Promoting Data Science

Volume 1 2021 Number 2

Journal of Behavioral Data Science

https://isdsa.org

JOURNAL OF BEHAVIORAL DATA SCIENCE

ISBN: 2575-8306 (Print) 2574-1284 (Online)

https://isdsa.org/jbds

JOURNAL OF BEHAVIORAL DATA SCIENCE

No Publication Charge and Open Access

jbds@isdsa.org

List of Articles

Journal of Behavioral Data Science, 2021, *1* (2), 1–30.
DOI: https://doi.org/10.35566/jbds/v1n2/p1

Bayesian Approach to Non-ignorable Missingness in Latent Growth Models

Zhenqiu (Laura) Lu[1][0000−0001−9482−1368] and Zhiyong Zhang[2][0000−0003−0590−2196]

[1] University of Georgia, Athens, GA 30602, USA
zlu@uga.edu
[2] University of Notre Dame, Notre Dame, IN 46530, USA
zhiyongzhang@nd.edu

Abstract. Latent growth curve models (LGCMs) are becoming increasingly important among growth models because they can effectively capture individuals' latent growth trajectories and also explain the factors that influence such growth by analyzing the repeatedly measured manifest variables. However, with the increase in complexity of LGCMs, there is an increase in issues on model estimation. This research proposes a Bayesian approach to LGCMs to address the perennial problem of almost all longitudinal research, namely, missing data. First, different missingness models are formulated. We focus on non-ignorable missingness in this article. Specifically, these models include the latent intercept dependent missingness, the latent slope dependent missingness, and the potential outcome dependent missingness. To implement the model estimation, this study proposes a full Bayesian approach through data augmentation algorithm and Gibbs sampling procedure. Simulation studies are conducted and results show that the proposed method accurately recover model parameters and the mis-specified missingness may result in severely misleading conclusions. Finally, the implications of the approach and future research directions are discussed.

Keywords: Bayesian Estimation · Missing Data · Latent Growth Curve Models · Non-ignorable Missingness · Longitudinal Analysis · Multilevel Modeling

1 Introduction

In social and behavioral sciences, there has been great interest in the analysis of change (e.g., Collins, 1991; Lu, Zhang, & Lubke, 2010; Singer & Willett, 2003). Growth modeling is designed to provide direct information of growth by measuring the variables of interest on the same participants repeatedly through time (e.g., Demidenko, 2004; Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2008; Fitzmaurice, Laird, & Ware, 2004; Hedeker & Gibbons, 2006; Singer

& Willett, 2003). Among the most popular growth models, *latent growth curve models (LGCMs)* are becoming increasingly important because they can effectively capture individuals' latent growth trajectories and also explain the latent factors that influence such growth by analyzing the repeatedly measured manifest variables (e.g., Baltes & Nesselroade, 1979). Manifest variables are evident in the data, such as observed scores; latent variables cannot be measured directly and are essentially hidden in the data, such as the latent initial levels and latent growth rates (Singer & Willett, 2003). We use the term "latent" because these variables are not directly observable but rather are assumed to be inferred, although they may be closely related to observed scores. For example, the latent intercept (i.e., the latent initial level) may be related to the test score at the first occasion, the prior knowledge of a course (such as mathematical knowledge), or other similar variables. The latent slope (i.e., the latent growth rate) may be related to the participant's learning ability, the attitude toward the course, the instructor's teaching methods, or other similar types of variables.

However, with an increase in complexity of LGCMs, comes an increase in difficulties estimating such models. First, missing data are almost inevitable with longitudinal data (e.g., Jelicic, Phelps, & Lerner, 2009; Little & Rubin, 2002). Second, conventional likelihood estimation procedures might fail for complex models with complicated data structures.

1.1 Missing Data

As LGCMs involve data collection on the same participants through multiple waves of surveys, tests, or questionnaires, missing data are almost inevitable. Research participants may drop out of a study, or some students may miss a test due to absence or fatigue (e.g., Little & Rubin, 2002; Schafer, 1997). Missing data can be investigated from their mechanisms, or why missing data occur. Little and Rubin (2002) distinguished *ignorable missingness mechanism* and *non-ignorable missingness mechanism*. For ignorable missingness mechanism, estimates are usually asymptotically consistent when the missingness is ignored (Little & Rubin, 2002), because the parameters that govern the missing process either are distinct from the parameters that govern the model outcomes or depend on the observed variables in the model. The non-ignorable missingness mechanism is also referred to as *missing not at random (MNAR)*, in which the missing data probability depends either on unobserved outcomes, or on latent variables that cannot be fully measured by the observed data, in other words, latent variables that depend on the missing values.

With the appearance of missing data comes the challenge in estimating growth model parameters. To address the challenge, statisticians have developed different approaches and models. Although there are a large amount of literature to address this problem in applied behavioral sciences (e.g., Acock, 2005; Schafer & Graham, 2002; Schlomer, Bauman, & Card, 2010), especially in longitudinal studies (e.g., Jelicic et al., 2009; Roth, 1994), the majority of the literature is on ignorable missingness. This is mainly because (1) analysis models or techniques for non-ignorable missing data are traditionally difficult to implement and not yet easy to use (e.g., Baraldi & Enders, 2010), and (2) missingness mechanisms are not testable (e.g., Little & Rubin, 2002). At the same time, however, non-ignorable missingness analysis is a crucial and a serious concern in applied research areas, in which participants may be dropping out for reasons directly related to the response being measured (e.g., Baraldi & Enders, 2010; Enders, 2011b; Hedeker & Gibbons, 1997). Not attending to the non-ignorable missingness may result in severely biased statistical estimates, standard errors, and associated confidence intervals, and thus poses substantial risk of leading researchers to incorrect conclusions (e.g., Little & Rubin, 2002; Schafer, 1997; Zhang & Wang, 2012).

In a study of latent growth models, Lu, Zhang, and Lubke (2011) investigated non-ignorable missingness in mixture models. However, the missingness in that study was only allowed to depend on latent class membership. In practice, even within one population, the missingness may depend on many other latent variables, such as latent initial levels and latent growth rates. When observed data are not completely informative about these latent variables, the missingness is non-ignorable. Furthermore, Lu et al. (2011) did not examine how to identify the missingness mechanisms. Accordingly, this study extends previous research to more general non-ignorable missingness and also investigates the influences of different types of non-ignorable missingness on model estimation.

1.2 Bayesian Approach

To implement the model estimation, we propose a full Bayesian approach. Traditionally, maximum likelihood methods have been widely used in most studies for estimating parameters of models in the presence of missing data (e.g., Enders, 2011a; Muthén, Asparouhov, Hunter, & Leuchter, 2011), and statistical inferences have been carried out using conventional likelihood procedures (e.g., Yuan & Lu, 2008). Recently, multiple imputation (MI) methods have been proposed as an alternative approach (e.g., Enders, 2011a; Muthén et al., 2011). MI is a Monte Carlo technique that replaces the missing values with multiple simulated values to generate multiple complete datasets. Each of these simulated datasets is then analyzed using methods that do not account for missingness, that is, using standard analytical methods. Results are then combined to produce estimates and confidence intervals that incorporate the uncertainty due to the missing-data (Enders, 2011a; Rubin, 1987; Schafer, 1997). Both ML and MI estimation methods typically assume that missing data mechanisms are MCAR or MAR. Further, using conventional estimation procedures may fail or may provide biased estimates (Yuan & Zhang, 2012) when estimating model parameters in complex models with complicated data structures such as GMMs with missing data and outliers. In addition, MI requires data to be imputed under a particular model (e.g., Allison, 2002; Newman, 2003). And literature also shows that multiple imputation is inappropriate as a general purpose methodology for complex problems or large datasets (e.g., Fay, 1992). When missingness is MNAR, most work augments the basic analysis using with a model that explains the probability of missing data (e.g., Enders, 2011a; Muthén et al., 2011).

In this article, a full Bayesian estimate approach (e.g., Lee, 2007; Muthén & Asparouhov, 2012) is proposed. There are several advantages. First, this approach involves Gibbs sampling methods (Geman & Geman, 1984). Gibbs sampling is especially useful when the joint distribution is complex or unknown but the conditional distribution of each variable is available. The sequence of samples constructs a Markov chain that can be shown to be ergodic (Geman & Geman, 1984). That is, once convergence is obtained, the samples can be assumed to be independent draws from the stationary distribution. Thus, after convergence the generated value is actually from the joint distribution of all parameters. Each variable from the Markov chain has also been shown to converge to the marginal distribution of that variable (Robert & Casella, 2004). Additional advantages of Bayesian methods include their intuitive interpretations of statistical results, their flexibility in incorporating prior information about how data behave in similar contexts and findings from experimental research, their capacity for dealing with small sample sizes (such as occur with special populations), and their flexibility in the analysis of complex statistical models with complicated data structure (e.g., Dunson, 2000; Scheines, Hoijtink, & Boomsma, 1999).

1.3 Goals and Structure

The goals of the paper are to propose latent growth curve models with non-ignorable missingness and to evaluate the performance of Bayesian methods to recover model parameters. The rest of the article consists of four sections. Section 2 presents and formulates three non-ignorable missingness selection models. Section 3 presents a full Bayesian method to estimate the latent growth models through the data augmentation and Gibbs sampling algorithms. Section 4 conducts a simulation study. Estimates from models with different non-ignorable missingness and different sample sizes are summarized, analyzed, and compared. Conclusions based on the simulation study are drawn. Section 5 discusses the implications and future directions of this study. Finally, the appendix presents technical details.

2 Non-ignorable Missingness in Latent Growth Models

In this section, we model the non-ignorable missingness in growth models. Before we introduce the three selection models, we first review the latent growth curve models (LGCMs).

2.1 Latent Growth Curve Models (LGCMs)

The latent growth curve models (LGCMs) can be expressed by a regression equation with latent variables being regressors. Specifically, for a longitudinal study with N subjects and T measurement time points, let $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{iT})'$ be a $T \times 1$ random vector, where y_{it} stands for the outcome or observation of individual i at occasion t ($i = 1, 2, ..., N$; $t = 1, 2, ..., T$), and let $\boldsymbol{\eta}_i$ be a $q \times 1$ random vector containing q continuous latent variables. A latent growth curve model for the outcome $\mathbf{y}_i$ related to the latent $\boldsymbol{\eta}_i$ can be written as

$$\mathbf{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \mathbf{e}_i \tag{1}$$

$$\boldsymbol{\eta}_i \ = \boldsymbol{\beta} + \boldsymbol{\xi}_i, \tag{2}$$

where $\boldsymbol{\Lambda}$ is a $T \times q$ matrix consisting of factor loadings, $\mathbf{e}_i$ is a $T \times 1$ vector of residuals or measurement errors that are assumed to follow multivariate normal distributions, i.e., $\mathbf{e}_i \sim MN_T(\mathbf{0}, \boldsymbol{\Theta})$ [1], and $\boldsymbol{\xi}_i$ is a $q \times 1$ vector that is assumed to follow a multivariate distribution, i.e., $\boldsymbol{\xi}_i \sim MN_q(\mathbf{0}, \boldsymbol{\Psi})$. In LGCMs, $\boldsymbol{\beta}$ is called *fixed effects* and $\boldsymbol{\xi}_i$ is called *random effects* (e.g., Fitzmaurice et al., 2004; Hedges, 1994; Luke, 2004; Singer & Willett, 2003). The vectors $\boldsymbol{\beta}$, $\boldsymbol{\eta}_i$, and the matrix $\boldsymbol{\Lambda}$ determine the growth trajectory of the model. For instance, when $q = 2$, $\boldsymbol{\beta} = (I, S)'$, $\boldsymbol{\eta}_i = (I_i, S_i)'$, and $\boldsymbol{\Lambda}$ is a $T \times 2$ matrix containing the first column of 1s and the second column of $(0, 1, ..., T-1)$. The corresponding model represents a linear growth model in which I is the latent population intercept (or latent random initial level), S is the latent population slope, I_i is individual i's latent random intercept and S_i is individual i's latent random slope. Furthermore, when $q = 3$, $\boldsymbol{\beta} = (I, S, Q)'$, $\boldsymbol{\eta}_i = (I_i, S_i, Q_i)'$, and $\boldsymbol{\Lambda}$ is a $T \times 3$ matrix containing the first column of 1s, the second column of $(0, 1, ..., T-1)$, and the third column of $(0, 1, ..., (T-1)^2)$. The corresponding model represents a quadratic growth curve model with Q and Q_i being latent quadratic coefficients for population and individual i, respectively.

2.2 Selection Models for Non-ignorable Missingness

To address the non-ignorable missingness, there are two general approaches, *pattern-mixture models* (Hedeker & Gibbons, 1997; Little & Rubin, 1987) and *selection models* (Glynn, Laird, & Rubin, 1986; Little, 1993, 1995). In both cases, the statistical analysis requires joint modelling of dependent variable and missing data processes. In this research, selection models are used, mainly for two reasons. First, substantively, it seems more natural to consider the behavior of the response variable in the full target population of interests, rather than in the sub-populations defined by missing data patterns (e.g., Fitzmaurice et al., 2008). Second, the selection model formulation leads directly to the joint distribution of both dependent variables and the missingness (e.g., Fitzmaurice et al., 2008) as follows,

$$f(\mathbf{y}_i, \mathbf{m}_i | \boldsymbol{\nu}, \boldsymbol{\phi}, \mathbf{x}_i) = f(\mathbf{y}_i | \boldsymbol{\nu}, \mathbf{x}_i)\, f(\mathbf{m}_i | \mathbf{y}_i, \boldsymbol{\nu}, \boldsymbol{\phi}, \mathbf{x}_i)$$

where $f(.)$ is a density function, $\mathbf{x}_i$ is a vector of covariates for individual i, $\mathbf{y}_i$ is a vector of individual i's outcome scores, $\boldsymbol{\Theta} = (\boldsymbol{\nu}, \boldsymbol{\phi})$ are all parameters in the model, in which $\boldsymbol{\nu}$ are parameters for the growth model and $\boldsymbol{\phi}$ are

[1] Throughout the article, $MN_n(\cdot)$ denotes a n-dimensional multivariate normal distribution, and $Mt_n(\cdot)$ denotes a n-dimensional multivariate t distribution.

parameters for the missingness, and $\mathbf{m}_i$ is a vector $\mathbf{m}_i = (m_{i1}, m_{i2}, ..., m_{iT})'$ that indicates the missingness status for $\mathbf{y}_i$. Specifically, if y_i is missing at time point t, then $m_{it} = 1$; otherwise, $m_{it} = 0$. Here, we assume the missingness is conditionally independent (e.g., Dawid, 1979), which means across different occasions the conditional distributions of missingness are independent with each other. Let $\tau_{it} = f(m_{it} = 1)$ be the probability that y_{it} is missing, then m_{it} follows a Bernoulli distribution of τ_{it}, and the density function of m_{it} is $f(m_{it}) = \tau_{it}^{m_{it}}(1 - \tau_{it})^{1-m_{it}}$. For different non-ignorable missingness patterns, the expressions of τ_{it} are different. Lu et al. (2011) investigated the non-ignorable missingness in mixture models. The τ_{it} in that article is a function of latent class membership, and thus the missingness is *Latent Class Dependent (LCD)*.

However, *LCD* was proposed in the framework of mixture models. Within each latent population, there is no class membership indicator. Consequently, the missingness is ignorable. In this article, we consider more complex non-ignorable missingness mechanisms within a population. In general, we assume $\mathbf{L}_i$ is a vector of latent variables that depend on the missing values. A general class of selection models for dealing with non-ignorable missing data in latent growth modelling can be formulated as

$$\begin{aligned} & f(\mathbf{y}_i, \mathbf{m}_i|\boldsymbol{\beta}, \boldsymbol{\xi}_i, \mathbf{L}_i, \boldsymbol{\gamma}_t, \mathbf{x}_i) = f(\boldsymbol{\eta}_i|\boldsymbol{\beta}, \boldsymbol{\xi}_i) f(\mathbf{y}_i|\boldsymbol{\eta}_i)\, \Phi(\boldsymbol{\omega}_i'\boldsymbol{\gamma}_t)^{m_{it}}[1 - \Phi(\boldsymbol{\omega}_i'\boldsymbol{\gamma}_t)]^{1-m_{it}} \\ & = f(\boldsymbol{\eta}_i|\boldsymbol{\beta}, \boldsymbol{\xi}_i) f(\mathbf{y}_i|\boldsymbol{\eta}_i) \Phi(\gamma_{0t} + \mathbf{L}_i\boldsymbol{\gamma}_{Lt} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt})^{m_{it}} \\ & \times[1 - \Phi(\gamma_{0t} + \mathbf{L}_i\boldsymbol{\gamma}_{Lt} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt})]^{1-m_{it}} \qquad (3) \end{aligned}$$

where $\mathbf{x}_i$ is an r-dimensional vector, $\boldsymbol{\omega}_i = (1, \mathbf{L}_i', \mathbf{x}_i')'$ and $\boldsymbol{\gamma}_t = (\gamma_{0t}, \boldsymbol{\gamma}_{Lt}', \boldsymbol{\gamma}_{xt}')'$. The missingness is non-ignorable because it depends on the latent variables $\mathbf{L}_i$ in the model and the observed data are not completely informative about these latent variables. Note that the vector $\boldsymbol{\gamma}_{Lt}$ here should be non-zero. Otherwise, the missingness becomes ignorable.

Specific sub-models under different situations can be derived from this general model. For example, missingness may be related to latent intercepts, latent growth rates, or potential outcomes. To show different types of non-ignorable missingness, we draw the path diagrams, as shown in Figures 1, 2, and 3, to illustrate the sub-models. These sub-models are based on three types of latent variables on which the missingness might depend. In these path diagrams, a square/rectangle indicates an observed variable, a circle/oval means a latent variable, a triangle represents a constant, and arrows show the relationship among them. y_t is the outcome at time t, which is influenced by latent effects such as I, S, and η_q. As the value of y_t might be missing, we use both circle and square in the path diagram. If y_t is missing, then the potential outcome cannot be observed and the corresponding missingness indicator m_t becomes 1. The dashed lines between y_t and m_t show the 1-1 relationship. In these sub-models, the value of m_t depends on the observed covariate x_r and some latent variables. The details of these three sub-models are described as follows.

2.2.1 Latent Intercept Dependent (LID) Missingness (Figure 1). It illustrates the situation where the missingness depends on individual's latent intercept, I_i. For example, a student's latent initial ability level of the knowledge of a course influences the likelihood of that participant dropping out of or staying in that course. If the latent initial ability of a course is not high, a student may choose to drop that course or even drop out a school. In the case of LID, the $\mathbf{L}_i$ in Equation (3) is simplified to a univariate I_i. Suppose that the missingness is also related to some observed covariates $\mathbf{x}_i$, such as parents' education or family income, then τ_{Iit} is expressed as a probit link function of I_i and $\mathbf{x}_i$

$$\tau_{Iit} = \Phi(\gamma_{0t} + I_i\gamma_{It} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt}) = \Phi(\boldsymbol{\omega}'_{Ii}\,\boldsymbol{\gamma}_{It}), \tag{4}$$

where $\boldsymbol{\omega}_{Ii} = (1, I_i, \mathbf{x}_i')'$ and $\boldsymbol{\gamma}_{It} = (\gamma_{0t}, \gamma_{It}, \boldsymbol{\gamma}_{xt}')'$.

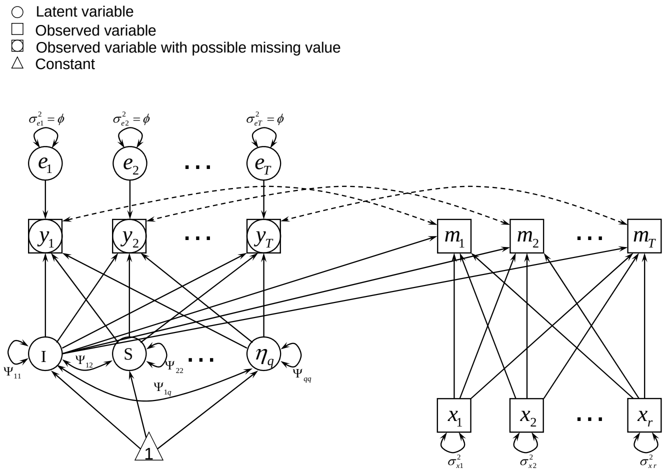


Figure 1. Path diagram of a latent growth model with latent intercept dependent missingness (LID) where $f(m_t)$ depends on covariates x_rs and latent intercept I.

2.2.2 Latent Slope Dependent (LSD) Missingness (Figure 2).

It describes the situation where the missingness depends on the latent slope, S_i. For example, a student's latent rate of change in a course influences the likelihood that the participant misses a test in the future. This might be the case, if the student didn't see any improvement over time, at which point he/she might choose to drop out. In the case of LSD, the $\mathbf{L}_i$ in Equation (3) becomes a univariate S_i. Together with some other observed covariates $\mathbf{x}_i$, for example, parents' education or family income, the missing data rate τ_{it} can be expressed as a probit link function of S_i and $\mathbf{x}_i$,

$$\tau_{Sit} = \Phi(\gamma_{0t} + S_i\gamma_{St} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt}) = \Phi(\boldsymbol{\omega}'_{Si}\,\boldsymbol{\gamma}_{St}), \tag{5}$$

with $\boldsymbol{\omega}_{Si} = (1, S_i, \mathbf{x}_i')'$ and $\boldsymbol{\gamma}_{St} = (\gamma_{0t}, \gamma_{St}, \boldsymbol{\gamma}_{xt}')'$.

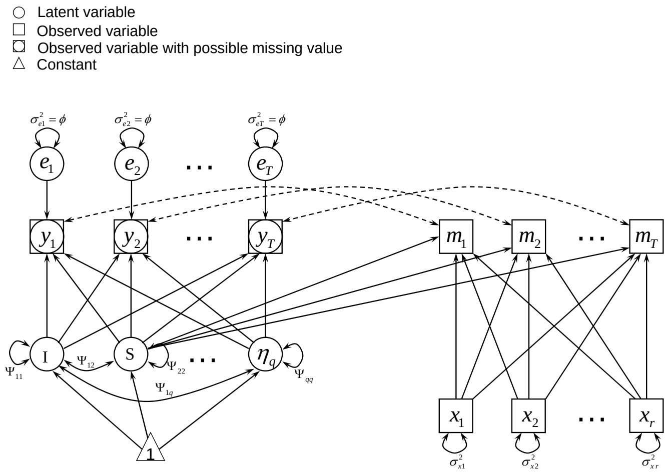


Figure 2. Path diagram of a latent growth model with latent slope dependent missingness (LSD) where $f(m_t)$ depends on covariates x_rs and latent slope S.

2.2.3 Latent Outcome Dependent (LOD) Missingness (Figure 3). It assumes that the missingness depends on potential outcomes y_{it}. For example, a student who feels not doing well on a test may be more likely to quit taking the rest of the test. As a result, the missing score is due to the perceived potential outcome of the test. In this case, the $\mathbf{L}_i$ in Equation (3) is the potential outcome y_{it}. With some covariates $\mathbf{x}_i$, we express τ_{it} as a probit link function as follows.

$$\tau_{yit} = \Phi(\gamma_{0t} + y_{it}\gamma_{yt} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt}) = \Phi(\boldsymbol{\omega}_{yit}'\boldsymbol{\gamma}_{yt}), \tag{6}$$

with $\boldsymbol{\omega}_{yit} = (1, y_{it}, \mathbf{x}_i')'$ and $\boldsymbol{\gamma}_{yt} = (\gamma_{0t}, \gamma_{yt}, \boldsymbol{\gamma}_{xt}')'$.

3 Bayesian Estimation

In this article, a full Bayesian estimation approach is used to estimate growth models. The algorithm is described as follows. First, model related latent variables are added via the data augmentation method (Tanner & Wong, 1987). By including auxiliary variables, the likelihood function for each model is obtained. Second, proper priors are adopted. Third, with the likelihood function and the priors, based on the Bayes' Theorem, the posterior distribution of

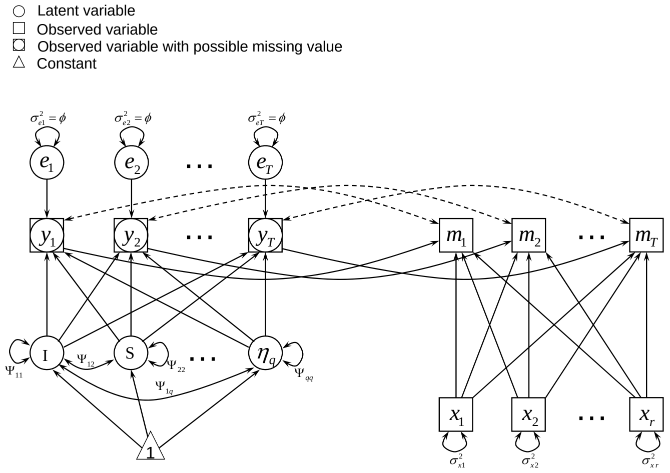


Figure 3. Path diagram of a latent growth model with latent outcome dependent missingness (LOD) where $f(m_t)$ depends on covariates x_rs and potential outcome y.

the unknown parameters is readily available. We obtain conditional posterior distributions instead of the joint posterior because the integrations of marginal posterior distributions of the parameters are usually hard to obtain explicitly for high-dimensional data. Fourth, with conditional posterior distributions, Markov chains are generated for the unknown model parameters by implementing a Gibbs sampling algorithm (Casella & George, 1992; Geman & Geman, 1984). Finally, statistical inference is conducted based on converged Markov chains.

3.1 Data Augmentation and Likelihood Functions

In order to construct the likelihood function explicitly, we use the data augmentation algorithm (Tanner & Wong, 1987). The observed outcomes $\mathbf{y}_i^{obs}$ can be augmented with the missing values $\mathbf{y}_i^{mis}$ such that $\mathbf{y}_i = (\mathbf{y}_i^{obs}, \mathbf{y}_i^{mis})'$ for individual i. Also, the missing data indicator variable $\mathbf{m}_i$ is added to models. Then the joint likelihood function of the selection model for the i^{th} individual can be expressed as $L_i(\boldsymbol{\eta}_i, \mathbf{y}_i, \mathbf{m}_i) = [f(\boldsymbol{\eta}_i)\, f(\mathbf{y}_i|\boldsymbol{\eta}_i)]\ f(\mathbf{m}_i|\mathbf{y}_i, \boldsymbol{\eta}_i, \mathbf{x}_i)$. For the whole sample, the likelihood function is specifically expressed as

$$
\begin{aligned}
L(\mathbf{y}, \boldsymbol{\eta}, \mathbf{m}) \propto \prod_{i=1}^{N} \Bigg\{ & |\boldsymbol{\Psi}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\eta}_i - \boldsymbol{\beta})' \boldsymbol{\Psi}^{-1} (\boldsymbol{\eta}_i - \boldsymbol{\beta})\right] \\
& \times |\phi|^{-T/2} \exp\left[-\frac{1}{2\phi}(\mathbf{y}_i - \boldsymbol{\Lambda}\boldsymbol{\eta}_i)'(\mathbf{y}_i - \boldsymbol{\Lambda}\boldsymbol{\eta}_i)\right] \\
& \times \prod_{t=1}^{T} \left[\tau_{it}^{m_{it}} (1 - \tau_{it})^{1-m_{it}}\right] \Bigg\},
\end{aligned}
\tag{7}
$$

where τ_{it} is defined by Equation (4) for the LID missingness, (5) for the LSD missingness, and (6) for the LOD missingness.

3.2 Prior and Posterior Distributions

The commonly used proper priors (e.g., Lee, 2007) are adopted in the study. Specifically, (1) an inverse Gamma distribution prior is used for $\phi \sim IG(v_0/2, s_0/2)$ where v_0 and s_0 are given hyper-parameters. The density function of an inverse Gamma distribution is $f(\phi) \propto \phi^{-(v_0/2)-1} \exp(-s_0/(2\phi))$. (2) An inverse Wishart distribution prior is used for $\boldsymbol{\Psi}$. With hyper-parameters m_0 and $\mathbf{V}_0$, $\boldsymbol{\Psi} \sim IW(m_0, \mathbf{V}_0)$, where m_0 is a scalar and $\mathbf{V}_0$ is a $q \times q$ matrix. Its density function is $f(\boldsymbol{\Psi}) \propto |\boldsymbol{\Psi}|^{-(m_0+q+1)/2} \exp[-tr(\mathbf{V}_0\boldsymbol{\Psi}^{-1})/2]$. (3) For $\boldsymbol{\beta}$ a multivariate normal prior is used, and $\boldsymbol{\beta} \sim MN_q(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ where the hyper-parameter $\boldsymbol{\beta}_0$ is a q-dimensional vector and $\boldsymbol{\Sigma}_0$ is a $q \times q$ matrix. (4) The prior for $\boldsymbol{\gamma}_t$ $(t = 1, 2, \ldots, T)$ is chosen to be a multivariate normal distribution $\boldsymbol{\gamma}_t \sim MN_{(2+r)}(\boldsymbol{\gamma}_{t0}, \mathbf{D}_{t0})$, where $\boldsymbol{\gamma}_{t0}$ is a $(2+r)$-dimensional vector, $\mathbf{D}_{t0}$ is a $(2+r) \times (2+r)$ matrix, and both are pre-determined hyper-parameters.

After constructing the likelihood function and assigning the priors, the joint posterior distribution for unknown parameters is readily available. Considering the high-dimensional integration for marginal distributions of parameters, the conditional distribution for each parameter is obtained instead. The derived conditional posteriors are provided in Equations (8) - (11) in the appendix. In addition, the conditional posteriors for the latent variable $\boldsymbol{\eta}_i$ and the augmented missing data $\mathbf{y}_i^{mis}$ $(i = 1, 2, ..., N)$ are also provided by Equations (12) and (13), respectively, in the appendix.

3.3 Gibbs Sampling

After obtaining the conditional posteriors, the Markov chain for each model parameter is generated by implementing a Gibbs sampling algorithm (Casella & George, 1992; Geman & Geman, 1984). Specifically, the following algorithm is used in the research.

1. Start with a set of initial values for model parameters $\phi^{(0)}$, $\boldsymbol{\Psi}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$, latent variable $\boldsymbol{\eta}^{(0)}$, and missing values $\mathbf{y}^{mis(0)}$.

2. At the s^{th} iteration, the following parameters are generated: $\phi^{(s)}$, $\mathbf{\Psi}^{(s)}$, $\boldsymbol{\beta}^{(s)}$, $\boldsymbol{\gamma}^{(s)}$, $\boldsymbol{\eta}^{(s)}$, and $\mathbf{y}^{mis(s)}$. To generate $\phi^{(s+1)}$, $\mathbf{\Psi}^{(s+1)}$, $\boldsymbol{\beta}^{(s+1)}$, $\boldsymbol{\gamma}^{(s+1)}$, $\boldsymbol{\eta}^{(s+1)}$, and $\mathbf{y}^{mis(s+1)}$, the following procedure is implemented:
(a) Generate $\phi^{(s+1)}$ from the distribution in Equation (8) in the appendix.
(b) Generate $\mathbf{\Psi}^{(s+1)}$ from the inverse Wishart distribution in Equation (9) in the appendix. iv. Generate $\boldsymbol{\beta}^{(s+1)}$ from the multivariate normal distribution in Equation (10) in the appendix.
(c) Generate $\boldsymbol{\gamma}^{(s+1)}$ from the distribution in Equation (11) in the appendix.
(d) Generate $\boldsymbol{\eta}^{(s+1)}$ from the multivariate normal distribution in Equation (12) in the appendix.
(e) Generate $\mathbf{y}^{mis(s+1)}$ from the normal distribution in Equation (13) in the appendix.

3.4 Statistical Inference

After passing convergence tests, the generated Markov chains can be viewed as from the joint and marginal distributions of all parameters. The statistical inference can then be conducted based on the generated Markov chains.

Suppose θ is an unknown parameter. For different loss functions of θ, the point estimates are different. For example, if a square loss function, $LF = (\theta - \hat{\theta})^2$, is used, then the posterior mean is the estimate of θ; but if an absolute loss function, $LF = |\theta - \hat{\theta}|$, is used, then its estimate is the posterior median. There are other function forms, such as 0-1 loss function, but in this research we use the square loss function.

Let $\mathbf{\Theta} = (\theta_1, \theta_2, ..., \theta_p)'$ denote a vector of all the unknown parameters in the model. Then the converged Markov chains can be recorded as $\mathbf{\Theta}^{(s)}, s = 1, 2, \ldots, S$, and each parameter estimate $\hat{\theta_j}$ $(j = 1, 2, ..., p)$ can be calculated as $\hat{\theta_j} = \sum_{s=1}^{S} \theta_j^{(s)}/S$ with standard error (SE) $s.e.(\hat{\theta_j}) = \sqrt{\sum_{s=1}^{S} (\theta_j^{(s)} - \hat{\theta_j})^2/(S-1)}$. To get the credible intervals, both percentile intervals and the highest posterior density intervals (HPD, Box & Tiao, 1973) of the Markov chains can be used. Percentile intervals are obtained by sorting $\theta_j^{(s)}$. HPD intervals may also be referred as minimum length confidence intervals for a Bayesian posterior distribution, and for symmetric distributions HPD intervals obtain equal tail area probabilities.

4 Simulation Studies

In this section, simulation studies are conducted to evaluate the performance of the proposed models estimated by the Bayesian method.

4.1 Simulation Design

In the simulation, we focus on linear LGCMs to simplify the presentation. Higher order LGCMs can be easily expanded by adding quadratic or higher order terms.

First, four waves of complete LGCM data $\mathbf{y}_i$ are generated based on Equations (1) and (2). The random effects consist of the intercept I_i and the slope S_i, with $Var(I_i) = 1$, $Var(S_i) = 4$, and $Cov(I_i, S_i) = 0$. The fix-effects are $(I, S) = (1, 3)$. The measurement errors are assumed to follow a normal distribution with mean 0 and standard deviation 1. In the simulation, we also assume there is one covariate X generated from a normal distribution, $X \sim N(1, sd = 0.2)$. Missing data are created based on different pre-designed missingness rates. We assume the true missingness is LSD (also noted as the XS missingness in this study because the missingness depends on the latent individual slope S and covariate X). With LSD, the bigger the slope is, the more the missing data. For the sake of simplicity in the simulation, the missingness rate is set the same for every occasion. Specifically, we set the missingness probit coefficients as $\gamma_0 = (-1, -1, -1, -1)$, $\gamma_x = (-1.5, -1.5, -1.5, -1.5)$, and $\gamma_S = (0.5, 0.5, 0.5, 0.5)$. With the setting, missingness rates are generated based on Equation (5). If a participant has a latent growth slope 3, with a covariate value 1, his or her missingness rate at each wave is $\tau \approx 16\%$; and if the slope is 5, with the same covariate value, the missing rate increases to $\tau \approx 50\%$; but when the slope is 1, the missingness rate decreases to $\tau \approx 2.3\%$.

Next, we fit data with LGCMs with different missingness. Specifically, the model design with different missingness is shown in Table 1, where the symbol "✓" shows the related factors on which the missing data rates depend. For example, when both "X" and "I" are checked, the missingness depends on the individual's latent intercept "I" and the observed covariate "X". Four types of missingness are studied: LID (also noted as XI in Table 1), LSD (XS), LOD (XY), and ignorable (X). The shaded model, LSD (XS), is the true model we used for generating the simulation data. Five levels of sample size (N=1000, N=500, N=300, N=200 and N=100) are investigated, and for each sample size. In total, 4×5=20 summary tables are combined and presented in Tables 2, 3, and 5-9 [2]. Each result table is summarized from 100 converged replications.

4.2 Simulation Implementation

The simulation studies are implemented by the following algorithm. (1) Set the counter $R = 0$. (2) Generate complete longitudinal growth data according to predefined model parameters. (3) Create missing data according to missing data mechanisms and missing data rates. (4) Generate Markov chains for model parameters through the Gibbs sampling procedure. (5) Test the convergence of generated Markov chains. (6) If the Markov chains pass the convergence test, set $R = R + 1$ and calculate and save the parameter estimates. Otherwise, set $R = R$ and discard the current replication of simulation. (7) Repeat the above process till $R = 100$ to obtain 100 replications of valid simulation.

In step 4, priors carrying little prior information are adopted (Congdon, 2003; Gill, 2002; Zhang, Hamagami, Wang, Grimm, & Nesselroade, 2007). Specifically,

[2] The summary table for the model with the latent intercept dependent (LID) missingness (XI), for N=100 is not included due to its low convergence rate.

Table 1. Simulation model design. N=1000, 500, 300, 200 and 100

Model	X[5]	I[6]	S[7]	Y[8]
Ignorable (X)	✓			
LID[2] (XI)	✓	✓		
LSD[3] (XS)[1]	✓		✓	
LOD[4] (XY)	✓			✓

Note. [1]The shaded model is the true model XS. [2]LID: Latent Intercept Dependent. [3]LSD: Latent Slope Dependent. [4]LOD: Latent Outcome Dependent. [5]X: Observed covariates. If X is the only item checked, the missingness is ignorable. [6]I: Individual latent intercept. If checked, the missingness is non-ignorable. [7]S: Individual latent slope. If checked, the missingness is non-ignorable. [8]Y: Individual potential outcome y. If checked, the missingness is non-ignorable.

for $\boldsymbol{\varphi}_1$, we set $\boldsymbol{\mu}_{\varphi_1} = \mathbf{0}_2$ and $\boldsymbol{\Sigma}_{\varphi_1} = 10^3\mathbf{I}_2$. For ϕ, we set $v_{0k} = s_{0k} = 0.002$. For $\boldsymbol{\beta}$, it is assumed that $\boldsymbol{\beta}_{k0} = \mathbf{0}_2$ and $\boldsymbol{\Sigma}_{k0} = 10^3\mathbf{I}_2$. For $\boldsymbol{\Psi}$, we define $m_{k0} = 2$ and $\mathbf{V}_{k0} = \mathbf{I}_2$. Finally, for $\boldsymbol{\gamma}_t$, we let $\boldsymbol{\gamma}_{t0} = \mathbf{0}_3$ and $\mathbf{D}_{t0} = 10^3\mathbf{I}_3$, where $\mathbf{0}_d$ and $\mathbf{I}_d$ denote a d-dimensional zero vector and a d-dimensional identity matrix, respectively. In step 5, the iteration number of burn-in period is set. The Geweke convergence criterion indicated that less than 10,000 iterations was adequate for all conditions in the study. Therefore, a conservative burn-in of 20,000 iterations was used for all iterations. And then the Markov chains with a length of $20,000$ iterations are saved for convergence testing and data analysis. After step 7, 12 summary statistics are reported based on 100 sets of converged simulation replications. For the purpose of presentation, let θ_j represent the j^{th} parameter, also the true value in the simulation. Twelve statistics are defined below. (1) The average estimate (est.$_j$) across 100 converged simulation replications of each parameter is obtained as est.$_j = \bar{\hat{\theta}}_j = \sum_{i=1}^{100} \hat{\theta}_{ij}/100$, where $\hat{\theta}_{ij}$ denotes the estimate of θ_j in the i^{th} simulation replication. (2) The simple bias (BIAS.smp$_j$) of each parameter is calculated as BIAS.smp$_j = \bar{\hat{\theta}}_j - \theta_j$. (3) The relative bias (BIAS.rel$_j$) of each parameter is calculated using BIAS.rel$_j = (\bar{\hat{\theta}}_j - \theta_j)/\theta_j$ when $\theta_j \neq 0$ and BIAS.rel$_j = \bar{\hat{\theta}}_j - \theta_j$ when $\theta_j = 0$. (4) The empirical standard error (SE.emp$_j$) of each parameter is obtained as SE.emp$_j = \sqrt{\sum_{i=1}^{100}(\hat{\theta}_{ij} - \bar{\hat{\theta}}_j)^2/99}$, and (5) the average standard error (SE.avg$_j$) of the same parameter is calculated by SE.avg$_j = \sum_{i=1}^{100} \hat{s}_{ij}/100$, where $\hat{s}_{ij}$ denotes the estimated standard error of $\hat{\theta}_{ij}$. (6) The average mean square error (MSE) of each parameter is obtained by MSE$_j = \sum_{i=1}^{100}$ MSE$_{ij}/100$, where MSE$_{ij}$ is the mean square error for the j^{th} parameter in the i^{th} simulation replication, MSE$_{ij} = (\text{Bias}_{ij})^2 + (\hat{s}_{ij})^2$. The average lower (7) and upper (8) limits of the 95% percentile confidence interval (CI.low$_j$ and CI.upper$_j$) are respectively defined as CI.low$_j = \sum_{i=1}^{100} \hat{\theta}_{ij}^l/100$ and CI.upper$_j = \sum_{i=1}^{100} \hat{\theta}_{ij}^u/100$ where $\hat{\theta}_{ij}^l$ and $\hat{\theta}_{ij}^u$ denote the 95% lower and upper limits of CI for the j^{th} parameter, respectively. (9) The coverage probability of the 95% percentile confidence interval (CI.cover$_j$) of each parameter is obtained using CI.cover$_j$ $= [\#(\hat{\theta}^l_{ij} \leq \theta_j \leq \hat{\theta}^u_{ij})]/100$. The average lower (10), upper (11) limits, and (12) the coverage probability of the 95% highest posterior density credible interval (HPD, Box & Tiao, 1973) of each parameter are similarly defined by HPD.low$_j$, HPD.upper$_j$, and HPD.cover$_j$, respectively.

4.3 Simulation Results

In this section, we show simulation results for the estimates obtained from the true model and misspecified models.

4.3.1 Estimates from the True Model. First, we investigate the estimates obtained from the true model. Tables 3, 4 and 5 in the appendix show the summarized estimates from the true model for N=1000, N=500, N=300, and N=100. From Tables 3 with the sample size 1000, first, one can see that all the relative estimate biases are very small, with the largest one being 0.067 for γ_{03}. Second, the difference between the empirical SEs and the average SEs is very small, which indicates the SEs are estimated accurately. Third, both CI and HPD interval coverage probabilities are very close to the theoretical percentage 95%, which means the type I error for each parameter is close to the specified 5% so that we can use the estimated confidence intervals to conduct statistical inference. Fourth, this true model has 100% convergence rate. When the sample sizes are smaller, the performance becomes worse as expected.

In order to compare estimates with different sample sizes, we further calculate the five summary statistics across all parameters, which are shown in Table 2. The first statistic is the average absolute relative biases (|Bias.rel|) across all parameters, which is defined as $|\text{Bias.rel}| = \sum_{j=1}^{p} |\text{Bias.rel}_j|/p$, where p is the total number parameters in a model. Second, we obtain the average absolute differences between the empirical SEs and the average Bayesian SEs (|SE.diff|) across all parameters by using $|\text{SE.diff}| = \sum_{j=1}^{p} |\text{SE.emp}_j - \text{SE.avg}_j|/p$. Third, we calculate the average percentile coverage probabilities (CI.cover) across all parameters by using $\text{CI.cover} = \sum_{j=1}^{p} \text{CI.cover}_j/p$. Fourth, we calculate the average HPD coverage probabilities (HPD.cover) across all parameters by using $\text{HPD.cover} = \sum_{j=1}^{p} \text{HPD.cover}_j/p$. Fifth, the convergence rate is calculated.

Table 2 shows that, except for the case for N=100, the true mode can recover model parameters very well, with small average absolute relative biases of estimates, |Bias.rel|, small average absolute differences between the empirical SEs and the average SEs, |SE.diff|, and almost 95% average percentile coverage probabilities (CI.cover), and the average HPD coverage probabilities (HPD.cover). With the increase of the sample size, both the point estimates and standard errors get more accurate.

4.3.2 Comparison of Different Models. We now compare the estimates obtained from the true model and different misspecified models. In this study,

Table 2. Summary and Comparison of the Results of True Model XS

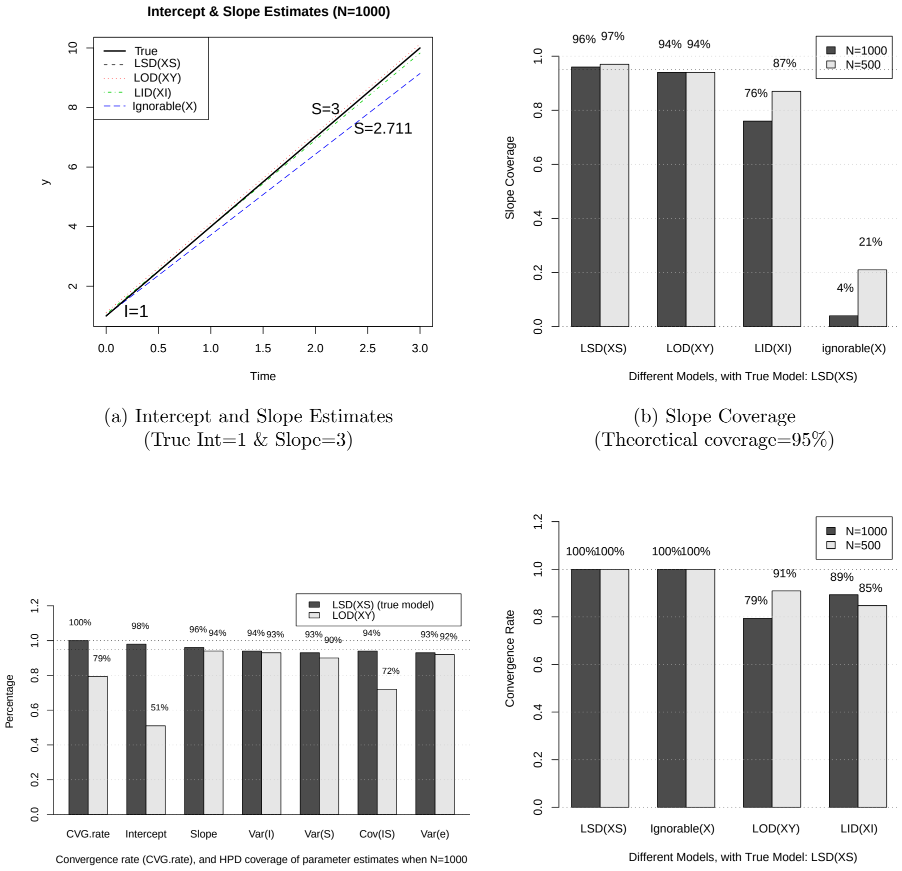
		$\lvert$Bias.rel$\rvert^1$	$\lvert$SE.diff$\rvert^2$	MSE^3	CI.cover^4	HPD.cover^5	CVG.rate^6
	1000	0.025	0.007	0.033	0.942	0.942	100%
	500	0.052	0.021	0.079	0.932	0.939	100%
N	300	0.089	0.031	0.150	0.922	0.930	100%
	200	0.160	0.090	0.366	0.909	0.924	94.34%
	100	1.202	2.664	23.743	0.869	0.893	70.42%

Note. [1]The average absolute relative bias across all parameters, defined by $|\text{Bias.rel}| = \sum_{j=1}^{p} |\text{Bias.rel}_j|/p$. The smaller, the better. [2]The average absolute difference between the empirical SEs and the average Bayesian SEs across all parameters, defined by $|\text{SE.diff}| = \sum_{j=1}^{p} |\text{SE.emp}_j - \text{SE.avg}_j|/p$. The smaller, the better. [3]The Mean Square Errors (MSE) across all parameters, defined by $MSE = \sum_{j=1}^{p}[(\text{Bias}_j)^2 + (\hat{s}_j)^2]/p$. The smaller, the better. [4]The average percentile coverage probability across all parameters, defined by $\text{CI.cover} = \sum_{j=1}^{p} \text{CI.cover}_j/p$, with a theoretical value of 0.95. [5]The average highest posterior density (HPD) coverage probability across all parameters, defined by $\text{HPD.cover} = \sum_{j=1}^{p} \text{HPD.cover}_j/p$, with a theoretical value of 0.95. [6]The convergence rate.

the true model is the LGCM with LSD (XS) missingness, and there are three mis-specified models, the LGCM with LID (XI) missingness, the LGCM with LOD (XY) missingness, and the LGCM with ignorable missingness (see Table 1 for the simulation design). Table 6 in the appendix shows the summarized estimates from the mis-specified model with LID (XI) missingness for N=1000, N=500, N=300, and N=200 (the summarized estimates for N=100 are unavailable due to a low convergence rate). Table 8 in the appendix provides the results for the mis-specified model with LOD (XY) missingness for N=1000, N=500, N=300, N=200, and N=100. Table 10 in the appendix is the summary table for the mis-specified model with ignorable (X) missingness for different sample sizes.

To compare estimates from different models, we further summarize and visualize some statistics. Figure 4 (a) compares the point estimates of intercept and slope for all models when N=1000. The true value of slope is 3 but the estimate is 2.711 when the missingness is ignored. Actually, for the model with ignorable missingness, the slope estimates are all less than 2.711 for all sample sizes in our study. Figure 4 (b) focuses on the coverage of slope. When the missingness is ignored, it is as low as 4% for N=1000, and 21% for N=500 (the coverage for N=1000 is lower because the SE for N=1000 is smaller than the SE for N=500). As a result, conclusions based on the model with ignorable missingness can be very misleading. Figure 4 (b) also shows that the slope estimate from the model with the mis-specified missingness, LID (XI), has low coverage, with 76% for N=1000 and 87% for N=500. So the conclusions based on this model may still be incorrect. Figure 4 (c) compares the true model and the model with another type of mis-specified missingness, LOD (XY) for N=1000.

For the wrong model, the coverage is 51% for intercept, and 72% for Cov(I,S). Finally, Figure 4 (d) compares the convergence rates for all models. One can see that the convergence rates of LOD (XY) and LID (XI) models are much lower than those of the true model LSD (XS) and the model with ignorable missingness. When the missingness is ignored, the number of parameters is smaller than that of non-ignorable models, and then convergence rate gets higher.



(a) Intercept and Slope Estimates
(True Int=1 & Slope=3)

(b) Slope Coverage
(Theoretical coverage=95%)

(c) Parameter Coverage for LSD(XS) and
LOD (XY) (Theoretical value=95%)

(d) Convergence Rates
(Closer to 100% is better.)

Figure 4. Comparison of four models.

In summary, the estimates from mis-specified models may result in misleading conclusions, especially when the missingness is ignored. Also, the convergence rate of a mis-specified model is usually lower than that of the true model.

4.4 Simulation Conclusions

Based on the simulation studies, we draw the following conclusions: (1) the proposed Bayesian method can accurately recover model parameters (both point estimates and standard errors), (2) the small difference between the empirical SE and the average SE indicates that the Bayesian method used in the study can estimate the standard errors accurately, (3) with the increase of the sample size, estimates get closer to their true values and standard errors become more accurate, (4) ignoring the non-ignorable missingness can lead to incorrect conclusions, (5) mis-specified missingness may also result in misleading conclusions, and (6) the non-convergence of models might be a sign of a misspecified model.

5 Discussion

The models proposed in this article have several implications for future research. First, the missingness in the simulation study is assumed to be independent across different times. If this assumption is violated, likelihood functions might be much more complicated. For example, if the missingness depends on the previous missingness, then the autocorrelation among missingness might be involved. A similar model is the Diggle and Kenward (1994)'s model, in which the probability of missing data at current wave depends directly on the current outcomes as well as on the preceding assessment. Another example is survival analysis (e.g., Klein & Moeschberger, 2003), in which censoring is the common form of missing data problem. In practice, the missingness can come from different sources and can be modeled as a combination of different types of missingness. Second, various model selection criteria could be considered (e.g., Cain & Zhang, 2019). It is an interesting topic for future work to propose new criteria. For example, observed-data and complete-data likelihood functions for random effects models can be used for $f(\mathbf{y}|\theta)$; information criterion can be proposed using other weighted combination of the growth model and the missing data model. Third, the data considered in the study are assumed to be normally distributed. However, in reality, data are seldom normally distributed, particularly in behavioral and educational sciences (e.g., Cain, Zhang, & Yuan, 2017; Micceri, 1989). When data have heavy tails, or contaminated with outliers, robust models (e.g., Hoaglin, Mosteller, & Tukey, 1983; Huber, 1996; Zhang, 2013; Zhang, Lai, Lu, & Tong, 2013) should be adopted to make models insensitive to small deviations from the assumption of normal distribution. Fourth, latent population heterogeneity (e.g., McLachlan & Peel, 2000) may exist in the collected longitudinal data. Growth mixture models (GMMs) can be considered to provide a flexible set of models for analyzing longitudinal data with latent or mixture distributions (e.g., Bartholomew & Knott, 1999).

References

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, *67*, 1012-1028. doi: https://doi.org/10.1111/j.1741-3737.2005.00191.x

Allison, P. D. (2002). *Missing data (sage university papers series on quantitative applications in the social sciences, 07-136).* Thousands Oaks, CA: sage: Thousand Oaks, CA: SAGE publications, Inc.

Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (p. 1-39). New York, NY: Academic Press.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5-37. doi: https://doi.org/10.1016/j.jsp.2009.10.001

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis: Kendall's library of statistics* (2nd ed., Vol. 7). New York, NY: Edward Arnold.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Hoboken, NJ: John Wiley & Sons.

Cain, M. K., & Zhang, Z. (2019). Fit for a bayesian: An evaluation of ppp and dic for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 39–50. doi: https://doi.org/10.1080/10705511.2018.1490648

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior research methods*, *49*(5), 1716–1735. doi: https://doi.org/10.3758/s13428-016-0814-1

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167-174. doi: https://doi.org/10.1080/00031305.1992.10475878

Collins, L. (1991). Measurement in longitudinal research. In J. L. Horn & L. Collins (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 137–148). Washington, DC: American Psychological Association.

Congdon, P. (2003). *Applied Bayesian modelling.* Chichester, UK: John Wiley & Sons.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*, 1-31. doi: https://doi.org/10.1111/j.2517-6161.1979.tb01052.x

Demidenko, E. (2004). *Mixed models: Theory and applications.* New York, NY: Wiley-Interscience.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *43*, 49-93. doi: https://doi.org/10.2307/2986113

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, B*, *62*, 355-366. doi: https://doi.org/10.1111/1467-9868.00236

Enders, C. K. (2011a). Analyzing longitudinal data with missing values. *Rehabilitation Psychology*, *56*, 1-22. doi: https://doi.org/10.1037/a0025579

Enders, C. K. (2011b). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*, 1-16. doi: https://doi.org/10.1037/a0022640

Fay, R. E. (1992). When are inferences from multiple imputation valid. In (Vol. 81, p. 227-232).

Fitzmaurice, G. M., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.). (2008). *Longitudinal data analysis.* Boca Raton, FL: Chapman & Hall/CRC Press.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis.* Hoboken, New Jersey: John Wiley & sons Inc.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721-741. doi: https://doi.org/10.1016/b978-0-08-051581-6.50057-x

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach.* Boca Raton, FL: Chapman & Hall/CRC.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Drawing inferences from self-selected samples. In H. Wainer (Ed.), (p. 115-142). New York: Springer Verlag.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*, 64-78. doi: https://doi.org/10.1037/1082-989x.2.1.64

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis.* (p. 285-299). NY, US: Russell Sage Foundation.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis.* New York: John Wiley & Sons.

Huber, P. (1996). *Robust statistical procedures* (2nd ed.). Philadelphia: SIAM.

Jelicic, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, *45*, 1195-1199. doi: https://doi.org/10.1037/a0015665

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data* (2nd, Ed.). Springer.

Lee, S. Y. (2007). *Structural equation modeling: A Bayesian approach.* Chichester, UK: John Wiley & Sons.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*, 125-134. doi: https://doi.org/10.2307/2290705

Little, R. J. A. (1995). Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112-1121. doi: https://doi.org/10.1080/01621459.1995.10476615

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York, N.Y.: Wiley.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, N.Y.: Wiley-Interscience.

Lu, Z., Zhang, Z., & Lubke, G. (2010). (Abstract) Bayesian inference for growth mixture models with non-ignorable missing data. *Multivariate Behavioral Research*, *45*, 1028-1029. doi: https://doi.org/10.1080/00273171.2010.534381

Lu, Z., Zhang, Z., & Lubke, G. (2011). Bayesian inference for growth mixture models with latent-class-dependent missing data. *Multivariate Behavioral Research*, *46*, 567-597. doi: https://doi.org/10.1080/00273171.2011.589261

Luke, D. A. (2004). *Multilevel modeling (quantitative applications in the social sciences).* Thousand Oaks, CA: Sage Publication, Inc.

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

Micceri, T. (1989). The unicorn, the normal curve and the other improbable creatures. *Psychological Bulletin*, *105*, 156-166. doi: https://doi.org/10.1037/0033-2909.105.1.156

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological methods*, *17*(3), 313–335. doi: https://doi.org/10.1037/a0026802

Muthén, B., Asparouhov, T., Hunter, A. M., & Leuchter, A. F. (2011). Growth modeling with nonignorable dropout: Alternative analyses of the STAR*D antidepressant trial. *Psychological Methods*, *16*, 17-33. doi: https://doi.org/10.1037/a0022634

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, *6*, 328-362. doi: https://doi.org/10.1177/1094428103254673

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods.* New York, NY: Springer Science+Business Media Inc.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, *47*, 537-560. doi: https://doi.org/10.1111/j.1744-6570.1994.tb01736.x

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* Wiley & Sons, New York.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* Boca Raton, FL: Chapman & Hall/CRC.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147-177. doi: https://doi.org/10.1037/1082-989x.7.2.147

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*, 37-52. doi: https://doi.org/10.1007/bf02294318

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, *57*, 1-10. doi: https://doi.org/10.1037/a0018082

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press, Inc.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528-540. doi: https://doi.org/10.1080/01621459.1987.10478458

Yuan, K.-H., & Lu, Z. (2008). SEM with missing data and unknown population using two-stage ML: theory and its application. *Multivariate Behavioral Research*, *43*, 621-652. doi: https://doi.org/10.1080/00273170802490699

Yuan, K.-H., & Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika*, *77(4)*, 803-826. doi: https://doi.org/10.1007/s11336-012-9282-4

Zhang, Z. (2013). Bayesian growth curve models with the generalized error distribution. *Journal of Applied Statistics*, *40*(8), 1779–1795. doi: https://doi.org/10.1080/02664763.2013.796348

Zhang, Z., Hamagami, F., Wang, L., Grimm, K. J., & Nesselroade, J. R. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development*, *31*(4), 374-383. doi: https://doi.org/10.1177/0165025407077764

Zhang, Z., Lai, K., Lu, Z., & Tong, X. (2013). Bayesian inference and application of robust growth curve models using student's t distribution. *Structural Equation Modeling*, *20(1)*, 47-78. doi: https://doi.org/10.1080/10705511.2013.742382

Zhang, Z., & Wang, L. (2012). A note on the robustness of a full bayesian method for non-ignorable missing data analysis. *Brazilian Journal of Probability and Statistics*, *26(3)*, 244-264. doi: https://doi.org/10.1214/10-bjps132

Appendix

Appendix A. The Derived Posteriors for LGCMs with Non-ignorable Missingness

Let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_N)$, and the conditional posterior distribution for ϕ can be easily derived as an Inverse Gamma distribution,

$$\phi|\boldsymbol{\eta}, \mathbf{y} \sim IG\left(a_1/2, b_1/2\right), \tag{8}$$

where $a_1 = v_0 + N\,T$, and $b_1 = s_0 + \sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{\Lambda}\boldsymbol{\eta}_i)'(\mathbf{y}_i - \mathbf{\Lambda}\boldsymbol{\eta}_i)$.

Notice that $tr(\mathbf{AB}) = tr(\mathbf{BA})$, so the conditional posterior distribution for $\mathbf{\Psi}$ is derived as an Inverse Wishart distribution,

$$\mathbf{\Psi}|\boldsymbol{\beta}, \boldsymbol{\eta} \sim IW\left(m_1, \mathbf{V}_1\right), \tag{9}$$

where $m_1 = m_0 + N$, and $\mathbf{V}_1 = \mathbf{V}_0 + \sum_{i=1}^{N}(\boldsymbol{\eta}_i - \boldsymbol{\beta})(\boldsymbol{\eta}_i - \boldsymbol{\beta})'$.

By expanding the terms inside the exponential part and combining similar terms, the conditional posterior distribution for $\boldsymbol{\beta}$ is derived as a multivariate normal distribution,

$$\boldsymbol{\beta}|\mathbf{\Psi}, \boldsymbol{\eta} \sim MN(\boldsymbol{\beta}_1, \mathbf{\Sigma}_1), \tag{10}$$

where $\boldsymbol{\beta}_1 = \left(N\,\mathbf{\Psi}^{-1} + \mathbf{\Sigma}_0^{-1}\right)^{-1}\left(\mathbf{\Psi}^{-1}\sum_{i=1}^{N}\boldsymbol{\eta}_i + \mathbf{\Sigma}_0^{-1}\boldsymbol{\beta}_0\right)$, and $\mathbf{\Sigma}_1 = \left(N\,\mathbf{\Psi}^{-1} + \mathbf{\Sigma}_0^{-1}\right)^{-1}$.

The conditional posterior for $\boldsymbol{\gamma}_t$, $(t = 1, 2, \ldots, T)$, is a distribution of

$$\begin{aligned} f(\boldsymbol{\gamma}_t|\boldsymbol{\omega}, \mathbf{x}, \mathbf{m}) \propto \exp\Bigg[&- \frac{1}{2}(\boldsymbol{\gamma}_t - \boldsymbol{\gamma}_{t0})'\mathbf{D}_{t0}^{-1}(\boldsymbol{\gamma}_t - \boldsymbol{\gamma}_{t0}) \\ &+ \sum_{i=1}^{N}\left\{m_{it}\log\Phi(\boldsymbol{\omega}_i'\boldsymbol{\gamma}_t) + (1 - m_{it})\log[1 - \Phi(\boldsymbol{\omega}_i'\boldsymbol{\gamma}_t)]\right\}\Bigg]. \end{aligned} \tag{11}$$

where $\Phi(\boldsymbol{\omega}_i'\boldsymbol{\gamma}_t)$ is defined by Equation (4), (5), or (6).

