5-2021

Applying Bayesian Growth Modeling In Machine Learning For Longitudinal Data

Alisa Udomvisawakul

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

APPLYING BAYESIAN GROWTH MODELING IN MACHINE LEARNING FOR LONGITUDINAL DATA

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

Alisa Udomvisawakul

College of Education and Behavioral Sciences
Department of Applied Statistics and Research Methods

May, 2021

This Dissertation by: Alisa Udomvisawakul

Entitled: *Applying Bayesian Growth Modeling in Machine Learning for Longitudinal Data*

has been approved as meeting the requirement for the Degree of Doctor of Philosophy
in College of Education and Behavioral Sciences in Department of Applied Statistics and
Research Methods.

Accepted by the Doctoral Committee

_____
Susan R. Hutchinson, Ph.D., Research Advisor

_____
Chia-Lin Tsai, Ph.D., Committee Member

_____
Han Yu, Ph.D., Committee Member

_____
James Reardon, Ph.D., Faculty Representative

Date of Dissertation Defense 4/2/2021_____

Accepted by the Graduate School

_____

Jeri-Anne Lyons, Ph.D.
Dean of the Graduate School
Associate Vice President for Research

ABSTRACT

Udomvisawakul, Alisa. *Applying Bayesian Growth Modeling in Machine Learning for Longitudinal Data.* Published Doctor of Philosophy dissertation, University of Northern Colorado, 2021.

There has been increasing interest in the use of Bayesian growth modeling in machine learning environment to answer the questions relating to the patterns of change in trends of social and human behavior in longitudinal data. It is well understood that machine learning works properly with "big data," because large sample sizes offer machines the better opportunity to "learn" the pattern/structure of data from a training data set to predict the performance in an unseen testing data set. Unfortunately, not all researchers have access to large samples and there is a lack of methodological research addressing the utility of using machine learning with longitudinal data based on small sample size. Additionally, there is limited methodological research conducted around moderation effect that priors have on other data conditions. Therefore, the purpose of the current study was to understand: (a) the interactive relationship between priors and sample sizes in longitudinal predictive modeling, (b) the interactive relationship between priors and number of waves of data, and (c) the interactive relationship between priors and the proportion of cases in the two levels of a dichotomous time-invariant predictor for Bayesian growth modeling in a machine learning environment. Monte Carlo simulation was adopted to answer assess the above aspects and data were generated based on

alumni donation data from a university in the mid-Atlantic region where model parameters were set to mimic “real life” data as closely as possible.

Results from the study show that although all main and interaction effects are statistically significant, only main effect of sample size, wave of data, and interaction between waves of data and sample sizes show meaningful effect size. Additionally, given the condition of prior of the study, informative priors did not show any higher prediction accuracy compared to non-informative priors. The reason behind indifferent between choices of informative and non-informative prior associated with model complexity, competition between strong informative and weakly informative prior. This study was one of the first known study to examine Bayesian estimation in the context of machine learning. Results of the current study suggest that capitalizing on the advantages offered jointly by these two modeling approaches shows promise. Although much is still unknown and in need of investigation regarding the conditions under which a combination of Bayesian modeling and machine learning affects prediction accuracy, the current dissertation provides a first step in that direction.

ACKNOWLEDGEMENT

I fully can say that without the love and support from my advisor, committee members, and family, the difficult journey of finishing dissertation would be harder or impossible to achieve. I am so lucky to have the best advisor and dissertation chair that one could have asked for, Dr. Susan Hutchinson. Everything she has done for me, I will always remember and be forever thankful. I can confidently say that she is the best professor I ever had. She truly cares for me, invests countless hours to help me grow professionally. She is always there to help me and guide me throughout my education at UNC since day one till I finished. She holds a very high standard for her students to not just thrive for good but excellent. I have never seen anyone gives thoroughly feedbacks as she does, and I sure learn a lot from them. She always gives her all to help me success. Her sincere dedication pushes me to work harder to meet up with her standard and makes sure that her time and effort do not go to waste. Dr. Hutchinson is overall an amazing lady, who I feel so privilege to have her in my life and have her as my advisor – you made a great impact to my life more than you know.

I also appreciated every one of my family members for the support. Particularly, Adam, who always be there in every step of my dissertation. Without you, I do not think that my dissertation would come together this nicely. You are always be my strength to help me carry on when I think things are impossible. Also, thank you, Narisara, who always supports, cheers me up and hears me vent --every time I hear your voice, it is brightened up my day and give me strength to fight forward. Additionally, thank you, Jenifer, who always check on me and believe in me-- your love and support mean so much and I deeply appreciated it. I feel so lucky!

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

INTRODUCTION

Answering questions relating to the patterns of change in trends of social and human behavior (i.e., examining developmental change, assessing long-term treatment effects, exploring market trends and brand awareness) has always been a topic of interest across research fields. And, in order to answer and gain a better understanding of those questions, longitudinal data are often used (McArdle, 1988; Zhang et al., 2007). Popular questions relating to longitudinal data include, but are not limited to: What is the average rate of change (trajectory) across time? How much do individual trajectories differ from one another? Can one predict the difference between individuals and their cohort's trajectories based on individual characteristics? (Curran et al., 2010; Meredith & Tisak, 1990). The insights from the above questions give researchers abilities to explore change in time-related patterns both within-person and across-persons (McArdle & Epstein, 1987).

Given that we are now living in the era where data are fast growing in volume, variety, and velocity due to the nonstop development in technology, Internet power, communication devices, and social media, the process of collecting and accessing longitudinal data has the potential to be easier (Kasun et al., 2013). When a substantial amount of data has been generated, operated, and stored mainly online, the term "big data" is used to describe the data (Raschka & Mirjalili, 2017). The sources of big data can be Internet search engines, social network interactions, comments posted on websites, applications on mobile phones, online transactions,

emails, videos, audio, images, health records, and science data (Sagiroglu & Sinanc, 2013). Although collecting longitudinal data for analysis existed before big data emerged (Cronbach & Furby, 1970; Rogosa et al., 1982), the ability to access the data collection where data from the same individuals are repeatedly recorded for an ongoing period of time is simpler from the help of big data (Konerman et al., 2015). Thus, several research fields (i.e., business, healthcare, sport, education, scientific research, politics, and law enforcement) take advantage of longitudinal and big data to develop research projects to analyze and gain better understanding of the data in their fields, as well as discover the hidden connections/patterns within the data and predict the upcoming trend (Bates et al., 2014; De Rosa & Aragona, 2017; Hatch, 2018; Mullainathan & Spiess, 2017). For example, in business, there are over 24 million active online shopping websites around the world, from which researchers use both longitudinal and big data to gain richer and deeper understanding of their customers' preferences and purchase behaviors. The data that companies gain from both longitudinal and big data help the companies to gain knowledge and stay ahead of their competition (Hatch, 2018). YouTube has over 1.9 billion logged-in users visit their website each month and people watch over a billion hours of video and generate billions of views per day. Thus, YouTube predicts the recommended videos and pop-up commercials based on their viewers' past viewing histories (Tufekci, 2018). In health care and public health, the power of longitudinal and big data analytics is used to understand and predict disease patterns in order to develop new cures (Bates et al., 2014). In education, educators use longitudinal and big data to improve students' experience, reach out to a wider range of students, customize educational programs, and reduce dropout rate (De Rosa & Aragona, 2017).

Approaches to Analysis of Longitudinal Data

There are various statistical methods that are used to analyze longitudinal data including repeated measures analysis of variance (ANOVA), multilevel modeling, latent growth curve analysis (based on structural equation modeling), piecewise linear growth models, and Bayesian growth curve models (Demidenko, 2004; Laird & Ware, 1982).

Repeated Measures Analysis of Variance

Repeated measures ANOVA is an extension of analysis of variance (ANOVA) that accounts for correlation between the repeated factors (Maxwell & Delaney, 1990). Generally, repeated measures ANOVA is used to answer questions about (a) changes in average scores on three or more time points, and (b) differences in average scores on two or more levels of the between-groups factor or testing conditions and the interactions effects (Von Ende, 2001). The repeated measures ANOVA allows users to assess the test over time, requires smaller sample size, and has good statistical power (Howitt & Cramer, 2011). However, its main drawbacks are that (a) all subjects are required to be measured on the dependent variable at the same time points and with the same number of time points and (b) the rate of growth is assumed to be the same across individuals and the starting point is the same for everyone. These shortcomings suggest that measurement times must be fixed for all participants and that repeated measures ANOVA treats intercept and slope variations as nuisance even though these values might differ across individuals (Field, 2011; Howitt & Cramer, 2011). With these shortcomings of repeated measure ANOVA, additional statistical models have been developed to attempt to address its limitations.

Hierarchical Linear Growth Modeling

Since longitudinal data have a nested structure (times nested within individuals), hierarchical linear growth modeling (HLGM), also known as multilevel modeling, has been

adopted to analyze growth across time (Bryk & Raudenbush, 1987; Muthén, 1997; Schonfeld, & Rindskopf, 2007; Singer, 1998). Unlike repeated measure ANOVA, HLGM does not require the same number of observations at each time point nor that data are collected at the same times for individuals. HLGM also allows users to assess individuals' trajectories and starting points (intercepts) and factors that might affect those trajectories and starting points. Moreover, it performs well with a large number of repeated measures and has the ability to be expanded to higher levels of nesting (Jackson, 2010; Lininger et al., 2015; Schonfeld & Rindskopf, 2007). Besides the great advantages listed above for HLGM, its complexity increases when the model is expanded to three levels (i.e., repeated measures at level-1, nested within students at level-2, and nested within schools at level-3). HLGM requires large sample sizes at level-3 (at least 30 data points for the level 3 units) to be sufficient to estimate regression coefficients and obtain unbiased estimates of variance components (Amatya & Bhaumik, 2018; Hox, 2002; Mok, 1995). Additionally, HLGM only allows data to be missing at level-1 and level-2 (time and individual level). For instance, if the performance data at one school (level-3) are missing, that particular school's data need to be entirely removed. Moreover, HLGM can get very complicated and the results hard to interpret when including multiple variables and complex growth trajectory shapes (Gentry & Martineau, 2010; Woltman et al., 2012).

Latent Growth Curve Modeling

Another well-known method for analyzing longitudinal data is latent growth curve modeling (LGM) within the structural equation modeling (SEM) framework. LGM is closely related to HLGM conceptually (Duncan & NetLibrary, 1999; Vasantha & Venkatesan, 2014; Willett & Bub, 2014). However, instead of treating the repeated measure as a nested structure within individuals, as is done in HLGM, LGM treats repeated measures as observed indicators

that define the underlying latent factors, and those latent factors represent the growth trajectory and intercept. In LGM, repeated measures are used as observed indicators to identify the structure of the latent variable; this way, repeated measurement is modeled by the latent intercept and slope variables of the latent growth curve (Duncan & NetLibrary, 1999). The strengths of LGM are similar to HLGM in the way that it does not require individuals to be measured at the same occasions or time intervals. LGM also treats the trajectories as a random effect; thus, the persons' differences in both intercept and slope can be measured. Additionally, LGM is a special case of SEM, so all the advantages of SEM also apply to LGM, which include the ability to assess the wide range of model fit (i.e., chi-square, root mean square error of approximation [RMSEA], comparative fit index [CFI]), the ability to examine change in latent variables, and the ability to examine the backgrounds and pathological conditions of change (Curran, 2003; Preacher, 2010). Some downsides for LGM include (a) complication of constructing multivariate structured data, (b) challenges when there is a large number of repeated measures, because each repeated measure becomes an indicator on the latent factor, (c) difficulty of testing interaction effects, and (d) requirement of large sample size: around 400, depending on the number of indicators (Curran, 2003; Tomarken & Waller, 2005).

Piecewise Growth Curve Modeling

When there exist nonlinear trajectories where the pace of change is faster in some periods than in others, piecewise growth curve modeling (PGCM) is usually adopted for the analysis (Chou et al., 2004; Cudeck & Harring, 2007; Flora, 2008; Grimm et al., 2011). Since, PGCM specially looks at patterns of nonlinear growth, it can be applied in combination with hierarchical linear modeling (HLM) growth modeling, structural equation modeling, and Bayesian modeling (Bollen & Curran, 2006; Flora, 2008; Raudenbush & Bryk, 2002). Instead of

looking at a single growth trajectory for the complete window of repeated measures of time, PGCM divides the nonlinear trajectory into separate series of linear trajectories or segments of different slopes, which are connected by turning points (or change points, also called a knot) at specific inflection points. If the change point is unknown, Bayesian inference can be used to estimate the growth (Chou et al., 2004). Since PGCM disintegrates the nonlinear trajectories into separate sections for the examination, it allows researchers to compare growth rates at different time periods (Kim et al., 2015). However, a main challenge of using PGCM is the specification of the turning point. For example, we usually specify the change point in the data based on theory or design (i.e., the starting point of the intervention); however, the changing point might happen after the intervention as a result of postponement in response to intervention. Consequently, misspecifying the changing point can lead to improper interpretation of growth traits (Chou et al., 2004).

Bayesian Growth Curve Modeling

As deliberated in the section above regarding the option of statistical models that can be used to analyze longitudinal data, each method has its own strengths and weaknesses. Moreover, structural equation modeling and multilevel modeling usually use maximum likelihood as a method for estimating growth curves. Therefore, it is meaningful to point out some requirements regarding the use of maximum likelihood. Maximum likelihood estimation (MLE) requires large sample size, as it is well understood that larger sample size can lead to better statistical power, model accuracy, reliability, effect size, and generalization (Greenland et al., 2000; Raudys & Jain, 1991). However, some research fields (i.e., biometrics, clinical study, environmental factors) do not have a luxury of accessing large sample size due to the cost and complexity of data collection or the infrequency of the situation of interest (Gagné & Hancock, 2002; Muthén

& Muthén, 2002). Additionally, most data in real life have a high tendency to be non-normal rather than normal (Muthén & Muthén, 2002). The problems associated with violation of the normality assumption are inaccurate estimation of parameters, standard errors, and confidence intervals, and unreliable statistical tests and inference. Moreover, variance components are biased in MLE, if the normality assumption is violated (Muthén & Asparouhov, 2012). Consequently, Bayesian inference methods are adopted to address the limitations about large sample sizes in maximum likelihood estimation. Bayesian modeling is not constructed on asymptotic concepts (a property that can be an interference) like maximum likelihood, so the restrictive conditions that are required in maximum likelihood do not apply in Bayesian inference (Wolfinger & Kass, 2000). Additionally, adoption of Bayesian inference has increased as a tool to estimate growth curve models, particularly in models that are complex and complicated to analyze using maximum likelihood estimation (Muthén & Asparouhov, 2012; Song & Lee., 2012; Wang & Preacher, 2015; Zhang et al., 2007). Moreover, there is evidence supporting that fitting growth curves within the Bayesian framework offers a lot of flexibility for estimating models with various levels of complexity (Curran et al., 2010; Dunson, 2000; Gelman, Carlin, et al., 2015; Grimm & Ram, 2009; Oravecz & Muthén, 2018).

Bayesian growth modeling offers several advantages over other estimation procedures for analyzing trajectories in longitudinal data. First Bayesian modeling requires less restrictive data requirements (i.e., the non-normally distributed repeated measures; Curran et al., 2010; Grimm & Ram, 2009). Second, Bayesian modeling offers an ability to incorporate prior information, which provides a natural and principled way to incorporate past information about a parameter. In turn, the uncertainty surrounding each parameter is taken into consideration (Kim et al., 2020).

Third, Bayesian inference is easily drawn based on the posterior distribution of model parameters, even if data have been transformed. Moreover, the result from the posterior distribution is easy to interpret in an intuitive way by describing it as a likely range of the parameter. For example, the true parameter of the data has a probability of .95 of falling within a 95% credible interval (Gelman, Carlin, et al., 2015). Fourth, Bayesian modeling can estimate complex models and data structures via the Markov chain Monte Carlo (MCMC) simulation-based algorithm, which calculates multiple data quantities, once samples from the posterior are retrieved from MCMC (Dunson, 2000). Fifth, Bayesian modeling works better with small sample size compared to MLE because Bayesian modeling is not based on asymptotic concepts and does not require restrictive estimations for posterior distributions (Berger, 1990; Hoogland & Boomsma, 1998; Schafer, 1997;). In addition, with small samples in growth data with specified priors, Bayesian modeling shows smaller prediction error (i.e., mean square error) compared to modeling based on MLE (Lee & Chang, 2000; Lee & Liu, 2000).

With all the above being said, Bayesian growth curve modeling was the main focus of the analysis for the current dissertation study; the full details of Bayesian growth curve modeling are discussed in Chapter 2. Although growth trajectories are not limited to straight-line functional growth (linear) and Bayesian inference can be used to analyze non-linear trajectories, the current dissertation only focused on using Bayesian inference to fit linear growth curve models (for which the function of change is linear).

Priors in Bayesian Inference. As mentioned above, one of the major benefits for Bayesian growth modeling is the ability of including prior knowledge into the model, which can help to increase the precision of the posterior distribution. Dunson (2001) stated that “The use of prior probability distributions represents a powerful mechanism for incorporating information

from previous studies and for controlling confounding" (p.1222). The additional knowledge in a prior distribution can also help to increase statistical power when the sample size is small and serves as mutual knowledge in the field of study. Priors are incorporated in the model as a part of posterior distribution calculation. In return, regardless of sample size, the model can be construed as a distribution presenting the probability of parameter values (Depaoli, Rus, et al., 2017; Zondervan-Zwijnenburg et al., 2017).

In longitudinal data, known information about the growth trajectory can be incorporated into the model to reveal knowledge about the cause or pattern of the growth, which in turn, can be useful for coming up with the sources for predicting the trajectory. Moreover, it is important to make the best use of prior knowledge to attain the parameters' posterior distribution (Gill, 2015). Several studies support that appropriately incorporating a prior distribution into Bayesian growth models can help to determine the ideal growth trajectory and improve model estimation accuracy (Depaoli & Boyajian, 2014; Depaoli, Rus, et al., 2017; Walls & Quigley, 2001).

There are four main types of priors which are non-informative, subjective, conjugate, and informative. A non-informative prior is the prior equation in a model that contain little explanatory information about the unknown parameters or hypotheses related to the model (Golchi, 2018). Non-informative priors are usually used when there is no dependable prior information about the hypotheses or model parameters, or an implication relying solely on the information at hand is preferable (Depaoli, Rus, et al., 2017). A subjective prior is the prior that informs opinion of the value of a parameter prior to the data collection. Subjective priors should be used when we do not have much information about model parameters but we might have an instinctive idea about the minimum, maximum, mean, and most probable value of the parameter (Choy et al., 2009; Hosack et al., 2017). A conjugate prior is the situation where both prior and

posterior distributions come from similar probability density functions and the posterior distribution has closed form. In this situation, matching the prior to the posterior is referred to as a conjugate prior for the likelihood function (conjugate to the likelihood). Some examples for conjugate pairs are Poisson-Gamm and Binomial-Beta. The conjugate prior is suitable to use when simplifying the model equations is appropriate (Alhamzawi & Yu, 2013). The informative prior is the prior that consists of existing information, knowledge, literature review, or a hypothesized parameter distribution associated with the model before the data are collected (Bolstad, 2007). When an informative prior is available, it should be adopted into the model over a non-informative prior, especially in longitudinal data (Depaoli, 2014; McArdle & Horn, 2004; Muthén & Asparouhov, 2012; Shi & Tong, 2017; Wolf, 1986). The main reason that informative priors should be used over non-informative priors is because non-informative priors show poor performance in parameter recovery and large bias in the posterior distribution (Richardson & Green, 1997; Roeder & Wasserman, 1997). In addition, further information about the model can be viewed through informative priors. Thus, by not employing an informative prior when it is available, important information can be wasted (Bolstad, 2007).

Since the prior plays an important role in Bayesian growth modeling, the choice of prior distribution has an important impact on the posterior distribution. This is because the posterior is the product of the multiplication of prior and likelihood functions and normalizing by integration over the parameter variables (i.e., if the prior that we chose is normally distributed, in turn, the posterior will become normal). However, there is no clear-cut method for how to choose a prior (Congdon, 2014). Furthermore, different choices of priors result in difference posterior inference; thus, the process of choosing priors can be confusing and frustrating, especially in the small sample research (Zondervan-Zwijnenburg et al., 2017). One well-known method used to

explore the effect of using different priors on posterior inference is through sensitivity analysis, which is the study of how different sources of the model input will result in the uncertainty of the model output (Gelman, Carlin, et al., 2015; Paruggia, 2006).

Machine Learning

As previously discussed, it is advantageous to be able to access longitudinal and large volume of data. However, when the data are collected at a faster pace than before, a major problem associated with collecting massive datasets is that data tend to have complex structures (i.e., nested, multidimensional structure), which can be problematic in terms of storing, analyzing, and visualizing for further analysis (Lum et al., 2013; Pedersen & Jensen, 1999). Thus, artificial intelligence (AI) was developed to help people to efficiently manage and analyze the fast-growing pace of data acquisition. Artificial intelligence is the replication of human intelligence through machines, especially computer systems, to learn the details of data and rules for using the data, apply methodologies to reach approximations or define conclusions from data, correct mistakes that arise during the data analysis process, and make decisions or accurate predictions in some problem domain (Nilsson, 1986). The following are examples of current use of AI: filtering incoming email and flagging spam emails (Raschka & Mirjalili, 2017) and personalization, with which online services like Amazon, Ebay, and Netflix learn their customers' purchase history and then develop new, related product recommendations to their customers (Chung, 2016). Fraud detection is another application of AI, in which credit card companies use AI to learn consumers' purchasing habits and can detect if there are unusual transactions (Raschka & Mirjalili, 2017).

Since AI is mainly operated through machines, often computers, machine learning (ML) is frequently mentioned when we talk about AI. Machine learning is a subset of AI that allows

machines to use computer algorithms and statistical techniques to learn and act like humans by feeding data and information into them without being explicitly programmed; then, these algorithms turn data into knowledge (Raschka & Mirjalili, 2017). For example, when we feed the data into the computer, the computer analyzes the data and eventually gets trained on that set of data and learns the data. When new data come in, the computer accurately makes decisions and predictions based on the past data. The main purpose of machine learning is to train the machines based on the provided data and algorithms. Consequently, the machine can learn how to make decisions according to the information that it processed. The machine learning algorithm tends to minimize the error and maximize the correct results/prediction. Additional to making the decision based on learned data, machine learning can modify the results (i.e., prediction, decision) when more data are fed into the machine (Jakhar & Kaur, 2019).

Types of Machine Leaning

Machine learning (ML) can be grouped into three categories which are supervised, unsupervised, and reinforcement learning. Supervised ML also can be categorized further into two main subsets which are classification and regression, while unsupervised ML can be divided further into clustering and association (Raschka & Mirjalili, 2017). Understanding the different types of ML is important because it helps researchers view the broader picture of AI and enables them to choose the ML algorithm that suits the business/research questions (Domingos, 2012). The discussion of the different type of ML is presented in greater detail in Chapter 2.

Problem Statement

Given the situation where answering questions regarding change over time has always been a topic of interest across research fields (McArdle, 1988; Zhang et al., 2007), and the increasing popularity of machine learning and Bayesian inference in longitudinal data (Chen, et

al., 2012; Cui & Gong, 2018; Walsh, et al., 2018), several applied researchers have adopted Bayesian growth modeling (also known as hierarchical Bayesian modeling) in a machine learning environment to help them answer research questions and make decisions about predictive modeling, which the majority of the researchers did with big data (Parslow et al., 2013; Schrodt et al., 2015; Wang & Preacher, 2015). This is because it is well understood that machine learning works properly with "big data," because large sample sizes offer machines the better opportunity to "learn" the pattern/structure of data from a training data set to predict the performance in an unseen testing data set (Wang & Gelfand, 2002).

Unfortunately, not all researchers have access to large samples. For example, in the medical field, researchers might be interested in applying ML to predict the risk of a rare disease. In the educational field, researchers might want to use machine learning to predict the graduation rate of Ph.D. students. Or researchers in environmental fields may want to adopt ML to predict behavior of endangered animal species. Hence, researchers who want to apply ML in their analyses, might wonder if ML will perform equally well with smaller sample size or how large a data set is "enough" for the training set in ML to produce a well performing model, particularly in longitudinal data. Moreover, there is a lack of methodological research addressing the utility of using ML with longitudinal data based on small sample size.

As discussed previously, Bayesian modeling is well-known in its ability to perform well with smaller sample size and allow researchers to incorporate prior knowledge into the model (Berger, 1990; Hoogland & Boomsma, 1998; Schafer, 1997; Scheines et al., 1999). However, one obstacle of applying Bayesian growth modeling to predictive models is how to choose the right priors. There is evidence in various studies that the choice of priors has an influential impact on the posterior distribution (Depaoli & Bovaiian, 2014; Depaoli, Rus, et al., 2017; Walls

& Quigley, 2001) and informative priors should be used over non-informative priors (Bolstad, 2007; Richardson & Green, 1997; Roeder & Wasserman, 1997; Zhang et al., 2012). Furthermore, there are few studies conducted around priors in Bayesian growth modeling. For example, Depaoli and Bovaiian (2014) studied the impact of inaccurate informative priors on Bayesian growth mixture models. Shi and Tong (2017) conducted a simulation study to evaluate the impact of different priors (including non-informative and informative priors with different level of accuracy and precision) and level of mis-specified models on parameter recovery and model estimation in a Bayesian latent growth model. Shi and Tong's study suggested that model misspecification has a much greater negative effect on parameter estimation than inaccurate priors. Additionally, the study from Zondervan-Zwijnenburg et al. (2017) presented guidelines on how to conduct Bayesian analysis with informative priors concerning a latent growth model.

Despite the potential advantages of Bayesian modeling and machine learning with longitudinal data, there is limited methodological research that has been conducted around the following aspects of Bayesian growth modeling, especially in a machine learning environment: (a) the potential moderator effect that priors have on sample sizes, (b) the moderator effect that priors have on number of waves of data, or (c) the moderator effect that priors have on proportion of dichotomous time-invariant predictors. The above aspects are important limitations of the existing research because assessing the potential moderator effect of priors on other model conditions, would allow us to determine whether the effect of priors on the prediction accuracy is different at different values of sample size, waves of data, and proportion of dichotomous time-invariant predictors (Cooper & Lanza, 2014).

Purpose of the Study

With all being said above, the purpose of the current study was to understand: (a) the interactive relationship between priors and sample sizes in longitudinal predictive modeling, (b) the interactive relationship between priors and number of waves of data, and (c) the interactive relationship between priors and the proportion of cases in the two levels of a dichotomous time-invariant predictor for Bayesian growth modeling in a machine learning environment. As mentioned above, correctly specifying priors can lead to higher prediction accuracy (Depaoli, Rus, et al., 2017; Zondervan-Zwijnenburg et al., 2017). Moreover, understanding how priors play a role in interacting with sample size, number of waves of data, and proportion of a dichotomous time-covarying predictor can help applied researchers to make decisions regarding what underlying conditions should be used when adopting Bayesian growth modeling in machine learning to yield acceptable model accuracy (Wang & Preacher, 2015).

Since this dissertation study was used to understand how prediction accuracy varies under the selected conditions noted above, Monte Carlo simulation was adopted to answer the research questions. A Monte Carlo simulation was chosen because it is suitable when the purpose of the research is to study theoretical outcomes of statistical properties (i.e., prediction accuracy, parameter estimate bias, standard errors) under different conditions from randomly generated and experimentally manipulated data that are not easily examined through "real data" (Graham & Talay, 2013).

The data used in this study were generated based on alumni donation data from a university in the mid-Atlantic region where model parameters were set to mimic "real life" data as closely as possible. The actual data will helped to guide selection of the possible values in each wave of data, information about the growth trajectory, ratio of cases in the two levels of a

dichotomous time-invariant predictor, correlations among variables, and prior distributions. Moreover, synthetic data allow researchers to know the correct values of the parameters and check whether those parameters can be recovered with the hypothesized models under varying data conditions (Martin, 2018).

Research Questions

The following are the research questions that were derived to investigate the effects of various data conditions on prediction accuracy.

Q1 Do the types of prior (informative and noninformative priors) moderate the effect of sample size on predictive accuracy for Bayesian growth modeling in a machine learning environment?

Q2 Do the types of prior (informative and noninformative priors) moderate the effect of number of waves of data on prediction accuracy for Bayesian growth modeling in a machine learning environment?

Q3 Do the types of prior (informative and noninformative priors) moderate the effect of proportion of cases in the two levels of a dichotomous time-invariant predictor on prediction accuracy for Bayesian growth modeling in a machine learning environment?

Limitations

As mentioned, the above research questions can be answered through Monte Carlo simulation; however, as would be true of any simulation study, my study did not address all possible conditions that might affect the outcomes of the predictive modeling procedure. Although, through this dissertation study I attempted to select study conditions that reflect situations commonly encountered by applied researchers, generalizability of this study may be limited by the variables that were held constant (i.e., linearity, level of non-normality) as well as by the specific levels of variables that were manipulated.

Chapter Summary

Chapter 1 included information related to the background of statistical procedures that can be used to analyze longitudinal data and the benefit of using Bayesian growth modeling to answer research/business questions in longitudinal data. Moreover, the concept of using Bayesian growth modeling in machine learning, and the brief concept of how to use a Monte Carlo study to assess the effect of prior sample size, waves of data, and proportion of dichotomous time-covarying predictor was introduced.

The following Chapter 2 presents a comprehensive review of the literature supporting the need and purpose for the proposed study, chapter 3 provides a detailed description of the proposed methods that used in this study, chapter 4 represents finding of the simulation results, and chapter 5 gives discussion of the findings.

CHAPTER II

REVIEW OF LITERATURE

Longitudinal data analysis has played an essential role in empirical research in the last few decades. The use of longitudinal data includes but is not limited to examining developmental change, assessing long-term treatment effects, exploring market trends and brand awareness, measuring employee engagement, and analyzing diary data (McArdle, 1988; Zhang et al., 2007). Longitudinal analyses allow researchers to explore the change in time-related patterns for both intra-individual and inter-individual data (McArdle & Epstein, 1987).

The ability to observe change over time makes longitudinal analysis stand out and there has been continuous research on developing analytic methods for longitudinal data. There is a broad range of statistical analysis procedures that are used to examine longitudinal data including repeated measures analysis of variance, multilevel modeling, latent growth curve analysis, piecewise linear growth models, latent basis growth models, and Bayesian growth curve models (Demidenko, 2004; Laird & Ware, 1982). Each of these types of longitudinal analysis and the methodological research associated with them is discussed in greater detail in this chapter. In addition, the details regarding priors when using Bayesian modeling, machine learning for longitudinal data, types of machine learning, and sample size requirements for predictive modeling are also addressed in this chapter.

Longitudinal Statistical Models

Growth Curve Model

One of the well-known analytical techniques that can be employed, within the above listed statistical frameworks, to investigate change over time (i.e., developmental changes, assessing treatment effects) is called growth curve modeling (GCM). Growth curve modeling allows us to take information that we observed (repeatedly measured) and make an inference about the existence of something that we believe to exist but did not directly observe, e.g., we did not observe the growth curve but we observed the individual measures around the growth curve (Curran et al., 2010). For example, we observed the change in students' reading scores from grades 3-5, and we try to account for the factors that influence the change in students' reading scores (e.g., gender, family background); however, these are not our primary factors that we directly observe. Sometimes growth curves can be viewed as latent growth curves, which are latent in the sense that we believe there are factors that affect the change in the growth curve, but we do not directly observe those factors (McArdle & Epstein, 1987).

The growth trajectory can be specified to be two main types which are linear and nonlinear. The linear trajectory is the situation when the measurement scores relating to time that define the trajectories usually increase evenly for equally spaced repeated measures (e.g., times are set to 0, 1, 2, 3 years). However, the space of repeated measures can be adjusted to allow for unequally spaced time measurements, but the slope of the change always refers to an equal change in the outcome per-unit change in time (Biesanz et al., 2004). In contrast, the growth in nonlinear trajectory has inconsistent change with respect to time with faster change in some stages than others (Cudeck & Harring, 2007). For example, the change might be greater at the earlier time intervals and then the rate of change becomes slower (e.g., rapid weight loss at the

beginning of a weight loss program followed by less weight loss later in the program), or the change starts slowly and accelerates with time (e.g., substance use in adolescence), or the change can slowly start, rapidly change, then slowly decrease (e.g., heavy alcohol drinking in young adults). In order to apply GCM techniques to evaluate changes occurring over time, the data structure has to be based on repeated measures where the same objects are being observed more than once over a period of time (McArdle & Nesselroade, 2003). Consequently, GCM estimates trajectories that are unique to each individual based on the set of repeated measures where the set of individual-specific trajectories becomes the unit of analysis. GCM can be used to assess both within-person effects (i.e., change over time for one person) and between-person effects (i.e., compare change over time across people; Curran et al., 2010). The ability of simultaneously analyzing both within- and across-individual changes has helped GCM to increase its popularity. Moreover, researchers in various fields have developed GCM to serve complex aspects of longitudinal data. For example, GCM has developed from being only able to fit a single trajectory curve for one individual to fitting multiple-level trajectory curves for multiple individuals (mixed-effects models) and has expanded from fitting only linear models to fitting nonlinear models (McArdle, 1988; McArdle & Nesselroade, 2003; Meredith & Tisak, 1990).

In applied research, GCM has been used to assess a wide range of complex longitudinal topics. To mention a few, Ferrer and McArdle (2004) verified new hypotheses about how cognitive ability from childhood to early adulthood impacts academic achievement. DeLucia and Pitts (2006) provided an introduction to apply GCM to analyze longitudinal pediatric psychology data. Brooks and Meltzoff (2008) used GCM to predict vocabulary growth in infants. Eggert et al. (2011) employed a multiple-group latent growth curve analysis to understand the causality between service infusion strategies and manufacturers' profit trajectories. Liu et al. (2014)

applied a GCM approach to investigate the dissemination of mobile digital content using a 149-week period of data from 31 regions in China. Kim et al. (2015) assessed the association among body mass index, physical activity, and healthy diet through GCM. Harris et al. (2018) applied logistic growth curve models on US Energy Information Agency data from 1949 to 2015 to predict US energy production and consumption up to 2040. MacAulay et al. (2018) developed a latent growth curve model of cognitive functioning by utilizing longitudinal data from the National Alzheimer's Disease Coordinating Center to investigate the relationship among executive attention/processing speed, episodic memory, language, and working memory functioning utilizing the neuropsychological test battery.

In growth modeling, the combination of fixed and random effects is used to explain the characteristics of the over-time growth both for individuals and across groups. In general terms, a fixed effect refers to a single value that occurs in the population (i.e., the population mean of women's height), and random effect refers to the random probability distribution around the fixed effect (i.e., the population variance in women's height). Fixed and random effects in growth modeling are interpreted along the same line with the previous general definitions. The fixed effects refer to the mean of the growth combined from all individuals within the sample (i.e., mean trajectory indicating the average growth in women's height from age 13 to 19), and random effect refers to the variation of the individual growth around the group's sample mean (i.e., the variation in individual women's change in height from the mean change in height for the overall sample).

