Estimation of Dynamic Diarrhea Effects on Childhood Growth with Latent Subgroup

Tonghao Zhang

B.S., Zhejiang University, P.R.China, 2016

A 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
May, 2021

_____

_____

_____

_____

Abstract

Childhood is a critical period for physical and cognitive development. Childhood diarrhea is not only associated with high mortality and morbidity rate, its complication can also lead to long term growth faltering.

Existing studies treat diarrhea effect on growth as constant over time, or assume diarrhea effect is the same across the population. Diarrhea episodes have different clinical severity, and the resulting growth shortfalls may not be the same. Under this rationale, we propose a semi-parametric model with latent subgroup to estimate the heterogeneous dynamic diarrhea effect. To accommodate subject differences, natural growth and diarrhea effect are treated as random effect curves and the mixture distributional assumption is adopted. The latent subgroup designation in the mixture model will help explain individual characteristics such as household socioeconomic status, hygiene level and other driving factors behind diarrhea. Parameter estimate and has been developed using the Expectation-Maximization algorithm. Asymptotic variance of the estimator is studied.

Simulation study shows that our model achieves simultaneous identification of subgroups in growth and diarrhea vulnerability and estimation of growth patterns and effect curves. The estimator variance is also validate by empirical coverage probability. The proposed models are applied to the data from NIH cohort study collected from children in Bangladesh. Results show that the diarrhea effect on children's HAZ score peaks at around 400 days with a decrease of HAZ -0.12, and will leave long-term growth loss for 85% of the cohort. Overall, our model provides new statistical tools to quantify the dynamic pattern of childhood growth and diarrhea affects, which helps us make better health policy and conduct more efficient interventions.

Acknowledgements

My time at UVa has been nothing short of amazing. I would like to thank everyone who has guided, helped and supported me along the way.

Firstly, I want to give my deepest thanks to Professor Jianhui Zhou, my advisor, for his enduring guidance and advice. Throughout my PhD years, I benefited tremendously from all the discussion with him. He has been the most approachable and patient, always able to answer my confusion with succinct and fundamental statistical knowledge. In addition to his academic advice, I learned from him how to work with others and treat others with kindness and honesty. I have been the most fortunate to work with him. As my teacher and mentor, he has taught me more than I could ever give him credit for here. He has shown me, by his example, what a good statistician (and person) should be.

I am deeply indebted to to Professor Xin Cynthia Tong, for her valuable guidance throughout my studies. Most of the methodology and computing tools I'm using today are acquired when working on projects with Professor Tong. I have learned a lot from her experience and expertise, and the opportunities she provided to me built my confidence and ability.

I am especially grateful to Professor Jennie Ma as my supervisor during my research assistant work at UVa Medical School. Her experience and expertise were vital in helping me bridge statistical theory and mediacal research. I learned from her the rigorous attitude towards data, when I was still a layman in real world application. Professor Ma cares about both student's study and life. I cannot thank her enough for inviting me to her thanksgiving dinner. All the stories she shared touched me deeply and inspire me to live with full positivity.

I am also grateful for Dr. William A. Petri for offering his expertises as a mem-

ber 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.

Most importantly, I want to give my deep gratitude to my parents, Rengang Zhang and Mei Liu. They have cherished with me every great moment and supported me whenever I needed it. Without their unconditional love, none of this would have been possible.

Let me end with a short Chinese verse I read when was just a kid. It just all came back to me when writing the thesis. 惟念当离别，恩情日以新。

Contents

Chapter 1

Introduction

1.1 Medical Background

1.1.1 Childhood Growth

Childhood is a critical period for both physical and cognitive development. Growth failure in childhood is the most prevalent form of undernutrition globally. In 2019, 144 million children are identified as stunting worldwide (UNICEF, WHO, World Bank Group joint malnutrition estimates, 2020), defined as height-for-age Z score (HAZ score) below -2. Even though stunted children are more than two standard deviations below in height compared to the median of an age and gender matched reference population, short stature is not usually in itself problematic. "Stunting syndrome" in which multiple pathological changes marked by linear growth retardation in early life are associated with increased morbidity and mortality. Stunting are also linked with reduced physical, cognitive capacity and an elevated risk of metabolic disease into adulthood (Prendergast and Humphrey, 2014). Moreover, stunting is a cyclical process, because women who were themselves stunted in child-

hood tend to have stunted offspring. Childhood stunting hinders economic productivity of adults, lead to poor living standard to children in the household (Derso et al., 2017). Stunting creates an intergenerational cycle of poverty and reduced human capital that is difficult to break. Due to its vast prevalence and dire short-, medium- and long-term sequelae, stunting has been identified as a top priority in public health, especially in low-income countries.

Despite the high global prevalence of stunting, the pathogenesis underlying stunting is surprisingly poorly understood. For this reason, the most tractable pathways for effective interventions to promote healthy growth remain unclear (Piwoz et al., 2012). Epidemiological studies show that suboptimal breastfeed, low meal frequency and micronutrient deficiencies are important proximal determinants of stunting. Stunting occurs within a complex interplay of more distal community and societal factors, such as access to healthcare and education, political stability, urbanisation, population density and social support networks (Prendergast and Humphrey, 2014). Recurrent infections such as diarrhea, are also linked to poor growth.

Infection control and growth promoting nutritional intervention are common practices to prevent stunting. Many intervention methods were studied and used in practice. However, there is an on-going debate over the timing for intervention. Victora et al. (2010) suggests -9 (9 months before birth) and 23 month is the optimal window of opportunity within which growth-promoting nutritional interventions should be focused, while Prentice et al. (2013) argues that substantial height catch-up occurs between 24 month and midchildhood and again between midchildhood and adulthood.

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. In this thesis, we study the pattern

of diarrhea effect on growth over time and the heterogeneity of change pattern within population.

1.1.2 Diarrhea and its Effect on Growth

Diarrhea has been one of the leading causes of death and illness for children under 5 years old, behind preterm birth complications, neonatal encephalopathy, and lower respiratory infections(GBD 2015). It is a major cause for mortality and mobility, especially in the developing countries. Troeger et al. (2017) shows that diarrheal diseases were responsible for half million deaths among children under 5 years old, representing 8.6% of the 5.82 million deaths in this age group. Diarrhea can also lead to increased risk of other infectious diseases due to diarrhea induced undernutrition.

Infection is the major causes of diarrhea. The most prevalent infectious cause is rotavirus (RV), followed by bacterial infections due to the presence of bacteria, such as Enterobacter, Escherischia coli and Shigella, in contaminated water or food. Parasitic infections are also common, caused by parasites from food or water such as Cryptosporidium. These three contributed over half of deaths caused by diarrhea (Tindyebwa, 2004). Treatment of childhood diarrhea should follow IMCI guidelines, which include management and correction of dehydration, aggressive nutritional management to minimise the occurrence of persistent diarrhea, and malnutrition and nutrition counselling, including a review of household hygienic practices.

Childhood diarrhea can even continue to have impact into adulthood. Childhood diarrhea is associated with elevated risk of adult chronic diseases such as cardiovascular disease, diabetes, obesity and hypertension (Wierzba and Muhib, 2018). Complications of diarrhea including dehydration and micronutrient deficiencies may not necessarily lead to death but could still have long-term effect on

children's growth. Studies have suggested diarrhea could also have long term effect on population health (Bowen et al, 2012, Troeger et al., 2018, Schnee et al, 2018).

Continued efforts to improve access to safe water, sanitation, and childhood nutrition will be important in reducing the global burden of diarrhea. Better understanding of diarrhea effect will be vital in improvement of policy making, optimize interventions and resource allocation.

1.1.3 NIH Study Cohort

During 2008-2012, a cohort of total 629 infants (332 boys and 297 girls) were enrolled after birth in Dhaka, Bangladesh under the NIH study. The enrolled infants are uniquely identified by their subject id. Every 3 months, the field team will take the growth measurements of each child untill the end of the study, resulting in 6381 total observations. Measures of growth include weight and 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 normalized within age and gender matched reference population. 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 (1.1) For each episodes of diarrhea, we tested the presence of Crypto, EH and Giardia, which are

Table 1.1: Summary of the NIH cohort data

NIH Cohort		
Data	Variables	Summary
Growth data	subject id	629 subjects
	gender	297 female, 332 male
	age	1 - 1560 days
	weight, height, WHZ, HAZ, WAZ, BAZ	the growth outcome measured at the visit date
Diarrhea data	subject id	521 subjects
	diarrhea episodes	2605 episodes, up to 1623 days after birth
	infection type	include EH_CTRT, GIA_CTRT, CRY_CTRT

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. Table (1.1.3) summarizes the data collected in the NIH study, it consist of growth outcome data and diarrhea related data.

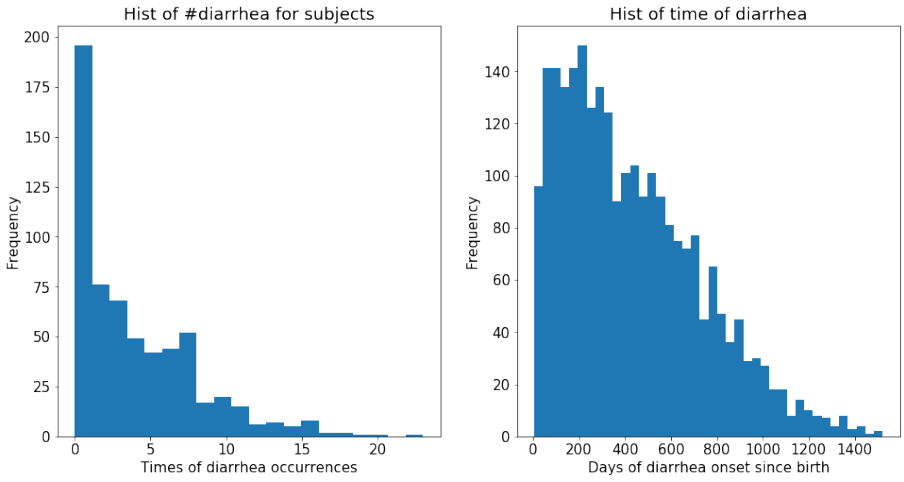

Figure 1.1: Histogram of recorded diarrhea

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. The first one being that, our goal is to study growth as a continuous curve but the growth outcomes in the data are measured at discrete time points. The second challenge is that, 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. Lastly, growth pattern and diarrhea effect on growth are different from subject to subject, the heterogeneity needs to be considered when proposing statistical model.

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 random effects are introduced to reflect different levels of diarrhea vulnerability. Mixture model distributional assumption is imposed on the random effects. Depending on the subject, diarrhea effect on growth ranges from mild digestion symptoms to severe malnutrition, or in statistical terms, the reaction is long-tailed. Mixture model is flexible enough to depict this long-tailed distribution. Moreover, mixture model divide subjects into latent groups. By studying the relationship between group membership and the available information on living hygiene level and societal factors, such as access to healthcare, education and population density, we can provide useful insight on better policy making and more efficient interventions. The next section reviews mixture models.

1.2 Mixture Model

1.2.1 Definition

Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ denote a random sample of size n, where each $\boldsymbol{Y}_k$ is a p-dimensional random vector with probability density function $f(\boldsymbol{Y}_k)$ on $\mathbb{R}^p$. In an application where we make p consecutive observations on subjects enrolled regarding some outcome of interests, $\boldsymbol{Y}_k$ will be the observed values on subject indexed by k.

The mixture model suppose that the density $f(\boldsymbol{Y}_k)$ of $\boldsymbol{Y}_k$ can be written in the form

$$f(\boldsymbol{Y}_k) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{Y}_k), \tag{1.1}$$

where the $f_g(\boldsymbol{Y}_k)$ are proper probability density functions. The π_g are non-negative quantities that sum to one; that is,

$$0 \leq \pi_g \leq 1, \ (g = 1, \ldots, G) \text{ and } \sum_{g=1}^{G} \pi_g = 1. \tag{1.2}$$

The quantities $\pi_1, \ldots, \pi_G$ are called the mixing proportions or weights. The $f_g(\boldsymbol{Y}_k)$ are called the component densities of the mixture. We will refer to the density (1.1) as a G-component finite mixture distribution. One popular choice of component distribution is to assume $f_g(\boldsymbol{Y})$s have Gaussian distribution, and the corresponding mixture model is called Gaussian mixture model. While there can be infinite number of components, in this study we will be focusing on finite number of mixture distributions, and hereafter we will refer to finite mixture model simply as mixture model, without causing confusion. In this formulation of the mixture model, the number of components G is considered fixed. But of course in any practical application, the value of G is unknown and has to be inferred from the available data,

along with the mixing proportions and the parameters in the specified forms of the component densities.

1.2.2 Interpretation of Mixture Models

The introduction of latent subgroup variable gives rise an intuitive interpretation of mixture model. Let $\boldsymbol{Z}_k$ be a categorical random variable taking on values $1, \ldots, G$ with probabilities $\pi_1, \ldots, \pi_G$, respectively, and suppose that the conditional density of $\boldsymbol{Y}_k$ given $\boldsymbol{Z}_k = i$ is $f_i(\boldsymbol{Y_k}),\ i = 1, \ldots, G$. Then the unconditional density of $\boldsymbol{Y}_k$ will be given by the previous summation equation (1.1). In this context, the variable Z_k can be interpreted as subgroup membership indicator for subject indexed by k. It is distributed according to a multinomial distribution consisting of one draw on G categories

$$Z_k \sim \text{Mult}_G(1, \boldsymbol{\pi}),\ \text{where } \boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)^\intercal. \tag{1.3}$$

Using the interpretation of mixture models above, the G-component mixture model assumption is an intuitive assumption on how is $\{\boldsymbol{Y}_k, k = 1, \ldots, \boldsymbol{Y}_N\}$ generated. Observed vector $\boldsymbol{Y}_k$ is drawn from a population consisting of G subgroups. The proportion of subgroups is given by $\pi_1, \ldots, \pi_G$. If the density of $\boldsymbol{Y}_k$ in group G_i is given by $f_i(\boldsymbol{y}_k)$ for $i = 1, \ldots, g$, then the density of $\boldsymbol{Y}_k$ has the G-component mixture form (1.1). In this situation, the G components of the mixture can be physically identified with the G externally existing groups. The concept of latent subgroups provides opportunity for examining subjects in greater details.

1.3 Function Approximation with B-spline

1.3.1 Function Approximation

Functions are the basic mathematical and statistical tools for describing and analyzing processes of interest. In practice, the need for function approximation on interval $[a, b]$ often arise: some finite amount of data on an unknown function is available, how to construct an approximation of this unknown function on interval $[a, b]$. The most common approach to finding approximations to such unknown functions proceeds in two steps

1. Select a reasonable class of functions, often based on smoothness requirements.

2. Search for plausible candidate within the given class, according to appropriate selection scheme.

The polynomial function class seems a natural candidate, defined in the form of $p(t) = \sum_{i=0}^{d-1} c_i t^i, c_i \in \mathbb{R}$. It can approximate any continuous function on an interval $[a, b]$ uniformly close, and precise convergence rate can be provided for the approximation scheme. But unfortunately it is not flexible enough, suffering from wide oscillation in certain scenarios.

The problem can be solved by extending polynomial function class to the space of piecewise polynomials (of order d and knots $\{\tau_i\}$). Here the term "piecewise" indicates that the function is defined by multiple sub-functions, where each sub-function applies to a different interval in the domain. Let $\{\tau_i\}_{i=0}^{k+1}$ be a sequence of strictly increasing real numbers such that

$$a = \tau_0 < \tau_1 < \ldots < \tau_k < \tau_{k+1} = b, \tag{1.4}$$

then sequence $\{\tau\}$ partitions interval $[a, b]$ into $k + 1$ sub-intervals. And the space spanned by all the polynomial function of order-d associated with $\{\tau\}$

$$\begin{aligned} &\mathcal{PS}_d(\{\tau\}) = \{f : f(t) = p_i(t) \text{ for } t \in I_i, i = 0, 1, \ldots, k\}, \\ &I_i = [\tau_i, \tau_{i+1}) \text{ for } i = 0, 1, \cdots, k-1.\ I_k = [\tau_k, \tau_{k+1}] \\ &p_i(t)\text{'s are polynomials of order-}d, \end{aligned} \tag{1.5}$$

is in fact a linear space of dimension $d + k$ (Schumaker, 1981). In practice, smoothness requirement will be specified, and the most common requirement is $\mathcal{C}^{d-2}$, meaning 0^{th} through $d - 2^{th}$ derivatives are continuous. A function belonging to $\mathcal{PS}_d(\boldsymbol{t}) \bigcap \mathcal{C}^{d-2}[a, b]$, a piecewise polynomial of order d that has continuous derivative of order $d - 2$ is called a polynomial spline of order d with the knots $\boldsymbol{t}$.

For $\mathcal{PS}_d(\{\tau\})$, a linear space of dimension $d + k$, a set of functions $\{B_i\}_{i=1}^{k+d}$ is a basis if and only if all the function belonging to the linear space can be represented as linear combinations of B_i. Suppose we have a collection of 2-dimensional data $\{x_i, y_i\}_{i=1}^n$ and would like to find a function to summarize the relationship between x and y, $f : X \mapsto Y$. Even though we do not know the underlying function, it can always be approximated close enough by a polynomial spline function, and then represented using the spline basis.

The number of knots is conventionally assumed to vary with number of observations n. With sufficient number of knots, any continuous function defined over the close interval $[a, b]$ can be well approximated by a set of polynomial spline functions of fixed order d, known as spline basis functions in $\mathcal{PS}(\boldsymbol{t})$. The approximation rate is well studied for functions with different smoothness in Schumaker (1981). Therefore, given a set of spline basis, approximation of a continuous function is essentially seeking an optimal coefficient vector of same length as the number of spline basis

such that the resulting linear combination achieves the minimal distance from the target continuous function within the polynomial space $\mathcal{PS}(\boldsymbol{t})$.

1.3.2 B-Splines

There are various equivalent basis functions to choose from $\mathcal{PS}(\boldsymbol{t})$, among which the normalized B-spline basis is commonly used and largely preferred due to its fast computation and other appealing features to be mentioned shortly.

Given a sequence of knots $\{\tau_i\}_{i=0}^{k_n+1}$, we define the augmented knot sequence $\boldsymbol{\xi}$ such that

1. additional knots $\xi_1 \leq \xi_2 \leq \cdots \leq \xi_d \leq t_0$,
2. $\xi_{i+d} = t_i, i = 1, \ldots, k_n$,
3. additional knots $t_{k_n+1} \leq \xi_{k_n+d+1} \leq \cdots \leq \xi_{k_n+2d}$.

The actual values of these additional knots beyond the boundary are arbitrary, and it is customary to make them all the same as t_0 and t_{k_n+1}, respectively. The subscript n in k_n indicates that the number of knots is conventionally assumed to vary with number of observations n.

The B-Splines of order 1 is defined as

$$B_{i,1}(x) = \begin{cases} 1, & \text{if } \xi_i \leq x < \xi_{i+1}; \\ 0, & \text{otherwise} \end{cases} \tag{1.6}$$

B-Spline functions of higher order are define by recurrence. Denote, by $B_{i,m}(x)$, the ith normalized B-spline basis function of order m define on knot-sequence $\boldsymbol{\xi}$, $1 \leq m \leq d$. The B-spline basis functions are defined recursively in terms of divided

differences as follows,

$$B_{i,m}(x) = \frac{x - \xi_i}{\xi_{i+m-1} - \xi_i} B_{i,m-1}(x) + \frac{\xi_{i+m} - x}{\xi_{i+m} - \xi_{i+1}} B_{i+1,m-1}(x), \tag{1.7}$$

for $i = 1, \ldots, k_n + d$.

Some desirable properties of the normalized B-spline basis functions are listed below; see Schumaker (1981) for details.

Property 1.3.1. *B-spline basis functions are non-negative with local supports:*

$$B_{i,m} = \begin{cases} > 0, & \xi_i < t < \xi_{i+d}; \\ = 0 & \textit{otherwise.} \end{cases} \tag{1.8}$$

Therefore, on any $I_i, i = 0, 1, \ldots, k_n$, there are at most d non-zero basis functions. This property is also referred to as local property.

Property 1.3.2. *B-spline basis functions form a parition of unity,*

$$\sum_{i=1}^{k_n+d} B_{k,d}(t) = 1, \forall t \in [a, b]. \tag{1.9}$$

This property is also referred to as unity partition property.

Property 1.3.3. *B-spline basis functions are linearly independent. That is, suppose there exists a set of constants $c_1, c_2, \ldots, c_{k_n+d}$ such that*

$$\sum_{k=1}^{k_n+d} c_k b_{k,d}(t) = 0, \forall t \in [a, b], \tag{1.10}$$

then all the c_k's must be zero.

Property 1.3.4. *Let f be a continuous function on $[a, b]$, and have bounded s-th derivative for some $s \geq 1$. Then for any n, there exists a vector $c^* = (c_1^*, c_2^*, \ldots, c_{k_n+d}^*)$ and a constant C_1 such that*

$$\sup_{t\in[a,b]} |f(t) - \sum_{k=1}^{k_n+d} c_k^* b_{k,d}(t)| \leq C_1 k_n^{-s}. \tag{1.11}$$

As mentioned earlier, n denotes the number of observations in data.

Along with fast computation due to recursive definition, the above properties together make B-spline a popular choice among its spline family for function approximation.

1.4 Expectation-Maximization Algorithm

The Expectation-Maximization (Dempster et al., 1977) algorithm is a popular method for parameter estimation in probabilistic models. Often, the data available for model training are incomplete. Missing value can occur, or some value are simply not observable, hidden Markov model being one example. The EM algorithm enables parameter estimation in probabilistic models with incomplete data.

Let γ denote unobserved data, ϕ the parameter of interests and y the observed data. The EM algorithm can be viewed as an iterative method for finding the maximizer of marginal likelihood $p(\phi|y)$, and is extremely useful for many common models for which is hard to maximize $p(\phi|y)$ directly but easy to work with $p(\gamma|\phi, y)$ and $p(\phi|\gamma, y)$. The algorithm proceeds in the following steps.

1. Start with a crude parameter estimate, ϕ^0.

2. For $t = 1, 2, \ldots$:

(a) E-step: Determine the expected log posterior density function,

$$E(log\ p(\gamma, \phi|y)) = \int p(\gamma|\phi^{t-1}, y) log\ p(\gamma, \phi|y) d\gamma,$$

where the expectation averages over the conditional distribution of γ, given the current estimate, ϕ^{t-1}

(b) M-step: Update ϕ. $\phi^t = argmax\ E(log\ p(\gamma, \phi|y))$

In summary, the expectation maximization algorithm alternates between the steps of finding the marginal log-likelihood over completions of missing data given the current model (known as the E-step) and then re-estimating the model parameters using these completions (known as the M-step). The name "E-step" comes from the fact that one does not usually need to form the probability distribution over completions explicitly, but rather need only compute "expected" sufficient statistics over these completions. Similarly, the name "M-step" comes from the fact that model re-estimation can be thought of as 'maximization' of the expected log-likelihood of the data.