By expanding the terms inside the exponential part and combining similar terms, the conditional posterior distribution for $\boldsymbol{\eta}_i$, $i = 1, 2, \ldots, N$, is derived as a Multivariate Normal distribution,

$$\boldsymbol{\eta}_i|\phi, \mathbf{\Psi}, \boldsymbol{\beta}, \mathbf{y}_i \sim MN(\boldsymbol{\mu}_{\eta i}, \mathbf{\Sigma}_{\eta i}), \tag{12}$$

where $\boldsymbol{\mu}_{\eta i} = \left(\frac{1}{\phi}\mathbf{\Lambda}'\mathbf{\Lambda} + \mathbf{\Psi}^{-1}\right)^{-1}\left(\frac{1}{\phi}\mathbf{\Lambda}'\mathbf{y}_i + \mathbf{\Psi}^{-1}\boldsymbol{\beta}\right)$, and $\mathbf{\Sigma}_{\eta i} = \left(\frac{1}{\phi}\mathbf{\Lambda}'\mathbf{\Lambda} + \mathbf{\Psi}^{-1}\right)^{-1}$.

The conditional posterior distribution for the missing data $\mathbf{y}_i^{mis}$, $i = 1, 2, \ldots, N$, is a normal distribution,

$$\mathbf{y}_i^{mis}|\boldsymbol{\eta}_i, \phi \sim MN\left[\mathbf{\Lambda}\boldsymbol{\eta}_i, \mathbf{I}\phi\right], \tag{13}$$

where $\mathbf{I}$ is a $T \times T$ identity matrix. The dimension and location of $\mathbf{y}_i^{mis}$ depend on the corresponding $\mathbf{m}_i$ value.

Appendix B. Simulation Results

Table 3. Summarized Estimates from True Model: LGCM with LSD Missingness (XS). N=1000 (convergence rate: 100/100 = 100%)

					BIAS		SE			CI[9]			HPD[13]		
		para.[1]	true[2]	est.[3]	smp.[4]	rel.[5]	emp.[6]	avg.[7]	MSE[8]	lower[10]	upper[11]	cover[12]	lower	upper	cover
Growth Curve		I	1	0.998	-0.002	-0.002	0.05	0.053	0.005	0.894	1.101	0.99	0.894	1.101	0.98
		S	3	3.003	0.003	0.001	0.079	0.077	0.012	2.853	3.155	0.97	2.853	3.154	0.96
		var(I)	1	1.011	0.011	0.011	0.105	0.102	0.022	0.82	1.22	0.94	0.814	1.213	0.94
		var(S)	4	3.99	-0.01	-0.003	0.232	0.232	0.107	3.56	4.468	0.94	3.545	4.449	0.93
		cov(IS)	0	0.001	0.001	0.001	0.119	0.112	0.026	-0.221	0.217	0.94	-0.218	0.218	0.94
		var(e)	1	1	0	0	0.043	0.042	0.004	0.92	1.086	0.92	0.918	1.084	0.93
Missingness Parameters	Wave 1	γ_{01}	-1	-1.025	-0.025	0.025	0.184	0.174	0.065	-1.375	-0.694	0.93	-1.365	-0.69	0.94
		γ_{x1}	-1.5	-1.541	-0.041	0.027	0.138	0.123	0.036	-1.795	-1.314	0.92	-1.783	-1.307	0.93
		γ_{S1}	0.5	0.515	0.015	0.03	0.066	0.062	0.008	0.4	0.641	0.9	0.397	0.636	0.92
	Wave 2	γ_{02}	-1	-1.038	-0.038	0.038	0.191	0.171	0.067	-1.385	-0.714	0.96	-1.376	-0.711	0.97
		γ_{x2}	-1.5	-1.551	-0.051	0.034	0.129	0.119	0.034	-1.798	-1.33	0.95	-1.786	-1.323	0.94
		γ_{S2}	0.5	0.521	0.021	0.042	0.066	0.06	0.008	0.41	0.643	0.95	0.408	0.639	0.94
	Wave 3	γ_{03}	-1	-1.067	-0.067	0.067	0.186	0.172	0.069	-1.417	-0.741	0.94	-1.407	-0.737	0.94
		γ_{x3}	-1.5	-1.557	-0.057	0.038	0.117	0.116	0.03	-1.796	-1.341	0.97	-1.785	-1.334	0.97
		γ_{S3}	0.5	0.529	0.029	0.058	0.063	0.058	0.008	0.42	0.648	0.89	0.418	0.643	0.91
	Wave 4	γ_{04}	-1	-1.034	-0.034	0.034	0.18	0.173	0.063	-1.384	-0.709	0.94	-1.374	-0.704	0.93
		γ_{x4}	-1.5	-1.539	-0.039	0.026	0.122	0.114	0.029	-1.773	-1.325	0.95	-1.763	-1.319	0.94
		γ_{S4}	0.5	0.514	0.014	0.027	0.058	0.057	0.007	0.407	0.63	0.95	0.405	0.625	0.95

Note. The results are summarized based on 100 converged replications with a convergence rate of 100/100 = 100%. [1]The estimated parameter. [2]The true value of the corresponding parameter. [3]The parameter estimate, defined by est.$_j = \bar{\hat{\theta}}_j = \sum_{i=1}^{100} \hat{\theta}_{ij}/100$. [4]The simple bias, defined by BIAS.smp$_j = \bar{\hat{\theta}}_j - \theta_j$. [5]The relative bias, defined by BIAS.rel$_j = (\bar{\hat{\theta}}_j - \theta_j)/\theta_j$ when $\theta_j \neq 0$ and BIAS.rel$_j = \bar{\hat{\theta}}_j - \theta_j$ when $\theta_j = 0$. [6]The empirical standard errors, defined by SE.emp$_j = \sqrt{\sum_{i=1}^{100}(\hat{\theta}_{ij} - \bar{\hat{\theta}}_j)^2/99}$. [7]The average standard errors, defined by SE.avg$_j = \sum_{i=1}^{100} \hat{s}_{ij}/100$. [8]The mean square error, defined by MSE$_j = \sum_{i=1}^{100}$ MSE$_{ij}/100$, where MSE$_{ij} = (\text{Bias}_{ij})^2 + (\hat{s}_{ij})^2$. [9]For percentile confidence interval. [10]The average lower 2.5% percentile. [11]The average upper 97.5% percentile. [12]The average 95% coverage of percentile confidence interval. [13]The lower,upper bounds, and coverage for HPD interval.

Table 4. Summarized Estimates from True Model: LGCM with LSD Missingness (XS) (con't)

					BIAS		SE			CI[9]			HPD[13]		
		para.[1]	true[2]	est.[3]	smp.[4]	rel.[5]	emp.[6]	avg.[7]	MSE[8]	lower[10]	upper[11]	cover[12]	lower	upper	cover
N=500 (convergence rate: 100/100 = 100%)															
Growth Curve		I	1	0.986	-0.014	-0.014	0.076	0.074	0.011	0.841	1.132	0.93	0.841	1.132	0.95
		S	3	3.001	0.001	0	0.097	0.109	0.021	2.789	3.216	0.97	2.788	3.213	0.97
		var(I)	1	0.976	-0.024	-0.024	0.146	0.144	0.042	0.712	1.274	0.97	0.7	1.26	0.97
		var(S)	4	4.001	0.001	0	0.388	0.329	0.258	3.403	4.691	0.9	3.373	4.652	0.9
		cov(IS)	0	-0.009	-0.009	-0.009	0.155	0.157	0.049	-0.324	0.294	0.96	-0.319	0.297	0.96
		var(e)	1	1.014	0.014	0.014	0.06	0.061	0.007	0.901	1.141	0.96	0.897	1.136	0.96
Missingness Parameters	Wave 1	γ_{01}	-1	-1.082	-0.082	0.082	0.254	0.255	0.137	-1.609	-0.608	0.95	-1.587	-0.596	0.97
		γ_{x1}	-1.5	-1.606	-0.106	0.071	0.181	0.186	0.079	-2.002	-1.275	0.95	-1.975	-1.258	0.97
		γ_{S1}	0.5	0.54	0.04	0.081	0.083	0.092	0.017	0.375	0.735	0.95	0.368	0.722	0.94
	Wave 2	γ_{02}	-1	-1.096	-0.096	0.096	0.281	0.252	0.152	-1.61	-0.624	0.89	-1.591	-0.615	0.89
		γ_{x2}	-1.5	-1.615	-0.115	0.077	0.204	0.18	0.088	-1.996	-1.291	0.91	-1.971	-1.275	0.94
		γ_{S2}	0.5	0.546	0.046	0.092	0.104	0.088	0.021	0.385	0.73	0.87	0.379	0.719	0.88
	Wave 3	γ_{03}	-1	-1.068	-0.068	0.068	0.32	0.248	0.169	-1.572	-0.602	0.93	-1.555	-0.594	0.93
		γ_{x3}	-1.5	-1.613	-0.113	0.075	0.279	0.174	0.123	-1.978	-1.295	0.9	-1.958	-1.283	0.93
		γ_{S3}	0.5	0.536	0.036	0.072	0.116	0.084	0.022	0.381	0.71	0.92	0.378	0.702	0.91
	Wave 4	γ_{04}	-1	-1.123	-0.123	0.123	0.261	0.257	0.15	-1.652	-0.647	0.94	-1.628	-0.633	0.95
		γ_{x4}	-1.5	-1.579	-0.079	0.053	0.174	0.168	0.066	-1.933	-1.274	0.95	-1.913	-1.261	0.96
		γ_{S4}	0.5	0.543	0.043	0.086	0.089	0.085	0.017	0.388	0.719	0.92	0.382	0.71	0.92
N=300 (convergence rate: 100/100 = 100%)															
Growth Curve		I	1	1.001	0.001	0.001	0.104	0.097	0.02	0.81	1.192	0.89	0.811	1.192	0.89
		S	3	2.984	-0.016	-0.005	0.149	0.14	0.042	2.712	3.262	0.93	2.71	3.259	0.93
		var(I)	1	1.014	0.014	0.014	0.183	0.19	0.07	0.673	1.418	0.96	0.654	1.392	0.96
		var(S)	4	3.975	-0.025	-0.006	0.416	0.425	0.354	3.22	4.886	0.96	3.174	4.82	0.96
		cov(IS)	0	0.054	0.054	0.054	0.212	0.205	0.09	-0.359	0.449	0.94	-0.351	0.454	0.93
		var(e)	1	1.011	0.011	0.011	0.073	0.08	0.012	0.867	1.179	0.96	0.86	1.17	0.96
Missingness Parameters	Wave 1	γ_{01}	-1	-1.094	-0.094	0.094	0.341	0.345	0.249	-1.822	-0.468	0.97	-1.778	-0.441	0.97
		γ_{x1}	-1.5	-1.65	-0.15	0.1	0.265	0.253	0.162	-2.209	-1.217	0.92	-2.155	-1.185	0.94
		γ_{S1}	0.5	0.548	0.048	0.097	0.121	0.124	0.033	0.331	0.82	0.97	0.318	0.794	0.97
	Wave 2	γ_{02}	-1	-1.106	-0.106	0.106	0.452	0.34	0.341	-1.819	-0.486	0.93	-1.782	-0.467	0.93
		γ_{x2}	-1.5	-1.692	-0.192	0.128	0.345	0.253	0.23	-2.243	-1.254	0.89	-2.196	-1.227	0.9
		γ_{S2}	0.5	0.566	0.066	0.132	0.158	0.121	0.046	0.354	0.827	0.93	0.343	0.807	0.92
	Wave 3	γ_{03}	-1	-1.139	-0.139	0.139	0.397	0.335	0.293	-1.845	-0.527	0.91	-1.801	-0.503	0.92
		γ_{x3}	-1.5	-1.648	-0.148	0.099	0.305	0.236	0.175	-2.152	-1.233	0.86	-2.115	-1.21	0.92
		γ_{S3}	0.5	0.566	0.066	0.132	0.141	0.115	0.038	0.361	0.811	0.9	0.352	0.794	0.91
	Wave 4	γ_{04}	-1	-1.217	-0.217	0.217	0.411	0.356	0.347	-1.976	-0.576	0.9	-1.932	-0.552	0.9
		γ_{x4}	-1.5	-1.681	-0.181	0.121	0.263	0.241	0.163	-2.203	-1.257	0.9	-2.161	-1.231	0.92
		γ_{S4}	0.5	0.583	0.083	0.165	0.138	0.118	0.041	0.372	0.839	0.88	0.363	0.82	0.91

Note. The same as Table 3

Table 5. Summarized Estimates from True Model: LGCM with LSD Missingness (XS) (con't)

					BIAS		SE			CI			HPD		
		para.	true	est.	smp.	rel.	emp.	avg.	MSE	lower	upper	cover	lower	upper	cover
							N=200 (convergence rate: 100/106 ≈ 94.34%)								
Growth Curve		I	1	1.011	0.011	0.011	0.099	0.119	0.024	0.779	1.244	0.98	0.779	1.243	0.98
		S	3	2.975	-0.025	-0.008	0.177	0.171	0.061	2.643	3.314	0.93	2.642	3.312	0.94
		var(I)	1	1.011	0.011	0.011	0.228	0.233	0.107	0.601	1.516	0.94	0.572	1.476	0.92
		var(S)	4	4	0	0	0.474	0.522	0.498	3.095	5.135	0.97	3.029	5.041	0.96
		cov(IS)	0	0.065	0.065	0.065	0.257	0.252	0.134	-0.447	0.549	0.92	-0.436	0.557	0.92
		var(e)	1	1.027	0.027	0.027	0.098	0.099	0.02	0.851	1.238	0.95	0.84	1.224	0.95
Missingness Parameters	Wave 1	γ_{01}	-1	-1.3	-0.3	0.3	0.671	0.5	0.901	-2.399	-0.449	0.93	-2.306	-0.402	0.94
		γ_{x1}	-1.5	-1.874	-0.374	0.249	0.745	0.424	1.113	-2.868	-1.227	0.88	-2.735	-1.169	0.91
		γ_{S1}	0.5	0.647	0.147	0.293	0.323	0.197	0.202	0.334	1.1	0.91	0.311	1.045	0.92
	Wave 2	γ_{02}	-1	-1.278	-0.278	0.278	0.69	0.468	0.838	-2.303	-0.463	0.87	-2.227	-0.426	0.89
		γ_{x2}	-1.5	-1.779	-0.279	0.186	0.456	0.349	0.451	-2.578	-1.209	0.91	-2.487	-1.163	0.9
		γ_{S2}	0.5	0.627	0.127	0.254	0.244	0.171	0.117	0.343	1.014	0.9	0.324	0.976	0.91
	Wave 3	γ_{03}	-1	-1.191	-0.191	0.191	0.505	0.436	0.5	-2.133	-0.419	0.91	-2.05	-0.377	0.93
		γ_{x3}	-1.5	-1.721	-0.221	0.147	0.502	0.314	0.426	-2.428	-1.193	0.9	-2.348	-1.15	0.94
		γ_{S3}	0.5	0.586	0.086	0.172	0.183	0.152	0.068	0.326	0.926	0.91	0.309	0.889	0.95
	Wave 4	γ_{04}	-1	-1.27	-0.27	0.27	0.594	0.467	0.67	-2.304	-0.457	0.86	-2.209	-0.404	0.9
		γ_{x4}	-1.5	-1.808	-0.308	0.205	0.397	0.336	0.382	-2.56	-1.24	0.82	-2.48	-1.195	0.89
		γ_{S4}	0.5	0.618	0.118	0.236	0.204	0.16	0.085	0.345	0.98	0.88	0.325	0.942	0.89
							N=100 (convergence rate: 100/142 ≈ 70.42%)								
Growth Curve		I	1	1.031	0.031	0.031	0.167	0.168	0.057	0.701	1.359	0.96	0.701	1.359	0.97
		S	3	2.983	-0.017	-0.006	0.236	0.242	0.115	2.514	3.467	0.95	2.51	3.46	0.94
		var(I)	1	0.933	-0.067	-0.067	0.305	0.323	0.206	0.408	1.665	0.93	0.355	1.574	0.91
		var(S)	4	3.965	-0.035	-0.009	0.829	0.747	1.261	2.743	5.656	0.91	2.623	5.458	0.91
		cov(IS)	0	0.069	0.069	0.069	0.333	0.357	0.246	-0.666	0.748	0.93	-0.646	0.762	0.95
		var(e)	1	1.078	0.078	0.078	0.157	0.151	0.054	0.82	1.409	0.93	0.801	1.38	0.94
Missingness Parameters	Wave 1	γ_{01}	-1	-3.257	-2.257	2.257	5.794	1.333	42.792	-6.264	-1.131	0.84	-5.922	-1.018	0.86
		γ_{x1}	-1.5	-4.314	-2.814	1.876	7.492	1.277	69.337	-7.171	-2.396	0.8	-6.739	-2.251	0.85
		γ_{S1}	0.5	1.626	1.126	2.252	2.881	0.55	10.353	0.788	2.857	0.8	0.746	2.698	0.84
	Wave 2	γ_{02}	-1	-3.011	-2.011	2.011	5.719	1.322	41.711	-6.062	-1.027	0.85	-5.696	-0.893	0.88
		γ_{x2}	-1.5	-3.772	-2.272	1.515	6.947	1.283	61.237	-6.811	-1.927	0.82	-6.385	-1.774	0.85
		γ_{S2}	0.5	1.436	0.936	1.871	2.57	0.549	8.564	0.653	2.71	0.81	0.586	2.527	0.86
	Wave 3	γ_{03}	-1	-2.877	-1.877	1.877	5.93	1.2	42.401	-5.493	-0.898	0.89	-5.233	-0.806	0.91
		γ_{x3}	-1.5	-3.86	-2.36	1.573	6.955	1.153	58.835	-6.508	-2.086	0.83	-6.125	-1.932	0.85
		γ_{S3}	0.5	1.388	0.888	1.776	2.567	0.467	7.977	0.641	2.428	0.85	0.596	2.289	0.89
	Wave 4	γ_{04}	-1	-2.831	-1.831	1.831	5.646	1.297	39.835	-5.902	-0.891	0.89	-5.522	-0.753	0.90
		γ_{x4}	-1.5	-3.386	-1.886	1.257	5.379	1.127	37.532	-6.048	-1.745	0.81	-5.622	-1.586	0.88
		γ_{S4}	0.5	1.222	0.722	1.444	1.944	0.457	4.854	0.552	2.312	0.84	0.491	2.152	0.88

Note. Abbreviations are as given in Table 3.

Table 6. Summarized Estimates from LGCM with LID Missingness (XI)

					BIAS		SE			CI			HPD		
		para.	true	est.	smp.	rel.	emp.	avg.	MSE	lower	upper	cover	lower	upper	cover
		N=1000 (convergence rate: 100/112 ≈ 89.29%)													
Growth Curve		I	1	1.064	0.064	0.064	0.052	0.044	0.009	0.977	1.151	0.66	0.977	1.150	0.66
		S	3	2.921	-0.079	-0.026	0.082	0.074	0.018	2.776	3.067	0.77	2.777	3.066	0.76
		var(I)	1	0.169	-0.831	-0.831	0.036	0.031	0.693	0.117	0.237	0	0.113	0.230	0
		var(S)	4	3.494	-0.506	-0.126	0.218	0.203	0.344	3.116	3.913	0.40	3.103	3.897	0.37
		cov(IS)	0	0.629	0.629	0.629	0.064	0.064	0.404	0.511	0.762	0	0.507	0.756	0
		var(e)	1	1.439	0.439	0.439	0.049	0.050	0.197	1.343	1.540	0	1.341	1.538	0
Missingness Parameters	Wave 1	γ_{01}	NA	-2.411	NA	NA	0.476	0.442	NA	-3.325	-1.619	NA	-3.27	-1.612	NA
		γ_{x1}	NA	-1.632	NA	NA	0.178	0.152	NA	-1.961	-1.362	NA	-1.939	-1.35	NA
		γ_{I1}	NA	2.794	NA	NA	0.479	0.442	NA	2.011	3.72	NA	2.009	3.661	NA
	Wave 2	γ_{02}	NA	-2.439	NA	NA	0.543	0.444	NA	-3.395	-1.667	NA	-3.321	-1.646	NA
		γ_{x2}	NA	-1.644	NA	NA	0.163	0.148	NA	-1.962	-1.382	NA	-1.938	-1.368	NA
		γ_{I2}	NA	2.826	NA	NA	0.546	0.437	NA	2.074	3.762	NA	2.045	3.682	NA
	Wave 3	γ_{03}	NA	-2.442	NA	NA	0.492	0.426	NA	-3.336	-1.678	NA	-3.277	-1.662	NA
		γ_{x3}	NA	-1.632	NA	NA	0.137	0.143	NA	-1.934	-1.375	NA	-1.915	-1.364	NA
		γ_{I3}	NA	2.819	NA	NA	0.482	0.422	NA	2.063	3.718	NA	2.048	3.646	NA
	Wave 4	γ_{04}	NA	-2.367	NA	NA	0.48	0.427	NA	-3.262	-1.596	NA	-3.199	-1.581	NA
		γ_{x4}	NA	-1.617	NA	NA	0.146	0.141	NA	-1.917	-1.362	NA	-1.896	-1.35	NA
		γ_{I4}	NA	2.733	NA	NA	0.448	0.422	NA	1.98	3.616	NA	1.978	3.569	NA
		N=500 (convergence rate: 100/118 ≈ 84.75%)													
Growth Curve		I	1	1.060	0.060	0.060	0.076	0.063	0.013	0.938	1.186	0.78	0.937	1.184	0.79
		S	3	2.914	-0.086	-0.029	0.099	0.105	0.028	2.710	3.120	0.88	2.709	3.118	0.87
		var(I)	1	0.197	-0.803	-0.803	0.046	0.048	0.650	0.121	0.309	0	0.114	0.294	0
		var(S)	4	3.448	-0.552	-0.138	0.315	0.284	0.484	2.934	4.043	0.56	2.909	4.012	0.54
		cov(IS)	0	0.633	0.633	0.633	0.074	0.088	0.414	0.474	0.819	0	0.466	0.808	0
		var(e)	1	1.425	0.425	0.425	0.079	0.072	0.192	1.289	1.571	0	1.286	1.567	0
Missingness Parameters	Wave 1	γ_{01}	NA	-2.471	NA	NA	0.726	0.665	NA	-3.903	-1.333	NA	-3.789	-1.292	NA
		γ_{x1}	NA	-1.765	NA	NA	0.298	0.269	NA	-2.385	-1.329	NA	-2.31	-1.294	NA
		γ_{I1}	NA	2.898	NA	NA	0.716	0.671	NA	1.762	4.339	NA	1.718	4.215	NA
	Wave 2	γ_{02}	NA	-2.393	NA	NA	0.77	0.631	NA	-3.746	-1.328	NA	-3.63	-1.281	NA
		γ_{x2}	NA	-1.723	NA	NA	0.265	0.239	NA	-2.257	-1.325	NA	-2.206	-1.297	NA
		γ_{I2}	NA	2.815	NA	NA	0.737	0.627	NA	1.759	4.162	NA	1.712	4.048	NA
	Wave 3	γ_{03}	NA	-2.425	NA	NA	0.779	0.644	NA	-3.804	-1.337	NA	-3.681	-1.293	NA
		γ_{x3}	NA	-1.761	NA	NA	0.352	0.257	NA	-2.336	-1.343	NA	-2.271	-1.309	NA
		γ_{I3}	NA	2.858	NA	NA	0.796	0.647	NA	1.775	4.235	NA	1.729	4.104	NA
	Wave 4	γ_{04}	NA	-2.396	NA	NA	0.782	0.655	NA	-3.86	-1.312	NA	-3.693	-1.24	NA
		γ_{x4}	NA	-1.687	NA	NA	0.294	0.24	NA	-2.223	-1.288	NA	-2.167	-1.259	NA
		γ_{I4}	NA	2.805	NA	NA	0.818	0.662	NA	1.713	4.275	NA	1.657	4.119	NA

Note. Abbreviations are as given in Table 3.

Table 7. Summarized Estimates from LGCM with LID Missingness (XI) (con't)

					BIAS		SE			CI			HPD		
		para.	true	est.	smp.	rel.	emp.	avg.	MSE	lower	upper	cover	lower	upper	cover
N=300 (convergence rate: 100/148 ≈ 67.57%)															
Growth Curve		I	1	1.077	0.077	0.077	0.11	0.083	0.025	0.916	1.242	0.78	0.915	1.24	0.81
		S	3	2.864	-0.136	-0.045	0.139	0.135	0.056	2.601	3.131	0.87	2.6	3.129	0.87
		var(I)	1	0.251	-0.749	-0.749	0.084	0.076	0.574	0.136	0.429	0.01	0.123	0.402	0.01
		var(S)	4	3.424	-0.576	-0.144	0.369	0.366	0.601	2.775	4.209	0.71	2.734	4.153	0.67
		cov(IS)	0	0.656	0.656	0.656	0.118	0.119	0.458	0.445	0.909	0	0.433	0.892	0
		var(e)	1	1.413	0.413	0.413	0.101	0.095	0.19	1.237	1.608	0	1.232	1.601	0
Missingness Parameters	Wave 1	γ_{01}	NA	-2.672	NA	NA	1.64	0.984	NA	-4.884	-1.166	NA	-4.637	-1.069	NA
		γ_{x1}	NA	-2.055	NA	NA	0.913	0.487	NA	-3.218	-1.343	NA	-3.058	-1.277	NA
		γ_{S1}	NA	3.108	NA	NA	1.68	1.008	NA	1.582	5.378	NA	1.502	5.129	NA
	Wave 2	γ_{02}	NA	-2.768	NA	NA	3.488	0.966	NA	-4.978	-1.265	NA	-4.741	-1.185	NA
		γ_{x2}	NA	-2.243	NA	NA	2.502	0.507	NA	-3.445	-1.474	NA	-3.282	-1.402	NA
		γ_{S2}	NA	3.409	NA	NA	4.711	0.994	NA	1.895	5.668	NA	1.809	5.422	NA
	Wave 3	γ_{03}	NA	-2.68	NA	NA	2.057	0.915	NA	-4.769	-1.249	NA	-4.567	-1.176	NA
		γ_{x3}	NA	-1.999	NA	NA	0.878	0.421	NA	-2.989	-1.348	NA	-2.861	-1.289	NA
		γ_{S3}	NA	3.118	NA	NA	1.884	0.936	NA	1.66	5.234	NA	1.59	5.008	NA
	Wave 4	γ_{04}	NA	-2.907	NA	NA	2.499	0.941	NA	-4.948	-1.426	NA	-4.744	-1.353	NA
		γ_{x4}	NA	-2.204	NA	NA	1.651	0.498	NA	-3.333	-1.449	NA	-3.196	-1.39	NA
		γ_{S4}	NA	3.371	NA	NA	2.766	0.98	NA	1.875	5.511	NA	1.804	5.296	NA
N=200 (convergence rate: 100/197 ≈ 50.76%)															
Growth Curve		I	1	1.052	0.082	0.082	0.219	0.1	0.03	0.858	1.248	0.79	0.857	1.247	0.79
		S	3	2.796	-0.114	-0.038	0.525	0.161	0.071	2.484	3.114	0.85	2.483	3.112	0.85
		var(I)	1	0.322	-0.648	-0.648	0.15	0.115	0.469	0.15	0.593	0.1	0.13	0.549	0.07
		var(S)	4	3.353	-0.527	-0.132	0.739	0.435	0.677	2.6	4.302	0.74	2.546	4.225	0.71
		cov(IS)	0	0.617	0.617	0.617	0.276	0.147	0.479	0.352	0.93	0.01	0.338	0.91	0.01
		var(e)	1	1.346	0.376	0.376	0.267	0.115	0.174	1.135	1.586	0.07	1.126	1.574	0.08
Missingness Parameters	Wave 1	γ_{01}	NA	-2.974	NA	NA	5.157	1.352	NA	-6.046	-0.844	NA	-5.701	-0.814	NA
		γ_{x1}	NA	-3.03	NA	NA	3.647	0.986	NA	-5.376	-1.659	NA	-5.055	-1.54	NA
		γ_{S1}	NA	3.622	NA	NA	6.034	1.465	NA	1.414	6.896	NA	1.354	6.57	NA
	Wave 2	γ_{02}	NA	-3.094	NA	NA	3.737	1.162	NA	-5.77	-1.267	NA	-5.452	-1.163	NA
		γ_{x2}	NA	-2.551	NA	NA	2.369	0.681	NA	-4.154	-1.547	NA	-3.906	-1.441	NA
		γ_{S2}	NA	3.652	NA	NA	4.116	1.194	NA	1.823	6.378	NA	1.709	6.06	NA
	Wave 3	γ_{03}	NA	-2.198	NA	NA	4.971	1.179	NA	-4.869	-0.328	NA	-4.545	-0.211	NA
		γ_{x3}	NA	-2.705	NA	NA	3.501	0.746	NA	-4.405	-1.534	NA	-4.189	-1.452	NA
		γ_{S3}	NA	2.627	NA	NA	5.346	1.21	NA	0.723	5.342	NA	0.631	4.989	NA
	Wave 4	γ_{04}	NA	-3.469	NA	NA	4.108	1.285	NA	-6.288	-1.421	NA	-6.014	-1.338	NA
		γ_{x4}	NA	-3.122	NA	NA	3.555	0.912	NA	-5.198	-1.739	NA	-4.895	-1.646	NA
		γ_{S4}	NA	4.199	NA	NA	4.895	1.378	NA	2.059	7.192	NA	1.978	6.86	NA
N=100 (unavailable due to low convergence rate)															

Note. Abbreviations are as given in Table 3.

Table 8. Summarized Estimates from LGCM with LOD Missingness (XY)

					BIAS		SE			CI			HPD		
		para.	true	est.	smp.	rel.	emp.	avg.	MSE	lower	upper	cover	lower	upper	cover
		N=1000 (convergence rate: 100/126 ≈ 79.37%)													
Growth Curve		I	1	1.12	0.12	0.12	0.062	0.06	0.022	1.002	1.238	0.52	1.002	1.237	0.51
		S	3	3.003	0.003	0.001	0.084	0.078	0.013	2.85	3.158	0.94	2.849	3.156	0.94
		var(I)	1	1.03	0.03	0.03	0.105	0.108	0.024	0.828	1.252	0.93	0.823	1.245	0.93
		var(S)	4	3.994	-0.006	-0.002	0.253	0.235	0.119	3.556	4.479	0.91	3.542	4.46	0.90
		cov(IS)	0	0.112	0.112	0.112	0.146	0.116	0.047	-0.118	0.337	0.74	-0.115	0.338	0.72
		var(e)	1	1.015	0.015	0.015	0.048	0.044	0.004	0.933	1.105	0.91	0.93	1.102	0.92
Missingness Parameters	Wave 1	γ_{01}	NA	0.164	NA	NA	0.134	0.117	NA	-0.072	0.387	NA	-0.066	0.39	NA
		γ_{x1}	NA	-1.106	NA	NA	0.076	0.071	NA	-1.249	-0.973	NA	-1.244	-0.97	NA
		γ_{S1}	NA	0.156	NA	NA	0.093	0.073	NA	0.014	0.299	NA	0.014	0.297	NA
	Wave 2	γ_{02}	NA	-1.156	NA	NA	0.185	0.196	NA	-1.557	-0.789	NA	-1.54	-0.781	NA
		γ_{x2}	NA	-1.468	NA	NA	0.127	0.108	NA	-1.69	-1.267	NA	-1.682	-1.262	NA
		γ_{S2}	NA	0.387	NA	NA	0.047	0.044	NA	0.304	0.477	NA	0.302	0.473	NA
	Wave 3	γ_{03}	NA	-1.235	NA	NA	0.196	0.186	NA	-1.611	-0.88	NA	-1.602	-0.878	NA
		γ_{x3}	NA	-1.515	NA	NA	0.117	0.109	NA	-1.739	-1.311	NA	-1.731	-1.306	NA
		γ_{S3}	NA	0.241	NA	NA	0.028	0.025	NA	0.193	0.292	NA	0.193	0.291	NA
	Wave 4	γ_{04}	NA	-1.179	NA	NA	0.189	0.182	NA	-1.547	-0.833	NA	-1.537	-0.831	NA
		γ_{x4}	NA	-1.511	NA	NA	0.111	0.109	NA	-1.735	-1.308	NA	-1.726	-1.302	NA
		γ_{S4}	NA	0.164	NA	NA	0.018	0.017	NA	0.131	0.2	NA	0.131	0.198	NA
		N=500 (convergence rate: 100/110 ≈ 90.91%)													
Growth Curve		I	1	1.121	0.121	0.121	0.098	0.085	0.032	0.956	1.288	0.68	0.956	1.287	0.69
		S	3	3.008	0.008	0.003	0.107	0.11	0.024	2.793	3.226	0.95	2.793	3.224	0.94
		var(I)	1	1.004	0.004	0.004	0.146	0.152	0.044	0.725	1.322	0.96	0.714	1.308	0.95
		var(S)	4	3.996	-0.004	-0.001	0.399	0.334	0.27	3.391	4.698	0.86	3.365	4.662	0.86
		cov(IS)	0	0.102	0.102	0.102	0.178	0.163	0.069	-0.223	0.417	0.88	-0.219	0.42	0.89
		var(e)	1	1.026	0.026	0.026	0.062	0.063	0.008	0.91	1.156	0.95	0.906	1.151	0.94
Missingness Parameters	Wave 1	γ_{01}	NA	0.131	NA	NA	0.2	0.175	NA	-0.228	0.459	NA	-0.214	0.467	NA
		γ_{x1}	NA	-1.143	NA	NA	0.119	0.107	NA	-1.366	-0.947	NA	-1.354	-0.939	NA
		γ_{S1}	NA	0.18	NA	NA	0.133	0.109	NA	-0.032	0.396	NA	-0.03	0.395	NA
	Wave 2	γ_{02}	NA	-1.246	NA	NA	0.344	0.292	NA	-1.854	-0.709	NA	-1.822	-0.691	NA
		γ_{x2}	NA	-1.52	NA	NA	0.171	0.162	NA	-1.859	-1.225	NA	-1.841	-1.215	NA
		γ_{S2}	NA	0.409	NA	NA	0.078	0.066	NA	0.288	0.547	NA	0.285	0.539	NA
	Wave 3	γ_{03}	NA	-1.282	NA	NA	0.332	0.278	NA	-1.858	-0.768	NA	-1.827	-0.753	NA
		γ_{x3}	NA	-1.577	NA	NA	0.257	0.166	NA	-1.922	-1.275	NA	-1.903	-1.262	NA
		γ_{S3}	NA	0.249	NA	NA	0.049	0.038	NA	0.18	0.327	NA	0.178	0.324	NA
	Wave 4	γ_{04}	NA	-1.277	NA	NA	0.277	0.269	NA	-1.833	-0.774	NA	-1.808	-0.763	NA
		γ_{x4}	NA	-1.546	NA	NA	0.171	0.159	NA	-1.878	-1.255	NA	-1.861	-1.244	NA
		γ_{S4}	NA	0.174	NA	NA	0.028	0.026	NA	0.126	0.227	NA	0.125	0.224	NA

Note. Abbreviations are as given in Table 3.

Table 9. Summarized Estimates from LGCM with LOD Missingness (XY) (con't)

					BIAS		SE			CI			HPD		
		para.	true	est.	smp.	rel.	emp.	avg.	MSE	lower	upper	cover	lower	upper	cover
N=300 (convergence rate: 100/107 ≈ 93.46%)															
Growth Curve		I	1	1.139	0.139	0.139	0.127	0.111	0.047	0.922	1.357	0.70	0.922	1.356	0.69
		S	3	2.988	-0.012	-0.004	0.157	0.144	0.046	2.708	3.274	0.90	2.707	3.272	0.90
		var(I)	1	1.045	0.045	0.045	0.196	0.204	0.082	0.682	1.479	0.94	0.661	1.451	0.95
		var(S)	4	4.04	0.04	0.01	0.463	0.44	0.41	3.261	4.982	0.92	3.212	4.915	0.95
		cov(IS)	0	0.153	0.153	0.153	0.256	0.215	0.135	-0.277	0.569	0.85	-0.27	0.574	0.84
		var(e)	1	1.021	0.021	0.021	0.079	0.082	0.013	0.873	1.195	0.94	0.865	1.184	0.94
Missingness Parameters	Wave 1	γ_{01}	NA	0.103	NA	NA	0.25	0.235	NA	-0.394	0.533	NA	-0.365	0.549	NA
		γ_{x1}	NA	-1.187	NA	NA	0.163	0.147	NA	-1.499	-0.924	NA	-1.478	-0.91	NA
		γ_{S1}	NA	0.201	NA	NA	0.178	0.144	NA	-0.075	0.491	NA	-0.075	0.487	NA
	Wave 2	γ_{02}	NA	-1.251	NA	NA	0.441	0.392	NA	-2.087	-0.547	NA	-2.028	-0.514	NA
		γ_{x2}	NA	-1.583	NA	NA	0.264	0.224	NA	-2.063	-1.189	NA	-2.026	-1.165	NA
		γ_{S2}	NA	0.416	NA	NA	0.098	0.088	NA	0.258	0.605	NA	0.251	0.591	NA
	Wave 3	γ_{03}	NA	-1.333	NA	NA	0.438	0.368	NA	-2.102	-0.666	NA	-2.058	-0.641	NA
		γ_{x3}	NA	-1.614	NA	NA	0.295	0.222	NA	-2.084	-1.218	NA	-2.054	-1.199	NA
		γ_{S3}	NA	0.259	NA	NA	0.063	0.05	NA	0.169	0.364	NA	0.166	0.358	NA
	Wave 4	γ_{04}	NA	-1.406	NA	NA	0.434	0.387	NA	-2.231	-0.712	NA	-2.169	-0.682	NA
		γ_{x4}	NA	-1.656	NA	NA	0.26	0.232	NA	-2.152	-1.245	NA	-2.117	-1.223	NA
		γ_{S4}	NA	0.188	NA	NA	0.042	0.037	NA	0.122	0.268	NA	0.119	0.261	NA
N=200 (convergence rate: 100/104 ≈ 96.15%)															
Growth Curve		I	1	1.154	0.154	0.154	0.141	0.135	0.062	0.892	1.421	0.75	0.891	1.419	0.76
		S	3	2.986	-0.014	-0.005	0.187	0.176	0.066	2.648	3.336	0.92	2.644	3.331	0.93
		var(I)	1	1.019	0.019	0.019	0.233	0.25	0.117	0.583	1.562	0.96	0.552	1.517	0.97
		var(S)	4	4.034	0.034	0.008	0.516	0.536	0.557	3.107	5.202	0.96	3.043	5.11	0.96
		cov(IS)	0	0.182	0.182	0.182	0.311	0.263	0.199	-0.347	0.691	0.85	-0.338	0.697	0.85
		var(e)	1	1.047	0.047	0.047	0.103	0.104	0.024	0.863	1.27	0.92	0.852	1.255	0.94
Missingness Parameters	Wave 1	γ_{01}	NA	0.043	NA	NA	0.375	0.32	NA	-0.654	0.608	NA	-0.594	0.642	NA
		γ_{x1}	NA	-1.269	NA	NA	0.266	0.212	NA	-1.739	-0.911	NA	-1.69	-0.883	NA
		γ_{S1}	NA	0.227	NA	NA	0.272	0.197	NA	-0.148	0.631	NA	-0.15	0.616	NA
	Wave 2	γ_{02}	NA	-1.46	NA	NA	0.675	0.542	NA	-2.674	-0.532	NA	-2.541	-0.471	NA
		γ_{x2}	NA	-1.683	NA	NA	0.373	0.311	NA	-2.387	-1.165	NA	-2.303	-1.125	NA
		γ_{S2}	NA	0.463	NA	NA	0.153	0.123	NA	0.253	0.738	NA	0.24	0.707	NA
	Wave 3	γ_{03}	NA	-1.442	NA	NA	0.608	0.502	NA	-2.541	-0.573	NA	-2.457	-0.535	NA
		γ_{x3}	NA	-1.718	NA	NA	0.632	0.322	NA	-2.445	-1.19	NA	-2.377	-1.153	NA
		γ_{S3}	NA	0.271	NA	NA	0.091	0.069	NA	0.154	0.421	NA	0.148	0.41	NA
	Wave 4	γ_{04}	NA	-1.454	NA	NA	0.597	0.478	NA	-2.471	-0.601	NA	-2.4	-0.563	NA
		γ_{x4}	NA	-1.757	NA	NA	0.37	0.304	NA	-2.414	-1.227	NA	-2.364	-1.194	NA
		γ_{S4}	NA	0.196	NA	NA	0.056	0.046	NA	0.114	0.295	NA	0.11	0.287	NA
N=100 (convergence rate: 100/138 ≈ 72.46%)															
Growth Curve		I	1	1.161	0.161	0.161	0.252	0.197	0.129	0.776	1.551	0.81	0.775	1.549	0.81
		S	3	3.028	0.028	0.009	0.254	0.259	0.133	2.535	3.548	0.97	2.528	3.539	0.97
		var(I)	1	0.937	-0.063	-0.063	0.332	0.354	0.246	0.375	1.751	0.92	0.315	1.637	0.90
		var(S)	4	4.136	0.136	0.034	0.845	0.809	1.414	2.817	5.971	0.93	2.686	5.757	0.94
		cov(IS)	0	0.15	0.15	0.15	0.453	0.394	0.39	-0.657	0.902	0.88	-0.633	0.918	0.88
		var(e)	1	1.153	0.153	0.153	0.34	0.176	0.184	0.847	1.529	0.86	0.825	1.494	0.89
Missingness Parameters	Wave 1	γ_{01}	NA	-0.711	NA	NA	5.806	1.29	NA	-3.446	1.305	NA	-3.079	1.381	NA
		γ_{x1}	NA	-3.4	NA	NA	6.975	1.599	NA	-7.14	-1.402	NA	-6.495	-1.259	NA
		γ_{S1}	NA	0.468	NA	NA	4.682	0.964	NA	-1.317	2.211	NA	-1.201	2.072	NA
	Wave 2	γ_{02}	NA	-3.803	NA	NA	9.225	1.602	NA	-7.571	-1.469	NA	-6.978	-1.269	NA
		γ_{x2}	NA	-3.378	NA	NA	6.747	1.042	NA	-5.814	-1.866	NA	-5.463	-1.739	NA
		γ_{S2}	NA	1.029	NA	NA	2.14	0.367	NA	0.49	1.883	NA	0.444	1.762	NA
	Wave 3	γ_{03}	NA	-3.148	NA	NA	6.553	1.128	NA	-5.681	-1.244	NA	-5.386	-1.19	NA
		γ_{x3}	NA	-3.039	NA	NA	4.774	0.759	NA	-4.767	-1.788	NA	-4.554	-1.735	NA
		γ_{S3}	NA	0.534	NA	NA	0.974	0.156	NA	0.275	0.888	NA	0.264	0.839	NA
	Wave 4	γ_{04}	NA	-2.582	NA	NA	4.213	1.27	NA	-5.474	-0.588	NA	-5.149	-0.48	NA
		γ_{x4}	NA	-3.259	NA	NA	5.963	0.989	NA	-5.478	-1.717	NA	-5.176	-1.583	NA
		γ_{S4}	NA	0.297	NA	NA	0.719	0.146	NA	0.064	0.611	NA	0.055	0.579	NA

Note. Abbreviations are as given in Table 3 .

Table 10. Summarized Estimates from LGCM with Ignorable Missingness (X)

				BIAS		SE			CI			HPD		
	para.	true	est.	smp.	rel.	emp.	avg.	MSE	lower	upper	cover	lower	upper	cover
						N=1000 (convergence rate: 100/100 = 100%)								
Growth Curve	I	1	1.009	0.009	0.009	0.051	0.052	0.005	0.906	1.111	0.94	0.906	1.111	0.93
	S	3	2.711	-0.289	-0.096	0.078	0.077	0.095	2.56	2.863	0.04	2.561	2.863	0.04
	var(I)	1	1.008	0.008	0.008	0.108	0.104	0.022	0.813	1.221	0.95	0.807	1.214	0.95
	var(S)	4	3.837	-0.163	-0.041	0.232	0.223	0.13	3.422	4.297	0.87	3.409	4.279	0.86
	cov(IS)	0	0.004	0.004	0.004	0.115	0.109	0.025	-0.214	0.214	0.96	-0.21	0.216	0.96
	var(e)	1	0.999	-0.001	-0.001	0.044	0.043	0.004	0.919	1.086	0.92	0.917	1.084	0.92
						N=500 (convergence rate: 100/100 = 100%)								
Growth Curve	I	1	0.999	-0.001	-0.001	0.073	0.074	0.011	0.854	1.143	0.98	0.855	1.143	0.98
	S	3	2.711	-0.289	-0.096	0.099	0.109	0.105	2.497	2.925	0.21	2.497	2.925	0.21
	var(I)	1	0.973	-0.027	-0.027	0.146	0.146	0.043	0.705	1.277	0.98	0.693	1.263	0.98
	var(S)	4	3.852	-0.148	-0.037	0.371	0.317	0.259	3.276	4.518	0.86	3.248	4.48	0.88
	cov(IS)	0	-0.008	-0.008	-0.008	0.154	0.154	0.047	-0.317	0.287	0.96	-0.311	0.292	0.96
	var(e)	1	1.014	0.014	0.014	0.06	0.062	0.008	0.9	1.141	0.96	0.895	1.136	0.95
						N=300 (convergence rate: 100/100 = 100%)								
Growth Curve	I	1	1.009	0.009	0.009	0.103	0.096	0.02	0.821	1.197	0.89	0.821	1.197	0.89
	S	3	2.687	-0.313	-0.104	0.139	0.141	0.137	2.411	2.964	0.34	2.411	2.963	0.35
	var(I)	1	1.006	0.006	0.006	0.189	0.194	0.073	0.657	1.416	0.94	0.639	1.391	0.94
	var(S)	4	3.816	-0.184	-0.046	0.412	0.41	0.372	3.091	4.694	0.93	3.045	4.631	0.92
	cov(IS)	0	0.045	0.045	0.045	0.214	0.2	0.088	-0.359	0.429	0.94	-0.351	0.435	0.94
	var(e)	1	1.01	0.01	0.01	0.075	0.08	0.012	0.864	1.179	0.96	0.857	1.17	0.94
						N=200 (convergence rate: 100/100 = 100%)								
Growth Curve	I	1	1.019	0.019	0.019	0.098	0.116	0.023	0.792	1.247	0.97	0.791	1.246	0.97
	S	3	2.69	-0.31	-0.103	0.178	0.173	0.157	2.352	3.029	0.52	2.351	3.027	0.52
	var(I)	1	0.99	-0.01	-0.01	0.232	0.236	0.11	0.576	1.5	0.94	0.548	1.461	0.95
	var(S)	4	3.884	-0.116	-0.029	0.47	0.509	0.495	3.004	4.992	0.96	2.938	4.898	0.96
	cov(IS)	0	0.066	0.066	0.066	0.25	0.246	0.127	-0.434	0.538	0.91	-0.422	0.546	0.92
	var(e)	1	1.02	0.02	0.02	0.094	0.1	0.019	0.843	1.233	0.95	0.833	1.219	0.97
						N=100 (convergence rate: 100/100 = 100%)								
Growth Curve	I	1	1.031	0.031	0.031	0.174	0.161	0.057	0.714	1.348	0.94	0.715	1.348	0.95
	S	3	2.699	-0.301	-0.1	0.239	0.248	0.209	2.21	3.187	0.78	2.212	3.187	0.78
	var(I)	1	0.863	-0.137	-0.137	0.275	0.315	0.197	0.354	1.579	0.94	0.302	1.487	0.86
	var(S)	4	3.951	-0.049	-0.012	0.815	0.753	1.247	2.726	5.66	0.92	2.601	5.456	0.92
	cov(IS)	0	0.063	0.063	0.063	0.35	0.35	0.25	-0.658	0.728	0.91	-0.637	0.744	0.94
	var(e)	1	1.063	0.063	0.063	0.137	0.149	0.045	0.808	1.39	0.94	0.788	1.361	0.95

Note. Abbreviations are as given in Table 3.

Journal of Behavioral Data Science, 2021, *1* (2), 31–53.
DOI: `https://doi.org/10.35566/jbds/v1n2/p3`

Tree-based Matching on Structural Equation Model Parameters

Sarfaraz Serang[1][0000−0002−7985−4951] and James Sears[2][0000−0002−0087−1354]

[1] Utah State University, Logan, UT 84322, USA
sarfaraz.serang@usu.edu
[2] University of California, Berkeley, CA 94720, USA
james.sears@berkeley.edu

Abstract. Understanding causal effects of a treatment is often of interest in the social sciences. When treatments cannot be randomly assigned, researchers must ensure that treated and untreated participants are balanced on covariates before estimating treatment effects. Conventional practices are useful in matching such that treated and untreated participants have similar average values on their covariates. However, situations arise in which a researcher may instead want to match on model parameters. We propose an algorithm, Causal M*plus* Trees, which uses decision trees to match on structural equation model parameters and estimates conditional average treatment effects in each node. We provide a proof of concept using two small simulation studies and demonstrate its application using COVID-19 data.

Keywords: Matching · Structural Equation Modeling · Decision Trees · Machine Learning

1 Introduction

Understanding the causal effect of a treatment has historically been of great scientific interest and remains one of the most frequently pursued objectives in scientific research today. The gold standard for evaluating treatment effects is the randomized controlled trial, where the researcher randomly assigns treatment status to each individual. The benefit of this approach is that the causal effect of the treatment can be estimated by simply comparing outcomes between those who were treated and those who were not (Greenland, Pearl, & Robins, 1999). Random assignment of treatment guarantees that, on average, the treated and untreated individuals will be equal on all potential confounding variables, both measured and unmeasured. Eliminating the possibility of confounding clears the way for a direct comparison to be made.

However, random assignment is not always possible. This can be for ethical reasons, since researchers cannot, for example, force participants to smoke to investigate the effects of smoking. It can also be for practical reasons, where the researcher cannot control the assignment of a treatment. For example, researchers cannot randomly assign depression to some participants, enact a law or policy in a randomly assigned jurisdiction, or choose where their participants live. An observational study, where treatment is not randomly assigned, may be the only available option in these cases. Unlike randomized controlled trials, direct comparisons between treated and untreated individuals in an observational study cannot be made as easily. This is because treated and untreated participants may not be equal in all other characteristics, creating the potential for confounding effects. In fact, it may be differences in these very characteristics that lead some participants to select treatment, making the estimation of the treatment's effect less straightforward. To estimate a treatment's effect, it must first be defined, which we do in the context of the potential outcomes framework.

1.1 Potential Outcomes Framework and Assumptions

The foundations for the potential outcomes framework were laid out by Neyman, Iwaszkiewicz, and Kolodziejczyk (1935) and further developed by Rubin (1974), resulting in it also being called the Rubin Causal Model, Neyman-Rubin Causal Model, and Neyman-Rubin counterfactual framework of causality. The model can be conceptualized as follows. Let Y_{1i} be the potential outcome of individual i if they received the treatment and Y_{0i} be the potential outcome of individual i if they did not receive the treatment. The observed score Y_i, can be written as

$$Y_i = W_i Y_{1i} + (1 - W_i) Y_{0i} \tag{1}$$

where $W_i = 1$ if the individual received treatment and $W_i = 0$ if they did not. W_i simply acts as an indicator variable denoting the receipt of treatment. The term *treatment* here and throughout the paper is used rather loosely and can be used interchangeably with *exposure*.

The effect of the treatment is simply $Y_{1i} - Y_{0i}$, the difference between the potential outcomes if the individual had received treatment and if they had not. The fundamental problem of causal inference, as stated by Holland (1986), is that it is impossible to observe both Y_{1i} and Y_{0i} for the same individual. If the individual received treatment, we can observe Y_{1i}, but not its counterfactual, Y_{0i} . The inverse is also true: if the individual did not receive treatment, we can observe Y_{0i} , but not its counterfactual, Y_{1i}. Therefore, it is impossible to observe the effect of the treatment on the individual. As an example, we can see that it is impossible to observe the effect of divorce on a child's academic test scores because at a given moment in time, the parents can either be divorced or not divorced, but not both. We cannot observe the test scores under both conditions, so we cannot observe the effect of divorce on that child's scores.

Though we cannot observe the effect of the treatment on a given individual, we can estimate the *average treatment effect* (ATE) on a population. The ATE is the average effect expected from taking a population where no individuals received the treatment and providing the treatment to all of them (Austin, 2011).

The ATE is defined as $ATE = E(Y_{1i} - Y_{0i}) = E(Y_{1i}) - E(Y_{0i})$, where $E(\cdot)$ is the expected value operator. Conceptually, this implies that although we cannot observe the treatment effect at the individual level, we can do so at a population level by using the average of the untreated participants as a proxy for the unobservable counterfactual (Guo & Fraser, 2010). A related effect of interest in this paper is the *conditional average treatment effect* , or *CATE*, (Abrevaya, Hsu, & Lieli, 2015), defined as $CATE = E(Y_{1i} - Y_{0i}|\boldsymbol{X}_i)$, where $\boldsymbol{X}_i$ is a vector of covariates. The CATE allows us to evaluate heterogeneity in treatment effects between subpopulations, for example, allowing for separate estimation of the ATE in males and females if they are believed to be different.

One important assumption of the potential outcomes framework is the Stable Unit Treatment Value Assumption, or SUTVA (Rubin, 1980, 1986). It represents the assumption that the potential outcomes would be the same no matter how an individual came to be assigned to a treatment, and no matter what treatments are received by other individuals. It assumes that neither treatment assignment mechanisms nor social interactions affect potential outcomes. Another assumption, one we give more attention due to the focus of this paper, is known as the *strong ignorability* assumption (Rosenbaum & Rubin, 1983). Treatment assignment is strongly ignorable if two conditions collectively hold. The first condition is $(Y_0, Y_1) \perp W|\boldsymbol{X}$, that treatment assignment is independent of the potential outcomes conditional on covariates. The second condition is $0 < P(W = 1|\boldsymbol{X}) < 1$, that every participant has a nonzero probability of receiving either treatment, conditional on covariates.

The necessity of the conditional independence piece of the strong ignorability assumption becomes evident when considering the necessary conditions for using untreated participants as a proxy for the counterfactual. To estimate the ATE by taking the difference between the averages of treated and untreated participants, we implicitly assume that the average scores produced by the untreated participants are an unbiased estimate of what the average scores produced by the treated participants would have been had they not received the treatment. In doing so, we must ensure that the treated and untreated participants are similar in relevant characteristics, so that the untreated participants can serve as a faithful representation for their treated counterparts. For example, if the treated group contained only males and the untreated group contained only females, using the untreated group as a proxy for the treated group might not produce a fair comparison, depending on what is being studied. This is why randomized controlled trials are considered the gold standard: random assignment ensures that, on average, all such possible confounders are balanced, making the treated and untreated participants comparable.

As pointed out by Thoemmes and Kim (2011), the strong ignorability assumption cannot be empirically tested. This is because treatment assignment must be conditionally independent of all relevant covariates both observed and unobserved, and it is not possible to empirically verify that variables that are not collected do not play a role. As such, researchers who attempt to justify this assumption are limited to making a convincing argument that they have measured the relevant covariates and showing that these are balanced across treated and untreated participants. The most common way of demonstrating balance in an observed covariate across groups is via a standardized mean difference. This takes the form of the mean difference in the covariate between groups (in absolute value) divided by either a pooled standard deviation or an unpooled standard deviation of one of the groups.

A standardized mean difference of 0 would indicate the covariate has the same mean across groups. However, there is no universally agreed upon metric for judging how small a nonzero standardized mean difference must be to be considered negligible enough for the groups to be considered balanced on the covariate for practical purposes. Many recommendations exist in the methodological literature. Harder, Stuart, and Anthony (2010) use a value less than 0.25, based on a suggestion by Ho, Imai, King, and Stuart (2007). Austin (2011) suggests a stricter value of less than 0.1, based on work by Normand et al. (2001). Leite, Stapleton, and Bettini (2018) point out that for educational research, the What Works Clearinghouse Procedures and Standards Handbook (version 4.0) requires a value less than 0.05 without additional covariate adjustment, or between 0.05 and 0.25 with additional regression adjustment (U.S. Department of Education, Institute of Education Sciences, & What Works Clearinghouse, 2017).

Analyzing standardized mean differences is reasonable when attempting to balance across demographic covariates such as sex, age, race, etc. Yet some characteristics do not lend themselves well to being assessed in this way. Consider an example where we are interested in evaluating the effects of a breakup from a romantic relationship (the treatment) on life satisfaction (the outcome). For simplicity, let us assume that we only collect data from one partner per couple. Putting demographics aside, affect might be an important covariate to balance on. However, ensuring that couples who do and do not break up have the same average affect might not be especially useful. *Stability* of affect has been shown to be predictive of whether couples remain together or break up (Ferrer, 2016; Ferrer, Steele, & Hsieh, 2012). That is, fluctuations in affect are what need to be balanced, not simply average affect. Consider the plot given in Figure 1 of two hypothetical individuals, J and K, and their affect over time. J has highly variable affect, whereas K has relatively stable affect. Based on the aforementioned research, J is more likely to experience a breakup, given their instability. However, both J and K have the same average affect. Imagine a treatment group filled with individuals like J and an untreated group filled with individuals like K. According to the standardized mean difference, these two groups would be balanced across affect, because they have the same mean affect. The fact that they have different patterns with regard to the variability would be entirely missed.

The literature does recommend that covariates should be balanced across groups on not just the mean, but the distribution of the variables (Austin, 2011; Ho et al., 2007). Researchers are encouraged to examine higher-order moments, as well as interactions between covariates. Graphical methods are often used

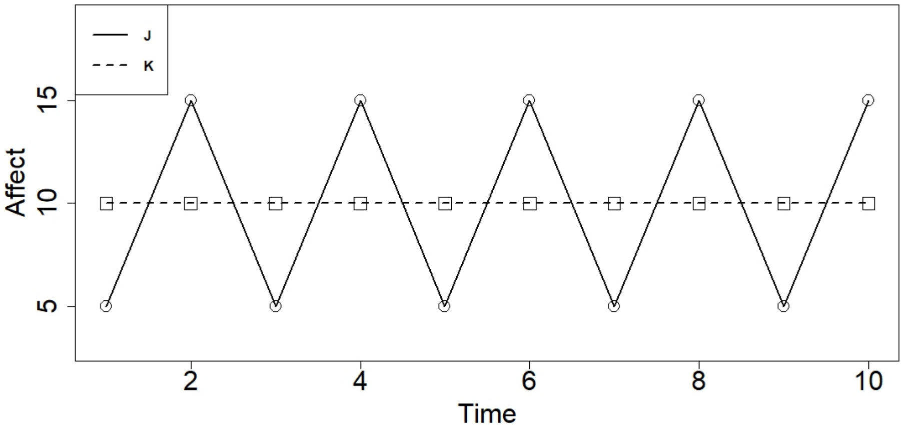


Figure 1. Stability of Affect in Two Hypothetical Individuals

to make these comparisons, including quantile-quantile plots, boxplots, density plots, etc. Though visualizations can be helpful for univariate or even bivariate data, they become less useful with higher-dimensional data, as in our example. Furthermore, in this case, they do not quite address the issue directly. We would like to balance on stability of affect, which is not entirely captured by either univariate higher order moments or interactions.

1.2 Purpose

Although conventional approaches can be useful when balancing on demographic variables and other such covariates, they are not as well suited for balancing on more complex functions of the data, such as stability of affect. This paper seeks to develop an approach that allows us to balance on more flexibly defined characteristics of interest. We begin by reviewing some classic and recent approaches to matching. We then provide an introduction to structural equation model trees and their variations. Drawing from these, we propose our own algorithm, Causal M*plus* Trees, and describe its implementation. We then conduct two small simulation studies demonstrating our algorithm's effectiveness and an empirical analysis of COVID-19 data. We conclude with a discussion of practical recommendations and future directions.

1.3 Propensity Score Matching

Thus far we have discussed ways to evaluate whether treated and untreated participants are balanced on covariates. If they are found to be unbalanced, we can turn to statistical approaches to balance them. A natural initial thought would be to use ordinary least squares regression, conditioning on covariates within the model. However, Berk (2004) points out that simply calculating a conditional distribution of the outcome is not sufficient to draw causal inference and that stronger assumptions are needed.

A popular alternative is to use propensity scores, defined as the probability of treatment conditional on observed covariates (Rosenbaum & Rubin, 1983). It has been shown that propensity scores can balance treated and untreated participants in the sample, and that both treatment assignment as well as observed covariates are conditionally independent given the propensity score (Rosenbaum & Rubin, 1983). This implies that for participants with the same propensity score, the mean difference in the outcome between treated and untreated participants is an unbiased estimate of the ATE at that propensity score (Guo & Fraser, 2010). Propensity scores are typically calculated using logistic regression, with the observed covariates predicting treatment status (W). The estimated regression coefficients are then used as weights in a model predicting the probability of treatment for each individual. The estimated propensity score is this predicted probability of treatment.

Once propensity scores have been calculated, they can be used in various ways, including propensity score matching (Rosenbaum & Rubin, 1985), stratification on the propensity score (Rosenbaum & Rubin, 1984), and inverse probability of treatment weighting using the propensity score (Hirano & Imbens, 2001). Of these three, propensity score matching seems to eliminate more of the systematic differences in covariates (Austin, 2009) and also seems to be the most popular (Thoemmes & Kim, 2011), so we limit our focus to propensity score matching. Propensity score matching involves finding treated and untreated participants with similar propensity scores to use as each other's counterfactuals. According to Austin (2011) and the systematic review conducted by Thoemmes and Kim (2011), the most commonly used form of matching is 1:1 matching, where each treated participant is matched with a single untreated participant, forming a pair. Thoemmes and Kim (2011) found that the most popular way to do this in the social sciences was to use greedy matching, in which a treated subject is selected at random and the untreated subject with the closest propensity score is paired with them. The process is repeated until all treated subjects have a match. This is in contrast to optimal matching, where matches are selected to optimize the distance between propensity scores for the entire sample, which has been shown to perform comparably to greedy matching (Gu & Rosenbaum, 1993). The 1:1 matching scheme produces pairs of treated and untreated participants who should in theory be balanced on the propensity scores. The ATE can then be estimated simply by performing a paired t test (Austin, 2011).