To put growth modeling in a practical example, a linear growth model for assessing students' reading improvement from grades 3-5 is used for illustration. In this particular case, the fixed effects are the average students' reading scores at the measurement starting point

(intercept), which is grade 3, along with the pooled average reading score change from grades 3 to 5 from all students in the sample. The random effects are estimates of between-student variability in reading scores around the intercept and slope. Smaller variation of students' reading scores in slopes (and therefore, smaller random effects) means that the trajectories in reading scores across time are more similar across all students. In the extreme situation where all students have the same trajectory of reading scores, the random effect will equal zero, whereas the low degree of variability in the intercept would indicate that the students' initial reading scores are similar. On the other hand, if there is larger variation in the slope associated with growth in students' reading scores (larger random effect), it implies there are larger differences in the magnitude of the growth in individual students' reading scores around the average rate of the growth in all students' reading scores. In other words, large slope variance reflects differences between students in how quickly (or slowly) they improve in reading performance from the third to fifth grades. Similarly, greater variation around the intercept indicates some students have higher or lower starting points on their reading scores.

Piecewise Growth Curve Model

Piecewise growth curve modeling (PGCM) is an option to use with nonlinear trajectories where the pace of change is faster in some periods than in others (Chou et al., 2004; Cudeck & Harring, 2007; Flora, 2008; Grimm et al., 2011). Piecewise growth curve modeling can be applied in conjunction with various statistical models, for example, hierarchical linear growth modeling (HLGM), structural equation modeling, and Bayesian modeling (Bollen & Curran, 2006; Flora, 2008; Raudenbush & Bryk, 2002). Instead of looking at a single growth for the complete window of repeated measure of time, PGCM divides the nonlinear trajectory into separate series of linear trajectories or segments of different slopes, which are connected by

turning points (or change points, also called a knot) at specific inflection points (Chou et al., 2004). For example, there might be a combination of linear growth that slowly increases in the beginning of the study, rapidly increases to the peak of the trajectory, then the growth decreases, and slowly increases again. Hence, there will be three linear trajectories that can be broken up into three phases to assess the initial increase, the subsequent decrease, and the increasing growth at the end, for which each line of the trajectory is tied to the next at the curve's turning point (inflection point). Since nonlinear trajectories are divided into linear pieces, the growth trajectory can be interpreted the same way as linear trajectories based on linear change in the dependent variable per unit change in time. Moreover, if the change points are unknown, Bayesian inference can be used to estimate the trajectory (Chou et al., 2004).

Piecewise growth curve modeling has been widely used in various fields of applied research including education, medicine, and psychology. For example, Jaggars and Xu (2016) used a piecewise growth curve model to study the earning trajectories of community college students. The study of Dagne and Ibrahimou (2017) incorporated a piecewise growth curve model using Bayesian analysis to estimate the time of the effect of an antiretroviral drug for HIV. Li et al. (2001) applied piecewise growth modeling via maximum likelihood estimation to assess the developmental change in older adults' daily activity based on associated demographic data (i.e., age, gender, educational level). Leroux (2019) examined mobile students' (students who attend multiple schools during the study) performance from kindergarten to second grade via three-level piecewise growth modeling, where repeated measures were nested within student and students were nested within school.

As mentioned above, PGCM breaks up the nonlinear trajectory into separate linear sections for the analysis; therefore, it offers users an advantage in ability to compare growth rates

during two or more different time periods. In other words, researchers can observe change that exhibits distinct phases of growth on the observed variable (Kim et al., 2015). The major challenge of using PGCM is the specification of the turning point. For example, we usually specify the change point in the data based on theory or design (i.e., the starting point of the intervention); however, the changing point might happen after the intervention as a result of postponement in response to intervention. Consequently, misspecifying the changing point can lead to improper interpretation of growth traits (Chou et al., 2004).

Repeated Measures Analysis of Variance

Repeated measures analysis of variance is also known as repeated measures ANOVA, within-subjects ANOVA, or ANOVA for correlated samples. Repeated measures ANOVA is one of the well-known traditional statistical methods to analyze longitudinal data (Vasey & Thayer, 1987). Repeated measures ANOVA is an extension of analysis of variance (ANOVA) that accounts for related, not independent, means and correlation between repeated factors. Since each individual is measured repeatedly over time using longitudinal data, the assumption of independent observations is violated for the traditional ANOVA because each participant contributes more than one measurement score to the data set (Maxwell & Delaney, 1990). The results of the analysis can be misrepresented if the assumption of independence is violated. Thus, we cannot just simply use one-way ANOVA to test whether the means for each time point differ. Moreover, in ANOVA, the discrepancy among individuals in their average measurement scores is treated as error. But in longitudinal data these differences show higher or lower values in traits/characteristics that we are measuring rather than error. Consequently, using one-way ANOVA with longitudinal data overlooks some informative details in the data set (Maxwell & Delaney, 1990; Vasey & Thayer, 1987).

Generally, repeated measured ANOVA is used to answer questions about (a) changes in average scores on three or more time points, and (b) differences in average scores on two or more levels of the between-groups factor or testing conditions and the interaction effects (Von Ende, 2001). For instance, the above example of studying students' reading improvement from grade 3-5 is extended to repeated measured ANOVA analysis. In this case, we are interested in comparing three methods for teaching style of reading in 50 students at four separate time points (beginning of grade 3 [initial reading scores], end of grade 3, end of grade 4, end of grade 5). The dependent variable is the change in students' average reading score and the independent variable is the teaching style of reading. This example is considered a mixed model with students as a random effect and time as a fixed effect. The 50 students are a random effect because they are randomly selected from a population of students who are in third grade and these 50 students represent each of 50 levels of the random factors (person). Times are a fixed effect because the study specifically selects the above four time points for the purpose of the study. All 50 students are tested on their reading score in the beginning of grade 3, end of grade 3, end of grade 4, end of grade 5.

The main advantages for repeated measure ANOVA are (a) increased statistical power compared to ANOVA and (b) requirement of fewer participants than ANOVA (Howitt & Cramer, 2011). Since the repeated measures ANOVA tests the differences within a person at different time points, which can exclude the effects of individual differences that possibly happened when measuring multiple people, it can control for factors that cause the variability (keep unexplained variability low). In turn, the statistical power is increased. Also, when the test has high statistical power, it requires smaller sample size to detect the magnitude of differences among conditions/times. When the study requires smaller sample size, it is also more cost

efficient (Field, 2011). While repeated measures ANOVA allows users to assess the test over time, requires smaller sample size, and has good statistical power, its major drawbacks are that (a) all subjects are required to be measured on the dependent variable at the same time points and with the same number of time points, (b) the rate of growth is assumed to be the same across individuals, and (c) the starting point is the same for everyone. These limitations indicate that measurement times must be fixed for all participants (constant variance and constant autocorrelation) and that repeated measures ANOVA treats intercept and slope variation and nuisance even though these values might differ across individuals (Field, 2011; Howitt & Cramer, 2011). With these shortcomings of repeated measures ANOVA, additional statistical models have been developed to attempt to address its drawbacks.

Multilevel Linear Growth Model

The nature of longitudinal data has a nested structure with times nested within individuals. This could occur, for example, when data are not collected at identical time points for all subjects and subjects might also differ in the number of non-missing data points. In such a case, hierarchical linear growth modeling (HLGM) has been adopted to analyze growth across time (Bryk & Raudenbush ,1987; Muthén, 1997; Schonfeld, & Rindskopf, 2007; Singer, 1998).

HLM growth modeling is considered a mixed effects model, for which group or condition that we are trying to assess can have multiple random effects, random intercept, random time, and random slopes. HLM growth also requires a minimum of three time points (cannot be used with a simple pretest and posttest design) and focuses on the change in individuals as a function of time (Raudenbush & Bryk, 2002; Schonfeld & Rindskopf, 2007).

The example above from repeated measured ANOVA about comparing the effectiveness of three methods of reading teaching styles in student grades 3-5 in the four time points

(beginning of grade 3 [initial reading scores], end of grade 3, end of grade 4, end of grade 5) can be extended to a 2-level HLM growth context. The dependent variable is the reading scores, and the independent variable is teaching reading methods. In a 2-level HLM growth model, time (or growth) is modeled at level 1 and students are modeled at level 2. The student is the clustering effect in that a student's growth is nested within that student. We can think of level 1 as a within person model (change within person across time) and level 2 is a between-person model (change across person and across time). Similar to repeated measure ANOVA, HLM growth is adopted to handle the correlated residual aspects in the data. When we test the students' reading scores at the end of their third grade, the residual variance from those scores is likely to be correlated with residual variance from reading scores measured at a later grade. The correlation in residual variance makes it impossible to separate growth from individual characteristics of students. On the other hand, if we look at time when we test students' reading scores as nested within the students, then we can control for which students the test occasions are nested. Thus, we can better control for these dependencies within the data (i.e., the correlated residuals) by calculating a different slope for each student, rather than a single average reading score change (slope) across students, which we can use to measure change over time.

Before assessing the effects of reading teaching methods across time, we can examine the extent to which reading teaching method has an effect on students' reading scores, starting with the students' beginning reading scores before being exposed to the reading instruction over the course of three years (grades 3 through 5). Since each student's starting reading score (intercept) and reading trajectory (slope) are estimated individually, both intercept and slope are considered as random effects in the HLM growth model. However, if the intercept and slope are not

different across students, they can be treated as a fixed effect in level 2 of the model (Bryk & Raudenbush, 1987).

The common first step to build an HLM growth model is specifying the unconditional growth model (null model), with time entered as a fixed effect predictor variable at level-1, and both the intercept and the slope for time as random factors at the level-2 (Raudenbush & Bryk, 2002). The unconditional model is used to determine if there is statistically significant variation in initial reading score and reading score trajectory. If either the intercept and/or slope variance in the unconditional model is statistically significant, it means there are differences in the students' initial reading scores and/or reading score trajectories. Consequently, we can add predictors to explain the reading intercept and trajectory across students in level 2. Generally, level 2 in an HLM model signifies interindividual difference in change; we theorize that students' reading growth depends on some predictors to make the change. In our example, the three methods of reading teaching styles are modeled at level 2. Moreover, if students' reading scores are assumed to differ according to genders or parents' economic status, for example, then gender and students' parents' economic status would also be added to the level-2 model to explain the initial level and rate of change in students' reading scores.

The advantages of the multilevel approach to assessing change across time are (a) having the flexibility to handle unbalanced structures (based on incomplete data which are assumed to be missing at random), (b) not requiring the same number of observations nor collecting data at the same times for individuals, (c) allowing researchers to assess individuals' trajectories and starting points (intercepts) and factors that might affect those trajectories and starting points, (d) allowing for non-constant variance and non-constant auto correlation, (e) working well with a large number of repeated measures and the ability to be expanded to higher levels of nesting, and

(f) serving a model comparison purpose, i.e., comparing the fixed effect coefficients between nested models (Jackson, 2010; Lininger et al., 2015; Schonfeld & Rindskopf, 2007).

In addition to advantages listed above, HLM growth modeling has its own limitations. For example, the situation where we expand an HLM growth model to three levels (i.e., repeated measures at level-1, nested within students at level-2, and nested within schools at level-3). HLM requires large sample sizes at level-3 (at least 30 data points for the level 3 units) to be sufficient for estimation of regression coefficients and to obtain unbiased estimates of variance components (Amatya & Bhaumik, 2018; Hox, 2002; Maas & Hox, 2005; Mok, 1995). Furthermore, it is more important to increase the number of level-3 units, which are groups (e.g., schools) than level-1 units, which are the observations per group (e.g., students). Moreover, HLM growth only allows data to be missing at level-1 and level-2 (time and individual level). For instance, if the performance data at one school (level-3) are missing, that particular school's data need to be entirely removed. Additionally, the model estimation can be bias if the trajectories are non-linear. HLM growth also can get very complicated and the results hard to interpret when including multiple variables and complex growth trajectory shapes (Gentry & Martineau, 2010; Woltman et al., 2012).

Structural Equation Modeling of Latent Growth Curve Modeling

Latent growth curve modeling (LGM) is one of the applications within the framework of structural equation modeling (SEM) to analyze change over time and is closely related to HLM growth modeling (Duncan & NetLibrary, 1999; Vasantha & Venkatesan, 2014; Willett & Bub, 2014). The approach for testing repeated measures in SEM is not based on specifying the repeated measures as nested within individuals like HLM growth models; instead, the repeated measure represents the observed indicators that define the underlying latent factors, and those

latent factors represent the growth trajectory and intercept. The repeated measures are viewed as latent in the sense that we do not directly observe the trajectory, but we infer the existence of the function as if we did observe the repeated measures. In LGM, repeated measures are used as observed indicators to identify the structure of the latent variable; this way, repeated measurement is modeled by the latent intercept and slope variables of the latent growth curve (Duncan & NetLibrary, 1999).

One of the unique characteristics of LGM is the data structure. A majority of longitudinal data analyses, except repeated measures ANOVA, require a person-period or univariate (long) format, while LGM requires data to be structured in multivariate (wide) format (Willett & Bub, 2014). For data in long format, every case has multiple rows in the record, with the time variant information in the vertical array and unique row to record time. In long format the rows represent different times whereas in wide format the columns represent different times. In contrast, in multivariate format, each person has a single row in the dataset, with multiple (multi-) variables (-variate) comprising the time invariant information in the horizontal array (different columns of data). For example, in comparing the effectiveness of the three methods of reading teaching styles in student grades 3-5 at the four time points, in multivariate format, the four time points require four columns to record each student's growth record and each column represents a measurement occasion. The time invariant variables of the reading trajectory (e.g., three methods of reading teaching styles and students' gender) have their own columns in the dataset. Additionally, there is no exclusive column to record time; the values in the reading scores' first column were measured at the beginning of grade 3, values in the second column were measured at the end of grade 3 and so on. The main reason that LGM requires multivariate data structure is that LGM uses the repeated measure as observed indicators to define an underlying latent factor

and uses covariance structure analysis (CSA) to compare sample and predicted covariance matrices (and mean vectors); thus, the data must be structured to support covariance matrices estimation (Willett & Bub, 2014). CSA is a method to study covariance structure, in which the sample covariance matrix (and mean vector) is assessed to estimate the relations among variables, including the various measurements of the dependent variable across the multiple measurement times. All values in the dependent variable of times and all parameter and time-specific residuals are formatted as vectors and matrices in which values representing the change are entered into columns of a factor loading matrix (Meredith & Tisak, 1990; Willett & Bub, 2014).

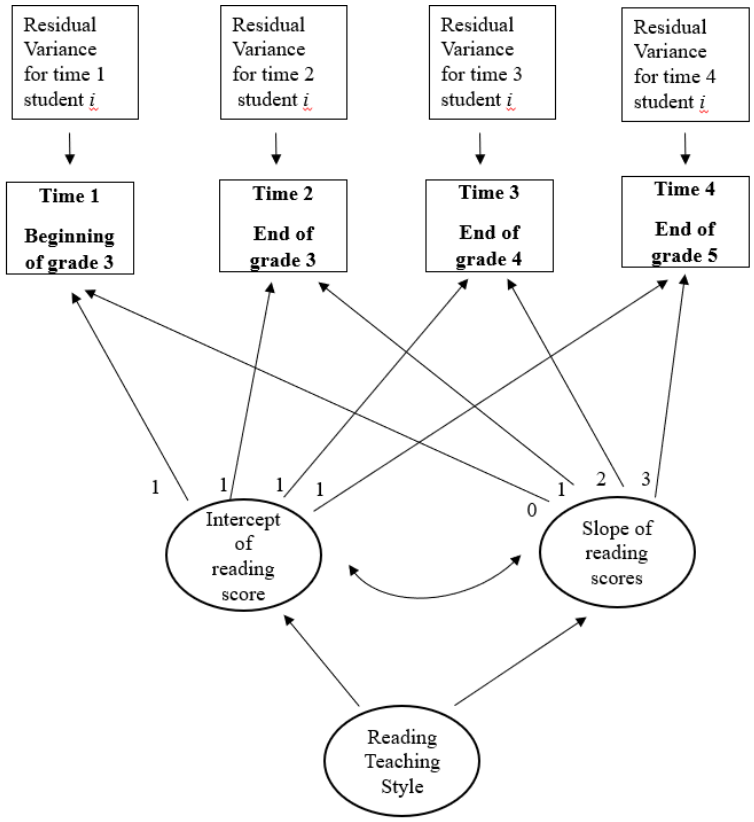
Preacher (2010) stated that "LGM is intended as a way to test the priori predictions of a theory of change against observed data. ……The lack of strong theorical predictions can lead to the misuse of LGM to generate theory from data in an exploratory, inductive fashion" (p.186). Using our example of comparing the effectiveness of the three methods of reading teaching styles in student grades 3-5 at the four time points, we hypothesize that the three different methods of reading teaching styles will influence students' reading ability. Thus, we can use LGM to test our theoretical reasons for identifying students' trajectories, which are characterized by an intercept and slope and the reading trajectories can be captured in terms of means, variances, and covariances.

The major components of latent growth modeling include the structural model, CSA Y-measurement, and X-measurement (Willett & Bub, 2014). The structural model is visualized through a path diagram to map out the relations between factor loadings associated with the students' reading scores and their constant and time-based values. The endogenous latent constructs are enforced to become the individual growth parameters (Curran, 2003). LGM in the

SEM framework is considered a single level analysis where dependent variables (reading score) become a multivariate outcome and the latent slope, and latent intercept variables are assessed by the multiple indicators of time repeated measures and the structural part linking the growth factors together (Curran & Hussong, 2002). To draw the paths as shown in Figure 1, there will be individual boxes representing each observed variable (reading score), repeated measures (four time points), single-headed arrows for path coefficients, and double-headed arrows for variance-covcariance (Preacher, 2010).

Figure 1

Latent Growth Curve Modeling Example



Note. The above example compares the effectiveness of three methods of reading teaching styles on reading scores in student grades 3 through 5 at four time points.

The students' reading score intercept and slope are considered as latent variables because they cannot be directly observed (Bollen & Curran, 2006). In fact, the change in reading score for each student is not what we are interested in; we are interested in using the repeated observations to identify the underlying latent trajectory that we believe to exist but cannot directly observe. The latent intercept represents the reading score in the beginning of grade 3 and the latent slope represents the trajectory in students' reading scores. If we find statistically significant variances for either latent intercept and/or latent slope, it suggests that we should add some variables that we believe can explain the changes in the intercept and/or slope. We want to take the information about what we did observe (reading scores) to infer the existence of what we believe to cause the change in the variable that we observed (for example, gender, parental style, family background). The CSA Y-measurement (equivalent to level-1 in HLM growth model) represents individual students' change and can be thought of as an unconditional model (model with no predictors). Students' reading trajectories can vary across students and covary with one another (i.e., initial reading scores may covary with rate of change). Moreover, the individual growth parameters (both intercept and slope) are transferred to an endogenous construct vector (known as latent growth vector).

The X-measurement model (equivalent to level-2 in an HLM growth model) is used to accommodate time-invariant predictors of change (in this case, the three teaching reading methods). If we have other indicators that are related to the predictors (i.e., students' gender, students' family background/parental style), we can add those indicators in the X-measurement model, which is done by expanding the exogenous indicator and construct score vectors to include sufficient elements to contain the indicators and construct (Curran & Hussong, 2002). In the X-measurement model, we can answer questions like: Do the students' initial reading scores

differ for boys and girls? Or do the students' reading scores change depending upon the teaching reading methods? The above X-measurement questions are addressed by specifying a CSA structural model because it is in the CSA structural model that the vector of unknown endogenous constructs, which now contains the all-important individual growth parameters, is hypothesized to vary across people (Willett & Bub, 2014).

As mentioned above, the HLM growth model is closely related to latent growth curve modeling in structural equation modeling; therefore, data in HLM growth can be mapped onto the general covariance structure model and all parameters can be estimated using standard covariance structure analysis (Curran & Hussong, 2002; Willett & Bub, 2014). Both LGM and HLM growth are similar in the way that (a) individuals' intercepts and slopes are viewed as random effects, (b) the unconditional univariate growth model (model with no predictors in the model) can have linear, quadratic, or piecewise trajectories, and (c) time-invariant covariates are used to explain the variation in the unconditional model. While LGM and HLM growth are similar in various aspects, the key difference between the two approaches are (a) time repeated measures are nested within individuals in the HLM growth model, while time repeated measures are treated as indicators of a latent trajectory variable in LGM, (b) HLM uses a univariate data format, while LGM uses a multivariate data format, (c) HLM normally uses maximum likelihood, or restricted maximum likelihood (REML) for model estimation, where LGM mainly uses full information maximum likelihood (FIML) for model estimation. In turn, use of REML makes it is hard for LGM to invert a large volume of complex dimension matrices. Another challenge in estimating latent growth curve models is the reliance on particular software packages compared to HLM (Curran, 2003; Hox, 2002; Swaminathan & Rogers, 2008).

The main advantages of LGM for observing growth trajectories include (a) not requiring individuals to be measured at the same occasions or time intervals which allows for missing data, (b) viewing the trajectories and intercepts as random effects, thus allowing the individuals' differences in both intercept and slope to be assessed, (c) having the ability to decompose total effects into direct and indirect effects, and to calculate standard errors of these effects. Also, LGM is a special case of SEM, so all the advantages of SEM also apply to LGM, which include the ability to assess the wide range of model fit (i.e., chi-square, root mean square error of approximation, comparative fit index), the ability to examine change in latent variables, and the ability to examine the backgrounds and pathological conditions of change (Curran, 2003; Preacher, 2010).

Although the above listed strengths have made LGM an increasingly popular analytic option for longitudinal data, LGM has its own limitations, including that it (a) needs several necessary steps to construct appropriate multivariate structured data, and (b) does not work well with a large number of repeated measures as each repeated measure becomes an indicator on the latent factor. Therefore, inverting a large number of repeated time measurements can become very challenging, Additional limitations include (c) the challenge of testing interaction effects and (d) requiring large sample size with the average required sample size at around 400, depending on number of indicators (Curran, 2003; Tomarken & Waller, 2005).

Bayesian Inference

As discussed in the sections above, growth curve modeling can be expressed through the repeated measures ANOVA, multilevel, structural equation modeling, or piecewise growth model frameworks, with each approach having its own strengths and weaknesses. Subsequently, Bayesian growth curve modeling was developed as one of the options to analyze growth

trajectories and was a main focus of the analysis for the current dissertation study. Several researchers support that fitting growth curves within the Bayesian framework offers a lot of flexibility to the model with various levels of complexity (Curran et al., 2010; Dunson, 2000; Gelman, Carlin, et al., 2015; Grimm & Ram, 2009; Oravecz & Muthén, 2018; Zhang, 2016). There are several features that make Bayesian growth modeling outstanding. First, Bayesian modeling allows for less restrictive data characteristics than other modeling procedures, including: unequally spaced occasions, and non-normally distributed repeated measures Curran et al., 2010; Grimm & Ram, 2009). Second, Bayesian modeling allows prior knowledge to be included in the model which provides a natural and principled way to incorporate past information about a parameter. Third, Bayesian inference is easily drawn based on the posterior distribution of model parameters, even if data have been transformed. Moreover, the result from the posterior distribution is easy to interpret in an intuitive way by describing it as a likely range of the parameter. For example, the true parameter of the data has a probability of .95 of falling within a 95% credible interval (Gelman, Carlin, et al., 2015). Fourth, Bayesian modeling can estimate complex models and data structures via the Markov chain Monte Carlo (MCMC) simulation-based algorithm in which the unobserved variables can be replaced by simulated variables (Dunson, 2000; Zhang, 2016). Fifth, Bayesian modeling works better with small sample size when a prior is specified compared to maximum likelihood estimation (MLE) because Bayesian modeling is not based on asymptotic concepts and restrictive estimations are not essential for posterior distribution (Berger, 1990; Hoogland & Boomsma, 1998; Schafer, 1997; Scheines et al., 1999). In addition, with small samples in growth data, Bayesian modeling shows smaller prediction error (i.e., mean square error) compared to modeling based on MLE (Lee & Liu, 2000).

Before diving into the concept of Bayesian growth curve modeling, it is helpful to understand the concept of Bayesian statistics, which are the foundation of Bayesian growth curve models. The use of Bayesian techniques has become increasingly popular in recent years, though Bayesian concepts have existed since 1770. Thomas Bayes, who was an English mathematician introduced the Bayesian concept known as the "Bayes Theorem" (also known as Bayes' rule): a mathematical technique that applies probability to statistical problems, which allows users to update their knowledge/belief in the evidence as a part of model (Gelman, 2004). The reasons that help Bayesian statistics to increase their popularity include (a) the ability to incorporate both sample and prior knowledge into the parameter estimate, (b) flexibility to include a wide range of data types, (c) easy implementation of the Bayesian approach with the help of statistical software development, and (d) good performance with smaller sample size (Zitzmann & Hecht, 2019).

Concept of Bayesian Statistics

Bayesian modeling consists of three main components: posterior distribution, prior distribution, and estimation methods of the posterior distribution (i.e., Markov chain Monte Carlo [MCMC], Gibbs, or Metropolis-Hasting). To build a Bayesian model, the majority of the time the researchers first specify a prior probability distribution for the unknowns in the model (e.g., parameters and latent variables), then they apprise the prior distribution to attain a posterior distribution for the unknown parameter by identifying the prior and the probability of the data (conditional on the unknown) into the Bayes' theorem.

Since Bayesian growth modeling is an extension of Bayes' theorem, it is useful to present the Bayes' theorem equation and walk through the notation to have a better understanding of each component. Equation 1 is the equation of Bayes' theorem.

$$p(H|E) = \frac{p(E|H)\,p(H)}{p(E)} \qquad (1)$$

To make the equation easy to digest, let us consider the situation where a non-profit organization has a major event that they have organized every year to raise money. So, the event organizers would like to assess the successfulness of the event by looking at the amount of donation money they received. Note that Equation 1 is the Bayes' Theorem for discrete probability where there is only one hypothesis in the research which, in this case, would be the event organizers only hypothesize that if people come to the event, they will make a donation. $P(H|E)$ is the posterior component (the outcome), which pronounces the probability of H given E (the conditional probability). $P(H|E)$ means the probability that event H happened, given that event E happened. In the current example, $P(H|E)$ indicates how likely it is that people will make a donation given that they come to an event. Conversely, $P(E|H)$ means the probability that event E happened, given that event H happened, which in this example indicates the probability of people coming to the event given that they make the donation. P(H) is the prior distribution (belief/knowledge). In this case we might know that the observed probability of donation of people attending the event was around 50%; thus, we can include this prior knowledge into the model. The last piece of the equation is P(E), which refers to the observed data, or in this example, the number of people who attend the event.

Equation 1 is normally used when there is one hypothesis in the research. However, if we have multiple hypotheses in the research, the Bayes' Theorem equation has to be restated as Equation 2 below**:**

$$p(H_i|E) = \frac{p(E|H_i)\,p(H_i)}{\sum_{i=1}^{n} P(E|H_i)P(H_i)} \qquad (2)$$

Let's extend the above situation of the annual event to the multiple hypotheses in Equation 2. Besides the hypothesis that people who come to the event also tend to make a donation, the event organizers might also hypothesize that if the potential donors have a record

of donating to other events, the people attending the event might be more likely to donate to this event. So, the additional hypothesis is represented in stage i which represents the people who have a record of donations at other events. $P(H_i|E)$ is the probability of donation given coming to the event for individual i. $P(E|H_i)$ is the probability of coming to the event given donation for the individual i. $P(H_i)$ is the prior knowledge that the event organizers might have about the event attendees who have a record of donations at other events. The denominator of the equation represents the total of the observed value of number of times they have attended other events about the people who attend the event.

The main difference between Equation 2 (for multiple hypotheses) and Equation 1 (for one hypothesis) is the posterior section. P$(H_i|E)$ in the multiple hypothesis equation (Equation 2) depends on the beliefs in each hypothesis; not only the observed value, E, as in Bayes' theorem for the one hypothesis equation. For example, we can update the prior belief $P(H_i)$ about the people who attend the event for each hypothesis to update the posterior $P(H_i|E)$, which is the probability of people making a donation given they come to the event. The ability to incorporate a prior belief/knowledge relating to the posterior for each hypothesis becomes useful when researchers have different information about the prior for each hypothesis.

The Bayes' theorem for discrete probability was considered above. To apply the Bayes' theorem to continuous probability requires a slight tweak to the discrete probability equation by replacing $p(H)$ with $p(\theta)$ where θ represents the hypotheses about one or more continuous parameters in the model. The Bayes' theorem for continuous data can be stated as in Equation 3 below:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int_\theta p(\theta)p(y|\theta)d\theta} = \frac{likelihood\ x\ prior}{normalising\ factor} \quad (3)$$

Where $p(\theta|y)$ is the posterior probability distribution which, after reviewing the data, is the researcher's belief about the model. The y denotation means the evidence shown by the data. p_θ refers to the prior probability distribution of θ (prior knowledge about θ), which is specified before seeing any data, and indicates the belief about the model (i.e., the distribution on parameter θ). The uncertainty about θ can be updated after the observations have been received because more information is revealed (Gelman, 2004). $p(y|\theta)$ is the probability of the data given the probability model (parameter θ), which can be represented as the likelihood $L(\theta; y)$ in maximum likelihood estimation (MLE). The $p(y)$ in Bayesian inference is the normalizing constant ($\int_\theta p(\theta)p(y|\theta)d\theta$), which is the same as the likelihood function in MLE for the equation of $p(\theta|y) \propto p(\theta)p(y|\theta) = p(\theta)\, L(\theta; y)$. The posterior distribution $p(\theta|y)$ is a function of the likelihood of the model based on the data, $L(y)$, and the prior distribution is in the parameters of θ (Gelman, 2004; Gill, 2015).

Prior Knowledge

The ability to account for prior knowledge into the model makes Bayesian modeling different from other statistical models. The prior distribution in Bayesian modeling reveals any information, prior belief, or knowledge regarding possible values of model parameters, θ, which combines with the probability distribution of data, and then results in a posterior distribution (McCarthy & Masters, 2005). The additional knowledge in a prior distribution can help to increase statistical power when the sample size is small and serves as mutual knowledge in the field of study (Depaoli, Yang, & Felt, 2017). In longitudinal data, known information about the growth trajectory can be incorporated into the model to reveal knowledge about the cause or pattern of the growth, which in turn, can be useful for coming up with the sources for predicting the trajectory. Moreover, it is important to make the best use of prior knowledge to attain the

parameters' posterior distribution (Gill, 2015). Several studies support that appropriately incorporating a prior distribution into Bayesian growth models can help to determine the ideal growth trajectory and improve model estimation accuracy (Depaoli, 2014; Depaoli, Yang, & Felt, 2017; Walls & Quigley, 2001).

Choosing the Right Prior Distribution

Although the ability of incorporating prior belief into the model is one of the biggest strengths of Bayesian modeling, there is no clear-cut method for how to choose a prior (Congdon, 2014). The choice of prior distribution has an important impact on the posterior distribution. This is because the posterior is the product of the multiplication of prior and likelihood functions and normalizing by integration over the parameter variables (i.e., if the prior that we chose is normally distributed, in turn, the posterior will become normal). In the case that we have large sample size and well-defined parameters in the model, the posterior distribution receives minimal effects from choosing reasonable priors as we have enough direct information from the parameters of interest. On the other hand, if we have small sample size, how we choose the priors becomes more critical because we have limited available data in hand about the parameter of interest (Gelman, Carlin, et al., 2015). Different choices of priors result in different posterior inference; thus, choosing priors can be confusing and challenging. One way to explore the effect of using different priors on posterior inference is through sensitivity analysis, which is the study of how different sources of the model input will result in the uncertainty of the model output (Gelman, Carlin, et al., 2015; Paruggia, 2006). The following section provides some detail about the different types of prior distribution and recommendations of choosing priors in the context that is commonly seen in practice.

Non-informative Prior. A non-informative prior (also known as a reference or objective) means the prior portion contains little explanatory information about the unknown parameters or hypotheses related to the model (Golchi, 2018). Non-informative priors are usually used when there is no dependable prior information about the hypotheses or model parameters, or an implication relying solely on the information at hand is preferable. For example, Shieh and Lee (2002) gave the reason of using non-informative priors to predict an unbalanced growth curve model in their study that "due to the complexity of the model, no analytic forms of the prediction densities are available" (p. 324). The study of Sun and Ni (2004) used a non-informative prior estimation technique called Jefferys prior (the concept of Jefferys prior is discussed below) to analyze the vector-autoregressive models coefficient.

The logic behind using non-informative priors is that we learn from the data as they go by letting the data explain for themselves. Consequently, the posterior distributions are apprehended by a likelihood function and the impact from the prior is diminished. In other words, the hypotheses or values in the model parameters cannot benefit if the non-informative prior is used (Gelman, 2006b). Moreover, Zhang et al. (2007) suggested that the result of using non-informative priors usually yields similar result to MLE.

Most of the time, a non-informative prior distribution is drawn from a uniform distribution function. However, the integral over the constant under a uniform distribution is not equal to 1 (non-normalizable). Consequently, the prior distribution drawn from the uniform distribution needs to be multiplied by the likelihood function because the likelihood function has the ability to make the posterior distribution become normalized (Gelman, Carlin, et al., 2015). Additionally, in case that the variance, σ^2 (which naturally cannot be a negative value) is the parameter of interest, the variance parameter needs to be log transformed ($x = ln\sigma^2$) to have a

uniform probability which can range from negative infinity to positive infinity. Another widely used non-informative prior in practical application is Jefferys' prior, which was developed from the Fisher information matrix that provides evidence about how much information in X is in the θ parameter, if X in $P(X|\theta)$ is a random variable.

Subjective Prior. Subjective prior refers to an informed opinion of the value of a parameter prior to the data collection. A subjective prior should be used when we do not have much information about model parameters but we might have an instinctive idea about the minimum, maximum, mean, and most probable value of the parameter. Therefore, we can incorporate that intuitive knowledge into the prior distribution. Incorporating expert opinion into the prior distribution is another well-known method of creating a subjective prior. For example, the study of Choy et al. (2009) explained how to quantify expert knowledge as an elicitation process for prior distribution in ecology research. Hosack et al. (2017) applied various types of available documented expert knowledge and uncertainty about a risk control option to the prior of their Bayesian generalized linear model in order to reduce the risk of ship collision in Australia's territorial sea and exclusive economic zone. Coussement et al. (2015) accounted for subjective expert opinions as the prior in their Bayesian model to enhance the decision support system of online consumer-satisfaction.

Conjugate Prior. Where both prior and posterior distributions come from the similar probability density functions, and the posterior has closed form, the matching prior to the posterior is referred to as a conjugate prior for the likelihood function (conjugate to the likelihood). For example, if we choose a normally distributed prior, the result for the posterior will derive a normal distribution based on Bayes' theorem. The advantages of using conjugate priors are (a) getting the closed-form expression for the posterior, and (b) reducing the

computation complexity of the posterior distribution (Alhamzawi & Yu, 2013). Thus, the conjugate prior is suitable to use when simplifying the model equations is appropriate. Let us consider the situation where we have three dimensional parameters and we want to estimate the integral to approximate our posterior $p(\theta|E)$ via quadrature (take 1,000 grid points for each direction). This type of situation can occur when there are multiple dimensional parameters (variables). This situation becomes complicated quickly because the computation for the grids results in $1{,}000^3$ or 10^9 data points. Calculating an integral based on 10^9 data points is difficult, especially in real time statistics. Choosing a conjugate prior for the above situation can reduce the computation of the posterior from complicated numerical integral to simple algebra in order to produce a tractable integral. The most commonly used family of distributions that have conjugate priors include the normal distribution, gamma distribution, and beta distribution.

Informative Prior. An informative prior means the prior comprises existing information, knowledge, or a hypothesized parameter distribution associated with the model before the data are collected. An informative prior can be derived from a literature review or previous data analysis which has relevance to the parameter of interest in the model (Bolstad, 2007). In the case where a research study or experiment is conducted for the first time, the non-informative prior, $p(\theta)$, can be applied to the model first because the knowledge about the parameter is still unknown. Once the information from the data, y_i, is revealed, the non-informative prior can be updated to an informative prior, with learned knowledge: $p(\theta|y_i)$. Moreover, additional information from the data (yi) can be obtained and researchers can use the posterior $p(\theta|y_i)$ from the previous research as the prior to update the knowledge about that parameter again.

