Estimating the Dynamic Diarrhea Effects on Childhood Growth with Statistical Models

Ye Lin

M.S., University of Virginia, US, 2017
B.S., Peking University, P.R.China, 2015

Thesis Proposal Presented to the Graduate Faculty
of University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Statistics

University of Virginia
April, 2020

_____

_____

_____

_____

Abstract

Diarrhea effects on children health have been a popular topic in biomedical research. Previous studies have focused on short term effect on mortality and morbidity. Recent evidence has raised interests on the long term effect on childhood growth. Instead of treating diarrhea effect as constant as what previous studies do, we propose a dynamic nonparametric model to estimate diarrhea effect as a function, studying how long the diarrhea effect lasts and its changes over time. Simulation study shows that our model can capture the length of diarrhea effect and quantify the effect curve simultaneously. In addition to the original proposed model, we also develop four extended models to estimate curves leveling off to a nonzero constant, to take into account additional covariates, to estimate multiple curves, and to model curves with more than one dimension. The proposed models are applied to the data from NIH cohort study collected from children in Bangladesh. Results of the original model show that the diarrhea effect on children's HAZ score starts to show up at 3 months, becomes most significant at around 9 months with a decrease of HAZ -0.013, and levels off to zero after 15 months. Overall, our models provide new statistical tools to quantify the relationship between diarrhea and childhood growth in a dynamic fashion, which gives us insights on the changing pattern and the effect window of diarrhea effect.

Acknowledgements

I want to give my deepest thank to my advisor, Professor Jianhui Zhou, for his enduring guidance and advice. My thesis work will not be possible without his guidance. I have benefited a lot from his vast knowledge and deep understanding of statistics during every step in my research. He encourages me to work on a topic that I am interested in and inspires me to find a valuable question to start with. When there is a challenge in research, he always offers me new ideas and points me to the right direction. In addition to his academic advices, he also shows me how to work with others and treat others with kindness and honesty. I don't feel pressured, but motivated, working with him. I am honored to be part of his academic family with all his brilliant students before and after me!

I would like to express my deep gratitude to my co-advisor, Professor Jennie Ma, for her guidance on my thesis work as well as my research assistant work in medical school. She teaches me how to collaborate with other team members from different backgrounds and apply statistical tools to solve real world problems, which is a valuable experience that not only equips me with the techniques to find my first internship in the US, but also inspires me to find my thesis topic. Prof. Ma cares about both student's study and life. I cannot thank her enough for inviting me to her thanksgiving dinner every year. It means a lot to me especially as an international student studying abroad.

I am also grateful for Dr. William A. Petri for offering his expertises as a member of my thesis committee. Dr. Petri gives me the opportunity to be part of his team working on public health sciences and infectious diseases. Especially during times like these, I would like to pay tribute to Dr. Petri, his team and all medical researchers and professionals who have done so much important works to

improve the global health.

My gratitude also goes to Prof. Daniel Keenan, for his help and advice on my thesis work being on my thesis committee. He offers new perspectives on the methodology and new ideas on how to extend our models, which has become an important and improved part of my thesis.

I would also like to thank my fellow PhD students and all the friends I have met in Charlottesville. Thank you for your support and company!

Most importantly, I want to thank my parents. Thank you for your love, and thank you for supporting me to pursue my PhD degree!

Contents

Chapter 1

Introduction

1.1 Background

1.1.1 Childhood growth

Childhood is a critical time for both physical and cognitive development. A poor growth outcome during this period of time is not only an indicator of malnutrition and recurrent infection, but also a linkage to mortality and morbidity (WHO database). Two forms of growth failure, stunting, defined as height-for-age Z-score (or HAZ, indicating how many standard deviations the height of the child is above the average of the children of the same age based on WHO standard) being below -2, and wasting, defined as weight-for-age Z-score (WAZ) being below -2, have been associated with poor physical and cognitive development, higher risk of mortality, and reduced economic development and productivity (Derso et al., 2017). Therefore, with the estimates of 155 million children under 5 years of age worldwide being stunted and of 52 million being wasted (UNICEF, WHO, World Bank Group joint malnutrition estimates, 2017), growth faltering, a slower rate of growth, has been

identified as a priority in global health, especially in low-income countries (Prendergast and Humphrey, 2014).

The causes of growth faltering have been well studied. Pre-lacteal feeding, non-exclusive breastfeeding, low meal frequency, dietary diversity, as well as socio-demographic and environmental factors (household wealth, sanitary practice, etc.), are associated with stunting and wasting. Recurrent infections such as diarrhea, as a likely result of poor sanitation conditions, have also been linked to poor growth. These infections could affect the absorption of nutritions and therefore lead to growth faltering (Derso et al., 2017).

Infection control and nutritional intervention are common practices to prevent stunting and wasting. Many intervention methods were studied and used in practice such as breastfeeding support, dietary advice and supplementation and non-nutritional interventions. However, the timing and intervention window haven't been studied extensively. There have been studies suggesting -9 (9 months before birth) to 24 months is the best window for nutritional interventions based on growth data, while Prentice et al. (2013) suggests other window like adolescence is also critical based on their longitudinal analysis of growth data. It is important to understand the dynamic pattern of childhood growth and quantify how different factors affect the growth outcome to make better health policy and conduct more efficient interventions.

1.1.2 Diarrhea

Diarrhea has been one of the leading causes of death and illness for children under 5 years old, especially for children in the developing countries. There are many causes of diarrhea, one of which is infections. There are 3 types of infections that causes

diarrhea: viral infections caused by viruses such as rotavirus and norovirus, bacterial infections due to the bacteria in contaminated water or food such as E.coli, Shigella, and parasitic infections caused by parasites from food or water such as Cryptosporidium. Those causes are closely related to the sanitary conditions especially in the developing countries where there is limited access to clean water. However, in many cases, diarrhea is preventable and treatable. Study has shown that common preventions like rotavirus vaccination, breastfeeding, safe water and improved sanitation are possible and also cost effective (Tindyebwa, 2004). Therefore it is important to have a better understanding of diarrhea effect to make better policies, optimize interventions and allocate resources.

Previous studies have focused on the mortality and morbidity of diarrhea. A study in 2015 (Troeger et al., 2017) shows that diarrheal diseases were responsible for half million deaths among children under 5 years old. Three most common pathogens associated with diarrhea mortality are rotavirus, Cryptosporidium spp, and Shigella spp, contributing to over half of deaths caused by diarrhea. In addition to mortality, diarrhea can also lead to increased risk of other infectious diseases due to diarrhea induced undernutrition. Childhood diarrhea can even continue to have impact into adulthood, with studies showing adult chronic diseases such as cardiovascular disease, diabetes, obesity and hypertension being linked to diarrhea in childhood (Wierzba and Muhib, 2018).

1.1.3 Diarrhea effect on growth curve

Despite these serious effects of diarrhea on morbidity and mortality, recent studies have suggested diarrhea could also have long term effect on population health (Bowen et al., 2012). Some of the complications of diarrhea including dehydration

and malabsorption may not necessarily lead to death but could still have long-term effect on children's growth.

Bowen et al. (2012) shows that less burden of diarrhea leads to better growth and cognitive development for children. Other studies have also shown the negative effect of diarrhea on children's growth outcome. Troeger et al. (2018) shows each day of diarrhea leads to -0.0033 HAZ score, a z-score that measures height for age based on WHO's standard. Schnee et al. (2018) also studies the effect of different types of diarrhea on LAZ (length-for-age z-score) score at 12 months. These studies try to quantify the diarrhea effect on children's growth scores. However, all of the above studies treat diarrhea effect as constant over time, instead of modeling it as a dynamic effect.

It is our goal in this thesis to study the diarrhea effect pattern over time, which could give us valuable information on when the effect is most significant and when it would level off to a certain level. In order to achieve our goal, we first need to treat growth outcome as a dynamic responses, instead of focusing on the growth at a certain time point alone. By doing that, we would be able to characterize how the effect of a certain diarrhea episode on growth changes over time. Figure 1.1 is an example of growth curves, showing the growth patterns of children from different percentiles based on WHO's standard on a population level. On an individual level, similarly, a smooth curve can also be obtained from growth outcomes measured at discrete time points, which are the responses we are interested in. In our study, the goal is to study the association between diarrhea and growth curves, and how each episode of diarrhea affect the growth curves in a dynamic manner.

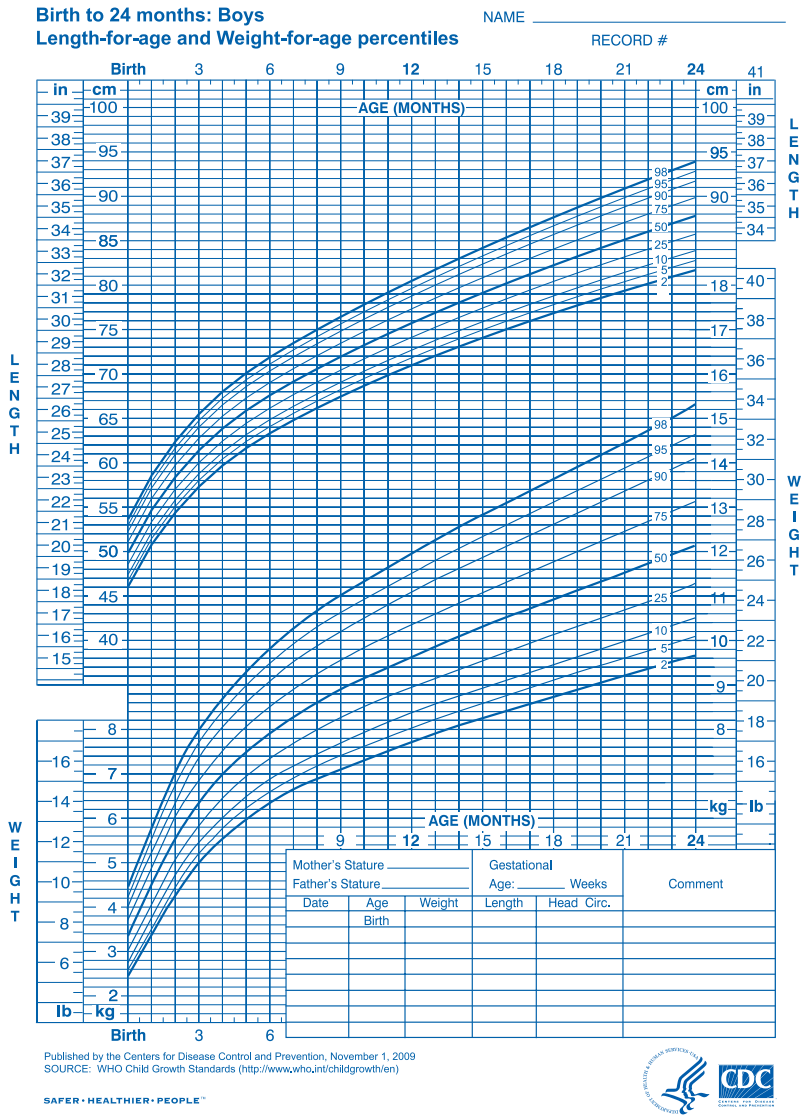

Figure 1.1: An example of growth curves

1.1.4 NIH study cohort

The data we use in this thesis comes from the NIH study. During 2008-2012, a cohort of total 629 infants (332 boys and 297 girls) were enrolled into the study after birth - in Dhaka, Bangladesh. The growth outcomes of each child were recorded every 3 months until the end of the study, resulting in 6831 observations. Measures of growth include weight, height, WHZ (weight-for-height z-score), HAZ (height-for-age z-score), WAZ (weight-for-age z-score), BAZ (BMI-for-age z-score), which are scores set by WHO based on children's growth worldwide. Children's health condition was monitored every two weeks by visits of a research staff. During the visits, questionnaire and follow-ups were given to record children's health condition. If there was an acute illness, the child would be sent to the study clinic for further evaluation. When a child had diarrhea symptoms, stool samples were taken to further decide if it was diarrhea or not and if there are any pathogens present in the stool sample. The information of the starting date of each diarrhea episode was also provided by the questionnaire. In our study, 521 out of 629 children had diarrhea of totally 2605 episodes. Most of the episodes we observed happened during first 3 years of life as shown in Figure 1.2. For each episodes of diarrhea, we tested the presence of Crypto, EH and Giardia, which are 3 common pathogens that could cause diarrhea. Out of all 2605 diarrhea episode samples, 197 of them contain Crypto, 243 of them contain EH, and 731 contain Giardia. Overall, majority of the samples (1601 samples) don't contain any of the three pathogens.

In the data from the NIH study cohort described above, we have both the diarrhea observations and growth outcomes which enables us to study the relationship between diarrhea and growth. Meanwhile, there exist 3 challenges to carry out the statistical analysis to achieve our study goal. First, our goal is to study growth as a

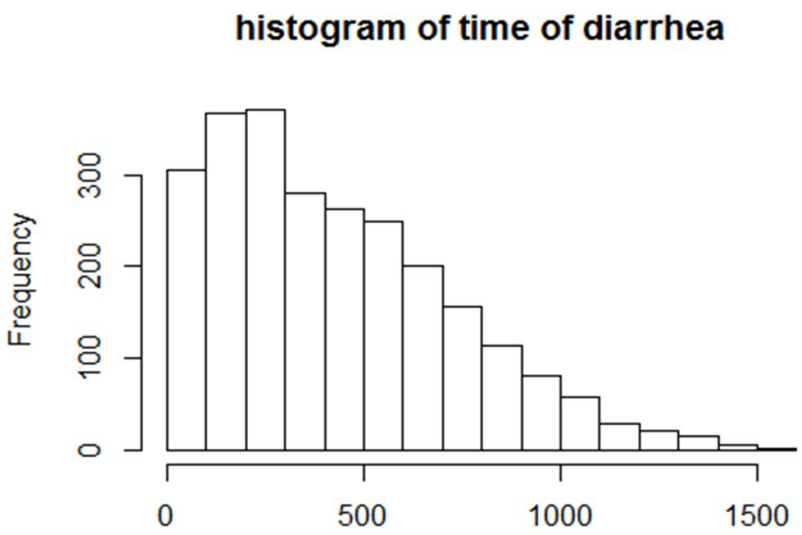

Figure 1.2: histogram of diarrhea onset time (in days)

continuous curve but the growth outcomes in the data are measured at discrete time points. Secondly, those growth outcomes are longitudinal data, which means they are correlated on the subject level. Ignoring the within-subject correlation leads to inefficient, and even biased, statistical estimation. Thirdly, for the diarrhea effect curve over time, we aim to obtain an estimated curve not only smooth, but with some interpretable features such as a curve being flat after a certain period of time indicating that the effect would level off after a period of time.

To address those three challenges, smoothing techniques shall be adopted to estimate the growth curve from discrete growth outcomes, longitudinal models are used to account for within subject correlations, and penalties are incorporated to control the shape of the curve. In the following section, literature on smoothing, longitudinal data analysis and penalized method are reviewed.

1.2 Introduction to smoothing

In our study, the growth outcome was measured repeatedly over discrete time points. To study the underlying smooth trajectory of growth and the diarrhea effect, smoothing is a necessary step. In other words, we want to fit a continuous function of age to estimate the underlying smooth trajectory of children growth, on which the discrete growth outcomes are observed. To fit into the nonparametric framework, we model the response $Y_i = f(X_i) + \epsilon_i$, where X_i represents the predictor such as age, Y_i is the corresponding response such as observed growth outcomes, and ϵ_i is the random error.

Nonparametric regression is to model the relationship between Y and X when we do not make any parametric assumptions of the function f. Two commonly used methods for the model fitting are kernel methods and basis function approximation.

1.2.1 Kernel smoothing

For the kernel methods, the value of the function f at each x is estimated based on the nearby observations with certain weights assigned by the kernel function $K(t)$. The fitted curve can be expressed as:

$$\begin{cases} \mu(x) = \sum_{i=1}^{n} l_i(x) Y_i \\ \\ l_i(x) = \frac{K(\frac{X - X_i}{h})}{\sum_{j=1}^{n} K(\frac{X - X_j}{h})}. \end{cases}$$

An extension of this idea is to use local polynomial regression to approximate f based on the kernel. The idea is to find the coefficients $\boldsymbol{\alpha}$ of the polynomial

approximation that minimize the weighted least squares:

$$\sum_{i=1}^{n} \left\{ Y_i - \alpha_0 - \sum_{j=1}^{p} \alpha_j (X_i - x)^j \right\}^2 K_h(X_i - x).$$

Let $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & X_1 - x & \dots & (X_1 - x)^p \\ 1 & X_2 - x & \dots & (X_2 - x)^p \\ \vdots & \vdots & \dots & \vdots \\ 1 & X_n - x & \dots & (X_n - x)^p \end{pmatrix}, \boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}, \boldsymbol{e} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$

and

$$\boldsymbol{K}_h = \begin{pmatrix} K_h(X_1 - x) & 0 & \dots & 0 \\ 0 & K_h(X_2 - x) & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \vdots & K_h(X_n - x) \end{pmatrix}.$$

Estimators can be calculated as follows:

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{X}^T \boldsymbol{K}_h \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{K}_h \boldsymbol{Y},$$

$$\widehat{\mu}(x) = e^T (\boldsymbol{X}^T \boldsymbol{K}_h \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{K}_h \boldsymbol{Y}.$$

One of the strengths of kernel smoothing is that the asymptotic properties have been well studied (Fan et al., 1996). However, the computational cost can be expansive. In addition, it is not very flexible for estimating curves of different shapes and features. Since in this thesis, our goal is to estimate a curve with flat regions, we use the following spline smoothing technique instead.

1.2.2 Spline methods

The other method is to use basis functions to approximate the function $f(x)$, where a set of basis functions are pre-specified and the target function $f(x)$ is approximated by a linear combination of basis functions, $\mu(x) \approx \boldsymbol{B}^T(x)\gamma$. By projecting observations onto the basis functions, it turns into a linear regression problem: $\boldsymbol{Y} = \boldsymbol{X}\gamma + \epsilon$, where

$$\boldsymbol{B}(x) = \{B_1(x), \ldots, B_{J+p}(x)\}^T, \quad \boldsymbol{X} = \{\boldsymbol{B}(X_1), \ldots, \boldsymbol{B}(X_n)\}^T.$$

The estimator of μ can be found by a least squares estimator:

$$\widehat{\mu}(x) = \boldsymbol{B}^T(x)\widehat{\gamma}, \quad \widehat{\gamma} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

B-spline is a common choice of basis functions. They are piecewise polynomials with order h jointed at a set of pre-specified knots as shown in Figure 1.3. One of the properties is that the target function between two knots is approximated by only h functions, which gives us flexibility to estimate curves with different pattern on different regions. Some asymptotic properties of polynomial regression splines are also developed in Stone (1994) and Zhou et al. (1998).

However, different choices of knots and order can lead to overfitting and unstable results. Ruppert et al. (2003) introduces a formulation of penalized spline, where a smoothness penalty is incorporated in the nonparametric regression. Coefficients of the basis functions $\hat{\gamma}$ is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \boldsymbol{B}^T(t_i)\gamma\}^2 + \lambda\gamma^T\Omega\gamma,$$

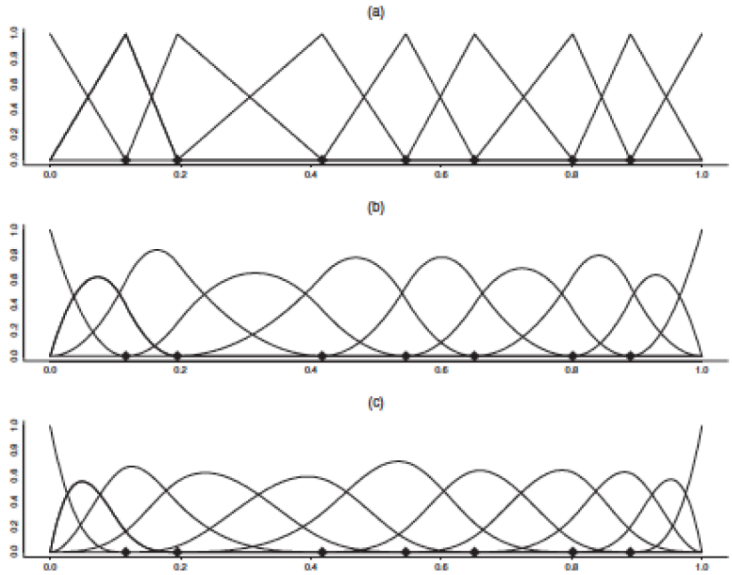

Figure 1.3: plots of B-splines (taken from Ruppert et al. (2003))

where $\boldsymbol{B}(t) = (B_1, \ldots, B_K)^T(t)$, $\quad \widehat{f_\lambda}(t) = \boldsymbol{B}^T(t)\widehat{\gamma}$. The term Ω is a matrix with (j, j')'s entry being $\int B_j^{(m)}(t) B_{j'}^{(m)}(t) dt$, penalizing the intergal of the derivative of the spline function, which would result in the fitted curve to be smooth. Tuning parameter λ can be selected based on model selection criterion such as cross-validation, generalized cross validation, Akaike information criterion (AIC), Bayesian information criterion (BIC), etc.

