STATISTICAL DEEP LEARNING OF MULTIVARIATE LONGITUDINAL DATA

Yunyi Li

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biostatistics and Health Data Science,
Indiana University

November 2024

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Sujuan Gao, Ph.D., Co-Chair

Hao Liu, Ph.D., Co-Chair

Doctoral Committee

Liana G. Apostolova, M.D., M.Sc., FAAN.

Xiaochun Li, Ph.D.

September 26, 2024

Yi Zhao, Ph.D.

DEDICATION

To my lovely family: my father (Jiguang Li), my mother (Ping Chen), my sister (Na Yang), my husband (Ruijie Yin), and my dog (Taffy). Your positive attitude and encouragement go beyond what words can adequately express. Your unwavering support has been my anchor throughout this journey, and I am deeply grateful for your love and belief in me.

ACKNOWLEDGMENTS

The completion of this study could not have been possible without the expertise of my thesis advisors, Dr. Sujuan Gao and Dr. Hao Liu. Foremost, I would like to express sincere gratitude to Dr. Gao and Dr. Liu for their constant guidance, encouragement, support, patience, and faith in me.

I would also like to thank my committee members, Dr. Liana G. Apostolova, Dr. Xiaochun Li, and Dr. Yi Zhao, for their brilliant comments and suggestions. Their insightful questions made me evaluate my study from various valuable angles. My warmest thanks also go to Dr. Liana G. Apostolova, for offering me great opportunities in her lab, and for providing funding support from the Early Onset Alzheimers Disease Consortium - The Longitudinal Early Onset Alzheimer's Disease Study (LEADS) (NIH U01AG057195), and for leading me to work on diverse exciting scientific projects. Her enthusiasm for research has been a constant source of encouragement.

It has been my great honor to spend five years in such a wonderful department. I thank all the lovely faculty members, staff, professors, and my best friends for all the help, stimulating academic discussions, the delicious food we shared, and the fun we had over the years.

Last but not least, I would like to thank my family: my parents, my sister, my husband, and my dog for their unwavering support and for bringing joy to my life every day.

Yunyi Li

STATISTICAL DEEP LEARNING OF MULTIVARIATE LONGITUDINAL DATA

Nowadays, various types of longitudinal data, including continuous, binary, and count data, are increasingly collected in numerous scientific research fields such as Alzheimer's disease studies. Despite the wealth of data, the complex structure of multivariate longitudinal data presents significant modeling challenges. For years, scientific research has been actively exploring dynamic interactions among multiple components and understanding how interventions can impact outcomes over time with complex underlying dynamics. However, statistical methods for modeling these dynamic changes and associations are still limited.

To address these gaps, we propose a novel nonparametric method to describe the mean temporal changes of sparsely and irregularly observed multivariate longitudinal data. This method is based on an Ordinary Differential Equation (ODE) system approximated by neural networks. Furthermore, we presented a novel approach to treat the initial values of ODEs as an unknown parameter vector, a departure from existing methods that either pre-specify the initial values or estimate them in an ad hoc manner.

In the second topic, we propose deep latent ODE models. These models nonparametrically model latent temporal trends by an unknown function of an ODE system and parametrically estimate the effects of covariates using Bayesian approaches. To address the intractability of the posterior distribution of initial values, we employ a variational autoencoder (VAE) algorithm. The approximate posterior distribution is characterized by a recurrent neural network (RNN), and high dimensional hy-

perparameters are estimated using the stochastic gradient descent method based on Kullback-Leibler (KL) divergence.

Lastly, we propose Bayesian generalized random effects models for modeling longitudinal data from various distributions, including longitudinal counts, and longitudinal binary outcomes. This model extends traditional generalized linear mixed effect models (GLMMs) to generalized semi-parametric mixed effect models. It assumes a nonparametric baseline function with a stochastic process prior, and parameters are estimated using the Bayesian approach. The proposed model is practical and can be applied to various types of longitudinal data, including longitudinal binary, and count data. Neural ODE, RNN, variational inference, and KL divergence techniques are also applied in this project.

Sujuan Gao, Ph.D., Co-Chair

Hao Liu, Ph.D., Co-Chair

Liana G. Apostolova, M.D., M.Sc., FAAN.

Xiaochun Li, Ph.D.

Yi Zhao, Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Overview

As data collection technologies advance, various types of longitudinal data are collected in scientific studies. Besides traditional visit records, such as blood pressure, cholesterol, and glucose levels, more data with various data types are collected, such as cognitive tests, gene expression, metabolomics, biomarkers, and imaging data. In longitudinal studies, each participant undergoes baseline tests, and subsequent evaluations are conducted at follow-up times over several years. Although follow-up time schedules were assigned, the actual follow-up times can be random and differ among participants. Additionally, the number of follow-ups for each participant and for each component of the same participant may vary, which results in sparsely observed multivariate longitudinal data within irregularly spaced follow-up times. Other than the irregularly spaced follow-up times, the complex correlations among multiple components of multivariate longitudinal data can be challenging to model. Moreover, in practical scientific studies, the temporal changes of longitudinal data can be influenced by unobservable latent dynamics. Such complex data structures and characteristics raise big challenges and enthusiasm for modeling temporal associations among multivariate longitudinal observed data and inferring intervention impacts by controlling latent dynamics or random effects for decades.

1.2 Ordinary Differential Equations for Multivariate Dynamics

1.2.1 Paramteric and Semi-parametric ODE Models

In the realm of hypothetical and mathematical research, ordinary differential equations (ODEs) have been successfully used in theoretical and mathematical research to model temporal interactions and associations among multivariate dynamics, such as biological systems, brain energy metabolic systems, brain networks, etc. (Garbarino et al., 2019; Ranjan et al., 2018; Tiveci et al., 2005). The dynamic changes of the system are described by a set of ODEs. For example, consider a system with K components, $x_1(t), x_2(t), \ldots, x_K(t)$, a general ODE model can be described by the following set of functions:

$$\begin{cases} \dfrac{dx_1(t)}{dt} = f_1(x_1(t), x_2(t), \ldots, x_K(t)) \\ \dfrac{dx_2(t)}{dt} = f_2(x_1(t), x_2(t), \ldots, x_K(t)) \\ \qquad\qquad \vdots \\ \dfrac{dx_K(t)}{dt} = f_n(x_1(t), x_2(t), \ldots, x_K(t)) \end{cases}$$

where $\boldsymbol{x}(t) = (x_1(t), x_2(t), \ldots, x_K(t))'$ is a vector representing the values of components $1, \ldots, K$ at a specific time t. $\frac{dx_1(t)}{dt}, \frac{dx_2(t)}{dt}, \ldots, \frac{dx_K(t)}{dt}$ represent the rates of change for each component in the system. The function set $\boldsymbol{f}(\cdot) = (f_1(\cdot), \ldots, f_K(\cdot))'$ acts on the rate of change of the current component and serves as the link function that quantifies the associations among all components. In general, $\boldsymbol{f}(\cdot)$ can take any linear or non-linear function forms.

In previous studies, researchers have proposed parametric (Lu et al., 2011) and semi-parametric (Henderson and Michailidis, 2014; Wu et al., 2014; Xue et al., 2019) methods to estimate the ODEs for handling practical problems. A simple linear ODE model (Lu et al., 2011) can be written as follows:

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{n} \theta_{ij} x_j(t), i = 1, \ldots, n$$

where θ_{ij} is the parameter of the linear ODE model. For low-dimensional systems, the linear ODE model can be estimated by standard statistical methods such as the standard least squares method or likelihood-based methods. However, for high-dimensional systems involving hundreds or even thousands of components, the linear ODE model can be difficult to estimate due to the curse-of-dimensionality problem. To deal with the high-dimensional problem, one method is to employ nonparametric smoothing-based approaches (James and Sugar, 2003; Luan and Li, 2004) and non-parametric mixed-effects smoothing spline model under the framework of a mixture distribution (Ma et al., 2006; Ma and Zhong, 2008) to cluster time course longitudinal data into groups to reduce the dimensionality of the system. However, the linear ODE models can be overly restrictive for practical applications. In many instances, there is little evidence to suggest that the associations among components take a linear form. So, linear semi-parametric and nonlinear parametric ODE models have been proposed (Chen and Wu, 2008a,b; Sakamoto and Iba, 2001; Spieth et al., 2006; Weaver et al., 1999) and have been extended into semi-parametric ODE models for

modeling high-dimensional nonlinear systems (Wu et al., 2014), as follows:

$$\frac{dx_k(t)}{dt} = \mu_k + \sum_{j=1}^{p} f_{kj}(x_j(t)), k = 1, \ldots, p$$

where μ_k is an intercept term and $f_{kj}(\cdot)$ is a smooth function to quantify the nonlinear associations among all components. In those models, two-stage smoothing estimation methods are often used. Nonparametric smoothing approaches such as smoothing splines, regression splines, penalized splines, or local polynomials are used to estimate the state variables $\boldsymbol{x}(t) = (x_1(t), x_2(t), \ldots, x_K(t))'$ or/ and their derivatives $\frac{dx_1(t)}{dt}, \frac{dx_2(t)}{dt}, \ldots, \frac{dx_K(t)}{dt}$.

However, specifying ODEs in practice can often be challenging due to a lack of complete knowledge of the system. Two-stage smoothing estimation methods can become time-consuming in high-dimensional scenarios. Particularly, making accurate assumptions about how the system works or applying smoothing techniques becomes impossible when dynamic changes are unobservable as latent variables.

1.2.2 Nonparametric Neural ODE Models

Inspired by recent advancements in deep learning and neural network techniques, Chen et al. (2018) proposed a novel technique, the Neural Ordinary Differential Equation (Neural ODE), to estimate ODEs using neural networks. Neural networks are computational models that, due to the universal approximation theorem (Lu and Lu, 2020), are widely used in machine learning and artificial intelligence to approximate complex functions. Each layer of a neural network can be considered as a simple linear regression model M that operates on a $b \times d$ matrix of X. A single-layer neural

network can be expressed as:

$$M_{W,B}(X) = X \times W + B$$

where W and B are parameter matrices for weight and bias, respectively. To build a model with non-linear relationships, an activate function σ is included in the neural network layer L to perform a non-linear transformation. The output of a single-layer neural network with activation function σ can be expressed as:

$$L(X) = \sigma(X \times W + B)$$

To model complicated functions, such as for complex ODEs, the neural network can be stacked with multiple layers to form a deep neural network. The output of the deep neural network can be expressed as:

$$f_\phi(X) = (L_L \circ L_{L-1} \circ \ldots \circ L_1)(X)$$

where L is the total number of layers in the deep neural network. The input X_i of the i-th layer is the output of the $(i-1)$-th layer. The deep neural network $f_\phi(X)$ can be considered as a nonparametric model, parametrized by ϕ. The parameter ϕ can be optimized either using frequentist methods, which minimize the loss function between the observed and predicted data or using Bayesian methods, which optimize the Evidence Lower Bound (ELBO).

In Neural ODEs, the rate of change of hidden states is approximated by a neural network f_ϕ:

$$\frac{d\boldsymbol{h}(i)}{di} = f_\phi(\boldsymbol{h}(i), i)$$

where ϕ is the parameter set of the neural network. In this way, the hidden state at any time point can be obtained by solving the ODE with the initial condition $\boldsymbol{h}(0)$ using any arbitrary numerical solving algorithms, denoted by $g(\cdot)$. These algorithms include the Runge-Kutta method, the Euler method, or the adaptive step-size control method (Butcher, 1987; Einkemmer, 2018; Nurujjaman, 2020). The hidden state at time t can be estimated as:

$$\boldsymbol{h}(t) = \boldsymbol{h}(0) + \int_0^t f_\phi(\boldsymbol{h}(i), i)di = g(\boldsymbol{h}(0), \phi, t)$$

In our first topic, we propose to use the Neural ODE technique to describe the dynamic changes and associations among multivariate longitudinal data. Our model is specifically designed to address challenges such as sparsely and irregularly observed multivariate longitudinal data with unknown initial values.

1.3 Stochastic model for normal distributed longitudinal outcomes

When outcomes follow a normal distribution, the Linear Mixed Model (LMM) is a commonly used method for making inferences about covariates in longitudinal data (Laird and Ware, 1982). Let $y_i(t_{ij})$ represent the response of subject i at time t_{ij}, $\boldsymbol{x}_{ij}$ be the covariate vector of subject i at time t_{ij}, $\boldsymbol{u}_{ij}$ be the covariate vector of subject i at time t_{ij} correspond to random effects, and $\boldsymbol{b}_i$ be the random effects of subject i.

The linear random effects model can be expressed as:

$$y_i(t_{ij}) = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{u}_{ij}\boldsymbol{b}_i + \epsilon_{ij}$$

where $\boldsymbol{\beta}$ represents the fixed effect, $\boldsymbol{b}_i$ is the random effect, and ϵ_{ij} is an independent, identically distributed ($i.i.d.$) error term. The random effect $\boldsymbol{b}_i$ is typically assumed to follow a normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Most of such models are estiamted by frequentist methods. Given the model assumptions and conditional on the random effects $\boldsymbol{b}_i$, the repeated measurements of response variable $\boldsymbol{y}_{ij}$ follow an independent normal distribution. For each individual subject i, the outcome vector $\boldsymbol{y}_i$ can be expressed as follows:

$$\boldsymbol{y}_i|\boldsymbol{b}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{U}_i\boldsymbol{b}_i, \sigma^2\boldsymbol{I})$$

where $\boldsymbol{X}_i$ and $\boldsymbol{U}_i$ represent covariate matrix of subject i for fixed and randome effects respectively. Marginally, the within-subject covariance structures are determined by random effect and covariance matrix $\boldsymbol{\Sigma}$, the response variable $\boldsymbol{y}_i$ follows a normal distribution as expressed below:

$$\boldsymbol{y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i), \quad \boldsymbol{V}_i = \boldsymbol{U}_i\boldsymbol{\Sigma}\boldsymbol{U}_i^T + \sigma^2\boldsymbol{I}$$

While Linear Mixed Models (LMMs) are widely used, the fitting trajectories of such models, which assume that individuals maintain the linear path, seem restricted for complex measurements such as subjective specific latent dynamics, pathological mechanisms, etc. To relax the constraints of linearity and parametric assumptions,

alternative models have been developed for handling dependent data, utilizing latent stochastic processes (Quintana et al., 2016). In these models, the linearity assumption is relaxed, and the within-subject covariance structures are determined by a stochastic process. Such a model can be represented as follows:

$$y_i(t_{ij}) = \boldsymbol{x}(t_{ij})\boldsymbol{\beta} + w_i(t_{ij}) + \epsilon_{ij}$$

where $w_i(t_{ij})$ represents a stochastic process. Most models of this kind are estimated using Frequentist methods. Often, a Gaussian Process (GP) or variant of GP, such as an integrated Ornstein-Uhlenbeck (IOU) process, is assumed (Taylor et al., 1994).

In our second topic, we propose to relax the specification of the stochastic process to model the stochastic process as a nonparametric function of the stochastic ODE process as follows:

$$y_i(t_{ij}) = f(\boldsymbol{z}_i(t_{ij})) + \boldsymbol{x}_i(t_{ij})\boldsymbol{\beta} + \epsilon_i(t)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of fixed effects associated with covariates $\boldsymbol{x}_i(t_{ij})$. Random effects are modeled by an unknown function $f(\cdot)$ is an unknown baseline function, $\boldsymbol{z}_i(t)$ is a multi-dimensional latent stochastic process, and the error term $\epsilon_i(t)$ is a process of random noise with mean zero.

By utilizing neural networks and neural ODE techniques. We estimate the fixed effects of interest covariates using Bayesian approaches, which provide the advantages of parameter estimation and uncertainty quantification.

1.4 Generalized mixed effect models for longitudinal data

In longitudinal data analysis, for making inferences about covariates when the outcomes are not normally distributed, generalized linear mixed models (GLMMs) are commonly used (McCulloch et al., 2001). The GLMMs are an extension of the LMMs, where the response variable follows a distribution from the exponential family, that is

$$p_{\boldsymbol{\theta}}(y_{ij}|\vartheta_{ij},\varphi) = exp\left[\frac{y_{ij}\vartheta_{ij} - b(\vartheta_{ij})}{a(\varphi)} + c(y_{ij},\varphi)\right]$$

where ϑ_{ij} is the canonical parameter and φ is a dispersion parameter, where

$$\mu_{ij} = E[y_{ij}|\vartheta_{ij},\varphi] = b'(\vartheta)$$

In the generalized linear mixed model, the canonical parameter ϑ_{ij} is linked to the covariates through a link function, as follows:

$$g(\mu_{ij}) = g(b'(\vartheta)) = \eta_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{u}_{ij}\boldsymbol{b}_i$$

where $g(\cdot)$ is a monotonic differentiable function, often referred to as the link function, μ_{ij} is the mean of the conditional outcome for individual i at time t_{ij}, and η_{ij} is called the linear predictor. At specific observation time point t_{ij}, $\boldsymbol{x}_{ij}$ is a $1 \times p$ vector of covariates, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects. $\boldsymbol{u}_{ij}$ is a $1 \times q$ vector of random effects covariates, and $\boldsymbol{b}_i$ is a $q \times 1$ vector of random effects for individual i. Therefore, we have:

$$p(y_{ij}|\vartheta,\varphi) \equiv p(y_{ij}|\boldsymbol{\beta},\boldsymbol{b}_i,\varphi)$$

To relax the constraints of the specific parametric baseline mean function of the response variable over time, Moyeed and Diggle (1994); Zeger and Diggle (1994); Zhang et al. (1998) proposed semiparametric mixed models (SPMMs). These models extend linear mixed models (LMMs) by incorporating stochastic processes into mixed models to model the time effect using a nonparametric function. However, when dealing with generalized models for outcomes from the exponential family, the complexity of high-dimensional integration, which is typically required when estimating using frequentist methods, makes the generation of SPMMs to model longitudinal data from a variety of outcome distributions significantly challenging.

In our third topic, we propose to combine the idea of GLMMs with the semiparametric mixed models, in which the specific parametric baseline mean function is replaced by a nonparametric function of time. The model can be expressed as follows:

$$g(\mu_i(t)) = \eta_i(t) = \gamma(t) + \boldsymbol{x}_i(t)\boldsymbol{\beta} + \boldsymbol{u}_i(t)\boldsymbol{b}_i$$

where $i = 1, \ldots, m$, $g(\cdot)$ is a monotonic link function, $\mu_i(t)$ is the mean of the conditional outcome for individual i. $\gamma(t)$ is a nonparametric baseline function that captures the time effects.

In our proposed model, rather than assuming $\gamma(t)$ to be a parametric or smooth function of time, as has been done in semi-parametric mixed models limited for continuous outcomes based on frequentist methods (MacKay et al., 1998; Moyeed and Diggle, 1994; Zeger and Diggle, 1994; Zhang et al., 1998), we propose to model $\gamma(t)$ with a stochastic process prior. This prior process is governed by an Ordinary Differential Equation (ODE) system and is modeled by nonparametric neural networks.

We estimate the parameters of interest using Bayesian approaches, which offer the flexibility for fitting generalized nonlinear mixed models, with the added ability to easily estimate parameter uncertainty, providing a probabilistic measure of confidence in the model's estimates.

Chapter 2

Deep Learning of Sparse Irregularly-Observed Multivariate Longitudinal Data

2.1 Introduction

Multivariate longitudinal data is commonly collected in many scientific studies, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI is a large research initiative for the early detection and tracking of Alzheimer's disease. It involves collecting multivariate longitudinal data, including clinical, imaging, genetic, and biochemical biomarkers, for each participant. The participants undergo baseline tests and subsequent tests at varying follow-up times over the course of several years. These follow-up times are random and can differ among participants. Additionally, the number of follow-ups for each participant may vary, resulting in sparsely observed multivariate longitudinal data with irregularly spaced follow-up times.

When analyzing longitudinal data, there are two schools of approaches. The first approach involves modeling and estimating the mean function of the longitudinal data separately for each variable (Diggle, 2002), often ignoring the potential dependence among the multivariate longitudinal data. The second approach models the temporal change of multivariate longitudinal data as a dynamic system using ordinary differential equations (ODEs) that incorporate current knowledge of the underlying biological mechanisms (Murray, 2002). One advantage of using ODEs is their ability to handle multivariate outcomes over time and provide a comprehensive understand-

ing of the interrelationships between variables. ODEs have been successfully applied in various biomedical applications, including the study of neurodegenerative biomarkers and cognitive symptoms in Alzheimer's disease (Jack et al., 2010; Jack Jr et al., 2013; Petrella et al., 2019), brain energy metabolic systems (Tiveci et al., 2005), signaling pathways (Ranjan et al., 2018), and brain networks (Garbarino et al., 2019). However, describing ODEs can be challenging due to the difficulty of specifying the underlying mechanisms and dynamics, particularly when complete knowledge of the system is lacking.

There has been significant interest in estimating ODEs based on observed longitudinal data, both parametrically and semiparametrically (Lu et al., 2011; Wu et al., 2014; Xue et al., 2019). However, these models still rely on partial assumptions about the underlying biological mechanisms and are typically developed for low-dimensional data. A recent approach, neural ODE, leverages deep learning techniques to infer underlying mechanisms directly from the data instead of relying on prior knowledge Chen et al. (2018). It can be seen as a nonparametric method that offers more flexibility in modeling compared to parametric and semiparametric models. However, the neural ODE was developed as a deep learning layer instead of directly modeling the dependence among multivariate longitudinal data. Additionally, the neural ODE was originally designed for regularly spaced and densely observed time series data, making it less suitable for sparsely observed multivariate longitudinal data with irregularly spaced follow-up times.

In this paper, we present a statistical approach based on neural ODEs to effectively handle sparsely observed multivariate longitudinal data with irregularly spaced follow-up times. Our proposed method is designed to be both straightforward and

computationally efficient, even for high-dimensional multivariate longitudinal data. Furthermore, we introduce a novel approach to simultaneously estimate the initial values as an unknown parameter with the estimation of the ODEs. This differs from existing methods that either require the initial values to be pre-specified or estimated in an ad hoc manner. We thoroughly evaluate the performance of our proposed method across various simulation scenarios, including those with complex variance-covariance structures for the underlying stochastic process.

In what follows, we describe the statistical method and the proposed estimation procedure. We then present comprehensive simulation studies to evaluate the performance of our proposed method across various temporal correlations. Additionally, we provide a practical demonstration of our approach using Alzheimer's disease data. Finally, we conclude the paper with a discussion of our methods and potential future research.

2.2 Statistical Model

2.2.1 Multivariate longitudinal data observed sparsely

Consider a scenario where there are n individuals with i.i.d. observable K-dimensional multivariate longitudinal data $\{X_i^{(1)}(t), \ldots, X_i^{(K)}(t)\}$ for $i = 1, \ldots, n$, and $t \geq 0$. Each individual would be followed over time and has the multivariate outcomes of interest measured at baseline and at random follow-up times. In many medical studies such as ADNI, the number of scheduled follow-ups may be small and follow-up times may be random, resulting in irregularly spaced intervals during follow-up. Additionally, it is possible that not all of the multivariate outcomes of an individual can be observed at

a follow-up time. When this happens, we call such data asynchronous observations. Figure 1 illustrates the difference between synchronous and asynchronous observations, where Figure 2.1a shows synchronous observations and Figure 2.1b shows the asynchronous observations in an example of bivariate longitudinal data from two subjects.

For each subject $i = 1, \ldots, n$, we denote the observation times separately for each k-th component of the multivariate outcomes to be $T_{ij}^{(k)}$ for $j = 1, \ldots, J_i^{(k)}$, where $J_i^{(k)}$ is the total number of observations for the k-th component of the multivariate outcomes. The number of observations $J_i^{(k)}$ can vary between subjects and outcomes. Such notations simplify the analysis of statistical methods for multivariate longitudinal data with synchronous and asynchronous observations. Furthermore, we assume that the observation times, $T_{ij}^{(k)}$, for $k = 1, \ldots, K$, $j = 1, \ldots, J_i^{(k)}$ and $i = 1, \ldots, n$ are independent and non-informative for the stochastic process $X_i^{(k)}(t)$. The observed data can be concisely written as $X_i^{(k)}(T_{ij}^{(k)})$ for $k = 1, \ldots, K$, $j = 1, \ldots, J_i^{(k)}$ and $i = 1, \ldots, n$.

2.2.2 Statistical models

The multivariate longitudinal data can be viewed as sampling from a K-dimensional stochastic process $\boldsymbol{X}_i(t) = \left\{X_i^{(1)}(t), \ldots, X_i^{(K)}(t)\right\}^\top$ for $i = 1, \ldots, n$. In order to analyze the pattern and temporal evolution of the multivariate longitudinal data, our primary focus is on estimating the mean functions of the stochastic processes $\{X_i^{(k)}(t), k = 1, \ldots, K\}$, denoted by

$$\mu_k(t) = E\{X_i^{(k)}(t)\}, \quad k = 1, \ldots, K, \qquad \text{for } t \geq 0. \tag{2.1}$$

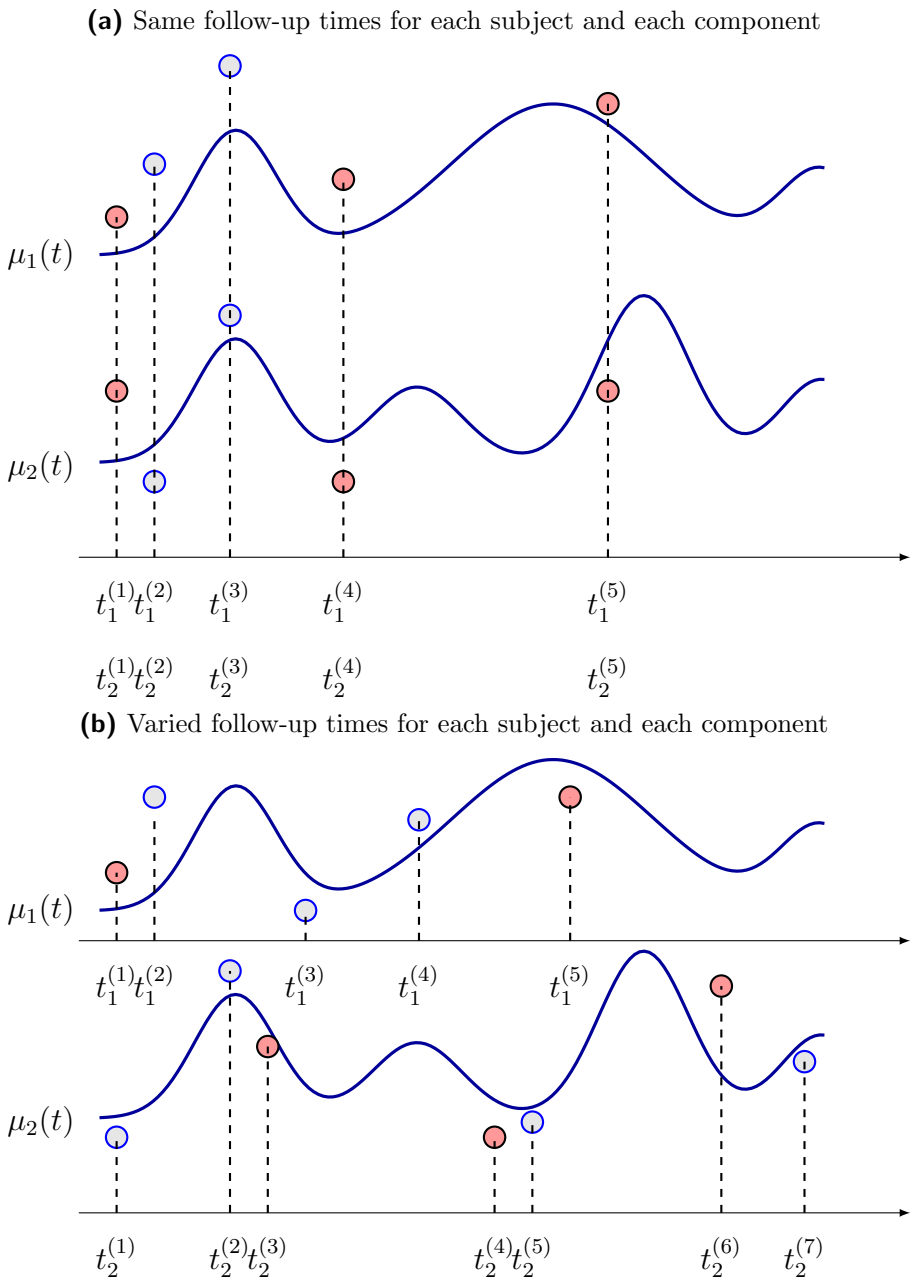


Figure 2.1: Demonstration of pooling the data with varied follow-up times for each subject and each component.

Denote $\boldsymbol{\mu}(t) = \{\mu_1(t), \ldots, \mu_K(t)\}^\top$. We then consider a general model that can capture the interrelationship and temporal progression of the mean process $\boldsymbol{\mu}(t) = \{\mu_1(t), \ldots, \mu_K(t)\}^\top$. We assume that the mean process $\boldsymbol{\mu}(t)$ is governed by a dynamic system composed of K-dimensional differential equations

$$\frac{d\boldsymbol{\mu}(t)}{dt} = \boldsymbol{f}\{\boldsymbol{\mu}(t)\}, \tag{2.2}$$

$$\boldsymbol{\mu}_0 \equiv \boldsymbol{\mu}(0), \tag{2.3}$$

where the K-dimensional function $\boldsymbol{f}(\cdot) = \{f_1(\cdot), \ldots, f_K(\cdot)\}^\top$ is unknown. The domain of the function $f_k(\cdot)$ is $\mathbb{R}^K$. The initial values $\boldsymbol{\mu}_0$ are also unknown parameters. The temporal changes of the mean process are thus directly described by the dynamic system (2.2). By the fundamental theorem of calculus, we have

$$\boldsymbol{\mu}(t) = \boldsymbol{\mu}_0 + \int_0^t \boldsymbol{f}\{\boldsymbol{\mu}(s)\}ds. \tag{2.4}$$

The interrelationship among the components of the mean process $\boldsymbol{\mu}(t)$ is determined by the unknown function $\boldsymbol{f}(\cdot)$ and the initial values $\boldsymbol{\mu}_0$. Given that $\boldsymbol{f}(\cdot)$ and the initial values $\boldsymbol{\mu}_0$ are known, the mean function $\boldsymbol{\mu}(t)$ can be computed using established numerical methods for solving Ordinary Differential Equations (ODEs)(Butcher, 2016).

In prior mathematical models utilizing ODEs, the function $\boldsymbol{f}(\cdot)$ was pre-specified based on prior knowledge. However, in scenarios where prior knowledge is insufficient, estimating $\boldsymbol{f}(\cdot)$ from observed data becomes highly desirable.

The existing statistical methods for ordinary differential equations (ODEs) are either parametric or semi-parametric (Lu et al., 2011; Wu et al., 2014; Xue et al.,

2019). Parametric methods presuppose a specific functional form for $\boldsymbol{f}(\cdot)$ based on prior scientific knowledge, while semi-parametric partially assume the function form. Both methodologies impose constraints on the functional form of $\boldsymbol{f}(\cdot)$, thereby limiting the potential applications of the ODE model. Furthermore, these methods are primarily designed for univariate or low-dimensional longitudinal data. This restricts the possible functional form of $\boldsymbol{f}(\cdot)$, thus greatly limiting the potential applications of the ODE model. Additionally, these methods are developed for univariate or low-dimensional longitudinal data.

We consider a novel method for estimating the function $\boldsymbol{f}(\cdot)$ nonparametrically using deep learning techniques. Neural networks (NN), a flexible nonparametric method, are adept at handling high-dimensional longitudinal data (LeCun et al., 2015). With interconnected layers of neurons applying nonlinear transformations to the input, neural networks can capture complex relationships within data. Moreover, the universal approximation theorem asserts that a neural network can approximate any continuous function given a sufficient number of layers and nodes (Hornik et al., 1989). Strategically, we choose to approximates $\boldsymbol{f}(\cdot)$ in model (2.2) using a neural network

$$\frac{d\boldsymbol{\mu}(t)}{dt} = NN(\boldsymbol{\mu}(t)).$$

Chen et al. (2018) also proposed to approximate ODEs using a neural network (NN) (LeCun et al., 2015). However, the neural ODE method is primarily designed to be used as a hidden layer within a deep neural network for analyzing regularly spaced and densely sampled time-series data. The effectiveness of this method in analyzing

multivariate longitudinal data, particularly with irregularly spaced follow-up times, remains unclear.

In a slight abuse of notation, let $\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot) = NN(\cdot)$. With an extension to the time-dependent ODE model, model (2.2) can be written as

$$\frac{d\boldsymbol{\mu}(t)}{dt} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{\mu}(t), t). \tag{2.5}$$

This is a straightforward extension since by the fundamental theorem of calculus,

$$\boldsymbol{\mu}(t) = \boldsymbol{\mu}_0 + \int_0^t \boldsymbol{f}_{\boldsymbol{\theta}}\{\boldsymbol{\mu}(s), s\}ds.$$

We aim to estimate $\boldsymbol{f}_{\boldsymbol{\theta}}$ using deep learning techniques, treating the high-dimensional parameter $\boldsymbol{\theta}$ and the initial values $\boldsymbol{\mu}_0$ as unknown parameters. We denote $\boldsymbol{\mu}(t; \boldsymbol{\mu}_0, \boldsymbol{\theta}) \equiv \boldsymbol{\mu}(t)$ to indicate its dependence on unknown parameters $(\boldsymbol{\mu}_0, \boldsymbol{\theta})$.