Informative priors sometimes are disapproved by some researchers because different priors can lead to different model results, which makes the model more subjective (Zhang et al.,

2007). However, when an informative prior is available, it should be adopted into the model over a non-informative prior. This is because non-informative priors show poor performance in parameter recovery and large bias in the posterior distribution (Richardson & Green, 1997; Roeder & Wasserman, 1997). Lee and Vanpaemel (2018) also stated that "Since, in the Bayesian approach, priors and likelihoods combine to form the predictive distribution over data that is the model, priors should also aim to be informative" (p. 115). Moreover, additional information about the model can be viewed through informative priors. Thus, by not employing an informative prior when it is available, important information can be wasted (Bolstad, 2007).

Several studies have been done on longitudinal data using a Bayesian approach and all of them recommended using informative over non-informative priors. The study from Zhang et al. (2007) revealed that informative priors provided "accumulated knowledge" (p. 381) in scientific research and the result from their study also showed that an informative prior increases statistical power and reduces bias in model parameter estimation, particularly when sample size is small. The study from Wolf (1986) viewed an informative prior as a meta-analysis, which is analogous to the combination of results from research on a closely related topic. Relatedly, the study from McArdle and Horn (2004) referred to an informative prior as a mega-analysis, which is the combination of raw data from research on a closely related topic.

Muthén and Asparouhov (2012) looked at growth modeling using a structural equation modeling approach and found that small variance in informative priors worked well to reflect the fundamental theory of the model. Depaoli (2014) examined the impact of incorrect informative priors in the context of growth mixture modeling and concluded that when the variance of hyper-parameters is large, growth mixture modeling is fairly robust to the use of incorrect mean hyper-parameters. Also, informative priors led to an affirmative effect on the model parameter

recovery. The study from Schafer (1997) incorporated an informative prior distribution using a Bayesian updating (BU) technique in a hierarchical linear growth model, where the informative prior was updated from a previous sample (i.e., it used results from the previous year to update the prior for the following year). Results from Schaper's study indicated that using an informative prior distribution improved overall model fit and showed higher accuracy in representing population parameters, compared with using a non-informative prior distribution. Shi and Tong (2017) assessed the influence of non-informative priors, informative priors with different levels of accuracy, and precision and data-dependent priors in a Bayesian latent growth model. Their result showed that (1) misspecified models have worse results in parameter estimation accuracy compared to correctly specified models with inaccurate priors; (2) Bayesian estimation was affected by sample size with higher sample size leading to decreasing bias in the model parameters, mean square errors (MSEs), and the impact of prior information; and (3) inaccurately incorporated prior knowledge leads to incompetent parameter estimates. Depaoli, Yang, & Felt (2017) conducted a sensitivity analysis of priors using Bayesian statistics to model uncertainty in growth mixture models to understand the role of priors in the final model estimations with both non-informative and informative priors. Their result showed that selection of priors greatly affected final model estimations; the more precise the information about the priors, the better the recovery in posterior distribution.

Interaction between Priors and Other Data Conditions. Thus far, it is apparent that researchers should use informative instead of non-informative priors in longitudinal research when they are is applicable (Depaoli, 2014; Shi & Tong, 2017). However, besides priors, there are other data conditions that affect model accuracy in Bayesian modeling (i.e. sample size, missing data, and proportion of cases at each level of a binary predictor). Unfortunately, aspects

relating to the extent to which the type of priors interacts with other model conditions have received little attention in the literature (Berger, 1990; Houghton, 1984; Jarociński & Marcet, 2019). Sigley (2003) noted that "interaction effects whereby the influence of some factor(s) is conditional on the values of other factor group, have received considerably less attention and, even when recognized, are rarely quantified" (p. 227). Understanding the moderation effect between priors and other data conditions is important because it can help applied researchers to know whether the effect of priors changes depending on the level of another data conditions, that is, if the effects of priors are not the same for all levels of the other data conditions. Consequently, understanding interaction effects can help researchers to make decisions regarding what underlying conditions should be considered while conducting Bayesian modeling to reach acceptable model accuracy (Wang & Preacher, 2015).

Houghton (1984) provides an example of interaction effects between priors and other model characteristics in Bayesian modeling. Specifically, he looked at the interaction effect between type of priors (informative versus noninformative) and age of data, that is, new data (most recent three years) versus old data (three year prior to most recent three years) on predicting business failure accuracy. Houghton's result showed that an informative prior with new data yielded the highest model accuracy (79%) compared to the combination of an informative prior and old data (70%), a noninformative prior and new data (69%), and a noninformative prior and old data (60%). Jarociński and Marcet (2019) examined the interaction effect between priors (informative versus noninformative) and coefficient value (0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2) on root mean square error (RMSE) in a vector auto aggressive model with a small sample (N = 25). Their result showed that RMSE values became smaller with the combination of informative prior and coefficient values greater than 0.5. Brutti et al. (2008)

assessed the interaction effect between priors (informative versus noninformative) and sample size (N = 50, 100, 150, 200, 250, 300) on model statistical power. Their result suggested that power significantly increased with the mixture of informative priors and increasing sample size (from 50 to 100 and 100 to 150). However, power flattened as the sample size reached 200. In a simulation study, Finch and Miller (2019) studied the interaction between informative prior accuracy and sample size (30, 40, 60, 80, 100, 120, 140, 160, 180, 200) on the absolute bias statistic (ARB) in Bayesian structural equation modeling. Their results suggest that inaccurate informative priors with large variances yield similar parameter estimates as accurate informative prior when sample sizes are larger than 140.

Although the above studies introduced the concept of interaction effects between priors and some data conditions, research gaps remain. For example, research from Brutti et al. (2008) only simulated data based on experimental designs. It is unknown whether the interaction effect observed between priors and sample size holds for data characteristics of non-experimental designs. Additionally, research from Finch and Miller (2019) only focused on investigating the interaction between inaccurate informative priors and sample size for Bayesian estimation of a multiple indicators, multiple causes (MIMIC) model. It is unknown how such an interaction impacts other types of models like Bayesian hierarchical linear modeling. The research gaps described above help form the direction of the current dissertation, which used simulated data based on real world correlational data. Moreover, through this current study I examined the interaction effect between priors and multiple data conditions (i.e., sample size, waves of data, and proportion of cases at each level of a binary predictor) in Bayesian hierarchical linear modeling predicting a continuous outcome.

Sample Size in Bayesian Modeling. It is well understood that larger sample sizes can lead to better model accuracy. However, not all research fields have the luxury of having large sample sizes due to the cost and complexity of data collection or the infrequency of observation (e.g., biometrics, clinical study, environmental factors). In addition, there are other features besides sample size (i.e., sample collection, data quality, model selection, model complexity, variables' distribution and reliability, and strength of variables' relationship) that have to be considered to determine model performance (Bayer et al., 2013; Cui & Gong, 2018; Gagné & Hancock, 2002; Marsh et al., 1998; Muthén & Muthén, 2002). Therefore, learning about sample size requirements is critical to reach desired model accuracy.

Unlike regression and SEM, which are often, though not always, based on maximum likelihood estimation, the asymptotic concepts and restrictive estimations are not critical for calculating posterior distributions in Bayesian modeling. Thus, Bayesian analysis works better with smaller sample sizes for complex models. Because Bayesian analysis performs well with small sample sizes, it has become a popular alternative to solving prediction problems (Berger, 1990; Hoogland & Boomsma, 1998; Levy, 2016; Schafer, 1997; Scheines et al., 1999).

There are several studies examining the sample size requirement for Bayesian modeling. A simulation study by Pezeshk (2003) revealed that the sample size needed to obtain the probability (power) for the data to fall in the critical region is around 50 subjects. More recently, Schnack and Kahn (2016) suggested that sample size for Bayesian analysis can be as low as 22 to reach reliable model accuracy.

For Bayesian analysis in longitudinal data, Bae and Mallick (2004) used a two-level hierarchical Bayesian model with a large number of variables (50 different genes) and small sample sizes (Leukemia dataset, $N = 38$; breast cancer dataset, $N = 22$). Their result showed that

Bayesian inference resulted in good prediction accuracy under these conditions. However, their study assumed data to be independent and considered only binary classifiers. White et al. (2018) showed that even with the case of incomplete follow-up (where the number of observations is not the same for all time points), a sample size of around 50 is required for minimizing prediction error, though they did not assess the relationship between sample size and change magnitude in each follow-up. Yin et al. (2008) examined sample size and number of replication measurement (waves of data) requirements for a continuous dependent variable in a Bayesian growth modeling framework. Findings suggested that to derive a precise inference for two replication measurements, a sample size of 58 individuals is required; for three replication measurements, a sample size of 34 individuals is required; and for four replication measurements, a sample size of 23 individuals is required. A major limitation for Yin et al.'s study is that they only used a noninformative distribution for their prior.

Waves of Data. One of the factors that needs to be considered when conducting longitudinal data analysis is the number of waves/replication measurements of data. It is well understood that the minimum of two measurements (i.e., pretest and posttest) is required for repeated measurement; however, Rogosa et al. (1982) argued that if data are limited to only two measurements, the information about change trajectory is severely limited. Yin et al., (2008) also pointed out that as number of measurements increases the sample size requirement decreases. Willett and Bub (2014) suggested that increasing the number of waves of data also helps to improve the reliability of scores when change is assessed. The recommended number of measurements in longitudinal data is four, with additional measurements often offering diminishing increases in model accuracy (Muthén, 1997; Muthén & Muthén, 2002). Although the recommended number of waves of data is known, there is limited research on the interaction

effect between priors and the number of waves of data in Bayesian inference. Since number of measurements is an important methodological consideration in Bayesian modeling with longitudinal data, learning about how the effect of priors on outcomes like model accuracy change, depending on the number of waves of data, can inform researchers' methodological decisions (Wang & Preacher, 2015). If the type of priors were found to moderate the effect of number of waves of data, such information could be beneficial to applied researchers who lack access to four or more waves of data.

Proportion of Cases in the Levels of a Binary Predictor. For the majority of the situations when binary predictors are used in Bayesian modeling, the proportion of cases within each binary value is not equal (Dixon et al., 2009). For example, the number of customers who buy versus not buy a certain product is not equal because many more customers usually do not buy any specific product than do; or, the number of alumni who stay engaged versus do not stay engaged with the university in philanthropy is not equal as far fewer alumni usually stay engaged than do not.

In regression analysis, the predicted mean value on an outcome variable for the group coded as 0 (on a binary predictor variable) is equal to the value of the intercept, whereas the slope indicates the mean difference in outcome values between the group coded as 1 and the group coded as 0 (Cohen et al., 2013). Let's use the example of alumni who stay engaged (coded as 1) versus do not stay engaged (coded as 0) with the university to predict donation amount. One would like to determine whether engagement categories of alumni differ in mean donation amounts. We would interpret an intercept value of 250 and slope value of 50 as meaning that the estimated mean donation amount among alumni who did not stay engaged with the university to be $250, and the estimated mean of donation amount among alumni who did stay engaged with

the university to be $50 higher than for alumni who do not stay engaged. Thus, the estimated mean donation amount among engaged alumni is $300. If the proportion of alumni who stay engaged and do not stay engaged is equal or even slightly different, we may not only conclude that these two groups differ in their donations by $50, on average, but we may also accurately conclude whether or not this mean difference is statistically significant at some predetermined level of alpha. However, if the proportions of alumni who stay engaged versus do not stay engaged are drastically different, one or more parameter estimates in our model may be biased because of unequal variances between the two groups (Babyak, 2004). Leonard (1975) and Aitkin (2001) showed Bayesian inference is robust to unequal variances. The common ratios of subject categorization on a dichotomous predictor used in Monte Carlo simulation studies have been 10:90, 25:75, and 50:50 (Shaw & Mitchell-Olds, 1993; Zahn, 2010). However, the moderator effect between type of priors and proportion of cases in the levels of a binary predictor is still unknown. Therefore, as part of the current dissertation I examined this aspect, because understanding how the effect of priors on outcomes like model accuracy changes depending on proportion of cases in the levels of a binary predictor can help researchers to precisely interpret binary predictors in Bayesian analysis (Bayer et al., 2013).

Markov Chain Monte Carlo Method for Estimating Posterior Distribution

Once we define our prior, calculation of the posterior distribution can begin. Calculating a posterior distribution in Bayesian inference can be complicated due to the computation of a normalizing constant, where each normalizing constant can have several numbers of dimension integrals and several parameters. Consequently, the difficulty in computing all dimensional integrals in the posterior distribution is a major reason that most researchers avoid using a Bayesian inference approach. However, with new developments in approaches to computing

posterior distributions, which are available in various statistical software, there has been an increasing number of researchers using Bayesian analysis (Koduvely, 2015).

Markov chain Monte Carlo (MCMC) is one of the techniques that is well known under Monte Carlo simulation's umbrella that has been developed for approximating posterior distributions. Ravenzwaaij et al. (2018) refer to MCMC as "a computer-driven sampling method. It allows one to characterize a distribution without knowing all of the distribution's mathematical properties by randomly sampling values out of the distribution" (p. 143). As stated in its name, Markov chain Monte Carlo comprises two components: Monte Carlo and Markov chain. Monte Carlo is a method of using random samples from a distribution to estimate the properties of a distribution. For example, if you wish to use a mean equation to find the mean of a normal distribution, the Monte Carlo approach draws a large number of random samples from a normal distribution, then calculates the mean from the random samples being drawn. As an example, using a Bayesian concept, MCMC draws values of θ from approximate distributions and then corrects those draws to better approximate the target posterior distribution $p(\theta|y)$.

While Monte Carlo represents the use of drawing random samples to estimate the posterior distribution, the Markov chain component in MCMC refers to the sequential process that creates the random samples. In the process of drawing multiple random samples from a normal distribution, each random sample is used as a base to create the next random sample (which can be thought of as a chain) in Markov's chain property. A distinct qualification of the Markov chain is that the newly generated sample only depends on the sample that immediately precedes it and does not rely on any samples before the immediately former one. MCMC can also be thought of as an approach that involves simulation via a pseudo-random number

generator to create samples that cover various possible outcomes of a given question to be answered (e.g., placement of children's reading level, coin toss outcome, height of women).

To give a clearer picture of MCMC, the Bayes' theorem in Equation 3 can be used as an example, which is $p(\theta|y) \propto p(\theta)p(y|\theta)$. The θ refers to a set of parameters of interest and y indicates the values in the data. The $p(\theta|y)$ denotes the posterior portion (the probability of the parameter given the data). The $(y|\theta)$ represents the likelihood or the probability of the data given the parameters, and $p(\theta)$ refers to the prior or the a priori probability of y. The symbol $\propto$ denotes "is proportional to." The main goal of deriving the posterior distribution in Bayes' theorem is to use the data to update the prior knowledge by exploring the likelihood of the data for the values in the given set of parameters. However, assessing the likelihood for every possible combination of the parameter value can be really challenging. In the case where analytical expressions for the likelihood function are available, they can be combined with the prior piece of the equation to calculate the posterior distribution analytically, though, a majority of the time, the expression of the likelihood function is not available. Therefore, MCMC is beneficial in calculating the posterior distribution because the MCMC has an ability to generate random distributions to estimate the characteristic of a posterior distribution (i.e., drawing the mean and range from the posterior of a random sample) with a reasonable calculating time, which most of the time is hard to directly assess through calculating manually (Andrieu et al., 2003; Ravenzwaaij et al., 2018).

Since MCMC is based on simulating a random distribution to estimate the posterior distribution, the estimation process is improved by each step if the simulated distribution result yields the target distribution, which can be visualized through iteration plots (Andrieu et al., 2003). For example, a teacher is interested in assessing the mean of the students' reading scores

in a student population. The teacher has an idea that the reading scores follow a normal distribution, have a standard deviation of 20, and a possible maximum score of 100. So far, the teacher has a reading score of one student (i.e., student A), which is 80. The teacher can use MCMC to draw samples from the target distribution, the posterior in this case, which represents the probability of each possible value of the population mean given this single observation. Although, MCMC can draw a sample without analytical expression, for the purpose of simplified illustration, the analytical expression for the posterior of (N [80,15]) is used, which means that the data follow a normal distribution (N) with starting score of 80 and standard deviation of 15. The first step of drawing a random sample in MCMC starts with choosing a plausible value that can be used to draw from the distribution. In this case, we know that the maximum reading score cannot exceed 100 points, so the initial starting value can be any number between 0-100. Once the initial starting number for drawing the distribution is stated (e.g., 75), MCMC then uses the starting number to produce a chain of new samples. Each new sample is created by (a) proposing a small random variation to the most recent sample, and (b) either accepting or rejecting the proposal of the previous step. The algorithm accepts the proposal if the new proposal has a higher posterior value, which means the new posterior proposal can bring the posterior closer to the target posterior distribution. If the new proposal has a lower posterior value than the most recent sample, the iteration will reject the proposal of the new random variation and retain the old sample (Roberts & Rosenthal, 2009). The above steps complete one iteration of MCMC. The steps for iteration are repeated until enough samples are drawn and the random sample chain converges to the target distribution. Additionally, for the reliable posterior result, several Markov chain properties are run with separate starting values to discover the posterior distribution for each parameter in the model (Gelman, Lee, & Guo, 2015).

Bayesian Growth Curve Model

Thus far, the concepts of growth curve model and Bayesian inference have been introduced. This section explains how those two concepts can be merged to use in the application of longitudinal data analysis to estimate trajectories. Although growth trajectories are not limited to straight-line functional growth (linear) and Bayesian inference can be used to analyze non-linear trajectories, this dissertation only focused on using Bayesian inference to fit linear growth curve models (for which the function of change is linear).

As mentioned, maximum likelihood is a commonly used method to estimate growth curve models in structural equation modeling and multilevel modeling. One thing to keep in mind when using maximum likelihood to estimate growth curves, besides requiring larger sample size, is that variance components can be biased if the assumption of normality (errors of the growth curve are assumed to be normally distributed) is violated; however, most data in real life have a high tendency to be non-normal rather than normal (Zhang et al., 2007). If the assumption of normality is violated, it can result in inaccurate estimation of parameters, standard errors, and confidence intervals, and unreliable statistical tests and inference (Leiby , 2006; Tong & Zhang, 2020; Zhang, 2016;). Consequently, the use of Bayesian inference to analyze growth curve models has increased as a tool to estimate growth curve models, particularly those that are complex and complicated to analyze using maximum likelihood estimation (Muthén & Asparouhov, 2012; Song & Lee, 2012; Zhang et al., 2007). For example, Wang and Gelfand (2002) studied unknown change points in growth curves through a Bayesian method. Song and Lee (2012) applied Bayesian inference to examine the dynamic change of longitudinal latent variables and their interaction effects, where nonlinear longitudinal latent effects presented. Leiby et al. (2014) explored mixed outcomes (combination of continuous, binary, and ordinal)

through Bayesian multivariate growth curve latent class models using simulation to validate the model estimation procedures to gather the information about the characteristics of disease or its severity. Tong and Zhang (2020) assessed non-normal error distributions using a Bayesian approach. Jana et al. (2019) created simulation data to study the usefulness of Bayesian growth curve models for high-dimensional (i.e., large number of variables) longitudinal data and found that the Bayesian growth curve model performed well with high-dimensionality even with small sample size or few study units. Dagne (2019) assessed the effect of time varying predictors that are measured with errors and missing values in Bayesian semiparametric growth models for patients who were potentially either progressors or nonprogressors to Acquired Immune Deficiency Syndrome (AIDS).

As described in the section about Bayesian inference, the nature of Bayesian inference is based on a probabilistic model and uses information from the observed data relating to parameters, formally the likelihood, to update the knowledge about the most likely value of the unknown parameters which then becomes the posterior portion of the parameters. Consequently, when fitting growth curve models in the Bayesian context, each model parameter for the growth curve needs to be assigned a probability distribution. Additionally, the model parameters are theorized as random variables with probability distributions (Oravecz & Muthén, 2018). The next step for building a Bayesian growth curve model is to specify the prior distribution. As mentioned in the section about how to choose prior distributions in Bayesian models, choice of priors directly impacts the posterior distribution and is integral to Bayesian modeling; thus, the prior distribution needs to be carefully chosen when fitting a growth curve model in the Bayesian context. Once priors are chosen, the posterior distribution of model parameters can be calculated from the product of the prior distributions and the probability of observed data, normalizing by

the marginal likelihood. The posterior distribution denotes the updated probability distribution assigned to the model parameters after conditioning on the data: a probability distribution referring to the most likely value of a growth curve model parameter included in the data and other model parameters.

To illustrate the Bayesian growth model, an example about studying students' reading trajectory is used. Suppose that we have a sample of 200 students, about whom we are interested in the improvement in their reading recognition ability (i.e., word recognition and pronunciation ability). Thus, the data about their reading scores are collected at four different time points: beginning of grade 3 (initial reading scores), end of grade 3, end of grade 4, and end of grade 5. The time measurement can be denoted as t, for which $t = 1,2,3,4$ in this case. In Bayesian growth modeling, not all students are required to have the same four-time measurements. Consequently, missing time measurements are allowed. Although, there are several types of missing patterns, to keep it simple, this example will assume that data are missing completely at random (MCAR). MCAR means that the students' missing reading scores have nothing to do with the values that are hypothesized about the reading scores or other values in the model. However, if the missing values are not MCAR, the missingness mechanism should be incorporated in the model to avoid misleading results (Little, 1999).

The reading score of student i at occasion t can be referred as $Y_{i,t}$. The change within-student over time can be examined at the initial reading score (intercept) and change trajectory (slope). Thus, when the straight lines are fitted to each student's four measurements, the dependent variable is reading scores and independent variable is time. The growth curve model can be written as:

$$Y_{i,t} \sim N(\beta_{i,1} + \beta_{i,2}T_t, \sigma^2_{eLevel1}), \quad (4)$$

$$\beta_{i,1} \sim N(\mu_{\beta,1}, \sigma^2_{e\beta 1}), \quad (5)$$

$$\beta_{i,2} \sim N(\mu_{\beta,2}, \sigma^2_{e\beta 2}) \quad (6)$$

Equation 4 refers to the effect of time at the student level (which also can be referred to as the level-1 equation), which is set up as a likelihood function. $Y_{i,t}$ is a function of student $i's$ initial reading parameter ($\beta_{i,1}$) and the product between student $i's$ reading score trajectory ($\beta_{i,2}$) and the time measurement (T_t). Thus, the conditional distribution can be specified as $Y_{i,t}$ given $\beta_{i,1}$) and $\beta_{i,2}$: the effect of time on each student given each student's initial reading (intercept) and reading score trajectory (slope). Although data can form several distributions, in this example, the normal distribution ($\sim N : distributed\ as$) is chosen with the time-specific residuals having variance $\sigma^2_{eLevel1}$. The variation of errors is allowed to be related to the predicted student-specific change. The $\sigma^2_{eLevel1}$ portion of the model can be revised to represent the autocorrelation in the error variation by adding additional parameters (random effect) to account for the autocorrelation and the time-dependency in the mean and variance associated with this additional parameter.

Equations 5 and 6 can be thought of as the level-2 or intraclass level, which represents between-student variability in initial students' reading score and reading score trajectory. Equation 5 refers to intraclass initial reading score. $\mu_{\beta,1}$ is the population mean, or the group parameter of reading scores shared across students. $\sigma^2_{e\beta 1}$ is the between-student reading variation of the intercept, indicating the magnitude of difference in reading scores of each student from the overall students' initial reading score. Equation 6 predominantly refers to rates of change in the reading score parameter, $\beta_{i,2}$. $\mu_{\beta,2}$ is the mean of all students' reading score trajectory and $\sigma^2_{e\beta 2}$ is the variation in reading score trajectory across students. Also, if the univariate priors on $\beta_{i,1}$

and $\beta_{i,2}$ are assumed to be independent, they can covary in the posterior distribution. Level-2 or hyperprior distribution pools information across students, which represents the information of a student's initial reading score. In addition, reading score trajectory is drawn from both individual students and all students in the sample. The information at the individual student level is pooled toward the group mean where each student's reading score contributes to the overall students' reading score mean, which in turn updates the student-specific terms.

Most of the time, what we are interested in based on a Bayesian growth curve model is the estimation of the variation in individual-specific intercept and slope. In turn, variance in the group level represents individual differences. It is common that the variance element of the model contains a lot of uncertainty as a result of latent construct variation of individual-specific variance being captured. The latent growth variable is usually estimated with uncertainty (Oravecz & Muthén, 2018). With the help of MCMC, we can address the uncertainty in all parameter values in order to estimate the posterior parameter distribution by assessing the posterior uncertainty of the latent and individual specific growth rate, which is the construct that impacts posterior uncertainty (Gelman, 2006a).

Equations 5 and 6 can be extended to more complex models by adding more variables to examine students' initial reading scores and rate of change in scores. For example, three methods of teaching reading (i.e., whole-word method, phonetic method, and context support method) can be added into the model to see if different teaching styles have different effects on the reading scores. In this example, the whole-word approach is used as the baseline group in order to evaluate between-student variability in initial students' reading score and reading score trajectory. The phonetic method (X_1) and context support method (X_2) are estimated in terms of

how much students' reading scores in these two methods deviate from the whole-word approach method used as the baseline.

Whether a student belongs to one of the three methods can be dummy coded into 0 and 1 values: $X_{i.1}$has a value 1 for student i who belongs to the phonetic method, and $X_{i.2}$ has a value of 1 for students belonging to the context support method. Students assigned to the whole-word approach method will have values of 0 for both $X_{i.1}$ and $X_{i.2}$. The Xs characterizes the time-invariant predictors (individual level). The following are the extended equations with the teaching reading methods groups:

$$\beta_{i,1} \sim N(\mu_{\text{whole-word Int}} + \beta_{\text{phonetic Int}}X_{i.1} + \beta_{\text{context support Int}}X_{i.2}, \sigma^2_{e\beta 1}) \quad (7)$$

$$\beta_{i,2} \sim N(\mu_{\text{whole-word slope}} + \beta_{\text{phonetic slope}}X_{i.1} + \beta_{\text{context support slope}}X_{i.2}, \sigma^2_{e\beta 2}) \quad (8)$$

As in Equations 7 and 8, initial students' reading score ($\beta_{i,1}$) and reading score trajectories ($\beta_{i,2}$) are individual student-specific values which refer to individual reading scores that can be different within group. Parameter $\mu_{\text{whole-word Int}}$ and $\mu_{\text{whole-word slope}}$ represent baseline initial reading score and growth trajectory for the reading scores for students assigned to the whole-word teaching reading method. $\beta_{\text{phonetic Int}}$ refers to the deviation in initial reading score of students from the phonetic reading methods group from baseline initial reading score of students in the whole-word teaching reading group, while, $\beta_{\text{context support Int}}$ refers to the deviation in initial reading score of students from the context support reading methods group and students in the whole-word teaching reading group. $\beta_{\text{phonetic slope}}$ is the deviation in rate of change in reading score of students from the phonetic reading methods group from the baseline (students in whole-word teaching reading group) rate of reading score change. $\beta_{\text{context support slope}}$ is the deviation in rate of change in reading scores of students from the context support reading methods group from students in the whole-word teaching reading group.

In Equations 7 and 8, the distributions of intercept and slope can be specified univariately and can be allowed to covary. Additionally, in growth curve models, the distribution of the individual-specific intercept and slope can be modeled bivariately (Oravecz & Muthén, 2018) as shown in Equation 9.

$$\begin{bmatrix}\beta_{i,1}\\ \beta_{i,2}\end{bmatrix} \sim \left(\begin{bmatrix}\mu_{\text{whole-word Int}} + \beta_{\text{phonetic Int}}X_{i.1} + \beta_{\text{context support Int}}X_{i.2}\\ \mu_{\text{whole-word slope}} + \beta_{\text{phonetic slope}}X_{i.1} + \beta_{\text{context support slope}}\end{bmatrix}, \begin{bmatrix}\sigma^2_{e\beta 1} & \sigma_{e\beta 12}\\ \sigma_{e\beta 12} & \sigma^2_{e\beta 2}\end{bmatrix}\right) \quad (9)$$

The variation of the bivariate mean in Equation 9 is denoted in terms of a covariance matrix, where $\sigma^2_{e\beta 12}$ represents covariation between the individual student-specific intercept and slope, and $\sigma^2_{e\beta 1}$and $\sigma^2_{e\beta 2}$, represent the variances of the terms. To get the student population-level correlation between intercepts and slopes of students' reading scores, we can divide the covariance by the product of the standard deviations.

After assigning a probability distribution to the parameters in the model, it is time to assign the prior distribution. As mentioned in the previous section, there are several choices for priors (i.e., informative, non-informative, subjective, etc.). For illustration purpose, the information with normally distributed priors is used. Assuming that reading scores can range from 0 to 100. The score values from 0 to 100 represent the minimally informative priors on the baseline value for the reading scores of students in the whole-word teaching reading group and on the overall student reading scores. Now, the priors can be set up using the normal distribution and assigned both mean and variance terms:

$$\mu_{\text{whole-word Int}} \sim N(0, 100),$$

$$\mu_{\text{whole-word Slope}} \sim N(0, 100),$$

$$\beta_{\text{phonetic Int}} \sim N(0, 100),$$

$$\beta_{\text{phonetic Slope}} \sim N(0, 100),$$

$$\beta_{\text{context support Int}} \sim N(0, 100),$$

$$\beta_{\text{context support Slope}} \sim N(0, 100) \quad (10)$$

The distribution for error terms can also be specified using the normal distribution (~N) to cover all possible reading score values (the same way that has been done in the intercept and slope). Once the parameter probability distribution and priors are specified, the posterior distribution can be estimated through the MCMC algorithm using statistical software.

So far, the concept of Bayesian growth modeling and the importance of priors have been discussed. Thus, the aspect of priors is carried on as a major part of this study. Particularly, how the effect of priors changes based on the level of another data conditions, for which further detail of how to conduct the study is discussed in Chapter 3.

Machine Learning

Machine Learning for Longitudinal Data

So far, I have talked about the concepts and traditional ways to answer business/research questions using longitudinal data, specifically using Bayesian growth curve models. In subsequent sections, information about analyzing longitudinal data in a machine learning environment is discussed.

As mentioned in Chapter 1, collecting longitudinal and large volumes of data becomes easier with the help of technology. Additionally, the increasing volume and complexity of big data encourages researchers to develop methodologies for handling data integration, data storage, and especially data mining, which is a method of automatically extracting knowledge from datasets with accuracy and coherent meaning (Konerman et al., 2015).

When data are coming in at a faster pace than in the past, human work itself might not be enough to handle the overflooded information. Thus, artificial intelligence has been developed to

efficiently process, organize, and analyze the data. Artificial intelligence (AI) is an innovation that is widely used to support data mining tasks. To recap, AI is the imitation of human intelligence using computer systems for learning, reasoning, solving problems, and making decisions/predictions from the data (Nilsson, 1986). When we talk about AI, machine learning (ML) is also often mentioned. This is because machine learning is one of the important branches under the AI umbrella. The key word for ML is "learn." ML is trained to understand the meaning, pattern, and structure of the data to which it is exposed. As a result, ML helps people to make correct predictions/decisions based on the data they uncover. ML also helps to reduce human labor from manually developing rules and building models based on analyzing massive amounts of data (Jakhar & Kaur, 2019).

The outstanding advantages of ML include, but are not limited to, performing well under large and complex datasets, being able to identify patterns and trends that might not be apparent to a human, and being able to automate methods for data analysis (Jakhar & Kaur, 2019). These characteristics of ML help promote its popularity, as shown by its extensive use as an analytical tool to answer research questions based on longitudinal data (Chen et al., 2018; Walsh et al., 2018). To give a few applied examples of ML for longitudinal data, Wu et al. (2010) predicted the likelihood, from electronic health records, of patients having heart failure through machine learning techniques. Chen et al. (2012) used machine learning as a tool to forecast customers' changing behavior using longitudinal behavioral data. Meng et al. (2016) used ML to analyze the dynamic brain trajectory development in infants, where missing time points were present. Walsh et al. (2018) applied a machine learning approach to longitudinal clinical data to predict suicide attempts in adolescents.

Types of Machine Learning

As seen in the above sections, ML has been widely used in various fields of research and has increased in popularity. Hence, it is beneficial to learn about types of machine learning and how it can be used in different circumstances. Machine learning can be broken down into three main types which are supervised, unsupervised, and reinforcement learning. Supervised ML can be divided into two main subsets which are classification and regression, while unsupervised ML can be divided further into clustering and association.

Supervised Learning

For supervised learning, the machine depends on former knowledge about the dataset to make a prediction with the help of a labeled dataset, which includes data for which we already know the target answers (Raschka & Mirjalili, 2017). The main goal of supervised learning is to predict the value of an outcome variable or label (both categorical and continuous variables) from the source of data in which the target answer is known (also known as training set). To achieve the prediction process, supervised learning uses the example from the labeled data, for which the responses of the outcome variables are known, to train the predictors (Schwaighofer et al., 2005). For example, the known pattern of a specific disease is used as the training set. Once the unknown pattern of disease is fed into the computer, the supervised machine learning algorithm can use the knowledge from the training set to predict the type of unknown disease. In terms of computer algorithms, supervised learning creates a function from the dataset and uses datapoints in the dataset as input vectors. It then creates a predicted value for each datapoint. For example, if we have input variables (x) and an output variable (y), we can use the computer algorithm methods to learn a function, f, that forecasts the output variable, y, from a vector, x, having M input variables. Thus, this process can be referred to as a mapping function $y = f(x)$.

The goal for supervised learning is to proficiently estimate the mapping function; thus, when we have new input data (X) we can predict the output variables (Y) for those data (Alpaydin, 2014). In other words, the machine learns from labeled training data, and then makes predictions about unseen or future data. The learning process stops when the algorithm achieves an acceptable level of performance. To illustrate the previous statements, which can tie back to the example about pattern of specific disease above, suppose we have colorized electron micrograph image of an influenza virus. Then we feed those influenza images into the machine. The machine will analyze and learn the association of those images which is labeled based on their features such as component, shape, size, etc. Consequently, when the new images of the influenza virus are fed into the machine without any labels, with the help of the past data, the machine is able to predict accurately and identify that it is an influenza virus; the algorithm has learned from the labeled examples; thus, is supervised.

To give a few examples of supervised machine learning in applied research, Ladds et al. (2016) employed supervised ML methods to interpret behavioral data of fur seals and sea lions, which was helpful to construct an activity budget for marine animals. Fabris et al. (2017) adopted supervised machine learning to understand the ageing process in the biology field, which used pre-annotated data about ageing (i.e., based on a known function protein) to extrapolate the explanation of new uncharacterized ageing characteristics. Heck et al. (2017) utilized supervised ML to predict ligand-binding affinity for a protein target, which is used as the determination of molecule discovery in the early stage of drug development and returned scoring functions that can assist the decision of drug development for a specific biological system. Grover et al. (2019) applied supervised ML technique as a predictive model to help with the verification activity for the performance-based financing in a healthcare facilities' incentive.

