# MULTIVARIATE SEMIPARAMETRIC REGRESSION MODELS FOR LONGITUDINAL DATA

Zhuokai Li

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

December 2014

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

Wanzhu Tu, Ph.D., Co-Chair

Doctoral Committee

Hai Liu, Ph.D., Co-Chair

Barry P. Katz, Ph.D.

October 31, 2014

J. Dennis Fortenberry, M.D., M.S.

# DEDICATION

*To my loving parents,*

*Jianping Li and Ying Zhang,*

*who have been the role models in my life.*

*To my dear husband,*

*Yu Wang,*

*who has always been supportive, patient and encouraging.*

*To our sweet little boy,*

*Jason Muxiao Wang,*

*who has made me want to be a better and stronger person.*

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Dr. Wanzhu Tu and Dr. Hai Liu, my co-advisors, for their tremendous guidance and support throughout my dissertation research. I am very fortunate to have Dr. Tu as my mentor in the last three years. His commitment to research and pursuit of quality has set an excellent example for me. He has always been willing to help me without reservation and to give me professional advice with his exceptional insight. Along with Dr. Tu, Dr. Liu has constantly guided me through the various challenges that I have encountered in my research. I truly appreciate the time and thoughts he has devoted in this research endeavor. I would also like to thank Dr. Barry P. Katz, Dr. J. Dennis Fortenberry and Dr. Terrell Zollinger for serving on my research and advisory committees. Their constructive feedback on my proposal and dissertation has motivated me to think deeper.

I am also thankful to all faculty members in the Biostatistics Program for offering students a stimulating academic environment. I acknowledge the staff members for providing all the support needed for successful completion of my degree. Finally, I would like to thank my fellow students for their continuous help and encouragement, which has made the five years of my graduate studies a wonderful experience.

Zhuokai Li

# MULTIVARIATE SEMIPARAMETRIC REGRESSION MODELS FOR LONGITUDINAL DATA

Multiple-outcome longitudinal data are abundant in clinical investigations. For example, infections with different pathogenic organisms are often tested concurrently, and assessments are usually taken repeatedly over time. It is therefore natural to consider a multivariate modeling approach to accommodate the underlying interrelationship among the multiple longitudinally measured outcomes. This dissertation proposes a multivariate semiparametric modeling framework for such data. Relevant estimation and inference procedures as well as model selection tools are discussed within this modeling framework. The first part of this research focuses on the analytical issues concerning binary data. The second part extends the binary model to a more general situation for data from the exponential family of distributions. The proposed model accounts for the correlations across the outcomes as well as the temporal dependency among the repeated measures of each outcome within an individual. An important feature of the proposed model is the addition of a bivariate smooth function for the depiction of concurrent nonlinear and possibly interacting influences of two independent variables on each outcome. For model implementation, a general approach for parameter estimation is developed by using the maximum penalized likelihood method. For statistical inference, a likelihood-based resampling procedure is proposed to compare the bivariate nonlinear effect surfaces across the outcomes. The final part of the dissertation presents a variable selection tool to facilitate model development in practical data analysis. Using the adaptive least absolute shrinkage and selection operator (LASSO) penalty, the variable selection tool simultaneously identifies important fixed effects and random effects,

determines the correlation structure of the outcomes, and selects the interaction effects in the bivariate smooth functions. Model selection and estimation are performed through a two-stage procedure based on an expectation-maximization (EM) algorithm. Simulation studies are conducted to evaluate the performance of the proposed methods. The utility of the methods is demonstrated through several clinical applications.

Wanzhu Tu, Ph.D., Co-Chair

Hai Liu, Ph.D., Co-Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# Chapter 1

# Introduction

## 1.1 Research Objectives

Multiple outcome data are frequently encountered in biomedical research. For example, randomized clinical trials use multiple outcomes to gauge therapeutic response, occurrence of adverse events and changes in quality of life. Observational studies use multiple outcomes to evaluate the overall health of study participants, disease exacerbation and care cost. Laboratory experiments measures different strains of microorganisms or various types of gene mutations. A quick examination of recent biomedical publications shows that single outcome studies are becoming rarer. When multiple outcome data are collected repeatedly over time, investigators face a challenge of incorporating not only the temporal dependency among repeated measures, but also correlations among the multiple outcomes. Traditionally, such data are analyzed one outcome at a time. Separate modeling of multiple outcome data, while being easy to implement, ignores the cross-outcome associations, and thus is prone to increased estimation bias and reduced inference efficiency. A multivariate modeling approach seems a logical alternative, if the temporal and cross-outcome data dependency can be appropriately accommodated.

The objectives of this dissertation are to propose a general and flexible modeling framework for multivariate longitudinal data and to develop relevant procedures for parameter estimation, statistical inference and model selection. Specifically, a general class of models is proposed with the following features: 1) simultaneous analysis of multiple longitudinally assessed outcomes through explicit specification of the correlation structure, 2) a general model formulation applicable to different types of data distributions, 3) accommodation

of joint nonlinear influences of two independent variables as well as their interaction, 4) efficient and robust algorithms for parameter estimation, 5) comparison of covariate effects across the outcomes through hypothesis testing, and 6) selection of relevant independent variables (both random and fixed effects) and determination of the cross-outcome correlation structure.

For clarity of presentation, I divide the dissertation into three interrelated pieces. The first part discusses the analysis of multiple binary outcome data in longitudinal studies. In this part, I present a multivariate semiparametric logistic regression model with mixed effects, as motivated by a sexually transmitted infection study of three different organisms. The second part extends the modeling structure to a broader class of data distributions. This extension results in a unified analytical framework applicable to a wide range of research areas. The focus of this part is the presentation of the parameter estimation and statistical inference procedures developed for the general model. The final part of the dissertation presents a variable selection tool that helps determine the inclusion of independent variables and interaction effects as well as the correlation structure of outcomes. Together, the three pieces present a very general and flexible modeling framework along with the necessary implementation tools for the analysis of multiple outcome data.

## 1.2 Background of the Proposed Research

Statistical methodology for analysis of longitudinal data has matured rapidly in the last three decades(Fitzmaurice et al., 2004). Most of the published methods, however, have focused on the modeling of single outcomes. The existing methods can be broadly categorized into two classes, i.e., generalized estimating equation (GEE)-based marginal models (Liang and Zeger, 1986) and likelihood-based mixed effects models (Laird and Ware, 1982). Extensions of the tradition longitudinal models for multiple outcomes fall neatly into these

two categories, including multivariate version of GEE models (Rochon, 1996) and multivariate random effects models (Reinsel, 1982). The multivariate mixed effect models, in particular, explicitly specify cross-outcome correlation structure, thus offering greater flexibility in model formulation. This said, the mixed model approach has only been applied to multivariate data with normally distributed outcomes.

In this dissertation I further extend the existing multivariate mixed models to other types of data, including the most often used binary and count data. The extension covers the entire exponential family of distributions, including the normal, Bernoulli, Poisson and gamma distributions. I present the general modeling framework for this extended distribution family. General-purpose model fitting and prediction algorithms are developed by using the traditional software packages for mixed models.

Unlike the existing multivariate regression models, an important feature of the proposed model is the nonparametric bivariate smoothing component. This new component is added to the model for accommodation of nonlinear independent variable effects and their interactions. In regression analysis, semiparametric components allow more flexible modeling of nonlinear functional relationships between covariates and outcomes, whereas traditional regression methods primarily assume linear independent variable effects. While linear regression models are easy to implement and interpret, they do not always provide a good fit to the data, especially when some covariates exert nonlinear influences on the outcomes. To remedy this issue, various semiparametric regression models have been proposed in which smooth functions are incorporated for depiction of nonlinear covariate effects; most of the published semiparametric regression methods are for analysis of single outcome data (Ruppert et al., 2003). There has been limited literature discussing semiparametric models for multivariate longitudinal data. Recent work by Liu and Tu (2012) explored the use of bivariate smooth functions in the semiparametric model for a pair of continuous longitudinal

outcomes.

In this dissertation, I extend the work of Liu and Tu (2012) to discrete data situations. Specifically, the multivariate semiparametric models proposed in this dissertation feature bivariate nonparametric functions to incorporate the concurrent nonlinear effects of two potentially interacting independent variables. Graphical tools are used to depict the bivariate effect surfaces for different outcomes. Since the multivariate modeling approach provides the opportunity for cross-outcome comparisons, a hypothesis testing procedure is developed to compare the joint nonlinear covariate effects across the outcomes.

The last topic of this dissertation concerns variable selection and structural discovery. When a large number of variables are available in the data, it is of critical importance to select the best subset of variables in order to develop an informative yet parsimonious model. A traditional approach for variable selection is to use information criteria (Keselman et al., 1998; Liang et al., 2008), but its feasibility is challenged when the number of candidate models becomes too large to handle. A popular class of selection methods is based on penalized likelihood (Tibshirani, 1996; Zou, 2006). In recent years, the regularization methods have been applied to traditional and semiparametric mixed models for selection of fixed and random effects (Ibrahim et al., 2011; Ni et al., 2010). However, variable selection tools have not been developed for multivariate semiparametric models.

Using the regularization methods, this dissertation presents the first variable selection tool for multivariate semiparametric mixed models. The selection process involves all three model components, fixed effects, random effects and bivariate nonparametric functions, with the respective intentions of selecting relevant independent variables, determining the associations of the outcomes (through selection of random effects) and examining the presence of interactions between the nonlinear covariates. A two-stage algorithm is proposed to ensure the accuracy of model selection and the unbiasedness of parameter estimation.

## 1.3 Dissertation Outline

The dissertation is organized as follows. Chapter 2 proposes a multivariate semiparametric model for binary longitudinal data and illustrates the method using data from an observational study of sexually transmitted infections in young women. Chapter 3 discusses the extension of the multivariate semiparametric model to non-normal data situations and develops relevant estimation and inference procedures. Chapter 4 presents a variable selection method for the proposed model. Chapter 5 summarizes the methodological contributions of this dissertation.

# Chapter 2

## A Multivariate Semiparametric Model for Longitudinal Binary Data

This chapter presents a multivariate semiparametric mixed model for binary longitudinal data. It starts with a brief introduction to the motivating research question about sexually transmitted infections (STIs) in young people and the rationale behind the proposed model. The construction of the model is then presented in detail, followed by an analysis of the STI data which demonstrates how the proposed method helps inform STI screening strategies.

### 2.1 Scientific motivation

*Chlamydia trachomatis* (CT), *Neisseria gonorrhoeae* (GC), and *Trichomonas vaginalis* (TV) are pathogenic organisms that cause sexually transmitted infections (STIs) chlamydia, gonorrhea, and trichomoniasis, respectively. Together, they are responsible for millions of new cases each year in the United States (Cates, 1999; Centers for Disease Control and Prevention, 2011). Adolescents and young adults have assumed much of the burden of these infections. For example, epidemiological data have shown that young people aged 15 to 24 account for nearly half of the new STI cases while representing only 25% of the sexually active population in the U.S. (Weinstock et al., 2004). Within this age range, infection risk tends to vary with age as sexual behavior changes during the transition from adolescence to adulthood. However, few studies have comparatively examined the age-specific incidence rates of these infections. Besides the age-related changes in infection risk, epidemiological evidence points to a strong partner effect (Bernstein et al., 1998; Faber et al., 2011). This said, no studies, to the best of my knowledge, have examined the concurrent influences of age and sexual partners on CT, GC, and TV infection risks, and whether partner effect

changes with age. Therefore, an improved understanding of the mutually interacting effects of age and sexual partners for different organisms may aid the development of organism-specific screening strategies that target adolescents at the most appropriate ages based on their risk profile.

In this chapter, I propose a model to assess the bivariate nonlinear effects of age and number of partners on the organism-specific probability of infection acquisition. The model is constructed with the following considerations: (1) The ability to account for the synergistic relationships among the three organisms. It has been well documented that co-infections with multiple organisms are common, especially between CT and GC. In certain populations, up to 70% of GC-positive youths were co-infected with CT (Dicker et al., 2003; Kahn et al., 2005). The fact that CT, GC and TV infections tended to cluster in adolescent women may be due to the organisms' biological synergy and their common mode of transmission (Fortenberry et al., 1999; Khan et al., 2005); this provides a compelling rationale to consider a joint modeling approach. (2) Accommodation of possibly nonlinear effects of risk factors on STI acquisition. Previous studies suggested that younger adolescents were at greater risk for STI, particular with CT (Weinstock et al., 2004). Nonetheless it is unclear whether a linear age effect is adequate to quantify an individual's STI risk. Similarly, having multiple sexual partners is a strong predictor for STI acquisition (Bernstein et al., 1998; Faber et al., 2011), but evidence suggested that STI risk did not increase linearly with the number of partners, possibly due to the increased prophylactic use in individuals with multiple partners (Yu et al., 2012). (3) Accommodation of potentially interacting influences. Effects of STI risk factors are unlikely to be additive. For example, a woman's infection risk depends not only on the behaviors that expose her to a source of infection, but also on her own immunological response to the disease pathogen (Tu et al., 2011). While the number of partners marks the level of exposure, strength of host immune response may be more related

to the biological condition such as age. It is therefore important that the model correctly depicts these interacting influences. Aggregating the aforementioned features into a statistical model, I envision a multivariate semiparametric regression model in which bivariate nonlinear independent variable effects are incorporated. The joint modeling structure is used to connect organism-specific infection outcomes; the nonparametric bivariate effects are used to accommodate the nonlinear and potentially interacting influences of age and partner.

Methodologically, constructing and fitting such a model is not trivial. To the best of my knowledge, no existing models have all of the desired features. This said, various components of the model have been developed in other contexts. For example, two general approaches in multivariate regression analysis of longitudinal data have been developed. One is based on the GEE techniques (Gray and Brookmeyer, 1998; Rochon, 1996). The other is the random-effects model (Reinsel, 1982; Shah et al., 1997). Various semiparametric models have also been proposed for nonlinear independent variable effects on multiple outcomes (Coull and Staudenmayer, 2004; Ghosh and Hanson, 2010; Ghosh and Tu, 2009). More recently, Liu and Tu (2012) developed a joint semiparametric model for paired continuous outcomes, which incorporated bivariate smooth components.

In this chapter, I extend the existing methods to a multivariate semiparametric regression model for binary longitudinal data, such as the infection status with different organisms. The model accounts for the correlations across the organisms and that among the repeated measurements of the same organism over time. Joint modeling of multiple outcomes is accomplished by specifying a covariance structure through the random effects. Additionally, bivariate smoothing components are incorporated into the model for nonlinear effects of age and partner as well as their potential interactions. Finally, the proposed model is used to quantify the organism-specific infection risks, and the predictive accuracy of the model is

assessed through a receiver operating characteristic (ROC) analysis.

## 2.2 Data Source

The data that motivated this research came from a longitudinal cohort study of inner-city young women, hereafter referred to as the Young Women's Project (YWP). The study was approved by a local Institutional Review Board and its protocol was described elsewhere (Tu et al., 2009). Briefly, young women aged 14 to 17 attending three primary care clinics were recruited for participation in this observational study. At enrollment, the participants were tested for CT, GC, and TV infections; those infected were treated promptly. They also completed an interview on their lifetime and most recent sexual behaviors, including the number of sex, condom use, and the number of sexual partners in the last three months. The participants returned to clinic every three months, at which time they had face-to-face interviews and received STI tests. Infections identified at all follow-up visits were considered as incident cases (i.e., newly acquired infections) because all prior infections were treated. The mean length of follow-up was approximately 3.2 years; the longest follow-up was 7.8 years. Of 5,213 follow-up visits of all participants, CT, GC and TV infection status were missing at only 20, 23 and 1 visit(s), respectively. A high completion rate for quarterly interviews was also achieved, with only 5% of possible follow-up interviews missing.

The study sample included 386 young women, consisting of 344 (89.1%) African Americans, 39 (10.1%) non-Hispanic Whites and 3 (0.8%) Hispanics. Co-infections with different organisms were common in the study sample. Of 193 cases of GC infection, 31.6% were co-infected with CT, and 14.5% were co-infected with TV; of 287 cases of TV infection, 16.0% were co-infected with CT. At enrollment, the participants were between 14 and 17 years of age, with a mean age of 15.8 years and a standard deviation of 1.1 years.

I examined the relationship between age and the number of sexual partners in the study

participants, and found that the number of partners increased with age in early and mid-adolescence until it peaked between 19 and 20 years of age. Figure 2.1 shows the infection rates of CT, GC and TV by age group and the number of partners in the last 3 months. Both age and the number of partners appear to have a nonlinear relationship with all three types of infections. The age patterns across the organisms are different, with the highest infection rates occurring at ages $16-17$, $18-19$ and $24-25$ for CT, GC and TV, respectively. These nonlinear patterns point to the need of nonparametric regression models. Furthermore, by introducing bivariate smooth functions into the analysis, I hope to capture the potential interactions between age and the number of partners, which are not available for assessment in additive models.

## 2.3 Model Formulation

### 2.3.1 Multivariate Semiparametric Model for Binary Outcomes

Let $Y_{ij}^k$ be the $i$th individual's infection status with sexually transmitted organism $k$ at the $j$th visit, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n_i$, and $k = 1, 2, \ldots, K$, where $m$ is the number of individuals, $n_i$ is the number of follow-up visits for the $i$th individual, and $K$ is the number of sexually transmitted organisms in the study. The infection status $Y_{ij}^k$ is a binary outcome with $Y_{ij}^k = 1$ and $Y_{ij}^k = 0$ indicating positive and negative test results, respectively, for organism $k$.

Assuming $Y_{ij}^k$ follows a Bernoulli distribution with parameter $p_{ij}^k$, we propose the following model

$$g(p_{ij}^k) = \boldsymbol{S}_i^T \boldsymbol{\beta}_1^k + \boldsymbol{T}_{ij}^T \boldsymbol{\beta}_2^k + \sum_{q=1}^{Q} \beta_{3q}^k Y_{i,j-q} + \boldsymbol{Z}_{ij}^T \boldsymbol{b}_i^k + f^k(u_{ij}, v_{ij}), \tag{2.1}$$

for $k = 1, \ldots, K$, where $g(\cdot)$ is a known invertible link function, e.g., logit link. The

parameter vectors $\boldsymbol{\beta}_1^k$ and $\boldsymbol{\beta}_2^k$ represent respectively the fixed effects regression coefficients associated with time-independent covariates $\boldsymbol{S}_i$ and time-dependent covariates $\boldsymbol{T}_{ij}$. The $q$th order autoregressive component $Y_{i,j-q}$ indicates the prior infection status with any of the $k$ organisms at the $(j-q)$th visit. Let $\boldsymbol{\beta}_3^k = (\beta_{31}^k, \ldots, \beta_{3Q}^k)$ denote the coefficient vector for the autoregressive component. When the follow-up visits are approximately regularly spaced without missing data, for example, if $Q = 1$, a fixed parameter $\beta_{31}^k$ is sufficient to characterize the effect of lag-1 infection on the current status. If some of the follow-up visits are irregularly spaced or missing, a time-varying coefficient $\beta_{3q}^k(t_{i,j} - t_{i,j-q})$ can be used, with $t_{i,j}$ and $t_{i,j-q}$ being the time at the $j$th and $(j-q)$th visits. The time-varying autoregressive structure is adopted in the analysis of the YWP data in Section 2.6. We also incorporate a bivariate function $f^k(u_{ij}, v_{ij})$ in order to capture the nonlinear effects of other risk factors, such as age and the number of sexual partner, and their potential interaction effects on STIs. To accommodate the interdependence of multiple organisms within an individual as well as the correlations among the repeated measurements, the random effects $\boldsymbol{b}_i^k$ are introduced into the model, which in general, can be a random vector with multivariate normal distribution. For simplicity, a simple, scalar random effects term $b_i^k$ is assumed in the context of this example. The vector of subject-specific random effects is denoted by $\boldsymbol{b}_i = (b_i^1, \cdots, b_i^K)^T$, assuming that it follows a multivariate normal distribution, i.e., $\boldsymbol{b}_i \sim \mathrm{N}_K(\boldsymbol{0}, \boldsymbol{\Omega}_b)$, with variance-covariance matrix $\boldsymbol{\Omega}_b$.

For each bivariate smooth function in the proposed model, a set of basis functions $h_l^k, l = 1, \ldots, M_k$ is specified, so it can be expressed as $f^k(u, v) = \sum_{l=1}^{M_k} \gamma_l^k h_l^k(u, v)$, and $\boldsymbol{\gamma}_k = (\gamma_1^k, \cdots, \gamma_{M_k}^k)$ denotes the vector of regression coefficients for $f^k$. Let $\boldsymbol{f}^k$ be a vector of smooth functions with elements $f^k(u_{ij}, v_{ij})$, for $j = 1, \ldots, n_i; i = 1, \ldots, m$, i.e., $\boldsymbol{f}^k = [f^k(u_{ij}, v_{ij})]_{1 \le j \le n_i; 1 \le i \le m}$, then it can be written in a matrix form $\boldsymbol{f}^k = \mathbf{X}_k \boldsymbol{\gamma}_k$, where the design matrix $\mathbf{X}_k = [h_1^k(u_{ij}, v_{ij}), \cdots, h_{M_k}^k(u_{ij}, v_{ij})]_{1 \le j \le n_i; 1 \le i \le m}$.