2.2.3 Estimating procedure

In our proposed method, the mean functions, represented as $\boldsymbol{\mu}(t)$, are characterized by the Ordinary Differential Equation (ODE) given by (2.2). While the primary objective is to estimate the unknown parameters $(\boldsymbol{\mu}_0, \boldsymbol{\theta})$, the ultimate goal is to make inference on the mean functions $\boldsymbol{\mu}(t; \boldsymbol{\mu}_0, \boldsymbol{\theta})$. Under the mean process model (2.1), we use the well-established method of minimizing the sum of squared errors between the observed data and the mean functions $\boldsymbol{\mu}(t)$ with LASSO type regulation term to penalize $\boldsymbol{\theta}$ and prevent overfitting (Hastie et al., 2015).

More specifically, by the method of ODE, if the function $\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot)$ and initial values $\boldsymbol{\mu}_0$ are given, then the function $\boldsymbol{\mu}(t)$ can be solved numerically through the ODE equation

(2.5). Denote the solution to the ODE by $\tilde{\boldsymbol{\mu}}(t;\boldsymbol{\mu}_0,\boldsymbol{\theta}) = \{\tilde{\mu}_1(t;\boldsymbol{\mu}_0,\boldsymbol{\theta}),\ldots,\tilde{\mu}_K(t;\boldsymbol{\mu}_0,\boldsymbol{\theta})\}^\top$. Given $\tilde{\mu}_k(t;\boldsymbol{\mu}_0,\boldsymbol{\theta})$, the estimating procedure is to minimize the following objective function with respect to $(\boldsymbol{\mu}_0,\boldsymbol{\theta})$:

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{J_i^{(k)}}\left\{\tilde{\mu}_k(T_{ij}^{(k)};\boldsymbol{\mu}_0,\boldsymbol{\theta}) - X_i^{(k)}(T_{ij}^{(k)})\right\}^2 + \lambda\|\boldsymbol{\theta}\|, \tag{2.6}$$

where $\|\cdot\|$ is a predefined norm corresponding to different penalties, such as ℓ_1 and ℓ_2-norm, and λ is a penalty term. The computation involves an iterative procedure with two main steps:

1. Given the neural network $\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot)$ with the current estimate of $\boldsymbol{\theta}$ and the current estimate of $\boldsymbol{\mu}_0$, solve the ODE (2.2) numerically to obtain $\tilde{\boldsymbol{\mu}}(t;\boldsymbol{\mu}_0,\boldsymbol{\theta})$.
2. Given $\tilde{\boldsymbol{\mu}}(t;\boldsymbol{\mu}_0,\boldsymbol{\theta})$, minimize (2.6) with respect to $(\boldsymbol{\mu}_0,\boldsymbol{\theta})$.

Considering the importance of the initial value $\boldsymbol{\mu}_0$ in the ODE solution, instead of using existing methods that either require the initial values to be pre-specified or estimated in an ad hoc manner, we iteratively minimize the objective function (2.6) separately for $\boldsymbol{\theta}$ and $\boldsymbol{\mu}_0$. The minimization of $\boldsymbol{\theta}$ involves training neural networks using the stochastic gradient descent (SGD) algorithm. This involves back-propagation, achieved by randomly sampling a small data batch and calculating the objective function's gradient with respect to $\boldsymbol{\theta}$. Specifically, we employ the adaptive moment (Adam) method, a variant of SGD known for its computational efficiency and low memory requirements (Kingma and Ba, 2014). This strategy has been shown to be practical and efficient for large amounts of high dimensional data in deep learning of neural networks (Goodfellow et al., 2016). Given the current estimate of $\boldsymbol{\theta}$, the minimization of (2.10) with respect to $\boldsymbol{\mu}_0$ can be solved using a standard optimiza-

tion algorithm such as the limited-memory BFGS (L-BFGS), a quasi-Newton method (Nocedal and Wright, 2006).

Pooling the data

To further refine the optimization procedure, rewrite the optimization objective function (2.6) by exchanging the summations over i and k,

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{J_i^{(k)}}\left\{\tilde{\mu}_k(T_{ij}^{(k)};\boldsymbol{\mu}_0,\boldsymbol{\theta})-X_i^{(k)}(T_{ij}^{(k)})\right\}^2+\lambda\|\boldsymbol{\theta}\|. \tag{2.7}$$

We pool the follow-up data from the n subjects for each $k-$th outcome for $k = 1,\ldots,K$. More specifically, for fixed k, pool the follow-up time $\{T_{ij}^{(k)}, j = 1,\ldots,J_i^{(k)}, i = 1,\ldots,n\}$ and denote their ordered values by

$$t_1^{(k)} \le t_2^{(k)} \le \ldots \le t_{L_k}^{(k)}, \tag{2.8}$$

where $L_k = \sum_{i=1}^{n} J_i^{(k)}$ is the total number of follow-up times in k-th outcome. Denote the corresponding observed outcome by

$$x_1^{(k)}, x_2^{(k)} \ldots, x_{L_k}^{(k)}. \tag{2.9}$$

Then, we can rewrite the objective function (2.7) as

$$\sum_{k=1}^{K}\sum_{l=1}^{L_k}\left\{\tilde{\mu}_k(t_l^{(k)};\boldsymbol{\mu}_0,\boldsymbol{\theta})-x_l^{(k)}\right\}^2+\lambda\|\boldsymbol{\theta}\|. \tag{2.10}$$

When the follow-up times $\{T_{ij}^{(k)}, j = 1, \ldots, J_i^{(k)}, i = 1, \ldots, n\}$ overlap, the objective function (2.10) simplifies that of (2.6). Consequently, pooling the data can significantly reduce computational costs, particularly when employing the stochastic gradient descent algorithm for neural network training.

We further illustrate the concept of data pooling in Figure 2.1a and Figure 2.1b. In these plots, observations from two subjects are marked by red and blue dots. The red dots correspond to the first subject's observations, while the blue dots correspond to the second subject's observations. For each component, observations from both subjects are pooled together. The bivariate mean process functions are displayed as blue solid curves. Figure 2.1a illustrates a scenario where each subject's multivariate outcomes are observed at identical follow-up times. Conversely, Figure 2.1b represents a case where the multivariate outcomes for each subject may not coincide with the same follow-up times.

Estimation of the ODE initial values $\boldsymbol{\mu}_0$

Given the current estimate of $\boldsymbol{\theta}$, the minimization of (2.10) with respect to $\boldsymbol{\mu}_0$ may be solved using a standard optimization algorithm such as the limited-memory BFGS (L-BFGS), a quasi-Newton method (Nocedal and Wright, 2006). However, the minimization of (2.10) with respect to $\boldsymbol{\mu}_0$ involves solving the ODE (2.2) repeatedly using a numerical method, which is computationally expensive. We propose an equivalent solution that is computationally efficient. Using the identity (2.4), taking the derivative of the objective function with respect to the k-th component of $\boldsymbol{\mu}_0 = \{\mu_{0,1}, \ldots, \mu_{0,K}\}$

and setting it to be equal to zero, we have

$$\sum_{l=1}^{L_k} \left\{ \mu_0^{(k)} + \int_0^{t_l^{(k)}} f_{\boldsymbol{\theta}}^{(k)}\{\tilde{\boldsymbol{\mu}}(s; \boldsymbol{\mu}_0, \boldsymbol{\theta})\} ds - x_l^{(k)} \right\} = 0,$$

where $L_k = \sum_{i=1}^{n} J_i^{(k)}$ is the total number of follow-up times in k-th outcome. This yields an iterative identity for solving $\mu_0^{(k)}$ for $k = 1, \ldots, K$,

$$\mu_0^{(k)} = \frac{1}{L_k} \sum_{l=1}^{L_k} \left\{ x_l^{(k)} - \int_0^{t_l^{(k)}} f_{\boldsymbol{\theta}}^{(k)}\{\tilde{\boldsymbol{\mu}}(s; \boldsymbol{\mu}_0, \boldsymbol{\theta})\} ds \right\}. \tag{2.11}$$

This equation implies that $\mu_{0,k}$ can be directly updated using the current estimates of the neural network indexed by $\boldsymbol{\theta}$ and the ODE initial value $\boldsymbol{\mu}_0$, without the need to repeatedly solve the ordinary differential equation (ODE) during an optimization algorithm. The integration in (2.11) can be numerically evaluated using the trapezoidal rule. We will use this method to update $\boldsymbol{\mu}_0$ in the computation algorithm.

2.2.4 Computational algorithms

The computational algorithms begin with the data pooling procedure. The algorithm works by iterating between learning the neural networks with respect to $\boldsymbol{\theta}$ and the ODE initial value $\boldsymbol{\mu}_0$ until a convergence criterion is satisfied. We propose two algorithms to minimize the objective function with respect to $\boldsymbol{\mu}_0$, one with a standard optimization algorithm and the other one with the explicit formula (2.11). The details of the algorithm using the standard optimization technique are described in Algorithm 1. A technical difficulty of training ODE with neural networks is to calculate the gradient of the objective function with respect to the neural network parameters $\boldsymbol{\theta}$; such a procedure is also known as the back-propagation derivative.

The solution proposed in Chen et al. (2018) is to use the adjoint method to calculate the gradient, which is a mathematical technique to define a new ODE so that its solution is the gradient of the loss function with respect to $\boldsymbol{\theta}$. Ma et al. (2021) used the simulation study to show that the automatic differentiation is computationally more efficient than the adjoint method. We will use the automatic differentiation to calculate the gradient of the objective function with respect to $\boldsymbol{\theta}$ in the algorithm.

To improve the computational efficiency, we propose an alternative algorithm based on the explicit formula (2.11). Let the current values of the parameters be $(\boldsymbol{\mu}_0^{(m)}, \boldsymbol{\theta}^{(m+1)})$. We can then update the ODE initial values as follows,

$$\left(\mu_0^{(k)}\right)^{(m+1)} \leftarrow \frac{1}{L_k}\sum_{l=1}^{L_k}\Big\{x_l^{(k)} - \int_0^{t_l^{(k)}} f_{\boldsymbol{\theta}}^{(k)}\{\tilde{\boldsymbol{\mu}}(s;\boldsymbol{\mu}_0^{(m)}, \boldsymbol{\theta}^{(m+1)})\}ds\Big\}. \tag{2.12}$$

To evaluate numerically the integration in (2.11), let τ be the maximum follow-up duration, where $\tau \geq \max\{t_{L_k}^{(k)}, k = 1, \ldots, K\}$. Divide the interval $[0, \tau]$ into small intervals of equal length Δ. Let the cutoff point denoted by $0 = r_0 < r_1 < \ldots < r_Q = \tau$. Let $r_{q(k,l)}$ be the largest among $\{r_1, \ldots, r_Q\}$ that is less than or equal to $\leq t_l^{(k)}$. Then, the integral in (2.11) can be approximated by the trapezoidal rule as follows,

$$\begin{aligned}\int_0^{t_l^{(k)}} f_{\boldsymbol{\theta}}^{(k)}\{\tilde{\boldsymbol{\mu}}(s;\boldsymbol{\mu}_0,\boldsymbol{\theta})\}ds \approx & \frac{\Delta}{2} f_{\boldsymbol{\theta}}^{(k)}(\boldsymbol{\mu}_0^{(m)}) + \Delta \sum_{i=1}^{r_{q(k,l)}-1} f_{\boldsymbol{\theta}}^{(k)}\{\tilde{\boldsymbol{\mu}}(r_i;\boldsymbol{\mu}_0,\boldsymbol{\theta})\} \\ & + \frac{\Delta + t_l^{(k)} - r_{q(k,l)}}{2} f_{\boldsymbol{\theta}}^{(k)}\{\tilde{\boldsymbol{\mu}}(r_{q(k,l)};\boldsymbol{\mu}_0,\boldsymbol{\theta})\} \\ & + \frac{t_l^{(k)} - r_{q(k,l)}}{2} f_{\boldsymbol{\theta}}^{(k)}\{\tilde{\boldsymbol{\mu}}(t_l^{(k)};\boldsymbol{\mu}_0,\boldsymbol{\theta})\}. \end{aligned} \tag{2.13}$$

Of note, because of the additive nature of the formula, the integral can be calculated efficiently as the follow-up time increases. The details of the alternative algorithm

based on the formula (2.11) are summarized in Algorithm 2. Figure 2.2 illustrates a single run of the algorithm, demonstrating the convergence of the algorithm for estimating ODE initial values when the number of training epochs increases.

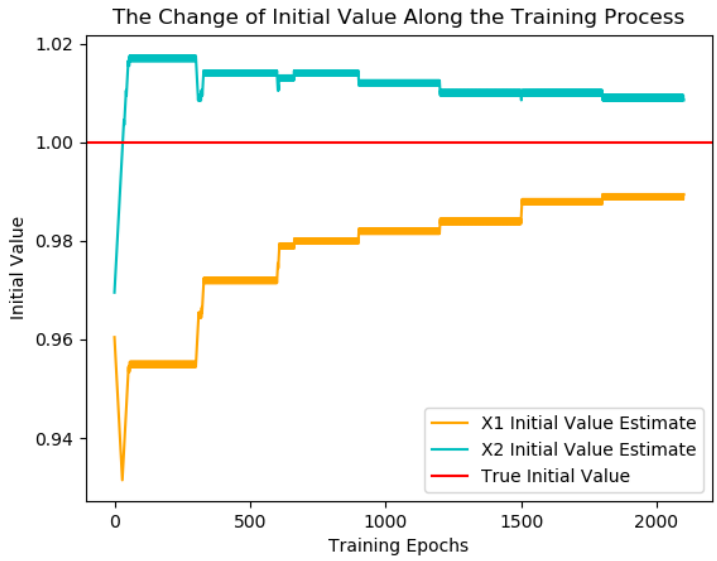


Figure 2.2: One single run of Algorithm 2.

The computational algorithm updates $\boldsymbol{\theta}$ and $\boldsymbol{\mu}_0$ alternatively until a convergence criterion is satisfied, for example, when $\|\boldsymbol{\mu}_0^{(m+1)} - \boldsymbol{\mu}_0^{(m)}\|_2$ is sufficiently small. By formula (2.12),

$$
\begin{aligned}
&\left\{\left(\mu_0^{(k)}\right)^{(m+1)} - \left(\mu_0^{(k)}\right)^{(m)}\right\}^2 \\
&\leq \frac{1}{L_k}\sum_{l=1}^{L_k}\int_0^{t_l^{(k)}}\left\{f^{(k)}_{\boldsymbol{\theta}^{(m+1)}}\{\tilde{\boldsymbol{\mu}}(s;\boldsymbol{\mu}_0^{(m)},\boldsymbol{\theta}^{(m+1)})\} - f^{(k)}_{\boldsymbol{\theta}^{(m)}}\{\tilde{\boldsymbol{\mu}}(s;\boldsymbol{\mu}_0^{(m-1)},\boldsymbol{\theta}^{(m)})\}\right\}^2 ds \\
&\leq \int_0^{\tau}\left\{f^{(k)}_{\boldsymbol{\theta}^{(m+1)}}\{\tilde{\boldsymbol{\mu}}(s;\boldsymbol{\mu}_0^{(m)},\boldsymbol{\theta}^{(m+1)})\} - f^{(k)}_{\boldsymbol{\theta}^{(m)}}\{\tilde{\boldsymbol{\mu}}(s;\boldsymbol{\mu}_0^{(m-1)},\boldsymbol{\theta}^{(m)})\}\right\}^2 ds.
\end{aligned}
$$

This suggests a global convergence criterion of convergence,

$$\frac{1}{K}\sum_{k=1}^{K}\int_0^\tau \left\{ f^{(k)}_{\boldsymbol{\theta}^{(m+1)}}\{\tilde{\boldsymbol{\mu}}(s;\boldsymbol{\mu}_0^{(m)},\boldsymbol{\theta}^{(m+1)})\} - f^{(k)}_{\boldsymbol{\theta}^{(m)}}\{\tilde{\boldsymbol{\mu}}(s;\boldsymbol{\mu}_0^{(m-1)},\boldsymbol{\theta}^{(m)})\}\right\}^2 ds \leq \epsilon_2, \quad (2.14)$$

for a sufficiently small ϵ_2.

The algorithm using the standard optimization algorithm L-BFGS can be computationally expensive. The algorithm based on the explicit formula depends on an integral approximation but is computationally efficient. We will conduct a simulation study to compare the two methods to evaluate their performance.

Algorithm 1: Learning the Mean Processes with Neural Network ODEs

Input: Observed multivariate longitudinal data

$$\{T_{ij}^{(k)}, X_i^{(k)}(T_{ij}^{(k)}), j = 1, \ldots, J_i^{(k)}, k = 1, \ldots, K, i = 1, \ldots, n\}$$

Output: Estimated mean process function $\hat{\boldsymbol{\mu}}(t; \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\theta}})$

data **step** Pooling the data: sorting (2.8) and reindexing the data (2.9)

$$data = \left\{t_1^{(k)} \leq \ldots \leq t_{L_k}^{(k)}, \quad x_1^{(k)}, \ldots, x_{L_k}^{(k)}, \quad L_k = \textstyle\sum_{i=1}^{n} J_i^{(k)}, k = 1, \ldots, K\right\};$$

Function `SSNeuralODE(`$\boldsymbol{\theta}$ *;* $\boldsymbol{\mu}_0$*, data)`**:**

Solve the ODE (2.2) to obtain $\tilde{\mu}_k(t; \boldsymbol{\mu}_0, \boldsymbol{\theta})$ for

$t \in \left\{t_1^{(k)} \leq \ldots \leq t_{L_k}^{(k)}\right\}, k = 1, \ldots, K;$

`/* only need the values of` $\tilde{\mu}_k$ `at the oberseved time points`

`*/`

return *the sum of sequares by formula (2.10)*;

Take a random initial values for the parameters $(\boldsymbol{\mu}_0^{(0)}, \boldsymbol{\theta}^{(0)})$;

repeat`/*` $1 \leq m \leq M$ `for a predefined large` M `*/`

set $\boldsymbol{\mu}_0^{(m)} = \boldsymbol{\mu}_0^{(m-1)}, \boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)}$;

repeat`/* Use SGD Adam method to minimize (2.10) */`

batchdata = randomly select a mini-batch of *data*;

Use automatic differentiation to obtain the gradient of

`SSNeuralODE(` $\boldsymbol{\theta}$*;* $\boldsymbol{\mu}_0^{(m)}$*, batchdata)* with respect to $\boldsymbol{\theta}$;

Use Adam to find $\boldsymbol{\theta}^{(m+1)}$;

until `SSNeuralODE(` $\boldsymbol{\theta}^{(m+1)}$*;* $\boldsymbol{\mu}_0^{(m)}$*, data)`* $\leq \epsilon_1$ *for a predefined small* ϵ_1;

$\boldsymbol{\mu}_0$ **step** Use L-BFGS to find a minimizer $\boldsymbol{\mu}_0^{(m+1)}$ of `SSNeuralODE(` $\boldsymbol{\theta}^{(m+1)}$*;* $\boldsymbol{\mu}_0$*, data)`* with respect to $\boldsymbol{\mu}_0$.

until *convergence* $\|\boldsymbol{\mu}_0^{(m+1)} - \boldsymbol{\mu}_0^{(m)}\|_2 \leq \epsilon_2$ *for a predefined small* ϵ_2;

Algorithm 2: Alternative Algorithm on Learning the Mean Processes with Neural Network ODEs

Modify Algorithm 1 as follows.

Right after the *data* **step** in Algorithm 1, **add**

> Set τ = maximum follow-up duration, where $\tau \geq \max\{t_{L_k}^{(k)}, k = 1, \ldots, K\}$;
>
> Divide $[0, \tau]$ into small intervals of equal length Δ. Denote the cutoff point $0 = r_0 \leq r_1 \leq \cdots \leq r_Q = \tau$;

Replace the $\boldsymbol{\mu}_0$ **step** in Algorithm 1 with the following

> With current values $(\boldsymbol{\mu}_0^{(m)}, \boldsymbol{\theta}^{(m+1)})$, solve the ODE (2.2) to obtain $f_{\boldsymbol{\theta}^{(m+1)}}^{(k)}\{\tilde{\boldsymbol{\mu}}(s; \boldsymbol{\mu}_0^{(m)}, \boldsymbol{\theta}^{(m+1)})\}$ for $s \in \{r_0, r_1, \ldots, r_Q\}$;
>
> Update $\boldsymbol{\mu}_0^{(m+1)}$ using formulas (2.12) and (2.13);

Optional: **Replace** convergence criterion on $\boldsymbol{\mu}_0^{(m)}$ in Algorithm 1 with

> the convergence criterion (2.14);

2.3 Simulations

2.3.1 Simulation setup

In the simulation study, we examined two setups of true mean process functions. In the first simulation setup, the true mean processes are characterized by a system of ODEs, specifically the Lotka-Volterra model:

$$\frac{d\mu_1(t)}{dt} = \mu_1(t)[1 - \mu_1(t)] - \frac{a\mu_1(t)\mu_2(t)}{\mu_1(t) + c}$$
$$\frac{d\mu_2(t)}{dt} = b \times \mu_2(t)\left[1 - \frac{\mu_2(t)}{\mu_1(t)}\right]$$

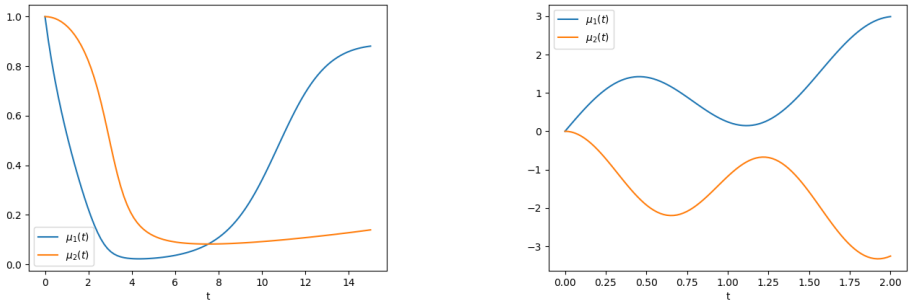
Figure 2.3a illustrates the true curves of the mean processes based on the Lotka-Volterra model with $(a, b, c) = (3/4, 0.1, 0.1)$ over the time interval $t \in [0, 15]$.

In the second simulation setup, the true mean processes do not originate from any well-known ODE system, but from two time-dependent functions:

$$\mu_1(t) = sin(4t) + t$$

$$\mu_2(t) = cos(t) + cos(5t) - 2.$$

Figure 2.3b shows the corresponding true curves of the mean processes, where $t \in [0, 2]$.



(a) **(b)**

Figure 2.3: True solutions of simulation studies

2.3.2 Variance-covariance structure of multivariate longitudinal data

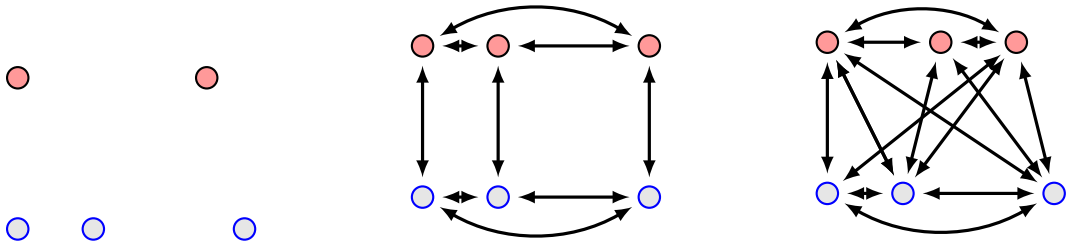
Multivariate longitudinal data dependencies can primarily be described in two ways. The first type of dependency arises when each component of the mean function, $\{\mu_1(t), \ldots, \mu_K(t)\}$, may be interdependent. This type of dependency has been modeled using the ODE approach in our study. The second type of dependency pertains to

the correlation among the random variables, which are sampled from the underlying stochastic processes.

We conduct a simulation study to evaluate the effects of various correlation structures on the proposed estimation procedure for the mean processes. Multivariate longitudinal data often exhibit complex correlations, including temporal correlations and correlations among the multivariate outcomes. Modeling these correlation structures for multivariate longitudinal data generally poses a significant challenge. Without the loss of generality, we consider the scenario of bivariate longitudinal data. We simulate the data by sampling from the following process:

$$X_i^{(1)}(t) = \mu_1(t) + \epsilon_i^{(1)}(t)$$
$$X_i^{(2)}(t) = \mu_2(t) + \epsilon_i^{(2)}(t),$$

where $\epsilon_i^{(1)}(t)$ and $\epsilon_i^{(2)}(t)$ are mean-zero stochastic processes. The correlation structures are characterized by $\{\epsilon_i^{(1)}(t), \epsilon_i^{(2)}(s)\}$ where $t > 0$ and $s > 0$. In the current paper, we discuss three scenarios illustrated in Figure 2.4. Each scenario represents various complexities involved in modeling and estimating the correlation structure for multivariate longitudinal data:



(a) No correlation

(b) Longitudinal and cross-sectional correlation

(c) Longitudinal, cross-sectional and mixed correlation

Figure 2.4: Association between observations

1. Independence case as illustrated in Figrue 2.4a: $\epsilon_i^{(1)}(t)$ and $\epsilon_i^{(2)}(t)$ are independent white noises with mean zero and variance σ_1^2 and σ_2^2, respectively.

2. A special correlation structure (Figure 2.4b): to allow within-component correlation and between-component correlation, assume that

   $\{\epsilon_i^{(1)}(t), \epsilon_i^{(2)}(s)\}$ for $t > 0$ and $s > 0$ have a multivariate normal distribution with the variance-covariance matrix constructed as follows:

$$cov\{\epsilon_i^{(1)}(t), \epsilon_i^{(1)}(s)\} = \sigma_1^2 \rho_1^{|t-s|}, \qquad cov\{\epsilon_i^{(2)}(t), \epsilon_i^{(2)}(s)\} = \sigma_2^2 \rho_1^{|t-s|}$$
$$cov\{\epsilon_i^{(1)}(t), \epsilon_i^{(2)}(t)\} = \sigma_1 \sigma_2 \rho_2, \qquad cov\{\epsilon_i^{(1)}(t), \epsilon_i^{(2)}(s)\} = 0 \text{ when } t \neq s,$$

   where ρ_1 is the correlation coefficient for random variables over time, and ρ_2 is the correlation coefficient between the bivariate outcomes.

   For example, consider three equal spaced follow-up times $t_1 < t_2 < t_3$, where $|t_2 - t_1| = |t_3 - t_2| = \delta$. Let us write the sampled random variables of the multivariate outcomes as

$$\{X_i^{(1)}(t_1), X_i^{(1)}(t_2), X_i^{(1)}(t_3), X_i^{(2)}(t_1), X_i^{(2)}(t_2), X_i^{(2)}(t_3)\}.$$

Then, the variance-covariance matrix has the following form:

$$\begin{pmatrix} \sigma_1^2 \begin{pmatrix} 1 & \rho_1^{\delta} & \rho_1^{2\delta} \\ \rho_1^{\delta} & 1 & \rho_1^{\delta} \\ \rho_1^{2\delta} & \rho_1 & 1 \end{pmatrix} & \sigma_1\sigma_2 \begin{pmatrix} \rho_2 & 0 & 0 \\ 0 & \rho_2 & 0 \\ 0 & 0 & \rho_2 \end{pmatrix} \\ \sigma_1\sigma_2 \begin{pmatrix} \rho_2 & 0 & 0 \\ 0 & \rho_2 & 0 \\ 0 & 0 & \rho_2 \end{pmatrix} & \sigma_2^2 \begin{pmatrix} 1 & \rho_1^{\delta} & \rho_1^{2\delta} \\ \rho_1^{\delta} & 1 & \rho_1^{\delta} \\ \rho_1^{2\delta} & \rho_1^{\delta} & 1 \end{pmatrix} \end{pmatrix}$$

In this scenario, the variance-covariance matrix is symmetric and needs to be positive definite. There are various methods to ensure the matrix is positive definite (Pinheiro and Bates, 1996). In Appendix 2.6.1, we show that by using the eigenvalue decomposition method, we can directly choose ρ_1 and ρ_2 so that the matrix is positive definite.

3. General correlation structure (Figure 2.4c): we consider a more general correlation structure where $\epsilon_i^{(1)}(t)$ and $\epsilon_i^{(2)}(s)$ are modeled as a Gaussian process indexed by time. For any two time points t and s, $\epsilon_i^{(k)}(t)$ and $\epsilon_i^{(k)}(s)$ has a multivariate normal distribution with mean zero and a covariance function that depends on t and s. In the current study, we employ the squared exponential covariance function (Wright, 2006):

$$\kappa_1(t,s) = \nu^2 \ \exp\left(- \frac{(t-s)^2}{\sigma^2} \right).$$

To construct a covariance structure for the bivari ate process $\{\epsilon_i^{(1)}(t), \epsilon_i^{(2)}(s)\}$, we propose a Gassian process indexed by $\{t, k\}$, where $t \in [0, \tau]$ and $k \in \{1, 2\}$.The covariance function is defined as follows:

$$\kappa_2(t, k, s, k') = \kappa_1(t, s)\kappa_0(k, k'),$$

where $\kappa_0(k, k')$ is a covariance function for the correlation between the two multivariate outcomes. Such a specification of multivariate Gaussian process has seen applications in multivariate time series data (Roberts et al., 2013). We simply specify $\kappa_0(k, k')$ as follows:

$$\kappa_0(k, k') = \begin{cases} 1, & \text{if } k = k'; \\ \rho, & \text{if } k \neq k'. \end{cases}$$

As an example, suppose that for $X_i^{(1)}$, there are two follow-up times $t_1^{(1)} < t_2^{(1)}$, and for $X_i^{(2)}$, there are three follow-up times $t_1^{(2)} < t_2^{(2)} < t_3^{(2)}$. Then, the covariance matrix for $\{X_i^{(1)}(t_1^{(1)}), X_i^{(1)}(t_2^{(1)})\}$ is $\nu^2 A$, where

$$A = \begin{pmatrix} 1 & e^{-\frac{\left\{t_2^{(1)} - t_1^{(1)}\right\}^2}{\sigma^2}} \\ e^{-\frac{\left\{t_2^{(1)} - t_1^{(1)}\right\}^2}{\sigma^2}} & 1 \end{pmatrix}.$$

The covariance matrix for $\{X_i^{(2)}(t_1^{(2)}), X_i^{(2)}(t_2^{(2)}), X_i^{(2)}(t_3^{(2)})\}$ is $\nu^2 B$, where

$$B = \begin{pmatrix} 1 & e^{-\frac{\left\{t_2^{(2)}-t_1^{(2)}\right\}^2}{\sigma^2}} & e^{-\frac{\left\{t_3^{(2)}-t_1^{(2)}\right\}^2}{\sigma^2}} \\ e^{-\frac{\left\{t_2^{(2)}-t_1^{(2)}\right\}^2}{\sigma^2}} & 1 & e^{-\frac{\left\{t_3^{(2)}-t_2^{(2)}\right\}^2}{\sigma^2}} \\ e^{-\frac{\left\{t_3^{(2)}-t_1^{(2)}\right\}^2}{\sigma^2}} & e^{-\frac{\left\{t_3^{(2)}-t_2^{(2)}\right\}^2}{\sigma^2}} & 1 \end{pmatrix}.$$

The covariance matrix between $\{X_i^{(1)}(t_1^{(1)}), X_i^{(1)}(t_2^{(1)})\}$ and $\{X_i^{(2)}(t_1^{(2)}), X_i^{(2)}(t_2^{(2)}), X_i^{(2)}(t_3^{(2)})\}$ is $\rho\nu^2 C$, where

$$C = \begin{pmatrix} e^{-\frac{\left\{t_1^{(2)}-t_1^{(1)}\right\}^2}{\sigma^2}} & e^{-\frac{\left\{t_2^{(2)}-t_1^{(1)}\right\}^2}{\sigma^2}} & e^{-\frac{\left\{t_3^{(2)}-t_1^{(1)}\right\}^2}{\sigma^2}} \\ e^{-\frac{\left\{t_1^{(2)}-t_2^{(1)}\right\}^2}{\sigma^2}} & e^{-\frac{\left\{t_2^{(2)}-t_2^{(1)}\right\}^2}{\sigma^2}} & e^{-\frac{\left\{t_3^{(2)}-t_2^{(1)}\right\}^2}{\sigma^2}} \end{pmatrix}.$$

Write the observed random variables as

$$\left\{X_i^{(1)}(t_1^{(1)}), X_i^{(1)}(t_2^{(1)}), X_i^{(2)}(t_1^{(2)}), X_i^{(2)}(t_2^{(2)}), X_i^{(2)}(t_3^{(2)})\right\}.$$

Then, its variance-covariance matrix is a 5×5 matrix with the following form:

$$\nu^2 \begin{pmatrix} A & \rho C \\ \rho C^\top & B \end{pmatrix}.$$

By the construction, the variance-covariance matrix is symmetric and positive definite.