Classification for Predicting Labeled Data. Supervised learning can be further divided into two main subcategories which are classification and regression. For classification, the output variables are discrete and unordered values that can be described as a group belonging to the data. They can be both binary classification (i.e., yes/no, true/false, male/female) and multiclass classification (i.e., orange, apple, or pear). The goal for classification is to predict the category class label of the newly fed data based on past data. Examples of statistical models used in ML classification are logistic regression, K nearest neighbor, decision trees, naïve Bayesian, neural network, and support vector machine (Raschka & Mirjalili, 2017). A practical example of classification for supervised ML is identifying spam emails (Yu & Xu, 2008). Email providers are working with a pool of incoming emails and use the machine to predict whether the incoming email is spam or not. To train the machine to classify spam email, the most important step is to teach the machine what to classify as spam emails and how they look alike (i.e., the labeled data). The processes of classifying spam email are done based on a large number of spam classification tasks. The first step is to review the contents of the email and email headers to see if they contain spam content information based on some key words (e.g., free, lottery, prize claim, etc.). The second step is to create blacklist filters to stop the email that comes from known, blacklisted spammers. The third step is to create spam scores based on the spam blacklist, content, and label; the lower the total spam score, the more likely that email will land in the inbox folder. The fourth and final step is to compute an algorithm that classifies whether the incoming email should be landing in the inbox or in spam folder (Robert, 2014; Yu & Xu, 2008).

Examples of using classification supervised machine learning in applied research include the study from Alghamdi et al. (2017), which compared the three ML predictive classification

methods (naïve Bayes, random forest, and logistic model) for predicting patients' incident diabetes based on clinical attributions that contribute to diabesity. The study from Ocharo and Hasegawa (2018) employed a support vector machine learning algorithm to classify the comments in academic drafts to be either content-related comments (meaningful global revision) or non-content-related comments (local revision) in order to guide students in their research article revision process and help them to enhance their article qualities. Hang and Banks (2019) applied classification supervised ML to classify packs (stock keeping units) in sale tracking audits, where the models were trained to classify the packs and could be used to categorize the new incoming packs. This method led to reducing human labor and production saving cost. Wahab and Jiang (2019) used ML-based techniques for non-parametric models to predict and classify motorcycle crash severity (i.e., fatal, hospitalized, injured, and damaged-only) along with investigating the effect of risk factors that are associated with the motorcycle crashes from the National Road Traffic Crash Database.

Regression for Predicting Labeled Data. While we use classification to assign categorical, unordered labeled data, we use ordinary least squares regression to predict continuous outcomes (e.g., salary base, years of work experience) where the predictive model can be used to show trends in the data based on one or more predictor/exploratory variables. Examples of the statistical models used in regression with supervised data in ML are linear and polynomial regression (Raschka & Mirjalili, 2017).

To illustrate the concept of linear regression in ML, which can be built up to more complex regression models, we have an outcome variable (y) and a predictor variable (x). We can plot the outcome and predictor variables on the x and y axes. The idea is to fit a straight line to these data to try to explain the relationship between the dependent and predictor variables.

What we would like to see is the minimized distance between the sample points and the fitted line where the closer the data points are to the line, the stronger the relationship. The average squared distance method is normally used to calculate the distance between the fitted line and actual data points. All that being said, the goal of regression is to find the linear relationship of the outcome and predictor variables, which also allows us to predict an outcome. The common questions being asked when performing regression analysis are: (a) Do the predictor variables perform a good job on predicting an outcome variable? (b) Out of the set of the predictors we have, which ones are best at predicting the outcome variable? and (c) If one unit of that predictor changes, what is the magnitude of change reflected in the outcome variable? (Robert, 2014).

To give an everyday example of linear regression, suppose we have two variables we are interested in, where the number of hours spent on studying is the predictor variable and test scores are the outcome variable. One would assume that when the number of hours spent on studying goes up, the test score should also increase; hence, they are positively correlated. When we input the data to the regression model in machine learning, the machine will try to understand the relationship between these two variables and determine how one variable depends on the other. After the machine is trained to learn a model that uses the hours of study to predict test scores, it can easily forecast future test score levels based on the given hours of study.

In applied research, regression-based machine learning is widely used to predict continuous outcomes in various fields. For example, Idowu et al. (2016) tried to develop a product realization plan for optimizing energy production by applying regression-based ML to predict temperature (heat load) in a district heating system, which helped them to understand the energy consumption in buildings. Amin and Riza (2018) adopted ML algorithms to enhance a lens optical distance sensor, the distance measurement technique used in an electronically

controlled variable focus lens for a camera, by training the machine with certain acquired features that correspond to target distance values and using polynomial regression-based to predict the accuracy of the operational sensor in lens. Gonzalez and Leboulluec (2019) used the University of California Irvine's communities and crime-supplied data source to predict the total number of violent crimes and to identify crime patterns based on socio-demographic factors (i.e., per capita income and education level). In their study they compared three multiple linear regression-based models (random forest regression, neural network regression, and Bayesian regression) in order to aid the crime prevention strategies.

Unsupervised Machine Learning

We have learned that researchers use supervised learning when they know the answer of the data beforehand (the data are labeled). However, sometimes we have to deal with data that are unlabeled or have an unknown structure. The main goal for unsupervised learning is to reveal underlying structures that are embedded within the data relationships, where the learning process is merely driven by the provided data with no prior knowledge about the data provided. Consequently, the computer is used to recognize and decide whether any obtainable latent patterns exist, and often unsupervised ML helps to reveal both answers and questions from data that have not been considered by researchers. Examples of unsupervised learning including grouping genetic species in biology (Escudero et al., 2011), differentiating groups of customers based on some traits (customer segmentation) in market research (Hang & Banks, 2019), recognizing patterns of behaviors and coming up with purchase recommendation in a business system (Herrman, 2016), and recognizing speech and synthesis in conversational user interface (Chawla et al., 2002). While supervised ML normally works with classification and regression problems, unsupervised ML mainly deals with clustering and dimensionality reduction. Although

unsupervised ML is applied to unlabeled data, the patterns that are used to identify clusters or dimensions still need to be evaluated by either human or computer application (Handelman et al., 2018; Raschka & Mirjalili, 2017).

Given that data can come in large volume, which increases the complexity and heterogeneity, identifying groups within the data can be challenging by intuition. Therefore, unsupervised ML can come in handy to help with figuring out the grouping of the underlying data, which is why it has increased its popularity in applied research. For example, Kallenberg et al. (2016) employed unsupervised ML to isolate features from breast mammograms in order to help with breast density segmentation and mammographic risk scoring, which is used to identify breast cancer risk. Vranas et al. (2017) used an unsupervised ML algorithm to help them to identify patients in the intensive care unit, where the patients shared similarity in diagnoses, into subgroups in order to use these patient subgroup data to further analyze patients' variables. Usama et al. (2019) applied unsupervised ML in unstructured raw Internet network data to assist finding some hidden structures that could be used to improve Internet network performance and provide services (i.e., traffic engineering, anomaly detection, Internet traffic classification, and quality if service optimization).

Clustering. Unsupervised learning can be grouped further into clustering or association. Clustering, sometimes called unsupervised classification, is an exploratory data analysis technique that can help group information into meaningful subgroups (clusters) without having any history about their group membership. Each cluster is formed based on a group of subjects that share certain similar characteristics but are dissimilar to other subjects in other clusters. The machine forms groups based on the behavior of the data (Suthaharan, 2014). The prevalent clustering techniques include, but are not limited to, k-means, hierarchical clustering, manifold

learning, density-based clustering, and latent class analysis (Xanthopoulos, 2014). To name a few examples of applied research that used clustering techniques via unsupervised ML, the study from Oluwadare and Cheng (2017) used a ML unsupervised clustering algorithm to develop a chromosomal confirmation capturing technique. The clustering algorithm helped to identify topologically associated domains in the chromosomes, which helped with studying gene regulation, genomic interaction, and genome function. Miller et al. (2018) applied clustering unsupervised ML to help them uncover structure and information about non-rapidly growing residential building in order to use the result to construct building performance regulation control. Yousefi et al. (2018) employed three unsupervised ML clustering techniques including principal component analysis (to reduce the dimensionality), manifold learning (to further reduce the selected input from principal components), and density-based clustering (for final component reduction) to identify and monitor keratoconus severity, which is the test to analyze a progressive eye disease. Sathiaraj et al. (2019) utilized unsupervised clustering techniques to classify climate types for regions across the United States, for which the result can further benefit the analysis in public health, environment, actuarial science, insurance, agriculture, and engineering.

Association. Another subfield of unsupervised machine learning is association. Association is the rule-based machine learning that discovers interesting relations between variables in a large dataset by discovering the probability of co-occurrence of items in a collection (Raschka & Mirjalili, 2017). For example, assume a researcher is analyzing data from a grocery store. An association can be used to determine which products the customers purchase together, whereas if the researcher were interested in identifying which customers make similar product purchases, the clustering method would be appropriate to use. If customer A bought bread, milk, fruits, and wheat and customer B bought bread, milk, rice, and butter, when

customer C comes into the store and buys bread, it is highly likely that he will buy milk too. Hence, the relationship is established based on customer behavior and the recommendations are made by computer algorithm. This type of analysis is also called market basket analysis (Alpaydin, 2014).

Reinforcement Machine Learning

The last type of machine learning is reinforcement learning (RL). Reinforcement learning is the method in ML that trains the computer to make a sequence of decisions, which is considered as a strategic learning process that progresses based on the situation (Cheng & Yu, 2019). In RL, the machine learns how to perform and make decisions from an interactive environment and consequences of its actions using feedback from the actions and experiences of experiments and errors (Raschka & Mirjalili, 2017).

One of the common terms used in reinforcement learning is “reward,” which is based on the same concept from behavioral psychology for which actions and performances indicate whether one gets a reward from a satisfying performance or punishment from bad performance (Dayan & Balleine, 2002). Since the machine learns from the consequences of correct and incorrect actions, the major goal of ML is to discover an appropriate action model that maximizes the total cumulative appropriate behavior (reward behavior) and minimizes punishment for users (Alpaydin, 2014). Without knowing which specific action to take, the computer tries different algorithms to find out which actions yield the most reward behaviors (shaping reward). It then adds an additional shaping reward trend to the domain knowledge to guide the next action of the computer algorithm; the computer algorithm changes the tactic as a result of the outcome it experiences to maximize the total amount of reward (specified by humans) over the long run.

RL involves autonomous agents (i.e., person, animal, robot) and the environment with which the agents interact. Agents learn to navigate the uncertain environment with the goal of maximizing the numeric reward. To compare RL with unsupervised learning, the main difference between them is the goal of building models to answer questions (Ghahramani, 2015). Unsupervised learning focuses on determining the similarities and differences within data points, whereas RL emphasizes getting better at finding rewards for agents (Ghahramani, 2015; Koduvely, 2015). RL has been adopted in various applied research fields, to mention a couple. In the security and stability control filed, Guo et al. (2004) applied RL algorithm to a voltage controller system to ensure safe and stable security system operation by using an adaptive experimental assessment to update the controller parameter based on the RL signal, while the voltage measurement was used to ensure that the voltage reached a safe level. Komorowski et al. (2018) employed RL to learn optimal treatment strategies for sepsis, which is considered a life-threatening infection and requires immediate proper medication to avoid fatality in the hospital. In their study, they recorded data from over 100,000 patients in the United States, where patients visited the intensive care unit and were admitted to the hospital afterward. The RL algorithm worked to assign a negative score to a patient who died in-hospital and a positive score to someone who was discharged. RL algorithms work with different prescription scenarios; the total score for a given RL scenario is based on the frequency with which a given prescription leads to patients surviving (rewarding path). The best possible scenario is decided by taking all possible dosing strategies into consideration and selecting the sequence of doses that leads to a maximal score from the dataset.

Evaluating Models

After running multiple models for comparison, the next step is to determine model performance for prediction. For regression problems (continuous outcome), the common measurements for model performance are mean square error (MSE) and mean absolute error (MAE).

Mean Squared Error. Mean square error (MSE) is one of the most well-known metrics used for determining prediction accuracy in regression problem. MSE calculates the average squared error between the predicted and actual values where the smaller MSE, the better prediction accuracy. MSE can be calculated as $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)$, where *N* refers to sample size, y_i refers to actual value at position i, and $\hat{y}_i$ refers to predicted value at position i. Sometimes, we take the square root of MSE, also called RMSE, for having the metric with scale as actual value. For example, for predicting gas price, RMSE represents what is the average deviation in the model predicted gas prices from the prices at which the gas is sold for (Willmott & Matsuura, 2005).

Mean Absolute Error. Mean absolute error, also known as mean absolute deviation, is another widely used metric for measuring prediction accuracy in regression problems. MAE is used to find the average absolute distance between the predicted and actual value; the smaller MAE, the better prediction accuracy. MAE has also been found to be more robust to outliers compared to MSE (references cited?). MAE can be calculated as $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$, where *N* refers to sample size, y_i refers to actual value, and $\hat{y}_i$ refers to predicted value, and $|y_i - \hat{y}_i|$ means the absolute difference between actual value at position i and predicted value at position i (Willmott & Matsuura, 2005).

Statistics in the Machine Learning

When it comes to data, the data themselves do not mean anything unless we put them in context and learn the meaning (data mining) behind them. There are several ways to learn from the data including but not limited to data clustering, data classification, regression, prediction, etc. The logic behind choosing each data approach depends on the types of problems/questions we are interested in examining. Regardless of the questions being answered, a common goal that most data mining techniques share is to learn the relationships with the underlying data and be able to express them into valuable and understandable information (Dehuri et al., 2011).

Theoretically, both ML and more traditional statistical models can be used for data mining processes, prediction, and inference. These different approaches have comparable methods that are used to answer business/research questions. The fundamental basis of the model algorithms that are used in ML is also built from a statistical framework with additional combinations of optimization and computer science. The theory behind ML is derived from mathematics and statistics, the algorithmic process stems from optimization, matrix algebra, and calculus, and the implementation phase originates from computer science and engineering concepts (Shu et al., 2013; Stamey et al., 2017). With technology and methodology development, there are several libraries/packages for statistical analysis (i.e., statmodel in Python, dply in R) and machine learning (i.e., sklearn in Python, MICE in R) that can be used to run the analysis. While these packages are helpful and easy to use, most people do not truly understand the concepts and logic behind those libraries, or the differences between statistical models and machine learning. Therefore, it is helpful to highlight the similarities and differences between ML and statistical models to help guide the appropriate approaches to use as a part of data analysis.

The main similarities between ML statistical models are their steps for mining data which are: (a) identifying the types of knowledge being mined, (b) identifying types of databases to work on, and (c) identifying types of techniques to be used. The key difference between ML and statistical modeling is the purpose of their being used (Stewart, 2019). Bzdok et al. (2018) mentioned the primary difference between ML and statistical models is that "Statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns" (p.233). Statistical models are mainly designed to support implications about the relationships among variables in the model. The data inference in statistical models uses mathematical models to validate beliefs or test hypotheses from the data. Statistical models offer an ability to calculate our confidence/belief in quantitative form while we learn about the relationship of the variables in the data (Hastie et al., 2005). Although statistical models can also be used for prediction (i.e., regression), ML is specifically used for prediction rather than explanation. Moreover, statistical models such as regression can apply cross-validation for justifying model accuracy, though they do not require cross-validation. On the other hand, ML relies heavily on cross-validation (split model to train and test model) to validate model accuracy (Boulesteix & Schmid, 2014).

Chapter Summary

To summarize, the detail of statistical methods that can be used to analyze longitudinal data is discussed including repeated measure analysis of variance, hierarchical linear growth modeling, latent growth curve modeling, and Bayesian growth curve models, including the advantages and disadvantages for each. One of the main focuses in this dissertation was to study Bayesian growth modeling; thus, in-depth detail about the components of the Bayesian model and examples are reviewed.

Another aspect of research that has been increasing in popularity in the past decade is machine learning. The well-known benefits for machine learning are its good performance under ongoing large and complex datasets, ability to identify patterns and trends that might not be apparent to a human, and ability to automate methods for data analysis (De Raedt & Kimmig, 2015). The details of ML that are discussed in this chapter include types of machine learning, roadmap for building machine learning, evaluating prediction accuracy in machine learning, and sample size in machine learning.

As mentioned earlier, machine learning has been shown to perform well with large sample size; however, not all fields of research have the luxury of having large datasets, even though researchers in these fields might be interested in using machine learning. In reviewing the literature about machine learning, one thing that stood out is the limited research about using machine learning with small sample size (Byrd et al., 2012). Moreover, research conducted on Bayesian growth curve modeling in the machine learning environment is limited (Depaoli, Rus, et al., 2017). As mentioned in the above section about interaction between priors and other data conditions, the limitations for the existing Bayesian inference research in longitudinal data are unknown interaction effect of priors and sample size in real life data and other types of models beside MIMIC models (Brutti et al., 2008; Finch & Miller, 2019).

Consequently, this dissertation study incorporates the concept of Bayesian growth modeling in hierarchical linear modeling within a machine learning environment, particularly looking at the different conditions of sample size, waves of data, and proportion of cases in the two levels of a dichotomous time invariant predictor, for which the simulated data were based on real data. Understanding the above conditions can help to guide applied researchers on how to make decisions regarding what underlying conditions should be used when adopting Bayesian

growth modeling in machine learning to yield acceptable model accuracy (Wang & Preacher, 2015). The following Chapter 3 presents details of how to conduct the proposed study to answer the research questions listed in Chapter 1.

CHAPTER III

METHODS

To validate the moderator effect of priors on the effects of different data conditions (sample size, waves of data, and proportion of cases in a dichotomous time-invariant predictor) on prediction accuracy of Bayesian growth modeling within a machine learning environment, a Monte Carlo simulation was adopted. Monte Carlo simulation is an appropriate tool in this dissertation study because it helps researchers to study theoretical outcomes of statistical properties (i.e., prediction accuracy, parameter estimate bias, standard errors) under different conditions from randomly generated and experimentally manipulated data that are not easily examined through "real data" (Graham & Talay, 2013). The contents of this chapter include description of the Monte Carlo simulation processes and Bayesian growth modeling analysis in a machine learning environment. A detailed description of the simulation processes was broken into research design, data generation, outcome and independent variables, procedures, and data analysis.

The following are the research questions that are used to set the direction of the proposed study.

Q1 Do the types of prior (informative and noninformative priors) moderate the effect of sample size on predictive accuracy for Bayesian growth modeling in a machine learning environment?

Q2 Do the types of prior (informative and noninformative priors) moderate the effect of number of waves of data on prediction accuracy for Bayesian growth modeling in a machine learning environment?

Q3 Do the types of prior (informative and noninformative priors) moderate the effect of proportion of cases in the two levels of a dichotomous time-invariant predictor on prediction accuracy for Bayesian growth modeling in a machine learning environment?

Model Parameters

This dissertation study used synthetic data, which offered an advantage in terms of allowing researchers to know the correct values of the parameters and check whether those parameters could recover with the hypothesized models under varying data conditions (Martin, 2018). To set the model parameters close to "real life" data, the parameters of the data in this dissertation study were based on alumni donation data from a university in the mid-Atlantic region. The actual data helped to guide selection of the possible values in each wave of data, information about the growth trajectory, ratio of cases in the two levels of a dichotomous time-invariant predictor, correlations among variables, and prior distributions.

The aspect of the alumni donation that was used as the guideline to set the model parameters in the proposed study was prediction of donors' donation amount for the next coming year, which was used to recommend the amount of donation suggested when the development officers contact the donors. In philanthropy, there are several predictors that can contribute to donation decisions for alumni, including, but not limited to, engagement (i.e., does the alumna/us come to university events) and communication (i.e., how often the university contacts the alumni; Sargeant, 2013; Thomas et al., 2015). Using the above situation as the guideline for data simulation in this dissertation study, the following was the structure of the data design:

(a) Existing alumni donors: Alumni who had donation records in the last five years (2015 - 2019).

(b) Donation history: This variable reflected five years of donation amount.

(c) Slope and intercept: A linear regression was fitted to each year's donation amount (dependent variable) and number of contacts in the last five years (independent variable) to obtain the intercept and slope for each donation year. The results showed the following: year 2015 had an intercept of $120 and slope of $700, year 2016 had an intercept of $140 and slope of $888; year 2017 had an intercept of $168 and slope of $900; year 2018 had an intercept of $192 and slope of $1,221; and year 2019 had an intercept of $210 and slope of $1,450. The information about the slopes and intercepts was incorporated into the prior knowledge of the Bayesian growth modeling model. These data were used as general guidelines for specifying prior distributions.

(d) Donation amount: The range of common donation amounts each year (continuous values). These values used were to guide to donation simulation values to reflect the real proportion donations.

(e) Donation distribution: In the real data, the donation distribution was right-skewed (mean > mode) toward less than $1,000. The details about the specific characteristics of the distributions that were generated in the Bayesian growth model are discussed in the model specification section.

(f) Number of contacts: Number of contacts that development officers had made to alumni in the last five years were used as a time-invariant predictor of growth in donation amounts, with the number of contacts ranging from 0 to 63, with a mean of 20 (*SD* = 12).

(g) Engagement with the university: Level of engagement with the university was defined as attending an event held by the university (1 = yes, 0 = no) and the ratio of alumni who attended versus did not attend the event (25:75). This ratio was one of the data conditions in proportion of cases in the levels of a binary predictor.

Table 1 is the correlation matrix of donation amounts from 2015 – 2019 and the number of contacts in the last five years.

Table 1

Correlation of Donation Amount from 2016-2019 and Number of Contacts

Variable	1	2	3	4	5	6
Year 2019	1.00					
Year 2018	.08	1.00				
Year 2017	.11	.03	1.00			
Year 2016	.08	.05	.31	1.00		
Year 2015	.12	.11	.30	.28	1.00	
Number of Contacts	.59	.47	.42	.42	.41	1.00

Population Growth Model

A first decision to make before conducting Bayesian growth modeling is choosing the population for the model. The Bayesian growth model in this study is viewed as a two-level hierarchical model with time nested within individual, where level-1 is the person level and level-2 is time nested within individuals. As mentioned in the model parameter section, this dissertation study used synthetic data for which the model parameters are based on alumni donation data from a university in the mid-Atlantic region in order to create simulated data to reflect real life as much as possible. The purpose of Bayesian growth modeling in this study was to predict the upcoming donation amount (continuous variables) based on the growth donation amount rate in the last five years (2015 to 2019), using number of contacts (continuous variable) and whether alumni attended any events held by the university after they graduated (dichotomous variable of yes = 1, no = 0) to explain the variation of growth in donation rate over time. Therefore, the dependent variable is the history of donation amount in the last five years

(2015- 2019), while the independent variables are number of contacts and whether alumni attended any events held by the university. To control the complication of the model, both number of contacts and event attendance variables (time-invariant variables) were specified to be the same within donors for all five years but different across donors. The level-1 (donor level) of the two-level of Bayesian growth modeling can be written as:

$$Y_{i,t} \sim N\left(\beta_{i,1} + \beta_{i,2}T_t, \sigma^2_{eLevel1}\right) \quad (11)$$

where $Y_{i,t}$ represents donation amounts by donor *i* at measurement occasion T_t. $\beta_{1,i}$ represents the predicted donation amount of donor *i* when $t = year\ 2015$ (intercept), $\beta_{i,2}$ represents the rate of change in donation amounts of donor *i* for one-unit change in T_t (slope). $\sigma^2_{eLevel1}$is the time-specific residual score at time t for individual *i.* To integrate Bayesian inference into the hierarchical growth model, the component of slope and intercept must be described in distribution form. Both slope and intercept are considered as random, which means that each donor has his/her own starting point for donation amount and his/her donation amount trajectory can be different from other donors. Moreover, the slope and intercept elements are chosen to be normal, tilde (∼) denotes "distributed as." Although the nature of the donation distribution is right skewed toward donation amounts of less than $1,000; the distribution in this study was specified to be normally distributed to control the complication of the model (Kosugi, 1996). Osvaldo et al. (2017) also recommended that normal distribution results in better prediction accuracy.

The simulation data in this study were designed to result in different variations of donation amount from year 2015 through 2019; thus, communication in term of number of contacts that alumni receive from the university and alumni engagement in terms of whether alumni come to the events the university holds are the common predictors added to explain the

variation of year over year donation amount (Sargeant, 2013; Thomas et al., 2015). Therefore, the level-1 model can be extended to a level-2 model:

$$\beta_{i,1} \sim N(\beta_{01} + \beta_{11}\,\#contact_{donor\,i} + \beta_{21}\,attend\ event_{donor\,i} + \sigma^2_{e\beta 1}) \quad (12)$$

$$\beta_{i,2} \sim N(\beta_{02} + \beta_{12}\,\#contact_{donor\,i} + \beta_{22}\,attend\ event_{donor\,i} + \sigma^2_{e\beta 2}). \quad (13)$$

Equation 12 is the between-person variability in the intercept and Equation 13 is the between-person variability in the slope with the number of contacts and university-based event attendance included in the model. The $\sigma^2_{e\beta 1}$ and $\sigma^2_{e\beta 2}$ components represent the magnitude of between-person differences (residuals) in the intercept ($\sigma^2_{e\beta 1}$) and slope ($\sigma^2_{e\beta 2}$) that are not explained by the time-invariant covariates. The final two-level hierarchical model is specified as follows:

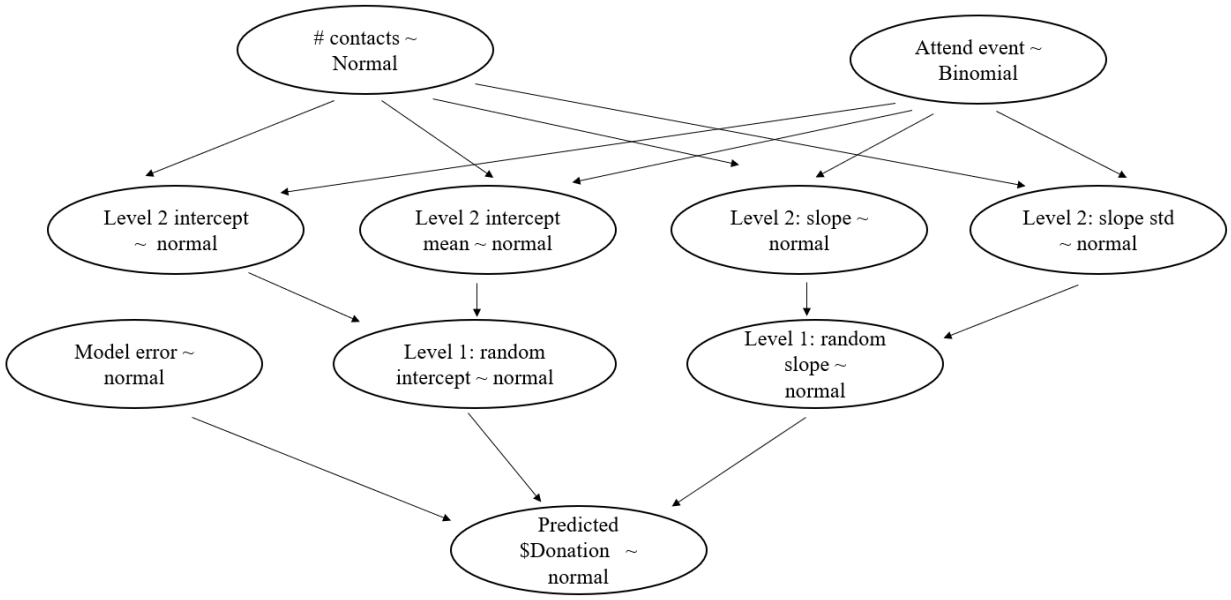
$$Y_{i,t} \sim N\left(\beta_{0i} + \beta_{1i}(year) + \beta_{2i}(\#contacts) + \beta_{2i}\,(attend\ event)\,, \sigma^2_{e\beta 12}\right) \quad (14)$$

The in-depth details of level-1 and level-2 equations as implemented in the proposed study are discussed in the model specification section.

Once the population growth model was specified in the form of a Bayesian growth model, the next step was to specify the distribution of each model parameter to prepare for running the model in probabilistic machine learning software, which was PyMC3, a library in Python (version 3.7.0). Along with PyMC3, ArviZ, a Python library that works collectively with PyMC3, was employed to facilitate the visualization and interpretation of the posterior distributions. The further details about PyMC3 are discussed in the Bayesian growth modeling in machine learning section. For the above Bayesian growth model, the prior distribution can be specified in PyMC3 as the workflow shown in Figure 2:

Figure 2

Distribution Specification for Bayesian Growth Model in PyMC3



After the population for the growth model was set, the next step was to integrate different data conditions into the model in order to examine whether different data conditions have different impacts on predictive accuracy in Bayesian growth modeling in a machine learning environment. In the following section I discuss the data conditions used to answer the research questions.

Design Factors

The four sets of sample size (N = 25, 50, 100, and 200), three waves of data (T = 3, 4, and 5), three different proportions of dichotomous time-invariant predictors (90:10, 75:25, 50:50), and two prior distributions (non-informative and informative prior) make up a $4\ x\ 3x\ 3\ x2 = 72$ condition design. Justification for these levels of the manipulated conditions is provided below.

Data Generation

Once the model parameters were set, the Monte Carlo simulation procedure was implemented using Python (version 3.7.0) and specifically the following routines within Python: random, scipy.stat, numpy, panda, and time library. The random.normal, random.choice, numpy, and multivariate_normal procedures within scipy.stat were particularly used in the data generation steps. The numpy function was used to specify an array of correlation matrices among waves of data. The multivariate_normal routine within scipy.stat was used to create the multivariate normal distribution for each wave of data with the specific correlation specified in the numpy array function.

Sample Size

Four sample size conditions were considered: N = 25, 50, 100, and 200. These values of sample size were selected after reviewing several applied and simulation studies regarding Bayesian growth curve modeling in both non-machine learning and machine learning environments (Oravecz & Muthén, 2018; Rasmussen & Ghahramani, 2003; Schwaighofer et al., 2005; Shi & Tong, 2017; White et al., 2018). The sample size of N = 25 was selected to reflect the interest of researchers who do not have access to large sample size but are attracted to applied Bayesian growth curve modeling in a machine learning environment. For example, the simulation study from Su et al. (2008), who only focused on performance of Bayesian statistics in small sample size (limited up to 50), suggested that Bayesian performance in terms of model accuracy is stable when the sample size reaches 25 observations. The sample size of 50 was selected based on the studies from Shu et al. (2013) and Stamey et al. (2017), who suggested that in order to use Bayesian modeling for longitudinal prediction in machine learning, the average sample size of around 50 observations in the training set is sufficient to reach sufficient model

accuracy. The sample sizes of 100 and 200 were selected to represent moderate sample size used in prediction problems in machine learning (Beleites et al., 2013; Stockwell & Peterson, 2002; Vabalas et al., 2019).

Waves of Data

It is well understood that two waves of data (i.e., pretest and posttest) is the minimum requirement for estimating change in longitudinal data analysis (Curran et al., 2010; Meredith & Tisak, 1990); however, the information relating to change occurring overtime can be limited if only two waves of data are collected (Rogosa et al., 1982). Hence, both applied and simulation longitudinal research commonly uses more than two waves of data to detect effect of change over time (Oravecz & Muthén, 2018; Shi & Tong, 2017; Willett, 1989). Moreover, the study from Willett (1989) suggested that the model reliability is increased up to 250% when using three waves of data instead of two. Consequently, three levels of waves of data were chosen, T = 3, 4, and 5, in this current dissertation, based on literature examining waves of data used in Bayesian growth modeling that indicated four waves of data have been found to be beneficial for longitudinal modeling (Curran, 2003; Oravecz & Muthén, 2018; Zondervan-Zwijnenburg et al., 2017).

Proportion of Dichotomous Predictor

It is common in the real-world prediction problem that the proportion of cases in each level of a dichotomous predictor is unequal (Dixon et al., 2009). Examples include the number of patients who receive versus do not receive a treatment; the number of clients who own versus do not own a certain product; and the number of alumni who stay engaged versus do not stay engaged with the university in philanthropy. The common ratios of subjects in the dichotomous predictor that were used in the current Monte Carlo simulation study are 10:90, 25:75, and 50:50

(Shaw & Mitchell-Olds, 1993; Zahn, 2010). Since the ratio of the cases per level of the dichotomous event attendance variable in the alumni donation data used to guide the model parameters is 25:75, three different ratios of the dichotomous time-invariant predictor chosen for this dissertation study are 10:90, 25:75, and 50:50. Here the first part of the ratio represents the proportion of cases coded as 1 (Yes, attends an event) and the second part of the ratio represents the proportion of cases coded 0 (No, does not attend an event).

Priors

In Bayesian statistics, previous knowledge about the model parameters can be combined in the analysis for estimating the values of unknown parameters (posterior distribution), where the prior needs to be assigned on the model parameter as a distribution, inferred as a distribution displaying the probability of parameter values. The additional knowledge in a prior distribution can help to increase statistical power when the sample size is small and serves as mutual knowledge in the field of study (Depaoli, Rus, et al., 2017). In this current study, informative and non-informative priors were used for comparison purpose when estimating prediction accuracy. To recap the definition of priors, non-informative priors mean the prior distribution contains little explanatory information about the unknown parameters or hypotheses related to the model, while informative priors mean the prior distribution comprises existing information, knowledge, or a hypothesized parameter distribution associated with the model before the data are collected (Bolstad, 2007; Golchi, 2018).

The informative priors in this study were derived from the average value of the intercept and slope after running separate linear regression models, described above, to predict each year's donation amount (from year 2015 through 2019) using number of contacts in the last five years as the predictor. The average intercept value is \$177 (*SD* = \$30), while the average slope value is

$1,093 (*SD* = $21). As mentioned above, the prior follows a normal distribution; therefore the informative prior distribution of slope and intercept of donation amount can be written as:

$$p(Intercept) \sim N\ (177, 30)$$

$$p(slope) \sim N\ (1093, 21)$$

For the prior distribution in a non-informative prior, in the general case, the prior is specified to cover as wide a range of parameter values as possible. Most of the time when precise information about the model parameter is not applied, the permissible parameter space is used to set the boundary of the non-informative prior. For example, a prior distribution for a mean could exclude values that are outside the range of the measurement scale (Depaoli, 2014; Depaoli, Yang, & Felt, 2017). From the alumni donation data that were used as the guideline to set the model parameters in the current study, the lowest value of average donation amount is $85 and the highest value of average donation amount is $8,548. Thus, to make the distribution of non-informative priors to cover as wide a range of parameter values as possible, the non-informative prior distribution of slope and intercept of donation amount can be written as:

$$p(Intercept) \sim uniform\ (0, 8548)$$

$$p(slope) \sim uniform\ (0, 85)$$

To summarize, the conditions representing four different sets of sample size, three waves of data, three different proportions of dichotomous outcome, and two prior distributions (non-informative, and informative prior) were crossed. Table 2 represents the design for the Bayesian growth model that was used to run the analysis in machine learning. The four-digit numbers (i.e., 1111, 1112, etc.) represent the level of each factor.