For some practical problems, we would like the fitted curve to have a certain shape or some properties other than smoothness for interpretability. Constraints need to be incorporated in the smoothing such as monotone smoothing (Ramsay, 1988; Gaylord and Ramirez, 1991) and smoothing for convex functions (Dierckx, 1980). In some applications, we hope to study the sparsity on the curve, which means we aim to identify some sub-regions on which the function is zero. By using B-spline basis functions to represent the target curve, this problem transforms to a variable selection problem where we investigate the sparsity among basis coefficients.

Zhou et al. (2013) proposes a two-step procedure using group penalties to identify the sub-regions on which the function has zero values in a functional linear regression model.

Besides univariate smoothing, bivariate smoothing and multivariate smoothing have also been studied. Bivariate P-splines (Marx and Eilers, 2005; Eilers and Marx, 2003) and thin plate splines (Wood, 2003) have been commonly used for bivariate case. Xiao et al. (2013) also proposed a sandwich smoother and developed asymptotic theories for bivariate P-splines. For estimating functions with more than two dimensions, tensor product splines (Huang, 2003) and multivariate form of P-splines (Currie et al., 2006) are also studied.

1.3 Introduction to longitudinal data

Growth outcomes are repeated measurements, also known as longitudinal data or clustered data. To model the growth curve, we have to take into consideration of the within-subject correlation among repeated measurements on the same subject. Measurements from different subjects are assumed independent. This type of data has been well studied over recent decades. There are mainly two ways of incorporating the within-subject correlation in the longitudinal structure: specifying a working correlation structure as in marginal models or employing random effects as in the mixed effects models.

The first method is used for marginal models. For normally distributed longitudinal responses, estimation and inference methods have been well developed for linear models (Ware, 1985). For generalized linear models, Liang and Zeger (1986) proposes Generalized Estimating Equations (GEE) approach where a working correlation matrix needs to be specified. An advantage of GEE approach is that the

estimator of regression coefficients is consistent even when the correlation structure is mis-specified. For mixed effects models, Laird and Ware (1982) studies random effect models for longitudinal data. Lindstrom and Bates (1990) proposes a non-linear mixed effects model. Mixed effect models have also been used in the context of analysis of variance (ANOVA) (Hedges and Vevea, 1998), and have been extended to the case of multi-level data (Hedeker and Gibbons, 2006), which are widely used in medicine and social sciences.

Variable selection plays an important role for high dimensional longitudinal data analysis to help reduce dimension and improve estimation accuracy. A variety of methods have been developed in the literature to select informative variables for high dimensional longitudinal data. Wang et al. (2012) proposes the Penalized Estimating Equation for variable selection by incorporating penalties in GEE framework and develops the asymptotic theories for the proposed penalized estimator. Fan and Li (2012) studies variable selection for mixed effect models with continuous response. Dziak et al. (2009) proposes to use SCAD-penalized quadratic inference function for variable selection; Xue et al. (2011) proposes a variable selection method for generalized additive model for longitudinal data.

Semi-parametric and nonparametric regression techniques can also be extended to both mixed effects model and marginal models with correlated errors for longitudinal data. Moyeed and Diggle (1994) incorporates random intercept in a semi-parametric model with smooth functions. Hoover et al. (1998) extends this model by allowing the random coefficient to be a random curve changing with time. Guo (2002) generalizes this model as a functional mixed effect model. For marginal models, Wang (2003) and Lin and Carroll (2000) use kernel methods for marginal models accounting for correlated errors. Wang et al. (2005) proposes kernel GEE for a semi-parametric model. Qu and Li (2006) uses quadratic inference functions

for estimation and inference for varying coefficient models with longitudinal data. In application, these semi-parametric models are often used by biomedical researchers to study growth curves because of the nature of the data.

However, those methods reviewed above cannot be applied to our NIH data directly due to the following reasons. First, none of these studies has considered the target curve to have the zero-value sub-regions. We aim to search for some sub-regions on which the diarrhea effect curve/surface levels off. This motivates us to incorporate penalties on sparsity in our longitudinal model. Zhou et al. (2013) considers this particular shape constraints in a functional linear model, but the method is not for longitudinal data. Secondly, none of those methods focuses on smoothing two curves with different features at the same time. For our study, we are interested in estimating the natural growth curve and diarrhea effect curve/surface simultaneously. The former should be monotone for some growth outcomes such as height growth, while the latter should be sparse in some regions, which requires us to estimate those two curves separately. Furthermore, the estimation methods and asymptotic theories for longitudinal models based on smoothing with group penalty on sparsity need to be developed. In our study, we propose a dynamic model based on Group Penalized Generalized Estimating Equations to estimate baseline growth curve and diarrhea effect curve simultaneously and identify the non-zero region of the diarrhea effect curve. We investigate the asymptotic properties of our developed estimator.

1.4 Introduction to penalized methods

Penalized methods have been widely used in both functional data and longitudinal data to regulate the smoothness of the estimates or to conduct variable selection.

The idea of imposing a penalty term to shrink estimator to zero or to control bias-variance trade-off is first used in linear regression models. Instead of finding β that minimizes the least squares, a penalty term is incorporated in the loss function:

$$min_{\boldsymbol{\beta}} \frac{1}{2n} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + P_\lambda(\boldsymbol{\beta}),$$

where $P_\lambda(\boldsymbol{\beta})$ is the penalty function, and λ is the tuning parameter.

Different penalty functions are proposed and studied in literature, such as Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001), Elastic Net (Zou and Hastie, 2005) and Adaptive LASSO (Zou, 2006). The SCAD penalty is defined through its derivative:

$$q_\lambda(\theta) = \lambda\{I(\theta < \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\},$$

for $\theta \geq 0$ and some $a > 2$. Usually we let $a = 3.7$ (Fan and Li 2001). Asymptotic results show that SCAD has Oracle property which includes model selection consistency and parameter estimator normality. An algorithm using iterative ridge regression is proposed to obtain SCAD estimator using local quadratic approximaton in Fan and Li (2001).

Another type of penalty method to achieve shrinkage is the Dantzig Selection, proposed in Candes and Tao (2007). Dantzig Selector is to minimize $||\beta||_{l1}$ subject to $||Xr||_{l_\infty} \leq (1+t^{-1})\sqrt{2log(p)}\sigma$, where $r = y - X\beta$. Estimator can be obtained by linear programming. The large sample properties of Dantzig selector are also developed. Under irrepresentable conditions, model selection consistency is guaranteed (Gai et al., 2013).

For some variable selection problems, predictors are grouped and the goal is to select variables in groups. Yuan and Lin (2006) proposes a group LASSO penalty which is defined as $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{J} ||\beta_j||_{K_j}$, where J is number of groups and K_j is the number of predictors for jth group. Group SCAD has also been adopted and studied in Zhou and Qu (2012).

In general, when penalty function is smooth, we can solve the equation of the derivative of the objective function to be 0 numerically by using Newton-Raphson algorithm. For non-convex penalty functions, Hunter and Li (2005) proposed a minorization-maximization (MM) algorithm, by introducing a small perturbation to render it differentiable.

Chapter 2

Methodology

2.1 Model setup

Our goal is to estimate diarrhea effect curve on growth in the NIH study cohort. As described in section 1.1.4, our dataset consists of two parts as shown in Table 2.1: the growth responses measured every 3 months for each subject, and the diarrhea onset times for all the episodes of diarrhea for each subject. The challenge is that most of the children develops multiple episodes of diarrhea. Since our assumption is that each episodes of diarrhea may have long lasting effect, we are not able to observe the effect of one single diarrhea directly from the data when the growth outcomes we measure include overlapping effects from multiple episodes of diarrhea. In addition, the cumulative effect we observe is hard to separate into individual effect curves because their starting points (diarrhea onset times) are different. To address those challenges, we propose the following model in order to separate individual effect curve that we are interested in from the cumulative effect we observe from the data.

We assume that once there is a diarrhea, the diarrhea will start to have an

Table 2.1: Data summary

NIH Cohort		
Data	Variables	Summary
Growth data	subject id	629 subjects
	gender	297 female, 332 male
	age	the length of enrollment ranges from 1 day to 1560 days
	biomarkers	weight, height, WHZ, HAZ, WAZ, BAZ scores.
Diarrhea data	subject id	521 subjects
	diarrhea episodes	2605 episodes, up to 1623 days since birth
	infection type	include EH_CTRT, GIA_CTRT, CRY_CTRT infections

effect $\beta(t)$ on the growth outcome. The effect $\beta(t)$ is the diarrhea effect curve which measures the effect of a single diarrhea on children's growth changing with elapsed time t. Let $\alpha(t)$ be the natural growth curve without diarrhea for a cohort of children. The growth outcomes $Y(t)$ we observe are modeled as the natural growth curve plus the effects of all the past diarrheas with random error. For model identifiability, we assume that the effect of each episode of diarrhea is the same regardless of the onset time, and the effects of different episode on growth outcomes are additive.

For kth subject, assuming diarrhea happened at time $T_{ki} : T_{k1}, T_{k2}, ...$, we model the response outcomes as

$$Y_k(t) = \alpha(t) + \sum_i \beta(t - T_{ki}) + \epsilon_k(t),$$

where $\epsilon_k(t)$ is the random error.

We observe longitudinal growth outcomes at time $t_{kj} : t_{k1}, t_{k2}, ...,$

$$Y_k(t_{kj}) = \alpha(t_{kj}) + \sum_i \beta(t_{kj} - T_{ki}) + \epsilon_{kj},$$

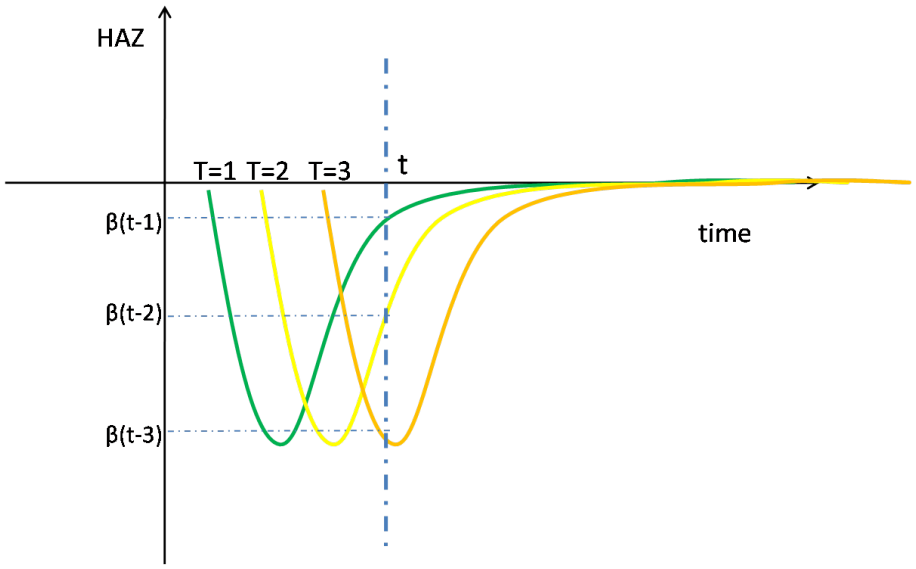

Figure 2.1: cumulative diarrhea effects

$$\boldsymbol{\epsilon_k} \sim N(0, \Sigma_k).$$

As shown above, we model the observed growth curve from the data actually as a summation of the natural growth curve $\alpha(t)$ and all the individual effect curves starting from different time points (shown in Figure 2.1). In other words, the model indicates that in order to estimate the growth outcome at time t, we need to look back to see how long it has been since each episode of diarrhea occurs and add all those effects on top of the growth outcome the child should have with no diarrhea $\alpha(t)$.

We use B-spline basis functions with evenly spaced knots and order h: $v_1(t), v_2(t), ..., v_{D_n}(t)$ and $w_1(t), w_2(t), ..., w_{d_n}(t)$. Those curves can be approximated as

$$\alpha(t) = \sum_{m=1}^{D_n} a_m v_m(t) + e^a(t),$$

$$\beta(t) = \sum_{m=1}^{d_n} b_m w_m(t) + e^b(t).$$

The model then becomes

$$Y_k(t_{kj}) = \sum_{m=1}^{D_n} a_m v_m(t_{kj}) + \sum_i \sum_{m=1}^{d_n} b_m w_m(t_{kj} - T_{ki}) + \widetilde{\epsilon_{kj}}$$

$$= \sum_{m=1}^{D_n} a_m v_m(t_{kj}) + \sum_{m=1}^{d_n} b_m \sum_i w_m(t_{kj} - T_{ki}) + \widetilde{\epsilon}_{kj},$$

where $\widetilde{\epsilon}_{kj} = \epsilon_{kj} + e_{kj}$, ϵ_{kj} is the random error and the approximation error $e_{kj} = e^a_{kj} + \sum_i e^b_{kji}$ is the summation of all the approximation errors from each curve. The term e^a_{kj} is the approximation error from $\alpha(t)$ at t_{kj} and e^b_{kji} is the approximation error from $\beta(t)$ at $t_{kj} - T_{ki}$. The model can be written as $\boldsymbol{Y} = (\boldsymbol{X}^a, \boldsymbol{X}^b)^T \boldsymbol{c} + \widetilde{\boldsymbol{\epsilon}} = \boldsymbol{X}^T \boldsymbol{c} + \widetilde{\boldsymbol{\epsilon}}$, $\boldsymbol{c} = (a_1, ..., a_{D_n}, b_1, ..., b_{d_n}) = (\boldsymbol{a}, \boldsymbol{b})$, where $\boldsymbol{X}^a$ and $\boldsymbol{X}^b$ are the covariates generated from B-spline functions for $\alpha(t)$ and $\beta(t)$ respectively, $\boldsymbol{a}$ is a D_n dimensional vector, $\boldsymbol{b}$ is a d_n dimensional vector and $\boldsymbol{c}$'s dimension is $p_n = D_n + d_n$.

It is straight forward to obtain an estimated $\beta(t)$ in the above approximated model. However, the actual estimated $\beta(t)$ by simply using spline approximation can be problematic, as shown in Figure 2.2. Even though we can see some patterns in this figure, for example, the effect is negative and levels off after around 2 years, the curve doesn't make sense at the beginning where it is positive suggesting diarrhea is good for the growth and at the end when the magnitude of effect starts to increasing. There are two reasons to cause the interoperability problem in practice. First, B-spline approximation has boundary effect, which means the fitted curve may not be accurate at the beginning and at the end of the smoothing interval where fewer observed observations are available and they are all on one side. Secondly, to obtain a reliable estimate for $\beta(t)$, we need observed values evenly spaced in the range of

t. However, the observations for $\beta(t)$ are concentrated at the beginning of age and become less and less as t increases, as shown in Figure 2.3. For instance, in order to observe the diarrhea effect on growth nearly four years after the diarrhea, there has to be a diarrhea right after birth, and at the same time, the child has to be enrolled in the study for whole four years, which is less likely to happen in reality. Therefore as t increases, we have less information available to estimate the value of $\beta(t)$, which would lead to an estimate of $\beta(t)$ with larger variation towards the end of the study.

Another reason that the above method can not be directly applied in our case is that we would like to study the pattern of the diarrhea effect on the growth, and our estimation method should be able to facilitate this pattern recognization goal. Specifically, in this thesis, we are interested to see if the effect of diarrhea on the body should only last for a certain period of time like other changing processes in nature triggered by a certain intervention. If so, it is reasonable to assume that the effect curve $\beta(t)$ would become flat after a certain period of time and level off to 0 (no permanent effect) or a constant level (permanent effect).

All the above motivates us to impose penalties in our model to achieve an interpretable pattern estimation of the diarrhea effect on growth. Penalized method is a statistical way to handle overfitting as well, which can deal with the unstable estimation near the boundary as shown in Figure 2.2. Certain penalties can ensure the fitted curve to be flat and allow us to identify the window of the effect. To identify the null region and estimate the coefficient function on the non-zero region in functional linear model, Zhou et al. (2013) proposed a two-step approach with B-splines approximation. However, their model is for functional linear regression and assumes independent random errors. Penalized models with structured functional estimation for longitudinal data have not been studied in the literature. To assume sparsity

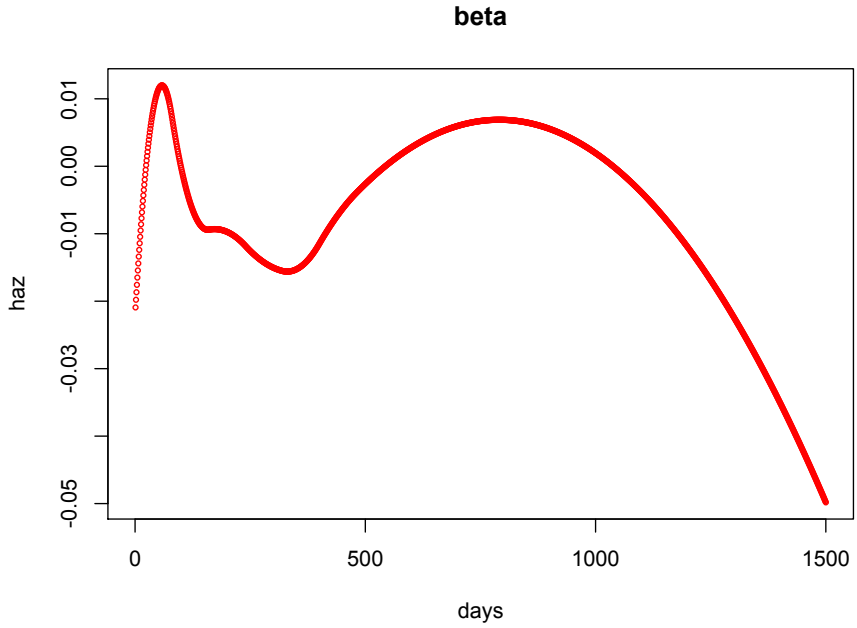

Figure 2.2: estimated effect curve without penalty

in longitudinal models for covariates, variable selection techniques have been well studied. Fan and Li (2012) studies variable selection for mixed effect models with continuous response. Dziak et al. (2009) proposes to use SCAD-penalized quadratic inference function. Wang et al. (2012) proposes the Penalized Estimating Equation for variable selection by incorporating penalties in GEE framework and the asymptotic theories are developed. However, the above methods are to select covariates, not non-zero regions of a function. In this work, we extend the two-step approach in Zhou et al. (2013) in PGEE framework of Wang et al. (2012) and develop a group penalties to select non-zero regions for our dynamic diarrhea effect model as follows.

