

Nonparametric Regression Models for the Analysis of Longitudinal Data

Colin O. Wu^a, Xin Tian^b

Office of Biostatistics Research

National Heart Lung and Blood Institute, Bethesda, MD 20892, U.S.A

^awuc@nhlbi.nih.gov, ^btianx@nhlbi.nih.gov

and

Kai F. Yu

Department of Mathematical Sciences and Mathematical Sciences Center

Tsinghua University, Beijing 100084, P.R.China

kfyu@math.tsinghua.edu.cn

1 Introduction

1.1 Scientific Objectives of Longitudinal Studies

In biomedical studies interests are often focused on evaluating the effects of treatments, medication dosage, risk factors or other biological or environmental covariates on the outcomes of interest, such as disease progression and change of health status, over time. Because the changes of outcomes and covariates over time within each subject usually provide important information of scientific relevance, longitudinal samples that contain repeated measurements of the subjects over time are often more preferable than the classical cross-sectional samples. By combining the characteristics of random sampling and time series observations, the usefulness of longitudinal samples goes far beyond biomedical studies, and their trace is often found in economics, psychology, sociology and many other fields of natural and social sciences.

In general, there are two main approaches for obtaining longitudinal observations in biomedical studies: (a) clinical trials with repeated measurements, and (b) epidemiological studies, which are often referred as observational cohort studies. The main difference between a clinical trial and an observational cohort study is at their designs. In a clinical trial,

the selection of subjects, randomized study treatments, length of the trial period, visiting times and methods of the measurement process are determined by the study investigators, although, in some occasions, non-randomized concomitant treatments or interventions may also be given to some subjects due to ethical and logistical reasons. An observational cohort study, on the other hand, is more complicated, because the risk factors, treatments and the measurement process depend on the participants of the study and are not controlled by the investigators.

In a longitudinal clinical trial, the main scientific objective is to evaluate the efficacy of the experimental treatments versus placebo or standard treatment on the primary outcomes, such as certain health status indicators, over time during the trial period. In many situations, a follow-up period is added at the end of the treatment period, so that time-to-event variables, such as time to hospitalization or death, may be included as a primary outcome in addition to the repeatedly measured health outcomes. In a particular analysis, the trial period may be defined based on the objectives of the analysis. For example, if the objective is to evaluate the treatment effects on the time-trend of a health indicator within the treatment period, it is appropriate to consider the treatment period as the trial period. On the other hand, if it is also of interest to consider certain time-to-event variables beyond the treatment period, it is then appropriate to include both the treatment period and the follow-up period into the trial period. Effects of the study treatments may be evaluated through the conditional means, conditional distributions or conditional quantiles of the outcome variables. Although regression models based on conditional means of the outcome variables are by far the most popular methods used in the analysis of longitudinal clinical trials, regression methods based on conditional distributions or conditional quantiles are often valuable statistical tools in situations when the outcome variables have highly skewed or other distributions that are not easily approximated by normality. In addition to the evaluation of randomized study treatments, an important secondary objective is to evaluate the effects of concomitant interventions or other covariates on the time-varying trends of the outcome variables. Regression analyses involving covariates other than the randomized study treatments are often useful for evaluating treatment-covariate interactions or identifying sub-groups of patient populations to whom the experimental treatments are effective.

In a longitudinal observational cohort study, there are no randomized experimental treatments to be tested, and the main objective is to evaluate the potential associations

of various covariates, such as demographic and environmental factors, with the outcome variables of interest and their trends over time. Observational cohort studies are often used for the purpose of data exploration, so that more specific scientific hypotheses may be generated and tested in a future properly designed clinical trial. Because the variables are repeatedly measured over time, long-term longitudinal observational cohort studies are useful for understanding the natural progression of certain diseases both on a population-wide level and for certain subsets of subjects. In practice, an observational cohort study usually involves a large number of subjects with sufficient numbers of repeated measurements over time, so that novel findings with adequate statistical accuracy can be obtained from the study. Similar to longitudinal clinical trials, the choices of statistical approaches for the analysis of data from observational cohort studies depend on the scientific objectives, and may involve regression models for the conditional means and conditional distributions.

1.2 Structures of Longitudinal Data

For a typical framework of longitudinal data, we define t to be a real-valued variable of time, $Y(t)$ a real-valued outcome variable and $\mathbf{X}(t) = (X^{(0)}(t), \dots, X^{(K)}(t))^T$, $K \geq 1$, a R^{K+1} -valued covariate vector at time t . Depending on the choice of origin, the time variable t is not necessarily non-negative. As part of the general methodology, interest of statistical analysis with regression models is often focused on modeling and determining the effects of $\{t, \mathbf{X}(t)\}$ on the population means, subject-specific deviations from the population means, conditional distributions or conditional quantiles of $Y(t)$. For n randomly selected subjects and each subject repeatedly measured over time, the longitudinal sample of $\{Y(t), t, \mathbf{X}(t)\}$ is denoted by $\{(Y_{ij}, t_{ij}, \mathbf{X}_{ij}); i = 1, \dots, n, j = 1, \dots, n_i\}$, where t_{ij} is the j th measurement time of the i th subject, Y_{ij} and $\mathbf{X}_{ij} = (X_{ij}^{(0)}, \dots, X_{ij}^{(D)})^T$ are the observed outcome and covariate vector, respectively, of the i th subject at time t_{ij} , and n_i is the i th subject's number of repeated measurements. The total number of measurements for the study is $N = \sum_{i=1}^n n_i$. In contrast to the independent identically distributed (i.i.d.) samples in classical cross-sectional studies, the measurements within each subject are possibly correlated, although the inter-subject measurements are assumed to be independent.

A longitudinal sample is said to have a balanced design if all the subjects have their measurements made at a common set of time points, i.e. $n_i = m$ for some $m \geq 1$ and all $i = 1, \dots, n$ and $t_{1j} = \dots = t_{nj}$ for all $j = 1, \dots, m$. An unbalanced design arises if the

design time points $\mathbf{t}_i = \{t_{ij}; 1 \leq j \leq n_i\}$ are different for different subjects. In practice, unbalanced designs may be caused by the presence of missing observations in an otherwise balanced design or by the random variations of the time design points. In long-term clinical trials or epidemiological studies, study subjects are often assigned to a set of prespecified “design visiting times”, but their actual visiting times could be different from the “design visiting times” because of missing visits or changing visiting times due to various reasons. It is possible to observe balanced longitudinal data from a well-controlled longitudinal clinical trial, because the randomized study treatments and clinical visiting times are determined by the study investigators. However, for various reasons that are out of the investigators’ control, most observational cohort studies have unbalanced longitudinal designs.

1.3 Examples of Longitudinal Studies

These real-life examples, three epidemiological studies and a longitudinal clinical trial, illustrate some typical features and scientific objectives of longitudinal studies. Although these examples share some similarities in study designs, they have different data structures and objectives.

Example 1. Alabama Small-for-Gestational-Age Study (ASGA)

This is a prospective study of risk factors and intrauterine growth retardation involving 1475 women who had their fetal anthropometry measurements made by ultrasound repeatedly during pregnancy. All women were scheduled to have their measurements made at approximately 17, 25, 31 and 36 weeks of gestation. In the dataset, however, their actual clinical visit times do not exactly follow this schedule and are scattered between 12 to 43 weeks of gestation. The numbers of visits range between 1 and 7 during pregnancy. This results in an unbalanced design that not all the subjects have measurements at the same time design points. Associated covariates that may affect fetal development include maternal behavioral factors, such as cigarette smoking, alcohol consumption and drug abuse, and maternal anthropometric measurements, such as pre-pregnancy weight, height and body mass index, and placental development measured by placental thickness at different stages of gestation. Some of these covariates, such as the maternal anthropometric measurements are time-dependent. But, the others, such as maternal behavioral factors, may be either time-dependent or time-invariant, depending on how these variables are defined in an analysis. The outcome variables of fetal development, which, of course, are all time-dependent,

include fetal abdominal circumference, biparietal diameter, femur length and other ultrasound measurements.

Figure 1: Relationship between fetal abdominal circumference (in centimeters) and fetal gestation (in weeks). Left Panel: individual measurement; Right Panel: sequences of measurements for some randomly selected subjects.

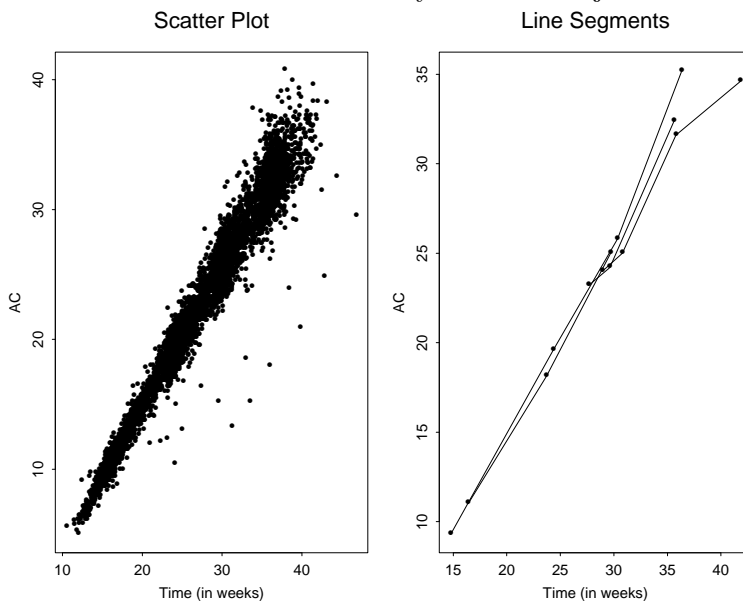


Figure 1 shows the observed fetal abdominal circumferences (in centimeters) at their corresponding gestational age in weeks. To see the temporal trend of the individual repeated measurements, the line segments indicate the measurement sequences for three randomly selected subjects. Heuristically, we observe a linear upward trend on the growth of fetal size. But, there has been no prior study which justifies the goodness of a linear growth model for this population or any other statistical model on the covariate effects on fetal growth. A statistical analysis should then focus on two objectives: establishing an appropriate statistical model so that the effects of these covariates on the outcome of interest can be clearly interpreted; developing estimation and inference procedures to adequately quantify the covariate effects based on the chosen statistical models. \square

Example 2. Baltimore Multicenter AIDS Cohort Study (BMACS)

This dataset is from the Baltimore site of the Multicenter AIDS Cohort Study (MACS), and includes 400 homosexual men who were infected by the human immunodeficiency virus