Of course, one must still ensure that the propensity scores are balanced across treated and untreated participants. If they are not, it is recommended that the logistic regression model be iteratively refined by including nonlinear terms and interactions between covariates until balance has been achieved (Austin, 2011; Rosenbaum & Rubin, 1984, 1985; West et al., 2014). Latent variable models can be used to calculate propensity scores by balancing on latent covariates whose scores are estimated via factor score estimation (Raykov, 2012), or by using structural equation modeling to estimate propensity scores directly (Leite et al., 2018). Machine learning techniques including bagging, boosting, trees, and random forests, have also been used for the estimation of propensity scores (Lee, Lessler, & Stuart, 2010).

1.4 Causal Trees

A recent alternative to propensity score matching is the causal tree approach proposed by Athey and Imbens (2016). Essentially, they use decision trees to partition the sample into groups of individuals who are similar on important dimensions. They then treat these groupings as matched, and use them to estimate the ATE. Decision trees (Breiman, Friedman, Olshen, & Stone, 1984), use recursive partitioning to separate a predictor space into regions that are as homogeneous as possible on a target variable of interest. Binary splits are made on predictors (e.g. female vs. male, age ≤ 60 vs. age > 60, etc.), splitting the sample into two nodes. All possible splits are made on all predictors, and the split that makes the resulting samples in each node as homogeneous as possible is presented as a candidate split. If this split exceeds a predetermined fit criterion, the split is made, partitioning the sample into the two daughter nodes. Otherwise, the split is not made, and the parent node becomes a terminal node. The process continues recursively on each daughter node until all nodes are terminal nodes. We refer readers to Serang et al. (2021) for additional description of the procedure.

Decision trees are most often used for prediction of a target variable. The critical insight of Athey and Imbens (2016) is that trees have a natural proclivity for creating homogeneous subgroups. Instead of trying to predict a target variable, we can substitute the vector of covariates, $\boldsymbol{X}$. The tree will then produce terminal nodes where the observations in each terminal node are as similar as possible on the covariates, achieving the same aim as matching. Each terminal node is characterized by splits on predictors (separate from $\boldsymbol{X}$) that define membership in that node. In what they call an *honest* approach to estimation, the authors recommend that these subgroup definitions be applied to a fresh holdout sample not involved in the construction of the tree, to create subgroups using the new data. CATEs (ATEs conditional on subgroup membership) can then be estimated in each subgroup via mean differences between treated and untreated participants within the subgroup. Causal inference can also be drawn using standard approaches, such as an independent-samples t test.

The advantage of causal trees over propensity score methods is that one need not worry about the estimation of or balancing on propensity scores. Propensity scores only serve as a middleman in propensity score matching, and causal trees use the properties of decision trees to bypass them entirely. Additionally, causal trees easily accommodate heterogeneity in causal effects. In our running example, we wish to match on stability of affect. If we use demographic variables as splitting variables in the tree, we can potentially find subgroups defined by these demographic characteristics (e.g. sex, age, etc.) that have different levels of affect stability. The causal tree approach would then allow us to estimate the causal effect of a breakup separately in each of these subgroups, as well as compare them to see if the causal effect differs by subgroup.

1.5 Structural Equation Model Trees

One limitation of causal trees as described is that they assume we wish to match on observed covariates. However, stability in our example is not an observed variable in the data: it is a characterization based on a pattern. One way to characterize stability for the data in our example would be to fit a simple intercept-only growth curve model and examine the residual variance. A model fit to individuals such as J would produce a large residual variance, whereas a model fit to individuals like K would yield a relatively small residual variance. Thus, stability of a group can be characterized by model-based parameter estimates, in lieu of observed variables.

To do this within the causal tree framework, we would need a mechanism to fit a model within each node. For longitudinal models, we can use an approach like the nonlinear longitudinal recursive partitioning algorithm proposed by Stegmann, Jacobucci, Serang, and Grimm (2018), which allows the user to fit linear and nonlinear longitudinal models within each node. A more general approach is the structural equation model tree (SEM Tree) proposed by Brandmaier, Oertzen, McArdle, and Lindenberger (2013), which allows for structural equation models (SEMs) to be fit within each node. A benefit of the latter is the flexibility of the SEM framework, which can accommodate a wide range of models, including many longitudinal models, via latent growth curve modeling (Meredith & Tisak, 1990).

The logic of SEM Trees is similar to that of standard decision trees, with some minor variations. A prespecified SEM is first fit to the full sample, and the minus two log-likelihood ($-2LogL$) is calculated. Then, the $-2LogL$ for the candidate split is calculated. Since the split can be conceptualized as a multiple group model (Jöreskog, 1971), the $-2LogL$ for the split is simply the sum of the $-2LogL$ values for each daughter node. A likelihood ratio test is then conducted with these two $-2LogL$ values. If it rejects, the split is made. As in other decision trees, this process is recursively repeated until all daughter nodes are terminal nodes. Unlike conventional decision trees, terminal nodes in SEM Trees do not provide a predicted proportion or mean. Rather, each terminal node is characterized by a set of parameter estimates for the SEM fit to the sample in that node. In this way, SEM Trees can be used to identify subgroups of people who are similar in that they can be represented by a set of parameter estimates that is distinct from the parameter estimates that characterize those in other nodes. SEM Trees can therefore identify subgroups with distinct patterns of stability, growth, or other patterns reflected in the parameter estimates.

1.6 M*plus* Trees

The SEM Trees algorithm is implemented in the semtree (Brandmaier, Prindle, & Arnold, 2021) package in R (R Core Team, 2020). The SEMs are fit in either the OpenMx package (Neale et al., 2016) or the lavaan package (Rosseel, 2012). The OpenMx package is flexible but challenging to use, especially for casual users, given the need to specify the entirety of the model with limited defaults. The lavaan package is much easier to use given the ease with which one can specify models, however it is currently more limited in the scope of the models it can fit. The MplusTrees package (Serang et al., 2021) is an implementation of SEM Trees which uses M*plus* (Muthén & Muthén, 1998-2017) to fit the models, the rpart package (Therneau & Atkinson, 2018) to perform the recursive partitioning needed to grow the trees, and the MplusAutomation package (Hallquist & Wiley, 2018) to interface between R and M*plus*. MplusTrees capitalizes on the wide variety of complex models that can be specified in M*plus*, the ease with which they can be specified, and the currently superior estimation algorithms it uses for fitting these models.

The M*plus* Trees algorithm itself (Serang et al., 2021) is very similar to the SEM Trees algorithm (Brandmaier et al., 2013). However, one key difference is the criterion used for splitting. Although the MplusTrees package also has the capability to split using the likelihood ratio test, this is not the primary method. Instead, M*plus* Trees uses a complexity parameter, *cp*. This *cp* parameter is a proportion specified in advance by the user. A split will be made if that split improves on the *-2LogL* of the full sample (the parent node) by at least *cp* times that *-2LogL*. Smaller values of *cp* result in more splits since a relatively smaller improvement in the *-2LogL* is needed for a split to be made, whereas larger values lead to fewer splits. As such, the use of *cp* serves more as a heuristic than a formal test based on statistical significance. Ideally, *cp* would be selected by cross-validation, and this functionality is available in the MplusTrees package. However, long computational times may require users to simply try a handful of *cp* values and select the most appropriate one given the context.

2 Causal M*plus* Trees

We now propose our own matching algorithm, Causal M*plus* Trees, using M*plus* Trees to create causal trees that match on parameters from an SEM, and estimating CATEs in a holdout sample. We begin by first randomly partitioning the dataset into two parts, one subsample to perform the matching and the other to perform the estimation of the CATEs. In most cases, the matching subsample will require more participants, since fitting an SEM and building a decision tree is more sample intensive than estimating a mean difference. We suggest devoting 80% of the sample to the matching subsample and 20% to the estimation subsample, though this ratio can be adjusted depending on the complexity of the SEM, the overall sample size, etc.

Beginning with the matching subsample, we can partition $\boldsymbol{X}$ into two parts: $\boldsymbol{X}_M$, the *modeled covariates* modeled in the SEM whose parameters we wish to match on, and $\boldsymbol{X}_S$, the *splitting covariates* we want to split on in the recursive partitioning process which define the subgroups of the tree's terminal nodes. Guidance for whether a covariate should be a modeled covariate or a splitting covariate is provided in the discussion. Let M be an SEM with parameters $\boldsymbol{\theta}$ that produces $\boldsymbol{X}_M$, so that $M(\boldsymbol{\theta}) = \boldsymbol{X}_M$. In our running example, M would be the intercept-only growth model and $\boldsymbol{\theta}$ would be its parameters. For properly specified M, $\boldsymbol{X}_M$ can be used to estimate $\boldsymbol{\theta}$, resulting in parameter estimates $\hat{\boldsymbol{\theta}}$. Using M*plus* Trees, we can build a tree that matches on $\hat{\boldsymbol{\theta}}$, with groups (terminal nodes) defined by their covariate patterns on $\boldsymbol{X}_S$. The treatment assignment information, W, is not provided to the recursive partitioning algorithm and so the tree is built blind to W. In the estimation subsample, we can divide participants into groups according to the splits found by the tree. Within each group, we can estimate the CATE as defined before by taking the difference between the means of the outcomes of the treated and untreated participants in each group. Since we are using a fresh sample, we can draw inference using hypothesis tests such as an independent-samples t test or another suitable alternative. We can also test whether the CATE differs by group by testing the interaction effect in a two-way independent ANOVA.

3 Simulation Studies

As a proof of concept for Causal Mplus Trees, we performed two small simulation studies. The simulation studies were conducted in R using the `lavaan` package to simulate data and the `MplusTrees` package for analysis. Readers are referred to the package documentation for details regarding the implementation of the algorithm in the software. Each simulation consisted of 1,000 replications.

3.1 Longitudinal Simulation

The first simulation mapped onto our running example regarding stability of affect. Each sample consisted of $N = 2{,}000$ individuals, 1,000 in each of two groups. The data were generated from an intercept-only (no growth) model with 10 time points. The intercept had a mean of 10 with a variance of 1. The only difference between the groups was in the residual variance, σ^2_ϵ. One group had a residual variance of 1 (the group with stable affect), and the other had a residual variance of 10 (the group with unstable affect). The group memberships were identified by a dichotomous covariate, used as a splitting variable. Thus, the tree matched on the growth curve, using the group membership to split. Within each group, treated and untreated participants were evenly split (500 each). A diagram of this population tree is given in Figure 2. For the stable affect group, outcomes were generated using a standard normal distribution, $N(0, 1)$, for the untreated group and a $N(0.5, 1)$ distribution for the treated group, to represent a medium-sized CATE. However, for the unstable affect group, the outcome distributions were flipped, with the untreated group's outcome being generated from a $N(0.5, 1)$ distribution, whereas the treated group's outcome was generated from a $N(0,1)$ distribution. In this way, although the ATE for the full sample was 0, the CATE for each group was 0.5 in absolute value.

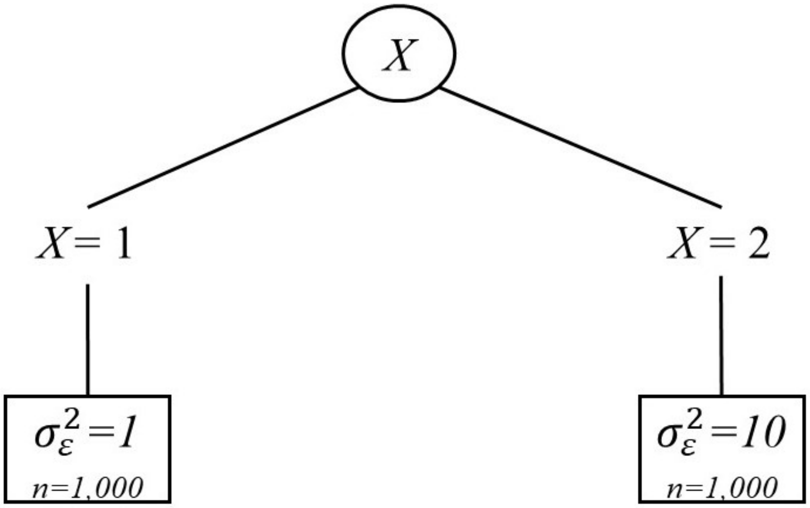


Figure 2. Population Tree for Longitudinal Simulation

It should be noted that these groups are, from the start, balanced on the modeled covariates. Since the growth curve variables were all generated to have a mean of 10, they would be considered balanced according to the standardized mean difference. Thus, if one were to follow conventional procedure, propensity scores would not be needed here, and the estimation of the ATE would consist of simply the mean difference between treated and untreated participants, which would be 0 on average.

The Causal M*plus* Trees algorithm was implemented as described in the prior section, with 80% of the sample (1,600 individuals) used for matching and 20% (400 individuals) used to estimate CATEs. A *cp* value of .01 was used to split, with a minimum of 100 individuals required to consider splitting on a node. Each terminal node was also required to have at least 100 individuals within it. For each replication, the CATE was estimated in each group using an independent samples t test. A two-way independent ANOVA was also conducted to determine if CATEs differed by group.

Overall, the results demonstrated the effectiveness of the algorithm. Across all replications, 94.5% of CATEs were detected. Additionally, 99.8% of the interactions from the two-way ANOVA were detected, showing that the algorithm can detect differences in CATEs by group. As a comparison, we also analyzed these data as they would have been analyzed using the conventional approach. Since the covariates were on average balanced according the standardized mean difference, the ATE would have been estimated by using the full sample to estimate the mean difference between treated and untreated participants. Despite a sample size of 2,000 to do this (relative to the only 400 available to Causal M*plus* Trees after performing the matching), only 3.4% of datasets yielded statistically significant ATEs, consistent with a nominal false positive rate of 5%.

3.2 Measurement Simulation

The second simulation study is similar to the first, but used a measurement model as opposed to a longitudinal model for the matching. For the second study, each sample consisted of $N = 3{,}000$ individuals, divided into three groups. One group (the small loading group) contained 1,500 individuals, while the remaining two groups (the medium and large loading group) each contained 750. Data were generated from a one-factor confirmatory factor analysis model with 15 items. Factor variances were fixed to 1, and uniquenesses were also simulated to be 1. As implied above, the only differences were in the loadings, λ. In the small loading group, all loadings were simulated to be 0.1, in the medium loading group they were 0.5, and in the large loading group they were 0.9. The model was generated to reflect the case where items are more related to a latent construct for some people than for others. If the latent variable were a psychological disorder, this would map onto the idea that the items better reflect the presence of that disorder in some groups relative to others.

As with the previous simulation study, a single splitting covariate denoting group membership was used as the splitting variable, albeit with three values given the three groups. Figure 3 shows a diagram for this population tree. As with the other simulation study, each group was evenly divided on treated and untreated participants. In the small loading group untreated participants had outcomes generated from a $N(0, 1)$ distribution, whereas the treated group's outcome was generated from a $N(0.5, 1)$ distribution. In the medium and large loading groups this was reversed: untreated participants had outcomes from a $N(0.5, 1)$ distribution whereas treated participants had outcomes from a $N(0, 1)$ distribution. In this way, these samples too had an average ATE of 0, in addition to being on average balanced on the modeled covariates according to the standardized mean difference, since all items had an average score of 0.

The algorithm again used 80% of each sample (2,400 participants) for matching and 20% (600 participants) for estimation. As before, a minimum of 100 individuals was required to consider splitting a node and in each terminal node, however this study used a *cp* value of .001. Unlike the previous study where the split was made in every replication, the algorithm had some slight trouble finding all the groups in this study. All three groups were found in 92.7% of simulations, but only two groups were found in the remaining 7.3%. Among all the groups found, 88.3% of the CATEs were detected, along with 99.7% of the interactions. Alternatively, when using the entire sample to calculate the ATE, only 3.5% of simulations yielded significant results. These results are similar to those found in the first simulation study. Taken together, they show that the Causal M*plus* Trees algorithm is able to estimate CATEs and support hypothesis testing to determine their statistical significance. It can also determine whether the CATEs differ by group. Notably, CATEs were found in the absence of ATEs, with modeled covariates already balanced across treated and untreated participants according to the standardized mean difference.

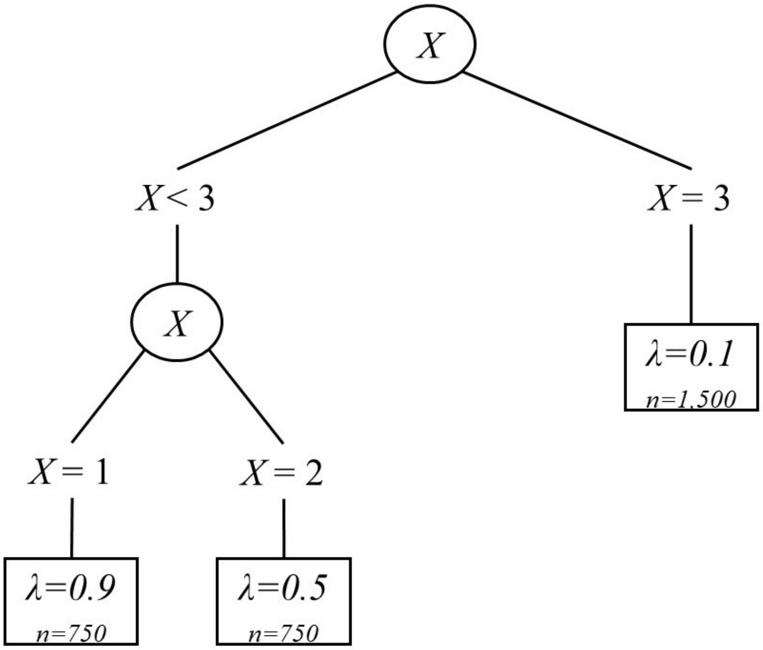


Figure 3. Population Tree for Measurement Simulation

4 Empirical Example

As an illustration of how Causal M*plus* Trees can be used in practice, we present an analysis of COVID-19 data. The dataset contains information from four different sources: public health data from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (Dong, Du, & Gardner, 2020), demographic data from the 2010 US Decennial Census (U.S. Census Bureau, 2010), governor's party information obtained from the National Governors Association Roster (National Governors Association, 2020), and mobility data from Unacast, a location data analytics company (Unacast, 2020).

To capture how individuals' travel activity patterns responded to the spread of COVID-19, we utilized Unacast's measure of the change in average distance traveled. Travel distance was measured using the GPS positions of millions of mobile devices and aggregated each day to a county-level average. For a detailed overview of variable construction and discussion of potential sources of bias, see Sears, Villas-Boas, Villas-Boas, and Villas-Boas (2020). The data were analyzed at the county level, consisting of 3,030 counties or county-equivalents from all 50 US states except Alaska. This represents over 95% of counties in the US.

The goal of this analysis was to estimate the CATE of the governor's party (Democrat or Republican) on mobility in counties matched on the trajectory of COVID-19 cases early in the pandemic. We sought to answer the question: "for counties with similar trajectories of the rise in COVID-19 cases from March through June 2020, could differences in mobility in July 2020 be attributed to the governor's party?" Prior studies reveal strong links between political partisanship and the adoption of stay-at-home and social distancing orders as well as changes in residents' travel behavior and time spent at home (Adolph, Amano, Bang-Jensen, Fullman, & Wilkerson, 2020; Allcott et al., 2020; Brzezinski, Deiana, Kecht, & Van Dijcke, 2020; Gadarian, Goodman, & Pepinsky, 2020). We provide a complementary analysis allowing us to understand whether the effect of gubernatorial political alignment extended beyond stay-at-home adoption timing to continued behavioral changes among constituents. Our analysis also examined how this effect differed across counties depending on demographic characteristics.

Case trajectories were modeled using the cumulative cases in the county divided by the population per 10,000 residents, hereafter referred to as *COVID rates*. COVID rates were calculated weekly from March 9, 2020 (around when states began reporting their first cases) until June 29, 2020, resulting in 17 time points of data per county. The SEM fit within each node of the tree was the logistic growth model given by

$$COVID_i = \frac{\beta_{1i}}{1 + e^{-(t-\gamma)\alpha}} + \epsilon_i \tag{2}$$

where $COVID_i$ is the COVID rate for county i, β_{1i} is the county-specific COVID rate when the "curve has flattened" (the upper asymptote), t is the number of weeks ($t = 1, 2, \ldots, 17$), γ is the inflection point, α is the rate of change, and ϵ is the residual. The model was specified using Taylor-series approximation (Browne & Toit, 1991; Grimm & Ram, 2009) with equal residual variances across time, σ^2_ϵ, to aid estimation.

We used six demographic splitting variables: *population* (the total population of the county), *white* (the percentage of non-Hispanic Whites), *age65_older* (the percentage of people ages 65 years and older), *median_inc* (the median household income), *bachelors* (the percentage of people with at least a bachelor's degree), and *rural* (the percentage of the population considered rural). To reduce the computational burden of the algorithm, we reassigned values from 1 to 4 to each of these splitting covariates depending on the quartile in which they fell relative to the other counties.

In implementing the Causal M*plus* Trees algorithm, we used 2,424 counties to match the data and the remaining 606 to estimate the CATEs. We required that a minimum sample size of 300 was required to both attempt a split and to remain in each terminal node. A *cp* value of .01 was used to split. The tree grown from the training data is given in Figure 4, with corresponding parameter estimates provided in Table 1. Group 1 consisted of those in the bottom three quartiles (<93.1%) on *white*, below the median (<17.2%) on *age65_older*, and in the bottom three quartiles of *median_inc* (<$53,601). It contained 29% of the counties, and was characterized by the highest asymptote, 61.33 cases per 10,000. Group 2 was made up of those in the bottom three quartiles (<93.1%) on *white*, below the median (<17.2%) on *age65_older*, but in the top quartile of *median_inc* (>$53,601). It represented 15% of the counties, and was characterized by the second highest asymptote, 50.19 cases per 10,000. Group 3 contained those in the bottom three quartiles (<93.1%) on *white*, but above the median (>17.2%) on *age65_older*. This group had 31% of counties, with the second lowest asymptote, 35.88 cases per 10,000. Group 4 consisted of those in the top quartile (>93.1%) on *white*, with 26% of counties and the lowest asymptote at 19.04 cases per 10,000. Group 4 also happened to be the most rural and least populated, potentially explaining the low asymptote.

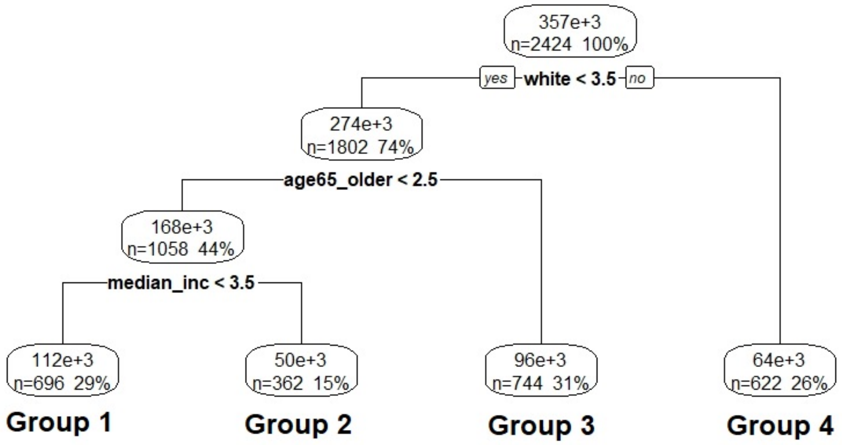


Figure 4. Tree from COVID-19 Data Matching Subsample

Governor's party (with Republican arbitrarily selected as the treatment) was used as the treatment variable in part because much of the policy, coordination, and messaging thus far has occurred via executive action at the state level. The outcome, mobility, was operationalized as the *change in average distance traveled*, or *CADT*. CADT for each day in July was calculated as the county-

Table 1. Parameter Estimates for SEMs from the Groups in Figure 4

	Group 1	Group 2	Group 3	Group 4
n	696	362	744	622
β_1 (Mean)	61.33	50.19	35.88	19.04
β_1 (Variance)	13,377.34	7,767.81	2,425.19	455.58
γ	9.33	6.94	10.23	10.44
α	0.69	0.56	0.44	0.40
σ^2_ϵ	554.05	139.86	88.02	18.73

day level percentage point change in average travel distance relative to that day-of-week's average in early 2020 (average for Feb 10 to March 8, prior to the presence of COVID-19 in the US). Accordingly, a value of –3 indicates a 3 percentage point decline in average travel distance relative to baseline levels. A positive value of CADT signals that residents of that county increased their travel distances relative to their pre-COVID-19 patterns, whereas a negative value indicates reduced travel distances (that can occur through reductions in both the distances traveled per trip as well as the overall number of trips taken). Each county's average CADT for July was estimated by taking the mean of the daily CADT for each day from July 1, 2020 until July 31, 2020. The estimate of the CATE in each group, along with corresponding information, is given in Table 2.

Table 2. CATEs and Significance Tests for COVID-19 Groups

	Group 1	Group 2	Group 3	Group 4
$n_{Rep}; n_{Dem}$	104; 59	58; 55	104; 94	76; 60
$\overline{CADT}_{Rep}$	-0.93%	-4.47%	-0.58%	-0.78%
$\overline{CADT}_{Dem}$	-2.47%	-10.92%	0.88%	-2.30%
CATE	1.54%	6.46%	-1.46%	1.53%
t test	$t(104.08) = 0.94$	$t(108.76) = 2.84$	$t(186.87) =$ -0.86	$t(122.99) = 1.15$
p value	.349	.006	.389	.252

Of the four groups, the only one with a statistically significant CATE was Group 2, where counties in states with Democratic governors had an average CADT that was 6.46 percentage points less than counties in states with Republican governors $t(108.76) =$ -2.84, $p = .006$. Group 2 was on average the most populous, least rural group of the four, as well as the most educated with highest median incomes. As such, Group 2 contained the country's more metropolitan areas. We interpret this result to mean that in metropolitan counties matched for COVID rates, people in counties in states with Democratic governors traveled 6.5 percentage points less in July than people in comparable counties in states with Republican governors. Of note, the two-way independent ANOVA found that in the estimation subsample, a significant main effect of party was not found $F(1, 598) = 3.76$, $p = .053$, whereas a main effect of Group $F(3, 598)$

= 13.45, p < .001, and an interaction $F(3,\ 598) = 3.41$, p = .017 were. This suggests that the party effect is more prominent for more metropolitan counties, but would be obscured if examining the country as a whole. The mean difference between parties in CADT for all 3,030 counties was only 0.60 percentage points, with a t test on the full dataset yielding $t(2422.6)$ = -1.50, p = .133, though this result should be read with the caveat that nearly all counties were represented in the sample. The value of Causal M*plus* Trees in analyzing these data is evident in its ability to find a group of counties exhibiting stronger party effects, while simultaneously matching on COVID-19 trajectories.

Our findings corroborate those of previous COVID-19 partisanship studies. Allcott et al. (2020) found evidence of 3.6 percent fewer point of interest visits associated with a 10 percentage point decrease in the Republican vote share (roughly equivalent to shifting from the median to the 25th percentile Republican vote share county for the 2010 presidential election). Brzezinski et al. (2020) estimated a 3 percentage point difference in the share of devices staying fully at home for the 90th vs 10th percentile Democrat vote share counties 15 days after a county's first case. Areas with relatively greater viewership of conservative news shows that initially downplayed the threat of coronavirus (versus those that accurately portrayed the pandemic) have also been linked to delayed behavior changes and higher initial occurrences of cases and deaths (Bursztyn, Rao, Roth, & Yanagizawa-Drott, 2020). Further, our Group 2 CATE is comparable in magnitude to the decline in travel distance attributable to statewide stay-at-home mandates (Sears et al., 2020). While prior studies employ traditional approaches for discussing treatment effect heterogeneity (i.e. running difference-in-differences or event study regressions on subgroups of interest), the Causal M*plus* Trees method provides a data-driven approach to identifying comparable groups on model fit and analyzing treatment effect heterogeneity.

5 Discussion

In this paper, we proposed the Causal M*plus* Trees algorithm, which matches on parameter estimates of an SEM using a tree-based approach and uses these groupings to estimate CATEs in a holdout sample. We used two small simulation studies to demonstrate a proof of concept for the approach. We also showed how it could be used to estimate party effects on mobility using COVID-19 data. We reiterate that we do not see Causal M*plus* Trees as a substitute for traditional matching methods. Propensity score matching and related methods have their place and can be effective in matching on covariates, both observed and latent. We believe that our approach offers an alternative option to those whose research questions would be better addressed by the ability to match on parameter estimates from an SEM.

5.1 Practical Recommendations

We encourage users of Causal M*plus* Trees to carefully consider how they select and differentiate between modeled covariates and splitting covariates. Although the procedure ultimately matches on both, the way it does so differs by covariate type. Matching is performed on modeled covariates indirectly through the parameter estimates produced by the model, whereas splitting covariates are matched more directly on the observed values of the scores. The choice of whether a covariate should be used as a modeled or splitting covariate depends upon what specifically the user wants to match, which can vary based on the research question, study design, and characteristics of the sample collected.

Another consideration for researchers using Causal M*plus* Trees is the depth to which the tree should be grown. Cross-validation is the most commonly used approach for this in the context of conventional decision trees. However, we believe that cross-validation may not be as well suited for our purposes primarily because it is designed to optimize predictive accuracy. In our algorithm, the goal of the tree is not to optimize predictive accuracy, but rather to partition the sample into groups that are matched well enough on $\hat{\boldsymbol{\theta}}$ to justify causal inference in the holdout sample. As in propensity score matching, there is no objective criterion for this, so the researcher must make a subjective judgment and make a case to justify it.

We urge researchers to take into account the following considerations. First, the sample size in each parent node must be large enough to estimate M in not only the parent node, but also each of the daughter nodes. SEMs can require larger sample sizes to estimate, so limits should be placed on the splitting procedure so as not to consider splitting on a sample that does not have a large enough sample to do this. Related to this is the need for a sufficient number of treated and untreated participants in each terminal node to be able to estimate the CATEs in the holdout sample. If a group has no treated (or no untreated) participants, the CATE cannot be estimated. Of course, it is possible that the mix in the tree differs from the mix in the holdout sample, but to the extent that the matching subsample is a reflection of the estimation subsample, the matching subsample can give a sense of the mix one would expect in the estimation subsample. If performing hypothesis tests, certain minimum sample sizes are required to meet the assumptions of the test as well as to detect the effects, so these must also be kept in mind when deciding how deep to grow the tree.

Parsimony is also important to consider, especially with respect to building a coherent narrative with policy implications. We are typically searching for groups with qualitative meaning given the relevant theoretical framework. If the tree were to produce a dozen groups, it may be challenging to map this onto available theory in order to interpret the results. The relative importance of parameters in characterizing a pattern should be taken into account as well. Theory may dictate that some parameters may be more important to match on than others for a given context (e.g., the residual variance in our stability example). As such, it could be justifiable to trim the tree earlier if splits begin resulting in differences in less relevant parameters. The size of parameter estimates may also play a role. For example, the algorithm could decide on a split that results in two daughter nodes with only small differences in their parameter estimates. Treating these as two separate groups for the purpose of estimating the CATE may not be worthwhile. Similar to the logic used in propensity score analysis, the treated and untreated participants in each node should be compared on their parameters estimates, to verify, even if only subjectively, that they are similar and therefore matched to some degree.

The choice for the depth of the tree depends on a trade-off between interpretability of a result and the validity of the causal inference. If one were to view the ability to draw causal inference as how well treated and untreated participants are matched, then the ability to draw causal inference can be conceptualized not as a dichotomy but as a continuum with perfectly matched participants on one end and perfectly unmatched participants on the other. The better matched participants are, the greater the ability to draw causal inference. However, better matching requires a deeper tree, which becomes less interpretable and generalizable as the depth grows. This trade-off exists in propensity score matching as well but is more apparent in the context of decision trees where such trade-offs are more apparent and a language with which to conceptualize and discuss them already exists.

5.2 Future Research and Conclusions

Plenty of opportunities exist to expand on this work. Although two simulation studies were conducted, they only served as a proof of concept. Additional simulations would be helpful in evaluating the effectiveness of the algorithm across a variety of conditions. The causal tree approach has been extended to use random forests (Wager & Athey, 2018), which are known to be more stable than decision trees. These causal forests have also been modified to accommodate multilevel data structures (Suk, Kang, & Kim, in press). SEM Trees have been expanded to SEM Forests (Brandmaier, Prindle, McArdle, & Lindenberger, 2016), so expanding our algorithm to use random forests would be a natural next step. Additionally, we note that our discussion of treatment effects was limited to mean differences in univariate outcomes. However, given that SEM is already being employed as well as the flexibility of Causal M*plus* Trees, it is possible that the outcome measure could be generalized to the multivariate context, with treated and untreated participants being compared on a model using, for example, a multiple group SEM. To conclude, we believe our proposed algorithm can provide researchers with the opportunity to match on SEM parameter estimates, thereby allowing them greater flexibility in what they can match on and the kinds of research questions they can address as a result.

References

Abrevaya, J., Hsu, Y.-C., & Lieli, R. (2015). Estimating conditional average treatment effects. *Journal of Business and Economic Statistics*, *33*, 485–505. doi: https://doi.org/10.1080/07350015.2014.975555

Adolph, C., Amano, K., Bang-Jensen, B., Fullman, N., & Wilkerson, J. (2020). Pandemic politics: Timing state-level social distancing responses to COVID-19. *medRxiv*. doi: https://doi.org/10.1101/2020.03.30.20046326

Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., & Yang, D. (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics*, *191*. doi: https://doi.org/10.1016/j.jpubeco.2020.104254

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*, 7353–7360. doi: https://doi.org/10.1073/pnas.1510489113

Austin, P. (2009). The relative ability of different propensity-score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, *29*, 661–677. doi: https://doi.org/10.1177/0272989X09341755

Austin, P. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399–424. doi: https://doi.org/10.1080/00273171.2011.568786

Berk, R. (2004). *Regression analysis: A constructive critique.* Sage. doi: https://doi.org/10.4135/9781483348834

Brandmaier, A., Oertzen, T., McArdle, J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*, 71–86. doi: https://doi.org/10.1037/a0030001.

Brandmaier, A., Prindle, J., & Arnold, M. (2021). *semtree: Recursive partitioning for structural equation models* [R package version 0.9.17.]. Retrieved from `https://CRAN.R-project.org/package=semtree`

Brandmaier, A., Prindle, J., McArdle, J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods*, *21*, 566–582. doi: https://doi.org/10.1037/met0000090

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* Chapman & Hall. doi: https://doi.org/10.1201/9781315139470

Browne, M., & Toit, S. (1991). Models for learning data. In L. Collins & J. Horn (Eds.), *Best methods for the analysis of change* (p. 47–68). American Psychological Association. doi: https://doi.org/10.1037/10099-004

Brzezinski, A., Deiana, G., Kecht, V., & Van Dijcke, D. (2020). The covid-19 pandemic: Government vs. community action across the united states. *INET Oxford Working Paper*. Retrieved from `https://www.inet.ox.ac.uk/publications/no-2020-06-the-covid-19-pandemic-government-vs-community-action-across-the-united-states/` (No. 2020-06.)

Bursztyn, L., Rao, A., Roth, C., & Yanagizawa-Drott, D. (2020). Misinformation during a pandemic. *NBER Working Paper*(27417). doi: https://doi.org/10.3386/w27417

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. *Lancet Infectious Disease*, *20*, 533–534. doi: https://doi.org/10.1016/S1473-3099(20)30120-1

Ferrer, E. (2016). Exploratory approaches for studying social interactions, dynamics, and multivariate processes in psychological science. *Multivariate Behavioral Research*, *51*, 240–256. doi:

https://doi.org/10.1080/00273171.2016.1140629

Ferrer, E., Steele, J., & Hsieh, F. (2012). Analyzing dynamics of affective dyadic interactions using patterns of intra- and inter-individual variability. *Multivariate Behavioral Research*, *47*, 136–171. doi: https://doi.org/10.1080/00273171.2012.640605

Gadarian, S., Goodman, S., & Pepinsky, T. (2020). Partisanship, health behavior, and policy attitudes in the early stages of the COVID-19 pandemic. *SSRN*. doi: https://doi.org/10.2139/ssrn.3562796

Greenland, S., Pearl, J., & Robins, J. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, *10*, 37–48. doi: https://doi.org/10.1097/00001648-199901000-00008

Grimm, K., & Ram, N. (2009). Nonlinear growth models in mplus and sas. *Structural Equation Modeling*, *16*, 676–701. doi: https://doi.org/10.1080/10705510903206055

Gu, X., & Rosenbaum, P. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*, 405–420. doi: https://doi.org/10.1080/10618600.1993.10474623

Guo, S., & Fraser, M. (2010). *Propensity score analysis: Statistical methods and applications*. Sage.

Hallquist, M., & Wiley, J. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling*, *25*, 621–638. doi: https://doi.org/10.1080/10705511.2017.1402334

Harder, V., Stuart, E., & Anthony, J. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*, 234–249. doi: https://doi.org/10.1037/a0019623

Hirano, K., & Imbens, G. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, *2*, 259–278. doi: https://doi.org/10.1023/A:1020371312283

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236. doi: https://doi.org/10.1093/pan/mpl013

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–60. doi: https://doi.org/10.1080/01621459.1986.10478354

Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409-426. doi: https://doi.org/10.1007/BF02291366

Lee, B., Lessler, J., & Stuart, E. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*, 337–346. doi: https://doi.org/10.1002/sim.3782

Leite, W., Stapleton, L., & Bettini, E. (2018). Propensity score analysis of complex survey data with structural equation modeling: A tutorial with mplus. *Structural Equation Modeling*, *3*, 448–469. doi:

https://doi.org/10.1080/10705511.2018.1522591

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107–122. doi: https://doi.org/10.1007/bf02294746

Muthén, L., & Muthén, B. (1998-2017). Mplus user's guide (8th ed.) [Computer software manual]. Muthén & Muthén.

National Governors Association. (2020). *Governors roster.* Retrieved from `https://www.nga.org/wp-content/uploads/2019/07/Governors-Roster.pdf`

Neale, M., Hunter, M., Pritikin, J., Zahery, M., Brick, T., Kirkpatrick, R., & Boker, S. (2016). Openmx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*, 535–549. doi: https://doi.org/10.1007/s11336-014-9435-8

Neyman, J., Iwaszkiewicz, K., & Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, *2*, 107–180. doi: https://doi.org/10.2307/2983637

Normand, S., Landrum, M., Guadagnoli, E., Ayanian, J., Ryan, T., Cleary, P., & McNeil, B. (2001). Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, *54*, 387–398. doi: https://doi.org/10.1016/S0895-4356(00)00321-8

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raykov, T. (2012). Propensity score analysis with fallible covariates: A note on a latent variable modeling approach. *Educational and Psychological Measurement*, *72*, 715–733. doi: https://doi.org/10.1177/0013164412440999

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55. doi: https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–24. doi: https://doi.org/10.2307/2288398

Rosenbaum, P., & Rubin, D. (1985). The bias due to incomplete matching. *Biometrics*, *41*, 103–16. doi: https://doi.org/10.2307/2530647

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48(2)*, 1–36. doi: https://doi.org/10.18637/jss.v048.i02

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701. doi: https://doi.org/10.1037/h0037350

Rubin, D. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, *75*, 591–593. doi: https://doi.org/10.2307/2287653

Rubin, D. (1986). What if's have causal answers. *Journal of the American Statistical Association*, *81*, 961–962. doi:

https://doi.org/10.1080/01621459.1986.10478355

Sears, J., Villas-Boas, S., Villas-Boas, M., & Villas-Boas, V. (2020). Are we #stayinghome to flatten the curve? *SSRN*. doi: https://doi.org/10.2139/ssrn.3569791

Serang, S., Jacobucci, R., Stegmann, G., Brandmaier, A., Culianos, D., & Grimm, K. (2021). Mplus Trees: Structural equation model trees using Mplus. *Structural Equation Modeling*, *28*, 127–137. doi: https://doi.org/10.1080/10705511.2020.1726179

Stegmann, G., Jacobucci, R., Serang, S., & Grimm, K. (2018). Recursive partitioning with nonlinear models of change. *Multivariate Behavioral Research*, *53*, 559–570. doi: https://doi.org/10.1080/00273171.2018.1461602

Suk, Y., Kang, H., & Kim, J.-S. (in press). Random forests approach for causal inference with clustered observational data. *Multivariate Behavioral Research*. doi: https://doi.org/10.1080/00273171.2020.1808437

Therneau, T., & Atkinson, B. (2018). *Rpart: Recursive partitioning and regression trees* [R package version 4.1-13.]. Retrieved from `https://CRAN.R-project.org/package=rpart`

Thoemmes, F., & Kim, E. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*, 90–118. doi: https://doi.org/10.1080/00273171.2011.540475

Unacast. (2020). *Unacast social distancing scoreboard dataset.* Retrieved from `https://www.unacast.com/data-for-good.`

U.S. Census Bureau. (2010). *Decennial census, 2010.* Retrieved from `https://data.census.gov/`

U.S. Department of Education, Institute of Education Sciences, & What Works Clearinghouse. (2017). *What works clearinghouse: Standards handbook (version 4.0).* Retrieved from `https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_\handbook_v4.pdf`

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*, 1228–1242. doi: https://doi.org/10.1080/01621459.2017.1319839

West, S., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology*, *82*, 906–919. doi: https://doi.org/10.1037/a0036387

Journal of Behavioral Data Science, 2021, *1* (2), 54–88.
DOI: https://doi.org/10.35566/jbds/v1n2/p4

Two-step growth mixture model to examine heterogeneity in nonlinear trajectories

Jin Liu[1], Le Kang[1], Roy T. Sabo[1], Robert M. Kirkpatrick[2], and Robert A. Perera[1]

[1] Department of Biostatistics, Virginia Commonwealth University, Richmond, VA
[2] Department of Psychiatry, Virginia Commonwealth University, Richmond, VA

Abstract. Empirical researchers are usually interested in investigating the impacts that baseline covariates have when uncovering sample heterogeneity and separating samples into more homogeneous groups. However, a considerable number of studies in the structural equation modeling (SEM) framework usually start with vague hypotheses in terms of heterogeneity and possible causes. It suggests that (1) the determination and specification of a proper model with covariates is not straightforward, and (2) the exploration process may be computationally intensive given that a model in the SEM framework is usually complicated and the pool of candidate covariates is usually huge in the psychological and educational domain where the SEM framework is widely employed. Following Bakk and Kuha (Bakk & Kuha, 2017), this article presents a two-step growth mixture model (GMM) that examines the relationship between latent classes of nonlinear trajectories and baseline characteristics. Our simulation studies demonstrate that the proposed model is capable of clustering the nonlinear change patterns, and estimating the parameters of interest unbiasedly, precisely, as well as exhibiting appropriate confidence interval coverage. Considering the pool of candidate covariates is usually huge and highly correlated, this study also proposes implementing exploratory factor analysis (EFA) to reduce the dimension of covariate space. We illustrate how to use the hybrid method, the two-step GMM and EFA, to efficiently explore the heterogeneity of nonlinear trajectories of longitudinal mathematics achievement data.

Keywords: Growth Mixture Models · Nonlinear Trajectories · Individual Measurement Occasions · Covariates · Simulation Studies · Exploratory Factor Analysis

1 Introduction

1.1 Motivating Example

Earlier studies have examined the impacts of time-invariant covariates (TICs) on nonlinear mathematics achievement trajectories. For example, Liu, Perera, Kang, Kirkpatrick, and Sabo (2019) associated nonlinear change patterns of mathematics IRT scaled scores to baseline covariates, including demographic information, socioeconomics factors, and school information. With the assumption that all covariates explain sample variability directly, this study showed that some baseline characteristics, such as sex, school type, family income, and parents' highest education, can explain the heterogeneity in the nonlinear trajectories of mathematics scores. However, Kohli, Hughes, Wang, Zopluoglu, and Davison (2015) showed that latent classes of change patterns of mathematics achievement exist. Accordingly, these covariates may also inform latent class formation. In this study, we want to investigate the indirect impacts the baseline characteristics have on sample heterogeneity.

1.2 Finite Mixture Model

The finite mixture model (FMM) represents heterogeneity in a sample by allowing for a finite number of latent (unobserved) classes. The idea of mixture models is to put multiple probability distributions together using a linear combination. Although researchers may want to consider two different or multiple different families for the different kernels in some circumstances, the assumption that all latent classes' probability density functions follow normal distributions with class-specific parameters is common in application.

This framework has gained considerable attention in the past twenty years among social and behavioral scientists due to its advantages over other clustering algorithms such as K-means for investigating sample heterogeneity. First, in the SEM framework, the FMM can incorporate any form of within-class models. For instance, Lubke and Muthén (2005) specified factor mixture models, where the within-class model is a factor model to investigate heterogeneity in common factors. In contrast, Muthén and Shedden (1999) defined growth mixture models (GMM), where the within-class model is a latent growth curve model to examine heterogeneity in trajectories. More importantly, the FMM is a model-based clustering method (Bouveyron, Celeux, Murphy, & Raftery, 2019) so that researchers can specify a model in this framework with domain knowledge: which parameters can be fixed to specific values, which need to be estimated, and which can be constrained to be equal (for example, invariance across classes). Additionally, the FMM is a probability-based clustering approach. Unlike other clustering methods, such as the K-means clustering algorithm, which aims to separate all observations into several clusters so that each entry belongs to one cluster without considering uncertainty, the FMM allows each element to belong to multiple classes simultaneously.

This article focuses on the GMM with a nonlinear latent growth curve model as the within-class model. Specifically, trajectories in each class in the proposed GMM is a linear-linear piecewise model (Harring, Cudeck, & du Toit, 2006; Kohli, 2011; Kohli & Harring, 2013; Kohli, Harring, & Hancock, 2013; Kohli et al., 2015; Sterba, 2014), also referred to as a bilinear growth model (Grimm, Ram, & Estabrook, 2016; Liu, 2019; Liu et al., 2019) with an unknown change-point (or knot). We decide to use the bilinear spline functional form for two considerations. First, in addition to examining the growth rate of each stage directly, this piecewise function allows for estimating the transition time from one stage to the other. Additionally, Kohli et al. (2015) and Liu et al. (2019), have shown that a growth model with this functional form can capture the underlying change patterns of mathematics achievement and outperforms several parametric functions: linear, quadratic, and Jenss-Bayley from the statistical perspective.

Similar to Liu et al. (2019), we propose the model in the framework of individual measurement occasions to account for possible heterogeneity in the measurement time in longitudinal studies (Cook & Ware, 1983; Finkel, Reynolds, Mcardle, Gatz, & Pedersen, 2003; Mehta & West, 2000). Earlier studies, for example, (Preacher & Hancock, 2015; Sterba, 2014) have demonstrated one possible solution to individual measurement occasions is to place the exact time to the matrix of factor loadings, termed the definition variable approach (Mehta & Neale, 2005; Mehta & West, 2000). Earlier studies have shown that the definition variable approach outperforms some approximate methods such as the time-bins approach (where the assessment period is divided into several bins, and the factor loadings are set as those time-bins) in terms of bias, efficiency, and Type I error rate (Blozis & Cho, 2008; Coulombe, Selig, & Delaney, 2015).

1.3 Challenges of Finite Mixture Models Implementation

Many studies in the SEM framework start from an exploratory stage where even empirical researchers only have vague assumptions about sample heterogeneity and possible reasons. It suggests that we usually have two challenges when implementing a FMM, deciding the number of latent classes and selecting which covariates need to be included in the model. To investigate which criterion can be used to decide the number of latent classes, Nylund, Asparouhov, and Muthén (2007) evaluated the performance of likelihood-based tests and the traditionally used information criteria and showed that the bootstrap likelihood ratio test is a consistent indicator while the Bayesian information criterion (BIC) performs the best among all information criteria. Note that in practice, the BIC, which is calculated from the estimated likelihood directly, is usually more favorable due to its computational efficiency.

It is also challenging to decide to include which covariates as predictors of class membership. Previous studies have shown that including subject-level predictors for latent classes can be realized by either one-step models (Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997; Clogg, 1981; Dayton & Macready, 1988; Goodman, 1974; Haberman, 1979; Hagenaars, 1993; Kamakura, Wedel, & Agrawal, 1994; Vermunt, 1997; Yamaguchi, 2000), two-step models (Bakk & Kuha, 2017) or three-step models (Asparouhov & Muthén, 2014; Bolck, Croon, & Hagenaars, 2004; Vermunt, 2010). The one-step model is suitable if a study is conducted in a confirmatory way or driven by answering a particular question, where specifying a proper mixture model for the covariates is usually a knowledge-driven process. On the contrary, the stepwise model is more suitable for an exploratory study in which empirical researchers usually have limited *a priori* knowledge about possible class structure. For such studies, the current recommended approach is to investigate the number and nature of the clusters without adding any covariates so that they do not inform class formation.

In this study, we utilize the two-step model given that a considerable number of studies investigated in the SEM framework start from the exploratory stage and that Bakk and Kuha (2017) has shown that the two-step procedure is consistently better than the three-step approach as it does not ignore the presence of uncertainty in the modal class assignments. Accordingly, by extending the method proposed in Bakk and Kuha (2017) to the FMM with a bilinear spline growth curve as the within-class model, we first group nonlinear trajectories and estimate class-specific parameters with a pre-specified number of clusters by fitting the measurement-model portion of the mixture model; we then investigate the associations between the 'soft clusters', where each sample is assigned with different posterior weights, and the individual-level covariates by fitting the measurement and structural model but fixing the measurement parameter estimates as their values from the first step. By utilizing the two-step model, we only need to refit the model in the second step rather than the whole model when adding or removing covariates, saving the computational budget.

However, the covariate space in the psychological and educational domains where the SEM framework is widely utilized is usually large, and some covariates are highly correlated. To address this issue, we propose to leverage a common multivariate data analysis approach in the SEM framework, exploratory factor analysis (EFA), to reduce the covariate space's dimension and address potential multicollinearity. Note that in this current study, it is not our aim to examine EFA comprehensively. We only want to demonstrate how to use the individual scores, for example, Thompson's scores (Thomson, 1939), or Bartlett's weighted least-squares scores (Bartlett, 1937), based on the output of EFA, with a basic understanding of its algorithm.

EFA is a useful multivariate data analysis approach to explain the variance-covariance matrix of the dataset by replacing a large set of manifest variables with a smaller latent variable set. In this approach, manifested variables are assumed to be caused by latent variables. When implementing EFA, we impose no constraints on the relationships between manifested and latent variables. Assuming that all manifested variables are related to all latent variables, this approach aims to determine the appropriate number of factors and factor loadings (i.e., correlations between observed variables and unobserved variables). Next, we calculate a score for each factor of each individual based on the factor loadings and standardized covariate values. We then view these individual-level scores instead of the covariates as baseline characteristics in the second step.

The proposed hybrid method aims to provide an analytical framework for examining heterogeneity in an exploratory study. We extend the two-step method proposed by Bakk and Kuha (2017) to investigate the heterogeneity in nonlinear trajectories in the framework of individually varying time points (ITPs). Specifically, we consider the bilinear spline growth curve with an unknown knot as the within-class model. We specify the model with truly individual measurement occasions, which are ubiquity in longitudinal studies, to avoid unnecessary inadmissible estimation. Additionally, we propose to use EFA to reduce the dimension of the covariate space.

The remainder of this article is organized as follows. We describe the model specification and model estimation of the two-step growth mixture model in the framework of ITPs in the method section. In the subsequent section, we describe the design of the Monte Carlo simulation for model evaluation. We evaluate the model performance through the performance measures, which include the relative bias, the empirical standard error (SE), the relative root-mean-squared-error (RMSE), and the empirical coverage for a nominal 95% confidence interval of each parameter of interest, as well as accuracy. We then introduce the dataset of repeated mathematics achievement scores from the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K: 2011), and demonstrate the implementation of the hybrid method in the application section. Finally, discussions are framed concerning methodological considerations and future directions.

2 Method

2.1 Model Specification

In this section, we specify the GMM with a bilinear spline growth curve as the within-class model. Harring et al. (2006) showed there are five parameters in the bilinear spline functional form: an intercept and slope of each linear piece and a change-point, yet the degree of freedom of the bilinear spline is four since two linear pieces join at the knot. In this study, we view the initial status, two slopes, and the knot as the four parameters. We construct the model with consideration of the variability of the initial status and two slopes, but assuming that the class-specific knot is the same across all individuals in a latent class though Liu et al. (2019); Preacher and Hancock (2015) have shown that the knot can also have a random effect by relaxing the assumption. Suppose the pre-specified number of latent classes is K, for $i = 1$ to n individuals and $k = 1$ to K latent classes, we express the model as

$$p(\boldsymbol{y}_i|z_i = k, \boldsymbol{x}_i) = \sum_{k=1}^{K} \pi(z_i = k|\boldsymbol{x}_i) \times p(\boldsymbol{y}_i|z_i = k), \tag{1}$$

$$\pi(z_i = k|\boldsymbol{x}_i) = \begin{cases} \frac{1}{1+\sum_{k=2}^{K}\exp(\beta_0^{(k)}+\boldsymbol{\beta}^{(k)T}\boldsymbol{x}_i)} & \text{Reference Group } (k=1) \\ \frac{\exp(\beta_0^{(k)}+\boldsymbol{\beta}^{(k)T}\boldsymbol{x}_i)}{1+\sum_{k=2}^{K}\exp(\beta_0^{(k)}+\boldsymbol{\beta}^{(k)T}\boldsymbol{x}_i)} & \text{Other Groups } (k=2,\dots,K) \end{cases}, \tag{2}$$

$$\boldsymbol{y}_i|(z_i = k) = \boldsymbol{\Lambda}_i(\gamma^{(k)})\boldsymbol{\eta}_i|(z_i = k) + \boldsymbol{\epsilon}_i|(z_i = k), \tag{3}$$

$$\boldsymbol{\eta}_i|(z_i = k) = \boldsymbol{\mu_\eta}^{(k)} + \boldsymbol{\zeta}_i|(z_i = k). \tag{4}$$

Equation (1) defines a FMM that combines mixing proportions, $\pi(z_i = k|\boldsymbol{x}_i)$, and within-class models, $p(\boldsymbol{y}_i|z_i = k)$, where $\boldsymbol{x}_i$, $\boldsymbol{y}_i$ and z_i are the covariates, $J \times 1$ vector of repeated outcome (where J is the number of measurements) and membership of the i^{th} individual, respectively. For Equation (1), we have two constratints: $0 \leq \pi(z_i = k|\boldsymbol{x}_i) \leq 1$ and $\sum_{k=1}^{K} \pi(z_i = k|\boldsymbol{x}_i) = 1$. Equation (2) defines mixing components as logistic functions of covariates $\boldsymbol{x}_i$, where $\beta_0^{(k)}$ and $\boldsymbol{\beta}^{(k)}$ are the class-specific logistic coefficients. These functions decide the membership for the i^{th} individual, depending on the values of the covariates $\boldsymbol{x}_i$.

Equations (3) and (4) together define a within-class model. Similar to all factor models, Equation (3) expresses the outcome $\boldsymbol{y}_i$ as a linear combination of growth factors. When the underlying functional form is bilinear spline growth curve with an unknown fixed knot, $\boldsymbol{\eta}_i$ is a 3×1 vector of growth factors ($\boldsymbol{\eta}_i = \eta_{0i}, \eta_{1i}, \eta_{2i}$, for an initial status and a slope of each stage of the i^{th} individual). Accordingly, $\boldsymbol{\Lambda}_i(\gamma^{(k)})$, which is a function of the class-specific knot $\gamma^{(k)}$, is a $J \times 3$ matrix of factor loadings. Note that the subscript i in $\boldsymbol{\Lambda}_i(\gamma^{(k)})$ indicates that it is a function of the individual measurement occasions of the i^{th} individual. The pre- and post-knot $\boldsymbol{y}_i$ can be expressed as

$$y_{ij} = \begin{cases} \eta_{0i} + \eta_{1i} t_{ij} + \epsilon_{ij} & t_{ij} \leq \gamma^{(k)} \\ \eta_{0i} + \eta_{1i}\gamma^{(k)} + \eta_{2i}(t_{ij} - \gamma^{(k)}) + \epsilon_{ij} & t_{ij} > \gamma^{(k)} \end{cases},$$

where y_{ij} and t_{ij} are the measurement and measurement occasion of the i^{th} individual at time j. Additionally, $\boldsymbol{\epsilon}_i$ is a $J \times 1$ vector of residuals of the i^{th} individual. Equation (4) further expresses the growth factors as deviations from their class-specific means. In the equation, $\boldsymbol{\mu_\eta}^{(k)}$ is a 3×1 vector of class-specific growth factor means and $\boldsymbol{\zeta}_i$ is a 3×1 vector of residual deviations from the mean vector of the i^{th} individual.

To unify pre- and post-knot expressions, we need to reparameterize growth factors. Earlier studies, for example, Grimm et al. (2016); Harring et al. (2006); Liu et al. (2019), presented multiple ways to realize this aim. Note that no matter which approach we follow to reparameterize growth factors, the reparameterized coefficients are not directly related to the underlying change patterns and need to be transformed back to be interpretable. In this article, we follow the reparameterized method in Liu et al. (2019) and define the class-specific reparameterized growth factors as the measurement at the knot, mean of two slopes, and the half difference of two slopes. Note that the expressions of the repeated outcome $\boldsymbol{y}_i$ using the growth factors in the original and reparameterized frames are equivalent. We also extend the (inverse-)transformation functions and matrices for the reduced model in Liu et al. (2019), with which we can obtain the original parameters efficiently for interpretation purposes. Detailed class-specific reparameterizing process and the class-specific (inverse-) transformation are provided in Appendix 6.2 and Appendix 6.2, respectively.

2.2 Model Estimation

To simplify the model, we assume that class-specific growth factors follow a multivariate Gaussian distribution, that is, $\boldsymbol{\zeta}_i|k \sim \text{MVN}(\boldsymbol{0}, \boldsymbol{\Psi_\eta}^{(k)})$. Note that $\boldsymbol{\Psi_\eta}^{(k)}$ is a 3×3 variance-covariance matrix of class-specific growth factors. We also assume that individual residuals follow identical and independent normal distributions over time in each latent class, that is, $\boldsymbol{\epsilon}_i|k \sim N(\mathbf{0}, \theta_\epsilon^{(k)}\boldsymbol{I})$, where $\boldsymbol{I}$ is a $J \times J$ identity matrix. Accordingly, for the i^{th} individual in the k^{th} unobserved group, the within-class model implied mean vector ($\boldsymbol{\mu}_i^{(k)}$) and variance-covariance matrix ($\boldsymbol{\Sigma}_i^{(k)}$) of repeated measurements are

$$\begin{aligned}\boldsymbol{\mu}_i^{(k)} &= \boldsymbol{\Lambda}_i \boldsymbol{\mu_\eta}^{(k)}, &(5)\\ \boldsymbol{\Sigma}_i^{(k)} &= \boldsymbol{\Lambda}_i \boldsymbol{\Psi_\eta}^{(k)} \boldsymbol{\Lambda}_i^T + \theta_\epsilon^{(k)}\boldsymbol{I}. &(6)\end{aligned}$$

Step 1 In the first step, we estimate the class-specific parameters and mixing proportions for the model specified in Equations (1), (2), (3) and (4) without considering the impact that covariates $\boldsymbol{x}_i$ have on the class formation. The parameters need to be estimated in this step include

$$\boldsymbol{\Theta}_{s1} = \{\mu_{\eta_0}^{(k)}, \mu_{\eta_1}^{(k)}, \mu_{\eta_2}^{(k)}, \gamma^{(k)}, \psi_{00}^{(k)}, \psi_{01}^{(k)}, \psi_{02}^{(k)}, \psi_{11}^{(k)}, \psi_{12}^{(k)}, \psi_{22}^{(k)}, \theta_\epsilon^{(k)}, \pi^{(2)}, \cdots, \pi^{(K)}\}.$$

We employ full information maximum likelihood (FIML) technique, which accounts for the potential heterogeneity of individual contributions to the likelihood, to estimate $\boldsymbol{\Theta}_{s1}$. The log-likelihood function of the model specified in Equations (1), (2), (3) and (4) without the effect of $\boldsymbol{x}_i$ is

$$\begin{aligned}\log lik(\boldsymbol{\Theta}_{\mathrm{s1}}) &= \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi(z_i = k) p(\boldsymbol{y}_i | z_i = k)\right)\\ &= \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi(z_i = k) p(\boldsymbol{y}_i | \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)})\right).\end{aligned} \tag{7}$$

Step 2 In the second step, we examine the associations between the ‘soft clusters’, where each trajectory is assigned with different posterior probabilities, and the baseline characteristics by fixing the class-specific parameters as their estimates from the first step, that is, the parameters need to be estimated in this step are those logistic coefficients, $\boldsymbol{\Theta}_{\mathrm{s2}} = \{\beta_0^{(k)}, \boldsymbol{\beta}^{T(k)}\}$ ($k = 2, \ldots, K$), in Equation (2). The log-likelihood function in Equation (7) also needs to be modified as

$$\begin{aligned}\log lik(\boldsymbol{\Theta}_{\mathrm{s2}}) &= \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi(z_i = k|\boldsymbol{x}_i) p(\boldsymbol{y}_i | z_i = k)\right)\\ &= \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi(z_i = k|\boldsymbol{x}_i) p(\boldsymbol{y}_i | \hat{\boldsymbol{\mu}}_i^{(k)}, \hat{\boldsymbol{\Sigma}}_i^{(k)})\right).\end{aligned} \tag{8}$$

We construct the proposed two-step GMM using the R package *OpenMx* with the optimizer *CSOLNP* (Boker et al., 2020; Hunter, 2018; Neale et al., 2016; Pritikin, Hunter, & Boker, 2015), with which we can fit the proposed GMM and implement the class-specific inverse-transformation matrices to obtain coefficients that are directly related to underlying change patterns as shown in

Appendix 6.2. In the online appendix (`https://github.com/Veronica0206/Dissertation_projects`), we provide the *OpenMx* code for the proposed model as well as a demonstration. For the researchers interested in using *Mplus*, we also provide *Mplus* 8 code for the model in the online appendix.