In our study, we are interested in identifying the zero-value region of the diarrhea effect curve and estimating the effect curve on the non-zero region, we penalize on the magnitudes of the B-spline coefficients, after spline approximation, to promote

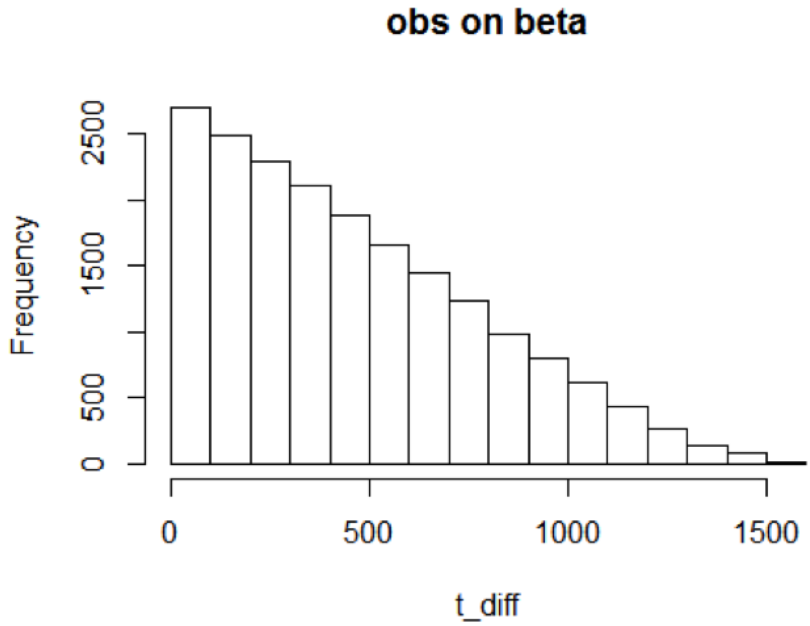

Figure 2.3: observed values on beta

group sparsity. The problem thus transforms into variable selection, in a group fashion, for the non-zero B-splines coefficients for longitudinal data.

2.2 Estimation procedure

The challenge in our study is that the boundary between non-zero region and zero region is unknown and to be identified. Without knowing where the boundary is, we will not be able to place the knots adaptively, which may result in inefficient functional estimation of the dynamic diarrhea effect. For instance, if the boundary is between two knots, even though the estimated B-spline coefficients can be shrunk to zero, the estimated zero region will not be accurate and the obtained estimated effect curve will be biased. Therefore our goal is to develop a procedure to place knots data-adaptively and adopt the appropriate penalties to obtain the estimation.

Ideally, to identify the true boundary T_0 (or the true null region $\mathcal{T} = [T_0, +\infty)$), we need to place more knots so that the estimated zero-value region is more accurate. However, too many knots will lead to overfitting. Therefore, to identify to null region and obtain a reliable estimate of the curve simultaneously, we propose a two-step estimating procedure, which is similar in Zhou et al. (2013), for the more complicated longitudinal data. We split the whole estimating process into two steps: the first step to identify the null region, and the second step to estimate the curves.

2.2.1 Step 1: Dantzig step

During this step, the goal is to obtain an initial estimate of the null region $\hat{\mathcal{T}}^{(0)} = [\hat{T}^{(0)}, +\infty)$, where $\hat{T}^{(0)}$ is the initial estimate of the boundary. We place as many knots as we can so that the estimated null region is more accurate. Since the number of B-spline coefficients to estimate is large, the penalty we adopt in this step should not only ensure sparsity, but also allow fast computation. Therefore we use Dantzig selector to screen for non-zero b_is in this step. Studies have shown the numerical advantages of Dantzig selector over Lasso (James et al., 2009) and also the computational advantage when tuning parameter selection is not necessary for Dantzig selector.

In the Dantzig step, we place $k_{0,n}$ knots over the whole study window in order to identify non-zero region more accurately and to estimate the corresponding B-spline coefficients $\boldsymbol{c}_0$. To estimate $\boldsymbol{c}_0$, we find $argmin_{\boldsymbol{c}_0}||\boldsymbol{c}_0||_{l_1}$ subject to $|\boldsymbol{X}_k^T(\hat{Y} - \boldsymbol{X}\boldsymbol{c}_0)| \leq \lambda$, with $\boldsymbol{X}_k^T$ being the kth column of $\boldsymbol{X}$, $\lambda = n^{\frac{1+(d_2-d_1)k_0}{2}}$ for some k_0, d_1, d_2 satisfying the regularity conditions in Section 2.3 and Chapter 5, where $k_{0,n}$ is the number of inner knots and h is the order. Based on the selected model, we refit the linear model using only the selected basis functions and shrink all coefficients whose absolute

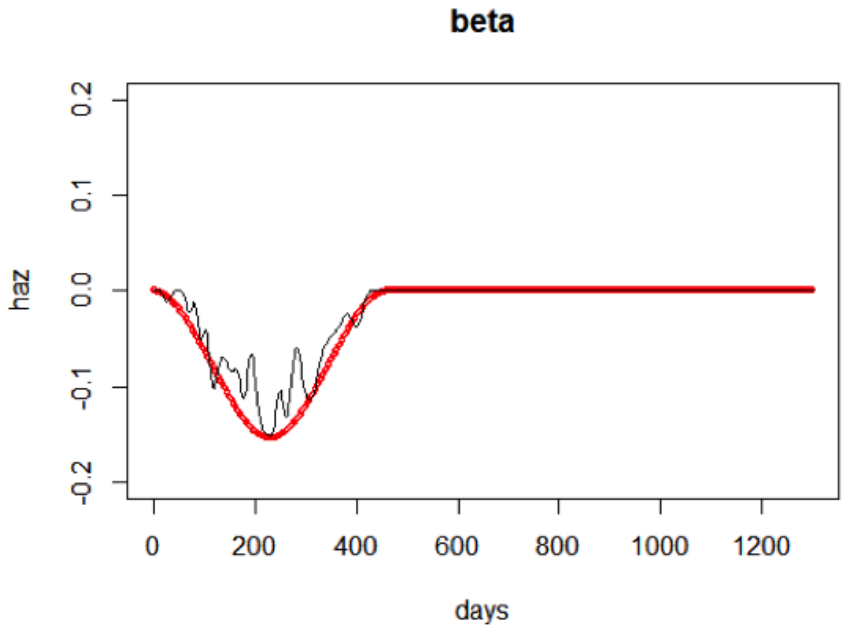

Figure 2.4: observed values on beta

value is less than a pre-specified threshold d to be exact 0. This way we obtain an initial estimate of the null region $\hat{\mathcal{T}}^{(0)} = [\hat{T}^{(0)}, +\infty)$, which should cover the true null region $\mathcal{T} = [T_0, +\infty)$ with probability tending to 1 as n goes to infinity, which is ensured in our Theorem 1 in Section 2.3.

The goal of this step is mainly to identify the boundary T_0, not to fit the curve, because many knots we place during this step would lead to overfitting of the curve. Figure 2.4 shows an example of the fitted curve in this step. Even though the boundary is clearly identified, the fitted curve on the non-zero region is overfitted, which is the reason we need to refine the estimate in the next step.

2.2.2 Step 2: SCAD step

During step 2, we refine the null region based on the one identified in step 1 and focus on estimating $\beta(t)$ on the non-null region $\hat{\mathcal{T}}^{(0),c} = [0, \hat{T}^{(0)})$. Starting with $\hat{T}^{(0)}$ from step 1, we keep refining the boundary to the right with step size Δ, placing $k_{1,n}$ knots on $[0, \hat{T}^{(0)} + \Delta i]$, for $i = 1, 2, ...$ As for $[\hat{T}^{(0)} + \Delta i, +\infty)$, it is treated as the working null region $\mathcal{T}_w$ with zero value and no knots is placed in it. To estimate the B-spline coefficients $\boldsymbol{c}_1$ for a given i, we use Group Penalized GEE with SCAD to select basis functions with none-zero coefficients. Then we select the optimal i that gives us the optimal working null region based on model selection criterion such as AIC, BIC, QIC, etc. Comparing to the Dantzig step, we place less knots on $[0, \hat{T}^{(0)} + \Delta i]$ to avoid over-fitting.

The resulting estimating equations are

$$U(\boldsymbol{c}_1) = S(\boldsymbol{c}_1) - n\mathbf{q}_\lambda^G(\boldsymbol{c}_1)sign(\boldsymbol{c}_1),$$

where

$$S(\boldsymbol{c}_1) = \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{A}_i^{1/2} \hat{\mathbf{R}}_i^{-1} \mathbf{A}_i^{-1/2} (\mathbf{Y}_i - \mu_i(\boldsymbol{c}_1)),$$

where $\hat{\mathbf{R}}_i$ is the estimated working correlation matrix for jth subject, $\boldsymbol{\Sigma}_i = \mathbf{A}_i^{1/2} \hat{\mathbf{R}}_i \mathbf{A}_i^{1/2}$, and μ_i is the identity link function in our model.

The term $\mathbf{q}_\lambda^G(\boldsymbol{c}_1)\mathbf{s}ign(\boldsymbol{c}_1)$ denotes the component-wise product, where

$$sign(\boldsymbol{c}_1) = (sign(c_1), ..., sign(c_p))$$

and

$$\mathbf{q}_\lambda^G(\boldsymbol{c}_1) = (q_\lambda(||\boldsymbol{c}_1^G||_1), ..., q_\lambda(||\boldsymbol{c}_p^G||_1)).$$

The term $q_\lambda(\theta)$ is the derivative of the SCAD penalty,

$$q_\lambda(\theta) = \lambda\{I(\theta < \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\},$$

for $\theta \geq 0$ and some $a > 2$. Usually we let $a = 3.7$ (Fan and Li 2001).

The term $q_\lambda(||\boldsymbol{c}_i^G||_1)$ denotes the group penalty for the ith covariate and $q_\lambda(\theta)$ is given by the SCAD penalty based on the L1 norm of the group vector $\boldsymbol{c}_i^G$ which is the vector of the group c_i (the ith covariates) belongs to. For example, if X_1, X_3, X_4 belong to the same group, then $\boldsymbol{c}_1^G = \boldsymbol{c}_3^G = \boldsymbol{c}_4^G = (c_1, c_3, c_4)$.

In our case, we divide $\boldsymbol{c}_1 = (\boldsymbol{a}_1, \boldsymbol{b}_{1,S,w}, \boldsymbol{b}_{1,N,w})$ into 3 groups based on working null regions $\mathcal{T}_w = [\hat{T}^{(0)} + \Delta i, +\infty)$ for a given i.

The algorithm for estimation is Newton-Raphson combined with minorization-maximization (MM) algorithm, similar to Wang et al. (2012),

$$\boldsymbol{c}_{1,n}^k = \boldsymbol{c}_{1,n}^{k-1} + [\boldsymbol{H}_n(\boldsymbol{c}_{1,n}^{k-1}) + n\boldsymbol{E}_n(\boldsymbol{c}_{1,n}^{k-1})]^{-1}[\boldsymbol{S}_n(\boldsymbol{c}_{1,n}^{k-1}) - n\boldsymbol{E}_n(\boldsymbol{c}_{1,n}^{k-1})\boldsymbol{c}_{1,n}^{k-1}],$$

where

$$\boldsymbol{H}_n(\boldsymbol{c}_{1,n}^{k-1}) = \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{c}_{1,n}^{k-1}) \hat{\mathbf{R}}_i^{-1} \mathbf{A}_i^{1/2}(\boldsymbol{c}_{1,n}^{k-1}) \mathbf{X}_i,$$

$$\boldsymbol{E}_n(\boldsymbol{c}_{1,n}^{k-1}) = diag\{\frac{q_{\lambda_n}(\hat{\boldsymbol{c}_{\boldsymbol{n1}}^{\boldsymbol{G}}})}{\epsilon + |\hat{c_{n1}}|}, ..., \frac{q_{\lambda_n}(\hat{\boldsymbol{c}_{\boldsymbol{np}}^{\boldsymbol{G}}})}{\epsilon + |\hat{c_{np}}|}\}.$$

Following the above iterative algorithm, we are able to obtain the estimator $\hat{\boldsymbol{c}}_1(i)$ for a given working null region $\mathcal{T}_w = [\hat{T}^{(0)} + \Delta i, +\infty)$ after the tuning parameter

selection. The fitted curves can be constructed as:

$$\hat{\alpha}(t) = \sum_{m=1}^{D_n} \hat{a}(i)_m v_m(t),$$

$$\hat{\beta}(t) = \sum_{m=1}^{d_n} \hat{b}(i)_m w_m(t),$$

and the criterion $C(\mathcal{T}_w, \hat{\boldsymbol{c}}_1(i))$ for model selection can be calculated, such as AIC and BIC.

We repeat this process for each working null region $\mathcal{T}_w = [\hat{T}^{(0)} + \Delta i, +\infty)$ for $i = 1, 2, ...,$ and select the best working null region $\hat{\mathcal{T}} = [\hat{T}^{(0)} + \Delta \hat{i}, +\infty)$, where $\hat{i} = argmin_i C(\mathcal{T}_w, \hat{\boldsymbol{c}}_1(i))$ based on the model selection criterion. With $\hat{\boldsymbol{c}}_1 = \hat{\boldsymbol{c}}_1(\hat{i})$, the refined fitted curves are

$$\hat{\alpha}(t) = \sum_{m=1}^{D_n} \hat{a}(\hat{i})_m v_m(t),$$

$$\hat{\beta}(t) = \sum_{m=1}^{d_n} \hat{b}(\hat{i})_m w_m(t).$$

2.3 Asymptotics

In this section, we show that our proposed estimator of the diarrhea effect function is consistent. Since the fitted curves are approximated by b-spline functions, the consistency of fitted curves depends on the asymptotic behaviors of the estimated b-spline coefficients, which is closely related to the asymptotic properties of Dantzig selector and SCAD in a regression setting. Since the asymptotic properties of Dantzig selector and SCAD in GEE have been studied (Gai et al., 2013; Zhou et al., 2013), we can extend those results for our model. There are two major dif-

ferences when we extend their results to our case. First, our data is longitudinal data with correlated responses on the subject level, whereas the previous proof of Dantzig selector is for independent cases. Secondly, there are approximation errors due to b-splines approximation as shown below:

$$Y_k(t_{kj}) = \sum_{m=1}^{D_n} a_m v_m(t_{kj}) + \sum_i \sum_{m=1}^{d_n} b_m w_m(t_{kj} - T_{ki}) + \widetilde{\epsilon}_{kj}$$

$$= \sum_{m=1}^{D_n} a_m v_m(t_{kj}) + \sum_{m=1}^{d_n} b_m \sum_i w_m(t_{kj} - T_{ki}) + \widetilde{\epsilon}_{kj},$$

where $\widetilde{\epsilon}_{kj} = \epsilon_{kj} + e_{kj}$, and e_{kj} is the approximation error for B-splines. Due to the B-spline approximation properties, $\widetilde{\epsilon}_{kj}$ doesn't follow a normal distribution with 0 mean and constant variance as in the linear regression settings. Our goal is to extend the established asymptotic results to incorporate approximation errors and with-subject correlation for longitudinal data.

We need the following conditions in Gai et al. (2013) for the proof of the consistency of the Dantzig estimate. Let $C = \frac{k_{0,n}}{n}(\boldsymbol{X})^T\boldsymbol{X}$, where $\boldsymbol{X}$ is the covariates generated from B-spline basis functions with $k_{0,n}$ inner knots and order h in the Dantzig step. For any subset $T \subset \{1, 2, ..., k_{0,n} + h\}$, $|T|$ denotes the number of elements in T, $\bar{T}$ is the complement of T in the set $\{1, 2, .., (k_{0,n}+h)\}$. Let $b_T = (b_j)_{j\in T}$ be the $|T|$ by 1 vector whose entries are those of b indexed by T. Similarly, $\boldsymbol{X}_T$ is defined as the n by $|T|$ matrix whose columns are those of $\boldsymbol{X}$ indexed by T. Given a $(k_{0,n} + h)$ by $(k_{0,n} + h)$ matrix C and subsets $T_1, T_2 \subset \{1, 2, ..., (k_{0,n} + h)\}$, let C_{T_1,T_2} be the $|T_1|$ by $|T_2|$ sub-matrix from C with rows corresponding to T_1 and columns corresponding to T_2. We also denote $T^* = \{j : \beta_j \neq 0\}$ and $q = |T^*|$. For longitudinal data, since $\boldsymbol{\epsilon}$ follows $N(\mathbf{0}, \boldsymbol{W})$, on the subject level $\boldsymbol{\epsilon}_i$ follows $N(\mathbf{0}, \boldsymbol{W}_i)$ and we define $M := \frac{1}{n}X^TWX = \frac{1}{n}\sum X_i^T W_i X_i$.

Let κ_{n1} be the largest eigenvalue of B, $B = k_{0,n}C_{E,T^*}^{-1}M_{E,E}C_{T^*,E}^{-1}$, with E satisfying the Irrepresentable Conditions proposed in Gai et al. (2013), and let τ_{n1} be the largest eigenvalue of the semi-positive definite matrix $(I-K)(I-K)^T$ with idempotent $K = \boldsymbol{X}_{T^*}(\boldsymbol{X}_E^T\boldsymbol{X}_{T^*})^{-1}\boldsymbol{X}_E^T W$, where $\boldsymbol{W}$ is the true covariance matrix.

Assume that there exist $0 \leq d_1 < d_2 \leq 1$ such that

(C1) $k_{0,n} = O(n^{d_1})$

(C2) $n^{(1-d_2)/2} min_{i\in T^*}|b_{0,i}| \geq M_1 > 0$,

(C3) $||C_{E,T^*}\alpha||_2^2 \geq M_2 > 0$ for any unit vector α, with E satisfying Irrepresentable Conditions,

(C4) $0 < \kappa_{n1} \leq \kappa_1 < \infty$,

(C5) $\tau_{n1} \leq \tau_1 < \infty$.

C1-C5 are necessary to prove the consistency of Dantzig estimator.

We also need the conditions from Zhou et al. (2013).

(AA1) $\beta(t)$ has rth ($r \geq 3$) bounded derivative on $[0, T]$

(AA2) For $k_{0,n}$, $n^{-1}k_{0,n}^{2r} \to 0$ and $n^{-1}k_{0,n}^{2r+2} \to \infty$.

(AA3) For threshold value d_n, $d_n n^{1/2}k_{0,n}^{-1} \to 0$ and $d_n k_{0,n}^{r-2} \to 0$.

In addition, the following conditions in Wang et al. (2012) are necessary for Theorem 2.

(A1) There exist two positive constants b_3 and b_4 such that

$$b_3 \leq \lambda_{min}(\frac{k_{1,n}}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{X}_i) \leq \lambda_{max}(\frac{k_{1,n}}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{X}_i) \leq b_4.$$

(A2) The common true correlation matrix $\boldsymbol{R}_0$ has eigenvalues bounded away from zero and $+\infty$. The estimated working correlation matrix $\hat{\boldsymbol{R}}$ satisfies $||\hat{\boldsymbol{R}}^{-1} - \bar{\boldsymbol{R}}^{-1}|| = O(\sqrt{k_{1,n}/n})$, where $\bar{\boldsymbol{R}}$ is a constant positive-definite matrix with eigenvalues bounded away from zero to $+\infty$. We do not require $\bar{\boldsymbol{R}}$ to be the true correlation

matrix $\boldsymbol{R}_0$.

(A3) The pdfs for t_{kj} and $t_{kj} - T_{ki}$ are bounded away from 0 and $+\infty$.

(A4) $\frac{k_{1,n}^2}{n} \to 0$, $\frac{n}{k_{1,n}^{r+1}} \to 0$, $\lambda_n \to 0$, $\frac{k_{1,n}}{n\lambda_n^2} \to 0$ and $\frac{k_{1,n} log(n)}{n\lambda_n} \to 0$.

A1-A2 are the assumptions from Wang (2011), which are needed to prove the consistency of the GEE estimator without the approximation error. A3 and A4 are additional assumptions to complete the proof when there are approximation errors. A5 suggests the approximation error goes to 0 as $k_{1,n}$, the number of knots, goes to infinity. A4 describes the convergence rates among the sample size, number of knots and the tuning parameter, suggesting that as n goes to infinity, $k_{1,n}$ and $1/\lambda$ also go to infinity but with a rate not too fast or too slow.

Theorem 1. *Let* $\tilde{\boldsymbol{b}}_0(n) = (\tilde{b}_{0,1}(n), ..., \tilde{b}_{0,k_{0,n}+h}(n))^T$ *be the post-selection estimate of* $\boldsymbol{b}_0(n)$ *from the Dantzig step, with* $\boldsymbol{\epsilon_k} \sim N(0, \Sigma_k)$. *If C1-C5 in Gai et al. (2013) hold, AA1-AA3 hold,* $k_{0,n} = o(n^{(d_2-d_1)k_0})$ *for some* $d_2 > d_1$, *and* λ *satisfies* $\lambda/\sqrt{n} = o(n^{(d_2-d_1)k_0})$, *and* $(\lambda/\sqrt{n})^{2k_0}/k_{0,n} \to \infty$, *we have*

(1) $||\tilde{\boldsymbol{b}}_0(n) - \boldsymbol{b}_0(n)||_{l2} = O_p(n^{-1/2}k_{0,n})$.

