of parameters and sample size. Thus, a comparative study between two

packages of `semtree` and `partykit` is also desired because each has distinct features in terms of how to split the subgroups and available options.

Lastly, there is a need to rigorously design a comparative study to compare GMM and MOB for different population models. Although the third part of this study dealt with whether GMM also recovers the true number of subgroups, the data generation under a given population model did not fully reflect the nature of heterogeneity of parameters for GMM. The next step is to compare the performance between two approaches under various population models as well as conditions.

APPENDIX

APPENDIX

R CODES CONNECTING LAVAAN AND PARTYKIT

```
# List of packages #
rm(list=ls())
library(MASS)
library(sandwich)
library(lavaan)
library(partykit)
library(strucchange)
library(plyr)
library(MplusAutomation)

# Quadratic model with free covariances and fixed residuals #
QLGM <- '
  inter =~ 1*GPA9 + 1*GPA10 + 1*GPA11 + 1*GPA12;
  slope =~ 0*GPA9 + 1*GPA10 + 2*GPA11 + 3*GPA12;
  quadr =~ 0*GPA9 + 1*GPA10 + 4*GPA11 + 9*GPA12;
  inter ~~ vi*inter; inter ~ mi*1;
  slope ~~ vs*slope; slope ~ ms*1;
  quadr ~~ vq*quadr; quadr ~ mq*1;
  inter ~~ cis*slope;
  inter ~~ ciq*quadr;
  slope ~~ csq*quadr;
  GPA9 ~~ res*GPA9; GPA9 ~ 0*1;
  GPA10 ~~ res*GPA10; GPA10 ~ 0*1;
  GPA11 ~~ res*GPA11; GPA11 ~ 0*1;
  GPA12 ~~ res*GPA12; GPA12 ~ 0*1;
'

# Fit function for SEM #
lavaan_fit <- function(model) {
  function(y, x = NULL, start = NULL, weights = NULL, offset = NULL,
..., estfun = FALSE, object = FALSE) {
    require(lavaan)
    lgcm <- lavaan::lavaan(model = model, data = y, start = start)
    list(
      coefficients = stats4::coef(lgcm), # coefficients
      objfun = -as.numeric(stats4::logLik(lgcm)), # negative log-
likelihood
      estfun = if(estfun) sandwich::estfun(lgcm) else NULL, # score
matrix including empirical estimating functions
      object = if(object) lgcm else NULL
    )
  }
}

# A function for concise results #
node_format <- function(node) {
  c("",
    sprintf("n = %s", node$nobs),

capture.output(print(cbind(node$coefficients[c("mi","ms","mq","vi","vs",
"vq","cis","ciq","csq","res")]), digits = 1L))[-1L])
}
```

```
# Latent growth curve model with mob algorithm #

pop.control=mob_control(alpha=0.05,bonferroni=TRUE,minsize=250,ytype="da
ta.frame",ordinal="max") # ordinal variable option = max

# Chapter 3 model with all covariates #
pop.mod <- mob(GPA9+GPA10+GPA11+GPA12 ~
SEX+RACE5+LOCA+HACT+SES+MATEFF+BEHAVSCH+SCHCLI+FLUNCH+MISBEHAV,
               data = dat,
               fit = lavaan_fit(QLGM),
               control = pop.control)

plot(tr, drop = TRUE, tnex = 2,FUN=node_format) # terminal nodes
```

```
# Parameters of population model
mean.isq=c(2.8,-0.1,0.06)
cov.isq=matrix(c(0.471,0.001,-0.012,0.001,0.057,-0.016,-0.012,-
0.016,0.007),3,3) # covariance matrix
par=c(mean.isq,diag(cov.isq),cov.isq[upper.tri(cov.isq, diag =
F)],0.084) # parameters for evaluation

# Store gmm results #
gmm.NG=list() # number of groups (sub-populations/classes)
gmm.M=list() # model fit information
gmm.C=list()# classification accuracy
gmm.E=list() # estimates of coefficients
gmm.times=list()

# Store mob results #
res.NG=list() # number of groups (sub-populations/classes)
res.cls=list() # classification accuracy
res.est=list() # estimates of coefficients
res.split=list() # split points
res.times=list()

# Simulation conditions #
size=c(1000,2000,5000,10000,20000) # sample size = 5