Table 2

The 4 x 3 x 3 x 2 Factorial Design for Bayesian Growth Model

			Sample Size			
Wave of Data			1	2	3	4
			Factor Level Combination			
3 = 3	1 = 10:90	1= Informative priors	3111	3112	3113	3114
3 = 3	2 = 25:75	2= Non-informative priors	3221	3222	3223	3224
3 = 3	3 = 50:50	1= Informative priors	3311	3312	3313	3314
3 = 3	1 = 10:90	2= Non-informative priors	3121	3122	3123	3124
3 = 3	2 = 25:75	1= Informative priors	3211	3212	3213	3214
3 = 3	3 = 50:50	2= Non-informative priors	3321	3322	3323	3324
4 = 4	1 = 10:90	1= Informative priors	4111	4112	4113	4114
4 = 4	2 = 25:75	2= Non-informative priors	4221	4222	4223	4224
4 = 4	3 = 50:50	1= Informative priors	4311	4312	4313	4314
4 = 4	1 = 10:90	2= Non-informative priors	4121	4122	4123	4124
4 = 4	2 = 25:75	1= Informative priors	4211	4212	4213	4214
4 = 4	3 = 50:50	2= Non-informative priors	4321	4322	4323	4324
5 = 5	1 = 10:90	1= Informative priors	5111	5112	5113	5114
5 = 5	2 = 25:75	2= Non-informative priors	5221	5222	5223	5224
5 = 5	3 = 50:50	1= Informative priors	5311	5312	5313	5314
5 = 5	1 = 10:90	2= Non-informative priors	5121	5122	5123	5124
5 = 5	2 = 25:75	1= Informative priors	5211	5212	5213	5214
5 = 5	3 = 50:50	2= Non-informative priors	5321	5322	5323	5324

Note. Sample sizes: (1) $N = 25$, (2) $N = 50$, (3) $N = 100$, (4) $N = 200$; waves of data (3) $T = 3$, (4) $T = 4$, (5) $T = 5$; proportions of cases per level of the dichotomous predictor (1) 10:90, (2) 25:75, (3) 50:50; prior distributions (1) non-informative, (2) informative prior.

Model Specification

Before applying Bayesian growth modeling in machine learning, details of the model need to be specified. As mentioned in the above section about model parameters, the data structure that was used as a guideline for data simulation was based on university alumni donation data, which were used to try to predict the upcoming year's donation amount. Since growth curve modeling can be viewed as a hierarchical (or multilevel) linear regression model, a two-level growth curve model can be applied. The outcome variable is the history of donation amount in the last five years (2015 - 2019). Two predictors (time-invariant covariates) that were included as factors to account for donation amount are (1) number of contacts (continuous value) from the development officer in the last years (x_1), and (2) attendance at events held by the university (0 = no, 1 = yes). The time-invariant covariates in this study help to explain whether the variability in the initial donation amounts (intercept) and change in the donation amount (slope) depend on the above two predictors.

In this study, the longitudinal data set had five ($T = 5$) measurements at occasions t, in this case year 2015-2019 ($t = 2015, 2016, 2017, 2018, 2019$); one measure each year, from an individual i, where i = ($i = 1, \dots, N$), with N representing the total number of donors in the sample. The time points are considered to be fixed across donors; however, each donor has his/her own characteristics (i.e., income, occupation, number of children). The donation amount for each donor in each year can be different from each other and can be looked at as a random intercept and random time. The measure of donation amount by donor *i* at measurement occasion *t* is denoted by $Y_{i,t}$. The within-person change over time can be articulated as initial donation amount in year 2015 (intercept) and rate of change in donation amount over time (slope). Thus, a straight line was fitted to each donor's four measurements, with x -axis (independent variable)

being the time, and y-axis (dependent variable) being the donation amounts. The two-level growth curve model can be broken down to level-1 model (Equation 15) and level-2 model (Equations 16 and 17). The level 1 model can be specified as follows:

$$Y_{i,t}|\beta_{i,1}, \beta_{i,2} \sim N\left(\beta_{i,1} + \beta_{i,2}T_t, \sigma^2_{eLevel1}\right). \quad (15)$$

The level-1 is described as the time effect at the donor (person) level, where $Y_{i,t}$ refers to donation amounts by donor *i* at measurement occasion *t*. $\beta_{1,i}$ represents the predicted donation amount of donor *i* when $t = year\ 2015$ (random intercept parameter), and $\beta_{i,2}$ represents the rate of change in donation amounts of donor *i* for one-unit change in t (random slope parameter). $\sigma^2_{eLevel1}$ is the time-specific residual score at time t for individual *i*.

For the level-1 equation, the mean of the donor's donation amount. $Y_{i,t}$ is a function of a donor *i*'s intercept parameter $\beta_{1,i}$ and the product between donor *i*'s slope parameter and the measurement occasion T_t at *t*. Consequently, the level-1 equation shows the conditional distribution of $Y_{i,t}$ given $\beta_{1,i}$ and $\beta_{2,i}$.

As mentioned, the distributional form of Equation 11 was selected to be normal, with the time-specific residuals having variance $\sigma^2_{eLevel1}$. The $\sigma^2_{eLevel1}$term can be marginalized by adding the parameter to account for random errors -- the time-dependency (time-invariant) in the mean and variance that predicts person-specific change, which in this case is based on number of contacts and attendance at events held by the university. Although the value of number of contacts and university-related event attendance can change over time, for the proposed study, the two simulated predictors were treated as time-invariant. Consequently, the parameter value estimated was assumed to be constant over time (McCoach & Kaniskan, 2010). For example, at every time point, the number of contacts that donors received in the last five years was held constant for each donor across all measurement years but varied across donors.

Thus, the model can be extended at the between-person level (level-2), for which the level-2 growth curve model can be specified as follows:

$$\beta_{i,1} \sim N(\beta_{01} + \beta_{11} x_{1i} + \beta_{21} x_{2i}, \sigma^2_{e\beta 1}), \qquad (16)$$

$$\beta_{i,2} \sim N(\beta_{02} + \beta_{12} x_{1i} + \beta_{22} x_{2i}, \sigma^2_{e\beta 2}). \qquad (17)$$

Equations 16 and 17 describe between-person variability in initial levels (intercepts) and rates of change (slopes). β_{01} and β_{02} are group-level parameters shared across donors showing the expected intercept (β_{01}) and slope (β_{02}) with the time-invariant covariates, number of contacts (x_{1i}) and university-based event attendance (x_{2i}), the predictor equal 0. β_{11} and β_{12} are the regression parameters representing the relation between the time-invariant covariates and the person-level intercept (β_{11}) and slope (β_{12}). The $\sigma^2_{e\beta 1}$ and $\sigma^2_{e\beta 2}$ components represent the magnitude of between-person differences (residuals) in the intercept ($\sigma^2_{e\beta 1}$) and slope ($\sigma^2_{e\beta 2}$) that are not explained by the time-invariant covariant. The level-2 parameters usually remain in the unstandardized form and are interpreted in the same manner as a standard regression model (Grimm et al., 2016). In this current study, the level-2 parameters can be interpreted as the expected difference in the donation amounts for a one-unit difference in number of contacts and donors' attendance (or non-attendance) at events held by the university.

In Equations 16 and 17, level-2 distributions on the intercept ($\beta_{i,1}$) and slope $(\beta_{i,2})$ are specified univariately. However, those terms in the intercept and slope can co-vary and the person-specific intercept and slope can be specified bivariately where the bivariate lognormal population hyperprior distributions are set on these parameters. Thus, Equations 16 and 17 can be rewritten as the following:

$$\begin{bmatrix}\beta_{i,1}\\ \beta_{i,2}\end{bmatrix} \sim N_2 \left(\begin{bmatrix}\beta_{01} + \beta_{11} x_{1i} + \beta_{21} x_{2i}\\ \beta_{02} + \beta_{12} x_{1i} + \beta_{22} x_{2i}\end{bmatrix}, \begin{bmatrix}\sigma^2_{e\beta 1} & \sigma^2_{e\beta 12}\\ \sigma^2_{e\beta 21} & \sigma^2_{e\beta 2}\end{bmatrix} \right) \qquad (18)$$

The mean vector of the bivariate distribution in Equation 18 reflects the function of regression coefficients and time-invariant covariates in Equations 16 and 17. The difference in Equation 18 and Equations 16 and 17 is that the elements of variation around the bivariate mean ($\sigma^2_{e\beta 1}\ and\ \sigma^2_{e\beta 2}$) are expressed in terms of a covariance matrix in Equation 18, instead of listed separately like in Equations 16 and 17. $\sigma^2_{e\beta 12}$ indicates covariation between person-specific intercepts ($\sigma^2_{e\beta 1}$) and slopes ($\sigma^2_{e\beta 2}$) which are the variances of these terms. Moreover, the population-level correlation between intercepts and slopes can be calculated by dividing the covariance with the product of the standard deviations. Now the level-1 and level-2 equations can be combined as

$$Y_{i,t} \sim N \begin{pmatrix} \beta_{0i} + \beta_{1i}(year) + \beta_{2i}(\#contacts) + \beta_{2i}(attend\ event), \\ \sigma^2_{e(level1)+} T_t^2 \sigma^2_{e\beta 2} + \sigma^2_{e\beta 1} + T_t \sigma^2_{e\beta 12} \end{pmatrix} \quad (19)$$

Once the model equations were laid out in terms of the growth model, the next step was to fit Bayesian inference into the growth model and specify priors to all model parameters. As mentioned in the priors section above, the model parameters were set based on the alumni donations data, for which the intercept and slope for informative priors of the donation amounts are:

$$p(Intercept) \sim normal\ (177, 30),$$

$$p(slope) \sim normal\ (1{,}093, 21)$$

The intercept and slope for non-informative priors of the donation amounts are:

$$p(Intercept) \sim uniform\ (0, 85),$$

$$p(slope) \sim uniform\ (0, 8548)$$

For the first time-invariant covariate, number of contacts, the information retrieved from the alumni donor data shows that the distribution is right skewed toward fewer than 30 contacts with mean of 20 and standard deviation of 12; thus, the distribution for priors was specified to a log normal distribution to bring the distribution to normal. The informative prior distribution for number of contacts is:

$$p(\#contact) \sim normal\ (20, 12)$$

The non-informative prior distribution for number of contacts is:

$$p(\#contact) \sim uniform\ (0, 100)$$

For the second time-invariant covariate, whether alumni attend any events held by the university (yes = 1, no = 0), the proportion for attendance versus non-attendance is .25: .75. Since the event attendance variable is dichotomous, the informative prior distribution can be described as a binomial distribution as follow:

$$p(attend\ event) \sim binomial\ (\ model\ sample\ size, .25)$$

The non-informative prior distribution for event attendance is:

$$p(attend\ event) \sim binomial\ (model\ sample\ size, .40)$$

For the error term, the standard deviations of the intercept and slope, and the correlation of these two terms for informative priors of the donation amounts are based on the values of the parameters in the alumni donation data which are:

$$\sigma^2_{eLevel2} \sim gamma\ (0, 10)$$

$$\sigma^2_{e\beta 1} \sim gamma\ (0, 10\ )$$

$$\sigma^2_{e\beta 2} \sim gamma\ (0, 10\ )$$

$$\rho_{e\beta 12} \sim gamma\ (0, 10)$$

The standard deviations of and the correlation between intercept and slope for non-informative priors of the donation amounts were set as uniform distributions to cover most of the possible values for the error terms, which are:

$$\sigma^2_{eLevel2} \sim uniform\ (0, 1000\ )$$

$$\sigma^2_{e\beta 1} \sim uniform\ (0, 1000\ )$$

$$\sigma^2_{e\beta 2} \sim uniform\ (0, 1000\ )$$

$$\rho_{e\beta 12}\ \sim uniform\ (0, 1000)$$

Once the priors were specified, the next step was to calculate the posterior distribution based on the specified data. The Bayesian growth curve model in a machine learning environment in this dissertation study was estimated using a built-in library in Python, for which the details about the analysis steps are discussed in the Bayesian growth modeling in machine leaning section below.

Number of Replications

To be consistent with other simulation designs for prediction accuracy in Bayesian growth modeling (Depaoli, 2104; Depaoli, Yang, & Felt, 2017; Shi & Tong, 2017), in each design condition, 1,000 replications were used in the current study.

Bayesian Model in Machine Learning

The analysis of Bayesian growth modeling in a machine learning environment for this current dissertation study was performed through PyMC version 3 (PyMC3), a popular built-in library for probabilistic machine learning in Python (version 3.7.0) to analyze Bayesian models. PyMC was developed in 2003 by Christopher Fonnesbeck, an associate professor in the Department of Biostatistics at Vanderbilt University in Nashville, Tennessee. PyMC was created as part of a Python module to fit Bayesian models and simplify the process of constructing

Metropolis-Hastings samplers, with a purpose to make Markov chain Monte Carlo (MCMC) easier to use for applied researchers. The development team of PyMC has been adding PyMC with new features and software iterations in order to offer more flexibility and better performance. Consequently, the latest version of PyMC, which is version 3, was released in January 2017 (Martin, 2018). The new features in PyMC version 3, including Gradient-based MCMC methods, Hamiltonian Monte Carlo (HMC), the No U-turn Sampler (NUTS), and Stein Variational Gradient Descent, perform well with high dimensional and complex posterior distributions without requiring any specialized knowledge about fitting algorithms. In turn, it is intuitive and simple to use. Along with PyMC3, ArviZ, a Python library that works together with PyMC3, was used in order to help with the visualization and interpretation of the posterior distributions. Another important library that helps to enhance the performance of PyMC3 is Theano. Theano helps to define, improve, and calculate mathematical expressions involving multidimensional arrays effectively (Salvatier et al., 2016).

Bayesian Growth Modeling in PyMC3

To recap, the data structure used for the data simulation in this study was based on the alumni donation data (year 2015 -2019) from a university in the mid-Atlantic region. Bayesian growth modeling in machine learning was applied to predict the donation amount for an upcoming year based on the variation in donation amounts from years 2015 -2019, and time-invariant covariates (number of contacts in the last five years and university-based event attendance). Since the Bayesian model is a probabilistic model, which is a tool to measure uncertainty, the outcome for the model is the distribution; not the point estimation (Oravecz & Muthén, 2018). For example, instead of the prediction for the donor *i* in X dollar amounts, the prediction is in probability distribution form and can be interpreted as with 95% probability,

indicating the average donation amount for donor *i* is between X dollar amounts and X dollar amounts.

To apply Bayesian growth modeling in PyMC3, the following steps were followed:

Step 1: Import the needed libraries to run the model, which were pandas that gave the data frame structure (i.e., import, export data);

- Numpy: helps with scientific computing (i.e., multidimensional array).
- Matplotlib.pyplot: plots a numerical mathematic extension of numpy.
- Scipy: helps with scientific programming (i.e., linear algebra, integration for calculus).
- Arviz: helps with visualization and interpretation inference problems.
- PyMC3; helps with probabilistic programming.
- Seaborn: helps with statistical data visualization.
- Sklearn: offers features of statistical models (i.e., classification, regression, clustering algorithms).
- Theano: helps with calculating mathematical expressions with multidimensional arrays (Salvatier et al., 2016).

Step 2: Import data, run descriptive statistics, and plot data to see how the donation growth rate changes over time to ensure that the simulated data come out as expected.

Step 3: Specify the likelihood and priors using probability distributions. As mentioned in the previous section, two sets of priors (informative and non-informative) were used, where an informative prior was derived from the average value of the intercept and slope after running separate linear regression models to predict each year's donation amount (from year 2015 to 2019) using number of contacts in the last five years as the predictor. To build a two-level growth curve model (time nested within individual) in PyMC3, prior distributions were specified

in each level, which can be broken down into level-1 intercept (α), level-1 slope (β), level-2 mean for intercept (α_μ), level-2 standard deviation for intercept (α_σ), level-2 mean for slope (β_μ), and level-2 standard deviation for slope (β_σ ; Martin, 2018). For informative priors in level-1 (person level):

- Prior distribution for intercept: $\alpha \sim Gamma\ (0, 2)$
- Prior distribution for slope $\beta \sim normal(0, 2)$

For informative prior in level-2 (between person level):

- Mean for intercept defined as α_μ $\sim normal\ (0, 2)$
- Standard deviation for intercept defined as α_μ $\sim Gamma\ (0, 2)$
- Mean for slope defined as $\beta_{\mu} \sim normal\ (0, 2)$
- Standard deviation for intercept defined as β_μ $\sim Gamma\ (0, 2)$

For noninformative prior in level-1 (person level):

- Prior distribution for intercept: $\alpha \sim uniform\ (0{,}100)$
- Prior distribution for slope $\beta \sim uniform\ (0, 100)$

For noninformative prior in level-2 (between person level):

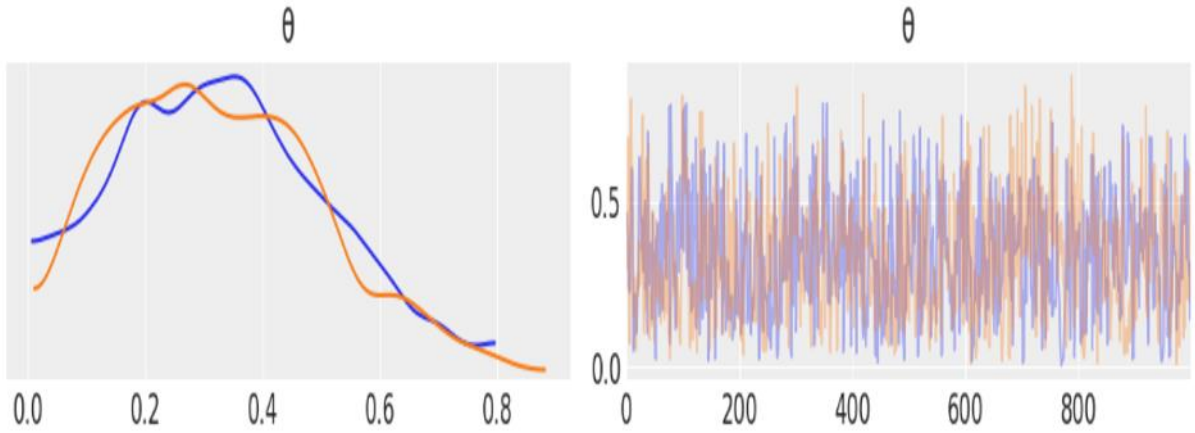
- Mean for intercept defined as $\alpha_{\mu} \sim uniform\ (0, 100)$
- Standard deviation for intercept defined as α_μ $\sim uniform\ (0, 100)$
- Mean for slope defined as $\beta_{\mu} \sim uniform\ (0, 100)$
- Standard deviation for intercept defined as β_μ $\sim uniform\ (0{,}100)$

Step 4: Once the priors were specified, the predicted dependent variable (y_pred) can be calculated by specifying the distribution for $Y_{i,t}$ and the above priors combined into the model.

Step 6: Plot the inference by specifying the number of samples (i.e., 2,000) from the posterior using MCMC sampling method. After sampling from the posterior, the ArviZ library was used to trace the posterior, which helps to summarize the posterior distribution. PyMC3 returns the posterior samples in a trace object, which looks similar to the following Figure 3:

Figure 3

Example of the Posterior Sample in a Trace Object



In Figure 3, the left plot is a Kernel Density Estimation (KDE) plot that shows the plausible value from the posterior. The right plot is the individually sampled values at each step during the sampling. The numerical summary of the trace can also be obtained.

Summarizing the Posterior Distribution

In Bayesian analysis, the result for the model is the posterior distribution, which contains all the information about the parameter given a dataset. Therefore, reviewing the posterior is equivalent to reviewing the logical results of a model. Moreover, the posterior distribution comes in the form of the plausible values given the data in the model instead of a single value. The common summarization values for a posterior distribution are the mean (or median or mode), to inform about the location of the distribution and standard deviation (*sd*) to inform about how far

away the values are from the mean. Or one can think of the standard deviation as uncertainty in the estimation. One thing to keep in mind is the standard deviation only informs the credible information for a normal-like distribution. If the distribution is non-normal (i.e., skewed), the standard deviation can be misleading.

When the distribution is non-normal, the common measure to summarize the spread of a posterior distribution is the Highest- Posterior Density (HPD) interval. HPD is the shortest interval covering a given part of the probability density. The common value used for reporting HPD is 95%. For example, if the result from a model has the 95% value of a HPD interval of 2-5, it means that corresponding to the model, the parameter number is between 2 and 5 with a probability of .95 (Osvaldo et al., 2017). The proposed dissertation used the ArviZ library in Python to help with visualization in the posterior distribution. ArviZ library computes and reports probability values of a .94 (94%) confidence interval instead of .95 (95%) level. For example, the default report value for the HPD interval using Arviz is .94.

To recall, the posterior in general Bayesian inference is $p(\theta|y)$ and describes a distribution of the parameters in a model conditioned on the observed samples. However, when the purpose of the model is for prediction the $p(\theta|y)$ can be used to generate prediction, $\tilde{y}$, based on the data, y, and the estimated parameter θ. The posterior predictive distribution is a distribution of the predicted samples and the posterior predictive distribution can be derived from the following equation:

$$p(\hat{y}|\theta) = \int p(\hat{y}|\theta)\, p(\theta|y) d\theta\,. \qquad (20)$$

Equation 20 indicates that the posterior predictive distribution comes from an average of conditional predictions over the posterior distribution of θ. Theoretically, Equation 20 is the integral of an iterative two-step procedure, which includes (1) obtaining sample values of θ from

the posterior $p(\theta|y)$ and (2) feeding the value of θ from step one to the likelihood, which is how data are introduced in the analysis (based on the chosen sample distribution, e.g., normal distribution). Then we attain the ỹ prediction. The prediction ỹ value can be used to compare to the value of observed data, y, in order to detect the difference between observed and predicted values. This process is known as posterior prediction checks or model validation. The major objective for comparing the observed versus predicted data values is for auto-consistency, i.e., the predicted data should look similar to the observed data. If the values of observed and predicted data are drastically different, it means that there might be some issue during the modeling or a problem while feeding data to the model. Model validation helps to make sure that we specify the model correctly, have a better understanding of model limitations, and gain the insight of how to improve the model (Osvaldo et al., 2017).

Dependent Variables

To access model performance with PyMC3, the widely applicable information criterion (WAIC), leave-one-out cross-validation (LOO), standard error of WAIC, and standard error of LOO were used in this dissertation study. Both WAIC and LOO are commonly used as cross-validation methods for assessing model accuracy in Bayesian machine learning (Gelman, Lee, & Guo, 2015; Vehtari et al., 2017).

WAIC indicates a fully Bayesian approach to evaluate the out-of-sample expectation employing the computed log pointwise posterior predictive density along with rectifying for the effective number of parameters to accommodate for overfitting (Watanabe, 2013). WAIC assesses how well the data fit the model and takes into consideration the model complexity and the lower values of WAIC represent better model predictive accuracy (Martin, 2018). Gelman, Lee, and Guo (2015) stated that WAIC is based on “starting with the computer log pointwise

posterior predictive density and then adding a correction for effective number of parameters to adjust for overfitting" (p. 8). WAIC information is part of a built-in function in the ArviZ library. The result of WAIC for each model can be directly compared in terms of the prediction accuracy by specifying the name of the data condition to the model in Arviz. Additionally, WAIC is the expected log pointwise predictive density ($\widehat{\text{elpd}}_{\text{waic}}$), which comprised from the following equation.

$$\text{WAIC} = \widehat{\text{elpd}}_{\text{waic}} = \widehat{lpd} - \hat{p}_{\text{waic}} \qquad (21)$$

Where $\widehat{\text{elpd}}_{\text{waic}}$ refers to expected log pointwise predictive density for a new dataset. $\widehat{lpd}$ refers to expected log pointwise predictive density, and $\hat{p}_{\text{waic}}$ refers to simulation-estimated effective number of parameters. The calculation of $\widehat{lpd}$ comes from;

$$\widehat{lpd} = \sum_{i=1}^{n} log\left(\frac{1}{S}\sum_{s=1}^{s} p(y_i|\theta^s)\right) \qquad (22)$$

In Equation 22, the lpd is an estimate of the expected log pointwise predictive density for new dataset in each observed data, y_i. *N* refers to sample size and θ^srefers to posterior simulations from s =1,…,*S*, $p(y_i|\theta^s)$ means probability of each observed data given posterior simulations. While $\hat{p}_{\text{waic}}$ is calculated from;

$$\hat{p}_{\text{waic}} = \sum_{i=1}^{n} V_{s=1}^{S} \left(log\, p(y_i|\theta^s)\right) \qquad (23)$$

In Equation 23, $V_{s=1}^{s}$ is a sample variance summing over all the observed data. To put WAIC in the measure of deviance, like other Bayesian prediction accuracy measures, for example, deviance information criterion (DIC) and Akaike information criterion (AIC), WAIC is classified as -2 times the expression. Therefore, WAIC is the negative of the average log pointwise predictive density (presuming the likelihood of a new data point) and is divided by sample size, *n* – larger sample size can dominant the variances explained in WAIC (Watanabe, 2013).

In addition to the WAIC, leave-one-out cross-validation (LOO) is another model prediction accuracy measure used in this study. LOO is an estimate of the out-of-sample predictive fit, which involves the process of data being partitioned into training and testing sets repetitively: reiteratively fitting the model with the training data set and assessing the model fit with the holdout data set. The model fit based on LOO is estimated using an approximation of the log predictive density of the holdout data and it is common for LOO values to have negative values. The approximation of the LOO method in machine learning adopts the Pareto smoothed importance sampling (PSIS), the approximation based on importance sampling, which provides the more precise and dependable estimate by fitting a Pareto distribution to the upper tail of the distribution of the importance weights. WAIC and LOO normally produce similar results (Vehtari et al., 2017). The equation for expected log pointwise predictive density in LOO ($\mathrm{elpd}_{\mathrm{loo}}$) is as follows.

$$\mathrm{elpd}_{\mathrm{loo}} = \sum_{i=1}^{n} \log p(y_{i|} y - i) \qquad (24)$$

Where $\sum_{i=1}^{n} \log p(y_{i|} y - i)$ is the leave-one-out predictive density given the data exclude the *i*th data point (Vehtari et al., 2017).

Besides WAIC and LOO, standard errors (SE) for WAIC and LOO were also examined to evaluate the uncertainty of the WAIC and LOO estimates. SE is useful for measuring the ambiguity of the WAIC and LOO estimates and smaller values of SE represent higher certainty in calculation (Gelman, Carlin, et al., 2015; Vehtari et al., 2017). Since each of the calculation of $\widehat{\mathrm{elpd}}_{\mathrm{waic}}$ and $\widehat{\mathrm{elpd}}_{\mathrm{loo}}$ is characterized as the sum of sample size (*n*) components, their standard errors are computed by calculating the standard deviation of the sample size (*n*) components and multiplying by square root of *n.* The following are the equations for standard error of expected

log pointwise predictive density on WAIC, se ($\widehat{\text{elpd}}_{\text{waic}}$), and standard error of expected log pointwise predictive density on LOO, se ($\widehat{\text{elpd}}_{\text{loo}}$), respectively.

$$\text{se}\left(\widehat{\text{elpd}}_{\text{waic}}\right) = \sqrt{n\, V_{i=1}^{n}\, \widehat{\text{elpd}}_{\text{waic,i}}} \quad (25)$$

$$\text{se}\left(\widehat{\text{elpd}}_{\text{loo}}\right) = \sqrt{n\, V_{i=1}^{n}\, \widehat{\text{elpd}}_{\text{loo,i}}} \quad (26)$$

Simulation Procedure

Python (version 3.7.0) was the primary programming language used to complete the Monte Carlo simulation for the current dissertation. The processes of the simulation were as follows.

Step 1. Imported Python libraries that helped with data generation processes, which were:

- Panda: function for data manipulation
- Numpy: function for scientific computing
- Matplotlib: function for data visualization
- Time: function for tracing simulation time
- Random: function for data generation
- Randint: subset of random function to generate integer values

Step 2. Set up the model parameter values, which were:

- Donation amounts: range from $0 to $1,000,000 for years 2015 - 2019
- Sample size: 25, 50, 100, 200
- Proportion of cases in the levels of a dichotomous predictor: dichotomous predictor of 1 and 0 with ratios of 10:90, 25:75, and 50:50

- Number of contacts: range from 0-63, with mean of 20 and standard deviation of 12
- Correlation among variables: used correlation matrix from Table 1 to set up correlation matrix
- Number of replications: number of Monte Carlo replications set to 1,000

Step 3. Randomly generated data based on data condition combinations in Table 2 (e.g., 5314 represents 5 waves of data, 50:50 proportion of dichotomous predictor, informative prior, sample size of 100) in wide data format (single row for each data point), which allowed for incorporation of correlations among variables:

- Randomly generated donation amounts with specified waves of data (i.e., for five waves of data presented the donation in 2015, 2016, 2017, 2018, 2019)
- Randomly generated number of contacts
- Randomly generated proportion in the two levels of dichotomous predictor
- Applied correlation matrix on continuous variables (donation amount from year 2015 – 2019 and number of contacts)
- Combined all variables together, resulting in donation amount within a certain year (number of years for amount depending on number of data waves, i.e., 3 waves equal 2019, 2018, 2017), number of contacts, and attendance at events held by the university (yes, no)

Step 4. Create new columns to specify the design conditions used in step 3 and save the file name to represent each data condition combination. There are separate

columns for each design variable (sample size, waves of data, etc.). Then the level of each condition was coded.

Step 5. Prepare data to run in Pymc3. To run Bayesian growth modeling in Pymc3, long data format (each data point has as many rows as the number of traits and each row consists of values of a trait for a given data point) is required. Thus, a wide data format in step 4 was transformed to a long data format. For example, a person has five rows representing different donation amounts in years 2015 -2019, with the same number of time-invariant predictors (number of contacts and event attendance value, 0 or 1) across rows.

Step 6. Read in the simulated data file to Pymc3 one replicated dataset at the time and followed the steps in applying Bayesian growth modeling with Pymc3 to calculate prediction accuracy for each file and make sure that data generation procedures for the training and testing samples that indicates the data were split based on the same model.

Step 7. Saved all the output of the WAIC, LOO WAIC SE, LOO SE from step 6

Step 8. WAIC, LOO WAIC SE, LOO SE output from all data conditions were merged into a single data set in a Python data frame in order to perform further analysis.

Pilot Study

A pilot study was performed to test the performance of the data generation process and to estimate the computing time. One of the concerns during a data simulation process is whether the simulated data follow the criteria that were specified in the model parameter and simulation process section above. Consequently, before running the actual simulation, data in the form of a pilot study were generated for validation purpose.

First, the combination of five waves of data, proportions of cases in the two levels of a dichotomous predictor of 10:90, sample size N = 100, and informative priors was analyzed in Python. The data generation process followed the steps described above in the simulation process section. The replications of 100 datasets were examined and the simulation time was five seconds. The descriptive statistics showed that the criteria specified in the model parameter sections were met, which are the proportion of donation amount for each wave, proportion of cases in the two levels of the dichotomous predictor, and minimum, maximum, mean, and standard deviation values of number of contacts. Moreover, correlation values of the simulation data also represented the desired correlation matrix.

Second, once the simulation data met the desired criteria, the simulated data were used for analysis in PyMC3. The steps in applying Bayesian growth modeling with PyMC3 were followed and the number of samples used for tracing the posterior inference was 2,000 (Osvaldo et al., 2017). The key factors for analyzing data in PyMC3, which require further data clean-up, were (a) data were in a long format, (b) all variables were integer, and (c) all time points were coded to be specified time point referred in the model. To make sure that the model was able to run properly before including all the variables in the model, each year's donation amounts from 2015 to 2019 with the informative priors were used as outcomes in the model whereas the variable representing whether or not alumni attended an event (yes = 1, no = 0) was used to predict the future donation amount. One of the challenging processes for running Bayesian growth modeling in PyMC3 was specifying the distribution of priors. The raw values of prior knowledge about donation amounts were entered into the model which took around seven minutes to complete each tracing posterior distribution process for sample size of 100. The informative prior distribution for the intercept was $\alpha \sim Gamma\ (0, 2)$, the prior distribution for

the slope was $\beta \sim Exponential\ (0, 2)$, and the prior distribution for the error term was $e \sim HalfCauchy\ (0, 5)$.

Third, once the model for each of the four year's donation amounts with the informative priors was successfully run, the predictive accuracy measure was assessed. As mentioned in the predictive accuracy measure section, WAIC and LOO were used to identify predictive performance and WAIC SE and LOO SE were used to examine the uncertainty in prediction accuracy, for which smaller values of WAIC and LOO indicated better predictive performance. Because the ArviZ library had a built-in function to calculate WAIC, the name of the model used in the pilot study was simply specified under the ArviZ function and the value of WAIC was calculated.

Data Analysis

After getting the results of WAIC, LOO, WAIC SE, and LOO SE for each combination of data condition listed in Table 2, data were ready to be analyzed to answer the research questions. SPSS (version 26) was used to analyze the data. The dependent variables for answering questions regarding prediction accuracy were WAIC, LOO, WAIC SE, and LOO SE while the independent variables were sample sizes, priors, waves of data, and proportion in the two levels of a dichotomous variable. Both descriptive and inferential statistics were used to examine the effects of the above independent variables on the dependent variables. Descriptive statistics were used to gain a general idea of the data values (i.e., minimum, maximum, mean, standard deviation). Factorial ANOVA was the main inferential statistical procedure used to answer the research questions with $\alpha = .05$. For this study I primarily wanted to see if different combinations of variables have different effects on prediction accuracy. Consequently WAIC, LOO, WAIC SE, and LOO SE for each data condition were compared. Moreover, because the

research questions are focused on the moderator effect of priors on the other independent variables, the interaction effects were assessed. If the interaction effects were statistically significant (at what alpha?) and if effect sizes were at least small in magnitude, the post hoc tests of simple main effects were further conducted to gain an understanding of the nature of the interactions. For example, if type of priors moderated the effect of waves of data and the tests of simple main effects indicated that the effect of waves of data was only significant for non-informative priors, then a Tukey test was conducted to determine which numbers of waves of data differ significantly when using non-informative priors. Where interactions were absent, but the main effects were significant, the Tukey test was conducted for the independent variables that had more than two levels. The corresponding main effects and interaction effects were examined at $\alpha = .05$.

This study had an orthogonal design, in which the four independent variables were unrelated to one another; therefore, eta-squared (η^2) was appropriate as the effect size measure to use to compare the magnitude of effect of group differences in the factorial ANOVA (Muthén & Muthén, 2002; Tabachnick & Fidell, 2001). The eta-squared can be derived from the following equation:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}. \qquad (27)$$

The SS_{effect} refers to the variation attributable to the factor, while SS_{total} refers to the total variation in the model. The values of eta-squared range from 0 to 1, because each SS_{effect} is calculated using the same value of SS_{total} as the denominator of Equation 27. Therefore, eta-squared is an accumulated measure in the dependent variable, where the non-error variation cannot be accounted for by other factors in the study (Pierce et al., 2004). According to Cohen (1988), the range of the value of eta-squared can also be looked at as the effect size indicating

magnitude of difference between groups, where $\eta^2 \geq .0099$ is classified as a small effect, $\eta^2 \geq .0588$ is classified as a medium effect, and $\eta^2 \geq .1379$ is classified as a large effect. If the interaction effect in the current study was statistically significant and had an effect size of at least $\alpha < .05$, tests of simple main effects were conducted, and interaction plots were examined as a follow-up.

Chapter Summary

In summary, this chapter covers the information about (a) criteria used to answer the research questions, for which the criteria were selected based on relevant literature and applied studies; (b) the source of parameters used in the model, which are based on real alumni donation data from a university in the mid-Atlantic region; (c) steps for data simulation; (d) steps to fit the simulated data to run Bayesian growth modeling; and (e) the process to assess the effect of data conditions on prediction accuracy. All the above processes were conducted through Python version (3.7.0).

CHAPTER IV

RESULTS

Analyses for a total of 72 designs were conducted on Bayesian growth modeling to assess prediction accuracy in a machine learning environment. The Monte Carlo simulation was conducted using Python to randomly generate data for the analysis. The model parameters were based on alumni donation data from a university in the mid-Atlantic region. The model outcome of interest was model accuracy (measured by WAIC, LOO, and their respective standard errors), and the independent variables were sample size (N = 25, 50, 100, and 200), number of waves of data (T = 3, 4, and 5), proportion of dichotomous time-invariant predictors (90:10, 75:25, 50:50), and prior distribution (non-informative and informative prior). The Monte Carlo simulation generated 1,000 replications for each model condition; therefore, there were 72,000 data points in the final analysis. The 72 Bayesian growth models were executed through the PyMc3 library via Python software and each model was set to draw 2,000 posterior samples.