(2) $sup|\tilde{b}_{0,j}(n)| = O_p(n^{-1/2}k_{0,n})$ *for* $b_{0,j}(n)$ *associated with true null region denoted as* $\mathcal{T}$.

(3) With probability tending to 1,

$$\mathcal{T} \subseteq \hat{\mathcal{T}}^{(0)} \text{ and } \hat{\mathcal{T}}^{(0)} \cap \mathcal{T}^c \subseteq \Omega(k_{0,n}),$$

where, with $r \geq 3$ *as in the condition AA1,* $\Omega(k_{0,n}) = \{t \in [0,T] : 0 < |\beta(t)| < k_{0,n}^{-r+2}\}$ *is a sub-region of* $[0,T]$, *converging to the empty set as* $n \to \infty$

Theorem 2. *Assuming the conditions* $A_1 - A_4$ *and an initial estimator with the rate* $||\tilde{\boldsymbol{b}}_1(n) - \boldsymbol{b}_1(n)||_{l2} = O_p(n^{-1/2}k_{1,n})$, *with* $e_i \sim N(0,1)$ *and* $\mu = 0$.

(1) For $t \in \mathcal{T}$, *we have* $\hat{\beta}(t) = 0$, *with probability tending to 1.*

(2) For $t \in \mathcal{T}^c$, we have $|\hat{\beta}(t) - \beta(t)| = O(n^{-1/2}k_{1,n}^{3/2})$.

Theorem 1 shows that the estimator from step 1 is consistent and that the estimated null region covers the true null region. The proof of Theorem 1 is based on Theorem 1 in Zhou et al. (2013), combined with our modified proof of the consistency of Dantzig selector for longitudinal data based on Gai et al. (2013) in Chapter 5. Theorem 2 shows the consistency for the estimated curve in the refinement step. The proof in Chapter 5 is an extension of the consistency for PGEE estimator (Wang et al., 2012).

2.4 Simulations

In this section, we evaluate the numerical performance of our proposed method in simulation studies. The sample sizes are set to be 100 and 500. For each subject, growth measures are collected around every 3 months. The number of total growth measures collected after birth ranges from 2 to 17 with equal probabilities. The histogram of the number of diarrhea episodes for each subject is shown in Figure 2.5, with diarrhea date uniformly distribution over the 4 years after birth. The growth outcome for the kth subject with diarrhea effect is modeled as follows:

$$Y_k(t_{kj}) = \alpha(t_{kj}) + \sum_i \beta(t_{kj} - T_{ki}) + \epsilon_{kj},$$

where $Y_k(t_{kj})$ is the growth outcome observed at time t_{kj}. The curve $\alpha(t)$ is the natural growth curve without diarrhea effect, which we set as the estimated natural growth curve from the real data (Figure 2.6). Diarrhea happens at time T_{ki} : $T_{k1}, T_{k2}, ...,$. The diarrhea effect curve $\beta(t)$ (Figure 2.7) is set as follows:

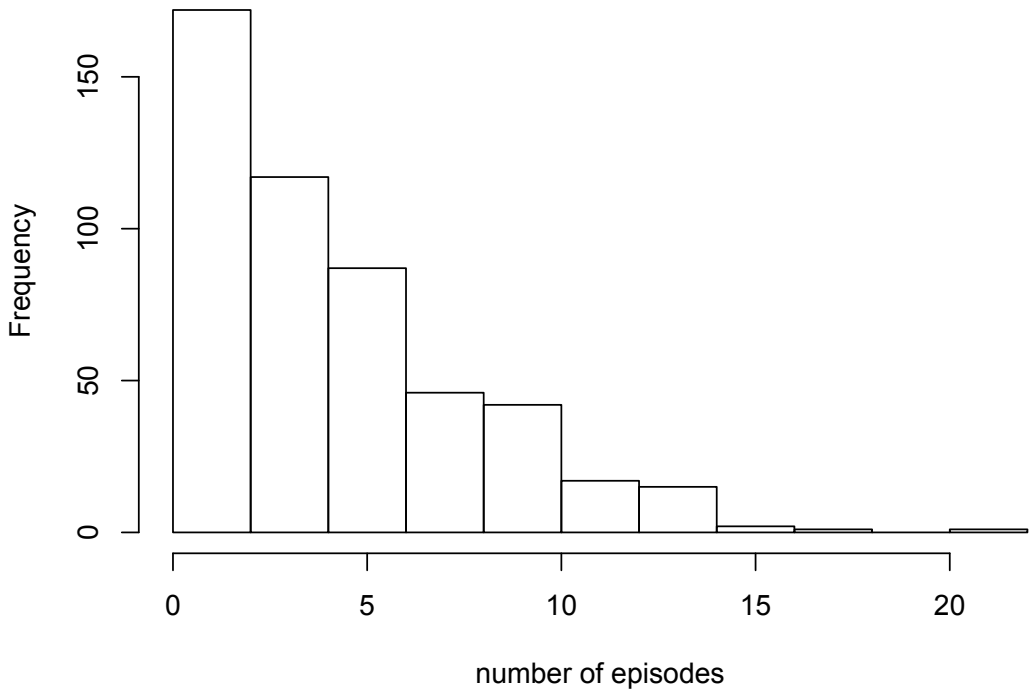

Figure 2.5: Histogram of number of diarrhea episodes for each child.

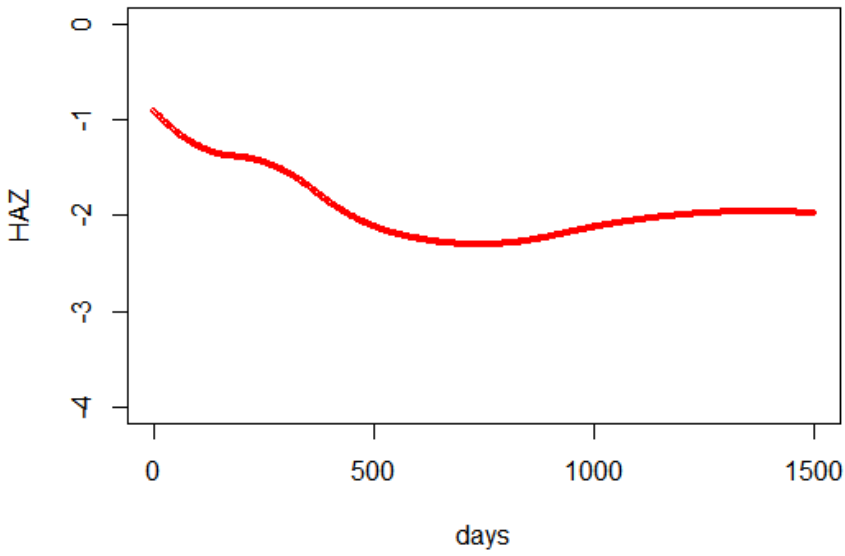

Figure 2.6: Natural growth curve $\alpha(t)$

$$\beta(t) = \begin{cases} 0.2(x/200 + 0.423)(x/200 - 0.577)(x/200 - 1.577) - 0.077 & 0 \leq x < 230.8 \\ 0.2((461.6 - x)/200 + 0.423)((461.6 - x)/200 - 0.577) & \\ \times((461.6 - x)/200 - 1.577) - 0.077 & 230.8 \leq x < 461.6 \\ 0 & 461.6 \leq x. \end{cases} \tag{2.1}$$

We specify that $\boldsymbol{\epsilon_k} \sim N(0, \Sigma_k)$ has a compound symmetry covariance structure with $\rho = 0.3$ and the marginal standard deviation σ.

For each generated dataset, we adopt the proposed two-step approach for the analysis. For the Dantzig step, we use 7 B-spline basis functions (4 inner knots with order 3) for the estimation of $\alpha(t)$ and 51 basis functions (48 inner knots with

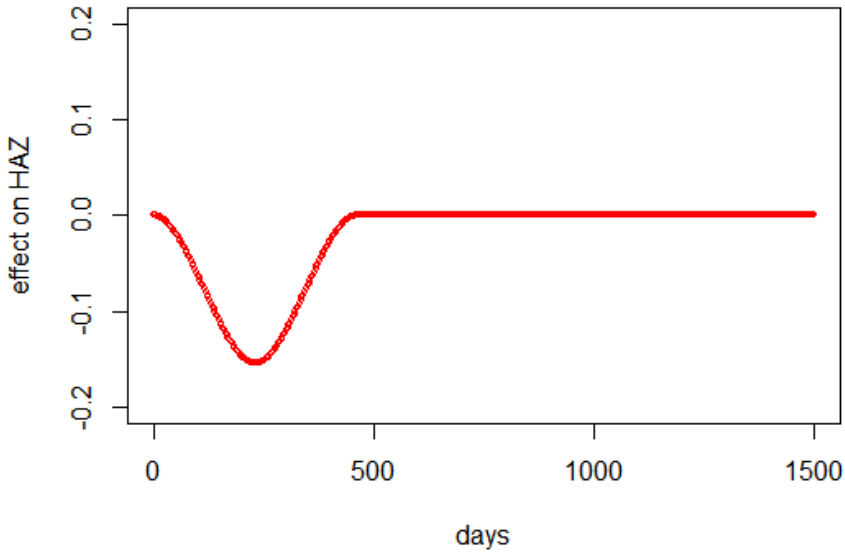

Figure 2.7: Diarrhea effect curve $\beta(t)$

order 3) for $\beta(t)$. The threshold d is set to be 0.05, according to the suggested order $O(n^{-0.25})$ in Zhou et al. (2013). Since Dantzig step is only to have an initial estimate of the non-zero region which would later be refined, a larger number of knots and larger threshold could be used and wouldn't affect the final results much after refinement.

During the PGEE step, we start with the boundary T from Dantzig step and refine the null region iteratively with step size equals to 3 weeks for computational efficiency. For each working boundary, different number of inner knots (2, 3, 4 respectively) are placed onto the working non-zero region to capture the trend while avoid overfitting and PGEE is applied for the estimation. AIC and BIC criterion for longitudinal data (Jones, 2011) are used for selection of best model with optimal working boundary and for the selection of tuning parameter and number of inner knots.

Table 2.2: Simulation results

Setting	Sample size	Criterion	σ	Boundary: Bias (SE)	MISE (SE)	failure rate
Setting 1	100	AIC	0.1	36.53 (26.55)	3.06 (1.16)	1%
Setting 2	100	BIC	0.1	38.55 (25.07)	3.03 (1.18)	1%
Setting 3	500	AIC	0.1	35.04 (9.29)	1.85 (0.51)	0%
Setting 4	100	AIC	0.3	62.71 (119.09)	7.51 (3.10)	18%

We try 4 simulation settings to evaluate the performance of our proposed estimator with different sample sizes, standard deviations of random error and model selection criteria. We repeat the simulation 200 times and obtain the fitted diarrhea effect curves and the boundaries between non-null region and null region. To evaluate the performance, we report the bias and standard error of the estimated boundary, and the mean and standard error of the Integrated Standard Error (ISE) of the estimated effect function $\beta(t)$, which is defined as $\mathrm{E}\int(\hat{\beta}(t)-\beta(t))^2dt$. We also report the percentage, out of the 200 generated data sets, we fail to identify the null region during the Dantzig step.

The simulation results are shown in Table 2.2. The true non-null region is $[0, 461.6]$ days. The results from setting 1 suggest that our method is able to identify the boundary (462 days) with a bias of roughly a month. Given that our step size is 3 weeks, this estimated boundary is rather close to the true value. Some of the fitted curves are shown in Figure 2.8, suggesting the estimated curves are close to the true curve and can capture the pattern of the diarrhea effect. If we use BIC instead of AIC for model selection as shown in setting 2, the results are very close, suggesting the robustness of our developed estimator to model selection criteria. We use AIC for other settings. Settings 3 and 4 show that larger sample size or smaller random error lead to better estimation with smaller bias and SE for both the estimated boundary and ISE, and smaller failure rate during the Dantzig step.

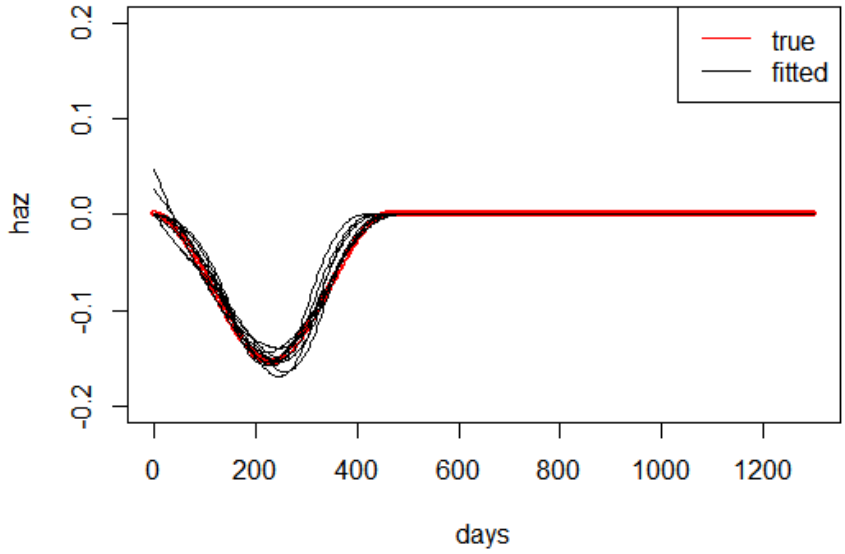

Figure 2.8: $\hat{\beta}(t)$ vs. $\beta(t)$

Chapter 3

Extended models

The model we propose in section 2 assume that the growth curve of each child is only affected by diarrhea and that those diarrhea episodes have the same pattern of effect. In reality, many other factors could affect growth or the diarrhea process. Even though it is reasonable to assume those other factors can be averaged out when there are enough samples and to model the effect of diarrhea marginally, we can easily extend the proposed model in Chapter 2 to take into account other factors of our interests. In this section, we present four extended models to show how our model can be applied in different scenarios.

3.1 With additional covariates

The model proposed in Chapter 2, $Y_k(t_{kj}) = \alpha(t_{kj}) + \sum_i \beta(t_{kj} - T_{ki}) + \epsilon_{kj}$, has several assumptions. One of the assumptions is that every child has the same $\alpha(t)$ which is the baseline growth curve without diarrhea, and other covariates associated with growth outcome such as social economic status, nutritional biomarkers, etc, are not considered in the modeling process. Based on the available data we have, we can

improve the efficiency of the estimation by including these additional covariates in the model as shown below:

$$Y_k(t_{kj}) = \boldsymbol{X}_k(t_{kj})\boldsymbol{u} + \alpha(t_{kj}) + \sum_i \beta(t_{kj} - T_{ki}) + \epsilon_{kj},$$

where $\boldsymbol{X}_k(t)$ is time-varying (or time independent) covariate and $\boldsymbol{u}$ is the coefficient of the corresponding effect. The proposed methods and estimation procedure still applies in this scenario. We can approximate the curves of interest in the same way:

$$\alpha(t) = \sum_{m=1}^{D_n} a_m v_m(t) + e^a(t),$$

$$\beta(t) = \sum_{m=1}^{d_n} b_m w_m(t) + e^b(t).$$

The model then becomes

$$Y_k(t_{kj}) = \sum_{m=1}^{D_n} a_m v_m(t_{kj}) + \sum_i \sum_{m=1}^{d_n} b_m w_m(t_{kj} - T_{ki}) + \widetilde{\epsilon_{kj}}$$

$$= \sum_{m=1}^{D_n} a_m v_m(t_{kj}) + \sum_{m=1}^{d_n} b_m \sum_i w_m(t_{kj} - T_{ki}) + \widetilde{\epsilon}_{kj},$$

where $\widetilde{\epsilon}_{kj} = \epsilon_{kj} + e_{kj}$, ϵ_{kj} is the random error and the approximation error $e_{kj} = e^a_{kj} + \sum_i e^b_{kji}$ is the summation of all the approximation errors from each curve. It can be written as $\boldsymbol{Y} = (\boldsymbol{X}^a, \boldsymbol{X}^b, \boldsymbol{X})^T\boldsymbol{c} + \widetilde{\boldsymbol{\epsilon}} = \boldsymbol{X}^T\boldsymbol{c} + \widetilde{\boldsymbol{\epsilon}}$, $\boldsymbol{c} = (a_1, ..., a_{D_n}, b_1, ..., b_{d_n}, u_1, ..., u_p) = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{u})$, where $\boldsymbol{a}$ is a D_n dimensional vector, $\boldsymbol{b}$ is a d_n dimensional vector and $\boldsymbol{u}$'s dimension is p which is the number of additional covariates.

Since we have control over whether to impose penalty on a specific coefficient or not, there are two options to estimate the additional parameters $\boldsymbol{u}$. If our goal is to estimate the curves with those additional covariates accounted for, we can only

penalize on $\boldsymbol{b}$ with no penalty imposed on $\boldsymbol{u}$. Or if our goal is to estimate the curves and identify the covariates that have significant impact on childhood growth, we can add penalty on both $\boldsymbol{b}$ and $\boldsymbol{u}$. Either way, the estimating procedure remain the same.

3.2 Leveling off to a constant

Besides including more covariates, we can also relax some of the assumptions we made about the effect curve $\beta(t)$ in Chapter 2. In the data analysis using the original model, we manage to identify the window of the diarrhea effect, after which the effect levels off to 0 suggesting there's no permanent effect. However, for other types of effect which could have a permanent effect, in stead of assuming it would level off the 0, we assume in general that it could level off to a constant. Our model can also be extended in this scenarios to identify the constant region (to estimate the window of temporary effect) and fit the curve on non-constant region at the same time.

The model remains the same:

$$Y_k(t_{kj}) = \alpha(t_{kj}) + \sum_i \beta(t_{kj} - T_{ki}) + \epsilon_{kj} = \sum_{m=1}^{D_n} a_m v_m(t_{kj}) + \sum_i \sum_{m=1}^{d_n} b_m w_m(t_{kj} - T_{ki}) + \widetilde{\epsilon_{kj}}$$

$$= \sum_{m=1}^{D_n} a_m v_m(t_{kj}) + \sum_{m=1}^{d_n} b_m \sum_i w_m(t_{kj} - T_{ki}) + \widetilde{\epsilon}_{kj},$$

where $\widetilde{\epsilon}_{kj} = \epsilon_{kj} + e_{kj}$, and e_{kj} is the approximation error for B-splines.

The difference is that instead of estimating b_is with penalty to promote $b_l = b_{l+1} = ... = b_{d_n} = 0$ for some l which would leads to a zero-value region at the end, we impose a different penalty to promote $b_l = b_{l+1} = ... = b_{d_n}$, which would lead to

a constant-value region. To achieve the goal, penalties cannot be imposed on those b_is directly. Instead, noticing that $\sum_{m=1}^{d_n} w_m(t) = 1$ for any t, which is the property of B-spline basis functions, we have

$$\beta(t) = \sum_{m=1}^{d_n} b_m w_m(t) + e(t) = b_{d_n} + \sum_{m=1}^{d_n-1} (b_m - b_{d_n}) w_m(t) + e(t).$$

Letting $h_0 = b_{d_n}, h_i = b_i - b_{d_n}$ for $i = 1, ..., d_n - 1$, we have

$$\beta(t) = h_0 + \sum_{m=1}^{d_n-1} h_m w_m(t) + e^b(t)$$

which becomes a linear model with intercept.

After the change of variables, identifying constant value region is equivalent to finding $h_l = h_{l+1} = ... = h_{d_n-1} = 0$ for some l, which becomes a variable selection problem that is essential the same as the original model. Traditionally, B-spline approximation requires the use of all d_n basis functions. In this case, by using only $d_n - 1$ basis functions with an intercept, we are able to make sure the fitted curve levels off the a constant if there's a permanent impact and the fitted h_0 (the intercept) is the constant level it levels off to.

3.3 Different types of effect curves

The original model also assumes that all the diarrhea episodes have the same effect function $\beta(t)$, which can also be relaxed in our framework. In practice, we are interested in modeling more than one type of effect curves at the same time, either because they are of different diseases, or because, in this case, they are different types of diarrhea. Our model can be easily extended for multiple curve estimation so that we can not only differentiate different types of effect but also estimate them

simultaneously.