In this chapter, thin plate regression splines are used to model the smooth functions, which provide good approximation to the full rank thin plate splines and significantly reduce the computational cost. Furthermore, truncated eigen-decomposition based approach is used to avoid choosing knot locations for thin plate regression splines (Wood, 2003, 2006). The smooth function estimators can be found by maximizing the penalized log-likelihood function of model (2.1),

$$\ell_p = \ell - \sum_{k=1}^{K} \lambda_k J(f^k), \tag{2.2}$$

where $\ell$ is the log-likelihood function of the model, and $\lambda_k$ is the smoothing parameter associated with $f^k$, which balances goodness-of-fit and smoothness of the model. In the case of bivariate smoothing, the roughness penalty $J(f)$ is defined as

$$J(f) = \iint_{\mathbb{R}^2} \left\{ \left( \frac{\partial^2 f}{\partial u^2} \right)^2 + 2 \left( \frac{\partial^2 f}{\partial u \partial v} \right)^2 + \left( \frac{\partial^2 f}{\partial v^2} \right)^2 \right\} \mathrm{d}u\mathrm{d}v,$$

which can be expressed as a quadratic form in regression coefficients $\boldsymbol{\gamma}_k$. For example, $J(f^k) = \boldsymbol{\gamma}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\gamma}_k / 2$, where $\boldsymbol{\Lambda}_k$ are positive semi-definite matrices of known coefficients. Therefore, the penalized log-likelihood function (2.2) can be rewritten as

$$\ell_p = \ell - \frac{1}{2} \sum_{k=1}^{K} \boldsymbol{\gamma}_k^T \mathbf{S}_k \boldsymbol{\gamma}_k, \tag{2.3}$$

where the penalty matrix $\mathbf{S}_k = \lambda_k \boldsymbol{\Lambda}_k$.

### 2.3.2 Mixed Model Representation

Semiparametric models using penalized splines can be represented by mixed effects models (Ruppert et al., 2003; Wood, 2006), and as a result, mixed model methodology and software can be adopted for the estimation of the proposed model. First, the quadratically penalized smooth functions, $\boldsymbol{f}^k$, are divided into fixed and random components of a mixed

effects model, which is achieved by using the eigen-decomposition of $\mathbf{S}_k$ (Wood, 2006). The regression coefficient vector of $\boldsymbol{f}^k$ is written as $\boldsymbol{\gamma}_k = (\boldsymbol{\gamma}_{k,F}^T, \boldsymbol{\gamma}_{k,R}^T)^T$, where $\boldsymbol{\gamma}_{k,F}$ represent unpenalized coefficients which are considered as fixed effects, and $\boldsymbol{\gamma}_{k,R}$ represent penalized coefficients which are considered as random effects. The penalty matrix corresponding to $\boldsymbol{\gamma}_{k,R}$ is denoted by $\mathbf{S}_{k,R}$ such that $\boldsymbol{\gamma}_k^T \mathbf{S}_k \boldsymbol{\gamma}_k = \boldsymbol{\gamma}_{k,R}^T \mathbf{S}_{k,R} \boldsymbol{\gamma}_{k,R}$. Accordingly, the design matrix of the smooth term $\boldsymbol{f}^k$ are partitioned into $\mathbf{X}_k = (\mathbf{X}_{k,F}, \mathbf{X}_{k,R})$.

Model (2.1) can now be rewritten as a generalized linear mixed model (GLMM). Let $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \cdots, \boldsymbol{Y}_K^T)^T$ be the response vector, where $\boldsymbol{Y}_k = [Y_{ij}^k]_{1 \le j \le n_i; 1 \le i \le m}$. The corresponding mean vector $\boldsymbol{p}$ is related to the linear predictor through a vector-valued link function $\boldsymbol{g}$. Defining $\tilde{\boldsymbol{\beta}}_k = ((\boldsymbol{\beta}_1^k)^T, (\boldsymbol{\beta}_2^k)^T, (\boldsymbol{\beta}_3^k)^T)^T$ and $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^T, \cdots, \tilde{\boldsymbol{\beta}}_K^T)^T$, the vector of fixed effects parameters is written as $\boldsymbol{\beta} = (\tilde{\boldsymbol{\beta}}^T, \boldsymbol{\gamma}_{1,F}^T, \cdots, \boldsymbol{\gamma}_{K,F}^T)^T$. Similarly, defining $\tilde{\boldsymbol{b}}_k = (b_1^k, \cdots, b_m^k)^T$ and $\tilde{\boldsymbol{b}} = (\tilde{\boldsymbol{b}}_1^T, \cdots, \tilde{\boldsymbol{b}}_K^T)^T$, the vector of random effects parameters is denoted by $\boldsymbol{b} = (\tilde{\boldsymbol{b}}^T, \boldsymbol{\gamma}_{1,R}^T, \cdots, \boldsymbol{\gamma}_{K,R}^T)^T$. The design matrix associated with $\tilde{\boldsymbol{b}}$ can be written as $\tilde{\mathbf{Z}} = \boldsymbol{I}_K \otimes \mathbf{Z}_b$ such that the components of $\mathbf{Z}_b \tilde{\boldsymbol{b}}_k$ corresponding to subject $i$ are equal to $b_i^k$. The design matrix associated with $\tilde{\boldsymbol{\beta}}$ is set up as follows: $\tilde{\mathbf{X}} = \boldsymbol{I}_K \otimes \mathbf{X}_\beta$, where $\mathbf{X}_\beta = (\mathbf{S}, \mathbf{T}, \mathbf{Y}_Q)$, and $\mathbf{S} = \mathbf{Z}_b[\boldsymbol{S}_i^T]_{1 \le i \le m}$, $\mathbf{T} = [\boldsymbol{T}_{ij}^T]_{1 \le j \le n_i; 1 \le i \le m}$ and $\mathbf{Y}_Q = [(\boldsymbol{Y}_{ij,Q}^k)^T]_{1 \le j \le n_i; 1 \le i \le m}$. Then model (2.1) can be written into a GLMM representation

$$\boldsymbol{g}(\boldsymbol{p}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b}, \tag{2.4}$$

where $\mathbf{X} = (\tilde{\mathbf{X}}, \mathrm{diag}(\mathbf{X}_{1,F}, \cdots, \mathbf{X}_{K,F}))$ and $\mathbf{Z} = (\tilde{\mathbf{Z}}, \mathrm{diag}(\mathbf{X}_{1,R}, \cdots, \mathbf{X}_{K,R}))$ are the design matrices associated with the fixed effects and the random effects, respectively. The random effects vector $\boldsymbol{b} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_b(\boldsymbol{\theta}))$, where $\boldsymbol{\Sigma}_b(\boldsymbol{\theta}) = \mathrm{diag}(\boldsymbol{\Omega}_b \otimes \boldsymbol{I}_m, \mathbf{S}_{1,R}^{-1}, \cdots, \mathbf{S}_{K,R}^{-1})$ with $\boldsymbol{\theta}$ being the variance components.

### 2.3.3 Estimation Procedure

The likelihood of the parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, given the observed data $\boldsymbol{y}$, can be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}) = |\boldsymbol{\Sigma}_b(\boldsymbol{\theta})|^{-1/2} \int \left[ \prod_{k=1}^{K} \prod_{i=1}^{m} \prod_{j=Q+1}^{n_i} (p_{ij}^k)^{y_{ij}^k} (1 - p_{ij}^k)^{1-y_{ij}^k} \right] \exp\left( -\frac{1}{2} \boldsymbol{b}^T \boldsymbol{\Sigma}_b^{-1}(\boldsymbol{\theta}) \boldsymbol{b} \right) \mathrm{d}\boldsymbol{b}, \tag{2.5}$$

where $p_{ij}^k$ is a function of $\boldsymbol{\beta}$ and $\boldsymbol{b}$, as defined in model (2.1). The integral in the likelihood function is tractable for linear mixed models where the outcome is normally distributed, but for binary outcomes it does not have a closed form expression. Instead it can be evaluated using a Laplace approximation (Barndorff-Nielsen and Cox, 1989). The approximate maximum likelihood estimators (MLEs) for parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be obtained by optimizing the Laplace approximation to the likelihood $L(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y})$.

An alternative estimation method for GLMMs is penalized quasi-likelihood (PQL) (Breslow and Clayton, 1993; Schall, 1991) in which the likelihood is replaced by a quasi-likelihood and maximized as in a linear mixed model to obtain the approximate MLEs. For binary outcomes, however, the parameter estimates for both fixed effects and variance components resulting from PQL tend to have a large bias toward zero (Goldstein and Rasbash, 1996; Ng et al., 2006; Rodriguez and Goldman, 1993). Therefore, the Laplace approximation method is used for fitting the proposed model, due to its more robust numerical performance. The details of the model fitting procedure are provided in Section 3.2.3 in the next chapter.

## 2.4 Statistical Inference

With model (2.1), one may be interested in inference on the fixed effects as well as the variance components. In the context of STI research, inference on the variance components (e.g., parameters in $\boldsymbol{\Omega}_b$) is usually of interest as they shed light on the correlations among infection outcomes associated with different organisms and the variability in STI risks in the

study population. However, a practical issue with statistical packages for fitting GLMMs is that it does not provide standard errors of variance components due to the violation of asymptotic normality assumption for Wald type confidence intervals (Bates, 2009). Therefore, the following procedure based on the bootstrap techniques (Efron, 1979) is proposed to generate a $(1-\alpha)100\%$ CI for a certain parameter (say $\theta$) in variance components:

1. Draw a bootstrap sample with replacement from the observed data. The sampling units are individuals, that is, either none or all of the records from an individual will be selected. If an individual is selected more than once, he/she will be treated as a different person each time by being assigned a new ID in the bootstrap data.
2. Fit model (2.1) to the bootstrap data and obtain a parameter estimate $\hat{\theta}^*$.
3. Repeat the above steps $B$ times to generate $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. Choose the $\frac{\alpha}{2}100\%$ and $(1-\frac{\alpha}{2})100\%$ quantiles of the bootstrap distribution $\{\hat{\theta}_b^*\}_{1\leq b\leq B}$ to form a $(1-\alpha)100\%$ CI of $\theta$.

This simple bootstrap procedure does not require any distributional assumptions on the data, while preserving the within-subject correlation structure. Percentile bootstrap CIs obtained from this procedure will always fall in their allowable ranges, which is especially desirable in our example where inference needs to be made on the correlation coefficients with a range of $[-1, 1]$.

## 2.5 Simulation Studies

### 2.5.1 Evaluation of Estimation Procedure

The first simulation study was conducted to evaluate the performance of the model estimation procedure. Two correlated binary variables $Y_{ij}^k | b_i^k, Y_{i,j-1}^k \sim \text{Bernoulli}(p_{ij}^k)$ for $i = 1, \ldots, m; j = 1, \ldots, n; k = 1, 2$ were generated using the following model

$$\text{logit}(p_{ij}^k) = \beta_0^k + \beta_1^k Y_{i,j-1}^k + b_i^k + \bar{f}_k(u_{ij}, v_{ij}), \tag{2.6}$$

where $(b_i^1, b_i^2)^T \sim \mathrm{N}(\mathbf{0}, \mathbf{\Omega}_b)$ with

$$\mathbf{\Omega}_b = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

In model (2.6), the autoregressive term $Y_{i,0}^k$ was generated from Bernoulli(0.5), $u_{ij}$ was generated from Uniform$(0, 30)$, and $v_{ij}$ was randomly sampled from $\{0, 1, \cdots, 10\}$. Two different nonlinear bivariate functions were considered: $f_1(u, v) = \exp[-(u-5)^2/200 - (v-10)^2/50 + (u+8)(v-10)/300]$ and $f_2(u, v) = \exp[-(u-18)^2/500 - (v-10)^2/40 + (u-15)(v-5)/200]$. The joint effects of $(u_{ij}, v_{ij})$ on the response variables had functional forms of $\bar{f}_1$ and $\bar{f}_2$, corresponding to the centered functions $f_1$ and $f_2$ over the simulated covariates, respectively. The fixed effects parameters were chosen as: $\beta_0^1 = -2.5$, $\beta_0^2 = -3.5$, $\beta_1^1 = 1$, and $\beta_1^2 = 0.7$. The parameters in the variance components were set to $\sigma_1 = 0.6$, $\sigma_2 = 1$, and $\rho = 0.7$.

The model performance was assessed under the following sample size settings: $m =$ 200, 400, and $n =$ 10, 20. The point estimates for the fixed effects parameters and the variance components were averaged over 200 simulation runs. The standard errors and the coverage probabilities of the 95% confidence intervals (CIs) for the parameter estimates were calculated using the proposed bootstrap procedure based on 200 bootstrap samples within each run. The mean squared errors (MSEs) of the smooth function estimates $\hat{f}_1$ and $\hat{f}_2$ (subject to a centering constraint) were also reported under each of the simulation settings.

The simulation results are presented in Table 2.1. In general, the estimation procedure performed well, and the parameter estimates approached the true values as the sample size (either the number of subjects or the number of repeated outcome measurements) increased. It can be noted that the estimation bias in the autoregressive coefficients was significantly

reduced when the number of repeated measurements increased. The coverage probabilities of the CIs were close to the nominal level 95%. The MSEs of both smooth functions steadily decreased as the sample size increased. In sum, the proposed model achieved a satisfactory performance in the estimation of parameters and bivariate smooth functions.

To further evaluate the model performance, another simulation study was conducted in a setting resembling the YWP data with three binary outcomes. Two hundred data sets were generated using a model similar to (2.6) with $m = 200$ and $n = 10$. An additional bivariate function was specified for the third outcome as $f_3(u, v) = (u-10)/120+v/30+\sqrt{35-u}/30$. In this simulation study, I compared the performance of two modeling approaches, the proposed multivariate model and the univariate models (i.e., fitting one model for each of the three outcomes). Table 2.2 provides the parameter estimates, bootstrap standard errors (SE) and coverage probabilities (CP) of the 95% bootstrap CIs based on 200 simulation runs. The multivariate model resulted in reduced estimation bias and better coverage probabilities of the CIs for most of the parameters. The standard errors estimated based on the multivariate model were consistently smaller. The efficiency improvement was more evident for the variance components. Overall, the multivariate model had improved performance in terms of estimation efficiency and accuracy as compared to the univariate models which ignored the correlations of the outcomes.

### 2.5.2 Assessment of Predictive Accuracy

In this section, a simulation study was performed to assess the predictive accuracy of the proposed model with a focus on the bivariate nonparametric components. A two-outcome setting was used with $m = 200$ and $n = 10$. The bivariate functions were defined as $f_1(u, v) = 9\exp[-4(u - 0.5)^2 - 5(v - 0.5)^2 + 4(u - 0.5)(v - 0.5)]$ and $f_2(u, v) = 4u + 3v$, where the covariates $u_{ij}$ and $v_{ij}$ were generated independently from Uniform$(0, 1)$. Here,

$f_1$ depicted the joint nonlinear effects of the two covariates with an interaction on the first outcome, whereas $f_2$ reflected the linear and additive covariate effects on the second outcome.

As an accuracy measure for prediction of binary outcomes, the area under the ROC curve (AUC) was calculated based on 10-fold cross validation (CV). Comparison of predictive performance was made for four multivariate mixed models, each using a different way to incorporate the effects of $u_{ij}$ and $v_{ij}$ in the mean structure. Model 1 was the proposed semiparametric model as specified in (2.6). Model 2 included the linear effects of $u$ and $v$ as well as their interactions, i.e., $\text{logit}(p_{ij}^k) = \beta_0^k + \beta_1^k Y_{i,j-1}^k + \beta_2^k u_{ij} + \beta_3^k v_{ij} + \beta_4^k u_{ij} v_{ij} + b_i^k$. Model 3 was similar to Model 2 except that the interaction term was removed, i.e., $\text{logit}(p_{ij}^k) = \beta_0^k + \beta_1^k Y_{i,j-1}^k + \beta_2^k u_{ij} + \beta_3^k v_{ij} + b_i^k$. Lastly, $u_{ij}$ and $v_{ij}$ were dichotomized at the medians, resulting in two categorical variables $\tilde{u}_{ij}$ and $\tilde{v}_{ij}$. Thus Model 4 was specified as $\text{logit}(p_{ij}^k) = \beta_0^k + \beta_1^k Y_{i,j-1}^k + \beta_2^k \tilde{u}_{ij} + \beta_3^k \tilde{v}_{ij} + \beta_4^k \tilde{u}_{ij} \tilde{v}_{ij} + b_i^k$. Prediction for the validation set was carried out in two steps: the fixed effects were predicted based on the estimation using the training set; the random effects were predicted by fitting a random effect model with only random intercepts on the validation set.

The simulation was repeated 200 times. For the prediction of the first outcome, the average AUC for Models $1-4$ were respectively 0.90, 0.67, 0.65 and 0.67; for the second outcome, the AUC were 0.85, 0.84, 0.85 and 0.81 respectively. The proposed model (Model 1) achieved high predictive accuracy under both situations. Compared to the other three parametric models, the semiparametric model was significantly better for predicting the first outcome when the two covariates had truly nonlinear effects and were interacting with each other; it still had an excellent performance comparable to Model 3 (with no interaction effect) in predicting linear and additive effects.

## 2.6 Real Data Analysis

### 2.6.1 Model Development

The YWP data described in Section 2.2 are used to construct the proposed model, which quantifies the organism-specific infection probability based on the risk factors including age, the number of sexual partners and an infection history. The data include 386 participants with a total of 5,213 follow-up visits. Let $Y_{ij}^{\text{ct}}$, $Y_{ij}^{\text{gc}}$, and $Y_{ij}^{\text{tv}}$ be the $i$th participant's infection status corresponding to CT, GC and TV at the $j$th visit, $i = 1, \ldots, 386$, $j = 1, \ldots, n_i$, and $n_i$ ranges from 1 to 30, with a median of 13 follow-up visits per participant.

Consider the following model

$$\begin{cases} \text{logit}(p_{ij}^{\text{ct}}) = \beta_0^{\text{ct}} + \beta_1^{\text{ct}}(t_{i,j} - t_{i,j-1})Y_{i,j-1} + b_i^{\text{ct}} + f^{\text{ct}}(u_{ij}, v_{ij}) \\ \text{logit}(p_{ij}^{\text{gc}}) = \beta_0^{\text{gc}} + \beta_1^{\text{gc}}(t_{i,j} - t_{i,j-1})Y_{i,j-1} + b_i^{\text{gc}} + f^{\text{gc}}(u_{ij}, v_{ij}) \\ \text{logit}(p_{ij}^{\text{tv}}) = \beta_0^{\text{tv}} + \beta_1^{\text{tv}}(t_{i,j} - t_{i,j-1})Y_{i,j-1} + b_i^{\text{tv}} + f^{\text{tv}}(u_{ij}, v_{ij}), \end{cases} \tag{2.7}$$

with the subject-specific random effects $\boldsymbol{b}_i = (b_i^{\text{ct}}, b_i^{\text{gc}}, b_i^{\text{tv}})^T \sim \text{N}(\mathbf{0}, \boldsymbol{\Omega}_b)$ where

$$\boldsymbol{\Omega}_b = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}.$$

In model (2.7), $p_{ij}^{\text{ct}}$, $p_{ij}^{\text{gc}}$, and $p_{ij}^{\text{tv}}$ are the corresponding means of the binary response variables conditional on the random effects $b_i^{\text{ct}}$, $b_i^{\text{gc}}$, and $b_i^{\text{tv}}$, respectively. Organism-specific intercepts are denoted by $\beta_0^{\text{ct}}$, $\beta_0^{\text{gc}}$, and $\beta_0^{\text{tv}}$, and $\beta_1^{\text{ct}}$, $\beta_1^{\text{gc}}$, and $\beta_1^{\text{tv}}$ are the organism-specific time-varying coefficients for the first-order autoregressive component $Y_{i,j-1}$, with $t_{i,j} - t_{i,j-1}$ being the lag time between the $(j-1)$th and $j$th visits. We only include the first-order

autoregressive terms because recurrent STIs can be regarded as a Markov process, where the current infection status depends on the infection status at the previous visit (Tu et al., 2011). Bivariate functions $f^{\text{ct}}$, $f^{\text{gc}}$, and $f^{\text{tv}}$ represent the joint effects of age ($u_{ij}$) and the number of partners in the last 3 months ($v_{ij}$) on CT, GC and TV, respectively.

Model (2.7) was fitted to the YWP data to obtain the parameter estimates for the fixed effects and the variance components. The standard errors and the 95% confidence intervals were computed based on 500 bootstrap samples. The estimated joint effects of age and the number of partners were depicted using colored contour plots.