As with most optimization methods for non-concave functions, the EM algorithm comes with guarantees only of convergence to a local maximum of the objective function, except in degenerate cases (Gelman et al., 2013). Running the procedure using multiple initial starting parameters is often helpful; similarly, initializing parameters with approximated estimate also helps. With this limited set of tricks, the EM algorithm provides a simple and robust tool for parameter estimation in models with incomplete data. In theory, other numerical optimization techniques, such as gradient descent or Newton-Raphson, could be used instead of expectation maximization; in practice, however, expectation maximization has the advantage of being simple,

robust and easy to implement.

1.5 Fisher Information Matrix and its Approximation

Let $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n]$ be a sequence of n independent but not necessarily identically distributed (i.n.i.d.) random vectors (variables) where each $\boldsymbol{X}_i$ may contain discrete or continuous components. The probability density/mass function of $\boldsymbol{X}_i$, say $p_i(\boldsymbol{x}_i, \boldsymbol{\theta})$, depends on a $p \times 1$ vector of unknown parameters $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{\Theta}$ is a p-dimensional parameter space. Let $\hat{\boldsymbol{\theta}}_n$ be an maximum likelihood estimator (MLE) for $\boldsymbol{\theta}$ based on $\boldsymbol{X}$ and the true value of $\boldsymbol{\theta}$ in the underlying distribution be $\boldsymbol{\theta}^*$. We use the notation t_i to denote the ith component of $\boldsymbol{\theta}$ because we reserve $\hat{\boldsymbol{\theta}}_n$ for the MLEs derived from a sample of size n. The joint probability density/mass function of $\boldsymbol{X}$ is $p(\boldsymbol{x}, \boldsymbol{\theta}) \equiv \prod_{i=1}^{n} p_i(\boldsymbol{x}_i, \boldsymbol{\theta})$. If we denote the negative log-likelihood function as $l(\boldsymbol{\theta}, \boldsymbol{x}) = -\ln p(\boldsymbol{x}, \boldsymbol{\theta})$, the $p \times p$ Fisher information matrix $\boldsymbol{F}_n(\boldsymbol{\theta})$ is defined as:

$$\begin{aligned} \boldsymbol{F}_n(\boldsymbol{\theta}) &\equiv E\left(\frac{\partial \ln p(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \times \frac{\ln p(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\intercal}}\right) \\ &= E\left(\frac{\partial - l(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta}} \times \frac{\partial - l(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta}^{\intercal}}\right) \\ &= E\left(\frac{\partial l(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta}} \times \frac{\partial l(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta}^{\intercal}}\right) \end{aligned} \tag{1.12}$$

where $\boldsymbol{\theta}^{\intercal}$ is the transpose of $\boldsymbol{\theta}$, all expectation are taken with respect to data $\boldsymbol{X}$ and are conditional on the true parameter $\boldsymbol{\theta}^*$. The $p \times p$ Hessian matrix of $l(\boldsymbol{\theta}, \boldsymbol{x})$, $\boldsymbol{H}_n(\boldsymbol{\theta})$,

is defined as the second derivative of $l(\boldsymbol{\theta}, \boldsymbol{x})$ with respect to $\boldsymbol{\theta}$:

$$\boldsymbol{H}_n(\boldsymbol{\theta}) \equiv \frac{\partial^2 l(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal}$$

Computation of $\boldsymbol{F}_n(\boldsymbol{\theta})$ according to its definition in (1.12) in often formidable because it involves direction calculation of expectation of an outer product form. Under some regularity conditions where the interchange of differentiation and integral is valid, $\boldsymbol{F}_n(\boldsymbol{\theta})$ has the following form equivalent to (1.12):

$$\boldsymbol{F}_n(\boldsymbol{\theta}) = E(\boldsymbol{H}_n(\boldsymbol{\theta})), \tag{1.13}$$

Expression (1.13) provides an alternative of computing $\boldsymbol{F}_n(\boldsymbol{\theta})$, which is often more computationally friendly than the definition in (1.12).

Standard statistical theory shows that the $\hat{\boldsymbol{\theta}}$ from either i.i.d or i.n.i.d samples is asymptotically Gaussian under some reasonable conditions (Ljung, 1999, pp. 215–218):

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{dist} N(0, \bar{\boldsymbol{F}}_n(\boldsymbol{\theta}^*)^{-1}), \tag{1.14}$$

where $\bar{\boldsymbol{F}}_n(\boldsymbol{\theta}^*) \equiv \lim_{n\to\infty} \boldsymbol{F}_n(\boldsymbol{\theta}^*)/n$, and the superscript "-1" denotes matrix inverse.

Given the asymptotic normality of MLEs, the problem of constructing confidence intervals reduces largely to the problem of determining the covariance matrix of MLEs. In practical applications, because the true parameter $\boldsymbol{\theta}^*$ is unknown, one of two matrices is commonly used to approximate the covariance matrix of MLE: $\bar{\boldsymbol{F}}_n(\hat{\boldsymbol{\theta}}_n)^{-1}$ or $\bar{\boldsymbol{H}}_n(\hat{\boldsymbol{\theta}}_n)^{-1}$. The first term $\bar{\boldsymbol{F}}_n(\hat{\boldsymbol{\theta}}_n)^{-1}$ is Fisher information matrix evaluated at the MLE. The second term $\bar{\boldsymbol{H}}_n(\hat{\boldsymbol{\theta}}_n)^{-1}$ is Hessian matrix evaluated at MLE. Because $\bar{\boldsymbol{H}}_n(\hat{\boldsymbol{\theta}}_n)^{-1}$ utilizes only the observed $\boldsymbol{y}$ instead of taking expectation over distribution of $y|\boldsymbol{\theta}$ as did in Fisher information matrix, it is referred to as observed

Fisher information matrix.

Efron and Hinkley (1978) demonstrate that for estimating scalar parameter θ, the conditional variance of $\hat{\theta}_n$ is better approximated by $\bar{H}_n(\hat{\theta}_n)^{-1}$ than by $\bar{F}_n(\hat{\theta}_n)^{-1}$. For the more general matrix case, there is not yet a solid theoretical validation for the better choice between $\bar{\boldsymbol{F}}_n(\hat{\boldsymbol{\theta}}_n)^{-1}$ and $\bar{\boldsymbol{H}}_n(\hat{\boldsymbol{\theta}}_n)^{-1}$. In fact, people in practice tend to choose one or the other, depending on which one is easier to obtain for their problems (Cao, 2013).

In the setting of mixture models, the structure of Gaussian log-likelihood often makes $\bar{\boldsymbol{F}}_n(\boldsymbol{\theta}^*)$ difficult to compute. The likelihood is often in the form of

$$L = \frac{p_1 f_1(y)}{\sum p_g f_g(y)} f_1(y) + \ldots + \frac{p_G f_G(y)}{\sum p_g f_g(y)} f_G(y),$$

where $\frac{p_1 f_1(y)}{\sum p_g f_g(y)}$ is the conditional probability of the observed subject is coming from subgroup 1. The second derivative of the log-likelihood will contain terms like

$$\frac{1}{(\sum p_g f_g(y))^2}$$

Fisher information matrix $\bar{\boldsymbol{F}}_n(\hat{\boldsymbol{\theta}}_n)^{-1}$ asks for the expectation of second derivatives. Taking expectation of such terms over Gaussian mixture is a tedious computation process, yet without guaranteed benefit. Instead, researchers often use $\bar{\boldsymbol{H}}_n(\hat{\boldsymbol{\theta}}_n)$ as an approximation of the covariance of MLE (Cavanaugh and Shumway 1996).

Chapter 2

Methodology

2.1 Model Description

Our goal is to study diarrhea effect on growth over time in the NIH study cohort. Most of the children suffer from multiple episodes of diarrhea. Since our assumption is that each episodes of diarrhea may have long lasting effect, at a given time t, all the diarrhea episodes happened before t will have impact on the observed growth outcome. Thus 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.

The observed growth outcomes $y(t)$ is an overlay of natural growth curve without diarrhea plus the effect induced by diarrheas, possibly blurred by measurement errors. For subject k, let $g_k(t)$ be the natural growth curve we should have observed without diarrhea. And let $d_{k,i}(t)$ be the i^{th} diarrhea's effect, with respect to elapsed time. Under such a formulation, the observed growth outcome for subject k can be

represented as

$$y_k(t) = g_k(t) + \sum_{i=1}^{N_k} d_{k,i}(t - l_{k,i}) + \epsilon_k(t), \tag{2.1}$$

where $l_{k,i}$ is the on set time for i^{th} diarrhea episode of subject k, N_k is the total number of diarrhea episodes happened to subject k, and $\epsilon(t)$ is the measurement error.

2.1.1 Multiplicative Curves Assumption

Different assumptions can be made when estimating the functions $\{g_k(t), d_{k,i}(t)\}$. We can assume all these curves are distinct for each subject, for each diarrhea episodes, and then set out to estimate each curve individually. This assumption is the most flexible, providing the largest search space that should accommodate observed data the best. But the introduction of too many free parameters makes the estimation process unstable. We also simply may not have enough data to perform accurate estimation of this many parameters.

If we are dealing with data collected from the cohort, or preliminary observation concludes that diarrhea effects among the population are homogeneous, one viable assumption is that all the curves are shared across subjects or different episodes, i.e., $g_k(t) = g(t), d_{k,i} = d(t)$, for all k and $i \in \{1, 2, \ldots, N_k\}$ (Lin et al., 2020). This model estimates diarrhea effect on a population level. If individual estimation of a specific subject is needed, or the data are collected from several cohort and homogeneity is not guarantee, this model may not work very well.

In fact, what actually happened most likely lies somewhere in the middle. The growth pattern and reaction to diarrhea episodes are different for different subjects, but in the meantime, these patterns share characteristics to certain degree. In this thesis, we attempt to propose a model in the middle ground. We propose that

$g_k(t)$'s share the same shape (temporal dynamics), but the magnitude of the growth is different among subjects. That is, $g_k(t) = s_{k,0} \cdot g(t)$. Here $s_{k,0}$ is a subject specific coefficient that represents personal magnitude of growth. For example, if a specific subject k's growth signal is strong, the corresponding $s_{k,0}$ will be on the larger side compared to group average. Similarly, $d_{k,i}(t) = s_{k,i} \cdot d(t)$, where $d(t)$ is the shared temporal dynamics of diarrhea effect, and $s_{i,k}$ is the subject and episode specific diarrhea multiplier. With introduction of such assumptions, the model is now,

$$y_k(t) = s_{k,0} \cdot g(t) + \sum_{i=1}^{N_k} s_{k,i} \cdot d(t - l_{k,i}) + \epsilon_k(t). \tag{2.2}$$

This model specification takes between-individual differences into consideration, so it can accommodate more data variability. The assumption of common temporal dynamics enables sharing of information across subjects, so we are doing estimation on too many parameters and ending up overfitting. The between-individual differences are modeled by coefficients $s_{k,0}, s_{k,i}$.

2.1.2 Distributional Assumption

We propose mixture of normal distribution for natural growth random coefficients $s_{k,0}$ and diarrhea multiplier $\boldsymbol{s}_{k,-0}$. Here the subscript "-0" in $\boldsymbol{s}_{k,-0}$ means "other than index 0", which are the coefficients associated with diarrhea effects. It is well known that patterns of natural growth has longer-than-normal tails. Depending on the subject, diarrhea effect on growth ranges from mild digestion symptoms to severe malnutrition, this wide spread of diarrhea reaction levels may not be captured by normal distribution well. Moreover, in less developed area where the lack of hygienic living condition and access healthcare makes diarrhea infection prevalent, growth shortfall induced by diarrhea can be predominantly severe, meaning the distribution

is skewed. Mixture of normal distributional assumption is flexible enough to capture the long-tailed and skewed behaviour.

Mixture of normal distributional assumption also introduces the idea of latent subgroups: given an individual, then his/her reaction to diarrhea will be varying around the subgroup average. Interpretation of this will be, subgroup membership individual characteristics, genetic or environmental. It will not be surprising children from household with more educated mother or better living standards will be consistently more resilient to diarrhea effects, indicating they come from the subgroup with better average reaction. By studying the relationship between group membership and the available information on living hygiene level and societal factors, such as access to healthcare, education and population density, we can provide useful insight on better policy making and more efficient interventions.

Suppose there are U subgroups of natural growth and V subgroups of diarrhea effect. The pdf functions are given by,

$$\begin{aligned} p(s_{k,0} = s) &= \sum_{u=1}^{U} p_u \ \varphi(s; \lambda_u, \nu_u^2) \\ p(\boldsymbol{s}_{k,-0} = \boldsymbol{s}) &= \sum_{v=1}^{V} \pi_v \ \varphi(\boldsymbol{s}; \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v). \end{aligned} \tag{2.3}$$

The dimension of $\boldsymbol{s}_{k,-0}$ is determined by the number of diarrhea episodes occurred to subject k, N_k. To ensure the model is scalable with different number of infections N_k, covariance $\boldsymbol{\Sigma}_i$ should have certain repeating patent, such as diagonal, compound symmetry or AR1.

In the mixture model, different normal components correspond to different subgroups of natural growth/diarrhea patent. As mentioned in section 1.2, we introduce a pair of latent class variables $[z_{k,0}, z_{k,1}]$. Indicator of natural growth subgroup

membership $z_{k,0}$ takes value from $1, 2, \ldots, U$. And $z_{k,1}$ indicates which subgroup the diarrhea loading factors $s_{k,i}$ are coming from, with a value range of $1, 2, \ldots, V$.

The normal components of mixture model corresponds to subgroups. Equation (2.3) states that the subject k has p_u possibility to be in the u^{th} growth group, which is a normal distribution parameterized by location λ_u and variance ν_i^2. Likewise, the subject-specific diarrhea response vector $\boldsymbol{s}_{k,-0}$ has π_v chance to be coming from the v^{th} subgroup, a normal distribution with mean vector $\boldsymbol{\mu}_v$ and variance structure $\boldsymbol{\Sigma}_j$.

Measurement error ϵ_k is given the conventional i.i.d normal distribution assumption

$$\boldsymbol{\epsilon}_k \sim \mathcal{N}(0, \text{diag}(\sigma^2)).$$

2.1.3 Model Identifiability Constraints

Additional constraints need to be introduced to make model (2.2) identifiable. Since, for any non-zero real constant r, $g^*(t) = rg(t), s^*_{k,0} = s_{k,0}/r$ balance the equation. Infinite pairs of $(\boldsymbol{M}, \boldsymbol{s})$ are thus equivalent, making the parameters unidentifiable. Without loss of generality, we impose the restriction that expectations of random coefficients $s_{k,0}$ equal to 1

$$\sum_{u=1}^{U} p_u \lambda_u = 1.$$

Similar statement can be made for $\boldsymbol{s}_{k,-0}$, the fisrt entry shoule be bounded by

$$\sum_{v=1}^{V} \pi_v \mu_{v1} = 1.$$

By fixing the mean location of mixture distribution as a whole at 1, the temporal dynamics $g(t)$ can be viewed as the baseline natural growth. Likewise, temporal $d(t)$

is the baseline diarrhea effect overtime.

2.2 Writing the Full Likelihood

2.2.1 Discretization and B-spline approximation of temporal dynamics

Since growth measurement and diarrhea onset are made on specific times, discretization of the model's functional representation is needed. For subject indexed k, let $\{t_i\}_{i=1}^{T_k}$ be the set of measurement times, $\{l_i\}_{i=1}^{N_k}$ the set of diarrhea onset time. Then equation (2.2) is now

$$\begin{bmatrix} y_{k,1} \\ y_{k,2} \\ \vdots \\ y_{k,T_k} \end{bmatrix} = \begin{bmatrix} g(t_1) & d(t_1-l_1) & d(t_1-l_2) & \dots & d(t_1-l_{N_k}) \\ g(t_2) & d(t_2-l_1) & d(t_2-l_2) & \dots & d(t_2-l_{N_k}) \\ & & \vdots & & \\ g(t_{T_k}) & d(t_{T_k}-l_1) & d(t_{T_k}-l_2) & \dots & d(t_{T_k}-l_{N_k}) \end{bmatrix} \begin{bmatrix} s_{k,0} \\ s_{k,1} \\ \vdots \\ s_{k,N_k} \end{bmatrix} + \begin{bmatrix} \epsilon_{k,1} \\ \epsilon_{k,2} \\ \vdots \\ \epsilon_{k,T_k} \end{bmatrix}, \tag{2.4}$$

or using matrix notation

$$\begin{aligned} \boldsymbol{y}_k &= \boldsymbol{M}_k \boldsymbol{s}_k + \boldsymbol{\epsilon}_k, \\ \boldsymbol{M}_k &:= [\boldsymbol{g}(\boldsymbol{t}_k), \boldsymbol{d}(\boldsymbol{t}_{k,1}), ..., \boldsymbol{d}(\boldsymbol{t}_{k,N_k})]. \end{aligned} \tag{2.5}$$

We propose to approximate temporal dynamics functions $g(t)$ and $d(t)$ in interval $[0, T]$ using B-splines, here T is the maximum followup time. Consider a strictly increasing knots sequence $\boldsymbol{u} = \{u_i\}_{i=0}^{N+1}$ that partitions interval $[0, T]$, i.e., $0 = u_0 < u_1 < \ldots < u_{N+1} = T$. Let $\boldsymbol{\alpha} = \{\alpha_1(t), \ldots, \alpha_p(t)\}$ be the set of normalized B-spline basis functions of order m associated with the partition $\boldsymbol{u}$. Note that $p = N + m$,

with N often being referred to as number of interior knots. The number of interior knots used is data-adaptive and can vary with respect to sample size to increase approximation accuracy. The common choices of N are two for linear splines, three for quadratic splines and four for cubic splines. Two popular types of knots are used in practice, equidistant knots and quantile knots. The former is referred to when the knots are equally spaced while quantile knots are selected as the quantiles from the empirical distribution of underlying variable and guarantee that the same number of sample observations fall into each interval. Due to the computation complexity involved in the smoothing parameters, it is however impractical to automatically select all three components simultaneously: basis order, knot placement and number of interior knots. In this thesis, we restrict our scope to quantile knots and fixed order of basis functions similar to Huang *et al.* (2004) and only select the number of interior knots as often done in the literature.

Now $g(t)$ can be approximated by $\boldsymbol{\alpha}$ as below

$$g(t) = \sum_{i=1}^{p} a_i \alpha_i(t) + e_g(t), \tag{2.6}$$

where $\boldsymbol{a} = (a_1, \ldots, a_p)$ is the p dimensional B-spline basis function approximation coefficient vector for $g(t)$ and e_g is the approximation error, which is shown to be uniformly bounded in $[0, T]$ as the number of knots goes to infinity with sample size, described in Property 1.3.4 of B-spline basis functions.

Similarly, we can construct partition $\boldsymbol{v}$ of $[0, T]$, associated B-spline basis $\{\beta_i\}_{i=1}^{q}$ such that $d(t)$ can be approximated as

$$d(t) = \sum_{i=1}^{q} b_i \beta_i(t) + e_d(t). \tag{2.7}$$

It's worth noting that, $t - l$ maybe negative, and consequently $d(t - l)$ is ill-defined. But to keep writings simple, we keep using such notation, and extend the domain of $d, \boldsymbol{\alpha}, \boldsymbol{\beta}$

$$\begin{cases} d(t-l) = 0, & \text{if } t - l < 0 \\ \alpha_i(t-l) = 0, & \text{if } t - l < 0, \forall i \\ \beta_j(t-l) = 0, & \text{if } t - l < 0, \forall j. \end{cases} \tag{2.8}$$

After the introduction of B-spline approximation, matrix $\boldsymbol{M}_k$ is now

$$\boldsymbol{M}_k = \begin{bmatrix} \sum_{i=1}^{p} a_i \alpha_i(t_1) & \sum_{i=1}^{q} b_i \beta_i(t_1 - l_1) & \dots & \sum_{i=1}^{q} b_i \beta_i(t_1 - l_{N_k}) \\ & \vdots & & \\ \sum_{i=1}^{p} a_i \alpha_i(t_{T_k}) & \sum_{i=1}^{q} b_i \beta_i(t_{T_k} - l_1) & \dots & \sum_{i=1}^{q} b_i \beta_i(t_{T_k} - l_{N_k}) \end{bmatrix} = \boldsymbol{X}_k \boldsymbol{C}_k. \tag{2.9}$$

Temporal dynamics matrix $\boldsymbol{X}_k$ is

$$\boldsymbol{X}_k = \begin{bmatrix} \boldsymbol{A}_k & \boldsymbol{B}_{k,1} & \dots & \boldsymbol{B}_{k,N_k} \end{bmatrix}, \tag{2.10}$$

where

$$\boldsymbol{A}_k = \begin{bmatrix} \alpha_1(t_1) & \dots & \alpha_p(t_1) \\ \vdots & \vdots & \vdots \\ \alpha_1(t_{T_k}) & \dots & \alpha_p(t_{T_k}) \end{bmatrix}, \ \boldsymbol{B}_{k,i} = \begin{bmatrix} \beta_1(t_1 - l_i) & \dots & \beta_q(t_1 - l_i) \\ \vdots & \vdots & \vdots \\ \beta_1(t_{T_k} - l_i) & \dots & \beta_q(t_{T_k} - l_i). \end{bmatrix} \tag{2.11}$$

The matrix $\boldsymbol{X}_k$ is known given the B-spline basis, measurement times $\{t_i\}_{i=1}^{T_k}$ and diarrhea on set times $\{l_i\}_{i=1}^{N_k}$.

The coefficient matrix $\boldsymbol{C}_k$ is:

$$\boldsymbol{C}_k = \begin{bmatrix} \boldsymbol{a} & \dots & \dots & \dots \\ \vdots & \boldsymbol{b} & \dots & \dots \\ \vdots & \vdots & \ddots & \dots \\ \vdots & \vdots & \vdots & \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \otimes \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{I}_{N_k} \end{bmatrix} \otimes \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_q \end{bmatrix}. \tag{2.12}$$

Now that we have our assumptions fully stated, the full likelihood of observed data can be constructed.