In addition to the single effect curve $\beta_1(t)$, we can add more effect curves $\beta_2(t), \beta_3(t), ...$ in the same way, shown as below:

$$Y_k(t_{kj}) = \alpha(t_{kj}) + \sum_{d=1}^{D} \sum_{i} \beta_d(t_{kj} - T_{kid}) + \epsilon_{kj},$$

where $T_{kd1}, T_{kd2}, ..., T_{kdi}$ are the onsite dates of type d for subject k. The estimation for this model is the same as our proposed model, after we use additional sets of basis functions to estimate different types of $\beta_d(t)$ as shown below,

$$\alpha(t) = \sum_{m=1}^{D_n} a_m v_m(t) + e^a(t),$$

$$\beta_d(t) = \sum_{m=1}^{d_n} b_m^d w_m^d(t) + e_d^b(t), d = 1, ..., D$$

Accordingly, instead of estimating the coefficient vector $\boldsymbol{c} = (\boldsymbol{a}, \boldsymbol{b})$ in the original model, we estimate the coefficients of different set of basis functions $\boldsymbol{c} = (\boldsymbol{a}, \boldsymbol{b}^1, \boldsymbol{b}^2, ..., \boldsymbol{b}^D)$. Since we approximate different types of curve separately, we can address different question of interests for different curves. This extended model is useful when multiple biological processes are involved, with each effect being a curve starting at different time points.

3.4 Two dimensional effect

Another assumption we made about the diarrhea effect curve $\beta(t)$ is that it is one dimensional, which means it only depends on time t. It is possible that the curve of our interests depends on more than 1 variable, for example, in the form of $\beta(t, s)$, where t could be the diarrhea onset time, and s could be the time elapsed after

the onset time. In this case, it is reasonable to believe that as children grow older, their ability to recover from the diarrhea can become stronger, which is the reason that researchers have been focusing on growth shortfall especially in early childhood. Therefore the age of the children should also be included in modeling the diarrhea effects.

We use a 2-dimensional function $\beta(t, s)$, $t > 0$ and $s > 0$, to model the effect of the diarrhea lasting s time after the onset time t. This matrix denoting the dynamic diarrhea effects is proposed to be estimated in the model $y_i(t_{ij}) = \alpha(t_{ij}) + \sum_{k=1}^{J} \beta(T_{ik}, t_{ij} - T_{ik}) + \epsilon_{ij}$, where $y_i(t_{ij})$ is the growth measurement, such as height, on child i taken at age t_{ij}, $\alpha(t_{ij})$ is the mean growth function of the study cohort, T_{ik} with $k = 1, 2, ..., K$ are the onset times of diarrhea episodes happened before time t_{ij} on child i, and ϵ_{ij} are random errors. Here, we assume that episodes of diarrhea have additive effects on the growth shortfalls. The estimation of this model will not only show us the pattern of diarrhea effect on childhood growth but also tell us how this pattern varies based on the age of the children when they develop diarrhea. Understanding the age range in which diarrhea affects children's growth most would further help us make better health policies, conduct more timely interventions, and improve the health of children globally.

The function $\beta(t, s)$ is a general form indicating that the effect curve depends on both the age and the length of time since diarrhea. However, in application, we can assume the function to have specific form to facilitate the estimation and to incorporate our prior understanding of the diarrhea process. We illustrate how to incorporate both time variables in the following example.

We assume $\beta(t, s) = b(t) * \tilde{\beta}(s)$, where $\tilde{\beta}(s)$ is the diarrhea effect s days after the onset of diarrhea, the shape of which remains the same for any age. However, the scale of the effect depends on the age of the subject when the diarrhea is onset.

Theoretically, we don't have to assume $\beta(t, s)$ to have any specific form by using two-dimensional splines for approximation, which certainly can be a future research direction. Here, we would start from a simpler model by assuming $\beta(t, s) = b(t) * \tilde{\beta}(s)$ to reduce the model complexity and increase the interpretability. By separating the scale function $b(t)$ and the shape function $\tilde{\beta}(s)$, we are able to study the pattern of the diarrhea effect, and investigate how the age factor fit into our study goals, for example, if as children grow older, their body responds better to the diarrhea and experience less effect.

The estimating procedure remains the same with the only difference being using additional set of b-spline functions to approximate $b(t)$.

To summarize, our original model is essentially a regression model for curve estimation, which can be extended the same way other regression model or curve estimation methods have been extended. We demonstrate four directions that are closely related to our question of interests in the NIH cohort study. Other extensions such as generalized linear model, mixed effects model and monotone function estimation can all be easily incorporated in our framework for other types of datasets or other question of interests.

Chapter 4

Real data analysis

The data we use in this thesis comes from the NIH study. During 2008-2012, a cohort of total 629 infants (332 boys and 297 girls) were enrolled into the study after birth - in Dhaka, Bangladesh. The growth outcomes of each child were recorded every 3 months until the end of the study, resulting in 6831 observations. Measures of growth include weight, height, WHZ (weight-for-height z-score), HAZ (height-for-age z-score), WAZ (weight-for-age z-score), BAZ (BMI-for-age z-score), which are scores set by WHO based on children's growth worldwide. Children's health condition was monitored every two weeks by visits of a research staff. During the visits, questionnaire and follow-ups were given to record children's health condition. If there was an acute illness, the child would be sent to the study clinic for further evaluation. When a child had diarrhea symptoms, stool samples were taken to further decide if it was diarrhea or not and if there are any pathogens present in the stool sample. The information of the starting date of each diarrhea episode was also provided by the questionnaire. In our study, 521 out of 629 children had diarrhea of totally 2605 episodes. For each episodes of diarrhea, we tested the presence of Crypto, EH and Giardia, which are 3 common pathogens that could

cause diarrhea. Out of all 2605 diarrhea episode samples, 197 of them contain Crypto, 243 of them contain EH, and 731 contain Giardia. Overall, majority of the samples (1601 samples) don't contain any of the three pathogens.

4.1 Original model fitting

To study the diarrhea effect on childhood growth, we choose HAZ as a better measure for overall growth because weight-associated score is highly sensible to diarrhea in short term. The reason we use HAZ instead of height (cm) is because as children grow older, their heights will increase and therefore leads to an increase of the variance of random errors. By using HAZ, we can adopt the constant variance assumption for the random errors. To apply the proposed model, we need to first specify the number of knots and the location of them. For estimating $\alpha(t)$, since this is mainly a baseline rather than the effect curve that we are mainly interests in, only 4 inner quantile knots are placed in both step 1 and 2. For estimating $\beta(t)$ in the Dantzig step, we place 49 evenly spaced inner knots over 4 years and obtain an estimate of the boundary around day 450. As explained before, the goal is to place as many knots as we can so we can narrow down the initial estimate of null region to begin with in the following step. Since the total length of the study is about 4 years, by breaking it down into 50 intervals with roughly 1 month each, we believe the number is a good balance between model complexity and estimation accuracy for the initial estimate of null region. We have tried increasing the number of knots and the results are consistent. Next, for the PGEE step, starting from the initial boundary we obtain in step 1, we move the boundary to the right every time with 7 days to refine the null region. For each fixed boundary, different number of knots are placed in the non-null region. We select the number of knots and the tuning

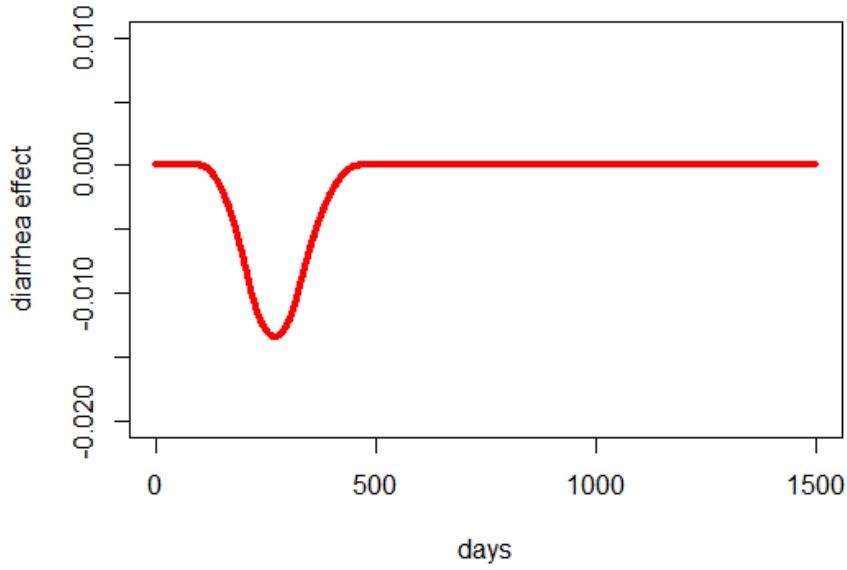

Figure 4.1: Fitted diarrhea effect curve $\beta(t)$

parameter at the same time based on AIC. The best model with 3 inner knots and a boundary of 464 (66.3 weeks) is selected with AIC=12560.31. The fitted $\alpha(t)$ and $\beta(t)$ are shown in Figure 2.6 and Figure 4.1.

The fitted natural growth curve shows that the growth of children in Bangladesh is below average in WHO standard. In addition, HAZ score keeps dropping over first two years after birth and then remain roughly the same level. The diarrhea effect curve shows that the HAZ score is not affected shortly after the diarrhea, but due the lack of nutrition, this diarrhea effect will start to show after 3 months and remain the negative effect until 15 months. The valley of the diarrhea effect is -0.013 around 9 months. After the catch-up growth starting from 9th month, children will return to the original level 15 months after the diarrhea. The above results provide a better understanding of the dynamic diarrhea effect on childhood growth along with the overall natural childhood growth in the study cohort.

These results confirm the association between diarrhea and childhood growth. Even though the effect will level off eventually due to the catch up growth, children can still experience growth shortfall for up to 15 months after a single episode of diarrhea. These can serve as the evidence to support the necessity of intervention for the welfare of children especially in the developing countries. Furthermore, the pattern estimated from our model could also give us a better understanding of how and when to intervene in order to prevent or reduce the negative effect of childhood growth due to diarrhea.

4.2 Extended model 1: additional covariates

The original model assumes the baseline ($\alpha(t)$) of growth for every child is the same. However, many factors can affect the childhood growth and certain groups of children do generally have better growth outcomes than other groups. For instance, we are all aware of the clear difference in growth patterns between boys and girls, which has been confirmed in some studies (Tumilowicz et al., 2015). Therefore, to illustrate how our model can incorporate additional covariates, we use gender as covariate in the following analysis.

The extended model can be expressed as

$$Y_k(t_{kj}) = gender * u + \alpha(t_{kj}) + \sum_i \beta(t_{kj} - T_{ki}) + \epsilon_{kj},$$

where u is the coefficient for gender (male=1, female=0).

Since our goal is to estimate the effect curves with the gender factor accounted for, we adopt the same estimating procedure with no penalty imposed on gender. The resulting estimated curves are shown in Figure 4.2 and Figure 4.3 and $\hat{u} =$

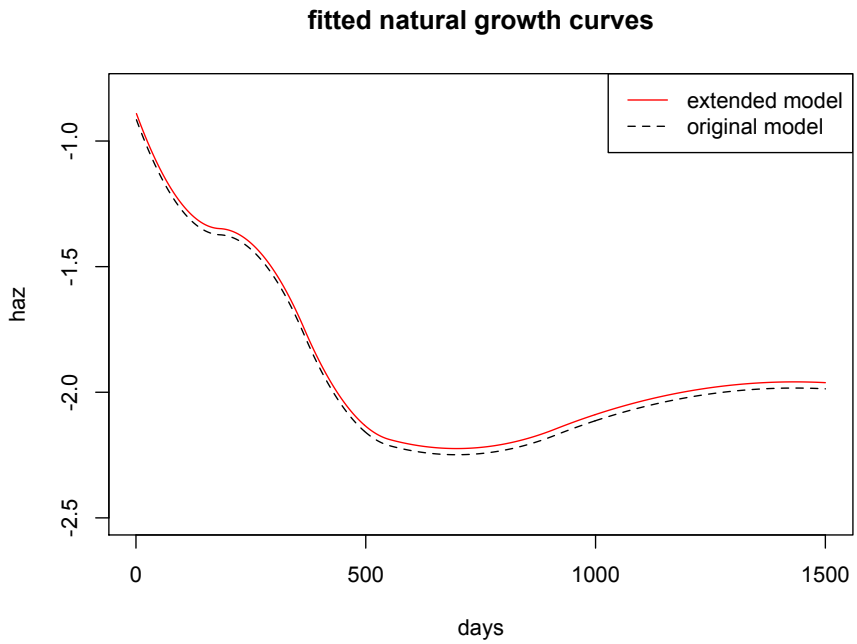

Figure 4.2: Fitted natural growth curves with extended model 1

-0.046. The fitted effect curves are very close, which suggests that incorporating gender doesn't affect the estimation of $\beta(t)$. However, boys have lower baseline growth than girls with a difference of 0.046 in HAZ, which can be seen in Figure 4.2 where the red curve (for girls) is slightly higher than the one from our original model where the estimated curve represent the average growth curve of all boys and girls.

4.3 Extended model 2: leveling off to a constant

In some cases, the curve to be estimated levels off to a constant that is different from 0. When estimating $\beta(t)$ with this property in the model $Y_k(t_{kj}) = \alpha(t_{kj}) + \sum_i \beta(t_{kj} - T_{ki}) + \epsilon_{kj}$, we include the intercept term for b-spline approximation. Since as shown above, there's no permanent effect of diarrhea based on the data.

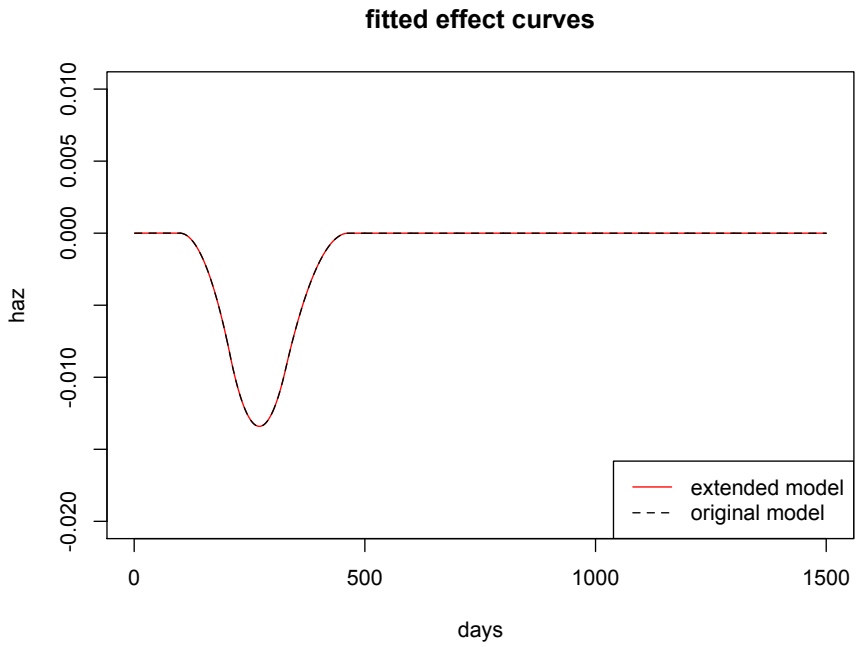

Figure 4.3: Fitted diarrhea effect curve with extended model 1

To illustrate how this extended model can be applied to identify nonzero constant-value region, we generate data based on a $\beta(t)$ with a permanent effect of -0.05. As shown in Figure 4.4, our fitted curve is flat after 300 days. In addition, the level it remains at is not 0, which suggests our extended model can shrink the curve to be flat, and detect a nonzero long term permanent effect. This model can be viewed as a generalization of our original model to estimate a long term constant effect.

4.4 Extended model 3: multiple curves

For our dataset, the question of interest here is whether the infection type of the diarrhea can be linked to the level of diarrhea effect. Crypto, EH and Giardia, which are 3 common pathogens that could cause diarrhea, were tested for the stool samples for each episode of diarrhea recored. We found one of the 3 pathogens in

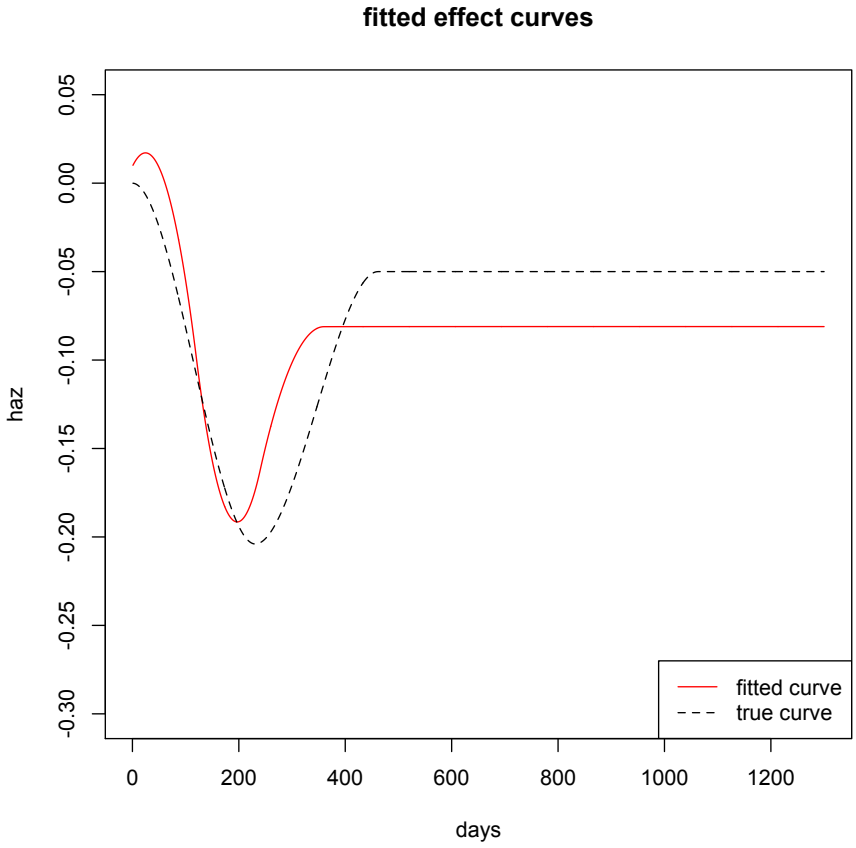

Figure 4.4: Fitted curve vs. true curve

1004 diarrhea episodes out of the total 2605 episodes of diarrhea recorded. To study whether the diarrhea with presence of one of those pathogens has greater impact on children's growth, we apply the above model with 2 types of effect curves: type 1, Crypto, EH and Giardia related; type 2, the rest,

$$Y_k(t_{k,j}) = \alpha(t_{k,j}) + \sum_i \beta_1(t_{k,j} - T_{k,i,1}) + \sum_i \beta_2(t_{k,j} - T_{k,i,2}) + \epsilon_{k,j}.$$

The effect of Crypto, EH and Giardia is denoted as $\beta_1(t)$ and $\beta_2(t)$ denotes the effect of diarrhea with no pathogen detected. Applying our estimation procedure, no signal can be detected for $\beta_2(t)$ ($\hat{\beta}_2(t) = 0$), which suggests that most of the growth shortfall is related to the diarrhea caused by pathogens (Crypto, EH or Giardia) and no growth shortfall is linked to the rest. Figure 4.5 and Figure 4.6 show the estimated baseline growth curve and the diarrhea effect curve for pathogen-related diarrhea. The shape is similar to that of the general diarrhea effect curve from our analysis above, but with a greater effect level (-0.023) and a shorter effect window (12 months). By differentiating different infection types, we are able to identify the pathogens associated with growth shortfall and estimate the corresponding effect curve more accurately, which could help to make better policies and conduct more efficient interventions.

4.5 Extended model 4: Two dimensional curves

Our next goal is to estimate the diarrhea effect curve by assuming the diarrhea effect curve is also dependent on the onset time of diarrhea. For example, diarrhea happens during first few months of life would have different effect from those in later years of life.