### 2.6.2 Analytical Results

The model fitting results are presented in Table 2.3. Interestingly, the within-subject pairwise correlations among the three organisms are strong, especially between CT and GC ($\hat{\rho}_{12} = 0.68$, 95% CI $= [0.44, 1.00]$), suggesting that young women at high risk for infection with one organism are very likely to be infected with other organisms. Such relationships among different organisms would not be captured if they were modeled individually, thus demonstrating the usefulness of the proposed multivariate modeling approach. Figure 2.2 displays the lag time effects of a prior infection of any type on the current infection status, from which it can seen that a prior infection significantly increases the risks of CT and TV infections.

In Figure 2.3, the estimated bivariate surfaces of age and the number of partners are plotted with (right panel) or without (left panel) a prior infection of any type at the previous visit. Several important observations can be drawn from the contour plots. First, the age effect has a nonlinear pattern for CT and GC. CT infection risk peaked at younger ages between 14 and 16, and then decreased steadily after age 18. GC infection risk increased until age 19, and then gradually decreased. In contrast, TV infection risk increased almost

linearly with age. Second, the number of partners is a highly significant risk factor for all of the three organisms, though its effect tends to depend on the age of the individual. Specifically, having multiple sexual partners had a stronger effect on CT infection at younger ages, which means younger girls having multiple partners were more vulnerable to CT infection than older ones with the same number of partners.

### 2.6.3 Predictive Accuracy Assessment

The proposed model can be considered as a model-based screening algorithm to target individuals at greater STI risk. The predicted values of the STI probabilities can be used to make screening decisions on whether an individual should be tested, and if so, for what organism(s).

An ROC analysis was performed to assess the predictive accuracy of the proposed model. The probability of organism-specific infections was predicted for each participant at each visit using model (2.7). Comparing to the observed infection status, the sensitivity and specificity of the model were calculated under different cutoff points of infection probabilities, and then an ROC curve was plotted for each type of infection.

The ROC curves are shown in Figure 2.4. The areas under the curve (AUC) for CT, GC and TV are respectively 0.80, 0.87 and 0.89, indicating that the proposed model achieved excellent predictive accuracy. As a targeted screening tool, the model was able to correctly identify most individuals at high risk for further STI testing. Table 2.4 provides the sensitivity and specificity of the model under different cutoff points for the three organisms, and the corresponding percentages of follow-up visits that meet those cutoff points. In general, one hopes to have a highly sensitive screening algorithm to target high-risk individuals for formal STI testing while letting the low specificity be compensated by the diagnostic test. Based on the proposed model, for example, if individuals who have a CT infection probabil-

ity of 0.075 or greater were targeted for testing, 84% of CT infection cases would be captured while the number of tests could be reduced by more than half. Similarly, with appropriately chosen cutoff points, desired levels of sensitivity can be achieved with greatly reduced number of testing for GC and TV infections. Therefore, the proposed model-based targeted screening algorithm had an excellent performance in attaining high level of sensitivity as well as reducing testing cost.

## 2.7 Discussion

In this chapter, I have proposed a multivariate semiparametric model for the analysis of multiple binary data in a longitudinal setting. The multivariate modeling approach has the flexibility to accommodate various types of dependency structure among multiple outcomes. The bivariate smoothing component allows the exploration of concurrent nonlinear effects of two independent variables as well as their interaction effects. As shown in the STI example, without such a flexible modeling tool, many of the important but nuanced observations could be lost in an oversimplified traditional analysis. Moreover, the method is generally applicable to a much wider class of biomedical applications where exploration of multiple biological influences is desired. The model has been proposed for binary outcomes, and it has the potential to be extended for other members in the exponential family, including multiple outcomes with different distributions. These extensions will further enhance the applicability of the proposed method.

Using the model, the risks of CT, GC and TV infections can be expressed as functions of age and the number of sexual partners in a comparative manner. Previous studies have examined the age trends of these common STIs (Datta et al., 2007; Sutton et al., 2007), but few studies have directly quantified age and organism-specific STI risks in longitudinal cohorts, possibly due to the lack of appropriate analytical tools. This research has

confirmed the differential timing of the peak risks of CT, GC and TV infections, with the respective peak ages at $14-16$, $18-19$, and $24-25$ years. Furthermore, the waning partner effect on CT over age once again raises an important question about the underlying causes of the early emergence of CT infections, in comparison to the relatively late surge of TV infections (Bernstein et al., 1998; Miller et al., 2005). While the prevalence of these STIs in the partner population may in part explain the organism-specific timing of infection acquisition, this does not exclude the possibility of additional contributing factors, for example, cervico-vaginal tissue immaturity, cervical ectopy, and immunological naïveté in younger women (Ethier and Orr, 2007). The latter explanation has become particularly attractive, considering the fact that clearly different partner effects between younger and older participants has been observed. Clinically, these results can help better define the risk profiles for those common STIs in young women and thus improving the efficiency of STI screening.

Table 2.1: Parameter estimates and mean squared errors of smooth functions with bootstrap standard errors (in parentheses) and coverage probabilities of 95% confidence intervals (in brackets).

| $m$ | $n$ | $\beta_0^1 = -2.5$ | $\beta_0^2 = -3.5$ | $\beta_1^1 = 1$ | $\beta_1^2 = 0.7$ | $\sigma_1 = 0.6$ | $\sigma_2 = 1$ | $\rho = 0.7$ | MSE($\hat{f}_1$) | MSE($\hat{f}_2$) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | −2.498 | −3.569 | 1.026 | 0.641 | 0.562 | 1.054 | 0.693 | | |
| 200 | 10 | (0.113) | (0.233) | (0.184) | (0.308) | (0.141) | (0.209) | (0.251) | 0.0260 | 0.0282 |
| | | [95.0%] | [91.0%] | [91.5%] | [93.5%] | [97.0%] | [92.5%] | [94.5%] | | |
| | | −2.505 | −3.529 | 1.011 | 0.674 | 0.572 | 1.009 | 0.699 | | |
| 200 | 20 | (0.082) | (0.149) | (0.134) | (0.223) | (0.086) | (0.129) | (0.158) | 0.0179 | 0.0160 |
| | | [93.0%] | [88.0%] | [94.0%] | [94.5%] | [91.0%] | [90.5%] | [93.0%] | | |
| | | −2.509 | −3.584 | 1.001 | 0.663 | 0.587 | 1.076 | 0.649 | | |
| 400 | 10 | (0.080) | (0.156) | (0.132) | (0.211) | (0.100) | (0.142) | (0.180) | 0.0182 | 0.0168 |
| | | [91.0%] | [87.0%] | [95.0%] | [93.5%] | [91.0%] | [86.5%] | [93.0%] | | |
| | | −2.499 | −3.515 | 0.999 | 0.701 | 0.584 | 0.991 | 0.681 | | |
| 400 | 20 | (0.058) | (0.102) | (0.095) | (0.155) | (0.061) | (0.091) | (0.116) | 0.0116 | 0.0083 |
| | | [94.0%] | [90.0%] | [94.0%] | [94.0%] | [93.5%] | [94.5%] | [91.5%] | | |

Table 2.2: Comparison of simulation results between the multivariate model and the univariate models.

| | Multivariate Model | | | Univariate Models | | |
|---|---|---|---|---|---|---|
| Parameter | Estimate | SE | CP(%) | Estimate | SE | CP(%) |
| $\beta_0^1 = -2.5$ | $-2.517$ | 0.115 | 92.0 | $-2.522$ | 0.117 | 90.0 |
| $\beta_0^2 = -3.5$ | $-3.576$ | 0.226 | 88.5 | $-3.601$ | 0.242 | 88.0 |
| $\beta_0^3 = -3.5$ | $-3.595$ | 0.273 | 91.0 | $-3.625$ | 0.300 | 90.5 |
| $\beta_1^1 = 1$ | 0.996 | 0.184 | 93.5 | 0.993 | 0.187 | 94.0 |
| $\beta_1^2 = 0.7$ | 0.636 | 0.315 | 90.5 | 0.624 | 0.317 | 90.5 |
| $\beta_1^3 = 1$ | 0.954 | 0.240 | 93.5 | 0.947 | 0.242 | 94.0 |
| $\sigma_1 = 0.6$ | 0.586 | 0.128 | 96.0 | 0.598 | 0.159 | 95.0 |
| $\sigma_2 = 1$ | 1.036 | 0.204 | 95.5 | 1.078 | 0.236 | 94.5 |
| $\sigma_3 = 1.5$ | 1.558 | 0.241 | 93.0 | 1.592 | 0.267 | 92.0 |
| $\rho_{12} = 0.8$ | 0.736 | 0.205 | 96.0 | – | – | – |
| $\rho_{13} = 0.6$ | 0.583 | 0.197 | 93.5 | – | – | – |
| $\rho_{23} = 0.5$ | 0.453 | 0.194 | 95.0 | – | – | – |

Table 2.3: Model fitting results for the YWP data.

| Parameter | Estimate | Std. Error | 95% CI |
|---|---|---|---|
| $\beta_0^{\text{ct}}$ | $-2.70$ | 0.10 | $(-2.93, -2.55)$ |
| $\beta_0^{\text{gc}}$ | $-3.86$ | 0.19 | $(-4.37, -3.63)$ |
| $\beta_0^{\text{tv}}$ | $-3.69$ | 0.17 | $(-4.12, -3.45)$ |
| $\sigma_1$ | 0.69 | 0.10 | $(0.47, 0.89)$ |
| $\sigma_2$ | 1.03 | 0.16 | $(0.70, 1.37)$ |
| $\sigma_3$ | 1.22 | 0.14 | $(0.98, 1.52)$ |
| $\rho_{12}$ | 0.68 | 0.15 | $(0.44, 1.00)$ |
| $\rho_{13}$ | 0.38 | 0.14 | $(0.16, 0.69)$ |
| $\rho_{23}$ | 0.51 | 0.16 | $(0.23, 0.84)$ |

Table 2.4: Sensitivity and specificity of the proposed model under different cutoff points of infection probability for CT, GC and TV, and percentages of follow-up visits that meet the cutoff points.

| Organism | Cutoff Points | Sensitivity | Specificity | Percentage of visits (%) |
|---|---|---|---|---|
| CT | 0.039 | 0.99 | 0.20 | 82 |
| | 0.056 | 0.94 | 0.40 | 63 |
| | 0.075 | 0.84 | 0.60 | 44 |
| | 0.090 | 0.76 | 0.76 | 35 |
| | 0.116 | 0.62 | 0.80 | 24 |
| GC | 0.011 | 1 | 0.20 | 81 |
| | 0.017 | 0.99 | 0.39 | 62 |
| | 0.026 | 0.94 | 0.59 | 43 |
| | 0.033 | 0.88 | 0.68 | 34 |
| | 0.055 | 0.72 | 0.84 | 19 |
| TV | 0.013 | 1 | 0.20 | 81 |
| | 0.019 | 0.99 | 0.40 | 62 |
| | 0.031 | 0.97 | 0.60 | 44 |
| | 0.047 | 0.89 | 0.72 | 31 |
| | 0.065 | 0.78 | 0.80 | 23 |

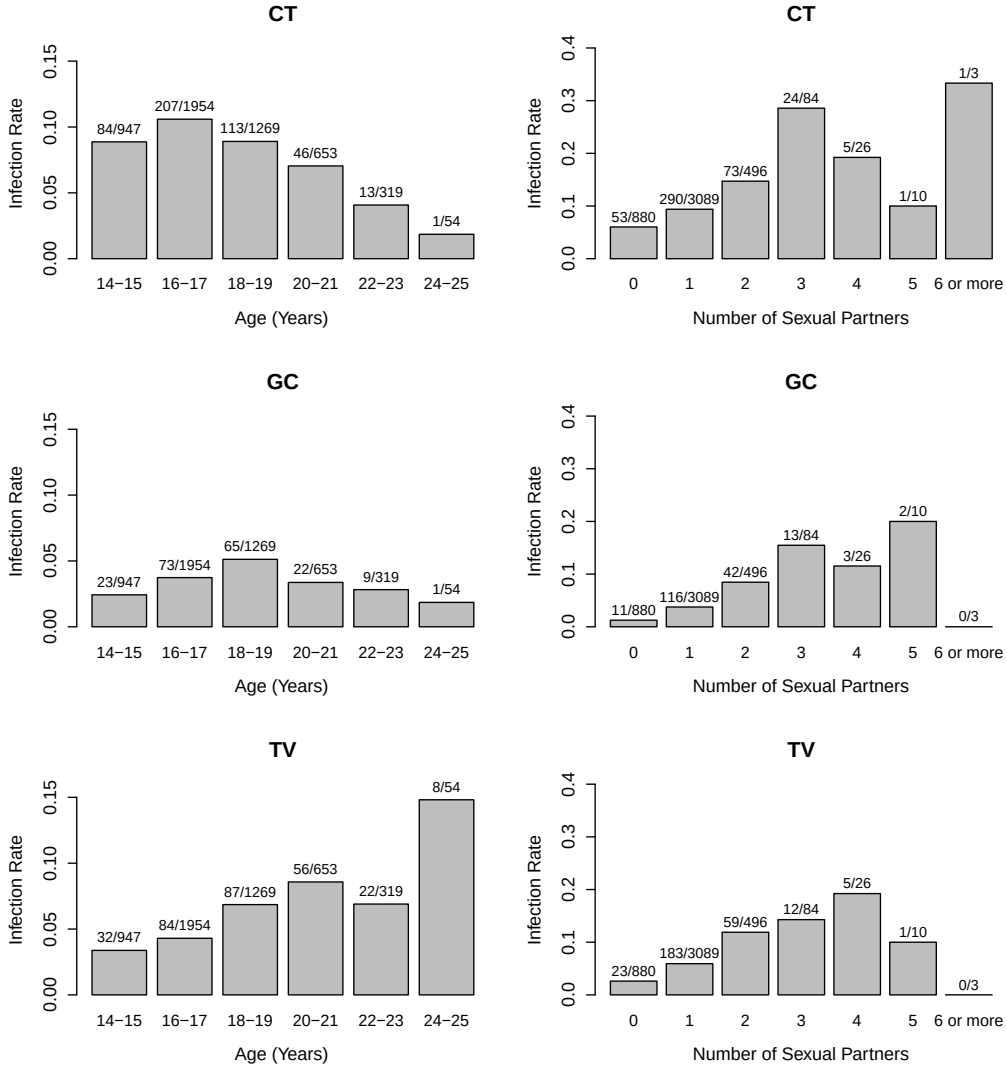


Figure 2.1: CT, GC and TV infection rates by age and the number of sexual partners.

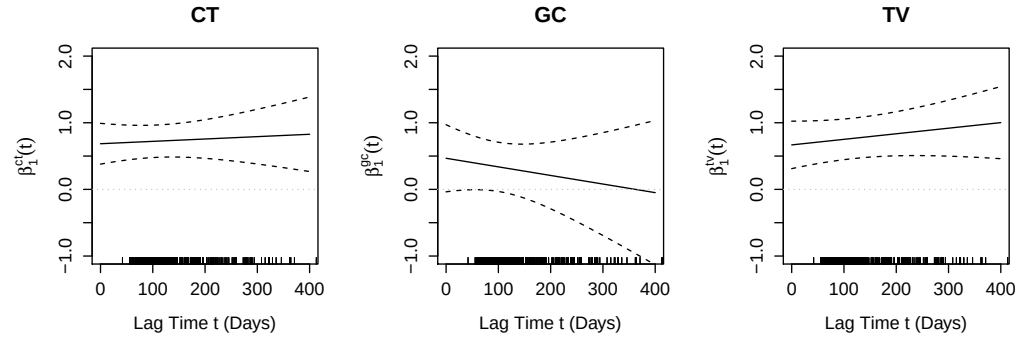


Figure 2.2: Lag time effects of a prior infection on current CT, GC and TV infections.

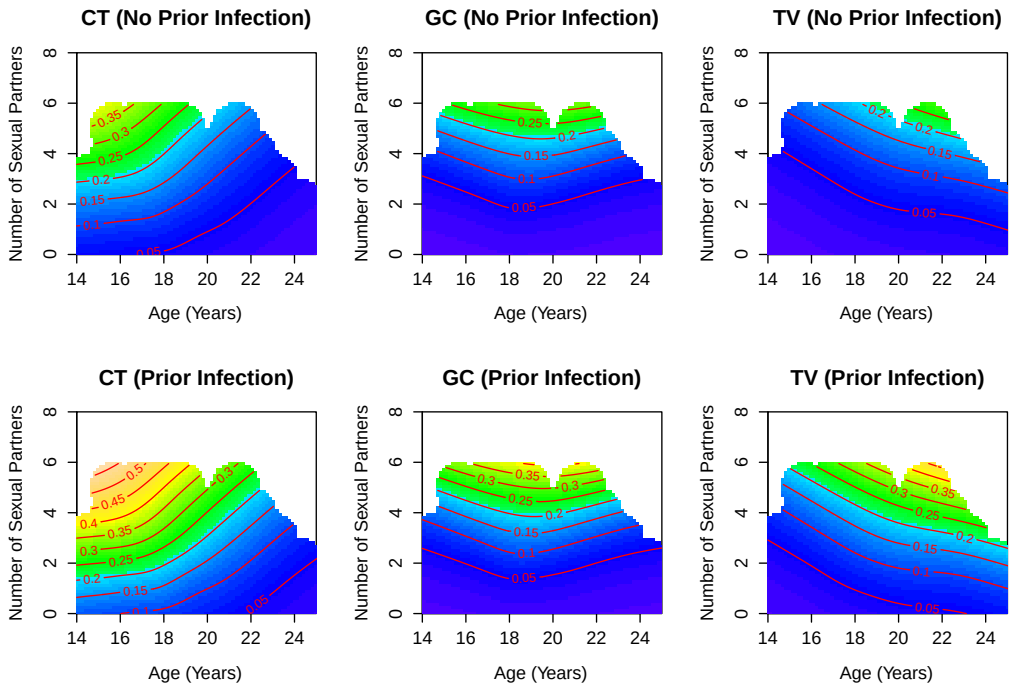


Figure 2.3: Bivariate surfaces showing the joint effects of age and the number of sexual partners on CT, GC and TV infections with or without a prior infection.

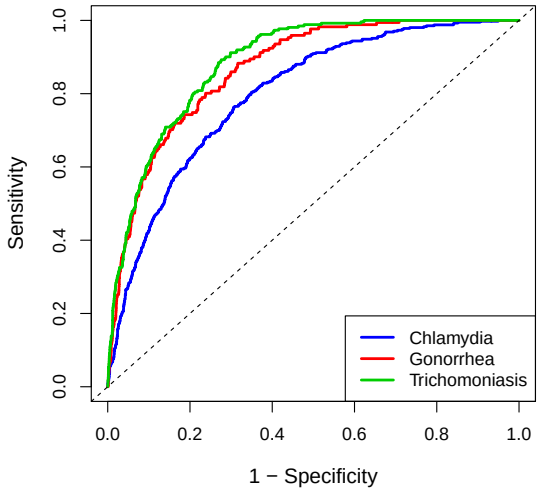


Figure 2.4: ROC curves for CT, GC and TV infections.

# Chapter 3

# A Generalized Semiparametric Mixed Model for Exponential Family of Distributions

This chapter presents a general multivariate semiparametric modeling framework by extending the model for binary data proposed in Chapter 2. The generalized model can accommodate different types of data following the exponential family of distributions. Details of the model fitting procedure are provided in this chapter. A relevant inference procedure is developed and illustrated by revisiting the YWP data described in the last chapter. Simulation studies and an analysis of real clinical data are conducted to demonstrate the generalized model.

## 3.1 Research Background

Multivariate longitudinal data are common in clinical investigations where multiple outcomes are measured repeatedly over time on each subject. Methods for univariate longitudinal data analysis have been well developed to take into account various data features, including the temporal correlations among the repeated measurements from the same subject (Laird and Ware, 1982; Liang and Zeger, 1986), and potential nonlinear independent variable effects (Lin and Carroll, 2001; Zhang et al., 1992). For studies with multiple outcomes, analytical options are generally more limited and analysts sometimes resort to univariate techniques that models the outcomes one at a time in separate models, at the expense of estimation bias and inefficiency.