2.3.3 Data generating process

In our simulation study, we generate bivariate longitudinal data characterized by irregular and sparse follow-up times. By 'irregular', we refer to the randomness of follow-up times, leading to non-equidistant intervals. 'Sparsity', implies that each subject has a relatively small number of follow-ups. For each subject i, we generate M_i follow-up times where each duration follows an exponential distribution with rate λ_i. The follow-up times are generated as $t_{i,1} < \cdots < t_{i,M_i}$, where $M_i = 5$ or $M_i = 10$. Equivalently, the number of observations within an interval $[0, \tau]$ follows a Poisson distribution whose mean equals $\tau \cdot \lambda_i$. In the synchronous case, we set $T_{ij}^{(k)} = t_{i,j}$ for all k, and $J_i^{(k)} = M_i$. In the asynchronous case, we independently generate $T_{i1}^{(k)} < \cdots < T_{i,J_i^{(k)}}^{(k)}$ for each k, where $J_i^{(k)} = M_i$.

Having the follow-up times for subject i, we then generate the random errors from a multivariate normal distribution with a mean-zero and a covariance structure as specified:

$$\big\{\epsilon_i^{(k)}(T_{ij}^{(k)}),\ j = 1, \ldots, J_i^{(k)},\ k = 1, 2\big\}.$$

We then consider them as a random vector of dimension $J_i^{(1)} + J_i^{(2)}$ and generate the bivariate longitudinal data as follows:

$$X_i^{(k)}(T_{ij}^{(k)}) = \mu_k(T_{ij}^{(k)}) + \epsilon_i^{(k)}(T_{ij}^{(k)}), \quad \text{for } j = 1, \ldots, J_i^{(k)},\ k = 1, 2,$$

where $\mu_k(\cdot)$ is the true mean processes.

2.3.4 Simulation scenarios

We conduct comprehensive simulation studies to evaluate the performance of our proposed method. Motivated by the ADNI study, we simulate data closely to realistic longitudinal measurements. The simulation scenarios are summarized as follows:

1. The number of subjects: (a) $n = 100$ and (b) $n = 300$;
2. The number of follow-ups: (a) $M_i = 5$ and (b) $M_i = 10$;
3. The rate of exponential distribution for the duration of follow-up: (a) $\lambda = 1/2$ and (b) $\lambda = 1/4$ ($\lambda = 5$ for simulation study 2);
4. Pattern of follow-ups: (a) synchronous and (b) asynchronous
5. Covariance structure of the stochastic processes:

   The standard deviation (SD) of the error: (a) $\sigma_1 = \sigma_2 = \nu = 0.3$; (b) $\sigma_1 = \sigma_2 = \nu = 1$; (c) $\sigma_1 = \sigma_2 = \nu = 2$;

   - independent;
   - a special case of correlation with longitudinal and cross-sectional correlation: (i) $\rho_1 = 0.3, \rho_2 = 0.1$; (ii) $\rho_1 = 0.3, \rho_2 = 0.5$;
   - general correlation with longitudinal, cross-sectional, and mixed correlation: $\rho = 1$.

Each simulation scenario is replicated 1000 times. The model is trained using a consistent number of training epochs across all scenarios to ensure comparability during the simulation studies. Each scenario's neural ODE comprises a two-layer fully connected neural network, with each layer containing 32 hidden units and utilizing a ReLU activation function. The Dormand-Prince-Shampine (DOPRI5) method, a Runge-Kutta method of order 5, is employed as the numerical ODE-solving method.

Prior to training, the neural ODE model $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\cdot)$ is initialized with parameters sampled from a uniform distribution. While the ℓ_2-norm penalty, such as weight decay, is a standard regularization option in the Adam optimizer, to avoid over-parameterized deep neural networks, ℓ_1-norm penalty is used and preferred in our proposed method.

2.3.5 Evaluation of the performance of the proposed method

The performance of an estimation method is commonly evaluated using the mean squared error, which, however, requires a fixed time definition. Given our objective to estimate the mean function, it is more appropriate to assess performance across the entire follow-up period. Therefore, we employ the mean integrated squared error (MISE) to evaluate the performance of our proposed estimation method for each component of $\boldsymbol{\mu}(t)$ across the entire follow-up period. The MISE is defined as follows:

$$\text{MISE} = \int_0^\tau E\big\{\hat{\mu}_k(t) - \mu_k(t)\big\}^2 dt = \int_0^\tau \text{Bias}_k(t)^2 dt + \int_0^\tau \text{Var}_k(t) dt, \tag{2.15}$$

where

$$\text{Bias}_k(t) = E(\hat{\mu}_k(t)) - \mu_k(t), \quad \text{Var}_k(t) = E\big\{\hat{\mu}_k(t) - E(\hat{\mu}_k(t))\big\}^2.$$

It is natural to define a single measure of bias and variance for all $t \in [0, \tau]$ as follows:

$$\text{Bias0} = \left(\int_0^\tau \text{Bias}_k(t)^2 dt \right)^{1/2}, \quad \text{Var} = \int_0^\tau \text{Var}_k(t) dt. \tag{2.16}$$

An alternative measure of bias is the integrated absolute difference between the estimated mean function and the true mean function

$$\text{Bias} = \int_0^\tau E\Big|\hat{\mu}_k(t) - \mu_k(t)\Big|dt. \tag{2.17}$$

Since $x^{1/2}$ is a convex function on $[0, \tau]$, by Jensen's inequality, we have

$$\begin{aligned}\text{Bias0} &= \left\{\int_0^\tau \big(E\hat{\mu}_k(t) - \mu_k(t)\big)^2 dt\right\}^{1/2} \leq \frac{1}{\sqrt{\tau}}\int_0^\tau \left\{\big(E\hat{\mu}_k(t) - \mu_k(t)\big)^2\right\}^{1/2} dt \\ &= \frac{1}{\sqrt{\tau}}\int_0^\tau \big|E\hat{\mu}_k(t) - \mu_k(t)\big|dt \leq \frac{1}{\sqrt{\tau}}\int_0^\tau E\big|\hat{\mu}_k(t) - \mu_k(t)\big|dt = \frac{\text{Bias}}{\sqrt{\tau}}.\end{aligned}$$

For $\tau \geq 1$, the bias definition in (2.17) is more stringent than the one in (2.16). Therefore, we use the bias definition in (2.17) to evaluate the bias of the proposed method.

We compute the integrals in the formulas using the trapezoidal rule, a method of numerical integration. Additionally, we substitute the expectations in formulas (2.15), (2.17), and (2.16) for MISE, bias, and variance with empirical values derived from the simulation replicates. The resulting values are referred to as empirical MISE (MISE), empirical bias (EB), and empirical variance (EV), respectively. We describe the details of the computation in the appendix.

2.3.6 Simulation results

For the first simulation setup, the performance evaluations of the proposed method on synchronous and asynchronous observations are presented in Table 2.1 and Table 2.2, respectively, when the follow-up duration adheres to an exponential distribution with

a rate of 1/2. When the rate of the exponential distribution for the follow-up duration is 1/4, the method's performance is displayed in Table 2.3 and Table 2.4. For the second simulation setup, the performance of our model is demonstrated in Table 2.5 and Table 2.6.

As the number of follow-ups increases, more data is available for model fitting, leading to varying effects. Table 2.1 and Table 2.2 show that 10 follow-ups outperform 5, as evidenced by decreasing EB, EV, and MISE. However, Table 2.3 and Table 2.4 indicate that more repeated measurements reduce prediction accuracy. In the second setup, a bias-variance trade-off is observed with changing repeated measurements (Table 2.5 and Table 2.6). More tests lead to higher empirical bias and lower variance, except in a scenario with 100 subjects and a 0.3 standard deviation of random error.

It's important to note that while the total number of observations remains fixed, the durations are randomly sampled from an exponential distribution. As λ decreases from 1/2 to 1/4, the average number of follow-ups within the estimation time interval $[0, 15]$ reduces from 7.5 to 3.75, potentially increasing the sparsity. Sparsity's impact is evident when comparing Table 2.1 with Table 2.3 and Table 2.2 with Table 2.4. In 5 follow-up scenarios, decreasing the exponential distribution rate from 1/2 to 1/4 improves model performance by evenly distributing observations. This trend doesn't hold when most observations fall outside the target range.

Increasing the sample size generally improves our method's performance, with EB, EV, and MISE decreasing. Exceptions exist, such as in scenarios with 10 follow-ups and a 0.3 standard deviation of random error for both X_1 and X_2 (Table 2.3 and Table 2.4), where increasing the sample size reduces EB, EV, and MISE. In the

second setup, a bias-variance trade-off is observed with 10 follow-ups per subject, leading to decreased model accuracy as the sample size increases.

In the current study, we examined three correlation complexities (Figure 2.4). Generally, model estimation performance declines as correlation complexity increases from Scenario 1 to 3, indicated by rising EB, EV, and MISE. However, some cases showed improved estimations with increased complexity, as seen in Table 2.3 for specific subjects and follow-up counts, and standard deviations of random errors. This is also supported by asynchronous observations in Table 2.4 and a similar trend in Study 2 (Table 2.5).

In the first setup, an increase in the standard deviation (σ) of Gaussian noise negatively impacts the model's performance, as shown in Table 2.1, Table 2.2, Table 2.3, and Table 2.4. However, this is not always the case in the second setup. An opposite trend can be observed when the number of follow-ups is 10, as depicted in Table 2.5 and Table 2.6.

Table 2.1: Simulation performance for explicitly bivariate ODE governed synchronous observations ($\lambda = 1/2$)

		$\boldsymbol{\sigma}$=[0.3, 0.3]			$\boldsymbol{\sigma}$=[1.0, 1.0]			$\boldsymbol{\sigma}$=[2.0, 2.0]		
		($\times K\Delta \times 10^{-3}$)			($\times K\Delta \times 10^{-3}$)			($\times K\Delta \times 10^{-3}$)		
# of follow-up		EB	EV	MISE	EB	EV	MISE	EB	EV	MISE
Scenario 1: Independent										
n=100										
X1, X2	k=[5, 5]	[44.37, 35.52]	[2.26, 1.53]	[3.46, 2.09]	[113.61, 100.00]	[22.36, 16.79]	[23.80, 17.79]	[224.90, 195.89]	[93.89, 73.39]	[96.10, 75.19]
X1, X2	k=[10, 10]	[32.00, 31.87]	[0.78, 0.85]	[1.57, 1.82]	[71.48, 66.00]	[6.88, 5.70]	[8.10, 7.23]	[134.72, 118.46]	[27.91, 20.62]	[29.75, 23.04]
n=300										
X1, X2	k=[5, 5]	[33.85, 27.45]	[0.89, 0.65]	[1.94, 1.23]	[69.94, 61.51]	[7.52, 5.65]	[8.72, 6.44]	[133.62, 117.69]	[34.39, 23.86]	[35.73, 25.21]
X1, X2	k=[10, 10]	[27.94, 28.03]	[0.41, 0.55]	[1.20, 1.47]	[49.25, 46.54]	[2.78, 2.42]	[3.79, 3.67]	[86.72, 78.34]	[10.05, 8.15]	[11.97, 9.97]
Scenario 2: Correlated within-variable and between-variable (partially)										
$\rho_1 = 0.3$, $\rho_2 = 0.1$										
n=100										
X1, X2	k=[5, 5]	[45.28, 39.09]	[2.49, 1.97]	[3.62, 2.57]	[123.42, 112.73]	[26.47, 21.63]	[27.71, 22.79]	[242.29, 219.41]	[107.71, 89.40]	[109.52, 91.99]
X1, X2	k=[10, 10]	[34.15, 33.73]	[1.02, 1.04]	[1.83, 2.01]	[81.40, 76.56]	[9.20, 7.85]	[10.46, 9.50]	[151.47, 139.73]	[34.96, 28.77]	[36.72, 31.61]
n=300										
X1, X2	k=[5, 5]	[34.32, 29.18]	[1.00, 0.79]	[1.99, 1.38]	[73.50, 69.82]	[8.43, 7.31]	[9.49, 8.26]	[142.59, 135.41]	[36.73, 34.47]	[37.99, 36.22]
X1, X2	k=[10, 10]	[28.84, 29.34]	[0.48, 0.63]	[1.28, 1.61]	[53.46, 52.42]	[3.43, 3.15]	[4.45, 4.61]	[95.99, 89.61]	[13.07, 10.95]	[14.51, 13.04]
$\rho_1 = 0.3$, $\rho_2 = 0.5$										
n=100										
X1, X2	k=[5, 5]	[46.36, 39.11]	[2.55, 1.95]	[3.82, 2.54]	[123.11, 114.58]	[25.59, 21.72]	[27.40, 22.81]	[237.26, 226.09]	[101.23, 90.50]	[104.04, 92.30]
X1, X2	k=[10, 10]	[33.42, 34.06]	[0.96, 1.09]	[1.72, 2.05]	[80.96, 77.36]	[9.30, 8.18]	[10.49, 9.82]	[155.81, 141.99]	[40.37, 30.09]	[42.26, 32.87]
n=300										
X1, X2	k=[5, 5]	[35.64, 29.46]	[1.07, 8.06]	[2.17, 1.41]	[77.87, 69.31]	[9.40, 7.24]	[10.74, 8.11]	[147.18, 133.68]	[36.80, 29.91]	[38.55, 31.48]
X1, X2	k=[10, 10]	[28.77, 29.19]	[0.47, 0.63]	[1.28, 1.58]	[54.41, 52.06]	[3.67, 3.20]	[4.71, 4.52]	[98.42, 91.53]	[13.97, 11.56]	[15.44, 13.58]
Scenario 3: Correlated within-variable and between-variable (fully)										
n=100										
X1, X2	k=[5, 5]	[49.16, 41.61]	[2.97, 2.27]	[4.17, 2.86]	[139.34, 129.07]	[32.42, 28.00]	[34.14, 29.53]	[274.68, 260.21]	[171.23, 128.13]	[173.70, 130.56]
X1, X2	k=[10, 10]	[37.43, 36.04]	[1.38, 1.34]	[2.18, 2.25]	[98.60, 89.76]	[13.91, 11.45]	[15.35, 13.18]	[186.06, 169.97]	[53.80, 43.96]	[56.12, 47.63]
n=300										
X1, X2	k=[5, 5]	[36.70, 30.18]	[1.23, 0.91]	[2.29, 1.49]	[84.22, 75.39]	[10.92, 8.84]	[12.21, 9.73]	[161.06, 152.10]	[44.03, 39.57]	[45.95, 41.51]
X1, X2	k=[10, 10]	[30.79, 30.58]	[0.60, 0.73]	[1.46, 1.73]	[62.91, 57.51]	[5.02, 4.25]	[6.23, 5.54]	[116.68, 104.57]	[19.85, 15.82]	[21.78, 18.07]

Table 2.2: Simulation performance for explicitly bivariate ODE governed asynchronous observations ($\lambda = 1/2$)

		$\boldsymbol{\sigma}$=[0.3, 0.3]			$\boldsymbol{\sigma}$=[1.0, 1.0]			$\boldsymbol{\sigma}$=[2.0, 2.0]		
		($\times K\Delta \times 10^{-3}$)			($\times K\Delta \times 10^{-3}$)			($\times K\Delta \times 10^{-3}$)		
# of follow-up		EB	EV	MISE	EB	EV	MISE	EB	EV	MISE
Scenario 1: Independent										
n=100										
X1, X2	k=[5, 5]	[43.58, 35.32]	[2.15, 1.51]	[3.35, 2.06]	[111.87, 97.97]	[21.42, 15.72]	[22.93, 16.71]	[217.61, 197.56]	[84.89, 68.87]	[87.13, 70.65]
X1, X2	k=[10, 10]	[31.44, 31.93]	[0.78, 0.88]	[1.52, 1.83]	[71.88, 67.99]	[7.09, 6.01]	[8.18, 7.61]	[132.37, 121.63]	[26.99, 21.70]	[28.46, 24.48]
n=300										
X1, X2	k=[5, 5]	[33.69, 27.88]	[0.88, 0.69]	[1.93, 1.27]	[69.49, 62.32]	[7.22, 5.77]	[8.48, 6.61]	[129.88, 118.21]	[29.25, 23.44]	[30.71, 24.84]
X1, X2	k=[10, 10]	[28.23, 28.39]	[0.42, 0.55]	[1.23, 1.51]	[48.47, 47.30]	[2.72, 2.45]	[3.74, 3.77]	[85.37, 79.60]	[10.43, 8.55]	[11.81, 10.48]
Scenario 3: Correlated within-variable and between-variable (fully)										
n=100										
X1, X2	k=[5, 5]	[47.84, 40.69]	[2.88, 2.17]	[3.98, 2.75]	[133.25, 121.15]	[30.92, 24.24]	[32.22, 25.32]	[256.88, 235.00]	[120.01, 98.82]	[121.79, 100.73]
X1, X2	k=[10, 10]	[37.13, 36.49]	[1.31, 1.31]	[2.14, 2.33]	[93.35, 86.39]	[12.59, 10.30]	[13.89, 12.14]	[178.25, 159.64]	[51.19, 38.26]	[53.08, 41.55]
n=300										
X1, X2	k=[5, 5]	[36.04, 30.09]	[1.10, 0.86]	[2.17, 1.47]	[80.76, 74.43]	[10.15, 8.44]	[11.32, 9.29]	[155.01, 144.34]	[41.14, 34.51]	[42.66, 35.95]
X1, X2	k=[10, 10]	[30.30, 31.10]	[0.60, 0.73]	[1.41, 1.78]	[61.09, 57.93]	[4.89, 4.04]	[5.92, 5.57]	[111.45, 102.51]	[18.56, 14.64]	[20.02, 17.01]

Table 2.3: Simulation performance for explicitly bivariate ODE governed synchronous observations ($\lambda = 1/4$)

		$\boldsymbol{\sigma}$=[0.3, 0.3]			$\boldsymbol{\sigma}$=[1.0, 1.0]			$\boldsymbol{\sigma}$=[2.0, 2.0]		
		($\times K\Delta \times 10^{-3}$)			($\times K\Delta \times 10^{-3}$)			($\times K\Delta \times 10^{-3}$)		
# of follow-up		EB	EV	MISE	EB	EV	MISE	EB	EV	MISE
Scenario 1: Independent										
n=100										
X1, X2	k=[5, 5]	[40.42, 38.14]	[1.52, 1.44]	[2.57, 2.49]	[100.57, 89.80]	[15.06, 11.28]	[16.53, 13.12]	[194.20, 164.53]	[60.94, 41.21]	[63.04, 44.50]
X1, X2	k=[10, 10]	[136.28 83.27]	[29.17, 5.11]	[43.56, 11.15]	[192.88, 108.04]	[45.30, 9.84]	[70.17, 17.44]	[268.51, 143.60]	[74.79, 24.09]	[115.13, 32.53]
n=300										
X1, X2	k=[5, 5]	[33.69, 31.73]	[0.67, 0.75]	[1.76, 1.78]	[64.73, 59.72]	[5.42, 4.32]	[6.69, 5.84]	[118.85, 105.12]	[21.16, 15.35]	[22.68, 17.91]
X1, X2	k=[10, 10]	[161.13, 92.84]	[33.13, 5.52]	[55.63, 13.25]	[179.32, 99.01]	[39.26, 7.05]	[64.89, 14.74]	[228.96, 119.65]	[52.64, 12.18]	[90.25, 21.00]
Scenario 2: Correlated within-variable and between-variable (partially)										
$\rho_1 = 0.3$, $\rho_2 = 0.1$										
n=100										
X1, X2	k=[5, 5]	[42.61, 39.52]	[1.69, 1.64]	[2.84, 2.66]	[108.98, 98.53]	[18.01, 13.89]	[19.78 15.90]	[208.44, 177.15]	[66.17, 48.38]	[70.08, 52.18]
X1, X2	k=[10, 10]	[145.45, 87.05]	[31.33, 5.43]	[48.03, 12.01]	[204.56, 112.36]	[50.40, 11.54]	[77.21, 19.08]	[285.93, 158.28]	[82.55, 30.61]	[127.72, 39.41]
n=300										
X1, X2	k=[5, 5]	[34.68, 31.55]	[0.77, 0.81]	[1.89, 1.77]	[68.62, 63.34]	[6.07, 4.98]	[7.42, 6.55]	[133.03, 121.41]	[228.00, 265.34]	[230.62, 268.00]
X1, X2	k=[10, 10]	[166.92, 93.98]	[34.10, 5.60]	[58.47, 13.57]	[187.49, 103.86]	[40.70, 7.82]	[68.60, 16.26]	[245.62, 125.64]	[55.77, 14.10]	[100.31, 23.29]
$\rho_1 = 0.3$, $\rho_2 = 0.5$										
n=100										
X1, X2	k=[5, 5]	[43.31, 40.01]	[1.82, 1.68]	[2.93, 2.72]	[109.94, 99.59]	[17.79, 14.17]	[19.35, 16.09]	[211.93, 182.92]	[69.07, 51.28]	[72.23, 55.13]
X1, X2	k=[10, 10]	[143.10, 85.63]	[30.26, 5.37]	[46.46, 11.59]	[208.10, 115.65]	[48.92, 11.94]	[78.00, 19.72]	[292.86, 160.70]	[82.91, 32.62]	[132.42, 41.39]
n=300										
X1, X2	k=[5, 5]	[35.14, 32.88]	[0.76, 0.85]	[1.91, 1.93]	[92.04, 85.08]	[11.95, 9.82]	[13.52, 11.68]	[132.96, 116.80]	[26.62, 19.28]	[28.73, 21.89]
X1, X2	k=[10, 10]	[164.62, 93.59]	[34.00, 5.64]	[57.52, 13.45]	[194.50, 106.57]	[40.68, 7.73]	[71.37, 16.52]	[241.26, 125.31]	[56.38, 14.62]	[97.41, 23.27]
Scenario 3: Correlated within-variable and between-variable (fully)										
n=100										
X1, X2	k=[5, 5]	[45.53, 41.11]	[2.06, 1.86]	[3.25, 2.86]	[121.34, 108.22]	[21.65, 17.00]	[23.57, 19.13]	[224.48, 202.85]	[78.23, 63.14]	[81.53, 67.93]
X1, X2	k=[10, 10]	[137.55, 85.69]	[28.76, 5.33]	[43.23, 11.54]	[205.04, 121.27]	[50.33, 13.96]	[73.73, 21.19]	[294.85, 179.01]	[92.11, 42.46]	[133.01, 50.69]
n=300										
X1, X2	k=[5, 5]	[35.60, 32.68]	[0.89, 0.90]	[1.99, 1.89]	[76.02, 68.64]	[7.90, 6.22]	[9.24, 7.66]	[143.29, 128.85]	[30.81, 23.87]	[32.98, 26.93]
X1, X2	k=[10, 10]	[167.30, 96.50]	[33.32, 5.86]	[57.73, 14.04]	[192.95, 109.01]	[40.73, 8.54]	[69.68, 17.00]	[249.70, 135.54]	[59.35, 16.86]	[101.41, 26.27]

Table 2.4: Simulation performance for explicitly bivariate ODE governed asynchronous observations ($\lambda = 1/4$)

		$\boldsymbol{\sigma}$=[0.3, 0.3]			$\boldsymbol{\sigma}$=[1.0, 1.0]			$\boldsymbol{\sigma}$=[2.0, 2.0]		
		($\times K\Delta \times 10^{-3}$)			($\times K\Delta \times 10^{-3}$)			($\times K\Delta \times 10^{-3}$)		
	# of follow-up	EB	EV	MISE	EB	EV	MISE	EB	EV	MISE
Scenario 1: Independent										
n=100										
X1, X2	k=[5, 5]	[40.57, 38.56]	[1.47, 1.41]	[2.55, 2.54]	[98.71, 91.06]	[14.27, 11.12]	[15.64, 13.37]	[195.06, 167.67]	[61.50, 45.59]	[63.75, 49.44]
X1, X2	k=[10, 10]	[145.33, 85.99]	[30.82, 5.19]	[47.89, 11.70]	[195.43, 109.24]	[45.55, 10.14]	[71.46, 17.95]	[268.61, 148.37]	[71.80, 25.66]	[114.38, 34.71]
n=300										
X1, X2	k=[5, 5]	[33.77, 32.58]	[0.68, 0.72]	[1.77, 1.86]	[63.33, 60.49]	[5.05, 4.28]	[6.29, 5.94]	[116.81, 105.97]	[20.95, 15.71]	[22.42, 18.28]
X1, X2	k=[10, 10]	[166.87, 95.47]	[33.49, 5.58]	[58.05, 13.84]	[190.29, 102.23]	[40.40, 6.89]	[70.60, 15.39]	[233.49, 118.37]	[53.62, 11.56]	[93.59, 20.82]
Scenario 3: Correlated within-variable and between-variable (fully)										
n=100										
X1, X2	k=[5, 5]	[43.91, 41.59]	[1.99, 1.90]	[3.04, 2.96]	[113.59, 105.47]	[19.26, 15.96]	[20.70, 18.01]	[219.25, 190.46]	[78.82, 56.49]	[81.53, 60.31]
X1, X2	k=[10, 10]	[135.70, 83.84]	[28.42, 5.20]	[42.34, 11.22]	[202.06, 115.74]	[50.43, 12.65]	[74.16, 19.81]	[286.00, 164.74]	[87.94, 34.99]	[127.34, 42.89]
n=300										
X1, X2	k=[5, 5]	[34.89, 32.97]	[0.81, 0.91]	[1.87, 1.94]	[72.59, 67.47]	[7.07, 5.99]	[8.44, 7.48]	[136.90, 122.74]	[28.81, 22.19]	[30.70, 24.51]
X1, X2	k=[10, 10]	[174.63, 96.88]	[35.08, 5.78]	[62.33, 14.31]	[195.99, 105.83]	[42.28, 7.66]	[73.15, 16.21]	[248.99, 128.98]	[60.01, 15.08]	[102.25, 24.01]

Table 2.5: Simulation performance for bivariate functional synchronous observations ($\lambda = 5.0$)

		$\boldsymbol{\sigma}$=[0.3, 0.3]			$\boldsymbol{\sigma}$=[1.0, 1.0]			$\boldsymbol{\sigma}$=[2.0, 2.0]		
		($\times K\Delta \times 10^{-3}$)			($\times K\Delta \times 10^{-3}$)			($\times K\Delta \times 10^{-3}$)		
	# of follow-up	EB	EV	MISE	EB	EV	MISE	EB	EV	MISE
Scenario 1: Independent										
n=100										
X1, X2	k=[5, 5]	[139.62, 151.60]	[47.78, 47.00]	[68.68, 71.89]	[244.49, 256.08]	[146.16, 141.93]	[163.28, 169.50]	[421.43, 434.41]	[508.84, 492.98]	[520.32, 534.03]
X1, X2	k=[10, 10]	[500.95, 526.11]	[55.78, 40.19]	[356.99, 396.79]	[492.84, 525.42]	[64.46, 49.91]	[348.59, 396.13]	[485.06, 521.99]	[91.00, 79.72]	[342.52, 394.27]
n=300										
X1, X2	k=[5, 5]	[128.61, 150.92]	[25.50, 29.33]	[37.13, 54.38]	[187.35, 210.21]	[69.20, 62.13]	[83.29, 95.22]	[281.69, 304.00]	[159.05, 141.55]	[174.00, 184.42]
X1, X2	k=[10, 10]	[578.62, 576.30]	[14.60, 13.88]	[440.84, 456.98]	[571.76, 571.93]	[24.67, 20.39]	[434.97, 452.78]	[565.57, 569.96]	[34.88, 32.22]	[430.29, 452.24]
Scenario 2: Correlated within-variable and between-variable (partially)										
$\rho_1 = 0.3$, $\rho_2 = 0.1$										
n=100										
X1, X2	k=[5, 5]	[145.74, 156.29]	[51.72, 51.82]	[73.74, 77.10]	[260.63, 266.41]	[169.11, 140.07]	[192.56, 173.31]	[435.01, 444.26]	[509.45, 491.01]	[526.34, 524.55]
X1, X2	k=[10, 10]	[511.58, 535.85]	[49.38, 34.97]	[367.54, 407.55]	[496.74, 524.84]	[69.08, 52.60]	[355.17, 396.52]	[492.87, 526.82]	[93.48, 81.17]	[354.51, 401.46]
n=300										
X1, X2	k=[5, 5]	[127.15, 152.22]	[26.05, 29.42]	[37.64, 54.79]	[187.23, 213.85]	[67.72, 60.68]	[81.91, 94.48]	[296.63, 316.13]	[167.66, 142.60]	[190.04, 189.29]
X1, X2	k=[10, 10]	[574.12, 573.58]	[19.32, 16.22]	[436.67, 454.51]	[563.32, 565.82]	[31.81, 25.53]	[427.07, 446.41]	[568.28, 572.52]	[35.86, 32.09]	[433.61, 455.68]
$\rho_1 = 0.3$, $\rho_2 = 0.5$										
n=100										
X1, X2	k=[5, 5]	[146.83, 156.90]	[54.93, 53.84]	[78.11, 78.95]	[264.79, 272.41]	[166.60, 173.33]	[195.76, 203.46]	[433.28, 445.15]	[541.11, 577.50]	[557.74, 609.04]
X1, X2	k=[10, 10]	[504.66, 530.78]	[53.19, 37.44]	[360.18, 401.76]	[508.75, 530.94]	[63.86, 47.44]	[367.51, 403.92]	[490.94, 522.54]	[93.55, 76.57]	[352.93, 395.00]
n=300										
X1, X2	k=[5, 5]	[127.30, 152.25]	[24.57, 30.18]	[36.02, 55.60]	[194.19, 216.54]	[77.27, 64.92]	[93.91, 99.88]	[296.01, 317.17]	[171.17, 145.19]	[190.45, 191.54]
X1, X2	k=[10, 10]	[570.93, 569.98]	[21.54, 17.64]	[433.62, 450.11]	[568.38, 570.22]	[25.54, 21.08]	[431.74, 450.53]	[564.83, 568.49]	[37.61, 31.81]	[430.70, 448.97]
Scenario 3: Correlated within-variable and between-variable (fully)										
n=100										
X1, X2	k=[5, 5]	[148.10, 161.90]	[56.50, 51.71]	[76.48, 79.94]	[293.45, 284.91]	[240.99, 169.43]	[261.17, 205.59]	[508.89, 493.93]	[987.42, 693.81]	[997.14, 735.28]
X1, X2	k=[10, 10]	[515.73, 539.35]	[48.84, 32.87]	[371.08, 411.31]	[515.59, 536.92]	[71.09, 47.56]	[376.30, 413.46]	[527.85, 552.93]	[121.12, 91.54]	[403.14, 443.56]
n=300										
X1, X2	k=[5, 5]	[133.30, 157.00]	[29.59, 31.26]	[41.74, 57.76]	[210.03, 230.93]	[89.01, 69.68]	[104.65, 107.36]	[330.14, 339.89]	[208.91, 175.40]	[227.41, 221.77]
X1, X2	k=[10, 10]	[578.39, 574.86]	[18.28, 15.17]	[441.93, 456.66]	[580.25, 577.27]	[23.52, 18.46]	[446.02, 460.38]	[588.03, 584.62]	[40.51, 33.25]	[462.79, 476.22]

Table 2.6: Simulation performance for bivariate functional asynchronous observations ($\lambda = 5.0$)

		$\boldsymbol{\sigma}$=[0.3, 0.3]			$\boldsymbol{\sigma}$=[1.0, 1.0]			$\boldsymbol{\sigma}$=[2.0, 2.0]		
		($\times K\Delta \times 10^{-3}$)			($\times K\Delta \times 10^{-3}$)			($\times K\Delta \times 10^{-3}$)		
	# of follow-up	EB	EV	MISE	EB	EV	MISE	EB	EV	MISE
Scenario 1: Independent										
n=100										
X1, X2	k=[5, 5]	[182.08, 174.00]	[111.19, 60.23]	[149.47, 90.81]	[273.61, 260.80]	[182.57, 138.55]	[227.69, 173.10]	[424.58, 427.11]	[444.88, 527.08]	[487.93, 570.92]
X1, X2	k=[10, 10]	[511.40, 537.10]	[49.51, 33.37]	[366.47, 408.25]	[494.70, 524.47]	[66.75, 51.48]	[351.75, 395.07]	[491.34, 525.12]	[87.99, 72.74]	[351.07, 398.39]
n=300										
X1, X2	k=[5, 5]	[139.30, 162.53]	[37.54, 35.52]	[51.29, 63.04]	[190.71, 212.84]	[74.43, 62.12]	[90.44, 95.83]	[284.43, 303.06]	[155.60, 134.77]	[175.17, 177.95]
X1, X2	k=[10, 10]	[578.27, 575.52]	[17.48, 14.84]	[441.71, 457.03]	[572.24, 573.36]	[25.97, 20.58]	[435.85, 455.14]	[571.23, 575.75]	[32.12, 28.09]	[436.19, 458.20]
Scenario 3: Correlated within-variable and between-variable (fully)										
n=100										
X1, X2	k=[5, 5]	[187.86, 174.62]	[111.91, 62.49]	[148.37, 90.91]	[316.34, 298.24]	[237.31, 169.50]	[281.28, 204.85]	[509.29, 504.46]	[675.63, 562.45]	[703.10, 605.39]
X1, X2	k=[10, 10]	[512.06, 539.04]	[47.39, 32.91]	[366.06, 410.23]	[516.66, 539.52]	[69.86, 50.17]	[376.45, 416.03]	[533.69, 558.27]	[119.42, 98.54]	[409.45, 451.66]
n=300										
X1, X2	k=[5, 5]	[140.02, 161.85]	[38.76, 36.12]	[52.92, 63.33]	[215.87, 233.13]	[89.45, 74.85]	[107.97, 111.41]	[335.24, 347.33]	[209.46, 183.74]	[232.28, 232.55]
X1, X2	k=[10, 10]	[586.88, 580.45]	[13.18, 11.97]	[450.57, 463.45]	[583.33, 579.52]	[21.65, 18.28]	[448.74, 463.46]	[586.47, 586.62]	[39.60, 34.04]	[457.36, 476.96]

2.4 Application

Alzheimer's disease (AD), a gradually advancing disorder, is currently ranked as the sixth leading cause of death in the United States (Kumar et al., 2022). Three primary categories of AD biomarkers have been well validated and incorporated into clinical diagnostic criteria, and are frequently used in therapeutic trials. These include: (i) amyloid (typically measured with amyloid Positron Emission Tomography (PET) (Drzezga, 2010)) (ii) tau (measured with cerebrospinal fluid (CSF) phosphorylated tau or tau PET)(Fagan et al., 2009) and (iii) neurodegeneration (fluorodeoxyglucose (FDG)-PET, or CSF total tau) (Jack Jr and Holtzman, 2013). It is important to understand how these biomarkers interact with each other and influence cognitive changes longitudinally.