(HIV) between 1984 and 1991. Because CD4 cells (T-helper lymphocytes) are vital for immune function, an important component of the study is to evaluate the effects of risk factors, such as cigarette smoking and drug use, and health status, such as CD4 cell levels before the infection, on the post-infection depletion of CD4 percent (CD4 percent of lymphocyte cells). Although all the individuals were scheduled to have their measurements made at semi-annual visits, the study has an unbalanced design because the subjects' actual visiting times did not exactly follow the schedule and the HIV infections happened randomly during the study. The numbers of repeated measurements range from 1 to 14 with a median of approximately 6. Compared with the ASGA Study of Example 1, this dataset has a smaller number of subjects and a wider range of repeated measurements. The covariates of interest in these data also include both time-dependent and time-invariant variables. Further details of the statistical design and scientific importance of the MACS data can be found in Kaslow *et al.* (1987). \square

Example 3. National Growth and Health Study (NGHS)

This is a 3-center population-based cohort study, which has a total of 2379 girls, among them 1213 (51%) African-Americans and 1166 (49%) Caucasians, who were 9 or 10 years old at enrollment and had their height, weight, blood pressure (BP) and other anthropometric and laboratory measurements recorded once a year until they reached 19 years old. The main objective of the study is to evaluate the patterns of obesity and cardiovascular risks during childhood and adolescence. The three centers, University of California, Berkeley, University of Cincinnati/Cincinnati Children's Hospital Medical Center, and Westat, Rockville, Maryland, were selected on the basis of census tract data to include a wide distribution of household income and parental education within each race. Observations were collected at clinical centers or home visits between 1986 and 1997 based on a standard protocol (NGHSRG, 1992), which contained the scientific objectives, justifications, study design and methods of measurements.

Major findings of the study have been reported, for example, in Daniels *et al.* (1998), Kimm *et al.* (2000, 2001, 2002), Thompson *et al.* (2007) and Obarzanek *et al.* (2010). Among these results, Daniels *et al.* (1998) studied the longitudinal mean effects of race, height and body mass index (BMI, measured in kg/m^2) on the levels of systolic blood pressure (SBP) and diastolic blood pressure (DBP). Kimm *et al.* (2000, 2001, 2002) studied

the effects of race on obesity, and the relationship between the changes in physical activity and BMI. Thompson et al. (2007) investigated the associations between childhood overweight and cardiovascular disease risk factors such as hypertension and unhealthy levels of low-density lipoprotein cholesterol (LDL) and triglyceride (TG), and Obarzanek et al. (2010) investigated the effects of race, obesity, and a range of socioeconomic and nutritional variables on the prevalence and incidence rates of hypertension and prehypertension as defined in NHBPEP Working Group (2004). The statistical methods used by these authors depend on summarizing the effects of childhood obesity and other variables on the cardiovascular risk variables through their conditional-means or the probabilities of unhealthy risk levels specified by certain threshold values. Although these results present some useful snapshots of the harmful effects of childhood obesity on cardiovascular risks, a further question is whether similar associations between childhood obesity and cardiovascular risks still hold if other definitions of unhealthy risk levels are used. It is desirable to have a flexible conditional-distribution based regression method so that the covariate effects on the entire distribution of a risk variable are described. This study gives a typical example that motivates the development of longitudinal estimation and inference methods based on conditional distributions and conditional quantiles, instead of the usual regression methods based on conditional means. \square

Example 4. Enhancing Recovery in Coronary Heart Disease (ENRICHD) Study

This study is a randomized clinical trial that evaluated the efficacy of a cognitive behavior therapy (CBT) versus usual cardiological care (UC) on survival and depression severity in 2481 patients who had depression and/or low perceived social support after acute myocardial infarction. Details of the study design, objectives, and major findings of the trial have been described in ENRICHD (2001, 2003). The primary objective of the study is to determine whether mortality and recurrent myocardial infarction are reduced by treatment of depression and low perceived social support with cognitive behavior therapy, supplemented with the use of selective serotonin reuptake inhibitor (SSRI) or other antidepressants as needed, in patients enrolled within 28 days after myocardial infarction (MI). The intervention of the trial consists of cognitive behavior therapy initiated at a median of 17 days after the index of MI for a median of 11 individual sessions throughout 6 months. Depression severity was measured by Hamilton Rating Scale for Depression (HRSD) and

Beck Depression Inventory (BDI) with higher HRSD and BDI scores indicating worsened depression severity. Group therapy was conducted when feasible, with antidepressants, such as SSRIs, as a pharmacotherapy for patients scoring higher than 24 on the HRSD or having a less than 50% reduction in BDI scores after 5 weeks. Antidepressants were also prescribed at the requests of the patients or their primary-care physicians, therefore, could be treated as a concomitant treatment in addition to the randomized CBT psychosocial treatment or usual cardiological care specified in the trial design. The main outcome measures consist of a composite primary end point of death or recurrent MI, and secondary outcomes including change in HRSD and BDI for depression or the ENRICHD Social Support Instrument scores for low perceived social support at 6 months (ENRICHD, 2001 and 2003).

In addition to the primary objective of the trial, an important question of secondary objective is whether the use of antidepressants has added benefits for reducing mortality and recurrent MI and lowering the BDI scores for this patient population. To evaluate the effects of pharmacotherapy with the use of antidepressants, Taylor et al. (2005) compared the survival rates for death and cardiovascular morbidity and mortality among 1834 depressed patients in this trial and found that the use of SSRIs seemed to reduce subsequent cardiovascular morbidity and mortality. Bang and Robins (2005) analyzed the same data but only with a cross-sectional component. By treating the patient’s starting time of pharmacotherapy as a subject-specific “change-point time” and evaluating the effects of pharmacotherapy on the patient’s BDI scores over time using different model patterns before and after the “change-point”, Wu, Tian and Bang (2008) and Wu, Tian and Jiang (2011) showed that the use of SSRIs was beneficial for lowering the patient’s depression severity measured by the BDI scores.

This study is used throughout Section 5 as a motivating example for the development of appropriate regression models and their estimation and inference procedures for situations when some of the time-dependent covariates are adaptive depending on the outcome trajectories at previous time points. In such situations, where some time-dependent covariates are adaptive to the previous outcome trajectories, the commonly used parametric or nonparametric mixed-effects models may be misspecified and lead to biased and erroneous conclusions. \square

1.4 Regression Models for Longitudinal Analysis

Generally speaking, a proper longitudinal analysis should achieve at least three objectives:

- (1) The model under consideration must give an adequate description of the scientific relevance of the data and be sufficiently simple and flexible to be practically implemented. In biomedical studies, an appropriate regression model should give a clear and meaningful biological interpretation and also has a simple mathematical structure.
- (2) The methodology must contain proper model diagnostic tools to evaluate the validity of a statistical model for a given data set. Two important diagnostic methods are confidence regions and tests of statistical hypotheses.
- (3) The methodology must have appropriate theoretical and practical properties, and can adequately handle the possible intra-subject correlations of the data. In practice, the intra-subject correlations are often completely unknown and difficult to be adequately estimated, so that it is generally preferred to use estimation and inference procedures that do not depend on modeling the specific correlation structures.

1.4.1 Parametric Models

Naturally, the most commonly used approach in longitudinal analysis is through parametric regression models, such as the generalized linear and nonlinear mixed-effects models. The simplest case of these models is the marginal linear model of the form

$$Y_{ij} = \sum_{l=0}^K \beta_l X_{ij}^{(l)} + \epsilon_i(t_{ij}), \quad (1.1)$$

where β_0, \dots, β_K are constant linear coefficients describing the effects of the corresponding covariates, $\epsilon_i(t)$ are realizations of a mean zero stochastic process $\epsilon(t)$ at t and \mathbf{X}_{ij} and $\epsilon_i(t_{ij})$ are independent. Similar to all regression models where a constant intercept term is desired, we set $X^{(0)} \equiv 1$, which produces a baseline coefficient β_0 , representing the mean value of $Y(t)$ when all the covariates $X^{(l)}(t)$ are set to zero. A popular special case of the error process is to take $\epsilon(t)$ to be a mean zero Gaussian stationary process. Although (1.1) appears to be overly simplified for many practical situations, its generalizations lead to many useful models which form the bulk of longitudinal analysis.

Estimation and inference methods based on parametric models, including the weighted least squares, the quasi-likelihoods and the generalized estimating equations, have been

extensively investigated in the literature. The details can be found, for example, in Laird and Ware (1982), Pantula and Pollock (1985), Ware (1985), Liang and Zeger (1986), Diggle (1988), Zeger, Liang and Albert (1988), Jones and Ackerson (1990), Jones and Boadi-Boteng (1991), Davidian and Giltinan (1995), Vonesh and Chinchilli (1997), Verbeke and Molenberghs (2000) and Diggle, Heagerty, Liang and Zeger (2002). The main advantage of parametric models is that they generally have simple and intuitive interpretations. User friendly computer programs are already available in most popular statistical software packages, such as SAS, STATA and R. However, these models suffer the potential shortfall of model misspecification, which may lead to erroneous conclusions. At least in exploratory studies, it is necessary to relax some of the parametric restrictions.

1.4.2 Semiparametric Models

A useful semiparametric model, investigated by Zeger and Diggle (1994) and Moyeed and Diggle (1994), is the partially linear model of the form

$$Y_{ij} = \beta_0(t_{ij}) + \sum_{l=1}^K \beta_l X_{ij}^{(l)} + \epsilon_i(t_{ij}), \quad (1.2)$$

where $\beta_0(t)$ is an unknown smooth function of t , β_l are unknown constants and $\epsilon_i(t)$ and \mathbf{X}_{ij} are defined in (1.1). This model is more general than the marginal linear model (1.1), because $\beta_0(t)$ is allowed to change with t , rather than setting to be a constant over time. On the other hand, by including the linear terms of $X_{ij}^{(l)}$, (1.2) is also more general than the nonparametric regression studied by Hart and Wehrly (1986), Altman (1990), Hart (1991), Rice and Silverman (1991), among others, which involves only (t_{ij}, Y_{ij}) . However, because (1.2) describes the effects of $X_{ij}^{(l)}$ on Y_{ij} through constant linear coefficients, this model is still, to some degree, based on mathematical convenience rather than scientific relevance. For example, there is no reason to expect that the influences of maternal risk factors on fetal development in the ASGA Study or the effects of cigarette smoking and pre-infection CD4 level on the post-infection CD4 cell percent in the HIV/CD4 Depletion Data are linear and constant throughout the study period. Thus, further generalization of (1.2) is needed in many situations.