3 Model Evaluation

We evaluate the proposed model using a Monte Carlo simulation study with two goals. The first goal is to evaluate the model performance by examining the relative bias, empirical SE, relative RMSE, and empirical coverage for a nominal 95% confidence interval (CI) of each parameter. Table 1 lists the definitions and estimates of these performance metrics.

Table 1. Performance Metrics: Definitions and Estimates

Criteria	**Definition**	**Estimate**
Relative Bias	$E_{\hat{\theta}}(\hat{\theta}-\theta)/\theta$	$\sum_{s=1}^{S}(\hat{\theta}_s-\theta)/S\theta$
Empirical SE	$\sqrt{Var(\hat{\theta})}$	$\sqrt{\sum_{s=1}^{S}(\hat{\theta}_s-\bar{\theta})^2/(S-1)}$
Relative RMSE	$\sqrt{E_{\hat{\theta}}(\hat{\theta}-\theta)^2}/\theta$	$\sqrt{\sum_{s=1}^{S}(\hat{\theta}_s-\theta)^2/S}/\theta$
Coverage Probability	$Pr(\hat{\theta}_{\text{low}}\le\theta\le\hat{\theta}_{\text{upper}})$	$\sum_{s=1}^{S}I(\hat{\theta}_{\text{low},s}\le\theta\le\hat{\theta}_{\text{upper},s})/S$

Note. θ: the population value of the parameter of interest; $\hat{\theta}$: the estimate of θ; S: the number of replications and set as $1,000$ in our simulation study; $s=1,\dots,S$: indexes the replications of the simulation; $\hat{\theta}_s$: the estimate of θ from the s^{th} replication; $\bar{\theta}$: the mean of $\hat{\theta}_s$'s across replications; $I()$: an indicator function

The second goal is to evaluate how well the clustering algorithm performs to separate the heterogeneous trajectories. To evaluate the clustering effects, we need to calculate the posterior probabilities for each individual belonging to the k^{th} unobserved group. The calculation is based on the class-specific estimates and mixing proportions obtained from the first step and realized by Bayes' theorem

$$p(z_i=k|\boldsymbol{y}_i)=\frac{\pi(z_i=k)p(\boldsymbol{y}_i|z_i=k)}{\sum_{k=1}^{K}\pi(z_i=k)p(\boldsymbol{y}_i|z_i=k)}.$$

We then assign each individual to the latent class with the highest posterior probability to which that observation most likely belongs. If multiple posterior probabilities equal to the maximum value, we break the tie among competing components randomly (McLachlan & Peel, 2000). We evaluate the clustering effects by accuracy and entropy. Since the true membership is available in simulation studies, we are able to calculate accuracy, which is defined as the fraction of all correctly labeled instances (Bishop, 2006). Entropy, which is given

$$\text{Entropy}=1+\frac{1}{n\log(K)}\bigg(\sum_{n=1}^{n}\sum_{k=1}^{K}p(z_i=k|\boldsymbol{y}_i)\log p(z_i=k|\boldsymbol{y}_i)\bigg),\qquad(9)$$

is a metric based on the average posterior probabilities (Stegmann & Grimm, 2018). It ranges from 0 to 1, where 0 and 1 suggesting no cluster separation and complete separation, respectively. It is an indicator of the quality of the mixture model. In the current study, entropy reflects separation only based on the trajectories as shown in Equation (9). Earlier studies, for example, Lubke and Muthén (2007), have demonstrated that entropy is a good indicator of accuracy when we exclude all covariates from the mixture model. It is our interest to test the robustness of this recommendation in the context of the growth mixture model with nonlinear trajectories.

We decided the number of repetitions $S = 1,000$ by an empirical approach proposed by Morris, White, and Crowther (2019) in the simulation design. The (relative) bias is the most important performance metric in our simulation, so we want to keep its Monte Carlo standard error[3] less than 0.005. We ran a pilot simulation study and noted that standard errors of all parameters except the intercept variances were less than 0.15, so we needed at least 900 replications to ensure the Monte Carlo standard error of bias is as low as we expected. We then decided to proceed with $S = 1,000$ to be more conservative.

3.1 Design of Simulation Study

The simulation study has two parts. As mentioned earlier, we propose the two-step model with a bilinear spline growth curve with an unknown knot as the within-class model, assuming that the change-point is roughly similar for all individuals in each latent class as the knot variance is not the primary interest of this study. In the first part, we restricted the knot to be identical for all trajectories in a latent class to evaluate the model performance when being specified correctly. We are also interested in examining how the proposed model works when relaxing the restriction. Accordingly, in the second part, by allowing for the individual difference in the knot, we investigated the robustness of the proposed model by assessing the model performance in the presence of knots with the standard deviation set as 0.3.

We list all conditions of simulation studies for Part 1 and Part 2 in Table 2. All conditions except the knot variance for both parts were set to be the same. For both parts, we fixed the conditions that are not of the primary interests of the current study. For example, we considered ten scaled and equally-spaced waves since Liu et al. (2019) has shown that the bilinear growth model had decent performance concerning the performance measures to a longitudinal data set with ten repeated measures and fewer number of measurements only affected model performance slightly. Similar to Liu et al. (2019), we allowed the time-window of individual measurement occasions ranging from −0.25 and +0.25, which was viewed as a 'medium' deviation, as an existing simulation study (Coulombe et al., 2015), around each wave. We also fixed the variance-covariance matrix of the class-specific growth factors that usually change with the time scale and the measurement scale in practice; accordingly, we kept the index of dispersion

[3] Monte Carlo SE(Bias) $= \sqrt{Var(\hat{\theta})/S}$ (Morris et al., 2019).

(σ^2/μ) of each growth factor at the one-tenth scale, guided by Bauer and Curran (2003); Kohli (2011); Kohli et al. (2015). Further, the growth factors were set to be positively correlated to a moderate degree ($\rho = 0.3$).

For both parts, the primary aim was to investigate how the separation between latent classes, the unbalanced class mixing proportion, and the trajectory shape affected the model performance. Utilizing a model-based clustering algorithm, we are usually interested in examining how well the model can detect heterogeneity in samples and estimate parameters of interest in each latent class. Intuitively, the model should perform better under those conditions with a larger separation between latent classes. We wanted to test this hypothesis. In the simulation design, we had two metrics to gauge the separation between clusters: the difference between the knot locations and the Mahalanobis distance (MD) of the three growth factors of latent classes. We set 1, 1.5 and 2 as a small, medium, and large difference between the knot locations. We chose 1 as the level of small difference to follow the rationale in Kohli et al. (2015) and considered the other two levels to investigate whether the more widely spaced knots improve the model performance. We considered two levels of MD, 0.86 (i.e., small distance) and 1.72 (i.e., large distance), for class separation. Note that both the small and large distance in the current simulation design was smaller than the corresponding level in Kohli et al. (2015) because we wanted to examine the proposed model under more challenging conditions in terms of cluster separation.

We chose two levels of mixing proportion, 1:1 and 1:2, for the conditions with two latent classes and three levels of mixing proportion, 1:1:1, 1:1:2 and 1:2:2, for the scenarios with three clusters. We selected these levels because we wanted to evaluate how the challenging conditions (i.e., the unbalanced allocation) affect performance measures and clustering effects. We also examined several common change patterns shown in Table 2 (Scenario 1, 2 and 3). We changed the knot locations and one growth factor under each scenario but fixed the other two growth factors to satisfy the specified MD. We considered $\theta = 1$ or $\theta = 2$ as two levels of homogeneous residual variances across latent classes to see the effect of the measurement precision, and we considered two levels of sample size.

3.2 Label Switching

All mixture models suffer from the label switching issue: inconsistent assignments of membership for multiple replications in simulation studies. The label switching does not hurt the model estimation in the frequentist framework since the likelihood is invariant to permutation of cluster labels; however, the estimates from the first latent class may be mislabeled as such from other latent classes (Class 2 or Class 3 in our case) (Tueller, Drotar, & Lubke, 2011). In this study, we utilized the column maxima switched label detection algorithm developed by Tueller et al. (2011) to check whether the labels were switched; and if it occurred, the final estimates were relabeled in the correct order before model evaluation.

Table 2. Simulation Design for the Proposed Two-step Growth Mixture Model

Fixed Conditions		
Variables	**Conditions**	
Variance of Intercept	$\psi_{00}^{(k)} = 25$	
Variance of Slopes	$\psi_{11}^{(k)} = \psi_{22}^{(k)} = 1$	
Correlations of GFs	$\rho^{(k)} = 0.3$	
Time (t)	10 scaled and equally spaced $t_j (j = 0, \cdots, J-1, J = 10)$	
Individual t	$t_{ij} \sim U(t_j - \Delta, t_j + \Delta)(j = 0, \cdots, J-1; \Delta = 0.25)$	
Manipulated Conditions		
Variables	**2 latent classes**	**3 latent classes**
Sample Size	$n = 500$ or 1000	$n = 500$ or 1000
Variance of Knots	$\psi_{\gamma\gamma}^{(k)} = 0.00 (k = 1, 2)$ $\psi_{\gamma\gamma}^{(k)} = 0.09 (k = 1, 2)$	$\psi_{\gamma\gamma}^{(k)} = 0.00 (k = 1, 2, 3)$ $\psi_{\gamma\gamma}^{(k)} = 0.09 (k = 1, 2, 3)$
Ratio of Proportions	$\pi^{(1)} : \pi^{(2)} = 1 : 1$ $\pi^{(1)} : \pi^{(2)} = 1 : 2$	$\pi^{(1)} : \pi^{(2)} : \pi^{(3)} = 1 : 1 : 1$ $\pi^{(1)} : \pi^{(2)} : \pi^{(3)} = 1 : 1 : 2$ $\pi^{(1)} : \pi^{(2)} : \pi^{(3)} = 1 : 2 : 2$
Residual Variance	$\theta_{\epsilon}^{(k)} = 1$ or 2	$\theta_{\epsilon}^{(k)} = 1$ or 2
Locations of knots	$\mu_{\gamma} = (4.00, 5.00)$ $\mu_{\gamma} = (3.75, 5.25)$ $\mu_{\gamma} = (3.50, 5.50)$	$\mu_{\gamma} = (3.50, 4.50, 5.50)$ $\mu_{\gamma} = (3.00, 4.50, 6.00)$
Mahalanobis distance	$d = 0.86$ or 1.72	$d = 0.86$
Scenario 1: Different means of initial status and (means of) knot locations		
Variables	**2 latent classes**	**3 latent classes**
Means of Slope 1's	$\mu_{\eta_1}^{(k)} = -5\ (k = 1, 2)$	$\mu_{\eta_1}^{(k)} = -5\ (k = 1, 2, 3)$
Means of Slope 2's	$\mu_{\eta_2}^{(k)} = -2.6\ (k = 1, 2)$	$\mu_{\eta_2}^{(k)} = -2.6\ (k = 1, 2, 3)$
Means of Intercepts	$\mu_{\eta_0} = (98, 102), (d = 0.86)$ $\mu_{\eta_0} = (96, 104), (d = 1.72)$	$\mu_{\eta_0} = (96, 100, 104)$
Scenario 2: Different means of slope 1 and (means of) knot locations		
Variables	**2 latent classes**	**3 latent classes**
Means of Intercepts	$\mu_{\eta_0}^{(k)} = 100\ (k = 1, 2)$	$\mu_{\eta_0}^{(k)} = 100\ (k = 1, 2, 3)$
Means of Slope 2's	$\mu_{\eta_2}^{(k)} = -2\ (k = 1, 2)$	$\mu_{\eta_2}^{(k)} = -2\ (k = 1, 2, 3)$
Means of Slope 1's	$\mu_{\eta_1} = (-4.4, -3.6), (d = 0.86)$ $\mu_{\eta_1} = (-5.2, -3.6), (d = 1.72)$	$\mu_{\eta_1} = (-5.2, -4.4, -3.6)$
Scenario 3: Different means of slope 2 and (means of) knot locations		
Variables	**2 latent classes**	**3 latent classes**
Means of Intercepts	$\mu_{\eta_0}^{(k)} = 100\ (k = 1, 2)$	$\mu_{\eta_0}^{(k)} = 100\ (k = 1, 2, 3)$
Means of Slope 1's	$\mu_{\eta_1}^{(k)} = -5\ (k = 1, 2)$	$\mu_{\eta_1}^{(k)} = -5\ (k = 1, 2, 3)$
Means of Slope 2's	$\mu_{\eta_2} = (-2.6, -3.4), (d = 0.86)$ $\mu_{\eta_2} = (-1.8, -3.4), (d = 1.72)$	$\mu_{\eta_2} = (-1.8, -2.6, -3.4)$

3.3 Data Generation and Simulation Step

For each condition listed in Table 2, we used two-step data generation to obtain a component label z_i for each individual and then generated data for each component. The general steps of the simulation for the proposed two-step model in the framework of individual measurement occasions were carried out as follows:

1. Create component label z_i for the i^{th} individual:
   (a) Generate data matrix of exogenous variables,
   (b) Calculate the probability vector for each entry with a set of specified regression coefficients using a multinomial logit link and assign a component label z_i to each observation,
2. Generate data for growth factors and a knot of each latent class using the R package *MASS* (Venables & Ripley, 2002),
3. Generate the time structure with J scaled and equally-spaced waves t_j and obtain individual measurement occasions: $t_{ij} \sim U(t_j - \Delta, t_j + \Delta)$ by allowing disturbances around each wave,
4. Calculate factor loadings, which are functions of ITPs and the knot, for each individual,
5. Calculate values of the repeated measurements based on the class-specific growth factors, corresponding factor loadings, and residual variances,
6. Apply the proposed model to the generated data set, estimate the parameters, and construct corresponding 95% Wald CIs, as well as calculate posterior probabilities that each individual belongs to each of the multiple latent classes, followed by accuracy and entropy,
7. Repeat the above steps until after obtaining $1,000$ convergent solutions to calculate the mean accuracy and mean entropy, perform the column maxima switched label detection algorithm, relabel the clusters if labels had been switched, and calculate the relative bias, empirical SE, relative RMSE and coverage probability of each parameter under investigation.

4 Result

4.1 Model Convergence

In this section, we first examine the convergence[4] rate of two steps for each condition. Based on our simulation studies, the convergence rate of the proposed two-step model achieved around 90% for all conditions, and the majority of non-convergence cases occurred in the first step. To elaborate, for the conditions with two latent classes, 96 out of total 288 conditions reported 100% convergence rate, while for the conditions with three latent classes, 12 out of total 144 conditions reported 100% convergence rate. Among all conditions with two latent classes, the worst scenario regarding the convergence rate was 121/1121, indicating that

[4] In our project, convergence is defined as to reach *OpenMx* status code 0, which indicates a successful optimization, until up to 10 attempts with different collections of starting values (Neale et al., 2016).

we need to replicate the procedure described in Section 3.3 $1,121$ times to have $1,000$ replications with a convergent solution. Across all scenarios with three latent classes, the worst condition was $134/1134$[5].

4.2 Performance Measures

Performance Measures of the First Part of Simulation Study In this section, we evaluate the performance measures of the proposed model across the conditions with fixed knots (i.e., knots without considering variability), under which the proposed model was specified correctly. In the result section, we named the latent classes from left to right as Class 1 (the left cluster) and Class 2 (the right cluster) and called them as Class 1 (the left cluster), Class 2 (the middle cluster) and Class 3 (the right cluster) for the model with two and three pre-specified clusters, respectively. We first calculated each performance metric across $1,000$ replications for each parameter of interest under each condition with two latent classes and fixed knots. We then summarized each metric across all conditions as the corresponding median and range.

Tables 3 and 4 present the median (range) of the relative bias and empirical SE for each parameter of interest of the two-step model, respectively. We observed that the proposed model generated unbiased point estimates with small empirical SEs when being specified correctly in the first step. Specifically, the magnitude of the relative biases of the growth factor means and growth factor variances across all conditions were under 0.016 and 0.038, respectively. In the second step, the median of relative bias of the logistic coefficients was around -0.010, although they may be underestimated under conditions with the small sample size (i.e., $n = 500$), the small difference in knot locations (i.e., the difference is 1) and less precise measurements (i.e., $\theta_{\epsilon} = 2$). From Table 4, the magnitude of empirical SE of all parameters except intercept means and variances were under 0.52 (i.e., the variances of estimates were under 0.25), though the median value of empirical SE of μ_{η_0} and ψ_{00} were around 0.40 and 2.50, respectively.

Table 5 list the median (range) of relative RMSE of each parameter, which assesses the point estimates holistically. From the table, the model was capable of estimating the parameters accurately in the first step. Under the conditions with two latent classes and fixed knots, the magnitude of the relative RMSEs of the growth factor means and variances were under 0.081 and 0.296, respectively. The relative RMSE of the logistic coefficients was relatively larger under some conditions due to their larger relative biases.

Table 6 shows the median (range) of the coverage probability for each parameter of interest of the two-step model with two latent classes under conditions with fixed knots. Overall, the proposed model performed well regarding empirical coverage under the conditions with the relatively large separation between two

[5] Conditions of these worst cases were the small sample size ($n = 500$), unbalanced allocation rate, small residual variance, small distance between the latent classes, and small or medium difference in the knot locations.

Table 3. Median (Range) of the Relative Bias over $1,000$ Replications of Parameters of Interest under the Conditions with Fixed Knots and 2 Latent Classes

	Parameters	Latent Class 1	Latent Class 2
Mean	μ_{η_0}	0.000 (0.000, 0.001)	0.000 (−0.001, 0.000)
	μ_{η_1}	0.000 (−0.008, 0.003)	0.001 (−0.001, 0.012)
	μ_{η_2}	0.000 (−0.009, 0.016)	−0.002 (−0.012, 0.003)
	μ_{γ}	0.000 (−0.001, 0.002)	0.000 (−0.001, 0.002)
Variance	ψ_{00}	−0.002 (−0.014, 0.006)	−0.005 (−0.031, 0.005)
	ψ_{11}	−0.005 (−0.028, 0.028)	−0.007 (−0.038, 0.003)
	ψ_{22}	−0.005 (−0.026, 0.031)	−0.007 (−0.037, 0.005)
Path Coef.	β_0	—	−0.009 (NA, NA)
	β_1	—	−0.012 (−0.225, 0.018)
	β_2	—	−0.010 (−0.218, 0.015)

Note. —: when fitting the proposed model, we set the first latent class as the reference group; accordingly, the coefficients of that class do not exist.
NA: Note that for the conditions with balanced allocation, the population value of $\beta_0 = 0$ and its relative bias goes infinity. The bias median (range) of β_0 is -0.002 (-0.070, 0.017).

Table 4. Median (Range) of the Empirical SE over $1,000$ Replications of Parameters of Interest under the Conditions with Fixed Knots and 2 Latent Classes

	Parameters	Latent Class 1	Latent Class 2
Mean	μ_{η_0}	0.422 (0.242, 0.933)	0.336 (0.198, 0.709)
	μ_{η_1}	0.101 (0.051, 0.276)	0.073 (0.042, 0.175)
	μ_{η_2}	0.100 (0.054, 0.276)	0.072 (0.042, 0.160)
	μ_{γ}	0.039 (0.017, 0.110)	0.046 (0.020, 0.134)
Variance	ψ_{00}	2.662 (1.692, 5.073)	2.173 (1.423, 3.942)
	ψ_{11}	0.124 (0.073, 0.296)	0.093 (0.059, 0.168)
	ψ_{22}	0.126 (0.072, 0.286)	0.095 (0.062, 0.178)
Path Coef.	β_0	—	0.168 (0.083, 0.516)
	β_1	—	0.120 (0.080, 0.200)
	β_2	—	0.124 (0.082, 0.198)

Note. —: when fitting the proposed model, we set the first latent class as the reference group; accordingly, the coefficients of that class do not exist.

Table 5. Median (Range) of the Relative RMSE over 1,000 Replications of Parameters of Interest under the Conditions with Fixed Knots and 2 Latent Classes

	Para.	**Latent Class 1**	**Latent Class 2**
Mean	μ_{η_0}	0.004 (0.002, 0.009)	0.003 (0.002, 0.007)
	μ_{η_1}	−0.021 (−0.063, −0.010)	−0.016 (−0.045, −0.009)
	μ_{η_2}	−0.045 (−0.112, −0.020)	−0.028 (−0.081, −0.012)
	μ_{γ}	0.010 (0.005, 0.028)	0.009 (0.004, 0.027)
Variance	ψ_{00}	0.106 (0.068, 0.203)	0.087 (0.057, 0.161)
	ψ_{11}	0.124 (0.074, 0.296)	0.093 (0.060, 0.172)
	ψ_{22}	0.126 (0.072, 0.288)	0.095 (0.062, 0.182)
Path Coef.	β_0	—	NA (0.121, NA)
	β_1	—	0.297 (0.197, 0.542)
	β_2	—	0.234 (0.155, 0.431)

Note. Para.: Parameters. —: when fitting the proposed model, we set the first latent class as the reference group; accordingly, the coefficients of that class do not exist. NA: Note that for the conditions with balanced allocation, the population value of $\beta_0 = 0$ and its relative RMSE goes infinity. The RMSE median (range) of β_0 is 0.168 (0.083, 0.521).

latent classes and the higher measurement precision. Specifically, coverage probability of all parameters except knots and intercept coefficient β_0 can achieve at least 90% across all conditions with a medium or large separation between the knot locations (i.e., 1.5 or 2) and small residual variance (i.e., $\theta_\epsilon = 1$).

Additionally, when being specified correctly, the model with three latent classes, similar to that with two clusters, performed well in terms of performance measures, though we noticed that the empirical SE of parameters in the middle cluster were slightly larger than those in the other two groups.

Performance Measures of the Second Part of Simulation Study In this section, we assess the robustness of the proposed model by examining the performance measures in the presence of random knots (i.e., the knots with the standard deviation set as 0.3), under which the model was underspecified. We noted that the relative biases increased slightly and that the empirical SE did not change meaningfully when the proposed model was misspecified, which decreased the performance of relative RMSE and coverage probability. For those conditions under which the model was underspecified, the summary of the relative bias and empirical SE were provided in 6.2.

4.3 Accuracy and Entropy

In this section, we evaluate the clustering effects across all conditions that we considered in the simulation design. We first calculated mean values of accuracy and entropy across 1,000 Monte Carlo replications for each condition. Of all the conditions we investigated, the mean entropy ranges from 0.3 to 0.8, while the mean accuracy ranges from 0.55 to 0.95. Factors such as the separation

Table 6. Median (Range) of the Coverage Probabilities over $1,000$ Replications of Parameters of Interest under the Conditions with Fixed Knots and 2 Latent Classes

	Small Separation between the Knots Locations			
	Latent Class 1		**Latent Class** 2	
	Small Residuals	Large Residuals	Small Residuals	Large Residuals
μ_{η_0}	.937 (.913, .961)	.915 (.866, .950)	.942 (.920, .971)	.919 (.867, .952)
μ_{η_1}	.919 (.861, .948)	.874 (.766, .942)	.936 (.901, .962)	.904 (.819, .941)
μ_{η_2}	.926 (.849, .949)	.893 (.747, .940)	.938 (.888, .956)	.913 (.855, .949)
μ_{γ}	.629 (.493, .724)	.476 (.290, .623)	.522 (.406, .685)	.355 (.227, .541)
ψ_{00}	.939 (.916, .954)	.932 (.896, .950)	.939 (.927, .957)	.925 (.888, .963)
ψ_{11}	.933 (.878, .950)	.921 (.831, .957)	.935 (.911, .966)	.927 (.877, .947)
ψ_{22}	.929 (.862, .950)	.904 (.809, .935)	.938 (.902, .961)	.930 (.888, .957)
β_0	—	—	.789 (.665, .854)	.643 (.502, .739)
β_1	—	—	.950 (.935, .960)	.936 (.891, .957)
β_2	—	—	.944 (.930, .959)	.933 (.873, .958)
	Medium Separation between the Knots Locations			
	Latent Class 1		**Latent Class** 2	
	Small Residuals	Large Residuals	Small Residuals	Large Residuals
μ_{η_0}	.944 (.918, .959)	.929 (.899, .951)	.943 (.923, .957)	.932 (.905, .955)
μ_{η_1}	.938 (.897, .957)	.922 (.833, .951)	.947 (.917, .959)	.932 (.884, .959)
μ_{η_2}	.935 (.910, .948)	.913 (.835, .947)	.940 (.913, .959)	.934 (.883, .954)
μ_{γ}	.814 (.786, .854)	.740 (.684, .800)	.767 (.721, .833)	.682 (.626, .780)
ψ_{00}	.940 (.925, .953)	.935 (.912, .955)	.944 (.927, .953)	.939 (.901, .950)
ψ_{11}	.939 (.905, .952)	.929 (.853, .953)	.939 (.914, .961)	.937 (.909, .952)
ψ_{22}	.930 (.906, .958)	.920 (.878, .951)	.939 (.917, .962)	.934 (.889, .951)
β_0	—	—	.858 (.782, .905)	.770 (.658, .839)
β_1	—	—	.954 (.937, .961)	.944 (.921, .965)
β_2	—	—	.949 (.934, .964)	.942 (.923, .961)
	Large Separation between the Knots Locations			
	Latent Class 1		**Latent Class** 2	
	Small Residuals	Large Residuals	Small Residuals	Large Residuals
μ_{η_0}	.946 (.931, .955)	.938 (.921, .965)	.946 (.932, .959)	.940 (.921, .967)
μ_{η_1}	.938 (.921, .959)	.936 (.875, .953)	.947 (.926, .958)	.937 (.893, .961)
μ_{η_2}	.939 (.907, .956)	.928 (.876, .951)	.949 (.937, .964)	.940 (.916, .955)
μ_{γ}	.952 (.935, .970)	.946 (.935, .961)	.950 (.933, .965)	.946 (.932, .960)
ψ_{00}	.946 (.929, .957)	.944 (.916, .963)	.943 (.916, .958)	.942 (.921, .959)
ψ_{11}	.938 (.917, .952)	.934 (.859, .951)	.942 (.918, .955)	.938 (.902, .956)
ψ_{22}	.935 (.910, .950)	.928 (.857, .951)	.946 (.925, .959)	.938 (.919, .953)
β_0	—	—	.892 (.825, .924)	.805 (.703, .865)
β_1	—	—	.950 (.927, .958)	.949 (.937, .960)
β_2	—	—	.950 (.934, .964)	.946 (.924, .958)

Note. —: when fitting the proposed model, we set the first latent class as the reference group; accordingly, the coefficients of that class do not exist.

between two latent classes and the precision of measurements were the primary determinants of entropy and accuracy.

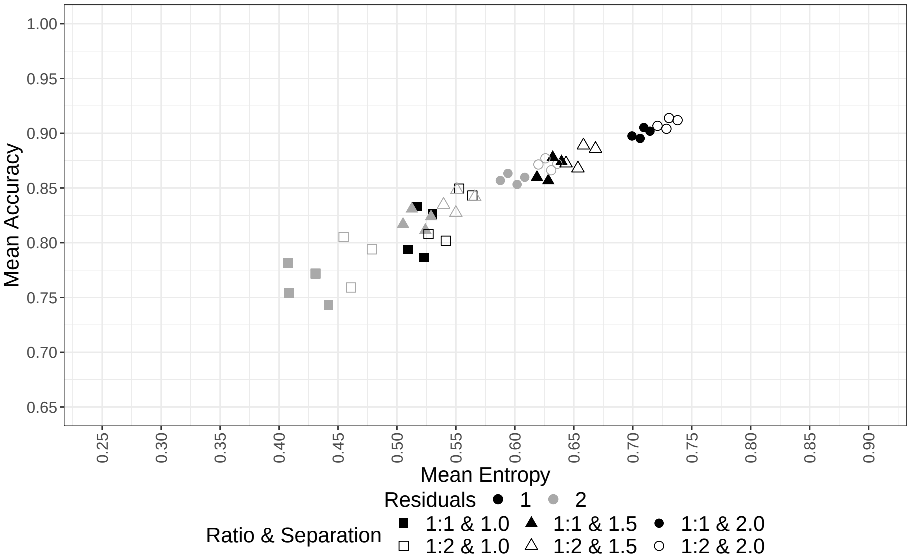


Figure 1. Accuracy vs Entropy of the Proposed Mixture Model (Step 1) with 2-Clusters and Small Mahalanobis Distance

Figure 1 depicts the mean accuracy against the mean entropy for each condition with two latent classes, the small Mahalanobis distance, and change patterns of Scenario 1 listed in Table 2. In the plot, we colored the conditions with the smaller and the larger residual variances black and grey, respectively. Squares, triangles, and circles are for the small, medium, and large differences between the locations of the knots. Additionally, we set solid and hollow shapes for the proportions 1:1 and 1:2, respectively. From the figure, we observed that both entropy and accuracy increased when the separation between two latent classes increased and as the residual variances were small. Additionally, unbalanced allocation tended to yield relatively larger accuracy and entropy. We also noticed that the scenario of change patterns only affected entropy and accuracy slightly, while other factors such as the knot standard deviation and the sample size did not have meaningful impacts on entropy and accuracy. We observed the same patterns between the mean accuracy and the mean entropy of conditions with three latent classes.

5 Application

In this section, we demonstrate how to fit the proposed model to separate nonlinear trajectories and associate the 'soft clusters' to the baseline characteristics using the motivating data. We extracted a random subsample ($n = 500$) from the Early Childhood Longitudinal Study Kindergarten Cohort: 2010-11 (ECLS-K: 2011) with complete records of repeated mathematics IRT scaled scores, demographic information (sex, race, and age in months at each wave), baseline school information (school location and baseline school type), baseline social-economic status (family income and the highest education level between parents), baseline teacher-reported social skills (including interpersonal skills, self-control ability, internalizing problem, externalizing problem), baseline teacher-reported approach to learning, and baseline teacher-reported children behavior question (including inhibitory control and attentional focus)[6].

ECLS-K: 2011 is a nationally representative longitudinal sample of US children enrolled in about 900 kindergarten programs beginning with $2010 - 2011$ school year, where children's mathematics ability was evaluated in nine waves: fall and spring of kindergarten ($2010 - 2011$), first ($2011 - 2012$) and second ($2012 - 2013$) grade, respectively as well as spring of 3^{rd} (2014), 4^{th} (2015) and 5^{th} (2016), respectively. Only about 30% students were assessed in the fall of 2011 and 2012 (Lê, Norman, Tourangeau, Brick, & Mulligan, 2011). In the analysis, we used children's age (in months) rather than their grade-in-school to obtain the time structure with individual measurement occasions. In the subset data, 52% of students were boys, and 48% of students were girls. Additionally, 50% of students were White, 4.8% were Black, 30.4% were Hispanic, 0.2% were Asian, and 14.6% were others. We dichotomized the variable race to be White (50%) and others (50%) for this analysis. At the beginning of the study, 87% and 13% students were from public and private schools, respectively. The covariates including school location (ranged between 1 and 4), family income (ranged between 1 and 18) and the highest parents' education (ranged between 0 and 8) were treated as a continuous variables, and the corresponding mean (SD) was 2.11 (1.12), 11.99 (5.34) and 5.32 (1.97), respectively.

Step 1

In the first step, we first fit a latent growth curve model with a linear-linear piecewise functional form and three GMMs with two-, three- and four-class and provided the obtained estimated likelihood, information criteria (AIC and BIC), residual of each latent class in Table 7. All four models converged. As introduced earlier, the BIC is a compelling information criterion for the enumeration process as it penalizes model complexity and adjusts for sample size (Nylund et al., 2007). The four fits led to BIC values of 31728.23, 31531.60, 31448.99, and 31478.35, respectively, which led to the selection of the GMM with three latent classes.

Table 8 presents the estimates of growth factors from which we obtained the model implied trajectory of each latent group, as shown in Figure 2. The estimated proportions in Class 1, 2 and 3 were 29.6%, 47.8% and 22.6%, respectively. On average, students in Class 1 had the lowest levels of mathematics

[6] The total sample size of ECLS-K: 2011 $n = 18174$. The number of entries after removing records with missing values (i.e., rows with any of NaN/-9/-8/-7/-1) is $n = 1853$.

Table 7. Summary of Model Fit Information For the Bilinear Spline Growth Models with Different # of Latent Classes

	1-Class	2-Class	3-Class	4-Class
-2LL	31659.87	31388.67	31231.48	31186.26
AIC	31681.87	31434.67	31301.48	31280.26
BIC	31728.23	31531.6	31448.99	31478.35
Residual 1	35.6	28.57	28.47	26.78
Residual 2	-	35.02	33.89	32.51
Residual 3	-	-	32.03	33.36
Residual 4	-	-	-	26.63

Note. − indicates that the metric was not available for the model.

achievement throughout the entire duration (the fixed effects of the baseline and two slopes were 24.133, 1.718 per month, and 0.841 per month, respectively). On average, students in Class 2 had a similar initial score and slope for the first stage but relatively lower slope in the second stage (the fixed effects of the baseline and two slopes were 24.498, 1.730 per month, and 0.588 per month, respectively) compared to the students in the Class 1. Students in Class 3 had the best mathematics performance on average (the fixed effects of the baseline and two slopes were 36.053, 2.123 per month, and 0.605 per month, respectively). For all three classes, post-knot development in mathematics skills slowed substantially, yet the change to the slower growth rate occurred earlier for Class 1 and 3 (around 8-year old: 91 and 97 months, respectively) than Class 2 (around 9-year old, 110 months). Additionally, for each latent class, the estimates of the intercept variance and first slope variance were statistically significant, indicating that each student had a 'personal' intercept and pre-knot slope, and then a 'personal' trajectory of the development in mathematics achievement.

Step 2

Table 9 summarizes the estimates of the second step of the GMM to associate 'soft clusters' of mathematics achievement trajectories to individual-level covariates. From the table, we noticed that the impacts of some covariates, such as baseline socioeconomic status and teacher-reported skills, may differ with or without other covariates. For example, higher family income, higher parents' education, higher-rated attentional focus, and inhibitory control increased the likelihood of being in Class 2 or Class 3 in univariable analyses, while these four baseline characteristics only associated with Class 3 in multivariable analyses. It is reasonable that the effect sizes of the Class 3 were larger than those of the Class 2, given its more evident difference from the reference group, as shown in Table 8 and Figure 2. However, it is still too rush to neglect that students from families with higher socioeconomic status and/or higher-rated behavior questions were more likely to be in Class 2 at the significant level of 0.05 in an exploratory study. Another possible explanation for this phenomenon is multicollinearity.

Table 8. Estimates of the Proposed Mixture Model with 3 Latent Classes (Step 1)

			Estimate (SE)	P value
Class 1	Mean	**Intercept**^{1}	24.133 (1.250)	$< 0.0001^*$
		Slope 1	1.718 (0.052)	$< 0.0001^*$
		Slope 2	0.841 (0.031)	$< 0.0001^*$
		Knot	90.788 (0.733)	$< 0.0001^*$
	Variance	**Intercept**	79.696 (17.419)	$< 0.0001^*$
		Slope 1	0.104 (0.023)	$< 0.0001^*$
		Slope 2	0.049 (0.011)	$< 0.0001^*$
Class 2	Mean	**Intercept**^{1}	24.498 (0.813)	$< 0.0001^*$
		Slope 1	1.730 (0.024)	$< 0.0001^*$
		Slope 2	0.588 (0.032)	$< 0.0001^*$
		Knot	109.653 (0.634)	$< 0.0001^*$
	Variance	**Intercept**	77.302 (11.973)	$< 0.0001^*$
		Slope 1	0.026 (0.007)	0.0002^*
		Slope 2	0.012 (0.011)	0.2753
Class 3	Mean	**Intercept**^{1}	36.053 (1.729)	$< 0.0001^*$
		Slope 1	2.123 (0.035)	$< 0.0001^*$
		Slope 2	0.605 (0.027)	$< 0.0001^*$
		Knot	97.610 (0.068)	$< 0.0001^*$
	Variance	**Intercept**	211.198 (36.057)	$< 0.0001^*$
		Slope 1	0.065 (0.017)	0.0001^*
		Slope 2	−0.002 (0.006)	0.7389

Note. ^{1}Intercept was defined as mathematics IRT scores at 60-month old in this case.

* indicates statistical significance at 0.05 level.

Table 9. Odds Ratio (OR) & 95% Confidence Interval (CI) of Individual-level Predictor of Latent Class in Mathematics Achievement(Reference group: Class 1)

		Class 2		
Predictor	**Uni-variable**		**Multi-variable**	
	OR	95% CI	OR	95% CI
Sex(0−Boy; 1−Girl)	0.435	(0.254, 0.745)*	0.332	(0.174, 0.633)*
Race(0−White; 1−Other)	0.764	(0.455, 1.281)	1.249	(0.624, 2.498)
School Location	1.407	(1.093, 1.811)*	1.357	(0.981, 1.877)
Parents' Highest Education	1.208	(1.051, 1.388)*	1.155	(0.933, 1.431)
Income	1.074	(1.023, 1.128)*	1.067	(0.987, 1.154)
School Type (0−Public; 1−Private)	0.573	(0.250, 1.317)	0.442	(0.149, 1.313)
Approach to Learning	1.305	(0.883, 1.929)	0.957	(0.384, 2.389)
Self-control	1.146	(0.764, 1.718)	0.663	(0.272, 1.616)
Interpersonal Skills	1.479	(0.959, 2.282)	1.276	(0.513, 3.175)
External Prob Behavior	0.858	(0.559, 1.319)	1.391	(0.571, 3.386)
Internal Prob Behavior	1.139	(0.658, 1.972)	1.190	(0.589, 2.406)
Attentional Focus	1.251	(1.035, 1.511)*	1.139	(0.764, 1.698)
Inhibitory Control	1.238	(1.007, 1.520)*	1.557	(0.915, 2.649)
		Class 3		
Predictor	**Uni-variable**		**Multi-variable**	
	OR	95% CI	OR	95% CI
Sex(0−Boy; 1−Girl)	0.379	(0.205, 0.700)*	0.212	(0.098, 0.459)*
Race(0−White; 1−Other)	0.397	(0.219, 0.721)*	0.943	(0.429, 2.073)
School Location	1.266	(0.957, 1.676)	1.211	(0.835, 1.755)
Parents' Highest Education	1.713	(1.418, 2.068)*	1.345	(1.043, 1.734)*
Income	1.241	(1.155, 1.334)*	1.195	(1.083, 1.318)*
School Type (0−Public; 1−Private)	1.437	(0.661, 3.124)	0.665	(0.234, 1.892)
Approach to Learning	2.624	(1.590, 4.332)*	5.363	(1.731, 16.612)*
Self-control	1.436	(0.903, 2.284)	0.414	(0.136, 1.265)
Interpersonal Skills	1.740	(1.057, 2.862)*	0.771	(0.269, 2.209)
External Prob Behavior	0.761	(0.451, 1.283)	1.565	(0.561, 4.367)
Internal Prob Behavior	0.787	(0.405, 1.532)	1.170	(0.488, 2.808)
Attentional Focus	1.601	(1.253, 2.045)*	1.095	(0.671, 1.787)*
Inhibitory Control	1.439	(1.116, 1.855)*	1.324	(0.720, 2.434)*

Note. * indicates 95% confidence interval excluded 1.

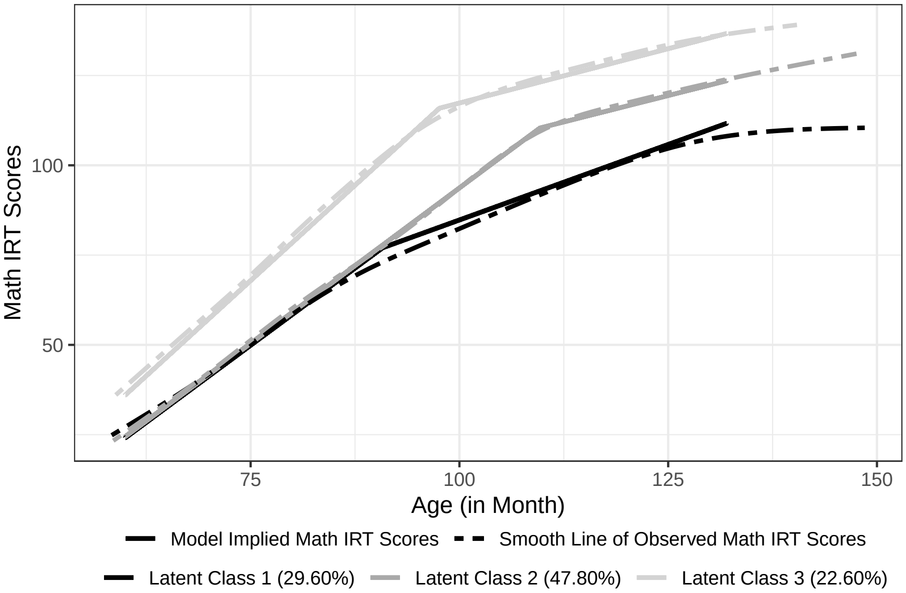
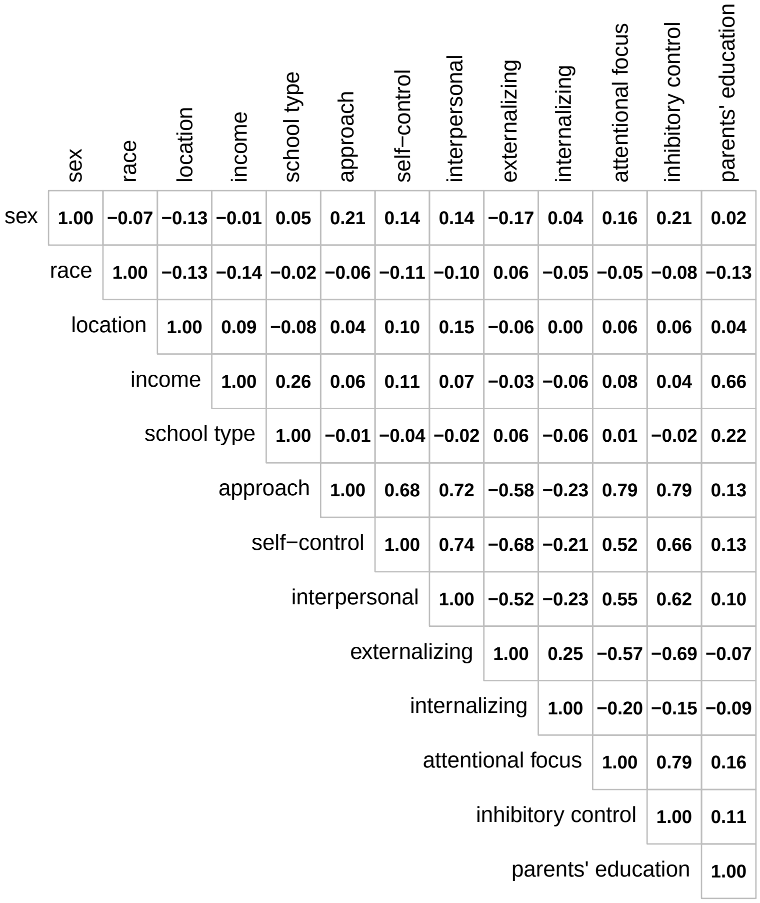


Figure 2. Three Latent Classes: Model Implied Trajectories and Smooth Lines of Observed Mathematics IRT Scores

Figure 3 visualizes the correlation matrix of all baseline characteristics, from which we can see that two socioeconomic variables, family income and parents' highest education, were highly correlated ($\rho = 0.66$). Additionally, teacher-rated baseline abilities were highly correlated; for example, the correlation of approach to learning with self-control, interpersonal ability, attentional focus, and inhibitory control was 0.68, 0.72, 0.79 and 0.79, respectively. We then conducted the exploratory factor analysis to address this collinearity issue for socioeconomic variables and teacher-reported abilities.

The exploratory factor analysis was conducted using the R function *factanal* in the *stats* package (R Core Team, 2020) with 2 specified factors as suggested by the eigenvalues greater than 1 (EVG1) component retention criterion, scree test (Cattell, 1966; Cattell & Jaspers, 1967), and parallel analysis (Horn, 1965; Humphreys & Ilgen, 1969; Humphreys & Montanelli, 1975). We employed the 'varimax' option to get a type of orthogonal rotation (Kaiser, 1958). By using Bartlett's weighted least-squares methods, we obtained the factor scores. Table 10 summarizes the results from the EFA. The first factor differentiates between teacher-rated abilities and teacher-reported problems; the second factor can be interpreted as general socioeconomic status. We then re-ran the second step with the two factors as well as demographic information and school information.

Table 11 summarizes the estimates obtained from the second step with factor scores, demographic information, and school information. From the table, we observed that boys with higher values of the first factor scores, and higher



	sex	race	location	income	school type	approach	self−control	interpersonal	externalizing	internalizing	attentional focus	inhibitory control	parents' education
sex	1.00	−0.07	−0.13	−0.01	0.05	0.21	0.14	0.14	−0.17	0.04	0.16	0.21	0.02
race		1.00	−0.13	−0.14	−0.02	−0.06	−0.11	−0.10	0.06	−0.05	−0.05	−0.08	−0.13
location			1.00	0.09	−0.08	0.04	0.10	0.15	−0.06	0.00	0.06	0.06	0.04
income				1.00	0.26	0.06	0.11	0.07	−0.03	−0.06	0.08	0.04	0.66
school type					1.00	−0.01	−0.04	−0.02	0.06	−0.06	0.01	−0.02	0.22
approach						1.00	0.68	0.72	−0.58	−0.23	0.79	0.79	0.13
self−control							1.00	0.74	−0.68	−0.21	0.52	0.66	0.13
interpersonal								1.00	−0.52	−0.23	0.55	0.62	0.10
externalizing									1.00	0.25	−0.57	−0.69	−0.07
internalizing										1.00	−0.20	−0.15	−0.09
attentional focus											1.00	0.79	0.16
inhibitory control												1.00	0.11
parents' education													1.00

Figure 3. Pairwise Correlation between Baseline Characteristics

Table 10. Exploratory Factor Analysis of Socioeconomic Variables and Teacher-reported Abilities

Factor Loadings		
Baseline Characteristics	**Factor 1**	**Factor 2**
Parents' Highest Education	0.10	0.76
Family Income	0.03	0.86
Approach to Learning	0.90	0.04
Self-control	0.77	0.08
Interpersonal Skills	0.76	0.05
External Prob Behavior	−0.72	0.00
Internal Prob Behavior	−0.24	−0.07
Attentional Focus	0.83	0.07
Inhibitory Control	0.89	0.01
Explained Variance		
	Factor 1	**Factor 2**
SS Loadings	4.04	1.34
Proportion Variance	0.45	0.15
Cumulative Variance	0.45	0.60

values of the second factor scores were more likely to be in Class 2[7] or Class 3[8]. It suggests that both socioeconomic variables and teacher-rated abilities were positively associated with mathematics performance, while externalizing/internalizing problems were negative associated with mathematics achievement.

Table 11. Odds Ratio (OR) & 95% Confidence Interval (CI) of Factor Scores, Demographic Information and School Information of Latent Class in Mathematics Achievement (Reference group: Class 1)

Predictor	**Class** 2		**Class** 3	
	OR	95% CI	OR	95% CI
Sex(0−Boy; 1−Girl)	0.345	(0.183, 0.651)*	0.234	(0.111, 0.494)*
Race(0−White; 1−Other)	1.221	(0.638, 2.339)	1.021	(0.486, 2.145)
School Type (0−Public; 1−Private)	0.439	(0.149, 1.291)	0.709	(0.244, 2.056)
School Location	1.333	(0.995, 1.786)	1.133	(0.806, 1.593)
Factor 1	1.454	(1.090, 1.939)*	2.006	(1.408, 2.858)*
Factor 2	1.656	(1.226, 2.235)*	3.410	(2.258, 5.148)*

Note. * indicates 95% confidence interval excluded 1.

[7] OR (95% CI) for sex, factor score 1 and factor score 2 was 0.345 (0.183, 0.651), 1.454 (1.090, 1.939) and 1.656 (1.226, 2.235), respectively.

[8] OR (95% CI) for sex, factor score 1 and factor score 2 was 0.234 (0.111, 0.494), 2.006 (1.408, 2.858) and 3.410 (2.258, 5.148), respectively.

6 Discussion

This article extends Bakk and Kuha (2017) study to conduct a stepwise analysis to investigate the heterogeneity in nonlinear trajectories. We fit a growth mixture model with a bilinear spline functional form to describe the underlying change pattern of nonlinear trajectories in the first step. In the second step, we investigated the associations between the 'soft' clusters and baseline characteristics. Although this stepwise method follows the recommended approach to fit a FMM model (i.e., separate the estimation of the class-specific parameters and that of the logistic coefficients), it is not our aim to show that this stepwise approach is universally preferred. Based on our understanding, this approach is more suitable for an exploratory study where empirical researchers only have vague assumptions in terms of sample heterogeneity and its possible causes.

On the one hand, the two-step model can save computational budget as we only need to refit the second-step model rather than the whole model when adding or removing covariates. On the other hand, our simulation study showed that the proposed model works well in terms of performance measures and accuracy, especially under preferable conditions, such as well-separated latent classes and precise measurements. This stepwise approach can also be utilized to analyze any other types of FMMs in the SEM framework to explore sample heterogeneity.

6.1 Methodological Consideration

Although this stepwise model can expedite the exploratory process, it is still challenging to decide which covariates should be added in the mixture model to inform the class formation. An additional challenge lies in that, in the psychological and behavioral research where the SEM framework is widely used, the candidate pool of covariates is huge, or some variables are highly correlated (i.e., collinearity issue), as shown in the application.

In the statistical and machine learning (ML) literature, multiple approaches have been proposed to reduce the number of covariates. These methods include greedy search, regularization to select covariates based on their corresponding coefficients, principal component analysis (PCA) to transform all features to space with fewer dimensions, and tree-based models (such as regression and classification trees, boosting, and bagging). In the SEM framework, the majority of counterparts of the above models have been proposed. For example, Marcoulides and Drezner (2003); Marcoulides, Drezner, and Schumacker (1998) proposed to conduct a heuristic specification search algorithm to identify an optimal set of models; Jacobucci, Grimm, and McArdle (2016); Scharf and Nestler (2019); Sun, Chen, Liu, Ying, and Xin (2016), demonstrated how to regularize parameters in the SEM framework to reduce the complexity of the model by selecting or removing paths (i.e., variables). Additionally, by applying a tree-based model Brandmaier, von Oertzen, McArdle, and Lindenberger (2013), Jacobucci, Grimm, and McArdle (2017) captured the heterogeneity in trajectories with respect to baseline covariates, where the FMM was compared with the tree-based model in terms of membership components and result interpretation.

This article proposes to employ the EFA to reduce the dimensions of covariates and address the multicollinearity issue. In this application, we applied the EFA in a process termed as 'feature engineering' in the ML literature, where researchers employ the PCA technique to reduce the covariate space and address the multicollinearity issue conventionally, as the interpretation of covariate coefficients is out of the primary interest in the ML literature. In this article, we decided to use the EFA rather than the PCA for two reasons. First, empirical researchers using the SEM framework are more familiar with the EFA as the idea behind it is very similar to another model in the SEM framework, the confirmatory factor analysis (CFA). More importantly, the factors (i.e., latent variables) obtained from the EFA are interpretable so that the estimated coefficients from the second step are interpretable, and we then gain valuable insights from an exploratory study. For example, in the application, we concluded that a student with a higher value of the difference between teacher-rated abilities and teacher-reported problems and/or from a family with higher socioeconomic status was more likely to achieve higher mathematics scores (i.e., in Class 2 and Class 3).

Although it is not our aim to comprehensively investigate the EFA, we still want to add two notes about factor retention criteria and factor rotation to empirical researchers. Following Fabrigar, Wegener, MacCallum, and Strahan (1999), we used multiple criteria in the application, including the EVG1 rule, scree test, and parallel analysis to decide the number of factors; fortunately, all these criteria gave the same decision. Patil, Singh, Mishra, and Todd Donavan (2008) also suggested conducting a subsequent CFA to evaluate the measurement properties of the factors identified by the EFA (if the number of factors is different from multiple criteria).

Additionally, several analytic rotation techniques have been developed for the EFA, with the most fundamental distinction lying in orthogonal and oblique rotation. Orthogonal rotations constrain factors to be uncorrelated, and the procedure, varimax, which we used in the application, is generally regarded as the best one and the most widely used orthogonal rotation in psychological research. One reason for this choice was its simplicity and conceptual clarity. More importantly, we assumed that the constructs (i.e., the factor of the socioeconomic variables and that of teacher-rated scores) identified from the covariates set are independent. However, many theoretical and empirical researchers provided the basis for expecting psychological constructs, such as personality traits, ability, and attitudes, to be associated with each other. Consequently, oblique rotations provide a more realistic and accurate picture of these factors.

One limitation of the proposed two-step model lies in that it only allows (generalized) linear models in the second step. If the linear assumption is invalid, we need to resort to other methods, such as structural equation model trees (SEM trees, Brandmaier et al. (2013)) or structural equation model forests (Brandmaier, Prindle, McArdle, & Lindenberger, 2016) to identify the most important covariates by investigating the variables on which the tree splits first (Brandmaier et al., 2013; Jacobucci et al., 2017) or the output named 'variable importance' (Brandmaier et al., 2016), respectively. Note that Jacobucci et al.

(2017) pointed out that the interpretations of the FMM and SEM trees are different, and the classes obtained from the SEM tree can be viewed as the clusters of associations between the covariates and trajectories.

6.2 Future Research

One possible future direction of the current study is to build its confirmatory counterpart. Conceptually, the confirmatory model consists of two measurement models, and there exists a unidirectional relationship between the factors of the EFA and the latent categorical variable. Additionally, driven by domain knowledge, the EFA can be replaced with the CFA in the confirmatory model. Additionally, the two-step model is proposed under the assumption that these covariates only indirectly impact the sample heterogeneity. It is also possible to develop a model that allows these baseline covariates to simultaneously explain between-group differences and within-group differences by relaxing the assumption.

References

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 329-341. Retrieved from `https://doi.org/10.1080/10705511.2014.915181`

Bakk, Z., & Kuha, J. (2017, 11 17). Two-step estimation of models between latent classes and external variables. *Psychometrika*, 1-22. Retrieved from `https://doi.org/10.1007/s11336-017-9592-7`

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *440*(92), 1375-1386. Retrieved from `https://doi.org/10.1080/01621459.1997.10473658`

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Educational Psychology*, *General Section, 28*, 97-104.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, *8(3)*, 338-363. Retrieved from `https://doi.org/10.1037/1082-989X.8.3.338`

Bishop, C. (2006). *Pattern recognition and machine learning.* Springer-Verlag.

Blozis, S. A., & Cho, Y. (2008). Coding and centering of time in latent curve models in the presence of interindividual time heterogeneity. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(3), 413-433. Retrieved from `https://doi.org/10.1080/10705510802154299`

Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., Spiegel, M., Brick, T. R., . . . Kirkpatrick, R. M. (2020). Openmx 2.17.2 user guide [Computer software manual].

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*(1), 3-27. Retrieved from `https://www.jstor.org/stable/25791751`

Bouveyron, C., Celeux, G., Murphy, T., & Raftery, A. (2019). *Model-based clustering and classification for data science: With applications in r (cambridge series in statistical and probabilistic mathematics).* Cambridge: Cambridge University Press. Retrieved from `https://doi.org/10.1017/9781108644181`

Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods*, *4*(21), 566-582. Retrieved from `https://doi.org/10.1037/met0000090`

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18(1)*, 71-86. Retrieved from `https://doi.org/10.1037/a0030001`

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245-276.

Cattell, R. B., & Jaspers, J. (1967). A general plasmode (no. 30-10-5-2) for factor analytic exercises and research. *Multivariate Behavioral Research Monographs*, *67*(3), 211.

Clogg, C. C. (1981). New developments in latent structure analysis. In D. J. Jackson & E. F. Borgotta (Eds.), *Factor analysis and measurement in sociological research: A multi-dimensional perspective* (p. 215-246). Beverly Hills, CA: SAGE Publications.

Cook, N. R., & Ware, J. H. (1983). Design and analysis methods for longitudinal research. *Annual Review of Public Health*, *4*(1), 1-23.

Coulombe, P., Selig, J. P., & Delaney, H. D. (2015). Ignoring individual differences in times of assessment in growth curve modeling. *International Journal of Behavioral Development*, *40*(1), 76-86. Retrieved from `https://doi.org/10.1177/0165025415577684`

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*(401), 173-178. Retrieved from `https://doi.org/10.2307/2288938`

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. Retrieved from `https://doi.org/10.1037/1082-989X.4.3.272`

Finkel, D., Reynolds, C. A., Mcardle, J. J., Gatz, M., & Pedersen, N. L. (2003, 06). Latent growth curve analyses of accelerating decline in cognitive abilities in late adulthood. *Developmental psychology*, *39*, 535-550. Retrieved from `https://doi.org/10.1037/0012-1649.39.3.535`

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. part i-a modified latent structure approach. *American Journal of Sociology*, *79*(5), 1179-1259. Retrieved from http://www.jstor.org/stable/2776792

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling.* Guilford Press.

Haberman, S. (1979). *Analysis of qualitative data. vol. 2: New developments.* New York: Academic Press.

Hagenaars, J. A. (1993). *Loglinear models with latent variables.* Newbury Park, CA: Sage.

Harring, J. R., Cudeck, R., & du Toit, S. H. C. (2006). Fitting partially nonlinear random coefficient models as sems. *Multivariate Behavioral Research*, *41*(4), 579-596. Retrieved from https://doi.org/10.1207/s15327906mbr4104_7

Horn, J. L. (1965). A rationale and technique for estimating the number of factors in factor analysis. *Psychometrika*, *30*, 179-185.

Humphreys, L. G., & Ilgen, D. R. (1969). Note on a criterion for the number of common factors. *Educational and Psychological Measurement*, *29*(3), 571–578.

Humphreys, L. G., & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, *10*(2), 193–205.

Hunter, M. D. (2018). State space modeling in an open source, modular, structural equation modeling environment. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(2), 307-324. Retrieved from https://doi.org/10.1080/10705511.2017.1369354

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555-566. Retrieved from https://doi.org/10.1080/10705511.2016.1154793

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. Retrieved from https://doi.org/10.1080/10705511.2016.1250637

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200. Retrieved from https://doi.org/10.1007/BF02289233

Kamakura, W. A., Wedel, M., & Agrawal, J. (1994). Concomitant variable latent class models for conjoint analysis. *International Journal of Research in Marketing*, *11*(5), 451-464. Retrieved from https://doi.org/10.1016/0167-8116(94)00004-2

Kohli, N. (2011). *Estimating unknown knots in piecewise linear-linear latent growth mixture models* (Doctoral dissertation, University of Maryland). Retrieved from http://hdl.handle.net/1903/11973

Kohli, N., & Harring, J. R. (2013). Modeling growth in latent variables using a piecewise function. *Multivariate Behavioral Research*, *48*(3), 370-397. Retrieved from https://doi.org/10.1080/00273171.2013.778191

Kohli, N., Harring, J. R., & Hancock, G. R. (2013). Piecewise linear-linear latent growth mixture models with unknown knots. *Educational and Psychological Measurement*, *73*(6), 935-955. Retrieved from https://doi.org/10.1177/0013164413496812

Kohli, N., Hughes, J., Wang, C., Zopluoglu, C., & Davison, M. L. (2015). Fitting a linear-linear piecewise growth mixture model with unknown knots: A comparison of two common approaches to inference. *Psychological Methods*, *20*(2), 259-275. Retrieved from https://doi.org/10.1037/met0000034

Lê, T., Norman, G., Tourangeau, K., Brick, J. M., & Mulligan, G. (2011). Early childhood longitudinal study: Kindergarten class of 2010-2011 - sample design issues. *JSM Proceedings*, 1629-1639. Retrieved from http://www.asasrms.org/Proceedings/y2011/Files/301090_66141.pdf

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation, 2nd edition.* Springer-Verlag New York, Inc.

Liu, J. (2019). *Estimating knots in bilinear spline growth models with time-invariant covariates in the framework of individual measurement occasions* (Doctoral dissertation, Virginia Commonwealth University). Retrieved from https://doi.org/10.25772/9WDR-9R85

Liu, J., Perera, R. A., Kang, L., Kirkpatrick, R. M., & Sabo, R. T. (2019). *Obtaining interpretable parameters from reparameterizing longitudinal models: transformation matrices between growth factors in two parameter-spaces.*

Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21–39. Retrieved from https://doi.org/10.1037/1082-989X.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. Retrieved from https://doi.org/10.1080/10705510709336735

Marcoulides, G. A., & Drezner, Z. (2003). Model specification searches using ant colony optimization algorithms. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 154-164. Retrieved from https://doi.org/10.1207/S15328007SEM1001_8

Marcoulides, G. A., Drezner, Z., & Schumacker, R. E. (1998). Model specification searches in structural equation modeling using tabu search. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(4), 365-376. Retrieved from https://doi.org/10.1080/10705519809540112

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* John Wiley & Sons, Inc.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259-284. Retrieved from https://doi.org/10.1037/1082-989x.5.1.23

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods*, *5*(1), 23-43.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074-2102. Retrieved from https://doi.org/10.1002/sim.8086

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*(2), 463-469. Retrieved from https://doi.org/10.1111/j.0006-341x.1999.00463.x

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. Retrieved from https://doi.org/10.1007/s11336-014-9435-8

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. Retrieved from https://doi.org/10.1080/10705510701575396

Patil, V. H., Singh, S. N., Mishra, S., & Todd Donavan, D. (2008). Efficient theory development and factor retention criteria: Abandon the 'eigenvalue greater than one' criterion. *Journal of Business Research*, *61*(2), 162-170. Retrieved from https://doi.org/10.1016/j.jbusres.2007.05.008

Preacher, K. J., & Hancock, G. R. (2015). Meaningful aspects of change as novel random coefficients: A general method for reparameterizing longitudinal models. *Psychological Methods*, *20*(1), 84-101. Retrieved from https://doi.org/10.1037/met0000028

Pritikin, J. N., Hunter, M. D., & Boker, S. M. (2015). Modular open-source software for Item Factor Analysis. *Educational and Psychological Measurement*, *75*(3), 458-474. Retrieved from https://doi.org/10.1177/0013164414554615

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Scharf, F., & Nestler, S. (2019). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 576-590. Retrieved from https://doi.org/10.1080/10705511.2018.1558060

Seber, G. A. F., & Wild, C. J. (2003). *Nonlinear regression.* John Wiley & Sons, Inc.