The Watanabe-Akaike information criterion (WAIC) and Leave-one-out (LOO) statistic were assessed as measures of model prediction accuracy in the models. Both WAIC and LOO are widely used as cross-validation methods for assessing model accuracy in Bayesian machine learning (Gelman, Lee, & Guo, 2015; Vehtari et al., 2017). To recap, WAIC refers to a fully Bayesian approach to evaluate the out-of-sample expectation employing the computed log pointwise posterior predictive density along with rectifying for the effective number of parameters to accommodate for overfitting (Watanabe, 2013). For Bayesian modeling, WAIC can be considered as an improvement on the deviance information criterion (DIC) and defined as

-2 times the expression to be on the deviance scale. Therefore, WAIC value refers to the negative of the average log pointwise predictive density (taking on the estimate of a specific new data value) and divided by sample size. It is common that WAIC returns negative values (Vehtari et al., 2017), as they were in the current study. Higher negative numbers reflect smaller values, which show better prediction accuracy (Vehtari et al., 2017).

LOO cross-validation is an estimate of the out-of-sample predictive fit, which involves the process of data being partitioned into training and testing sets repetitively: reiteratively fitting the model with the training data set and assessing the model fit with the holdout data set. The model fit based on LOO is estimated using an approximation of the log predictive density of the holdout data and it is common for LOO values to have negative values (Vehtari et al., 2017), as they were in the current study. The approximation of the LOO method in machine learning adopts the Pareto smoothed importance sampling (PSIS), the approximation based on importance sampling, which provides the more precise and dependable estimate by fitting a Pareto distribution to the upper tail of the distribution of the importance weights. WAIC and LOO normally produce similar results (Vehtari et al., 2017). Besides WAIC and LOO, standard errors for WAIC and LOO were also examined to assess the uncertainty of the WAIC and LOO estimates. Once all the models were run through Python, the results of WAIC, LOO, and standard errors for WAIC and LOO were exported to Excel, then read into SPSS version 26 for ANOVA procedures, which were used to answer the following research questions.

Q1 Do the types of prior (informative and noninformative priors) moderate the effect of sample size on predictive accuracy for Bayesian growth modeling in a machine learning environment?

Q2 Do the types of prior (informative and noninformative priors) moderate the effect of number of waves of data on prediction accuracy for Bayesian growth modeling in a machine learning environment?

Q3 Do the types of prior (informative and noninformative priors) moderate the effect of proportion of cases in the two levels of a dichotomous time-invariant predictor on prediction accuracy for Bayesian growth modeling in a machine learning environment?

The remaining chapter describes the model descriptive statistics and ANOVA tests to answer the above research questions.

Model Descriptive Statistics

According to the WAIC and LOO results, as mentioned above, it is common for the WAIC and LOO value to be negative, and the larger the negative number the better prediction accuracy (Vehtari et al., 2017). From the model results, both WAIC and LOO showed similar minimum, maximum, mean, and standard deviation values across all conditions. For sample size, the model prediction increased as the sample size increased. For number of waves, the model prediction increased as the number of waves increased. Proportions of cases per level of the dichotomous predictor of 10:90 and 25:75 showed similar prediction accuracy performance and proportion of 50:50 showed slightly lower prediction accuracy performance compared to the proportions of 10:90 and 25:75. In the case of priors, the informative prior showed marginally higher prediction accuracy compared to the non-informative prior.

For WAIC standard errors and LOO standard error results, lower values represented better accuracy of the WAIC and LOO estimates. From the model results, similar mean and standard deviation values are shown across all conditions in WAIC and LOO. Based on sample size, the accuracy of WAIC and LOO estimation decreased as the sample size increased. For number of waves, the accuracy of WAIC and LOO estimation was highest at three waves of data and lowest at four waves of data. The accuracy of WAIC and LOO estimation was similar across all proportions of cases per level of the dichotomous predictor. In the case of priors, the accuracy of WAIC and LOO estimation was similar across priors.

Tables 3 through10 present the minimum, maximum, mean, and standard deviation of WAIC and LOO by sample size, waves of data, proportion of cases in the two levels of a dichotomous time-invariant covariate predictor, and priors. Tables 11 through 18 represent the minimum, maximum, mean, and standard deviation of WAIC standard errors and LOO standard errors by sample size, waves of data, proportion of cases in the two levels of a dichotomous time-invariant covariates predictor, and priors.

Table 3

Descriptive Statistics of WAIC by Sample Sizes

Model *N*	*N*	Minimum	Maximum	Mean	Std. Deviation
25	18000	-603.93	-1089.92	-834.36	174.50
50	18000	-1198.64	-2126.82	-1664.11	332.18
100	18000	-2456.57	-4348.52	-3308.85	673.25
200	18000	-4846.53	-8564.80	-6582.55	1334.92
Total	72000	-603.93	-8564.80	-3097.47	2331.47

Table 4

Descriptive Statistics of WAIC by Waves of Data

Wave	*N*	Minimum	Maximum	Mean	Std. Deviation
3	24000	-603.93	-5187.62	-2334.82	1665.07
4	24000	-791.14	-6817.89	-3092.94	2190.46
5	24000	-995.73	-8564.80	-3864.64	2750.78
Total	7,000	-603.93	-8564.80	-3097.47	2331.46

Table 5

Descriptive Statistics of WAIC by Proportions

Proportion	*N*	Minimum	Maximum	Mean	Std. Deviation
10:90	24000	-611.75	-8564.80	-3110.80	2357.45
25:75	24000	-611.66	-8490.75	-3105.14	2344.98
50:50	24000	-603.93	-8091.04	-3076.46	2291.39
Total	72000	-603.93	-8564.80	-3097.47	2331.46

Table 6

Descriptive Statistics of WAIC by Priors

Prior	*N*	Minimum	Maximum	Mean	Std. Deviation
Non-informative	36000	-623.92	-8136.03	-3092.21	2276.39
Informative	36000	-603.93	-8564.80	-3102.72	2385.29
Total	72000	-603.93	-8564.80	-3097.47	2331.46

Table 7

Descriptive Statistics of LOO by Sample Sizes

Model *N*	*N*	Minimum	Maximum	Mean	Std. Deviation
25	18000	-604.59	-1090.11	-835.22	174.22
50	18000	-1201.44	-2126.78	-1665.27	331.93
100	18000	-2458.23	-4327.66	-3310.72	672.32
200	18000	-4846.52	-8572.42	-6582.30	1335.78
Total	72000	-604.59	-8572.42	-3098.37	2331.07

Table 8

Descriptive Statistics of LOO by Waves of Data

Wave	*N*	Minimum	Maximum	Mean	Std. Deviation
3	24000	-604.59	-5165.07	-2335.54	1663.45
4	24000	-792.22	-6817.90	-3094.25	2190.75
5	24000	-995.86	-8572.42	-3865.33	2750.53
Total	72000	-604.59	-8572.42	-3098.37	2331.07

Table 9

Descriptive Statistics of LOO by Proportions

Proportion	*N*	Minimum	Maximum	Mean	Std. Deviation
10:90	24000	-612.96	-8572.42	-3111.89	2357.35
25:75	24000	-610.70	-8490.09	-3105.23	2343.94
50:50	24000	-604.59	-8073.56	-3078.00	2291.35
Total	72000	-604.59	-8572.42	-3098.37	2331.07

Table 10

Descriptive Statistic of LOO by Priors

Prior	*N*	Minimum	Maximum	Mean	Std. Deviation
Non-informative	36000	-623.92	-8139.00	-3092.19	2276.41
Informative	36000	-604.59	-8572.42	-3104.56	2384.49
Total	72000	-604.59	-8572.42	-3098.37	2331.07

Table 11

Descriptive Statistics of WAIC Standard Errors by Sample Sizes

Model *N*	*N*	Minimum	Maximum	Mean	Std. Deviation
25	18000	2.61	19.23	8.68	1.85
50	18000	8.15	14.74	11.41	1.75
100	18000	0.01	26.07	16.94	2.75
200	18000	0.91	41.45	25.23	5.29
Total	72000	0.01	41.45	15.57	7.11

Table 12

Descriptive Statistics of WAIC Standard Errors by Waves of Data

Wave	*N*	Minimum	Maximum	Mean	Std. Deviation
3	24000	0.91	26.00	13.97	5.83
4	24000	7.43	33.25	17.02	7.77
5	24000	0.01	41.45	15.70	7.24
Total	72000	0.01	41.45	15.57	7.11

Table 13

Descriptive Statistics of WAIC Standard Errors by Proportions

Proportion	*N*	Minimum	Maximum	Mean	Std. Deviation
10:90	24000	2.61	41.45	15.66	7.14
25:75	24000	0.91	33.79	15.61	7.21
50:50	24000	0.00	31.32	15.42	6.97
Total	72000	0.00	41.45	15.57	7.11

Table 14

Descriptive Statistics of WAIC Standard Errors by Priors

Prior	*N*	Minimum	Maximum	Mean	Std. Deviation
Non-informative	36000	0.01	41.45	15.65	8.25
Informative	36000	0.02	34.04	33.25	15.48
Total	72000	0.01	41.45	15.57	7.11

Table 15

Descriptive Statistics of LOO Standard Errors by Sample Sizes

Model *N*	*N*	Minimum	Maximum	Mean	Std. Deviation
25	18000	2.25	25.75	9.01	2.36
50	18000	8.16	16.20	11.58	1.84
100	18000	0.01	28.62	17.17	2.92
200	18000	0.02	41.02	24.95	6.02
Total	72000	0.01	41.02	15.68	7.13

Table 16

Descriptive Statistics of LOO Standard Errors by Waves of Data

Wave	*N*	Minimum	Maximum	Mean	Std. Deviation
3	24000	0.02	34.04	14.06	6.01
4	24000	4.74	33.71	17.15	7.71
5	24000	0.01	41.02	15.83	7.23
Total	72000	0.01	41.02	15.68	7.13

Table 17

Descriptive Statistics of LOO Standard Errors by Proportions

Proportion	*N*	Minimum	Maximum	Mean	Std. Deviation
10:90	24000	2.25	41.02	15.79	7.10
25:75	24000	0.02	34.33	15.60	7.34
50:50	24000	0.01	30.80	15.65	6.95
Total	72000	0.01	41.02	15.68	7.13

Table 18

Descriptive Statistics of Standard Errors LOO by Priors

Prior	*N*	Minimum	Maximum	Mean	Std. Deviation
Non-informative	36000	0.02	34.04	15.70	5.83
Informative	36000	0.01	41.02	15.66	8.23
Total	72000	0.01	41.02	15.68	7.13

Model Results

After reviewing the descriptive statistics, factorial ANOVA was used as the main inferential statistical procedure to answer the research questions, with $\alpha = .05$. Along with the inferential statistical test value, the magnitude of effect size (η^2) was used to estimate the effect of group differences in the factorial ANOVA. According to Cohen (1988), a small effect based on η^2 is $\geq .0099$, the medium effect is $\geq .0588$, and the large effect is $\geq .1379$. In this study, only medium and large effect size were interpreted to be meaningful. Therefore, the statistically significant main and interaction effect with at least medium effect size were further analyze as

post-hoc analysis. WAIC and LOO were used as dependent variables (in two separate ANOVA models) and independent variables were sample size (N = 25, 50, 100, 200), waves of data (T = 3, 4, 5), proportions of cases per level of the dichotomous predictor (10:90, 25:75, 50:50), and priors (non-informative, informative).

Besides testing the main effects of each of these four independent variables, all possible interaction effects were also estimated (including the 4-way interaction). The three research questions were tested by examining the 2-way interactions between priors and each of the other three independent variables – that is, to determine whether the relationship between dependent variables and any of the other independent variables (sample size, waves of data, and proportion of cases per level of the dichotomous predictor) differ as a function of whether noninformative or informative priors were used.

Table 19 presents the results of the ANOVA model with WAIC as the dependent variable. The results show that all main effects and interaction effects were statistically significant at $p < .05$. However, only a subset of effects produced effect size values greater than zero. These are the main effects of sample size and waves of data and interaction between sample size and waves of data. The main effect of sample sizes had a large effect size ($\eta^2 = .89$) and waves of data had medium effect size ($\eta^2 = .07$). The large effect size estimate for the sample size variable indicates that the magnitude differences among sample size on prediction accuracy are large. In addition, the medium effect size associated with waves of data represents that there is a meaningful magnitude of difference among waves of data on prediction accuracy. The interaction between waves of data and sample sizes also produced a medium effect size of ($\eta^2 = .07$).

Table 20 presents the results of the ANOVA model with LOO as dependent variable. Model estimates using LOO as the dependent variable are similar to those using WAIC as the dependent variable. Again, all main effects and interaction effects were statistically significant. The effect size values of main and interaction effects were also close to those reported using WAIC as the dependent variable, with the exception of the effect size for the sample size by waves of data interaction which fell between a small and medium effect ($\eta^2 = .04$).

Table 19

ANOVA Table for WAIC as Dependent Variable

Source	Type III Sum of Squares	df	Mean Square	*F*	Sig.	η^2
Corrected Model	391364656534.97	71	5512178261.06	171422215.55	<.001	1.00
Intercept	69078953409.27	1	690789534091.27	21482736370.81	<.001	.17
Wave	28084920914.51	2	14042460457.25	436703888.97	<.001	.07
Proportion	16272015.85	2	8136007.93	253020.21	<.001	.00
Prior	1987978.43	1	1987978.43	61823.77	<.001	.00
N	348599895378.68	3	116199965126.23	3613681293.52	<.001	.89
Wave * Proportion	28470126.07	4	7117531.52	221346.80	<.001	.00
Wave * Prior	11812187.04	2	5906093.52	183672.51	<.001	.00
Wave * N	14119479269.16	6	2353246544.86	73183180.47	<.001	.07
Proportion * Prior	3438325.04	2	1719162.52	53463.92	<.001	.00
Proportion * N	32102049.66	6	5350341.61	166389.29	<.001	.00
Prior * N	241208153.33	3	80402717.78	2500429.30	<.001	.00
Wave * Proportion * Prior	31157422.07	4	7789355.51	242239.73	<.001	.00
Wave * Proportion * N	67907315.13	12	5658942.93	175986.43	<.001	.00
Wave * Prior * N	17501018.98	6	2916836.50	90710.16	<.001	.00
Proportion * Prior * N	22295394.71	6	3715899.12	115560.06	<.001	.00
Wave * Proportion * Prior * N	86208986.33	12	7184082.19	223416.45	<.001	.00
Error	2312885.51	71928	32.16			
Total	1082156503511.78	72000				
Corrected Total	391366969420.47	71999				

Table 20

ANOVA Table for LOO as Dependent Variable

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	η^2
Corrected Model	391231232780. 60	71	5510299053.25	167060634.16	<.001	1.00
Intercept	69119407704.42	1	691194077040.42	20955545193.19	<.001	.17
Wave	28083547283.62	2	14041773641.81	425716932.36	<.001	.07
Proportion	15480852.12	2	7740426.06	234673.38	<.001	.00
Prior	2757060.93	1	2757060.93	83588.27	<.001	.00
N	348452728642.73	3	116150909547.58	3521450364.10	<.001	.89
Wave * Proportion	27380274.19	4	6845068.55	207528.03	<.001	.00
Wave * Prior	11382061.42	2	5691030.71	172540.04	<.001	.00
Wave * N	14150755646.58	6	2358459274.43	71503506.11	<.001	.04
Proportion * Prior	2892902.62	2	1446451.31	43853.35	<.001	.00
Proportion * N	31773331.84	6	5295555.31	160550.06	<.001	.00
Prior * N	235263885.14	3	78421295.05	2377568.11	<.001	.00
Wave * Proportion * Prior	30276313.75	4	7569078.44	229478.48	<.001	.00
Wave * Proportion * N	64349339.91	12	5362444.99	162578.01	<.001	.00
Wave * Prior * N	17936813.54	6	2989468.92	90634.39	<.001	.00
Proportion * Prior * N	21928403.78	6	3654733.96	110803.82	<.001	.00
Wave * Proportion * Prior * N	82779968.45	12	6898330.70	209142.82	<.001	.00
Error	2372460.71	71928	32.98			
Total	1082427682281.78	72000				
Corrected Total	391233605241.30	71999				

ANOVA results from Tables 19 and 20 provide the information needed to answer Research Questions 1-3. Regarding Research Question 1, despite the statistically significant interaction between sample size and prior, the negligible effect size ($\eta^2 = .00$) indicates that the type of prior does not moderate the effect of sample size on predictive accuracy (using either WAIC or LOO). Regarding Research Question 2, despite the statistically significant interaction between waves of data and type of prior, the negligible effect size ($\eta^2 < .001$) indicated that types of prior did not moderate the effect of waves of data on predictive accuracy. Regarding Research Question 3, despite the statistically significant interaction between proportion of cases in the two levels of a dichotomous time-invariant covariates and type of prior, the negligible effect size ($\eta^2 < .001$) indicated that types of prior did not moderate the effect of proportion of cases in the two levels of a dichotomous time-invariant covariates on predictive accuracy.

Test of Simple Main Effects

The ANOVA results in Tables 19 and 20 showed that the only interaction exhibiting a non-zero effect size was between sample size and waves of data. Therefore, tests of simple main effects were performed as a post hoc analysis and the interaction between sample size and waves of data was plotted. For any statistically significant simple main effects, Tukey's test was used to provide insights regarding comparisons among groups. It is worth pointing out that sample size is a part of the equation in both WAIC and LOO; therefore, in the current study, the substantial amounts of variance explained by sample size resulted in little variance left to explain for other variables (Vehtari et al., 2017). To test simple main effects, sample size was held constant, and a one-way ANOVA test was conducted to compare WAIC and LOO values across waves of data.

Results for both WAIC and LOO showed that, regardless of sample size, the prediction accuracy increased as number of waves increased. Moreover, the results indicated that for

sample size of 25, 50, 100, and 200, there was a statistically significant difference in prediction accuracy (for both WAIC and LOO) among waves of data, with effect sizes (η^2) of < .001, < .001, .01, and .03, respectively. Since sample sizes of 100 and 200 showed more meaningful magnitudes of differences in prediction accuracy across waves ($\eta^2 = .01$, and $\eta^2 = .03$), Tukey tests were conducted as pairwise comparisons of prediction accuracy across waves of data for those two sample sizes. The results suggest that for both sample sizes of 100 and 200, prediction accuracy was highest with five waves of data and lowest with three waves of data. Prediction accuracy with four waves of data was statistically significantly higher than with three waves of data, and prediction accuracy with five waves of data was statistically significantly higher than with four waves of data. The descriptive statistics for WAIC values across waves of data are shown, for sample sizes of 25, 50, 100, and 200 in Tables 21 and 22 and for WAIC and LOO, respectively.

Table 21

Descriptive Statistics of WAIC across Waves of Data for Sample Size of 25, 50, 100, and 200

Sample Size	Wave	*N*	Mean	Std. Deviation
25	3	6000	-622.47	13.3
25	4	6000	-835.75	27.37
25	5	6000	-1044.86	34.88
50	3	6000	-1253.78	32.05
50	4	6000	-1675.99	38.76
50	5	6000	-2062.56	34.81
100	3	6000	-2492.7	32.5
100	4	6000	-3299.15	59.8
100	5	6000	-4134.69	83.04
200	3	6000	-4970.34	124.03
200	4	6000	-6560.85	119.66
200	5	6000	-8216.47	216.06

Table 22

Descriptive Statistics of LOO across Waves of Data for Sample Size of 25, 50, 100, and 200

Sample Size	Wave	*N*	Mean	Std. Deviation
25	3	6000	-623.47	12.37
25	4	6000	-836.85	27.05
25	5	6000	-1045.33	34.39
50	3	6000	-1255.21	30.68
50	4	6000	-1676.94	37.88
50	5	6000	-2063.65	34.08
100	3	6000	-2495.79	29.67
100	4	6000	-3300.58	58.36
100	5	6000	-4135.78	82.14
200	3	6000	-4967.6883	116.78
200	4	6000	-6562.6445	119.57
200	5	6000	-8216.5551	215.73

Along with the test of simple main effects, the interaction between waves of data and sample sizes was plotted to explore how the relationship between number of waves and prediction accuracy, based on WAIC (see Figure 2) and LOO (see Figure 3), differed by sample size. According to the plots, the relationship between number of waves and prediction accuracy is stronger as sample size increases, with the strongest relationship occurring with a sample size of 200. Prediction accuracy is greatest with a sample size of 200 and five waves of data, whereas prediction accuracy is least with a sample size was 25 and three waves of data.

Figure 4

Waves of Data and Sample Sizes Interaction Effect on WAIC

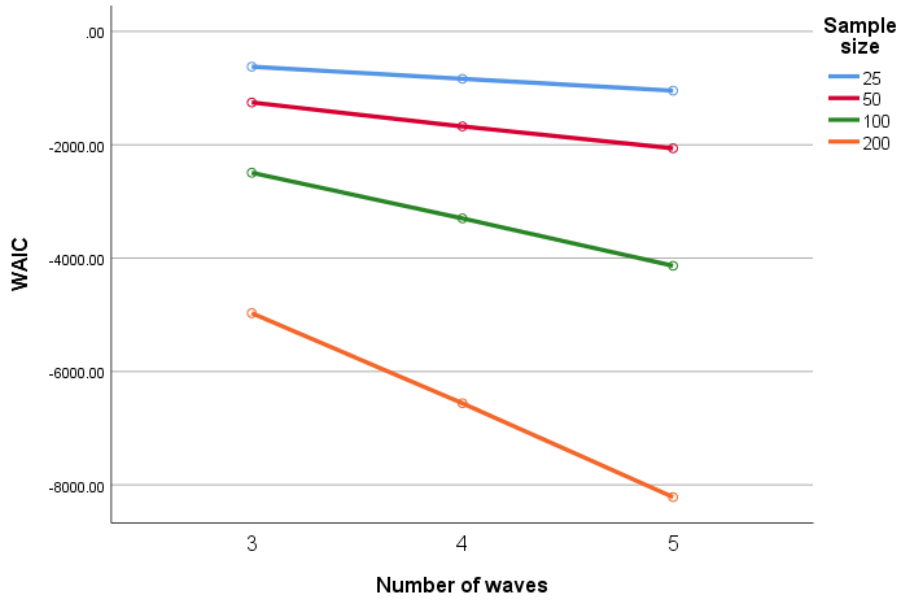
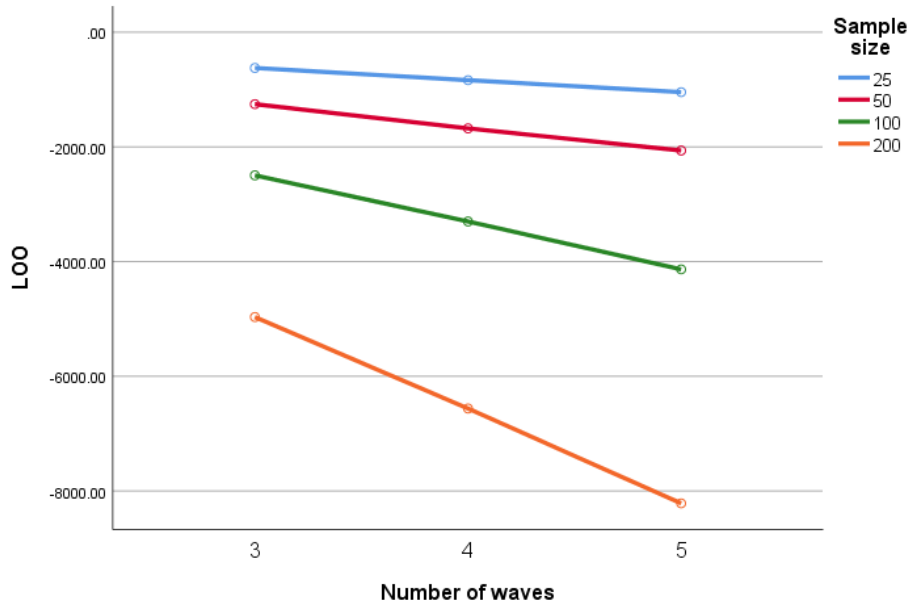


Figure 5

Waves of Data and Sample Sizes Interaction Effect on LOO



Model Standard Errors

In addition to ANOVA models using sample size, prior, waves of data, and proportions of cases per level of the dichotomous predictor to estimate prediction accuracy (WAIC and LOO), a separate set of ANOVA models was analyzed using those same predictors to estimate standard errors associated with WAIC and LOO values. This was done to understand the uncertainty in WAIC and LOO estimations across all variables. Thus, standard errors of WAIC and LOO values were used as dependents variables (in two separate ANOVA models), and sample size (N = 25, 50, 100, 200), waves of data (T = 3, 4, 5), proportions of cases per level of the dichotomous predictor (10:90, 25:75, 50:50), and priors (non-informative, informative) were used as independent variables in both models. Besides testing the main effects of each of these four independent variables, all possible interaction effects were also estimated (including the 4-way interaction).

Table 23 presents the results of the ANOVA model with WAIC standard errors as the dependent variable. The results showed that all main effects and interaction effects were statistically significant at $p < .05$. However, only a subset of effects showed effect size values close to or greater than a medium-sized effect ($\eta^2 \geq .0588$). These were the main effects of sample size and the interaction effect between sample size and prior. The main effect of sample size had a large effect size ($\eta^2 = .79$) and the interaction effect between sample size and prior had a value approaching a medium effect size of ($\eta^2 = .05$). These results suggest that sample size and the interaction between sample size and prior meaningfully impact the estimation of prediction accuracy. Table 24 presents the results of the ANOVA model with LOO standard errors as the dependent variable. Model estimates using LOO standard errors as the dependent variable were similar to those described above using WAIC standard errors as the dependent

variable. Again, all main effects and interaction effects were statistically significant. The main effect of sample size produced a large effect size ($\eta^2 = .74$). While the interaction between sample size and prior produced an effect size approaching medium size in the model used to predict WAIC standard errors, the interaction between sample size and prior surpassed the threshold for a medium effect ($\eta^2 = .07$) in predicting LOO standard errors.

Table 23

ANOVA Table of Standard Error for WAIC as Dependent Variable

Source	Type III Sum of Squares	df	Mean Square	*F*	Sig.	η^2
Corrected Model	3571017.92	71	50296.03	55325.39	<.001	0.98
Intercept	1 7443752.73	1	17443752.73	19188043.83	<.001	0.48
Wave	112513.31	2	56256.66	61882.05	<.001	0.03
Proportion	739.59	2	369.79	406.77	<.001	0.00
Prior	564.52	1	564.52	620.97	<.001	0.00
N	2879181.63	3	959727.22	1055695.30	<.001	0.79
Wave * Proportion	21048.94	4	5262.23	5788.43	<.001	0.01
Wave * Prior	14552.79	2	7276.40	8004.00	<.001	0.00
Wave * N	105138.07	6	17523.01	19275.23	<.001	0.03
Proportion * Prior	12808.47	2	6404.24	7044.63	<.001	0.00
Proportion * N	4136.30	6	689.39	758.32	<.001	0.00
Prior * N	194200.89	3	64733.63	71206.68	<.001	0.05
Wave * Proportion * Prior	29897.85	4	7474.46	8221.87	<.001	0.01
Wave * Proportion * N	69601.25	12	5800.10	6380.09	<.001	0.02
Wave * Prior * N	19867.22	6	3311.20	3642.31	<.001	0.01
Proportion * Prior * N	28196.83	6	4699.47	5169.40	<.001	0.01
Wave * Proportion * Prior * N	78570.29	12	6547.52	7202.25	<.001	0.02
Error	65389.38	71928	0.91			
Total	21080160.03	72000				
Corrected Total	3636407.30	71999				

Table 24

ANOVA Table of Standard Error for LOO as Dependent Variable

Source	Type III Sum of Squares	df	Mean Square	*F*	Sig.	η^2
Corrected Model	3565442.47	71	50217.50	37634.51	<.001	0.97
Intercept	1744375.73	1	1769800.97	13263420.77	<.001	0.48
Wave	115224.20	2	57612.10	43176.25	<.001	0.03
Proportion	459.21	2	229.61	172.07	<.001	0.00
Prior	27.29	1	27.29	20.45	<.001	0.00
N	2693098.65	3	897699.55	672763.15	<.001	0.74
Wave * Proportion	24482.32	4	6120.58	4586.95	<.001	0.01
Wave * Prior	14969.70	2	7484.85	5609.37	<.001	0.00
Wave * N	129356.07	6	21559.35	16157.22	<.001	0.04
Proportion * Prior	20662.35	2	10331.18	7742.50	<.001	0.01
Proportion * N	9424.74	6	1570.79	1177.20	<.001	0.00
Prior * N	245669.03	3	81889.68	61370.60	<.001	0.07
Wave * Proportion * Prior	36585.47	4	9146.37	6854.56	<.001	0.01
Wave * Proportion * N	100592.95	12	8382.75	6282.28	<.001	0.03
Wave * Prior * N	24677.84	6	4112.97	3082.39	<.001	0.01
Proportion * Prior * N	28763.44	6	4793.91	3592.70	<.001	0.01
Wave * Proportion * Prior * N	121449.21	12	10120.77	7584.81	<.001	0.03
Error	95976.92	71928	1.33			
Total	21359426.36	72000				
Corrected Total	3661419.39	71999				

The ANOVA results in Tables 23 and 24 indicate that the only interaction showing a medium effect size or greater in estimating WAIC and LOO standard errors was between sample size and prior. Therefore, a test of simple main effects of these variables was performed as post hoc analysis, and their interaction was plotted. To run the test of simple main effects for the interaction between sample size and prior, sample size was held constant and a one-way ANOVA test was run to compare WAIC and LOO standard error values across type of prior (informative vs non-informative). Simple main effect test results showed informative priors had statistically significantly higher uncertainty (larger standard error value) in estimating prediction accuracy than did non-informative priors. However, the effect size (η^2) for the prior variable in both the WAIC standard error and LOO standard error model was < .001. On the other hand, for sample size of 200, non-informative priors showed higher uncertainty in estimating prediction accuracy than did informative priors, with $\eta^2 = .04$ for WAIC standard errors as dependent variable and $\eta^2 = .05$ for LOO standard errors as dependent variable. The descriptive statistics of WAIC standard errors across priors for sample size of 25, 50, 100, and 200 are shown in Tables 25. Descriptive statistics for LOO standard errors are shown in Tables 26.

Table 25

Descriptive Statistics of WAIC Standard Error across Prior for Sample Size of 25, 50, 100, and 200

Sample size	Prior	*N*	Mean	Std. Deviation
25	Informative	9000	9.67	1.99
25	Non-informative	9000	7.69	0.96
50	Informative	9000	12.24	1.47
50	Non-informative	9000	10.58	1.62
100	Informative	9000	17.70	1.68
100	Non-informative	9000	16.19	3.33
200	Informative	9000	22.30	5.18
200	Non-informative	9000	28.16	3.47

Table 26

Descriptive Statistics of LOO Standard Error across Prior for Sample Size of 25, 50, 100, and 200

Sample size	Prior	*N*	Mean	Std. Deviation
25	Informative	9000	10.27	2.66
25	Non-informative	9000	7.74	0.94
50	Informative	9000	12.57	1.46
50	Non-informative	9000	10.59	1.63
100	Informative	9000	18.17	1.92
100	Non-informative	9000	16.18	3.38
200	Informative	9000	21.78	6.38
200	Non-informative	9000	28.13	3.43

The interaction between priors and sample sizes was plotted to explore how the relationship between priors and the uncertainty in estimating prediction accuracy, based on WAIC standard errors (see Figure 4) and LOO standard errors, differ across sample sizes. According to the plots, non-informative priors demonstrated greater precision in prediction accuracy than did informative prior for sample sizes ≤ 100. However, when sample size was 200, informative priors showed greater precision in estimating prediction accuracy than did non-informative priors.

Figure 6

Priors and Sample Size Interaction Effect on WAIC Standard Error

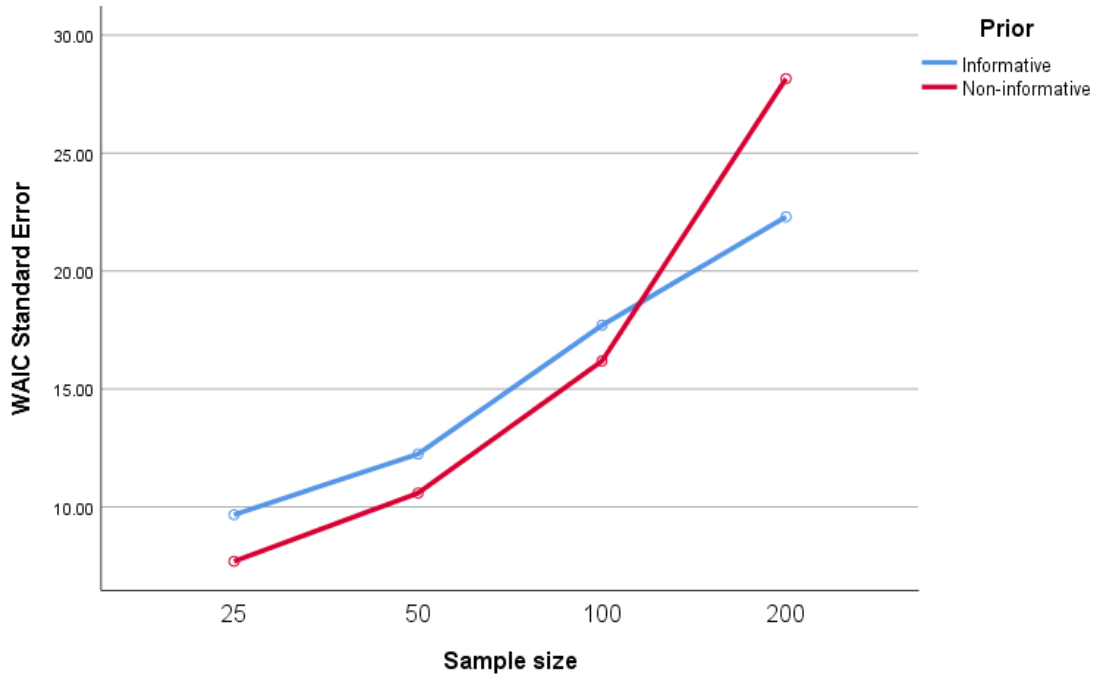
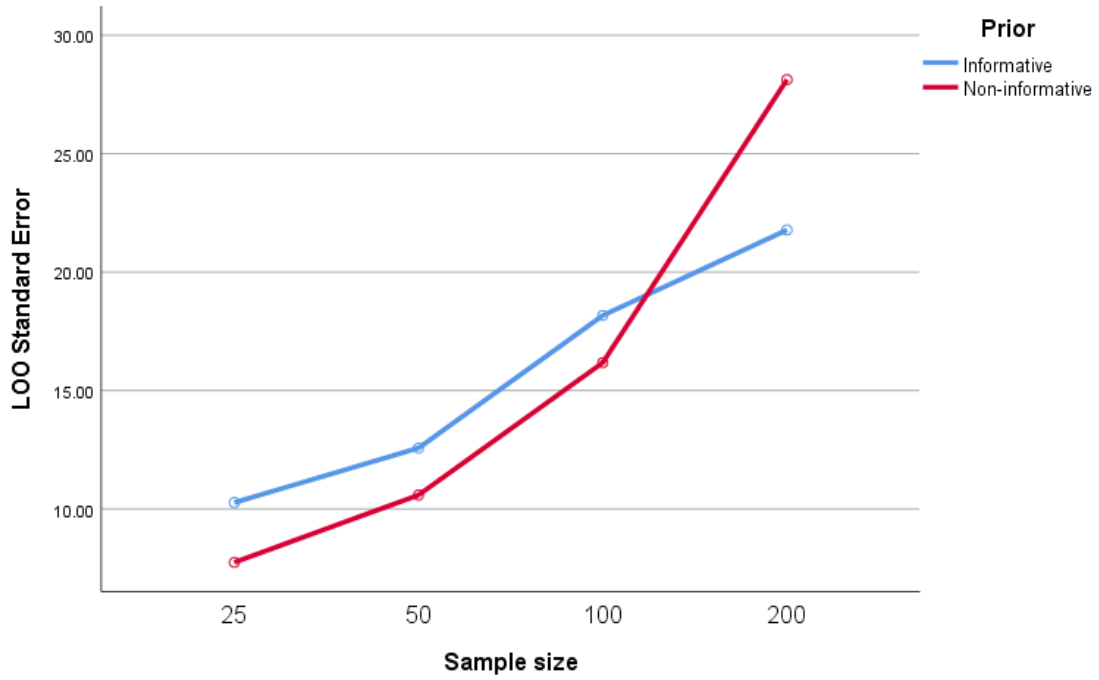


Figure 7

Priors and Sample Size Interaction Effect on LOO Standard Error



Summary of Results

Descriptive and inferential statistics were used to answer the three research questions regarding prediction accuracy (WAIC and LOO). Independent variables were waves of data, sample size, proportions of cases per level of the dichotomous predictor, and priors. Both main and interaction effects were tested. Means and standard deviations of WAIC and LOO values were similar to each other. Tables 3 through 10 provide the descriptive statistics of WAIC and LOO.