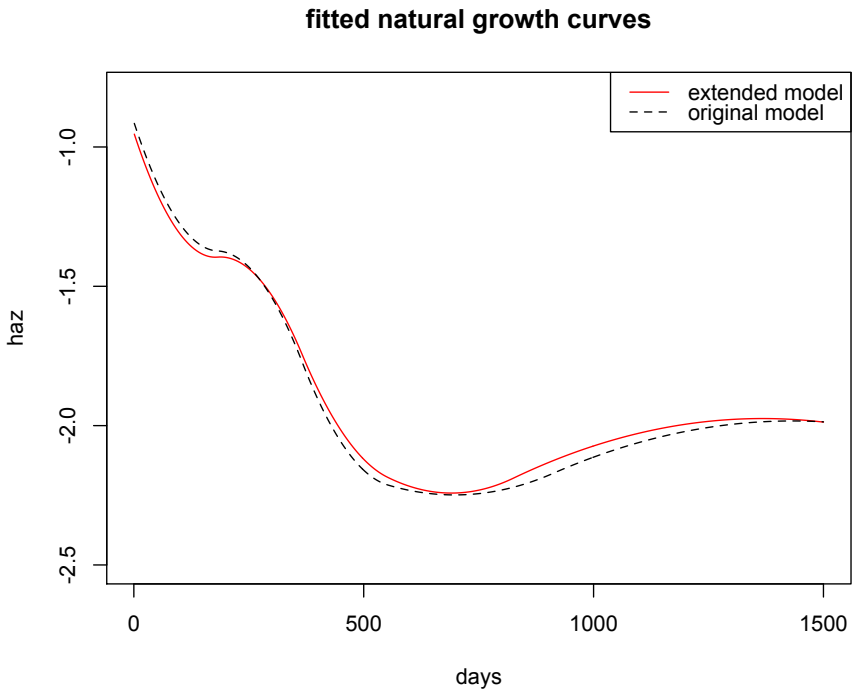

Figure 4.5: Fitted natural growth curves with extended model 3

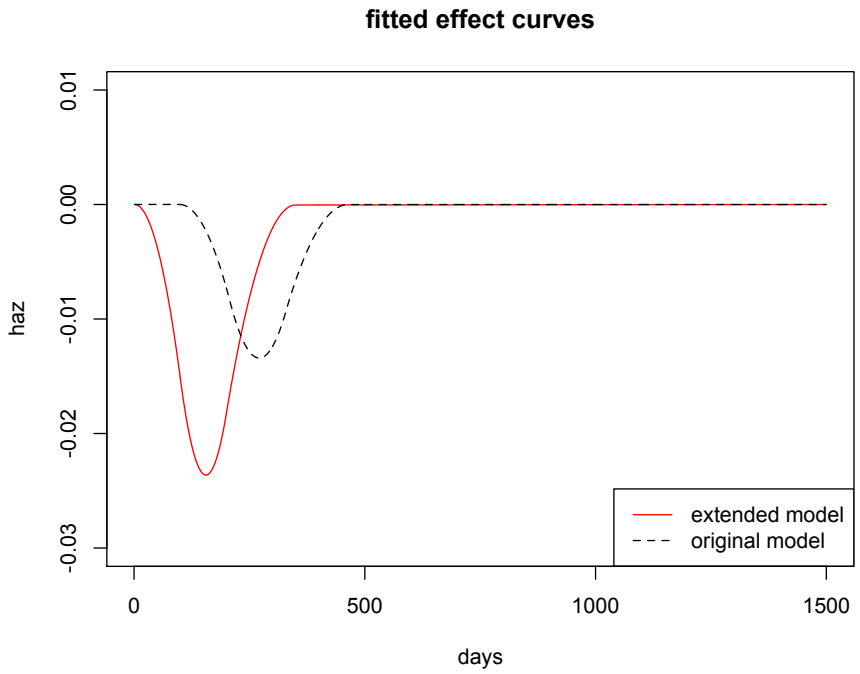

Figure 4.6: Fitted diarrhea effect curve with extended model 3

We use a 2-dimensional functional matrix $\beta(t, s)$, $t > 0$ and $s > 0$, to model the effect of the diarrhea lasting s time after the onset time t. We assume $\beta(t, s) = b(t) * \tilde{\beta}(s)$, where $\tilde{\beta}(s)$ is the diarrhea effect s days after the onset of diarrhea, the shape of which remains the same for any age. However, the scale of the effect depends on the age of the subject when having diarrhea, and is denoted by $b(t)$ where t refers to the onset age.

The estimated shape of the diarrhea effect and the scale function with onset time are shown in Figure 4.7 and Figure 4.8. The estimated shape function shows that the pattern of the diarrhea is similar to the one estimated from the marginal model in Chapter 2. However, the scale function provides additional information for understanding the impact of diarrhea on early childhood. The effect levels off after roughly the age of 700 days, suggesting that diarrhea mainly affect the growth

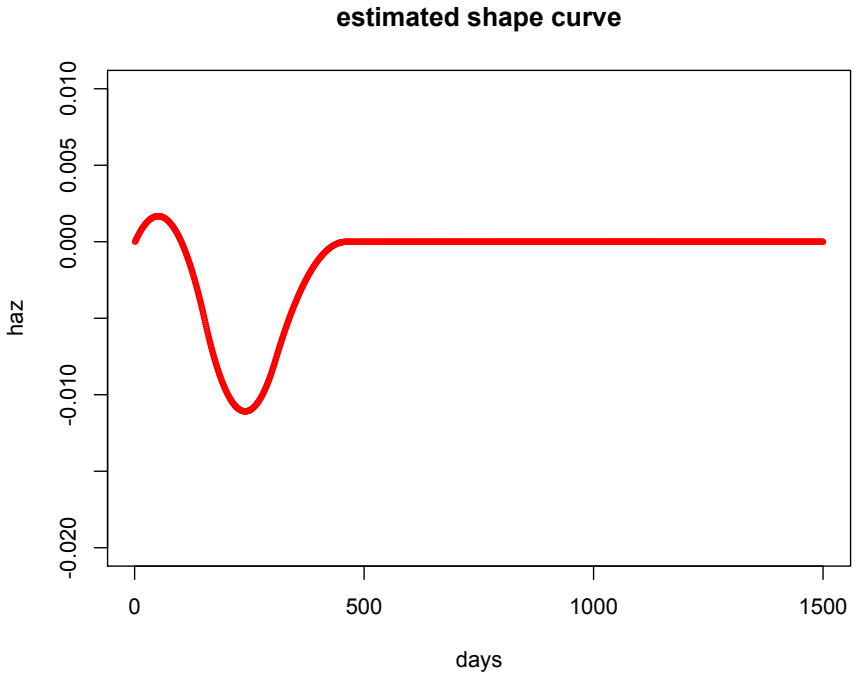

Figure 4.7: Estimated shape curve

of children under the age of 2. In addition, the effect is most significant to children around 8 months old. These preliminary results not only provide evidence on the association between diarrhea and growth shortfall in early childhood, but also give us a better understanding of the pattern of the effect and its change with age.

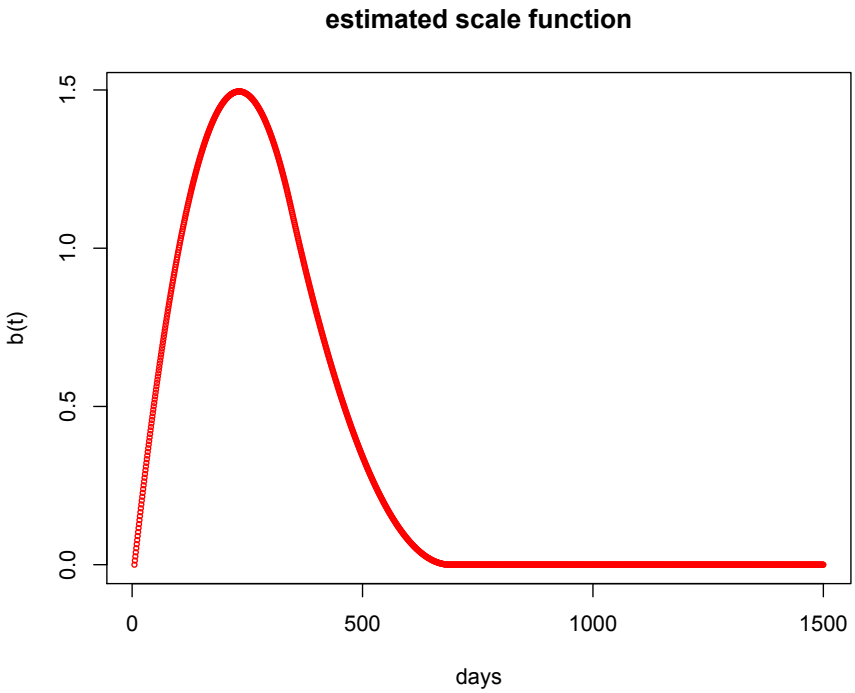

Figure 4.8: Estimated scale curve

Chapter 5

Detailed proofs

5.1 Proof of Theorem 1

To prove Theorem 1, it is sufficient to prove the consistency of Dantzig selector in our case (with approximation errors for longitudinal data). We modify the proves in Gai et al. (2013) to incorporate approximation errors and covariance matrix. As follows, Lemma 1 proves the sign consistency and Lemma 2 proves the post-selection consistency based on Lemma 1.

Lemma 1. (Sign consistency). Assume the $\widetilde{\epsilon}_{ij}$ are random variables with $E(\widetilde{\epsilon}_{ij})^{2k_0} < \infty$ for some interger $k_0 > 0$. If (C1)-(C3) in Gai et al. (2013) hold, $k_{0,n} = o(n^{(d_2-d_1)k_0})$ for $d_2 > d_1$, and λ satisfies $\lambda/\sqrt{n} = o(n^{(d_2-d_1)/2})$, and $(\lambda/\sqrt{n})^{2k_0}/k_{0,n} \to \infty$, then the Irrepresentable Conditions imply that

$$P(\tilde{\boldsymbol{b}}_0(\lambda) =_s \boldsymbol{b}_0) \geq 1 - O(\frac{k_{0,n}n^{k_0}}{\lambda^{2k_0}}) \to 1, n \to \infty.$$

To prove Lemma 1, we need to following lemmas.

Lemma A.1. For longitudinal data of m subjects, d_i repeated measurements

for the *ith* subject ($d_i < D$). Total number of observations $n = \sum_{i=1}^{m} d_i$. Let $\theta = (\boldsymbol{\theta_1}^T, ..., \boldsymbol{\theta_m}^T) = (\theta_{11}, ..., \theta_{1d_1}, ..., \theta_{md_m})$ be a vector of random variables with same marginal normal distribution (mean=0) and such that $E(\theta_{11})^{2k_0} < \infty$ for some integer $k_0 > 0$. Then, for constant vector α,

$$E(\alpha^T\theta)^{2k_0} \leq D^{2k_0}(2k_0 - 1)!!||\alpha||_2^{2k_0}E(\theta_{11})^{2k_0}.$$

Proof:

We know for normal distribution with zero mean, $E(\alpha^T\theta)^{2k_0} = (2k_0-1)!!var(\alpha^T\theta)^{k_0}$. We also have $var(\alpha^T\theta) = ||\alpha||_2^2E(\theta_{11})^2 + \sum cov(\alpha_i\theta_i, \alpha_j\theta_j)$. Since different subjects $\boldsymbol{\theta_i}$ are independent to each other, we only need to account for within subject correlation, which will lead to $var(\alpha^T\theta) \leq D^2||\alpha||_2^2E(\theta_{11})^2$. Therefore $E(\alpha^T\theta)^{2k_0} \leq D^{2k_0}(2k_0 - 1)!!||\alpha||_2^{2k_0}E(\theta_{11})^{2k_0}$.

Lemma A.2. In addition to Lemma A.1, with the approximation error $e = (\boldsymbol{e_1}^T, ..., \boldsymbol{e_m}^T) = (e_{11}, ..., e_{1d_1}, ..., e_{md_m})$ where $|e_{ij}| = O(k_{0,n}^{-r})$ and $k_{0,n}, n \to 0$,

$$E(\alpha^T(\theta + e))^{2k_0} \leq CD^{2k_0}(2k_0 - 1)!!||\alpha||_2^{2k_0}E(\theta_{11})^{2k_0}.$$

Proof: $E(\alpha^T(\theta + e))^{2k_0} = E(\alpha^T\theta)^{2k_0} + \sum_{i=1}^{2k_0}(\alpha^Te)^iE(\alpha^T\theta)^{2k_0-i}$. We know $|\alpha^Te| \to 0$ as $n, p \to 0$ and that $E(\alpha^T\theta)^{2k_0-i} = O(||\alpha||_2^{2k_0})$.

Proof for Lemma 1: The proof is similar to the proof of theorem 3 in Gai et al. (2013) using Lemma A.1 to substitute the original Lemma A.1.

Let $\zeta = (\zeta_1, ..., \zeta_{p-q})^T = C_{\bar{E},T^*}C_{E,T^*}^{-1}Z_E - Z_{\bar{E}}$, $\xi = (\xi_1, ..., \xi_q)^T = DC_{E,T^*}^{-1}Z_E$, and $h = (h_1, ..., h_n)^T = DC_{E,T^*}^{-1}sign(\tilde{\mu}_E)$, where $D = k_{0,n}diag(sign(\beta_{T^*}))$.

Follow the proof in (Gai et al.), $1 - P(\hat{\beta}^D =_s \beta) \leq \sum_{i=1}^{p-q} P(|\zeta_i| \geq \frac{\lambda}{\sqrt{n}}\eta_i) + \sum_{j=1}^{q} P(|\xi_j| \geq \sqrt{n}(|\beta_j| - \frac{\lambda}{n}h_j))$

Since $\boldsymbol{\epsilon}$ follows $N(\mathbf{0}, \boldsymbol{W})$, on the subject level $\boldsymbol{\epsilon}_i$ follows $N(\mathbf{0}, \boldsymbol{W}_i)$ and we can prove $M := \frac{1}{n}X^T W X = \frac{1}{n}\sum X_i^T W_i X_i \to M^*$ converges to a matrix M^*.

Therefore $\zeta = G^T\widetilde{\epsilon} = G^T(\epsilon + e)$. $G^T G = \frac{1}{n}(C_{\bar{E},\bar{T}^*}C_{E,T^*}^{-1}\boldsymbol{X}_E^T - \boldsymbol{X}_{\bar{E}}^T)W(\boldsymbol{X}_E C_{T^*,E}^{-1}C_{\bar{T}^*,\bar{E}} - \boldsymbol{X}_{\bar{E}}) = C_{\bar{E},\bar{T}^*}^*(C^*)_{E,T^*}^{-1}M_{E,E}(C^*)_{T^*,E}^{-1}C_{T^*,\bar{E}}^* - C_{\bar{E},\bar{T}^*}^*(C^*)_{E,T^*}^{-1}M_{E,\bar{E}} - M_{\bar{E},E}(C^*)_{T^*,E}^{-1}C_{T^*,\bar{E}}^* + M_{\bar{E},\bar{E}} = \frac{1}{n}\boldsymbol{X}_{\bar{E}}^T(I-K)(I-K)^T\boldsymbol{X}_{\bar{E}}$, where $K = \boldsymbol{X}_{T^*}(\boldsymbol{X}_E^T\boldsymbol{X}_{T^*})\boldsymbol{X}_E^T W$.

Therefore, by (C5), we have $||G_i||_2^2 \leq \tau_1 < \infty$. Lemma A.1 implies

$$E(\zeta_i)^{2k_0} < \infty$$

. Similarly, we can show $E(\xi_j)^{2k_0} < \infty$, which is the results of A.20 in Gai et al. (2013). Therefore the rests of the proof are the same as Theorem 3 in Gai et al. (2013).

Next, we further prove the post-selection consistency.

Let $\boldsymbol{X}$ denotes the design matrix after using B-splines and $C = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$. Write the $n \times |\hat{T}|$ design matrix as $\boldsymbol{X}_{\hat{T}} = (X_{1\hat{T}}, ..., X_{n\hat{T}})^T$, where $\hat{T}$ is the set of selected column indexes. The post-selection least squares estimator $\tilde{\boldsymbol{b}}_0$ of $\boldsymbol{b}_0$ is

$$\tilde{\boldsymbol{b}}_{\hat{T}} = C_{\hat{T},\hat{T}}^{-1}\{\frac{1}{n}X_{\hat{T}}^T Y\} = C_{\hat{T},\hat{T}}^{-1}\{\frac{1}{n}\sum_i^n X_{i\hat{T}}^T Y_i\}, and\ \tilde{\boldsymbol{b}}_{\bar{\hat{T}}} = 0.$$

Lemma 2 (Post-selection consistency of Dantzig selector). Assume that the Dantzig selector estimator is strongly sign consistent, and the $max_{1\leq i\leq n, 1\leq j\leq q}x_{iT^*,j}^2 < \infty$ holds, where $X_{iT^*}^T = (x_{iT^*,1}, ..., x_{iT^*,q})$ corresponding to the ith row of matrix $\boldsymbol{X}_{T^*}$. Then

$$||\tilde{\boldsymbol{b}}_0 - \boldsymbol{b}_0||_2 = O_p(\sqrt{\frac{q^2}{n}}).$$

Proof of Lemma 2: We modify the proof given in Gai et al. (2013). Take

$\Delta_n = \{sign(\tilde{\boldsymbol{b}}_0) = sign(\boldsymbol{b}_0)\}$. On Δ_n, $\hat{T} = T^*$ is fixed. For any $\tau > 0$,

$$P(|\tilde{\boldsymbol{b}}_0 - \boldsymbol{b}_0| > \tau) = P(|\tilde{\boldsymbol{b}}_{\hat{T}} - \boldsymbol{b}_{T^*}| > \tau) \leq P(|\tilde{\boldsymbol{b}}_{\hat{T}} - \boldsymbol{b}_{T^*}| > \tau, \Delta_n) + P(\Delta_n^c)$$

$$= P(|\tilde{\boldsymbol{b}}_{\hat{T}} - \boldsymbol{b}_{T^*}| > \tau | \Delta_n) * P(\Delta_n) + P(\Delta_n^c).$$

Under sign consestency, we know that $P(\Delta_n^c)$ goes to zero as n tends to infinity. Thus it is sufficient to prove $P(|\tilde{\boldsymbol{b}}_{\hat{T}} - \boldsymbol{b}_{T^*}| > \tau | \Delta_n)$ tends to zero. For this, it suffices to prove that, on Δ_n,

$$||k_{0,n} C^{-1}_{T^*,T^*} \{\frac{1}{n} \sum_{i=1}^{n} X_{iT^*} \widetilde{\epsilon}_i\}||_2 = O_p(\sqrt{\frac{q^2}{n}}).$$

Since $max_{1 \leq i \leq n, 1 \leq j \leq q} x^2_{iT^*,j} < \infty$, we have

$$E||\{\frac{1}{n} \sum_{i=1}^{n} X_{iT^*} \widetilde{\epsilon}_i\}||_2^2 = \frac{1}{n^2} \{\sum_{i=1}^{n} ||X_{iT^*}||_2^2 E(\widetilde{\epsilon}_i^2) + \sum_{u,v \in same \ subject} X^T_{uT^*} X_{vT^*} cov(\widetilde{\epsilon}_u, \widetilde{\epsilon}_v)\} \tag{5.1}$$

$$\leq \frac{D^2}{n^2} \sum_{i=1}^{n} ||X_{iT^*}||_2^2 E(\widetilde{\epsilon}_i^2) = O(\frac{q}{n}) \tag{5.2}$$

By the Markov inequality, we get that

$$||\frac{1}{n} \sum_{i=1}^{n} X_{iT^*} \widetilde{\epsilon}_i||_2^2 = O_p(\frac{q}{n}).$$

And it follows that

$$||k_{0,n}C^{-1}_{T^*,T^*}\{\frac{1}{n}\sum_{i=1}^{n}X_{iT^*}\widetilde{\epsilon}_i\}||_2^2 = trace((\frac{1}{n}\sum_{i=1}^{n}X_{iT^*}\widetilde{\epsilon}_i)^T C^{-1}_{T^*,T^*}C^{-1}_{T^*,T^*}(\frac{1}{n}\sum_{i=1}^{n}X_{iT^*}\widetilde{\epsilon}_i)) \tag{5.3}$$

$$= k_{0,n}O(||\frac{1}{n}\sum_{i=1}^{n}X_{iT^*}\widetilde{\epsilon}_i||_2^2) = O_p(\frac{q^2}{n}). \tag{5.4}$$

Therefore, Theorem 1 part (1) is proved. Part (2) & (3) can be proved based on part (1) following the proof of Theorem 1 in Zhou et al. (2013).