1.4.3 Nonparametric Models

Although it is possible in principle to model $\{Y_{ij}, t_{ij}, \mathbf{X}_{ij}\}$ through a completely nonparametric high dimensional function, such approach is often impractical due to the well-known

problem of “curse of dimensionality”. Moreover, numerical results obtained from high dimensional nonparametric fittings are often difficult to interpret. These problems motivate the consideration of nonparametric models that have certain meaningful structures. An important class of structural nonparametric regression models is the varying-coefficient models of the form

$$Y_{ij} = \mathbf{X}_{ij}^T \beta(t_{ij}) + \epsilon_i(t_{ij}), \quad (1.3)$$

where $\beta(t) = (\beta_0(t), \dots, \beta_K(t))^T$ is a $(K + 1)$ -vector of smooth functions of t and $\epsilon_i(t)$ and \mathbf{X}_{ij} are defined as in (1.1). Because (1.3) gives a linear model between $Y(t)$ and $\mathbf{X}(t)$ at each fixed t , the linear coefficients $\beta_l(t)$, $l = 0, \dots, K$, can be interpreted the same way as in (1.2). Taking $X_{ij}^{(0)} \equiv 1$, $\beta_0(t)$ represents the intercept at time t . On the other hand, because all the linear coefficients may change with t , we may obtain different linear models at different time points. Model (1.3) is a special case of the general varying-coefficient models discussed by Hastie and Tibshirani (1993).

Methods of estimation and inferences based on this class of models have been subjected to intense investigation in the literature. A number of different smoothing methods for the estimation of $\beta(t)$ have been proposed. These include the ordinary least squares local polynomials, the penalized least squares, the two-step and componentwise methods and the basis approximation approaches. Targeted to specific types of longitudinal designs, each of these methods has its own advantages and disadvantages in practice. We will present in Section 3.2 and Section 3.3 an overview of the above estimation and inference methods and demonstrate in Section 3.4 the applications of these methods.

1.4.4 Regression Models for Conditional Distributions

In addition to the regression models described above, which are mostly based on modeling the conditional means of the repeatedly measured outcome variables, regression models based on conditional distributions are often used in longitudinal analysis when the response variables are not normally distributed. Well-known approaches with non-Gaussian data, such as the generalized linear models (e.g., Molenberghs and Verbeke, 2005), are often used for discrete outcome variables or the discretized versions of continuous outcome variables defined by some prespecified threshold values. In some situations when the outcome variable has non-Gaussian conditional distributions, a known transformation, such as the Box-Cox transformation, may be applied so that the conditional distribution of the transformed

variable is approximately normal. In many situations, however, there are no appropriate threshold values or transformations for the outcome variables being considered, and the scientific objectives are better studied by directly modeling the conditional distributions.

Nonparametric estimation of conditional distribution functions for independent identically distributed data and certain time series samples without incorporating any modeling structures has been studied by Hall, Wolff and Yao (1999) based on two kernel-based smoothing methods, namely the local logistic estimation method and the adjusted Nadaraya-Watson method. Since the estimation methods based on unstructured nonparametric models could be numerically unstable when the number of covariates is large and the corresponding results are often difficult to interpret, Wu, Tian and Yu (2010) proposed a dimension-reduction strategy by extending the linear transformation modeling structures studied by Cheng, Wei and Ying (1995, 1997) to longitudinal data, and studied a time-varying linear transformation model with a two-step local polynomial method for the estimation of covariate effects and conditional distributions. As a result, the time-varying transformation models proposed by Wu, Tian and Yu (2010) give a flexible and convenient structural nonparametric framework for characterizing the conditional distributions of the outcome variables, when the conditional-mean based regression models are insufficient for the scientific objectives of the study. Applying this conditional-distribution based structural nonparametric modeling approach to the NGHS data of Example 3, the statistical estimates, predictions and their inferences presented in Section 4 yield useful insights on the temporal trends of blood pressure for children and adolescents.

2 An Overview of Parametric and Semiparametric Methods

2.1 Linear Mixed-Effects Models

As a popular regression approach for modeling the covariate effects on the longitudinal outcome variables, the mixed-effects models generally serve two purposes: (i) describing the effects of the treatments and other factors on the mean response profile; (ii) describing the differences in response profiles between individuals. A regression model serving the first purpose is generally classified as a marginal model or a population average model. A regression model serving the second purpose is a random effects model or a subject specified model (e.g. Zeger, Liang and Albert, 1988). A mixed effects model then combines both the marginal and random effects. In particular, a linear mixed effects model is obtained when

the marginal and random effects are additive and follow a linear relationship.

It is convenient to describe the model through a matrix representation. Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$ be the $[n_i \times 1]$ vector of the response for the i th subject, $t_i = (t_{i1}, \dots, t_{in_i})^T$ be the subject's time design points and \mathbf{X}_i be the corresponding $[n_i \times (k + 1)]$ covariate matrix whose j th row, for $j = 1, \dots, n_i$, is $(1, X_{ij}^{(1)}, \dots, X_{ij}^{(k)})$. Assuming that the error term $\epsilon_i(t)$ of (1.1) is a mean zero Gaussian process with covariate matrix $\mathbf{V}_i(t_i)$, the responses \mathbf{Y}_i are then independent Gaussian random vectors such that

$$\mathbf{Y}_i \sim \mathbf{N}(\mathbf{X}_i\beta, \mathbf{V}_i(t_i)), \quad (2.1)$$

where $\beta = (\beta_0, \dots, \beta_k)^T$ with β_j being defined in (1.1) and $\mathbf{N}(\mathbf{a}, \mathbf{b})$ denotes a multivariate normal distribution with mean vector \mathbf{a} and covariance matrix \mathbf{b} . Note that, because (2.1) represents the conditional mean of \mathbf{Y}_i at \mathbf{X}_i through $\mathbf{X}_i\beta$, it is a marginal model.

The covariance structures of (2.1) are usually influenced by three factors: random effect, serial correlation and measurement error. The random effects characterize the stochastic variations between subjects within the population. In particular, we may view that, when the covariates affect the response linearly, some of the linear coefficients may vary from subject to subject. The serial correlations are the results of time-varying associations between different measurements of the same subject. Such correlations are typically positive in biomedical studies, and become weaker as the time interval between the measurements increases. Finally, the measurement errors, which are normally assumed to be independent both between and within the subjects, are induced by the measurement process or random variations within the subjects.

Suppose that, for each subject i , there is a $[r \times 1]$ vector of explanatory variables \mathbf{U}_{ij} measured at time t_{ij} , which may or may not overlap with the original covariate vector \mathbf{X}_{ij} . Using the additive decomposition of random effects, serial correlations and measurement errors, $\epsilon_i(t_{ij})$ can be expressed as

$$\epsilon_i(t_{ij}) = \mathbf{U}_{ij}^T \mathbf{b}_i + W_i(t_{ij}) + Z_{ij}, \quad (2.2)$$

where \mathbf{b}_i is the $[r \times 1]$ random vector with multivariate normal distribution $\mathbf{N}(\mathbf{0}, \mathbf{D})$, \mathbf{D} is a $[r \times r]$ covariate matrix with (p, q) th element $d_{pq} = d_{qp}$, $W_i(t_{ij})$ for $i = 1, \dots, n$ are independent copies of a mean zero Gaussian process whose covariance at time points t_{ij_1} and t_{ij_2} is $\rho_W(t_{ij_1}, t_{ij_2})$, and Z_{ij} for $i = 1, \dots, n$ and $j = 1, \dots, n_i$ are independent identically

distributed random variables with $N(0, \tau^2)$ distribution. Writing $\delta_i(t_{ij}) = W_i(t_{ij}) + Z_{ij}$, $\delta_i = (\delta_i(t_{i1}), \dots, \delta_i(t_{in_i}))^T$ and \mathbf{U}_i the $[n_i \times r]$ matrix whose j th row is \mathbf{U}_{ij}^T , (2.1) and (2.2) reduce to the linear mixed effects model of Laird and Ware (1982)

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{U}_i b_i + \delta_i. \quad (2.3)$$

The marginal effect β represents the influence of \mathbf{X}_i on the population average of the response profile, while b_i describes the variation of the i th subject from the population conditioning on the given explanatory variable \mathbf{U}_i . Thus, conditioning on \mathbf{X}_i and \mathbf{U}_i , (2.3) implies that \mathbf{Y}_i for $i = 1, \dots, n$ are independent Gaussian vectors such that

$$\mathbf{Y}_i \sim \mathbf{N}(\mathbf{X}_i\beta, \mathbf{U}_i\mathbf{D}\mathbf{U}_i^T + \mathbf{P}_i + \tau^2\mathbf{I}_i), \quad (2.4)$$

where \mathbf{P}_i is the $[n_i \times n_i]$ covariance matrix whose (j_1, j_2) th element is $\rho_W(t_{ij_1}, t_{ij_2})$ and \mathbf{I}_i is the $[n_i \times n_i]$ identity matrix.

A number of special cases can be derived for the variance-covariance structure of (2.2). The classical linear models for the independent cross-sectional data (or the independent identically distributed data) is a special case of (2.4) where $\epsilon_i(t_{ij})$ are only affected by the measurement errors Z_{ij} . When neither random effects nor measurement errors are present, the error term is of pure serial correlation $\epsilon_i(t_{ij}) = W_i(t_{ij})$. Moreover, if $W_i(t_{ij})$ are from a mean zero stationary Gaussian process, the covariance of $\epsilon_i(t_{ij_1})$ and $\epsilon_i(t_{ij_2})$, hence, Y_{ij_1} and Y_{ij_2} , can be specified by

$$\text{Cov}(\epsilon_i(t_{ij_1}), \epsilon_i(t_{ij_2})) = \sigma^2 \rho(|t_{ij_1} - t_{ij_2}|), \quad (2.5)$$

where σ is a positive constant and $\rho(\cdot)$ is a continuous function. Useful choices of $\rho(\cdot)$ include the exponential correlation $\rho(s) = \exp(-as)$ for some constant $a > 0$ and the Gaussian correlation $\rho(s) = \exp(-as^2)$, among others. When $\epsilon_i(t_{ij})$ are affected by a mean zero stationary Gaussian process and a mean zero Gaussian white noise (measurement error), the variance of Y_{ij} is $\sigma^2\rho(0) + \tau^2$, while the covariance of Y_{ij_1} and Y_{ij_2} , for $j_1 \neq j_2$, is $\sigma^2\rho(|t_{ij_1} - t_{ij_2}|)$, for some $\sigma > 0$, $\tau > 0$ and continuous correlation function $\rho(\cdot)$. When serial correlations are not present, the intra-subject correlations are only induced by the random effects, so that \mathbf{P}_i is not present in (2.4).