2.2.2 Full Likelihood

Even though random multiplier $\boldsymbol{s}_k$ and subgroup membership $\boldsymbol{z}_k$ is not observed, they are treated as if they were in the full likelihood. Based on model (2.2) and the distribution specified in (2.3), the full likelihood can be written down as

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{Z}; \boldsymbol{\theta}) &= \prod_k p(\boldsymbol{y}_k | \boldsymbol{s}_k, \sigma^2) p(\boldsymbol{s}_k | \boldsymbol{z}_k) p(\boldsymbol{z}_k) \\ &= \prod_k \varphi(\boldsymbol{y}_k; \boldsymbol{X}_k \boldsymbol{C}_k, \Sigma_k) \varphi(s_{k,0}; \lambda_{z_{k,0}}, \nu^2_{z_{k,0}}) \varphi(\boldsymbol{s}_{k,-0}; \boldsymbol{\mu}_{z_{k,1}}, \boldsymbol{\Sigma}_{z_{k,1}}) p_{z_{k,0}} \pi_{z_{k,1}}, \end{aligned}$$

where $\boldsymbol{Y} = \{\boldsymbol{y}_k\}, \boldsymbol{S} = \{\boldsymbol{s}_k\}$ and $\boldsymbol{Z} = \{\boldsymbol{z}_k\}$. Vector $\boldsymbol{\theta}$ denotes all the parameter to be estimated, namely $\{\sigma^2, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu}^2, \boldsymbol{p}, \boldsymbol{\pi}\}$. The log likelihood would be

$$\begin{aligned} log\ p(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{Z}; \boldsymbol{\theta}) = \sum_{k=1}^{N} & \log \varphi(\boldsymbol{y}_k; \boldsymbol{X}_k \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \otimes \boldsymbol{a} + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{I}_{N_k} \end{bmatrix} \otimes \boldsymbol{b} \right) \boldsymbol{s}_k, \boldsymbol{\Sigma}_k) \\ & + \log \varphi(s_{k,0}; \lambda_{z_{k,0}}, \nu^2_{z_{k,0}}) + \log \varphi(\boldsymbol{s}_{k,-0}; \boldsymbol{\mu}_{z_{k,1}}, \boldsymbol{\Sigma}_{z_{k,1}}) \\ & + \log p_{z_{k,0}} + \log \pi_{z_{k,1}}. \end{aligned} \tag{2.13}$$

Next, the EM algorithm will be performed to maximize the marginal log likelihood. The corresponding maximizer is the parameter estimate.

2.3 Estimation by Expectation-Maximization

Since latent variable $\boldsymbol{Z}$ and personal loading factor $\boldsymbol{s}$ in the complete log likelihood cannot be observed, estimation of parameters $\boldsymbol{\theta}$ need to be carried out by maximizing the marginal log likelihood $log\ p(\boldsymbol{Y};\boldsymbol{\theta})$. The EM algorithm is a perfect candidate for such a task, since all the conditional distribution involved are normal. In this section, we present the derived iteration procedure of the exact EM algorithm that provides explicit E-step and M-step.

2.3.1 Expectation-step

At the E-step, we derive the expectation of full log likelihood by integrating it over the conditional distribution $\boldsymbol{S}, \boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta}^{(t)}$.

Given the hidden state $[z_{k,0}, z_{k,1}] = [z_{k,0}^{(r)}, z_{k,1}^{(r)}]$, here the superscript r stands for realization, $\boldsymbol{s}$ follows a multivariate normal distribution

$$\boldsymbol{s}_k|\boldsymbol{z}_k = \boldsymbol{z}_k^{(r)}, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_k^{(r)}, \boldsymbol{\Gamma}_k^{(r)}), \tag{2.14}$$

where $\boldsymbol{\mu}_k^{(r)} = [\lambda_{z_{k,0}^{(r)}}, \mu_{z_{k,1}^{(r)}}, \ldots, \mu_{z_{k,1}^{(r)}}]$, $\boldsymbol{\Gamma}_k^{(r)} = \text{diag}(\nu^2_{z_{k,0}^{(r)}}, \sigma^2_{z_{k,1}^{(r)}}, \ldots, \sigma^2_{z_{k,1}^{(r)}})$. We can obtain be conditional distribution of $\boldsymbol{y}_k$ as

$$\boldsymbol{y}_k|\boldsymbol{z}_k = \boldsymbol{z}_k^{(r)}, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{X}_k\boldsymbol{C}_k\boldsymbol{\mu}_k^{(r)}, \boldsymbol{X}_k\boldsymbol{C}_k\boldsymbol{\Gamma}_k^{(r)}\boldsymbol{C}_k^\intercal\boldsymbol{X}_k^\intercal + \boldsymbol{\Sigma}_k), \tag{2.15}$$

Then using the Bayes rule, the posterior subgroup classification probability given

observed data is

$$\begin{aligned} p(\boldsymbol{z}_k = \boldsymbol{z}_k^{(r)}|\boldsymbol{y}_k, \boldsymbol{\theta}) &= \frac{p(\boldsymbol{y}_k|\boldsymbol{z}_k = \boldsymbol{z}_k^{(r)})p(\boldsymbol{z}_k = \boldsymbol{z}_k^{(r)})}{p(\boldsymbol{y}_k)} \\ &= \frac{p(\boldsymbol{y}_k|\boldsymbol{z}_k = \boldsymbol{z}_k^{(r)})p(\boldsymbol{z}_k = \boldsymbol{z}_k^{(r)})}{\sum_{i=[1,1]}^{[U,V]} p(\boldsymbol{y}_k|\boldsymbol{z}_k = i)p(\boldsymbol{z}_k = i)} \\ &= \frac{g_k(\boldsymbol{z}_k^{(r)})}{\sum_i g_k(i)}, \end{aligned} \tag{2.16}$$

where

$$g_k(\boldsymbol{z}_k^{(r)}) = p_{z_{k,0}^{(r)}} \pi_{z_{k,1}^{(r)}} \varphi(\boldsymbol{y}_k; \boldsymbol{X}_k \boldsymbol{C}_k \boldsymbol{\mu}_k^{(r)}, \boldsymbol{X}_k \boldsymbol{C}_k \boldsymbol{\Gamma}_k^{(r)} \boldsymbol{C}_k^\intercal \boldsymbol{X}_k^\intercal + \boldsymbol{\Sigma}_k).$$

Similarly, given the hidden state $\boldsymbol{z}_k$, model (2.5) is now a simple normal additive model, the conditional probability $P(\boldsymbol{s}_k|\boldsymbol{y}_k, \boldsymbol{z}_k = \boldsymbol{z}_k^{(r)}, \boldsymbol{\theta})$ is found by

$$\begin{aligned} \boldsymbol{s}_k|\boldsymbol{y}_k, \boldsymbol{\theta}, \boldsymbol{z}_k = \boldsymbol{z}_k^{(r)} &\sim \mathcal{N}(\boldsymbol{\omega}_k(\boldsymbol{z}_k^{(r)}), \boldsymbol{\Omega}_k(\boldsymbol{z}_k^{(r)})) \\ \boldsymbol{\Omega}_k(\boldsymbol{z}_k^{(r)}) &= \left[(\boldsymbol{X}_k \boldsymbol{C}_k)^\intercal \boldsymbol{\Sigma}_k^{-1} \boldsymbol{X}_k \boldsymbol{C}_k + {\boldsymbol{\Gamma}_k^{(r)}}^{-1} \right]^{-1} \\ \boldsymbol{\omega}_k(\boldsymbol{z}_k^{(r)}) &= \boldsymbol{\Omega}_k(\boldsymbol{z}_k^{(r)}) \left[(\boldsymbol{X}_k \boldsymbol{C}_k)^\intercal \boldsymbol{\Sigma}_k^{-1} \boldsymbol{y}_k + {\boldsymbol{\Gamma}_k^{(r)}}^{-1} \boldsymbol{\mu}_k^{(r)} \right]. \end{aligned} \tag{2.17}$$

Combining (2.16) and (2.17), we show that

$$p(\boldsymbol{s}_k|\boldsymbol{y}_k, \boldsymbol{\theta}) = \sum_{z=[1,1]}^{[U,V]} \frac{g_k(z)}{\sum_i g_k(i)} \varphi(\boldsymbol{s}_k; \boldsymbol{\omega}_k(z), \boldsymbol{\Omega}_k(z)). \tag{2.18}$$

Based on the conditional distribution, the expectation of the full log likelihood is found to be

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E_{\boldsymbol{S},\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}^{(t)}} \left[\log p(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{Z}; \boldsymbol{\theta}) \right] \\ &= Q_1(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + Q_2(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + Q_3(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \end{aligned} \tag{2.19}$$

with the terms being

$$
\begin{aligned}
Q_1(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E_{\boldsymbol{S}|\boldsymbol{Y},\boldsymbol{\theta}^{(t)}}\left[\sum_{k=1}^{N}\log\varphi(\boldsymbol{y}_k;\boldsymbol{X}_k\boldsymbol{C}_k\boldsymbol{s}_k,\boldsymbol{\Sigma}_k)\right] \\
&= -\frac{1}{2}\sum_{k=1}^{N}\Bigg[\log|\boldsymbol{\Sigma}_k| + \boldsymbol{y}_k^{\intercal}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{y}_k - 2\boldsymbol{y}_k\boldsymbol{\Sigma}_k^{-1}\boldsymbol{X}_k\boldsymbol{C}_k E(\boldsymbol{s}_v|\boldsymbol{y}_v,\boldsymbol{\theta}^{(t)}) \\
&\qquad + \mathrm{Tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{X}_k\boldsymbol{C}_k E(\boldsymbol{s}_k\boldsymbol{s}_k^{\intercal}|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\boldsymbol{C}_k^{\intercal}\boldsymbol{X}_k^{\intercal})\Bigg],
\end{aligned}
$$

$$
\begin{aligned}
Q_2(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E_{\boldsymbol{S},\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}^{(t)}}\left[\ \log\varphi(s_{k,0};\lambda_{z_{k,0}},\nu^2_{z_{k,0}}) + \log\varphi(s_{k,-0};\mu_{z_{k,1}},\boldsymbol{\Sigma}_{z_{k,1}})\right] \\
&= -\frac{1}{2}\sum_{k=1}^{N}\Bigg\{\sum_{i=1}^{U}p(z_{k,0}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\Bigg[\log\nu_i^2 + \frac{1}{\nu_i^2}\Big(E(s_{k,0}^2|z_{k,0}=i,\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}) \\
&\qquad - 2\lambda_i E(s_{k,0}|z_{k,0}=i,\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}) + \lambda_i^2\Big)\Bigg] \\
&\quad + \sum_{i=1}^{V}p(z_{k,1}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\sum_{j=1}^{N_k}\Bigg[\log\sigma_i^2 + \frac{1}{\sigma_i^2}\Big(E(s_{k,j}^2|z_{k,1}=i,\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}) \\
&\qquad - 2\mu_i E(s_{k,j}|z_{k,1}=i,\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}) + \mu_i^2\Big)\Bigg]\Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
Q_3(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}^{(t)}}\left[\sum_{k=1}^{N}\left(\log p_{z_{k,0}} + \log\pi_{z_{k,1}}\right)\right] \\
&= \sum_{k=1}^{N}\left[\sum_{i=1}^{U}\log p_i\ p(z_{k,0}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}) + \sum_{i=1}^{V}\log\pi_i\ p(z_{k,1}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\right].
\end{aligned}
\tag{2.20}
$$

The posterior marginal probabilities in (2.20) can be calculated as

$$
\begin{aligned}
p(z_{k,0}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}) &= \sum_{j=1}^{V}p(\boldsymbol{z}_k=(i,j)|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}), \\
p(z_{k,1}=j|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^{U}p(\boldsymbol{z}_k=(i,j)|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}),
\end{aligned}
$$

and the posterior conditional expectations are given as follows

$$E(s_{k,0}|z_{k,0}=i,\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}) = \sum_{j=1}^{V} E(s_{k,0}|\boldsymbol{z}_k=(i,j),\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})p(\boldsymbol{z}_k=(i,j)|z_{k,0}=i,\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})$$
$$= \sum_{j=1}^{V} E(s_{k,0}|\boldsymbol{z}_k=(i,j),\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\frac{p(\boldsymbol{z}_k=(i,j)|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})}{\sum_j p(\boldsymbol{z}_k=(i,j)|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})}.$$

Now that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is obtained, $\boldsymbol{\theta}$ of next iteration $t+1$ will be found by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$.

2.3.2 Maximization step

In the maximization step, the EM algorithm proceeds with updating $\boldsymbol{\theta}$ estimate by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$. Explicit maximizer can be obtained. More specifically, parameters can by updated by solving corresponding score functions. Measurement error σ^2 in the next iteration can be shown to be

$$\sigma^{2(t+1)} = \frac{\sum_{k=1}^{N}\left[\boldsymbol{y}_k^\intercal\boldsymbol{y}_k - 2\boldsymbol{y}_k^\intercal\boldsymbol{X}_k\boldsymbol{C}_k^{(t)}E(\boldsymbol{s}_k|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}) + \mathrm{Tr}(\boldsymbol{X}_k\boldsymbol{C}_kE(\boldsymbol{s}_k\boldsymbol{s}_k^\intercal|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\boldsymbol{C}_k^\intercal\boldsymbol{X}_k^\intercal)\right]}{\sum_{k=1}^{N}T_k}. \tag{2.21}$$

Iteration relationship regarding $\boldsymbol{a},\boldsymbol{b}$ can be obtained by taking derivatives of Q_1. Since $\boldsymbol{C}_k$ is a structured matrix, its entries are not "free paremeters". Score functions of $\boldsymbol{a},\boldsymbol{b}$ need to be derived using

$$\begin{aligned}\frac{\partial Q_1}{\partial a_i} &= \sum_{k=1}^{N}\sum_{xy}\frac{\partial Q_1}{\partial C_{k,xy}}\frac{\partial C_{k,xy}}{a_i},\\ \frac{\partial Q_1}{\partial b_j} &= \sum_{k=1}^{N}\sum_{xy}\frac{\partial Q_1}{\partial C_{k,xy}}\frac{\partial C_{k,xy}}{b_j}.\end{aligned} \tag{2.22}$$

The resulted equations are

$$\begin{aligned}\left[\sum_k A_k^\intercal A_k E(s_{k,0}\cdot s_{k,0}|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\right]\boldsymbol{a}+\left[\sum_k\sum_i A_k^\intercal B_{k,i}E(s_{k,0}\cdot s_{k,i}|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\right]\boldsymbol{b}\\ =\sum_k A_k^\intercal \boldsymbol{y}_k E(s_{k,0}|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}),\\ \left[\sum_k\sum_i B_{k,i}^\intercal A_k E(s_{k,i}\cdot s_{k,0}|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\right]\boldsymbol{a}+\left[\sum_k\sum_i\sum_j B_{k,i}^\intercal B_{k,j}E(s_{k,i}\cdot s_{k,j}|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\right]\boldsymbol{b}\\ =\sum_k\sum_i B_{k,i}^\intercal \boldsymbol{y}_k E(s_{k,i}|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}).\end{aligned} \tag{2.23}$$

Adapting a block matrix denotation can make things clearer

$$\begin{bmatrix}M_1 & M2\\ M_3 & M4\end{bmatrix}\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix}=\begin{bmatrix}M_5\\ M_6\end{bmatrix},\ \text{thus}\ \begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix}=\begin{bmatrix}M_1 & M2\\ M_3 & M4\end{bmatrix}^{-1}\begin{bmatrix}M_5\\ M_6\end{bmatrix}$$

,

$$\begin{aligned}M_1&=\sum_k A_k^\intercal A_k E(s_{k,0}\cdot s_{k,0}|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}),\\ M_2&=\sum_k\sum_i A_k^\intercal B_{k,i}E(s_{k,0}\cdot s_{k,i}|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}),\\ M_3&=\sum_k\sum_i B_{k,i}^\intercal A_k E(s_{k,i}\cdot s_{k,0}|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}),\\ M_4&=\sum_k\sum_i\sum_j B_{k,i}^\intercal B_{k,j}E(s_{k,i}s_{k,j}|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}),\\ M_5&=\sum_k A_k^\intercal \boldsymbol{y}_k E(s_{k,0}|\boldsymbol{y}_v,\boldsymbol{\theta}^{(t)}),\\ M_6&=\sum_k\sum_i B_{k,i}^\intercal \boldsymbol{y}_k E(s_{k,i}|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)}).\end{aligned} \tag{2.24}$$

We can also derive the following explicit estimates for the parameters of normal

components

$$
\begin{aligned}
\lambda_i^{(t+1)} &= \frac{\sum_{k=1}^{N}\left[p(z_{k,0}=i|\boldsymbol{y}_v,\boldsymbol{\theta}^{(t)})E(s_{k,0}|z_{k,0}=i,\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\right]}{\sum_{k=1}^{N}p(z_{k,0}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})},\\
\mu_i^{(t+1)} &= \frac{\sum_{k=1}^{N}\left[p(z_{k,1}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\sum_{j=1}^{N_k}E(s_{k,j}|z_{k,1}=i,\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\right]}{\sum_{k=1}^{N}N_k\cdot p(z_{k,1}=i|\boldsymbol{y}_v,\boldsymbol{\theta}^{(t)})},\\
\nu_i^{(t+1)} &= \frac{\sum_{k=1}^{N}\left[p(z_{k,0}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})E(s_{k,0}^2|z_{k,0}=i,\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\right]}{\sum_{k=1}^{N}p(z_{k,0}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})}-[\lambda_i^{(t+1)}]^2,\\
\delta_i^{(t+1)} &= \frac{\sum_{k=1}^{N}\left[p(z_{k,1}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\sum_{j=1}^{N_k}E(s_{k,j}^2|z_{k,1}=i,\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})\right]}{\sum_{k=1}^{N}N_k\cdot p(z_{k,1}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})}-[\mu_i^{(t+1)}]^2.
\end{aligned}
\tag{2.25}
$$

Iteration relationship of mixing probability of normal distribution can be found by taking derivatives of Q_3

$$
\begin{aligned}
p_i^{(t+1)} &= \frac{\sum_{k=1}^{N}p(z_{k,0}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})}{N}\\
\pi_i^{(t+1)} &= \frac{\sum_{k=1}^{N}p(z_{k,1}=i|\boldsymbol{y}_k,\boldsymbol{\theta}^{(t)})}{N}
\end{aligned}
\tag{2.26}
$$

So far we have established the EM algorithm iteration relationship for our proposed model, connecting parameter estimate of iteration t and iteration $t+1$. In summary, the EM algorithm alternates between the steps of guessing a probability distribution over completions of subgroup membership and random multiplier given the current parameter estimate (known as the E-step) and then re-estimating the model parameters using these completions (known as the M-step). Start with initial guess $\boldsymbol{\theta}^{(0)}$, and iterate until convergence.

2.4 Model Selection

In the process of deriving EM algorithm, the numbers of components U, V in mixture distribution (2.3) are considered fixed constants. While in practice, they correspond to number of growth patterns, or number of diarrhea vulnerability clusters respectively, and need to be selected. Identifying the number of latent subgroups in both natural growth and diarrhea effects is of great practical interest. The number of Bspline knots used p, q also have substantial impact on estimation outcome.

- U: number of natural growth subgroups,
- V: number of diarrhea vulnerability subgroups,
- p: number of inner knots used in the B-spline basis for estimating $g(t)$,
- q: number of inner knots used in the B-spline basis for estimating $d(t)$

We proposed to select U, V, p, q based on the Bayesian information criterion (Schwarz 1978). A grid search will be performed on $\{U, V, p, q\}$. After estimating model parameters using the EM algorithm at each combination of U, V, p, q, the BIC is computed. The model with the lowest BIC will be selected.

The model proposed in this thesis is a mixed-effect model. Bsplines in the model are fixed effects, and random multipliers are random effects. The BIC criterion for mixed effect model (Delattre et al., 2014) is define as

$$\begin{aligned} BIC_{mixed} &= -2\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + dim(\boldsymbol{\theta}_R)\ \log N + dim(\boldsymbol{\theta}_F)\ \log n_{tot} \\ \dim(\boldsymbol{\theta}_R) &= 3(U+V) - 4 \\ \dim(\boldsymbol{\theta}_F) &= p + q + 1. \end{aligned} \tag{2.27}$$

The dimensionality associated with random effect is $3(U + V) - 2$, as each latent

subgroup is parameterized by mixture proportion p_u, subgroup mean λ_u and subgroup variance ν_u^2, resulting in $3(U+V)$ dimensions. There parameters are restricted by summation of mixture proportion equals to 1 (two constrains, one for natural growth and one for diarrhea), and two identifiability constrains. So the number of free parameters is $3(U+V)-4$. The random effects are penalized by a factor of $\log N$, N is the number of subjects.

The dimensionality associated with fixed effect is $p+q+1$, p for length of B-spline coefficients $\boldsymbol{a}$ estimating natural growth $g(t)$, q for length of $\boldsymbol{b}$ estimating diarrhea effect $d(t)$, measurement error σ^2 is another fixed effect. The fixed effects are penalized by a factor of $\log n_{tot} = \sum_k T_k$, total number of observations.

2.5 Asymptotic Variance of the Estimator

The Expectation-Maximization algorithm is used to obtain the maximum likelihood estimates of parameters. It maximize the marginal likelihood, integrated over latent data which was never intended to be observed in the first place. Given the asymptotic normality of MLE, the problem of finding asymptotic variance of the estimator reduces largely to the problem of determining the Fisher information matrix.

In this section we demonstrate the derived Fisher information matrix in the special case of $U = V = 1$. In this case, both natural growth and diarrhea multiplier follows a normal distribution. When $U \neq 1$ or $V \neq 1$, the structure of mixture of normal distribution makes it difficult to compute the Fisher information matrix, so we use the observed Fisher information matrix as an approximation. The observed Fisher information matrix for the general case of $U \neq 1$ or $V \neq 1$ has a more complex form, the results can be found in Chapter 6.