This said, several approaches have been proposed for analysis of repeatedly measured multiple outcome data. For example, Rochon (1996) used the generalized estimating equa-

tions (GEE) to analyze bivariate repeated outcomes. Along this line, similar models have been developed for the analysis of multiple continuous and binary outcomes (Gray and Brookmeyer, 1998; O'Brien and Fitzmaurice, 2004). With the GEE approach, there is no need to explicitly specify the covariance structure of the data, and thus the approach is more useful when one is interested in characterizing the population-averaged covariate effects on the outcomes, as opposed to evaluating the underlying associations across the outcomes. An alternative approach is the latent variable models which assume that the observed outcomes are surrogate measures of a non-observable endpoint of real interest (Sammel and Ryan, 1996; Sammel et al., 1997). Such a modeling approach has been applied to a variety of situations including bivariate clustered outcomes, multiple continuous longitudinal outcomes and a mixture of longitudinal outcomes (Catalano and Ryan, 1992; Miglioretti, 2003; Roy and Lin, 2000). Another approach accounts for the covariance structure of the outcomes through multivariate mixed effects models (Reinsel, 1982; Shah et al., 1997). Correlated outcomes are naturally linked together by a prespecified joint distribution of the random effects. A pairwise model fitting approach was then developed to resolve the computational problems due to high-dimensionality of the joint covariance structure (Fieuws and Verbeke, 2006). Most of the existing literature on multivariate mixed models has thus far focused on continuous data.

Another underdeveloped modeling feature is the accommodation of nonlinear independent variable effects in the multivariate setting. Without knowing the true functional form of the independent variable effect, one convenient way of incorporating a potential nonlinear effect is to use semi-parametric regression models. Existing work on semiparametric regression models in the multivariate setting, however, is rather limited. Most of the published methods have focused on the depiction of nonlinear time effect (Coull and Staudenmayer, 2004; Ghosh and Hanson, 2010; Ghosh and Tu, 2009). Liu and Tu (2012) considered bi-

variate smooth components in a semiparametric model for a pair of continuous outcomes, and estimated the smooth functions using penalized regression splines.

In this chapter, I extend the bivariate semiparametric model proposed by Liu and Tu (2012) to a general exponential family setting. Specifically, a generalized semiparametric mixed model is constructed for multivariate longitudinal data. The proposed model provides a unified framework for common types of data that follow the exponential family of distributions. The correlation structure of the outcomes are specified through the random effects. Bivariate smooth functions are incorporated to accommodate nonlinear influences of two potentially interacting independent variables. Model parameters are estimated by using the maximum penalized likelihood method. Simulation studies are conducted to evaluate the performance of the estimation method. Finally in this chapter, the proposed method is illustrated by analyzing data collected from a real clinical investigation.

## 3.2 Methods

### 3.2.1 Generalized Multivariate Semiparametric Model

Suppose that there are $K$ outcomes of interest. Let $Y_{ijk}$ be the $i$th subject's response on the $k$th outcome at the $j$th time point, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, and $k = 1, \ldots, K$. A $q \times 1$ vector of random effects $\mathbf{b}_{ik}$ is introduced to accommodate the correlations among the repeated measurements of outcome $k$ within subject $i$, and $\boldsymbol{b}_{ik}$ is assumed to be normally distributed, i.e., $\boldsymbol{b}_{ik} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_k)$ where $\boldsymbol{\Sigma}_k$ is a $q \times q$ variance-covariance matrix. It is further assumed that the conditional distribution of $Y_{ijk}$ given $\boldsymbol{b}_{ik}$ belongs to an exponential family of canonical form, i.e., for $k = 1, \ldots, K$,

$$f_k(y_{ijk}|\boldsymbol{b}_{ik}) = \exp\left\{\frac{y_{ijk}\eta_{ijk} - d(\eta_{ijk})}{a(\phi_k)} + c(y_{ijk}, \phi_k)\right\}, \tag{3.1}$$

where the natural parameter $\eta_{ijk} = g(\mu_{ijk})$, with $\mu_{ijk}$ being the conditional mean of $Y_{ijk}$ given $\boldsymbol{b}_{ik}$ and $g(\cdot)$ being a monotone, invertible link function. For example, $\eta_{ijk} = \mu_{ijk}$ for normally distributed data; binary data use a logit link function defined as $\eta_{ijk} = \log(\frac{\mu_{ijk}}{1-\mu_{ijk}})$; for count data following a Poisson distribution, a log link function is used, i.e., $\eta_{ijk} = \log(\mu_{ijk})$.

Consider the following semiparametric mixed model,

$$\eta_{ijk} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k + \boldsymbol{z}_{ij}^T \boldsymbol{b}_{ik} + s_k(t_{1ij}, t_{2ij}), \tag{3.2}$$

where $\boldsymbol{\beta}_k$ is a $p \times 1$ vector of coefficients for fixed effect covariates $\boldsymbol{x}_{ij}$, $\boldsymbol{z}_{ij}$ is a vector of random effect covariates which is usually a subset of $\boldsymbol{x}_{ij}$, and $s_k$ is a bivariate smooth function for independent variables $t_{1ij}$ and $t_{2ij}$ associated with outcome $k$, which is incorporated to capture the joint nonlinear effects of two independent variables on each outcome. To account for the within-subject correlations across the outcomes, we define a vector of subject-specific random effects as $\boldsymbol{b}_i = (\boldsymbol{b}_{i1}^T, \ldots, \boldsymbol{b}_{iK}^T)^T$ following a multivariate normal distribution $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_b)$, where the variance-covariance matrix $\boldsymbol{\Sigma}_b$ consists of diagonal blocks $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$ and off-diagonal elements which accommodate the between-outcome correlations.

For the smooth functions, a set of basis functions $h_l$, $l = 1, \ldots, L$ is specified so that $s_k(t_{1ij}, t_{2ij}) = \sum_{l=1}^{L} \alpha_{kl} h_l(t_{1ij}, t_{2ij})$ with $\alpha_{kl}$ being the corresponding coefficients. Then it can be written compactly as $s_k(t_{1ij}, t_{2ij}) = \boldsymbol{T}_{ij}^T \boldsymbol{\alpha}_k$, where $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{kL})^T$ and $\boldsymbol{T}_{ij} = [h_1(t_{1ij}, t_{2ij}), \ldots, h_L(t_{1ij}, t_{2ij})]^T$. After combining $s_k(t_{1ij}, t_{2ij})$, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$ into a vector $\boldsymbol{s}_k$, it follows that

$$\boldsymbol{s}_k = \mathbf{T}_k \boldsymbol{\alpha}_k, \tag{3.3}$$

where $\boldsymbol{\alpha}_k$ is the coefficient vector and $\mathbf{T}_k$ is the basis function matrix consisting of row vectors $\boldsymbol{T}_{ij}^T$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$.

For convenience, model (3.2) can be written into a matrix form. Denote the outcome vector by $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \cdots, \boldsymbol{Y}_K^T)^T$ where $\boldsymbol{Y}_k^T = (Y_{11k}, \ldots, Y_{1n_ik}, \ldots, Y_{m1k}, \ldots, Y_{mn_ik})^T$ for $k = 1, \ldots, K$. The natural parameter vector $\boldsymbol{\eta}$ is defined in a similar way. Then model (3.2) is rewritten as

$$\boldsymbol{\eta} = \mathbf{X}_\beta \tilde{\boldsymbol{\beta}} + \mathbf{Z}_b \tilde{\boldsymbol{b}} + \mathbf{T}\boldsymbol{\alpha}, \tag{3.4}$$

where $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T$ is the vector of fixed effect coefficients for the design matrix $\mathbf{X}_\beta$, $\tilde{\boldsymbol{b}} = (\boldsymbol{b}_{11}^T, \ldots, \boldsymbol{b}_{m1}^T, \ldots, \boldsymbol{b}_{1K}^T, \ldots, \boldsymbol{b}_{mK}^T)^T$ is the vector of random effects for the design matrix $\mathbf{Z}_b$ and it follows a multivariate normal distribution $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_b \otimes \boldsymbol{I}_m)$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_K^T)^T$ is the coefficient vector for the basis function matrix $\mathbf{T}$.

### 3.2.2 Penalized Likelihood

In this chapter, thin plate regression splines are used to model the bivariate smooth function $s_k$ in model (3.2). The estimation of the thin plate regression splines can be unified into a mixed model framework (Ruppert et al., 2003; Wood, 2006). Let $\boldsymbol{\psi} = (\tilde{\boldsymbol{\beta}}^T, \boldsymbol{\xi}^T, \boldsymbol{\alpha}^T)^T$ be a vector of all unknown parameters, where $\boldsymbol{\xi}$ denotes the variance components in $\boldsymbol{\Sigma}_b$. The proposed model can be estimated by maximizing the following penalized log-likelihood function

$$p\ell(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}) - \sum_{k=1}^{K} \lambda_k J(s_k), \tag{3.5}$$

where $\ell(\boldsymbol{\psi})$ is the log-likelihood function of the model, $J(s_k)$ is the penalty function measuring the roughness of $s_k$, and $\lambda_k$ is the corresponding smoothing parameter which balances the goodness-of-fit of the model and the smoothness of $s_k$. A commonly used form of roughness penalty for bivariate smoothers is $J(s) = \iint_{\mathbb{R}^2} \{(\frac{\partial^2 s}{\partial t_1^2})^2 + 2(\frac{\partial^2 s}{\partial t_1 \partial t_2})^2 + (\frac{\partial^2 s}{\partial t_2^2})^2\} \mathrm{d}t_1 \mathrm{d}t_2$. Based on the observed data, it can be further written as a quadratic form in the coefficients of the smooth function, i.e., $J(s_k) = \boldsymbol{\alpha}_k^T \mathbf{S}_k \boldsymbol{\alpha}_k$ where the penalty matrix $\mathbf{S}_k$ is a positive

semi-definite matrix of known coefficients.

The quadratically penalized smooth function $s_k$ can be partitioned into conventional fixed effects and random effects in the generalized linear mixed model (GLMM) by using eigen decomposition of the penalty matrix $\mathbf{S}_k$. Through reparameterization, the coefficient vector $\boldsymbol{\alpha}_k$ is divided into the fixed effect coefficients $\boldsymbol{\alpha}_{k,F}$ and the random effects $\boldsymbol{\alpha}_{k,R}$, such that $\boldsymbol{\alpha}_k^T \mathbf{S}_k \boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k,R}^T \mathbf{S}_{k,R} \boldsymbol{\alpha}_{k,R}$ where $\mathbf{S}_{k,R}$ is a diagonal matrix with all positive eigenvalues of $\mathbf{S}_k$ on the diagonal. Therefore the fixed effect coefficients $\boldsymbol{\alpha}_{k,F}$ are unpenalized. Equation (3.3) now has the following mixed model representation

$$\boldsymbol{s}_k = \mathbf{T}_{k,F} \boldsymbol{\alpha}_{k,F} + \mathbf{T}_{k,R} \boldsymbol{\alpha}_{k,R}, \tag{3.6}$$

where $\mathbf{T}_{k,F}$ and $\mathbf{T}_{k,R}$ are the design matrices for fixed effects and random effects respectively, and $\boldsymbol{\alpha}_{k,R} \sim \mathrm{N}(\mathbf{0}, \mathbf{S}_{k,R}^{-1}/\lambda_k)$. Applying equation (3.6) to model (3.4) gives the following GLMM representation,

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b}, \tag{3.7}$$

where $\mathbf{X} = (\mathbf{X}_\beta, \mathrm{diag}(\mathbf{T}_{1,F}, \ldots, \mathbf{T}_{K,F}))$ and $\boldsymbol{\beta} = (\tilde{\boldsymbol{\beta}}^T, \boldsymbol{\alpha}_{1,F}^T, \ldots, \boldsymbol{\alpha}_{K,F}^T)^T$ are the design matrix and coefficient vector of fixed effects, $\mathbf{Z} = (\mathbf{Z}_b, \mathrm{diag}(\mathbf{T}_{1,R}, \ldots, \mathbf{T}_{K,R}))$ is the random effect design matrix, $\boldsymbol{b} = (\tilde{\boldsymbol{b}}^T, \boldsymbol{\alpha}_{1,R}^T, \ldots, \boldsymbol{\alpha}_{K,R}^T)^T$ is the vector of random effects, and $\boldsymbol{b} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ where $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathrm{diag}(\boldsymbol{\Sigma}_b \otimes \boldsymbol{I}_m, \mathbf{S}_{1,R}^{-1}/\lambda_1, \ldots, \mathbf{S}_{K,R}^{-1}/\lambda_K)$ is the variance-covariance matrix and $\boldsymbol{\theta} = (\boldsymbol{\xi}^T, \lambda_1, \ldots, \lambda_K)^T$ denotes the variance components.

### 3.2.3 Estimation Algorithm

Since the proposed model can be formulated into a GLMM representation, the parameters can be estimated conveniently using existing approaches for mixed models. In particular, the smoothing parameters can be estimated simultaneously with other variance components,

which makes the estimation procedure computationally efficient.

Based on equation (3.1), the likelihood function of model (3.7) given the observed data $\boldsymbol{y}$ is

$$L(\boldsymbol{\beta},\boldsymbol{\theta}) \propto |\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{-1/2} \int \exp\left\{ \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \left[ \frac{y_{ijk}\eta_{ijk} - d(\eta_{ijk})}{a(\phi_k)} + c(y_{ijk},\phi_k) \right] - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{b} \right\} \mathrm{d}\boldsymbol{b}, \tag{3.8}$$

The integral in the likelihood function is usually intractable except when the outcomes are continuous and follow normal distributions. As discussed in Section 2.3.3, the Laplace approximation method is used to evaluate this integral. Note that the integrand in equation (3.8) is the unnormalized conditional density of the random effects $\boldsymbol{b}$ given $\boldsymbol{Y} = \boldsymbol{y}$. For given $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the conditional mode of $\boldsymbol{b}$ is

$$\hat{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{\theta}) = \arg\max_{\boldsymbol{b}} \left\{ \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \left[ \frac{y_{ijk}\eta_{ijk} - d(\eta_{ijk})}{a(\phi_k)} + c(y_{ijk},\phi_k) \right] - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{b} \right\}.$$

which can be determined by using a penalized iteratively reweighted least squares (PIRLS) algorithm (Bates, 2010). By replacing the logarithm of the integrand with its second-order Taylor expansion at the conditional mode $\hat{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{\theta})$, the Laplace approximation to the likelihood $L(\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{y})$ can be optimized to obtain the approximate maximum likelihood estimators (MLEs) for parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ (Breslow and Clayton, 1993).

The estimation algorithm is developed based on R (R Development Core Team, 2011) packages for fitting traditional mixed models (e.g., `gamm4` (Wood, 2011)). Again, the standard errors of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ can be obtained by using the bootstrap procedure described in Section 3.3. The 95% confidence intervals (CIs) of the parameter estimates and the coverage probabilities of the CIs can also be calculated based on the bootstrap samples.

### 3.3 Statistical Inference

One of the advantages of the multivariate modeling approach is that it allows the comparison of independent variable effects across the outcomes. For example, one may be interested in testing whether the bivariate nonlinear effects vary for different outcomes. The question can be formulated into the following hypothesis about the functional forms of $s_k$,

$$H_0 : s_1 = \cdots = s_K \quad \text{vs.} \quad H_1 : \text{otherwise.} \tag{3.9}$$

Zhang and Lin (2003) considered testing the equivalence of two nonparametric univariate functions in semiparametric additive mixed models for two groups. They constructed a test statistic based on the integrated squared difference of two functions and approximated the distribution of the test statistic by a scaled chi-square distribution. However, it is difficult to apply the test they developed to compare bivariate smooth functions. Herein, a likelihood ratio test (LRT) is proposed based on the test statistic $\Delta = -2[\ell(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\theta}}_0) - \ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})]$, where $\ell(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\theta}}_0)$ is the maximized value of the log-likelihood for the null model under $H_0$, and $\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ is the maximized log-likelihood for the unrestricted model under $H_1$. Theoretically, it is very challenging to derive the asymptotic distribution of the test statistic $\Delta$ under the null hypothesis. The asymptotic properties of LRT based on the large sample chi-squared mixture approximations are not satisfactory when applied to penalized splines models (Crainiceanu and Ruppert, 2004). Therefore, resampling techniques are employed to approximate the sampling distribution of $\Delta$. Härdle et al. (2004) have shown that bootstrap can be applied to componentwise hypothesis testing in semiparametric generalized additive models. Roca-Pardiñas et al. (2008) also used a bootstrap method to test factor-by-surface interactions in a logistic generalized additive model. Liu and Tu (2012) extended the bootstrap test for comparing the bivariate surfaces among different groups of subjects

to a longitudinal data setting with paired outcomes. Here a similar strategy is used to compare the bivariate effects across the outcomes.

To test the hypothesis in (3.9), a resampling procedure is proposed which combines bootstrap and permutation techniques:

1. Fit model (3.2) under the null hypothesis to obtain the effective degrees of freedom (EDF) for the penalized splines estimates.
2. Draw a bootstrap sample with replacement from the observed data. The sampling unit is the subject, that is, either none or all of the observations from a subject will be selected. If a subject is selected more than once, he/she will be treated as a different person each time by being assigned a new ID in the bootstrap sample.
3. Permute the labels indicating the 1st, 2nd, ... and $K$th outcomes within each subjects in the bootstrap sample, preserving the order of the repeated measurements for each outcome.
4. For the bootstrap data with permuted labels, refit the null and unrestricted models using regression splines with the degrees of freedom (DF) fixed at the EDF estimated in step 1, and calculate the likelihood ratio test statistic $\Delta^*$.
5. Repeat steps $2-4$ for $B$ times to generate a sample of test statistic $\{\Delta_b^*\}_{1\leq b\leq B}$, representing an empirical distribution of $\Delta$ under the null hypothesis. The p-value of the test can be calculated as $p = \#_{b=1}^{B}\{\Delta_b^* \geq \Delta\}/B$.

In the absence of asymptotic results, the resampling procedure provides a valid alternative to the traditional inference based on large sample theories. The performance of the procedure is assessed in a simulation study described in Section 3.4.

## 3.4 Simulation Studies

### 3.4.1 Evaluation of Estimation Procedure

A simulation study was conducted to evaluate the performance of the estimation procedure for the proposed model. Bivariate outcomes $(Y_{ij1}, Y_{ij2})$, $i = 1, \ldots, m$, $j = 1, \ldots, n$ were generated from Poisson distributions with means $\mu_{ij1}$ and $\mu_{ij2}$ respectively. Two settings were considered with different degrees of between-outcome correlation in the random effects.

In Setting 1, a strong between-outcome correlation was assumed with correlation coefficient $\rho = 0.7$. The data were simulated using the following semiparametric mixed model

$$\begin{cases} \log(\mu_{ij1}) = \beta_{01} + \beta_{11} x_{ij} + b_{i1} + \bar{s}_1(t_{1ij}, t_{2ij}) \\ \log(\mu_{ij2}) = \beta_{02} + \beta_{12} x_{ij} + b_{i2} + \bar{s}_2(t_{1ij}, t_{2ij}), \end{cases} \tag{3.10}$$

where $(b_{i1}, b_{i2})^T \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_b)$ with

$$\mathbf{\Sigma}_b = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

the smooth functions were defined as $s_1(t_1, t_2) = e^{t_1} \sin(\pi t_2)$ and $s_2(t_1, t_2) = 2\sqrt{t_1} e^{(t_2 - 0.4)^2}$, and $\bar{s}_1$ and $\bar{s}_2$ were centered over the observed values of the covariates $(t_{ij1}, t_{ij2})$ which were generated from $\mathrm{N}(0, 0.25)$ independently. Other parameters were chosen as $\beta_{01} = 1.5$, $\beta_{02} = 1$, $\beta_{11} = 0.3$, $\beta_{12} = -0.2$, $\sigma_1 = 0.7$ and $\sigma_2 = 0.7$.

In Setting 2, the correlation was assumed to be moderate with $\rho = 0.4$. Model (3.10) was used again to generate the data. The smooth functions remained the same. Other parameters were $\beta_{01} = -0.8$, $\beta_{02} = -1.2$, $\beta_{11} = 2$, $\beta_{12} = 1$, $\sigma_1 = 0.5$ and $\sigma_2 = 1$.

Different sample sizes were examined under each setting with $m = 200, 500$ and $n =$

5, 10. The parameter estimates were averaged over 200 replications. For each simulated data set, 200 bootstrap samples were drawn to calculate the standard errors and coverage probabilities of the 95% bootstrap CIs. The MSEs of the smooth function estimates were also calculated for each setting.

The simulation results under the two settings are presented in Table 3.1 and Table 3.2. In both settings, the estimation procedure achieved excellent performance for fitting the proposed model. All parameter estimates had very small bias. The coverage probabilities of the 95% CIs were close to the nominal level. The MSEs of the smooth functions gradually decreased while the sample size increased.

### 3.4.2 Evaluation of Inference Procedure

Another simulation study was conducted to assess the performance of the proposed likelihood-based resampling procedure. Binary data were generated using the same setting as in the three-outcome simulation study described in Section 2.5.1, except that the bivariate functions were assumed to have the same functional form for all outcomes. The size of the test was assessed based on 200 simulation runs, each including 200 bootstrap samples. Under a sample size of 200 subjects with 10 repeated measurements on each outcome per subject, the resampling test achieved a size of 0.04, which was close to the nominal level 0.05.