2.2 Likelihood Based Estimation and Inferences

2.2.1 Conditional Maximum Likelihood Estimation

Suppose that the variance-covariance matrix $\mathbf{V}_i(t_i)$ of (2.2) is determined by a R^q -valued parameter vector α . Denote $\mathbf{V}_i(t_i; \alpha)$ to be the variance-covariance matrix parametrized by α . The log-likelihood function for (2.1) is

$$L(\beta, \alpha) = c + \sum_{i=1}^n \left\{ -\frac{1}{2} \log |\mathbf{V}_i(t_i; \alpha)| - \frac{1}{2} (\mathbf{Y}_i - \mathbf{X}_i \beta)^T \mathbf{V}_i^{-1}(t_i; \alpha) (\mathbf{Y}_i - \mathbf{X}_i \beta) \right\}, \quad (2.6)$$

where $c = \sum_{i=1}^n [(-n_i/2) \log(2\pi)]$. For a given α , (2.6) can be maximized by

$$\hat{\beta}(\alpha) = \left[\sum_{i=1}^n (\mathbf{X}_i^T \mathbf{V}_i^{-1}(t_i; \alpha) \mathbf{X}_i) \right]^{-1} \left[\sum_{i=1}^n (\mathbf{X}_i^T \mathbf{V}_i^{-1}(t_i; \alpha) \mathbf{Y}_i) \right]. \quad (2.7)$$

It is easy to verify that, under (2.1), $\hat{\beta}(\alpha)$ is an unbiased estimator of β . Direct calculation also shows that the covariance matrix of $\hat{\beta}(\alpha)$ is

$$\begin{aligned} & \text{Cov} \left[\hat{\beta}(\alpha) \right] \\ &= \left[\sum_{i=1}^n (\mathbf{X}_i^T \mathbf{V}_i^{-1}(t_i; \alpha) \mathbf{X}_i) \right]^{-1} \left[\sum_{i=1}^n (\mathbf{X}_i^T \mathbf{V}_i^{-1}(t_i; \alpha) \text{Cov}(\mathbf{Y}_i) \mathbf{V}_i^{-1}(t_i; \alpha) \mathbf{X}_i) \right] \\ & \quad \times \left[\sum_{i=1}^n (\mathbf{X}_i^T \mathbf{V}_i^{-1}(t_i; \alpha) \mathbf{X}_i) \right]^{-1} \\ &= \left[\sum_{i=1}^n (\mathbf{X}_i^T \mathbf{V}_i^{-1}(t_i; \alpha) \mathbf{X}_i) \right]^{-1}. \end{aligned} \quad (2.8)$$

It is interesting to note that the second equality sign of (2.8) does not hold when the structure of the variance-covariance matrix is not correctly specified. Further derivation using (2.1), (2.7) and (2.8) shows that $\hat{\beta}(\alpha)$ has a multivariate Normal distribution,

$$\hat{\beta}(\alpha) \sim \mathbf{N} \left\{ \beta, \left[\sum_{i=1}^n (\mathbf{X}_i^T \mathbf{V}_i^{-1}(t_i; \alpha) \mathbf{X}_i) \right]^{-1} \right\}. \quad (2.9)$$

When α is known, this result can be used to develop inference procedures, such as confidence regions and test statistics, for β .

2.2.2 Maximum Likelihood Estimation

When α is unknown, as in most practical situations, a consistent estimate of α has to be used. An intuitive approach is to estimate β and α by maximizing (2.6) with respect to β

and α simultaneously. Maximum likelihood estimators (MLE) of this type can be computed by substituting (2.7) into (2.6) and then maximizing (2.6) with respect to α . Denote the resulting ML estimators by $\widehat{\beta}_{ML}$ and $\widehat{\alpha}_{ML}$. The asymptotic distributions of $(\widehat{\beta}_{ML}, \widehat{\alpha}_{ML})$ can be developed using the standard approaches in large sample theory.

Although $(\widehat{\beta}_{ML}, \widehat{\alpha}_{ML})$ has some justifiable statistical properties, as for most likelihood-based methods, it may not be desirable in practice. To see why an alternative estimation method might be warranted in some situations, we consider the simple linear regression with independent errors and $n_1 = \dots = n_n = m$,

$$\mathbf{Y}_i \sim \mathbf{N}(\mathbf{X}_i\beta, \sigma^2\mathbf{I}_m), \quad (2.10)$$

where \mathbf{I}_m is the $[m \times m]$ identity matrix. The parameters involved in the model are β and σ . Let $\widehat{\beta}_{ML}$ and $\widehat{\sigma}_{ML}$ be the MLEs of β and σ , respectively, and RSS be the residual sum of squares defined by

$$\text{RSS} = \sum_{i=1}^n (\mathbf{Y}_i - \mathbf{X}_i\widehat{\beta}_{ML})^T (\mathbf{Y}_i - \mathbf{X}_i\widehat{\beta}_{ML}).$$

The MLE of σ^2 is $\widehat{\sigma}_{ML}^2 = \text{RSS}/(nm)$. However, it is well-known that, for any finite n and m , $\widehat{\sigma}_{ML}^2$ is a biased estimator of σ^2 . On the other hand, a slightly modified estimator $\widehat{\sigma}_{REML}^2 = \text{RSS}/[nm - (k + 1)]$ is unbiased for σ^2 . Here, $\widehat{\sigma}_{REML}^2$ is the restricted maximum likelihood (REML) estimator for the model (2.10).

2.2.3 Restricted Maximum Likelihood Estimation

This class of estimators was introduced by Patterson and Thompson (1971) for the purpose of estimating variance components in the linear models. The main idea is to consider a linear transformation of the original response variable so that the distribution of the transformed variable does not depend on β . Let $\mathbf{Y} = (\mathbf{Y}_1^T, \dots, \mathbf{Y}_n^T)^T$, $\mathbf{X} = (\mathbf{X}_1^T, \dots, \mathbf{X}_n^T)^T$ and \mathbf{V} be the block-diagonal matrix with $\mathbf{V}_i(t_i)$ on the i th main diagonal and zeros elsewhere. Then, with \mathbf{V} parameterized by α , model (2.1) is equivalent to

$$\mathbf{Y} \sim \mathbf{N}(\mathbf{X}\beta, \mathbf{V}(\alpha)). \quad (2.11)$$

The REML estimator of α , the variance component of (2.11), is obtained by maximizing the likelihood function of $\mathbf{Y}^* = \mathbf{A}^T\mathbf{Y}$, where \mathbf{A} is a $[N \times (N - k - 1)]$, $N = \sum_{i=1}^n n_i$, full rank matrix such that $\mathbf{A}^T\mathbf{X} = \mathbf{0}$. A specific construction of \mathbf{A} can be found in Diggle,

Heagerty, Liang and Zeger (2002, Section 4.5). It follows from (2.11) that \mathbf{Y}^* has a mean zero multivariate Gaussian distribution with covariance matrix $\mathbf{A}^T \mathbf{V}(\alpha) \mathbf{A}$. Harville (1974) showed that the likelihood function of \mathbf{Y}^* is proportional to

$$L^*(\alpha) = \left| \sum_{i=1}^n \mathbf{X}_i^T \mathbf{X}_i \right|^{1/2} \left| \sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1}(t_i; \alpha) \mathbf{X}_i \right|^{-1/2} \left\{ \prod_{i=1}^n |\mathbf{V}_i(t_i; \alpha)|^{-1/2} \right\} \quad (2.12)$$

$$\times \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta}(\alpha))^T \mathbf{V}_i^{-1}(t_i; \alpha) (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta}(\alpha)) \right\}.$$

The REML estimator $\hat{\alpha}_{REML}$ of α maximizes (2.12). The REML estimator $\hat{\beta}_{REML}$ of β is obtained by substituting α of (2.7) with $\hat{\alpha}_{REML}$. Because (2.12) does not depend on the choice of \mathbf{A} , the resulting estimators $\hat{\beta}_{REML}$ and $\hat{\alpha}_{REML}$ are free of the specific linear transformations.

The log-likelihood of \mathbf{Y}^* , $\log[L^*(\alpha)]$, differs from the log-likelihood $L(\hat{\beta}, \alpha)$ only through a constant, which does not depend on α , and

$$-\frac{1}{2} \log \left| \sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1}(t_i; \alpha) \mathbf{X}_i \right|,$$

which does not depend on β . Because both REML and ML methods are based on the likelihood principle, they all have important theoretical properties such as consistency, asymptotic normality and asymptotic efficiency. In practice, neither one is uniformly superior to the other for all the situations. Their numerical values are also computed from different algorithms. For the ML method, the fixed effects and the variance components are estimated simultaneously, while for the REML method, only the variance components are estimated.

2.2.4 Inferences

The results established in the previous sections are useful to construct inference procedures for β . For the purpose of illustration, only a few special cases are presented here. A more complete account of inferential and diagnostic tools may be found in Diggle (1988), Zeger, Liang and Albert (1988), Diggle, Heagerty, Liang and Zeger (2002) or Vonesh and Chinchilli (1997), among others.

Suppose that there is a consistent estimator $\hat{\alpha}$ of α , which may be either the ML estimator $\hat{\alpha}_{ML}$ or the REML estimator $\hat{\alpha}_{REML}$. Substituting α of (2.9) with $\hat{\alpha}$, the distribution of $\hat{\beta}(\hat{\alpha})$ can be approximated, at least when n is large, by

$$\hat{\beta}(\hat{\alpha}) \sim \mathbf{N}(\beta, \hat{\mathbf{V}}), \quad (2.13)$$

where $\widehat{\mathbf{V}} = [\sum_{i=1}^n (\mathbf{X}_i^T \mathbf{V}_i^{-1}(t_i; \widehat{\alpha}) \mathbf{X}_i)]^{-1}$. Suppose that \mathbf{C} is a known $[r \times (k + 1)]$ matrix with full rank. It follows immediately from (2.13) that, when n is sufficiently large, the distribution of $\mathbf{C}\widehat{\beta}(\widehat{\alpha})$ can be approximated by

$$\mathbf{C}\widehat{\beta}(\widehat{\alpha}) \sim \mathbf{N}(\mathbf{C}\beta, \mathbf{C}\widehat{\mathbf{V}}\mathbf{C}^T). \quad (2.14)$$

Consequently, an approximate $[100 \times (1 - a)]\%$, $0 < a < 1$, confidence interval for $\mathbf{C}\beta$ can be given by

$$\mathbf{C}\widehat{\beta}(\widehat{\alpha}) \pm Z_{1-a/2} (\mathbf{C}\widehat{\mathbf{V}}\mathbf{C}^T)^{1/2}.$$

Taking \mathbf{C} to be the $(k + 1)$ row vector with 1 at its l th place and zero elsewhere, an approximate $[100 \times (1 - a)]\%$ confidence interval for β_l can be given by

$$\widehat{\beta}_l(\widehat{\alpha}) \pm Z_{1-a/2} \sqrt{\widehat{V}_l}, \quad (2.15)$$

where \widehat{V}_l is the l th diagonal element of $\widehat{\mathbf{V}}$.