In this study, we aim to investigate the longitudinal dynamic relationships among all three primary categories of biomarkers along with cognitive progression. Additionally, we examine the potential interaction effect of the apolipoprotein $\epsilon 4$ ($ApoE4$) gene by conducting separate analyses for $ApoE4$ carriers and non-carriers. We utilize longitudinal observed data, which includes PET imaging, CSF biomarker measurements, and cognitive measurements.

2.4.1 Data

The data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (adni.loni.usc.edu). The measurements analyzed consist of irregular, sparse longitudinal data that were observed asynchronously.

To investigate the dynamics of biomarkers and cognitive function over time, all measurements were aligned to years from the onset of Alzheimer's disease. The dynamic system analyzed in this study includes seven components: (i) Amyloid PET: "WHOLECEREBELLUM SUVR" AV45-PET; (ii) AV45 ratio: AV45-PET (cortical grey matter/the whole cerebellum); (iii) Total Tau PET: "META TEMPORAL SUVR" AV1451-PET; (iv) FDG: the average fluorodeoxyglucose (FDG)-PET of angular, temporal, and posterior cingulate regions; (v) CSF Tau: cerebrospinal fluid (CSF) total tau; (vi) CSF PTau: CSF phosphorylated tau (pTau); (vii) ADAS13: Alzheimer's Disease Assessment Scale-Cognitive Subscale 13 scores.

The average number of repeated measurements for each component was 3.00, 1.94, 1.62, 1.83, 1.94, 1.94, and 3.37, respectively. The data for each variable were standardized by its mean and standard deviation. A total of 245 subjects were included in the analyses, comprising 95 $ApoE4$ carriers, 127 $ApoE4$ non-carriers, and 23 indi-

viduals whose *ApoE*4 status was missing. The latter were excluded from the neural ODEs fittings but included for data standardization.

2.4.2 Neural ODEs estimate

We fitted two separate time-invariant neural ODE systems for *ApoE*4 carriers and non-carriers. Given that the initial values were unknown, we used the average observed values of each component within the first year of onset as the initial guess. To refine the estimate of initial values, we conducted the Algorithm (1) until convergence. We then plotted the estimated dynamic changes of the biomarkers and the cognitive score of interest in Figure 2.5. The results revealed complex interactions among biomarkers and cognitive performance in both populations. However, the *ApoE*4 carrier population demonstrated more intense and intricate entanglement associations among the biomarkers and cognitive performance compared to *ApoE*4 non-carriers.

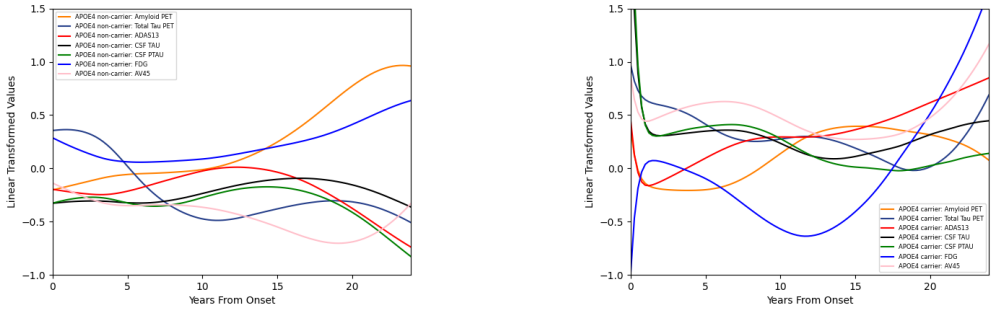
In *ApoE*4 non-carriers, the population mean of Amyloid PET showed a continuous increase over 20 years from the onset. For the first 5 years, other measurements either remained stable or slightly decreased, then started to follow the increasing trend of Amyloid PET until around 15 years from the onset. After this point, ADAS13 scores began to decline. Five years later, tau-related biomarkers (Total tau PET, CSF tau, CSF Ptau) decreased, followed by the decline of amyloid PET. On the other hand, FDG kept a continuously increasing trend. The measurement of florbetapir cortical normalized by the whole cerebellum (AV45) exhibited a trend almost opposite to that of Amyloid PET, which measures the florbetapir mean of the whole cerebellum.

The curves fitted in Figure 2.5b suggest that the *ApoE*4 carrier population was governed by a distinct dynamic system from the *ApoE*4 non-carriers. Unlike the clear divergence trends observed in the *ApoE*4 non-carriers, all components in the *ApoE*4 carrier population alternately fluctuated up and down over time. Except for AV45 and ADAS13, the other variables showed a slight but steady downward trend until 15 years. FDG was the first to turn upward, occurring within 10 to 15 years after disease onset, with the other components following behind. Given that AV45 and Amyloid PET almost mirrored each other. Along with interactions with other biomarkers, an overall upward trend of ADAS13 can be identified in *ApoE*4 carriers, which typically indicates greater clinical cognitive dysfunctions. However, such a trend is not observed in the non-carrier population.

The refined method of initial values is presented in Figure 2.6 for the model of *ApoE*4 carriers and non-carriers. In the non-carrier group, the refinement results remained relatively consistent with the initial guesses, which were calculated using the average of observations within the first year of onset. On the other hand, significant refinement changes were observed in the carrier group. However, after a sufficient number of training epochs, the estimates converged to stable values.

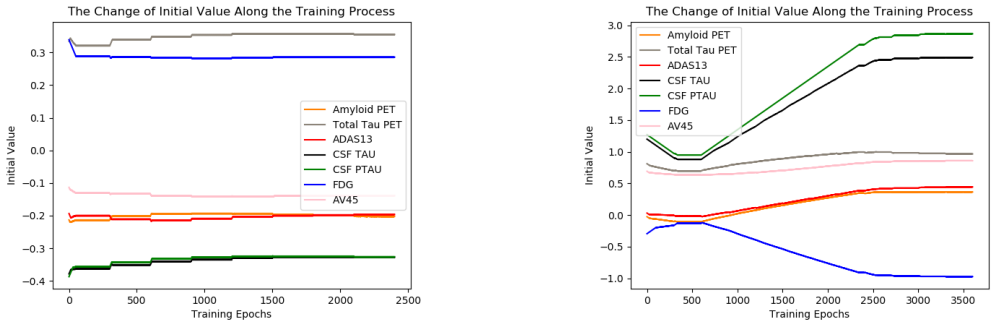
In this study, the bootstrap method was used to assess the uncertainty of the model. The same number of subjects as in the original observed data were randomly re-sampled with replacement at the subject level. Fitted curves based on the original observed values (thick curves) and 100 fitted curves (thin curves) using bootstrap resampling data are displayed in the appendix (Figure 2.7 and Figure 2.8). It is not surprising that the variance of estimation increases at the time of disease onset and as we move to the tail end, given that not all individuals have corresponding

measurements available at the time of disease onset and the average lifespan after a diagnosis of Alzheimer's disease is four to eight years, resulting in extremely sparse observations beyond that range.



(a) $\hat{\boldsymbol{\mu}}(t)$ of *ApoE*4 Non-carriers **(b)** $\hat{\boldsymbol{\mu}}(t)$ of *ApoE*4 Carriers

Figure 2.5: Neural ODE fitted dynamic change



(a) $\hat{\boldsymbol{\mu}}_0$ of *ApoE*4 Non-carriers **(b)** $\hat{\boldsymbol{\mu}}_0$ of *ApoE*4 Carriers

Figure 2.6: The refine training of initial value

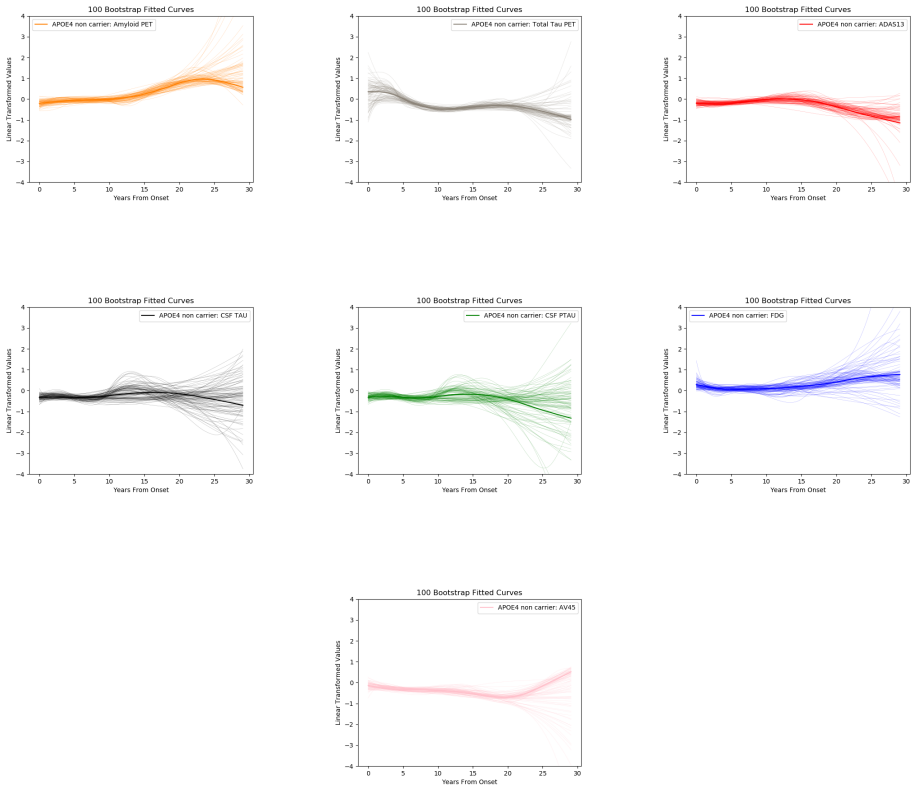


Figure 2.7: Bootstrap results of $\hat{\boldsymbol{\mu}}(t)$ for ApoE non-carriers

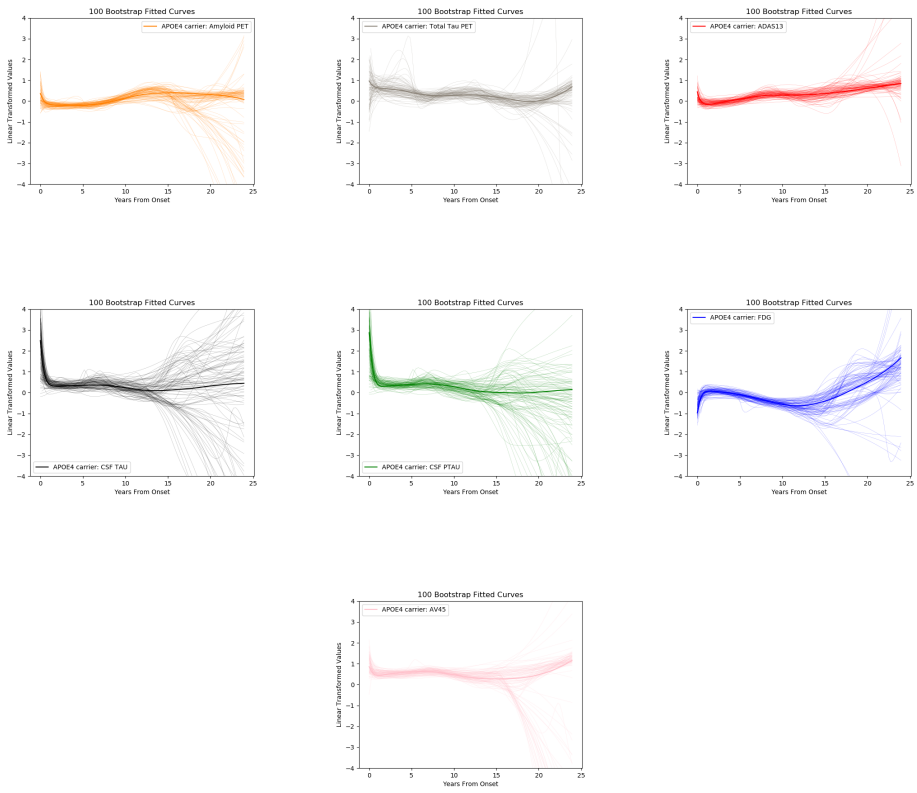


Figure 2.8: Bootstrap results of $\hat{\boldsymbol{\mu}}(t)$ for ApoE carriers

2.5 Conclusion

This paper makes a methodological contribution in proposing neural network-based ODE estimators for modeling the dynamic change of sparse multivariate longitudinal observed data by estimating the instantaneous changing rates of a dynamic system. Neural networks as a non-parametric approach, are specifically structured to exploit the complex associations among multivariate components. This method is particularly useful when ODEs cannot be formulated as parametric or semi-parametric models, which is often the case in real-world biological studies. In simulation studies, we evaluated the fitted dynamic systems with explicit and implicit ODEs and found

that neural ODEs perform well in both studies, in terms of the empirical estimation of the cumulative integral differences between the actual functional curves $\boldsymbol{\mu}(t)$ and the estimated curves $\hat{\boldsymbol{\mu}}(t)$, empirical variance (EV) and the mean squared error (MSE). Model performance can be improved by gathering more useful information, such as increasing the number of observations, enlarging sample sizes, and data density. Also, the more the observations spread out within the target time range the better the method performed. The quality of data can also improve the model performance by reducing noise and complexity correlations. In the application study, we modeled the change of 6 potentially interactively associated biomarkers with 1 cognitive measurement of Alzheimer's disease. They were sparsely and asynchronously measured depending on patients' visits. We trained two separate neural ODEs for *ApoE*4 carriers and non-carriers and observed different dynamic changes between those two subgroups. As the functional form of ODEs can be flexibly modeled and estimated using neural networks, our method is particularly attractive in analyzing multivariate longitudinal data. Our methods can be straightforwardly extended to more complicated including longitudinally observed imaging data that change over time in a dynamic fashion.

2.6 Appendix

2.6.1 Positive definite variance-covariance matrix

We show how to pick ρ_1 and ρ_2 so that the following variance-covariance matrix is positive definite:

$$\begin{pmatrix} \sigma_1^2 \begin{pmatrix} 1 & \rho_1^\delta & \rho_1^{2\delta} \\ \rho_1^\delta & 1 & \rho_1^\delta \\ \rho_1^{2\delta} & \rho_1 & 1 \end{pmatrix} & \sigma_1\sigma_2 \begin{pmatrix} \rho_2 & 0 & 0 \\ 0 & \rho_2 & 0 \\ 0 & 0 & \rho_2 \end{pmatrix} \\ \sigma_1\sigma_2 \begin{pmatrix} \rho_2 & 0 & 0 \\ 0 & \rho_2 & 0 \\ 0 & 0 & \rho_2 \end{pmatrix} & \sigma_2^2 \begin{pmatrix} 1 & \rho_1^\delta & \rho_1^{2\delta} \\ \rho_1^\delta & 1 & \rho_1^\delta \\ \rho_1^{2\delta} & \rho_1^\delta & 1 \end{pmatrix} \end{pmatrix}.$$

A matrix is positive definite if and only if all of its eigenvalues are positive. Assuming $\sigma_1 = \sigma_2 = \sigma$, the eigenvalues of the matrix are

$$\sigma^2 \begin{pmatrix} 1 - \rho_2 + \frac{\rho_1^\delta}{2}\left(\rho_1^\delta - \sqrt{\rho_1^{2\delta} + 8}\right) \\ 1 - \rho_2 + \frac{\rho_1^\delta}{2}\left(\rho_1^\delta + \sqrt{\rho_1^{2\delta} + 8}\right) \\ 1 - \rho_2 - \rho_1^{2\delta} \\ 1 + \rho_2 - \rho_1^{2\delta} \\ 1 + \rho_2 + \frac{\rho_1^\delta}{2}\left(\rho_1^\delta - \sqrt{\rho_1^{2\delta} + 8}\right) \\ 1 + \rho_2 + \frac{\rho_1^\delta}{2}\left(\rho_1^\delta + \sqrt{\rho_1^{2\delta} + 8}\right) \end{pmatrix}$$

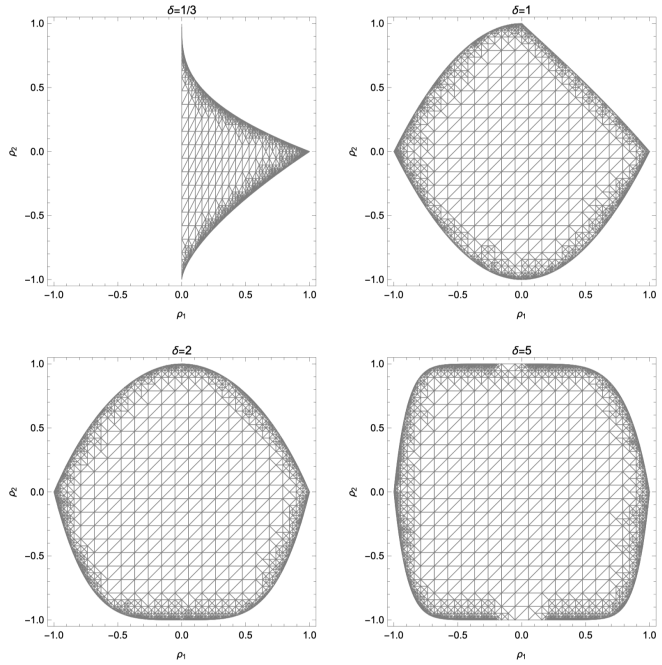


Figure 2.9: Regions in (ρ_1, ρ_2) where the variance-covariance matrix is positive definite

Suppose $-1 < \rho_2 < 1$. Then $1 \pm \rho_2 > 0$. When $\delta \geq 1$, a set of sufficient conditions for the eigenvalues to be positive is

$$1 - \rho_2 + \frac{\rho_1^\delta}{2}\left(\rho_1^\delta - \sqrt{\rho_1^{2\delta} + 8}\right) > 0, \qquad 1 - \rho_2 + \frac{\rho_1^\delta}{2}\left(\rho_1^\delta + \sqrt{\rho_1^{2\delta} + 8}\right) > 0,$$
$$1 - \rho_2 - \rho_1^{2\delta} > 0, \qquad 1 + \rho_2 - \rho_1^{2\delta} > 0.$$

When $1 > \delta > 0$, ρ_1 has to be positive to make the eigenvalues a real number. Consequently, $1 - \rho_2 + \frac{\rho_1^\delta}{2}\left(\rho_1^\delta + \sqrt{\rho_1^{2\delta} + 8}\right)$ is always positive. The set of sufficient conditions for the eigenvalues to be positive becomes

$$1 - \rho_2 + \frac{\rho_1^\delta}{2}\left(\rho_1^\delta - \sqrt{\rho_1^{2\delta} + 8}\right) > 0, \qquad 1 - \rho_2 - \rho_1^{2\delta} > 0, \quad 1 + \rho_2 - \rho_1^{2\delta} > 0.$$

We plot the regions in $\{\rho_1, \rho_2\}$ that satify the sufficient conditions for $\delta = 1/3, 1, 2, 5$ respectively in Figure 2.9. With increasing δ, the region of ρ_1 and ρ_2 that satisfy

the sufficient conditions becomes larger. Both positive and negative correlations are allowed. We chose $\{\rho_1, \rho_2\}$ for the following scenarios:

- Strong within-component correlation and weak positive between-outcome correlation: $\rho_1 = 0.3, \rho_2 = 0.1$;
- Strong within-component correlation and strong negative between-outcome correlation: $\rho_1 = 0.3, \rho_2 = 0.5$

2.6.2 Evaluate of MISE, Bias, and Variance

Let $\hat{\mu}_k^{(m)}(t)$ be the estimate of $\mu_k(t)$ using the mth replicate. The empirical bias (EB) of the estimate $\hat{\mu}_k(t)$ is

$$\text{EB} = \int_0^\tau \frac{1}{M} \sum_{m=1}^{M} \left| \hat{\mu}_k^{(m)}(t) - \mu_j(t) \right| dt.$$

The empirical variance (EV) of the estimate $\hat{\mu}_k(t)$ is

$$\text{EV} = \int_0^\tau \frac{1}{M} \sum_{m=1}^{M} \left\{ \hat{\mu}_k^{(m)}(t) - \text{EM}(t) \right\}^2 dt,$$

where

$$\text{EM}(t) = \frac{1}{M} \sum_{m=1}^{M} \hat{\mu}_k^{(m)}(t).$$

The empirical mean integrated squared error (MISE) is evaluated as

$$\text{MISE} = \int_0^\tau \frac{1}{M} \sum_{m=1}^{M} \left\{ \hat{\mu}_k^{(m)}(t) - \mu_k(t) \right\}^2 dt.$$

We use the trapezoidal method to evaluate the integrals in MISE, Bias, and variance. We evaluate the expectations in the formula empirically using M simulated replicates.

Chapter 3

Deep Latent ODE Models for Longitudinal Data

3.1 Introduction

Longitudinal data are increasingly commonly collected in many fields, such as clinical trials, epidemiology, and medical science. In many longitudinal studies, including Alzheimer's research, researchers are interested in understanding how risk factors could influence the longitudinal progression of disease-related biomarkers while complex unobservable pathological mechanisms are considered. To make inferences about the effects of interest interventions with individual variation, parametric linear mixed models (LMMs) are usually applied for longitudinal data analysis, in which fixed effects are used to estimate the effect of interventions and random effects account for individual deviations from the average effect and introduce within-subject covariance (Laird and Ware, 1982). Although LMMs are widely used, they are parametrically modeled using maximum likelihood methods, which rely on linear parametric specifications for both the baseline function and random effects. Furthermore, LMMs assume that random effects follow normal distributions. These assumptions can be challenging in practice, especially when the underlying dynamic is complex and nonlinear.

To relax the linearity assumption, stochastic models were proposed. These models specify the within-subject covariance structure and individual variation using a stochastic process (Taylor et al., 1994). Other semiparametric mixed models extend

LMMs by modeling time effect using a nonparametric function, while other covariates are parametrically modeled (Zeger and Diggle, 1994; Zhang et al., 1998). The inference on fixed effects has been demonstrated to be robust to the nonnormality of the random effects (Butler and Louis, 1992; Verbeke and Lesaffre, 1997). To relax the normality assumption for random effects, nonparametric maximum likelihood estimation has been widely used (Laird, 1978). This algorithm does not make parametric assumptions on the random effects distribution. However, the nonparametric maximum likelihood estimate of the distribution is discrete and has been criticized due to its unrealistic. Some researchers offered flexible random effects distribution by approximating the random effects density by Hermite series (Chen et al., 2002; Zhang and Davidian, 2001). Others modeled the random effects with a finite mixture of Gaussian density and developed a heterogeneity model (Verbeke and Lesaffre, 1996). A penalized Gaussian mixture approach was also proposed to smooth the density of random effects (Ghidey et al., 2004). All these approaches make inferences using frequentist methods, which rely on observable random covariates. Unobserved latent covariates are common in many real-world scenarios.

Motivated by the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, active research is interested in disentangling the effects of treatments, risk factors, and genetic status in the longitudinal progression of Alzheimer's diseases (Bernal-Rusiel et al., 2013; Chen et al., 2021; Cho et al., 2013; Mielke et al., 2011). Evidence from Alzheimer's disease studies indicates that statistical methods with strict assumptions may not be flexible enough to accurately capture the intricate dynamic in their longitudinal data. Also, as complex pathological mechanisms are hard to fully observe, the latent variable analytic statistical method for characterizing AD progression is promis-

ing (Green et al., 2011). This motivated us to propose a novel statistical model that incorporates a flexible nonparametric stochastic process to account for unobservable latent complex dynamic covariates. Taking advantage of Bayesian inference, which allows for the relaxation of the distribution restrictions on latent variables, we propose our model based on Bayesian approaches.

Although LMMs are typically not identified as latent variable models, the random effects in these models can clearly be regarded as latent variables (Skrondal and Rabe-Hesketh, 2007). Bayesian approaches for parametric LMMs have been well-studied. Under the normality assumption of random effects, posterior inference can be easily conducted using the standard Gibbs sampling method. To relax the normality assumption, models based on Bayesian approaches that incorporate nonparametric random effect have been considered (Kleinman and Ibrahim, 1998; Mukhopadhyay and Gelfand, 1997; Müller and Rosner, 1997; Pennell and Dunson, 2007).

In this article, we propose a novel deep latent ODE model for longitudinal data analysis. This model is based on Bayesian methods and models the latent stochastic process as an unknown function of multivariate stochastic processes, governed by an Ordinary Differential Equation (ODE) system. Leveraging the universal approximation property of neural networks (Lu and Lu, 2020), we approximate both the unknown function and the ODE system using neural networks (Chen et al., 2018). Given the involvement of neural networks and ODEs, obtaining closed forms of posteriors is typically unfeasible. Therefore, we utilize approximation techniques from variational Inference (VI) (Gelman et al., 1995; Kingma and Welling, 2013; Rezende et al., 2014) to optimize hyperparameters. Inferences of fixed effects in our model are implemented using a Gibbs-type sampler, based on non-informative priors.

In Section 2, we provide details of our proposed statistical methods. We demonstrate the robustness of our approach through comprehensive simulated experiments in Section 3. In Section 4, we demonstrate our proposed method by applying it to Alzheimer's disease data. Finally, we conclude the paper with remarks and potential future research in Section 5.

3.2 Statistical methods

3.2.1 Statistical models

Let $y_i(t)$ be the outcome of interest and $\boldsymbol{x}_i(t)$ be a vector of p-dimensional covariates for the i-th subject at time t in a longitudinal follow-up study. Suppose that $y_i(t)$ satisfy the following semi-parametric model for the i-th subject,

$$y_i(t) = f\{\boldsymbol{z}_i(t)\} + \boldsymbol{x}_i(t)'\boldsymbol{\beta} + \epsilon_i(t), \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a p-dimentional fixed effects associated with covariates $\boldsymbol{x}_i(t)$, $f(\cdot)$ is an unknown baseline function, $\boldsymbol{z}_i(t) = \{z_i^{(1)}(t), z_i^{(2)}(t), \ldots, z_i^{(q)}(t)\}$ is a q-dimensional latent stochastic process, and $\epsilon_i(t)$ is a process of random noise with mean zero.

The parameter of interest is $\boldsymbol{\beta}$, which represents the effect of covariates on the outcome. The latent stochastic processes $\boldsymbol{z}_i(t)$ are an unobservable random process that affects the outcome through an unknown function $f(\cdot)$. Assume that $\boldsymbol{z}_i(t)$ is independent of the covariates $\boldsymbol{x}_i(t)$. Then the population version of the model (3.1) is

$$E\Big\{y_i(t)|\boldsymbol{x}_i\Big\} = g(t) + \boldsymbol{x}_i(t)\boldsymbol{\beta}, \tag{3.2}$$

where

$$g(t) = E\Big\{ f\{\boldsymbol{z}_i(t)\}\Big\}.$$

Model (3.1) is thus can be considered as a generalization of the semiparametric additive model, where the outcome is a function of a linear combination of covariates and a baseline nonparametric function (Zeger and Diggle, 1994; Zhang et al., 1998). If $f(\cdot)$ is given, model (3.1) can also be seen as a generalization of the Linear Mixed Model (LMM) (Laird and Ware, 1982), where the outcome is a function of a linear combination of fixed effects and random effects through $\boldsymbol{z}_i(t)$. E.g., if $f(x) = x$,

$$y_i(t) = \boldsymbol{z}_i(t) + \boldsymbol{x}_i(t)\boldsymbol{\beta} + \epsilon_i(t).$$

where a nonparametric or flexible distribution for random effects density can be offered (Taylor et al., 1994). In this paper, we propose to model the unknown function $f(\cdot)$ nonparametrically, using a neural network parametrized by ζ, denoted as $f_\zeta(\cdot)$.

3.2.2 The observsed data

Consider the situation where there are n i.i.d. subjects, for each subject i, longitudinal outcome, and covariates are observed at random follow-up times, denoted to be $\boldsymbol{t}_i = \{t_{i1}, \ldots, t_{iJ_i}\}'$ where J_i is the total number of observations for subject i. So that, the observed data can be concisely written as $\boldsymbol{y}_i = \{y_{i1}, y_{i2}, \ldots, y_{iJ_i}\}'$ and $\boldsymbol{X}_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{iJ_i}\}' \in \mathbb{R}^{J_i \times p}$ where $y_{ij} \equiv y_i(t_{ij})$ and $\boldsymbol{x}_{ij} \equiv \boldsymbol{x}_i(t_{ij})$, for $j = 1, 2, \ldots, J_i$.

3.2.3 Model for the latent process

In the current paper, we modeled the latent variables $\boldsymbol{z}_i(t)$ as stochastic processes in latent space (demonstrated in Figure 3.1). The unobservable latent processes are assumed to be governed by an ODE system composed of q-dimensional differential equations with initial values $\boldsymbol{z}_{i0} = (z_{i0}^{(1)}, z_{i0}^{(2)}, \ldots, z_{i0}^{(q)})'$. For each subject i, we have:

$$\frac{d\boldsymbol{z}_i(t)}{dt} = \boldsymbol{f}_{\boldsymbol{\eta}}(\boldsymbol{z}_i(t)),$$
$$\boldsymbol{z}_{i0} \equiv \boldsymbol{z}_i(0),$$

where $\boldsymbol{f}_{\boldsymbol{\eta}}(\cdot) = (f^{(1)}(\cdot), f^{(2)}(\cdot), \ldots, f^{(q)}(\cdot))'$ are unknown functions and parameterized by $\boldsymbol{\eta}$. To specify the ODE system nonparametrically, we approximate ODE functions $\boldsymbol{f}_{\boldsymbol{\eta}}(\cdot)$ as a neural network that provides the most flexibility for both the dimension and the modeling structure of unobservable latent space.

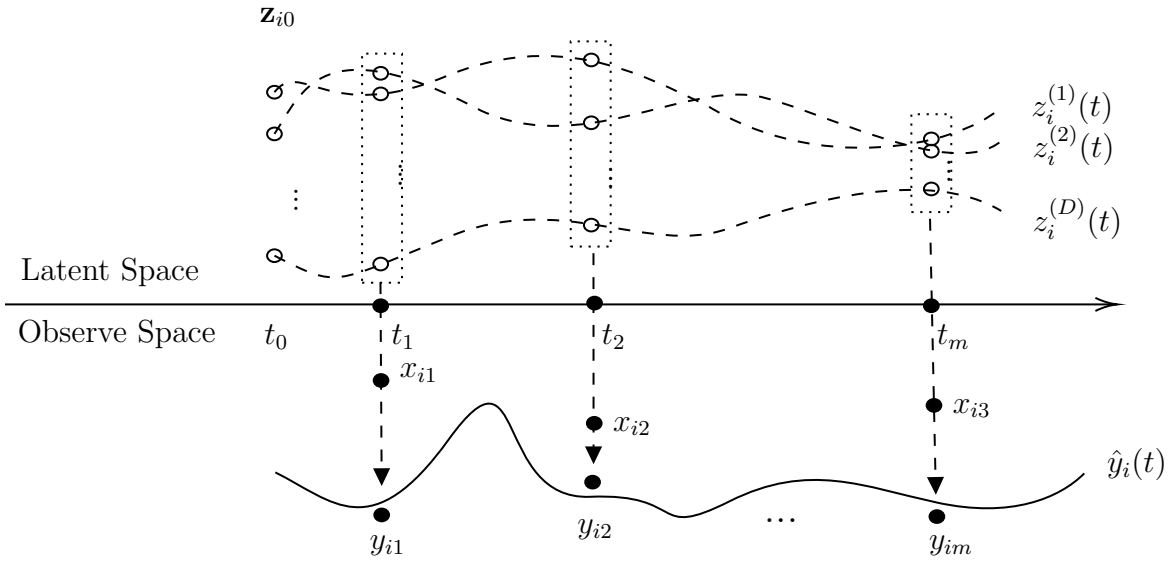


Figure 3.1: Demonstration of the proposed model

3.2.4 Bayesian methods

Posterior distribution of fixed effects $\boldsymbol{\beta}$

Given the value of latent dynamic at specific time point $\boldsymbol{z}_i(t_{ij})$ and unknown function $f_{\boldsymbol{\zeta}}(\cdot)$, posterior distributions of our interest fixed effects $\boldsymbol{\beta}$ and the variance of error term σ^2 are carried out by iterative Markov Monte Carlo (MCMC) method. The noninformative uniform prior used in the standard Bayesian linear regression model on $(\boldsymbol{\beta}, \log\sigma)$ is applicable in our model or equivalently,

$$\boldsymbol{\beta}, \sigma^2 \propto \sigma^{-2}$$

Given a current estimate of $f_{\boldsymbol{\zeta}}(\boldsymbol{z}_i(t_{ij}))$, we can reformulate Model (3.1) in the form of linear regression as follows:

$$y_{ij} - f_{\boldsymbol{\zeta}}(\boldsymbol{z}_i(t_{ij})) = \tilde{y}_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \epsilon_{ij}$$

So that $\tilde{y}_{ij}$ follows a normal distribution with mean $\boldsymbol{x}_{ij}\boldsymbol{\beta}$ and variance σ^2, the conditional posterior distribution of $\boldsymbol{\beta}$ also follows a normal distribution, with mean and variance as follows:

$$\begin{aligned}
\boldsymbol{\beta} \mid \sigma, \boldsymbol{y}, \boldsymbol{z}_0, \boldsymbol{X} &\sim N(\hat{\boldsymbol{\beta}}, V_\beta\sigma^2), \\
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\tilde{\boldsymbol{y}}, \\
V_\beta &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}.
\end{aligned}$$

The conditional posterior distribution of σ^2 has a scaled inverse-χ^2 form:

$$\sigma^2 \mid \boldsymbol{y}, \boldsymbol{z}_0, \boldsymbol{x}, \boldsymbol{\beta} \sim Inv - \chi^2(N-p, s^2)$$

where

$$s^2 = \frac{1}{N-p}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\tilde{\boldsymbol{y}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

and $N = \sum_{i=1}^{n} J_i$ which is the total number of all observations.