From (2.17), the marginal distribution of $\boldsymbol{y}_k$ is multivariate normal

$$\begin{aligned}
\boldsymbol{y}_k &\sim N(\boldsymbol{XC\mu}, \boldsymbol{V}) \\
\boldsymbol{V} &= \boldsymbol{XC\Gamma}_{u,v}(\boldsymbol{XC})^\intercal + \boldsymbol{\Sigma} \\
\boldsymbol{\mu} &= [\lambda, \mu, \ldots, \mu]^\intercal \\
\boldsymbol{\Gamma} &= \text{diag}(\nu^2, \sigma^2, \ldots, \sigma^2) \\
\boldsymbol{\Sigma} &= \text{diag}(\sigma_0^2, \ldots, \sigma_0^2),
\end{aligned}$$

The log-likelihood for $\boldsymbol{y_k}$ is

$$\log\varphi_k(\boldsymbol{y}_k|\boldsymbol{\mu}, \boldsymbol{V}) = -\frac{1}{2}\log|2\pi V| - \frac{1}{2}(y - M\mu)^\intercal V^{-1}(y - M\mu).$$

And the Fisher information matrix, with respect to Bspline coefficients $\boldsymbol{a}, \boldsymbol{b}$, random multiplier parameters ν^2, σ^2 and measurement error σ_0^2, for $\log\varphi_k$ is

$$\boldsymbol{F}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix}
E\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \boldsymbol{a}^\intercal} & E\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \boldsymbol{b}^\intercal} & E\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \nu^2} & E\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \sigma^2} & E\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \sigma_0^2} \\
 & E\frac{\partial^2 l}{\partial \boldsymbol{b}\partial \boldsymbol{b}^\intercal} & E\frac{\partial^2 l}{\partial \boldsymbol{b}\partial \nu^2} & E\frac{\partial^2 l}{\partial \boldsymbol{b}\partial \sigma^2} & E\frac{\partial^2 l}{\partial \boldsymbol{b}\partial \sigma_0^2} \\
 & & E\frac{\partial^2 l}{\partial \nu^2\partial \nu^2} & E\frac{\partial^2 l}{\partial \nu^2\partial \sigma^2} & E\frac{\partial^2 l}{\partial \nu^2\partial \sigma_0^2} \\
 & & & E\frac{\partial^2 l}{\partial \sigma^2\partial \sigma^2} & E\frac{\partial^2 l}{\partial \sigma^2\partial \sigma_0^2} \\
 & & & & E\frac{\partial^2 l}{\partial \sigma_0^2\partial \sigma_0^2}
\end{bmatrix}.$$

where the blocks are

$$
\begin{aligned}
E(\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \boldsymbol{a}^\intercal}) &= \nu^4 A^\intercal V^{-1} A \boldsymbol{a}\boldsymbol{a}^\intercal A^\intercal V^{-1} A + \nu^4 (\boldsymbol{a}^\intercal A^\intercal V^{-1} A \boldsymbol{a}) A^\intercal V^{-1} A + \mu_0^2 A^\intercal V^{-1} A \\
E(\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \boldsymbol{b}^\intercal}) &= \nu^2\sigma^2 A^\intercal V^{-1} A \boldsymbol{a}\boldsymbol{b}^\intercal B^\intercal V^{-1} B + \nu^2\sigma^2 (\boldsymbol{a}^\intercal A^\intercal V^{-1} B \boldsymbol{b}) A^\intercal V^{-1} B + \mu_0\mu_1 A^\intercal V^{-1} B \\
E(\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \nu^2}) &= \nu^2 (\boldsymbol{a}^\intercal A^\intercal V^{-1} A \boldsymbol{a}) A^\intercal V^{-1} A \boldsymbol{a} \\
E(\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \sigma^2}) &= \nu^2 (\boldsymbol{a}^\intercal A^\intercal V^{-1} B \boldsymbol{b}) A^\intercal V^{-1} B \boldsymbol{b} \\
E(\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \sigma_0^2}) &= \nu^2 A^\intercal V^{-1} V^{-1} A \boldsymbol{a} \\
E(\frac{\partial^2 l}{\partial \nu^2 \partial \nu^2}) &= 0.5 (\boldsymbol{a}^\intercal A^\intercal V^{-1} A \boldsymbol{a})^2 \\
E(\frac{\partial^2 l}{\partial \sigma^2 \partial \sigma^2}) &= 0.5 (\boldsymbol{b}^\intercal B^\intercal V^{-1} B \boldsymbol{b})^2 \\
E(\frac{\partial^2 l}{\partial \sigma_0^2 \partial \sigma_0^2}) &= 0.5 \mathrm{Tr}(V^{-1} V^{-1}) \\
E(\frac{\partial^2 l}{\partial \nu^2 \partial \sigma^2}) &= 0.5 (\boldsymbol{a}^\intercal A^\intercal V^{-1} B \boldsymbol{b})^2 \\
E(\frac{\partial^2 l}{\partial \nu^2 \partial \sigma_0^2}) &= 0.5 \mathrm{Tr}(V^{-1} A \boldsymbol{a}\boldsymbol{a}^\intercal A^\intercal V^{-1}) \\
E(\frac{\partial^2 l}{\partial \sigma^2 \partial \sigma_0^2}) &= 0.5 \mathrm{Tr}(V^{-1} B \boldsymbol{b}\boldsymbol{b}^\intercal B^\intercal V^{-1}).
\end{aligned}
\tag{2.28}
$$

All the expectations are evaluated at the parameter estimate $\hat{\boldsymbol{\theta}}$.

Standard statistical theory shows that the $\hat{\boldsymbol{\theta}}$ from i.n.i.d samples is asymptotically normal under some reasonable conditions (Ljung, 1999, pp. 215–218)

$$
\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{dist} N(0, \bar{\boldsymbol{F}}_n(\boldsymbol{\theta}^*)^{-1}), \tag{2.29}
$$

The estimated covariance matrix of parameter esimate is then $\boldsymbol{F}(\hat{\boldsymbol{\theta}})^{-1}$. The asymptotic covariance matrix enable us to perform inference on parameter of interests, such as finding the 95% pointwise confidence interval of estimated diarrhea effect

curve.

2.6 Model Variants

The model we proposed in Section 2.1,

$$y_k(t) = s_{k,0} \cdot g(t) + \sum_{i=1}^{N_k} s_{k,i} \cdot d(t - l_{k,i}) + \epsilon_k(t),$$

assumes that growth curve of each child is only affected by diarrhea and that those diarrhea episodes have the similar but not necessarily the same pattern of effect. In reality, the size of available data affects the choice of model assumption. When not enough samples are collected to estimate the differences between diarrhea episodes of the same subject, regularization should be introduced and restrict the degree of freedom so that the model is not overfitting. On the other hand, many factors could affect growth or the diarrhea process. Even though it is reasonable to assume those other factors can be averaged out when estimating diarrhea effect marginally, when enough samples are available, including additional covariates can improve the efficiency of estimation. In this section, we present two model variants to show how our model can be adjusted in different scenarios.

2.6.1 Fixed Shape Diarrhea Effect within Individual

The model we proposed in Section 2.1 assumes natural growth differs among children and a given subject's diarrhea episodes has different random multiplier. While the general assumption allows for different shapes of curve for all natural growth and all diarrhea effects over time, estimating those differences requires more data. When not enough samples are available, regularize the model by fixing $s_{k,i} \equiv s_{i,1}$ can be

a viable option to prevent overfitting. In this case, diarrhea effect $d_k(t)$ is subject dependent, but within the subject, diarrhea effects from different episodes have the same pattern over time:

$$y_k(t) = s_{k,0} \cdot g(t) + s_{k,1} \sum_{i=1}^{N_k} d(t - l_{k,i}) + \epsilon_k(t)$$

The proposed methods and estimation procedure still apply in this scenario. We can approximate the curves of interest in the same way

$$g(t) = \sum_{i=1}^{p} a_i \alpha_i(t) + e_g(t),$$
$$d(t) = \sum_{j=1}^{q} b_j \beta_j(t) + e_d(t).$$

The model then becomes

$$\begin{aligned} y_k(t) &= s_{k,0} \cdot g(t) + s_{k,1} \sum_{n=1}^{N_k} \cdot d(t - l_{k,n}) + \epsilon_k(t) \\ &= s_{k,0} \cdot \sum_{i=1}^{p} a_i \alpha_i(t) + s_{k,1} \sum_{j=1}^{q} \cdot b_j [\sum_{n=1}^{N_k} \beta_j(t - l_{k,n})] + \epsilon_k(t). \end{aligned}$$

The matrix representation is

$$\begin{aligned} \boldsymbol{y}_k &= \boldsymbol{M}_k^* \boldsymbol{s}_k + \boldsymbol{\epsilon}_k \\ \boldsymbol{M}_k^* &: = [\boldsymbol{g}(\boldsymbol{t}_k), \boldsymbol{d}(\boldsymbol{t}_{k,1}) + ... + \boldsymbol{d}(\boldsymbol{t}_{k,N_k})] \\ \boldsymbol{s}_k &= [s_{k,0}, s_{k,1}]^\intercal. \end{aligned}$$

Noticing that

$$\boldsymbol{M}_k^* = [\boldsymbol{A}_k, \sum_n \boldsymbol{B}_{k,n}]\boldsymbol{C}_k = [\boldsymbol{A}_k, \boldsymbol{B}_{k,1}^*] \begin{bmatrix} \boldsymbol{a} & 0 \\ 0 & \boldsymbol{b} \end{bmatrix},$$

estimation of the regularized model can be still achieved by the EM algorithm proposed in Section 2.3. The regularized model can be viewed as if only one diarrhea episode occurred and the corresponding $\boldsymbol{B}_{k,1}^*$ matrix is computed by the sum of $\boldsymbol{B}_{k,n} : \boldsymbol{B}_{k,1}^* = \sum_n \boldsymbol{B}_{k,n}$. The estimation procedure remain the same.

2.6.2 With Additional Covariates

One of the assumptions underlies model proposed in Section 2.1 is that the differences in natural growth pattern can be sufficiently explained by a multiplicative factor: $d_k(t) = s_{k,0} \cdot d(t)$. Other covariates associated with growth outcome such as social economic status, nutritional biomarkers, etc, are not considered in the modeling process.

Even though it is reasonable to assume those other factors can be averaged out when there are enough samples to model the natural growth and diarrhea effects marginally, we can easily extend the proposed model in Section 2.1 to take into account other factors of our interests. Depending on the size of available data, we can improve the efficiency of the estimation by including these additional covariates in the model

$$y_k(t) = \boldsymbol{X}_k(t)^\intercal \boldsymbol{c} + s_{k,0} \cdot g(t) + \sum_{i=1}^{N_k} s_{k,i} \cdot d(t - l_{k,i}) + \epsilon_k(t), \tag{2.30}$$

where $\boldsymbol{X}_k(t)$ is the time-varying (or fixed) covariate of interest and $\boldsymbol{c}$ is the coefficient.

The estimation can be carried out the same way by artificially associating one degenerate coefficient $s_{k,c} \equiv 1$ or $s_{k,c} \sim N(1,0)$ with the covariate component. The equation (2.30) in vector form is now

$$\begin{aligned}\boldsymbol{y}_k &= s_{k,c} \cdot \boldsymbol{X}_k(t)^\intercal \boldsymbol{c} + s_{k,0} \cdot g(t) + \sum_{i=1}^{N_k} s_{k,i} \cdot d(t - l_{k,i}) + \epsilon_k(t) \\ &= \boldsymbol{M}_k^* \boldsymbol{C}_k^* \begin{bmatrix} s_{k,c} \\ \boldsymbol{s}_k \end{bmatrix} + \boldsymbol{\epsilon}_k,\end{aligned}$$

where

$$\boldsymbol{M}_k^* = \left[\boldsymbol{X}_k(t), \boldsymbol{A}_k, \boldsymbol{B}_{k,1}, \ldots, \boldsymbol{B}_{k,N_k}\right], \qquad \boldsymbol{C}_k^* = \begin{bmatrix} \boldsymbol{c} & 0 \\ 0 & \boldsymbol{C}_k \end{bmatrix}.$$

In the EM algorithm iteration update,

$$E(s_{k,c}|\boldsymbol{y}_k, \boldsymbol{\theta}) = 1, E(s_{k,c}^2|\boldsymbol{y}_k, \boldsymbol{\theta}) = 1.$$

So the the estimation procedure is similar, with the distinction that one degenerate normal component is added.

Chapter 3

Simulation

In this chapter, simulation studies are conducted to evaluate the numerical performance of the proposed estimation method in analyzing generated children growth data with heterogeneous natural growth trend and diarrhea effects. BIC criterion is shown to select the correct model with 100% accuracy. When classifying subjects into subgroups, the misclassification rate remains low throughout different data scenarios. The asymptotic variance of model estimates presented in Section 2.5 is validated by studying the empirical coverage probability of confidence interval for parameter estimates.

3.1 Data Generation

The proposed model is evaluated with generated data sets. Each dataset contains 700 subjects. For each subjects, growth measurements are collected around every three months. The number of followup times of a subject is uniformly generated between 2 and 17. The number of diarrhea episodes N_k is generated according to a Poisson distribution with a mean of 3. The onset time $\{l_{k,i}\}$ for diarrhea episodes

are uniformly generated within the followup period.

We conduct simulation in three scenarios, where the number of natural growth subgroups, the number of diarrhea vulnerability subgroups, the separation between subgroups are manipulated.

Table 3.1: Mixture distribution parameters used in the three simulation scenarios

Scenario 1	
Number of subjects	700
Natural growth multiplier distribution	$N(4, 1^2)$
Diarrhea multiplier distribution	$.5MVN(2, 1^2) + .5MVN(6, 1^2)$
Scenario 2	
Number of subjects	700
Natural growth multiplier distribution	$N(1, 1^2)$
Diarrhea multiplier distribution	$.5MVN(2.5, 1^2) + .5MVN(5.5, 1^2)$
Scenario 3	
Number of subjects	700
Natural growth multiplier distribution	$.5N(2, 1^2) + .5N(6, 1^2)$
Diarrhea multiplier distribution	$.5MVN(2, 1^2) + .5MVN(6, 1^2)$

For each scenario, 500 date sets are generated. Scenario 1 is used as the baseline for studying the finite sample performance of the proposed estimation method. One group of natural growth and two subgroups of diarrhea vulnerability are used for data generation. Subgroups of diarrhea vulnerability are placed at 2 and 6, the standard deviation for both groups is 1. In scenario 2, diarrhea subgroups are 3 standard deviation apart, closer than the 4 standard deviation in scenario 1. The estimation performance can show how well can the proposed method separate subgroups. Natural growth in scenario 3 is assume to follow a mixture of two normal distribution. This scenario is used to evaluate performance of the proposed method work under more complex data generation setting.

3.1.1 Data Generation Procedure

First, we focus on the detailed data generation procedure for scenario 1. The growth outcomes for the kth subject with diarrhea effect are generated according to

$$y_k(t) = s_{k,0} \cdot g(t) + \sum_{i=1}^{N_k} s_{k,i} \cdot d(t - l_{k,i}) + \epsilon_k(t). \tag{3.1}$$

Since the natural growth loading factor has only one cluster, the natural growth group membership is always 1

$$z_{k,0} = \Big\{ 1, \quad \text{with probability } 1 \ , \forall k. \tag{3.2}$$

And the natural growth multiplier $s_{k,0}$ is given by

$$s_{k,0} | z_{k,0} = 1 \sim N(4, 1^2). \tag{3.3}$$

Likewise, let $z_{k,1}$ be subject k's diarrhea subgroup membership indicator variable, then

$$c_{k,1} = \begin{cases} 1, & \text{with probability } 0.5 \\ 2, & \text{with probability } 0.5 \end{cases}, \forall k. \tag{3.4}$$

And the diarrhea susceptibility coefficients $s_{k,i}$ in each subgroup will follow a normal distribution, respectively

$$s_{k,i} | c_{k,1} = 1 \sim N(2, 1^2), \quad s_{k,i} | c_{k,1} = 2 \sim N(6, 1^2), \ \forall i. \tag{3.5}$$

The function $g(t)$ is the natural growth pattern over time, which we set as the estimated natural growth curve from the NIH cohort data. The true curve $g(t)$

specified the simulation is plotted in dash line at the left panel of Figure (3.1), along with the blue line representing the initial value used for the EM algorithm, which we will explain in Section 3.1.2.

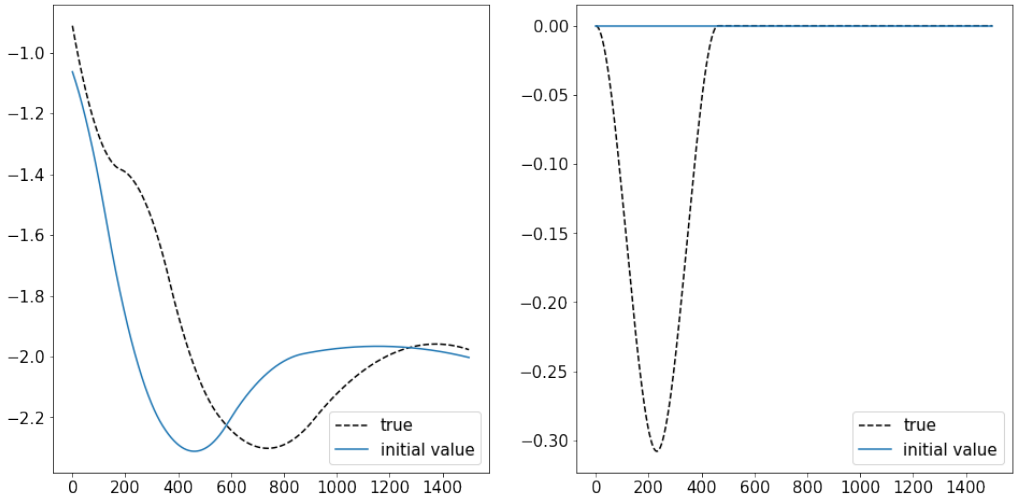

Figure 3.1: Left panel: dashed line is the natural growth pattern used in simulation study, blue line is the initial value supplied in model estimation. Right panel: true diarrhea effects pattern is plotted in dash line , blue line is the initial value for the EM algorithm

The diarrhea effect pattern $d(t)$ is specified as

$$d(t) = \begin{cases} \frac{1}{5} \cdot (\frac{t}{200} + 0.423)(\frac{t}{200} - 0.577)(\frac{t}{200} - 1.577) - 0.077, & \text{if } t \in [0, 230.8) \\ \frac{1}{5} \cdot (\frac{t-230.8}{200} + 0.423)(\frac{t-230.8}{200} - 0.577)(\frac{t-230.8}{200} - 1.577) - 0.077, & \text{if } t \in [230.8, 461.6) \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

We specify that $\epsilon_k \sim N(0, \mathbf{\Sigma}_k)$ has a diagonal covariance structure with standard deviation $\sigma = 0.1$.

Figure (3.2) is the trajectory plot of one dataset generated based on the procedures above, which connects the consecutive measurements on each individual.

Data sets for scenario 2 and scenario 3 are generated following similar procedure.

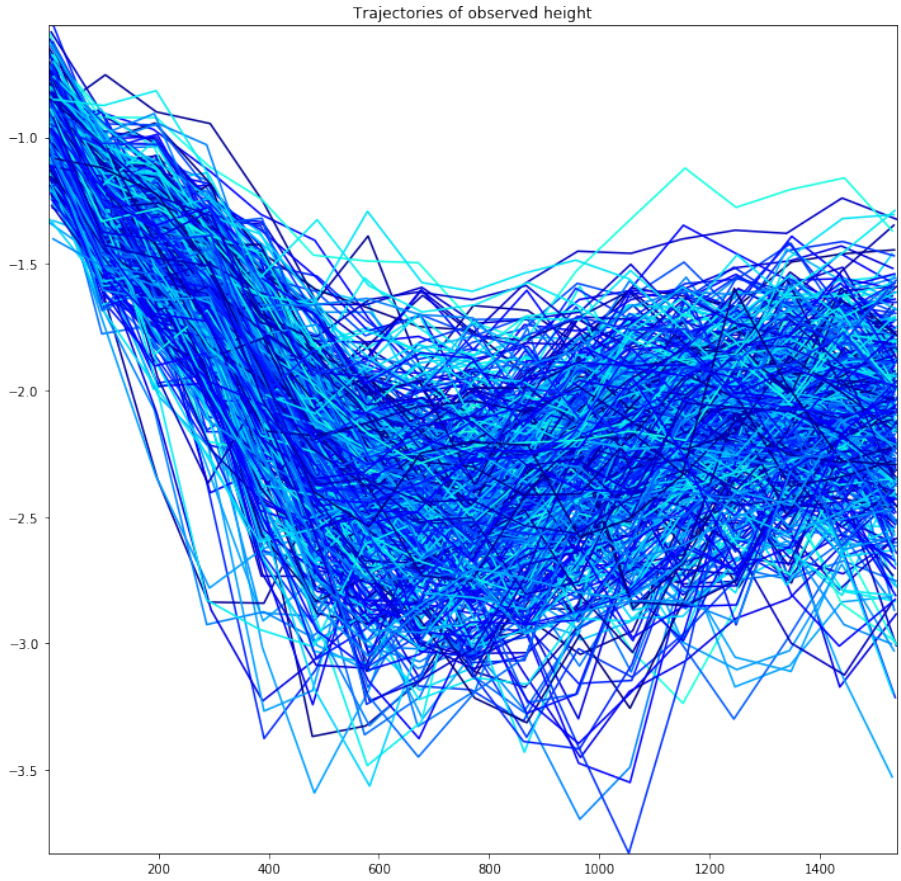

Figure 3.2: Trajectory plot of one dataset generated based on the procedures above. Each lines is formed by connecting the consecutive measurements on the same individual.

3.1.2 Initial Values

The B-spline basis for natural growth and diarrhea effect temporal dynamics are constructed using quantile knots. As mentioned in Section 1.3, we restrict our scope to quantile knots and fixed order of B-spline basis to be 4, cubic spline that is commonly used in practice. The initial value of natural growth B-spline coefficient $\boldsymbol{a}$ is obtained by assuming working independence among repeated measurements of the same subject and fit B-spline approximation. In other works, disregard the correlation between measurements of the same person, simply treat the dataset as

independent points measurement. For starting value of diarrhea coefficient $\boldsymbol{b}$, we simply choose $\mathbf{0}$, the zero vector. Figure (3.1) plots the initial values in blue lines.

3.2 Evaluation Criteria

Identifying the numbers of latent subgroups in both natural growth and diarrhea effects is of interests in our study. The numbers of latent subgroups need to be inferred from the data, and we proposed in Section 2.4 to do model selection by BIC criterion. In the simulation study, we evaluate the probability of correctly selecting the underlying mixture model structure, by examine the percentage of repetition where the true Model is selected

$$P_{\text{model}} = \frac{\text{\# times when correct model selected}}{500}.$$

We obtain parameter estimates from the proposed the model estimation method. Estimation bias, empirical standard error (ESE) and mean squared error (MSE) for each parameter are calculated and used to evaluate numerical performance.

Let θ denote a parameter and also its population value, and let $\hat{\theta}_r$ and SE_r denote its estimate and the corresponding estimated standard error in the rth replication. Then the parameter estimate of θ, $\hat{\theta}$, is calculated as the average of parameter estimates of 500 simulation replications

$$\hat{\theta} = \frac{1}{500}\sum_{r=1}^{500}\hat{\theta}_r.$$

With the introduction of subgroup indicator, given a subject, we are able to calculate the posterior probability of it belonging to each subgroup combination. Under the 1-0 loss commonly used in categorical prediction, the predicted subgroup

membership will be the one with the highest posterior probability. The misclassification rate for one generated data set is defined as the empirical probability of classifying subjects into the wrong group membership combination is evaluated as

$$P_{\text{mis}} = \frac{\#\text{ subjects that } \boldsymbol{z}_{pred} \neq \boldsymbol{z}_k}{700}.$$

The integrated squared error (ISE) between function f and its estimate $\hat{f}$ on region $[0, T]$ is defined as

$$ISE(\hat{f}, f) = \frac{1}{T}\int (f(t) - \hat{f}(t))^2 dt, \tag{3.7}$$

and mean integrated squared error (MISE) is the average ISE

$$MISE = \frac{1}{500}\sum_{r=1}^{500} ISE(\hat{f}_r, f).$$

MISE is used as the evaluation metric for performance of the $g(t)$ and $d(t)$ curve estimates.

Lastly, we compute the variance of the $\boldsymbol{b}$, B-spline coefficients for $d(t)$, and construct the pointwise 95% confidence interval according to asymptotic normality for 100 evenly spaced t on the interval $[0, 1500] : t = 15, 30, \ldots, 1500$. At each point, the B-spline estimate for $d(t)$ is $\hat{b}(t) = [B_1(t), \ldots, B_q(t)]\boldsymbol{b} = B(t)^\intercal \boldsymbol{b}$. The variance of $\hat{d}(t)$ is given by the formula $Var(\hat{d}(t)) = B(t)^\intercal Cov(\hat{\boldsymbol{b}})B(t)$. The true value of $d(t)$ is compared to the pointwise 95% confidence interval, and the empirical coverage probability of the confidence intervals are computed over the 500 generated data sets.