Stegmann, G., & Grimm, K. J. (2018). A new perspective on the effects of covariates in mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(2), 167-178. Retrieved from https://doi.org/10.1080/10705511.2017.1318070

Sterba, S. K. (2014). Fitting nonlinear latent growth curve models with individually varying time points. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 630-647. Retrieved from https://doi.org/10.1080/10705511.2014.919828

Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). Latent variable selection for multidimensional item response theory models via l1 regularization.

Psychometrika, *81*, 921-939. Retrieved from https://doi.org/10.1007/s11336-016-9529-6

Thomson, G. (1939). The factorial analysis of human ability. *British Journal of Educational Psychology*, *9*, 188-195.

Tishler, A., & Zang, I. (1981). A new maximum likelihood algorithm for piecewise regression. *Journal of the American Statistical Association*, *76*(376), 980-987. Retrieved from https://doi.org/10.2307/2287599

Tueller, S. J., Drotar, S., & Lubke, G. H. (2011). Addressing the problem of switched class labels in latent variable mixture model simulation studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(1), 110–131. Retrieved from https://doi.org/10.1080/10705511.2011.534695

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer.

Vermunt, J. K. (1997). *Advanced quantitative techniques in the social sciences series, vol. 8. log-linear models for event histories.* Thousand Oaks, CA, US: Sage Publications, Inc.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450-469. Retrieved from https://www.jstor.org/stable/25792024

Yamaguchi, K. (2000). Multinomial logit latent-class regression models: An analysis of the predictors of gender-role attitudes among japanese women. *American Journal of Sociology*, *105*(6), 1702-1740. Retrieved from https://doi.org/10.1086/210470

Appendix A. Formula Derivation

A.1. The Reparameterizing Procedure for a Fixed Knot

In the original setting of the bilinear spline model, we have three growth factors: an intercept at t_0 (η_0) and one slope of each stage (η_1 and η_2, respectively). To estimate knots, we may reparameterize the growth factors. For the i^{th} individual, according to Seber and Wild (Seber & Wild, 2003), we may re-expressed them as the measurement at the knot (i.e., $\eta_{0i} + \eta_{1i}\gamma^{(k)}$), the mean of two slopes (i.e., $\frac{\eta_{1i}+\eta_{2i}}{2}$), and the half difference between two slopes (i.e., $\frac{\eta_{2i}-\eta_{1i}}{2}$).

Tishler and Zang (1981) and Seber and Wild (2003) showed that the regression model with two linear stages can be written as either the minimum or maximum response value of two trajectories. Liu et al. (2019) extended such expressions to the latent growth curve modeling framework and showed two forms of bilinear spline for the i^{th} individual in Figure A.1. In the left panel ($\eta_{1i} > \eta_{2i}$), the measurement y_{ij} is always the minimum value of two lines; that is, $y_{ij} = \min(\eta_{0i} + \eta_{1i}t_{ij}, \eta_{02i} + \eta_{2i}t_{ij})$. To unify the formula of measurements

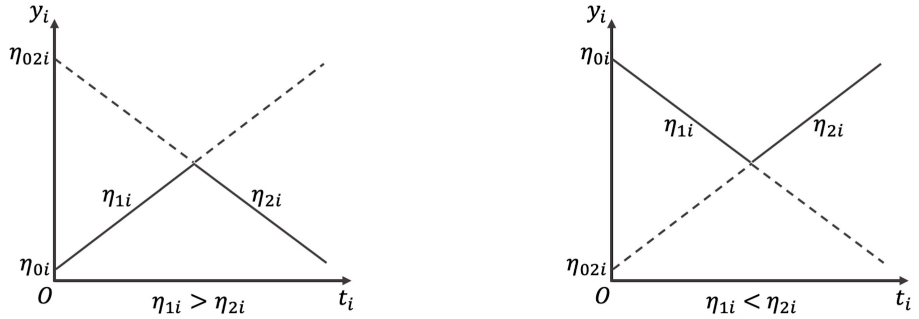


Figure A.1. Reparameterizing growth factors for Estimating a Fixed Knot

pre- and post-knot, we express y_{ij} as

$$\begin{aligned}
y_{ij} &= \min\left(\eta_{0i}+\eta_{1i}t_{ij}, \eta_{02i}+\eta_{2i}t_{ij}\right)\\
&= \frac{1}{2}\big(\eta_{0i}+\eta_{1i}t_{ij}+\eta_{02i}+\eta_{2i}t_{ij}-|\eta_{0i}+\eta_{1i}t_{ij}-\eta_{02i}-\eta_{2i}t_{ij}|\big)\\
&= \frac{1}{2}\big(\eta_{0i}+\eta_{1i}t_{ij}+\eta_{02i}+\eta_{2i}t_{ij}\big)-\frac{1}{2}\big(|\eta_{0i}+\eta_{1i}t_{ij}-\eta_{02i}-\eta_{2i}t_{ij}|\big)\\
&= \frac{1}{2}\big(\eta_{0i}+\eta_{02i}+\eta_{1i}t_{ij}+\eta_{2i}t_{ij}\big)-\frac{1}{2}\big(\eta_{1i}-\eta_{2i}\big)|t_{ij}-\gamma^{(k)}|\\
&= \eta_{0i}^{'}+\eta_{1i}^{'}\big(t_{ij}-\gamma^{(k)}\big)+\eta_{2i}^{'}|t_{ij}-\gamma^{(k)}|\\
&= \eta_{0i}^{'}+\eta_{1i}^{'}\big(t_{ij}-\gamma^{(k)}\big)+\eta_{2i}^{'}\sqrt{(t_{ij}-\gamma^{(k)})^2},
\end{aligned} \tag{A.1}$$

where $\eta_{0i}^{'}$, $\eta_{1i}^{'}$ and $\eta_{2i}^{'}$ are the measurement at the knot, the mean of two slopes, and the half difference between two slopes. Similarly, the measurement y_{ij} of the bilinear spline in the right panel, in which the measurement y_{ij} is always the maximum value of two lines, has the identical final form in Equation A.1.

A.2. Class-specific Transformation and Inverse-transformation between Two Parameter-spaces

Suppose $\boldsymbol{f}: \mathcal{R}^3 \rightarrow \mathcal{R}^3$ is a function, which takes a point $\boldsymbol{\eta}_i \in \mathcal{R}^3$ as input and produces the vector $\boldsymbol{f}(\boldsymbol{\eta}_i) \in \mathcal{R}^3$ (i.e., $\boldsymbol{\eta}_i^{'} \in \mathcal{R}^3$) as output. By the multivariate delta method (Lehmann & Casella, 1998, Chapter 1), for an individual in the k^{th} class

$$\boldsymbol{\eta}_i^{'} = \boldsymbol{f}(\boldsymbol{\eta}_i) \sim N\Bigg(\boldsymbol{f}(\boldsymbol{\mu_\eta}^{[k]}), \boldsymbol{\nabla_f}(\boldsymbol{\mu_\eta}^{[k]})\boldsymbol{\Psi_\eta}^{[k]}\boldsymbol{\nabla}^T_{\boldsymbol{f}}(\boldsymbol{\mu_\eta}^{[k]})\Bigg), \tag{A.2}$$

where $\boldsymbol{\mu_\eta}^{[k]}$ and $\boldsymbol{\Psi_\eta}^{[k]}$ are the mean vector and variance-covariance matrix of original class-specific growth factors, respectively, and $\boldsymbol{f}$ is defined as

$$\boldsymbol{f}(\boldsymbol{\eta}_i) = \begin{pmatrix}\eta_{0i}+\gamma^{[k]}\eta_{1i} & \frac{\eta_{1i}+\eta_{2i}}{2} & \frac{\eta_{2i}-\eta_{1i}}{2}\end{pmatrix}^T.$$

Similarly, suppose $\boldsymbol{h} : \mathcal{R}^3 \to \mathcal{R}^3$ is a function, which takes a point $\boldsymbol{\eta}_i^{'} \in \mathcal{R}^3$ as input and produces the vector $\boldsymbol{h}(\boldsymbol{\eta}_i^{'}) \in \mathcal{R}^3$ (i.e., $\boldsymbol{\eta}_i \in \mathcal{R}^3$) as output. By the multivariate delta method,

$$\boldsymbol{\eta}_i = \boldsymbol{h}(\boldsymbol{\eta}_i^{'[k]}) \sim N\Bigg(\boldsymbol{h}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{'[k]}), \boldsymbol{\nabla}_{\boldsymbol{h}}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{'[\boldsymbol{k}]})\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{'[k]}\boldsymbol{\nabla}_{\boldsymbol{h}}^{T}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{'[\boldsymbol{k}]})\Bigg), \qquad \text{(A.3)}$$

where $\boldsymbol{\mu}_{\boldsymbol{\eta}}^{'[k]}$ and $\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{'[k]}$ are the mean vector and variance-covariance matrix of class-specific reparameterized growth factors, respectively, and $\boldsymbol{h}$ is defined as

$$\boldsymbol{h}(\boldsymbol{\eta}_i^{'}) = \left(\eta_{0i}^{'} - \gamma^{[k]}\eta_{1i}^{'} + \gamma^{[k]}\eta_{2i}^{'} \ \ \eta_{1i}^{'} - \eta_{2i}^{'} \ \ \eta_{1i}^{'} + \eta_{2i}^{'}\right)^{T}.$$

Based on Equations (A.2) and (A.3), we can make the transformation between the growth factor means of two parameter-spaces by $\boldsymbol{\mu}_{\boldsymbol{\eta}}^{'[k]} = \boldsymbol{f}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{[k]})$ and $\boldsymbol{\mu}_{\boldsymbol{\eta}}^{[k]} = \boldsymbol{h}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{'[k]})$, respectively. We can also define the transformation matrix $\boldsymbol{\nabla}_{\boldsymbol{f}}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{[k]})$ and $\boldsymbol{\nabla}_{\boldsymbol{h}}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{'[\boldsymbol{k}]})$ between the variance-covariance matrix of two parameter-spaces as

$$\begin{aligned}\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{'[k]} &= \boldsymbol{\nabla}_{\boldsymbol{f}}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{[k]})\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[k]}\boldsymbol{\nabla}_{\boldsymbol{f}}^{T}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{[k]}) \\ &= \begin{pmatrix} 1 & \gamma^{[k]} & 0 \\ 0 & 0.5 & 0.5 \\ 0 & -0.5 & 0.5 \end{pmatrix} \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[k]} \begin{pmatrix} 1 & \gamma^{[k]} & 0 \\ 0 & 0.5 & 0.5 \\ 0 & -0.5 & 0.5 \end{pmatrix}^{T}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Psi}_{\boldsymbol{\eta}}{}^{[k]} &= \boldsymbol{\nabla}_{\boldsymbol{h}}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{'[\boldsymbol{k}]})\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{'[k]}\boldsymbol{\nabla}_{\boldsymbol{h}}^{T}(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{'[\boldsymbol{k}]}) \\ &= \begin{pmatrix} 1 & -\gamma^{[k]} & \gamma^{[k]} \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix} \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{'[k]} \begin{pmatrix} 1 & -\gamma^{[k]} & \gamma^{[k]} \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix}^{T},\end{aligned}$$

respectively.

B. More Results

Table B.1. Median (Range) of the Relative Bias over 1,000 Replications of Parameters of Interest under the Conditions with Random Knots of the Standard Deviation of 0.3 and 2 Latent Classes

	Para.	**Latent Class** 1	**Latent Class** 2
Mean	μ_{η_0}	−0.003 (−0.009, 0.003)	0.002 (0.000, 0.007)
	μ_{η_1}	0.008 (−0.009, 0.029)	−0.009 (−0.024, 0.007)
	μ_{η_2}	0.033 (0.007, 0.098)	−0.019 (−0.060, 0.001)
	μ_{γ}	−0.005 (−0.016, 0.004)	0.003 (−0.005, 0.013)
Variance	ψ_{00}	−0.001 (−0.069, 0.037)	−0.016 (−0.055, 0.006)
	ψ_{11}	−0.076 (−0.126, −0.040)	−0.030 (−0.083, −0.008)
	ψ_{22}	−0.015 (−0.061, 0.137)	−0.057 (−0.089, 0.179)
Path Coef.	β_0	—	−0.055 (NA, NA)
	β_1	—	−0.042 (−0.332, 0.013)
	β_2	—	−0.038 (−0.332, 0.019)

Note. —: when fitting the proposed model, we set the first latent class as the reference group; accordingly, the coefficients of that class do not exist. NA: Note that for the conditions with balanced allocation, the population value of $\beta_0 = 0$ and its relative bias goes infinity. The bias median (range) of β_0 is −0.015 (−0.204, 0.118).

Table B.2. Median (Range) of the Empirical SE over 1,000 Replications of Parameters of Interest under the Conditions with Random Knots of the Standard Deviation of 0.3 and 2 Latent Classes

	Para.	**Latent Class** 1	**Latent Class** 2
Mean	μ_{η_0}	0.432 (0.243, 0.892)	0.350 (0.200, 0.707)
	μ_{η_1}	0.106 (0.053, 0.294)	0.074 (0.042, 0.174)
	μ_{η_2}	0.103 (0.052, 0.280)	0.079 (0.042, 0.164)
	μ_{γ}	0.055 (0.024, 0.167)	0.062 (0.024, 0.198)
Variance	ψ_{00}	2.652 (1.731, 4.817)	2.201 (1.400, 3.789)
	ψ_{11}	0.123 (0.069, 0.272)	0.092 (0.057, 0.170)
	ψ_{22}	0.128 (0.071, 0.333)	0.101 (0.062, 0.219)
Path Coef.	β_0	—	0.182 (0.084, 0.592)
	β_1	—	0.120 (0.079, 0.186)
	β_2	—	0.124 (0.083, 0.199)

Note. —: when fitting the proposed model, we set the first latent class as the reference group; accordingly, the coefficients of that class do not exist.

Journal of Behavioral Data Science, 2021, *1* (2), 127–155.
DOI: https://doi.org/10.35566/jbds/v1n2/p5

GPS2space: An Open-source Python Library for Spatial Measure Extraction from GPS Data

Shuai Zhou[1], Yanling Li[1], Guangqing Chi[1], Junjun Yin[1], Zita Oravecz[1], Yosef Bodovski[1], Naomi P. Friedman[2], Scott I. Vrieze[3], and Sy-Miin Chow[1]

[1] The Pennsylvania State University, University Park, PA 16801, USA
sxz217@psu.edu
[2] University of Colorado Boulder, Boulder, CO
[3] University of Minnesota, Minneapolis, MN

Abstract. Global Positioning System (GPS) data have become one of the routine data streams collected by wearable devices, cell phones, and social media platforms in this digital age. Such data provide research opportunities in that they may provide contextual information to elucidate where, when, and why individuals engage in and sustain particular behavioral patterns. However, raw GPS data consisting of densely sampled time series of latitude and longitude coordinate pairs do not readily convey meaningful information concerning intra-individual dynamics and inter-individual differences; substantial data processing is required. Raw GPS data need to be integrated into a Geographic Information System (GIS) and analyzed, from which the mobility and activity patterns of individuals can be derived, a process that is unfamiliar to many behavioral scientists. In this tutorial article, we introduced GPS2space, a free and open-source Python library that we developed to facilitate the processing of GPS data, integration with GIS to derive distances from landmarks of interest, as well as extraction of two spatial features: activity space of individuals and shared space between individuals, such as members of the same family. We demonstrated functions available in the library using data from the Colorado Online Twin Study to explore seasonal and age-related changes in individuals' activity space and twin siblings' shared space, as well as gender, zygosity and baseline age-related differences in their initial levels and/or changes over time. We concluded with discussions of other potential usages, caveats, and future developments of GPS2space.

Keywords: Spatial Measure · Twins · Behavior Genetics · Latent Growth Curve Model · Python

1 Introduction

Spatial analysis is used to explain locations, attributes, and relationships of features in spatial data and has increasingly become a subject of interest in many social and behavioral science disciplines including psychology, sociology, demography, and environmental science (Chi & Zhu, 2019; Sui & Goodchild, 2011). The past three decades have witnessed the emergence and substantial growth of using spatial analysis to investigate environmental effects on behavioral changes and population dynamics. Many earlier analyses of spatial and mobility patterns were based mostly on self-reports, surveys, or administrative data (Chi & Marcouiller, 2013; Kestens et al., 2012; Vallée, Cadot, Roustit, Parizot, & Chauvin, 2011). For example, participants were usually asked to draw a map displaying their daily mobility patterns or provide locations they frequently visited in their daily routines. Recent advances in mobile technology tools (e.g., smartphones, wearable sensors) now allow researchers to collect physical location data in real-time over very short intervals (e.g., across seconds or minutes) (Kerr, Duncan, & Schipperjin, 2011; Kestens, Thierry, Shareck, Steinmetz-Wood, & Chaix, 2018; Russell, Almeida, & Maggs, 2017). Such intensive and continuous location data streams provide contextual information to elucidate the context in which (e.g., where, when, and why) individuals engage in and sustain particular behavioral and lifestyle patterns. However, the central focus of many studies in the social and behavioral sciences not only examines individuals' short-term spatial activities over hours or days, but also those that may extend over weeks, months, or even years, as well as across large populations. In such scenarios, as in the case of the Colorado Online Twin Study (CoTwins) used for demonstration in this study, the sheer quantity and density of the longitudinal Global Positioning System (GPS) data (approximately 6.65 million points from June 2016 to December 2018) make the spatial measure extraction via conventional and non-programmable spatial analysis tools highly impractical, inefficient, and irreproducible. In this article, we introduced GPS2space, a user-friendly Python package that can be used to facilitate and automate the processes of spatial data building, activity and shared space measure extraction, and fast distance query.

Myriad spatial and aspatial measures can be extracted from raw physical location data or social network data. One measure that has been found to be a useful lifestyle indicator is activity space, which has been used in studies of obesity, substance use, and mental health. Generally, these studies treat activity space as the space within which an individual engages in routine activities. This space measure may be quantified subjectively via individuals' self-reports (Buchowski, Townsend, Chen, Acra, & Sun, 1999), or objectively via location data (N. C. Lee et al., 2016). For example, using a representative sample from the Paris metropolitan area of France, Vallée et al. (2011) explored the relationship between depression and activity space as measured by individuals' daily activities. They found that depression was related to limited activity space and neighborhood characteristics such as deprivation status. Mason et al. (2010) constructed activity space from 301 Philadelphia adolescents' place-based social networks, and found that adolescents' substance use depended on their activity space, as moderated by participants' age and gender.

Another measure is shared space, which can be spatial or aspatial depending on disciplines and research questions. From a social science perspective, shared space refers to the socio-psychological or physical space within which individuals share a common identity and social belonging (Cleaveland & Kelly, 2008; Fine, 2012), or a common physical area. Studies have shown that shared space, such as coworking space shared by independent professionals, can provide social support (Gerdenitsch, Scheel, Andorfer, & Korunka, 2016). Shared space also increases neighborhood satisfaction and sense of community (Kearney, 2006).

In this study, we define an individual's activity space as the area of the minimum bounding geometry consisting of routine locations visited by the individual over a specific period of time (i.e., daily, weekly, or monthly). Accordingly, we define shared space as the overlapping areas of two individuals' activity spaces. Activity space depends on the spatial distributions of the geolocations: geolocations spanning larger areas and broader geographical regions would give rise to higher values of activity space. In contrast, geolocations that are concentrated around certain places such as home and working place would yield smaller activity space. Shared space is not necessarily linearly related to activity space because the latter is determined by the extent to which two individuals' activity spaces overlap with each other, in other words, how much they share the same area within their activity spaces.

Despite the richness of information available in location data, the mapping of raw data consisting of latitude and longitude coordinate pairs to landmarks of inferential interest requires reverse geocoding. Reverse geocoding is the process of converting machine-readable GPS coordinates into location information for geoprocessing, such as the nearest distance query, as well as specialized spatial feature extraction procedures (Yin et al., 2020). These procedures are typically implemented via specialized spatial software that may not be familiar or accessible to many social and behavioral scientists (McCormick, Lee, Cesare, Shojaie, & Spiro, 2017; Shelton, 2017; Shelton, Poorthuis, & Zook, 2015). Commercial software such as ArcGIS, TransCAD, and MapInfo (Drummond & French, 2008; Murray, Xu, Wang, & Church, 2019) are available and relatively easy to use. However, licensing restrictions may prevent broad dissemination of methodological advances and reproducibility of analytic results, and these programs are not readily available on High Performance Computing (HPC) platforms used to process data and perform large-scale analyses. ArcGIS and an open-source software, QGIS, are programmable, but their programming environments are not well developed. In contrast, R is an open-source programmable statistical language whose usage has been increasing in social and environmental sciences (Bivand, 2006). However, R poses known challenges in handling very large data sets, and often performs less satisfactorily in terms of memory management and computational speed (Patil, 2016). Taking into consideration computational speed, ease of usage, and open-source availability, we developed GPS2space in Python, a popular open-source programming language among researchers and data scientists.

The objectives of this tutorial are to introduce and demonstrate the use of GPS2space, a new, open-source Python library that we created to facilitate the construction of spatial data, simplify extraction of mobility-related measures such as activity space and shared space, and boost the nearest distance query for big data. GPS2space builds upon existing functions and includes all the necessary, tunable parameters as arguments for generating spatial measures in a straightforward and well-documented package that can be readily implemented by newer users. We used the terms library, package, and toolbox interchangeably throughout the article, as these terms all refer to reusable chunks of code but are used differently in different conventions. Likewise, we used the terms methods and functions interchangeably, in that they both refer to snippets of a library/package/toolbox that are used for specific purposes.

The remainder of the article proceeds as follows. First, we briefly introduce commonly used Python libraries for managing and analyzing GPS data and highlight the contributions of GPS2space. Then, we illustrate the utility of the GSP2space library using the CoTwins data to extract the twin siblings' activity space and shared space. These measures are used to address questions related to seasonal, age-based, gender, and zygosity effects in shaping individuals' activity space and shared space. Finally, we conclude with discussions on other potential usages, caveats, and future developments of GPS2space.

2 Contributions of GPS2space Relative to Other Commonly Used Spatial Python Packages

Like many data analysis procedures, geospatial analyses involve data reading and writing, data managing and processing, and visualization. Beyond that, geospatial analyses also deal with spatial projection and operation, Exploratory Spatial Data Analysis (ESDA), and spatial modeling. There are existing Python libraries that focus on certain specific functions useful for geospatial analysis – a brief overview is provided next.

Geospatial Data Abstraction Library (GDAL/OGR contributors, 2020) specializes in reading and writing raster and vector data, which are the two commonly used data types in GIS. It supports 168 raster data formats and 99 vector data formats at the time of writing (October 2020). Fiona (Gillies et al., 2011) and Rasterio (Gillies et al., 2013), two other popular libraries in Python, focus on reading, writing, and manipulating vector and raster data, respectively. Pyproj exclusively focuses on cartographic projections and coordinate transformations (Crickard, Toms, & Rees, 2018). Shapely specializes in spatial operations such as distance query and intersecting and overlapping analyses (Gillies et al., 2007). Python Spatial Analysis Library (PySAL) is the most commonly used library in conducting ESDA and spatial modeling (Rey, 2019; Rey & Anselin, 2007). GeoPandas, on the other hand, combines Pandas, a widely used Python data analysis library, and GIS science, providing a wide array of geospatial functions such as spatial operation, spatial projection transformation, and visualization (Jordahl, 2014). These packages are often used together to conduct a series of data managing, manipulation, visualization, and modeling tasks. For example, GeoPandas relies on Fiona to read and write spatial data and PyProj to perform spatial projection transformations. Rasterio also uses PyProj for its projection functionalities.

The packages reviewed thus far do have limitations, especially for novices who do not have strong background in programming and GIS. For example, Shapely does not provide options for coordinate system transformations, so the original units of distance and area measures are usually degrees, which may not be intuitive for non-specialist audiences. GeoPandas incorporates many useful geoprocessing methods and spatial analysis techniques and provides foundational functions for such spatial operations; however, it assumes users have GIS and programming background to perform the analyses. For example, to calculate the area of a polygon from GPS data with latitude and longitude coordinate pairs using GeoPandas, a researcher has to first build a spatial data set, project it to an appropriate coordinate reference system (CRS), and then calculate the area.

Even though we did not provide an exhaustive list of all the Python packages that can perform geospatial manipulation and analysis, we highlighted that almost all of these packages are tailored for experts with considerable spatial data handling and GIS experience, and require function customizations in multiple steps. For novices such multi-step data pre-processing and function customization processes can be challenging and error-prone. In addition, none of the above packages provides immediately available functions for constructing activity space and shared space.

In this article, we introduced GPS2space with the aim to facilitate and automate, whenever possible, the processes of spatial data building, activity and shared space measure extraction, and distance query. Specifically, GPS2space has three functionalities: (1) building unprojected spatial data from geolocations with latitude and longitude coordinate pairs using the geodf function; (2) constructing buffer- and convex hull-based activity space and shared space at different timescales using the space function; and (3) performing nearest distance query using the dist function, which incorporates cKDTree [1] and spatial indexing and R-Tree [2] algorithms to decrease execution time. GPS2space provides an easily replicable and open-source solution to building spatial data directly from latitude and longitude coordinate pairs. It also provides default parameterizations suited for many longitudinal spatial data streams that can be used to simplify and reduce the specification steps needed for extraction of activity- and shared-space-related and distance measures included in the package. GPS2space enables transparent and easily replicable ways to change these default options for experienced GIS scientists and programmers to perform custom specifications.

[1] cKDTree is a function from SciPy, a commonly used library for scientific computing in Python. cKDTree is used to rapidly look up the nearest neighbors of any point and can dramatically reduce the time needed for such processes.

[2] GeoPandas incorporated spatial indexing using the R-tree algorithm to boost the performance of spatial queries. R-tree is a tree-like data structure that groups nearby objects together along with their minimum bounding box. In this tree-like data structure, spatial queries such as finding the nearest neighbor does not have to travel through all geometries, dramatically increasing performance, especially for two data sets with different bounding boxes.

These spatial measures provide additional contextual information and expand the usages of GPS data. In sum, GPS2space provides an open-source tool to consolidate, simplify, and automate data processing and spatial measure extraction from large (e.g., intensive longitudinal) GPS data sets. In this way, replicability and reproducibility of results can be greatly enhanced – for veteran and novice researchers alike.

3 Motivating Data: The CoTwins Study

We used data from the CoTwins study to illustrate the utility of GPS2space and demonstrate how spatial activity measures can shed light on individual and dyadic activity patterns between twin siblings. Twin studies have the advantage of disentangling genetic and environmental factors for the trait of interest (Newman, Freeman, & Holzinger, 1937). Despite the increasing application of spatial thinking and spatial data in social and behavioral research, few twin studies have been designed to collect twins' location data, which often convey valuable information concerning social contexts. For instance, shared activity space and time spent with each other reflect opportunities for relationship bonding, and may thus convey the extent of emotional closeness between two individuals (Ben-Ari & Lavee, 2007). Furthermore, with twins' location data, it would be interesting to investigate how monozygotic (MZ; identical) twins and dizygotic (DZ; fraternal) twins differ in their shared activity space.

The CoTwins study comprises data on substance use among 670 twins. Twins were initially recruited at ages 14 to 17 and followed from 2015 to 2018. Throughout 2016 to 2018, the twins' geolocations were recorded and reported via their GPS enabled smartphones. iOS devices used the built-in significant-change location service to record and report geolocations whenever they detected a significant position change of 500 meters or more. Android devices recorded and reported geolocations every five minutes as long as the device was in use. Over the course of the study, the twins' spatial footprints covered locations within and outside of the United States. In this article, we only used locations in the contiguous United States, which includes the District of Columbia but excludes Alaska and Hawaii.

Figure 1 shows the spatial distribution of the twins' footprints in 2016, 2017, and 2018 across Colorado and the contiguous United States. The CoTwins study began collecting locations in June 2016 so the figure shows fewer data points in 2016. Throughout 2017 and 2018, the twins set foot in almost every state of the contiguous United States and showed a consistent pattern of footprints concentrated in Colorado and all over parts of the contiguous US, with North Dakota, Arkansas, and Alabama as the least visited states. In Colorado in 2017 and 2018 they showed consistent mobility patterns with geolocations clustered around metropolitan areas such as Denver and Colorado Springs and along major roads within the state. The border counties in Colorado such as Moffat, Rio Blanco, Yuma, Cheyenne, Kiowa, and Baca were rarely visited. The code for Figure 1 can be found in Supplementary Material.

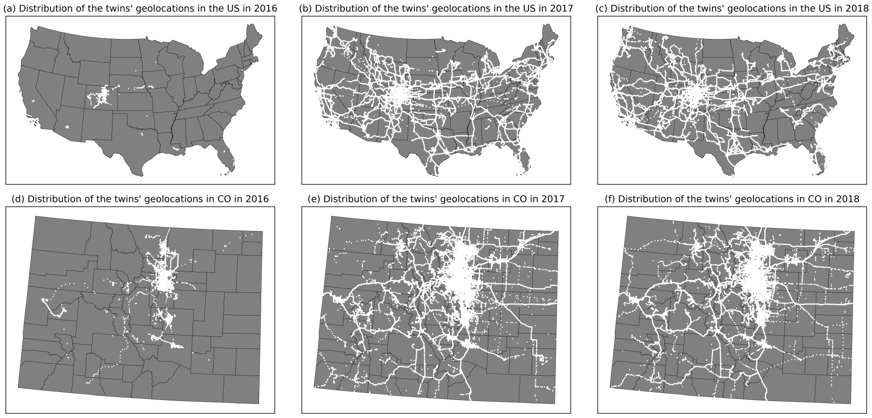

(a) Distribution of the twins' geolocations in the US in 2016
(b) Distribution of the twins' geolocations in the US in 2017
(c) Distribution of the twins' geolocations in the US in 2018
(d) Distribution of the twins' geolocations in CO in 2016
(e) Distribution of the twins' geolocations in CO in 2017
(f) Distribution of the twins' geolocations in CO in 2018

Figure 1. Distribution of geolocations in the contiguous United States and Colorado across 2016, 2017, and 2018 in the CoTwins study.

Many related works have demonstrated the spatial aspects of activity space and shared space and their impact on human behaviors such as substance use (Mason et al., 2010) and social support in a specific setting such as working space (Gerdenitsch et al., 2016); however, the temporal variations of such spatial measures and interindividual differences therein have not been thoroughly explored. Hence, we employed passive sensor (GPS) data to investigate whether meaningful seasonal, time- (e.g., weekend), and age-based variations, as well as between-individual differences in these intra-individual changes, could be meaningfully inferred from individuals' spatial measures as extracted using GPS2space. In particular, we examined (1) whether there were seasonal effects in twins' activity space/shared space; (2) whether there were weekend effects in twins' activity space/shared space; (3) inter-individual differences in initial levels of activity space/shared space, and possible associations with gender, baseline age, and twin type (MZ vs. DZ twins); and (4) age-related changes in activity space/shared space, and possible roles of gender as correlates of interindividual differences in these age-based changes.

4 Example I: Buffer- and Convex hull-based Activity Space and Shared Space

As previously defined, activity space refers to the area of individuals' routine locations over a specific time period. Practically, ellipses, convex hulls, and density kernels are often used to construct the activity space (Huang & Wong, 2016). The GPS2space library currently includes two commonly used methods for constructing activity space: the buffer method and the convex hull method. The buffer method uses a user-specified buffer distance as the radius in determining activity space, while the convex hull method lines up the outermost points to a minimum bounding geometry (J. H. Lee, Davis, Yoon, & Goulias, 2016) to represent activity space. Both buffer- and convex hull-based activity space approaches are associated with their own pros and cons. For buffer-based activity space, users have to specify a buffer distance to group and dissolve points into polygons to enable extraction of activity space. The choice of buffer distance can be arbitrary and application-specific, and it affects the sizes of activity space and shared space. However, this approach provides interpretable mobility estimates even with only one data point. In this case, activity space for that one data point is simply the area of the circle whose radius is the buffer distance. Importantly, it is less sensitive to extreme geolocations that are beyond the clusters of geolocation. Convex hull-based activity space does not require any arbitrary parameter. However, convex hull-based activity space computations require at least three non-collinear points to form an enclosed convex hull. In addition, convex hull-based activity space is sensitive to extreme geolocations, giving extreme activity space values in the presence of outliers. For example, instances where individuals travel via cars or flights from one main location to another would be outliers. The convex hull method would yield extreme activity space values in trying to construct a convex hull containing all the data points prior to, during, and after such travels, whereas the buffer-based method would use the user-specified buffer value to "group" the data points into clusters of points and compute activity and other spatial activity measures accordingly. We recommend that users consider their respective applications and contexts in detail when choosing between these two methods.

To illustrate how buffer- and convex hull-based activity space and shared space are obtained from raw GPS data with latitude and longitude coordinate pairs, we used one randomly selected twin pair, denoted herein as TwinX, and their geolocations on May 12, 2017. For buffer-based activity space, we used a buffer distance of 1000 meters based on common choices of buffer distance in other published studies (Perchoux, Chaix, Brondeel, & Kestens, 2016; Stewart et al., 2015). The process of computing activity and shared spaces can be grouped largely into 3 steps. We described each step and provided the associated code as organized by these steps.

Step 1: Conversion of raw GPS data into spatial data.

To perform spatial operations, we need to first convert raw GPS data with latitude and longitude coordinate pairs to spatial data using the *df_to_gdf* function in the GPS2space library. The *df_to_gdf* function takes three parameters: the first one is the Pandas dataframe [3] that contains GPS data with geolocation information as represented by latitude and longitude coordinate pairs; the second one is the column name of the longitude information; the third one is the column name of the latitude information. The *df_to_gdf* function returns an un-

[3] Pandas is a commonly used library for data manipulation analysis in Python. A Pandas dataframe is a 2-dimensional data structure with rows representing observations and columns representing variables. A column can have different data types in a Pandas dataframe.

projected GeoPandas dataframe [4] in the World Geodetic System 84 (WGS84). The following code imports the required libraries for the process, then reads in latitude and longitude coordinate pairs stored in two csv files comprising the two twin members' respective data, TwinXa_512.csv and TwinXb_512.csv, and finally converts the non-spatial dataframe to spatial data using the *df_to_gdf* function. One important note is that users must pass the longitude column name to x and the latitude column name to y.

```
# Import required libraries for the analyses.
import pandas as pd
import geopandas as gpd
from gps2space import geodf, space, dist

# Read TwinXa_512 and TwinXb_512 as Pandas dataframes.
df_twinXa_512 = pd.read_csv(‘./data/TwinXa_512.csv’)
df_twinXb_512 = pd.read_csv(‘./data/TwinXb_512.csv’)

# Convert Pandas dataframes to GeoPandas dataframes.
gdf_twinXa_512 = geodf.df_to_gdf(df_twinXa_512, x=‘
    longitude’, y=‘latitude’)
gdf_twinXb_512 = geodf.df_to_gdf(df_twinXb_512, x=‘
    longitude’, y=‘latitude’)
```

Step 2: Spatial projection and spatial measure extraction of activity space.

After successful data conversion, the next step is to project the spatial data and calculate buffer- and convex hull-based activity space using the *space.buffer_space* and *space.convex_space* functions, respectively. The *buffer_space* takes four parameters: the first is the unprojected GeoPandas dataframe; the second is a user-defined buffer distance dist, where the default value is 0; the third is dissolve, the user-specified level of timescale at which the geolocations are aggregated to form polygons, where the default value is "week"; the fourth is *proj*, the user-specified EPSG identifier [5] based on the selected spatial data for projection. The default value for *proj* is 2163 (US National Atlas Equal Area projection), a commonly used projection for the US. The *buffer_space* function returns a GeoPandas dataframe with a "buff_area" column representing the buffer-based activity space. The *proj* parameter specifies the unit for activity space, shared space, and buffer distance in the *buffer_space* function. For instance, the unit of EPSG 2163 is meter, so the unit for dist is meter; accordingly, the unit for activity space and shared space is square meter. We recommend that users choose a meter-based projection system because it provides more intuitive measurement

[4] A GeoPandas dataframe is an extension of Pandas dataframe with a "geometry" column storing geolocation information.

[5] EPSG identifiers are codes representing different spatial reference systems that can be used to project, reproject, and transform between different spatial reference systems. For example, the EPSG: 4326 is the default spatial reference system used by GPS, the EPSG: 3857 is used by Google Map and OpenStreetMap.

units than a degree-based projection system. [6] As mentioned above, the buffer distance in the *buffer_space* function is an application-specific parameter, users can refer to Browning and Lee (2017), K. Lee and Kwan (2019), Sugiyama, Kubota, Sugiyama, Cole, and Owen (2019), and Prins et al. (2014) for discussion on selecting buffer distances and their impacts on the study involved.

The *convex_space* takes three parameters: the first is the unprojected GeoPandas dataframe; the second is *group*, the level of timescale at which users want to group geolocations to form polygons, where the default value is "week"; the third is the EPSG identifier, where the default value is 2163. The *convex_space* function returns a GeoPandas dataframe with a "convx_area" column representing the convex hull-based activity space. When constructing activity space, the timescale should either be one of the variables in the dataframe, or it can be inferred and included as a variable in the dataframe from the timestamp when the geolocations are recorded. In the following example, we constructed TwinXa and TwinXb's daily activity space on May 12, 2017, and the variable "day" is inferred from the twin pairs' timestamps ranging from 5/12/2017 at 07:25 to 5/12/2017 at 20:10.

```
# Project spatial data.
gdf_twinXa_512 = gdf_twinXa_512.to_crs('epsg:2163')
gdf_twinXb_512 = gdf_twinXb_512.to_crs('epsg:2163')

# Buffer- and convex hull-based activity space.
buff_twinXa_512 = space.buffer_space(gdf_twinXa_512,
   dist=1000, dissolve='day', proj=2163)
buff_twinXb_512 = space.buffer_space(gdf_twinXb_512,
   dist=1000, dissolve='day', proj=2163)
convex_twinXa_512 = space.convex_space(gdf_twinXa_512,
   group='day', proj=2163)
convex_twinXb_512 = space.convex_space(gdf_twinXb_512,
   group='day', proj=2163)
```

Step 3: Extraction of shared space by overlaying activity space features.

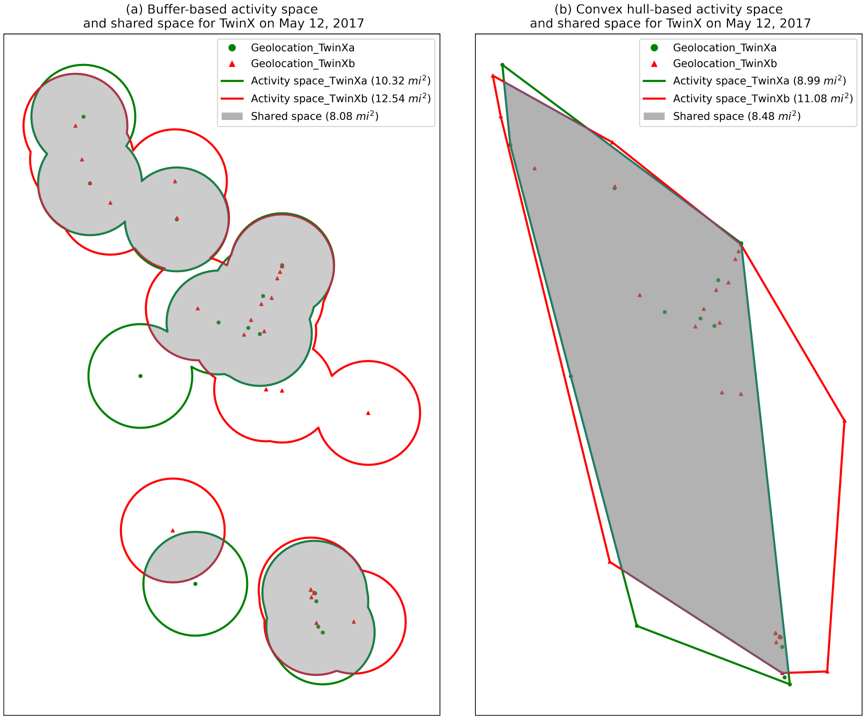
Once we have the activity space, we can utilize the *overlay* function from GeoPandas to calculate shared space by overlaying and intersecting the activity spaces of two individuals. For instance, in the following code example, we overlaid the buffer- and convex hull-based activity space. We specified "intersection" for the how parameter to extract the intersection area between the twins' activity space. We then invoked the *area* function to obtain a column named "share_space," representing the areas of the twins' shared space. A loop to iterate over multiple activity space features to obtain shared space between one another is provided in Supplementary Material.

```
# Calculate shared space from activity space.
buff_share = gpd.overlay(buff_twinXa_512,
   buff_twinXb_512, how='intersection')
```

[6] For the unit of different projection systems, see https://epsg.io/.

```
buff_share[‘share_space’] = buff_share[‘geometry’].area

convex_share = gpd.overlay(convex_twinXa_512,
    convex_twinXb_512, how=‘intersection’)
convex_share[‘share_space’] = convex_share[‘geometry’].
    area
```

Figure 2 shows the buffer- and convex hull-based activity space and shared space for TwinX on May 12, 2017. The buffer-based approach using 1000 meters as buffer distance gives an activity space of 10.32 and 12.54 square miles [7] for TwinXa and TwinXb, and a shared space of 8.08 square miles between them. The convex hull-based approach produces an activity space of 8.99 and 11.08 square miles for each individual of TwinX and a shared space of 8.48 square miles between them. The code for Figure 2 can be found in Supplementary Material.


(a) Buffer-based activity space and shared space for TwinX on May 12, 2017

(b) Convex hull-based activity space and shared space for TwinX on May 12, 2017

Geolocation_TwinXa
Geolocation_TwinXb
Activity space_TwinXa (10.32 mi^2)
Activity space_TwinXb (12.54 mi^2)
Shared space (8.08 mi^2)
Geolocation_TwinXa
Geolocation_TwinXb
Activity space_TwinXa (8.99 mi^2)
Activity space_TwinXb (11.08 mi^2)
Shared space (8.48 mi^2)

Figure 2. (a) Buffer-based activity space and shared space for TwinX on May 12, 2017 in Colorado. (b) Convex hull-based activity space and shared space for TwinX on May 12, 2017 in Colorado.

[7] For illustration purposes, we converted area measurement in square meters to square miles.

5 Example II. The Nearest Distance Query

The nearest distance measure is a useful indicator of accessibility of infrastructures and places that would influence behavioral and socioeconomic outcomes. For example, research has shown that the distance to the ballot drop box influences voters' turnout (McGuire, O'Brien, Baird, Corbett, & Collingwood, 2020), and access to highways affects population distribution (Chi, 2010). However, the nearest distance query can be computationally demanding and time-consuming, especially for processing data in large volumes. To boost the nearest distance query, the *dist* function in the GPS2space library incorporates two types of spatial indices to rapidly look up the nearest neighbor and calculate the distance. When the geometries of target features are points, *dist_to_point* in the dist function utilizes the cKDTree from SciPy to search for nearest neighbors; when the geometries of target features are polygons, *dist_to_poly* in the dist function utilizes the R-Tree from Geopandas to search for nearest neighbors. Both cKDTree and R-Tree algorithms create tree-like data structures from the Geopandas dataframe which enable fast nearest neighbor searching, therefore working efficiently with data sets in large volumes.

We used TwinXa's geolocations on May 12, 2017 to demonstrate the utility of the dist function and calculated the distance from each unique location to its nearest supermarket (represented as points) and park (represented as polygons) in Colorado. The supermarket and park data were obtained from OpenStreetMap (OSM). The OSM started in 2004 and its main goal is to collect and provide free access to geospatial data. The initial focus was on transportation infrastructure (streets, highways, railways, etc.), but data collection has expanded to multiple points of interest, such as buildings and community landmarks. Since most commercial data sources are expensive and have data sharing restrictions, OSM has quickly become a popular data source for geospatial-related research.

We downloaded and compiled the OSM data from Geofabrik [8], a Germany-based company specializing in processing and reorganizing free geodata created by projects like OSM. There are some concerns, however, about the quality of OSM data. For example, studies have shown that there were some disparities in data quality between urban/densely populated areas and rural/sparsely populated areas in OSM (Barron, Neis, & Zipf, 2014). In this study we compared OSM data with a high quality commercial data source called Infogroup Business Dataset, which contains more than 15 million geocoded business locations in the US. We found that the OSM data provided solid coverage when it came to major retail chains and good positional accuracy for corresponding locations. For example, comparing Infogroup and OSM data for the major Colorado supermarket chain "Safeway," 94% of the Safeway locations contained in the OSM data were also found in Infogroup. We also found similar results for two other major retail chains – "King Soopers" and "Whole Foods."

The *dist_to_point* function takes three parameters: the first one is the source GeoPandas dataframe; the second one is the target GeoPandas dataframe; and

[8] See https://www.geofabrik.de/geofabrik/geofabrik.html.

the third one is the EPSG identifier, with a default value of 2163. When *dist_to_point* function is called, the nearest neighbor search is then performed by traversing the cKDTree created on the spatial points in the target data set, which only deals with a subset of the points for the distance calculation. As shown in the following code example, we first constructed the spatial data set for the supermarket data, then we provided three parameters to the *dist_to_point* function for the nearest distance query from the TwinXa to supermarkets. The “dist” is the outcome GeoPandas dataframe with a “dist2point” column showing the distance from the source point to its nearest supermarket. All the columns from both the source and target dataframes are preserved in the outcome dataframe.

```
# Read market data into Pandas dataframes.
df_market = pd.read_csv(‘./data/market.csv’)

# Convert Pandas dataframes to GeoPandas dataframes.
gdf_market = geodf.df_to_gdf(df_market, x=‘longitude’,
    y=‘latitude’)

# The nearest distance from twinXa_512 to supermarket.
dist = dist.dist_to_point(gdf_twinXa_512, gdf_market,
    proj=2163)
```

The *dist_to_poly* function is similar to the *dist_to_point* function, except that the nearest neighbors in the *dist_to_poly* function are polygons. The *dist_to_poly* function takes four parameters: the first one is the source GeoPandas dataframe; the second one is the target GeoPandas dataframe; the third one is the EPSG identifier, with a default value of 2163; and the fourth one is a search radius in meters, with a default value of None. If the search radius is not specified, the *dist_to_poly* function employs a brute-force search to find the nearest distance, and the computation time increases significantly as the number of polygons grows. If the search radius is specified, R-tree is implemented by creating a minimum bounding box (MBR) for each target polygon. Instead of calculating the distance from the source point to every polygon in the target dataframe, the *dist_to_poly* function takes advantage of the R-tree index to only consider those polygons whose MBRs intersected with the search radius and calculate the minimum distance. If no polygon is within the search radius, then the *dist_to_poly* function returns a *NaN* value, a common way to represent missing values in Python. The *dist_to_poly* function works efficiently in calculating the nearest distance by specifying a search radius, but at the expense of missing values for points with no neighbors within the search radius. We recommend choosing an appropriate search radius based on how it can affect specific research designs.

As shown in the following code example, we read the park data in the form of shapefiles as GeoPandas dataframe, then we provided the parameters to the *dist_to_poly* function for the nearest distance query from the TwinXa to parks. The “dist_no_radius” and “dist_with_radius” are the outcome GeoPandas dataframes with a “dist2poly” column showing the distance from the source point to its nearest park.

```
# Read parks as GeoPandas dataframes.
gdf_parks = gpd.read_file ('./data/parks.shp')

# The nearest distance without search radius.
dist_no_radius = dist.dist_to_poly (gdf_twinXa_512,
    gdf_parks, proj=2163)

# The nearest distance with search radius of 5000m.
dist_with_radius = dist.dist_to_poly(gdf_twinXa_512,
    gdf_parks, proj=2163, search_radius=5000)
```

The two functions, *dist_to_point* and *dist_to_poly*, serve to provide distance measures geared respectively toward places of interest that are adequately represented as points (typically places covering smaller geographical regions such that the centroids of their enclosing polygon provide a reasonable representation, such as supermarkets, transportation terminals, and health facilities) vs. polygons (typically geographically dispersed places of interest or places that require precise definitions of boundaries, such as parks, water bodies, and administrative boundaries). Results from *dist_to_poly* and *dist_to_point* do not always agree, mainly because *dist_to_poly* and *dist_to_point* treat points within polygons differently. To illustrate the differences, we calculated the nearest distance from TwinX to the nearest park, playground, and supermarket (represented as polygons, search radius not specified) and their centroids (represented as points). Table 1 shows the results. Overall, the two functions produce similar results except for differences in minimum distance, where *dist_to_poly* may produce 0 values while *dist_to_point* rarely produces 0 values. The main reason for the differences in the minimum distance is that once *dist_to_poly* detects the point is within the polygon it assigns 0 to the nearest distance, while *dist_to_point* calculates the Euclidean distance between the two points and only returns 0 if the geolocations of the two points are identical. In sum, the distance measure between *dist_to_point* and *dist_to_poly* depends on the source data's relative position to the target polygon and the shape of the target polygon. The code for Table 1 can be found in Supplementary Material.

Table 1. Comparison between the nearest distance from TwinX to polygon boundary and polygon centroid for parks, playgrounds, and supermarkets in Colorado

Nearest distance to landmark measure	Mean	SD	Min	Median	Max
Distance to park (point)	0.57	0.39	0.01	0.46	6.60
Distance to park (polygon)	0.50	0.38	0.00	0.42	6.46
Distance to playground (point)	0.84	0.53	0.01	0.93	8.29
Distance to playground (polygon)	0.83	0.53	0.00	0.92	8.29
Distance to supermarket (point)	1.30	1.06	0.01	0.98	9.36
Distance to supermarket (polygon)	1.27	1.06	0.00	0.95	9.33

Note. The original distance measures were in meters, we converted them to miles for illustration purposes.

6 Example III. Growth Curve Analysis of Activity and Shared Spaces

6.1 Data Pre-Processing

Before extracting the daily activity space and shared space for all participants using the functions presented above, we pre-processed the GPS data following procedures implemented in the previous study (Li et al., in press). First, we excluded records with fewer than 20 valid data points within a week because these unusually low numbers of GPS points lacked sufficient variability. Then we excluded data points showing atypical travel trajectories as detected by *dbscan* (Density-Based Spatial Clustering of Applications with Noise), an R package that is commonly used to identify clusters and outlying points (Hahsler, Piekenbrock, & Doran, 2019). Then the daily activity space was calculated using a buffer distance of 1000 meters and transformed from square meters to square miles for illustrative purposes. The activity space was then log transformed to reduce skewness in the data. The log transformed activity space was referred to hereafter as LAS. For each participant, we focused on the proportion of shared space, referred to as PSS hereafter and defined as the proportion of one's daily activity space that overlapped with his/her twin sibling's daily activity space. The distributions of LAS and PSS were shown in Figure 3. The final data set consisted of 558 participants with baseline ages between 14 and 20 (mean = 17), followed between 1 to 3 years (mean = 2). 43% of the participants were males. In terms of twin types, 33% were MZ twins, 41% were DZ twins of the same sex, and 26% were DZ twins of opposite sex.

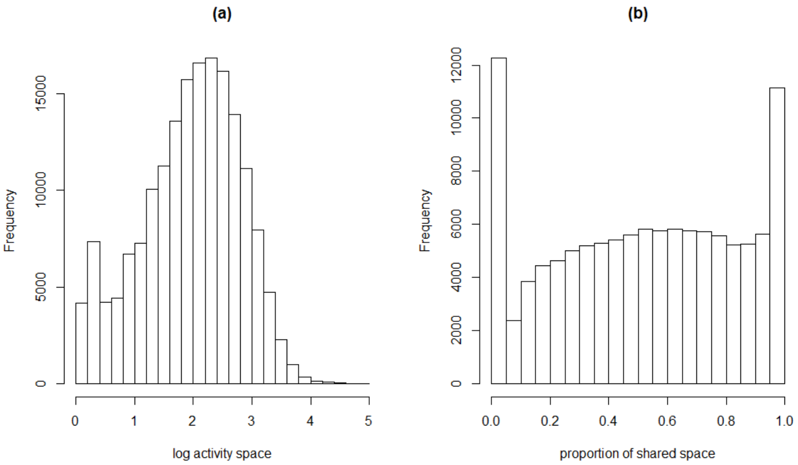


Figure 3. Distributions of (a) log activity spaces (LAS) and (b) proportions of shared space (PSS) across participants.

6.2 Data Analytic Plans

As mentioned before, we were interested in exploring within-individual changes of LAS and PSS and inter-individual differences in their initial levels and changes over time, including both between-individual and between-family differences. At the within-individual level, we sought to address the seasonal effect (research question 1), the weekend effect (research question 2), and age-related changes (research question 4) in LAS and PSS; at the between-individual level, we sought to explore gender differences in the initial levels and changes of LAS and PSS, as well as the effect of baseline ages on the initial levels (research questions 3-4); at the between-family level, we investigated the effect of twin zygosity (MZ vs. DZ twins) on the initial levels of PSS (research question 3). Therefore, we used three-level growth curve models (see, e.g., Enders & Tofighi, 2007; Hoffman, 2015) as implemented using the R package, *brms* (Bürkner, 2017), to study these temporal changes and levels of nesting within this data set, namely, time nested within individuals within family. In particular, we used seasonal and weekend indicators, as well as participants' ages as within-individual (or so-called level-1) predictors, gender and baseline age as between-individual (level-2) predictors, and twin zygosity as a between-family (level-3) predictor when relevant to address our questions of interest. The R code for model fitting can be found in Supplementary Material.

We first introduced the model for LAS, as shown below.

Level-1 model:

$$LAS_{itk} = \beta_{0ik}+\beta_{1ik}Age_{itk}+\beta_2 Weekend_t+\beta_3 Summer_t+\beta_4 Fall_t+\beta_5 Winter_t+e_{itk} \tag{1}$$

Level-2 model:

$$\beta_{0ik} = \gamma_{00k} + \gamma_{01k}Gender_{ik} + \gamma_{02k}Age_{i0k} + u_{0ik} \tag{2}$$

$$\beta_{1ik} = \gamma_{10k} + \gamma_{11k}Gender_{ik} + u_{1ik} \tag{3}$$

Level-3 model:

$$\gamma_{00k} = \delta_{000} + v_{0k} \tag{4}$$

$$\gamma_{01k} = \delta_{010} + v_{1k} \tag{5}$$

$$\gamma_{02k} = \delta_{020} + v_{2k} \tag{6}$$

$$\gamma_{10k} = \delta_{100} + v_{3k} \tag{7}$$

$$\gamma_{11k} = \delta_{110} + v_{4k} \tag{8}$$

with,

$$
\begin{aligned}
e_{itk} &\sim N(0, \sigma^2),\\
\begin{bmatrix} u_{0ik} \\ u_{1ik} \end{bmatrix} &\sim MN(\mathbf{0}, T = \begin{bmatrix} \tau_0^2 & \\ \tau_{01} & \tau_1^2 \end{bmatrix}),\\
\begin{bmatrix} v_{0k} \\ v_{1k} \\ v_{2k} \\ v_{3k} \end{bmatrix} &\sim MN(\mathbf{0}, \mathbf{\Phi} = \begin{bmatrix} \varphi_0^2 & & & & \\ \varphi_{01} & \varphi_1^2 & & & \\ \varphi_{02} & \varphi_{12} & \varphi_2^2 & & \\ \varphi_{03} & \varphi_{13} & \varphi_{23} & \varphi_3^2 & \\ \varphi_{04} & \varphi_{14} & \varphi_{24} & \varphi_{34} & \varphi_4^2 \end{bmatrix})
\end{aligned}
$$

The seasonal effect, weekend effect, and age-based changes in LAS were modeled in the level-1 model, where LAS_{itk} was the LAS of person i in family k on day t, and Age_{itk} was the age of person i in family k on day t, centered by subtracting the baseline age from each age instance so that 0 corresponded to the baseline age. The *Weekend*, *Summer*, *Fall*, and *Winter* variables were dummy-coded, with 1 each representing weekend, summer (June 1 to August 30), fall (September 1 to November 30), and winter (December 1 to February 28 or 29). Based on the definitions of these variables, β_{0ik} represented person i's initial LAS at baseline age on Spring weekdays; β_{1ik} was the effect of age on the LAS for person i; and β_j $(j = 2, \ldots, 5)$ represented weekend or seasonal effects, which were not set as person-specific since we focused on the overall seasonal and weekend effects in this study. Finally, the level-1 error e_{itk} followed a normal distribution with a zero mean and a variance of σ^2.

In the level-2 model, the level-1 parameters, β_{0ik} and β_{1ik}, were regressed on a *person-specific* variable, $Gender_{ik}$ (1 = male; -1 = female), to explore gender differences in the initial levels and age-based changes of LAS. In addition, β_{0ik} was regressed on the baseline age, Age_{i0k}, centered by subtracting the mean of baseline ages so that 0 corresponded to the average baseline age. Thus, the corresponding coefficient γ_{02k} represented the effect of baseline ages on the initial LAS, and γ_{00k} and γ_{10k} represented the overall initial level and growth rate of LAS across individuals, respectively, while $2\gamma_{01k}$ and $2\gamma_{11k}$ represented the corresponding gender differences, respectively. The level-2 random effects were denoted as u_{0ik} and u_{1ik}, which described person i's deviations in the values of β_{0ik} and β_{1ik} not accounted for by the predictors. Finally, the variance and covariance structure of level-2 random effects was defined in $\mathbf{T}$. For instance, the variance of β_{0ik}, denoted as τ_0^2, described the extent of between-individual difference in the initial LAS; the covariance between β_{0ik} and β_{1ik}, denoted as τ_{01}, described the relationship between initial levels and growth rates of LAS.

The level-3 model was built to capture between-family differences. Specifically, we would like to investigate whether twins from different families would have different initial levels and growth rates of LAS and whether the effects of gender and baseline age on the initial levels and/or growth rates of LAS would differ across families as well. Note that twin type was not included as a predictor in the level-3 model because the magnitudes of activity space were not expected to be significantly different between MZ and DZ twins (although they might be expected to differ in the degree to which they share space with their siblings, which was addressed below in the model for PSS). Among parameters in the level-3 model, δ_{010} and δ_{110} were of particular interest because they reflected the differences between males and females in terms of their average initial levels and growth rates of LAS, respectively. The level-3 random effects, v_{0k} - v_{4k}, followed a multivariate normal distribution with zero means and a covariance matrix, $\mathbf{\Phi}$, where the variances, denoted as φ_0^2 - φ_4^2, captured the extent of between-family differences in the overall initial LAS, the effects of gender and baseline age on the initial LAS, the overall growth rate of LAS and gender differences therein, respectively.

In terms of the model for PSS, some slight modeling adaptations were needed to capture characteristics of the PSS data. As noted, PSS was defined as the proportion of one's activity space that overlapped with his/her twin sibling's activity space, thus yielding a value ranging from 0 to 1. The model presented above, which assumed that the error term followed a normal distribution with a constant variance, might not be appropriate for the data in this scenario. However, the beta distribution is known for its flexibility in modeling proportions because its density can display different shapes as decided by the values of α and β. The beta density can be expressed as:

$$f(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1},\ 0 < y < 1, \alpha > 0, \beta > 0 \tag{9}$$

Thus, in the generalized growth curve model with PSS as the dependent variable, PSS was specified to conform to a beta distribution. Consistent with the beta regression specification proposed by Ferrari and Cribari-Neto (2004), which is similar to that of the well-known class of generalized linear models (McCullagh & Nelder, 1989), we defined $\mu = \alpha/(\alpha + \beta)$ and $\phi = \ \alpha + \beta$, then $E(y) = \mu$ and $Var(y) = \mu(1 - \mu)/(1 + \phi)$, where μ was the mean and ϕ was called the precision parameter. In our case, we assumed that the PSS, PSS_{itk}, followed a beta distribution with person-specific means (i.e., $E(PSS_{itk}) = \mu_{itk}$). Then we implemented a logit transformation of μ_{itk} and built a three-level growth curve model on the transformed value (i.e., η_{itk}). The level-1 model for PSS was specified as:

$$\begin{aligned} \eta_{itk} &= log(\frac{\mu_{itk}}{1-\mu_{itk}}) \\ &= \beta_{0ik} + \beta_{1ik} Age_{itk} + \beta_2 Weekend_t + \beta_3 Summer_t + \beta_4 Fall_t + \beta_5 Winter_t \end{aligned} \tag{10}$$

where $\frac{\mu_{itk}}{1-\mu_{itk}}$, denoted below as the odds of PSS, represented the average level of PSS for individual i in family k at time t relative to not sharing space with twin siblings, and $\eta_{itk} = log(\frac{\mu_{itk}}{1-\mu_{itk}})$ represented the corresponding log odds. The independent variables were as summarized in Equation 1. Note that the regression coefficients had different interpretations due to the logit transformation. For instance, β_{0ik} represented the log-odds of PSS for person i in family k at the baseline age on Spring weekdays; β_{1ik} was the age-related log-odds ratio, which means that the odds of PSS would multiply by $e^{\beta_{1ik}}$ for every 1-unit increase in Age_{itk}. Other parameters (e.g., seasonal and weekend effects) can be interpreted in a similar way.