The approximation in (2.14) can also be used to construct test statistics for linear statistical hypotheses. For example, suppose that we would like to test the null hypothesis of $\mathbf{C}\beta = \theta_0$ for a known vector θ_0 against the general alternative that $\mathbf{C}\beta \neq \theta_0$. A natural test statistic would be

$$\widehat{T} = [\mathbf{C}\widehat{\beta}(\widehat{\alpha}) - \theta_0]^T (\mathbf{C}\widehat{\mathbf{V}}\mathbf{C}^T)^{-1} [\mathbf{C}\widehat{\beta}(\widehat{\alpha}) - \theta_0], \quad (2.16)$$

which has approximately a χ^2 -distribution with r degrees of freedom, denoted by χ_r^2 , under the null hypothesis. A level $(100 \times a)\%$ test based on \widehat{T} then rejects the null hypothesis when $\widehat{T} > \chi_r^2(a)$ with $\chi_r^2(a)$ being the $[100 \times (1 - a)]$ th percentile of χ_r^2 . For the special case of testing $\beta_l = 0$ versus $\beta_l \neq 0$, a simple procedure equivalent to (2.16) is to reject the null hypothesis when

$$|\widehat{\beta}_l(\widehat{\alpha})| > Z_{1-a/2} \sqrt{\widehat{V}_l},$$

where $Z_{1-a/2}$ and \widehat{V}_l are defined in (2.15).

2.3 Partially Linear Models

As discussed in Section 1.3.3, this class of models has been studied by Zeger and Diggle (1994) and Moyeed and Diggle (1994) as a means to generalize the marginal linear models. With further restrictions on the error process, (1.2) is equivalent to

$$Y(t) = \beta_0(t) + \sum_{l=1}^D \beta_l X^{(l)}(t) + \epsilon(t), \quad (2.17)$$

where $\epsilon(t)$ is a mean zero stochastic process with variance σ^2 and correlation function $\rho(t)$, and $X^{(l)}(t)$, $l = 1, \dots, D$, and $\epsilon(t)$ are independent. The errors $\epsilon_i(t_{ij})$ specified in (1.2) are then independent copies of $\epsilon(t)$. A useful way to view $\epsilon_i(t_{ij})$ is through the decomposition

$$\epsilon_i(t_{ij}) = W_i(t_{ij}) + Z_{ij}, \quad (2.18)$$

where $W_i(t)$ are independent copies of a mean zero stationary process $W(t)$ with covariance function $\sigma_W^2 \rho(t)$ and Z_{ij} are independent identically distributed measurement errors with mean zero and variance σ_Z^2 . The covariance structure of the measurements Y_{ij} for $i = 1, \dots, n$ and $j = 1, \dots, n_i$ are

$$\text{Cov}(Y_{i_1 j_1}, Y_{i_2 j_2}) = \begin{cases} \sigma_Z^2 + \sigma_W^2, & \text{if } i_1 = i_2 \text{ and } j_1 = j_2, \\ \sigma_W^2 \rho(t_{i_1 j_1} - t_{i_2 j_2}), & \text{if } i_1 = i_2 \text{ and } j_1 \neq j_2, \\ 0, & \text{otherwise.} \end{cases} \quad (2.19)$$

Although the above models can be classified as a special case of (1.3), a class of the structural nonparametric models to be discussed in later sections, their estimation methods are quite different, a fact owing to the structural differences between these two classes of models. The rest of this section focuses on an iteration procedure for the estimation of $\beta_0(t), \beta_1, \dots, \beta_k$. Inferential and alternative estimation methods, which constitute some major research activities in longitudinal analyses, are still not well-understood and warrant considerable effort in further investigation.

2.4 Unstructured Smoothing Methods for Mean Responses

Suppose for the moment that no covariate other than time is considered in modeling the mean response. The model (2.17) then reduces to

$$Y(t) = \beta_0(t) + \epsilon(t). \quad (2.20)$$

Equivalently, with $\epsilon(t)$ defined in (2.17), $\beta_0(t)$ is the mean response of $Y(t)$ conditioning on time t ; that is, $\beta_0(t) = E[Y(t)|t]$.

A natural approach for estimating $\beta_0(t)$ nonparametrically is to borrow smoothing techniques from the classical independent identically distributed (i.i.d.) setting, while evaluating the statistical performances of the resulting estimators by taking the influences of the intra-subject correlations into account. A simple method is to use kernel smoothing, which amounts to estimate $\beta_0(t)$ through a weighted average using the measurements obtained within a neighborhood of t defined by a kernel function. Let $K(u)$ be a continuous

kernel function, usually a continuous probability density function, defined on the real line, and h a positive bandwidth sequence which shrinks to zero as n tends to infinity. A kernel estimator similar to the well-known Nadaraya-Watson type kernel estimators in the i.i.d. setting is

$$\widehat{\beta}_0^K(t) = \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} \{Y_{ij} K[(t - t_{ij})/h]\}}{\sum_{i=1}^n \sum_{j=1}^{n_i} K[(t - t_{ij})/h]}. \quad (2.21)$$

Here, (2.21) uses uniform weight on each measurement, hence, makes no distinction between the subjects that have unequal numbers of repeated measurements. Thus subjects with more repeated measurements are used more often than those with fewer repeated measurements. A general formulation is to assign a specific weight to each subject and estimate $\beta_0(t)$ by

$$\widehat{\beta}_0^K(t; w) = \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} \{Y_{ij} w_i K[(t - t_{ij})/h]\}}{\sum_{i=1}^n \sum_{j=1}^{n_i} \{w_i K[(t - t_{ij})/h]\}}, \quad (2.22)$$

where the weights, $w = (w_1, \dots, w_n)$, satisfy $w_i \geq 0$ for all $i = 1, \dots, n$ with strict inequality for some $1 \leq i \leq n$. Clearly, (2.22) reduces to (2.21) when $w_i = 1/N$. An intuitive weight choice other than $w_i = 1/N$ is to uniformly weight each subject, rather than each measurement, so that the resulting kernel estimator is (2.22) with $w_i = 1/(nn_i)$.

Other approaches for the estimation of (2.19) have also been studied by Hart and Wehrly (1986), Müller (1988), Altman (1990), Hart (1991), Rice and Silverman (1991), among others. These methods are not discussed here, and their details can be found in these original articles. These methods, including (2.22) and the above alternative approaches, are essentially based on the fundamental spirit of local smoothing, hence, often lead to similar results in practice. This is in contrast to the smoothing methods to be discussed in the next section, where, because of the model complexity, different smoothing methods often produce very different results.

A crucial step in obtaining an adequate kernel estimator for $\beta_0(t)$ is to select an appropriate bandwidth h , while the choices of kernel functions are relatively less important. For estimation methods other than kernel smoothing, such as splines, this amounts to selecting an appropriate smoothing parameter. Rice and Silverman (1991) suggested a simple cross-validation for selecting a data-driven smoothing parameter which does not depend on the intra-subject correlation structures of the data. Applying their cross-validation to the kernel estimator (2.22), we first define $\widehat{\beta}_0^{(-i,K)}(t; w)$ to be the estimator computed using (2.22) and the remaining data after deleting the entire set of repeated measurements of the i th subject. Predicting the i th subject's outcome at time t by $\widehat{\beta}_0^{(-i,K)}(t; w)$, the cross-validation

score of (2.22) is

$$\text{CV}(h) = \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ w_i \left[Y_{ij} - \widehat{\beta}_0^{(-i,K)}(t_{ij}; w) \right]^2 \right\}. \quad (2.23)$$

Suppose that (2.23) can be uniquely minimized. The “leave-one-subject-out” cross-validated bandwidth h_{cv} is the minimizer of (2.23). Heuristically, the use of h_{cv} can be justified because, by minimizing (2.23), it approximately minimizes an average prediction error of (2.22). More details for the implementations and generalizations of this cross-validation will be discussed in Section 4.7.

Direct calculation of (2.23) can often be time consuming, as the algorithm repeats itself each time a new subject is deleted. Denote $K_{ij} = K[(t - t_{ij})/h]$,

$$K_{ij}^* = \frac{w_i K[(t - t_{ij})/h]}{\sum_{i=1}^n \sum_{j=1}^{n_i} w_i K[(t - t_{ij})/h]} \quad \text{and} \quad K_i^* = \sum_{j=1}^{n_i} K_{ij}^*$$

for $i = 1, \dots, n$. A computationally simpler approach, also suggested by Rice and Silverman (1991), is to compute $[Y_{ij} - \widehat{\beta}_0^{(-i,K)}(t_{ij}; w)]$ using the following expression:

$$\begin{aligned} & Y_{ij} - \widehat{\beta}_0^{(-i,K)}(t_{ij}; w) \\ &= Y_{ij} - \left[\widehat{\beta}_0^K(t_{ij}; w) - \sum_{j=1}^{n_i} (Y_{ij} K_{ij}^*) \right] \left(1 + \frac{K_i^*}{1 - K_i^*} \right) \\ &= [Y_{ij} - \widehat{\beta}_0^K(t_{ij}; w)] + \sum_{j=1}^{n_i} (Y_{ij} K_{ij}^*) - \left[\widehat{\beta}_0^K(t_{ij}; w) - \sum_{j=1}^{n_i} (Y_{ij} K_{ij}^*) \right] \left(\frac{K_i^*}{1 - K_i^*} \right) \\ &= [Y_{ij} - \widehat{\beta}_0^K(t_{ij}; w)] + \left(\frac{K_i^*}{1 - K_i^*} \right) \left[\frac{\sum_{j=1}^{n_i} (Y_{ij} K_{ij}^*)}{K_i^*} - \widehat{\beta}_0^K(t_{ij}; w) \right]. \end{aligned} \quad (2.24)$$

The above expression, as currently stated, is specifically targeted to kernel estimators defined in (2.22). When other smoothing methods, such as splines, are used, we may not get an explicit expression as the right side of (2.24), hence, direct calculation of (2.23) has to be carried out by deleting the subjects one at a time.