3.3 Simulation Result

Numerical results from the three data scenarios are aggregated and evaluated in this section. The proposed estimation method works well. The correct number of latent subgroup is identified with 100% accuracy in all scenarios. The individual membership misclassfication rate 3.53% in scenario 1, and less than 6% for scenario 2&3.

3.3.1 Scenario 1

Figure 3.3 shows the estimated curves for the one of the 500 data sets generated. The doted black lines represent true curves specified in data generation, the red ones are our estimates. Visually the estimated ones are very close to the true ones. The ISE for estimated natural growth curve is 2.77×10^{-10}, and for estimated diarrhea effect is 1.31×10^{-12}.

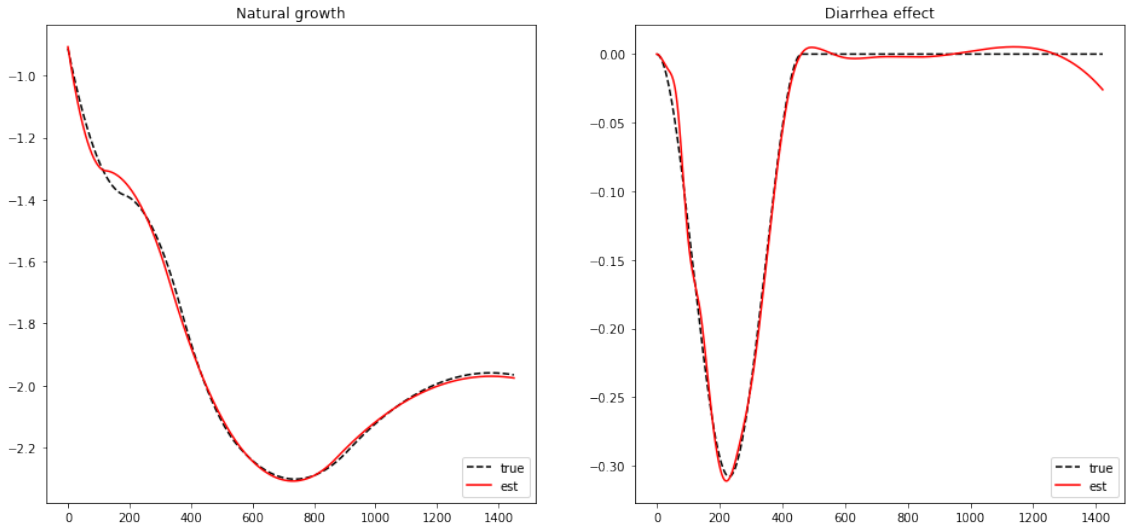

Figure 3.3: Red lines are estimated curves for the one of generated data sets. The doted black lines represent true curves specified in data generation. The ISE for estimated natural growth curve is 2.77×10^{-10}, and for estimated diarrhea effect is 1.31×10^{-12}.

Table 3.2 presents the summarized statistics over 500 repetitions. The misclassification rate ranges from 1.57% to 6.14%, and the average misclassification rate is 3.53%. Overall the estimated mixture distribution parameters are close to the true values. It is worth mentioning that due to identifiability constraint, the mixing probability is forced to be 100.00%, and mean forced to 4.00.

Table 3.2: Summary of estimation, scenario 1. Overall the estimates are close to the true value. Due to identifiability constraint, the natural growth probability is forced to be 100.00%, and mean forced to 4.00.

	Estimate
P_{model}	100%
Mean of P_{mis}	3.53%
$MISE_a$	6.27×10^{-10}
$MISE_b$	2.09×10^{-11}
Median time	19.78s

	Prob		Mean		Std	
	True	Est	True	Est	True	Est
Natural growth multiplier group 1	100%	100%	4	4	1	0.998
Diarrhea multiplier group 1	50%	49.95%	2	1.982	1	1.007
Diarrhea multiplier group 2	50%	50.05%	6	6.018	1	1.003

3.3.2 Scenario 2

Figure 3.4 shows the estimated curves for the one of the 500 data sets generated. The doted black lines represent true curves specified in data generation, the red ones are our estimates. Visually the estimated ones are very close to the true ones. The ISE for estimated natural growth curve is 2.76×10^{-10}, and for estimated diarrhea effect is 2.23×10^{-10}.

Table 3.3 presents the summarized statistics over 500 repetitions. The misclassification rate ranges from 3.42% to 8.71%, and the average misclassification rate is 5.96%. Overall the estimated mixture distribution parameters are close to the

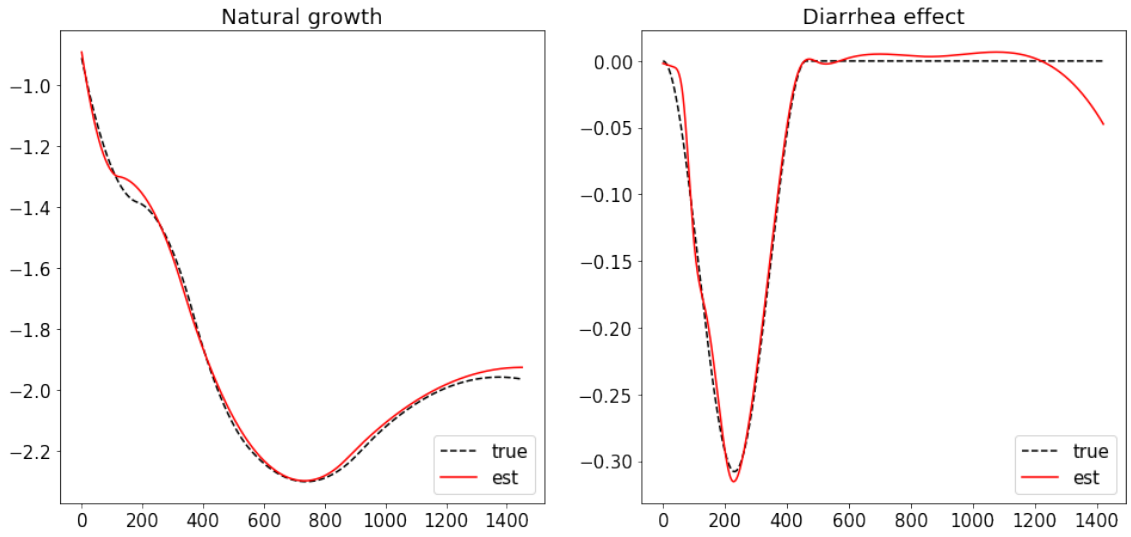

Figure 3.4: Red lines are estimated curves for the one of generated data sets. The doted black lines represent true curves specified in data generation. The ISE for estimated natural growth curve is 2.76×10^{-10}, and for estimated diarrhea effect is 2.23×10^{-10}.

true values. The distance between diarrhea subgroups is now 3 standard deviation, compared to the 4 standard deviation in scenario 2. It is more difficult to separate the two subgroups, which explains the increase in misclassification rate. When the difference between subgroups is less obvious, it takes more time for the EM algorithm to converge. Meanwhile, mixture distribution parameter estimates remains similar to that of scenario 1.

Table 3.3: Summary of estimation, scenario 2. The probability and mean of natural growth group 1 is fixed according to identifiability constraints.

	Estimate
P_{model}	100%
Mean of P_{mis}	5.96%
$MISE_a$	6.25×10^{-10}
$MISE_b$	1.50×10^{-11}
Median time	23.57s

	Prob		Mean		Std	
	True	Est	True	Est	True	Est
Natural growth multiplier group 1	100%	100%	4	4	1	0.998
Diarrhea group multiplier 1	50%	50.05%	2.5	2.488	1	1.007
Diarrhea group multiplier 2	50%	49.95%	5.5	5.516	1	1.003

3.3.3 Scenario 3

Figure 3.5 shows the estimated curves for the one of the 500 data sets generated. The doted black lines represent true curves specified in data generation, the red ones are our estimates. Visually the estimated ones are very close to the true ones. The ISE for estimated natural growth curve is 2.18×10^{-9}, and for estimated diarrhea effect is 6.33×10^{-13}.

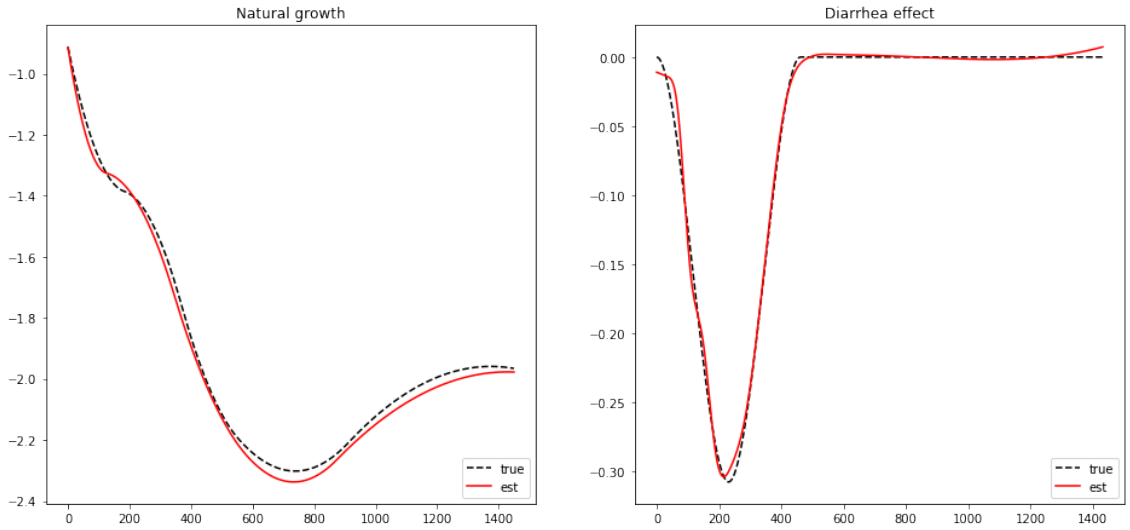

Figure 3.5: Red lines are estimated curves for the one of generated data sets. The doted black lines represent true curves specified in data generation. The ISE for estimated natural growth curve is 2.18×10^{-9}, and for estimated diarrhea effect is 6.33×10^{-13}.

Table 3.4 presents the summarized statistics over 500 repetitions. The misclassification rate ranges from 3.86% to 9.29%, and the average misclassification rate is 5.99%. With the introduction of the second natural growth group, scenario 3 is more complex in terms of subgroup structure. The median runtime increased. The misclassification rate increased to 5.96%. As there is more variance associated with natural growth curve in the generated data set, MISE$(\hat{d}(t), d(t))$ increase to 1.94×10^{-9}, while MISE for diarrhea curve remains almost the same to that of

scenario 1.

Table 3.4: Summary of estimation, scenario 3.

	Estimate
P_{model}	100%
Mean of P_{mis}	5.91%
$MISE_a$	1.94×10^{-9}
$MISE_b$	2.01×10^{-11}
Median time	80.57s

Group	Prob		Mean		Std	
	True	Est	True	Est	True	Est
Natural growth multiplier group 1	50%	50.01%	2	2.002	1	0.995
Natural growth multiplier group 2	50%	49.99%	6	6.002	1	1.000
Diarrhea group multiplier 1	50%	50.05%	2	1.982	1	1.007
Diarrhea group multiplier 2	50%	49.95%	6	6.018	1	1.003

3.3.4 Empirical Coverage Probability

In this subsection, evaluation on the asymptotic variance derived in Section 2.5 is performed. We computed the variance of the $d(t)$ B-spline coefficients and constructed the pointwise 95% confidence interval according to asymptotic normality for 100 evenly spaced t on the interval $[0, 1500] : t = 15, 30, \ldots, 1500$. At each point, the B-spline estimate for $d(t)$ is $\hat{b}(t) = [B_1(t), \ldots, B_q(t)]\boldsymbol{b} = B(t)^\intercal \boldsymbol{b}$. The variance of $\hat{d}(t)$ is given by the formula $Var(\hat{d}(t)) = B(t)^\intercal Cov(\hat{\boldsymbol{b}})B(t)$.

Table 3.5: Setting for data generation used in validating asymptotic variance.

Number of subjects	700
Natural growth multiplier distribution	$N(4, 1^2)$
Diarrhea multiplier distribution	$.5MVN(2, 1^2) + .5MVN(6, 1^2)$
Measurement error std	0.1

Under setting of Table 3.5, we generated 500 data sets with sample size 700. Mixture distribution parameters used in data generation are the same as those of scenario 1. The true value of $d(t)$ is compared to the pointwise 95% confidence interval, and the empirical coverage probability of the confidence intervals are computed over the 500 generated data sets, and plotted in Figure (3.6).

The empirical coverage probability stays at 95% for interval $[150, 1300]$, validating that the asymptotic variance derived in Section 2.5. Quantiles knots are used in constructing the B-spline basis, and the first knot is placed at around 100 for a typical data set. The boundary effect of B-spline causes bias in the interval $[0, \text{first knot}]$, thus coverage probability is lower near the left boundary. Not many diarrhea episodes with elapsed time of 1500 are observed. With limited information at the right boundary, the computed asymptotic variance is large. Boundary bias is offsetted by the wider confidence interval, so that the coverage probability stays relatively close to 95% for $[1300, 1500]$.

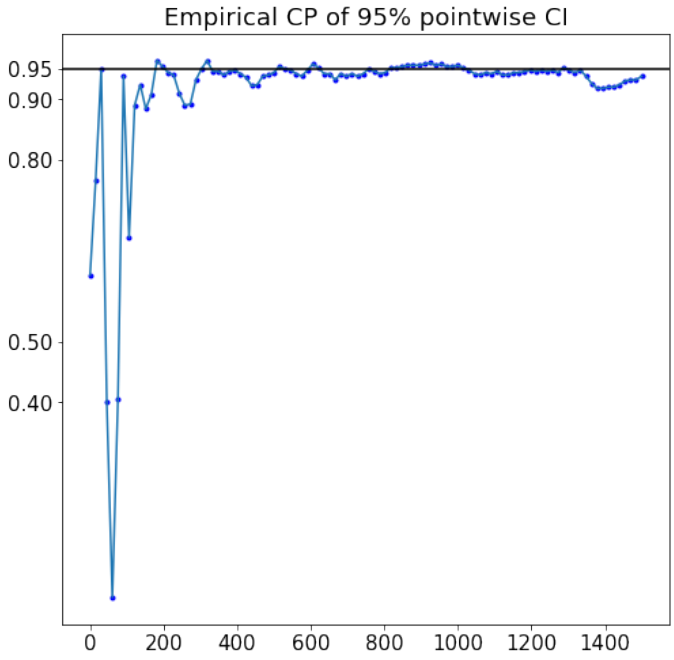

Figure 3.6: Empirical coverage probability of 95% pointwise confidence interval. Date generated under settings of scenario 1, 500 data sets are generated with sample size = 700.

Boundary effect can be mitigated by increasing sample size, the low CP at the left boundary and slightly lower CP at the right boundary will be corrected as more data becomes available.

In summary, simulation studies show satisfactory performance in estimating model parameters. BIC criterion is shown to select the correct model with 100% accuracy. When classifying subjects into subgroups, the misclassification rate remains low throughout different data scenarios. The validity of derived asymptotic variance is tested by studying the empirical coverage probability of confidence interval for parameter estimates.

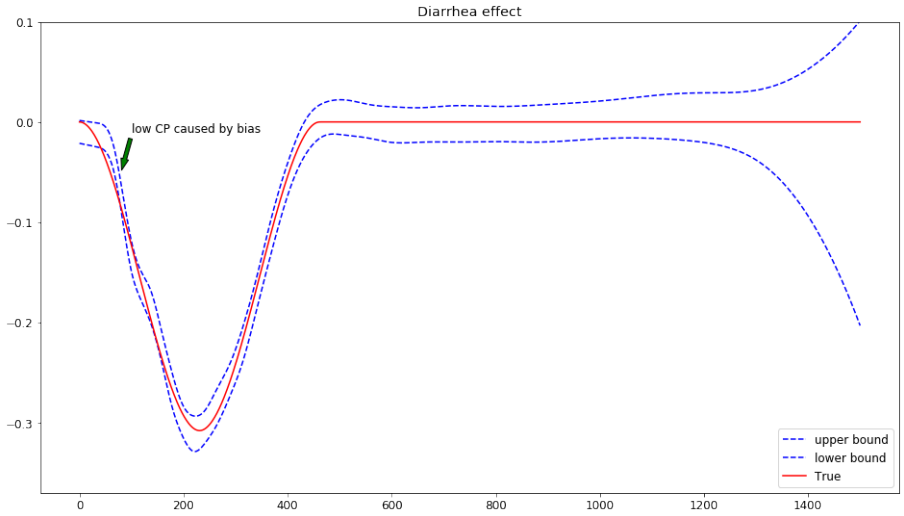

Figure 3.7: Pointwise CI for one of the 500 simulation runs when sample size = 700. The boundary bias often causes the true curve (in red) to be outside of the confidence interval

Chapter 4

Real Data Application

This thesis is motivated by the NIH cohort study data set. 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 enrolled infants are uniquely identified by their subject id. The growth outcomes of each child were recorded every 3 months until the end of the study, resulting in 6381 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. Let $t_{k,j}$ denote the j^{th} follow up date of subject indexed by k, and $l_{k,i}$ denote the onset time of subject k's i^{th} diarrhea episode, then $t_{k,j} - l_{k,i}$ is the elapsed time for diarrhea effect at time $t_{k,j}$. Figure (4.1) shows the distribution of observed elapsed time of diarrhea effect $t_{k,j} - l_{k,i}$.

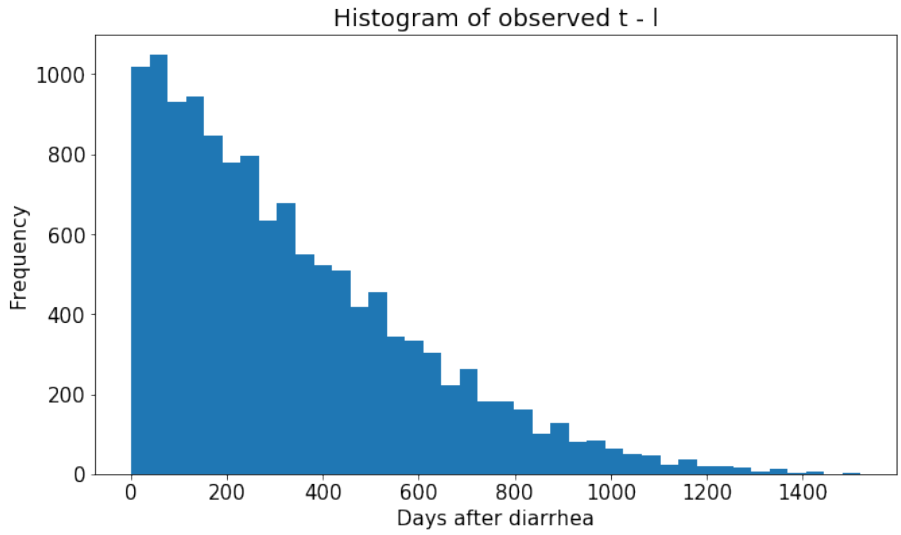

Figure 4.1: Distribution of observed diarrhea elapsed time.

4.1 Choose the Best Response Variable

We choose ΔHAZ, the change of HAZ score since enrollment, as the response variable when studying the diarrhea effect on childhood growth.

Weight-associated growth measurements are not selected as the response, for weight is highly sensitive to diarrhea in the short term. The original height (cm) is

not used, because when children grows taller, the range of their height grows wider. This makes it hard to justify the constant variance assumption for the random errors.

Then we discuss the drawbacks of using HAZ as the response. Indeed, we can adopt the constant variance assumption for random error if HAZ, the age adjusted z-score for height, were to be used as the response variable. But if we do, fitting the starting value becomes difficult. At enrollment, the starting value is

$$\begin{aligned} y_k(0) &= s_{k,0} \cdot g(0) + s_{k,1} \sum_i d(0 - l_{k,i}) + \epsilon(0) \\ &= s_{k,0} \cdot g(0) + s_{k,1} \sum_i 0 + \epsilon(0) \\ &= s_{k,0} \cdot g(0) + \epsilon(0), \end{aligned}$$

because there is no diarrhea effect yet at enrollment. In fact, $g(0)$ is "fixed", in the sense that

$$\begin{aligned} \bar{y}(0) &= \bar{s}_{k,0} g(0) + \bar{\epsilon}(0) \\ &= 1 \cdot g(0) + 0. \end{aligned}$$

The second equal sign is based on identifiability constraint. So $g(0)$ should be close to the average of starting values, that leaves differences in HAZ score at the start to be modeled only by $s_{k,0}$. The natural growth multiplier $s_{k,0}$ is largely determined by the starting HAZ. The signals in natural growth is much stronger than that of diarrhea effect. When $s_{k,0}$s are determined, the relative small diarrhea will be masked.

Therefore, to avoid the above concerns, we decide to use ΔHAZ as the response variable. It is height-associated, which is less sensitive to diarrhea in the short term; it is the difference between two z-scores, so the variance is stable hence justifying

the constant variance assumption; most importantly, the starting value is always 0, so $s_{k,0}$s are not restricted solely by the starting value and can adjust accordingly to provide a better fit to the data and more meaningful results.

4.2 Original Model Fitting

In order to apply the proposed model to the NIH cohort data, we need to first specify the numbers of subgroup for both random effects, the number of knots and the location of them. We find the $d(t)$ estimate robust to the number of knots. The growth signal is strong and there are plenty of evenly spaced observations, besides that natural growth is not the main study interest, 4 inner quantile knots are placed and we did not search for knot numbers. A three-way grid search of U, V and number of $d(t)$ Bspline knots are conducted. We search U in $\{1, 2\}$, V in $\{1, 2, 3\}$ and number of knots in $\{5, 6, \ldots, 16\}$. The model with u latent natural growth subgroups, v latent diarrhea vulnerability subgroups and q inners for $d(t)$ will be denoted as $M_{u,v,q}$.

The combination of 1 natural group, 3 diarrhea groups and 8 inner knots $M_{1,3,8}$ is selected by the BIC criterion. $M_{1,3,8}$ has the lowest BIC = 9352.50.

The mixture model parameters estimate is summarized in Table (4.2), and estimated curves are plotted in Figure (4.2):

Table 4.1: Diarrhea group 3 has small probability. Out of the 521 children with at least one diarrhea infection, 0.2% corresponds to exactly one child. The standard deviation is small, the mean value is far from the other two subgroups.