The level-2 model for PSS was identical to the level-2 model for LAS (see Equations 2 - 3), but the regression coefficients had different interpretations for the reason stated above. For instance, the level-2 intercept, γ_{00k}, represented the overall log-odds of PSS.

In terms of the level-3 model, we hypothesized that MZ and DZ twins might have different levels of space sharing to the extent that these spatial measures reflect genetically influenced behavior/preferences. To evaluate this hypothesis, we added a predictor, twin type, to Equations 4 - 6 (i.e., the level-3 model for γ_{00k}, γ_{01k}, and γ_{02k}, which were the coefficients in the level-2 model for β_{0ik}, the log-odds of initial levels of PSS), to investigate zygosity differences in PSS and how these differences might affect the effects of gender and baseline age on PSS, as shown below.

$$\gamma_{00k} = \delta_{000} + \delta_{001} DZSS_k + \delta_{002} DZOS_k + v_{0k} \tag{11}$$

$$\gamma_{01k} = \delta_{010} + \delta_{011} DZSS_k + \delta_{012} DZOS_k + v_{1k} \tag{12}$$

$$\gamma_{02k} = \delta_{020} + \delta_{021} DZSS_k + \delta_{022} DZOS_k + v_{2k} \tag{13}$$

Specifically, we set MZ twins as the reference and added two dummy-coded, *family-specific* variables, $DZSS_k$(1 = DZ twins of the same sex) and $DZOS_k$ (1 = DZ twins of opposite sex). Thus, δ_{000}, δ_{001}, and δ_{002} represented the average log-odds of PSS for MZ twins, DZ twins of the same sex, and DZ twins of the opposite sex, respectively; δ_{010}, δ_{011}, and δ_{012} represented the corresponding gender differences in each twin type group; and δ_{020}, δ_{021}, and δ_{022} represented the effect of the baseline age on the average log-odds of PSS in each twin type group. The models for other level-2 parameters (i.e., γ_{10k}, γ_{11k}) were identical to the level-3 model for LAS (see Equations 7 - 8).

6.3 Results

With the *brms* package, the models were fitted in a Bayesian framework using Markov chain Monte Carlo (MCMC) methods. Specifically, we ran two chains, each with 5000 iterations in total and a burn-in of 2000 (discarded) iterations. On an Intel i5-8350U, 16GB RAM, Windows 10 computer, it took about 40 hours to run each model. Two diagnostic statistics were used to check the sampling quality (Gelman et al., 2013): (1) the effective sample size (ESS), which describes how many posterior draws in the MCMC procedure can be regarded as independent, and (2) $\hat{R}$, which describes the ratio of the overall variance of posterior samples across chains to the within-chain variance. The diagnostic criteria for adequate sampling and convergence were set as ESS greater than 800 and $\hat{R}$ below 1.1, respectively. Results showed that ESS was greater than 800 for most parameters, except for some random effect standard deviation parameters (e.g., $\varphi_1 - \varphi_4$), for which the average ESS was about 400, which can be deemed satisfactory. $\hat{R}$ was below 1.1 for all parameters in both models.

Table 2. Parameter estimates of the model for LAS from the CoTwins study, 2016-2018

Parameter	Estimate	SE	95% CI
Fixed effects			
Intercept, δ_{000}	1.81	0.03	[1.76, 1.86]
Gender, δ_{010}	-0.07	0.02	[-0.11, -0.01]
Baseline age, δ_{020}	0.13	0.02	[0.09, 0.16]
Age, δ_{100}	-0.01	0.01	[-0.03, 0.02]
Age*Gender, δ_{110}	0.01	0.01	[-0.01, 0.03]
Weekend, β_2	0.06	0.00	[0.05, 0.07]
Summer, β_3	0.07	0.00	[0.06, 0.07]
Fall, β_4	-0.12	0.00	[-0.13, -0.11]
Winter, β_5	-0.08	0.01	[-0.09, -0.07]
Level-2 random effects			
Intercept standard deviation, τ_0	0.25	0.01	[0.22, 0.28]
Age standard deviation, τ_1	0.19	0.01	[0.16, 0.22]
Intercept-Age correlation, $\tau_{01}/(\tau_0 * \tau_1)$	-0.31	0.08	[-0.46, -0.16]
Level-3 random effects			
Intercept standard deviation, φ_0	0.37	0.02	[0.33, 0.42]
Age standard deviation, φ_3	0.11	0.03	[0.06, 0.16]
Residual standard deviation, σ	0.72	0.00	[0.71, 0.72]

Note. SE = standard errors estimated by standard deviations of the posterior samples; CI = credible interval. N = 558 participants. The number of time points for each participant ranged from 3 to 569.

Table 2 shows the parameter estimates for LAS. In terms of the fixed effects, weekend and seasonal effects were found in the trajectory of LAS. Specifically, the participants showed greater LAS values on weekends than on weekdays ($\beta_2 = 0.06,\ \ 95\%\ CI = [0.05,\ 0.07]$), which was reasonable since most of the participants were supposed to be spending most of their time in school on weekdays, thus yielding limited activity space. Seasonally, the participants tended to display greater LAS in summer ($\beta_3 = 0.07,\ \ 95\%\ CI = [0.06,\ 0.07]$), which was likely due to summer break as well as the warmer weather. Gender differences were found in the initial levels of LAS ($\delta_{010} = -0.07,\ \ 95\%\ CI = [-0.11,\ -0.01]$), although the upper bound of the 95% credible interval was close to 0. No gender differences were found in the growth rates of LAS. Finally, older participants tended to have higher levels of LAS at baseline ($\delta_{020} = 0.13,\ \ 95\%\ CI = [0.09,\ 0.16]$), but when it comes to within-individual changes over time, participants' ages were not found to be credibly linked to their levels of LAS, as indicated by the 95% credible interval including 0.

In terms of the random effects, we found between-individual and between-family differences in both initial levels and age-based changes of LAS. These differences were indicated by the relatively high magnitude of random effect standard deviations and the credible intervals whose lower bounds were far from 0 (see, τ_0, τ_1, φ_0, and φ_3; random effect standard deviation parameters whose credible intervals were close to 0 were not shown in Table 2). We also found negative associations between the initial levels and growth rates at the individual level, indicating that individuals who had higher initial levels of activity space tended to experience larger decreases in activity space with age.

Table 3. Parameter estimates of the model for PSS from the CoTwins study, 2016-2018

Parameter	Estimate	SE	95% CI
Fixed effects			
Intercept, δ_{000}	0.74	0.08	[0.58, 0.89]
Gender, δ_{010}	0.03	0.08	[-0.12, 0.19]
Baseline age, δ_{020}	-0.30	0.06	[-0.42, -0.18]
Age, δ_{100}	-0.38	0.03	[-0.44, -0.31]
Age*Gender, δ_{110}	-0.04	0.03	[-0.09, 0.01]
DZSS, δ_{001}	-0.29	0.10	[-0.49, -0.09]
DZOS,δ_{002}	-0.49	0.11	[-0.71, -0.27]
DZSS*Gender, δ_{011}	0.04	0.10	[-0.17, 0.24]
DZOS*Gender, δ_{012}	0.08	0.09	[-0.09, 0.25]
DZSS*Baseline age, δ_{021}	-0.13	0.08	[-0.28, 0.02]
DZOS*Baseline age, δ_{022}	-0.04	0.09	[-0.22, 0.13]
Weekend, β_2	-0.12	0.01	[-0.13, -0.10]
Summer, β_3	-0.31	0.01	[-0.33, -0.29]
Fall, β_4	-0.46	0.01	[-0.48, -0.44]
Winter, β_5	-0.09	0.01	[-0.11, -0.07]
Level-2 random effects			
Intercept standard deviation, τ_0	0.34	0.02	[0.30, 0.38]
Age standard deviation, τ_1	0.20	0.02	[0.17, 0.24]
Intercept-Age correlation, $\tau_{01}/(\tau_0 * \tau_1)$	-0.35	0.09	[-0.52, -0.16]
Level-3 random effects			
Intercept standard deviation, φ_0	0.58	0.11	[1.08, 1.50]
Age standard deviation, φ_3	0.40	0.03	[0.32, 0.42]
Precision parameter, ϕ	1.91	0.01	[1.89, 1.92]

Note. SE = standard errors estimated by standard deviations of the posterior samples; CI = credible interval. N = 484 participants (or 242 pairs of twins). The number of time points for each participant ranged from 3 to 569.

Table 3 shows the parameter estimates for PSS. In terms of the fixed effects, weekend and seasonal effects were found in the trajectory of PSS. Specifically, participants shared more activity space on weekdays than on weekends ($\beta_2 =$

$-0.12, \ \ 95\% \ CI = [-0.13, \ 0.10]$). This pattern might be due to the restricted daily routines on weekdays during which twin siblings in this age range tended to spend most of their time in school and thus, showed greater PSS. Participants tended to have the largest PSS in spring, followed by winter, summer, and fall. In addition, older twins tended to share less activity space at baseline ($\delta_{020} = -0.30, \ \ 95\% \ CI = [-0.42, \ -0.18]$), and when it comes to within-individual changes over time, in contrast to the lack of age-related changes in LAS, PSS was found to decrease as twins grew older ($\delta_{100} = -0.38, \ 95\% \ CI = [-0.44, \ -0.31]$). Note that a small portion of twins were in the transition from high school to college, so the reduction in PSS might also reflect some of the inevitable life transitions that occur with age, such as attending colleges or working at different geographical locations. In terms of zygosity differences, both DZ twins of the same sex and opposite sex were found to share less activity space than MZ twins ($\delta_{001} = -0.29, \ \ 95\% \ CI = [-0.49, \ -0.09]$; $\delta_{002} = -0.49, \ \ 95\% \ CI = [-0.71, \ -0.27]$), indicating that there might be genetically influenced differences in PSS. Finally, no gender differences were found in the initial levels and growth rates of PSS.

Results for random effects were similar to those in the LAS model. We found between-individual and between-family differences in both initial levels and age-based changes of PSS. We also found negative associations between the initial levels and growth rates at the individual level, indicating that twins who had higher initial levels of PSS tended to show more declines in PSS with age. In other words, the participants' GPS data suggested that higher physical closeness at younger ages might not persist as the twins grew older.

Finally, we conducted sensitivity analysis by re-running the analysis with the full data set (i.e., keeping the records with fewer than 20 valid data points within a week in the final data set). Results were detailed in Table S1 and Table S2 in Supplementary Material, which showed only slight differences in the magnitude of point estimates and standard errors. Both data sets yielded consistent conclusions across all parameters in terms of whether they were credibly different from zero based on their 95% credible intervals.

7 Discussion

The proliferation of real-time and longitudinal GPS data provides excellent opportunities to study human behavior (Osorio-Arjona & García-Palomares, 2019). At the same time, the GPS data also pose challenges for consolidating, automating, and analyzing data that are not only massive in their quantities but also contain spatial features that require expertise in GIS. Commercial software packages make these studies easier but may have license and reproducibility issues, and analyses with commercial software cannot be readily deployed to HPC platforms to facilitate research procedures. In this article, we reviewed and compared existing commonly used Python libraries for spatial analysis with GPS2space, our newly developed open-source Python library. GPS2space can build spatial data from GPS data with latitude and longitude coordinate pairs, construct buffer-

and convex hull-based activity space and shared space, and perform the nearest distance query from user-specified locations. We demonstrated how to process spatial data and calculate buffer- and convex hull-based activity space and shared space, as well as the nearest distance, with code examples. We also discussed the pros and cons of buffer- and convex hull-based approaches and illustrated different scenarios when the two approaches could be appropriately applied. Lastly, using data from the CoTwins study, we explored intra-individual changes and between-individual differences in daily activity space and shared space with twin siblings; and gender, zygosity and baseline age-related differences in their initial levels and/or changes, using growth curve modeling techniques. We found different patterns of seasonal effects in the trajectories of LAS and PSS, less activity space shared between DZ twins compared with MZ twins, and a decrease of PSS with increasing age.

There are several limitations to the current data analysis. First, we did not allow for individual differences in the seasonal effects, so our results only provided a general description of seasonal patterns of LAS and PSS. In practice, the seasonal effects might vary across individuals and need to be considered in model specifications. Second, some other factors might affect individuals' activity space, such as time of the year (e.g., school days versus holidays) and weather (e.g., snow). Similarly, the magnitude of shared space between twin siblings depends on whether they live together or not. These factors need to be included in the models to better explain the temporal pattern of LAS and PSS as well as individual differences in these patterns. Finally, in our example, some participants were assessed for fewer than three years, while typically at least three repeated measures per individual are required in the growth curve analysis. Therefore, participants need to be followed for several more years to better investigate age-related changes at the year level. We may also assess changes of finer granularity (e.g., at the month level) based on the current data.

Although we illustrated usage of GPS2space with data from a twin study, the functions available in this package are applicable to a broad range of studies that rely on GPS data or geolocation data with latitude and longitude coordinate pairs. For example, GPS2space can be used to quantify individuals' mobility patterns using data from social media platforms. Health studies investigating the spread of contagious diseases can examine individuals' physical movements and interaction patterns with other individuals using activity space and shared space measures as derived from GPS2space. From demographic and sociological perspectives, activity space and shared space obtained using GPS2space can provide important information regarding people's sense of place, social segregation, and their impacts on a series of socioeconomic outcomes such as educational attainment and occupational status. In addition, the nearest distance measure from GPS2space can also be used to examine the effects of accessibility to food and healthcare providers. Meanwhile, researchers have shown disagreements in mobility or trajectory measures between self-reported data and GPS/Sensor data (Fillekes, Kim, et al., 2019; Fillekes, Röcke, Katana, & Weibel, 2019). GPS2space can provide information for researchers to validate and compare mobility or trajectory measures from different data sources.

Many other extensions are possible within GPS2space to circumvent some of its current limitations. For example, constructing activity space and shared space involves topological structuring, which can take other forms besides convex hull and buffer, the two methods currently available in GPS2space. Some researchers use hexagon methods to measure territorial control based on road data (Tao, Strandow, Findley, Thill, & Walsh, 2016); others also use the concave hull method to estimate crown volumes of trees from remote sensing data (Yan et al., 2019). Those approaches are useful and beneficial for certain research questions but are currently unavailable in GPS2space. To extend the GPS2space, one could include concave hull, hexagon, and network-based methods in constructing activity space and parameterize the column name variables for the spatial measures in GPS2space so that users have control of naming their desired outcomes.

With rapid developments of spatial economics, readily available spatial data sets, and the computational power of personal computer and cloud computing, spatial analyses have gained popularity in areas such as social, behavioral, and environmental studies. We provided a timely open-source solution to work with GPS data and extract spatial measures with code snippets and empirical examples using GPS2space. Overall, we have demonstrated that GPS2space can be a versatile, handy, and extendable tool for researchers to harness the spatialities of GPS data to investigate a wide array of research questions regarding spatial-temporal variations of human behavioral changes and environment-population linkages.

References

Barron, C., Neis, P., & Zipf, A. (2014). A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, *18*(6), 877–895. doi: https://doi.org/10.1111/tgis.12073

Ben-Ari, A., & Lavee, Y. (2007). Dyadic closeness in marriage: From the inside story to a conceptual model. *Journal of Social and Personal Relationships*, *24*(5), 627–644. doi: https://doi.org/10.1177/0265407507081451

Bivand, R. (2006). Implementing spatial data analysis software tools in R. *Geographical Analysis*, *38*(1), 23–40. doi: https://doi.org/10.1111/j.0016-7363.2005.00672.x

Browning, M., & Lee, K. (2017). Within what distance does "greenness" best predict physical health? A systematic review of articles with gis buffer analyses across the lifespan. *International Journal of Environmental Research and Public Health*, *14*(7), 675. doi: https://doi.org/10.3390/ijerph14070675

Buchowski, M. S., Townsend, K. M., Chen, K. Y., Acra, S. A., & Sun, M. (1999). Energy expenditure determined by self-reported physical activity is related to body fatness. *Obesity Research*, *7*(1), 23–33. doi: https://doi.org/10.1002/j.1550-8528.1999.tb00387.x

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: https://doi.org/10.18637/jss.v080.i01

Chi, G. (2010). The impacts of highway expansion on population change: An integrated spatial approach. *Rural Sociology*, *75*(1), 58–89. doi: https://doi.org/10.1111/j.1549-0831.2009.00003.x

Chi, G., & Marcouiller, D. W. (2013). Natural amenities and their effects on migration along the urban-rural continuum. *Annals of Regional Science*, *50*(3), 861–883. doi: https://doi.org/10.1007/s00168-012-0524-2

Chi, G., & Zhu, J. (2019). *Spatial Regression Models for the Social Sciences.* Thousand Oaks: SAGE Publications.

Cleaveland, C., & Kelly, L. (2008). Shared Social Space and Strategies to Find Work: An Exploratory Study of Mexican Day Laborers in Freehold, N.J. *Social Justice*, *35*, 51–65.

Crickard, P., Toms, S., & Rees, E. v. (2018). *Mastering geospatial analysis with Python: explore GIS processing and learn to work with GeoDjango, CARTOframes and MapboxGL-Jupyter.* Birmingham: Packt Publishing Ltd.

Drummond, W. J., & French, S. P. (2008). The future of GIS in planning: Converging technologies and diverging interests. *Journal of the American Planning Association*, *74*(2), 161–174. doi: https://doi.org/10.1080/01944360801982146

Enders, C. K., & Tofighi, D. (2007). Centering Predictor Variables in Cross-Sectional Multilevel Models: A New Look at an Old Issue. *Psychological Methods*, *12*(2), 121–138. doi: https://doi.org/10.1037/1082-989X.12.2.121

Ferrari, S. L., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. doi: https://doi.org/10.1080/0266476042000214501

Fillekes, M. P., Kim, E.-k., Trumpf, R., Zijlstra, W., Giannouli, E., & Weibel, R. (2019). Assessing older adults' daily mobility: a comparison of GPS-derived and self-reported mobility indicators. *Sensors*, *19*(20), 4551.

Fillekes, M. P., Röcke, C., Katana, M., & Weibel, R. (2019). Self-reported versus GPS-derived indicators of daily mobility in a sample of healthy older adults. *Social Science and Medicine*, *220*(October 2018), 193–202. doi: https://doi.org/10.1016/j.socscimed.2018.11.010

Fine, G. A. (2012). Group culture and the interaction order: Local sociology on the meso-level. *Annual Review of Sociology*, *38*, 159–179. doi: https://doi.org/10.1146/annurev-soc-071811-145518

GDAL/OGR contributors. (2020). *GDAL/OGR Geospatial Data Abstraction software Library.*

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian Data Analysis* (Third Edit ed.). New York: Chapman and Hall/CRC.

Gerdenitsch, C., Scheel, T. E., Andorfer, J., & Korunka, C. (2016). Coworking spaces: A source of social support for independent professionals. *Frontiers in Psychology*, *7*, 581. doi: https://doi.org/10.3389/fpsyg.2016.00581

Gillies, S., et al. (2007). *Shapely: manipulation and analysis of geometric objects.* Retrieved from https://github.com/Toblerity/Shapely

Gillies, S., et al. (2011). *Fiona is ogr's neat, nimble, no-nonsense api.* Retrieved from https://github.com/Toblerity/Fiona

Gillies, S., et al. (2013). *Rasterio: geospatial raster i/o for Python programmers.* Retrieved from https://github.com/rasterio/rasterio

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, *91*(1), 1–30. doi: https://doi.org/10.18637/jss.v091.i01

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change.* New York: Routledge.

Huang, Q., & Wong, D. W. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, *30*(9), 1873–1898. doi: https://doi.org/10.1080/13658816.2016.1145225

Jordahl, K. (2014). *GeoPandas: Python tools for geographic data.*

Kearney, A. R. (2006). Residential development patterns and neighborhood satisfaction: Impacts of density and nearby nature. *Environment and Behavior*, *38*(1), 112–139. doi: https://doi.org/10.1177/0013916505277607

Kerr, J., Duncan, S., & Schipperjin, J. (2011). Using global positioning systems in health research: A practical approach to data collection and processing. *American Journal of Preventive Medicine*, *41*(5), 532–540. doi: https://doi.org/10.1016/j.amepre.2011.07.017

Kestens, Y., Lebel, A., Chaix, B., Clary, C., Daniel, M., Pampalon, R., … p Subramanian, S. V. (2012). Association between activity space exposure to food establishments and individual risk of overweight. *PLoS ONE*, *7*(8), e41418. doi: https://doi.org/10.1371/journal.pone.0041418

Kestens, Y., Thierry, B., Shareck, M., Steinmetz-Wood, M., & Chaix, B. (2018). Integrating activity spaces in health research: Comparing the VERITAS activity space questionnaire with 7-day GPS tracking and prompted recall. *Spatial and Spatio-temporal Epidemiology*, *25*, 1–9. doi: https://doi.org/10.1016/j.sste.2017.12.003

Lee, J. H., Davis, A. W., Yoon, S. Y., & Goulias, K. G. (2016). Activity space estimation with longitudinal observations of social media data. *Transportation*, *43*, 955–977. doi: https://doi.org/10.1007/s11116-016-9719-1

Lee, K., & Kwan, M.-P. (2019). The Effects of GPS-Based Buffer Size on the Association between Travel Modes and Environmental Contexts. *ISPRS International Journal of Geo-Information*, *8*(11), 514. doi: https://doi.org/10.3390/ijgi8110514

Lee, N. C., Voss, C., Frazer, A. D., Hirsch, J. A., McKay, H. A., & Winters, M. (2016). Does activity space size influence physical activity levels of adolescents?-A GPS study of an urban environment. *Preventive Medicine Reports*, *3*, 75–78. doi: https://doi.org/10.1016/j.pmedr.2015.12.002

Li, Y., Oravecz, Z., Zhou, S., Bodovski, Y., Barnett, I. J., Chi, G., … Chow, S.-M. (in press). Bayesian forecasting with a regime-switching zero-inflated multilevel poisson regression model: An application to adolescent alcohol use with spatial covariates. *Psychometrika*.

Mason, M. J., Valente, T. W., Coatsworth, J. D., Mennis, J., Lawrence, F., & Zelenak, P. (2010). Place-based social network quality and correlates of substance use among urban adolescents. *Journal of Adolescence*, *33*(3), 419–427. doi: https://doi.org/10.1016/j.adolescence.2009.07.006

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2017). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. *Sociological Methods and Research*, *46*(3), 390–421. doi: https://doi.org/10.1177/0049124115605339

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall.

McGuire, W., O'Brien, B. G., Baird, K., Corbett, B., & Collingwood, L. (2020). Does Distance Matter? Evaluating the Impact of Drop Boxes on Voter Turnout. *Social Science Quarterly*, *101*(5), 1789–1809. doi: https://doi.org/10.1111/ssqu.12853

Murray, A. T., Xu, J., Wang, Z., & Church, R. L. (2019). Commercial GIS location analytics: capabilities and performance. *International Journal of Geographical Information Science*, *33*(5), 1106–1130. doi: https://doi.org/10.1080/13658816.2019.1572898

Newman, H. H., Freeman, F. N., & Holzinger, K. J. (1937). *Twins: a study of heredity and environment.* Chicago: University of Chicago Press.

Osorio-Arjona, J., & García-Palomares, J. C. (2019). Social media and urban mobility: Using twitter to calculate home-work travel matrices. *Cities*, *89*, 268–280. doi: https://doi.org/10.1016/j.cities.2019.03.006

Patil, S. (2016). Big Data Analytics Using R. *International Research Journal of Engineering and Technology*, *3*(7), 78–81.

Perchoux, C., Chaix, B., Brondeel, R., & Kestens, Y. (2016). Residential buffer, perceived neighborhood, and individual activity space: New refinements in the definition of exposure areas - The RECORD Cohort Study. *Health and Place*, *40*, 116–122. doi: https://doi.org/10.1016/j.healthplace.2016.05.004

Prins, R. G., Pierik, F., Etman, A., Sterkenburg, R. P., Kamphuis, C. B., & van Lenthe, F. J. (2014). How many walking and cycling trips made by elderly are beyond commonly used buffer sizes: Results from a GPS study. *Health and Place*, *27*, 127–133. doi: https://doi.org/10.1016/j.healthplace.2014.01.012

Rey, S. J. (2019). PySAL: the first 10 years. *Spatial Economic Analysis*, *14*(3), 273–282. doi: https://doi.org/10.1080/17421772.2019.1593495

Rey, S. J., & Anselin, L. (2007). PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies*, *37*(1), 7–27.

Russell, M. A., Almeida, D. M., & Maggs, J. L. (2017). Stressor-related drinking and future alcohol problems among university students. *Psychology of Addictive Behaviors*, *31*(6), 676–687. doi: https://doi.org/10.1037/adb0000303

Shelton, T. (2017). Spatialities of data: mapping social media 'beyond the geotag'. *GeoJournal*, *82*(4), 721–734. doi: https://doi.org/10.1007/s10708-016-9713-3

Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, *142*, 198–211. doi: https://doi.org/10.1016/j.landurbplan.2015.02.020

Stewart, T., Duncan, S., Chaix, B., Kestens, Y., Schipperijn, J., & Schofield, G. (2015). A Novel Assessment of Adolescent Mobility: A Pilot Study. *International Journal of Behavioral Nutrition and Physical Activity*, *12*, 18. doi: https://doi.org/10.1186/s12966-015-0176-6

Sugiyama, T., Kubota, A., Sugiyama, M., Cole, R., & Owen, N. (2019). Distances walked to and from local destinations: Age-related variations and implications for determining buffer sizes. *Journal of Transport and Health*, *15*, 100621. doi: https://doi.org/10.1016/j.jth.2019.100621

Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: Challenges for GIScience. *International Journal of Geographical Information Science*, *25*(11), 1737–1748. doi: https://doi.org/10.1080/13658816.2011.604636

Tao, R., Strandow, D., Findley, M., Thill, J. C., & Walsh, J. (2016). A hybrid approach to modeling territorial control in violent armed conflicts. *Transactions in GIS*, *20*(3), 413–425. doi: https://doi.org/10.1111/tgis.12228

Vallée, J., Cadot, E., Roustit, C., Parizot, I., & Chauvin, P. (2011). The role of daily mobility in mental health inequalities: The interactive influence of activity space and neighbourhood of residence on depression. *Social Science and Medicine*, *73*(8), 1133–1144. doi: https://doi.org/10.1016/j.socscimed.2011.08.009

Yan, Z., Liu, R., Cheng, L., Zhou, X., Ruan, X., & Xiao, Y. (2019). A Concave Hull Methodology for Calculating the Crown Volume of Individual Trees Based on Vehicle-Borne LiDAR Data. *Remote Sensing*, *11*(6), 623. doi: https://doi.org/10.3390/rs11060623

Yin, Z., Goldberg, D. W., Hammond, T. A., Zhang, C., Ma, A., & Li, X. (2020). A probabilistic framework for improving reverse geocoding output. *Transactions in GIS*, *24*(3), 656–680. doi: https://doi.org/10.1111/tgis.12623

Supplementary Material

Supplementary material including links to the source code and documentation of GPS2space, code for replicating the examples, and results from sensitivity analysis are available at `https://github.com/shuai-zhou/GPS2space_SupMaterial/blob/main/Supplementary_Material_V2.pdf`. To replicate Example I and Example II and explore data structures of the input and output data sets, please follow the Jupyter Notebook at `https://github.com/shuai-zhou/GPS2space_SupMaterial/blob/main/Example%20I%20and%20II.ipynb`.

Journal of Behavioral Data Science, 2021, *1* (2), 119–126.
DOI:https://doi.org/10.35566/jbds/v1n2/p2

A Note on Wishart and Inverse Wishart Priors for Covariance Matrix

Zhiyong Zhang[0000−0003−0590−2196]

University of Notre Dame

Abstract. For inference involving a covariance matrix, inverse Wishart priors are often used in Bayesian analysis. To help researchers better understand the influence of inverse Wishart priors, we provide a concrete example based on the analysis of a two by two covariance matrix. Recommendations are provided on how to specify an inverse Wishart prior.

Keywords: Wishart distribution · inverse Wishart distribution · prior distribution · covariance matrix

In Bayesian analysis, an inverse Wishart (IW) distribution is often used as a prior for the variance-covariance parameter matrix (e.g., Barnard, McCulloch, & Meng, 2000; Gelman et al., 2014; Leonard, Hsu, et al., 1992). The IW prior is very popular because it is conjugate to normal data. For best illustration, consider a multivariate normal (MN) variable. Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ denote a vector of p variables

$$\mathbf{X}|\mathbf{\Sigma} \sim MN(\mathbf{0}, \mathbf{\Sigma})$$

with the mean vector $\boldsymbol{\mu} = \mathbf{0}$ and the variance-covariance matrix $\mathbf{\Sigma}$. The density function is

$$p(\mathbf{x}|\mathbf{\Sigma}) = (2\pi)^{-p/2}|\mathbf{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right).$$

Given a sample $\mathbf{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with n being the sample size, the likelihood function for $\mathbf{\Sigma}$ is

$$\begin{aligned} L(\mathbf{\Sigma}|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{\Sigma}) &\propto |\mathbf{\Sigma}|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) \\ &= |\mathbf{\Sigma}|^{-n/2} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\right)\right] \\ &= |\mathbf{\Sigma}|^{-n/2} \exp\left[-\frac{n}{2}\mathrm{tr}(\mathbf{S}\mathbf{\Sigma}^{-1})\right], \end{aligned}$$

where $\mathbf{S} = \sum_i^n \mathbf{x}_i\mathbf{x}_i^T/n$ is the biased sample covariance matrix (the sample is centered at 0). Note that this is also the maximum likelihood estimate of $\mathbf{\Sigma}$. To get the posterior distribution of $\mathbf{\Sigma}$ for Bayesian inference, one needs to specify a prior distribution $p(\mathbf{\Sigma})$ for it. With the prior, the posterior distribution can be obtained through the Bayes' Theorem:

$$p(\mathbf{\Sigma}|\mathbf{D}) = \frac{p(\mathbf{D}|\mathbf{\Sigma})p(\mathbf{\Sigma})}{p(\mathbf{D})}.$$

1 The Inverse Wishart Prior

The most commonly used prior for $\mathbf{\Sigma}$ is probably the inverse Wishart conjugate prior. The density function of an inverse Wishart distribution $IW(\mathbf{V}, m)$ with the scale matrix $\mathbf{V}$ and the degrees of freedom m for a $p \times p$ variance-covariance matrix $\mathbf{\Sigma}$ is

$$p(\mathbf{\Sigma}) = \frac{|\mathbf{V}|^{m/2}|\mathbf{\Sigma}|^{-(m+p+1)/2}\exp\left[-\mathrm{tr}(\mathbf{V}\mathbf{\Sigma}^{-1})/2\right]}{2^{mp/2}\Gamma(m/2)}.$$

The inverse Wishart distribution is a multivariate generalization of the inverse Gamma distribution. The mean of it is

$$E(\mathbf{\Sigma}) = \frac{\mathbf{V}}{m-p-1} \tag{1}$$

and the variance of each element of $\mathbf{\Sigma} = (\sigma_{ij})$ is

$$Var(\sigma_{ij}) = \frac{(m-p+1)v_{ij}^2 + (m-p-1)v_{ii}v_{jj}}{(m-p)(m-p-1)^2(m-p-3)}.$$

Especially,

$$Var(\sigma_{ii}) = \frac{2v_{ii}^2}{(m-p-1)^2(m-p-3)}. \tag{2}$$

With an inverse Wishart prior $IW(\mathbf{V}_0, m_0)$ based on known $\mathbf{V}_0$ and m_0, the posterior distribution of $\mathbf{\Sigma}$ is

$$\begin{aligned}
p(\mathbf{\Sigma}|\mathbf{D}) &\propto p(\mathbf{D}|\mathbf{\Sigma})p(\mathbf{\Sigma}) \\
&= |\mathbf{\Sigma}|^{-n/2}\exp\left[-\frac{n}{2}\mathrm{tr}(\mathbf{S}\mathbf{\Sigma}^{-1})\right]|\mathbf{\Sigma}|^{-(m_0+p+1)/2}\exp\left[-\mathrm{tr}(\mathbf{V}_0\mathbf{\Sigma}^{-1})/2\right] \\
&= |\mathbf{\Sigma}|^{-(n+m_0+p+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[(n\mathbf{S}+\mathbf{V}_0)\,\mathbf{\Sigma}^{-1}\right]\right\}.
\end{aligned}$$

From it, we can get the posterior distribution for $\mathbf{\Sigma}$, also an inverse Wishart distribution:

$$\mathbf{\Sigma}|\mathbf{D} \sim IW(n\mathbf{S}+\mathbf{V}_0, n+m_0) = IW(\mathbf{V}_1, m_1) \tag{3}$$

with the updated scale matrix and degrees of freedom.

1.1 Information in an inverse Wishart prior

The posterior mean of $\mathbf{\Sigma}$ is

$$\begin{aligned} E(\mathbf{\Sigma}|\mathbf{D}) &= \frac{n\mathbf{S}+\mathbf{V}_0}{n+m_0-p-1} \\ &= \frac{n}{n+m_0-p-1}\mathbf{S} + \left(1-\frac{n}{n+m_0-p-1}\right)\frac{\mathbf{V}_0}{m_0-p-1}. \end{aligned} \tag{4}$$

Therefore, the posterior mean is a weighted average of the sample covariance matrix $\mathbf{S}$ and the prior mean $\mathbf{V}_0/(m_0-p-1)$. When the sample size $n \to \infty$, the posterior mean approaches the sample mean given fixed m_0 and p.

The information in a prior can be connected to data. For example, if we specify the prior $IW(\mathbf{V}_0, m_0)$ as $\mathbf{V}_0 = n_0\mathbf{S}$ and $m_0 = n_0$, then the informative in the prior is equivalent to n_0 participants in the sample. Note that if we set $\mathbf{V}_0 = (m_0-p-1)\mathbf{S}$, then $E(\mathbf{\Sigma}|\mathbf{D}) = \mathbf{S}$, meaning the posterior mean is the same as the sample covariance matrix.

2 Precision Matrix and the Wishart Prior

In practice, the BUGS program is probably the most widely used software for Bayesian analysis (e.g., Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012; Ntzoufras, 2009). BUGS uses the precision matrix, defined as the inverse of the covariance matrix, to specify the multivariate normal distribution. Let $\mathbf{P} = \mathbf{\Sigma}^{-1}$, then the normal density function can be written as

$$p(\mathbf{x}|\mathbf{P}) = (2\pi)^{-p/2}|\mathbf{P}|^{1/2}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{x}\right).$$

The use of the precision matrix has the computational advantage by avoiding the inverse of matrix in the density calculation in certain situations.

For the precision matrix $\mathbf{P}$, a Wishart prior $W(\mathbf{U}_0, w_0)$ with the scale matrix $\mathbf{U}_0$ and degrees of freedom w_0 is used (e.g., Lunn et al., 2012). The density function of the prior is

$$p(\mathbf{P}) = \frac{|\mathbf{P}|^{(w_0-p-1)/2}\exp\left[-\mathrm{tr}(\mathbf{U}_0^{-1}\mathbf{P})/2\right]}{2^{w_0p/2}\mathit{\Gamma}(w_0/2)|\mathbf{U}_0|^{w_0/2}}.$$

Given the sample $\mathbf{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, the posterior distribution of $\mathbf{P}$ is

$$\begin{aligned} p(\mathbf{P}|\mathbf{D}) &\propto \prod_{i=1}^{n}\left[|\mathbf{P}|^{1/2}\exp\left(-\frac{1}{2}\mathbf{x}_i^T\mathbf{P}\mathbf{x}_i\right)\right]|\mathbf{P}|^{(w_0-p-1)/2}\exp\left[-\mathrm{tr}(\mathbf{U}_0^{-1}\mathbf{P})/2\right] \\ &= |\mathbf{P}|^{(n+w_0-p-1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[(n\mathbf{S}+\mathbf{U}_0^{-1})\mathbf{P}\right]\right\}. \end{aligned}$$

Therefore, the posterior is also a Wishart distribution $W(\mathbf{U}_1, w_1)$ with $\mathbf{U}_1 = \left(n\mathbf{S} + \mathbf{U}_0^{-1}\right)^{-1}$ and $w_1 = n + w_0$. The posterior mean of $\mathbf{P}$ is

$$E(\mathbf{P}|\mathbf{D}) = w_1\mathbf{U}_1 = (n + w_0)\left(n\mathbf{S} + \mathbf{U}_0^{-1}\right)^{-1}.$$

Based on the relationship between Wishart and inverse Wishart distributions (Mardia, Bibby, & Kent, 1982),

$$\mathbf{\Sigma}|\mathbf{D} = \mathbf{P}^{-1}|\mathbf{D} \sim IW(\mathbf{U}_1^{-1}, w_1) = IW(n\mathbf{S} + \mathbf{U}_0^{-1}, n + w_0). \tag{5}$$

The posterior mean of $\mathbf{\Sigma}$ is

$$E(\mathbf{\Sigma}|\mathbf{D}) = \frac{\mathbf{U}_1^{-1}}{w_1 - p - 1} = \frac{n\mathbf{S} + \mathbf{U}_0^{-1}}{n + w_0 - p - 1}. \tag{6}$$

Comparing the posterior distributions in Equation (3) and (5), giving an inverse Wishart distribution $IW(\mathbf{V}_0, m_0)$ prior to the covariance matrix $\mathbf{\Sigma}$ is the same as giving a Wishart distribution $W(\mathbf{V}_0^{-1}, m_0)$ prior to the precision matrix $\mathbf{P} = \mathbf{\Sigma}^{-1}$. However, note that

$$[E(\mathbf{P}|\mathbf{D})]^{-1} = \frac{n\mathbf{S} + \mathbf{U}_0^{-1}}{n + w_0} \neq E(\mathbf{\Sigma}|\mathbf{D}) = \frac{n\mathbf{S} + \mathbf{U}_0^{-1}}{n + w_0 - p - 1}.$$

Therefore, one cannot simply invert the posterior mean of the precision matrix to get the posterior mean of the covariance matrix.

3 Numerical Examples

For illustration, we look at a concrete experiment. Suppose we have a sample of size $n = 100$ with the sample covariance matrix ($p = 2$)

$$\mathbf{S} = \begin{pmatrix} 5 & 2 \\ 2 & 10 \end{pmatrix}.$$

The aim is to estimate $\mathbf{\Sigma}$ through Bayesian method. We now consider the use of different priors and evaluate their influence. Given the connection between the Wishart and inverse Wishart distributions, we focus our discussion on the specification of an inverse Wishart prior for the covariance matrix $\mathbf{\Sigma}$.

3.1 Priors based on an identity scale matrix

For an inverse Wishart prior $IW(\mathbf{V}_0, m_0)$, we need to specify its scale matrix and degrees of freedom. In practice, an identity matrix has been frequently used as the scale matrix. Therefore, we first set $\mathbf{V}_0 = \mathbf{I}$ and vary the degrees of freedom by letting $m_0 = 2, 5, 10, 50, 100$. Note that when $m_0 = 2$, the prior is not a proper distribution but the posterior is still a proper distribution. The mean and variance of the posterior distribution are given in Table 1. First, when

$m_0 = 2$ or 5, the posterior means are close to the sample covariance matrix. With the increase of m_0, the posterior means become smaller and the posterior variances also become smaller. This can be easily explained by Equation (4) – the posterior mean is a weighted average between the sample mean and the prior mean. Take the element Σ_{11} as an example. From the data, $S_{11} = 5$. The mean of the inverse Wishart prior is $V_{0,11}/(m_0 - 3) = 1/(m_0 - 3)$. When $m_0 = 5$, the prior mean is 0.5 and when $m_0 = 100$, the prior mean is about 0.01. Furthermore, when $m_0 = 5$, the weight for the prior mean is about 0.05 but when $m_0 = 100$, the weight increases to about 0.5. Therefore, with the increase of m_0, the posterior mean is pulled towards the prior mean since the prior mean has a greater weight.

Table 1. Posterior inference of the covariance matrix parameter based on the inverse Wishart prior with the scale matrix specified based on an identity matrix.

		Mean					Variance				
	S	2	5	10	50	100	2	5	10	50	100
						$IW(\mathbf{I}, m_0)$					
Σ_{11}	5	5.06	4.91	4.68	3.41	2.54	0.528	0.483	0.418	0.160	0.066
Σ_{12}	2	1.96	1.96	1.87	1.36	1.02	0.516	0.516	0.447	0.172	0.071
Σ_{22}	10	10.11	9.81	9.36	6.81	5.08	2.108	1.926	1.667	0.640	0.265
						$IW[(m_0 - p - 1)\mathbf{I}, m_0]$					
Σ_{11}	5	5.04	4.92	4.74	3.72	3.03	0.524	0.484	0.428	0.191	0.094
Σ_{12}	2	1.96	1.96	1.87	1.36	1.02	0.518	0.518	0.454	0.194	0.091
Σ_{22}	10	10.09	9.82	9.41	7.12	5.57	2.100	1.930	1.687	0.700	0.318

In the above specification, since $\mathbf{V}_0 \equiv \mathbf{I}$, the prior mean also changes along the change of m_0. In practice, e.g., in sensitivity analysis, it can be helpful to fix the prior mean. To achieve this, one can set $\mathbf{V}_0 = (m_0 - p - 1)\mathbf{I}$. Therefore, when $m_0 = 5$, the scale matrix will be $2\mathbf{I}$, and when $m_0 = 100$, the scale matrix will be $m_0 = 97\mathbf{I}$. With such specification, the prior mean is always $\mathbf{I}$.

3.2 Priors with the scale matrix formed from data

Another way to specify the prior is to construct the scale matrix for the inverse Wishart distribution based on the sample data. Intuitively, we can set $\mathbf{V}_0 = \mathbf{S}$ and change m_0. From the top of Table 2, with the increase of m_0, the posterior mean deviates from the sample covariance matrix. This is again because that the prior mean becomes smaller with the increase of m_0 since the prior mean is equal to $\mathbf{S}/m_0$. To maintain the same prior mean while changing the information in the prior, we set $\mathbf{V}_0 = (m_0 - p - 1)\mathbf{S}$. With such specification, the prior mean is always $\mathbf{S}$ and the posterior mean is also $\mathbf{S}$ as we can see from the bottom part of Table 2. With the increase of the degrees of freedom, more information is supplied through the prior and we can observe the decrease in the posterior variance.

Table 2. Posterior inference of the covariance matrix parameter based on the priors with the scale matrix constructed from data.

		Mean					Variance				
	S	2	5	10	50	100	2	5	10	50	100
						$IW(\mathbf{S}, m_0)$					
$\varSigma_{11}$	5	5.10	4.95	4.72	3.44	2.56	0.537	0.490	0.424	0.163	0.067
$\varSigma_{12}$	2	1.98	1.98	1.89	1.37	1.03	0.525	0.525	0.455	0.175	0.072
$\varSigma_{22}$	10	10.20	9.90	9.44	6.87	5.13	2.146	1.961	1.697	0.651	0.270
						$IW[(m_0 - p - 1)\mathbf{S}, m_0]$					
$\varSigma_{11}$	5	5.00	5.00	5.00	5.00	5.00	0.515	0.500	0.476	0.345	0.256
$\varSigma_{12}$	2	2.00	2.00	2.00	2.00	2.00	0.536	0.536	0.510	0.370	0.276
$\varSigma_{22}$	10	10.00	10.00	10.00	10.00	10.00	2.062	2.000	1.905	1.379	1.026

3.3 Other types of specifications

We now consider several other types of specifications of the scale matrix to illustrate the influence of the prior. In all the the specifications, we maintain the same prior mean by setting the prior in the form of $IW[(m_0 - p - 1)\mathbf{V}_0, m_0]$. The priors considered and the associated posterior mean and variance are summarized in Table 3.

For prior P1, it assumes that $\varSigma_{11}$ is 10 times of $\varSigma_{22}$, which is not consistent with the sample data. As expected, the posterior mean is pulled towards prior mean with the increase of m_0. Notably, the variance of $\varSigma_{11}$ does not monotonously decrease with the increase of m_0 as one might incorrectly assume that the use of prior information will lead to more precise results. This is because the variance of the inverse Wishart distribution is related to its mean as shown in Equation (2), and the prior is not consistent with data.

For Priors P2, P3, P4, and the one at the bottom of Figure 2, the scale matrices have the same diagonal values and different off-diagonal values. Note that changing the values on the off-diagonals influences neither the posterior means nor variances on the diagonals, which can also be seen in Equations (1) and (2). As expected, changing the off-diagonal values influences both the posterior means and variances. However, the posterior variances are relatively stable.

3.4 Using priors for a precision matrix P

The influence of the priors on the precision matrix is the same as for the covariance matrix because of the connection of Wishart and inverse Wishart distribution – if $\mathbf{\Sigma} \sim IW(\mathbf{V}_0, m_0)$, $\mathbf{P} = \mathbf{\Sigma}^{-1} \sim W(\mathbf{V}_0^{-1}, m_0)$. If the prior $IW(\mathbf{I}, m_0)$ is specified for the covariance matrix, it is equivalent to use $W(\mathbf{I}, m_0)$ for the precision matrix. As discussed earlier, to maintain the same prior mean, we can use $IW[(m_0 - p - 1)\mathbf{I}, m_0]$ for $\mathbf{\Sigma}$. In this case, the prior for the precision matrix should be $W[\mathbf{I}/(m_0 - p - 1), m_0]$. Similarly, if we specify a prior for $\mathbf{\Sigma}$ based on the data using $IW[(m_0 - p - 1)\mathbf{S}, m_0]$, then the prior for the precision matrix would be $W[\mathbf{S}^{-1}/(m_0 - p - 1), m_0]$.

Table 3. Posterior inference of the covariance matrix parameter with additional specifications of inverse Wishart priors $IW[(m_0 - p - 1)\mathbf{V}_0, m_0]$.

		Mean					Variance				
	S	2	5	10	50	100	2	5	10	50	100
P1: $\mathbf{V}_0 = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}$											
Σ_{11}	5	4.95	5.10	5.33	6.60	7.46	0.505	0.520	0.541	0.601	0.571
Σ_{12}	2	1.96	1.96	1.87	1.36	1.02	0.535	0.535	0.507	0.335	0.217
Σ_{22}	10	10.09	9.82	9.41	7.12	5.57	2.100	1.930	1.687	0.700	0.318
P2: $\mathbf{V}_0 = \begin{pmatrix} 5 & -2 \\ -2 & 10 \end{pmatrix}$											
Σ_{11}	5	5.00	5.00	5.00	5.00	5.00	0.515	0.500	0.476	0.345	0.256
Σ_{12}	2	1.92	1.92	1.74	0.72	0.03	0.532	0.532	0.501	0.346	0.255
Σ_{22}	10	10.00	10.00	10.00	10.00	10.00	2.062	2.000	1.905	1.379	1.026
P3: $\mathbf{V}_0 = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix}$											
Σ_{11}	5	5.00	5.00	5.00	5.00	5.00	0.515	0.500	0.476	0.345	0.256
Σ_{12}	2	1.96	1.96	1.87	1.36	1.02	0.534	0.534	0.505	0.355	0.260
Σ_{22}	10	10.00	10.00	10.00	10.00	10.00	2.062	2.000	1.905	1.379	1.026
P4: $\mathbf{V}_0 = \begin{pmatrix} 5 & -5 \\ -5 & 10 \end{pmatrix}$											
Σ_{11}	5	5.00	5.00	5.00	5.00	5.00	0.515	0.500	0.476	0.345	0.256
Σ_{12}	2	1.86	1.86	1.54	-0.24	-1.45	0.530	0.530	0.495	0.343	0.266
Σ_{22}	10	10.00	10.00	10.00	10.00	10.00	2.062	2.000	1.905	1.379	1.026

4 Discussion

Although not without issues, Wishart and inverse Wishart distributions are still commonly used prior distributions for Bayesian analysis involving a covariance matrix (Alvarez, Niemi, & Simpson, 2014; Liu, Zhang, & Grimm, 2016). As we have shown, the use of the inverse Wishart prior has the advantage of conjugate, which simplifies the posterior distribution. By using an inverse Wishart prior, the posterior distribution is also an inverse Wishart distribution given normally distributed data. The posterior mean can be conveniently expressed as a weighted average of the prior mean and the sample covariance matrix. The influence of the prior can also be clearly quantified.

When reliable information is available, an informative inverse Wishart prior can be constructed. For example, previous estimates on the covariance matrix could be available. In this situation, such covariance matrix estimates can be used to construct the scale matrix. If the variance estimates of the covariance matrix is also available, one can determine the degrees of freedom for the inverse Wishart prior based on the variance expression in Equation (2), which can be done using the R package discussed in the Appendix. The degrees of freedom based on each individual element may vary. The overall degrees of freedom for the inverse Wishart distribution can be determined based on the practical research question.

When no reliable information is available, an identity matrix has often been suggested to use as the scale matrix for the inverse Wishart distribution for the covariance matrix and Wishart distribution for the precision matrix (e.g., Congdon, 2014). But as one can see from the numerical example, how much information such a prior has is related to the covariance matrix. We believe a better way to specify an uninformative prior is to determine the scale matrix based on the sample covariance matrix. Therefore, we recommend the prior $IW[(m_0 - p - 1)\mathbf{S}, m_0]$. As for the precision matrix, one can use $W[\mathbf{S}^{-1}/(m_0 - p - 1), m_0]$.

Appendix

The R package wishartprior is developed and made available on GitHub to help understand the Wishart and inverse Wishart priors. The URL to the package is `https://github.com/johnnyzhz/wishartprior`. The package can be used to generate random numbers from an inverse Wishart distribution. It can calculate the mean and variance of Wishart and inverse Wishart distributions. Using the package, one can investigate the influence of priors.

References

Alvarez, I., Niemi, J., & Simpson, M. (2014). Bayesian inference for a covariance matrix. In *Anual conference on applied statistics in agriculture* (pp. 71–82). Retrieved from `arXiv:1408.4050`

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*, 1281–1311.

Congdon, P. (2014). *Applied bayesian modeling* (2nd ed.). John Wiley & Sons.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (2nd ed.). CRC press.

Leonard, T., Hsu, J. S., et al. (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics*, *20*(4), 1669–1696. doi: https://doi.org/10.1214/aos/1176348885

Liu, H., Zhang, Z., & Grimm, K. J. (2016). Comparison of inverse wishart and separation-strategy priors for bayesian estimation of covariance parameter matrix in growth curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 354–367. doi: https://doi.org/10.1080/10705511.2015.1057285

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The bugs book: A practical introduction to bayesian analysis.* CRC Press.

Mardia, K., Bibby, J., & Kent, J. (1982). *Multivariate analysis.* Academic Press.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS.* John Wiley & Sons.

Journal of Behavioral Data Science, 2021, *1* (2), 89–118.
DOI: https://doi.org/10.35566/jbds/v1n2/p6

A Weighted Residual Bootstrap Method for Multilevel Modeling with Sampling Weights

Wen Luo[1] and Hok Chio Lai[2]

[1] Texas A&M University, College Station, TX 77843, USA
wluo@tamu.edu
[2] University of South California, Los Angeles, CA 90089, USA
hokchiol@usc.edu

Abstract. Multilevel modeling is often used to analyze survey data collected with a multi-stage sampling design. When the selection is informative, sampling weights need to be incorporated into the estimation. We propose a weighted residual bootstrap method as an alternative to the multilevel pseudo-maximum likelihood (MPML) estimators. In a Monte Carlo simulation using two-level linear mixed-effects models, the bootstrap method showed advantages over MPML for the estimates and the statistical inferences of the intercept, the slope of the level-2 predictor, and the variance components at level-2. The impact of sample size, selection mechanism, intraclass correlation (ICC), and distributional assumptions on the performance of the methods was examined. The performance of MPML was suboptimal when sample size and ICC were small and when the normality assumption was violated. The bootstrap estimates generally performed well across all the simulation conditions but had notably suboptimal performance in estimating the covariance component in a random slopes model when sample size and ICCs were large. As an illustration, the bootstrap method is applied to the American data of the OECD's Program for International Students Assessment (PISA) survey on math achievement using the R package *bootmlm*.

Keywords: Bootstrap · Informative Selection · Multilevel Modeling · Sampling Weights · Pseudo-maximum Likelihood

1 Introduction

Multi-stage sampling design is often used in survey data collection. For example, in order to obtain a nationally representative sample of kindergartners, a two-stage sample design may be used in which a representative set of schools are sampled in the first stage and students within schools are sampled in the second stage. Besides the advantage of cost-effectiveness and convenience, data obtained by multi-stage sampling allow researchers to answer multilevel research questions. For example, researchers could examine how students' achievement is related to individual student socioeconomic status (SES) on average, how this association varies across schools, how school socioeconomic composition (i.e., school SES) affects student achievement, and how school SES affects the association between student achievement and their SES. One challenge in analyzing complex survey data is the non-independence of observations (or clustering effect) because individuals in the same cluster usually share the same environment and tend to be more alike. Another challenge arises when there are unequal selection probabilities at one or more stages of the sampling process, which is often the case due to the necessity of oversampling certain underrepresented groups or accounting for non-response.

To answer multilevel research questions and to handle the nested data structures, multilevel modeling (MLM) is frequently used. MLM allows researchers to decompose the variance into the between-cluster and within-cluster components and investigate the variability of within-cluster effects across clusters. For example, using MLM researchers could examine not only the average association between individual student achievement and their SES, but also how this association may vary across schools. Established estimation methods for MLM include maximum likelihood (ML) and iterative generalized least squares (IGLS), which are equivalent under normality (Goldstein, 1986). When there are unequal selection probabilities in the stage of selecting schools and/or the stage of selecting students within schools, in order to obtain accurate estimate of the mean outcome and/or the average association between a predictor and the outcome in the population of students, methods were developed to incorporate sampling weights in estimation, such as multilevel pseudo-maximum-likelihood (MPML)(e.g., Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006) and probability-weighted IGLS (PWIGLS; Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998). It has been shown that PWIGLS could result in biased standard error estimates for weighted multilevel data (Asparouhov, 2005). Hence we only considered MPML in our study.

MPML has two crucial underlying assumptions. First, it assumes that the sample size is sufficiently large at both the within-cluster (e.g., number of students per school) and the between-cluster level (e.g., number of schools), especially the latter. In practical research, even if it is possible to obtain a large number of clusters, the sample size within each cluster is often small. To reduce bias in the estimates of the standard errors of fixed effects and the estimates of variance components due to small cluster sizes, scaling of level-1 weights has been used as the major tool. However, the performances of the various scaling methods depend on a host of factors such as cluster size, intraclass correlation (ICC), the degree of informativeness of the selection mechanism, and so forth (Asparouhov, 2006). Applied researchers should select the appropriate scaling method based on the specific sampling design of a study, which could be challenging due to the lack of information. Second, MPML assumes that the error term and random effects follow a distribution of a specified class. In multilevel models, each level has its own error term and random effects; therefore the distributional assumptions should be met at each level. For example, in a two-level linear model, the level-1 errors are assumed to follow a univariate normal distribution, and the level-2 random effects are assumed to follow a multivariate normal distribution. It has been documented that although ML estimators for fixed effects and variance components are consistent even when the random-effects distribution is not normal, the standard error estimated by the inverse Fisher information matrix may be biased, especially for variance components (Verbeke & Lesaffre, 1997). The more sophisticated Huber-White robust standard errors are more accurate for the variance component estimates, but require at least 100 clusters (Maas & Hox, 2004). To our knowledge, the performance of MPML with robust standard errors under distributional misspecification has not been studied yet.

Bootstrap resampling methods for multilevel data have been developed as an alternative to ML estimation in the case where the general assumptions mentioned above are violated. In general, there are three main approaches to bootstrap: (1) the parametric bootstrap, (2) the nonparametric residual bootstrap, and (3) the case bootstrap. The parametric bootstrap has the strongest assumptions, which require that the specifications of the functional form and the distributions of the residuals are both correct. The residual bootstrap only requires the correct specification of the functional form. Finally, the case bootstrap has minimum assumptions and only requires the hierarchical structure to be correctly specified. Van der Leeden, Meijer, and Busing (2008) provided a detailed discussion of the systematic development of bootstrap resampling methods for multilevel models. It has been shown that bootstrap methods could provide accurate confidence intervals for fixed effect estimates when the distribution of the residuals are highly skewed at all levels (Carpenter, Goldstein, & Rasbash, 2003). In addition, applications to small area estimation showed that the bootstrap method could produce sensible estimates for standard errors for shrinkage estimates of small area means based on generalized linear mixed models (e.g., Booth, 1995; Hall & Maiti, 2006; Lahiri, 2003).

Given the advantages of multilevel bootstrap resampling under conditions with distributional assumption violation and small sample sizes, it is useful to extend the method to accommodate multilevel data with sampling weights. Research in this area is limited and existing methods only use the case bootstrap approach (Grilli & Pratesi, 2004; Kovacevic, Huang, & You, 2006; Wang & Thompson, 2012) . Although the case bootstrap is more robust to assumption violations than residual bootstrap, it is typically less efficient. Some studies have shown that case bootstrap performed worse than residual bootstrap even when the assumptions were violated (Efron & Tibshirani, 1993; Van der Leeden et al., 2008). Hence the purpose of this paper is to propose a weighted nonparametric residual bootstrap procedure for multilevel modeling with sampling weights. The proposed procedure is an extension of the nonparametric residual bootstrap procedure developed by Carpenter et al. (2003). With a Monte Carlo simulation, we examined the performance of the proposed bootstrap method in terms of parameter estimates and statistical inferences under a variety of conditions.

The outline of the paper is as follows. First, we briefly discuss sampling weights for multilevel models, followed by a review of existing bootstrap methods for multilevel data. Next, we provide details of the proposed procedure followed by a demonstration of the method using real data. Then we present the simulation study to examine the performance of the proposed bootstrap method. Finally, the findings are summarized and discussed.

2 Sampling Weights and Pseudo-Maximum-Likelihood Estimation for Multilevel Models

Multilevel data are often collected using a multi-stage sampling design which involves sampling clusters in the first stage and then sampling units within selected clusters in the subsequent stages. Due to the clustering, observations in multilevel data often have some degree of dependence among them, which makes the traditional methods based on a simple random sample design inappropriate. Therefore, MLM is often used to account for the dependency among the observations. More importantly, MLM not only allows researchers to examine the average association between a predictor and an outcome, but also to address questions on how the associations among variables within clusters vary across clusters, such as how the association between individual student achievement and their SES varies across schools. In this section, we consider a two-level model with students nested within schools to provide a background for sampling weights in multilevel models.

Let Y_{ij} be the achievement scores, $\mathbf{X}_{ij}$ be the scores on the level-1 predictors (e.g., individual student SES, gender, etc.) associated with student $i(i = 1, \ldots, n_j)$ within school $j(j = 1, \ldots, J)$, and $\mathbf{X}_j$ be the scores on the level-2 predictors (e.g., school SES, school sector, etc.) associated with school j. A two-level model can be specified as

$$Y_{ij} = \boldsymbol{\beta}_1 \mathbf{X}_{ij} + \boldsymbol{\beta}_2 \mathbf{X}_j + \boldsymbol{\mu}_j \mathbf{Z}_{ij} + \varepsilon_{ij} \tag{1}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are row vectors of regression coefficients associated with student-level and school-level predictors respectively, which represent the average effects of the predictors in the population of students. The row vector $\boldsymbol{\mu}_j$ contains random effects associated with school j, which could be a random intercept, or a random slope of a student-level predictor, or both. The design vector $\mathbf{Z}_{ij}$ usually includes the constant 1 (for the random intercept) and the student-level predictors that have random slopes across schools. Finally, ε_{ij} is the level-1 error. The main parameters of interest in MLM are usually the fixed effects (i.e., $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$) and the variance and covariance components (i.e., the variances and covariances of the random effects $\boldsymbol{\mu}_j$). The conventional maximum likelihood estimates of the parameters are obtained by maximizing the likelihood function $L(\theta) = \prod_{j=1}^{J} [\int \prod_{i=1}^{n_j} f(Y_{ij} | \mathbf{X}_{ij}, \boldsymbol{\mu}_j, \boldsymbol{\beta}_1) q(\boldsymbol{\mu}_j | \mathbf{X}_j, \boldsymbol{\beta}_2) d\boldsymbol{\mu}_j]$ where $f\left(Y_{ij} | \mathbf{X}_{ij}, \boldsymbol{\mu}_j, \boldsymbol{\beta}_1\right)$ is the density function of Y_{ij} and $q(\boldsymbol{\mu}_j | \mathbf{X}_j, \boldsymbol{\beta}_2)$ is the density function of $\boldsymbol{\mu}_j$.

Suppose that schools and students within schools are selected with unequal probabilities. Let the probability of selecting school j be p_j and the probability of selecting student i given that school j is sampled be $p_{i|j}$. The sampling weight for school j is $w_j = 1/p_j$. The conditional sampling weight for student i within school j is $w_{i|j} = 1/p_{i|j}$. The unconditional sampling weight for an individual student is $w_{ij} = w_j \times w_{i|j}$. If the sampling weights are related to the dependent variable after conditioning on the covariates in the model, they are called informative weights (Pfeffermann, 1993). For example, if students with lower achievement have a higher probability of being sampled controlling for the predictors $\mathbf{X}_{ij}$ and $\mathbf{X}_j$, then the sampling weights are informative. Informative sampling weights should be incorporated in statistical inferences to avoid bias in estimates or poor performance of test statistics and confidence intervals. For multilevel models, the sampling weights at each level need to be taken into account when they are informative, to ensure that the average association between the predictors and the outcome in the population of students as well as the variance and covariance components of school random effects can be accurately estimated. One approach to incorporate the sampling weights is to use multilevel pseudo maximum likelihood estimation (MPML), which defines the likelihood function as $l\left(\theta\right) = \prod_{j=1}^{J}(\int \prod_{i=1}^{n_j} f\left(Y_{ij}|\mathbf{X}_{ij}, \boldsymbol{\mu}_j, \boldsymbol{\beta}_1\right)^{w_{i|j}} q(\boldsymbol{\mu}_j|\mathbf{X}_j, \boldsymbol{\beta}_2)d\boldsymbol{\mu}_j)^{w_j}$.