Tables 19 and 20 show ANOVA results for the full model and answer research questions 1 through 3. For Research Question 1, although the interaction between sample size and prior was statistically significant, the effect size associated with the interaction was negligible.

Consequently, I could not conclude the type of prior (informative/noninformative) moderated the effect of sample size on predictive accuracy.

For Research Question 2, while the interaction between waves of data and prior was statistically significant, there was no meaningful effect size. Therefore, I could not conclude the type of prior (informative/noninformative) moderated the effect of wavs of data on predictive accuracy.

For Research Question 3, whereas the interaction between proportion of cases in the two levels of a dichotomous time-invariant covariate and type of prior was statistically significant, there was no meaningful effect size. Hence, I could not conclude the type of prior (informative/noninformative) moderated the effect of proportion on predictive accuracy.

Although results show that all main effects and interaction effects were statistically significant at $p < .05$. Only the interaction between sample size and waves of data shows effect size values greater than zero. Therefore, the test of simple main effects was conducted as a post-hoc analysis. Post hoc results revealed that regardless of sample size, the prediction accuracy (i.e., WAIC and LOO) increased as the number of waves increased. Moreover, for sample sizes of 25, 50, 100, and 200, there was a statistically significant difference in prediction accuracy (both WAIC and LOO) across waves of data. Only sample sizes of 100 and 200 showed meaningful magnitudes in differences in prediction accuracy. Consequently, a pairwise comparison test was conducted on prediction accuracy across waves of data for those two sample sizes. The pairwise test suggested that, for both sample sizes ($N = 100$ and $N = 200$), prediction accuracy was highest with five waves of data and lowest with three waves of data. Prediction accuracy with four waves of data was statistically significantly higher than with three waves of

data, and prediction accuracy with five waves of data was statistically significantly higher than with four waves of data.

Along with assessing prediction accuracy using WAIC and LOO values, the precision in estimating prediction accuracy was also examined using WAIC and LOO standard error values. The ANOVA model for estimating precision in estimating prediction accuracy indicated that although all main effects and interaction effects were statistically significant, only the main effect of sample size and interaction effect between sample size and prior showed meaningful effect sizes. Hence, the test of simple main effects was performed to understand whether the relationship between precision in estimating prediction accuracy and sample size differed across type of priors. Results showed that non-informative priors demonstrated greater precision in estimating prediction accuracy than did informative priors with sample sizes of 25, 50, and 100. However, informative priors showed greater precision in estimating prediction accuracy with a sample size of 200. The interpretation and implications of study results are discussed in Chapter 5.

CHAPTER V

DISCUSSION

Through this simulation study, I aimed to understand the effect of priors' interaction with sample size, number of waves of data, and the proportion of cases in the two levels of a dichotomous time-invariant predictor on model prediction accuracy in Bayesian growth modeling in a machine learning environment. Analyses were conducted using the PyMC3 program in Python. Prediction accuracy was operationalized using WAIC and LOO indices. Moreover, measures of prediction accuracy uncertainty were also utilized as outcome variables using WAIC and LOO standard error values. In this chapter I summarize and discuss study results in the context of existing literature. The finding for each variable is discussed, along with implications for practice, limitations of the present study, and recommendations for future research.

Discussion of Findings

Priors

One of the major benefits of Bayesian growth modeling is the ability to include prior knowledge to inform model estimation, which can help to increase the precision of the posterior distribution (Depaoli, Rus, et al., 2017; Dunson, 2001; Zondervan-Zwijnenburg et al., 2017). A prior is also a primary focus of this dissertation. There are several studies supporting how appropriately incorporating a prior distribution into Bayesian growth models can help to determine the ideal growth trajectory and improve model estimation accuracy (Depaoli, 2014; Depaoli, Yang, & Felt, 2017;

Walls & Quigley, 2001). In addition, when an informative prior is available, it should be adopted into the model over a non-informative prior, especially in longitudinal data. Various studies support that non-informative priors show poor performance in parameter recovery and large bias in the posterior distribution (Richardson & Green, 1997; Roeder & Wasserman, 1997). Additional information about the model can also be accessed from informative priors. Thus, by not employing an informative prior when it is available, important information can be wasted (Bolstad, 2007).

Given the condition of informative priors of the study, informative priors did not show any higher prediction accuracy compared to non-informative priors. However, with a sample size of $N = 200$, models using non-informative priors showed higher uncertainty in estimating prediction accuracy than did models using informative priors, with meaningful effect sizes observed for both WAIC and LOO standard errors. The fact that the difference in prediction accuracy between models using informative and non-informative priors was not practically significant may be due to the complexity of the model (Depaoli, 2014; Shi & Tong, 2017; Zhang et al., 2007).

The models tested in this study were quite simple: one continuous dependent variable, one continuous independent variable, and one dichotomous independent variable in conjunction with multiple waves of data and large sample size, which might have resulted in overfitting the model. Therefore, one of the reasons it would be important to study the use of Bayesian machine learning in more complex models, such as those with additional predictors, interaction effects, non-linear terms, etc., is that the model in the current study might have overfit the data. Shi and Tong (2017) considered model complexity as a predictor and results showed that, holding all else constant, priors matter more with more complex models. That is, informative priors showed

higher prediction accuracy in more complex models than in less complex models. Their result also showed that model recovery and estimation have less effect whether an informative or non-informative prior is applied, if the parameter in the model is specified correctly.

Although there is no absolute way to choose priors, choosing types of priors mainly depends on the researcher's knowledge regarding which model parameters should be given more weight (Congdon, 2014). By adding knowledge that we know about priors can help with narrowing the posterior distribution, therefore, increasing prediction accuracy (Alzubi et al., 2018; Depaoli, 2014; Shi & Tong, 2017; Zhang et al., 2007).

The non-informative prior used in this dissertation was a uniform distribution, as opposed to the normal distribution that was used for the informative prior. For the uniform distribution used in this study, it was one uniform distribution of many possible uniform distributions. Therefore, the conclusion of the result of the non-informative prior from only one type of uniform distribution in this study cannot be generalized to suggest that the results hold true for all types of non-informative priors. In other words, it is not that the results from non-informative priors did not differ from those of informative priors in a general sense; my results only indicated a lack of effect of priors for the specific priors that I used. Using different sets of informative or non-informative priors might have yielded different results.

Relative to the uniform distribution which was used as non-informative priors in this study, the plausible range for the model parameters still needed to be specified. The lower and upper bound was set based on the range of available data of a donation amount. Thus, the range or variance of uniform distribution can be viewed as weakly informative, where a narrow range of priors would correspond with a stronger belief in the values of parameters in the model (Golchi, 2018). Stated differently, by specifying the lower and upper bound of the values based

on available data, in return the non-informative priors used in this study were not totally naïve. Weakly informative priors are generally used to keep data inferences in a reasonable range. Consequently, a weakly informative prior represents partial information about a variable and some amount of variance is explained via a weakly informative prior (Zhang et al., 2007). Moreover, the weakly informative prior can be viewed as regularization, which means the technique of including information in the parameter to solve the overfitting problem in a statistical model (Chung et al., 2015; Röver et al., 2021). Overfitting in statistics, especially in machine learning, means the combination of the analysis that represents values too close with the set of data, and resulting in deteriorating fit to additional data or predicting future data reliably (Dietterich, 1995). With the context of how the prior distributions were set in the current study, it seems that the prior comparison was between strongly informative priors and weakly informative priors rather than informative and non-informative prior (Alzubi et al., 2018; Chung et al., 2015; Röver et al., 2021). Therefore, given the combination of the data conditions in this study, the choice of non-informative priors chosen in this study, which more likely reflected weakly informative priors rather than non-informative priors, did not perform differently in term of prediction accuracy compared to the choice of informative priors.

Another reason that possibly explains why the choice of informative and non-informative priors chosen in this study did not show any difference in prediction accuracy is the amount of variance explained in each level of prior (Gelman, 2006b). For a hierarchical model, as in this study, prior distribution is required in each level (Daniels, 1999). In the study study, the amount of variance explained was assumed to be equal within each type of prior and each data level. According to a study from Gelman (2006b), different amounts of variance in the prior parameter result in different prediction accuracy.

Sample Size

One of the benefits of Bayesian modeling is working well with smaller sample size if appropriate priors are specified (Berger, 1990; Hoogland & Boomsma, 1998; Schafer, 1997; Scheines et al., 1999). On the other hand, machine learning works well under complex and large data sets (Chen et al., 2018; Cui & Gong., 2018; Walsh et al., 2018). With the contradictory sample size concepts between Bayesian modeling and machine learning, applied researchers tend to use Bayesian modeling in machine learning with respect to accounting for uncertainty of model parameter in posterior distribution via prior rather than in terms of the small sample size requirement aspect (Sambasivan et al., 2020; Zeng & Luo, 2017). Therefore, the descriptive results for sample size were as expected, as prediction accuracy in Bayesian model evaluation increased along with sample size (Martin, 2018; Vehtari, et al., 2017).

One observation of WAIC and LOO mean and standard deviation values observed across sample sizes was that as WAIC and LOO mean values increased, their standard deviations also increased. Standard deviation measures variability; in this case in WAIC and LOO values. Typically, as sample size increases, variability in variable values is expected to decrease and yield better results. For example, when comparing data between a treatment versus non-treatment control group, a smaller standard deviation in the treatment group indicates consistent results of the treatment, which is good if we assume that the treatment worked (Altman & Bland, 2005; Hess & Hess, 2016). In this dissertation, results suggest models with larger sample sizes produced greater mean prediction accuracy than models with smaller sample sizes, but there was also greater variability in the prediction accuracy among models with larger sample sizes than among models with smaller sizes. Thus, rather than findings indicating that models with larger sample sizes consistently show higher predictive accuracy, as is typically the case, findings

indicted that models with larger sample sizes were less consistent though more accurate than for smaller sample sizes. Or another way to say this would be that contrary to expectations, larger sample size was associated with more accurate but less precise prediction whereas smaller sample size produced findings that were less accurate but more precise. Stated differently, findings suggest larger sample sizes to be associated with lower certainty of higher prediction accuracy and smaller sample sizes to be associated with higher certainty of lower prediction accuracy.

The finding of increasing standard error with increasing sample size is unexpected, as the standard error is expected to decrease as sample size increases. Standard errors in both WAIC and LOO derive from taking N data points into account as a sample from a larger population or, consistently, as independent completions of an error model (Vehtari et al., 2017). The standard error for WAIC and LOO is calculated assuming normality and may not be reliable when the sample size is low (Vehtari et al., 2017). Moreover, the standard deviation of the standard error is usually smaller as sample size is larger (Hess & Hess, 2016). Despite the unexpected finding that standard errors increased with sample size, it is possible for the model with better prediction accuracy (smaller WAIC/LOO) to have larger WAIC/LOO standard errors (Martin, 2018).

The full ANOVA models, including sample size as an independent variable and WAIC and, separately, LOO as dependent variables, were used to answer Research Question 1. While the interaction between sample size and prior was statistically significant, the effect size associated with the interaction was negligible. Subsequently, there was not enough evidence to support that the type of prior (informative/noninformative) moderated the effect of sample size on predictive accuracy. Although the effect size associated with the interaction between sample size and prior was not practically meaningful, the main effect of sample size was associated with

a large effect size. This suggests that sample size meaningfully affected prediction accuracy. This effect was anticipated and is partly due to sample size being used as a part of the calculation of both WAIC and LOO estimates. However, the considerable amount of variance in WAIC and LOO values explained by sample size consequently means that there is little variance in WAIC and LOO values left to be explained by other variables (Gelman, Lee, & Guo, 2015; Vehtari et al., 2017).

In machine learning, there are three main components that affect prediction accuracy, which are problem complexity, model complexity, and sample size (Oravecz & Muthén, 2018; Shi & Tong, 2017). The models tested in this dissertation lacked complexity, which may have caused the sample size to explain the overwhelming amount of variance in prediction accuracy (Oravecz & Muthén, 2018). For example, the simulated data in this study included only two predictors (one binary and one continuous). Also, both continuous predictors and the dependent variable were generated to be normally distributed and, other than manipulating the proportions in the two groups on the binary predictor, the model was tested under near-ideal conditions. It is possible that when a model is relatively simple and the data are normally distributed, neither machine learning nor Bayesian estimation offer any real advantage. That is, model complexity is associated with more variance explained in the model and, in turn, better prediction accuracy (Kwok et al., 2008; Salakhutdinov & Mnih, 2008; Spiegelhalter et al., 2002). Furthermore, Kass and Raftery (1995) recommended specifying the number of parameters in each model when comparing prediction accuracy in Bayesian information criterion. Ghaffari et al. (2019) also recommended that in a scenario where the problem and model complexity are both low, the majority of the weight in the model will be on sample size, with larger sample sizes returning greater prediction accuracy, as was the case in the current study. In addition, when both problem

and model complexity are great and sample size is large, most machine learning models still produce reliable prediction accuracy (Vabalas et al., 2019).

Waves of Data

The descriptive results for waves of data were as expected, with models having more waves of data producing higher prediction accuracy (Oravecz & Muthén, 2018; Shi & Tong, 2017; Willett, 1989). One important observation regarding these results is that the standard deviations of WAIC and LOO increased as waves of data increased. The interpretation for this finding is similar to that for sample size described above. That is, results showed greater certainty of low prediction accuracy among models with fewer waves of data, while results showed less certainty of high prediction among models with more waves of data. Although the prediction accuracy increased along with waves of data, the uncertainty indicated by the WAIC and LOO standard error estimates for both four and five waves of data was slightly higher than for three waves of data. However, the WAIC and LOO standard errors for five waves of data were lower than for four waves of data. The value of higher standard error in four and five waves of data compared to three waves of data is quite surprising since the standard error is expected to decrease as waves of data decrease (Gibbons et al., 2010). However, as mentioned in the discussion of sample size findings above, it is possible for models with better prediction accuracy (smaller WAIC/LOO) to have larger WAIC/LOO standard errors (Martin, 2018).

The full ANOVA models including wave as an independent variable and WAIC and, separately, LOO values as dependent variables were used to answer Research Question 2. Despite the statistically significant interaction between waves of data and type of prior, the negligible effect size indicated that types of prior did not moderate the effect of waves of data on predictive accuracy. However, the main effect of waves of data had a medium effect size,

indicating a meaningful effect of waves of data on prediction accuracy. The finding regarding higher waves of data associated with increased prediction accuracy was anticipated and several studies suggest that more waves of data are associated with greater prediction accuracy and more reliable estimates for individual growth (Gibbons et al., 2010; Kwok et al., 2008; Long & Mills, 2018). One of the reasons that could explain why priors did not moderate the effect of waves of data on prediction accuracy is the number of waves of data used in the current study. Evidence suggests that machine learning provides an advantage when using complex datasets; the combination of informative priors and more waves of data provide better information in machine learning compared to the combination of informative or non-informative priors and fewer waves of data (Fawcett et al., 2017; Harris & Rice, 2013). Unfortunately, only three, four, and five waves of data were tested in this dissertation, which may not have been sufficient for machine learning using informative priors to produce meaningfully improved prediction accuracy over models using non-informative priors.

Proportions of Cases per Level of the Dichotomous Predictor

The descriptive results for the proportion of cases per level of the dichotomous predictor showed that proportions of 10:90 and 25:75 exhibited similar prediction accuracy, which was slightly higher than the prediction accuracy attained with a proportion of 50:50. Despite the statistically significant main effect in ANOVA and the apparent pattern seen in the descriptive statistics, the associated effect size was negligible for the main effect of proportions in the dichotomous predictor. The interaction effect between prior and proportion of cases in the two levels of the dichotomous predictor did not show any practical effect size. Therefore, regarding Research Question 3, the types of prior did not moderate the effect of proportion of cases in the two levels of a dichotomous time-invariant covariates on predictive accuracy.

The ANOVA test for WAIC and LOO on proportions of cases per level of the dichotomous predictor also showed only negligible effect sizes. According to these findings, performance of prediction accuracy was similar regardless of disparity in proportions of cases per level of the dichotomous predictor. The finding with respect to negligible effect sizes on proportions of cases per level of the dichotomous predictor is not surprising because in machine learning, prediction accuracy is more likely to be lower when dependent variable values, as opposed to independent variable values, are imbalanced, that is, the algorithm is more likely to classify data into the class with more cases (i.e., the majority class), in turn, giving the inaccurate assumption of a highly accurate model (Barella et al., 2021). Imbalanced data on dependent variables has led to poor prediction of unusual events (i.e., the minority class) and distorted the predictive models (Barella et al., 2021; Luque et al., 2019; Yap Bee Wah et al., 2016). Although imbalanced data on the dependent variable tends to reduce prediction more than does imbalanced data on independent variable values, as a good practice, whether imbalanced data are in the dependent or independent variables, researcher should follow techniques to handle unbalanced data, presented in several studies (Ali et al., 2019; Kim et al., 2019). These include under-sampling (to reduce the ratio of cases in the majority and minority groups) and oversampling (to create synthetic data of the minority class based on the available data). These techniques are common to find in statistical packages such as Python and R (Ali et al., 2019; Kim et al., 2019; Yap Bee Wah et al., 2016).

Interaction Effect between Sample Sizes and Waves of Data

Along with meaningful effect sizes for the main effects of sample size and waves of data, the only interaction effect exhibiting a non-zero effect size was between sample size and waves of data. Hence, the tests of simple main effects and the interaction plot were assessed as a post

hoc analysis of effects of these variables. Results revealed that regardless of sample size, the prediction accuracy increased as number of waves increased. Moreover, sample sizes of 100 and 200 showed greater prediction accuracy across waves than did sample sizes of 25 and 50. This result is not surprising. As Kwok et al. (2008) suggested, the combination of increasing sample size and waves of data not only can increase prediction accuracy but also enhances the statistical power to uncover the effects of higher-level predictors and the cross-level interaction effects between the within- and between-individual predictors in longitudinal hierarchical models.

In my design, sample size and waves of data were not independent factors; as waves of data increased, so did total sample size. Consequently, the design of this study included confounding between number of waves of data and total sample size. To be more specific, waves of data and sample size potentially have a cause-and-effect relationship to each other and can introduce bias into the result (VanderWeele, 2019). For example, the results in this study showed that higher waves of data were associated with higher prediction accuracy; however, the sample size might be a confounding variable. This is because, in the current study, total sample size increased with increasing waves of data. Consequently, when number of waves increased, the effect found for waves on prediction accuracy might not have been solely due to the increased number of waves but was also likely due to the increased overall sample size. As a result, the confounding between number of waves of data and total sample size might have affected the results on prediction accuracy, particularly in terms of the sample size and waves main effects as well as their interaction (Stadtfeld et al., 2018; VanderWeele, 2019).

Research Implications

The main components for this dissertation study are machine learning, Bayesian growth modeling, and variable conditions, especially type of priors, that affect prediction accuracy. Machine learning has increased its popularity in the last decade due to its well-known advantage of performing well with large and complex data sets, particularly those used to answer research questions with longitudinal data (Chen et al., 2018; Cui & Gong, 2018; Walsh et al., 2018). However, not all researchers have the luxury of assessing large samples even when testing complex models. Alternatively, Bayesian modeling is well-known in its ability to perform well with smaller sample sizes, and it allows researchers to incorporate prior knowledge into the model (Berger, 1990; Hoogland & Boomsma, 1998; Schafer, 1997; Scheines et al., 1999). Therefore, in this dissertation study I examined the combination of Bayesian modeling and machine learning to see at which level sample size starts to make a difference in prediction accuracy in machine learning and how type of prior might moderate the sample size effect.

The findings of this dissertation show that if reasonable values for model parameters in non-informative priors are specified, similar prediction accuracy can be achieved as with informative priors in Bayesian machine learning. Additionally, the performance of prediction accuracy was similar regardless of the proportion of cases per level of the dichotomous predictor.

Key take-aways from this dissertation are: (1) larger sample sizes result in higher prediction accuracy regardless of other model conditions; (2) more waves of data result in higher prediction accuracy; (3) prediction accuracy is similar regardless of number of proportions of cases per level of the dichotomous predictor; (4) the relationship between prediction accuracy and sample size does not differ based on the type of prior used; (5) the relationship between prediction accuracy and waves of data does not differ based on type of prior used; (6) the

relationship between prediction accuracy and proportions of cases per level of the dichotomous predictor does not differ based on type of prior used; and (7) the relationship between sample size and prediction accuracy differs by waves of data. From a practical perspective, applied researchers who might consider using Bayesian modeling in a machine learning context with longitudinal data should take into consideration the key takeaways, described above, in the model design process in order to yield reliable prediction accuracy.

Limitations of the Study

Consistent with other studies, the range of plausible conditions needed to be constrained to keep the research controllable. Therefore, not all possible aspects were considered in the models tested in the current study. Although in this dissertation study I tried to mimic the real-world data as much as possible, the data were still simulated under close to ideal settings. For example, the data were generated as normal distributions, whereas most real-world data are non-normal (Witten et al., 2011).

Given the conditions in this study, the choice of informative and non-informative prior chosen in this study showed similar prediction accuracy. However, only one type of informative prior (normal distribution) and one type of non-informative prior (uniform distribution) were used in this study. Consequently, the conclusion of the result regarding the non-informative priors from only one type of uniform distribution in this study cannot be generalized as if the results hold true for all types of non-informative priors. By using different sets of informative or non-informative priors might have yielded different results; therefore, in this study, I was not be able to determine to what extent priors might have affected predictive accuracy over a wider range of informativeness (Gelman, 2006b; Van Erp et al., 2018).

The confounding of the sample size and waves of data is also another limitation that affected the results particularly in terms of the sample size and waves main effects as well as their interaction on prediction accuracy. Consequently, when number of waves increased, the effect found for waves on prediction accuracy was not exclusively due to the increased number of waves but also due to the increased overall sample sizes (Stadtfeld et al., 2018; VanderWeele, 2019).

Additionally, there were no missing data in the current study, while longitudinal data normally show missing cases. Also, only certain sample sizes were set (25, 50, 100, and 200), although the difference in prediction accuracy might also occur with a smaller range of sample size (Kwok et al., 2008; Long & Mills, 2018). For example, if the range had been closer, such as 20, 25, 50, 75, and 100, the large sample size effect found in the current study would likely be reduced, and other factors might have appeared to make a greater difference in prediction accuracy (Shirzadi et al., 2019). Moreover, the complexity of the model in the current study was relatively simple, though there is evidence to support that prediction accuracy matters more with greater model complexity (Shi & Tong, 2017; Spiegelhalter et al., 2002).

Another limitation in this study is that it is possible that the non-informative priors used were not totally non-informative. For example, the uniform distribution was used as non-informative priors; however, the possible ranges of upper and lower bounds were specified to cover the conceivable range of ranges of donation data that were used as simulation parameter.

In another aspect, the uniform prior in the above situation tends to be viewed as weakly informative prior since it reflects some values of the magnitude of the event of interest more likely than others (Seaman et al., 2012; Seltzer et al., 1996). Additionally, the actual donation data have a tendency to be skewed, but I chose to use a normal distribution when I simulated

data to be able to control number of data conditions. Consequently, normal distribution might not be an appropriate prior to this particular data setting, resulting in priors that did not appear to moderate any of the other data conditions. Nevertheless, the judgment of the right priors is hard to pin down in applied situations (Congdon, 2014).

Recommendations for Future Research

As mentioned in the limitation section, this dissertation study did not cover all possible aspects in the longitudinal modeling environment. Therefore, the effects of several data conditions on prediction accuracy in Bayesian machine learning still warrant consideration. First is the consideration of the level of non-normality (e.g., skewness, kurtosis). For example, researchers should consider different levels of non-normality (varying amounts of skewness and kurtosis) on a continuous dependent variable, a continuous independent variable, or both (Fernández & Steel, 1998). Varying the amount of skewness and kurtosis can be useful to understanding whether non-normality affects prediction accuracy in Bayesian machine learning and, if so, by how much and at which level of non-normality (Maniruzzaman et al., 2017).

Second, beside normal distribution for informative prior and uniform distribution for non-informative prior, future researchers should possibly consider comparing different distributions of informative and non-informative priors, or informative prior versus different levels of weakly non-informative prior (Gelman, 2006b; Van Erp et al., 2018). For example, with informative priors, the Gamma distribution can be used for informative distribution and the inverse-gamma family for non-informative priors. With weakly informative priors, like the uniform distribution, different lower and upper bound values can be used as a point of comparison. Additionally, different amount of variance in each type of prior can be considered. For example, researchers could compare various ranges of negative variance parameters with several range of positive

variance parameters in prior distributions (Gelman, 2006b). Accessing different characteristics of informative and non-informative priors (i.e., type of prior, amount of variance in priors, level of weakly informative priors) helps to determine to what extent a prior's characteristics might have affected predictive accuracy in different data conditions (Gelman, 2006b; Van Erp et al., 2018).

Third, regarding the aspect of confounding between the sample size and waves of data, future research should be designed so that the effects of sample size versus waves of data can be disentangled, for example, by including conditions in which sample size is held constant while varying number of waves. Holding sample size constant and varying number of waves helps to understand to what extent waves of data might have affected predictive accuracy, controlling for sample sizes (Stadtfeld et al., 2018; VanderWeele, 2019).

Fourth, future research should consider the effect of missing data on the dependent variable, independent variable, or both. This could include comparing different types (or mechanisms) of missing data and/or comparing the percentage of data missing particularly with increasing waves of data (Hong & Lynn, 2020). Effects of missing data in longitudinal research shown in other studies included, but were not limited to, bias in the parameter estimation, reduction in representativeness of the samples, bias in interpreting results, and limited understanding of the change over time (Allen, 2017; Laird & Ware, 1982; Raghunathan, 2015). Research that incorporates missing data into longitudinal models based on Bayesian estimation and machine learning should also consider the mechanisms of missing data (MCAR, MAR, and MNAR; Laird & Ware, 1982; Rubin, 1976). Regarding the percentage of missing data, different rates of missing data can be considered (i.e., 15%, 20%, 25%). Incorporating missing data could enhance the understanding of how different patterns of missing data and/or percentages of missing data impact prediction accuracy in Bayesian machine learning (Daniels & Hogan, 2008).

Fifth, future research may consider the effect of differing levels of model complexity in the simulation model on prediction accuracy in Bayesian machine learning by adding a greater number of predictors, interaction effects, non-linear terms. In addition to adding more variables to increase the model complexity, researchers might consider adding one or more time-varying covariates as only two time-invariant covariates were included in this study. There is evidence supporting that increasing the number of time-invariant covariates results in increasing prediction accuracy in machine learning (González-Recio & Forni, 2011; Shi & Tong, 2017). Understanding how model complexity influences prediction accuracy would help applied researchers to make more informed decisions about model specification. Additionally, inclusion of more complex models would provide an examination of the utility of informative versus non-informative priors under more demanding circumstances where priors might play a bigger role (Shi & Tong, 2017; Spiegelhalter et al., 2002).

Sixth, reducing the incremental range of sample size to see whether lower increments of sample size (i.e., 20, 25, 30) alter the prediction accuracy may warrant future research. Assessing smaller increments of sample sizes can give detailed information regarding the point at which sample size begins to influence prediction accuracy in Bayesian machine learning (Shirzadi et al., 2019).

Seventh, comparison of the prediction accuracy results based on Bayesian estimation should be made with results obtained using other estimators in machine learning (i.e., Maximum likelihood). This way, it would be possible to determine if Bayesian estimation is truly beneficial in the context of machine learning (Malhotra, 2015).

Conclusion

Researchers interested in studying longitudinal data have various analytic options available, with two of the newer and more flexible options including Bayesian estimation and machine learning. Despite the potential benefits of Bayesian modeling, many applied researchers shy away from using Bayesian techniques. Difficulties that discourage applied researchers from choosing Bayesian modeling are (a) specifying the right priors and (b) the complexity of the concept of Bayesian modeling (Zitzmann & Hecht, 2019). Fortunately, there is a growing number of resources that show potential users how to apply Bayesian modeling in machine learning (Martin, 2018; Oravecz & & Muthén, 2018).

Through this simulation study I aimed to understand the effect of priors' interaction with sample size, number of waves of data, and the proportion of cases in the two levels of a dichotomous time-invariant predictor on model prediction accuracy in Bayesian growth modeling in a machine learning environment.

Given the limited research on this emerging technique, through this dissertation I sought to provide researchers with a better understanding of the effects of these factors on prediction accuracy, as well as predictive accuracy certainty. Even though the findings in this study generally did not support there being a moderating effect of type of priors chosen in this study on the relationship between other predictive factors (i.e., sample sizes, waves of data, and the effect of proportion of cases in the two levels of a dichotomous time-invariant covariates) and predictive accuracy, there were a few highlighted takeaways for applied researchers who might consider using Bayesian modeling in a machine learning context with longitudinal. First, for machine learning to work well, if it is feasible, consider using larger sample size and more waves of data, even with Bayesian modeling by utilizing the strength of Bayesian modeling in term of

accounting for uncertainty of posterior distributions via priors rather than based on the small sample size aspect of Bayesian modeling (Sambasivan et al., 2020; Zeng & Luo, 2017). Second, although there are no clear rules for choosing priors, if it is achievable, researchers should use their knowledge as a guide for selecting priors and determine which model parameters should be given more weight (Richardson & Green, 1997; Roeder & Wasserman, 1997; Shi & Tong, 2017).

This study was one of the first known studies to examine Bayesian estimation in the context of machine learning. Results of the current study suggest that capitalizing on the advantages offered jointly by these two modeling approaches shows promise. Although much is still unknown and in need of investigation regarding the conditions under which a combination of Bayesian modeling and machine learning affects prediction accuracy, the current dissertation provides a first step in that direction.

References

Aitkin, M. (2001). Likelihood and bayesian analysis of mixtures. *Statistical Modeling, 1*(4), 287-304. doi:10.1177/1471082X0100100404

Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse testing (FIT) project. *PloS One, 12*(7), e0179805. doi:10.1371/journal.pone.0179805

Alhamzawi, R., & Yu, K. (2013). Conjugate priors and variable selection for Bayesian quantile regression. *Computational Statistics and Data Analysis, 64*, 209-219. doi:10.1016/j.csda.2012.01.014

Ali, H., Salleh, M. N. M., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: A review. *Indonesian Journal of Electrical Engineering and Computer Science, 14*(3), 1560-1571.

Allen, J. (2017). A Bayesian hierarchical selection model for academic growth with missing data. *Applied Measurement in Education, 30*(2), 147-162. doi:10.1080/08957347.2017.1283318

Alpaydin, E. (2014). *Introduction to machine learning*. Retrieved from https://ebookcentral.proquest.com.

Altman, D. G., & Bland, J. M. (2005). Standard deviations and standard errors. *Bmj, 331*(7521), 903-903. https://doi.org/10.1136/bmj.331.7521.903

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: An overview. *Journal of Physics: Conference Series, 1142*, 12012. doi:10.1088/1742-6596/1142/1/012012

Amatya, A., & Bhaumik, D. K. (2018). Sample size determination for multilevel hierarchical designs using generalized linear mixed models: Sample size determination for multilevel hierarchical designs. *Biometrics, 74*(2), 673-684. https://doi.org/10.1111/biom.12764

Amin, M. J., & Riza, N. A. (2018). Machine learning enhanced optical distance sensor. *Optics Communications, 407*, 262-270. doi:10.1016/j.optcom.2017.09.028

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning, 50*(1), 5-43. doi:10.1023/A:1020281327116

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411-421. doi:10.1097/00006842-200405000-00021

Bae, K., & Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics, 20*(18), 3423-3430.

Barella, V. H., Garcia, L. P. F., De Souto, M. C. P., Lorena, A. C., & De Carvalho, A. C. P. L. F. (2021). Assessing the data complexity of imbalanced datasets. *Information Sciences, 553*, 83 - 109.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs, 33*(7), 1123-1131.

Bayer, I., Groth, P., & Schneckener, S. (2013). Prediction errors in learning drug response from gene expression data – influence of labeling, sample size, and machine learning algorithm. *Plos One, 8*(7), e70294-e70294. doi:10.1371/journal.pone.0070294

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica chimica acta, 760*, 25-33.

Berger, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference, 25*(3), 303-328.

Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods, 9*(1), 30.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol. 467). John Wiley & Sons.

Bolstad, W. (2007). *Introduction to Bayesian statistics* (2nd ed.). New York, NY: John Wiley.

Boulesteix, A. L., & Schmid, M. (2014). Machine learning versus statistical modeling. *Biometrical Journal, 56*(4), 588-593.

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language, 35*(1), 207-220. doi:10.1017/S030500090700829X

Brutti, P., De Santis, F., & Gubbiotti, S. (2008). Robust Bayesian sample size determination in clinical trials. *Statistics in Medicine*, *27*(13), 2290-2306.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological bulletin*, *101*(1), 147.

Byrd, R. H., Chin, G. M., Nocedal, J., & Wu, Y. (2012). Sample size selection in optimization methods for machine learning. *Mathematical Programming*, *134*(1), 127-155. doi:10.1007/s10107-012-0572-5

Bzdok, B., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*. Retrieved from https://www.nature.com/articles/nmeth.4642?source=post_page 1-5----------

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357.

Chen, C., Lee, M., Weng, C., & Leong, M. K. (2018). Theoretical prediction of the complex P-glycoprotein substrate efflux based on the novel hierarchical support vector regression scheme. *Molecules (Basel, Switzerland)*, *23*(7), 326-389

Chen, Z. Y., Fan, Z. P., & Sun, M. (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of Operational Research*, *223*(2), 461-472.