Given observed data, prior information, and posterior distribution of $\boldsymbol{z}_0$, we applied a Gibbs-type sampler to obtain a draw $(\boldsymbol{z}_0, \sigma^2, \boldsymbol{\beta})$ from the intractable joint distribution $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{z}_0|\boldsymbol{y}, \boldsymbol{X})$ as demonstrated in Algorithm (3) below.

Algorithm 3: Gibbs Sampler of $\boldsymbol{\beta}$ and σ^2

Step 1. Specify initial values $\boldsymbol{\beta}^{(0)}, \sigma^{2^{(0)}}, \boldsymbol{z}_0^{(0)}$

Step 2. Cycle through the conditional drawing

(a) $\boldsymbol{z}_0^{(q)}$ from $p(\boldsymbol{z}_0 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}^{(q-1)}, \sigma^{2^{(q-1)}})$

(b) $\sigma^{2^{(q)}}$ from $p(\sigma^2|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}_0^{(q)}, \boldsymbol{\beta}^{(q-1)})$

(c) $\boldsymbol{\beta}^{(q)}$ from $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}_0^{(q)}, \sigma^{2^{(q)}})$

Step 3. Set $q = q + 1$, and go to step 2.

However, because the posterior distribution of $\boldsymbol{z}_0$ depends on the unknown function $f_{\boldsymbol{\zeta}}(\cdot)$, which is parameterized by high-dimensional hyperparameters $\boldsymbol{\zeta}$ and the complex unobserved latent variables $\boldsymbol{z}(t)$, it can not be analytically solved in our model. Therefore, we proposed to find an approximation posterior distribution of $\boldsymbol{z}_0$ based on the variational inference (VI) technique to conduct the Gibbs sampling method.

Posterior distribution of initial values $\boldsymbol{z}_0$ of latent dynamics

Based on the Bayes Rule, the posterior distribution of $\boldsymbol{z}_0$ can be formulated as below,

$$p(\boldsymbol{z}_0 \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = \frac{p(\boldsymbol{z}_0, \boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)}$$

which is equivalent to

$$p(\boldsymbol{z}_0 \mid \boldsymbol{y}^*, \sigma^2) = \frac{p(\boldsymbol{z}_0, \boldsymbol{y}^* | \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{y}^* | \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)}$$

where $\boldsymbol{y}^* = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$.

Based on Model (3.1), data points y_{ij}^* are assumed to be conditionally independently distributed, which can be expressed as follows,

$$y_{ij}^* = y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta} = f_{\boldsymbol{\zeta}}(\boldsymbol{z}_i(t_{ij})) + \epsilon_{ij}$$

According to our model assumptions, the conditional distribution of the data y_{ij}^* follows a normal distribution,

$$p(y_{ij}^* \mid \boldsymbol{z}_i(t_{ij}), \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \sigma^2) \sim \mathcal{N}(f_{\boldsymbol{\zeta}}(\boldsymbol{z}_i(t_{ij})), \sigma^2)$$

where the first-moment parameters are approximated by a function of the latent stochastic process, denoted as $f_{\boldsymbol{\zeta}}(\boldsymbol{z}_i(t_{ij}))$ and the second-moment parameter σ^2 is updated by the Gibbs sampler algorithm in Section 2.3.

Assuming that the latent processes $\boldsymbol{z}(t)$ are governed by an ODE system, and by applying the fundamental theory of calculus, we can numerically solve the fitted

latent process $\tilde{\boldsymbol{z}}i(t)$ given subject-specific initial values $\boldsymbol{z}_{i0}$ and hyperparameter $\boldsymbol{\eta}$. This process can be characterized by its initial values $\boldsymbol{z}_0$, as below,

$$\tilde{\boldsymbol{z}}_i(t) = g_{\boldsymbol{\eta}}(\boldsymbol{z}_{i0}, t) = \boldsymbol{z}_{i0} + \int_0^t f_{\boldsymbol{\eta}}(g(\boldsymbol{z}_{i0}, s))ds \tag{3.3}$$

where function $g(\cdot)$ can be any arbitrary ODE numeric solving method chosen from widely used approaches, such as Dormand–Prince (RKDP) method (Dormand and Prince, 1980), Euler's method (Euler, 1792), Midpoint method, etc. To provide a high level of accuracy while controlling the error and adaptive step size, Runge-Kutta of order 5 of the Dormand-Prince-Shampine method is used as the default choice in the current paper. So that $p(\boldsymbol{y}^* \mid \boldsymbol{z}(t), \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)$ is equivalent to $p(\boldsymbol{y}^* \mid \boldsymbol{z}_0, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)$.

Under the assumption that the responses $\boldsymbol{y}^*$ are conditionally independent given $\boldsymbol{z}(t), \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2$, we can calculate the log-likelihood as follows:

$$\begin{aligned}
\log p(\boldsymbol{y}^* \mid \boldsymbol{z}(t), \boldsymbol{\beta}, \boldsymbol{X}, \sigma^2) &= \log p(\boldsymbol{y}^* \mid \boldsymbol{z}_0, \boldsymbol{\beta}, \boldsymbol{X}, \sigma^2) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{J_i} \log p(y_{ij}^* \mid \boldsymbol{z}_{i0}, \boldsymbol{\beta}, \boldsymbol{x}_{ij}, \sigma^2) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{J_i} \log N(y_{ij}^*; f_\zeta(g_{\boldsymbol{\eta}}(\boldsymbol{z}_{i0}, t_{ij})), \sigma^2)
\end{aligned}$$

Given $\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2$ and under the assumption that the prior distribution of $\boldsymbol{z}_0$ does not depend on $\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2$, the marginal distribution of $\boldsymbol{y}^*$ can be calculated by,

$$p(\boldsymbol{y}^* \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = \int p(\boldsymbol{y}^* \mid \boldsymbol{z}_0, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{z}_0) d\boldsymbol{z}_0$$

However, as unknown functional approximations, neural networks, and ODE become involved, the marginal distribution of $\boldsymbol{y}^*$ becomes intractable. Given that the posterior distribution of the initial value is dependent on the marginal distribution of $\boldsymbol{y}^*$, it follows that the posterior distribution $p_{\boldsymbol{\theta}}(\boldsymbol{z}_0 \mid \boldsymbol{y}^*, \sigma^2)$ is also intractable, where $\boldsymbol{\theta} = [\boldsymbol{\zeta}', \boldsymbol{\eta}']'$.

Variational inference

To approximate the posterior distribution of $\boldsymbol{z}_0$, we propose to use the variational inference (VI) technique which relies on minimizing the Kullback-Leibler (KL) divergence between the approximation posterior distribution of $\boldsymbol{z}_0$, parameterized by hyperparameters $\boldsymbol{\phi}$, denoted as $q_{\boldsymbol{\phi}}(\boldsymbol{z}_0 \mid \boldsymbol{y}^*, \sigma^2)$ and the true posterior density of $\boldsymbol{z}_0$, parameterized by hyperparameters $\boldsymbol{\theta}$, denoted as $p_{\boldsymbol{\theta}}(\boldsymbol{z}_0 \mid \boldsymbol{y}^*, \sigma^2)$. The minimization is performed with respect to $\boldsymbol{\theta} = [\boldsymbol{\zeta}', \boldsymbol{\eta}']'$ and $\boldsymbol{\phi}$. By definition, KL can be written as below,

$$\begin{aligned} KL(q_{\boldsymbol{\phi}}, p_{\boldsymbol{\theta}} | \boldsymbol{y}^*, \sigma^2) &= -\int q_{\boldsymbol{\phi}}(\boldsymbol{z}_0 | \boldsymbol{y}^*, \sigma^2) \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{z}_0 | \boldsymbol{y}^*, \sigma^2)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_0 | \boldsymbol{y}^*, \sigma^2)} d\boldsymbol{z}_0 \\ &\geq -\log \left\{ \int q_{\boldsymbol{\phi}}(\boldsymbol{z}_0 | \boldsymbol{y}^*, \sigma^2) \frac{p_{\boldsymbol{\theta}}(\boldsymbol{z}_0 | \boldsymbol{y}^*, \sigma^2)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_0 | \boldsymbol{y}^*, \sigma^2)} d\boldsymbol{z}_0 \right\} = 0, \end{aligned}$$

where due to the convexity of $-\log(\cdot)$ and by Jensen's inequality, the KL divergence is always non-negative. Because of the intractability of $p_{\boldsymbol{\theta}}(\boldsymbol{z}_0 | \boldsymbol{y}^*, \sigma^2)$, the KL divergence is intractable. However, the so-called variational lower bound, or the evidence lower bound (ELBO) (Kingma et al., 2019; Neal and Hinton, 1998) can be computationally

evaluated. ELBO is defined as below:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}^*, \sigma^2) = \int q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*, \sigma^2) \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{y}^*, \boldsymbol{z}_0|\boldsymbol{\beta}, \sigma^2)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*, \sigma^2)} d\boldsymbol{z}_0.$$

Importantly, maximizing ELBO with respect to $(\boldsymbol{\theta}, \boldsymbol{\phi})$ would approximately maximize the likelihood function of $p(\boldsymbol{y}^*|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)$ and minimizing the KL divergence of $q_{\boldsymbol{\phi}}(\boldsymbol{z}_0 \mid \boldsymbol{y}^*, \sigma^2)$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z}_0 \mid \boldsymbol{y}^*, \sigma^2)$.

The optimization of the ELBO needs to be carried out jointly for both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ using methods such as stochastic gradient descent (SGD). However, calculating the gradient of ELBO with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, denoted as $\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{y}^* \mid \boldsymbol{\beta}, \sigma^2)$ is generally challenging. This is particularly due to the expectation with respect to $q_{\boldsymbol{\phi}}(\boldsymbol{z}_0 \mid \boldsymbol{y}^*, \sigma^2)$.

The gradient of the ELBO with respect to $\boldsymbol{\theta}$ can be calculated as follows:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{y}^* \mid \boldsymbol{\beta}, \sigma^2) &= \nabla_{\boldsymbol{\theta}} \Big[E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*, \sigma^2)} \{\log p_{\boldsymbol{\theta}}(\boldsymbol{y}^*, \boldsymbol{z}_0|\boldsymbol{\beta}, \sigma^2)\} - E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*, \sigma^2)} \{\log q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*, \sigma^2)\} \Big] \\
&= E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*, \sigma^2)} \big\{ \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}^*, \boldsymbol{z}_0|\boldsymbol{\beta}, \sigma^2) \big\}.
\end{aligned}$$

which can be evaluated using the Monte Carlo method, given that the *generative* model parameter $\boldsymbol{\theta}$ does not overlap with $\boldsymbol{\phi}$. However, the gradients with respect to $\boldsymbol{\phi}$ are not the case.

$$\begin{aligned}
\nabla_{\boldsymbol{\phi}} \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{y}^* \mid \boldsymbol{\beta}, \sigma^2) &= \nabla_{\boldsymbol{\phi}} E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*, \sigma^2)} \Big[\log p_{\boldsymbol{\theta}}(\boldsymbol{y}^*, \boldsymbol{z}_0|\boldsymbol{\beta}, \sigma^2) - \log q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*, \sigma^2) \Big] \\
&\neq E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*, \sigma^2)} \Big\{ \nabla_{\boldsymbol{\phi}} \Big[\log p_{\boldsymbol{\theta}}(\boldsymbol{y}^*, \boldsymbol{z}_0|\boldsymbol{\beta}, \sigma^2) - \log q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*, \sigma^2) \Big] \Big\},
\end{aligned}$$

To go around the problem, we transformed the density of $\boldsymbol{z}_0$ into a function of the density of a known random variable $\boldsymbol{\epsilon}$ by applying the re-parameterization trick (Kingma et al., 2019). Assuming the density of $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{\phi}$, $\boldsymbol{y}^*$ and $\boldsymbol{z}_0 \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*, \sigma^2)$ can be expressed as a straightforward function of $\boldsymbol{\epsilon}, \boldsymbol{z}_0$, and $\boldsymbol{\phi}$,

$$\boldsymbol{z}_0 = u(\boldsymbol{\epsilon}, \boldsymbol{y}^*, \boldsymbol{\phi}).$$

given $\boldsymbol{\phi}$ and $\boldsymbol{y}^*$, the density of approximate posterior distribution of $\boldsymbol{z}_0$ can be written as a function of the density of $\boldsymbol{\epsilon}$ and the determinant of the Jacobian matrix of the transformation u. By a change of variable, for given $\boldsymbol{\phi}$ and $\boldsymbol{y}^*$,

$$E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*,\sigma^2)} f(\boldsymbol{z}_0) = E_{p(\boldsymbol{\epsilon})} f(\boldsymbol{z}_0),$$

this allows the gradients with respect to $\boldsymbol{\phi}$ to be moved inside the expectation as follows. The resulting integration can be evaluated using the Monte-Carlo method,

$$\nabla_{\boldsymbol{\phi}} E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}^*,\sigma^2)} f(\boldsymbol{z}_0) = E_{p(\boldsymbol{\epsilon})} \nabla_{\boldsymbol{\phi}} f(\boldsymbol{z}_0).$$

To be specific, in the current article, we assume that the approximated posterior distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}_0 \mid \boldsymbol{y}^*, \sigma^2)$ follows a Gaussian distribution. This distribution is characterized by parameters $\boldsymbol{\nu}_{\boldsymbol{\phi}}(\boldsymbol{y}^*, \sigma^2)$ and $\boldsymbol{\tau}^2_{\boldsymbol{\phi}}(\boldsymbol{y}^*, \sigma^2)$. For the i-th subject, the approximate posterior distribution specific to this individual is as follows,

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}_{i0} \mid \boldsymbol{y}_i^*, \sigma^2) \sim \mathcal{N}(\boldsymbol{\nu}_{\boldsymbol{\phi}}(\boldsymbol{y}_i^*, \sigma^2), diag(\boldsymbol{\tau}^2_{\boldsymbol{\phi}}(\boldsymbol{y}_i^*, \sigma^2)))$$

where $\boldsymbol{\nu}_{\boldsymbol{\phi}}(\boldsymbol{y}_i^*, \sigma^2)$ and $\boldsymbol{\tau}_{\boldsymbol{\phi}}^2(\boldsymbol{y}_i^*, \sigma^2)$ are two D dimensional vectors. These vectors are approximated by Recurrent Neural Network (RNN) with D representing the dimension of latent variables.

Approximation posterior distribution of initial values by RNN

To infer the parameters for the approximate posterior distribution of initial value, given $\boldsymbol{y}_i^*$ and σ^2, we utilized a Recurrent Neural Network (RNN). The RNN recursively processes the observed sequence $\boldsymbol{y}^* = \{\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_n^*\}'$ and σ^2 which is updated by Gibbs-sampler. Throughout this process, the RNN maintains its internal hidden state $\boldsymbol{h}$. Rather than processing the sequential data in its original order as in classical RNN, we fed it into the RNN in reverse. At each timestamp m, the RNN takes the corresponding input $\boldsymbol{y}_m^* \in \boldsymbol{R}^d$ (where d is the batch size) and the previous hidden state $\boldsymbol{h}_{m+1}$. It then updates its hidden state $\boldsymbol{h}_m \in \boldsymbol{R}^k$ (where k is the dimension of hidden state) as follows:

$$\boldsymbol{h}_m = f_{W,U}(\boldsymbol{y}_t^*, \sigma^2, \boldsymbol{h}_{m+1}),$$

with $f_{\boldsymbol{\phi}}(\cdot)$ is a deterministic non-linear neural network function, and $\boldsymbol{\phi} = [W, U, V]'$ is the parameter set of f.

In longitudinal data, it's common to encounter random follow-up observations. This can lead to varying numbers of follow-up times for each participant, resulting in irregularly timed observations. One approach to handle this is to use padding techniques and incorporate the time change between observations, denoted as $\Delta t_{i,m} = t_{i,m} - t_{i,m-1}$, into the RNN's update function (Rubanova et al., 2019):

$$\boldsymbol{h}_m = f_{W,U}(\boldsymbol{y}_m^*, \sigma^2, \boldsymbol{h}_{m+1}, \Delta t),$$

As demonstrated in Figure (3.2), we utilize a many-to-one RNN architecture. The initial hidden state is given as h_{m+1}. The parameters W, U, and V which are shared across each timestamp, will be optimized during the training epochs. The final hidden state of the RNN is used as input to estimate the parameters of the surrogate posterior distribution over $\boldsymbol{z}_0$, which can be expressed as below,

$$(\boldsymbol{\nu}_{\boldsymbol{\phi}}(\boldsymbol{y}^*, \sigma^2), \boldsymbol{\tau}^2_{\boldsymbol{\phi}}(\boldsymbol{y}^*, \sigma^2)) = f_V(h_1),$$

where function $f_V(\cdot)$ is an unknown function parametrized by V. It maps the hidden state from R^k to R^{2D}. The dimensions of the outputs are determined by the dimension of latent variables $\boldsymbol{z}(t)$.

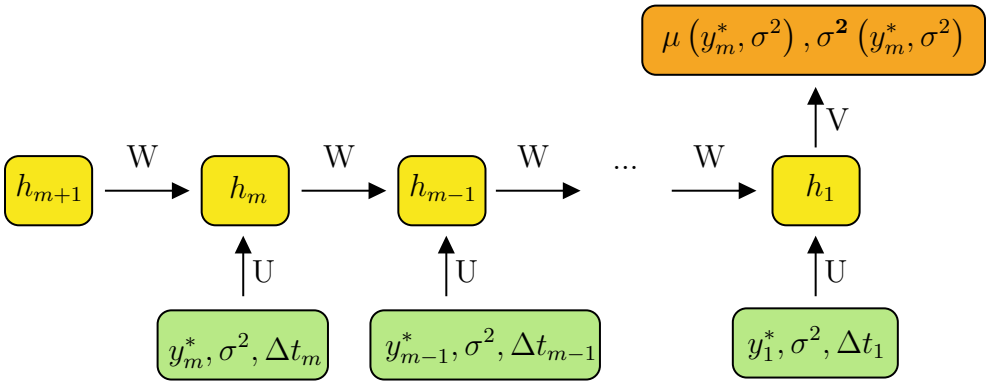


Figure 3.2: Demonstration of Recurrent Neural Network (RNN) for approximating the posterior distribution of initial values

3.2.5 Computation

Estimation of hyperparameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$

In our proposed model, both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are high dimensional parameters for the deep neural networks. They are considered nuisance parameters for which we do not aim

to make inferences on them. However, they play a crucial role in approximating inferences for the key parameters of interest, $\boldsymbol{\beta}$ and σ^2. Given $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{x}_i, \boldsymbol{z}_{i0})$, we assume that y_{ij}^* are independent for $j = 1, \ldots, J_i$. The prior distribution of $\boldsymbol{z}_{i0}$ is not dependent on $\boldsymbol{\beta}$ and σ^2. In line with the independence assumption, given $\boldsymbol{\beta}$, σ^2 and the dataset, the ELBO objective is computed as the average of individual ELBOs,

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{y}^* \mid \boldsymbol{\beta}, \sigma^2) &= \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(y_{ij} \mid \boldsymbol{\beta}, \sigma^2) \\ &= \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \left[\log \left[\frac{p_\theta(y_{ij}^* \mid \boldsymbol{z}_{i0}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{z}_{i0})}{q_\phi(\boldsymbol{z}_{i0} \mid y_{ij}^*, \sigma^2)} \right] \right]\end{aligned}$$

where $N = \sum_{i=1}^{n} J_i$, N is the total number of observations. The details of the algorithm are described in Algorithm 4.

Estimating of fixed effects $\boldsymbol{\beta}$

Once the parameter $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are optimized using Algorithm 4, and $\boldsymbol{z}_0$ is sampled from its approximate posterior distribution, we estimate our parameters of interest, $\boldsymbol{\beta}$ and σ^2 using the Bayesian approach. The posterior distribution of each parameter is constructed by Gibbs samplers, as shown in Algorithm 3, 4 and Figure 3.3. After a burn-in period, we compute the posterior mean of the p-th fixed effect, $\hat{\beta}_p$ by taking every r-th sample obtained during the sampling procedure, where r is a pre-defined arbitrary value.

Algorithm 4: Estimating of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$

Input: Observed longitudinal data

$$\{t_{ij}, \boldsymbol{x}_i(t_{ij}), y_i(t_{ij}), \text{ for } i = 1, \ldots, n, j = 1, \ldots, J_i\}$$

Output: Posterior distribution of $\boldsymbol{\beta}$ and σ^2

Take random initial values for the parameters

$(\boldsymbol{\beta}^{(0)}, \sigma^{2^{(0)}}, \boldsymbol{\phi}^{(0)}, \boldsymbol{\theta}^{(0)} = [\boldsymbol{\zeta}^{(0)}, \boldsymbol{\eta}^{(0)}])$

for *iteration* $\leftarrow 1$ **to** Q **do**

RNN **step** Approximate $\boldsymbol{\nu}_{\boldsymbol{\phi}}(\boldsymbol{y}^*, \sigma^2), \boldsymbol{\tau}_{\boldsymbol{\phi}}(\boldsymbol{y}^*, \sigma^2)$ using RNN based on Algorithm 5

Reparametrized **step** Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$

Reparameterized $\boldsymbol{z}_0$ as $\boldsymbol{z}_0 = \boldsymbol{\nu}_{\boldsymbol{\phi}}(\boldsymbol{y}^*, \sigma^2) + \boldsymbol{\tau}_{\boldsymbol{\phi}}(\boldsymbol{y}^*, \sigma^2) \odot \boldsymbol{\epsilon}$

ODE **step** **for** $i \leftarrow 1$ **to** n **do**

$\boldsymbol{t}_i = (t_{i1}, t_{i2}, \ldots, t_{iJ_i})$

Estimate $\boldsymbol{z}_i(\boldsymbol{t}_i) = g_{\boldsymbol{\eta}^{(q-1)}}(\boldsymbol{z}_{i0}, \boldsymbol{t}_i)$

Compute $\hat{y}^*_{ij} = f_{\boldsymbol{\zeta}^{(q-1)}}(z_i(t_{ij}))$

ELBO **step** The first part of loss:

$\mathcal{L}_1 = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \log p_{\boldsymbol{\theta}^{(q-1)}}(y^*_{ij} | \boldsymbol{z}_{i0}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \sigma^2)$

The second part of loss:

$\mathcal{L}_2 = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \log p(\boldsymbol{z}_{i0}) - \log q_{\boldsymbol{\phi}^{(q-1)}}(\boldsymbol{z}_{i0} | y^*_{ij}, \sigma^2)$

Compute ELBO: $\mathcal{L}_{\text{ELBO}} = \mathcal{L}_1 + \mathcal{L}_2$

Update $\boldsymbol{\theta}^{(q)}$ and $\boldsymbol{\phi}^{(q)}$ optimizing the objective function $\mathcal{L}_{\text{ELBO}}$ by SGD

Gibbs **step** `/* Conduct Gibbs Sampling in Algorithm 3 */`

Sample $\boldsymbol{z}_0^{(q)}$ from surrogate posterior distribution $q_{\boldsymbol{\phi}^{(q)}}(\boldsymbol{z}_0 | \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\beta}, \sigma^2)$

Sample $\sigma^{2^{(q)}}$ from $p(\sigma^2 | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}_0^{(q)}, \boldsymbol{\beta}^{(q-1)})$

Sample $\boldsymbol{\beta}^{(q)}$ from $p(\boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}_0^{(q)}, \sigma^{2^{(q)}})$

return *posterior distribution of* $\boldsymbol{\beta}$ *and* σ^2

Algorithm 5: RNN with Padding

Input: Observed longitudinal data

$\{t_{ij}, \boldsymbol{x}_i(t_{ij}), y_i(t_{ij}),$ for $i = 1, \ldots, n, j = 1, \ldots, J_i\}$ Updated

parameters $\boldsymbol{\beta}^{(q-1)}$ and $\sigma^{2^{(q-1)}}$

Output: $\boldsymbol{\nu}_{\boldsymbol{\phi}}(\boldsymbol{y}^*, \sigma^2), \boldsymbol{\tau}_{\boldsymbol{\phi}}(\boldsymbol{y}^*, \sigma^2)$

Function `pad_sequence(`*sequence,max_length*`):`

- `/* Pad the sequence with zeros to match the max_length */`
- Padding_length = max_length - len(sequence)
- Padded_sequence = sequence+[0]* padding_length
- **return** *padded_sequence*

Padding **step** `/* Find the maximum sequence length in the batch */`

$J_{max} = max(J_1, J_2, \ldots, J_n)$;

for $i \leftarrow 1$ **to** n **do**

- padded_$\boldsymbol{t}_i$ = pad_sequence($\boldsymbol{t}_i, J_{max}$)
- padded_$\boldsymbol{y}_i$ = pad_sequence($\boldsymbol{y}_i, J_{max}$)
- **for** $k \leftarrow 1$ **to** p **do**
  - padded_$\boldsymbol{x}_{ik}$=pad_sequence($\boldsymbol{x}_{ik}, J_{max}$)

RNN **step** h_{m+1} = initial_hidden_state

for $m \leftarrow J_{max}$ **to** 1 **do**

- `/* make sure all inputs are 0 if others are padded */`
- **if** *padded_*$\boldsymbol{t}_i[m] = t_{i,m} = 0$ *and* $m! = 1$ **then**
  - $\sigma_{i,m}^{2^{(q-1)}} = 0$
- `/* calculate` $\Delta t_{i,m}$ `for each subject */`
- $\Delta t_{i,m} = t_{i,m} - t_{i,m-1}$
- $h_m = f_{W,U}(\boldsymbol{y}^*, \sigma^{2^{(q-1)}}, h_{m+1}, \Delta t)$

$(\boldsymbol{\nu}_{\boldsymbol{\phi}}(\boldsymbol{y}^*, \sigma^2), \boldsymbol{\tau}_{\boldsymbol{\phi}}^2(\boldsymbol{y}^*, \sigma^2)) = f_V(h_1)$

return $\boldsymbol{\nu}_{\boldsymbol{\phi}}(\boldsymbol{y}^*, \sigma^2), \boldsymbol{\tau}_{\boldsymbol{\phi}}^2(\boldsymbol{y}^*, \sigma^2)$

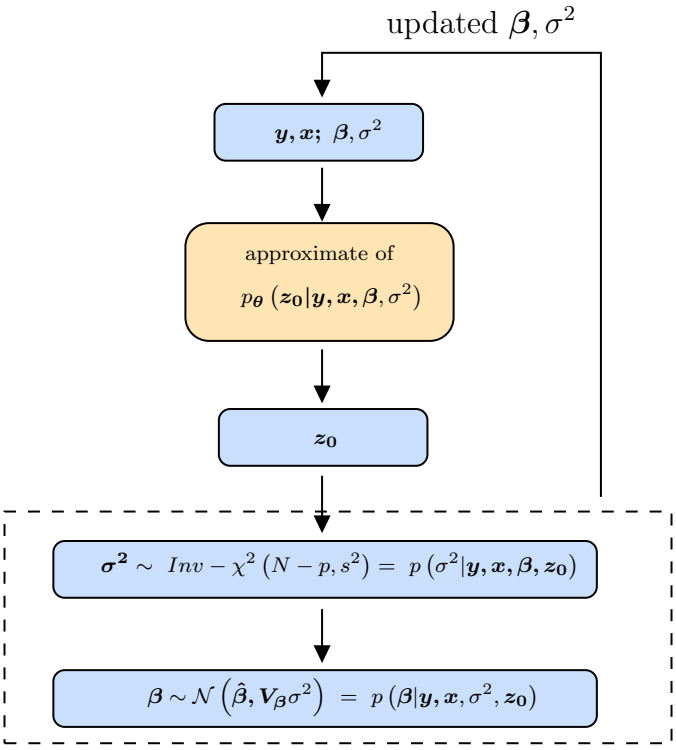


Figure 3.3: Demonstration of the Gibbs sampler for estimating $\boldsymbol{\beta}$ and σ^2

3.3 Simulation study

To evaluate the performance of our proposed model, we conducted a systematic simulation study under various situations. Without loss of generality, we considered two covariates $\boldsymbol{X}_i(t) = [\boldsymbol{x}_{i1}(t), \boldsymbol{x}_{i2}(t)]$ with their respective coefficients $\boldsymbol{\beta} = [\beta_1, \beta_2]'$. In the first setup of our simulation studies, we examined a continuous time process $Y_i(t)$, which is influenced by a latent Ordinary Differential Equation (ODE) system, observed covariates, and independently, identically distributed (i.i.d.) random error components. The observed outcome is expressed as:

$$Y_i(t) = z_{i1}(t)sin(t) + z_{i2}(t)cos(t) + \boldsymbol{x_i}(t)\boldsymbol{\beta} + \epsilon_i$$

where the latent variables $z_{i1}(t)$ and $z_{i2}(t)$ are driven by a dynamic system with initial values $z_{i1}^{(0)}, z_{i2}^{(0)} \sim N(1,1)$. The random noise ϵ_i is assumed to be $i.i.d.$ with a Gaussian distribution $\epsilon_i \sim N(0, \sigma^2)$.

The latent variables are governed by ODEs, which are formulated as follows. Figure 3.4 provides a demonstration of the true latent variable curves with initial values $z_1^{(0)} = z_2^{(0)} = 1$.

$$\frac{dz_{i1}(t)}{dt} = z_{i2}(t)$$
$$\frac{dz_{i2}(t)}{dt} = a \times sin(z_{i1}(t))$$

with $a = -5.0$.

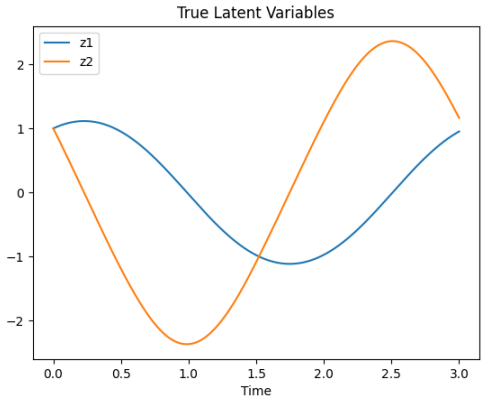


Figure 3.4: The true ODE solution of latent variables

In the second simulation setting, we introduced a time-dependent auto-regression (AR) random error component, in contrast to the first one, where an $i.i.d.$ random error was considered. To determine the robustness of our proposed method under the $i.i.d.$ assumption, the temporal correlations of repeated measurements arise not only from latent variables but also from error terms. In this study, we maintained

the same data generation process as in the first setting, but with an AR(1) process assumed for the random measurement error. This can be represented as follows,

$$Y_i(t) = z_{i1}(t)sin(t) + z_{i2}(t)cos(t) + \boldsymbol{x_i}(t)\boldsymbol{\beta} + \epsilon_i(t)$$

where $\boldsymbol{z}(t)$ is generated from the same ODE system as in the previous simulation setting. The random errors are assumed to follow an AR(1) process, as shown below,

$$\epsilon_i(t_{ij}) = \rho \cdot \epsilon_i(t_{i,j-1}) + \varepsilon_i(t_{ij}), \text{ for } j = 2, \cdots, J_i$$

where $\rho = 0.5$, and $\varepsilon_i(t_{ij})$ is a white noise process with mean zero and constant variance σ_ε^2. The mean of the random error process can be expressed as follows:

$$E(\epsilon_i(t_{ij})) = \rho E(\epsilon_i(t_{ij-1})) + E(\varepsilon_i(t_{ij}))$$

which can be easily derived that $E(\epsilon_i(t)) = 0$.