Large sample inferences of $\widehat{\beta}_0^K(t; w)$ can be derived based on the asymptotic expressions of its means and variances and its asymptotic distributions. Because $\widehat{\beta}_0^K(t; w)$ is a linear statistic of Y_{ij} , its means and variances can be directly computed and, consequently, its asymptotic distributions can be easily established by checking the triangular array central limit theorem after taking the intra-subject correlations into account; see, for example, Wu, Chiang and Hoover (1998) and Wu and Chiang (2000). Because $\widehat{\beta}_0^K(t; w)$ is a special case

of the kernel estimators of Section 4, details of pointwise and simultaneous inferences for $\beta_0(t)$ are discussed in Section 5.1.

2.5 Estimation of Covariate Effects in Partially Linear Models

With covariates other than time entered into the model, the estimation of $(\beta_0(t), \beta_1, \dots, \beta_k)$ can be proceeded by an iteration that combines smoothing with parametric estimation techniques. Suppose that the error terms $\epsilon_i(t)$ of (2.18) have known variance-covariance matrices $\mathbf{V}_i(t_i)$ for $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})$ and all $i = 1, \dots, n$. The iteration can be proceeded as follows:

- (a) Set $\beta_0(t)$ to zero and calculate an initial estimate of $(\beta_1, \dots, \beta_k)^T$ using (2.7), an expression also for the generalized least squares, with $\mathbf{V}_i(\mathbf{t}_i; \alpha)$ replaced by $\mathbf{V}_i(\mathbf{t}_i)$.
- (b) Based on the current estimate $(\hat{\beta}_1, \dots, \hat{\beta}_k)$, calculate the residual $r_{ij} = Y_{ij} - \sum_{l=1}^k \hat{\beta}_l X_{ij}^{(l)}$ and compute the kernel estimator $\hat{\beta}_0^K(t; w)$ of $\beta_0(t)$ using (2.21) with Y_{ij} replaced by r_{ij} .
- (c) Based on the current kernel estimator $\hat{\beta}_0^K(t; w)$, calculate the residual $r_{ij} = Y_{ij} - \hat{\beta}_0^K(t_{ij}; w)$ and update the estimate of $(\beta_1, \dots, \beta_k)$ using (2.7) with $(\mathbf{V}_i(\mathbf{t}_i; \alpha), Y_{ij})$ replaced by $(\mathbf{V}_i(\mathbf{t}_i), r_{ij})$.
- (d) Repeat steps (b) and (c) until the estimates converge.

This algorithm is a special case of the more general backfitting algorithm described in Hastie and Tibshirani (1993).

The assumption of having a known correlation structure is unrealistic and can be relaxed. Although an incorrectly specified correlation structure may cost the efficiency of the estimators, it generally does not affect the consistency. When the variance-covariance matrix is parametrized by a parameter α and the error terms are from a mean zero Gaussian stationary process, the above iteration algorithm can be used in conjunction with the likelihood and restricted likelihood methods of the previous section, i.e. the generalized least squares estimators used in Steps (a) and (c) can be replaced by the likelihood based estimators $\hat{\beta}_{ML}$ or $\hat{\beta}(\hat{\alpha}_{REML})$. Further computational details, statistical properties of the resulting estimators and a modified estimation procedure can be found in Zeger and Diggle (1994) and Moyeed and Diggle (1994). Inferences based on the resulting estimators have not been systematically investigated, hence, warrant substantial further development.

3 Nonparametric Time-Varying Coefficient Models

This section presents a series of different smoothing methods for estimating the coefficient curves $\beta(t) = (\beta_0(t), \dots, \beta_k(t))^T$ of (1.2) and some asymptotic and bootstrap inference procedures based on the smoothing estimators of $\beta(t)$. Applications of the estimation and inference procedures are illustrate through the Alabama Small-for-Gestational-Age Study (ASGA) of Example 1 and the Baltimore Multicenter AIDS Cohort Study (BMACS) of Example 2.

3.1 Some Useful Expressions

In observational studies, the covariates are usually random as the subjects are randomly chosen, although they could in principle be either random or fixed. For generality, it is assumed throughout that $\mathbf{X}(t)$ is random and the matrix $E[\mathbf{X}(t)\mathbf{X}^T(t)] \equiv E_{\mathbf{X}\mathbf{X}^T}(t)$ exist. With a proper change of the notation, the methods here can be modified to accommodate the case of nonrandom covariates. An equivalent expression of (1.3) is then

$$Y(t) = \mathbf{X}^T(t)\beta(t) + \epsilon(t), \quad (4.1)$$

where $\epsilon(t)$ is a mean zero stochastic process and $\epsilon(t)$ and $\mathbf{X}(t)$ are independent. Suppose that $E_{\mathbf{X}\mathbf{X}^T}(t)$ is invertible and its inverse is $E_{\mathbf{X}\mathbf{X}^T}^{-1}(t)$. It directly follows from (4.1) that $\beta(t)$ uniquely minimizes the second moment of $\epsilon(t)$ in the sense that

$$E \left\{ \left[Y(t) - \mathbf{X}^T(t)\beta(t) \right]^2 \right\} = \inf_{\text{all } b(\cdot)} E \left\{ \left[Y(t) - \mathbf{X}^T(t)b(t) \right]^2 \right\}, \quad (4.2)$$

and is given by

$$\beta(t) = E_{\mathbf{X}\mathbf{X}^T}^{-1}(t)E[\mathbf{X}(t)Y(t)]. \quad (4.3)$$

When the covariates are time-invariant, we have $\mathbf{X}(t) \equiv \mathbf{X}$ and $E_{\mathbf{X}\mathbf{X}^T}(t) \equiv E_{\mathbf{X}\mathbf{X}^T}$, so that the equation (4.3) reduces to

$$\beta_r(t) = E \left[\left(\sum_{l=0}^k e_{rl}X^{(l)} \right) Y(t) \right], \quad (4.4)$$

where e_{rl} is the element of $E_{\mathbf{X}\mathbf{X}^T}^{-1}$ at the r th row and l th column.

3.2 Smoothing Based on Least Squares

3.2.1 General Formulation

Intuitively, (4.2) suggests that $\beta(t)$ can be estimated by a method of local least squares using the measurements observed within a neighborhood of t . Assume that, for each l and some integer $p \geq 0$, $\beta_l(t)$ is p times differentiable and its p th derivative is continuous. Approximating $\beta_l(t_{ij})$ by a p th order polynomial $\sum_{r=0}^p \{b_{lr}(t)(t_{ij} - t)^r\}$ for all $l = 0, \dots, k$, a local polynomial estimator of $\beta(t) = (\beta_0(t), \dots, \beta_k(t))^T$ based on a kernel neighborhood is $\widehat{b}_0(t) = (\widehat{b}_{00}(t), \dots, \widehat{b}_{k0}(t))^T$, where $\{\widehat{b}_{lr}(t); l = 0, \dots, k, r = 0, \dots, p\}$ minimizes

$$L_p(t) = \sum_{i=1}^n \sum_{j=1}^{n_i} w_i \left\{ Y_{ij} - \sum_{l=0}^k \left[X_{ij}^{(l)} \left(\sum_{r=0}^p b_{lr}(t)(t_{ij} - t)^r \right) \right] \right\}^2 K \left(\frac{t_{ij} - t}{h} \right), \quad (4.5)$$

where w_i are the non-negative weights as in (2.22), $K(\cdot)$ is a kernel function, usually chosen to be a probability density function, and h is a non-negative bandwidth. As a by-product of (4.5), $(r!) \widehat{b}_{lr}(t)$ may be used to estimate the r th derivative $\beta_l^{(r)}(t)$ of $\beta_l(t)$, $r = 1, \dots, p$.

3.2.2 Least Squares Kernel Estimators

The simplest case of (4.5) is the ordinary least squares kernel estimator, also known as the local constant fit, obtained by minimizing (4.5) with $p = 0$. Using the matrix representation $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$,

$$\mathbf{X}_i = \begin{pmatrix} 1 & X_{i1}^{(1)} & \cdots & X_{i1}^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{in_i}^{(1)} & \cdots & X_{in_i}^{(k)} \end{pmatrix} \quad \text{and} \quad \mathbf{K}_i(t) = \begin{pmatrix} K_{i1} & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & K_{in_i} \end{pmatrix}$$

with $K_{ij} = K[(t_{ij} - t)/h]$, if $\sum_{i=1}^n \mathbf{X}_i^T \mathbf{K}_i(t) \mathbf{X}_i$ is invertible, then (4.5) with $p = 0$ can be uniquely minimized and its minimizer, the kernel estimator of $\beta(t)$, is given by

$$\widehat{\beta}^{LSK}(t) = \left(\sum_{i=1}^n w_i \mathbf{X}_i^T \mathbf{K}_i(t) \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^n w_i \mathbf{X}_i^T \mathbf{K}_i(t) \mathbf{Y}_i \right). \quad (4.6)$$

When the model incorporates no covariate other than time, i.e. $k = 0$, (4.6) reduces to a Nadaraya-Watson type kernel estimator of the conditional expectation $E[Y(t)|t]$; see, for example, Härdle (1990).

3.2.3 Least Squares Local Linear Estimators

Although (4.6) has a simple mathematical expression, it often leads to significant bias when t is at the boundary of its support. An automatic procedure to reduce such boundary bias

is to use higher order local polynomial fits. But, a high order local polynomial fit can be impractical in some applications because it usually requires large sample sizes and may be computationally intensive. A practical approach that provides automatic boundary bias adjustment is to use local linear fit that minimizes (4.5) with $p = 1$. Denote

$$\mathcal{N}_{lr} = \begin{pmatrix} \sum_{i,j} \left[w_i X_{ij}^{(l)} X_{ij}^{(r)} K_{ij} \right] & \sum_{i,j} \left[w_i X_{ij}^{(l)} X_{ij}^{(r)} (t_{ij} - t) K_{ij} \right] \\ \sum_{i,j} \left[w_i X_{ij}^{(l)} X_{ij}^{(r)} (t_{ij} - t) K_{ij} \right] & \sum_{i,j} \left[w_i X_{ij}^{(l)} X_{ij}^{(r)} (t_{ij} - t)^2 K_{ij} \right] \end{pmatrix},$$

$$\mathcal{N}_r = (\mathcal{N}_{0r}, \dots, \mathcal{N}_{kr}), \mathcal{N} = (\mathcal{N}_0^T, \dots, \mathcal{N}_k^T)^T,$$

$$\mathcal{M}_r = \left(\sum_{i,j} \left[w_i X_{ij}^{(r)} Y_{ij} K_{ij} \right], \sum_{i,j} \left[w_i X_{ij}^{(r)} (t_{ij} - t) Y_{ij} K_{ij} \right] \right)^T,$$

$\mathcal{M} = (\mathcal{M}_0^T, \dots, \mathcal{M}_k^T)^T$, $b_l(t) = (b_{l0}(t), b_{l1}(t))^T$ and $b(t) = (b_0(t), \dots, b_k(t))^T$ for $r, l = 0, \dots, k$. Setting the partial derivatives of $L_1(t)$ with respect to $b_{lr}(t)$ to zero, the normal equation of (4.5) with $p = 1$ is

$$\mathcal{N}b(t) = \mathcal{M}. \tag{4.7}$$

Suppose that the matrix \mathcal{N} is invertible at t . The solution of (4.7) exists and is uniquely given by $\hat{b}(t) = \mathcal{N}^{-1}\mathcal{M}$. The least squares local linear estimator $\hat{\beta}_l^{LSL}(t)$ of $\beta_l(t)$ is then

$$\hat{\beta}_l^{LSL}(t) = e_{2l+1}^T \hat{b}(t), \tag{4.8}$$

where e_q is the $[2(k+1) \times 1]$ column vector with 1 at its q th place and zero elsewhere. Explicit expressions for the general higher order least squares local polynomial estimators can be similarly derived; see Hoover *et al.* (1998). Details of these general higher order estimators are omitted, since a local linear fitting is sufficiently satisfactory in almost all the biomedical studies that have appeared in the literature.