Cheng, L., & Yu, T. (2019). A new generation of AI: A review and perspective on machine learning technologies applied to smart energy and electric power systems. *International Journal of Energy Research, 43*(6), 1928-1973. doi:10.1002/er.4333

Chou, C. P., Yang, D., Pentz, M. A., & Hser, Y. I. (2004). Piecewise growth curve modeling approach for longitudinal prevention study. *Computational Statistics & Data Analysis, 46*(2), 213-225.

Choy, S. L., O'Leary, R., & Mengersen, K. (2009). Elicitation by design in ecology: Using expert opinion to inform priors for Bayesian statistical models. *Ecology*, *90*(1), 265-277. doi:10.1890/07-1886.1

Chung, K. (2016). Generating recommendations at Amazon scale with apache spark and Amazon DSSTNE. *AWS*. Retrieved from https://aws.amazon.com/blogs/big-data/generating-recommendations-at-amazon-scale-with-apache-spark-and-amazon-dsstne/

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Journal of Educational and Behavioral Statistics*, *40*(2), 136-157.

Cohen, J. (1988). The effect size index: d. *Statistical Power Analysis for the Behavioral Sciences*, *2*, 284-288.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Congdon, P. (2014). *Applied Bayesian modelling* (Second ed.). Chichester, [England]: Wiley.

Cooper, B. R., & Lanza, S. T. (2014). Who benefits most from Head Start? Using latent class moderation to examine differential treatment effects. *Child Development, 85*(6), 2317-2338.

Coussement, K., Benoit, D. F., & Antioco, M. (2015). A Bayesian approach for incorporating expert opinions into decision support systems: A case study of online consumer-satisfaction detection. *Decision Support Systems, 79,* 24-32. doi:10.1016/j.dss.2015.07.006

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we?. *Psychological bulletin, 74*(1), 68.

Cudeck, R., & Harring, J. R. (2007). Analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology, 58,* 615-637.

Cui, Z., & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *Neuroimage, 178*, 622-637. doi:10.1016/j.neuroimage.2018.06.001

Curran, P. J. (2003). Have multilevel models been structural equation models all along?. *Multivariate Behavioral Research, 38*(4), 529-569.

Curran, P. J., & Hussong, A. M. (2002). Structural equation modeling of repeated measures data: Latent curve analysis. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Multivariate applications book series. Modeling intraindividual variability with repeated measures data: Methods and applications* (pp. 59-85). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognitive Development, 11*, 121-136.

Dagne, G. A. (2019). Bayesian semiparametric growth models for measurement error and missing data in CD4/CD8 ratio: Application to AIDS study. *Statistical Methods in Medical Research*, 96228021982640. doi:10.1177/0962280219826403.

Dagne, G. A., & Ibrahimou, B. (2017). Bayesian analysis of piecewise growth mixture models with skew-t distributions: Application to AIDS studies. *Journal of Biopharmaceutical Statistics*, *27*(4), 691-704. doi:10.1080/10543406.2016.1269782

Daniels, M. J. (1999). A prior for the variance in hierarchical models. Canadian Journal of Statistics 27, 569–580.

Daniels, M. J., & Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. Chapman and Hall/CRC.

Dayan, P., & Balleine, B. W. (2002). *Reward, motivation, and reinforcement learning*. United States: Elsevier Inc. doi:10.1016/S0896-6273(02)00963-7

Dehuri, S., Ghosh, S., & Cho, S. (2011). *Integration of swarm intelligence and artificial neutral network*. London; Hackensack, N.J;: World Scientific.

DeLucia, C., & Pitts, S. C. (2006). Applications of individual growth curve modeling for pediatric psychology research. *Journal of Pediatric Psychology*, *31*(10), 1002-1023. doi:10.1093/jpepsy/jsj074.

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

Depaoli, S. (2014). The impact of inaccurate informative priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 21*, 239-252.

Depaoli, S., & Boyajian, J. (2014). Linear and nonlinear growth models: Describing a Bayesian perspective. *Journal of Consulting and Clinical Psychology, 82*(5), 784-802. doi:10.1037/a0035147.

Depaoli, S., Rus, H. M., Clifton, J. P., van de Schoot, R., & Tiemensma, J. (2017). An introduction to Bayesian statistics in health psychology. *Health Psychology Review, 11*(3), 248-264.

Depaoli, S., Yang, Y., & Felt, J. (2017). Using Bayesian statistics to model uncertainty in mixture models: A sensitivity analysis of priors. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(2), 198-215. doi:10.1080/10705511.2016.1250640.

De Raedt, L., & Kimmig, A. (2015). Probabilistic (logic) programming concepts. *Machine Learning, 100*(1), 5-47. doi:10.1007/s10994-015-5494-z

De Rosa, R., & Aragona, B. (2017). Unpacking big data in education. A research framework. *Statistics, Politics and Policy, 8*(2), 123-137. doi:10.1515/spp-2017-0014

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR), 27*(3), 326-327.

Dixon, S. J., Heinrich, N., Holmboe, M., Schaefer, M. L., Reed, R. R., Trevejo, J., & Brereton, R. G. (2009). Application of classification methods when group sizes are unequal by incorporation of prior probabilities to three common approaches: Application to simulations and mouse urinary chemo signals. *Chemometrics and Intelligent Laboratory Systems, 99*(2), 111-120. doi:10.1016/j.chemolab.2009.07.016

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

Duncan, T. E., & NetLibrary, I. (1999). *An introduction to latent variable growth curve modeling*: Concepts, issues, and applications. Mahwah, N.J: L. Erlbaum Associates.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62*, 355-366.

Dunson, D. B. (2001). Commentary: Practical advantages of Bayesian analysis of epidemiologic data. *American journal of Epidemiology, 153*(12), 1222-1226.

Eggert, A., Hogreve, J., Ulaga, W., & Muenkhoff, E. (2011). Industrial services, product innovations, and firm profitability: A multiple-group latent growth curve analysis. *Industrial Marketing Management, 40*(5), 661-670. doi:10.1016/j.indmarman.2011.05.007

Escudero, J., Zajicek, J. P., & I feachor, E. (2011). Machine learning classification of MRI features of Alzheimer's disease and mild cognitive impairment subjects to reduce the sample size in clinical trials. Conference Proceedings: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference, 2011*, 7957.

Fabris, F., Magalhães, J. P. d., & Freitas, A. A. (2017). A review of supervised machine learning applied to ageing research. *Biogerontology*, *18*(2), 171-188. doi:10.1007/s10522-017-9683-y

Fawcett, L., Thorpe, N., Matthews, J., & Kremer, K. (2017). A novel Bayesian hierarchical model for road safety hotspot prediction. *Accident Analysis & Prevention*, *99*, 262-271.

Fernández, C., & Steel, M. F. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, *93*(441), 359-371.

Ferrer, E., & McArdle, J. J. (2004). An experimental analysis of dynamic hypotheses about cognitive abilities and achievement from childhood to early adulthood. *Developmental Psychology*, *40*(6), 935.

Field, A. (2011). *Discovering Statistics Using SPSS.* (3rd ed.). (pp. 15-18). Thousand Oaks, : SAGE Publications

Finch, W. H., & Miller, J. E. (2019). The use of incorrect informative priors in the estimation of MIMIC model parameters with small sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 497-508.

Flora, D. B. (2008). Specifying piecewise latent trajectory models for longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal, 15*(3), 513-533.

Gagné, P., & Hancock, G. (2002). *Relation of sample size and solution propriety in latent variable models as a function of construct reliability.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. April, 2002.

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association, 99*(466), 537-545.

Gelman, A. (2006a). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics, 48*(3), 432-435.

Gelman, A. (2006b). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis, 1*(3)https://doi.org/10.1214/06-BA117A

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2015). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: CRC Press.

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics, 40*(5), 530–543. https://doi.org/10.3102/1076998615606113.

Gentry, W. A., & Martineau, J. W. (2010). Hierarchical linear modeling as an example for measuring change over time in a leadership development evaluation context. *The Leadership Quarterly, 21*(4), 645-656. doi:10.1016/j.leaqua.2010.06.007

Ghaffari, M. H., Jahanbekam, A., Sadri, H., Schuh, K., Dusel, G., Prehn, C., & Sauerwein, H. (2019). Metabolomics Meets Machine Learning: Longitudinal Metabolite Profiling in Serum of Normal Versus Overconditioned Cows and Pathway Analysis. *Journal of Dairy Science, 102*(12), 11561 - 11585.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452-459. doi:10.1038/nature14541

Gibbons, R. D., Hedeker, D., & DuToit, S. (2010). Advances in analysis of longitudinal data. *Annual review of clinical psychology, 6*, 79-107.

Gill, J. (2015). *Bayesian Methods: A Social and Behavioral Sciences Approach* (Third ed.). Boca Raton, FL: CRC Press, Taylor & Francis Group.

Golchi, S. (2018). Informative priors in Bayesian inference and computation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, doi:10.1002/sam.11371

Gonzalez, J., & Leboulluec, A. (2019). Crime prediction and socio-demographic factors: A comparative study of machine learning regression-based algorithms. *Journal of Applied Computer Science & Mathematics, 13*(1), 13-18. doi:10.4316/JACSM.201901002

González-Recio, O., & Forni, S. (2011). Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genetics Selection Evolution, 43*(1), 1-12.

Graham, C., & Talay, D. (2013). *Stochastic simulation and monte carlo methods: Mathematical foundations of stochastic simulation* (2013th ed.). New York; Berlin;: Springer. doi:10.1007/978-3-642-39363-1

Greenland, S., Schwartzbaum, J. A., & Finkle, W. D. (2000). Problems due to small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology, 151*(5), 531-539.

Grimm, K., & Ram, N. (2009). Non-linear growth models in M*plus* and SAS. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 676-701.

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. Retrieved from https://ebookcentral-proquest-com.proxy01.its.virginia.edu

Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child Development, 82,* 1357-1371.

Grover, D., Bauhoff, S., & Friedman, J. (2019). Using supervised learning to select audit targets in performance-based financing in health: An example from Zambia. *PloS One, 14*(1), e0211262. doi:10.1371/journal.pone.0211262

Guo, H. X., Liu, Y. Q., Wu, J., & Yang, J. M. (2004). A reinforcement learning approach to STATCOM controller. In *2004 IEEE International Conference on Electric Utility Deregulation, Restructuring and Power Technologies. Proceedings* (Vol. 2, pp. 638-642). IEEE.

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Lee, M. J., & Asadi, H. (2018). eDoctor: Machine learning and the future of medicine. *Journal of Internal Medicine, 284*(6), 603-619. doi:10.1111/joim.12822

Hang, W., & Banks, T. (2019). Machine learning applied to pack classification. *International Journal of Market Research, 61*(6), 601-620. doi:10.1177/1470785319841217

Harris, G. T., & Rice, M. E. (2013). Bayes and base rates: what is an informative prior for actuarial violence risk assessment?. *Behavioral Sciences & the Law*, *31*(1), 103-124.

Harris, T. M., Devkota, J. P., Khanna, V., Eranki, P. L., & Landis, A. E. (2018). Logistic growth curve modeling of US energy production and consumption. *Renewable and Sustainable Energy Reviews, 96*, 46-57. doi:10.1016/j.rser.2018.07.049

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83-85.

Hatch, C. (2018). Be in the Know: 2018 Ecommerce Statistics You Should Know. Retrieved May 15, 2019, from https://www.disruptiveadvertising.com/ppc/ecommerce/2018-ecommerce-statistics/

Heck, G. S., Pintro, V. O., Pereira, R. R., de Ávila, M. B., Levin, N. M. B., & de Azevedo, W. F. (2017). Supervised machine learning methods applied to predict ligand- binding affinity. *Current Medicinal Chemistry, 24*(23), 2459.

Herrman, J. (2016). Media website battle faltering ad revenue and traffic. *New York Time.* Retrieved from https://www.nytimes.com/2016/04/18/business/media-websites-battle-falteringad-revenue-and-traffic.html

Hess, A. S., & Hess, J. R. (2016). Understanding standard deviations and standard errors: SDs AND SEs. *Transfusion (Philadelphia, Pa.), 56*(6), 1259-1261. https://doi.org/10.1111/trf.13625

Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC medical research methodology*, *20*(1), 1-12.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329-367.

Hosack, G. R., Hayes, K. R., & Barry, S. C. (2017). Prior elicitation for Bayesian generalised linear models with application to risk control option assessment. *Reliability Engineering and System Safety, 167,* 351-361. doi:10.1016/j.ress.2017.06.011.

Houghton, K. A. (1984). Accounting data and the prediction of business failure: the setting of priors and the age of data. *Journal of accounting Research*, 361-368.

Howitt, D., & Cramer, D. (2011). *Introduction to Research Methods in Psychology*. (3rd ed.). (pp. 164, 179-181). Harlow, Essex: Pearson Education Limited

Hox, J. (2002). Multilevel analysis: *Techniques and applications*. Mahwah,NJ: Erlbaum

Idowu, S., Saguna, S., Åhlund, C., & Schelén, O. (2016). Applied machine learning: Forecasting heat load in district heating system. *Energy & Buildings*, *133*, 478-488. doi:10.1016/j.enbuild.2016.09.068

Jackson, D. L. (2010). Reporting results of latent growth modeling and multilevel modeling analyses: Some recommendations for rehabilitation psychology. *Rehabilitation Psychology*, *55*(3), 272.

Jaggars, S. S., & Xu, D. (2016). Examining the earnings trajectories of community college students using a piecewise growth curve modeling approach. *Journal of Research on Educational Effectiveness*, *9*(3), 445-471

Jakhar, D., & Kaur, I. (2019). Artificial intelligence, machine learning & deep learning: Definitions and differences. *Clinical and Experimental Dermatology*, doi:10.1111/ced.14029

Jana, S., Balakrishnan, N., & Hamid, J. S. (2019). Bayesian growth curve model useful for high-dimensional longitudinal data. *Journal of Applied Statistics*, *46*(5), 814-834. doi:10.1080/02664763.2018.1517145.

Jarociński, M., & Marcet, A. (2019). Priors about observables in vector autoregressions. *Journal of Econometrics, 209*(2), 238-255. https://doi.org/10.1016/j.jeconom.2018.12.023

Kallenberg, M., Petersen, K., Nielsen, M., Ng, A. Y., Diao, P., I gel, C., Vachon, C., Holland, K., Winkel, R., Karssemeijer, N., & Lillholm, M. (2016). Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Transactions on Medical Imaging*, *35*(5), 1322 - 1330.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773-795.

Kasun, L. L. C., Zhou, H., Huang, G. B., & Vong, C. M. (2013). Representational learning with extreme learning machine for big data. *IEEE Intelligent Systems, 28*(6), 31-34.

Kim, S. Y., Huh, D., Zhou, Z., & Mun, E. Y. (2020). A comparison of Bayesian to maximum likelihood estimation for latent growth models in the presence of a binary outcome. *International Journal of Behavioral Development*0165025419894730.

Kim, Y., Lee, J., Kim, J., Soliman, G., & Wehbi, N. K. (2015). Longitudinal associations between BMI, physical activity, and healthy diet: A parallel latent growth curve modeling. *Medicine & Science in Sports & Exercise, 47*, 176 - 190. doi:10.1249/01.mss.0000476899.36821.42

Kim, Y. G., Kwon, Y., & Paik, M. C. (2019). Valid oversampling schemes to handle imbalance. *Pattern Recognition Letters*, *125*, 661-667.

Koduvely, H. M. (2015). *Learning Bayesian Models With R: Become an Expert in Bayesian Machine Learning Methods Using R and Apply Them to Solve Real-World Big Data Problems*. Packt Publishing.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine, 24*(11), 1716-1720. doi:10.1038/s41591-018-0213-5

Konerman, M. A., Zhang, Y., Zhu, J., Higgins, P. D., Lok, A. S., & Waljee, A. K. (2015). Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data. *Hepatology*, *61*(6), 1832-1841.

Kosugi, K. I. (1996). Lognormal distribution model for unsaturated soil hydraulic properties. *Water Resources Research*, *32*(9), 2697-2703.

Kwok, O. M., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R., & Yoon, M. (2008). Analyzing longitudinal data with multilevel models: an example with individuals living with lower extremity intra-articular fractures. *Rehabilitation psychology*, *53*(3), 370.

Ladds, M. A., Thompson, A. P., Slip, D. J., Hocking, D. P., & Harcourt, R. G. (2016). Seeing it all: evaluating supervised machine learning methods for the classification of diverse otariid behaviours. *PloS One*, *11*(12), e0166898.

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

Lee, J. C., & Chang, C. H. (2000). Bayesian analysis of a growth curve model with a general autoregressive covariance structure. *Scandinavian Journal of Statistics*, *27*, 703–713.

Lee, J. C., & Liu, K.-C. (2000). Bayesian analysis of a general growth curve model with predictions using power transformations and AR(1) autoregressive dependence. *Journal of Applied Statistics*, *27*, 321–336.

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*(1), 114-127.

Leiby, B. E. (2006). *Bayesian multivariate growth curve latent class models* (Doctoral dissertation, University of Pennsylvania).

Leiby, B. E., Ten Have, T. R., Lynch, K. G., & Sammel, M. D. (2014). Bayesian multivariate growth curve latent class models for mixed outcomes. *Statistics in Medicine, 33*(20), 3434-3452. doi:10.1002/sim.5596

Leonard, T. (1975). A Bayesian approach to the linear model with unequal variances. Technometrics, *17*(1), 95-102.

Leroux, A. J. (2019). Student mobility in multilevel growth modeling: A multiple membership piecewise growth model. *The Journal of Experimental Education, 87*(3), 430-448. doi:10.1080/00220973.2018.1465384.

Levy, R. (2016). Advances in Bayesian Modeling in Educational Research. *Educational Psychologist, 51*(3/4), 368 - 380.

Li, F., Duncan, T. E., Duncan, S. C., & Acock, A. (2001). Latent growth modeling of longitudinal data: A finite growth mixture modeling approach. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(4), 493-530. doi:10.1207/S15328007SEM0804_01.

Lininger, M., Spybrook, J., & Cheatham, C. C. (2015). Hierarchical linear model: Thinking outside the traditional repeated-measures analysis-of-variance box. *Journal of Athletic Training, 50*(4), 438-441. doi:10.4085/1062-6050-49.5.09

Little, R. J. (1999). Methods for handling missing values in clinical trials. *The Journal of Rheumatology, 26*(8), 1654.

Liu, A. X., Wang, Y., Chen, X., & Jiang, X. (2014). Understanding the diffusion of mobile digital content: A growth curve modelling approach. *Information Systems and e-Business Management, 12*(2), 239-258. doi:10.1007/s10257-013-0224-1

Long, J. D., & Mills, J. A. (2018). Joint modeling of multivariate longitudinal data and survival data in several observational studies of Huntington's disease. *BMC medical research methodology*, *18*(1), 1-15.

Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., & Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. *Scientific Reports*, *3*, 1236.

Luque, A., Carrasco, A., Martin, A., & De las Heras, A. (2019). The impact of Class imbalance in classification performance metrics based on the binary bonfusion matrix. *Pattern Recognition*, *91*, 216 - 231.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 86-92. doi:10.1027/1614-2241.1.3.86

MacAulay, R. K., Calamia, M. R., Cohen, A. S., Daigle, K., Foil, H., Brouillette, R., Bruce-Keller, A. J., & Keller, J. N. (2018). Understanding heterogeneity in older adults: Latent growth curve modeling of cognitive functioning. *Journal of Clinical and Experimental Neuropsychology, 40*(3), 292-302. doi:10.1080/13803395.2017.1342772

Malhotra, R. (2015). A systematic review of machine learning techniques for software fault prediction. *Applied Soft Computing, 27*, 504-518.

Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Computer methods and programs in biomedicine, 152*, 23-34.

Marsh, H., Hau, K., Balla, J., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181-220.

Martin, O. (2018). Bayesian analysis with Python: *introduction to statistical modeling and probabilistic programming using PyMC3 and ArviZ*. Packt Publishing Ltd.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data.* Pacific Grove, CA: Brooks/Cole.

McArdle, J. J. (1988). *Dynamic but structural equation modeling of repeated measures data*. In J. R. Nesselroade & R. B. Cattell (Eds.), Handbook of Multivariate Experimental Psychology (pp. 561–614). New York: Plenum Press, 2nd edition.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58*, 110–133.

McArdle, J. J., & Horn, J. L. (2004). *A mega analysis of the WAIS: Adult intelligence across the life-span*. Mahwah, NJ: Lawrence Erlbaum.

McArdle, J. J., & Nesselroade, J. R. (2003). Growth curve analysis in contemporary psychological research. In J. Schinka & W. Velicer (Eds.), *Comprehensive handbook of psychology: Research methods in psychology* (Vol. 2, p. 447–480).

New York: Wiley.

McCarthy, M. A., & Masters, P. (2005). Profiting from prior information in Bayesian analyses of ecological data. *Journal of Applied Ecology, 42*(6), 1012-1019. doi:10.1111/j.1365-2664.2005.01101.x

McCoach, D. B., & Kaniskan, B. (2010). Using time-varying covariates in multilevel growth models. *Frontiers in psychology, 1*, 17 -30.

Meng, Y., Li, G., Gao, Y., Lin, W., & Shen, D. (2016). Learning-based subject-specific estimation of dynamic maps of cortical morphology at missing time points in longitudinal infant studies. *Human Brain Mapping, 37*(11), 4129-4147. doi:10.1002/hbm.23301

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107–122.

Miller, C., Nagy, Z., & Schlueter, A. (2018). A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. *Renewable and Sustainable Energy Reviews, 81*, 1365-1377. doi:10.1016/j.rser.2017.05.124

Mok, M. (1995). Sample size requirements for 2-level designs in educational research. *Multilevel Modelling Newsletter, 7*(2), 11–15.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *The Journal of Economic Perspectives, 31*(2), 87-106. doi:10.1257/jep.31.2.87

Muthén, B. (1997). Latent variable modeling of longitudinal and multilevel data. *Sociological Methodology, 27*(1), 453-480.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological methods*, *17*(3), 313.

Muthén, B., & Muthén, L. (2002). How to use a Monte Carlo study to decide on sample size and determine power. Preliminary version of a paper scheduled to appear in *Structural Equation Modeling. [What was the source of this paper? E.g., was there a website where you found it?]*

Nilsson, N. J. (1986). *Principles of artificial intelligence*. Los Altos, Calif: Morgan Kaufmann Publishers.

Ocharo, H. N., & Hasegawa, S. (2018). Using machine learning to classify reviewer comments in research article drafts to enable students to focus on global revision. *Education and Information Technologies, 23*(5), 2093-2110. doi:10.1007/s10639-018-9705-7

Oluwadare, O., & Cheng, J. (2017). ClusterTAD: An unsupervised machine learning approach to detecting topologically associated domains of chromosomes from hi-C data. *BMC Bioinformatics, 18*(1), 480-14. doi:10.1186/s12859-017-1931-2

Oravecz, Z., & Muthén, C. (2018). Fitting growth curve models in the Bayesian framework. *Psychonomic Bulletin & Review, 25*(1), 235-255. doi:10.3758/s13423-017-1281-0

Osvaldo, S. S., Lopes, D., Silva, A. C., & Abdelouahab, Z. (2017). Developing software systems to Big Data platform based on MapReduce model: An approach based on Model Driven Engineering. *Information and Software Technology*, *92*, 30-48.

Parslow, J., Cressie, N., Campbell, E. P., Jones, E., & Murray, L. (2013). Bayesian learning and predictability in a stochastic nonlinear dynamical model. *Ecological Applications, 23*(4), 679-698. doi:10.1890/12-0312.1

Paruggia, M. (2006). *Sensitivity analysis in practice: A guide to assessing scientific models*. Alexandria: Taylor & Francis. doi:10.1198/jasa.2006.s80

Pedersen, T. B., & Jensen, C. S. (1999). Multidimensional data modeling for complex data. In *Proceedings 15th International Conference on Data Engineering (Cat. No. 99CB36337)* (pp. 336-345). IEEE.

Pezeshk, H. (2003). Bayesian techniques for sample size determination in clinical trials: A short review. *Statistical Methods in Medical Research, 12*(6), 489-504. doi:10.1191/0962280203sm345oa

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and psychological measurement, 64*(6), 916-924.

Preacher, K. J. (2010). Latent growth curve models. *The reviewer's guide to quantitative methods in the social sciences, 1*, 185-198.

Raghunathan, T. (2015). *Missing data analysis in practice*. CRC press.

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning.* Birmingham, UK: Packt Publishing Ltd.

Rasmussen, C. E., & Ghahramani, Z. (2003). Bayesian Monte Carlo. *Advances in Neural Information Processing Systems*, 30(84) 505-512.

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

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 112* (3), 252-264.

Ravenzwaaij, D. V., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review, 25*(1), 143-154.

Richardson, S., & Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 59*, 731-792.

Robert, C. (2014). *Machine learning, a probabilistic perspective*. Abingdon: Taylor & Francis. doi:10.1080/09332480.2014.914768

Roberts, G. O., & Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics, 18*(2), 349-367. doi:10.1198/jcgs.2009.06134

Roeder, K., & Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association, 92*, 894-902.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*(3), 726.

Röver, C., Bender, R., Dias, S., Schmid, C. H., Schmidli, H., Sturtz, S., Weber, S., & Friede, T. (2021). On weakly informative prior distributions for the heterogeneity parameter in

bayesian random-effects meta-analysis. *Research Synthesis Methods*, https://doi.org/10.1002/jrsm.1475

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 42-47). IEEE.

Salakhutdinov, R., & Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning* (pp. 880-887).

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Sambasivan, R., Das, S., & Sahu, S. K. (2020). A bayesian perspective of statistical machine learning for big data. *Computational Statistics*, *35*(3), 893-930. https://doi.org/10.1007/s00180-020-00970-8

Sargeant, A. (2013). *Donor Retention: What Do We Know & What Can We Do about It?* (pp. 12-23). Nonprofit Quarterly.

Sathiaraj, D., Huang, X., & Chen, J. (2019). Predicting climate types for the continental united states using unsupervised clustering techniques. *Environmetrics*, *30*(4), e2524-n/a. doi:10.1002/env.2524

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*(1), 37-52.

Schnack, H. G., & Kahn, R. S. (2016). Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Frontiers in Psychiatry [E]*, *7*, 50-50. doi:10.3389/fpsyt.2016.00050

Schonfeld, I. S., & Rindskopf, D. (2007). Hierarchical linear modeling in organizational research: Longitudinal data outside the context of growth modeling. *Organizational Research Methods*, *10*(3), 417-429.

Schrodt, F., Kattge, J., Shan, H., Fazayeli, F., Joswig, J., Banerjee, A., & Reich, P. B. (2015). BHPMF – a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Global Ecology and Biogeography*, *24*(12), 1510-1521. doi:10.1111/geb.12335

Schwaighofer, A., Tresp, V., & Yu, K. (2005). Learning Gaussian process kernels via hierarchical Bayes. In *Advances in neural information processing systems* (pp. 1209-

Seaman, J. W., Seaman, J. W. Jr., & Stamey, J. D. (2012). Hidden dangers of specifying noninformative priors. The American Statistician, 66(2), 77-84.

Seltzer, M. H., Wong, W. H. M., & Bryk, A. S. (1996). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational Statistics*, *21*, 131–167.

Shaw, R. G., & Mitchell-Olds, T. (1993). ANOVA for unbalanced data: an overview. *Ecology*, *74*(6), 1638-1645.

Shi, D., & Tong, X. (2017). The Impact of Prior Information on Bayesian Latent Basis Growth Model Estimation. *SAGE Open*. https://doi.org/10.1177/2158244017727039

Shieh, G., & Lee, J. C. (2002). Bayesian prediction analysis for growth curve model using noninformative priors. *Annals of the Institute of Statistical Mathematics, 54*(2), 324-337. doi:10.1023/A:1022474018976

Shirzadi, A., Solaimani, K., Roshan, M. H., Kavian, A., Chapi, K., Shahabi, H., & Bui, D. T. (2019). Uncertainties of prediction accuracy in shallow landslide modeling: Sample size and raster resolution. *Catena*, *178*, 172-188.

Shu, Z., Henson, R., & Willse, J. (2013). Using neural network analysis to define methods of DINA model estimation for small sample sizes. *Journal of Classification, 30*(2), 173-194.

Sigley, R. (2003). The importance of interaction effects. *Language Variation and Change*, *15*(2), 227-253.

Singer, J. D. (1998). Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. *Journal of Educational and Behavioral Statistics*, *23*(4), 323–355. https://doi.org/10.3102/10769986023004323

Song, X., & Lee, S. (2012). *Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences*. Chichester, West Sussex: John Wiley.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*(4), 583-639.

Stadtfeld, C., Snijders, T. A. B., Steglich, C., & van Duijn, M. (2018). Statistical power in longitudinal network studies. *Sociological Methods & Research, 49*(4), 1103-1132. https://doi.org/10.1177/0049124118769113

Stamey, J. D., Beavers, D. P., & Sherr, M. E. (2017). Bayesian analysis and design for joint modeling of two binary responses with misclassification. *Sociological Methods & Research, 46*(4), 772-792. doi:10.1177/0049124115605340

Stewart, M. (2019). The Actual Difference Between Statistics and Machine Learning. Retrieved from https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3

Stockwell, D. R., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological modelling*, *148*(1), 1-13.

Su, J., Zhang, H., Ling, C. X., & Matwin, S. (2008). Discriminative parameter learning for bayesian networks. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 1016-1023). ACM.

Sun, D., & Ni, S. (2004). Bayesian analysis of vector-autoregressive models with noninformative priors. *Journal of Statistical Planning and Inference, 121*(2), 291-309. doi:10.1016/S0378-3758(03)00116-2

Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review, 41*(4), 70-73.

Swaminathan, H., & Rogers, J. H. (2008). Estimation procedures for hierarchical linear models. In A. A.

Tabachnick, B. G., & Fidell, L. S. (2001). *Experimental designs using ANOVA* (p. 724). Belmont, CA: Thomson/Brooks/Cole.

Thomas, S. A., Feng, S., & Krishnan, T. V. (2015). To retain? To upgrade? The effects of direct mail on regular donation behavior. *International Journal of Research in Marketing*, *32*(1), 48-63.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, *1*(1), 31-65. doi:10.1146/annurev.clinpsy.1.102803.144239

Tong, X., & Zhang, Z. (2020). Robust Bayesian Approaches in Growth Curve Modeling: Using Student'st Distributions versus a Semiparametric Method. Structural Equation Modeling: *A Multidisciplinary Journal*, *27*(4), 544-560.

Tufekci, Z. (2018). YouTube, the great radicalizer. *The New York Times*, *10*, 23.

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. A., Elkhatib, Y., Hussain, A., & Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE Access*, *7*, 65579-65615. doi:10.1109/ACCESS.2019.2916648

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, *14*(11), e0224365.

VanderWeele, T. J. (2019). Principles of confounder selection. *European journal of epidemiology*, *34*(3), 211-219.

Van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default bayesian structural equation modeling. *Psychological Methods, 23*(2), 363-388. https://doi.org/10.1037/met0000162

Vasantha, M., & Venkatesan, P. (2014). Structural equation modeling of latent growth curves of weight gain among treated tuberculosis patients. *PloS One, 9*(3), e91152-e91152. doi:10.1371/journal.pone.0091152.

Vasey, M. W., & Thayer, J. F. (1987). The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. *Psychophysiology, 24*(4), 479-486.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and computing, 27(5), 1413-1432.

Von Ende, C. N. (2001). Repeated-measures analysis. *Design and Analysis of Ecological Experiments, 8,* 134-157.

Vranas, K. C., Jopling, J. K., Sweeney, T. E., Ramsey, M. C., Milstein, A. S., Slatore, C. G., & Liu, V. X. (2017). Identifying Distinct Subgroups of Intensive Care Unit Patients: a Machine Learning Approach. *Critical Care Medicine, 45*(10), 1607.

Wahab, L., & Jiang, H. (2019). A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. *PloS One, 14*(4), e0214966. doi:10.1371/journal.pone.0214966

Walls, L., & Quigley, J. (2001). Building prior distributions to support bayesian reliability growth modelling using expert judgement. *Reliability Engineering and System Safety, 74*(2), 117-128. doi:10.1016/S0951-8320(01)00069-2

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2018). Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *Journal of Child Psychology and Psychiatry, 59*(12), 1261-1270. doi:10.1111/jcpp.12916

Wang, F., & Gelfand, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, *17*(2), 193-208.

Wang, L., & Preacher, K. J. (2015). Moderated mediation analysis using Bayesian methods. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(2), 249-263.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research 14*, 867–897.

White, S. R., Muniz-Terrera, G., & Matthews, F. E. (2018). Sample size and classification error for Bayesian change-point models with unlabelled sub-groups and incomplete follow-up. *Statistical Methods in Medical Research, 27*(5), 1476-1497. doi:10.1177/0962280216662298

Willett, J. B., & Bub, K. L. (2014). Structural equation modeling: Latent growth curve analysis. *Wiley StatsRef: Statistics Reference Online*.

Willett, L. H. (1989). Are two-year college students first-generation college students? *Community College Review*, 17(2), 48-52.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79-82.

Witten, I. H., Frank, E., Hall, M. A., & Holmes, G. (2011). Data mining: Practical machine learning tools and techniques (3rd ed.). Amsterdam: Elsevier/Morgan Kaufmann.

Wolf, F. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA: Sage.

Wolfinger, R. D., & Kass, R. E. (2000). Nonconjugate Bayesian analysis of variance component models. *Biometrics*, *56*(3), 768-774.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology, 8*(1), 52-69. doi:10.20982/tqmp.08.1.p052

Wu, J., Roy, J., & Stewart, W. F. (2010). Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical Care*, S106-S113.

Xanthopoulos, P. (2014). A review on consensus clustering methods. In T. M. Rassias, C. A. Floudas & S. Butenko (Eds.), *Optimization in Science and Engineering* (pp. 553–566). New York: Springer.

Yap Bee Wah, H. A. A., Haibo, H., & Bulgiba, A. (2016). Handling Imbalanced Dataset Using SVM and k-NN Approach. *AIP Conference Proceedings*, 1750(1), 1 - 8.

Yin, K., Choudhary, P. K., Varghese, D., & Goodman, S. R. (2008). A Bayesian approach for sample size determination in method comparison studies. *Statistics in Medicine, 27*(13), 2273-2289. doi:10.1002/sim.3124

Yousefi, S., Yousefi, E., Takahashi, H., Hayashi, T., Tampo, H., I noda, S., Arai, Y., & Asbell, P. (2018). Keratoconus severity identification using unsupervised machine learning. *PloS One, 13*(11), e0205998. doi:10.1371/journal.pone.0205998

Yu, B., & Xu, Z. (2008). A comparative study for content-based dynamic spam classification using four machine learning algorithms. *Knowledge-Based Systems, 21*(4), 355-362. doi:10.1016/j.knosys.2008.01.001

Zahn, I. (2010). Working with unbalanced cell sizes in multiple regression with categorical predictors. *Education*,112(84)1-18.

Zeng, X., & Luo, G. (2017). Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection. Health Information Science and Systems, 5(1), 2 - 1. Retrieved from https://doi.org/10.1007/s13755-017-0023-z

Zhang, Y., Dukic, V., & Guszcza, J. (2012). A Bayesian non-linear model for forecasting insurance loss payments. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 175*(2), 637-656. doi:10.1111/j.1467-985X.2011.01002.x

Zhang, Z. (2016). Modeling error distributions of growth curve models through Bayesian methods. *Behavior Research Methods, 48*(2), 427-444. doi:10.3758/s13428-015-0589-9

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*, 374-383.

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(4), 646-661. doi:10.1080/10705511.2018.1545232

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., & Van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development, 14*(4), 305-320. doi:10.1080/15427609.2017.1370966