Extant literature has shown that the level-1 weights should be scaled in order to reduce the bias of variance component estimates and standard error estimates of fixed effects when cluster sizes are not large (e.g., Pfeffermann et al., 1998; Potthoff, Woodbury, & Manton, 1992; Stapleton, 2002). There are two commonly used scaling methods: relative vs. effective sample size scaling. In relative sample size rescaling, the level-1 weights $w_{i|j}$ are multiplied by a scaling factor $s_{1j} = \frac{n_j}{\sum_{i=1}^{n_j} w_{i|j}}$ so that the sum of the rescaled level-1 weights within a cluster equals the actual cluster size. In effective sample size rescaling, the scaling factor $s_{1j} = \frac{\sum_{i=1}^{n_j} w_{i|j}}{\sum_{i=1}^{n_j} w_{i|j}^2}$ is used such that the sum of the rescaled level-1 weights within a cluster equals the effective cluster size which is defined as $\frac{\left(\sum_{i=1}^{n_j} w_{i|j}\right)^2}{\sum_{i=1}^{n_j} w_{i|j}^2}$. Some simulation studies showed that relative sample size rescaling works better for informative weights, whereas effective sample size rescaling is more appropriate for non-informative weights (Pfeffermann et al., 1998). Some researchers argue that non-informative weights should not be used in multilevel analyses because they tend to result in a loss of efficiency and even bias in parameter estimates under some conditions. For example, Asparouhov (2006) found bias in the estimation of multilevel models when cluster sample size is small and non-informative within-cluster weights are used.

However, in practical applications, choosing the right scaling method may be challenging. Pfeffermann (1993) described a general method for testing the informativeness of the weights. Asparouhov (2006) proposed a simpler method based on the informative index, and recommended to consider both the value of the informative index and Pfeffermann's test, the invariance of selection mechanism across clusters, and the average cluster size when determining weighting in multilevel modeling.

3 Bootstrap for Multilevel Data

Depending on whether and what parametric assumptions are involved, there are multiple approaches to do bootstrapping (Davison & Hinkley, 1997), and additional care is needed to address the dependencies in the data when resampling with multilevel data (Van der Leeden et al., 2008). Below we first provide a brief summary of the common bootstrap procedures for multilevel data in general (i.e., the parametric bootstrap, the residual bootstrap, and the case bootstrap) and then focus on the bootstrap method for multilevel data with sampling weights. Readers should consult Davison and Hinkley (1997), Goldstein (2011), and Van der Leeden et al. (2008) for more detailed reviews of the statistical theory of multilevel bootstrapping methods.

3.1 Parametric Bootstrap

As described in Goldstein (2011), with parametric bootstrap, researchers first fit a multilevel model to obtain fixed effect estimates, and the random effect variance estimates, $\hat{\boldsymbol{\tau}}$ and $\hat{\sigma}$. Then, for each bootstrap sample, a new set of N level-1 errors, ε^*_{ij}, and a new set of J level-2 random effects, $\boldsymbol{\mu}^*_j$, are drawn from independent $N(0, \hat{\boldsymbol{\tau}})$ and $N(0, \hat{\sigma})$ distributions to form a new set of responses, y^*_{ij}. The multilevel model is then refitted to the new bootstrap data, and the target statistics (e.g., fixed effects) are computed. The resampling process is repeated for a large number of B bootstrap samples (e.g., $B = 1,999$) to obtain bootstrap sampling distributions of the target statistics.

3.2 Non-parametric Residual Bootstrap

The (nonparametric) residual bootstrap is similar to the parametric bootstrap except that, when forming new responses, the new errors and random effects were obtained by sampling with replacement the residuals of the multilevel fitted model. In this paper, the resampled residuals were denoted as $\tilde{\boldsymbol{\mu}}_j$ and $\tilde{\varepsilon}_{ij}$ to distinguish them from the counterparts in the parametric bootstrap. In addition, because the sampling variance of $\tilde{\boldsymbol{\mu}}_j$ is generally smaller than $\hat{\boldsymbol{\tau}}$, and so is the sampling variance of $\tilde{\varepsilon}_{ij}$ smaller than $\hat{\sigma}$ (albeit to a lesser extent). Carpenter et al. (2003) and Goldstein (2011) recommended to first "reflate" the residuals so that the sample variances of the reflated residuals were exactly $\hat{\boldsymbol{\tau}}$ and $\hat{\sigma}$, respectively. Finally, as in parametric bootstrap, a new set of response $\tilde{y}_{ij}$ is formed, and the target statistics are computed, and then the process is repeated B times to obtain a bootstrap sampling distribution of the target statistics.

3.3 Case Bootstrap

With the case bootstrap, each bootstrap sample consisted of observations (i.e., "cases") sampled with replacement from the original data. When there are two levels in the data so that a case can mean a cluster or a unit within a cluster, there are two variants of the case bootstrap (Davison & Hinkley, 1997): (a) to resample with replacement intact clusters but no resampling within a cluster, and (b) to first resample the clusters, and within each cluster resample with replacement the units. Both Davison and Hinkley (1997) and Goldstein (2011) recommended (a) over (b).

A few previous studies have examined these three bootstrap methods for multilevel analyses. Seco, García, García, and Rojas (2013) showed that the residual bootstrap produced more precise estimates, in terms of smaller root mean squared errors, for fixed effects than restricted maximum likelihood. On the other hand, because the case bootstrap makes fewer assumptions than the parametric and the residual bootstraps, it requires more information from the data. As such, previous literature found that its performance was poor compared to the other two methods, even when the assumptions for the latter two methods were violated (Efron & Tibshirani, 1993; Van der Leeden et al., 2008). On the other hand, Thai, Mentré, Holford, Veyrat-Follet, and Comets (2014) found that in longitudinal linear-mixed models where cluster size is constant, residual bootstrap and case bootstrap performed similarly when there were at least 100 individuals (i.e., $J = 100$).

3.4 Bootstrap for Multilevel Data with Sampling Weights

For multilevel data with sampling weights, the extant literature documents two types of bootstrap methods, both of which can be viewed as modifications to case bootstrap. One type involves generating a pseudo (or artificial) population that mimics the population from which the original sample is selected, and then selecting bootstrap samples from the pseudo population based on the sampling weights in the original sample (Grilli & Pratesi, 2004; Wang & Thompson, 2012). As described in Grilli and Pratesi (2004), when generating the pseudo population, the ith unit ($i = 1, \ldots, n_j$) in the jth cluster ($j = 1, \ldots, J$) is duplicated $w_{i|j}$ times, rounding the weight to the nearest integer to form J artificial clusters. Then each of the J artificial clusters is replicated w_j times, rounding the weight to the nearest integer, to obtain the artificial population. From the artificial population, bootstrap samples are obtained by first selecting J clusters with probability proportional to $1/w_j$ and then selecting n_j units with probability proportional to $1/w_{i|j}$ from the jth resampled cluster. Wang and Thompson (2012)'s procedure is similar except that they added an additional step to account for the potential biases caused by rounding the weights when generating the pseudo population.

The other type of bootstrap for multilevel data with sampling weights involves a two-stage resampling and rescaling of weights at each level. As described in Kovacevic et al. (2006), $J - 1$ clusters are first drawn from the original sample using simple random sampling with replacement (SRSWR). Then w_j is rescaled to obtain the cluster bootstrap weights $w_j^* = w_j \frac{J}{J-1} t_j$ where t_j is the number of times that cluster j is included in the bootstrap sample. From each resampled cluster, $n_j - 1$ units are drawn using SRSWR and the unadjusted conditional bootstrap weights are calculated for level-1 units as $b^*_{i|j} = w_{i|j}\left(\frac{n_j}{n_j-1}\right)\left(\frac{t_{i|j}}{t_j}\right)$ where $t_{i|j}$ is the total number of times that the ith unit is resampled. Based on the rescaled cluster bootstrap weights and the unadjusted conditional bootstrap weights, the unadjusted unconditional bootstrap weights are computed as $b^*_{ij} = b^*_{i|j}w^*_j$. The adjusted unconditional bootstrap weights (w^*_{ij}) are obtained after applying all the same adjustments done in the process of calculating the original full sample unconditional weights. If no adjustment is made, then $w^*_{ij} = b^*_{ij}$. Finally, the within-cluster conditional weights are calculated as $w^*_{i|j} = w^*_{ij}/w^*_j$.

Both Grilli and Pratesi (2004) and Kovacevic et al. (2006) noted that the steps concerning the level-1 units in their procedures can be omitted when the sampling fraction is low at the cluster level. Kovacevic et al. (2006) also showed that the accuracy and stability of variance estimation improved when using the relative within-cluster weights (i.e., the sum of the rescaled level-1 weights within a cluster equals the actual cluster size) as compared to the original unscaled within-cluster weights. However, to the best of our knowledge, these methods have not been developed into statistical packages that can be easily accessed by applied researchers.

4 The Proposed Weighted Residual Bootstrap

4.1 Algorithm

The weighted residual bootstrap method was developed based on an idea similar to the one outlined in Goldstein, Carpenter, and Kenward (2018). Without loss of generality, we present the weighted nonparametric residual bootstrap algorithm for a two-level model. An extension to a model with more levels is straightforward.

Step 1: Obtain parameter estimates for model 1 (i.e., $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$) based on sample data using unweighted maximum likelihood and restricted maximum likelihood, and compute level-1 residuals ε_{ij} and level-2 residuals $\boldsymbol{\mu}_j$.

Step 2: Obtain reflated level-1 and level-2 residuals ($\varepsilon^{'}_{ij}$and $\boldsymbol{\mu}^{'}_j$) using Carpenter et al. (2003)'s procedure.

Step 3: Sample independently with replacement from the set of reflated level-1 residuals using level-1 unconditional weights and from the set of reflated level-2 residuals using level-2 weights, obtaining two new sets of residuals $\varepsilon^{'b}_{ij}$ and $\boldsymbol{\mu}^{'b}_j$, where b is the index of bootstrap samples. It is noted that the level-1 unconditional weights are used instead of the conditional weights to resample level-1 residuals, because the new set of level-1 residuals are selected from the entire sample across clusters rather than within clusters. This approach makes it unnecessary to scale the within-cluster weights.

Step 4: The new response of the bth bootstrap sample is then obtained by $Y^{'b}_{ij} = \hat{\boldsymbol{\beta}}_1\boldsymbol{X}_{ij} + \hat{\boldsymbol{\beta}}_2\boldsymbol{X}_j + \boldsymbol{\mu}^{'b}_j\boldsymbol{Z}_{ij} + \varepsilon^{'b}_{ij}$.

Step 5: Refit the model to the bootstrap sample to obtain one set of bootstrap parameter estimates using either unweighted maximum likelihood or restricted maximum likelihood.

Step 6: Repeat steps 2-5 to obtain B set sets of bootstrap parameter estimates.

4.2 Illustration

As a demonstration, we applied the proposed procedure to examine the associations between student math achievement and student gender and school SES among 15-year-old students in the United States using the 2000 PISA data Organization for Economic Co-operation and Development (2000) . PISA used a cluster sampling design with unequal selection probabilities. Specifically, schools with more than 15% of minority students were oversampled, and minority students were oversampled within those schools. The data include weights at the school level (named WNRSCHBW) and unconditional weights at the student level (named W_FSTUWT). We used a two-level random intercept model with students' math test scores (Y_{ij}) as the dependent variable, student gender ($Gender_{ij} = 0$ for females and 1 for males) and school mean ISEI (*ISEI_m*) as the school-level predictor (Equation 2),

$$Y_{ij} = \beta_0 + \beta_1 Gender_{ij} + \beta_2 ISEI\_m_j + u_{0j} + e_{ij} \tag{2}$$

where i indexes students and j indexes schools, u_{0j} represents random effects associated with the intercept. The main parameters of interest are the average effects of gender (β_1) and school SES (β_2) on students' math achievement in the population of 15-year-old students in the United States. Although we used a random intercept model in this demonstration, researchers could further examine whether the association between student gender and achievement varies across schools by adding a random effect associated with the slope of gender that varies across schools (i.e., a random slope model).

The US sample consists of 2135 students from 145 schools. 74% students had complete data on both *ISEI* and *Math* while 26% had at least one missing value on the two variables. After removing cases with missing data, the final sample of analysis consists of 1578 students from 145 schools. The cluster size ranged from 1 to 20, with the first quartile of 8, median of 12, and the third quartile of 14. To determine the degree to which the weights were informative, we followed the recommendation by Asparouhov (2006) and computed the informative index by $|\widehat{\mu_w} - \widehat{\mu_0}| / \sqrt{\upsilon_0}$ where $\widehat{\mu_w}$ is the weighted mean of the dependent variable, $\widehat{\mu_0}$ is the unweighted mean, and υ_0 is the unweighted variance. The informative index for math was 0.03, indicating that the sampling weights were very slightly informative.

The bootstrap estimates were obtained using researcher developed R package *bootmlm* (see Appendix for the R code). As a comparison, the model was also estimated using unweighted ML, and MPML with relative and effective weights respectively. The MPML estimates were obtained using M*plus* 8.2 Muthén and Muthén (1998, see Appendix B for the Mplus code). The ML estimates were obtained using the *lme4* package in R (Bates, Maechler, Bolker, & Walker, 2015). Percentile confidence intervals were computed in the bootstrap method (i.e., $\alpha/2$

and $1-\alpha/2$ quantiles of the bootstrap distribution), profile likelihood confidence intervals were computed in *lme4* for the ML estimates, and the delta method[3] was used to construct approximate confidence intervals for the MPML variance component estimates. The MPML results based on relative weights were almost identical to those based on effective weights, thus we only reported the latter.

Table 1. ML, MPML, and Bootstrap Results Based on the PISA Data

		Estimate	SE	95% CI
Unweighted ML	Intercept	74.33	2.49	[69.45, 79.20]
	Gender	-1.6	0.66	[-2.88, -0.31]
	ISEI_m	0.16	0.05	[0.06, 0.26]
	Variance			
	School	9.43	3.02	[4.35, 16.48]
	Residual	162.4	6.06	[151.07, 174.87]
	Conditional ICC	0.06		
MPML Effective Weights	Intercept	80.42	5.52	[69.59, 91.24]
	Gender	-2.43	1.16	[-4.70, -0.16]
	ISEI_m	0.06	0.12	[-0.17, 0.28]
	Variance			
	School	10.86	9.03	[2.12, 55.41]
	Residual	152.47	24.3	[111.56, 208.38]
	Conditional ICC	0.07		
Bootstrap	Intercept	74.94	2.51	[70.17, 80.18]
	Gender	-1.56	0.67	[-2.85, -0.17]
	ISEI_m	0.16	0.05	[0.05, 0.26]
	Variance			
	School	7.42	2.68	[2.23, 13.02]
	Residual	162.51	9.95	[144.5, 184.0]
	Conditional ICC	0.04		

Before looking at the parameter estimates, we examined the distribution of the residuals. The level-1 residuals based on the ML estimates were slightly non-normal with skewness of -1.45 and kurtosis of 6.77. The distribution of the level-2 residuals was close to normal with skewness of -0.46 and kurtosis of 3.39. Table 1 shows the parameter estimates, standard error estimates, 95% confidence intervals, and conditional ICCs. There was little difference between the ML estimates and the bootstrap estimates. However, the MPML results showed different point estimates and standard error estimates, especially for the slope of school mean ISEI (i.e., ISEI_m). As a result, the statistical inference also reached different conclusions regarding the slope of school mean ISEI, which

[3] The 1- α confidence interval of a variance component θ is given by $\exp\left[\ln\left(\hat{\theta}\right) \pm z_{1-\frac{\alpha}{2}} \frac{\sqrt{Var\left(\hat{\theta}\right)}}{\hat{\theta}}\right]$ where $\hat{\theta}$ is the MPML estimate of θ, $Var\left(\hat{\theta}\right)$ is the asymptotic variance of $\hat{\theta}$.

was statistically significant based on the ML and the bootstrap results, but non-significant based on MPML.

From this particular sample and model, we obtained inconsistent results from the bootstrap and the MPML methods. We suspected that the MPML results might not be trustworthy because the specific condition of this sample (i.e., small cluster size, low ICC, and very slight informativeness) has been shown to be unfavorable to MPML (e.g., Asparouhov, 2006). However, it is unknown whether the performance of the bootstrap method is acceptable, thus a Monte Carlo simulation is needed to assess the performance of these methods under various conditions.

5 Simulation

5.1 Data Generation

To evaluate the performance of the weighted bootstrap procedure in accounting for nonrandom sampling, we used R 3.5.0 (R Core Team, 2018) to simulate two-level data mimicking the data structure of students nested in schools. The population models were either (a) a random intercept model or (b) a random slopes model. The models include one level-1 predictor such as student SES (denoted as $X1_{ij}$) and one level-2 predictor such as school SES (denoted as $X2_j$). Because multilevel modeling is a model-based technique usually justified by a superpopulation model (Cochran, 1977; Lohr, 2010), the data generating model is treated as the superpopulation, and in each replication, we first generated a finite population with $J_{pop} = 500$ clusters and $n_{pop} = 100$ observations for each cluster.

When generating a finite population based on the random intercept model (see Equation 2), we simulated $X2_j$ from $N(0,\, 1)$ distributions and the cluster-level random intercept effect u_{0j} from either normal distributions or scaled $\chi^2(df = 2)$ distributions with mean 0 and variance τ, depending on the simulation condition described in the next section. We then simulated $n_{pop} \times J_{pop}$ values of $X1_{ij}$ from $N(0,\, 1)$ distributions and e_{ij} from either normal distributions or scaled $\chi^2(df = 2)$ distributions with mean 0 and variance σ, depending on the simulation condition. For all simulation conditions, we set β_0= 0.5, $\beta_1 = \beta_2 = 1$, and the total variance $\tau{+}\sigma = 2.5$. The outcome was computed based on Equation (2).

When generating a finite population based on the random slopes model, the following equation was used

$$Y_{ij} = \beta_0 + \beta_1 X1_{ij} + \beta_2 X2_j + u_{0j} + u_{1j} X1_{ij} + e_{ij} \tag{3}$$

where u_{0j} and u_{1j} represent the random effects associated with the intercept and the slope of $X1_{ij}$ respectively. We simulated u_{0j} and u_{1j} from a bivariate normal distribution with mean of 0 and variance-covariance of $\begin{bmatrix} \tau_{00} & \\ \tau_{01} & \tau_{11} \end{bmatrix}$ in which τ_{00} represents the variance of the random intercept, τ_{11} the variance of the random slope of $X1_{ij}$, and τ_{01} the covariance between the random intercept and the random slope. The magnitude of τ_{00} depends on the simulation condition, and the magnitude of τ_{11} is half of τ_{00} because the variance of random slopes is typically smaller than the variance of random intercepts. The covariance τ_{01} is computed as $\rho\sqrt{\tau_{00}\tau_{11}}$ where ρ denotes the correlation between the random intercepts and the random slopes and was set at 0.5 to represent a moderate correlation.

After simulating the finite populations, we first sampled J clusters with a sampling fraction f according to a certain selection mechanism depending on the simulation condition. Then in each cluster we randomly sampled n observations with the same sampling fraction f according to a certain selection mechanism depending on the simulation condition.

5.2 Design Factors

We considered 5 design factors to generate a variety of experimental conditions. First, the variance of the random intercepts: 0.125, 0.5, and 1.25. They correspond to small, medium, and large conditional ICCs (i.e., ICC = 0.05, 0.2, and 0.5) commonly seen in multilevel data. Second, sampling fraction (f): 0.1 and 0.5. Similar levels were used in previous simulations such as 0.12 in Grilli and Pratesi (2004) and 0.6 in Rabe-Hesketh and Skrondal (2006). Under the 0.1 sampling fraction condition, the cluster size was 10 and the number of clusters was 50. This was considered a small sample size condition. Under the 0.5 sampling fraction condition, the cluster size was 50 and the number of clusters was 250, which was considered a large sample size. Third, normality of random effects. For the random intercept model, we considered the normal distribution vs. the scaled $\chi^2(df = 2)$ distribution for the random effects and the level-1 errors. The $\chi^2(df = 2)$ distribution has skewness $= \sqrt{8/2} = 2$ and kurtosis $= 12/2 = 6$. For the random slopes model, we only considered normal distribution.

Fourth, between-cluster selection mechanism: non-informative vs. informative. For non-informative selection, simple random sampling (SRS) was used. For the random intercept model with informative sampling, we first divided the clusters into two strata: $\mu_{0j} > 0$ (stratum 1) and $\mu_{0j} < 0$ (stratum 2), and then sampled without replacement in each stratum such that the sampling probability of each cluster is $1.4f$ for stratum 1 and $0.6f$ for stratum 2. In other words, it was expected that for each replication, 70% of the sampled units came from stratum 1, and 30% of the sampled units came from stratum 2. For the random slopes model with informative sampling, we divided the clusters into four strata: $\mu_{0j} > 0$ and $\mu_{1j} > 0$ (stratum 1), $\mu_{0j} > 0$ and $\mu_{1j} < 0$ (stratum 2), $\mu_{0j} < 0$ and $\mu_{1j} > 0$ (stratum 3), and $\mu_{0j} < 0$ and $\mu_{1j} < 0$ (stratum 4), with sampling probabilities of $1.96f$, $0.84f$, $0.84f$, and $0.36f$, respectively. It was expected that for each replication, 49% of the sampled units came from stratum 1, 21% from stratum 2, 21% from stratum 3, and 9% from stratum 4.

Finally, within-cluster selection mechanism: non-informative vs. informative. For non-informative selection, within-cluster units were sampled using SRS. For informative selection, units in each cluster were first divided into two strata: $e_{ij} >$

0 (stratum 1) and $e_{ij} < 0$ (stratum 2), and then sampled without replacement according to the 7:3 ratio of sampling probability. The informative index was about 0.17 when informative selection occurred at level-1 only, 0.09 when at level-2 only, and 0.27 when at both levels based on the random intercept models. These values represent slight to moderate informativeness according to Asparouhov (2006).

Combining the five design factors, there are a total of 48 data conditions (3 ICCs × 2 sampling fractions × 2 distributions × 2 between-cluster selection mechanisms × 2 within-cluster selection mechanisms) for the random intercept models and 24 conditions (3 ICCs × 2 sampling fractions × 2 between-cluster selection mechanisms × 2 within-cluster selection mechanisms) for the random slopes models. We conducted 500 replications for each simulation condition. For each generated data set, three estimators were applied: the proposed bootstrap method (using the R package *bootmlm*), MPML with effective weights (using M*plus* 8.2 for the random intercept models and Stata 16 for the random slopes models), and unweighted maximum likelihood (using the R package *lme4*).

5.3 Analysis

For each parameter in the models (including both fixed effects and variance components), we examined the relative bias of the point estimate and the coverage rate of the 95% confidence intervals. For the bootstrap method, we used the 2.5 and 97.5 percentile of the empirical sampling distribution as the lower and upper boundaries of the 95% confidence interval. Following Hoogland and Boomsma (1998), relative biases of point estimates are considered acceptable if their magnitudes are less than 0.05. The coverage rate of a 95% confidence interval should be approximately equal to 95%, with a margin of error of 1.9% based on 500 replications. Hence coverage rates between 93% and 97% are acceptable.

5.4 Results

5.4.1 Random intercept models Tables 2 to 5 show the relative bias and coverage rate for parameter estimates under all conditions based on the random intercept models. The relative biases for the slope of the level-1 predictor *X1* and the slope of the level-2 predictor *X2* are not shown in the tables because they were close to zero for all conditions. In addition, the coverage rate for the slope of *X1* was close to 95% under all conditions, therefore it was not included in the tables.

Intercept. As shown by the relative biases of the ML estimates, ignoring sampling weights when the selection mechanism was informative caused moderate to large relative biases, ranging from 0.14 to 1.38 (see Table 2 and 3). As a result of biased point estimate, the coverage rates of the confidence intervals for the ML estimates were also poor under those conditions ranging from 0.00 to 0.85 (see Table 4 and 5).

MPML successfully reduced the relative biases to an acceptable level under the majority of conditions, however, there were still small to moderate relative biases under 11 conditions where the sample size was small and the selection mechanism was informative at level 1 or both levels (relative bias ranging from 0.07 to 0.13). As a result, there was slight under-coverage (ranging from 0.88 to 0.92) in about half of those conditions (6 out of 11), mainly when there was informative selection at both levels.

The bootstrap method performed the best in terms of relative biases because they were below 0.05 under all conditions. However, the advantage of the bootstrap method over MPML was less obvious in terms of the coverage rate because the bootstrap method also had slightly low coverage rate (ranging from 0.88 to 0.92) under similar conditions.

Slope of *X2*. The relative bias of the estimated slope of *X2* was acceptable for all methods under all conditions. However, the MPML confidence intervals suffered from slight under-coverage (89%-92%) in 18 conditions, mainly when sample size was small and selection was informative at level 2 or both levels.

Variance component of the random intercepts (τ). ML estimates had small relative biases under 18 conditions when there was informative sampling at level-2 or at both levels. The biases were negative ranging from -0.07 to -0.11 when the distribution was normal, and were positive ranging from 0.10 to 0.12 when the distribution was skewed. MPML suffered from small to moderate biases (-0.10 to 0.27) under 10 conditions when small sample size was combined with small to moderate ICCs. It was noted that the two moderately large relative biases (i.e., 0.25 and 0.27) both occurred when there was informative selection at level-1 or at both levels. The bootstrap method performed better with only small positive biases (0.08 to 0.11) under 5 conditions where both ICC and sample size were small. It was noted that out of the 5 conditions where relative biases were obvious, one was under the normal distribution and four under the skewed distribution, indicating that the performance of the bootstrap method might be sensitive to skewed distributions.

In general, all three methods tended to have under-coverage, with ML being the worst and bootstrap being the best. Where the distribution was normal, 15 conditions had under-coverage ranging from 0.87 to 0.92 for ML, 14 conditions ranging from 0.86 to 0.92 for MPML, and 11 conditions ranging from 0.89 to 0.92 for bootstrap. When data were skewed, 23 conditions had under-coverage ranging from 0.67 to 0.92 for ML, 22 conditions ranging from 0.76 to 0.92 for MPML, and 15 conditions ranging from 0.81 to 0.92 for bootstrap. For both MPML and bootstrap, the coverage rate tended to worsen as the sample size decreased. In addition, when data were skewed, larger ICCs led to lower coverage rate for MPML.

Level-1 residual variance (σ). Only ML estimates had small negative relative biases when there was informative selection at level-1 or at both levels. As a result, ML estimates had severe under-coverage under those conditions, especially when sample size was large. The performance of ML deteriorated when the distribution was skewed as there were severe under-coverage across all conditions.

Although MPML and bootstrap estimates had minimum relative biases, both had slight under-coverage under certain conditions. Specifically, when the distribution was normal, under-coverage mainly occurred when sample size was small combined with informative selection at both levels. When the distribution was skewed, under-coverage mainly occurred when sample size was small and when the selection was non-informative or only informative at level-2.

Table 2. Relative Bias for the Random Intercept Model Under Normal Distribution

ICC	Selection Mechanism	Sampling Fraction	Intercept			TAU			SIGMA		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-informative	0.1	-0.01	-0.01	-0.01	0.03	**0.07**	**-0.07**	0.00	0.00	0.00
		0.5	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00
	Informative at level-1	0.1	**0.93**	0.01	**0.10**	-0.04	0.00	**0.25**	**-0.08**	0.00	-0.02
		0.5	**0.70**	0.00	0.01	0.01	0.01	0.02	-0.05	0.00	0.00
	Informative at level-2	0.1	**0.22**	-0.04	0.01	**-0.07**	0.00	**-0.10**	0.01	0.01	0.00
		0.5	**0.16**	-0.02	0.00	-0.05	-0.01	-0.01	0.00	0.00	0.00
	Informative at both levels	0.1	**1.15**	-0.02	**0.12**	-0.10	-0.03	**0.27**	**-0.09**	0.00	-0.02
		0.5	**0.86**	-0.02	0.01	-0.05	-0.01	0.01	-0.05	0.00	0.00
0.2	Non-informative	0.1	-0.01	-0.01	-0.01	0.00	0.01	-0.05	0.00	0.00	0.00
		0.5	-0.01	-0.01	-0.01	0.00	0.00	-0.01	0.00	0.00	0.00
	Informative at level-1	0.1	**0.85**	0.02	**0.09**	-0.01	-0.01	0.02	**-0.08**	0.00	-0.02
		0.5	**0.63**	-0.01	0.00	0.00	0.00	0.00	-0.05	0.00	0.00
	Informative at level-2	0.1	**0.44**	-0.04	0.03	**-0.09**	-0.02	**-0.06**	0.01	0.01	0.00
		0.5	**0.32**	-0.01	0.00	-0.05	0.00	-0.01	0.00	0.00	0.00
	Informative at both levels	0.1	**1.30**	-0.01	**0.13**	**-0.11**	-0.04	0.01	**-0.09**	0.00	-0.02
		0.5	**0.97**	-0.01	0.01	-0.05	-0.01	-0.01	-0.05	0.00	0.00
0.5	Non-informative	0.1	-0.01	-0.01	-0.01	0.00	0.01	-0.04	0.00	0.00	0.00
		0.5	-0.01	-0.01	-0.01	0.00	0.00	-0.01	0.00	0.00	0.00
	Informative at level-1	0.1	**0.67**	0.02	**0.07**	0.00	0.00	-0.03	**-0.08**	0.00	-0.02
		0.5	**0.49**	-0.01	-0.01	0.00	0.00	0.00	-0.05	0.00	0.00
	Informative at level-2	0.1	**0.70**	0.00	0.05	**-0.09**	-0.01	-0.05	0.01	0.01	0.00
		0.5	**0.51**	0.00	0.00	-0.05	0.00	-0.01	0.00	0.00	0.00
	Informative at both levels	0.1	**1.38**	0.03	**0.13**	**-0.10**	-0.02	-0.04	**-0.09**	-0.01	-0.02
		0.5	**1.02**	0.00	0.01	-0.05	0.00	-0.01	-0.05	0.00	0.00

Note. Values in bold represent unacceptably large relative bias (i.e., absolute value > 0.05)

5.4.2 Random slopes models Tables 6 to 9 show the relative biases and coverage rates for parameter estimates under all conditions based on the random slopes models. Notably, while convergence was not an issue for ML and the bootstrap method, MPML estimation suffered from a low convergence rate (ranging between 0.59 and 0.76) when both ICC and sample size were small.

Intercept. Similar to the pattern under the random intercept models, ML estimates of the intercept suffered from moderate to large relative biases (ranging

Table 3. Relative Bias for the Random Intercept Model Under $\chi^2(2)$ Distribution

ICC	Selection Mechanism	Sampling Fraction	Intercept			TAU			SIGMA		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-	0.1	.00	.00	.00	0.03	**0.07**	**-0.07**	0.00	0.00	0.00
	informative	0.5	.00	.00	.00	0.00	0.00	-0.02	0.00	0.00	0.00
	Informative	0.1	**0.81**	0.01	**0.08**	0.04	**0.09**	-0.05	**0.12**	0.01	0.02
	at level-1	0.5	**0.60**	0.00	0.00	0.00	0.00	-0.04	**0.10**	0.00	0.00
	Informative	0.1	**0.18**	-0.04	0.01	**0.12**	**0.11**	**-0.09**	-0.01	-0.01	-0.01
	at level-2	0.5	**0.14**	-0.02	0.00	**0.10**	0.02	-0.01	0.00	0.00	0.00
	Informative	0.1	**1.00**	-0.02	**0.10**	**0.11**	**0.10**	**-0.08**	**0.11**	0.00	0.01
	at both levels	0.5	**0.75**	-0.02	0.01	**0.10**	0.03	-0.04	**0.10**	0.00	0.00
0.2	Non-	0.1	-0.01	-0.01	-0.01	0.00	-0.01	**-0.06**	0.00	0.00	0.00
	informative	0.5	-0.01	-0.01	-0.01	-0.01	0.01	-0.02	0.00	0.00	0.00
	Informative	0.1	**0.74**	0.01	**0.07**	0.01	0.01	-0.05	**0.12**	0.01	0.02
	at level-1	0.5	**0.55**	-0.01	0.00	-0.01	-0.01	-0.02	**0.10**	0.00	0.00
	Informative	0.1	**0.37**	-0.04	0.01	**0.11**	0.03	**-0.06**	-0.01	-0.01	-0.01
	at level-2	0.5	**0.28**	-0.01	0.00	**0.10**	0.01	-0.01	0.00	0.00	0.00
	Informative	0.1	**1.12**	-0.02	0.10	**0.10**	0.03	-0.05	**0.11**	0.00	0.01
	at both levels	0.5	**0.83**	-0.01	0.01	**0.10**	0.01	-0.04	**0.10**	0.00	0.00
0.5	Non-	0.1	-0.01	-0.01	-0.01	-0.01	0.00	-0.05	0.00	0.00	0.00
	informative	0.5	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	0.00	0.00	0.00
	Informative	0.1	**0.57**	0.01	0.05	0.00	0.00	-0.04	**0.12**	0.01	0.02
	at level-1	0.5	**0.43**	-0.01	0.00	-0.01	-0.01	-0.02	**0.10**	0.00	0.00
	Informative	0.1	**0.58**	-0.02	0.02	**0.11**	0.01	-0.04	-0.01	-0.01	-0.01
	at level-2	0.5	**0.44**	0.00	0.00	**0.10**	0.00	-0.01	0.00	0.00	0.00
	Informative	0.1	**1.17**	0.00	**0.09**	**0.11**	0.02	-0.04	**0.11**	0.01	0.01
	at both levels	0.5	**0.88**	-0.01	0.01	**0.10**	0.00	-0.01	**0.10**	0.00	0.00

Note. Values in bold represent unacceptably large relative bias (i.e., absolute value > 0.05)

Table 4. Coverage Rate for the Random Intercept Model Under Normal Distribution

ICC	Selection Mechanism	Sampling Fraction	Intercept			X2 Slope			TAU			SIGMA		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-	0.1	0.93	0.94	0.93	0.95	0.95	0.94	**0.91**	**0.92**	**0.91**	0.97	0.96	0.94
	informative	0.5	0.96	0.97	0.95	0.94	0.94	0.94	0.95	0.96	0.94	0.95	0.95	0.95
	Informative	0.1	**0.05**	0.96	0.97	0.95	0.96	0.93	**0.92**	0.95	0.96	**0.73**	0.93	0.93
	at level-1	0.5	**0.49**	0.96	0.96	0.93	0.93	**0.91**	0.94	0.94	0.93	**0.01**	0.95	0.96
	Informative	0.1	**0.74**	**0.92**	0.94	0.94	0.94	**0.91**	0.93	0.93	**0.92**	**0.98**	0.94	0.94
	at level-2	0.5	**0.14**	0.93	0.96	0.95	0.94	0.94	0.91	0.93	0.94	0.96	0.95	0.96
	Informative	0.1	**0.00**	**0.88**	**0.88**	0.94	0.95	**0.89**	**0.92**	0.95	0.93	**0.71**	**0.91**	**0.92**
	at both levels	0.5	**0.00**	0.95	0.95	0.94	0.94	0.94	**0.90**	**0.92**	0.94	**0.02**	0.94	0.94
0.2	Non-	0.1	0.94	0.94	0.94	0.95	0.95	0.94	**0.90**	**0.91**	**0.86**	0.96	0.96	0.93
	informative	0.5	0.95	0.96	0.95	0.93	0.94	**0.92**	0.94	0.94	0.94	0.95	0.96	0.95
	Informative	0.1	**0.00**	0.94	**0.90**	0.95	0.96	0.93	**0.92**	**0.92**	**0.90**	**0.71**	**0.92**	0.93
	at level-1	0.5	**0.00**	0.95	0.95	0.93	0.93	**0.92**	0.94	0.94	0.94	**0.01**	0.96	0.96
	Informative	0.1	**0.51**	0.93	0.93	0.94	0.96	0.93	**0.91**	**0.92**	**0.88**	0.95	0.93	0.94
	at level-2	0.5	**0.06**	0.96	0.96	0.94	0.94	0.94	**0.88**	**0.92**	**0.92**	0.96	0.95	0.96
	Informative	0.1	**0.00**	**0.90**	**0.90**	0.95	0.95	**0.91**	**0.88**	**0.89**	**0.90**	**0.69**	**0.91**	**0.92**
	at both levels	0.5	**0.00**	0.96	0.97	0.95	0.96	0.93	**0.87**	0.93	**0.92**	**0.02**	0.94	0.94
0.5	Non-	0.1	0.95	0.94	0.95	0.95	0.94	0.94	0.94	0.93	**0.87**	0.96	0.95	0.93
	informative	0.5	0.95	0.96	0.95	0.93	0.93	0.93	0.93	0.94	0.94	0.95	0.96	0.95
	Informative	0.1	**0.04**	0.95	0.94	0.95	0.95	0.93	**0.91**	**0.91**	**0.87**	**0.71**	**0.92**	0.93
	at level-1	0.5	**0.00**	0.95	0.95	0.93	0.93	0.93	0.94	0.95	0.94	**0.01**	0.95	0.96
	Informative	0.1	**0.41**	0.93	0.94	0.95	0.96	**0.92**	**0.89**	**0.91**	**0.87**	0.95	0.94	0.94
	at level-2	0.5	**0.05**	0.96	0.97	0.94	0.95	0.94	0.97	0.93	**0.92**	0.96	0.95	0.96
	Informative	0.1	**0.02**	**0.92**	**0.91**	0.95	0.95	**0.92**	**0.88**	**0.89**	**0.87**	**0.69**	**0.91**	**0.92**
	at both levels	0.5	**0.00**	0.96	0.97	0.94	0.95	0.94	**0.87**	**0.92**	**0.92**	**0.02**	0.93	0.94

Note. Values in bold represent under-coverage or over-coverage (i.e., coverage rate < 0.93 or > 0.97)

Table 5. Coverage Rate for the Random Intercept Model Under $\chi^2(2)$ Distribution

ICC	Selection Mechanism	Sampling Fraction	Intercept			X2 Slope			TAU			SIGMA		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-informative	0.1	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.91**	**0.90**	0.93	**0.90**	**0.78**	**0.91**	**0.91**
		0.5	0.97	0.97	0.97	0.94	0.94	0.94	**0.79**	**0.89**	**0.92**	**0.67**	0.94	0.94
	Informative at level-1	0.1	**0.00**	0.94	0.94	0.94	0.94	**0.92**	**0.92**	**0.92**	**0.92**	**0.82**	0.96	0.94
		0.5	**0.00**	0.96	0.97	0.94	0.94	0.93	**0.81**	**0.90**	**0.91**	**0.00**	**0.98**	0.95
	Informative at level-2	0.1	**0.85**	**0.92**	0.93	0.95	0.95	**0.90**	**0.92**	0.96	0.95	**0.77**	**0.89**	**0.90**
		0.5	**0.25**	0.95	0.96	0.95	0.95	0.96	**0.77**	0.94	**0.92**	**0.69**	0.93	0.94
	Informative at both levels	0.1	**0.00**	0.93	**0.91**	0.95	0.94	0.93	0.93	0.93	0.95	**0.82**	0.95	0.95
		0.5	**0.00**	0.95	0.97	0.95	0.95	0.94	**0.78**	**0.91**	**0.89**	**0.00**	0.96	0.96
0.2	Non-informative	0.1	0.93	0.93	**0.92**	0.94	0.94	**0.92**	**0.77**	**0.83**	**0.78**	**0.66**	**0.91**	**0.91**
		0.5	0.97	0.97	0.97	0.93	0.93	**0.92**	**0.72**	**0.91**	**0.91**	**0.67**	0.94	0.94
	Informative at level-1	0.1	**0.10**	0.95	0.94	**0.92**	**0.92**	**0.92**	**0.77**	**0.83**	**0.79**	**0.58**	0.96	0.94
		0.5	**0.00**	0.97	0.97	0.93	0.93	0.94	**0.73**	**0.92**	**0.91**	**0.00**	**0.98**	0.95
	Informative at level-2	0.1	**0.71**	**0.92**	**0.92**	0.94	0.94	0.93	**0.83**	**0.90**	**0.83**	**0.69**	**0.89**	**0.90**
		0.5	**0.17**	0.97	0.96	0.95	0.94	0.96	**0.70**	0.94	**0.91**	**0.69**	0.93	0.94
	Informative at both levels	0.1	**0.00**	0.94	**0.92**	0.95	0.95	0.93	**0.85**	**0.90**	**0.86**	**0.62**	0.95	0.94
		0.5	**0.00**	0.96	0.96	0.95	0.94	0.95	**0.71**	0.95	**0.90**	**0.00**	0.97	0.96
0.5	Non-informative	0.1	0.93	0.93	**0.92**	0.94	0.94	**0.92**	**0.71**	**0.81**	**0.76**	**0.66**	**0.91**	**0.91**
		0.5	0.97	0.97	0.96	**0.92**	0.93	**0.92**	**0.70**	**0.92**	**0.91**	**0.67**	0.93	0.94
	Informative at level-1	0.1	**0.63**	0.93	0.93	0.93	**0.92**	**0.92**	**0.70**	**0.81**	**0.77**	**0.58**	0.95	0.94
		0.5	**0.11**	0.97	0.97	0.93	0.93	**0.92**	**0.70**	0.93	**0.91**	**0.00**	**0.98**	0.95
	Informative at level-2	0.1	**0.64**	**0.92**	**0.92**	0.95	0.95	0.94	**0.76**	**0.88**	**0.82**	**0.69**	**0.88**	**0.90**
		0.5	**0.14**	0.97	0.96	0.95	0.94	0.95	**0.67**	0.94	**0.90**	**0.69**	0.93	0.94
	Informative at both levels	0.1	**0.04**	0.93	0.94	0.95	0.94	0.93	**0.77**	**0.89**	**0.81**	**0.62**	0.95	0.94
		0.5	**0.00**	0.96	0.96	0.94	0.93	0.95	**0.67**	0.94	**0.90**	**0.00**	0.97	0.96

Note. Values in bold represent under-coverage or over-coverage (i.e., coverage rate < 0.93 or > 0.97)

from 0.23 to 1.64) when the selection mechanism was informative (see Table 6). The relative biases based on MPML estimates were acceptable under the majority of conditions, except for 6 conditions where the sample size was small and the selection mechanism was informative at level 1 or both levels (relative bias ranging from 0.12 to 0.14). The bootstrap method performed the best in terms of relative biases because there were only 3 conditions where small biases were found (ranging from -0.06 to -0.10).

Table 6. Relative Bias for Fixed Effects Estimates from the Random Slopes Model Under Normal Distribution

ICC	Selection Mechanism	Sampling Fraction	Intercept			X1		
			ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-informative	0.1	0.01	0.01	0.00	0.00	0.00	0.00
		0.5	0.00	0.00	0.00	0.00	0.00	0.00
	Informative at level-1	0.1	**0.93**	0.04	**0.12**	0.00	0.00	0.01
		0.5	**0.70**	0.00	0.01	0.00	0.00	0.00
	Informative at level-2	0.1	**0.30**	**-0.06**	0.00	**0.11**	0.06	0.00
		0.5	**0.23**	-0.03	0.00	**0.08**	0.02	0.00
	Informative at both levels	0.1	**1.23**	-0.03	**0.13**	**0.11**	**0.06**	0.01
		0.5	**0.92**	-0.03	0.01	**0.08**	0.02	0.00
0.2	Non-informative	0.1	0.00	0.01	0.00	0.00	0.00	0.00
		0.5	-0.01	0.00	0.00	0.00	0.00	0.00
	Informative at level-1	0.1	**0.85**	0.04	**0.13**	0.00	0.00	0.01
		0.5	**0.64**	0.00	0.01	0.00	0.00	0.00
	Informative at level-2	0.1	**0.61**	**-0.10**	0.00	**0.22**	**0.06**	0.00
		0.5	**0.45**	-0.02	0.00	**0.16**	0.01	0.00
	Informative at both levels	0.1	**1.46**	**-0.07**	**0.14**	**0.22**	**0.06**	0.01
		0.5	**1.09**	-0.02	0.01	**0.16**	0.01	0.00
0.5	Non-informative	0.1	0.00	0.01	0.01	0.00	0.00	0.00
		0.5	-0.01	0.00	0.00	0.00	0.00	0.00
	Informative at level-1	0.1	**0.66**	0.03	**0.13**	0.00	0.00	0.01
		0.5	**0.50**	0.00	0.01	0.00	0.00	0.00
	Informative at level-2	0.1	**0.96**	**-0.07**	0.00	**0.35**	0.04	0.00
		0.5	**0.71**	-0.01	0.00	**0.25**	0.01	0.00
	Informative at both levels	0.1	**1.64**	-0.04	**0.14**	**0.35**	0.04	0.01
		0.5	**1.22**	-0.01	0.01	**0.25**	0.01	0.00

Note. Values in bold represent unacceptably large relative bias (i.e., absolute value > 0.05)

As a result of the biased point estimate based on ML, the coverage rates of the confidence intervals for the ML estimates were also poor (ranging from 0.00 to 0.61) under informative selection mechanisms (see Table 6). On the other hand, both MPML and the bootstrap method had the issue of over-coverage (coverage rate above 0.98) in the majority of the conditions, indicating that the estimated confidence intervals were wider than expected.

Table 7. Coverage Rate for Fixed Effects Estimates from the Random Slopes Model Under Normal Distribution

ICC	Selection Mechanism	Sampling Fraction	Intercept			X1			X2		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-informative	0.1	0.96	**0.99**	**0.98**	0.96	0.96	0.96	0.95	0.95	0.93
		0.5	0.96	**1.00**	**1.00**	0.97	**0.99**	**0.99**	0.95	0.95	0.96
	Informative at level-1	0.1	**0.00**	0.97	**0.89**	0.96	0.97	0.95	0.94	0.95	**0.91**
		0.5	**0.00**	**1.00**	**1.00**	0.96	**0.99**	**0.98**	0.95	0.95	0.94
	Informative at level-2	0.1	**0.61**	0.95	**0.98**	**0.72**	**0.85**	0.95	0.95	0.95	**0.86**
		0.5	**0.01**	0.97	**0.99**	**0.02**	**0.86**	**0.98**	0.96	0.96	0.95
	Informative at both levels	0.1	**0.00**	**0.90**	**0.89**	**0.70**	**0.84**	0.96	0.95	0.96	**0.89**
		0.5	**0.00**	0.97	0.99	**0.02**	**0.88**	**0.99**	0.96	0.96	0.95
0.2	Non-informative	0.1	0.95	**0.99**	**0.99**	0.96	0.97	0.97	0.94	0.94	**0.92**
		0.5	0.96	**1.00**	**1.00**	0.97	**1.00**	**1.00**	0.96	0.96	0.95
	Informative at level-1	0.1	**0.06**	**0.99**	0.96	0.95	**0.99**	0.98	0.94	0.94	**0.91**
		0.5	**0.00**	**1.00**	**1.00**	0.96	**1.00**	**1.00**	0.95	0.95	0.95
	Informative at level-2	0.1	**0.23**	0.97	**0.99**	**0.36**	**0.88**	0.96	0.94	0.95	**0.88**
		0.5	**0.00**	**1.00**	**1.00**	**0.00**	**0.98**	**1.00**	0.97	0.97	0.95
	Informative at both levels	0.1	**0.00**	0.94	0.93	0.33	0.86	0.96	0.94	0.96	**0.88**
		0.5	**0.00**	**1.00**	**0.99**	**0.00**	**0.98**	**1.00**	0.96	0.96	0.95
0.5	Non-informative	0.1	0.94	**0.99**	**0.99**	0.97	**1.00**	**1.00**	0.94	0.94	**0.91**
		0.5	0.95	**1.00**	**1.00**	0.96	**1.00**	**1.00**	0.96	0.96	0.95
	Informative at level-1	0.1	**0.49**	**1.00**	**0.99**	0.94	**1.00**	**1.00**	0.94	0.94	0.91
		0.5	**0.07**	**1.00**	**1.00**	0.96	**1.00**	**1.00**	0.95	0.97	0.95
	Informative at level-2	0.1	**0.14**	**0.98**	**0.99**	**0.17**	0.96	**0.98**	0.94	0.94	**0.87**
		0.5	**0.00**	**1.00**	**1.00**	**0.00**	**1.00**	**1.00**	0.96	0.97	0.95
	Informative at both levels	0.1	**0.00**	0.96	0.97	**0.18**	0.95	**0.98**	0.95	0.96	**0.88**
		0.5	**0.00**	**1.00**	**1.00**	**0.00**	**1.00**	**1.00**	0.97	0.97	0.96

Note. Values in bold represent under-coverage or over-coverage (i.e., coverage rate $<$ 0.93 or $>$ 0.97)

Table 8. Relative Bias for Variance Components Estimates from the Random Slopes Model Under Normal Distribution

ICC	Selection Mechanism	Sampling Fraction	TAU00			TAU11			TAU01			SIGMA		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-	0.1	0.02	0.01	**-0.87**	**0.16**	**0.20**	**0.55**	**-0.09**	**-0.35**	**-0.57**	-0.01	0.03	0.01
	informative	0.5	0.01	**-0.10**	**-0.35**	0.01	**-0.07**	**-0.09**	0.01	**-0.26**	**-0.40**	0.00	0.02	0.01
	Informative	0.1	-0.01	-0.02	**-0.61**	**0.17**	**0.25**	**1.24**	-0.04	**-0.31**	**-0.48**	**-0.09**	0.02	-0.01
	at level-1	0.5	0.00	**-0.10**	**-0.34**	0.01	**-0.07**	-0.01	0.01	**-0.27**	**-0.41**	-0.05	0.02	0.01
	Informative	0.1	**-0.11**	-0.05	**-0.83**	-0.01	**0.19**	**0.69**	**-0.34**	**-0.47**	**-0.50**	-0.01	0.03	0.01
	at level-2	0.5	**-0.09**	**-0.11**	**-0.35**	**-0.08**	**-0.10**	**-0.09**	**-0.15**	**-0.30**	**-0.40**	0.00	0.02	0.01
	Informative	0.1	**-0.18**	**-0.10**	**-0.62**	**0.06**	**0.26**	**1.61**	**-0.36**	**-0.47**	**-0.61**	**-0.09**	0.03	-0.02
	at both levels	0.5	**-0.09**	**-0.11**	**-0.34**	**-0.09**	**-0.10**	-0.01	**-0.16**	**-0.30**	**-0.41**	-0.05	0.02	0.01
0.2	Non-	0.1	0.00	**-0.08**	**-0.38**	0.01	**-0.25**	-0.01	0.01	-0.05	**-0.39**	0.00	**0.08**	0.05
	informative	0.5	0.00	**-0.11**	**-0.16**	0.01	**-0.36**	**-0.10**	0.00	**-0.09**	**-0.40**	0.00	0.02	0.01
	Informative	0.1	0.00	**-0.09**	**-0.27**	0.04	**-0.23**	**0.17**	0.02	-0.04	**-0.39**	**-0.09**	**0.07**	0.00
	at level-1	0.5	0.00	**-0.11**	**-0.16**	0.01	**-0.37**	**-0.08**	0.00	**-0.09**	**-0.41**	-0.05	0.02	0.01
	Informative	0.1	**-0.15**	**-0.12**	**-0.38**	**-0.24**	**-0.30**	-0.01	**-0.16**	**-0.12**	**-0.38**	0.00	**0.09**	0.05
	at level-2	0.5	**-0.09**	**-0.11**	**-0.16**	**-0.16**	**-0.37**	**-0.10**	**-0.09**	**-0.10**	**-0.41**	0.00	0.02	0.01
	Informative	0.1	**-0.17**	**-0.13**	**-0.31**	**-0.24**	**-0.32**	**0.23**	**-0.12**	**-0.07**	**-0.41**	**-0.09**	**0.08**	0.00
	at both levels	0.5	**-0.09**	**-0.11**	**-0.15**	**-0.16**	**-0.37**	**-0.08**	**-0.09**	**-0.10**	**-0.41**	-0.05	0.02	0.01
0.5	Non-	0.1	0.01	**-0.09**	**-0.18**	0.01	**-0.33**	**-0.07**	0.01	**-0.07**	**-0.40**	0.00	**0.11**	0.04
	informative	0.5	0.00	**-0.10**	**-0.12**	0.01	**-0.39**	**-0.10**	0.00	**-0.10**	**-0.41**	0.00	0.03	0.01
	Informative	0.1	0.00	**-0.09**	**-0.16**	0.02	**-0.32**	-0.03	0.01	**-0.07**	**-0.39**	**-0.09**	**0.09**	0.00
	at level-1	0.5	0.00	**-0.11**	**-0.12**	0.01	**-0.40**	**-0.10**	0.00	**-0.10**	**-0.41**	-0.05	0.02	0.01
	Informative	0.1	**-0.15**	**-0.10**	**-0.20**	**-0.23**	**-0.33**	**-0.09**	**-0.15**	**-0.10**	**-0.37**	0.00	**0.12**	0.05
	at level-2	0.5	**-0.09**	**-0.10**	**-0.12**	**-0.16**	**-0.40**	**-0.10**	**-0.09**	**-0.10**	**-0.41**	0.00	0.03	0.01
	Informative	0.1	**-0.16**	**-0.10**	**-0.18**	**-0.23**	**-0.34**	-0.01	**-0.14**	**-0.08**	**-0.39**	**-0.09**	**0.10**	0.00
	at both levels	0.5	**-0.09**	**-0.10**	**-0.12**	**-0.16**	**-0.40**	**-0.10**	**-0.09**	**-0.10**	**-0.41**	-0.05	0.02	0.01

Note. Values in bold represent unacceptably large relative bias (i.e., absolute value > 0.05)

Table 9. Coverage rate for Variance Components Estimates from the Random Slopes Model Under Normal Distribution

ICC	Selection Mechanism	Sampling Fraction	TAU00			TAU11			TAU01			SIGMA		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-	0.1	0.95	0.93	**0.71**	**0.98**	**1.00**	**0.71**	0.97	0.95	**0.87**	0.94	0.93	0.94
	informative	0.5	0.96	**0.85**	**0.09**	0.94	**0.92**	**0.41**	0.95	**0.76**	**0.92**	0.95	**0.70**	**0.84**
	Informative	0.1	0.95	**0.95**	**0.86**	**0.99**	**1.00**	**0.81**	0.96	0.94	**0.66**	**0.68**	**0.88**	0.95
	at level-1	0.5	0.96	**0.85**	**0.11**	0.95	**0.92**	**0.39**	0.96	**0.75**	0.95	**0.02**	**0.71**	**0.91**
	Informative	0.1	0.96	0.95	**0.68**	**0.98**	**0.99**	**0.70**	0.94	**0.92**	**0.86**	0.96	**0.83**	0.94
	at level-2	0.5	**0.85**	**0.80**	**0.16**	**0.90**	**0.87**	**0.44**	**0.87**	**0.68**	0.93	0.94	**0.68**	**0.86**
	Informative	0.1	0.94	0.93	**0.84**	**0.98**	**0.99**	**0.77**	0.96	**0.91**	**0.66**	**0.68**	**0.76**	**0.91**
	at both levels	0.5	**0.85**	**0.83**	**0.20**	**0.90**	**0.86**	**0.46**	**0.85**	**0.65**	0.96	**0.03**	**0.67**	**0.89**
0.2	Non-	0.1	0.94	0.93	**0.78**	0.96	0.94	**0.85**	0.93	0.94	0.97	0.95	**0.79**	**0.91**
	informative	0.5	0.96	**0.77**	**0.56**	0.94	**0.83**	**0.16**	0.97	**0.31**	**0.82**	0.95	**0.54**	0.84
	Informative	0.1	0.95	0.93	**0.84**	0.96	0.96	**0.85**	0.93	0.93	**0.92**	**0.73**	**0.78**	0.94
	at level-1	0.5	0.96	0.77	**0.57**	0.94	**0.82**	**0.15**	0.96	**0.28**	**0.86**	**0.02**	**0.60**	**0.91**
	Informative	0.1	**0.86**	**0.87**	**0.77**	**0.90**	**0.88**	**0.83**	**0.92**	**0.87**	0.95	0.96	**0.70**	**0.91**
	at level-2	0.5	**0.80**	**0.75**	**0.64**	**0.84**	**0.81**	**0.24**	**0.80**	**0.29**	**0.87**	0.94	**0.55**	**0.86**
	Informative	0.1	**0.84**	**0.86**	**0.85**	**0.91**	**0.91**	**0.82**	**0.90**	**0.87**	**0.88**	**0.72**	**0.70**	**0.92**
	at both levels	0.5	**0.80**	**0.75**	**0.65**	**0.84**	**0.79**	**0.25**	**0.80**	**0.30**	**0.89**	**0.03**	**0.57**	**0.89**
0.5	Non-	0.1	0.95	0.90	**0.85**	0.96	0.93	**0.75**	0.94	**0.86**	0.94	0.95	**0.66**	**0.91**
	informative	0.5	0.96	**0.75**	**0.72**	0.93	**0.77**	**0.11**	0.97	**0.17**	**0.75**	0.95	**0.50**	**0.84**
	Informative	0.1	0.96	**0.89**	**0.86**	0.95	**0.92**	**0.76**	0.93	**0.86**	0.94	**0.73**	**0.71**	0.94
	at level-1	0.5	**0.85**	**0.75**	**0.70**	0.93	**0.77**	**0.10**	0.95	**0.17**	**0.78**	**0.02**	**0.57**	**0.91**
	Informative	0.1	**0.83**	**0.82**	**0.78**	**0.87**	**0.85**	**0.75**	**0.88**	**0.78**	**0.90**	0.97	**0.62**	**0.91**
	at level-2	0.5	**0.79**	**0.75**	**0.75**	**0.82**	**0.77**	**0.19**	**0.77**	**0.19**	**0.83**	0.94	**0.52**	**0.85**
	Informative	0.1	**0.80**	**0.83**	**0.82**	**0.88**	**0.85**	**0.75**	**0.84**	**0.76**	**0.91**	**0.72**	**0.64**	**0.92**
	at both levels	0.5	**0.79**	**0.75**	**0.76**	**0.81**	**0.75**	**0.20**	**0.77**	**0.20**	**0.82**	**0.02**	**0.54**	**0.89**

Note. Values in bold represent under-coverage or over-coverage (i.e., coverage rate < 0.93 or > 0.97)

Slope of *X1*. As expected, the ML estimates of the slope of *X1* were biased when the selection mechanism was informative at level 2 or both levels (relative bias ranging between 0.08 and 0.35). The magnitude of the biases increased as ICC increased. On the other hand, both the MPML and the bootstrap estimation methods successfully reduced the biases to an acceptable level, although the MPML method performed slightly better than the bootstrap method when sample size was small and the selection mechanism was informative at level 2 or both levels.

Similarly, due to the biased point estimates, the coverage rates of the confidence intervals for the ML estimates were also poor (ranging from 0.00 to 0.72) under informative selection mechanisms. The MPML confidence intervals demonstrated over-coverage, especially when sample size and ICC were large. The bootstrap confidence intervals demonstrated slight under-coverage (ranging between 0.84 and 0.88) when informative selection occurred at level 2 or both levels, but showed a similar over-coverage pattern as the MPML confidence intervals in the other conditions.

Slope of *X2*. The relative bias of the estimated slope of *X2* was acceptable for all methods under all conditions. However, the MPML confidence intervals suffered from slight under-coverage (0.86 to 0.92) in about half of the conditions, mainly when sample size was small. The performance of the ML and the bootstrap confidence intervals was acceptable under all conditions.

Variance of the random intercepts (τ_{00}). ML estimates had small relative bias (-0.09 to -0.18), mainly when there was informative sampling at level-2 or at both levels. MPML suffered from moderate to large biases (-0.34 to -0.87) when ICC was small. The magnitude of the relative biases decreased as ICC or sample size increased, but was still more than 0.12 when ICC and sample size were large. The bootstrap method showed small negative biases (-0.08 to -0.13) across all conditions consistently and had the greatest advantages over MPML when ICC was small.

The coverage rates of the confidence intervals based on the three methods showed similar patterns. The ML-based confidence intervals had slight under-coverage (0.79 to 0.85) when there was informative sampling at level-2 or at both levels. The MPML-based confidence intervals suffered from severe under-coverage (0.09 to 0.20) under conditions where small ICCs were combined with large sample sizes. The bootstrap confidence intervals had slight under-coverage (0.75 to 0.90) in the majority of the conditions. It is noted that when ICC was large, MPML and bootstrap confidence intervals performed similarly.

Variance of the random slopes (τ_{11}). Similar to τ_{00}, ML estimate of τ_{11} showed small to moderate relative bias (-0.16 to 0.17), mainly when there was informative sampling at level-2 or at both levels. The MPML estimates had large positive biases (0.55 to 1.61) when ICC and sample size were both small, and mostly small negative biases (-0.08 to -0.10) under the other conditions. The bootstrap estimates had small positive biases (0.19 to 0.26) when ICC and sample size were both small, and moderate negative biases (-0.23 to -0.40) when

ICC was moderate and large. Comparing the three methods, ML showed the least amount of bias across all conditions.

In terms of the confidence intervals, MPML had the worst performance because of the severe under-coverage (0.10-0.46) when sample size was large. The bootstrap confidence intervals had somewhat under-coverage (0.77-0.92) across the conditions. The ML confidence intervals had the best performance, showing slight under-coverage (0.81 to 0.91) when there was informative sampling at level-2 or at both levels.

Covariance of the random intercepts and the random slopes (τ_{01}). The ML estimate of τ_{01} showed small to moderate negative biases (-0.09 to -0.36) when there was informative sampling at level-2 or at both levels. The MPML estimates showed moderate negative biases across all conditions, ranging from -0.37 to -0.61. The bootstrap estimates showed small to moderate negative biases, with the magnitude decreasing from -0.34 to -0.09 as ICC increased from 0.05 to 0.5.

The ML confidence intervals had slight under-coverage (0.77 to 0.92) when there was informative sampling at level-2 or at both levels. Despite the moderate negative biases in the point estimates, MPML confidence intervals only showed slight under-coverage in most of the conditions (0.66 to 0.92). In general, the bootstrap confidence intervals suffered from under-coverage (0.17 to 0.92), and the degree of under-coverage was severe (0.17 to 0.31) when sample sizes were large and ICCs were moderate to large.