	Prob	Mean	Std
Natural growth multiplier group 1	100%	1	1.326
Diarrhea multiplier group 1	85.76%	0.658	0.842
Diarrhea multiplier group 2	14.04%	4.134	1.691
Diarrhea multiplier group 3	0.20 %	-72.801	0.557

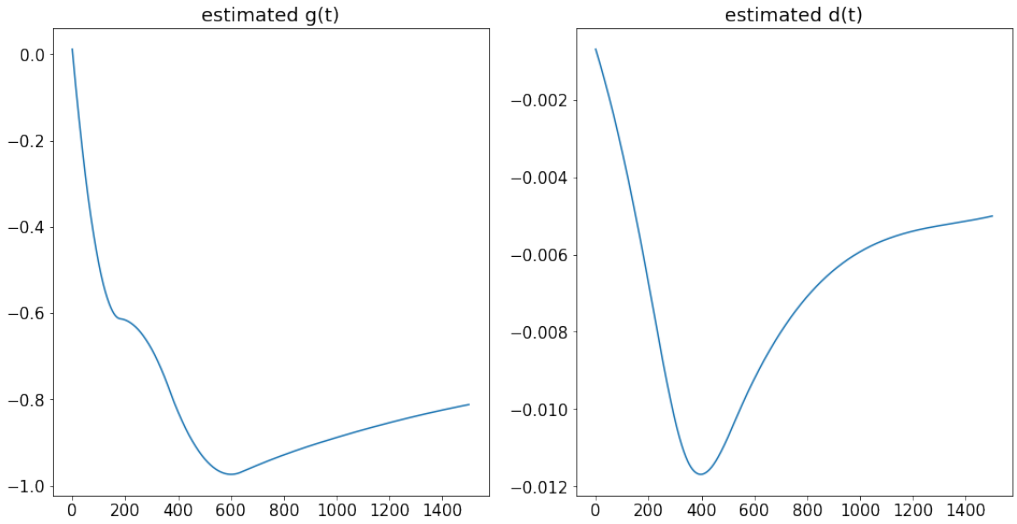

Figure 4.2: Estimated curves for NIH cohort.

Estimated diarrhea effect on growth indicates that the first stage of negative effect happens right after diarrhea infection. After 450 days of infection, the diarrhea effect is at its worst, causing a 0.013 loss in HAZ. The catch up growth begin after around 450 days of infection, but does not fully recover the previous growth deficit. One episode of diarrhea infection will cause permanent HAZ loss. For 86% of the NIH cohort, the permanent loss is -0.0518, and for 14% of the cohort, and permanent loss is more severe: -0.3. The finding of long-term growth deficit is in line with Troeger et al. (2018), which shows each day of diarrhea leads to -0.0033 HAZ score. 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.

The identification of diarrhea group 3 shows the possibility that our proposed model can be applied in outlier detection. Group 3 has small probability, out of the 521 children with at least one diarrhea infection, 0.2% corresponds to exactly one child. The standard deviation is small, and the mean value is far from the other

two subgroups. These signs indicate this subject is an outlier in the data, and it is are picked up by the model and recognized as a subgroup of its own. Group 3 corresponds to subject #8228, its growth trajectory is plotted in Figure (4.3)

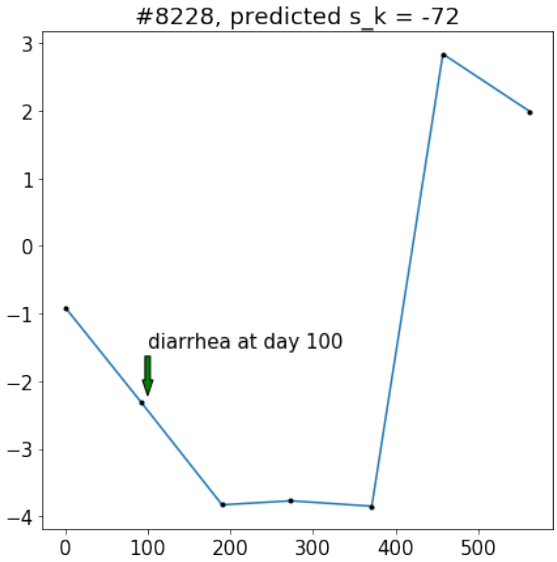

Figure 4.3: Growth trajectory of #8228, the line is formed by connecting the consecutive measurements on the individual.

For the cohort, the overall diarrhea effect on growth is estimated to be negative over time. Child #8228 contracted diarrhea 100 days after enrollment. Its HAZ at the time is -2.31 standard deviation below average. Its HAZ continues to drop to -4, then bounced back to 3, as if diarrhea had a positive impact on growth. The pattern is far different from the common trend, so its diarrhea multiplier is predicted to be -70. We believe this abnormal pattern of growth is an outlier to the cohort.

4.2.1 Identify Outlying Observations

The growth pattern of subject #8223 is for sure very different from the typical trajectory in NIH study cohort, and identification of #8228 further motivate us to study the possibility of the existence of other outliers. While examining the pre-

dicted diarrhea multipliers, we observe that several subjects have extreme values, and this pattern is found consistently through out the top models ranked by BIC, the smaller BIC the higher the ranking. Figure (4.4) are the histograms of predicted diarrhea multipliers in the best model $M_{1,3,8}$ with BIC = 9352.50, and the second best model $M_{1,3,7}$ with BIC = 9354.13. Not only are these children identified simultaneously, their relative ranks in predicted values are preserved as well. Figure (4.5) provides a closer look at #7088 and #8371. All three children contracted only one episode of diarrhea. Subjects #7088 and #8371 experienced shape drop after diarrhea episode, both lost over 3 HAZ in 500 days. Their diarrhea multipliers are estimated to be around 20, signifying high vulnerability compared to the population.

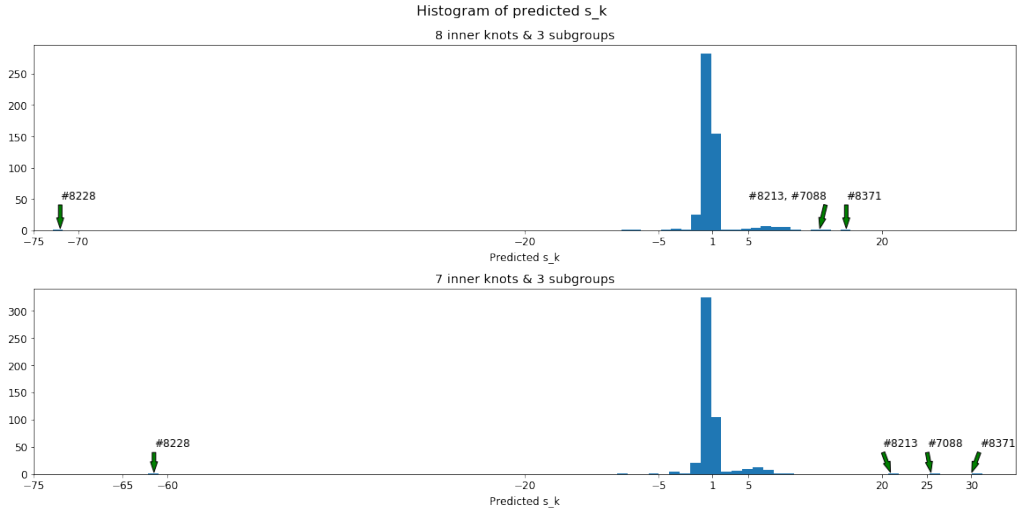

Figure 4.4: Subjects with high predicted diarrhea multiplier matched in the best models by BIC ranking.

Table 4.2: Params est for $M_{1,3,7}$

	Prob	Mean	Std
Natural growth multiplier group 1	100%	1	1.323
Diarrhea multiplier group 1	79.78%	0.564	0.704
Diarrhea multiplier group multiplier2	18.99%	2.862	1.116
Diarrhea multiplier group multiplier 3	1.23%	0.478	31.196

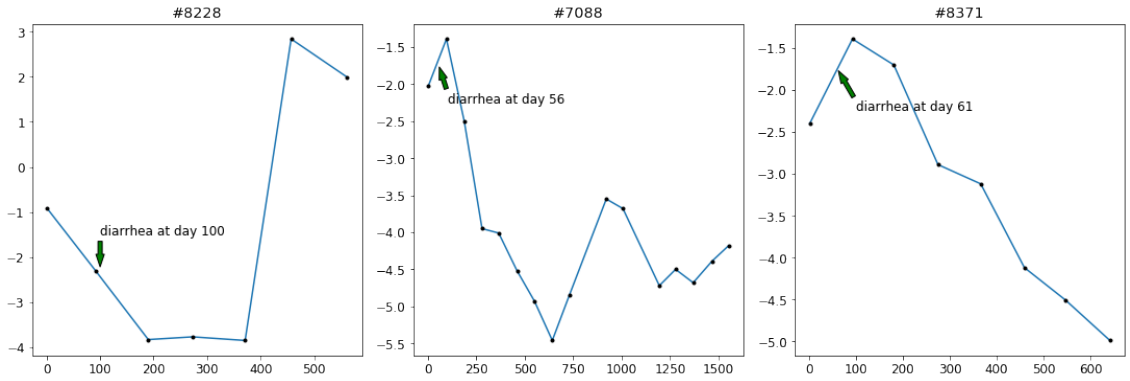

Figure 4.5: Trajectories of subjects with extreme predicted diarrhea multiplier. The lines are formed by connecting consecutive measurements of the same subject.

The model proposed is mean-based and existence of outliers will have large impact on the final model estimate. All the evidence suggests that we should try to remove outliers. We proceed by excluding these subjects in the order of how far are they being outliers, and stop when the model behaves moderately, in an attempt to preserve as much data as possible. Starting with #8228, then #8371 and #7088 (see Fig 4.6).

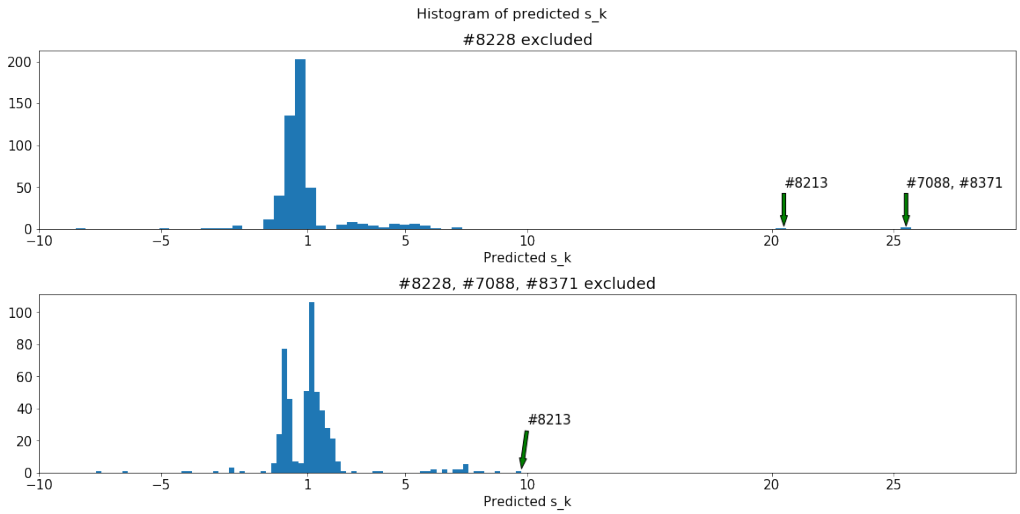

Figure 4.6: Predicted diarrhea multiplier when removing outliers recursively

- If only 8228 is removed, there are still large values further at the right side

(upper panel).

- Once 7088 and 8371 are excluded, the predicted multiplier now look much better (lower panel).

- Since #8213 is not lying too far away from other subjects, we decide to stop removing at this step.

Outlying observations in data could result in unreliable parameter estimates, incorrect standard errors and confidence intervals, and misleading statistical inferences. Diagnosing multivariate outlying observations in longitudinal data is challenging due to the well-known "curse of dimensionality". Because of the masking effect, high-dimensional outlying observations in a given sample can be well hidden and are barely all detected (Tong et al, 2020). As illustrated in Figure (4.7), besides subject #8228, the other children is hiding in plain sight. This section shows the possibility to use our proposed model for outlier detection in non-linear growth models. The ill-behaving subjects will be group together, and low probability or huge subgroup variance can be strong indicator of existence of outliers. Plotting the predicted random effect will give confirmation about whether the growth pattern and diarrhea reaction of a certain subject is very different compared to the rest.

4.2.2 Model Fit with Outliers Excluded

Due to the existence of outliers, especially #8228, BIC criterion prefers models with 3 diarrhea subgroups. Since classifying #8228 on its own can provide more gains in likelihood than penalty incurred by adding one more subgroup. The highest four ranking models by BIC all come with 3 diarrhea subgroups. Now that we have excluded the influential observations, model selection concerning number of

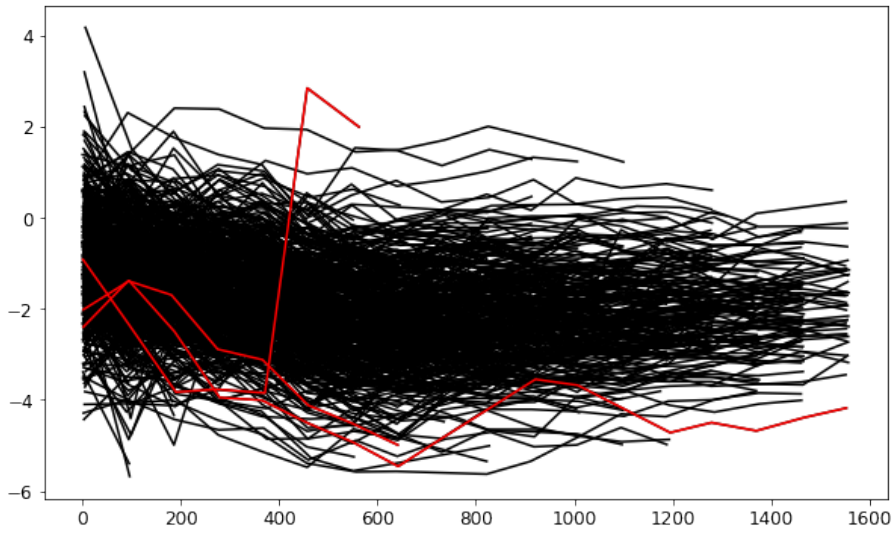

Figure 4.7: Trajectories of HAZ score change over time. Red lines indicates outliers identified.

natural growth groups, diarrhea groups and number of $d(t)$ B-spline inner knots is needed. The best model selected this time is 1 natural growth group, 2 diarrhea subgroups and 6 inner knots $M_{1,2,6}$. The model estimate is summarized in Table (4.3). And Figure (4.8) visualizes the curve estimates along with 95% pointwise confidence interval. Since in the case of having two latent subgroups for diarrhea, the Fisher information matrix can not be easily obtained, the confidence interval is calculated based on observed Fisher information matrix for the proposed model. See the Chapter 6 for more details.

Table 4.3: Summary of Model Estimates

	Prob	Mean	Std
Natural growth multiplier group 1	100%	1	1.411
Diarrhea multiplier group 1	85.38%	0.721	0.817
Diarrhea multiplier group 2	14.62%	2.630	1.281

Figure (4.9) are the histograms of the predicted diarrhea multipier in different latent subgroups. When cross referencing Table (4.3) and Figure (4.9), we can see

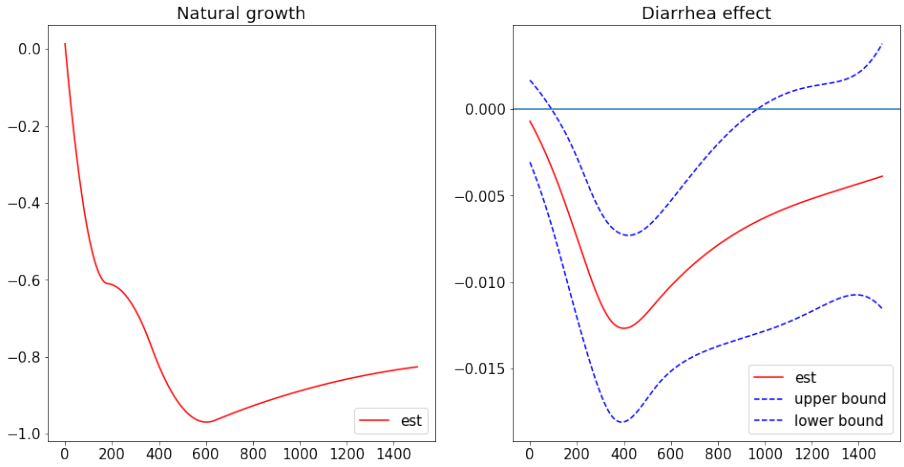

Figure 4.8: Estimated growth curve and diarrhea effects, with 95% pointwise confidence interval.

that the mixture model distribution here is working to approximate the long tail behavior of diarrhea multiplier $s_{k,i}$. The moderate values is classified into group 1, and the more extreme values are bracketed under group 2.

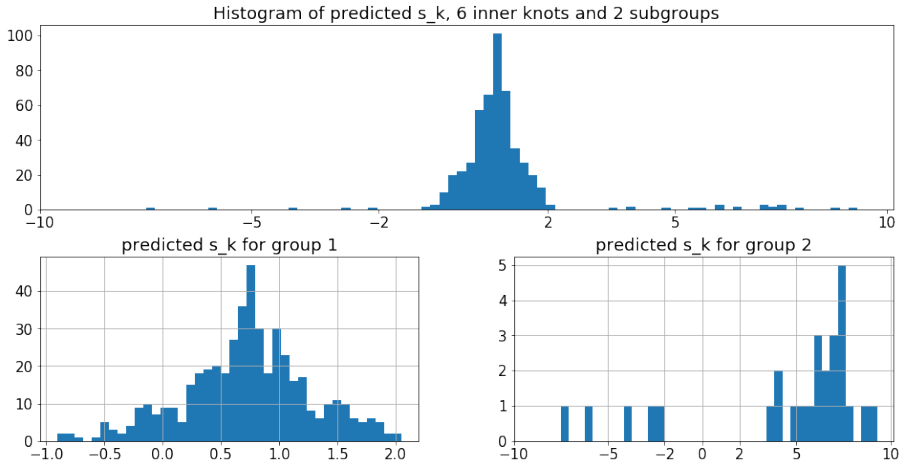

Figure 4.9: Subjects are divided into two groups, group 1 for moderate diarrhea multiplier and group 2 for more extreme values.

The fitted natural growth curve shows that the HAZ score of children in Bangladesh keeps dropping over first two years after birth and then remain roughly the same level. The diarrhea effect pattern shows that the HAZ score is negatively affected by diarrhea. For subjects in subgroup 1, which constitutes 85% of the cohort, HAZ score will drop at an even pace, reaching the valley of the diarrhea effect $-0.013 \times 0.658 = -0.0085$ at around 400 days. The catch-up growth starts after 400 days of infection, at roughly an even speed, 25% that of the dropping rate. The recovering trend is not shown to stop at the longest follow up time. The long-term HAZ loss after 3 year is $-0.004 \times 0.658 = -0.0026$. And since the 95% confidence interval covers 0, so we are unable to conclude whether diarrhea will leave a permanent long term negative effect on growth. 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.

It is worth mentioning that the indecisive conclusion about whether the catch up growth will cover growth deficit, i.e., whether the long-term loss is significant, is caused by the more extreme growth patterns in group 2. If only well-behaving subjects are used in model building process, estimated curves and the 95% CI will be Figure (4.10). Diarrhea effect seems to level off at -0.005. With confidence interval not covering 0, we can conclude that this negative long term effect is significant. How subjects in group 2 inflated estimation variance can be visualized more clearly with bootstrap. Depending on which and how many group 2 subjects are included in the bootstrapped sample, the estimated diarrhea effects have entirely different outlook (see Figure 4.11). Too many subjects with negative predicted diarrhea multiplier selected when sampling with replacement, the end result will looks like "Fast, almost full recovery", or to many large positive predicted selected, the end result will be "No recovery".

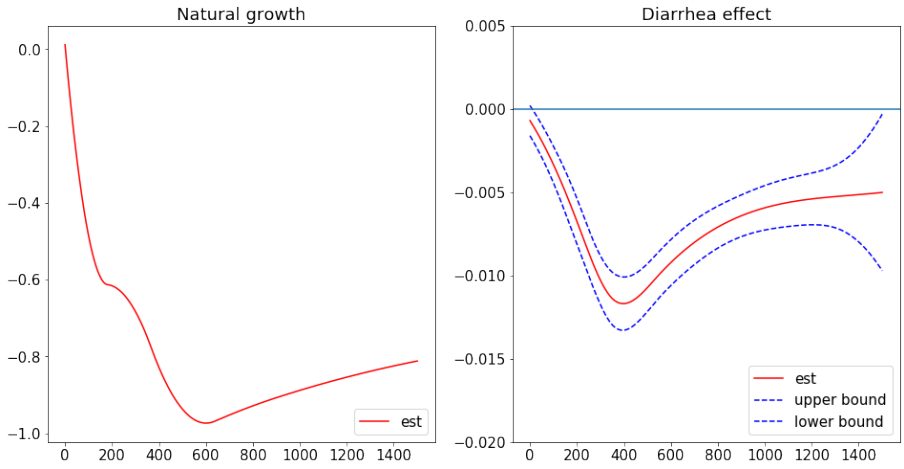

Figure 4.10: Only subjects in group 1 from previous results used. 95% pointwise confidence interval.

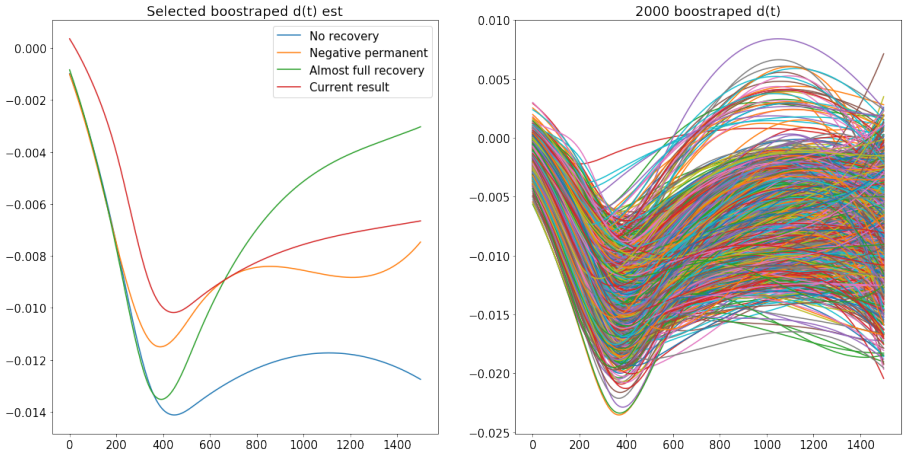

Figure 4.11: Depending on which and how many influential subjects are subjected in the bootstrapped sample, the estimated diarrhea effects have entirely different outlook.