5.2 Proof of Theorem 2

To prove Theorem 2, it is sufficient to prove the consistency of the estimator $\hat{\boldsymbol{b}}_1(n)$ (Theorem 3).

Theorem. 3.

For the estimated b-splines coefficients from step 2, we have

(1) $\hat{b}_{1,j}(n) = 0$*, with probability tending to 1, for any* $\hat{b}_{1,j}(n)$ *associated with* $\hat{\mathcal{T}}^{(0)}$*.*

(2) $|\hat{b}_{1,j}(n) - b_{1,j}(n)| = O(n^{-1/2}k_{1,n}^{1/2})$ *for any* $\hat{b}_{1,j}(n)$ *associated with* $\hat{\mathcal{T}}^{(0),c}$*.*

Theorem 2.1 can be proved based on the consistency of PGEE estimator established by Wang et al. (2012). The only difference in our case is that there are approximation errors from B-splines approximation. Our goal is to show that as n goes to infinity, the PGEE estimator with approximation errors is very close to the PGEE estimator which is consistent. The following theorems complete the proof of theorem 3.

5.3 Proof of Theorem 3

Our estimator $\hat{\boldsymbol{c}}_n$ is the root of the following equations:

$$U(\boldsymbol{c}) = S(\boldsymbol{c}) - n\mathbf{q}_\lambda^G(\boldsymbol{c})sign(\boldsymbol{c}) = 0,$$

where

$$S(\boldsymbol{c}) = \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{A}_i^{1/2}(c)\hat{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(c)(\mathbf{Y}_i - \mu_i(\boldsymbol{c})).$$

However, $S(\boldsymbol{c})$ is not a GEE because the existence of approximation errors. If the response is subtracted by the approximation error, then $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{e}$ can be used as the response in regular GEE.

We define

$$\widetilde{U}(\boldsymbol{c}) = \widetilde{S}(\boldsymbol{c}) - n\mathbf{q}_\lambda^G(\boldsymbol{c})sign(\boldsymbol{c}),$$

where

$$\widetilde{S}(\boldsymbol{c}) = \sum_{i=1}^{n} \mathbf{X}_i^T \widetilde{\boldsymbol{A}}_i^{1/2}(\mathbf{c})\widetilde{\mathbf{R}}^{-1}\widetilde{\mathbf{A}}_i^{-1/2}(\mathbf{c})(\widetilde{\mathbf{Y}_i} - \mu_i(\boldsymbol{c})).$$

Now we have our estimator $\hat{\boldsymbol{c}}_n$ from $U(\boldsymbol{c}) = 0$ and $\widetilde{\boldsymbol{c}}_n$ from $\widetilde{U}(\boldsymbol{c}) = 0$ after adjusting for approximation error.

Based on Wang et al.(2012), $\widetilde{\boldsymbol{c}}_n$ is consistent and converges to the true parameter $\boldsymbol{c}_{n0}$ in our model. Our goal is to extend the results to $\hat{\boldsymbol{c}}_n$ with approximation error.

Theorem. 3.1

There exists an approximate penalized GEE solution $\widetilde{\mathbf{c}}_n = (\widetilde{\mathbf{c}}_{n1}^T, \widetilde{\mathbf{c}}_{n2}^T)^T$ *which satisfies the following properties:*

(1)

$$P(|\widetilde{U}_{nj}(\widetilde{\boldsymbol{c}}_n)| = 0, j = 1, ..., s_n) \to 1,$$

$$P(|\widetilde{U}_{nj}(\widetilde{\boldsymbol{c}}_n)| \leq \frac{\lambda_n}{\log n}, j = s_n + 1, ..., k_{1,n} + h) \to 1.$$

(2)

$$P(\widetilde{\boldsymbol{c}}_{n2} = 0) \to 1,$$

$$\widetilde{\boldsymbol{c}}_{n1} \xrightarrow{p} \boldsymbol{c}_{10}.$$

Theorem 3.1 is the asymptotic theories for $\widetilde{\boldsymbol{c}}$ already established in Wang et al. (2012), which suggests that the oracle estimator is an approximation solution for $\widetilde{U}(\boldsymbol{c}) = 0$.

Theorem. 3.2

There exists an approximate penalized GEE solution $\hat{\boldsymbol{c}}_n = (\hat{\boldsymbol{c}}_{n1}^T, \hat{\boldsymbol{c}}_{n2}^T)^T$ *which satisfies the following properties:*

(1.1) $P(|U_{nj}(\hat{\boldsymbol{c}}_n)| = 0, j = 1, ..., s_n) \to 1.$

(1.2) $P(|U_{nj}(\hat{\boldsymbol{c}}_n)| \leq \frac{\lambda_n}{\log n}, j = s_n + 1, ..., k_{1,n} + h) \to 1.$

(2) $P(\hat{\boldsymbol{c}}_{n2} = 0) \to 1.$

(3) $\hat{\boldsymbol{c}}_{n1} \to \boldsymbol{c}_{10}.$

Theorem 3.2 is the main results of this section. Properties (2) and (3) together suggest our proposed estimator is an oracle estimator, which means asymptotically, those truly zero coefficients are estimated as zero by our methods and those non-zero coefficients are estimated as efficiently as if the true model is known in advance.

Theorem. 3.3

For a regular GEE with no penalty or approximation error, we have the GEE estimator $\widetilde{\boldsymbol{c}}_n$ *from* $\widetilde{S}_n(\boldsymbol{c})$*. And with the approximation error but no penalty, we have estimator* $\hat{\boldsymbol{c}}_n$ *from* $\hat{S}_n(\boldsymbol{c})$*. We have the following results (*$\boldsymbol{c}_{n0}$ *is the true parameter):*

(1) $||\widetilde{\boldsymbol{c}}_n - \boldsymbol{c}_{n0}|| = O(\sqrt{k_{1,n}/n})$.

(2) $||\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0}|| = O(\sqrt{k_{1,n}/n})$.

(3) $||\widetilde{\boldsymbol{c}}_n - \hat{\boldsymbol{c}}_{n0}|| = O(\sqrt{k_{1,n}/n})$.

Theorem 3.3 establishs the convergence rates among true parameters, estimated parameters from GEE and estimated parameters from GEE with approximation error. The results are needed to complete the proof of Theorem 3.2 where penalty is incorporated.

5.4 Proofs of Theorem 3.1, 3.2 and 3.3

Theorem 3.1: There exists an approximate penalized GEE solution $\widetilde{\boldsymbol{c}}_n = (\widetilde{\boldsymbol{c}}_{n1}^T, \widetilde{\boldsymbol{c}}_{n2}^T)^T$ which satisfy the following properties:

(1)

$$P(|\widetilde{U}_{nj}(\widetilde{\boldsymbol{c}}_n)| = 0, j = 1, ..., s_n) \to 1$$

$$P(|\widetilde{U}_{nj}(\widetilde{\boldsymbol{c}}_n)| \leq \frac{\lambda_n}{\log n}, j = s_n + 1, ..., k_{1,n} + h) \to 1$$

(2)

$$P(\widetilde{\boldsymbol{c}}_{n2} = 0) \to 1$$

$$\widetilde{\boldsymbol{c}}_{n1} \to \boldsymbol{c}_{10}$$

Proof by construction: (Wang et al., 2012) Let $\widetilde{\boldsymbol{c}}_n = (\widetilde{\boldsymbol{c}}_{n1}^T, 0^T)^T$ be the oracle estimator. ($\widetilde{\boldsymbol{c}}_{n1}$ is the solution to $\widetilde{S}_{n1}(\boldsymbol{c}_{n1}) = 0$). And they shows that the constructed estimator satisfy both (1) and (2).

Theorem 3.2: There exists an approximate penalized GEE solution $\hat{\boldsymbol{c}}_n = (\hat{\boldsymbol{c}}_{n1}^T, \hat{\boldsymbol{c}}_{n2}^T)^T$ which satisfy the following properties:

(1.1)

$$P(|U_{nj}(\hat{\boldsymbol{c}}_n)| = 0, j = 1, ..., s_n) \to 1$$

(1.2)

$$P(|U_{nj}(\hat{\boldsymbol{c}}_n)| \leq \frac{\lambda_n}{\log n}, j = s_n + 1, ..., k_{1,n} + h) \to 1$$

(2)

$$P(\hat{\boldsymbol{c}}_{n2} = 0) \to 1$$

(3)

$$\hat{\boldsymbol{c}}_{n1} \to \boldsymbol{c}_{10}$$

Proof by construction: Let $\hat{\boldsymbol{c}}_n = (\hat{\boldsymbol{c}}_{n1}^T, 0^T)^T$ be the oracle estimator. ($\hat{\boldsymbol{c}}_{n1}$ is the solution to $S_{n1}(\boldsymbol{c}_{n1}) = 0$)

We will show $\hat{\boldsymbol{c}}_n$ satisfy (1.1) and (1.2). And (2) is always true. (3) can be proved by a lemma in Theorem 3.3. (GEE estimator with approximation error is still consistent)

(1.1) can be proved based on assumption 4 and the proof of theorem 1 in Wang et al. (2012) which shows that the GEE estimator we constructed above as $\widetilde{\boldsymbol{c}}_{n1}$ satisfies $P(|\widetilde{c}_{nj}| \geq a\lambda_n, j = 1, ..., s_n) \to 1$, which is sufficient to prove the rest of (1.1). Based on theorem 3, both $\widetilde{\boldsymbol{c}}_{n1}$ and $\hat{\boldsymbol{c}}_{n1}$ converges to the true parameter $\boldsymbol{c_{n1}}$ $O(\sqrt{k_{1,n}/n})$ with a rate of $O(\sqrt{k_{1,n}/n})$. Combined with assumption 4, we have $P(|\hat{c_{nj}}| \geq a\lambda_n, j = 1, ..., s_n) \to 1$ which completes the proof of (1.1).

Proof of (1.2):

Same as in Wang (2011), we only need to prove

$$P(\max_{s_n+1 \leq k \leq k_{1,n}} \frac{1}{n}|S_{nk}(\hat{\boldsymbol{c}}_n)| \leq \frac{\lambda_n}{\log(n)}) \to 1$$

Without approximation error, it is already been proven that

$$P(\max_{s_n+1\leq k\leq k_{1,n}} \frac{1}{n}|\widetilde{S}_{nk}(\widetilde{\boldsymbol{c}}_n)| \leq \frac{\lambda_n}{\log(n)}) \to 1.$$

Next, we will show $||S(\hat{\boldsymbol{c}}_n) - \widetilde{S}(\widetilde{\boldsymbol{c}}_n)|| = O(k_{1,n})$ dominated by $\frac{n\lambda_n}{log(n)}$ which will complete the proof.

Using the decomposition similar in Theorem 3.6 and Lemma 3.3, 3.4, 3.5 in Wang (2011), we can easily show that $||S(\hat{\boldsymbol{c}}_n) - S(\boldsymbol{c}_{n0})|| = O(k_{1,n})$, $||S(\boldsymbol{c}_{n0}) - \bar{S}(\boldsymbol{c}_{n0})|| = O(k_{1,n})$ (our Lemma 3.1 as shown later). Similar results can be derived for GEE with no approximation error: $||\widetilde{S}(\hat{\boldsymbol{c}}_n) - \widetilde{S}(\boldsymbol{c}_{n0})|| = O(k_{1,n})$, $||\widetilde{S}(\boldsymbol{c}_{n0}) - \widetilde{\bar{S}}(\boldsymbol{c}_{n0})|| = O(k_{1,n})$ (See the definition in the following Lemma 3.1)

We can also prove that $||\bar{S}(\boldsymbol{c}_{n0}) - \widetilde{\bar{S}}(\boldsymbol{c}_{n0})|| = O(nk_{1,n}^{-r})$, which completes the proof.

Theorem 3.3: The two constructed estimators are asymptotically equal. In another word, for a regular GEE with no penalty or approximation error, we have the GEE estimator $\widetilde{\boldsymbol{c}}_n$ from $\widetilde{S}_n(\boldsymbol{c})$. And with the approximation error but no penalty, we have estimator $\hat{\boldsymbol{c}}_n$ from $\hat{S}_n(\boldsymbol{c})$. We have the following results ($\boldsymbol{c}_{n0}$ is the true parameter):

(1) $||\widetilde{\boldsymbol{c}}_n - \boldsymbol{c}_0|| = O(\sqrt{k_{1,n}/n})$

(2) $||\hat{\boldsymbol{c}}_n - \boldsymbol{c}_0|| = O(\sqrt{k_{1,n}/n})$

(3) $||\widetilde{\boldsymbol{c}}_n - \hat{c}_0|| = O(\sqrt{k_{1,n}/n})$

Proof: (1) is the Theorem 3.6 in Wang (2011). (3) is a direct result from (1) and (2). So we only need to prove (2). In what follows, we prove (2) by similarly following the proof of Theorem 3.6 in Wang (2011).

Lemma 1 (Similar to (3.3) in Wang (2011)): The initial estimator from working independence correlation structure $\bar{\boldsymbol{c}}_n$ satisfies $||\bar{\boldsymbol{c}}_n - \boldsymbol{c}_{n0}|| = O(\sqrt{k_{1,n}/n})$.

Proof: We modify the proof given in Wang (2011) and only need to show that

$E[||\bar{S}_n(\boldsymbol{c}_{n0})||^2] = O(nk_{1,n})$, where $\bar{S}_n(\boldsymbol{c}_{n0}) = \sum_{i=1}^n \boldsymbol{X}_i^T(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{c}_{n0}))$ with the approximation errors included. Since $E[||\bar{S}_n(\boldsymbol{c}_{n0})||^2] \leq E[\sum \lambda_{max}(\mathbf{X}_i\mathbf{X}_i^T)||\mathbf{Y}_i - \pi_i(c_{n0})||^2] \leq Ctr(\sum \mathbf{X}_i\mathbf{X}_i^T) = O(nk_{1,n})$

Lemma 2 (Similar to (3.4) in Wang (2011)): If we use $\hat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^n \mathbf{A}_i^{-1/2}(\bar{\boldsymbol{c}}_n)(\mathbf{Y}_i - \boldsymbol{\pi}_i(\bar{\boldsymbol{c}}_n))(\mathbf{Y}_i - \boldsymbol{\pi}_i(\bar{\boldsymbol{c}}_n))^T\mathbf{A}_i^{-1/2}(\bar{\boldsymbol{c}}_n)$, we can prove that

$$||\hat{\boldsymbol{R}}^{-1} - \boldsymbol{R}_0^{-1}|| = O(\sqrt{k_{1,n}/n}),$$

where $\boldsymbol{R}_0$ denotes the true common correlation matrix.

Proof: Wang (2011) has proved this results for regular GEE without approximation errors:

$$\widetilde{S}(\boldsymbol{c}) = \sum_{i=1}^n \mathbf{X}_i^T\widetilde{\boldsymbol{A}}_i^{1/2}(\mathbf{c})\widetilde{\mathbf{R}}^{-1}\widetilde{\mathbf{A}}_i^{-1/2}(\mathbf{c})(\widetilde{\mathbf{Y}}_i - \mu_i(\boldsymbol{c}))$$

where $\widetilde{\boldsymbol{Y}}_i = \boldsymbol{Y}_i - \boldsymbol{e}_i$, $\widetilde{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^n \mathbf{A}_i^{-1/2}(\bar{\boldsymbol{c}}_n^0)(\widetilde{\boldsymbol{Y}}_i - \boldsymbol{\pi}_i(\bar{\boldsymbol{c}}_n^0))(\widetilde{\boldsymbol{Y}}_i - \boldsymbol{\pi}_i(\bar{\boldsymbol{c}}_n^0))^T\mathbf{A}_i^{-1/2}(\bar{\boldsymbol{c}}_n^0)$ and $\bar{\boldsymbol{c}}_n^0$ is the estimator assuming working independence.

They defines

$$\boldsymbol{R}^{**} = \frac{1}{n}\sum_{i=1}^n \mathbf{A}_i^{-1/2}(\boldsymbol{c}_{n0})(\widetilde{\boldsymbol{Y}}_i - \boldsymbol{\pi}_i(\boldsymbol{c}_{n0}))(\widetilde{\boldsymbol{Y}}_i - \boldsymbol{\pi}_i(\boldsymbol{c}_{n0}))^T\mathbf{A}_i^{-1/2}(\boldsymbol{c}_{n0}),$$

and shows that $||\widetilde{\mathbf{R}} - \mathbf{R}^{**}|| = O(1/\sqrt{n})$ and $||\mathbf{R}^{**} - \mathbf{R}_0|| = O(\sqrt{k_{1,n}/n})$. Therefore $||\widetilde{\mathbf{R}} - \mathbf{R}_0|| = O(\sqrt{k_{1,n}/n})$, which is sufficient to prove (3.4) in Wang (2011), which is $||\widetilde{\mathbf{R}}^{-1} - \mathbf{R}_0^{-1}|| = O(\sqrt{k_{1,n}/n})$.

With the approximation errors, to prove Lemma 2, we modify the proof above

and define

$$\boldsymbol{R}^* = \frac{1}{n}\sum_{i=1}^{n} \mathbf{A}_i^{-1/2}(\boldsymbol{c}_{n0})(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{c}_{n0}))(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{c}_{n0}))^T \mathbf{A}_i^{-1/2}(\boldsymbol{c}_{n0}).$$

First, we show $||\mathbf{R}^* - \mathbf{R}_0|| = O(\sqrt{k_{1,n}/n})$. That is because $||\mathbf{R}^* - \mathbf{R}^{**}|| \leq 2||\frac{1}{n}\sum \mathbf{A}_i^{-1/2}\mathbf{e}_i(\mathbf{Y}_i - \pi_i(\mathbf{c}_{n0}))\mathbf{A}_i^{-1/2})|| + ||\frac{1}{n}\sum \mathbf{A}_i^{-1/2}\mathbf{e}_i\mathbf{e}_i^T\mathbf{A}_i^{-1/2}|| = O(k_{1,n}^{-r})$, which is dominated by $O(\sqrt{k_{1,n}/n})$. Given that $||\mathbf{R}^{**} - \mathbf{R}_0|| = O(\sqrt{k_{1,n}/n})$, we have $||\mathbf{R}^* - \mathbf{R}_0|| = O(\sqrt{k_{1,n}/n})$.

Next, we show $||\hat{\mathbf{R}} - \mathbf{R}^*|| = O(\sqrt{k_{1,n}/n})$.

$$\begin{aligned}
|\hat{\mathbf{R}}_{kj} - \mathbf{R}^*_{kj}| \leq & |\frac{1}{n}\sum \frac{(Y_{ik} - \hat{\pi}_{ik})(Y_{ij} - \hat{\pi}_{ij}) - (Y_{ik} - \pi^0_{ik})(Y_{ij} - \pi^0_{ij})}{\sqrt{\mathbf{A}^0_{ik}}\sqrt{\mathbf{A}^0_{ij}}}| \\
& + |\frac{1}{n}\sum \frac{(Y_{ik} - \hat{\pi}_{ik})(Y_{ij} - \hat{\pi}_{ij})}{\sqrt{\mathbf{A}^0_{ik}}\sqrt{\mathbf{A}^0_{ij}}}\hat{\delta}_{ijk}| \\
\leq & |\frac{1}{n}\sum \frac{(\widetilde{Y}_{ik} - \widetilde{pi}_{ik})(\widetilde{Y}_{ij} - \widetilde{\pi}_{ij}) - (\widetilde{Y}_{ik} - \pi^0_{ik})(\widetilde{Y}_{ij} - \pi^0_{ij})}{\sqrt{\mathbf{A}^0_{ik}}\sqrt{\mathbf{A}^0_{ij}}}| \\
& + |\frac{1}{n}\sum \frac{(\widetilde{Y}_{ik} - \widetilde{\pi}_{ik})(\widetilde{Y}_{ij} - \widetilde{\pi}_{ij})}{\sqrt{\mathbf{A}^0_{ik}}\sqrt{\mathbf{A}^0_{ij}}}\widetilde{\delta}_{ijk}| + O(k_{1,n}^{-r})
\end{aligned}$$

where $\pi^0_{ik} = \pi_{ik}(\mathbf{c}_{n0})$, $\widetilde{\pi}_{ik} = \pi_{ik}(\widetilde{\mathbf{c}}_{n0})$, $\hat{\pi}_{ik} = \pi_{ik}(\hat{\mathbf{c}}_{n0})$, $\mathbf{A}^0_{ik} = \pi^0_{ik}(1 - \pi^0_{ik})$, $\widetilde{\mathbf{A}_i k} = \widetilde{\pi}_{ik}(1 - \widetilde{\pi}_{ik})$, and $\hat{\mathbf{A}_{ik}} = \hat{\pi}_{ik}(1 - \hat{\pi}_{ik})$. $\widetilde{\delta}_{ijk} = [\mathbf{A}^0_{ik}\mathbf{A}^0_{ij}]^{1/2}[\widetilde{\mathbf{A}}_{ik}\widetilde{\mathbf{A}}_{ij}]^{-1/2} - 1$ and $\hat{\delta}_{ijk} = [\mathbf{A}^0_{ik}\mathbf{A}^0_{ij}]^{1/2}[\hat{\mathbf{A}}_{ik}\hat{\mathbf{A}}_{ij}]^{-1/2} - 1$.