## 3.5 Real Data Applications

### 3.5.1 Revisit of YWP Data

In this section, the YWP data introduced in the last chapter was revisited in order to illustrate the proposed likelihood-based resampling test. In the YWP example, an important question that one may be interested in is whether the concurrent influences of age and the number of partners differ for the three organisms. In model (2.1), the joint effect of the two

risk factors on the infection risk of the $k$th organism is represented by bivariate function $f^k$, and thus the question of interest can be answered by testing the hypothesis (3.9). After repeating the proposed resampling procedure 200 times, the test statistic was obtained as $\Delta = 59.6$ with a p-value $< 0.001$, indicating a highly significant difference in the joint effects of age and the number of partners across the three organisms. The significant test result was supported by the fact that the estimated bivariate effect surfaces have very different shapes as shown in Figure 2.3.

### 3.5.2 Analysis of Health Care Utilization Data

To illustrate the aforementioned model generalization, I used data from a clinical trial of a care management intervention, namely Geriatric Resources for Assessment and Care of Elders (GRACE), for low-income elderly patients. The detailed study protocol has been described by (Counsell et al., 2007). Briefly, patients were recruited from community-based health centers and were assigned to either the GRACE intervention or usual care group based on the randomization of their primary care physicians. Patients in the intervention group received home-based care management that was individualized based on their geriatric conditions, while the usual care group had access to all primary and specialty care services as usual. Multiple outcomes were assessed at baseline and then semiannually for 2 years, including health-related quality of life, activities of daily living, emergency department (ED) visits (not resulting in hospital admission) and hospital admissions in the last 6 months. In this analysis, ED visit and hospital admission counts were considered as bivariate outcomes because they both characterize acute health care utilization. The primary objective was to examine whether and how patients' physical and mental health affect their acute care utilization. Only the control group was used in the analysis as it was representative of the general population who received usual care.

Among the 477 patients in the control group, 365 (77%) were females and 292 (61%) were blacks. They contributed a total of 2037 observations. At baseline, the median age of this group was 70 years (range: $65 - 97$ years; standard deviation: 6 years). A few comorbid conditions were assessed, including hypertension, angina, congestive heart failure, heart attack, stroke, chronic lung disease, arthritis, diabetes and cancer. Every 6 months, the patients were interviewed on their quality of life and health status, and their ED visit and hospital admission records were obtained from a regional health information exchange. Specifically, quality of life was evaluated using the Medical Outcomes 36-Item Short-Form (SF-36; Brazier et al., 1992), and it was aggregated into a single measure by averaging the Physical Component Score (PCS) and Mental Component Score (MCS), with higher score indicating better health (score range: $0 - 100$); depression status was measured using the Patient Health Questionnaire-9 (PHQ-9; Martin et al., 2006), with higher score indicating more severe depression (score range: $0 - 27$).

To examine the concurrent effects of health-related qualify of life (SF-36) and depression severity (PHQ-9) on ED visit and hospital admission rates, I considered the following model

$$\begin{cases} \log(\mu_{ij}^{\mathrm{ED}}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}^{\mathrm{ED}} + b_i^{\mathrm{ED}} + s^{\mathrm{ED}}(t_{1ij}, t_{2ij}) \\ \log(\mu_{ij2}^{\mathrm{HA}}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}^{\mathrm{HA}} + b_i^{\mathrm{HA}} + s^{\mathrm{HA}}(t_{1ij}, t_{2ij}), \end{cases} \tag{3.11}$$

where $\mu_{ij}^{\mathrm{ED}}$ and $\mu_{ij}^{\mathrm{HA}}$ are the mean numbers of ED visits and hospital admissions in the last 6 months; $\boldsymbol{\beta}^{\mathrm{ED}} = (\boldsymbol{\beta}_0^{\mathrm{ED}}, \boldsymbol{\beta}_1^{\mathrm{ED}}, \boldsymbol{\beta}_2^{\mathrm{ED}}, \boldsymbol{\beta}_3^{\mathrm{ED}}, \boldsymbol{\beta}_4^{\mathrm{ED}})^T$ and $\boldsymbol{\beta}^{\mathrm{HA}} = (\boldsymbol{\beta}_0^{\mathrm{HA}}, \boldsymbol{\beta}_1^{\mathrm{HA}}, \boldsymbol{\beta}_2^{\mathrm{HA}}, \boldsymbol{\beta}_3^{\mathrm{HA}}, \boldsymbol{\beta}_4^{\mathrm{HA}})^T$ are the outcome-specific regression coefficients for the following covariates: intercept, gender (female vs male), race (black vs others), age and the number of comorbidities; $(b_i^{\mathrm{ED}}, b_i^{\mathrm{HA}})^T \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_b)$ with $\boldsymbol{\Sigma}_b$ defined as in model (3.10); $s^{\mathrm{ED}}$ and $s^{\mathrm{HA}}$ are the bivariate smooth functions of SF-36 and PHQ-9 to capture the joint effects of overall health and depression status on ED visits and hospital admissions.

Using the proposed model fitting procedure, it took about 30 minutes to fit model (3.11) on a PC with dual 2.13GHz CPUs and 2 GB memory. The parameter estimates as well as the bootstrap SEs and 95% CIs are provided in Table 3.3. Females tended to have fewer ED visits and hospital admissions although the associations were not statistically significant. Older patients had significantly higher hospital admission rates. A ten-year increase in age would result in a 35% increase in hospital admission rate. The number of comorbidities was a significant predictor for both outcomes. An additional comorbidity was associated with 14% and 23% increase in ED visit and hospital admission rates, respectively. We also noted a strong within-subject correlation between ED visit and hospital admission rates ($\rho = 0.86$, 95% CI $= (0.77, 0.98)$) and a greater variability in hospital admission rate ($\sigma_2 = 1.71$, 95% CI $= (0.91, 1.80)$) than in ED visit rate ($\sigma_1 = 1.04$, 95% CI $= (0.79, 1.34)$).

The fitted bivariate surfaces showing the concurrent influences of SF-36 and PHQ-9 scores are presented using colored contour plots in Figure 3.1. Note that warmer color represents higher rates, but the color scales are different for ED visit and hospital admission. As expected, lower SF-36 scores and higher PHQ-9 scores, indicating worse physical and mental health, were associated with increased utilization of both ED and inpatient care. Nonetheless, the two scores interacted very differently as the joint effect surfaces for ED visit and hospitalization have distinct patterns. Depression status (measured by PHQ-9) dominated ED visit rate when patients were in poor general health (e.g., SF-36 score $<$ 40), while general health status had the dominating effect when patients were healthier overall (e.g., SF-36 score $>$ 45). On the other hand, the effects of overall health and depression status on hospital admission appeared more linear with little interaction. This is not surprising because when people have poor health conditions, they are more likely to feel depressed which may result in more frequent visit to ED, but whether or not they are hospitalized mostly depends on other illnesses instead of depression alone.

## 3.6 Discussion

In this chapter, a generalized semiparametric modeling framework has been proposed for multivariate longitudinal data in which multiple outcomes of interest are measured repeatedly over time. The model can be used for data following an exponential family of distributions. A mixed model approach is adopted for explicit specification of the correlation structure of the outcomes within each subject. The concurrent nonlinear influences and potential interaction effects of two independent variables are incorporated through bivariate nonparametric functions. Using thin plate regression splines as the smoother, the proposed model can be formulated into a generalized linear mixed model so that traditional mixed model packages can be utilized for parameter estimation. With this modeling framework, one can explore the covariate effects in a flexible way as well as examine the within-subject correlations among the outcomes. The multivariate modeling approach provides an opportunity to compare the covariate effects through hypothesis testing. Testing procedures have been developed based on the likelihood ratio and bootstrap techniques to compare the nonlinear covariate effects among different subgroups of the population (Liu and Tu, 2012). Here, a likelihood-based resampling procedure has been proposed to compare the bivariate nonparametric functions across multiple outcomes, which may advance the understanding of the concurrent nonlinear influences on the outcomes in a comparative manner.

For univariate longitudinal data, likelihood-based methods provide valid inference on fixed effects when data are missing at random as long as the joint distribution of the outcomes is specified correctly (Fitzmaurice et al., 2004). This holds for the multivariate semiparametric models presented in this paper since the proposed estimation procedure is also based on (penalized) likelihood. In the GRACE trial, the control group had very few intermittent missing values for ED visit and hospital admission; most of the missing data came from dropout. About 32% of the patients did not complete the 2-year trial. Further

research is needed to examine the possibility of nonignorable missing data mechanism and to develop approaches for handling data not missing at random within the proposed modeling framework.

Currently the proposed method works for the situation in which the multiple outcomes follow the same type of distribution (e.g., normal, Bernoulli and Poisson distributions). An important extension will be to accommodate mixed types of data such as continuous and discrete outcomes. Another area of methodological development is to provide practical tools for evaluating the goodness-of-fit of the proposed model.

Table 3.1: Parameter estimates and mean squared errors of smooth functions under Setting 1 with bootstrap standard errors (in parentheses) and coverage probabilities of 95% confidence intervals (in brackets).

| $m$ | $n$ | $\beta_{01} = 1.5$ | $\beta_{02} = 1$ | $\beta_{11} = 0.3$ | $\beta_{12} = -0.2$ | $\sigma_1 = 0.7$ | $\sigma_2 = 0.7$ | $\rho = 0.7$ | MSE($\hat{s}_1$) | MSE($\hat{s}_2$) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1.501 | 1.006 | 0.300 | $-0.199$ | 0.697 | 0.694 | 0.701 | | |
| 200 | 5 | (0.056) | (0.057) | (0.028) | (0.036) | (0.053) | (0.057) | (0.045) | 0.0070 | 0.0058 |
| | | [96.5%] | [94.0%] | [90.0%] | [93.5%] | [94.5%] | [96.5%] | [95.5%] | | |
| | | 1.503 | 1.002 | 0.298 | $-0.198$ | 0.698 | 0.698 | 0.701 | | |
| 200 | 10 | (0.053) | (0.053) | (0.019) | (0.024) | (0.051) | (0.053) | (0.040) | 0.0055 | 0.0040 |
| | | [93.5%] | [93.0%] | [95.5%] | [95.0%] | [94.0%] | [93.5%] | [94.0%] | | |
| | | 1.504 | 1.002 | 0.300 | $-0.199$ | 0.699 | 0.696 | 0.698 | | |
| 500 | 5 | (0.035) | (0.036) | (0.018) | (0.023) | (0.034) | (0.037) | (0.028) | 0.0054 | 0.0039 |
| | | [96.0%] | [94.0%] | [95.0%] | [96.5%] | [94.5%] | [94.0%] | [95.0%] | | |
| | | 1.507 | 1.005 | 0.300 | $-0.201$ | 0.700 | 0.699 | 0.700 | | |
| 500 | 10 | (0.033) | (0.033) | (0.012) | (0.015) | (0.033) | (0.034) | (0.025) | 0.0047 | 0.0031 |
| | | [94.0%] | [96.5%] | [94.5%] | [95.5%] | [95.0%] | [95.5%] | [96.5%] | | |

Table 3.2: Parameter estimates and mean squared errors of smooth functions under Setting 2 with bootstrap standard errors (in parentheses) and coverage probabilities of 95% confidence intervals (in brackets).

| $m$ | $n$ | $\beta_{01} = -0.8$ | $\beta_{02} = -1.2$ | $\beta_{11} = 2$ | $\beta_{12} = 1$ | $\sigma_1 = 0.5$ | $\sigma_2 = 1$ | $\rho = 0.4$ | MSE($\hat{s}_1$) | MSE($\hat{s}_2$) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | −0.803 | −1.204 | 2.002 | 1.004 | 0.491 | 0.977 | 0.389 | | |
| 200 | 5 | (0.069) | (0.102) | (0.079) | (0.095) | (0.047) | (0.155) | (0.123) | 0.0232 | 0.0203 |
| | | [93.0%] | [92.0%] | [93.0%] | [94.5%] | [93.0%] | [93.0%] | [91.5%] | | |
| | | −0.798 | −1.193 | 1.997 | 0.999 | 0.495 | 0.984 | 0.395 | | |
| 200 | 10 | (0.053) | (0.086) | (0.052) | (0.062) | (0.037) | (0.130) | (0.091) | 0.0138 | 0.0109 |
| | | [96.0%] | [96.0%] | [92.5%] | [94.5%] | [93.5%] | [94.0%] | [92.5%] | | |
| | | −0.799 | −1.200 | 2.000 | 1.002 | 0.495 | 0.984 | 0.390 | | |
| 500 | 5 | (0.043) | (0.065) | (0.050) | (0.058) | (0.030) | (0.099) | (0.075) | 0.0128 | 0.0099 |
| | | [94.0%] | [95.0%] | [93.0%] | [94.0%] | [94.5%] | [96.0%] | [95.5%] | | |
| | | −0.796 | −1.195 | 1.999 | 0.998 | 0.495 | 0.990 | 0.398 | | |
| 500 | 10 | (0.034) | (0.055) | (0.033) | (0.039) | (0.024) | (0.085) | (0.056) | 0.0083 | 0.0063 |
| | | [93.0%] | [94.0%] | [95.0%] | [94.0%] | [94.0%] | [95.5%] | [95.0%] | | |

Table 3.3: Model fitting results for the GRACE trial data.

| Outcome | Covariate | Estimate | SE | 95% CI |
|---|---|---|---|---|
| ED Visit | Intercept | $-1.58$ | 0.90 | $(-3.24, 0.29)$ |
| | Female | $-0.12$ | 0.15 | $(-0.43, 0.17)$ |
| | Black | $-0.03$ | 0.13 | $(-0.22, 0.31)$ |
| | Age | 0.0002 | 0.01 | $(-0.03, 0.02)$ |
| | Number of Comorbidities | 0.13 | 0.04 | $(0.06, 0.21)$ |
| Hospital Admission | Intercept | $-4.83$ | 1.11 | $(-7.13, -2.89)$ |
| | Female | $-0.37$ | 0.22 | $(-0.79, 0.11)$ |
| | Black | 0.08 | 0.20 | $(-0.48, 0.35)$ |
| | Age | 0.03 | 0.01 | $(0.001, 0.06)$ |
| | Number of Comorbidities | 0.21 | 0.06 | $(0.10, 0.34)$ |

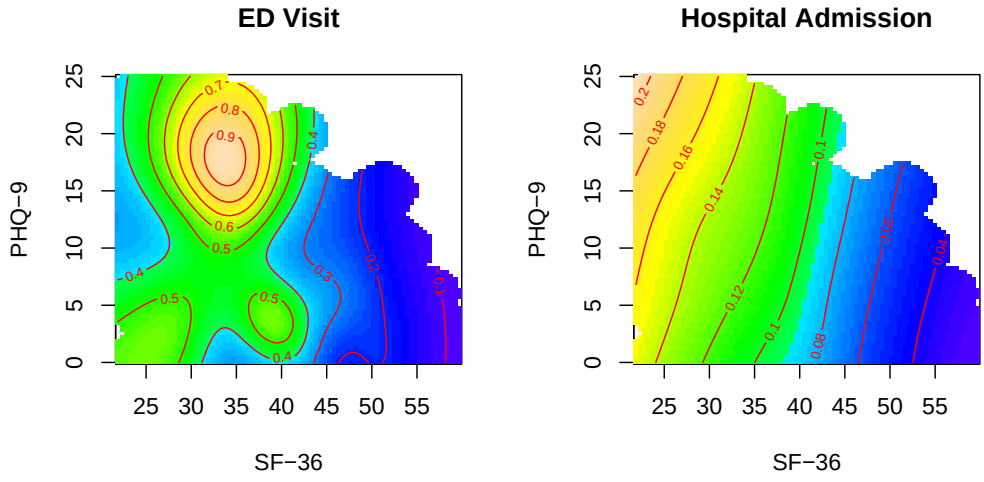


Figure 3.1: Contour plots of estimated joint effects of SF-36 and PHQ-9 scores on ED visit and hospital admission rates.

# Chapter 4

## Variable Selection in Multivariate Semiparametric Models

This chapter discusses variable selection in the proposed multivariate semiparametric mixed models. The situation where multiple outcomes are normally distributed is considered here. A two-stage method is proposed for simultaneous selection of the fixed and random effects as well as the interaction effects in the bivariate nonparametric functions. An expectation-maximization algorithm is developed to implement the method. The performance of the method is evaluated through simulation studies, followed by an illustration using data from a clinical investigation.

### 4.1 Research Background

Longitudinally assessed multiple outcome data are abundant in clinical investigations. For example, systolic and diastolic blood pressure readings are always measured in pairs. Together, they quantify the arterial pressure that circulating blood exerts on the walls of blood vessels during a cardiac cycle. Although one could choose to analyze systolic and diastolic readings in separate models, simultaneous modeling of the two measures provides a more complete picture of the systemic circulation; it also affords an opportunity to test and compare the unique contributing factors to these outcomes.

Among the existing methods, multivariate semiparametric mixed effects models provide perhaps the most general analytical framework for such data. Among other things, the inclusion of nonparametric independent variable effects has greatly enhanced the modeling flexibility for accommodating nonlinear effects. Structurally, these models are extensions of the traditional mixed effects models (Laird and Ware, 1982), where the fixed effects

characterize the influences of independent variables, and the random effects reflect the dependency of repeated measures within each outcome, as well as the correlations across outcomes within each subject (Reinsel, 1982). The nonparametric components are reserved for depiction of the nonlinear independent variable effects (Coull and Staudenmayer, 2004; Ghosh and Tu, 2009). Recently, this modeling framework has been extended to include bivariate smooth functions to describe the interacting influences of nonlinear independent variables (Liu and Tu, 2012).

These techniques have been successfully used to disseminate the concurrent influences of biological regulators of blood pressure (Tu et al., 2014, 2012; Yu et al., 2013). Editorial commentaries of these studies noted their contributions to the understanding of the pathogenesis of essential hypertension (Falkner and Gidding, 2011; Funder, 2014). These findings would not have been made without the advancement of statistical methodology.

What remains unavailable is a systematic approach that helps investigators to determine the inclusion of independent variables and the functional forms with which key variables enter the model. This is practically important because including unnecessary variables reduces model efficiency and creates numerical instability. Similarly, correct specification of the random effects ensures the validity of variance estimation and statistical inference (Lange and Laird, 1989); it also determines the correlation structure from which multiple outcomes arise. Along the same line, the inclusion of specific interactions (as depicted by the bivariate surfaces) need to be justified in a more objective manner.

Such a methodological need fits nicely into the context of variable selection. A traditional approach for variable selection is to perform likelihood-based model comparisons, using the Akaike or the Bayesian information criteria (AIC or BIC) (Akaike, 1973; Schwarz, 1978). While the information criterion-based methods are being used in practice, computational challenges are often formidable, especially when the model is complex, as the

number of competing models increases at a much faster rate than the number of predictors. Additionally, effectiveness of AIC and BIC is debatable for mixed models especially when the focus is on selection of random effects (Greven and Kneib, 2010; Keselman et al., 1998; Liang et al., 2008). An alternative approach that has gained increasing popularity in recent years is regularization. It includes various methods based on the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001), the least angle regression (LARS; Efron et al., 2004), and the adaptive LASSO (Zou, 2006). Different variations of the penalized methods have been used for simultaneous selection of fixed and random effects in linear mixed models (Bondell et al., 2010; Fan and Li, 2012) and in generalized linear mixed models (Ibrahim et al., 2011). Extension of the variable selection methods to semiparametric models for longitudinal data has been limited. Among the published methods, Fan and Li (2004) employed the SCAD penalty to select parametric covariate effects in a class of semiparametric models which did not require explicit specification of the correlation structure in longitudinal data. Following a similar vein, Ni et al. (2010) proposed a double-penalized likelihood method for semiparametric mixed models, in which a shrinkage penalty was imposed for fixed effect selection and a roughness penalty was applied for smooth function estimation. More recently, Zhang et al. (2011) proposed a data-driven method for determining the adequacy of linear effects of independent variables. To the best of my knowledge, none of these selection tools are available in a multivariate semiparametric modeling setting.

In this chapter, I present a variable selection tool for determining the inclusion of fixed and random effects in multivariate semiparametric models. Additionally, the proposed procedure helps to determine the cross-outcome correlations and to select the interaction effects in the form of bivariate smooth functions. Specifically, a two-stage model selection and estimation method is developed. In Stage 1, the regularization methods are used

to simultaneously select the fixed and random effects, and the interaction effects in the nonparametric components. In Stage 2, the unbiased estimates for selected parameters are obtained by maximizing the observed likelihood function. The performance of the proposed method is demonstrated in simulation studies. Finally, the method is illustrated by analyzing research data from a blood pressure study.