Chapter 5

Discussion

Literature on the dynamic effect of diarrhea on children growth is limited. In this study, we propose a semi-parametric model to study the pattern of diarrhea effect over time and the heterogeneity of change pattern within cohort. Mixture of Gaussian distribution assumption is adopted, for reaction to diarrhea effect is often long-tailed. Since subgroup membership is latent and unobserved, Expectation-Maximization algorithm is used to find the maximum likelihood estimator, and asymptotic variance of the estimator is investigated. Simulation studies show that the proposed model is capable of identifying the correct mixture distribution structure, classifying subjects into the right subgroups, and obtaining reliable estimate of the functional effect of diarrhea over time. Based on the data from the NIH study in Bangladesh, our analysis results separate children into two groups based on their vulnerability to diarrhea. Diarrhea effect will reach its peak at 400 days after onset. At the maximum follow up time, the slow catch up growth still continues. It is inconclusive whether diarrhea effect is permanent for the cohort at large, but if only the data of subjects from the "moderate" group is used, the permanent effect is statistically significant. Two model variants are developed, and depending on the

size of available data, these model variants can improve estimation efficiency. Overall, our models provide new statistical tools to quantify the relationship between diarrhea and childhood growth in a dynamic fashion, which helps us make better health policy and conduct more efficient interventions.

Several limitations exist when modeling the effect of a complicated biological process such as diarrhea, in a statistical model with underlying model assumptions. For example, inter-individual difference happens in both magnitude and pattern, the assumption that inter-individual differences can be captured by multiplicative factor may not be flexible enough. In our model, we assume that different diarrhea episodes may have different effect curves, and 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. B-spline method is often applied in conjunction with penalty so that null-region of the curve can be identified and smoothness achieved. But because the exact EM algorithm is an optimization algorithm using marginal likelihood as the object function, it is difficult to incorporate penalty term into the framework.

On the other hand, limitations of the estimating process can also be improved. If the penalty terms is carefully constructed, we might be able to used generalized EM algorithm to find the estimator. The asymptotic variance of the estimator is inflated by the observations in the extreme value group. While the main focus of the study is not outlier detection and robust estimation method, it is possible to develop robust estimator of the variance of B-spline estimates. That would be another direction for future research.

Chapter 6

Asymptotic Variance of Model Estimator

In this section, we derive the Fisher information matrix of subject k, FIM_k. We start the appendix with matrix derivatives tool necessary for computing the asymptotic covariance matrix. Following the notation in Magnus and Neudecker (1999), for any function $y(x)$, the differential $dy(x)$ is define to be that part of $y(x + dx) - y(x)$ which is linear in dx. Unlike the classical definition relying on limits, the equation:

$$y(\boldsymbol{x} + d\boldsymbol{x}) = y(\boldsymbol{x}) + \boldsymbol{A}d\boldsymbol{x} + (\text{higher order terms})$$

holds even when x or y are not scalars. Matrix $\boldsymbol{A}$ is defined as the derivative.

Therefore, the derivative of any expression involving matrices can be computed in two steps:

1. compute the differential

2. rearrange the result into one of the canonical forms

after which the derivative is immediately read off as the coefficient of $dx, d\boldsymbol{x}$, or $d\boldsymbol{X}$. The canonical forms is one of the following:

$dy = adx$	$d\boldsymbol{y} = \boldsymbol{a}x$	$d\boldsymbol{Y} = \boldsymbol{A}dx$
$dy = \boldsymbol{a}d\boldsymbol{x}$	$d\boldsymbol{y} = \boldsymbol{A}d\boldsymbol{x}$	
$dy = tr(\boldsymbol{A}d\boldsymbol{X})$		

Here are some useful rules for our derivation. These rules can be iteratively applied because of the chain rule:

1. $d(\alpha\boldsymbol{X}) = \alpha d\boldsymbol{X}$;
2. $d(\mathrm{Tr}(\boldsymbol{X})) = \mathrm{Tr}(d\boldsymbol{X})$;
3. $d(\boldsymbol{X}\boldsymbol{Y}) = (d\boldsymbol{X})\boldsymbol{Y} + \boldsymbol{X}d\boldsymbol{Y}$;
4. $d\boldsymbol{X}^{-1} = -\boldsymbol{X}^{-1}(d\boldsymbol{X})\boldsymbol{X}^{-1}$;
5. $d\ln|\boldsymbol{X}| = \mathrm{Tr}(\boldsymbol{X}^{-1}d\boldsymbol{X})$;
6. $d\boldsymbol{X}^{\intercal} = (d\boldsymbol{X})^{\intercal}$.

Let L denote the marginal likelihood, and l log marginal likelihood. For clarity in notation, we drop the subject index subscript k. We can write down the marginal likelihood and the log of marginal likelihood as

$$\begin{aligned} L &= \prod_{i=[u,v]}^{[U,V]} p(\boldsymbol{y}, z = i|\boldsymbol{\theta}) \\ &= \prod p(\boldsymbol{y}|z = i, \boldsymbol{\theta})p(z = i|\boldsymbol{\theta}) \qquad \text{combine (2.15) and (2.16)} \\ &= \prod_{i=[u,v]} \psi_{u,v} \frac{p_u \pi_v \psi_{u,v}}{\sum_{u,v} p_u \pi_v \psi_{u,v}}, \\ l &= \ln(\sum_{u,v} p_u \pi_v \psi_{u,v}^2) - \ln(\sum_{u,v} p_u \pi_v \psi_{u,v}), \end{aligned} \tag{6.1}$$

where $\psi_{u,v}$ is the conditional pdf of $\boldsymbol{y}_i$ given s_0 comes from the uth subgroup and s_1 the vth subgroup, respectively. Then

$$\begin{aligned}
\boldsymbol{y}_i|z=[u,v] &\sim N(\boldsymbol{XC}\boldsymbol{\mu}_{u,v}, \boldsymbol{V}_{u,v}) \\
\boldsymbol{V}_{u,v} &= \boldsymbol{XC}\boldsymbol{\Gamma}_{u,v}(\boldsymbol{XC})^\intercal + \boldsymbol{\Sigma} \\
\boldsymbol{\mu}_{u,v} &= [\lambda_u, \mu_v, \ldots, \mu_v]^\intercal \\
\boldsymbol{\Gamma}_{u,v} &= \mathrm{diag}(\nu_u^2, \sigma_v^2, \ldots, \sigma_v^2), \boldsymbol{\Sigma} = \mathrm{diag}(\sigma^2, \ldots, \sigma^2), \\
\ln\psi_{u,v} &= -\frac{1}{2}\ln|2\pi V| - \frac{1}{2}(y-M\mu)^\intercal V^{-1}(y-M\mu).
\end{aligned}$$

For any function $y(x)$, the differential $dy(x)$ is defined to be that part of $y(x+dx)-y(x)$ which is linear in dx. This definition applies even when x or y are not scalars. We start our derivation from computing the differential of $\ln\psi_{u,v}$. The equation:

The differential of the first term is:

$$d\ln|V| = \mathrm{Tr}(V^{-1}dV).$$

Differential for the second term is:

$$\begin{aligned}
d(y-M\mu)^\intercal V^{-1}(y-M\mu) &= -(dM\mu)^\intercal V^{-1}(y-M\mu) \\
&\quad + (y-M\mu)^\intercal d(V^{-1})(y-M\mu) + (y-M\mu)^\intercal V^{-1}(-dM\mu) \\
&= (y-M\mu)^\intercal V^{-1} d(M\mu) + \frac{1}{2}\mathrm{Tr}(V^{-1}(y-M\mu)(y-M\mu)^\intercal V^{-1}dV).
\end{aligned}$$

By dividing matrix M and V into blocks, we have

$$\begin{aligned}
\boldsymbol{V} &= \nu^2 A\boldsymbol{a}(A\boldsymbol{a})^\intercal + \sigma^2 \sum B_i\boldsymbol{b}(B_i\boldsymbol{b})^\intercal \\
\boldsymbol{M\mu} &= \lambda A\boldsymbol{a} + \mu\sum_i B_i\boldsymbol{b}.
\end{aligned}$$

Depending on which variable is being differentiate, $d\boldsymbol{V}$ and $d\boldsymbol{M\mu}$ will be different. Take $\boldsymbol{a}$ for example:

$$d\boldsymbol{V} = \nu^2 A(d\boldsymbol{a}\boldsymbol{a}^\intercal + \boldsymbol{a}d\boldsymbol{a}^\intercal)A^\intercal$$
$$d(\boldsymbol{M\mu}) = \lambda A d\boldsymbol{a}.$$

Use Y to denote $y - M\mu$, plug in the differential term, we arrive at

$$\begin{aligned} d\ln\psi = &\frac{1}{2}\mathrm{Tr}(V^{-1}A(da\ a^\intercal + a(da)^\intercal)A^\intercal) + \frac{1}{2}\nu^2 Y^\intercal V^{-1}A(da\ a^\intercal + a(da)^\intercal)A^\intercal V^{-1}Y \\ &- \frac{1}{2}\lambda(Ada)^\intercal V^{-1}Y - \frac{1}{2}\lambda Y^\intercal V^{-1}Ada. \end{aligned} \tag{6.2}$$

Now aggregate terms associated with $d\boldsymbol{a}, (d\boldsymbol{a})^\intercal$, and the derivative is obtained as

$$\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}} = \nu_u^2 A^\intercal V^{-1}(y - M\mu)(y - M\mu)^\intercal V^{-1}A\boldsymbol{a} + \lambda_u A^\intercal V^{-1}(y - M\mu) - \nu_u^2 A^\intercal V^{-1}A\boldsymbol{a}.$$

Similarly, the derivative with respect to other parameters can be obtained:

$$\begin{aligned}
\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{b}} &= \sum_k [\sigma_u^2 B_n^\intercal V^{-1}(y - M\mu)(y - M\mu)^\intercal V^{-1}B_n\boldsymbol{b} + \mu_u B_n^\intercal V^{-1}(y - M\mu) - \sigma_u^2 B_n^\intercal V^{-1}B_n\boldsymbol{a}]; \\
\frac{\partial \ln\psi_{u,v}}{\partial \lambda_u} &= -\lambda_u(A\boldsymbol{a})^\intercal V^{-1}A\boldsymbol{a} + (A\boldsymbol{a})^\intercal V^{-1}(y - B\boldsymbol{b}\mu_v); \\
\frac{\partial \ln\psi_{u,v}}{\partial \mu_v} &= -\mu_v(B\boldsymbol{b})^\intercal V^{-1}B\boldsymbol{b} + (B\boldsymbol{b})^\intercal V^{-1}(y - A\boldsymbol{a}\lambda_u); \\
\frac{\partial \ln\psi_{u,v}}{\partial \nu_u^2} &= -\frac{1}{2}\mathrm{Tr}(V^{-1}A\boldsymbol{a}\boldsymbol{a}^\intercal A^\intercal) - \frac{1}{2}(y - M\mu)^\intercal A\boldsymbol{a}\boldsymbol{a}^\intercal A^\intercal(y - M\mu); \\
\frac{\partial \ln\psi_{u,v}}{\partial \sigma_v^2} &= -\frac{1}{2}\mathrm{Tr}(V^{-1}B\boldsymbol{b}\boldsymbol{b}^\intercal B^\intercal) - \frac{1}{2}(y - M\mu)^\intercal B\boldsymbol{b}\boldsymbol{b}^\intercal B^\intercal(y - M\mu); \\
\frac{\partial \ln\psi_{u,v}}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} - \frac{1}{2}(y - M\mu)^\intercal(y - M\mu).
\end{aligned}$$

The Hessian matrix of $l(\boldsymbol{\theta}, \boldsymbol{y}), \boldsymbol{H}(\boldsymbol{\theta})$ is defined as the second derivative of $l(\boldsymbol{\theta}, \boldsymbol{y})$ with respect to $\boldsymbol{\theta}$:

$$\boldsymbol{H}(\boldsymbol{\theta}) \equiv \frac{\partial^2 l(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal}.$$

And the Hessian matrix under block representation is:

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 l}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\intercal} & \frac{\partial^2 l}{\partial \boldsymbol{a} \partial \boldsymbol{b}^\intercal} & \frac{\partial^2 l}{\partial \boldsymbol{a} \partial \nu^2} & \frac{\partial^2 l}{\partial \boldsymbol{a} \partial \sigma^2} & \frac{\partial^2 l}{\partial \boldsymbol{a} \partial \sigma_0^2} \\ & \frac{\partial^2 l}{\partial \boldsymbol{b} \partial \boldsymbol{b}^\intercal} & \frac{\partial^2 l}{\partial \boldsymbol{b} \partial \nu^2} & \frac{\partial^2 l}{\partial \boldsymbol{b} \partial \sigma^2} & \frac{\partial^2 l}{\partial \boldsymbol{b} \partial \sigma_0^2} \\ & & \frac{\partial^2 l}{\partial \nu^2 \partial \nu^2} & \frac{\partial^2 l}{\partial \nu^2 \partial \sigma^2} & \frac{\partial^2 l}{\partial \nu^2 \partial \sigma_0^2} \\ & & & \frac{\partial^2 l}{\partial \sigma^2 \partial \sigma^2} & \frac{\partial^2 l}{\partial \sigma^2 \partial \sigma_0^2} \\ & & & & \frac{\partial^2 l}{\partial \sigma_0^2 \partial \sigma_0^2} \end{bmatrix}. \tag{6.3}$$

Second derivative is obtained, again by computing differential on the first derivatives. Take $\frac{\partial \ln \psi_{u,v}}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\intercal}$ for example. The first derivative can be divided into three parts:

$$\begin{aligned} \frac{\partial \ln \psi}{\partial \boldsymbol{a}} &= \nu^2 A^\intercal V^{-1}(y - M\mu)(y - M\mu)^\intercal V^{-1} A\boldsymbol{a} + \lambda A^\intercal V^{-1}(y - M\mu) - \nu^2 A^\intercal V^{-1} A\boldsymbol{a} \\ &= \nu^2 T_1 \quad + \lambda T_2 \quad - \nu^2 T_3, \\ dT_1 &= A^\intercal d(V^{-1}) Y Y^\intercal V^{-1} A\boldsymbol{a} + (dY) Y^\intercal V^{-1} A\boldsymbol{a} + A^\intercal Y (dY)^\intercal d(V^{-1}) A\boldsymbol{a}, \\ &\quad + A^\intercal V^{-1} Y Y^\intercal V^{-1} A d\boldsymbol{a}, \\ dT_2 &= A^\intercal d(V^{-1}) Y + A^\intercal V^{-1} dY, \\ dT_3 &= A^\intercal d(V^{-1}) A\boldsymbol{a} + A^\intercal V^{-1} A d\boldsymbol{a}. \end{aligned}$$

Still, combine terms with $d\boldsymbol{a}$ and $(d\boldsymbol{a})^\intercal$. Every term containing $d\boldsymbol{a}$ will be in the form of $P d\boldsymbol{a} Q$, where P is $p \times p$ matrix and Q is scalar. Then part of Hessian matrix contributed by this term will be $Q \times P$. Term associated with $(d\boldsymbol{a})^\intercal$ will be in the form of $P(d\boldsymbol{a})^\intercal Q$, where P is a column vector and Q a column vector. Hessian matrix contributed by this term should be outer product $PQ^\intercal$.

Thus we are able to derive all block matrices on the diagonal in equation (6.3):

$$
\begin{aligned}
\frac{\partial \ln \psi_{u,v}}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\intercal} &= \nu_u^2 (M_1 + M_2 + M_3) + M_4 - M_5 \\
M_1 &= \nu_u^2 A^\intercal V^{-1} A \boldsymbol{a} \boldsymbol{a}^\intercal A^\intercal V^{-1} (y - M\mu)(y - M\mu)^\intercal V^{-1} A \\
&\quad - \lambda_u A^\intercal V^{-1} (y - M\mu) \boldsymbol{a}^\intercal A^\intercal V^{-1} A + \nu_u^2 A^\intercal V^{-1} (y - M\mu)(y - M\mu)^\intercal V^{-1} A \boldsymbol{a} \boldsymbol{a}^\intercal A^\intercal V^{-1} A \\
M_2 &= \nu_u^2 A^\intercal V^{-1} A \boldsymbol{a}^\intercal A^\intercal V^{-1} (y - M\mu)(y - M\mu)^\intercal V^{-1} A \boldsymbol{a} \\
&\quad - \lambda_u A^\intercal V^{-1} A (y - M\mu)^\intercal V^{-1} A \boldsymbol{a} + \nu_u^2 A^\intercal V^{-1} (y - M\mu)(y - M\mu)^\intercal V^{-1} A \boldsymbol{a}^\intercal A^\intercal V^{-1} A \boldsymbol{a} \\
M_3 &= A^\intercal V^{-1} (y - M\mu)(y - M\mu)^\intercal V^{-1} A \\
M_4 &= \lambda_u \nu_u^2 [A^\intercal V^{-1} A \boldsymbol{a} (y - M\mu)^\intercal A + A^\intercal V^{-1} A \boldsymbol{a}^\intercal A^\intercal (y - M\mu)] - \lambda_u^2 A^\intercal V^{-1} A \\
M_5 &= \nu_u^4 [A^\intercal V^{-1} A \boldsymbol{a} \boldsymbol{a}^\intercal A^\intercal V^{-1} A + A^\intercal V^{-1} A \boldsymbol{a}^\intercal A^\intercal V^{-1} A \boldsymbol{a}] + \nu_u^2 A^\intercal V^{-1} A \\
\frac{\partial \ln \psi_{u,v}}{\partial \boldsymbol{b} \partial \boldsymbol{b}^\intercal} &= \sigma_v^2 (M_1 + M_2 + M_3) - M_4 + M_5 \\
M_1 &= \sigma_v^2 B^\intercal V^{-1} B \boldsymbol{b} \boldsymbol{b}^\intercal B^\intercal V^{-1} (y - M\mu)(y - M\mu)^\intercal V^{-1} B \\
&\quad - \mu_v B^\intercal V^{-1} (y - M\mu) \boldsymbol{b}^\intercal B^\intercal V^{-1} B + \sigma_v^2 B^\intercal V^{-1} (y - M\mu)(y - M\mu)^\intercal V^{-1} B \boldsymbol{b} \boldsymbol{b}^\intercal B^\intercal V^{-1} B \\
M_2 &= \sigma_v^2 B^\intercal V^{-1} B \boldsymbol{b}^\intercal B^\intercal V^{-1} (y - M\mu)(y - M\mu)^\intercal V^{-1} B \boldsymbol{b} \\
&\quad - \mu_v B^\intercal V^{-1} B (y - M\mu)^\intercal V^{-1} B \boldsymbol{b} + \sigma_v^2 B^\intercal V^{-1} (y - M\mu)(y - M\mu)^\intercal V^{-1} B \boldsymbol{b}^\intercal B^\intercal V^{-1} B \boldsymbol{b} \\
M_3 &= B^\intercal V^{-1} (y - M\mu)(y - M\mu)^\intercal V^{-1} B \\
M_4 &= \mu_v \sigma_v^2 [B^\intercal V^{-1} B \boldsymbol{b} (y - M\mu)^\intercal B + B^\intercal V^{-1} B \boldsymbol{b}^\intercal B^\intercal (y - M\mu)] - \mu_v^2 B^\intercal V^{-1} B \\
M_5 &= \sigma_v^4 [B^\intercal V^{-1} B \boldsymbol{b} \boldsymbol{b}^\intercal B^\intercal V^{-1} B + B^\intercal V^{-1} B \boldsymbol{b}^\intercal B^\intercal V^{-1} B \boldsymbol{b}] + \sigma_v^2 B^\intercal V^{-1} B \\
\frac{\partial \ln \psi_{u,v}}{\partial \lambda_u \partial \lambda_u} &= -(A\boldsymbol{a})^\intercal V^{-1} A \boldsymbol{a} \\
\frac{\partial \ln \psi_{u,v}}{\partial \mu_v \partial \mu_v} &= -(B\boldsymbol{b})^\intercal V^{-1} B \boldsymbol{b} \\
\frac{\partial \ln \psi_{u,v}}{\partial \nu_u^2 \partial \nu_u^2} &= -\frac{1}{2} (\boldsymbol{a}^\intercal A^\intercal V^{-1} A \boldsymbol{a})^2 \\
\frac{\partial \ln \psi_{u,v}}{\partial \sigma_v^2 \partial \sigma_v^2} &= -\frac{1}{2} (\boldsymbol{b}^\intercal B^\intercal V^{-1} B \boldsymbol{b})^2 \\
\frac{\partial \ln \psi_{u,v}}{\partial \sigma^2 \partial \sigma^2} &= \frac{n}{2\sigma^4}
\end{aligned}
\tag{6.4}
$$