The variance of the random error process can be calculated as follows:

$$\begin{aligned} var(\epsilon_i(t_{ij})) &= E(\epsilon_i(t_{ij}^2)) \\ &= \rho^2 var(\epsilon_i(t_{ij-1})) + \sigma_\varepsilon^2 \\ &= \frac{\sigma_\varepsilon^2}{1-\rho^2} \end{aligned}$$

Meanwhile, the covariance between two random errors at different time points can be calculated as follows:

$$E(\epsilon_i(t_{ij})\epsilon_i(t_{ij-1})) = \frac{\sigma_\varepsilon^2}{1-\rho^2}\rho^{|n|}$$

In the third simulation setting, we deviated from assuming explicit ODE functions to govern the latent variables. Instead, we modeled the latent variables as the product of a time-dependent deterministic component and a stochastic component, represented by a Weiner Process. The observed curves can be represented as follows,

$$Y_i(t) = z_i(t)sin(t) + z_i(t)cos(t) + \boldsymbol{x_i}(t)\boldsymbol{\beta} + \epsilon_i$$

where the random noise ϵ_i is assumed to follow an $i.i.d.$ Gaussian distribution, denoted as $\epsilon_i \sim N(0, \sigma^2)$. The latent variables can be represented as below,

$$z_i(0) \sim N(0, 1)$$

$$z_i(t) - z_i(s) \sim N(0, t-s), t > s$$

In our simulation study, we evaluate two time-independent covariates $\boldsymbol{X}_i(t) = [\boldsymbol{x}_{i1}(t), \boldsymbol{x}_{i2}(t)]$ for $i = 1, \ldots, n$. Each covariate is drawn from either discrete or continuous distributions. In the discrete case, the covariates are assumed to be generated from a Bernoulli distribution with a probability of 0.5. In the continuous case, the covariates

are assumed to follow a normal distribution $N(0,1)$. The true values of the coefficients are $\beta_1 = 0$ and $\beta_2 = 1$.

We conducted a comprehensive exploration of various scenarios by considering all possible combinations of the following choices:

1. The number of subjects: (a) n=100 and (b) n=300;
2. The variance for the error: (a) $\sigma^2 = 1$, (b) $\sigma^2 = 4$;
3. The number of follow-ups: (a) $k = 5$, (b) $k = 10$, and (c) $k = 30$;
4. The distributions of covariates x_1 and x_2: (a) Bernoulli and Bernoulli (BB) (b) Normal and Normal (NN) distributions.

For each scenario in the simulation study, we conducted $M = 500$ replicates. In each replicate, we generated n subjects, each with k observations. To control the influence of random follow-ups, we assumed that all subjects adhered to the same schedule.

In the simulation studies, we trained our model using a fixed number of training epochs and burn-in periods, rather than setting a cutoff loss value. This approach offered more comparable model performance across different replications and simulation scenarios. As the number of follow-up observations increased, leading to a larger volume of information to process, we correspondingly increased the number of hidden nodes in our model. We used the initial values of each parameter as the starting points for the estimation process. Parameters $\boldsymbol{\beta}$ and σ^2 were initialized with arbitrary values, specifically $\beta_1^{(0)} = \beta_2^{(0)} = -1$ and $\sigma^{2^{(0)}} = 1$. Concurrently, the parameters of neural networks, denoted as $\boldsymbol{\phi}$, $\boldsymbol{\theta} = [\boldsymbol{\zeta}', \boldsymbol{\eta}']'$ were randomly initialized prior to model training.

3.3.1 Evaluation of the performance of the proposed model

We determined the posterior mean for each parameter, denoted as $\hat{\beta}_p$, by computing the average of every 10-th sample obtained following the burn-in period. The bias of the estimates for the p-th coefficient can be assessed by calculating the expectation of the absolute difference between the average estimate and the true value,

$$E\left\{|\hat{\beta}_p - \beta_p|\right\},$$

and we utilize the standard deviation (std) of the estimation for the p-th component as a measure to evaluate the uncertainty associated with the corresponding estimate. This can be calculated as follows,

$$\sqrt{E\left\{\hat{\beta}_p - E(\hat{\beta}_p)\right\}^2},$$

We empirically estimate the expectation using M replicated simulations. Consequently, the empirical bias (Ebias) for the p-th coefficient is evaluated as follows:

$$Ebias = \frac{1}{M}\sum_{m=1}^{M}|\hat{\beta}_{p,m} - \beta_p| \tag{3.4}$$

The empirical standard deviation (Estd) for the p-th coefficient is calculated as follows,

$$Estd = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left\{\hat{\beta}_{p,m} - \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_{p,m}\right\}^2} \tag{3.5}$$

We also computed the average of the standard deviation (std) and median of the posterior distribution. Additionally, to measure the uncertainty around parameter

estimates, we compute the coverage probability (CP). The CP represents the proportion of intervals, defined from the 2.5-th percentile to the 97.5-th percentile of the posterior distribution, that contains the true parameter value of interest.

3.4 Results

To offer a comprehensive summary of our model's performance across various simulation scenarios, we have presented the performance evaluation results from simulation studies 1 to 3 in tables 3.1, 3.2, and 3.3. These tables present key evaluation statistics, including empirical bias (Ebias), empirical standard deviation (Estd), average posterior std (APstd), average posterior median (APmedian), and coverage probability (CP).

The results demonstrate a strong alignment between the posterior mean estimates and the average posterior median, aligning closely with the true values used to generate the data. Each table provides insight into the effects of varying the number of subjects and follow-ups. When other factors are held constant, increasing the number of subjects from 100 to 300, or the number of follow-ups from 5 to 30, leads to a decrease in empirical bias, empirical standard deviation, and the average posterior standard deviation. We also investigated the impact of the variance of observed noise on our model's performance. In each study, we noted an increase in the variance of observed noise from 1 to 4 resulting in a corresponding increase in empirical bias, empirical standard deviation, and the average posterior standard deviation. We also evaluated our model's performance when covariates $\boldsymbol{x}$ were drawn from different distributions. Specifically, we compared scenarios where both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ followed Bernoulli distributions to those where they followed continuous normal distributions.

In most study scenarios, the estimation results showed notable improvement when $\boldsymbol{x}$ was drawn from normal distributions, as evidenced by reduced empirical bias, empirical standard deviation, and the average posterior standard deviation. However, in Table 2, we observed a few exceptions when 300 subjects each had 30 follow-ups with random noise of variance 1. In this scenario, the empirical bias did not improve when covariates were drawn from two Normal distributions compared to two Bernoulli distributions. Despite this, other statistics, including empirical standard deviation and average posterior standard deviation, showed improvements.

In terms of the coverage probabilities (CP) for the 95% credible intervals, they closely aligned with the desired value of 0.95. Table 1 shows that variations in the number of subjects, follow-ups, and covariate distributions have a limited impact on coverage probabilities. However, an increase in the variance of errors results in a decrease in coverage probabilities. Conversely, in Table 2, an increase in follow-ups corresponds to a decrease in coverage probability, while the number of subjects has a limited impact on CP. In most cases, sampling covariates from continuous distributions leads to an increase in coverage probability. Exceptions are observed for β_2 estimates in scenarios with 100 and 300 subjects each having 30 follow-ups. As the variance of errors increases from 1 to 4, a slight decrease in coverage probability is noted, except when covariates are drawn from Normal distributions. Notably, exceptions in coverage probabilities are seen for β_1 estimates with 100 and 300 subjects each having 30 follow-ups and β_2 estimates with 300 subjects each having 10 follow-ups. In Table 3, when covariates are sampled from continuous normal distributions, a substantial increase in coverage probabilities is observed. However, a slight decrease is noted for β_2 estimates in scenarios with 100 subjects each having 30 follow-ups.

Overall, the trends in coverage probabilities do not exhibit a consistent pattern across different scenarios in simulation study 3.

Table 3.1: Simulation performance for ODE governed latent variables with *i.i.d.* noise

			$\sigma^2 = 1$					$\sigma^2 = 4$				
	Follow-ups	Distributions	Ebias	Estd	APstd	APmedian	CP	Ebias	Estd	APstd	APmedian	CP
n=100			$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$
	5	BB	[0.09, 0.10]	[0.11, 0.13]	[0.12, 0.12]	[-0.01, 1.00]	[0.96, 0.92]	[0.16, 0.18]	[0.20, 0.23]	[0.19, 0.19]	[-0.02, 1.00]	[0.94, 0.90]
	10	BB	[0.06, 0.06]	[0.08, 0.08]	[0.07, 0.07]	[-0.02, 1.00]	[0.93, 0.92]	[0.12, 0.11]	[0.14, 0.14]	[0.13, 0.13]	[-0.03, 1.01]	[0.90, 0.92]
	30	BB	[0.03, 0.03]	[0.04, 0.04]	[0.04, 0.04]	[-0.02, 1.00]	[0.94, 0.92]	[0.06, 0.06]	[0.07, 0.08]	[0.07, 0.07]	[-0.03, 1.00]	[0.91, 0.92]
n=300												
	5	BB	[0.05, 0.06]	[0.06, 0.07]	[0.07, 0.07]	[-0.01, 1.00]	[0.96, 0.91]	[0.09, 0.09]	[0.11, 0.12]	[0.11, 0.11]	[-0.01, 1.01]	[0.92, 0.92]
	10	BB	[0.04, 0.04]	[0.04, 0.05]	[0.04, 0.04]	[-0.01, 0.99]	[0.93, 0.92]	[0.06, 0.06]	[0.08, 0.08]	[0.07, 0.07]	[-0.01, 1.01]	[0.91, 0.91]
	30	BB	[0.02, 0.02]	[0.02, 0.02]	[0.02, 0.02]	[-0.01, 0.99]	[0.94, 0.94]	[0.04, 0.03]	[0.04, 0.04]	[0.04, 0.04]	[-0.02, 1.00]	[0.93, 0.93]
n=100												
	5	NN	[0.05, 0.06]	[0.06, 0.08]	[0.07, 0.07]	[0.00, 1.01]	[0.97, 0.94]	[0.08, 0.09]	[0.10, 0.12]	[0.11, 0.12]	[0.01, 1.00]	[0.96, 0.93]
	10	NN	[0.03, 0.04]	[0.04, 0.04]	[0.04, 0.04]	[0.00, 1.00]	[0.97, 0.94]	[0.06, 0.07]	[0.07, 0.08]	[0.08, 0.08]	[0.01, 0.99]	[0.96, 0.93]
	30	NN	[0.02, 0.02]	[0.02, 0.02]	[0.02, 0.02]	[-0.01, 0.99]	[0.94, 0.93]	[0.03, 0.04]	[0.04, 0.04]	[0.04, 0.04]	[-0.00, 0.99]	[0.96, 0.93]
n=300												
	5	NN	[0.03, 0.03]	[0.03, 0.04]	[0.04, 0.04]	[0.00, 1.00]	[0.97, 0.93]	[0.05, 0.05]	[0.06, 0.07]	[0.06, 0.06]	[0.00, 1.00]	[0.96, 0.94]
	10	NN	[0.02, 0.02]	[0.02, 0.02]	[0.02, 0.02]	[0.00, 1.00]	[0.97, 0.94]	[0.03, 0.04]	[0.04, 0.05]	[0.04, 0.04]	[0.01, 0.99]	[0.97, 0.95]
	30	NN	[0.01, 0.01]	[0.01, 0.01]	[0.01, 0.01]	[-0.00, 0.99]	[0.95, 0.93]	[0.02, 0.02]	[0.02, 0.02]	[0.02, 0.02]	[0.00, 0.99]	[0.96, 0.94]

Table 3.2: Simulation performance for ODE governed latent variables with AR(1) noise

			$\sigma^2 = 1$					$\sigma^2 = 4$				
	Follow-ups	Distributions	Ebias	Estd	APstd	APmedian	CP	Ebias	Estd	APstd	APmedian	CP
n=100			$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$
	5	BB	[0.10, 0.11]	[0.13, 0.14]	[0.13, 0.13]	[-0.02, 1.01]	[0.95, 0.93]	[0.19, 0.21]	[0.24, 0.26]	[0.22, 0.22]	[-0.04, 1.02]	[0.94, 0.92]
	10	BB	[0.07, 0.07]	[0.09, 0.09]	[0.08, 0.08]	[-0.03, 1.01]	[0.91, 0.91]	[0.13, 0.14]	[0.16, 0.17]	[0.15, 0.15]	[-0.04, 1.02]	[0.91, 0.89]
	30	BB	[0.04, 0.05]	[0.05, 0.06]	[0.04, 0.04]	[-0.03, 1.00]	[0.87, 0.86]	[0.09, 0.09]	[0.10, 0.11]	[0.08, 0.08]	[-0.05, 1.02]	[0.84, 0.84]
n=300												
	5	BB	[0.06, 0.06]	[0.07, 0.08]	[0.07, 0.07]	[-0.01, 1.01]	[0.95, 0.92]	[0.10, 0.11]	[0.13, 0.13]	[0.12, 0.12]	[-0.02, 1.03]	[0.93, 0.91]
	5	BB	[0.04, 0.04]	[0.05, 0.05]	[0.05, 0.05]	[-0.02, 1.00]	[0.91, 0.91]	[0.07, 0.07]	[0.09, 0.09]	[0.08, 0.08]	[-0.02, 1.02]	[0.91, 0.91]
	30	BB	[0.03, 0.02]	[0.03, 0.03]	[0.03, 0.03]	[-0.02, 1.00]	[0.89, 0.90]	[0.05, 0.05]	[0.05, 0.06]	[0.05, 0.05]	[-0.03, 1.02]	[0.87, 0.88]
n=100												
	5	NN	[0.05, 0.07]	[0.07, 0.08]	[0.08, 0.08]	[0.01, 1.00]	[0.97, 0.93]	[0.09, 0.11]	[0.12, 0.14]	[0.13, 0.13]	[0.01, 1.00]	[0.96, 0.90]
	10	NN	[0.04, 0.04]	[0.05, 0.06]	[0.05, 0.05]	[0.00, 0.99]	[0.96, 0.91]	[0.07, 0.08]	[0.08, 0.11]	[0.09, 0.09]	[0.01, 0.99]	[0.96, 0.88]
	30	NN	[0.02, 0.03]	[0.03, 0.03]	[0.02, 0.03]	[-0.00, 0.99]	[0.91, 0.83]	[0.04, 0.05]	[0.05, 0.07]	[0.05, 0.05]	[0.00, 0.98]	[0.93, 0.83]
n=300												
	5	NN	[0.03, 0.04]	[0.04, 0.05]	[0.04, 0.04]	[0.01, 1.00]	[0.97, 0.93]	[0.05, 0.06]	[0.07, 0.08]	[0.07, 0.07]	[0.01, 0.99]	[0.97, 0.94]
	10	NN	[0.02, 0.02]	[0.02, 0.03]	[0.03, 0.03]	[0.01, 0.99]	[0.96, 0.92]	[0.04, 0.04]	[0.04, 0.05]	[0.05, 0.05]	[0.01, 0.98]	[0.97, 0.92]
	30	NN	[0.01, 0.02]	[0.01, 0.02]	[0.01, 0.01]	[0.00, 0.99]	[0.95, 0.81]	[0.02, 0.03]	[0.03, 0.03]	[0.03, 0.03]	[0.01, 0.98]	[0.93, 0.84]

Table 3.3: Simulation performance for Weiner process latent variables with *i.i.d.* noise

			$\sigma^2 = 1$					$\sigma^2 = 4$				
	Follow-ups	Distributions	Ebias	Estd	APstd	APmedian	CP	Ebias	Estd	APstd	APmedian	CP
n=100			$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$	$[\beta_1, \beta_2]$
	5	BB	[0.16, 0.19]	[0.20, 0.24]	[0.18, 0.18]	[-0.06, 0.96]	[0.92, 0.86]	[0.21, 0.24]	[0.26, 0.30]	[0.22, 0.22]	[-0.07, 0.96]	[0.90, 0.85]
	10	BB	[0.14, 0.16]	[0.16, 0.19]	[0.15, 0.15]	[-0.07, 0.96]	[0.90, 0.88]	[0.17, 0.18]	[0.20, 0.22]	[0.18, 0.18]	[-0.08, 0.96]	[0.90, 0.86]
	30	BB	[0.09, 0.09]	[0.11, 0.12]	[0.15, 0.14]	[-0.04, 0.99]	[0.92, 0.92]	[0.11, 0.12]	[0.13, 0.15]	[0.14, 0.14]	[-0.06, 0.98]	[0.93, 0.91]
n=300												
	5	BB	[0.09, 0.11]	[0.11, 0.12]	[0.11, 0.11]	[-0.02, 0.94]	[0.93, 0.91]	[0.11, 0.13]	[0.14, 0.15]	[0.13, 0.13]	[-0.03, 0.95]	[0.92, 0.88]
	10	BB	[0.08, 0.10]	[0.10, 0.11]	[0.18, 0.15]	[-0.03, 0.94]	[0.93, 0.86]	[0.09, 0.11]	[0.11, 0.13]	[0.11, 0.11]	[-0.03, 0.95]	[0.92, 0.88]
	30	BB	[0.07, 0.07]	[0.08, 0.08]	[0.07, 0.08]	[-0.05, 0.97]	[0.81, 0.88]	[0.08, 0.08]	[0.08, 0.09]	[0.09, 0.09]	[-0.06, 0.95]	[0.89, 0.90]
n=100												
	5	NN	[0.07, 0.08]	[0.09, 0.10]	[0.10, 0.10]	[-0.01, 1.04]	[0.96, 0.94]	[0.10, 0.11]	[0.12, 0.13]	[0.13, 0.13]	[-0.01, 1.04]	[0.97, 0.93]
	10	NN	[0.06, 0.07]	[0.07, 0.08]	[0.08, 0.08]	[-0.01, 1.03]	[0.97, 0.94]	[0.07, 0.09]	[0.09, 0.11]	[0.10, 0.10]	[-0.01, 1.03]	[0.97, 0.95]
	30	NN	[0.03, 0.05]	[0.04, 0.06]	[0.05, 0.05]	[-0.00, 1.02]	[0.98, 0.91]	[0.04, 0.06]	[0.05, 0.07]	[0.07, 0.07]	[-0.00, 1.03]	[0.99, 0.93]
n=300												
	5	NN	[0.04, 0.05]	[0.05, 0.05]	[0.06, 0.06]	[-0.01, 1.03]	[0.96, 0.94]	[0.05, 0.06]	[0.07, 0.07]	[0.08, 0.08]	[-0.01, 1.02]	[0.96, 0.94]
	10	NN	[0.03, 0.03]	[0.04, 0.04]	[0.05, 0.05]	[-0.01, 1.02]	[0.96, 0.94]	[0.04, 0.04]	[0.05, 0.05]	[0.06, 0.06]	[-0.01, 1.02]	[0.99, 0.97]
	30	NN	[0.02, 0.02]	[0.02, 0.03]	[0.03, 0.03]	[-0.00, 1.00]	[0.98, 0.97]	[0.02, 0.03]	[0.03, 0.04]	[0.05, 0.09]	[-0.00, 1.00]	[1.00, 0.98]

3.5 Application

Alzheimer's disease (AD) is the most common progressive neurodegenerative disease which affects a large population. It is currently ranked as the top 10 leading cause of death in the United States (Kumar et al., 2022). However, due to the involvement of various unobservable complex pathological mechanisms, the factors influencing the longitudinal progression of AD-related biomarkers are still not fully understood.

According to the World Health Organization (WHO), total tau is recognized as one of the most remarkable pathological characteristics of AD and has been widely used as a valuable biomarker for identifying individuals at risk of AD progression (WHO, 2023). In 2011, the National Institute on Aging and Alzheimer's Association introduced diagnostic recommendations that incorporate biomarkers, including cerebrospinal fluid (CSF) total Tau, for the preclinical stage of AD (Jack Jr et al., 2018). However, the influence of cognitive performance, gender, and gene status on tau levels remains controversial (Blomberg et al., 1996; Li et al., 2016; Williams et al., 2011), particularly among participants with different clinical diagnosis statuses (Bertens et al., 2015; Rolstad et al., 2011).

Conflicting findings have been reported regarding the relationship between cerebrospinal fluid (CSF) tau and disease severity. Some studies have shown that CSF tau levels are significantly higher in the AD group compared to mild cognitive impairment (MCI) and other dementing diseases (OD) (Blomberg et al., 1996). However, other studies have suggested that there is no significant overall change in CSF tau levels with disease progression (Sunderland et al., 1999). In addition to the varying findings of the association between CSF tau and disease status, divergent perspectives

have emerged regarding the potential impact of the apolipoprotein $\epsilon 4$ allele (*ApoE*4) on tau regulation. While some researchers suggest that there may be *ApoE*4-specific differences in tau progression in AD (Blomberg et al., 1996), others have concluded that temporal changes in CSF tau levels are not influenced by *ApoE*4 status (Arai et al., 1997). There is evidence suggesting a potential association between escalated tau levels and the *ApoE*4 genotype (Nagy et al., 1995; Tapiola et al., 1998). However, the interaction between *ApoE*4 status and underlying CSF tau pathology is still poorly understood. More importantly, factors affecting the progression of CSF tau levels should be taken into account such as disease status, *ApoE*4 as well effects of age, gender, etc.

Cross-sectional epidemiological studies have observed gender differences in the risk of developing AD (Andersen et al., 1999; Kukull et al., 2002; Miech et al., 2002). However, the underlying mechanisms remain unclear due to the limited number of studies focusing on sex differences. Hua et al. (2010) found a correlation between changes in CSF tau levels and 1-year atrophy rates, as well as *ApoE*4 status. They also reported that annual atrophy rates were slower in males than in females (Hua et al., 2010). Research has further explored whether genetic and hormonal mechanisms contribute to gender differences in Alzheimer's pathology. Some studies have reported that elevated levels of CSF tau in *ApoE*4 carriers are specifically associated with females (Damoiseaux et al., 2012). Altmann et al. (2014) found that average levels of CSF tau in MCI patients were higher in female *ApoE*4 carriers than in male *ApoE*4 carriers. However, they found no gender differences in normal controls, regardless of ApoE genotype.

The results mentioned earlier are all based on cross-sectional studies. Our study, to the best of our knowledge, is the first longitudinal study examining the associations between CSF tau levels and AD-related demographics, genetic characteristics, and clinical cognitive status. We specifically focus on the impact of baseline age, gender, *ApoE*4 status, baseline diagnosis, and Mini-Mental State Examination (MMSE). Furthermore, we model individual-specific high-dimensional unobserved latent random effects using a generative deep neural network ODE system, taking into account the complex pathological mechanism behind the observed space.

3.5.1 Data

Our study included 2430 participants, aged 50.4 to 91.4 years, from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (adni.loni.usc.edu). To ensure that all variables are on the same scale, we used data from all participants for standardization. We standardized the observed CSF tau and MMSE values using the mean and standard deviation of overall baseline CSF tau and MMSE values respectively. Similarly, we standardized the participants' ages using their mean and standard deviation. Participants with missing baseline measurements were excluded from further data analyses. We included a total of 1215 participants, comprising 367 cognitive normal (CN) controls, 619 individuals with mild cognitive impairment (MCI), and 229 AD patients. The MCI group includes participants with early mild cognitive impairment (EMCI) and late mild cognitive impairment (LMCI). The CN group includes participants with cognitive normal (CN) and stable mild cognitive impairment (SMCI). We coded the three diagnostic categories as 0, 1, and 2, corresponding to increasing disease severity from NC to AD. The sample includes 564 *ApoE*4 car-

riers and 651 non-carriers, with 543 females and 672 males. All participants were required to have baseline CSF tau measurements and at least one follow-up measurement. The median number of repeated CSF tau measurements was 2.00. As per the ADNI study protocol, CSF biomarkers were collected at baseline visits and month 12 or every two years (https://adni.loni.usc.edu/data-samples/data-types/biospecimen-data/). Of the 2298 valid CSF tau measurements for all subjects, the majority (1215) were taken at the baseline visit.

3.5.2 Application of proposed model and results

The analysis was carried out using noninformative uniform priors for the parameters $\boldsymbol{\beta}$ and σ^2. The initial value of the latent process was assumed to follow a prior distribution $\mathcal{N}(0, 1)$. We used the Gibbs sampler algorithm, running it for 500 iterations. After a burn-in period of 200 iterations, we extracted every 10th sample from the posterior distribution. All the programs were written in Python and run on Indiana University's high-throughput computing cluster.

The primary objective of this longitudinal study was to investigate the impact of AD-related factors on CSF tau levels, while controlling unobservable latent dynamic mechanisms. Based on the results in Table (3.4), baseline age, *ApoE*4 status, and baseline clinical diagnosis significantly and positively influence CSF tau levels at a $\alpha = 5\%$ significance level, as their 95% confidence intervals do not include zero. Conversely, gender has a significant negative impact on tau levels, while the MMSE score does not significantly influence tau levels over time. When controlling for unobserved latent dynamics, for participants of the same gender, *ApoE*4 genotype, and baseline diagnosis, a one-unit increase in age corresponds to a 0.11 increase in the standard-

ized CSF tau value. Controlling for baseline age, gender, and baseline diagnosis, *ApoE*4 carriers have an average CSF tau value that is 0.39 higher than non-carriers. In contrast, for participants of the same *ApoE*4 status and gender at the same age, a change in baseline clinical status from normal cognitive to AD, corresponds to a 0.14 increase in the standardized CSF tau unit. Compared to females, males have a 0.37 lower CSF tau value when other covariates are held constant. Additionally, a one-unit decrease in the MMSE score corresponds to a 0.15 increase in the standardized CSF tau value when other factors are controlled.

Given the sparsity of repeated CSF specimen measurements, most CSF tau studies are cross-sectional. These studies often overlook temporal influences and underlying unobserved variables due to the limited use of repeated measurement information, potentially leading to a vulnerable conclusion. In contrast, our longitudinal study examined the effects of baseline age, gender, *ApoE*4 status, baseline diagnosis status, and MMSE score on CSF tau levels, while controlling for other complex mechanisms as unobservable latent random effects. We found that the CSF tau levels increased with the participant's baseline age. We also observed a positive effect on baseline clinical cognitive status and a negative effect on longitudinal MMSE scores. As the severity of the baseline AD diagnosis increased, so did the CSF tau level. Furthermore, we observed gender differences; when controlling for other covariates, female participants had higher mean CSF tau levels than males. *ApoE*4 status was also a significant factor, with *ApoE*4 carriers showing increased CSF tau levels.

Table 3.4: Summary of posterior fixed effect estimates

	Mean	Std	2.5%	Median	97.5%
Baseline Age	0.11	0.01	0.09	0.11	0.14
*ApoE*4 (ref: non-carrier)	0.39	0.02	0.36	0.39	0.42
Baseline Diagnosis (ref: CN)	0.14	0.02	0.11	0.15	0.17
Gender (ref: female)	-0.37	0.03	-0.42	-0.36	-0.32
MMSE	-0.15	0.01	-0.16	-0.15	-0.13

3.6 Conclusion

In this paper, we proposed a novel latent ODE model with Bayesian inference in longitudinal data analysis. This model addresses two important gaps in mixed-effects models: the handling of unobservable latent processes and the adjustment of inference for fixed effects through the control of non-parametrically modeled random effects. Furthermore, our model relaxes the assumption of linearity of the mean function over time.

Moreover, the approach we proposed benefits from all the advantages of Bayesian methods. Specifically, it allows for posterior inference on all fixed effects of interest and random error variance. Uncertainties can be easily characterized by summaries of the posterior distributions. Our methods accommodate irregularly timed observations. Even in the presence of skipping observations, the proposed inference procedures remain valid.

The proposed approach performs well in various simulation scenarios, including i.i.d observations, and temporally correlated time-dependent observations with various stochastic latent processes. Model performance, in terms of the empirical bias, empirical standard deviation, and the average posterior standard deviation, can be enhanced by gathering more useful information. This includes increasing the sample size, the number of repeated measurements, and incorporating continuous covariates. In our application study, we used our model to analyze the impact of AD-related demographic, genetic, and clinical characteristics on a crucial biomarker, CSF tau longitudinally. By well-controlling unobservable latent mechanisms, we identified significant positive influences of age, *ApoE*4 status, and baseline diagnosis status. Our model also captured significant gender differences.

Our current model assumes that the outcomes are continuous and normally distributed. Future research could extend the proposed model to handle outcomes from a variety of distributions. We also assume that the latent process is continuous and differentiable. However, in practice, the latent process may be discrete. Future work could extend the proposed model to accommodate non-differentiable latent processes.

Chapter 4

Bayesian Generalized Random Effects Models

4.1 Introduction

A wide range of longitudinal outcomes are being collected in an increasing number of research areas and studies. For instance, the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, which has been ongoing for around a decade, has gathered repeated measurements from cognitive tests, genetic data, biomarker expression counts, disease onset status, and more. Promising statistical models for this extensive data collection can be crucial in aiding researchers to identify key impact characteristics associated with Alzheimer-related outcomes, which could include longitudinal binary and count outcomes. As the types of collected data diversified, generalized linear models (GLMs) (McCullagh, 2019) were proposed. These models, which are based on likelihood approaches to regression analysis are typically used for a variety of outcome measures when the outcomes are independent. However, the assumption of independence is often not reasonable in many longitudinal studies.

For continuous outcomes that follow a normal distribution, Laird and Ware (1982) introduced linear mixed models (LMMs) for longitudinal data analysis. In this model, it is assumed that each individual has subject-specific regression coefficients, known as random effects, which are distributed around the mean regression coefficients for the population, referred to as the fixed effects. To generalize the LMMs to accommodate various distributions in the exponential family, generalized linear mixed models

(GLMMs) were developed. These models incorporate random terms into the linear predictor of GLMs and extend the capabilities to handle various types of outcomes in linear mixed models (LMMs) (Diggle, 2002; McCulloch et al., 2001; Molenberghs and Verbeke, 2005; Verbeke et al., 1997). Moreover, by incorporating random effects, GLMMs can accommodate the overdispersion often observed among Binomial (Williams, 1982) or Poisson (Breslow, 1984) distributions. However, both LMMs and GLMMs employ parametric random effects to represent covariate effects and model within-subject correlation.

To relax the specific parametric baseline mean function of the response variable over time, Moyeed and Diggle (1994); Zeger and Diggle (1994); Zhang et al. (1998) introduced semiparametric mixed models (SPMMs). These models extend linear mixed models (LMMs) by modeling the time effect using a nonparametric function. To handle a more general covariance structure over time, a stochastic process component was incorporated into mixed models. In existing semiparametric mixed models for longitudinal data analysis, the baseline function is typically modeled as an unspecified smooth function of t (He et al., 2002; Moyeed and Diggle, 1994; Zeger and Diggle, 1994) or an arbitrary function of time (MacKay et al., 1998). This model assumption facilitates parameter estimation using frequentist approaches. However, when the outcomes are in the form of proportions or counts, a full maximum likelihood analysis based on their joint marginal distribution requires numerical integration techniques for the calculation of the log-likelihood, score equation, and information matrix. While this method has been successfully applied to relatively simple problems involving binomial (Crouch and Spiegelman, 1990) and Poisson (Hinde, 1982) distri-

butions with a high degree of independence among the observations, it has proved to be intractable for more complicated problems involving high-dimensional integrals.

To avoid the need for numerical integration and to provide a more flexible model that incorporates prior knowledge,Zeger and Karim (1991) proposed to model GLMMs using a Bayesian approach and fitting it using a Gibbs sampler. Bush and MacEachern (1996) introduced a semi-parametric Bayesian version of LMMs, in which the normal assumption on the random effects is relaxed. Kleinman and Ibrahim (1998) applied non-parametric Bayesian techniques to GLMMs, demonstrating that the GLMM can be freed from the parametric assumption for the random effects. A notable feature of the Bayesian approach is its flexibility in evaluating the uncertainty in the estimated model parameters.

In this paper, we extend the GLMMs by allowing the mean baseline function to have a non-parametric prior distribution. We generalize the SPMMs and replace the typical Gaussian Process assumption in frequentist methods with a flexible stochastic process prior governed by ordinary differential equations (ODEs). Our proposed model is based on the Bayesian approach, and we perform computations for the model estimation using the Gibbs sampler algorithm.

In the following sections, we describe our statistical models and the proposed estimation procedure. We then conduct comprehensive simulation studies to evaluate the performance of our proposed method in various scenarios. Furthermore, we illustrate the practical application of our approach using data related to Alzheimer's disease. We conclude the paper with a discussion of our methods and potential avenues for future research.

4.2 Statistical methods

4.2.1 Longitudinal observed data

Consider we have m *i.i.d.* subjects followed longitudinally, with time-dependent outcomes $y_i(t)$ and p-dimensional covariates $\boldsymbol{x}_i(t) = \{x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t)\}$, where $i = 1, \ldots, m$ and $t \geq 0$. For each subject i, we observe longitudinal outcomes and covariates at random follow-up times, denoted as t_{ij}, where $j = 1, \ldots, n_i$ and n_i is the total number of observations for subject i. The observed data at specific time points can be represented as $\boldsymbol{y}_i = \{y_{i1}, y_{i2}, \ldots, y_{in_i}\}^T$ and $\boldsymbol{X}_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{in_i}\}^T \in \mathbb{R}^{n_i \times p}$ where $y_{ij} \equiv y_i(t_{ij})$ and $\boldsymbol{x}_{ij} \equiv \boldsymbol{x}_i(t_{ij})$, for $j = 1, 2, \ldots, n_i$.