## 4.2 Methods

### 4.2.1 Model Formulation

Suppose that there are $m$ subjects in a longitudinal study and K outcomes are measured at each visit. For the $i$th subject, let $Y_{ijk}$ be the $k$th outcome observed at the $j$th time point of repeated measurements, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, and $k = 1, \ldots, K$. Consider the following multivariate semiparametric mixed model

$$Y_{ijk} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k + \boldsymbol{z}_{ij}^T \boldsymbol{u}_{ik} + s_k(t_{1ij}, t_{2ij}) + \epsilon_{ijk}, \tag{4.1}$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp})^T$ is a $p \times 1$ coefficient vector for the fixed effects $\boldsymbol{x}_{ij}$, $\boldsymbol{u}_{ik}$ is a $q \times 1$ vector of subject- and outcome-specific random effects for the corresponding covariates $\boldsymbol{z}_{ij}$, which could be a subset of $\boldsymbol{x}_{ij}$. It is assumed that the random effects $\boldsymbol{u}_{ik}$ follow a multivariate normal distribution $\mathrm{N}_q(\boldsymbol{0}, \boldsymbol{D}_{kk})$, the measurement errors $\epsilon_{ijk}$ independently follow a normal distribution $\mathrm{N}(0, \sigma_k^2)$, and $\boldsymbol{u}_{ik}$ and $\epsilon_{ijk}$ are independent. Let $t_{1ij}$ and $t_{2ij}$ be the continuous covariates that potentially have nonlinear influences on the outcomes. Without loss of generality, a bivariate nonparametic smooth function $s_k$ of $t_1$ and $t_2$ is included in the model, which can be easily extended to multiple nonparametric components. The primary objective is to simultaneously select important fixed and random effects from $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ respectively, as well as examine to whether $t_{1ij}$ and $t_{2ij}$ interact with each other.

For the bivariate smooth function $s_k$, a tensor product basis (Ruppert et al., 2003) is specified, which consists of marginal basis functions $\phi_{l_1}(t_1)$, $l_1 = 1, \ldots, L_1$ for $t_1$, $\psi_{l_2}(t_2)$, $l_2 = 1, \ldots, L_2$ for $t_2$, and all of their pairwise products. Examples of the marginal basis functions include truncated polynomials and B-splines. The possible interactions between $t_{1ij}$ and $t_{2ij}$ are incorporated through the product terms in the tensor product basis, and therefore they can be selected while keeping the main effects intact. Assuming $\phi_1(t_{1ij}) = \psi_1(t_{2ij}) = 1$, $s_k$ can be written as $s_k(t_{1ij}, t_{2ij}) = \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \alpha_{l_1,l_2,k} \phi_{l_1}(t_{1ij}) \psi_{l_2}(t_{2ij})$ where $\alpha_{l_1,l_2,k}$ are the coefficients associated with the corresponding basis functions. Using a vector form, $s_k(t_{1ij}, t_{2ij}) = \boldsymbol{T}_{ij}^T \boldsymbol{\alpha}_k$ where $\boldsymbol{\alpha}_k$ is a vector of the coefficients $\alpha_{l_1,l_2,k}$ for $l_1 = 1, \ldots, L_1$ and $l_2 = 1, \ldots, L_2$, and $\boldsymbol{T}_{ij}$ is the corresponding vector of the tensor product basis functions.

For convenience, model (4.1) is rewritten into a matrix form. Define the response vector as $\boldsymbol{Y}_i = (Y_{i11}, \cdots, Y_{in_i1}, \cdots, Y_{i1K}, \cdots, Y_{in_iK})^T$, the fixed effects coefficient vector as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_K^T)^T$, the subject-specific random effects as $\boldsymbol{u}_i = (\boldsymbol{u}_{i1}^T, \cdots, \boldsymbol{u}_{iK}^T)^T$, the coefficient vector for the smooth functions as $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_K^T)^T$, and the vector of measurement errors as $\boldsymbol{\epsilon}_i = (\epsilon_{i11}, \cdots, \epsilon_{in_i1}, \cdots, \epsilon_{i1K}, \cdots, \epsilon_{in_iK})^T$. Then model (4.1) can be rewritten as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{u}_i + \boldsymbol{T}_i \boldsymbol{\alpha} + \boldsymbol{\epsilon}_i, \tag{4.2}$$

where $\boldsymbol{u}_i$ follows a multivariate normal distribution $\mathrm{N}_{Kq}(\boldsymbol{0}, \boldsymbol{D})$. The covariance matrix $\boldsymbol{D}$ accommodates the within-subject correlations among the repeated measurements (through

the diagonal blocks) and across the multiple outcomes. Specifically, $\boldsymbol{D}$ can be written as

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} & \cdots & \boldsymbol{D}_{1K} \\ \boldsymbol{D}_{21} & \boldsymbol{D}_{22} & \cdots & \boldsymbol{D}_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{D}_{K1} & \boldsymbol{D}_{K2} & \cdots & \boldsymbol{D}_{KK} \end{pmatrix}, \tag{4.3}$$

where $\boldsymbol{D}_{jk}$ for $j, k = 1, \ldots, K$ are $q \times q$ submatrices. The diagonal submatrices $\boldsymbol{D}_{jj}$ are the covariance matrices of the random effects within each of the outcomes, and the off-diagonal ones indicate the potential correlations across the outcomes. If the outcomes are not correlated, then $\boldsymbol{D} = \text{diag}(\boldsymbol{D}_{11}, \cdots, \boldsymbol{D}_{KK})$. In addition, the measurement errors $\boldsymbol{\epsilon}_i \sim \text{N}_{Kn_i}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$ where $\boldsymbol{\Sigma}_i = \text{diag}(\sigma_1^2 \boldsymbol{I}_{n_i}, \cdots, \sigma_K^2 \boldsymbol{I}_{n_i})$.

Cholesky decomposition of the covariance matrix $\boldsymbol{D}$ is a key step for the selection of random effects as it ensures that $\boldsymbol{D}$ is positive semidefinite (Bondell et al., 2010; Chen and Dunson, 2003; Ibrahim et al., 2011; Kinney and Dunson, 2007). Using similar approach as in Ibrahim et al. (2011), $\boldsymbol{D}$ can be decomposed as $\boldsymbol{D} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$ where $\boldsymbol{\Gamma}$ is a $Kq \times Kq$ lower triangular matrix. Accordingly, the random effects $\boldsymbol{u}_i$ can be reparameterized as $\boldsymbol{u}_i = \boldsymbol{\Gamma}\boldsymbol{b}_i$, and $\boldsymbol{b}_i \sim \text{N}_{Kq}(\boldsymbol{0}, \boldsymbol{I}_{Kq})$. Then model (4.2) becomes

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{b}_i + \boldsymbol{T}_i\boldsymbol{\alpha} + \boldsymbol{\epsilon}_i. \tag{4.4}$$

### 4.2.2 Penalized Likelihood

A penalized likelihood method is used for simultaneous selection of fixed effects, random effects, and interaction effects between the two covariates in the smooth functions. A specific aim is to determine whether there are within-subject correlations across the out-

comes by identifying the nonzero elements of $\boldsymbol{D}$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\alpha}^T, \boldsymbol{\sigma}^T)^T$ be a vector of all unknown parameters, where $\boldsymbol{\gamma}$ is a $\frac{Kq(Kq+1)}{2} \times 1$ vector of parameters of $\boldsymbol{\Gamma}$, and $\boldsymbol{\sigma} = (\sigma_1^2, \ldots, \sigma_K^2)$. I propose to maximize the following penalized (observed) log-likelihood function:

$$p\ell_o(\boldsymbol{\theta}) = \ell_o(\boldsymbol{\theta}) - \eta_{\lambda_1}(\boldsymbol{\beta}) - \eta_{\lambda_2}(\boldsymbol{\gamma}) - \sum_{k=1}^{K} \eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k), \tag{4.5}$$

where $\ell_o(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \log f_o(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta})$ is the observed log-likelihood function, and $\eta_{\lambda_j}(\cdot)$ for $j = 1, \ldots, K+2$ are nonnegetive and nondecreasing penalty functions for fixed effects, random effects and smooth functions, with $\lambda_j > 0$ being the tuning parameters which control the amount of shrinkage.

Many options for the penalty functions can be considered as discussed in Section 4.1. Here, the adaptive LASSO penalty is adopted for easy implementation in practice. For the fixed effects, the penalty function is defined as $\eta_{\lambda_1}(\boldsymbol{\beta}) = \lambda_1 \sum_{k=1}^{K}\sum_{l=1}^{p} |\widetilde{\beta}_{kl}|^{-1}|\beta_{kl}|$, where $\widetilde{\beta}_{kl}$ are the unpenalized maximum likelihood estimators (MLEs). Note that it may not be necessary to penalize all of the fixed effects coefficients, for example, the intercept can be left out of the penalty function.

Selecting random effects in a multivariate model involves identifying important random effects for each outcome and determining the correlation structures across the outcomes. Similar to equation (4.3),the lower triangular matrix $\boldsymbol{\Gamma}$ is partitioned as

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_{11} & & & \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} & & \\ \vdots & \vdots & \ddots & \\ \boldsymbol{\Gamma}_{K1} & \boldsymbol{\Gamma}_{K2} & \cdots & \boldsymbol{\Gamma}_{KK} \end{pmatrix},$$

where $\boldsymbol{\Gamma}_{jk}$ for $j = 1, \ldots, K$ and $k = 1, \ldots, j$ are $q \times q$ submatrices, and the diagonal submatri-

ces $\mathbf{\Gamma}_{jj}$ are also lower triangular. For a single outcome, the penalty is placed on the row vectors of $\mathbf{\Gamma}_{jj}$ in a grouped manner to ensure the positive semidefiniteness of $\boldsymbol{D}$ (Ibrahim et al., 2011; Yuan and Lin, 2006). If the elements of a certain row of $\mathbf{\Gamma}_{jj}$ are all shrunk to zero, then the corresponding row and column vectors of $\boldsymbol{D}_{jj}$ will also be zero and thus the corresponding random effect will be removed from the model. In a multivariate setting, I propose to use the following penalty function: $\eta_{\lambda_2}(\boldsymbol{\gamma}) = \lambda_2 \sum_{j=1}^{K} \sum_{k=1}^{j} \sum_{l=1}^{p} \sqrt{c_{jkl}} ||\widetilde{\gamma}_{jkl}||^{-1} ||\gamma_{jkl}||$, where $\gamma_{jkl}$ for $l = 1, \ldots, p$ are the $l$th rows of the submatrix $\mathbf{\Gamma}_{jk}$, $\widetilde{\gamma}_{jkl}$ are the unpenalized MLEs, and $c_{jkl}$ are normalizing constants to adjust for the varying sizes of $\gamma_{jkl}$ (e.g., $c_{jkl} = dim(\gamma_{jkl})$). Penalizations of the diagonal and off-diagonal submatrices in $\mathbf{\Gamma}$ are separated so that the non-random elements in the within- and between-outcome variance components can be identified individually.

To select the interaction terms in the smooth functions, grouped penalties are imposed on the corresponding product terms in the tensor product basis. For $s_k$, the penalty function is defined as $\eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k) = \lambda_{2+k} ||\widetilde{\alpha}_k^*||^{-1} ||\alpha_k^*||$, where $\alpha_k^*$ is a $(L_1-1)(L_2-1)\times 1$ vector consisting of $\alpha_{l_1,l_2,k}$, $l_1 = 2, \ldots, L_1$, $l_2 = 2, \ldots, L_2$. Note that different degrees of penalty are allowed for smooth functions $s_k$ in penalized likelihood (4.5) through different tuning parameters $\lambda_{2+k}$, $k = 1, \ldots, K$.

### 4.3 Computational Algorithm

The model selection and estimation procedure is implemented in two stages. In Stage 1, model selection is performed by maximizing the penalized likelihood function. Given a set of tuning parameters $\boldsymbol{\lambda} = \{\lambda_j\}_{j=1}^{K+2}$, I use an EM algorithm to optimize (4.5) and obtain the maximum penalized likelihood estimator (MPLE) $\hat{\boldsymbol{\theta}}_\lambda$. The optimization procedure is carried out for different values of $\boldsymbol{\lambda}$, and the optimal $\boldsymbol{\lambda}$ is selected based on a certain criterion. Covariates corresponding to the nonzero elements in $\hat{\boldsymbol{\theta}}_\lambda$ will be selected. In Stage

2, I refit the model with selected fixed effects, random effects and smooth functions (with or without interactions) to obtain the MLE $\hat{\boldsymbol{\theta}}$. Penalized regression splines are used as the smoother.

### 4.3.1 EM algorithm

Consider $(\boldsymbol{Y}_i, \boldsymbol{b}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ and $(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ for $i = 1, \ldots, m$ as the complete data and observed data, respectively. With the same penalty functions in (4.5), the penalized complete log-likelihood function can be written as

$$p\ell_c(\boldsymbol{\theta}) = \ell_c(\boldsymbol{\theta}) - \eta_{\lambda_1}(\boldsymbol{\beta}) - \eta_{\lambda_2}(\boldsymbol{\gamma}) - \sum_{k=1}^{K} \eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k), \tag{4.6}$$

where $\ell_c(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \log f_c(\boldsymbol{Y}_i, \boldsymbol{b}_i | \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta})$ is the complete log-likelihood function.

Denote the estimates of $\boldsymbol{\theta}$ at the $s$th iteration by $\boldsymbol{\theta}^{(s)} = (\boldsymbol{\beta}^{(s)T}, \boldsymbol{\gamma}^{(s)T}, \boldsymbol{\alpha}^{(s)T}, \boldsymbol{\sigma}^{(s)T})^T$. In the E-step, for fixed tuning parameter $\boldsymbol{\lambda}$, the expectation of the penalized complete log-likelihood (4.6) given the observed data and $\boldsymbol{\theta}^{(s)}$ can be calculated as follows:

$$\begin{aligned}
Q_\lambda(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) &= E[p\ell_c(\boldsymbol{\theta})|(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)_{i=1}^m, \boldsymbol{\theta}^{(s)}] \\
&= \frac{1}{m}\sum_{i=1}^{m} E[\log f_c(\boldsymbol{Y}_i, \boldsymbol{b}_i|\boldsymbol{\theta})|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}] \\
&\quad - \eta_{\lambda_1}(\boldsymbol{\beta}) - \eta_{\lambda_2}(\boldsymbol{\gamma}) - \sum_{k=1}^{K} \eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k) \\
&= \frac{1}{m}\sum_{i=1}^{m} E[\log f(\boldsymbol{Y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}] + \frac{1}{m}\sum_{i=1}^{m} E[\log f_b(\boldsymbol{b}_i)|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{(s)}] \\
&\quad - \eta_{\lambda_1}(\boldsymbol{\beta}) - \eta_{\lambda_2}(\boldsymbol{\gamma}) - \sum_{k=1}^{K} \eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k),
\end{aligned} \tag{4.7}$$

where $f(\boldsymbol{Y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}) = \mathrm{N}_{Kn_i}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{b}_i + \boldsymbol{T}_i\boldsymbol{\alpha}, \boldsymbol{\Sigma}_i)$, and $f_b(\boldsymbol{b}_i) = \mathrm{N}_{Kq}(\boldsymbol{0}, \boldsymbol{I}_{Kq})$. Let $g_1(\boldsymbol{b}_i, \boldsymbol{\theta}) = \log f(\boldsymbol{Y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})$, and $g_2(\boldsymbol{b}_i) = \log f_b(\boldsymbol{b}_i)$. The two expectation terms in (4.7)

can be written as

$$E[g_1(\boldsymbol{b}_i,\boldsymbol{\theta})|\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)}] = \int g_1(\boldsymbol{b}_i,\boldsymbol{\theta})h(\boldsymbol{b}_i|\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)})d\boldsymbol{b}_i, \tag{4.8}$$

and

$$E[g_2(\boldsymbol{b}_i)|\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)}] = \int g_2(\boldsymbol{b}_i)h(\boldsymbol{b}_i|\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)})d\boldsymbol{b}_i, \tag{4.9}$$

where

$$\begin{aligned} h(\boldsymbol{b}_i|\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)}) &= \frac{f_c(\boldsymbol{Y}_i,\boldsymbol{b}_i|\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)})}{f_o(\boldsymbol{Y}_i|\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)})} \\ &= \frac{f(\boldsymbol{Y}_i|\boldsymbol{b}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)})f_b(\boldsymbol{b}_i|\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)})}{\int f(\boldsymbol{Y}_i|\boldsymbol{b}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)})f_b(\boldsymbol{b}_i|\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)})d\boldsymbol{b}_i}. \end{aligned} \tag{4.10}$$

Since the $q$-dimensional integrals in (4.8), (4.9) and the denominator of (4.10) are usually intractable, multivariate Gauss-Hermite quadrature rules can be used to approximate them (Pinhero and Bates, 1995). Denote the number of quadrature nodes for each dimension by $n$. Let $\boldsymbol{b}_d$ and $w_d$, $d = 1,\ldots,N_{\text{GH}}$ be the pre-specified quadrature nodes and weights respectively, where the total number of quadrature nodes is $N_{\text{GH}} = n^q$. Then the first expectation term (4.8) can be approximated as

$$E[g_1(\boldsymbol{b}_i,\boldsymbol{\theta})|\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)}] \approx \sum_{d=1}^{N_{\text{GH}}} w_d \exp(||\boldsymbol{b}_d||^2) g_1(\boldsymbol{b}_d,\boldsymbol{\theta})h(\boldsymbol{b}_d|\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)}). \tag{4.11}$$

Since the second expectation term does not involve $\boldsymbol{\theta}$, it can be omitted in the M-step from the penalized Q-function (4.7), and thus $\boldsymbol{\theta}^{s+1}$ can be found by maximizing

$$Q^*_{\boldsymbol{\lambda}}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) - \eta_{\lambda_1}(\boldsymbol{\beta}) - \eta_{\lambda_2}(\boldsymbol{\gamma}) - \sum_{k=1}^{K} \eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k), \tag{4.12}$$

where $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \frac{1}{m}\sum_{i=1}^{m} E[g_1(\boldsymbol{b}_i,\boldsymbol{\theta})|\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i,\boldsymbol{\theta}^{(s)}]$.

Considering that maximizing (4.12) with respect to $\boldsymbol{\theta}$ involves high-dimensional optimization, I propose the following expectation-conditional maximization (ECM; Meng and Rubin (1993)) algorithm which breaks down the M-step into several conditional maximization (CM) steps:

1. Given $\boldsymbol{\gamma}^{(s)}$, $\boldsymbol{\alpha}^{(s)}$ and $\boldsymbol{\sigma}^{(s)}$, find $\boldsymbol{\beta}^{(s+1)} = \arg\max_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(s)}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\sigma}^{(s)} | \boldsymbol{\beta}^{(s)}, \boldsymbol{\gamma}^{(s)}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\sigma}^{(s)}) - m\eta_{\lambda_1}(\boldsymbol{\beta})$.
2. Given $\boldsymbol{\beta}^{(s+1)}$, $\boldsymbol{\alpha}^{(s)}$ and $\boldsymbol{\sigma}^{(s)}$, find $\boldsymbol{\gamma}^{(s+1)} = \arg\max_{\boldsymbol{\gamma}} Q(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\sigma}^{(s)} | \boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}^{(s)}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\sigma}^{(s)}) - m\eta_{\lambda_2}(\boldsymbol{\gamma})$.
3. Given $\boldsymbol{\beta}^{(s+1)}$, $\boldsymbol{\gamma}^{(s+1)}$ and $\boldsymbol{\sigma}^{(s)}$, find $\boldsymbol{\alpha}^{(s+1)} = \arg\max_{\boldsymbol{\alpha}} Q(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}^{(s+1)}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^{(s)} | \boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}^{(s+1)}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\sigma}^{(s)}) - m \sum_{k=1}^{K} \eta_{\lambda_{2+k}}(\boldsymbol{\alpha}_k)$.
4. Given $\boldsymbol{\beta}^{(s+1)}$, $\boldsymbol{\gamma}^{(s+1)}$ and $\boldsymbol{\alpha}^{(s+1)}$, find $\boldsymbol{\sigma}^{(s+1)} = \arg\max_{\boldsymbol{\sigma}} Q(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}^{(s+1)}, \boldsymbol{\alpha}^{(s+1)}, \boldsymbol{\sigma} | \boldsymbol{\beta}^{(s+1)}, \boldsymbol{\gamma}^{(s+1)}, \boldsymbol{\alpha}^{(s+1)}, \boldsymbol{\sigma}^{(s)})$.
5. Steps $1 - 4$ are iterated until convergence to obtain the MPLE $\hat{\boldsymbol{\theta}}_\lambda$.

The optimization procedure is started by fitting the full model with all covariates and using the parameter estimates as the initial values.