effectsize=c(0.2,0.4,0.6,0.8,1.0) # Cohen's d effect size = 5
mean.diff=round(effectsize*sqrt(cov.isq[1,1]),2);mean.diff # mean
difference based on Cohen's effect size
ordinal.option=c("chisq","L2","max") # ordinal = 3
minsize=c("min.No","min.250") # minimum size of node = 2
prune=c("prune.No","prune.BIC") # pruning = 2

# data generation #
set.seed(10)
REP=1:100 # replication number

# Simulation start #
for(s in 1:length(mean.diff)){
  # parameters matrix; intercept only difference
  # Scenario 1: Four groups having different intercepts
  pars=matrix(c(c(par[1]-mean.diff[s],par[-1]), # G1 parameters
                c(par[1]+0*mean.diff[s],par[-1]), # G2 parameters
                c(par[1]+mean.diff[s],par[-1]), # G3 parameters
                c(par[1]+2*mean.diff[s],par[-
1])),nrow=4,ncol=10,byrow=T) # G4 parameters

colnames(pars)=c("mi","ms","mq","vi","vs","vq","cis","ciq","csq","res")
  for(i in 1:length(size)){
    for(kk in 1:length(minsize)){
      for(pp in 1:length(prune)){
        for(o in 1:length(ordinal.option)){
         # replication start #
             for(r in REP){
        tem=NA
        N=size[i]
        factors<-mvrnorm(N/4,mean.isq,cov.isq) # four subgroups
        G1<-as.data.frame(matrix(NA,N/4,4)) # four time points
        G2<-as.data.frame(matrix(NA,N/4,4)) # four time points
        G3<-as.data.frame(matrix(NA,N/4,4)) # four time points
        G4<-as.data.frame(matrix(NA,N/4,4)) # four time points
```

```
        colnames(G1)=c(paste0("GPA",9:12))
        colnames(G2)=c(paste0("GPA",9:12))
        colnames(G3)=c(paste0("GPA",9:12))
        colnames(G4)=c(paste0("GPA",9:12))

        # intercept (fixed effect) only different models depending on
effect size
        for(t in 1:4){
          G1[,t]=1*(factors[,1]-mean.diff[s])+(t-1)*factors[,2]+(t-
1)^2*factors[,3]+rnorm(N/4,0,sqrt(0.084))
          G2[,t]=1*factors[,1]+(t-1)*factors[,2]+(t-
1)^2*factors[,3]+rnorm(N/4,0,sqrt(0.084))
          G3[,t]=1*(factors[,1]+mean.diff[s])+(t-1)*factors[,2]+(t-
1)^2*factors[,3]+rnorm(N/4,0,sqrt(0.084))
          G4[,t]=1*(factors[,1]+2*mean.diff[s])+(t-1)*factors[,2]+(t-
1)^2*factors[,3]+rnorm(N/4,0,sqrt(0.084))
        }

        # 3 Informative covariates
        G1$ORDINAL=as.ordered(sample(c(1:2),size=N/4,replace=T))
        G1$CONTINUOUS=sample(seq(from=-3.5,to=-
0.7,by=0.1),size=N/4,replace=T) # continuous
        G1$CATEGORY=as.factor(sample(c(10,20,30,40),size=N/4,replace=T))
        G2$ORDINAL=as.ordered(sample(c(1:2),size=N/4,replace=T))
        G2$CONTINUOUS=sample(seq(from=-
0.6,to=3.5,by=0.1),size=N/4,replace=T) # continuous
        G2$CATEGORY=as.factor(sample(c(10,20,30,40),size=N/4,replace=T))
        G3$ORDINAL=as.ordered(sample(c(3:5),size=N/4,replace=T))
        G3$CONTINUOUS=sample(seq(from=-
3.5,to=3.5,by=0.1),size=N/4,replace=T) # continuous
        G3$CATEGORY=as.factor(sample(c(10,20),size=N/4,replace=T))
        G4$ORDINAL=as.ordered(sample(c(3:5),size=N/4,replace=T))
        G4$CONTINUOUS=sample(seq(from=-
3.5,to=3.5,by=0.1),size=N/4,replace=T) # continuous
        G4$CATEGORY=as.factor(sample(c(30,40),size=N/4,replace=T))
        tem=rbind(G1,G2,G3,G4)

        # 4 NOISE VARIABLES #
        tem$N.ORDINAL=as.ordered(sample(c(1:5),size=N,replace=T))
        tem$N.CONTINUOUS1=sample(seq(from=-
3.5,to=3.5,by=0.1),size=N,replace=T)
        tem$N.CONTINUOUS2=round(rnorm(N, mean = 0, sd = 1),1)#
continuous

tem$N.CATEGORY=as.factor(sample(c(10,20,30,40,50),size=N,replace=T))
        True.G=as.factor(rep(c("a","b","c","d"),each=N/4))