Level-1 residual variance (σ^2). ML estimates had small negative relative biases (-0.09) when there was informative selection at level-1 or at both levels. The bootstrap estimates showed small positive relative biases (0.07 to 0.12) when sample size was small and ICC was moderate to large. MPML estimates had the best performance with little bias across all conditions.

The ML-based confidence intervals showed under-coverage when there was informative selection at level-1 or at both levels. The degree of under-coverage was severe (0.02 to 0.03) when sample size was large. The bootstrap confidence interval had moderate under-coverage across all conditions, ranging from 0.50 to 0.88. The MPML confidence intervals had slight under-coverage across all conditions, ranging from 0.84 to 0.91.

6 Discussion and Conclusion

We proposed a weighted residual bootstrap method for multilevel modeling of data from complex sampling designs. Unlike previously proposed bootstrap methods (e.g., Grilli & Pratesi, 2004; Kovacevic et al., 2006; Wang & Thompson, 2012), our method does not require generating a pseudo population or rescaling weights. The performance of the proposed bootstrap method for linear two-level models was investigated under various conditions, and compared with the multilevel pseudo maximum likelihood (MPML) approach and the unweighted ML approach using Monte Carlo simulations.

In general, the proposed weighted bootstrap method performed similar to or better than the MPML method in random intercept models and had mixed results in random slopes models. As expected, for the random intercept model, unweighted ML resulted in biased intercept estimate when there were informative selections. Both the bootstrap and the MPML estimates of the slopes for the level-1 and level-2 predictors (*X1* and *X2*) had acceptable performance. However, the bootstrap showed advantages over MPML for the estimate of the level-2 variance component when sample size is small (i.e., 50 clusters and 10 units per cluster), selection mechanism is informative, and ICC is low (i.e., 0.05). As a result, the confidence interval of the slope of the level-2 predictor (*X2*) based on the bootstrap method also had a better coverage rate compared to MPML under those conditions. It has been demonstrated in the literature that MPML estimates have increased biases as ICC decreases (Asparouhov, 2006; Kovacevic & Rai, 2003). As Asparouhov (2006) explained, the weakness of MPML is in the estimation on the individual level, therefore as ICC decreases the individual level becomes more influential, which exacerbates the problem.

For the random slopes model, the ML estimates of both the intercept and the slope of the level-1 predictor (*X1*) showed moderate to severe biases when there are informative selections. The bootstrap and the MPML approaches performed similarly in terms of the estimates of the fixed effects, with the bootstrap method slightly better for the estimate of the intercept and the slope of the level-2 predictor, while the MPML slightly better for the slope of the level-1 predictor. While convergence was not an issue for the bootstrap method, MPML suffered from a high rate of non-convergence when ICC is low. As a result, MPML had severe biases in the estimates of the level-2 variance components when ICC is low. The performance of the bootstrap estimate of the variance components was not ideal either as small to moderate biases existed across the conditions. However, the bootstrap confidence intervals performed much better than the MPML approach, especially when sample size is large. The only drawback of the bootstrap method is in the estimation of the covariance between the random intercept and the random slope, which showed severe under-coverage when sample size is large.

Another advantage of the bootstrap method is that it is more robust to the distributional violation. Previous simulation studies on MPML for linear models only considered normally distributed random effects and residuals. Our findings showed that when the normality assumption was violated, the coverage rate of the MPML confidence interval for the level-2 variance component in a random intercept model became much worse with 8 more conditions showing under-coverage. The bootstrap method was also affected by the distributional violation, but to a lesser degree because only 4 more conditions showed under-coverage when the distributions were skewed.

As a demonstration, the weighted residual bootstrap method was applied to the American 2000 PISA data on math achievement. Based on the random intercept model, the bootstrap and the MPML results showed some inconsistency, especially for the slope of the level-2 predictor. We believe that the bootstrap results were more trustworthy in this case because conditions in the simulation study that were similar to the specific condition of this sample (i.e., small cluster size, low ICC, very slightly informative, and slight distributional violation) have shown favorable results in the bootstrap than the MPML method.

6.1 Implications

The weighted residual bootstrap method provides a robust alternative to MPML. Applied researchers can use the bootstrap approach when the traditional MPML estimation fails to converge or when there is severe violation of the normality assumption. In analyses of random intercept models, the weighted residual bootstrap method is preferred to MPML when the effect of level-2 predictors (e.g., school SES), or the variance of the random intercept (e.g., variance of school mean achievement) are of interest and when both sample sizes and ICCs are small. In random slopes models, the bootstrap method has advantages over MPML in the point estimates and the confidence interval estiamtes of the slopes of level-2 predictors, as well as the variance component estimates associated with the random intercept and the random slopes (e.g., variance of the association of student SES and student achievement across schools). However, the statistical inferences for the covariance component (e.g., the covariance between school mean achievement and the slope of student SES and student achievement) based on the bootstrap method might not be trustworthy.

It is recommended that researchers conduct sensitivity analyses using different methods. Discrepancies among the results may indicate that the conditions for MPML to work properly are not satisfied. The weighted residual bootstrap method is implemented in the developmental version of the R package *bootmlm*, which has the capacity to analyze two-level linear random intercept and random coefficients models with sampling weights.

6.2 Limitations and Future Directions

The findings of the study should be interpreted in light of the limitations. First, there is still room for improvement in terms of the bootstrap confidence interval for level-2 variance and covariance components. We used percentile confidence interval for its simplicity. Future research may be conducted to investigate whether more sophisticated methods such as bias-corrected and accelerated confidence intervals and studentized intervals could further improve the performance. Second, the proposed bootstrap method was only applied to multilevel linear models. Although it is possible to extend it to generalized multilevel models (Goldstein et al., 2018), Monte Carlo experiments should be conducted to examine the performance of the method for generalized multilevel models such as multilevel ordinal and binary models. Third, this study only compared the performance of the proposed method with MPML. Future studies could compare the proposed method with other bootstrap methods for multilevel data with sampling weights.

References

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling*, *12*(3), 411–434. doi: https://doi.org/10.1207/s15328007sem1203_4

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics—Theory and Methods*, *35*(3), 439–460. doi: https://doi.org/10.1080/03610920500476598

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: https://doi.org/10.18637/jss.v067.i01

Booth, J. (1995). Bootstrap methods for generalized linear mixed models with applications to small area estimation. In G. Seeber, J. Francis, R. Hatzinger, & G. Steckel-Berger (Eds.), *Statistical modelling* (pp. 43–51). New York, NY: Springer. doi: https://doi.org/10.1007/978-1-4612-0789-4_6

Carpenter, J. R., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *52*(4), 431–443. doi: https://doi.org/10.1111/1467-9876.00415

Cochran, W. G. (1977). *Sampling techniques (3rd ed.).* New York, NY: Wiley.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application.* Cambridge, UK: Cambridge University. doi: https://doi.org/10.1017/cbo9780511802843

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York, NY: Chapman and Hall. doi: https://doi.org/10.1201/9780429246593

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, *73*(1), 43–56. doi: https://doi.org/10.1093/biomet/73.1.43

Goldstein, H. (2011). Bootstrapping in multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (p. 163–171). New York, NY: Routledge.

Goldstein, H., Carpenter, J., & Kenward, M. G. (2018). Bayesian models for weighted data with missing values: a bootstrap approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*(4), 1071–1081. doi: https://doi.org/10.1111/rssc.12259

Grilli, L., & Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Statistics Canada*, *30*(1), 93–103.

Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(2), 221–238. doi: https://doi.org/10.1111/j.1467-9868.2006.00541.x

Hoogland, J. J., & Boomsma, A. (1998). *Robustness studies in covariance structure modeling.* Sociological Methods and Research, 26:329–367. doi: https://doi.org/10.1177/0049124198026003003

Kovacevic, M. S., Huang, R., & You, Y. (2006). *Bootstrapping for variance estimation in multi-level models fitted to survey data.* ASA Proceedings of the Survey Research Methods Section.

Kovacevic, M. S., & Rai, S. N. (2003). *A pseudo maximum likelihood approach to multi-level modeling of survey data.* Communications in Statistics—Theory and Methods, 32:103–121. doi: https://doi.org/10.1081/sta-120017802

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science*, *18*(2), 199–210. doi: https://doi.org/10.1214/ss/1063994975

Lohr, S. L. (2010). *Sampling: Design and analysis (2nd ed.).* Boston, MA: Cengage.

Maas, C. J., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*, 427–440. doi: https://doi.org/10.1016/j.csda.2003.08.006

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide (8th ed.).* Los Angeles, CA: Muthén & Muthén.

Organization for Economic Co-operation and Development. (2000). Manual for the pisa 2000 database [Computer software manual]. Retrieved from `http://www.pisa.oecd.org/dataoecd/53/18/33688135.pdf`

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistics Review.* doi: https://doi.org/10.2307/1403631

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multi-level models. *Journal of the Royal Statistics Society: Series B (Statistical Methodology).* doi: https://doi.org/10.1111/1467-9868.00106

Potthoff, R. F., Woodbury, M. A., & Manton, K. G. (1992). "equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of American Statistical Association.* doi: https://doi.org/10.2307/2290269

R Core Team. (2018). *R: A language and environment for statistical computing.* Vienna, Austria.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(4), 805–827. doi: https://doi.org/10.1111/j.1467-985x.2006.00426.x

Seco, G. V., García, M. A., García, M. P. F., & Rojas, P. E. L. (2013). Multilevel bootstrap analysis with assumptions violated. *Psicothema*, *25*(4), 520–528.

Stapleton, L. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling.* doi: https://doi.org/10.1207/s15328007sem0904_2

Thai, H. T., Mentré, F., Holford, N. H. G., Veyrat-Follet, C., & Comets, E. (2014). Evaluation of bootstrap methods for estimating uncertainty of parameters in nonlinear mixed-effects models: a simulation study in population pharmacokinetics. *Journal of Pharmacokinetics and Pharmacodynamics*, *41*(1).

Van der Leeden, R., Meijer, E., & Busing, F. M. (2008). Resampling multilevel models. In *Handbook of multilevel analysis* (pp. 401–433). New York, NY: Springer.

Verbeke, G., & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, *23*. doi: https://doi.org/10.1016/s0167-9473(96)00047-3

Wang, Z., & Thompson, M. E. (2012). A resampling approach to estimate variance components of multilevel models. *Canadian Journal of Statistics*, *40*(1), 150–171. doi: https://doi.org/10.1002/cjs.10136

Appendix A. R Code for the Analysis of PISA Data using Weighted Residual Bootstrap

```
# Check if devtools were installed
if (!require("devtools")) {
   install.packages("devtools")
}
# Install developmental version of the bootmlm package
devtools::install_github("marklhc/bootmlm",
  ref = "weighted_boot")

# Load required packages
library(bootmlm)
library(boot)
library(lme4)

# Unweighted ML
m1 <- lmer(SC17Q01 ~ ISEI_m + male + (1 | Sch_ID),
           data = PISA, REML = FALSE)

# Weighted semi-parameteric bootstrap
boo <- bootstrap_mer(
  m1,
  FUN = function(x) {
    c(x@beta,
      c(x@theta ^ 2, 1) * sigma(x) ^ 2)
  },
  nsim = 999L,
  type = "residual_cgr",
w1 = PISA$ W_FSTUWT,
w2 = unique(PISA[c("Sch_ID", "WNRSCHBW")])$WNRSCHBW
)

# Print the output
boo  # bootstrap results
colMeans(boo$t)  # parameter estimates
apply(boo$t, 2, sd)  # bootstrap SE

# Percentile intervals for the six parameters
boot.ci(boo, type = "perc", index = 1L)
boot.ci(boo, type = "perc", index = 2L)
boot.ci(boo, type = "perc", index = 3L)
boot.ci(boo, type = "perc", index = 4L)
boot.ci(boo, type = "perc", index = 5L)
boot.ci(boo, type = "perc", index = 6L)
```

Appendix B. Mplus Code for the Analysis of PISA Data using MPML

```
Data:         File=pisa.csv;
Variable:     Names are math ISEI_m male Sch_ID
              W_FSTUWT WNRSCHBW lv1_con_wt;
              Usevariables are math ISEI_m male;
              Between = ISEI_m;
              Within = male;
              Cluster = Sch_ID;
              Weight = lv1_con_wt;  !lv1_con_wt=
                 W_FSTUWT/WNRSCHBW;
              Bweight = WNRSCHBW;
Analysis:     Type = twolevel;
Model:        %within%
              math on male;
              %between%
              math on ISEI_m;
Output:       Cinterval;
```

Journal of Behavioral Data Science, 2020, *1* (2), 156–177.
DOI: https://doi.org/10.35566/jbds/v1n2/p7

Structural Equation Modeling using Stata

Meghan K. Cain[1] [0000−0003−4790−4843]

StataCorp LLC, College Station, TX 77845, USA
mcain@stata.com

Abstract. In this tutorial, you will learn how to fit structural equation models (SEM) using Stata software. SEMs can be fit in Stata using the sem command for standard linear SEMs, the gsem command for generalized linear SEMs, or by drawing their path diagrams in the SEM Builder. After a brief introduction to Stata, the sem command will be demonstrated through a confirmatory factor analysis model, mediation model, group analysis, and a growth curve model, and the gsem command will be demonstrated through a random-slope model and a logistic ordinal regression. Materials and datasets are provided online, allowing anyone with Stata to follow along.

Keywords: Structural Equation Modeling · Growth Curve Modeling · Mediation · Software.

1 Introduction

Structural equation modeling (SEM) is a multivariate statistical analysis framework that allows simultaneous estimation of a system of equations. SEM can be used to fit a wide range of models, including those involving measurement error and latent constructs. This tutorial will demonstrate how to fit a variety of SEMs using Stata statistical software (StataCorp, 2021). Specifically, we will fit models in Stata with both measurement and structural components, as well as those with random effects and generalized responses. We will assess model fit, compute modification indices, estimate mediation effects, conduct group analysis, and more. First, however, we will begin with an introduction to Stata itself. Familiarity with SEM theory and concepts is assumed.

Stata is a complete, integrated software package that provides tools for data manipulation, visualization, statistics, and automated reporting. The Data Editor, Variables window, and Properties window can be used to view and edit your dataset and to manage variables, including their names, labels, value labels, notes, formats, and storage types. Commands can be typed into the Command window, or generated through the point-and-click interface. Log files keep a record of every command issued in a session, while do-files save selected commands to allow users to replicate their work. To learn more about a command, you can type `help` followed by the command name in the Command window and the Viewer window will open with the help file and provide links to further documentation. Stata's documentation consists of over 17,000 pages detailing each feature in Stata including the methods and formulas and fully worked examples.

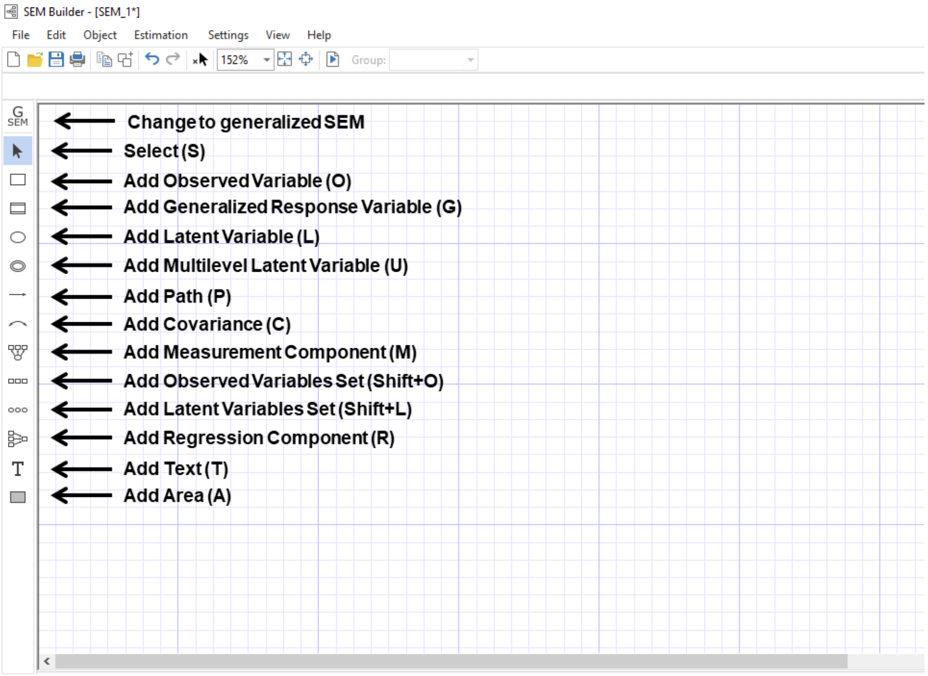

SEM Builder - [SEM_1*]
File Edit Object Estimation Settings View Help
152%
Group:
G SEM
Change to generalized SEM
Select (S)
Add Observed Variable (O)
Add Generalized Response Variable (G)
Add Latent Variable (L)
Add Multilevel Latent Variable (U)
Add Path (P)
Add Covariance (C)
Add Measurement Component (M)
Add Observed Variables Set (Shift+O)
Add Latent Variables Set (Shift+L)
Add Regression Component (R)
Add Text (T)
Add Area (A)

Figure 1. SEM Builder

There are three ways to fit SEMs in Stata: the `sem` command, the `gsem` command, and through the SEM Builder. The `sem` command is for fitting standard linear SEMs. It is quicker and has more features for testing and interpreting results than `gsem`. The `gsem` command is for fitting models with generalized responses, such as binary, count, or categorical responses, models with random effects, and mixture models. Both `sem` and `gsem` models can be fit via path diagrams using the SEM Builder. You can open the SEM Builder window by typing `sembuilder` into the Command window. See the interface in Figure 1; click the tools you need on the left, or type their shortcuts shown in the parentheses. To fit `gsem` models, the GSEM button must first be selected. Estimation and diagram settings can be changed using the menus at the top. The Estimate button fits the model. Path diagrams can be saved as `.stsem` files to be modified later, or can be exported to a variety of image formats (for example see Figure 2). Although this tutorial will focus on the `sem` and `gsem` commands, the Builder shares the same functionality. You can watch a demonstration with the SEM Builder on the StataCorp YouTube Channel: `https://www.youtube.com/watch?v=Xj0gBlqwYHI`

To download the datasets, do-file, and path diagrams, you can type the following into Stata's Command window:

```
. net from http://www.stata.com/users/mcain/JBDS_SEM
```

Clicking on the `SEMtutorial` link will download the materials to your current working directory. To open the do-file with the commands we'll be using, you can type

```
. doedit SEMtutorial
```

Commands can either be executed from the do-file or typed into the Command window. We'll start by loading and exploring our first dataset. These data contain observations on four indicators for socioeconomic status of high school students as well as their math scores, school types (private or public), and the student-teacher ratio of their school. Alternatively, we could have used a summary statistics dataset containing means, variances, and correlations of the variables rather than observations.

```
. use math
. codebook, compact
Variable   Obs Unique      Mean  Min  Max  Label
-------------------------------------------------------------------------------
schtype    519      2    .61079    0    1  School type
ratio      519     14  16.75723   10   28  Student-Teacher ratio
math       519     42  51.72254   30   71  Math score
ses1       519      5  1.982659    0    4  SES item 1
ses2       519      5  2.003854    0    4  SES item 2
ses3       519      5  2.003854    0    4  SES item 3
ses4       519      5  2.003854    0    4  SES item 4
-------------------------------------------------------------------------------
```

2 Fitting models with the `sem` command

2.1 Path Analysis

Let's start our analysis by fitting the one-factor confirmatory factor analysis (CFA) model shown in Figure 2. Using the `sem` command, paths are specified in parentheses and the direction of the relationships are specified using arrows, i.e. `(x->y)`. Arrows can point in either direction, `(x->y)` or `(y<-x)`. Paths can be specified individually, or multiple paths can be specified within a single set of parentheses, `(x1 x2 x3 -> y)`. By default, Stata assumes that all lower-case variables are observed and uppercase variables are latent. You can change these settings using the `nocapslatent` and the `latent()` options. In Stata, options are always added after a comma. We'll see plenty of examples of this later.

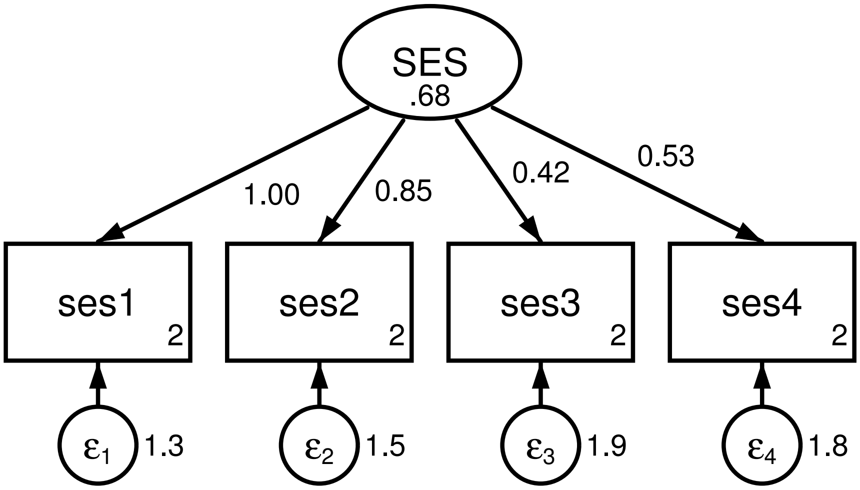
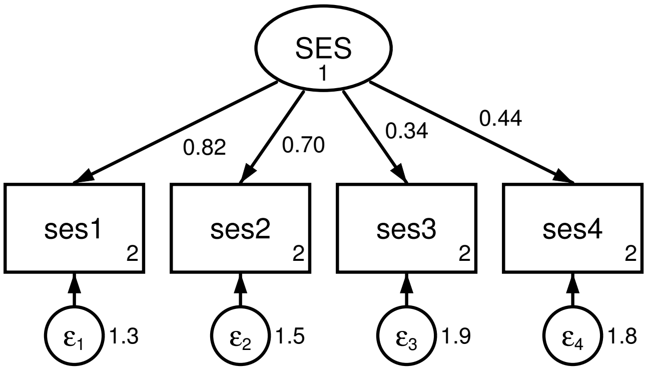


Figure 2. One-factor CFA measuring socioeconomic status (SES)

```
. sem (SES -> ses1-ses4)
Endogenous variables
  Measurement: ses1 ses2 ses3 ses4
Exogenous variables
  Latent: SES
Fitting target model:
Iteration 0:   log likelihood = -3621.9572
Iteration 1:   log likelihood = -3621.5801
Iteration 2:   log likelihood = -3621.5573
Iteration 3:   log likelihood =  -3621.557
Structural equation model                              Number of obs = 519
Estimation method: ml
Log likelihood = -3621.557
 ( 1)  [ses1]SES = 1
------------------------------------------------------------------------------
             |                 OIM
             | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
Measurement  |
  ses1       |
         SES |          1  (constrained)
       _cons |   1.982659   .0620424    31.96   0.000     1.861058     2.10426
  -----------+----------------------------------------------------------------
  ses2       |
         SES |   .8481035   .1962358     4.32   0.000     .4634884    1.232719
       _cons |   2.003854   .0620169    32.31   0.000     1.882303    2.125404
  -----------+----------------------------------------------------------------
  ses3       |
         SES |    .416385   .1331306     3.13   0.002     .1554539    .6773161
       _cons |   2.003854    .062017    32.31   0.000     1.882302    2.125405
  -----------+----------------------------------------------------------------
  ses4       |
         SES |   .5315065   .1517342     3.50   0.000      .234113    .8289001
       _cons |   2.003854    .062017    32.31   0.000     1.882302    2.125405
----------------+----------------------------------------------------------------
  var(e.ses1)  |   1.317579   .1855509                      .9997798    1.736397
  var(e.ses2)  |   1.506881   .1493285                      1.240872    1.829916
  var(e.ses3)  |   1.878203   .1257611                      1.647204    2.141595
  var(e.ses4)  |   1.803979   .1287389                      1.568507    2.074801
     var(SES)  |   .6801844   .1908617                      .3924434    1.178898
---------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(2) = 11.03          Prob > chi2 = 0.0040
```

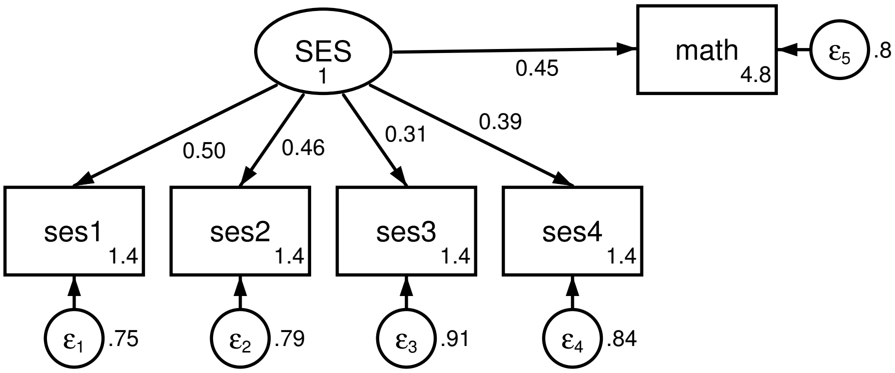
Viewing the results, we see that by default Stata constrained the first factor loading to be 1 and estimated the variance of the latent variable. If, instead, we would like to constrain the variance and estimate all four factor loadings, we could use the `var()` option. Constraints in any part of the model can be specified using the `@` symbol. To save room, syntax and results for this and the remaining models will be shown on their path diagrams; see Figure 3.



```
. sem (SES -> ses1-ses4), var(SES@1)
```

Figure 3. One-factor CFA with constrained variance.

Specifying structural paths is no different from specifying measurement paths. We can add math score to our model and hypothesize that socioeconomic status influences expected math performance. This model is shown in Figure 4; we've added the `standardized` option to get standardized coefficients. With every increase of one standard deviation in SES, math score is expected to increase by 0.45 standard deviations.



```
. sem (SES -> ses1-ses4 math), standardized
```

Figure 4. SES influences math scores.

To get fit indices for our model, we can use the postestimation command `estat gof` after any `sem` model. Add the `stats(all)` option to see all fit indices.

```
. estat gof, stats(all)

----------------------------------------------------------------------------
Fit statistic        |      Value   Description
---------------------+------------------------------------------------------
Likelihood ratio     |
         chi2_ms(5)  |     17.689   model vs. saturated
           p > chi2  |      0.003
        chi2_bs(10)  |    150.126   baseline vs. saturated
           p > chi2  |      0.000
---------------------+------------------------------------------------------
Population error     |
              RMSEA  |      0.070   Root mean squared error of approximation
 90% CI, lower bound |      0.037
        upper bound  |      0.107
             pclose  |      0.147   Probability RMSEA <= 0.05
---------------------+------------------------------------------------------
Information criteria |
                AIC  |  11157.441   Akaike´s information criterion
                BIC  |  11221.219   Bayesian information criterion
---------------------+------------------------------------------------------
Baseline comparison  |
                CFI  |      0.909   Comparative fit index
                TLI  |      0.819   Tucker-Lewis index
---------------------+------------------------------------------------------
Size of residuals    |
               SRMR  |      0.040   Standardized root mean squared residual
                 CD  |      0.532   Coefficient of determination
----------------------------------------------------------------------------
```

Satorra-Bentler adjusted model fit indices can be obtained by adding the `vce(sbentler)` option to our model statement and recalculating the model fit indices. This option still uses maximum likelihood estimation, the default, but adjusts the standard errors and the fit indices. Alternatively, estimation can be changed to asymptotic distribution-free or full-information maximum likelihood for missing values using the `method(adf)` or `method(mlmv)` options, respectively. For this example, we'll use the Satorra-Bentler adjustment (Satorra & Bentler, 1994). First, we'll store the current model to use again later.

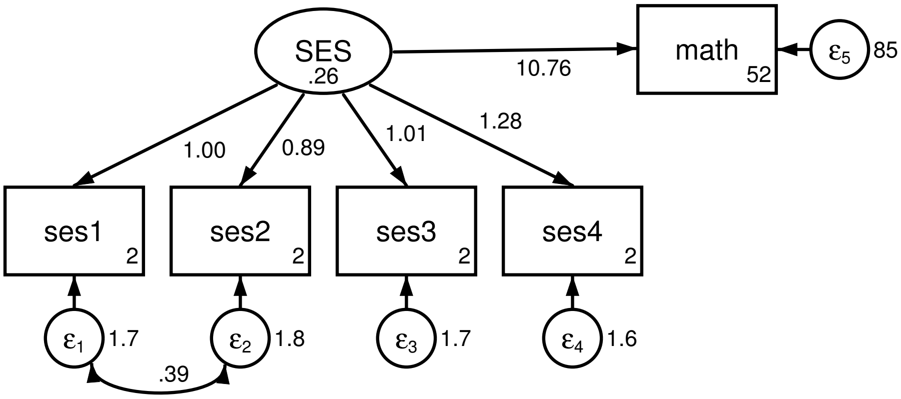
```
. estimates store m1
. sem (SES -> ses1-ses4 math), vce(sbentler)
Endogenous variables
  Measurement: ses1 ses2 ses3 ses4 math
Exogenous variables
  Latent: SES
Fitting target model:
Iteration 0:   log pseudolikelihood = -5564.2324
Iteration 1:   log pseudolikelihood = -5563.7459
Iteration 2:   log pseudolikelihood = -5563.7204
Iteration 3:   log pseudolikelihood = -5563.7204
Structural equation model                            Number of obs = 519
Estimation method: ml
Log pseudolikelihood = -5563.7204
 ( 1)  [ses1]SES = 1
------------------------------------------------------------------------------
             |              Satorra-Bentler
             |  Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
Measurement  |
  ses1       |
         SES |          1  (constrained)
       _cons |   1.982659   .0621024    31.93   0.000     1.860941    2.104377
  -----------+----------------------------------------------------------------
  ses2       |
         SES |   .9278593    .169484     5.47   0.000     .5956767    1.260042
       _cons |   2.003854   .0620769    32.28   0.000     1.882185    2.125522
  -----------+----------------------------------------------------------------
  ses3       |
         SES |    .620192   .1438296     4.31   0.000     .3382912    .9020928
       _cons |   2.003854   .0620769    32.28   0.000     1.882185    2.125522
  -----------+----------------------------------------------------------------
  ses4       |
         SES |   .7954927   .1580751     5.03   0.000     .4856712    1.105314
       _cons |   2.003854   .0620769    32.28   0.000     1.882185    2.125522
  -----------+----------------------------------------------------------------
  math       |
         SES |   6.858402   1.335695     5.13   0.000     4.240488    9.476315
       _cons |   51.72254   .4700825   110.03   0.000      50.8012    52.64389
-------------+----------------------------------------------------------------
  var(e.ses1)|   1.506551   .1203549                        1.2882    1.761913
  var(e.ses2)|   1.573228   .1228219                      1.350014    1.833348
  var(e.ses3)|   1.807189   .0933725                      1.633143    1.999783
  var(e.ses4)|   1.685282   .1047979                      1.491906    1.903724
  var(e.math)|   91.36045   6.594622                       79.3079    105.2447
     var(SES)|   .4912213   .1193158                      .3051572    .7907347
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(5) = 17.69          Prob > chi2 = 0.0034
Satorra-Bentler scaled test:    chi2(5) = 17.80          Prob > chi2 = 0.0032
```

```
. estat gof, stats(all)

─────────────────────┬────────────────────────────────────────────────────────────
Fit statistic        │        Value   Description
─────────────────────┼────────────────────────────────────────────────────────────
Likelihood ratio     │
          chi2_ms(5) │       17.689   model vs. saturated
            p > chi2 │        0.003
         chi2_bs(10) │      150.126   baseline vs. saturated
            p > chi2 │        0.000
                     │
  Satorra-Bentler    │
        chi2sb_ms(5) │       17.804
            p > chi2 │        0.003
       chi2sb_bs(10) │      153.258
            p > chi2 │        0.000
─────────────────────┼────────────────────────────────────────────────────────────
Population error     │
               RMSEA │        0.070   Root mean squared error of approximation
 90% CI, lower bound │        0.037
         upper bound │        0.107
              pclose │        0.147   Probability RMSEA <= 0.05
                     │
  Satorra-Bentler    │
            RMSEA_SB │        0.070   Root mean squared error of approximation
─────────────────────┼────────────────────────────────────────────────────────────
Information criteria │
                 AIC │    11157.441   Akaike´s information criterion
                 BIC │    11221.219   Bayesian information criterion
─────────────────────┼────────────────────────────────────────────────────────────
Baseline comparison  │
                 CFI │        0.909   Comparative fit index
                 TLI │        0.819   Tucker-Lewis index
                     │
  Satorra-Bentler    │
              CFI_SB │        0.911   Comparative fit index
              TLI_SB │        0.821   Tucker-Lewis index
─────────────────────┼────────────────────────────────────────────────────────────
Size of residuals    │
                SRMR │        0.040   Standardized root mean squared residual
                  CD │        0.532   Coefficient of determination
─────────────────────┴────────────────────────────────────────────────────────────
```

The SB-adjusted CFI is still rather low, 0.91, indicating poor fit. We can use **estat mindices** to compute modification indices that can be used to check for paths and covariances that could be added to the model to improve fit. First, we'll need to restore our original model.

```
. estimates restore m1
. estat mindices
Modification indices

─────────────────────────────────────────────────────────────────────────────
                  │                                               Standard
                  │         MI      df    P>MI         EPC             EPC
──────────────────┼──────────────────────────────────────────────────────────
cov(e.ses1,e.ses2)│     16.565       1    0.00    .4818524         .312987
cov(e.ses2,e.ses3)│      5.404       1    0.02   -.2203899       -.1307056
cov(e.ses3,e.ses4)│      4.956       1    0.03    .2033998          .11655
─────────────────────────────────────────────────────────────────────────────
EPC is expected parameter change.
```

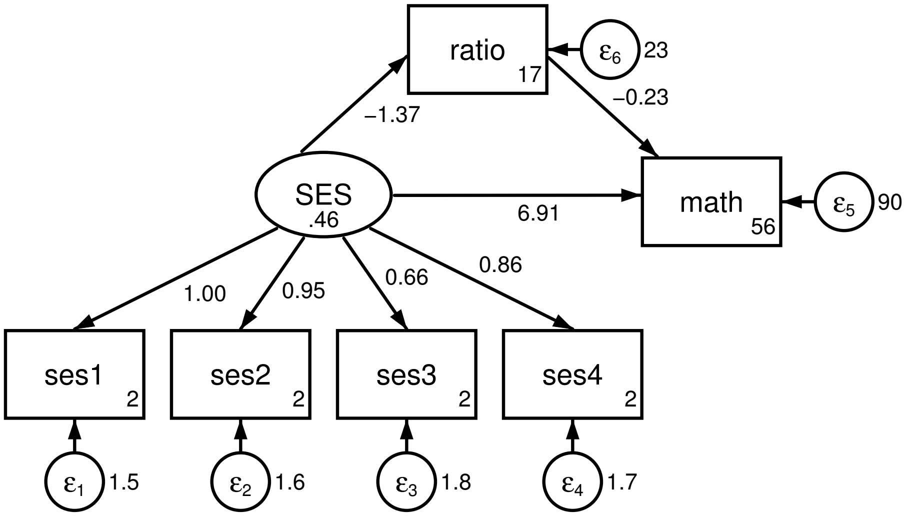
The `MI`, `df`, and `P>MI` are the estimated chi-squared test statistic, degrees of freedom, and p value of the score test testing the statistical significance of the constrained parameter. By default, only parameters that would significantly ($p < 0.05$) improve the model are reported. The `EPC` is the amount that the parameter is expected to change if the constraint is relaxed. According to these results, we see that there is a stronger relationship between the first and second indicator for SES than would be expected given our model, $\texttt{MI} = 16.57, p < 0.001$. We could consider adding a residual covariance between these two indicators to our model using the `cov()` option. We use the `e.` prefix to refer to a residual variance of an endogenous variable; see Figure 5.



```
. sem (SES -> ses1-ses4 math), cov(e.ses1*e.ses2)
```

Figure 5. CFA with residual covariance.

One potential explanation of the effect that SES has on math score is that students of higher SES attend schools with smaller student to teacher ratios. We can test this hypothesis using the mediation model shown in Figure 6. Here, we get estimates of the direct effects between each of our variables, but what we would really like to test is the indirect effect between SES and math through ratio. We can get direct effects, indirect effects, and total effects of mediation models with the postestimation command estat teffects.



```
. sem (SES -> ses1-ses4 ratio math) (ratio -> math)
```

Figure 6. Student-teacher ratio mediates the effect of SES on math score.

```
. estat teffects

Direct effects
-------------------------------------------------------------------------------
             |                 OIM
             | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+-----------------------------------------------------------------
Structural   |
  ratio      |
         SES |  -1.367306   .5562429    -2.46   0.014    -2.457522   -.2770903
-------------+-----------------------------------------------------------------
  math       |
       ratio |  -.2256084   .1026128    -2.20   0.028    -.4267257    -.024491
         SES |   6.908564   1.583778     4.36   0.000     3.804417    10.01271
-------------+-----------------------------------------------------------------
Measurement  |
  ses1       |
         SES |          1  (constrained)
  -----------+-----------------------------------------------------------------
  ses2       |
         SES |   .9450302   .1643867     5.75   0.000     .6228382    1.267222
  -----------+-----------------------------------------------------------------
  ses3       |
         SES |   .6632608   .1725434     3.84   0.000     .3250819     1.00144
  -----------+----------------------------------------------------------------
  ses4       |
         SES |   .8574695   .2012317     4.26   0.000     .4630625    1.251876
------------------------------------------------------------------------------

Indirect effects
------------------------------------------------------------------------------
             |                 OIM
             | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
Structural   |
  ratio      |
         SES |          0  (no path)
  -----------+----------------------------------------------------------------
  math       |
       ratio |          0  (no path)
         SES |   .3084758   .1451257     2.13   0.034     .0240346    .5929169
-------------+----------------------------------------------------------------
Measurement  |
  ses1       |
         SES |          0  (no path)
  -----------+----------------------------------------------------------------
  ses2       |
         SES |          0  (no path)
  -----------+----------------------------------------------------------------
  ses3       |
         SES |          0  (no path)
  -----------+----------------------------------------------------------------
  ses4       |
         SES |          0  (no path)
------------------------------------------------------------------------------
Total effects
------------------------------------------------------------------------------
             |                 OIM
             | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
Structural   |
  ratio      |
         SES |  -1.367306   .5562429    -2.46   0.014    -2.457522   -.2770903
  -----------+----------------------------------------------------------------
  math       |
       ratio |  -.2256084   .1026128    -2.20   0.028    -.4267257    -.024491
         SES |   7.217039   1.599953     4.51   0.000     4.081189    10.35289
-------------+----------------------------------------------------------------
Measurement  |
  ses1       |
         SES |          1  (constrained)
  -----------+----------------------------------------------------------------
  ses2       |
         SES |   .9450302   .1643867     5.75   0.000     .6228382    1.267222
  -----------+----------------------------------------------------------------
  ses3       |
         SES |   .6632608   .1725434     3.84   0.000     .3250819     1.00144
  -----------+----------------------------------------------------------------
  ses4       |
         SES |   .8574695   .2012317     4.26   0.000     .4630625    1.251876
------------------------------------------------------------------------------
```

In the second group of the output, we see that the mediation effect is not statistically significant, $z = 1.48, p = 0.138$. We may consider bootstrapping this effect to get a more powerful test. We can do this with the `bootstrap` command. First, we need to get labels for the effects we would like to test. We can get these by replaying our model results with the `coeflegend` option. We can use these labels to construct an expression for the mediation effect that we're calling `indirect`. We put this expression in parentheses after `bootstrap` and put any bootstrapping options after a comma; then, we put the model and its options after a colon. Multiple expressions can be included using multiple parentheses sets.

```
. sem, coeflegend

Structural equation model                              Number of obs = 519
Estimation method: ml

Log likelihood = -7117.1959

 ( 1)  [ses1]SES = 1
------------------------------------------------------------------------------
             |  Coefficient  Legend
-------------+----------------------------------------------------------------
Structural   |
  ratio      |
         SES |  -1.367306  _b[ratio:SES]
       _cons |   16.75723  _b[ratio:_cons]
  -----------+----------------------------------------------------------------
  math       |
       ratio |  -.2256084  _b[math:ratio]
         SES |   6.908564  _b[math:SES]
       _cons |   55.50311  _b[math:_cons]
-------------+----------------------------------------------------------------
Measurement  |
  ses1       |
         SES |          1  _b[ses1:SES]
       _cons |   1.982659  _b[ses1:_cons]
  -----------+----------------------------------------------------------------
  ses2       |
         SES |   .9450302  _b[ses2:SES]
       _cons |   2.003854  _b[ses2:_cons]
  -----------+----------------------------------------------------------------
  ses3       |
         SES |   .6632608  _b[ses3:SES]
       _cons |   2.003854  _b[ses3:_cons]
  -----------+----------------------------------------------------------------
  ses4       |
         SES |   .8574695  _b[ses4:SES]
       _cons |   2.003854  _b[ses4:_cons]
-------------+----------------------------------------------------------------
  var(e.ses1)|   1.541523  _b[/var(e.ses1)]
  var(e.ses2)|   1.588663  _b[/var(e.ses2)]
  var(e.ses3)|   1.795421  _b[/var(e.ses3)]
  var(e.ses4)|   1.660672  _b[/var(e.ses4)]
 var(e.ratio)|   23.41179  _b[/var(e.ratio)]
  var(e.math)|   89.51067  _b[/var(e.math)]
     var(SES)|   .4562495  _b[/var(SES)]
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(8) = 21.72           Prob > chi2 = 0.0055
```

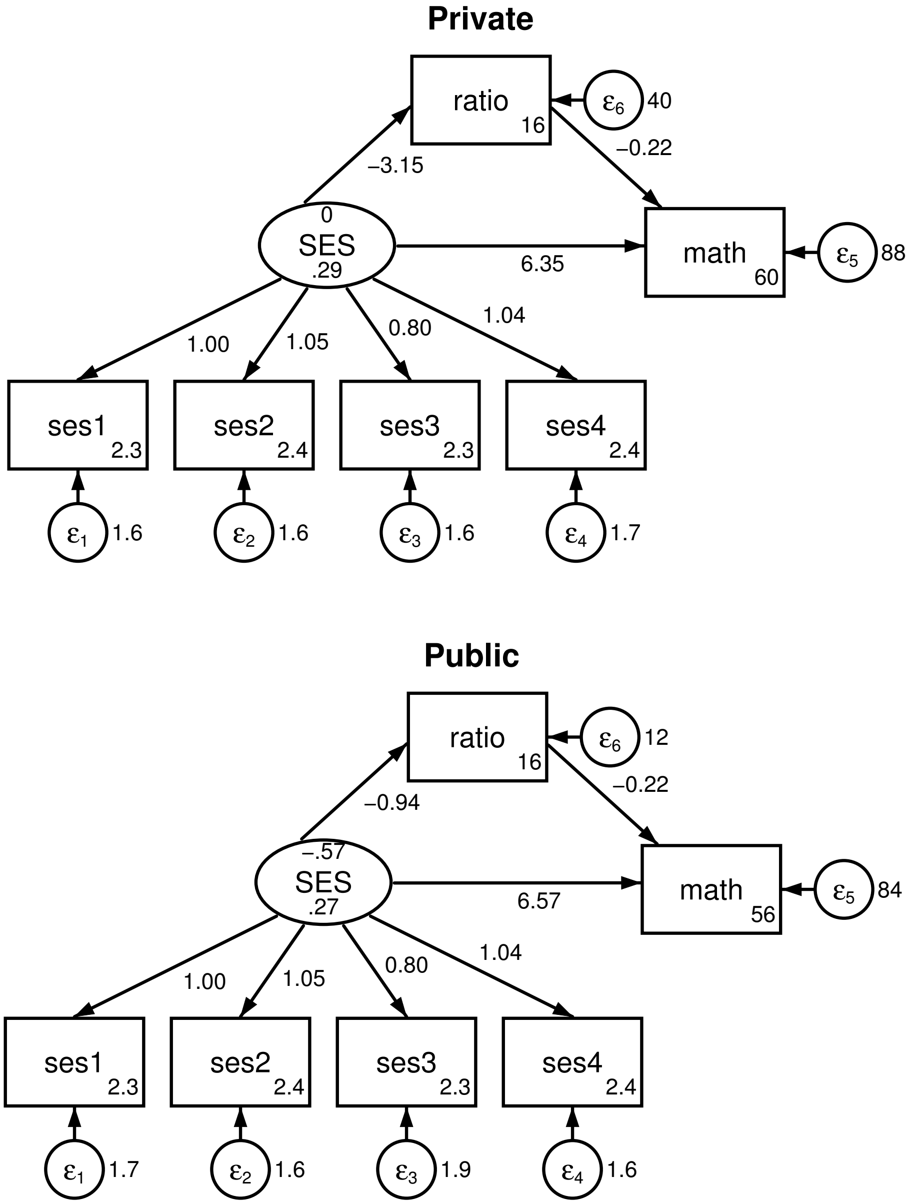
```
. bootstrap indirect=(_b[ratio:SES]*_b[math:ratio]), reps(1000) nodots: ///
> sem (SES -> ses1-ses4 ratio math) (ratio -> math)
Bootstrap results                                       Number of obs =   519
                                                        Replications  = 1,000

      Command: sem (SES -> ses1-ses4 ratio math) (ratio -> math)
     indirect: _b[ratio:SES]*_b[math:ratio]

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |  coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    indirect |   .3084758   .1932632     1.60   0.110     -.070313    .6872646
------------------------------------------------------------------------------
```

We've added the `reps(1000)` option to compute 1,000 bootstrap replications and the `nodots` option to suppress displaying a dot for each replication. To get 95 percentile confidence intervals based on our bootstrap sampling distribution, we can follow with the postestimation command `estat bootstrap` using the `percentile` option. The resulting confidence interval contains zero so we cannot reject the null hypothesis.

```
. estat bootstrap, percentile
Bootstrap results                               Number of obs     =        519
                                                Replications      =       1000

      Command: sem (SES -> ses1-ses4 ratio math) (ratio -> math)
     indirect: _b[ratio:SES]*_b[math:ratio]

------------------------------------------------------------------------------
             |    Observed               Bootstrap
             |  coefficient       Bias    std. err.  [95% conf. interval]
-------------+----------------------------------------------------------------
    indirect |   .30847577  -.0307326   .19326315   -.0707015   .6837121   (P)
------------------------------------------------------------------------------
Key: P: Percentile
```

2.2 Group Analysis

Finally, we may consider comparing our mediation across groups. Group analysis can be done in Stata by adding the `group()` option. We would like to compare students in public schools versus private schools so we will specify `schtype` as our grouping variable. Then, we can use `ginvariant()` to specify the types of parameters we would like to constrain across groups. All other variables will be estimated separately for each group. The `ginvariant()` options are listed in Table 1. If we don't specify any `ginvariant` option, by default Stata will constrain measurement coefficients and measurement intercepts, `ginvariant(mcoef mcons)`. See the model in Figure 7. Now when we run `estat teffects`, we will get a separate estimated mediation effect for each group.



```
. sem (SES -> ses1-ses4 math) (ratio -> math), group(schtype)
```

Figure 7. Group analysis.

```
. estat teffects, nodirect nototal compact

Indirect effects
─────────────┬──────────────────────────────────────────────────────────────────
             │                 OIM
             │ Coefficient  std. err.      z    P>|z|     [95% conf. interval]
─────────────┼──────────────────────────────────────────────────────────────────
Structural   │
  math       │
         SES │
    Private  │   .7043843   .4184641     1.68   0.092    -.1157902    1.524559
     Public  │   .2035724   .1710134     1.19   0.234    -.1316076    .5387525
  ───────────┼──────────────────────────────────────────────────────────────────
  ratio      │
  ───────────┼──────────────────────────────────────────────────────────────────
  ses1       │
  ───────────┼──────────────────────────────────────────────────────────────────
  ses2       │
─────────────┼──────────────────────────────────────────────────────────────────
Measurement  │
  ses3       │
  ───────────┼──────────────────────────────────────────────────────────────────
  ses4       │
─────────────┴──────────────────────────────────────────────────────────────────
```

Table 1. `ginvariant()` suboptions

Option	Description
`mcoef`	measurement coefficients
`mcons`	measurement intercepts
`merrvar`	covariances of measurement errors
`scoef`	structural coefficients
`scons`	structural intercepts
`serrvar`	covariances of structural errors
`smerrcov`	covariances between structural and measurement errors
`meanex`	means of exogenous variables
`covex`	covariances of exogenous variables
`all`	all the above
`none`	none of the above

To test whether these mediation effects significantly differ, we can conduct a Wald test with the `test` or `testnl` postestimation commands, again using the labels from the `coeflegend` option. Because mediation effects are nonlinear, we will use `testnl`. The mediation effects do not significantly differ between groups, $\chi^2(1) = 1.27, p = 0.260$.

```
. testnl _b[ratio:0bn.schtype#c.SES]*_b[math:0bn.schtype#c.ratio]= ///
> _b[ratio:1.schtype#c.SES]*_b[math:1.schtype#c.ratio]
  (1)  _b[ratio:0bn.schtype#c.SES]*_b[math:0bn.schtype#c.ratio]
> _b[ratio:1.schtype#c.SES]*_b[math:1.schtype#c.ratio]

               chi2(1) =        1.27
           Prob > chi2 =        0.2599
```

```
. estat ginvariant
Tests for group invariance of parameters

                         Wald test                     Score test
                  chi2         df    P>chi2       chi2         df    P>chi2

Structural
  math
       ratio     0.001          1    0.9709          .          .         .
         SES     0.005          1    0.9441          .          .         .
       _cons     1.314          1    0.2516          .          .         .

  ratio
         SES     1.825          1    0.1768          .          .         .
       _cons     0.011          1    0.9147          .          .         .

Measurement
  ses1
         SES         .          .         .      1.832          1    0.1759
       _cons         .          .         .      5.997          1    0.0143

  ses2
         SES         .          .         .      0.072          1    0.7882
       _cons         .          .         .      0.341          1    0.5592

  ses3
         SES         .          .         .      0.049          1    0.8253
       _cons         .          .         .      0.634          1    0.4259

  ses4
         SES         .          .         .      1.945          1    0.1632
       _cons         .          .         .      1.149          1    0.2838

  var(e.ses1)    0.189          1    0.6640          .          .         .
  var(e.ses2)    0.063          1    0.8023          .          .         .
  var(e.ses3)    1.011          1    0.3146          .          .         .
  var(e.ses4)    0.090          1    0.7641          .          .         .
  var(e.math)    0.065          1    0.7982          .          .         .
 var(e.ratio)   36.627          1    0.0000          .          .         .
     var(SES)    0.042          1    0.8383          .          .         .
```

To test group differences in each direct path, we can use the postestimation command `estat ginvariant`. These results show us Wald tests evaluating constraining parameters that were allowed to vary across groups and score tests evaluating relaxing constraints. Both are testing whether individual paths significantly differ across groups.

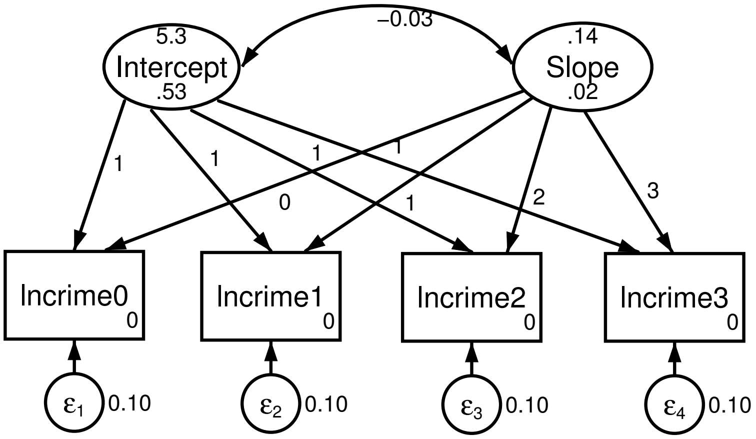
2.3 Growth Curve Modeling

The last model we will fit using `sem` is a growth curve model. This will require a new dataset.

```
. use crime
. describe
Contains data from crime.dta
 Observations:           359
    Variables:             4                  4 Oct 2012 16:22
                                              (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
lncrime0        float   %9.0g                 ln(crime rate) in Jan & Feb
lncrime1        float   %9.0g                 ln(crime rate) in Mar & Apr
lncrime2        float   %9.0g                 ln(crime rate) in May & Jun
lncrime3        float   %9.0g                 ln(crime rate) in Jul & Aug
-------------------------------------------------------------------------------
Sorted by:
```

These data are from Bollen and Curran (2006); they contain crime rates collected in two-month intervals for the first eight months of 1995 for 359 communities in New York state. We would like to fit a linear growth curve to these data to model how crime rate changed over time. In our model, we can set constraints using the @ symbol as we did before. To constrain all intercepts to 0, we can add the nocons option. We will also need the means() option. By default, Stata constrains the means of latent variables to 0. For this model, we would like to estimate them so we need to specify the latent variable names inside means(). We may also consider constraining all the residual variances to equality by constraining each of them to the same arbitrary letter or word, in this case eps. See the model in Figure 8.

The estimated mean log crime rate at the beginning of the study was 5.33 and it increased by an average of 0.14 every two months. We could have fit this same model using gsem. One way we can do this is to simply replace sem with gsem in the command in Figure 8. Alternatively, we can can think of this as a multilevel model, and fit it using gsem's notation for random effects. Let's do that next.



```
. sem (Intercept@1 -> lncrime0-lncrime3) (Slope -> lncrime0@0 lncrime1@1
lncrime2@2 lncrime3@3), nocons means(Intercept Slope)
var(e.lncrime0-lncrime3@eps)
```

Figure 8. Growth curve model on crime rate.

3 Fitting models with the gsem command

3.1 Models with Random Effects

The gsem command implements generalizations to the standard linear structural equation model implemented in sem, such as models with generalized-linear response variables, random effects, and categorical latent variables (latent classes). Its syntax is the same as sem, with some different options and postestimation commands. We will start by fitting a random-slope model to the crimes dataset, reproducing the results we obtained with the growth curve model using sem. First, we need to create an observation identification variable and reshape the data into long format.

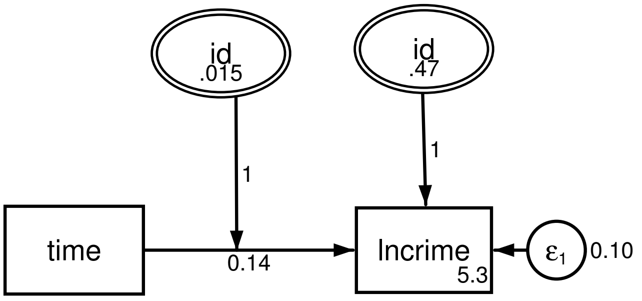
```
. gen id = _n

. reshape long lncrime, i(id) j(time)
(j = 0 1 2 3)

Data                               Wide   ->   Long
-----------------------------------------------------------------------------
Number of observations              359   ->   1,436
Number of variables                   5   ->   3
j variable (4 values)                     ->   time
xij variables:
         lncrime0 lncrime1 ... lncrime3   ->   lncrime
-----------------------------------------------------------------------------
```

```
. summarize
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
          id |      1,436         180    103.6701          1        359
        time |      1,436         1.5    1.118423          0          3
     lncrime |      1,436    5.551958    .7856259   2.415164   9.575166
```

We now have long-format data in which we have several rows of observations for each individual; we're ready to fit our random-slope model. We specify random effects in `gsem` by adding brackets enclosing the clustering variable to the latent variable, i.e. `Intercept[id]`. This tells Stata to include a latent variable in the model called `Intercept` that has variability at the `id` level. As with other latent variables, it will have a mean of 0 and an initial factor loading of 1, so the only parameter this term introduces is a level-2 variance. Random coefficients can be added to any term by interacting a latent random effect with that variable, i.e. `c.time#Slope[id]`.

Interactions in Stata are specified using `#`; interaction terms are assumed to be factor variables unless prefixed by `c.` to indicate that they are continuous variables. Contrarily, main-effect terms are assumed to be continuous unless prefixed by `i.` to indicate that they are factor variables. We'll see this in the next example. This factor variable notation is not available using `sem`.

See the syntax and results of the random slope model in Figure 9; these results replicate those by `sem`. In the SEM Builder, random effects are represented as double-bordered ovals labeled with the clustering variable to indicate that they represent variability at the cluster level.



```
. gsem (Intercept[id] time c.time#Slope[id] -> lncrime)
```

Figure 9. Random-slope model on crime rate.

3.2 Models with Generalized Responses

The gsem command can also be used to fit generalized linear SEMs; that is, SEMs in which an endogenous variable is distributed according to some distribution family and is related to the linear prediction of the model through a link function. See Table 2 for a list of available distribution families and links. Either the family and link can be specified, i.e. family(bernoulli) link(logit), or some combinations have shortcuts that you can specify instead, i.e. logit. For this example, we will return to the first dataset.

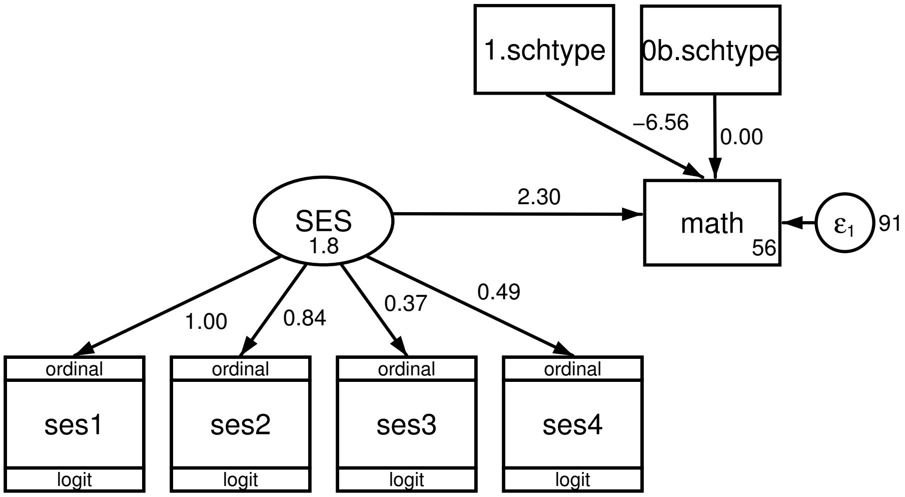
Table 2. gsem distribution families and link functions

family() options	link() options				
	identity	log	logit	probit	cloglog
gaussian	X	X			
bernoulli			logit	probit	cloglog
beta			X	X	X
binomial			X	X	X
ordinal			ologit	oprobit	ocloglog
multinomial			mlogit		
Poisson		poisson			
negative binomial		nbreg			
exponential		exponential			
Weibull		weibull			
gamma		gamma			
loglogistic		loglogistic			
lognormal		lognormal			

Note: X indicates possible combinations. Where applicable, regression names that imply that family/link combination are shown. If no family/link are provided, family(gaussian) link(identity) is assumed.

```
. use math

. codebook, compact

Variable   Obs Unique      Mean  Min  Max  Label
─────────────────────────────────────────────────────────────────────────
schtype    519      2    .61079    0    1  School type
ratio      519     14  16.75723   10   28  Student-Teacher ratio
math       519     42  51.72254   30   71  Math score
ses1       519      5  1.982659    0    4  SES item 1
ses2       519      5  2.003854    0    4  SES item 2
ses3       519      5  2.003854    0    4  SES item 3
ses4       519      5  2.003854    0    4  SES item 4
─────────────────────────────────────────────────────────────────────────
```

In our previous analysis, we had treated each socioeconomic status Likert item as continuous. Now, we will treat them as ordinal using gsem. Adding the ologit option will fit the measurement model using the ordinal family with a logistic link. We will also use factor variable notation to include indicator variables for school type in our analysis. See figure Figure 10. By adding `schtype` as a factor variable, a dummy variable for each level of `schtype` is included in the model. The path coefficient for the base level, by default the lowest, is constrained to zero. To get exponentiated coefficients, we can follow with the postestimation command `estat eform`.



```
. sem (SES -> ses1-ses4, ologit) (SES i.schtype -> math)
```

Figure 10. Ordinal logistic regression model.

```
. estat eform ses1 ses2 ses3 ses4
------------------------------------------------------------------------------
             |     exp(b)   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ses1         |
         SES |   2.718282  (constrained)
-------------+----------------------------------------------------------------
ses2         |
         SES |   2.311549    .483485     4.01   0.000     1.534141    3.482899
-------------+----------------------------------------------------------------
ses3         |
         SES |   1.449492    .180061     2.99   0.003     1.136257    1.849077
-------------+----------------------------------------------------------------
ses4         |
         SES |   1.628133   .2474222     3.21   0.001     1.208748    2.193029
------------------------------------------------------------------------------
```

4 Conclusion

In this tutorial, we've shown the basics of fitting SEMs in Stata using the `sem` and `gsem` commands, and have provided example datasets and syntax online to follow along. We demonstrated confirmatory factor analysis, mediation, group analysis, growth curve modeling, and models with random effects and generalized responses. However, there are many possibilities and options not included in this tutorial, such as latent class analysis models, nonrecursive models, reliability models, mediation models with generalized responses, multivariate random-effects models, and much more. Visit Stata's documentation to see all the available options for these commands, their methods and formulas, and many more examples online at `https://www.stata.com/manuals/sem.pdf`.

References

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol. 467). John Wiley & Sons.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In *Latent variables analysis: Applications for developmental research.* (pp. 399–419). Sage Publications, Inc.

StataCorp. (2021). *Stata statistical software: Release 17.* StataCorp LLC.