### 4.3.2 Tuning Parameter Selection

The performance of the proposed method depends on the appropriate selection of the tuning parameters. Selection criteria that have been used extensively include cross validation (CV), generalized cross-validation (GCV) and information criterion such as AIC and BIC. It has been shown that GCV tends to select overfitted models, while BIC can identify the true model consistently (Shao, 1997; Wang et al., 2009, 2007). Therefore, the following BIC-type criterion is proposed to select the optimal tuning parameters:

$$BIC_\lambda = -2\ell_o(\hat{\boldsymbol{\theta}}_\lambda) + \log(N) df_\lambda, \tag{4.13}$$

where $\ell_o(\hat{\boldsymbol{\theta}}_\lambda)$ is the value of the observed log-likelihood at the MPLE $\hat{\theta}_\lambda$ obtained through the proposed EM algorithm for a given $\boldsymbol{\lambda}$. In practice, $\ell_o(\hat{\boldsymbol{\theta}}_\lambda)$ is approximated using the Gauss-Hermite quadrature rules described in Section 4.3.1. The sample size $N$ in a multivariate setting is defined as $N = K\sum_{i=1}^{m} n_i$. The degrees of freedom $df_\lambda$ is defined as the number of nonzero elements of $\hat{\boldsymbol{\theta}}_\lambda$. The EM algorithm proposed above is repeated over a grid of tuning parameters, and the one that minimizes $BIC_\lambda$ is considered optimal.

### 4.3.3 Implementation

The proposed computational algorithm is developed in R. The M-steps in the ECM algorithm are implemented using the `optim` function in the `stats` package (R Development Core Team, 2011). The initial values of the parameters are obtained by fitting the full multivariate semiparametric model using the `gamm4` function in the `gamm4` package (Wood, 2011).

## 4.4 Simulation Studies

To evaluate the performance of the proposed method, two settings are considered in the simulation study. For each setting, bivariate outcomes were generated from the following model

$$\begin{cases} Y_{ij1} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_1 + \boldsymbol{z}_{ij}^T\boldsymbol{u}_{i1} + \bar{s}_1(t_{1ij}, t_{2ij}) + \epsilon_{ij1} \\ Y_{ij2} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2 + \boldsymbol{z}_{ij}^T\boldsymbol{u}_{i2} + \bar{s}_2(t_{1ij}, t_{2ij}) + \epsilon_{ij2}, \end{cases} \tag{4.14}$$

for $i = 1, \ldots, 200$ and $j = 1, \ldots, 5$.

In Setting 1, the fixed effect coefficients were specified as $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15})^T = (1, 1, 3, 0, -1, 0)^T$ and $\boldsymbol{\beta}_2 = (\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25})^T = (1, 2, 0, -2, 0, 0)^T$. The corresponding covariates $\boldsymbol{x}_{ij} = (x_{ij0}, x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, x_{ij5})^T$ were generated independently from N$(0, 1)$ except that the intercept $x_{ij0} = 1$. The subject-specific random effects were

$(\boldsymbol{u}_{i1}^T, \boldsymbol{u}_{i2}^T)^T = (u_{i10}, u_{i11}, u_{i12}, u_{i20}, u_{i21}, u_{i22})^T \sim \mathrm{N}_6(\boldsymbol{0}, \boldsymbol{D})$ with

$$\boldsymbol{D} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0.5 & 1.25 & 0.75 & 0 & 0 & 0 \\ 0.5 & 0.75 & 1.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and the corresponding covariates $\boldsymbol{z}_{ij} = (z_{ij0}, z_{ij1}, z_{ij2})^T = (x_{ij0}, x_{ij1}, x_{ij2})^T$. Note that the outcomes were independent of each other since the $3 \times 3$ off-diagonal submatrices in $\boldsymbol{D}$ were $\boldsymbol{0}$. The smooth functions were given by $s_1(t_1, t_2) = t_1 + t_2$ and $s_2(t_1, t_2) = t_1 + t_2 + 2\exp(t_1)/(1.2 - t_2)$ with $t_1, t_2 \sim \mathrm{Uniform}(0, 1)$, and $\bar{s}_1(t_{1ij}, t_{2ij})$ and $\bar{s}_2(t_{1ij}, t_{2ij})$ were the values of corresponding smooth functions centered over $(t_{1ij}, t_{2ij})$ for $i = 1, \ldots, 200$ and $j = 1, \ldots, 5$. The measurement errors $\epsilon_{ij1} \sim \mathrm{N}(0, \sigma_1^2)$ and $\epsilon_{ij2} \sim \mathrm{N}(0, \sigma_2^2)$ with $\sigma_1 = 1$ and $\sigma_2 = 1.5$.

In Setting 2, the setup was the same except that the outcomes were correlated with the

covariance matrix $\boldsymbol{D}$ given by

$$\boldsymbol{D} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.25 & 0.25 & 0 \\ 0.5 & 1.25 & 0.75 & 0.375 & 0.375 & 0 \\ 0.5 & 0.75 & 1.5 & 0.5 & 0.5 & 0 \\ 0.25 & 0.375 & 0.5 & 1.1875 & 0.6875 & 0 \\ 0.25 & 0.375 & 0.5 & 0.6875 & 0.6875 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

These two settings were chosen to assess whether the proposed method could correctly determine the correlation structure between the outcomes.

The simulation was repeated 100 times under each setting using the proposed two-stage method. In Stage 1, important fixed effects and random effects were identified; the existence of interaction effects in the bivariate surfaces were determined. The between-outcome correlation structure was determined as part of the random effects selection. Then the selected effects and surfaces were estimated in Stage 2. For estimation, the same algorithm was used as described in Section 4.3.1, with the penalty terms removed from the likelihood function (4.6).

Table 4.1 summarizes the selection results for the full model and its components, including the fixed effects, the random effects, the interaction effects in the smooth functions, and the correlation between the outcomes. It presents the percentages of times the correct model and components are identified, as well as the incorrect selection (an unimportant effect being selected) and incorrect exclusion (an important effect being excluded) rates. Under both settings, the selection algorithm achieved a high rate of correct selection of the true model and components. Specifically, it was able to identify the true fixed and random

effects, as well as the interactions, with correct selection rates $\geq 99\%$. The performance in terms of determining the between-outcome correlation was also satisfactory, although errors mostly occurred when the outcomes were truly correlated.

Tables 4.2 and 4.3 show the estimated fixed effect coefficients and variance components from Stage 2. The magnitude of biases in the estimation of non-zero parameters was generally small. Only one of the unimportant covariates was incorrectly included in the model. In addition, the mean squared errors (MSE) of the estimated smooth functions were calculated as follows: in Setting 1, $MSE(\hat{s}_1) = 0.010$, and $MSE(\hat{s}_2) = 0.347$; in Setting 2, $MSE(\hat{s}_1) = 0.039$, and $MSE(\hat{s}_2) = 0.382$. These results support the notion that the two-stage method worked well for both selection and estimation.

The simulation study was performed on a Dell PowerEdge R820 server with Linux operating system. The server has 4 Intel Xeon CPU ES-4620 8-core processors and 128 GB memory (shared by multiple users). The computing time increased rapidly with the number of random effects and the number of quadrature nodes. In both settings, 4 quadrature nodes was used for each random effect. Given a set of tuning parameters, it took approximately 2 hours to complete model selection in Stage 1. Parameter estimation for the selected model in Stage 2 took approximately 1.5 hours.

### 4.5 Real Data Analysis

This research is motivated by a long running cohort study of blood pressure development in children. In this section, I illustrate the proposed variable selection method by analyzing the study data. The recruitment protocol of the original study can be found in Pratt et al. (1989) and the follow-up protocol in Tu et al. (2011). Briefly, study subjects were recruited from schools in Indianapolis selected to provide a range of socioeconomic status. They were followed up twice a year to measure blood pressure, height, weight and upper arm circum-

ference. When measuring blood pressure, three readings were obtained at least two minutes apart, and the average of the last two was taken as the final measurement. Body mass index (BMI) was calculated based on height and weight as follows: BMI = weight/height$^2$. Urine samples were also collected to determine the urine volume and excretion rates of sodium and potassium.

A subset of the blood pressure data was used for this analysis. Of the 250 randomly selected subjects, 117 are males and 80 are blacks. The selected data include a total of 1776 follow-up visits, with an average of 7.1 visits per subject. The mean age at enrollment is 9.8 years (SD = 2.7 years).

In this analysis, systolic and diastolic blood pressure are considered as paired outcomes. The proposed method is used to identify covariates that are associated with the outcomes. The model selection process is started with the following full model

$$\begin{cases} Y_{ij1} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_1 + \boldsymbol{z}_{ij}^T\boldsymbol{u}_{i1} + s_1(t_{1ij}, t_{2ij}) + \epsilon_{ij1} \\ Y_{ij2} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2 + \boldsymbol{z}_{ij}^T\boldsymbol{u}_{i2} + s_2(t_{1ij}, t_{2ij}) + \epsilon_{ij2}, \end{cases} \tag{4.15}$$

where $Y_{ij1}$ and $Y_{ij2}$ are the systolic and diastolic blood pressure respectively for the $i$th subject measured at the $j$th visit, for $i = 1, \ldots, 250$ and $j = 1, \ldots, n_i$ with $n_i$ ranging from 1 to 18, $\boldsymbol{x}_{ij}$ is a vector of fixed effect covariates including intercept, gender (male or female), race (black or other), birth weight (pound), mother's length of pregnancy (month), upper arm circumference (cm), urine volume (L), urinary sodium excretion rate (mmol/mg creatinine) and urinary potassium excretion rate (mmol/mg creatinine), $\boldsymbol{z}_{ij}$ is a vector of random effect covariates including intercept, birth weight and mother's length of pregnancy, $(\boldsymbol{u}_{i1}^T, \boldsymbol{u}_{i2}^T)^T \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$ are the subject-specific random effects, $s_1$ and $s_2$ are bivariate smooth functions of age ($t_{1ij}$) and BMI ($t_{2ij}$), and $\epsilon_{ij1} \sim \mathrm{N}(0, \sigma_1^2)$ and $\epsilon_{ij2} \sim \mathrm{N}(0, \sigma_2^2)$ are the independent measurement errors. Age and BMI were chosen as the bivariate nonparametric

components because a preliminary analysis (Figure 4.1) showed that both had strong non-linear effects on blood pressure. Here, the interest is in examining whether they interact with each other.

In Stage 1, the outcomes and the continuous covariates were standardized to ensure numerical stability before selection was performed. In Stage 2, the selected covariates were estimated in the original scale so that the coefficient estimates could be easily interpreted. The model selection and estimation results are summarized in Table 4.4. Zero estimates indicate that the corresponding covariates were not selected. For the systolic blood pressure, the selected fixed effect covariates were gender, race and upper arm circumference; race and upper arm circumference were also selected for the diastolic blood pressure. Based on the coefficient estimates, males had significantly higher systolic blood pressure than females. Comparing to other races, blacks tended to have higher systolic and diastolic blood pressure. Upper arm circumference, an indicator of obesity, was positively associated with both the systolic and diastolic blood pressure.

As to random effects, neither of the covariates, birth weight and mother's length of pregnancy, were selected for the systolic or diastolic blood pressure. Table 4.4 provides the variance component estimates, i.e., square roots of the diagonal elements of $\boldsymbol{D}$, $\sigma_1$ and $\sigma_2$. In addition, the systolic and diastolic blood pressure were highly correlated within each subject ($\rho = 0.78, \mathrm{SE} = 0.074$), as suggested by the non-zero estimate of the off-diagonal element of $\boldsymbol{D}$.

The estimated bivariate smooth functions $s_1$ and $s_2$ are presented using the contour plots in Figure 4.2. Generally speaking, the systolic and diastolic blood pressure increased with both age and BMI. It can also be noted that there were substantial interactions between the two, and specifically, BMI effects on blood pressure were stronger in older children over 12 years of age than in younger children. These observations lend support to the model

selection results in which the interaction terms in both bivariate smooth functions were selected.

## 4.6 Discussion

Variable selection plays a fundamental role in scientific investigation. The ability to determine the relevance of independent variables to outcomes of interest is of vital importance to scientific inquiry. For a long time, variable selection has presented many practical challenges to analysts, who often struggled to find appropriate selection methods and implementation programs. The situation has improved significantly in the last two decades since the publication of the least absolute shrinkage and selection operator by Tibshirani (1996, 1997), which have provided a theoretical foundation for regularization methods. Important applications of LASSO to various modeling situations have since alleviated barriers for performing variable selections in most standard modeling settings. This said, for newly developed statistical models the challenge remains.

In this chapter, variable selection has been considered in multivariable semiparametric models, a class of models that have been shown to be useful, yet for which selection methods have not been available. To remedy, a two-stage model selection and estimation method has been presented for random and fixed effect selection and for determining the presence of interactions in the form of bivariate smooth functions. To the best of my knowledge, the proposed variable selection method is the first in this model setting. In fact, there are few formal discussions of variable selection in the context of multivariate models for repeated measurements. The selection of random effects in multivariate models is important, because the correlations across the outcomes are accommodated by the random effects. Therefore, by selecting random effects, one will be able to decide whether simultaneous modeling of multiple outcomes is truly necessary. The selection of interactions is equally important

because it facilitates an understanding of the concurrent influences of two continuous independent variables. Scientific investigations have repeatedly demonstrated the scarcity of true linear effects in biological research and the danger of over-simplification with linear approximations. The proposed method has excellent performance, as indicated by the simulation studies. This said, future extensions may be needed to make the method more widely applicable in data situations where non-normal outcomes are of interest. The extension is anticipated to be straightforward, although not trivial. Notwithstanding such limitations, the proposed method could be of use for a wide variety of investigations.

Table 4.1: Variable selection results based on the 100 simulation runs.

| | Setting 1 | | | Setting 2 | | |
|---|---|---|---|---|---|---|
| Model Component | Correct | Incorrect Selection | Incorrect Exclusion | Correct | Incorrect Selection | Incorrect Exclusion |
| Model | 99 | 1 | 0 | 94 | 1 | 6 |
| Fixed Effects | 99 | 1 | 0 | 99 | 1 | 0 |
| Random Effects | 100 | 0 | 0 | 100 | 0 | 0 |
| Correlation | 100 | 0 | – | 94 | – | 6 |
| Interaction in $s_1$ | 100 | 0 | – | 100 | 0 | – |
| Interaction in $s_2$ | 100 | – | 0 | 100 | – | 0 |

Table 4.2: Estimates of the fixed effect coefficients with empirical standard errors.

| | | Setting 1 | | Setting 2 | |
|---|---|---|---|---|---|
| Parameter | True Value | Estimate | SE | Estimate | SE |
| $\hat{\beta}_{10}$ | 1 | 0.941 | 0.055 | 0.985 | 0.037 |
| $\hat{\beta}_{11}$ | 1 | 1.013 | 0.041 | 1.079 | 0.072 |
| $\hat{\beta}_{12}$ | 3 | 2.957 | 0.058 | 2.955 | 0.052 |
| $\hat{\beta}_{13}$ | 0 | 0 | 0 | 0 | 0 |
| $\hat{\beta}_{14}$ | $-1$ | $-0.997$ | 0.021 | $-0.993$ | 0.023 |
| $\hat{\beta}_{15}$ | 0 | 0 | 0 | 0 | 0 |
| $\hat{\beta}_{20}$ | 1 | 1.062 | 0.056 | 0.904 | 0.059 |
| $\hat{\beta}_{21}$ | 2 | 2.077 | 0.051 | 1.923 | 0.064 |
| $\hat{\beta}_{22}$ | 0 | $-0.001$ | 0.008 | 0.001 | 0.006 |
| $\hat{\beta}_{23}$ | $-2$ | $-1.953$ | 0.034 | $-1.908$ | 0.213 |
| $\hat{\beta}_{24}$ | 0 | 0 | 0 | 0 | 0 |
| $\hat{\beta}_{25}$ | 0 | 0 | 0 | 0 | 0 |

Table 4.3: Estimates of the variance components with empirical standard errors.

| | Setting 1 | | | Setting 2 | | |
|---|---|---|---|---|---|---|
| Parameter | True Value | Estimate | SE | True Value | Estimate | SE |
| $\hat{D}_{11}$ | 1 | 0.923 | 0.058 | 1 | 0.864 | 0.137 |
| $\hat{D}_{22}$ | 1.25 | 1.306 | 0.084 | 1.25 | 1.364 | 0.098 |
| $\hat{D}_{33}$ | 1.5 | 1.771 | 0.165 | 1.5 | 1.776 | 0.127 |
| $\hat{D}_{44}$ | 1 | 1.044 | 0.062 | 1.1875 | 1.103 | 0.141 |
| $\hat{D}_{55}$ | 0.5 | 0.406 | 0.070 | 0.6875 | 0.742 | 0.074 |
| $\hat{D}_{66}$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $\hat{\sigma}_1$ | 1 | 0.998 | 0.018 | 1 | 1.029 | 0.013 |
| $\hat{\sigma}_2$ | 1.5 | 1.595 | 0.047 | 1.5 | 1.553 | 0.040 |

Table 4.4: Model selection and estimation results for the blood pressure data.

| | Systolic blood pressure | | Diastolic blood pressure | |
|---|---|---|---|---|
| Variable | Estimate | SE | Estimate | SE |
| Fixed effects | | | | |
| Intercept | 90.52 | 1.25 | 48.13 | 1.34 |
| Male | 2.69 | 0.45 | 0 | – |
| Black | 2.23 | 0.51 | 2.01 | 0.55 |
| Birth weight | 0 | – | 0 | – |
| Mother's length of pregnency | 0 | – | 0 | – |
| Upper arm circumference | 0.44 | 0.054 | 0.54 | 0.058 |
| Urine volume | 0 | – | 0 | – |
| Urinary sodium excretion | 0 | – | 0 | – |
| Urinary potassium excretion | 0 | – | 0 | – |
| Variance components | | | | |
| Intercept | 6.09 | 0.46 | 5.99 | 0.49 |
| Birth weight | 0 | – | 0 | – |
| Mother's length of pregnency | 0 | – | 0 | – |
| Error term | 9.28 | 0.16 | 10.04 | 0.17 |

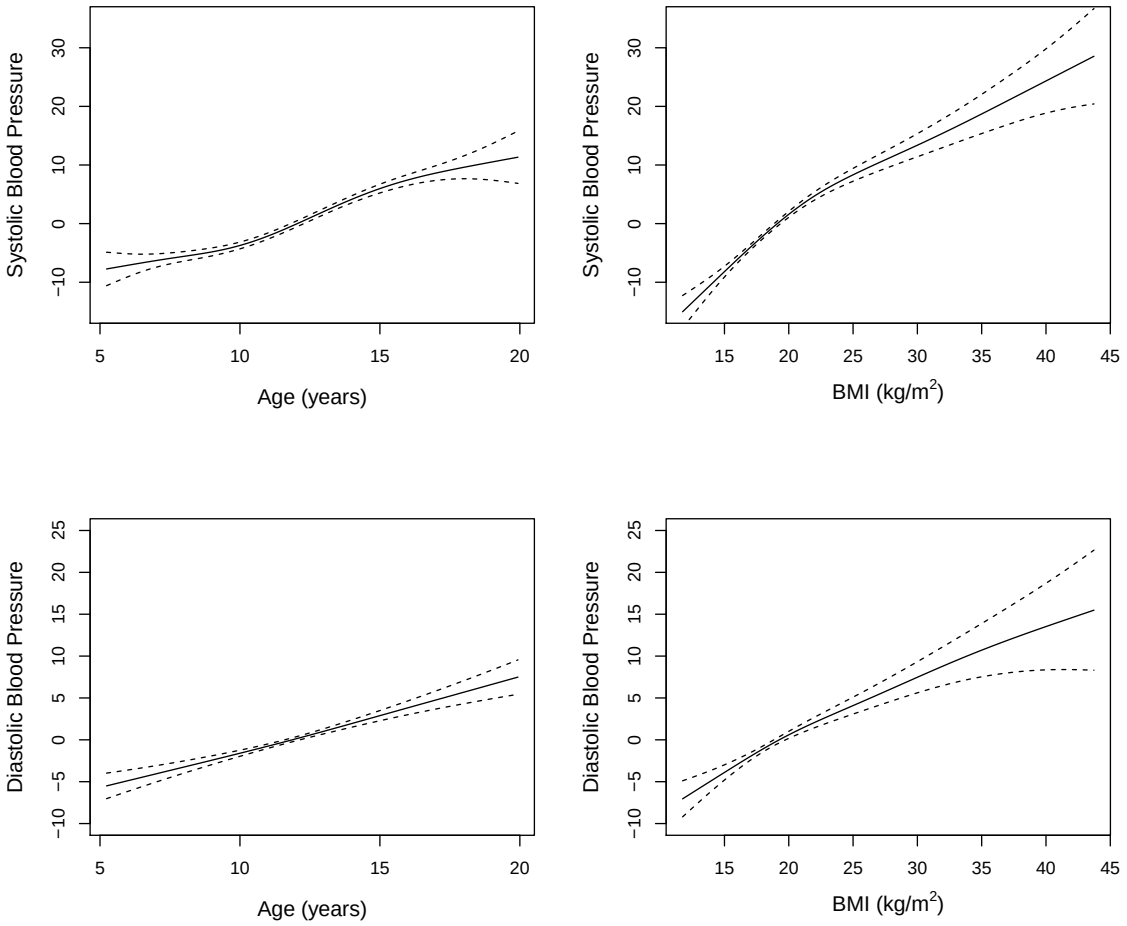


Figure 4.1: Marginal effects of age and BMI on systolic and diastolic blood pressure (subject to the centering constraint) (solid lines) with 95% confidence bands (dashed lines).