And these are the off-diagonal terms:

$$
\begin{aligned}
\frac{\ln\psi_{u,v}}{\partial \boldsymbol{a}\partial \boldsymbol{b}^\intercal} &= \nu_u^2(M_1+M_2)-M_3+M_4,\\
M1 &= \sigma_v^2 A^\intercal V^{-1}B\boldsymbol{b}\boldsymbol{a}^\intercal AV^{-1}(y-M\mu)(y-M\mu)^\intercal V^{-1}B\\
&\quad -\mu_v A^\intercal V^{-1}(y-M\mu)\boldsymbol{a}^\intercal A^\intercal V^{-1}B+\sigma_v^2 A^\intercal V^{-1}(y-M\mu)(y-M\mu)^\intercal V^{-1}B\boldsymbol{b}\boldsymbol{a}^\intercal A^\intercal V^{-1}B,\\
M_2 &= \sigma_v^2 A^\intercal V^{-1}B\boldsymbol{b}^\intercal B^\intercal V^{-1}(y-M\mu)(y-M\mu)^\intercal V^{-1}A\boldsymbol{a}\\
&\quad -\mu_v A^\intercal V^{-1}B(y-M\mu)^\intercal V^{-1}A\boldsymbol{a}+\sigma_v^2 A^\intercal V^{-1}(y-M\mu)(y-M\mu)^\intercal V^{-1}B\boldsymbol{b}^\intercal B^\intercal V^{-1}A\boldsymbol{a},\\
M_3 &= \lambda_u\sigma_v^2[A^\intercal V^{-1}B\boldsymbol{b}(y-M\mu)^\intercal V^{-1}B+A^\intercal V^{-1}B\boldsymbol{b}^\intercal B^\intercal V^{-1}(y-M\mu)]-\lambda_u\mu_v A^\intercal V^{-1}B,\\
M_4 &= \nu_u^2\sigma_v^2[A^\intercal V^{-1}B\boldsymbol{b}\boldsymbol{a}^\intercal A^\intercal V^{-1}B+A^\intercal V^{-1}B\boldsymbol{b}^\intercal B^\intercal V^{-1}A\boldsymbol{a}],\\
\frac{\partial\ln\psi_{u,v}}{\partial\boldsymbol{a}\partial\lambda_u} &= 2\nu_u^2A^\intercal V^{-1}[\lambda_u A\boldsymbol{a}\boldsymbol{a}^\intercal A^\intercal-A\boldsymbol{a}(y-\lambda_u B\boldsymbol{b})^\intercal]V^{-1}A\boldsymbol{a}+A^\intercal V^{-1}(y-M\mu)-A^\intercal V^{-1}A\boldsymbol{a},\\
\frac{\partial\ln\psi_{u,v}}{\partial\boldsymbol{a}\partial\mu_v} &= 2\nu_u^2A^\intercal V^{-1}[\mu_v B\boldsymbol{b}\boldsymbol{b}^\intercal B^\intercal-B\boldsymbol{b}(y-\mu_v A\boldsymbol{a})^\intercal]V^{-1}A\boldsymbol{a}-A^\intercal V^{-1}B\boldsymbol{b},\\
\frac{\partial\ln\psi_{u,v}}{\partial\boldsymbol{a}\partial\nu_u^2} &= A^\intercal V^{-1}(y-M\mu)(y-M\mu)^\intercal V^{-1}A\boldsymbol{a},\\
&\quad +\nu_u^2A^\intercal V^{-1}A\boldsymbol{a}\boldsymbol{a}^\intercal A^\intercal V^{-1}(y-M\mu)(y-M\mu)^\intercal V^{-1}A\boldsymbol{a}\\
&\quad +\nu_v^2A^\intercal V^{-1}(y-M\mu)(y-M\mu)^\intercal V^{-1}A\boldsymbol{a}\boldsymbol{a}A^\intercal V^{-1}A\boldsymbol{a}\\
&\quad +\lambda_u A^\intercal V^{-1}A\boldsymbol{a}\boldsymbol{a}^\intercal A^\intercal V^{-1}(y-M\mu)-A^\intercal V^{-1}A\boldsymbol{a}-\nu_u^2A^\intercal V^{-1}A\boldsymbol{a}\boldsymbol{a}^\intercal A^\intercal V^{-1}A\boldsymbol{a},\\
\frac{\partial\ln\psi_{u,v}}{\partial\boldsymbol{a}\partial\sigma_v^2} &= \nu_u^2A^\intercal V^{-1}B\boldsymbol{b}\boldsymbol{b}^\intercal B^\intercal V^{-1}(y-M\mu)(y-M\mu)^\intercal V^{-1}A\boldsymbol{a}\\
&\quad +\nu_v^2A^\intercal V^{-1}(y-M\mu)(y-M\mu)^\intercal V^{-1}B\boldsymbol{b}\boldsymbol{b}^\intercal B^\intercal V^{-1}A\boldsymbol{a}\\
&\quad +\lambda_u A^\intercal V^{-1}B\boldsymbol{b}\boldsymbol{b}^\intercal B^\intercal V^{-1}(y-M\mu)-\nu_u^2A^\intercal V^{-1}B\boldsymbol{b}\boldsymbol{b}^\intercal B^\intercal V^{-1}A\boldsymbol{a},\\
\frac{\partial\ln\psi_{u,v}}{\partial\boldsymbol{a}\partial\sigma^2} &= \nu_u^2A^\intercal V^{-1}V^{-1}(y-M\mu)(y-M\mu)^\intercal V^{-1}A\boldsymbol{a}\\
&\quad +\nu_v^2A^\intercal V^{-1}(y-M\mu)(y-M\mu)^\intercal V^{-1}V^{-1}A\boldsymbol{a}\\
&\quad +\lambda_u A^\intercal V^{-1}V^{-1}(y-M\mu)-\nu_u^2A^\intercal V^{-1}V^{-1}A\boldsymbol{a}.
\end{aligned}
\tag{6.5}
$$

In Special case of $U = V = 1$, both natural growth and diarrhea multiplier follows a Gaussian distribution. Then matrix (6.3) is in fact the Hessian matrix of marginal likelihood defined in (6.1). As mentioned in Section 1.5, under some regular conditions, $\boldsymbol{F}(\boldsymbol{\theta}) = E(\boldsymbol{H}(\boldsymbol{\theta}))$. And since

$$Y \sim N(\mathbf{0}, \boldsymbol{V}),$$

the Fisher information matrix is determined by these block matrices:

$$\begin{aligned}
E(\frac{\partial^2 l}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\intercal}) &= \nu^4 A^\intercal V^{-1} A \boldsymbol{a}\boldsymbol{a}^\intercal A^\intercal V^{-1} A + \nu^4 (\boldsymbol{a}^\intercal A^\intercal V^{-1} A \boldsymbol{a}) A^\intercal V^{-1} A + \mu_0^2 A^\intercal V^{-1} A, \\
E(\frac{\partial^2 l}{\partial \boldsymbol{a} \partial \boldsymbol{b}^\intercal}) &= \nu^2 \sigma^2 A^\intercal V^{-1} A \boldsymbol{a}\boldsymbol{b}^\intercal B^\intercal V^{-1} B + \nu^2 \sigma^2 (\boldsymbol{a}^\intercal A^\intercal V^{-1} B \boldsymbol{b}) A^\intercal V^{-1} B + \mu_0 \mu_1 A^\intercal V^{-1} B, \\
E(\frac{\partial^2 l}{\partial \boldsymbol{a} \partial \nu^2}) &= \nu^2 (\boldsymbol{a}^\intercal A^\intercal V^{-1} A \boldsymbol{a}) A^\intercal V^{-1} A \boldsymbol{a}, \\
E(\frac{\partial^2 l}{\partial \boldsymbol{a} \partial \sigma^2}) &= \nu^2 (\boldsymbol{a}^\intercal A^\intercal V^{-1} B \boldsymbol{b}) A^\intercal V^{-1} B \boldsymbol{b}, \\
E(\frac{\partial^2 l}{\partial \boldsymbol{a} \partial \sigma_0^2}) &= \nu^2 A^\intercal V^{-1} V^{-1} A \boldsymbol{a}, \\
E(\frac{\partial^2 l}{\partial \nu^2 \partial \nu^2}) &= 0.5 (\boldsymbol{a}^\intercal A^\intercal V^{-1} A \boldsymbol{a})^2, \\
E(\frac{\partial^2 l}{\partial \sigma^2 \partial \sigma^2}) &= 0.5 (\boldsymbol{b}^\intercal B^\intercal V^{-1} B \boldsymbol{b})^2, \\
E(\frac{\partial^2 l}{\partial \sigma_0^2 \partial \sigma_0^2}) &= 0.5 \mathrm{Tr}(V^{-1} V^{-1}), \\
E(\frac{\partial^2 l}{\partial \nu^2 \partial \sigma^2}) &= 0.5 (\boldsymbol{a}^\intercal A^\intercal V^{-1} B \boldsymbol{b})^2, \\
E(\frac{\partial^2 l}{\partial \nu^2 \partial \sigma_0^2}) &= 0.5 \mathrm{Tr}(V^{-1} A \boldsymbol{a}\boldsymbol{a}^\intercal A^\intercal V^{-1}), \\
E(\frac{\partial^2 l}{\partial \sigma^2 \partial \sigma_0^2}) &= 0.5 \mathrm{Tr}(V^{-1} B \boldsymbol{b}\boldsymbol{b}^\intercal B^\intercal V^{-1}).
\end{aligned}$$

For the more general cases where $U \neq 1$ or $V \neq 1$, the gradient of l with respect to Bspline coefficient $\boldsymbol{a}, \boldsymbol{b}$ is obtained by taking derivative to all combinations of

u, v. Take $\boldsymbol{a}$ for example:

$$\frac{\partial l}{\partial \boldsymbol{a}} = \frac{\sum_{u,v} 2p_u\pi_v\psi_{u,v}^2 \frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}}}{\sum_{u,v} p_u\pi_v\psi_{u,v}^2} - \frac{\sum_{u,v} p_u\pi_v\psi_{u,v} \frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}}}{\sum_{u,v} p_u\pi_v\psi_{u,v}} \tag{6.6}$$

The Hessian matrix can still be divided into blocks. These are the diagonal blocks:

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \boldsymbol{a}^\intercal} = & 2\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2 (\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}\partial \boldsymbol{a}^\intercal} + 2[\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}}][\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}}]^\intercal)}{\sum_{u,v} p_u\pi_v\psi_{u,v}^2} \\
& - 4\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2 \frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}} \sum_{u,v} p_u\pi_v\psi_{u,v}^2 [\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}}]^\intercal}{(\sum_{u,v} p_u\pi_v\psi_{u,v}^2)^2} \\
& - \frac{\sum_{u,v} p_u\pi_v\psi_{u,v} (\frac{\partial^2 \ln\psi_{u,v}}{\partial \boldsymbol{a}\partial \boldsymbol{a}^\intercal} + [\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}}][\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}}]^\intercal)}{\sum_{u,v} p_u\pi_v\psi_{u,v}} \\
& + \frac{\sum_{u,v} p_u\pi_v\psi_{u,v} \frac{\ln\psi_{u,v}}{\partial \boldsymbol{a}} \sum_{u,v} p_u\pi_v\psi_{u,v} [\frac{\ln\psi_{u,v}}{\partial \boldsymbol{a}}]^\intercal}{(\sum_{u,v} p_u\pi_v\psi_{u,v})^2} \\
\frac{\partial^2 l}{\partial \boldsymbol{b}\partial \boldsymbol{b}^\intercal} = & 2\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2 (\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{b}\partial \boldsymbol{b}^\intercal} + 2[\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{b}}][\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{b}}]^\intercal)}{\sum_{u,v} p_u\pi_v\psi_{u,v}^2} \\
& - 4\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2 \frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{b}} \sum_{u,v} p_u\pi_v\psi_{u,v}^2 [\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{b}}]^\intercal}{(\sum_{u,v} p_u\pi_v\psi_{u,v}^2)^2} \\
& - \frac{\sum_{u,v} p_u\pi_v\psi_{u,v} (\frac{\partial^2 \ln\psi_{u,v}}{\partial \boldsymbol{b}\partial \boldsymbol{b}^\intercal} + [\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{b}}][\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{b}}]^\intercal)}{\sum_{u,v} p_u\pi_v\psi_{u,v}} \\
& + \frac{\sum_{u,v} p_u\pi_v\psi_{u,v} \frac{\ln\psi_{u,v}}{\partial \boldsymbol{b}} \sum_{u,v} p_u\pi_v\psi_{u,v} [\frac{\ln\psi_{u,v}}{\partial \boldsymbol{b}}]^\intercal}{(\sum_{u,v} p_u\pi_v\psi_{u,v})^2} \\
\frac{\partial^2 l}{\partial \nu^2\partial \nu^2} = & 2\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2 (\frac{\partial \ln\psi_{u,v}}{\partial \nu^2\partial \nu^2} + 2[\frac{\partial \ln\psi_{u,v}}{\partial \nu^2}][\frac{\partial \ln\psi_{u,v}}{\partial \nu^2}])}{\sum_{u,v} p_u\pi_v\psi_{u,v}^2} \\
& - 4\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2 \frac{\partial \ln\psi_{u,v}}{\partial \nu^2} \sum_{u,v} p_u\pi_v\psi_{u,v}^2 [\frac{\partial \ln\psi_{u,v}}{\partial \nu^2}]}{(\sum_{u,v} p_u\pi_v\psi_{u,v}^2)^2} \\
& - \frac{\sum_{u,v} p_u\pi_v\psi_{u,v} (\frac{\partial^2 \ln\psi_{u,v}}{\partial \nu^2\partial \nu^2} + [\frac{\partial \ln\psi_{u,v}}{\partial \nu^2}][\frac{\partial \ln\psi_{u,v}}{\partial \nu^2}])}{\sum_{u,v} p_u\pi_v\psi_{u,v}} \\
& + \frac{\sum_{u,v} p_u\pi_v\psi_{u,v} \frac{\ln\psi_{u,v}}{\partial \nu^2} \sum_{u,v} p_u\pi_v\psi_{u,v} [\frac{\ln\psi_{u,v}}{\partial \nu^2}]}{(\sum_{u,v} p_u\pi_v\psi_{u,v})^2}
\end{aligned}$$

$$\frac{\partial^2 l}{\partial\sigma^2\partial\sigma^2} = 2\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2(\frac{\partial \ln\psi_{u,v}}{\partial\sigma^2\partial\sigma^2} + 2[\frac{\partial \ln\psi_{u,v}}{\partial\sigma^2}][\frac{\partial \ln\psi_{u,v}}{\partial\sigma^2}])}{\sum_{u,v} p_u\pi_v\psi_{u,v}^2}$$
$$- 4\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2 \frac{\partial \ln\psi_{u,v}}{\partial\sigma^2}\sum_{u,v} p_u\pi_v\psi_{u,v}^2[\frac{\partial \ln\psi_{u,v}}{\partial\sigma^2}]}{(\sum_{u,v} p_u\pi_v\psi_{u,v}^2)^2}$$
$$- \frac{\sum_{u,v} p_u\pi_v\psi_{u,v}(\frac{\partial^2 \ln\psi_{u,v}}{\partial\sigma^2\partial\sigma^2} + [\frac{\partial \ln\psi_{u,v}}{\partial\sigma^2}][\frac{\partial \ln\psi_{u,v}}{\partial\sigma^2}])}{\sum_{u,v} p_u\pi_v\psi_{u,v}}$$
$$+ \frac{\sum_{u,v} p_u\pi_v\psi_{u,v}\frac{\ln\psi_{u,v}}{\partial\sigma^2}\sum_{u,v} p_u\pi_v\psi_{u,v}[\frac{\ln\psi_{u,v}}{\partial\sigma^2}]}{(\sum_{u,v} p_u\pi_v\psi_{u,v})^2}$$

$$\frac{\partial^2 l}{\partial\sigma_0^2\partial\sigma_0^2} = 2\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2(\frac{\partial \ln\psi_{u,v}}{\partial\sigma_0^2\partial\sigma_0^2} + 2[\frac{\partial \ln\psi_{u,v}}{\partial\sigma_0^2}][\frac{\partial \ln\psi_{u,v}}{\partial\sigma_0^2}])}{\sum_{u,v} p_u\pi_v\psi_{u,v}^2}$$
$$- 4\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2 \frac{\partial \ln\psi_{u,v}}{\partial\sigma_0^2}\sum_{u,v} p_u\pi_v\psi_{u,v}^2[\frac{\partial \ln\psi_{u,v}}{\partial\sigma_0^2}]}{(\sum_{u,v} p_u\pi_v\psi_{u,v}^2)^2}$$
$$- \frac{\sum_{u,v} p_u\pi_v\psi_{u,v}(\frac{\partial^2 \ln\psi_{u,v}}{\partial\sigma_0^2\partial\sigma_0^2} + [\frac{\partial \ln\psi_{u,v}}{\partial\sigma_0^2}][\frac{\partial \ln\psi_{u,v}}{\partial\sigma_0^2}])}{\sum_{u,v} p_u\pi_v\psi_{u,v}}$$
$$+ \frac{\sum_{u,v} p_u\pi_v\psi_{u,v}\frac{\ln\psi_{u,v}}{\partial\sigma_0^2}\sum_{u,v} p_u\pi_v\psi_{u,v}[\frac{\ln\psi_{u,v}}{\partial\sigma_0^2}]}{(\sum_{u,v} p_u\pi_v\psi_{u,v})^2}$$

The the off-diagonal blocks will follow similar pattern, we only list three typical components in matrix, vector, scalar form respectively. The rest of block matrix can be obtained using symmetry between paramters, and replace $\boldsymbol{a}$ for the corresponding

variable.

$$\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \boldsymbol{b}^\intercal} = 2\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2(\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}\partial \boldsymbol{b}^\intercal} + 2[\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}}][\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{b}}]^\intercal)}{\sum_{u,v} p_u\pi_v\psi_{u,v}^2} - 4\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2 \frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}} \sum_{u,v} p_u\pi_v\psi_{u,v}^2[\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{b}}]^\intercal}{(\sum_{u,v} p_u\pi_v\psi_{u,v}^2)^2} - \frac{\sum_{u,v} p_u\pi_v\psi_{u,v}(\frac{\partial^2 \ln\psi_{u,v}}{\partial \boldsymbol{a}\partial \boldsymbol{b}^\intercal} + [\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}}][\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{b}}]^\intercal)}{\sum_{u,v} p_u\pi_v\psi_{u,v}} + \frac{\sum_{u,v} p_u\pi_v\psi_{u,v} \frac{\ln\psi_{u,v}}{\partial \boldsymbol{a}} \sum_{u,v} p_u\pi_v\psi_{u,v}[\frac{\ln\psi_{u,v}}{\partial \boldsymbol{b}}]^\intercal}{(\sum_{u,v} p_u\pi_v\psi_{u,v})^2}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{a}\partial \nu^2} = 2\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2(\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}\partial \nu^2} + 2[\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}}][\frac{\partial \ln\psi_{u,v}}{\partial \nu^2}])}{\sum_{u,v} p_u\pi_v\psi_{u,v}^2} - 4\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2 \frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}} \sum_{u,v} p_u\pi_v\psi_{u,v}^2[\frac{\partial \ln\psi_{u,v}}{\partial \nu^2}]}{(\sum_{u,v} p_u\pi_v\psi_{u,v}^2)^2} - \frac{\sum_{u,v} p_u\pi_v\psi_{u,v}(\frac{\partial^2 \ln\psi_{u,v}}{\partial \boldsymbol{a}\partial \nu^2} + [\frac{\partial \ln\psi_{u,v}}{\partial \boldsymbol{a}}][\frac{\partial \ln\psi_{u,v}}{\partial \nu^2}])}{\sum_{u,v} p_u\pi_v\psi_{u,v}} + \frac{\sum_{u,v} p_u\pi_v\psi_{u,v} \frac{\ln\psi_{u,v}}{\partial \boldsymbol{a}} \sum_{u,v} p_u\pi_v\psi_{u,v}[\frac{\ln\psi_{u,v}}{\partial \nu^2}]}{(\sum_{u,v} p_u\pi_v\psi_{u,v})^2}$$

$$\frac{\partial^2 l}{\partial \sigma^2\partial \nu^2} = 2\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2(\frac{\partial \ln\psi_{u,v}}{\partial \sigma^2\partial \nu^2} + 2[\frac{\partial \ln\psi_{u,v}}{\partial \sigma^2}][\frac{\partial \ln\psi_{u,v}}{\partial \nu^2}])}{\sum_{u,v} p_u\pi_v\psi_{u,v}^2} - 4\frac{\sum_{u,v} p_u\pi_v\psi_{u,v}^2 \frac{\partial \ln\psi_{u,v}}{\partial \sigma^2} \sum_{u,v} p_u\pi_v\psi_{u,v}^2[\frac{\partial \ln\psi_{u,v}}{\partial \nu^2}]}{(\sum_{u,v} p_u\pi_v\psi_{u,v}^2)^2} - \frac{\sum_{u,v} p_u\pi_v\psi_{u,v}(\frac{\partial^2 \ln\psi_{u,v}}{\partial \sigma^2\partial \nu^2} + [\frac{\partial \ln\psi_{u,v}}{\partial \sigma^2}][\frac{\partial \ln\psi_{u,v}}{\partial \nu^2}])}{\sum_{u,v} p_u\pi_v\psi_{u,v}} + \frac{\sum_{u,v} p_u\pi_v\psi_{u,v} \frac{\ln\psi_{u,v}}{\partial \sigma^2} \sum_{u,v} p_u\pi_v\psi_{u,v}[\frac{\ln\psi_{u,v}}{\partial \nu^2}]}{(\sum_{u,v} p_u\pi_v\psi_{u,v})^2}$$

Efron and Hinkley (1978) demonstrate that for estimating scalar parameter θ,

the conditional variance of $\hat{\theta}_n$ is better approximated by $\bar{H}_n(\hat{\theta}_n)^{-1}$ than by $\bar{F}_n(\hat{\theta}_n)^{-1}$. For the more general matrix case, there is not yet a solid theoretical validation for the better choice between $\bar{\boldsymbol{F}}_n(\hat{\boldsymbol{\theta}}_n)^{-1}$ and $\bar{\boldsymbol{H}}_n(\hat{\boldsymbol{\theta}}_n)^{-1}$. In fact, people in practice tend to choose one or the other, depending on which one is easier to obtain for their problems (Cao, 2013). Due to the structure introduced by Gaussian mixture, taking expectation of the Hessian matrix is difficult in the general case. Instead, the observed Fisher information matrix can be utilized here: $\bar{\boldsymbol{H}}_n(\boldsymbol{\theta}^*)^{-1}$. It is the Hessian matrix evaluated at the MLE $\hat{\boldsymbol{\theta}}_n$.

Bibliography

[1] Schumaker, L.L. (1981). *Spline functions: basic theory*. John Wiley and Sons, New York.

[2] de Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory* **6** 50-62

[3] J. O. Ramsay. (2006). *Functional data analysis*. Wiley Online Library.

[4] Gelman, A., Carlin, J., Stern, H., Dunson,D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*. CRC Press

[5] 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.

[6] Derso, T., Tariku, A., Biks, G. A., and 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

[7] Prentice, A. M., Ward, K. A., Goldberg, G. R., Jarjou, L. M., Moore, S. E., Fulford, A. J., and Prentice, A. (2013). Critical windows for nutritional

interventions against stunting. *The American of Clinical Nutrition*, 97(5), 911-918

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

[9] Schnee, A. E., Haque, R., Taniuchi, M., Uddin, M. J., Alam, M. M., Liu, J., ... and 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.

[10] Troeger, C., Forouzanfar, M., Rao, P. C., Khalil, I., Brown, A., Reiner Jr, R. C., ... and 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.

[11] Troeger, C., Colombara, D. V., Rao, P. C., Khalil, I. A., Brown, A., Brewer, T. G., ... and 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.

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

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

[14] Victora, C.G., de Onis, M., Hallal, P.C., Blossner, M., and Shrimpton, R. (2015). Worldwide timing of growth faltering: revisiting implications for interventions. *Pediatrics* 125, e473–80.

[15] Piwoz, E., Sundberg, S., Rooke, J. (2012). Promoting healthy growth: what are the priorities for research and action? *Advances in Nutrition*, 3:234–41.

[16] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1), 1-38.

[17] Ljung, L. (1999), *System Identification Theory for the User (2nd edition)*, Prentice Hall PTR, Upper Saddle River, NJ.

[18] Cavanaugh, J. E. and Shumway, R. H. (1996), On Computing the Expected Fisher Information Matrix for State-Space Model Parameters, *Statistics and Probability Letters*, vol. 26, no. 4, pp. 347–355.

[19] Cao, X. (2013). Relative Performance of Expected and Observed Fisher Information in Covariance Estimation for Maximum Likelihood Estimates, John Hopkins U, PhD dissertation.

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

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

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

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

[24] Efron, B., and Hinkley, D. (1978). Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information. Biometrika, 65(3), 457-482.

[25] Schwarz, G. (1978). Estimating the Dimension of a Model, The Annals of Statistics, 6(2), 461-464

[26] Delattre, M., Lavielle, M., and Poursat, M., (2014). A note on BIC in mixed-effects models, Electron. J. Statist. 8 (1) 456 - 475, 2014.

[27] Magnus, J. and Neudecker, H., (1999). Matrix Differential Calculus with Applications in Statistics and Econometrics, John Wiley, Second edition

[28] Tong, X., Zhang, T., and Zhou, J. (2020). Robust Bayesian growth curve modelling using conditional medians. The British journal of mathematical and statistical psychology.