        tem=data.frame(tem,True.G) # generated data
        #
        # # GMM using Mplus #
        write.table(tem,"mydat.dat",col.names = F,row.names = T,
quote=F)
        runModels(dir) # running GMM through Mplus
        out=readModels(dir,what="all") # reading Mplus output
        gmm.res=do.call("rbind.fill",sapply(out,"[","summaries"))

mod.fit=gmm.res[,c("BIC","Entropy","T11_LMR_PValue","NLatentClasses")]
        mod.fit.best=mod.fit[order(mod.fit$BIC),] # the lowest BIC
        bc=mod.fit.best$NLatentClasses[1]

        # Number of classes
```

```
gmm.M[[paste0(effectsize[s],"_",size[i],"_",ordinal.option[o],"_",minsiz
e[kk],"_",prune[pp])]][[r]]=mod.fit.best

gmm.NG[[paste0(effectsize[s],"_",size[i],"_",ordinal.option[o],"_",minsi
ze[kk],"_",prune[pp])]][r]=bc

        # gmm class membership
        gmm.cls=out[[bc-1]]$savedata
        gmm.cls$tg=tem$True.G

        # classification accuracy of gmm
        cm=table(gmm.cls$tg,gmm.cls$C) # confusion matrix
        n = sum(cm) # number of instances
        nc = nrow(cm) # number of classes
        diag = diag(cm) # number of correctly classified instances per
class
        rowsums = apply(cm, 1, sum) # number of instances per class
        colsums = apply(cm, 2, sum) # number of predictions per class
        p = rowsums / n # distribution of instances over the actual
classes
        q = colsums / n # distribution of instances over the predicted
classes

        # Overall classification accuracy
        accuracy = sum(diag)/ n # the total number of correct
predictions divided by the total number of predictions made for a
dataset.

        # Per-class precision, recall, and F-1
        precision = diag / colsums # the number of positive class
predictions that actually belong to the positive class.
        recall = diag / rowsums # the number of positive class
predictions made out of all positive examples in the dataset.
        f1 = 2 * precision * recall / (precision + recall) # a single
score that balances both the concerns of precision and recall in one
number.

        #Macro-averaged metrics
        macroPrecision=mean(precision)
        macroRecall=mean(recall)
        macroF1=mean(recall)
        CA=c(accuracy,macroPrecision,macroRecall,macroF1)

gmm.C[[paste0(effectsize[s],"_",size[i],"_",ordinal.option[o],"_",minsiz
e[kk],"_",prune[pp])]][[r]]=CA

        # gmm estimates
        gmm.est=out[[bc-1]]$parameters$unstandardized
        gmm.est2=matrix(NA,nrow=bc,ncol=11)
        for(gm in 1:bc){
          m=subset(gmm.est,LatentClass==gm & paramHeader=="Means")[,3]
          v=subset(gmm.est,LatentClass==gm &
paramHeader=="Variances")[,3]
          sw=subset(gmm.est,LatentClass==gm & paramHeader=="S.WITH")[,3]
          qw=subset(gmm.est,LatentClass==gm & paramHeader=="Q.WITH")[,3]
          rv=subset(gmm.est,LatentClass==gm &
paramHeader=="Residual.Variances")[,3]
          gmm.tem=c(m,v,sw,qw,rv[1],gm)
          gmm.est2[gm,]=gmm.tem
```

```
        }
        gmm.est3=as.data.frame(gmm.est2[order(gmm.est2[,1]),])

gmm.E[[paste0(effectsize[s],"_",size[i],"_",ordinal.option[o],"_",minsiz
e[kk],"_",prune[pp])]][[r]]=gmm.est3
        End=Sys.time()

gmm.times[[paste0(effectsize[s],"_",size[i],"_",ordinal.option[o],"_",mi
nsize[kk],"_",prune[pp])]][[r]]=difftime(End, Start, units = "secs")
        rm(out)
        ###############
        # MOB fitting #
        ###############

        # control option here #
        if(kk==1 & pp==1)

pop.control=mob_control(alpha=0.05,bonferroni=TRUE,ytype="data.frame",or