Sine the first two terms are calculated in Wang (2011), we have $||\hat{\mathbf{R}} - \mathbf{R}^*|| \leq O(k_{1,n}/n) + O(k_{1,n}^{-r}) = O(k_{1,n}/n)$, which complete the proof for Lemma 2.

Lemma 3.1. Define

$$\widetilde{\mathbf{S}}_n(\mathbf{c}_{n0}) = \sum \mathbf{X}_i \mathbf{A}_i^{1/2} \widetilde{\mathbf{R}}^{-1} \mathbf{A}_i^{-1/2} (\widetilde{\mathbf{Y}}_i - \pi_i)$$

$$\widetilde{\bar{\mathbf{S}}}_n(\mathbf{c}_{n0}) = \sum \mathbf{X}_i \mathbf{A}_i^{1/2} \bar{\mathbf{R}}^{-1} \mathbf{A}_i^{-1/2} (\widetilde{\mathbf{Y}}_i - \pi_i)$$

$$\mathbf{S}_n(\mathbf{c}_{n0}) = \sum \mathbf{X}_i \mathbf{A}_i^{1/2} \hat{\mathbf{R}}^{-1} \mathbf{A}_i^{-1/2} (\mathbf{Y}_i - \pi_i)$$

$$\bar{\mathbf{S}}_n(\mathbf{c}_{n0}) = \sum \mathbf{X}_i \mathbf{A}_i^{1/2} \bar{\mathbf{R}}^{-1} \mathbf{A}_i^{-1/2} (\mathbf{Y}_i - \pi_i)$$

Then we have $||\mathbf{S}_n(\mathbf{c}_{n0}) - \bar{\mathbf{S}}_n(\mathbf{c}_{n0})|| = O(k_{1,n})$.

Proof: Following the proof of Lemma 3.1 in Wang (2011) which proves $||\widetilde{\mathbf{S}}_n(\mathbf{c}_{n0}) - \widetilde{\bar{\mathbf{S}}}_n(\mathbf{c}_{n0})|| = O(k_{1,n})$, we incorporate the approximation error and modify the proof accordingly.

Let $\mathbf{Q} = \hat{\mathbf{R}}^{-1} - \bar{\mathbf{R}}^{-1}$, it is already been proven that $q_{ij} = O(\sqrt{k_{1,n}/n})$.

$$\begin{aligned}
\mathbf{S}_n(\mathbf{c}_{n0}) - \bar{\mathbf{S}}_n(\mathbf{c}_{n0}) &= \sum_i \sum_{j_1} \sum_{j_2} q_{j_1,j_2} A_{ij_1}^{1/2} A_{ij_2}^{-1/2} (Y_{ij_2} - \pi_{ij_2}) X_{ij_1} \\
&= \sum_{j_1}^{m} \sum_{j_2}^{m} q_{j_1 j_2} [\sum_{i}^{n} A_{ij_1}^{1/2} \eta_{ij_2} X_{ij_1}]
\end{aligned}$$

where $\eta_{ij_2} = A_{ij_2}^{-1/2} (Y_{ij_2} - \pi_{ij_2})$. Note that

$$E[||\sum_{i}^{n} A_{ij_1}^{1/2}\eta_{ij_2}X_{ij_1}||^2] = \sum_{i}^{n} A_{ij_1}E[\widetilde{\eta}_{ij_2} + O(p^{-r})]^2 X_{ij_1}^T X_{ij_1}$$
$$\leq \sum_{i}^{n} X_{ij_1}^T X_{ij_1} = O(nk_{1,n})$$

where $\widetilde{\eta}_{ij_2} = A_{ij_2}^{-1/2}(\widetilde{Y}_{ij_2} - \pi_{ij_2})$, and the rest of the proof is same as in Wang (2011).

Lemma 3.2, 3.3, 3.4, 3.5 in Wang (2011) can also be extended to GEE with approximation error as follows, and the proof is the same as in Wang (2011).

Lemma 3.2.

$$\bar{\mathbf{D}}(\mathbf{c}_n) = \bar{\mathbf{H}}(\mathbf{c}_n) + \bar{\mathbf{E}}(\mathbf{c}_n) + \bar{\mathbf{G}}(\mathbf{c}_n)$$

where

$$\bar{\mathbf{D}}(\mathbf{c}_n) = -\frac{\partial \bar{\mathbf{S}}(\mathbf{c}_n)}{\partial \mathbf{c}_n^T}$$
$$\bar{\mathbf{H}}(\mathbf{c}_n) = \sum_{i}^{n} \mathbf{X}_i^T \mathbf{A}_i^{1/2}(\mathbf{c}_n)\bar{\mathbf{R}}^{-1}\mathbf{A}_i^{1/2}(\mathbf{c}_n)\mathbf{X}_i$$
$$\bar{\mathbf{E}}(\mathbf{c}_n) = \frac{1}{2}\sum_{i}^{n}\sum_{j}^{m}(1 - 2\pi_{ij}(\mathbf{c}_n))\eta_{ij}(\mathbf{c}_n)\mathbf{X}_i^T \mathbf{A}_i^{1/2}(\mathbf{c}_n)\bar{\mathbf{R}}^{-1}\hat{\mathbf{e}}_j\hat{\mathbf{e}}_j^T\mathbf{X}_i$$
$$\bar{\mathbf{G}}(\mathbf{c}_n) = -\frac{1}{2}\sum_{i}^{n}\sum_{j}^{m}(1 - 2\pi_{ij}(\mathbf{c}_n))\mathbf{A}_i^{1/2}(\mathbf{c}_n)\mathbf{X}_{ij}\mathbf{X}_{ij}^T\hat{\mathbf{e}}_j^T\bar{\mathbf{R}}^{-1}\eta_i(\mathbf{c}_n)$$

where $\hat{\mathbf{e}}_i$ denotes a unit vector of length m whose jth entry is one and all the other entries are zero.

Lemma 3.3. For any $\Delta > 0$, for $\mathbf{b}_n \in R^{k_{1,n}+h}$, we have

$$\sup_{||\mathbf{c}_n - \mathbf{c}_{n0}|| \leq \Delta\sqrt{k_{1,n}^2/n}} \sup_{||\mathbf{b}_n||=1} |\mathbf{b}_n^T[\mathbf{D}(\mathbf{c}_n) - \bar{\mathbf{D}}(\mathbf{c}_n)]\mathbf{b}_n| = O(\sqrt{nk_{1,n}})$$

Lemma 3.4. For any $\Delta > 0$, for $\mathbf{b}_n \in R^{k_{1,n}+h}$, we have

$$\sup_{||\boldsymbol{c}_n - \boldsymbol{c}_{n0}|| \le \Delta\sqrt{k_{1,n}^2/n}} \sup_{||\mathbf{b}_n||=1} |\mathbf{b}_n^T[\bar{\mathbf{D}}(\mathbf{c}_n) - \bar{\mathbf{H}}(\mathbf{c}_n)]\mathbf{b}_n| = O(\sqrt{nk_{1,n}})$$

Lemma 3.5. For any $\Delta > 0$, for $\mathbf{b}_n \in R^{k_{1,n}+h}$, we have

$$\sup_{||\boldsymbol{c}_n - \boldsymbol{c}_{n0}|| \le \Delta\sqrt{k_{1,n}^2/n}} \sup_{||\mathbf{b}_n||=1} |\mathbf{b}_n^T[\bar{\mathbf{H}}(\mathbf{c}_n) - \bar{\mathbf{H}}(\mathbf{c}_{n0})]\mathbf{b}_n| = O(\sqrt{nk_{1,n}})$$

Theroem 3.6. The root $\hat{\boldsymbol{c}}_n$ of $\boldsymbol{S}_n(\boldsymbol{c}_n) = 0$ satisfies

$$||\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0}|| = O(\sqrt{k_{1,n}/n})$$

Proof: Similar to the proof of Theroem 3.6 in Wang (2011), we only need to prove that for any $\epsilon > 0$, there exists a constant $\Delta > 0$ such that for all n sufficiently large,

$$P(\sup_{||\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0}|| = \Delta\sqrt{k_{1,n}^2/n}} (\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0})^T \mathbf{S}_n(\hat{\boldsymbol{c}}_n) < 0) \ge 1 - \epsilon$$

We have

$$(\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0})^T \mathbf{S}_n(\hat{\boldsymbol{c}}_n) = (\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0})^T \mathbf{S}_n(\boldsymbol{c}_{n0}) - (\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0})^T \mathbf{D}_n(\hat{\boldsymbol{c}}_n^*)(\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0}) = I_{n1} - I_{n2}$$

where $\hat{\boldsymbol{c}}_n^*$ is between $\hat{\boldsymbol{c}}_n$ and $\boldsymbol{c}_{n0}$.

And $I_{n1} = (\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0})^T \bar{\mathbf{S}}_n(\boldsymbol{c}_{n0}) + (\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0})^T [\mathbf{S}_n(\boldsymbol{c}_{n0}) - \bar{\mathbf{S}}_n(\boldsymbol{c}_{n0})] = I_{n11} + I_{n12}$

$I_{n2} = -(\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0})^T \bar{\mathbf{D}}_n(\hat{\boldsymbol{c}}_n^*)(\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0}) - (\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0})^T [\mathbf{D}_n(\hat{\boldsymbol{c}}_n^*) - \bar{\mathbf{D}}_n(\hat{\boldsymbol{c}}_n^*)](\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0}) = I_{n21} + I_{n22}$

And

$$\begin{aligned} I_{n21} =& -(\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0})^T \bar{\mathbf{H}}_n(\hat{\boldsymbol{c}}_{n0})(\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0}) \\ & -(\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0})^T [\bar{\mathbf{H}}_n(\hat{\boldsymbol{c}}_n^*) - \bar{\mathbf{H}}_n(\hat{\boldsymbol{c}}_{n0})](\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0}) \\ & -(\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0})^T [\bar{\mathbf{D}}_n(\hat{\boldsymbol{c}}_n^*) - \bar{\mathbf{H}}_n(\hat{\boldsymbol{c}}_n^*)](\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0}) \\ =& I_{n21}^a + I_{n21}^b + I_{n21}^c \end{aligned}$$

Based on the proof in Wang (2011), we can show that $I_{n11} = \Delta O(k_{1,n})$, $I_{n12} = \Delta o(k_{1,n})$, $I_{n22} = \Delta^2 O(k_{1,n})$, $I_{n21}^b = \Delta^2 O(k_{1,n})$, $I_{n21}^c = \Delta^2 O(k_{1,n})$ and $I_{n21}^a \leq -C\Delta^2 k_{1,n}$. Therefore $(\hat{\boldsymbol{c}}_n - \boldsymbol{c}_{n0})^T \mathbf{S}_n(\hat{\boldsymbol{c}}_n)$ is negative for Δ large enough, which completes the proof of Theroem 3.6.

Chapter 6

Discussion

Literature on the dynamic effect of diarrhea on children growth is limited. In this study, we propose a non-parametric model to study the pattern of diarrhea effect over time. A two-step estimation approach is developed to identify the null region of the effect and estimate the effect on non-null region. The approach is developed by imposing group penalty on estimating equations for longitudinal data and the asymptotic properties of the developed estimators are investigated. Simulation studies show that the proposed method is capable of identifying null region and obtain a reliable estimate of the functional effect of diarrhea on the non-null region. Based on the data from NIH study in Bangladesh, our data analysis results show that the diarrhea effect on children growth becomes significant 3 months after diarrhea onsets and vanishes after 15 months. Besides the original model, 4 of the extended model are developed to estimate curves leveling off to a constant, to take into account additional covariates, to estimate multiple curves and to model curves with more than one dimension. Overall, our models provide new statistical tools to quantify the relationship between diarrhea and childhood growth in a dynamic fashion, which gives us insights on the pattern and the effect window of diarrhea

effect.

There are limitations of modeling the effect of a complicated biological process, such as diarrhea, in a statistical model with underlying model assumptions. For example, in our model, we assume that, even though different diarrhea episodes may have different effect curves in our extended models, the effects are additive. However, it is possible that when children have diarrhea frequently, they might experience more growth shortfall than the simple addition of each individual episode effect. Also, in addition to the diarrhea onset time that we include as additional factor affecting the diarrhea effect curve, many other factors may play a role as well. For instance, it is reasonable to believe that the severity of the diarrhea also affects the effect curve. Those variables measuring the severity may include the duration of the episodes, the symptoms developed during the episodes, etc. Therefore, a more complex statistical model is preferred to include more relevant covariates in the future studies when those variable are available.

On the other hand, some of the limitations of the estimating process can also be improved. Noted that in the simulation study, even though the fitted curve is very close to the true curve, the estimated null region is not exactly same as the true null region. In addition, due to the boundary effect, the fitted curve may not be reliable right after diarrhea onsets. The reason of causing both issues is that we place only a few knots to avoid overfitting. The penalty imposed in our model promotes the sparsity of the curve, but does not guarantee the smoothness of the estimated function. It is possible to impose two penalties at the same time, one for sparsity and one for smoothness, which can be another future research direction.

Bibliography

[1] Bowen, A., Agboatwalla, M., Luby, S., Tobery, T., Ayers, T., Hoekstra, RM. (2012). Association between intensive handwashing promotion and child development in Karachi, Pakistan: a cluster randomized controlled trial. *Archives of Pediatrics and Adolescent Medicine*, 166, 1037-–44.

[2] Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6), 2313-2351.

[3] Currie, I. D., Durban, M., & Eilers, P. H. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2), 259-280.

[4] Derso, T., Tariku, A., Biks, G. A., & Wassie, M. M. (2017). Stunting, wasting and associated factors among children aged 6-24 months in Dabat health and demographic surveillance system site: A community based cross-sectional study in Ethiopia. *BMC pediatrics*, 17(1), 96. doi:10.1186/s12887-017-0848-2

[5] Dierckx, P. (1980). An algorithm for cubic spline fitting with convexity constraints.*Computing*, 24, 349—371.

[6] Dziak, J. J., Li, R., and Qu, A. (2009). An overview on quadratic inference function approaches for longitudinal data. *Frontiers of Statistics*, Volume 1: New Developments in Biostatistics and Bioinformatics, 49–72.

[7] Eilers, P. and B. Marx (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, 66, 159-–174.

[8] Fan, J., Gijbels, I., Hu, T. C., and Huang, L. S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica*, 113-127.

[9] Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), 1348-1360.

[10] Fan, Y., & Li, R. (2012). Variable selection in linear mixed effects models. *Annals of statistics*, 40(4), 2043.

[11] Gai, Y., Zhu, L., & Lin, L. (2013). Model selection consistency of Dantzig selector. *Statistica Sinica*, 615-634.

[12] Gaylord, C. K., & Ramirez, D. E. (1991). Monotone regression splines for smoothed bootstrapping. *Computational Statistics Quarterly*, 6(2), 85-97.

[13] Guo, W. (2002). Functional mixed effects models, *Biometrics*, 58, 121–128.

[14] Hedeker, D., & Gibbons, R. D. (2006). Longitudinal data analysis (Vol. 451). *John Wiley & Sons*.

[15] Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological methods*, 3(4), 486.

[16] Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. P. (1998), Nonparametric Smoothing Estimates of Time-Varying Coefficient Models With Longitudinal Data, *Biometrika*, 85, 809–822.

[17] Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31(5), 1600-1635.

[18] Hunter, D. R., & Li, R. (2005). Variable selection using MM algorithms. *Annals of statistics*, 33(4), 1617.

[19] Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in medicine*, 30(25), 3050-3056.

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

[21] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.

[22] Lin, X., & Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American statistical Association*, 95(450), 520-534.

[23] Lindstrom, M. J., and Bates, D. M. (1990). Nonlinear Mixed Effects Models for Repeated Measures Data, *Biometrics*, 46, 673–6.

[24] Marx, B. and P. Eilers (2005). Multidimensional Penalized Signal Regression. *Technometrics*, 47, 13-–22.

[25] Moyeed, R. A., and Diggle, P. J. (1994). Rates of Convergence in Semi-Parametric Modeling of Longitudinal Data, *Alustralian Journal of Stactistics*, 36, 75–93.

[26] Prentice, A. M., Ward, K. A., Goldberg, G. R., Jarjou, L. M., Moore, S. E., Fulford, A. J., & Prentice, A. (2013). Critical windows for nutritional interventions against stunting. *The American of Clinical Nutrition*, 97(5), 911-918.

[27] Prendergast, A. J., & Humphrey, J. H. (2014). The stunting syndrome in developing countries. *Paediatrics and international child health*, 34(4), 250–265. doi:10.1179/2046905514Y.0000000158

[28] Qu, A. and Li, R. (2006). Quadratic inference functions for varying coefficient models with longitudinal data. *Biometrics*, 62, 379—391.

[29] Ramsay, J. O. (1988). Monotone regression splines in action (with discussion). *Statistical Science*, 3, 425—461.

[30] Ruppert, D., Wand, MP., Carroll, RJ. (2003) Semiparametric Regression. *New York: Cambridge University Press.*

[31] Schnee, A. E., Haque, R., Taniuchi, M., Uddin, M. J., Alam, M. M., Liu, J., ... & Platts-Mills, J. A. (2018). Identification of etiology-specific diarrhea associated with linear growth faltering in Bangladeshi infants. *American journal of epidemiology*, 187(10), 2210-2218.

[32] Troeger, C., Forouzanfar, M., Rao, P. C., Khalil, I., Brown, A., Reiner Jr, R. C., ... & Alemayohu, M. A. (2017). Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Infectious Diseases*, 17(9), 909-948.

[33] Troeger, C., Colombara, D. V., Rao, P. C., Khalil, I. A., Brown, A., Brewer, T. G., ... & Petri Jr, W. A. (2018). Global disability-adjusted life-year estimates of

long-term health burden and undernutrition attributable to diarrhoeal diseases in children younger than 5 years. *The Lancet Global Health*, 6(3), e255-e269.

[34] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

[35] Tindyebwa, D. (2004). Common clinical conditions associated with HIV. *Handbook on Paediatric AIDS in Africa*.

[36] Tumilowicz, A., Habicht, J. P., Pelto, G., & Pelletier, D. L. (2015). Gender perceptions predict sex differences in growth patterns of indigenous Guatemalan infants and young children. *The American journal of clinical nutrition*, 102(5), 1249-1258.

[37] Wang N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, 90, 43-–52.

[38] Wang, N., Carroll, R. J., & Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, 100(469), 147-157.

[39] Wang, L., Zhou, J. and Qu, A. (2012). Penalized generalized estimating equations for high dimensional longitudinal data analysis. *Biometrics*, 68, 353-–360.

[40] WARE, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician*, 39, 95–101.

[41] Wierzba, T. F., & Muhib, F. (2018). Exploring the broader consequences of diarrhoeal diseases on child health. *The Lancet Global Health*, 6(3), e230-e231.

[42] Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B*, 65, 95-–114.

[43] Xiao, L., Li, Y., & Ruppert, D. (2013). Fast bivariate P-splines: the sandwich smoother. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3), 577-599.

[44] Xue, L., Qu, A. and Zhou, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association*, 105, 1518–1530.

[45] Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49-67.

[46] Zhou, J., and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistical Association*, 107(498), 701-710.

[47] Zhou, J., Wang, N. Y., & Wang, N. (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica*, 23(1), 25.

[48] Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net.*Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.

[49] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418-1429.