3.2.4 Least Squares with Centered Covariates

In some situations, some of the covariates used in (4.1) can not have values at zero, so that the baseline coefficient curve $\beta_0(t)$ does not have a practical interpretation. Strictly positive covariates appear naturally both in the ASGA Study (Section 1.2.1), such as the mother's placental thickness and pre-pregnancy height, and the HIV/CD4 Depletion Data (Section 1.2.2), such as the subject's pre-infection CD4 level. A useful remedy when such a situation arises is to use a centered version of the covariates in the model, so that the corresponding

baseline coefficient can be interpreted as the conditional mean of $Y(t)$ when the centered covariates are set to zero.

Let $X^{(*l)}(t) = X^{(l)}(t) - E[X^{(l)}(t)]$ be the centered version of $X^{(l)}(t)$ and $\mathbf{X}^{(*)}(t)$ be the covariate vector with some or all of its components being centered. An equivalent form of (4.1) is

$$Y(t) = \left(\mathbf{X}^{(*)}(t)\right)^T \beta^*(t) + \epsilon(t), \quad (4.9)$$

where $\beta^*(t) = (\beta_0^*(t), \beta_1(t), \dots, \beta_k(t))^T$. Note that $\beta_0^*(t)$, the baseline coefficient curve of (4.9), represents the mean of $Y(t)$, when $X^{(*l)}(t)$, rather than $X^{(l)}(t)$, for $l = 1, \dots, k$ are set to zero. Other coefficient curves of (4.9) can be interpreted the same way as those of (4.1).

The estimation of $\beta^*(t)$ can be obtained by first estimating the centered covariates $X_{ij}^{(*l)}$ of $X_{ij}^{(l)}$ and then minimizing (4.5) with $X_{ij}^{(l)}$ replaced by $X_{ij}^{(*l)}$. If $X^{(l)}(t)$ is a time-dependent covariate, then, using a kernel smoothing, a centered version of $X_{ij}^{(l)}$ can be estimated by $X_{ij}^{(*l)} = X_{ij}^{(l)} - \hat{\mu}_l(t_{ij})$ with

$$\hat{\mu}_l(t) = \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} \{w_i X_{ij}^{(l)} \Gamma_l[(t - t_{ij})/\gamma_l]\}}{\sum_{i=1}^n \sum_{j=1}^{n_i} \{w_i \Gamma_l[(t - t_{ij})/\gamma_l]\}}, \quad (4.10)$$

where $(\Gamma_l(\cdot), \gamma_l)$ is a set of kernel and bandwidth. On the other hand, if $X^{(l)}(t) \equiv X^{(l)}$ is time-invariant, then $X_{ij}^{(l)} \equiv X_i^{(l)}$ for all $j = 1, \dots, n_i$, and $X_i^{(*l)}$ can be taken as $X_i^{(l)} - \bar{X}^{(l)}$, where $\bar{X}^{(l)} = n^{-1} \sum_{i=1}^n X_i^{(l)}$ is the weighted sample mean for $X^{(l)}$. Let $\mathbf{X}_i^{(*)}$ be the $n_i \times (k+1)$ centered covariate vector whose j th row is $(1, X_{ij}^{(*1)}, \dots, X_{ij}^{(*k)})$. A least squares kernel estimator of $\beta^*(t)$ is

$$\hat{\beta}^{*LSK}(t) = \left[\sum_{i=1}^n w_i \left(\mathbf{X}_i^{(*)}\right)^T \mathbf{K}_i(t) \left(\mathbf{X}_i^{(*)}\right)^T \right]^{-1} \left[\sum_{i=1}^n w_i \left(\mathbf{X}_i^{(*)}\right)^T \mathbf{K}_i(t) \mathbf{Y}_i \right], \quad (4.11)$$

where $\mathbf{K}_i(t)$ and \mathbf{Y}_i are defined as in (4.6).

Wu, Yu and Chiang (2000) investigated the large sample properties of $\hat{\beta}^{*LSK}(t)$. Their results suggest that neither $\hat{\beta}^{LSK}(t)$ nor $\hat{\beta}^{*LSK}(t)$ is uniformly superior to the other. In particular, when the covariates are time-invariant, $\hat{\beta}^{LSK}(t)$ and $\hat{\beta}^{*LSK}(t)$ are asymptotically equivalent. However, when $X^{(l)}(t)$ for $l \geq 1$ changes significantly with t , theoretically and practically superior estimators of $\beta_l(t)$ may be obtained by centering $X^{(l)}(t)$.

Of course, after a covariate is centered, the baseline coefficient curve of the model is changed. The decision on whether a covariate should be centered or not primarily depends

on the biological interpretations of the corresponding baseline coefficient curve. Such a decision should be made based on the statistical properties of the estimators only if the effects of the covariates, rather than the baseline coefficient curve, is of primary interest in the investigation. Clearly, methods other than kernel smoothing may also be applied to the estimation with centered covariates. But, because of the complication caused by smoothing the covariates, statistical properties for estimators other than (4.11) have not been investigated in the literature.

3.2.5 A Simple Modification

The estimators mentioned above, both with and without covariate centering, rely on a single bandwidth to estimate all $(k + 1)$ coefficient curves. This simple approach may work well when all the curves roughly belong to the same smoothness family. However, such an idealized scenario is often not anticipated in practice. A flexible method which automatically adjusts for the possibly different smoothing needs for different coefficient curves is always preferred.

In the literature, the potential deficiency associated with the use of a single bandwidth has been reported by Hoover *et al.* (1998), Fan and Zhang (2000), Wu, Yu and Chiang (2000), among others. These authors have also proposed a number of alternative approaches (see Sections 4.3-4.6) to overcome this potential drawback. A simple method suggested by Wu, Yu and Chiang (2000) is to use a linear combination of the form

$$\widehat{\beta}(t; \mathbf{K}, \mathbf{h}) = \sum_{l=0}^k e_{l+1}^T \widehat{\beta}(t; K_l, h_l), \quad (4.12)$$

where $\mathbf{K}(\cdot) = (K_0(\cdot), \dots, K_k(\cdot))$, $\mathbf{h} = (h_0, \dots, h_k)$, e_p is the $[(k + 1) \times 1]$ vector with 1 at its p th place and zero elsewhere and $\widehat{\beta}(t; K_l, h_l)$ is the kernel estimator of $\beta(t)$ or $\beta^*(t)$ obtained from (4.6) or (4.11), respectively, using kernel $K_l(\cdot)$ and bandwidth h_l . Intuitively, $\widehat{\beta}(t; \mathbf{K}, \mathbf{h})$ relies on a specific pair of kernel and bandwidth to estimate the corresponding component of $\beta(t)$ or $\beta^*(t)$. As a general methodology, (4.12) is not limited to kernel estimators and may be applied to other local polynomial estimators as well.

3.2.6 Choices of w_i

An important factor that affects the theoretical and practical behaviors of the least squares local polynomial estimators of $\beta(t)$ is the choice of w_i in (4.5). For cross-sectional studies

with independent identically distributed data, a uniform weight choice, $w_i \equiv 1/N$, is often desirable. For the current sampling, it is conceivable that a proper choice of w_i may depend on the intra-subject correlation structures and the numbers of repeated measurements n_i . In practice, however, the correlation structures of the data are often completely unknown and may be difficult to estimate, so that subjective choices such as $w_i = 1/N$ and $w_i = 1/(nn_i)$ are often considered. Intuitively, $w_i = 1/N$ assigns equal weight to each observation point, while $w_i = 1/(nn_i)$ assigns equal weight to each subject. Theoretically, the choice of $w_i = 1/N$ may produce inconsistent least squares kernel estimators when some n_i are much larger than the others. On the other hand, the least squares kernel estimators based on $w_i = 1/(nn_i)$ are always consistent regardless the choices of n_i (Hoover *et al.*, 1998, and Wu and Chiang, 2000).

3.3 Penalized Least Squares

Suppose that all the components of $\beta(t)$ are twice continuously differentiable and have bounded and square integrable second derivatives with respect to t . A natural penalized least squares criterion is to minimize

$$J(\beta, \lambda) = \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ Y_{ij} - \sum_{l=0}^k X_{ij}^{(l)} \beta_l(t_{ij}) \right\}^2 + \sum_{l=0}^k \lambda_l \int [\beta_l''(t)]^2 dt \quad (4.13)$$

with respect to $\beta_l(t)$, where $\lambda = (\lambda_0, \dots, \lambda_k)^T$ and λ_l are positive smoothing parameters. The existence and uniqueness of the minimizer of (4.13) depend on t_{ij} and $X_{ij}^{(l)}$. Suppose that (4.13) can be uniquely minimized. The penalized least squares estimator $\widehat{\beta}^{PLS}(t) = (\widehat{\beta}_0^{PLS}(t), \dots, \widehat{\beta}_k^{PLS}(t))^T$ of $\beta(t)$ is then defined to be the unique minimizer of (4.13). Using similar techniques as in univariate smoothing, it can be shown that $\widehat{\beta}_l^{PLS}(t)$ are natural cubic splines with knots at the distinct values of $\{t_{ij} : i = 1, \dots, n, j = 1, \dots, n_i\}$ and can be expressed as linear functions of $\{Y_{ij} : i = 1, \dots, n, j = 1, \dots, n_i\}$.