dinal=ordinal.option[o],vcov="opg") # minsize option
        if(kk==1 & pp==2)

pop.control=mob_control(alpha=0.05,bonferroni=TRUE,ytype="data.frame",pr
une="BIC",ordinal=ordinal.option[o],vcov="opg") # minsize option
        if(kk==2 & pp==1)

pop.control=mob_control(alpha=0.05,bonferroni=TRUE,ytype="data.frame",mi
nsize=250,ordinal=ordinal.option[o],vcov="opg") # minsize option
        if(kk==2 & pp==2)

pop.control=mob_control(alpha=0.05,bonferroni=TRUE,ytype="data.frame",mi
nsize=250,prune="BIC",ordinal=ordinal.option[o],vcov="opg") # minsize
option
        # Fit the data and run MOB
        Start <- Sys.time()
        tr <- mob(GPA9+GPA10+GPA11+GPA12 ~
ORDINAL+CONTINUOUS+CATEGORY+N.ORDINAL+N.CONTINUOUS1+N.CONTINUOUS2+N.CATE
GORY,
                  data = tem,
                  fit = lavaan_fit(QLGM),
                  control = pop.control)

        # number of subgroups #

res.NG[[paste0(effectsize[s],"_",size[i],"_",ordinal.option[o],"_",minsi
ze[kk],"_",prune[pp])]][r]=width(tr)
        # estimates #
        if (width(tr)==1)

est=round(coef(tr)[c("mi","ms","mq","vi","vs","vq","cis","ciq","csq","re
s")],3) else

est=round(coef(tr)[,c("mi","ms","mq","vi","vs","vq","cis","ciq","csq","r
es")],3)

res.est[[paste0(effectsize[s],"_",size[i],"_",ordinal.option[o],"_",mins
ize[kk],"_",prune[pp])]][[r]]=est

        # split points #
        ni=nodeids(tr)
        ni_terminal=nodeids(tr, terminal = TRUE) # terminal node ids
```

```
        ni_inner=ni[!ni %in% ni_terminal] # inner node ids
        a=sapply(ni_inner, function(x)
split_node(node_party(tr[[x]]))$breaks)

res.split[[paste0(effectsize[s],"_",size[i],"_",ordinal.option[o],"_",mi
nsize[kk],"_",prune[pp])]][[r]]=unlist(a)

        # Classification accuracy #
        cm=table(tem[,"True.G"],predict(tr,newdata=tem,type="node")) #
confusion matrix
        n = sum(cm) # number of instances
        nc = nrow(cm) # number of classes
        diag = diag(cm) # number of correctly classified instances per
class
        rowsums = apply(cm, 1, sum) # number of instances per class
        colsums = apply(cm, 2, sum) # number of predictions per class
        p = rowsums / n # distribution of instances over the actual
classes
        q = colsums / n # distribution of instances over the predicted
classes

        # Overall classification accuracy
        accuracy = sum(diag)/ n # the total number of correct
predictions divided by the total number of predictions made for a
dataset.

        # Per-class precision, recall, and F-1
        precision = diag / colsums # the number of positive class
predictions that actually belong to the positive class.
        recall = diag / rowsums # the number of positive class
predictions made out of all positive examples in the dataset.
        f1 = 2 * precision * recall / (precision + recall) # a single
score that balances both the concerns of precision and recall in one
number.

        #Macro-averaged metrics
        macroPrecision=mean(precision)
        macroRecall=mean(recall)
        macroF1=mean(recall)
        CA=c(accuracy,macroPrecision,macroRecall,macroF1)

res.cls[[paste0(effectsize[s],"_",size[i],"_",ordinal.option[o],"_",mins
ize[kk],"_",prune[pp])]][[r]]=CA
        End <- Sys.time()

res.times[[paste0(effectsize[s],"_",size[i],"_",ordinal.option[o],"_",mi
nsize[kk],"_",prune[pp])]][[r]]=difftime(End, Start, units = "secs")
        rm(tr)
        print(paste0(Sys.time(),
effectsize[s],"_",size[i],"_",ordinal.option[o],"_",minsize[kk],"_",prun
e[pp],"_REP = ",r))
        } # one replication
}}}}}
```