4.2.2 Statistical models

Suppose the individual mean outcome trajectory, denoted as $\mu_i(t) = E(y_i(t)|\boldsymbol{x}_i(t), \boldsymbol{u}_i(t))$, is linked to a non-linear predictor $\eta_i(t)$ through a link function. This non-linear predictor consists of a nonparametric baseline function, fixed effects, and random effects, as follows:

$$g(\mu_i(t)) = \eta_i(t) = \gamma(t) + \boldsymbol{x}_i(t)\boldsymbol{\beta} + \boldsymbol{u}_i(t)\boldsymbol{b}_i \tag{4.1}$$

where $i = 1, \ldots, m$, $g(\cdot)$ is a monotonic differentiable function, often referred to as the link function, $\mu_i(t)$ is the mean of the conditional outcome for individual i, $\gamma(t)$ is a nonparametric baseline function that captures the time effects. $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector of regression coefficients, commonly referred to as fixed effects. $\boldsymbol{u}_i(t) = \{u_{i1}(t), u_{i2}(t), \ldots, u_{iq}(t)\}$ is q-dimensional covariates for the random effects $\boldsymbol{b}_i$.

At specific observation time point t_{ij}, suppose the outcome variable y_{ij} is a member of the exponential family, denoted as $y_{ij}|\vartheta_{ij}, \varphi \sim p_{\boldsymbol{\theta}}(\cdot)$, where $p_{\boldsymbol{\theta}}(\cdot)$ is a member of the exponential family, that is

$$p_{\boldsymbol{\theta}}(y_{ij}|\vartheta_{ij}, \varphi) = exp\left[\frac{y_{ij}\vartheta_{ij} - b(\vartheta_{ij})}{a(\varphi)} + c(y_{ij}, \varphi)\right]$$

where ϑ_{ij} is the canonical parameter and φ is a dispersion parameter, where

$$\mu_{ij} = E[y_{ij}|\vartheta_{ij}, \varphi] = b'(\vartheta)$$

therefore, the canonical parameter ϑ_{ij} is linked to the covariates through the link function, as follows:

$$g(\mu_{ij}) = g(b'(\vartheta)) = \eta_{ij} = \gamma(t_{ij}) + \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{u}_{ij}\boldsymbol{b}_i$$

where $\gamma(t_{ij})$ is the time effect at specific time t_{ij}, $\boldsymbol{x}_{ij}$ is a $1 \times p$ vector of covariates, and $\boldsymbol{u}_{ij}$ is a $1 \times q$ vector of random effects covariates for individual i. Therefore, we have:

$$p(y_{ij}|\vartheta, \varphi) \equiv p(y_{ij}|\gamma(t_{ij}), \boldsymbol{\beta}, \boldsymbol{b}_i, \varphi)$$

In semiparametric mixed models (SPMMs), researchers have proposed to model the nonparametric baseline function of the response variable over time, denoted by $\gamma(t)$, as an unspecified or an arbitrary smooth function of t using frequentist methods (MacKay et al., 1998; Moyeed and Diggle, 1994; Zeger and Diggle, 1994; Zhang et al., 1998). However, previous research indicated that fitting this model using fre-

quentist approaches is often constrained by the need for multi-dimensional numerical integrations in most cases. When the baseline function is parametrically specified, as in Generalized Linear Mixed Models (GLMMs), Bayesian approaches have been proposed for modeling (Kleinman and Ibrahim, 1998; Zeger and Karim, 1991).

In this paper, we propose to use Bayesian methods to extend GLMMs and SPMMs into Generalized Semiparametric Mixed Models (GSPMMs). We assume that a stochastic process denoted as $f_{\boldsymbol{\zeta}}(\boldsymbol{z}(t))$ serves as the prior process for the nonparametric baseline function $\gamma(t)$,

$$\gamma(t) \sim f_{\boldsymbol{\zeta}}(\boldsymbol{z}(t)),$$

where $\boldsymbol{z}(t)$ is a d-dimensional stochastic process that is governed by an ODE system. As per the definition of ODEs, given ordinary differential equations, $\boldsymbol{z}(t)$ is determined by its initial condition $\boldsymbol{z}^{(0)}$, which is independent of other parameters. The function $f_{\boldsymbol{\zeta}}(\cdot)$ is an unknown function that maps $\boldsymbol{z}(t)$ from $\mathbb{R}^d$ into a space with the same dimension as the outcome. This can be expressed as $f_{\boldsymbol{\zeta}}(\boldsymbol{z}(t)) = h_{\boldsymbol{\theta}}(\boldsymbol{z}^{(0)}, t)$, where $\boldsymbol{\theta} = (\boldsymbol{\zeta}, \boldsymbol{\omega})$ are parameters from the unknown function $f_{\boldsymbol{\zeta}}(\cdot)$, parameterized by $\boldsymbol{\zeta}$ and the unknown ODE process of $\boldsymbol{z}(t)$, parameterized by $\boldsymbol{\omega}$. We nonparametrically approximate both of these by using neural networks.

We further assigned normal priors to $\boldsymbol{z}^{(0)}$, $\boldsymbol{\beta}$, and $\boldsymbol{b}_i|\sigma$. Moreover, $\boldsymbol{b}_i|\sigma \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, where the variance-covariance matrix of random effects is dependent on parameters σ^2. Various noninformative prior distributions for σ^2 have been suggested in Bayesian literature, including proper distribution like InverseGamma (Gelman, 2006; Spiegel-

halter et al., 1996, 2004). The priors of our model are as follows:

$$\gamma(t) \sim h_{\boldsymbol{\theta}}(\boldsymbol{z}^{(0)}, t), \quad \boldsymbol{z}^{(0)} \sim \prod_{l=1}^{d} N(0, b_{0l})$$

$$\boldsymbol{\beta} \sim \prod_{l=1}^{p} N(0, a_{0l})$$

$$\boldsymbol{b}_i|\sigma \overset{i.i.d.}{\sim} N(\mathbf{0}, \sigma^2 \boldsymbol{I}_{q\times q}); \quad \sigma \sim p(\sigma; \boldsymbol{c}_0)$$

$$\varphi \sim p(\varphi; \boldsymbol{d}_0)$$

where $a_{01}, a_{02}, \ldots, a_{0p}, b_{01}, b_{02}, \ldots, b_{0d}, \boldsymbol{c}_0, \boldsymbol{d}_0$ are hyperparameters. The functions $p(\sigma; \boldsymbol{c}_0)$ and $p(\varphi; \boldsymbol{d}_0)$ represent the prior density function for σ and φ respectively. The dimensions of $\boldsymbol{c}_0$ and $\boldsymbol{d}_0$ are determined by the number of parameters used to characterize the distributions. When the dispersion parameter φ of the exponential family is a constant, there is no need to assign priors to it. In our proposed model, we make an assumption of a priori independence among the parameter components. Under the hierarchical model, the data points y_{ij}, where $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ are conditionally independently distributed within each of the i clusters. The total count of observations is represented by $N = \sum_{i=1}^{m} n_i$. The joint posterior, given these conditions, is subsequently derived as follows:

$$\pi(\varphi, \boldsymbol{\beta}, \boldsymbol{b}, \sigma, \gamma(t)|\boldsymbol{y}) \propto \pi(\varphi)\pi(\boldsymbol{\beta}) \prod_{i=1}^{m} \pi(b_i|\sigma)\pi(\sigma) \prod_{i=1}^{m} \prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{\beta}, b_i, \gamma(t_{ij})) \times \pi(\gamma(t_{ij}))$$

Posterior Inferences

To approximate the joint posterior, we propose to use the Gibbs sampling algorithm. This method requires the full conditional distribution for each parameter. These are

represented as follows:

$$p(\boldsymbol{\beta}|\text{else}) \quad p(\gamma(t)|\text{else}) \quad p(\sigma|\text{else}) \quad p(\boldsymbol{b}_i|\text{else}) \quad p(\varphi|\text{else})$$

where "else" represents all other parameters and data except the parameter in question.

Posterior Inference of $\boldsymbol{\beta}, \boldsymbol{b}_i, \sigma, \varphi$

Given $\gamma(t)$, we use standard Markov chain Monte Carlo (MCMC) methods with NUTS steps to sample the full conditional distribution of $\boldsymbol{\beta}, \boldsymbol{b}_i, \sigma, \varphi$. The full conditional distribution of these parameters is as follows:

$$\begin{aligned}
p(\boldsymbol{\beta}|\text{else}) &\propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{m} \prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{\beta}, b_i, \gamma(t_{ij})) \\
p(\boldsymbol{b}_i|\text{else}) &\propto \pi(\boldsymbol{b}_i|\sigma) \prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{\beta}, b_i, \gamma(t_{ij}), \sigma) \\
p(\sigma|\text{else}) &\propto \pi(\sigma) \prod_{i=1}^{m} p(b_i|\sigma) \\
p(\varphi|\text{else}) &\propto \pi(\varphi) \prod_{i=1}^{m} \prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{\beta}, b_i, \gamma(t_{ij}))
\end{aligned}$$

Posterior Inference of $\gamma(t)$

Given all other parameters and the deterministic nature of the ODE process, where the state of the system at any future time is only determined by its initial conditions when the ODE functions are given, the full conditional distribution of $\gamma(t)$ is

determined by the full conditional distribution of $\boldsymbol{z}^{(0)}$ which is as follows:

$$p(\boldsymbol{z}^{(0)}|\text{else}) \propto \pi(\boldsymbol{z}^{(0)}) \prod_{i=1}^{m} \prod_{j=1}^{n_i} p(y_{ij}|\varphi, \boldsymbol{\beta}, b_i, \gamma(t_{ij}))$$

Given that $\gamma(t)$ is the process we need to make an inference, and it involves complex ODEs with multiple neural networks, the posterior distribution of $p(\boldsymbol{z}^{(0)}|else)$ is not as straightforward as in traditional Bayesian inference situations. In the current paper, we proposed using Variational Inference (VI) to approximate the posterior distribution of $\boldsymbol{z}^{(0)}$. By applying the principles of VI, we aim to minimize the Kullback-Leibler (KL) divergence between the approximated posterior distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)} \mid \boldsymbol{y}, \boldsymbol{\beta}, b_i, \sigma, \varphi)$ and the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}^{(0)} \mid \boldsymbol{y}, \boldsymbol{\beta}, b_i, \sigma, \varphi)$ with respect to $\boldsymbol{\theta} = (\boldsymbol{\zeta}, \boldsymbol{\omega})$ and $\boldsymbol{\phi}$. The KL divergence is defined as follows:

$$KL(q_{\boldsymbol{\phi}}, p_{\boldsymbol{\theta}}|\boldsymbol{y}, \boldsymbol{\beta}, b_i, \sigma, \varphi) = -\int q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)}|\boldsymbol{y}, \boldsymbol{\beta}, b_i, \sigma, \varphi) \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{z}^{(0)}|\boldsymbol{y}, \boldsymbol{\beta}, b_i, \sigma, \varphi)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_0|\boldsymbol{y}, \varphi, \boldsymbol{\beta}, b_i, \sigma)} d\boldsymbol{z}^{(0)}$$

Due to the intractability of $p_{\boldsymbol{\theta}}(\boldsymbol{z}^{(0)}|\boldsymbol{y}, \varphi, \boldsymbol{\beta}, b_i, \sigma)$, the KL divergence is also intractable. However, we can computationally evaluate the variational lower bound, also known as the evidence lower bound (ELBO)(Kingma et al., 2019; Neal and Hinton, 1998) as follows:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}, \varphi, \boldsymbol{\beta}, b_i, \sigma) &= \int q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)}|\boldsymbol{y}, \varphi, \boldsymbol{\beta}, b_i, \sigma) \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{z}^{(0)}|\varphi, \boldsymbol{\beta}, b_i, \sigma)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)}|\boldsymbol{y}, \varphi, \boldsymbol{\beta}, b_i, \sigma)} d\boldsymbol{z}^{(0)} \\
&= E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)}|\boldsymbol{y}, \varphi, \boldsymbol{\beta}, b_i, \sigma)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z}^{(0)}, \varphi, \boldsymbol{\beta}, b_i, \sigma) p(\boldsymbol{z}^{(0)})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)}|\boldsymbol{y}, \varphi, \boldsymbol{\beta}, b_i, \sigma)} \right]
\end{aligned}$$

By maximizing the ELBO with respect to $(\boldsymbol{\theta}, \boldsymbol{\phi})$, we can approximate the maximization of the marginal likelihood $p(\boldsymbol{y}|\varphi, \boldsymbol{\beta}, b_i, \sigma)$. At the same time, this process also minimizes the KL divergence between $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)}|\varphi, \boldsymbol{y}, \boldsymbol{\beta}, b_i, \sigma)$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z}^{(0)}|\boldsymbol{y}, \varphi, \boldsymbol{\beta}, b_i, \sigma)$. The KL divergence becomes zero if and only if when $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)}|\varphi, \boldsymbol{y}, \boldsymbol{\beta}, b_i, \sigma)$ is equal to the true posterior distribution, otherwise, it remains non-negative. To maximize the ELBO, we jointly optimize for both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ using techniques such as stochastic gradient descent (SGD) in conjunction with the re-parameterization trick.

In practice, the Variational Inference algorithm allows for the selection of a family of tractable distributions to serve as the approximated posterior distribution. In this paper, we make the assumption that $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)}|\boldsymbol{y}, \varphi, \boldsymbol{\beta}, b_i, \sigma) \sim \mathcal{N}(\boldsymbol{\nu}_{\boldsymbol{\phi}}(\boldsymbol{y}), diag(\boldsymbol{\tau}^2_{\boldsymbol{\phi}}(\boldsymbol{y})))$, where $\boldsymbol{\nu}_{\boldsymbol{\phi}}(\boldsymbol{y})$ and $\boldsymbol{\tau}^2_{\boldsymbol{\phi}}(\boldsymbol{y})$ are d-dimensional parameters approximated by a Recurrent Neural Network (RNN) parameterized by $\boldsymbol{\phi}$. By taking advantage of the RNN's capacity to handle sequential data and its adaptable input/output dimensions, our model provides the ability to approximate $\boldsymbol{z}^{(0)}$ as a population parameter and $\gamma(t)$ as a population baseline function. Additionally, it also offers the flexibility to model $\gamma_i(t)$ as individual or cluster baseline functions, which can be adjusted based on the research purpose. In this paper, our primary focus is on the demonstration of the scenario of approximating $\boldsymbol{z}^{(0)}$ as a population parameter and $\gamma(t)$ as a population baseline function.

4.2.3 Computation

Estimation of hyperparameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$

In our proposed model, we utilize high-dimensional parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ to characterize neural networks. Although we consider them as nuisance parameters we do not

make inferences on them, they are crucial in approximating inferences for parameters of our interest. In this study, we estimate $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ by maximizing the ELBO objective function. Given the independent assumption that y_{ij} is independent conditional on $(\boldsymbol{\beta}, \gamma(t), \boldsymbol{b}_i, \sigma, \varphi)$, and given $\boldsymbol{\beta}$, $\boldsymbol{b}_i$, σ, φ with the dataset, we sample $\boldsymbol{z}^{(0)}$ from its approximated posterior distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)}|\boldsymbol{y}, \boldsymbol{\beta}, b_i, \sigma, \varphi)$. Subsequently, we approximate the ELBO objective by Monte Carlo estimate as the average of individual ELBOs. The ELBO is calculated as follows:

$$
\begin{aligned}
\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{y} \mid \boldsymbol{\beta}, b_i, \sigma, \varphi) &\approx \frac{1}{N+L} \sum_{l=1}^{L} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(y_{ij} \mid \boldsymbol{\beta}, b_i, \sigma, \varphi) \\
&= \frac{1}{N+L} \sum_{l=1}^{L} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[\log \left(\frac{p_{\boldsymbol{\theta}}(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{\beta}, b_i, \sigma, \gamma^{(l)}(t_{ij}), \varphi) p(\boldsymbol{z}^{(0)})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)} \mid y_{ij}, \boldsymbol{\beta}, b_i, \sigma, \varphi)} \right) \right] \\
&= \frac{1}{N+L} \sum_{l=1}^{L} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[\frac{y_{ij}\vartheta_{ij} - b(\vartheta_{ij})}{a(\varphi)} + c(y_{ij}, \varphi) \right. \\
&\qquad \left. + \log \left(\frac{p(\boldsymbol{z}^{(0)})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)} \mid y_{ij}, \boldsymbol{\beta}, b_i, \sigma, \varphi)} \right) \right]
\end{aligned}
$$

where $N = \sum_{i=1}^{m} n_i$, N is the total number of observations. L denotes the number of samples for each data point drawn from the surrogate posterior distribution. The function $p_{\boldsymbol{\theta}}(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\beta}, b_i, \sigma, \gamma(t_{ij}), \varphi)$ represents the density function of outcomes, which is modeled as an exponential family distribution. ϑ_{ij} is the canonical parameter, and $\mu_{ij} = E[y_{ij}|\gamma(t), \boldsymbol{\beta}, \boldsymbol{b}_i, \varphi] = b'(\theta_{ij})$. $p(\boldsymbol{z}^{(0)})$ is the prior density of $\boldsymbol{z}^{(0)}$, assumed to be independent of other parameters. Finally, $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)}|\boldsymbol{y}, \boldsymbol{\beta}, b_i, \sigma, \varphi)$ is a known distribution which is a Gaussian distribution in the current paper.

Estimating procedure

The Gibbs sampler is used to establish the joint posterior distribution and the inference of each parameter. With the hyperparameters $\boldsymbol{\phi}$ estimates, we draw samples of $\boldsymbol{z}^{(0)}$ from its approximate posterior distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(0)} \mid \boldsymbol{y}, \boldsymbol{\beta}, b_i, \sigma, \varphi)$. Once $\boldsymbol{z}^{(0)}$ and the optimal estimate of hyperparameter $\boldsymbol{\theta}$ are known, $\gamma(t)$ can be estimated by $h_{\boldsymbol{\theta}}(\boldsymbol{z}^{(0)}, t)$, in accordance with the deterministic property of ODEs. Given $\gamma(t)$ estimate, we perform Gibbs sampling of the parameters $\boldsymbol{\beta}$, b_i, σ, and φ through the MCMC estimation of each of their full conditional posterior distributions. After a pre-defined burn-in period, we compute the posterior mean of the p-th fixed effect, denoted as $\hat{\beta}_p$, by selecting every r-th sample from the sampling procedure, where r is a pre-determined arbitrary value.

4.2.4 Specific models

In this section, we provide specific priors and link functions for the longitudinal count and longitudinal binary outcomes to demonstrate our proposed model for different types of longitudinal data.

Models for longitudinal counts

For identical independent counts data, the most basic regression model is the linear Poisson regression model without random effects (Cameron and Trivedi, 2013; Scott Long, 1997; Winkelmann, 2008). A significant limitation of this model is its assumption of equal dispersion. However, in practice, counts data often exhibit overdispersion due to two factors: contagion, which refers to the dependence between event occurrences, and heterogeneity, which represents the difference between individuals

(Winkelmann, 2008). This overdispersion results in a variance that is often much larger than the mean. To address this overdispersion, a modification to the linear Poisson regression model has been proposed (Fahrmeir and Osuna Echavarría, 2006). This modification incorporates an individuals-specific nonnegative random-effect term to model individual heterogeneity and extends the parametric linear predictor to a semiparametric structured additive predictor. In this paper, we propose a mixed-effect Poisson regression model with a nonparametric baseline function to handle longitudinal counts and take into account the dispersion of counts data. The model is defined as follows:

$$y_i(t)|\mu_i(t) \sim Pois(\mu_i(t)), \quad \mu_i(t) = exp(\gamma(t) + \boldsymbol{x}_i(t)\boldsymbol{\beta} + \boldsymbol{u}_i(t)\boldsymbol{b}_i)$$

where $\boldsymbol{b}_i$ is a Gaussian distributed random effect vector with mean $\mathbf{0}$.

When the random effects $\boldsymbol{b}_i$ are given along with $\boldsymbol{x}_i(t)$ and $\gamma(t)$, the conditional density of outcomes $y_i(t)$ follows a Poisson distribution with equal dispersion:

$$p_{\boldsymbol{\theta}}(y_i(t) \mid \boldsymbol{x}_i(t), \boldsymbol{\beta}, b_i, \sigma, \gamma(t)) = \exp[y_i(t)\log(\mu_i(t)) - \mu_i(t) - \log(y_i(t)!)]$$

using a log link function, we have $g(\mu_i(t)) = \log(\mu_i(t)) = \gamma(t) + \boldsymbol{x_i}(t)\boldsymbol{\beta} + \boldsymbol{u}_i(t)\boldsymbol{b}_i$, where $\mu_i(t) = E[y_i(t) \mid \boldsymbol{x}_i(t), \boldsymbol{\beta}, \boldsymbol{b}_i, \sigma, \gamma(t)] = Var[y_i(t) \mid \boldsymbol{x}_i(t), \boldsymbol{\beta}, \boldsymbol{b}_i, \sigma, \gamma(t)]$. However, the marginal density of $y_i(t)$ is not the case. By applying the law of total expectation and the law of total variance, it can be shown that $Var[y_i(t)|\gamma(t), \boldsymbol{x}_i(t)] \geq E[y_i(t)|\gamma(t), \boldsymbol{x}_i(t)]$ as follows:

$$E[y_i(t)|\gamma(t), \boldsymbol{x}_i(t)] = exp(\gamma(t) + \boldsymbol{x}_i(t)\boldsymbol{\beta}) \cdot E[exp(\boldsymbol{u}_i(t)\boldsymbol{b}_i)]$$

$$Var[y_i(t)|\gamma(t), \boldsymbol{x}_i(t)] = E[y_i(t)|\gamma(t), \boldsymbol{x}_i(t)] + \frac{Var[exp(\boldsymbol{u}_i(t)\boldsymbol{b}_i)] \cdot E^2[y_i(t)|\gamma(t), \boldsymbol{x}_i(t)]}{E^2[exp(\boldsymbol{u}_i(t)\boldsymbol{b}_i)]}$$

Furthermore, in the model we propose, the covariance between repeated measurements for the same individual is calculated as follows:

$$\begin{aligned}
Cov(y_i(t_{ij}), y_i(t_{ik})) &= E[y_i(t_{ij}) \cdot y_i(t_{ik})] - E[y_i(t_{ij})|\gamma(t_{ij}), \boldsymbol{x}_i(t_{ij})] \cdot E[y_i(t_{ik})|\gamma(t_{ik}), \boldsymbol{x}_i(t_{ik})] \\
&= E\left[E[y_i(t_{ij}) \cdot y_i(t_{ik})|b_i]\right] - E[y_i(t_{ij})|\gamma(t_{ij}), \boldsymbol{x}_i(t_{ij})] \cdot E[y_i(t_{ik})|\gamma(t_{ik}), \boldsymbol{x}_i(t_{ik})] \\
&= E\left[E[y_i(t_{ij})|b_i] \cdot E[y_i(t_{ik})|b_i]\right] \\
&\quad - E[y_i(t_{ij})|\gamma(t_{ij}), \boldsymbol{x}_i(t_{ij})] \cdot E[y_i(t_{ik})|\gamma(t_{ik}), \boldsymbol{x}_i(t_{ik})] \\
&= \exp\{\gamma(t_{ij}) + \gamma(t_{ik}) + \boldsymbol{x}_i(t_{ij})\boldsymbol{\beta} + \boldsymbol{x}_i(t_{ik})\boldsymbol{\beta}\} \\
&\quad \cdot \left(E\left[\exp(\boldsymbol{u}_i(t_{ij}) + \boldsymbol{u}_i(t_{ik}))\boldsymbol{b}_i\right] - E\left[\exp(\boldsymbol{u}_i(t_{ij})\boldsymbol{b}_i)\right] E\left[\exp(\boldsymbol{u}_i(t_{ik})\boldsymbol{b}_i)\right]\right)
\end{aligned}$$

The Poisson distribution has a constant dispersion parameter, φ. As such, there's no need for the prior distribution for this parameter. Our proposed model is then specified as follows:

$$\begin{aligned}
&y_i(t)|\boldsymbol{\beta}, \gamma(t), \boldsymbol{b}_i, \sigma \overset{ind.}{\sim} \text{Pois}\left(\exp\{\gamma(t) + \boldsymbol{x_i}(t)\boldsymbol{\beta} + \boldsymbol{u}_i(t)\boldsymbol{b}_i\}\right) \\
&\boldsymbol{\beta} \sim \prod_{l=1}^{p} N(0, a_{0l}) \\
&\gamma(t) \sim h_{\boldsymbol{\theta}}(\boldsymbol{z}^{(0)}, t), \quad \boldsymbol{z}^{(0)} \sim \prod_{l=1}^{d} N(0, b_{0l}) \\
&\boldsymbol{b}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \sigma^{-1}\boldsymbol{I}), \quad \sigma \sim \text{Gamma}(c_0, 1/d_0)
\end{aligned}$$

where $a_{01}, \ldots, a_{0p}, b_{01}, \ldots, b_{0d}, c_0$ are hyperparameters. In our simulation studies, we set $a_{01} = a_{02} = \ldots = a_{0p} = 100, b_{01} = b_{02} = \ldots = b_{0d} = 1$, $c_0 = d_0 = 0.01$.

Model for binary longitudinal classification

Binary outcomes and corresponding longitudinal classification problems are often encountered in research studies. In terms of prediction, these outcomes are crucial as they assist researchers in predicting disease onset, recovery status, or the probability of certain changes over time. The flexibility provided by the assumption of a stochastic process prior to the nonparametric baseline function enhances such predictions. In terms of inference, making inferences about time-dependent covariates for longitudinal binary outcomes is essential for scientific research as well. It helps in understanding the effectiveness of interventions, treatments, and risk factors over time. To model this type of data longitudinally, we assume that the binary responses follow a Bernoulli distribution:

$$p_{\boldsymbol{\theta}}(y_i(t) \mid \boldsymbol{x}_i(t), \boldsymbol{\beta}, b_i, \sigma, \gamma(t)) = \exp\left[y_i(t) \log \frac{p_i(t)}{1-p_i(t)} + \log(1-p_i(t))\right]$$

By given a *logit* link function, we have $g(\mu_i(t)) = \log(\frac{p_i(t)}{1-p_i(t)}) = \gamma(t) + \boldsymbol{x_i}(t)\boldsymbol{\beta} + \boldsymbol{u}_i(t)\boldsymbol{b}_i$. Priors for each parameter are defined as follows:

$$\begin{aligned}
&y_i(t)|\boldsymbol{\beta}, \gamma(t), \boldsymbol{b}_i, \sigma \overset{ind.}{\sim} \text{Bern}\left(\frac{\exp\{\gamma(t) + \boldsymbol{x_i}(t)\boldsymbol{\beta} + \boldsymbol{u}_i(t)\boldsymbol{b}_i\}}{1+\exp\{\gamma(t) + \boldsymbol{x_i}(t)\boldsymbol{\beta} + \boldsymbol{u}_i(t)\boldsymbol{b}_i\}}\right) \\
&\boldsymbol{\beta} \sim \prod_{l=1}^{p} N(0, a_{0l}) \\
&\gamma(t) = h_{\boldsymbol{\theta}}(\boldsymbol{z}^{(0)}, t), \quad \boldsymbol{z}^{(0)} \sim \prod_{l=1}^{d} N(0, b_{0l}) \\
&\boldsymbol{b}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \sigma^{-1}\boldsymbol{I}), \quad \sigma \sim \text{Gamma}(c_0, 1/d_0)
\end{aligned}$$

where $a_{01}, \ldots, a_{0p}, b_{01}, \ldots, b_{0d}, c_0$ are hyperparameters. In the simulation, we set $a_{01} = a_{02} = \ldots = a_{0p} = 100, b_{01} = b_{02} = \ldots = b_{0d} = 1$,$c_0 = d_0 = 0.01$.

4.3 Simulation

In our simulation study, we evaluated the performance of our model for both longitudinal counts and longitudinal binary outcomes. Based on our model assumptions, we constructed the link function's transformed conditional expectation of observed curves, $g(\mu_i(t))$, as a linear function. This function is composed of the baseline function $\gamma(t)$, the random effect $\boldsymbol{b}_i$, and observed covariates $\boldsymbol{x}_i(t)$. Without loss of generality, we evaluate random intercept models with two time-dependent covariates, denoted as $\boldsymbol{x}_i(t) = [x_{i1}(t), x_{i2}(t)]'$ for $i = 1, \ldots, n$. Each covariate is independently drawn from normal distributions $N(0, 1)$ at specific time points t. The follow-up times t are simulated randomly in the study duration $(0, \tau)$. The true coefficient values used in this simulation are $\beta_1 = 0$ and $\beta_2 = 1$. Given $\gamma(t)$, $\boldsymbol{x}_i(t)$, true coefficient values, and random effects drawn from a normal distribution with a mean zero, we generated the longitudinal counts and binary observed data $y_i(t)$ for each subject i at each time point t from the corresponding distributions:

Longitudinal count outcomes are generated from the following distributions:

$$y_i(t)|\boldsymbol{\beta}, \gamma(t), \boldsymbol{b}_i, \sigma \overset{ind.}{\sim} \text{Pois}\left(\exp\{\gamma(t) + \boldsymbol{x_i}(t)\boldsymbol{\beta} + b_i\}\right)$$

Longitudinal Binary outcomes are generated from the following distributions:

$$y_i(t)|\boldsymbol{\beta}, \gamma(t), \boldsymbol{b}_i, \sigma \overset{ind.}{\sim} \text{Bern}\left(\frac{\exp\{\gamma(t) + \boldsymbol{x_i}(t)\boldsymbol{\beta} + b_i\}}{1 + \exp\{\gamma(t) + \boldsymbol{x_i}(t)\boldsymbol{\beta} + b_i\}}\right)$$

4.3.1 Simulation setting 1

In the first simulation setup, the baseline function is determined by a time-dependent process, represented as $\gamma(t) = z_1(t)sin(t) + z_2(t)cos(t)$. Here $z_1(t)$ and $z_2(t)$ are governed by an ODE system with initial values $z_1^{(0)}, z_2^{(0)} = 1$. This setting can be represented as follows:

$$y_i(t) \sim p(y_i(t)|\mu_i(t), \varphi)$$

where $\mu_i(t) = E(y_i(t)|\gamma(t), \boldsymbol{\beta}, b_i, \varphi)$

$$g(\mu_i(t)) = z_1(t)sin(t) + z_2(t)cos(t) + \boldsymbol{x_i}(t)\boldsymbol{\beta} + b_i$$

The ODE system is specifically formulated as:

$$\frac{dz_1(t)}{dt} = z_2(t)$$
$$\frac{dz_2(t)}{dt} = a \times sin(z_1(t))$$

with a = -5.0, $b_i \sim N(0, \sigma^2)$. The true curve of $\gamma(t)$ is displayed in Figure 4.1.

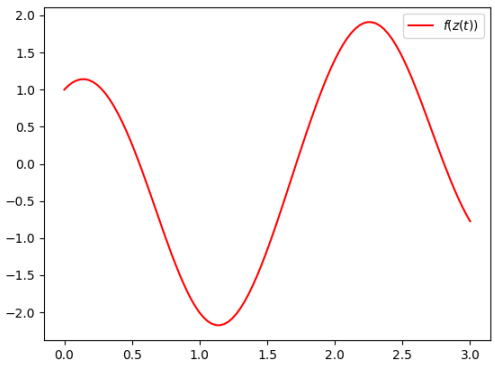


Figure 4.1: The curve of $\gamma(t)$ in simulation setting 1

4.3.2 Simulation setting 2

In the second simulation setting, a linear baseline function is assigned, $\gamma(t) = t$, and the true curve of $\gamma(t)$ is displayed in Figure 4.2. The data is generated as follows

$$g(\mu_i(t)) = t + \boldsymbol{x_i}(t)\boldsymbol{\beta} + b_i,$$

where $\boldsymbol{x}_i(t)$ are time-dependent normal random variables with mean 0 and variance 1. The true coefficient values are $\beta_1 = 0$ and $\beta_2 = 1$, and $b_i \sim N(0, \sigma^2)$.

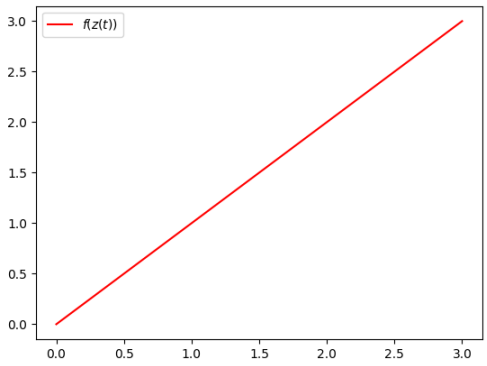


Figure 4.2: The curve of $\gamma(t)$ in simulation setting 2

To assess the effectiveness and robustness of our proposed model, we examined various scenarios by considering all possible combinations of the following choices:

1. The number of subjects: (a) n=100 and (b) n=300;
2. The number of follow-ups: (a) $k = 30$
3. Distributions of x_1 and x_2: (a) Normal and Normal (NN) Distributions
4. The variance of random effect: (a) 0.5 and (b) 1.0

For each simulation study, we conducted $M = 500$ replicates. In each replicate, we generated n subjects, each with k observations. Each individual has follow-up times

that are uniform, denoted as $t_{ij} \sim U(0, \tau)$, where $U(a, b)$ represents the uniform distribution on the interval (a, b). Instead of setting a cutoff loss value during the simulation studies, we trained our model over a fixed number of training epochs and a fixed number of burn-in periods. This approach ensures more comparable model performances across different replications and simulation scenarios.