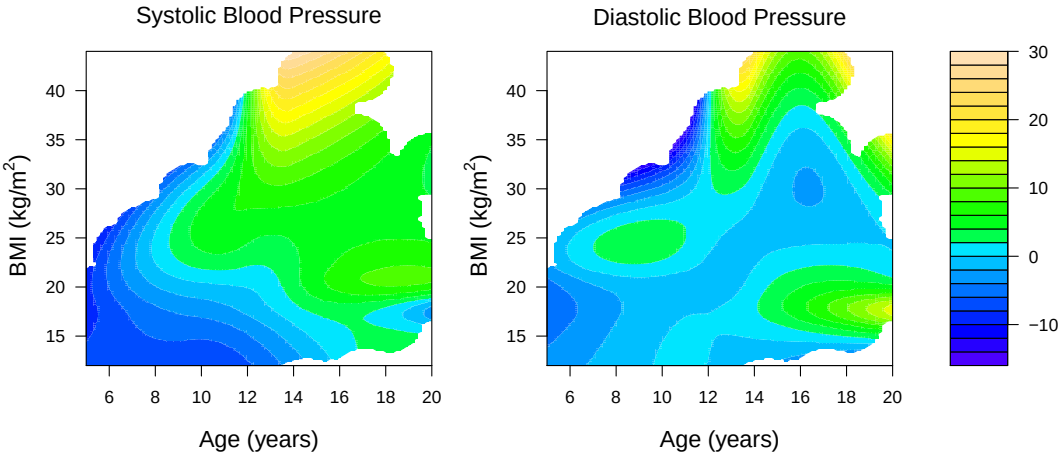


Figure 4.2: Estimated joint effects of age and BMI on systolic and diastolic blood pressure (subject to the centering constraint).

# Chapter 5

## Discussion

This dissertation describes a general class of regression models for analysis of longitudinally measured multiple outcome data. In this chapter, I briefly summarize the main methodological contributions of my work and reflect upon its practical impact.

First, this dissertation has presented the first generalized multivariate model that incorporates the nonlinear and possibly interacting influences of independent variables. The proposed model combines the strengths of univariate semiparametric models and those of multivariate linear models, to achieve much enhanced modeling flexibility. In the meantime, retaining the basic structure of the traditional mixed models allows for the use of existing inference and computational tools. For example, the inclusion of standard linear model components, such as the fixed effects and autoregressive terms, in additive forms, allows for traditional inferences on intervention efficacy, as expected in clinical trials. The mixed effects model representation of the proposed framework also makes it possible for model fitting by using existing computational software, thus greatly enhancing the practical applicability of the proposed methodology.

Second, the proposed model offers a way to accommodate complex correlation structures through a simple formulation. A central challenge to the analysis of longitudinal data is the incorporation of temporal dependency among repeated measures from the same individual. In multivariate data analysis, this difficulty is significantly magnified because of the presence of correlations among multiple outcomes at each time point of data collection. How to formulate a structure that is able to accommodate both temporal within-outcome dependency and cross-outcome correlations through a limited number of parameters becomes

a key challenge. In this research, an indirect approach has been taken to circumvent the difficulty of explicitly specifying a large and complex correlation matrix, by introducing a random effect vector and letting it be shared by multiple outcomes within the same individual. In the simplest form, this represents a case of the shared random intercept model, which contains a manageable number of parameters, resulting in a significant alleviation of the computational burden and assurance of model identifiability. Such a simple formulation, I contend, has nonetheless induced enough data interdependency both within an outcome and across the outcomes for a given individual.

Third, a unified modeling framework has been constructed, along with relevant parameter estimation and inference procedures for a very broad class of data following the exponential family of distributions. By following the tradition of generalized linear models in this regard, I have presented the proposed methodology for the exponential family, so that the model is maximally applicable to the most commonly encountered data distributions. Along the same line, the estimation and inference procedures are presented in the most general form to assure universal applicability.

Finally, I have developed a set of variable selection tools for practical data analysis. Variable selection, or more generally model selection, plays a fundamental role in scientific inquiry. Misspecification of the analytical model may introduce estimation bias, reduce analytical efficiency, and/or lead to erroneous inference. For complex models such as the ones presented here, model selection plays an additional role of determining the necessity of each model component. Herein, a penalized likelihood approach has been taken to achieve the goal of variable selection. This approach is very much in line with the current literature on variable selection. For instance, simultaneous selection of the fixed and random effects in traditional mixed models has been conducted by imposing penalties on the parameters. A similar penalization method has been used here with a few critical modifications. In this

research context, because multiple outcomes are incorporated in the same model through shared random effects, random effects selection helps determine the cross-outcome correlation structure. This is of particular importance in this modeling setting as it helps decide whether a multivariate modeling approach is justified. Another significant extension is the selection of interaction effects in the bivariate nonparametric components, which can help justify the inclusion of bivariate independent variable effects. The bias in parameter estimation introduced by the use of penalty terms is corrected through a two-stage algorithm to ensure both selection and estimation accuracy.

Multiple real data examples have been used to illustrate the application of the proposed methods. At the conclusion of this dissertation, I remain hopeful that the potential applicability of this new class of models will grow in time, when more analysts become familiar with the newly developed techniques. At the same time, future modifications and variations are going to be inevitable to meet the demand of specific analytical situations. Further extensions of the model to accommodate complex data distributions, such as zero-inflated counts or mixture distributions, and handling of missing data are all worthy objectives. Further improvement of the computational efficiency of the model selection procedure is also of great practical importance. In summary, there is no shortage of topics for future extension, which I take as a sign of methodological vitality. All things considered, I hope that the more widespread use of this modeling approach will stimulate new thoughts for its continued improvement in years to come.

# BIBLIOGRAPHY


Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Proceedings of the 2nd International Symposium Information Theory*, Budapest: Akademia Kiado, pp. 267–281.

Barndorff-Nielsen, O. E. and D. R. Cox (1989). *Asymptotic Techniques for Use in Statistics*. London: Chapman and Hall.

Bates, D. M. (2009). Assessing the precision of estimates of variance components. http://lme4.r-forge.r-project.org/slides/2009-07-21-Seewiesen/4PrecisionD.pdf.

Bates, D. M. (2010). lme4: Mixed-effects modeling with R. http://lme4.r-forge.r-project.org/book/.

Bernstein, G. R., C. A. Gaydos, M. Diener-West, M. R. Howell, J. M. Zenilman, and T. C. Quinn (1998). Incident *Chlamydia trachomatis* infections among inner-city adolescent females. *Journal of the American Medical Association 280*(6), 521–526.

Bondell, H. D., A. Krishna, and S. K. Ghosh (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics 66*(4), 1069–1077.

Brazier, J. E., R. Harper, N. M. Jones, A. O'Cathain, K. J. Thomas, T. Usherwood, and L. Westlake (1992). Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *British Medical Journal 305*(6846), 160–164.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*(421), 9–25.

Catalano, P. J. and L. M. Ryan (1992). Bivariate latent variable models for clustered discrete

and continuous outcomes. *Journal of the American Statistical Association 87*(419), 651–658.

Cates, Jr., W. (1999). Estimates of the incidence and prevalence of sexually transmitted diseases in the United States. *Sexually Transmitted Diseases 26*(4), S2–S7.

Centers for Disease Control and Prevention (2011). *Sexually Transmitted Disease Surveillance 2010*. Atlanta: U.S. Department of Health and Human Services.

Chen, Z. and D. B. Dunson (2003). Random effects selection in linear mixed models. *Biometrics 59*(4), 762–769.

Coull, B. A. and J. Staudenmayer (2004). Self-modeling regression for multivariate curve data. *Statistica Sinica 14*, 695–711.

Counsell, S. R., C. M. Callahan, D. O. Clark, W. Tu, A. B. Buttar, T. E. Stump, and G. D. Ricketts (2007). Geriatric care management for low-income seniors: a randomized controlled trial. *Journal of the American Medical Association 298*(22), 2623–2633.

Crainiceanu, C. M. and D. Ruppert (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*, 165–185.

Datta, S. D., M. Sternberg, R. E. Johnson, S. Berman, J. R. Papp, G. McQuillan, and H. Weinstock (2007). Gonorrhea and Chlamydia in the United States among persons 14 to 39 years of age, 1999 to 2002. *Annals of Internal Medicine 147*(2), 89–96.

Dicker, L. W., D. J. Mosure, S. M. Berman, W. C. Levine, and the Regional Infertility Prevention Program (2003). Gonorrhea prevalence and coinfection with Chlamydia in women in the United States, 2000. *Sexually Transmitted Diseases 30*(5), 572–576.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–499.

Ethier, K. A. and D. P. Orr (2007). *Behavioral Interventions for Prevention and Control of STDs Among Adolescents*. New York: Springer.

Faber, M. T., A. Nielsen, M. Nygård, P. Sparén, L. Tryggvadottir, B. T. Hansen, K.-L. Liaw, and S. K. Kjaer (2011). Genital Chlamydia, genital herpes, *Trichomonas vaginalis* and Gonorrhea prevalence, and risk factors among nearly 70,000 randomly selected women in 4 nordic countries. *Sexually Transmitted Diseases 38*(8), 727–734.

Falkner, B. and S. Gidding (2011). Childhood obesity and blood pressure: Back to the future? *Hypertension 58*(5), 754–755.

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

Fan, J. and R. Li (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association 99*(467), 710–723.

Fan, Y. and R. Li (2012). Variable selection in linear mixed effects models. *The Annals of Statistics 40*(4), 2043–2068.

Fieuws, S. and G. Verbeke (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics 62*(2), 424–431.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2004). *Applied Longitudinal Analysis*. New Jersey: John Wiley & Sons, Inc.

Fortenberry, J. D., E. J. Brizendine, B. P. Katz, K. K. Wools, M. J. Blythe, and D. P. Orr (1999). Subsequent sexually transmitted infections among adolescent women with genital infection due to *Chlamydia trachomatis*, *Neisseria gonorrhoeae*, or *Trichomonas vaginalis*. *Sexually Transmitted Diseases 26*(1), 26–32.

Funder, J. W. (2014). Sensitivity to aldosterone: Plasma levels are not the full story. *Hypertension 63*(6), 1168–1170.

Ghosh, P. and T. Hanson (2010). A semiparametric bayesian approach to multivariate longitudinal data. *Australian & New Zealand Journal of Statistics 52*(3), 275–288.

Ghosh, P. and W. Tu (2009). Assessing sexual attitudes and behaviors of young women: A joint model with nonlinear time effects, time varying covariates, and dropouts. *Journal of the American Statistical Association 104*(486), 474–485.

Goldstein, H. and J. Rasbash (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 159*(3), 505–513.

Gray, S. M. and R. Brookmeyer (1998). Estimating a treatment effect from multidimensional longitudinal data. *Biometrics 54*(3), 976–988.

Greven, S. and T. Kneib (2010). On the behavior of marginal and conditional AIC in linear mixed models. *Biometrika 97*(4), 773–789.

Härdle, W., S. Huet, E. Mammen, and S. Sperlich (2004). Bootstrap inference in semiparametric generalized additive models. *Econometric Theory 20*, 265–300.

Ibrahim, J. G., H. Zhu, R. I. Garcia, and R. Guo (2011). Fixed and random effects selection in mixed effects models. *Biometrics 67*(2), 495–503.

Kahn, R. H., D. J. Mosure, S. Blank, C. K. Kent, J. M. Chow, M. R. Boudov, J. Brock, S. Tulloch, and the Jail STD Prevalence Monitoring Project (2005). *Chlamydia trachomatis* and *Neisseria gonorrhoeae* prevalence and coinfection in adolescents entering selected US juvenile detention centers, 1997-2002. *Sexually Transmitted Diseases 32*(4), 255–259.

Keselman, H. J., J. Algina, R. K. Kowalchuk, and R. D. Wolfinger (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics - Simulation and Computation 27*(3), 591–604.

Khan, A., J. D. Fortenberry, B. E. Juliar, W. Tu, D. P. Orr, and B. E. Batteiger (2005). The prevalence of Chlamydia, Gonorrhea, and Trichomonas in sexual partnerships: Implications for partner notification and treatment. *Sexually Transmitted Diseases 32*(4), 260–264.

Kinney, S. K. and D. B. Dunson (2007). Fixed and random effects selection in linear and logistic models. *Biometrics 63*(3), 690–698.

Laird, N. M. and J. H. Ware (1982). Random-effects melection for longitudinal data. *Biometrics 38*(4), 963–974.

Lange, N. and N. M. Laird (1989). The effect of covariance structures on variance estimation in balance growth-curve models with random parameters. *Journal of the American Statistical Association 84*(405), 241–247.

Liang, H., H. Wu, and G. Zou (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika 95*(3), 773–778.

Liang, K. Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Lin, X. and R. J. Carroll (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association 96*(455), 1045–1056.

Liu, H. and W. Tu (2012). A semiparametric regression model for paired longitudinal outcomes with application in childhood blood pressure development. *The Annals of Applied Statistics 6*(4), 1861–1882.

Martin, A., W. Rief, A. Klaiberg, and E. Braehler (2006). Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. *General Hospital Psychiatry 28*(1), 71–77.

Meng, X. and D. B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika 80*(2), 267–278.

Miglioretti, D. L. (2003). Latent transition regression for mixed outcomes. *Biometrics 59*(3), 710–720.

Miller, W. C., H. Swygard, M. M. Hobbs, C. A. Ford, M. S. Handcock, M. Morris, J. L. Schmitz, M. S. Cohen, K. M. Harris, and J. R. Udry (2005). The prevalence of Trichomonas in young adults in the United States. *Sexually Transmitted Diseases 32*(10), 593–598.

Ng, E. S. W., J. R. Carpenter, H. Goldstein, and J. Rasbash (2006). Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. *Statistically Modelling 6*(1), 23–42.

Ni, X., D. Zhang, and H. H. Zhang (2010). Variable selection in semiparametric mixed models in longitudinal studies. *Biometrics 66*(1), 79–88.

O'Brien, L. M. and G. M. Fitzmaurice (2004). Analysis of longitudinal multiple-source binary data using generalized estimating equations. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 53*(1), 177–193.

Pinhero, J. C. and D. M. Bates (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics 4*(1), 12–35.

Pratt, J. H., J. J. Jones, J. Z. Miller, M. A. Wagner, and N. S. Fineberg (1989). Racial differences in aldosterone excretion and plasma aldosterone concentrations in children. *New England Journal of Medicine 321*(17), 1152–1157.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Reinsel, G. (1982). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association 77*(377), 190–195.

Roca-Pardiñas, J., C. Cadarso-Suárez, P. G. Tahoces, and M. J. Lado (2008). Assessing continuous bivariate effects among different groups through nonparametric regression models: An application to breast cancer detection. *Computational Statistics & Data Analysis 52*, 1958–1970.

Rochon, J. (1996). Analyzing bivariate repeated measures for discrete and continuous outcome variables. *Biometrics 52*(2), 740–750.

Rodriguez, G. and N. Goldman (1993). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 158*(1), 73–89.

Roy, J. and X. Lin (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics 56*(4), 1047–1054.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. New York, NY: Cambridge University Press.

Sammel, M. D. and L. M. Ryan (1996). Latent variable models with fixed effects. *Biometrics 52*(2), 650–663.

Sammel, M. D., L. M. Ryan, and J. M. Legler (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59*(3), 667–678.

Schall, N. (1991). Estimation in generalized linear models with random effects. *Biometrika 78*(4), 719–727.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Shah, A., N. Laird, and D. Schoenfeld (1997). A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association 92*(438), 775–779.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica 7*, 221–264.

Sutton, M., M. Sternburg, E. H. Koumans, G. McQuillan, S. Berman, and L. Markowitz (2007). The prevalence of *Trichomonas vaginalis* infection among reproductive-age women in the United States, 2001-2004. *Clinical Infectious Diseases 45*(10), 1319–1326.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58*(1), 267–288.

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine 16*(4), 385–395.

Tu, W., B. E. Batteiger, S. Wiehe, S. Ofner, B. Van Der Pol, B. P. Katz, D. P. Orr, and J. D. Fortenberry (2009). Time from first intercourse to first sexually transmitted infection diagnosis among adolescent women. *Archives of Pediatrics & Adolescent Medicine 163*(12), 1106–1111.

Tu, W., G. J. Eckert, T. S. Hannon, H. Liu, L. M. Pratt, M. A. Wagner, L. A. DiMeglio, J. J., and J. H. Pratt (2014). Racial differences in sensitivity of blood pressure to aldosterone. *Hypertension 63*(6), 1212–1218.

Tu, W., G. J. Eckert, J. H. Pratt, and A. H. Jan Danser (2012). Plasma levels of prorenin and renin in blacks and whites: their relative abundance and associations with plasma aldosterone concentration. *American Journal of Hypertension 25*(9), 1030–1034.

Tu, W., P. Ghosh, and B. P. Katz (2011). A stochastic model for assessing *Chlamydia trachomatis* transmission risk using longitudinal observational data. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 174*(4), 975–989.

Wang, H., B. Li, and C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Biometrika 94*(3), 553–568.

Wang, H., R. Li, and C. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*(3), 553–568.

Weinstock, H., S. Berman, and W. Cates, Jr. (2004). Sexually transmitted diseases among american youth: Incidence and prevalence estimates, 2000. *Perspectives on Sexual and Reproductive Health 36*(1), 6–10.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(1), 95–114.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.

Wood, S. N. (2011). *gamm4: Generalized additive mixed models using mgcv and lme4*. R package version 0.1-5.

Yu, Z., G. J. Eckert, H. Liu, J. H. Pratt, and W. Tu (2013). Adiposity has unique influence on the renin-aldosterone axis and blood pressure in black children. *Journal of Pediatrics 163*(5), 1317–1322.

Yu, Z., X. Lin, and W. Tu (2012). Semiparametric frailty models for clustered failure time data. *Biometrics 68*(2), 429–436.

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

Zhang, D., X. Lin, J. Raz, and M. Sowers (1992). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association 93*(442), 710–719.

Zhang, D. W. and X. H. Lin (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics 4*(1), 57–74.

Zhang, H. H., G. Cheng, and Y. Liu (2011). Linear or nonlinear? automatic structure discovery for partially linear models. *Journal of the American Statistical Association 106*(495), 1099–1112.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

# CURRICULUM VITAE

Zhuokai Li

## EDUCATION

- Ph.D. in Biostatistics, Minor in Epidemiology, Indiana University, Indianapolis, IN, 2014
- M.S. in Mathematics - Applied Statistics, Purdue University, Indianapolis, IN, 2013
- B.S. in Biological Sciences, Fudan University, Shanghai, China, 2008

## WORKING EXPERIENCE

- Research Assistant, Department of Biostatistics, Indiana University, Indianapolis, IN, Jan. 2012 - present
- Research Assistant, Center for Computational Biology and Bioinformatics, Indiana University, IN, Aug. 2010 - Dec. 2011
- Teaching Assistant, Department of Biostatistics, Indiana University, Indianapolis, IN, Aug. 2010 - Dec. 2010

## HONORS AND AWARDS

- Student Travel Award, 10th International Conference on Health Policy Statistics, Chicago, IL, 2013
- Outstanding Advanced Graduate Student Award, Purdue University School of Science, Indianapolis, IN, 2013
- Fellowship, Purdue University School of Science, Indianapolis, IN, 2009 - 2010
- Merit Scholarships, Fudan University, Shanghai, China, 2005 - 2008
- Excellent Student Leader, Fudan University, Shanghai, China, 2006

SELECTED PUBLICATIONS

- Li Z, Liu H, Tu W. (In revision). A Sexually Transmitted Infection Screening Algorithm Based on Multivariate Semiparametric Regression Models. *Statistics in Medicine*.
- Hannon TS, Gupta S, Li Z, Eckert GJ, Carroll AE, Pratt JH, Tu W. (In press). The Effect of Body Mass Index on Blood Pressure Varies by Race among Obese Children. *Journal of Pediatric Endocrinology and Metabolism*.
- Hannon TS, Li Z, Tu W, Huber JN, Carroll AE, Lagges AM, Gupta S. (2014). Depressive Symptoms are Associated with Fasting Insulin Resistance in Obese Youth. *Pediatric Obesity*. doi: 10.1111/ijpo.237.
- Watson SE, Li Z, Tu W, Jalou H, Brubaker JL, Gupta S, Huber JN, Carroll AE, Hannon TS. (2013). Obstructive Sleep Apnea in Obese Adolescents and Cardiometabolic Risk Markers. *Pediatric Obesity*. doi: 10.1111/j.2047-6310.2013.00198.x.