REFERENCES

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723.

Allensworth, E. M., & Clark, K. (2020). High school GPAs and ACT scores as predictors of college completion: Examining assumptions about consistency across high schools. *Educational Researcher, 49*(3), 198-211.

Andrews, D.W. K. (1993), Tests for parameter instability and structural change with unknown change point, *Econometrica*, *61*(4), 821–856.

Arnold, M., Oberski, D. L., Brandmaier, A. M., & Voelkle, M. C. (2020). Identifying heterogeneity in dynamic panel models with individual parameter contribution regression. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(4), 613-628.

Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2021). Score-guided structural equation model trees. *Frontiers in Psychology*, *11*, 3913. https://doi.org/10.3389/fpsyg.2020.564403

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using M plus. *Structural equation modeling: A multidisciplinary Journal*, *21*(3), 329-341.

Baker, R. S., Martin, T., & Rossi, L. M. (2016). Educational data mining and learning analytics. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 379-396.

Bauer, D. J., & Shanahan, M. J. (2007). Modeling complex interactions: Person-centered and variable-centered approaches. In T. Little, J. Bovaird, & N. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 255-283). Routledge.

Bowers, A. J., & Sprott, R. (2012). Examining the multiple trajectories associated with dropping out of high school: A growth mixture model analysis. *The Journal of educational research*, *105*(3), 176-195.

Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological methods*, *21*(4), 566.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological methods*, *18*(1), 71.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2014). Exploratory data mining with structural equation model trees. *Contemporary issues in exploratory data mining in the behavioral sciences*, 96-127.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Bürgin, R., & Ritschard, G. (2015). Tree-based varying coefficient regression for longitudinal ordinal responses. *Computational Statistics & Data Analysis*, *86*, 65-80.

Chen, Q., Luo, W., Palardy, G. J., Glaman, R., & McEnturff, A. (2017). The efficacy of common fit indices for enumerating classes in growth mixture models when nested data structure is ignored: A Monte Carlo study. *Sage Open*, *7*(1), 2158244017700459.

Chow, G. C. (1960), Tests of equality between sets of coefficients in two linear regressions, *Econometrica*, 28, 591-605.

Cohen, P., West, S. G., & Aiken, L. S. (2014). *Applied multiple regression/correlation analysis for the behavioral sciences*. Psychology press.

De Gonzalez, A. B., & Cox, D. R. (2007). Interpretation of interaction: A review. *The Annals of Applied Statistics*, *1*(2), 371-385.

Diallo, T. M., Morin, A. J., & Parker, P. D. (2014). Statistical power of latent growth curve models to detect quadratic growth. *Behavior research methods*, *46*(2), 357-371.

Dimitrov, D. M., Al-Saud, F. A. A. M., & Alsadaawi, A. S. (2015). Investigating population heterogeneity and interaction effects of covariates: The case of a large-scale assessment for teacher licensure in Saudi Arabia. *Journal of Psychoeducational Assessment*, *33*(7), 674-686.

Eo, S. H., & Cho, H. (2014). Tree-structured mixed-effects regression modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, *23*(3), 740-760.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The journal of machine learning research*, *15*(1), 3133-3181.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior research methods*, *50*(5), 2016-2034.

Fu, W., & Simonoff, J. S. (2015). Unbiased regression trees for longitudinal and clustered data. *Computational Statistics & Data Analysis*, *88*, 53-74.

Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika, 24*(3), 229-252.

Grimm, K. J., & Jacobucci, R. (2020). Reliable Trees: Reliability Informed Recursive Partitioning for Psychological Data. *Multivariate behavioral research*, 1-13. https://doi.org/10.1080/00273171.2020.1751028

Grün, B., Kosmidis, I., & Zeileis, A. (2012). Extended Beta Regression in R: Shaken. *Stirred, Mixed, and Partitioned*, *2012*(48), 25.

Hajjem, A., Larocque, D., & Bellavance, F. (2017). Generalized mixed effects regression trees. *Statistics & Probability Letters*, *126*, 114-118.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: an R package for facilitating large-scale latent variable analyses in M plus. *Structural equation modeling: a multidisciplinary journal*, *25*(4), 621-638.

Hansen, B. E. (1997). Approximate asymptotic p values for structuras-change tests. *Journal of Business & Economic Statistics*, *15*(1), 60-67.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, *21*(4), 447.

Hjort, N. L., & Koning, A. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, *14*(1-2), 113-132.

Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research*, *16*(1), 3905-3909.

Hu, J., Leite, W. L., & Gao, M. (2017). An evaluation of the use of covariates to assist in class enumeration in linear growth mixture modeling. *Behavior Research Methods*, *49*(3), 1179-1190.

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809-816.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2017). A comparison of methods for uncovering sample heterogeneity: Structural equation model trees and finite mixture models. *Structural equation modeling: a multidisciplinary journal*, *24*(2), 270-282.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409-426.

Kim, J., & Cho, H. (2019). Seemingly unrelated regression tree. *Journal of Applied Statistics*, *46*(7), 1177-1195.