As the quantity of follow-up observations grows, providing more information to process, we correspondingly increase the number of hidden nodes in our model. The initial values of each parameter act as the starting points for the estimation process. Parameters $\boldsymbol{\beta}$, $\boldsymbol{b}_i$ are initialized with arbitrary values, specifically $\beta_1^{(0)} = \beta_2^{(0)} = 0$ and $\boldsymbol{b}_1 = \ldots = \boldsymbol{b}_m = \mathbf{0}$ where m represents the total number of subjects or clusters. Concurrently, the hyperparameters of neural networks, denoted as $\boldsymbol{\phi}$, $\boldsymbol{\theta} = [\boldsymbol{\zeta}, \boldsymbol{\eta}]$ are randomly initialized prior to training the model.

4.3.3 Evaluation of the performance of the proposed model

We calculate the posterior mean for each parameter $\hat{\beta}_p$ by averaging every 10th sample obtained after the burn-in period. The bias of the estimates for the p-th coefficient is assessed as the expected absolute difference between the average estimate and the true value, calculated as:

$$E\left\{|\hat{\beta}_p - \beta_p|\right\},$$

We use the standard deviation (std) of the estimation for the p-th component to evaluate the uncertainty of the corresponding estimate, which can be computed as:

$$\sqrt{E\left\{\hat{\beta}_p - E(\hat{\beta}_p)\right\}^2},$$

In the current study, the expectations are estimated empirically based on the M replicated simulations. Hence, the empirical bias (Ebias) for the p-th coefficient is evaluated as:

$$Ebias = \frac{1}{M}\sum_{m=1}^{M}|\hat{\beta}_{p,m} - \beta_p| \tag{4.2}$$

The empirical standard deviation (Estd) for the p-th coefficient is evaluated as:

$$Estd = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left\{\hat{\beta}_{p,m} - \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_{p,m}\right\}^2} \tag{4.3}$$

We also calculate the average of the standard deviation and median of the posterior (AP) distribution.

4.3.4 Simulation results

To offer a summary of our model's performance across various simulation scenarios when outcomes are from Poisson and Bernoulli distributions, we present the empirical bias (Ebias), empirical standard deviation (Estd), the average standard deviation of the posterior distribution (APstd) and the average median of the posterior distribution (APmedian) in Tables 4.1 and 4.2.

Similar patterns are observed for both Poisson and Bernoulli distributed outcomes. The results demonstrate a strong alignment between the posterior mean estimates and

the average posterior median, aligning closely with the true values used to generate the data. The empirical bias is close to zero, indicating that the model provides precise estimates of the true parameter values. When other factors are held constant, increasing the number of subjects from 100 to 300 results in a slight decrease in the empirical bias and empirical standard deviation. We also investigated the impact of the variance of the random effect on the model's performance. The results show that when the variance of the random effect increases from 0.5 to 1, the empirical bias and empirical standard deviation increase slightly. However, the model still performs well. Comparing the simulation results of setting 1 with setting 2, there's no significant model performance difference between the two settings. Although a significantly more complex structure is present in a simulation setting 1, the model still performs as well as a simpler structure setting. Overall, our model performs well across all scenarios, demonstrating its effectiveness in estimating the true parameter values for both Poisson and Bernoulli distributed outcomes.

Table 4.1: Simulation performance for Poisson distribution

			$\sigma = 0.5$				$\sigma = 1$			
	Follow-ups	Setting	Ebias	Estd	APstd	APmedian	Ebias	Estd	APstd	APmedian
n=100										
	30	Setting 1	[0.01, 0.03]	[0.02, 0.06]	[0.03, 0.10]	[0.00, 1.01]	[0.02, 0.04]	[0.02, 0.04]	[0.04, 0.13]	[0.00, 1.02]
n=300										
	30	Setting 1	[0.00, 0.02]	[0.00, 0.05]	[0.02, 0.10]	[0.00, 1.01]	[0.02, 0.03]	[0.02, 0.02]	[0.03, 0.11]	[-0.01, 1.02]
n=100										
	30	Setting 2	[0.01, 0.02]	[0.01, 0.03]	[0.02, 0.10]	[-0.01, 1.02]	[0.02, 0.06]	[0.02, 0.05]	[0.05, 0.16]	[-0.01, 1.05]
n=300										
	30	Setting 2	[0.01, 0.03]	[0.01, 0.05]	[0.02, 0.14]	[-0.00, 1.02]	[0.02, 0.05]	[0.02, 0.03]	[0.04, 0.13]	[-0.02, 1.05]

Table 4.2: Simulation performance for Bernoulli distribution

			$\sigma = 0.5$				$\sigma = 1$			
	Follow-ups	Setting	Ebias	Estd	APstd	APmedian	Ebias	Estd	APstd	APmedian
n=100										
	30	Setting 1	[0.04, 0.02]	[0.04, 0.03]	[0.05, 0.08]	[0.04, 0.99]	[0.05, 0.03]	[0.06, 0.04]	[0.05, 0.08]	[0.04, 0.98]
n=300										
	30	Setting 1	[0.03, 0.01]	[0.02, 0.01]	[0.03, 0.07]	[0.03, 0.98]	[0.05, 0.02]	[0.03, 0.01]	[0.03, 0.07]	[0.05, 0.97]
n=100										
	30	Setting 2	[0.04, 0.05]	[0.04, 0.05]	[0.05, 0.06]	[0.02, 1.03]	[0.02, 0.05]	[0.03, 0.05]	[0.05, 0.06]	[0.00, 1.03]
n=300										
	30	Setting 2	[0.02, 0.03]	[0.01, 0.02]	[0.03, 0.04]	[0.02, 1.02]	[0.01, 0.02]	[0.01, 0.01]	[0.03, 0.04]	[0.01, 1.02]

4.4 Application

In this section, we present an analysis of longitudinal repeated count measurements in Alzheimer's disease (AD). Cognitive scales are used frequently in clinical practice to identify the presence of probable AD. One of the most commonly used scales for screening cognitive ability is the Mini-Mental State Examination (MMSE). The MMSE, developed by Folstein et al. in 1975 (Folstein et al., 1975), is a widely used tool for evaluating cognitive impairment. It is often used in clinical trials for early Alzheimer's Disease (AD) diagnosis, monitoring disease progression, and determining optimal treatment strategies. The MMSE score, ranging from 0 to 30, assesses five areas of cognitive function: orientation, registration, attention and calculation, recall, and language. While a lower MMSE score generally indicates more severe cognitive impairment, the MMSE has traditionally been used to distinguish between patients with cognitive impairment and dementia using a cutoff value. However, various factors can influence MMSE performance (Foreman et al., 1996), leading to

suggestions for different cutoff values. There is growing research interest in investigating the impact of factors on MMSE scores. There is clear evidence that the presence of the apolipoprotein $\epsilon 4$ allele ($ApoE4$) is associated with an increased risk of AD, and $ApoE4$ carriers have been shown to perform worse on the MMSE compared to non-carriers (Christensen et al., 2008). Age-related changes in MMSE scores begin at age 55-60 and these declines increase more rapidly in individuals over the age of 75 (Dufouil et al., 2000). A study of 331 older adults, aged between 77 to 101, showed a significant linear decline in mean MMSE scores from baseline over time Armstrong Esther et al. (2004). There are mixed suggestions for age adjustment of cognitive scores. Some suggest that age adjustment of MMSE is necessary (Kittner et al., 1986), while others argue that adjusting cut-off scores for age fails to identify individuals in the early stages of dementia (Morris et al., 1996). Moreover, previous cross-sectional studies have observed a significant drop in the MMSE score at initial diagnosis with increasing baseline age. Additionally, women under 90 had lower MMSE scores (Pradier et al., 2014). Therefore, comparing raw MMSE scores without controlling other factors might introduce bias. Despite the MMSE score's extensive use, prior studies investigating the longitudinal trajectory of MMSE scores and their associations with AD-related risk factors remain limited. In this study, we aimed to investigate the influence of AD-related factors on the MMSE score over time. Given that MMSE scores do not follow a normal distribution, we adapted the MMSE error counts using our proposed model for longitudinal counts, which is based on the Poisson distribution.

The data used in this study was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (adni.loni.usc.edu). Our study included 2430 par-

ticipants, aged 50.4 to 91.4 years. The outcome of interest in this study was the MMSE score. We transformed the raw score into error counts of MMSE, where a higher value indicates worse cognitive impairment. The covariates included in the model were baseline age, diagnostics, gender, apolipoprotein E4 allele (*ApoE*4) status. Participants with invalid baseline measurements were excluded from further data analyses. In total, 2205 participants were included in the analysis. They include 798 normal cognitive (NC) controls, 1018 mild cognitive impairment (MCI) participants, and 389 Alzheimer's disease (AD) participants. The MCI group includes participants with early mild cognitive impairment (EMCI) and late mild cognitive impairment (LMCI). The NC group includes participants with normal cognitive (NC) and Stable mild cognitive impairment (SMCI). We coded the three diagnostic categories as 0, 1, and 2, corresponding to increasing disease severity from NC to MCI to AD. The group includes 1012 *ApoE*4 carriers and 1193 non-carriers, with 1032 females and 1173 males. All participants had at least one measurement. The average number of repeated measurements per participant was 5.05, with a range of 1 to 18. In total 16033 MMSE measurements were included in the analysis.

4.4.1 Application of proposed model and results

The same noninformative prior distributions outlined in the simulation section were utilized for the parameters. The Gibbs sampler was run for 2500 iterations, with the first 500 iterations discarded as burn-in. Every 10th sample from the subsequent 2000 iterations was used to estimate the posterior distribution of the parameters. The model was developed in Python and implemented on Indiana University's high-performance computing cluster. The results are presented in Table 4.3. By care-

fully controlling for other risk factors, our study results demonstrated that baseline age, *ApoE*4 status, baseline diagnostic categories, and gender significantly impacted the MMSE error counts. Increasing baseline age significantly increased MMSE error counts. *ApoE*4 carriers had significantly higher MMSE error counts than non-carriers. Male participants had significantly lower MMSE error counts than females. AD participants had significantly higher MMSE error counts than normal control participants.

Table 4.3: Posterior summary of fixed effects

	Mean	Std	2.5%	Median	97.5%
Baseline Age	0.16	0.02	0.12	0.15	0.20
*ApoE*4 (ref: non-carrier)	0.14	0.07	0.01	0.16	0.26
Baseline Diagnosis (ref: CN)	0.94	0.06	0.83	0.95	1.04
Gender (ref: female)	-0.19	0.10	-0.37	-0.18	-0.01

4.5 Conclusion

In this paper, we proposed novel generalized semiparametric mixed effect models (GSPMMs). We utilized general techniques for Bayesian inference to extend generalized linear mixed models (GLMMs) and semiparametric mixed models (SPMMs). Our model involved assuming a stochastic process prior to the distribution of non-parametric baseline function, with the prior distribution specified to be governed by a stochastic ordinary differential equations (ODEs) system. We then fitted the model with a Gibbs sampler. We also demonstrated how our model can be effectively ap-

plied to longitudinal data following the Poisson and Bernoulli distributions, which are often collected in popular research areas such as Alzheimer's disease studies.

We conducted a simulation study to assess the performance of our model in estimating the true parameter values. The results showed that our model performed well across all scenarios, demonstrating its effectiveness in estimating the true parameter values for both Poisson and Bernoulli distributed outcomes. We also applied our model to the Alzheimer's Disease Neuroimaging Initiative (ADNI) study dataset and successfully identified scientifically meaningful factors that longitudinally influence the performance of one of the most widely used cognitive assessments, the Mini-Mental State Examination (MMSE).

The important contributions of this article include relaxing the parametric specification of the baseline mean function of the outcome over time in GLMMs to allow for more flexibility. By applying Bayesian approaches, we avoided the need for high-dimensional numerical integration in the frequentist method, thereby extending semi-parametric mixed models to accommodate outcomes from various distributions in the exponential family. We also demonstrated the practical utility of applying stochastic ODE processes as prior distribution for the baseline function. Our model provides a flexible framework for modeling longitudinal data and can be applied to a wide range of research areas.

In the current paper, we focused on exponential family distributions, especially the longitudinal Poisson and Bernoulli distributions. Future research could extend our model to other distributions, such as Tweedie, negative binomial, Beta, shifted t-distributions.

4.6 Appendix

4.6.1 Distribution Definitions

distribution	density/probability function in y	abbreviation
Poisson	$\frac{\mu^y e^{-\mu}}{y!}; \quad y = 0, 1, \ldots$	Poisson(μ)
Negative Binomial	$\frac{\Gamma(\varphi+y)}{\Gamma(\varphi)y!}\left(\frac{\varphi}{\mu+\varphi}\right)^{\varphi}\left(\frac{\mu}{\mu+\varphi}\right)^{y}; \quad y = 0, 1, \ldots$	NB(μ, φ)
Bernoulli	$p^y(1-p)^{1-y}; \quad y = 0, 1$	Bern(p)
Normal	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}; \quad y \in \mathbb{R}$	N(μ, σ^2)
Half-Cauchy	$\frac{2}{\pi\sigma((y/\sigma)^2+1)}; \quad y > 0; \sigma > 0$	Half-Cauchy(σ)
Gamma	$\frac{\beta^\alpha}{\Gamma(\alpha)y^{\alpha-1}e^{-\beta y}}; \quad y > 0$	Gamma(α, β)

References

Altmann, A., L. Tian, V. W. Henderson, M. D. Greicius, and A. D. N. I. Investigators (2014). Sex modifies the apoe-related risk of developing alzheimer disease. *Annals of neurology 75*(4), 563–573.

Andersen, K., L. J. Launer, M. E. Dewey, L. Letenneur, A. Ott, J. Copeland, J.-F. Dartigues, P. Kragh-Sorensen, M. Baldereschi, C. Brayne, et al. (1999). Gender differences in the incidence of ad and vascular dementia: The eurodem studies. *Neurology 53*(9), 1992–1992.

Arai, H., M. Terajima, M. Miura, S. Higuchi, T. Muramatsu, S. Matsushita, N. Machida, T. Nakagawa, V. M.-Y. Lee, J. Q. Trojanowski, et al. (1997). Effect of genetic risk factors and disease progression on the cerebrospinal fluid tau levels in alzheimer's disease. *Journal of the American Geriatrics Society 45*(10), 1228–1231.

Armstrong Esther, C., B. Hagen, M. Sandilands, and C. Smith (2004). Assessing cognitive impairment in older people: the watson clock drawing test. *British journal of community nursing 9*(8), 350–355.

Bernal-Rusiel, J. L., D. N. Greve, M. Reuter, B. Fischl, M. R. Sabuncu, A. D. N. Initiative, et al. (2013). Statistical analysis of longitudinal neuroimage data with linear mixed effects models. *Neuroimage 66*, 249–260.

Bertens, D., D. L. Knol, P. Scheltens, P. J. Visser, A. D. N. Initiative, et al. (2015). Temporal evolution of biomarkers and cognitive markers in the asymptomatic, mci, and dementia stage of alzheimer's disease. *Alzheimer's & Dementia 11*(5), 511–522.

Blomberg, M., M. Jensen, H. Basun, L. Lannfelt, and L.-O. Wahlund (1996). Increasing cerebrospinal fluid tau levels in a subgroup of alzheimer patients with apolipoprotein e allele $\varepsilon 4$ during 14 months follow-up. *Neuroscience letters 214*(2-3), 163–166.

Breslow, N. E. (1984). Extra-poisson variation in log-linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 33*(1), 38–44.

Bush, C. A. and S. N. MacEachern (1996). A semiparametric bayesian model for randomised block designs. *Biometrika 83*(2), 275–285.

Butcher, J. C. (1987). *The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods*. Wiley-Interscience.

Butcher, J. C. (2016). *Numerical methods for ordinary differential equations*. John Wiley & Sons.

Butler, S. M. and T. A. Louis (1992). Random effects models with non-parametric priors. *Statistics in medicine 11*(14-15), 1981–2000.

Cameron, A. C. and P. K. Trivedi (2013). *Regression analysis of count data*. Number 53. Cambridge university press.

Chen, J. and H. Wu (2008a). Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to hiv-1 dynamics. *Journal of the American Statistical Association 103*(481), 369–384.

Chen, J. and H. Wu (2008b). Estimation of time-varying parameters in deterministic dynamic models. *Statistica Sinica*, 987–1006.

Chen, J., D. Zhang, and M. Davidian (2002). A monte carlo em algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics 3*(3), 347–360.

Chen, R. T., Y. Rubanova, J. Bettencourt, and D. K. Duvenaud (2018). Neural ordinary differential equations. *Advances in neural information processing systems 31*.

Chen, X. R., Y. Shao, M. J. Sadowski, A. D. N. Initiative, et al. (2021). Segmented linear mixed model analysis reveals association of the apoe $\varepsilon 4$ allele with faster rate of alzheimer's disease dementia progression. *Journal of Alzheimer's Disease 82*(3), 921–937.

Cho, H., S. W. Seo, J.-H. Kim, C. Kim, B. S. Ye, G. H. Kim, Y. Noh, H. J. Kim, C. W. Yoon, J.-K. Seong, et al. (2013). Changes in subcortical structures in early-versus late-onset alzheimer's disease. *Neurobiology of Aging 34*(7), 1740–1747.

Christensen, H., P. J. Batterham, A. J. Mackinnon, A. F. Jorm, H. A. Mack, K. A. Mather, K. J. Anstey, P. S. Sachdev, and S. Easteal (2008). The association of apoe genotype and cognitive decline in interaction with risk factors in a 65–69 year old community sample. *BMC geriatrics 8*, 1–10.

Crouch, E. A. and D. Spiegelman (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t)\exp(-t^2)\,dt$: Application to logistic-normal models. *Journal of the American Statistical Association 85*(410), 464–469.

Damoiseaux, J. S., W. W. Seeley, J. Zhou, W. R. Shirer, G. Coppola, A. Karydas, H. J. Rosen, B. L. Miller, J. H. Kramer, and M. D. Greicius (2012). Gender modulates the apoe $\varepsilon 4$ effect in healthy older adults: convergent evidence from functional brain connectivity and spinal fluid tau levels. *Journal of Neuroscience 32*(24), 8254–8262.

Diggle, P. (2002). *Analysis of longitudinal data*. Oxford university press.

Dormand, J. R. and P. J. Prince (1980). A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics 6*(1), 19–26.

Drzezga, A. (2010). Amyloid-plaque imaging in early and differential diagnosis of dementia. *Annals of nuclear medicine 24*, 55–66.

Dufouil, C., D. Clayton, C. Brayne, L.-Y. Chi, T. R. Dening, E. Paykel, D. W. O'Connor, A. Ahmed, M. McGee, and F. A. Huppert (2000). Population norms for the mmse in the very old: estimates based on longitudinal data. *Neurology 55*(11), 1609–1613.

Einkemmer, L. (2018). An adaptive step size controller for iterative implicit methods. *Applied Numerical Mathematics 132*, 182–204.

Euler, L. (1792). *Institutiones calculi integralis*, Volume 1. impensis Academiae imperialis scientiarum.

Fagan, A. M., M. A. Mintun, A. R. Shah, P. Aldea, C. M. Roe, R. H. Mach, D. Marcus, J. C. Morris, and D. M. Holtzman (2009). Inversely proportional csf amyloid levels and amyloid quotient (aq): method and validation of a novel csf biomarker for alzheimer's disease. *Annals of neurology 65*(2), 122–131.

Fahrmeir, L. and L. Osuna Echavarría (2006). Structured additive regression for overdispersed and zero-inflated count data. *Applied Stochastic Models in Business and Industry 22*(4), 351–369.

Folstein, M. F., S. E. Folstein, and P. R. McHugh (1975). "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research 12*(3), 189–198.

Foreman, M. D., K. Fletcher, L. C. Mion, L. Simon, and N. Faculty (1996). Assessing cognitive function: The complexities of assessment of an individual's cognitive status are important in making an accurate and comprehensive evaluation. *Geriatric Nursing 17*(5), 228–232.

Garbarino, S., M. Lorenzi, A. D. N. Initiative, et al. (2019). Modeling and inference of spatio-temporal protein dynamics across brain networks. In *International Conference on Information Processing in Medical Imaging*, pp. 57–69. Springer.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper).

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Ghidey, W., E. Lesaffre, and P. Eilers (2004). Smooth random effects distribution in a linear mixed model. *Biometrics 60*(4), 945–953.

Goodfellow, I. J., Y. Bengio, and A. Courville (2016). *Deep Learning*. Cambridge, MA, USA: MIT Press.

Green, C., J. Shearer, C. W. Ritchie, and J. P. Zajicek (2011). Model-based economic evaluation in alzheimer's disease: a review of the methods available to model alzheimer's disease progression. *Value in health 14*(5), 621–630.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability 143*(143), 8.

He, X., Z.-Y. Zhu, and W.-K. Fung (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika 89*(3), 579–590.

Henderson, J. and G. Michailidis (2014). Network reconstruction using nonparametric additive ode models. *PloS one 9*(4), e94003.

Hinde, J. (1982). Compound poisson regression models. In *Glim 82: Proceedings of the international conference on generalised linear models*, pp. 109–121. Springer.

Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural networks 2*(5), 359–366.

Hua, X., D. P. Hibar, S. Lee, A. W. Toga, C. R. Jack Jr, M. W. Weiner, P. M. Thompson, A. D. N. Initiative, et al. (2010). Sex and age differences in atrophic

rates: an adni study with n= 1368 mri scans. *Neurobiology of aging 31*(8), 1463–1480.

Jack, C. R. J., D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski (2010). Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology 9*(1), 119–128.

Jack Jr, C. R., D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish, et al. (2018). Nia-aa research framework: toward a biological definition of alzheimer's disease. *Alzheimer's & dementia 14*(4), 535–562.

Jack Jr, C. R. and D. M. Holtzman (2013). Biomarker modeling of alzheimer's disease. *Neuron 80*(6), 1347–1358.

Jack Jr, C. R., D. S. Knopman, W. J. Jagust, R. C. Petersen, M. W. Weiner, P. S. Aisen, L. M. Shaw, P. Vemuri, H. J. Wiste, S. D. Weigand, et al. (2013). Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The lancet neurology 12*(2), 207–216.

James, G. M. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association 98*(462), 397–408.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and M. Welling (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P., M. Welling, et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning 12*(4), 307–392.

Kittner, S. J., L. R. White, M. E. Farmer, M. Wolz, E. Kaplan, E. Moes, J. A. Brody, and M. Feinleib (1986). Methodological issues in screening for dementia: the problem of education adjustment. *Journal of chronic diseases 39*(3), 163–170.

Kleinman, K. P. and J. G. Ibrahim (1998). A semi-parametric bayesian approach to generalized linear mixed models. *Statistics in Medicine 17*(22), 2579–2596.

Kukull, W. A., R. Higdon, J. D. Bowen, W. C. McCormick, L. Teri, G. D. Schellenberg, G. Van Belle, L. Jolley, and E. B. Larson (2002). Dementia and alzheimer disease incidence: a prospective cohort study. *Archives of neurology 59*(11), 1737–1746.

Kumar, A., J. Sidhu, A. Goyal, and et al. (2022). Alzheimer disease. *StatPearls [Internet]*.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association 73*(364), 805–811.

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

LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep learning. *nature 521*(7553), 436–444.

Li, J.-Q., L. Tan, H.-F. Wang, M.-S. Tan, L. Tan, W. Xu, Q.-F. Zhao, J. Wang, T. Jiang, and J.-T. Yu (2016). Risk factors for predicting progression from mild

cognitive impairment to alzheimer's disease: a systematic review and meta-analysis of cohort studies. *Journal of Neurology, Neurosurgery & Psychiatry 87*(5), 476–484.

Lu, T., H. Liang, H. Li, and H. Wu (2011). High-dimensional odes coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association 106*(496), 1242–1258.

Lu, Y. and J. Lu (2020). A universal approximation theorem of deep neural networks for expressing probability distributions. *Advances in neural information processing systems 33*, 3094–3105.

Luan, Y. and H. Li (2004). Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics 20*(3), 332–339.

Ma, P., C. I. Castillo-Davis, W. Zhong, and J. S. Liu (2006). A data-driven clustering method for time course gene expression data. *Nucleic acids research 34*(4), 1261–1269.

Ma, P. and W. Zhong (2008). Penalized clustering of large-scale functional data with multiple covariates. *Journal of the American Statistical Association 103*(482), 625–636.

Ma, Y., V. Dixit, M. J. Innes, X. Guo, and C. Rackauckas (2021). A comparison of automatic differentiation and continuous sensitivity analysis for derivatives of differential equation solutions. In *2021 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–9. IEEE.

MacKay, D. J. et al. (1998). Introduction to gaussian processes. *NATO ASI series F computer and systems sciences 168*, 133–166.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

McCulloch, C. E., S. R. Searle, and J. M. Neuhaus (2001). *Generalized, linear, and mixed models*, Volume 325. Wiley Online Library.

Miech, R., J. C. Breitner, P. Zandi, A. Khachaturian, J. Anthony, L. Mayer, et al. (2002). Incidence of ad may decline in the early 90s for men, later for women: The cache county study. *Neurology 58*(2), 209–218.

Mielke, M. M., N. J. Haughey, V. V. R. Bandaru, D. D. Weinberg, E. Darby, N. Zaidi, V. Pavlik, R. S. Doody, and C. G. Lyketsos (2011). Plasma sphingomyelins are associated with cognitive progression in alzheimer's disease. *Journal of Alzheimer's disease 27*(2), 259–269.

Molenberghs, G. and G. Verbeke (2005). Models for discrete longitudinal data.

Morris, J., M. Storandt, D. McKeel Jr, E. Rubin, J. Price, E. Grant, and L. Berg (1996). Cerebral amyloid deposition and diffuse plaques in "normal" aging: Evidence for presymptomatic and very mild alzheimer's disease. *Neurology 46*(3), 707–719.

Moyeed, R. and P. Diggle (1994). Rates of convergence in semi-parametric modelling of longitudinal data. *Australian Journal of Statistics 36*(1), 75–93.

Mukhopadhyay, S. and A. E. Gelfand (1997). Dirichlet process mixed generalized linear models. *journal of the American Statistical Association 92*(438), 633–639.

Müller, P. and G. L. Rosner (1997). A bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association 92*(440), 1279–1292.

Murray, J. D. (2002). *Mathematical Biology I. An Introduction* (3 ed.). New York: Springer.

Nagy, Z., M. Esiri, K. Jobst, C. Johnston, S. Litchfield, E. Sim, and A. Smith (1995). Influence of the apolipoprotein e genotype on amyloid deposition and neurofibrillary tangle formation in alzheimer's disease. *Neuroscience 69*(3), 757–761.

Neal, R. M. and G. E. Hinton (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pp. 355–368. Springer.

Nocedal, J. and S. J. Wright (2006). *Numerical Optimization* (2e ed.). New York, NY, USA: Springer.

Nurujjaman, M. (2020). Enhanced euler's method to solve first order ordinary differential equations with better accuracy. *Journal of Engineering Mathematics & Statistics 4*(1), 1–13.

Pennell, M. L. and D. B. Dunson (2007). Fitting semiparametric random effects models to large data sets. *Biostatistics 8*(4), 821–834.

Petrella, J. R., W. Hao, A. Rao, and P. M. Doraiswamy (2019). Computational causal modeling of the dynamic biomarker cascade in alzheimer's disease. *Computational and mathematical methods in medicine 2019*.

Pinheiro, J. C. and D. M. Bates (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing 6*, 289–296.

Pradier, C., C. Sakarovitch, F. Le Duff, R. Layese, A. Metelkina, S. Anthony, K. Tifratene, and P. Robert (2014). The mini mental state examination at the time of alzheimer's disease and related disorders diagnosis, according to age, education, gender and place of residence: a cross-sectional study among the french national alzheimer database. *PloS one 9*(8), e103630.

Quintana, F. A., W. O. Johnson, L. E. Waetjen, and E. B. Gold (2016). Bayesian nonparametric longitudinal data analysis. *Journal of the American Statistical Association 111*(515), 1168–1181.

Ranjan, B., K. H. Chong, and J. Zheng (2018). Composite mathematical modeling of calcium signaling behind neuronal cell death in alzheimer's disease. *BMC systems biology 12*(1), 61–74.

Rezende, D. J., S. Mohamed, and D. Wierstra (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR.

Roberts, S., M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371*(1984), 20110550.

Rolstad, S., A. I. Berg, M. Bjerke, K. Blennow, B. Johansson, H. Zetterberg, and A. Wallin (2011). Amyloid-β 42 is associated with cognitive impairment in healthy

elderly and subjective cognitive impairment. *Journal of Alzheimer's Disease 26*(1), 135–142.

Rubanova, Y., R. T. Chen, and D. K. Duvenaud (2019). Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems 32*.

Sakamoto, E. and H. Iba (2001). Inferring a system of differential equations for a gene regulatory network by using genetic programming. In *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No. 01TH8546)*, Volume 1, pp. 720–726. IEEE.

Scott Long, J. (1997). Regression models for categorical and limited dependent variables. *Advanced quantitative techniques in the social sciences 7*.

Skrondal, A. and S. Rabe-Hesketh (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics 34*(4), 712–745.

Spiegelhalter, D., A. Thomas, N. Best, and W. Gilks (1996). Bugs 0.5: Bayesian inference using gibbs sampling manual (version ii). *MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK*, 1–59.

Spiegelhalter, D. J., K. R. Abrams, and J. P. Myles (2004). *Bayesian approaches to clinical trials and health-care evaluation*, Volume 13. John Wiley & Sons.

Spieth, C., N. Hassis, and F. Streichert (2006). Comparing mathematical models on the problem of network inference. In *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pp. 279–286.

Sunderland, T., B. Wolozin, D. Galasko, J. Levy, R. Dukoff, M. Bahro, R. Lasser, R. Motter, T. Lehtimäki, and P. Seubert (1999). Longitudinal stability of csf tau levels in alzheimer patients. *Biological psychiatry 46*(6), 750–755.

Tapiola, T., M. Lehtovirta, J. Ramberg, S. Helisalmi, K. Linnaranta, P. Riekkinen, and H. Soininen (1998). Csf tau is related to apolipoprotein e genotype in early alzheimer's disease. *Neurology 50*(1), 169–174.

Taylor, J. M., W. Cumberland, and J. Sy (1994). A stochastic model for analysis of longitudinal aids data. *Journal of the American Statistical Association 89*(427), 727–736.

Tiveci, S., A. Akın, T. Çakır, H. Saybaşılı, and K. Ülgen (2005). Modelling of calcium dynamics in brain energy metabolism and alzheimer's disease. *Computational biology and chemistry 29*(2), 151–162.

Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association 91*(433), 217–221.

Verbeke, G. and E. Lesaffre (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis 23*(4), 541–556.

Verbeke, G., G. Molenberghs, and G. Verbeke (1997). *Linear mixed models for longitudinal data*. Springer.

Weaver, D. C., C. T. Workman, and G. D. Stormo (1999). Modeling regulatory networks with weight matrices. In *Biocomputing'99*, pp. 112–123. World Scientific.

WHO, W. H. O. (2023).

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 31*(2), 144–148.

Williams, J. H., G. K. Wilcock, J. Seeburger, A. Dallob, O. Laterza, W. Potter, and A. D. Smith (2011). Non-linear relationships of cerebrospinal fluid biomarker levels with cognitive function: an observational study. *Alzheimer's research & therapy 3*, 1–11.

Winkelmann, R. (2008). *Econometric analysis of count data*. Springer Science & Business Media.

Wright, S. J. (2006). Numerical optimization.

Wu, H., T. Lu, H. Xue, and H. Liang (2014). Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association 109*(506), 700–716.

Xue, H., A. Kumar, and H. Wu (2019). Parameter estimation for semiparametric ordinary differential equation models. *Communications in Statistics-Theory and Methods 48*(24), 5985–6004.

Zeger, S. L. and P. J. Diggle (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, 689–699.

Zeger, S. L. and M. R. Karim (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American statistical association 86*(413), 79–86.

Zhang, D. and M. Davidian (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics 57*(3), 795–802.

Zhang, D., X. Lin, J. Raz, and M. Sowers (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association 93*(442), 710–719.

Curriculum Vitae

Yunyi Li

EDUCATION

- Ph.D. in Biostatistics, Indiana University, Indianapolis, IN, 2024

  (minor in Computer Science)
- M.S. in Biostatistics, Georgetown University, Washington, DC, 2016
- B.S. in Applied Psychology, Sun Yat-sen University, Guangzhou, China, 2012

WORKING EXPERIENCE

- Summer Intern, U.S. Food and Drug Administration, Silver Spring, MD, 2022
- Research Assistant, Indiana University, Indianapolis, IN, 2019 - 2024
- Biostatistician, Roche Tissue Diagnostics, Tucson, AZ, 2016 - 2018

AWARD

- 2019-2020 Richard M. Fairbanks Fellowship Award
- 2008-2009 China's National Scholarship
- 2008-2009 Sun Yat-Sen University First-class Scholarship
- 2010-2011 National Student Innovative Experiment Award