Kleiber, C., Hornik, K., Leisch, F., & Zeileis, A. (2002). strucchange: An r package for testing for structural change in linear regression models. *Journal of statistical software*, *7*(2), 1-38.

Kundu, M. G., & Harezlak, J. (2019). Regression trees for longitudinal data with baseline covariates. *Biostatistics & epidemiology*, *3*(1), 1-22.

Le, T. T., & Moore, J. H. (2021). treeheatr: an R package for interpretable decision tree visualizations. *Bioinformatics*, *37*(2), 282-284.

Liu, J. (2020). Extending mixture of experts model to investigate heterogeneity of trajectories: when, where and how to add which covariates. *arXiv preprint arXiv:2007.02432*.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal

mixture. *Biometrika*, *88*(3), 767-778.

Loh, W. Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329-348.

Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods*, *10*(1), 21.

Lubke, G. H., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(1), 26-47.

Magidson, J., & Vermunt, J. (2004). Latent class models. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 175–198). Newbury Park, CA: Sage.

McCutcheon, A. C. (1987). *Latent class analysis.* Beverly Hills, CA: Sage.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* (Vol. 38). New York: M. Dekker.

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, *6*, 355-378.

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

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research*, *13*(2), 127-143.

Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika*, *78*(1), 59-82.

Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika, 79*(4), 569-584.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, *58*(302), 415-434.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*(2), 463-469.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus: Statistical Analysis with Latent Variables: User's Guide (Version 8)*. Los Angeles, CA: Authors.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., ... & Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*(2), 535-549.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling: A multidisciplinary Journal*, *14*(4), 535-569.

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 782-797.

Nylund-Gibson, K., Grimm, R. P., & Masyn, K. E. (2019). Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 967-985.

Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., & Baker, R. (2020). Who's Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining*, *12*(3), 1-30.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). SAGE.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of statistical software*, *48*(2), 1-36.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271-282.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461-464.

Serang, S. (2021). A comparison of three approaches for identifying correlates of heterogeneity in change. *New Directions for Child and Adolescent Development*.

Serang, S., Jacobucci, R., Stegmann, G., Brandmaier, A. M., Culianos, D., & Grimm, K. J. (2020). Mplus Trees: structural equation model trees using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-11.

Stegmann, G., Jacobucci, R., Serang, S., & Grimm, K. J. (2018). Recursive partitioning with nonlinear models of change. *Multivariate behavioral research*, *53*(4), 559-570.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289-316.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, *14*(4), 323.

Strobl, C., Wickelmaier, F., & Zeileis, A. (2011). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics*, *36*(2), 135-153.

Su, X., Wang, M., & Fan, J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, *13*(3), 586-598.

Usami, S., Hayes, T., & McArdle, J. (2017). Fitting structural equation model trees and latent growth curve mixture models in longitudinal designs: The influence of model misspecification. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 585-598.

Usami, S., Jacobucci, R., & Hayes, T. (2019). The performance of latent growth curve model-based structural equation model trees to uncover population heterogeneity in growth trajectories. *Computational Statistics*, *34*(1), 1-22.

van Wie, M. P., Li, X., & Wiedermann, W. (2019). Identification of confounded subgroups using linear model-based recursive partitioning. *Psychological Test and Assessment Modeling*, *61*(4), 365-387.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis*, *18*(4), 450-469.

Wang, T., Merkle, E. C., Anguera, J. A., & Turner, B. M. (2021). Score-based tests for detecting heterogeneity in linear mixed models. *Behavior Research Methods*, *53*(1), 216-231.

Wang, T., Strobl, C., Zeileis, A., & Merkle, E. C. (2018). Score-based tests of differential item functioning via pairwise maximum likelihood estimation. *psychometrika*, *83*(1), 132-155.

Wang, Y., Kim, E., Ferron, J. M., Dedrick, R. F., Tan, T. X., & Stark, S. (2020). Testing measurement invariance across unobserved groups: The role of covariates in factor mixture modeling. *Educational and Psychological Measurement*, 0013164420925122.

Zeileis, A. (2020, September 7). Structural equation model trees with partykit and lavaan. *Research homepage of Achim Zeileis*. https://www.zeileis.org/news/lavaantree/

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488-508.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492-514.

Zeileis, A., Strobl, C., Wickelmaier, F., Komboz, B., and Kopf, J. (2020). Psychotree: recursive partitioning based on psychometric models (Version 0.15- 3) [Computer software].