One feature that distinguishes $\widehat{\beta}^{PLS}(t)$ from the estimators obtained from (4.5) is the use of multiple smoothing parameters λ_l in the penalty term. In (4.13), all $(k+1)$ smoothing parameters λ_l , $l = 0, \dots, k$, can be adjusted in the penalty term. Numerical results presented in Hoover *et al.* (1998) demonstrated that the extra flexibility created by multiple smoothing parameters could indeed lead to better estimators than the least squares local polynomials that rely on a single smoothing parameter. However, because $\widehat{\beta}^{PLS}(t)$ has knots at all the distinct time points, it can be extremely computationally intensive when the

number of distinct time points is large, a case often happened in unbalanced longitudinal studies.

3.4 Two-Step Smoothing Method

In an attempt to provide flexible smoothing estimators that are computationally accessible with large longitudinal data, Fan and Zhang (2000) proposed to estimate $\beta(t)$ by a two-step smoothing method which uses $(k + 1)$ smoothing parameters in a different way from (4.12) and (4.13). Their procedure calls for the following two steps:

- (i) computing the raw estimates $\widehat{\beta}^{RAW}(s)$ of $\beta(s)$ at a set of distinct time points, say s_1, \dots, s_m , where m may depend on n and n_i , $i = 1, \dots, n$;
- (ii) estimating each coefficient curve $\beta_l(t)$ by smoothing the raw estimates $\widehat{\beta}_l^{RAW}(s_r)$, $r = 1, \dots, m$.

Although Fan and Zhang (2000) used local polynomials to illustrate the method, other smoothing methods such as splines may in principle be used.

For the special case of balanced longitudinal data where all the subjects are observed at a same set of time points $\{s_j; j = 1, \dots, m\}$ with $m = n_i$, $i = 1, \dots, n$, the raw estimates can be computed by fitting linear models between Y_{ij} and X_{ij} at s_j for all $j = 1, \dots, m$. However, when the design is unbalanced and the numbers of subjects on some time points are sparse, as in most practical situations, it may be necessary to computing the raw estimates by grouping the observations from the adjacent time points. In particular, we can first compute $\widehat{\beta}_l^{RAW}(s_r)$, $l = 0, \dots, k$, using the local polynomial method (4.5) with a small bandwidth, and then, treating $\widehat{\beta}_l^{RAW}(s_r)$ as the new data, estimate $\beta_l(t)$ by minimizing

$$L_{p,l}^{TS}(t) = \sum_{j=1}^m \left\{ \widehat{\beta}_l^{RAW}(s_j) - \sum_{r=0}^p b_{lr}(t)(s_j - t)^r \right\}^2 K_l \left(\frac{s_j - t}{h_l} \right) \quad (4.14)$$

with respect to $b_{lr}(t)$, where $(K_l(\cdot), h_l)$ is a set of kernel and bandwidth. Similar to (4.5), if $\widehat{b}_{lr}^{TS}(t)$ for $r = 0, \dots, k$ uniquely minimize (4.14), $\widehat{b}_{l0}^{TS}(t)$ is the two-step p th order local polynomial estimator of $\beta_l(t)$, while $(r!) \widehat{b}_{lr}^{TS}(t)$ can be used to estimate the r th derivative of $\beta_l(t)$.

In contrast to the estimators obtained from (4.5) where a single bandwidth must be used for all $\beta_l(t)$, the two-step method has in principle the flexibility to adjust for the specific smoothing need of each coefficient curve. However, a main difficulty in current version of

two-step smoothing is that it lacks a specific and practical guideline to construct the raw estimates for unbalanced longitudinal data. Certain data-driven bandwidth procedures would be desirable for computing both the raw and the final estimates. Impacts of different raw estimates on the theoretical and practical properties of the final two-step estimators are still not well-understood and require substantial further development.

3.5 Component-wise Smoothing with Time-Invariant Covariates

When the covariates of interest are time-invariant, such as in clinical trials when the treatments are kept fixed throughout the study periods, an effective way motivated by (4.3) to provide flexible and computational feasible estimators of $\beta(t)$ is to smooth each component of $\beta(t)$ separately.

Let $Z^{(r)}(t) = [\sum_{l=0}^k e_{rl}X^{(l)}]Y(t)$, $\mathbf{X}_i = (1, X_i^{(1)}, \dots, X_i^{(k)})^T$ be the covariate vector of the i th subject and \hat{e}_{rl} be the (r, l) th element of the matrix $(\hat{E}_{\mathbf{X}\mathbf{X}^T})^{-1}$, the inverse of the sample mean $\hat{E}_{\mathbf{X}\mathbf{X}^T} = (1/n) \sum_{i=0}^n \mathbf{X}_i \mathbf{X}_i^T$. A natural estimator of $Z^{(r)}(t)$ is $Z_{ij}^{(r)} = [\sum_{l=0}^k \hat{e}_{rl} X_i^{(l)}] Y_{ij}$. By (4.3), a componentwise smoothing estimator of $\beta_r(t)$ can be obtained by smoothing $Z_{ij}^{(r)}$ for $i = 1, \dots, n$ and $j = 1, \dots, n_i$. Specifically, a local polynomial estimator of $\beta_r(t)$ with order $p \geq 0$ is $\hat{b}_{r0}^{COM}(t)$, such that $\hat{b}_{rl}^{COM}(t)$, $l = 0, \dots, p$, uniquely minimize

$$L_{p,r}^{COM}(t) = \sum_{i=1}^n \sum_{j=1}^{n_i} w_i \left\{ Z_{ij}^{(r)} - \sum_{l=0}^p b_{rl}(t)(t_{ij} - t)^l \right\}^2 K_r \left(\frac{t_{ij} - t}{h_r} \right), \quad (4.15)$$

with respect to $b_{rl}(t)$. For the local constant fitting with $p = 0$, (4.15) leads to the componentwise kernel estimator

$$\hat{\beta}_r^{COM}(t) = \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ w_i Z_{ij}^{(r)} K_r [(t_{ij} - t)/h_r] \right\}}{\sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ w_i K_r [(t_{ij} - t)/h_r] \right\}}. \quad (4.16)$$

Wu and Chiang (2000) established the large sample mean squared errors of $\hat{\beta}_r^{COM}(t)$, while Wu, Yu and Yuan (2000) developed a procedure for constructing approximate asymptotic pointwise and simultaneous confidence regions for $\beta_r(t)$. These results shed some light on the asymptotic behaviors of the higher order estimators $\hat{b}_{r0}^{COM}(t)$, although specific asymptotic risks and asymptotic distributions have not been established for the case with $p \geq 1$. The results of Wu and Chiang (2000) and Wu, Yu and Yuan (2000) indicate some clear advantages of $\hat{\beta}_r^{COM}(t)$ over the kernel estimator (4.6) both in terms of theoretical con-

vergence rates and practical flexibilities. Similar advantages over the least squares method of (4.5) are also expected for the componentwise local polynomial estimators.

Obviously, minimizing (4.15) is not the only componentwise smoothing approach. Suppose that the support of the design time points is contained in a compact set $[a, b]$ and $\beta_r(t)$ is twice differentiable with respect to t in $[a, b]$. A viable alternative is to estimate $\beta_r(t)$ by penalized least squares estimator $\tilde{\beta}_r^{COM}(t)$, where $\tilde{\beta}_r^{COM}(t)$ minimizes

$$J_r^{COM}(\beta_r, \lambda_r) = \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ w_i \left[Z_{ij}^{(r)} - \beta_r(t_{ij}) \right]^2 \right\} + \lambda_r \int_a^b \left[\beta_r''(s) \right]^2 ds, \quad (4.17)$$

with λ_r being a non-negative smoothing parameter. By the same rationale as in Section 2.3, it is easy to verify that $\tilde{\beta}_r^{COM}(t)$ is a natural cubic spline with knots at the distinct points of $\{t_{ij}; i = 1, \dots, n, j = 1, \dots, n_i\}$. Furthermore, using the approach of equivalent kernels, Chiang, Rice and Wu (2001) derived the asymptotic mean squared errors and the asymptotic distributions of $\tilde{\beta}_r^{COM}(t)$. In contrast to the multiple penalized least squares of (4.13) whose solution is obtained by solving a large linear system involving all $(k+1)$ components, (4.17) significantly simplifies the computation by solving $(k+1)$ separate linear systems. This computational advantage ensures the practical implementability of (4.17) in many situations, while the intensive computational needs often make the optimization of (4.13) impracticable.

3.6 Smoothing via Basis Approximations

All the smoothing methods described above depend on local smoothing in the sense that only the measurements obtained within some neighborhood of t are effectively used to estimate $\beta(t)$. Although local smoothing works well when all the coefficient curves $\beta_r(t)$ are nonparametric, it is not adequate when some of the coefficient curves have known parametric forms, as in the partially linear model (1.1).

Compared with local smoothing, estimation using basis approximations has three important advantages. First, it can be used to estimate $\beta(t)$ whether its components are parametric or nonparametric, hence, is suitable for both nonparametric and semiparametric varying-coefficient models. Second, when a random effect is desired, it provides a natural means to incorporate random effects into a nonparametric or semiparametric model. Third, because popular basis estimators, such as truncated polynomials or B-splines, often rely on far fewer knots or approximation terms than smoothing splines, they often enjoy consider-

able computationally advantage over smoothing splines or even local polynomials. Although estimation with mixed effects is of great interest in various settings, we only discuss here the case of marginal models. Extension to mixed effects models can be found in Rice and Wu (2001).

The main idea is to first approximate $\beta_r(t)$ by a basis function expansion with K_r terms, where K_r may or may not tend to infinity as n tends to infinity, and then estimate $\beta_r(t)$ by estimating the coefficients of this expansion. For each $r = 0, \dots, k$, let $B_{rs}(t)$, $s = 1, \dots, K_r$, be a set of basis functions. If $\beta_r(t)$ can be approximated by an expansion based on $B_{rs}(t)$, $s = 1, \dots, K_r$, there is a set of constants γ_{rs} so that

$$\beta_r(t) \approx \sum_{s=1}^{K_r} \gamma_{rs} B_{rs}(t). \quad (4.18)$$

Substituting (4.18) into (1.2), an approximation of the varying-coefficient model is

$$Y_{ij} \approx \sum_{r=0}^k \sum_{s=1}^{K_r} X_{ij}^{(r)} \gamma_{rs} B_{rs}(t) + \epsilon_i(t_{ij}). \quad (4.19)$$

The approximation sign in (4.19) will be replaced by the equality sign if, for all $r = 0, \dots, k$, $\beta_r(t)$ belongs to a linear space spanned by $\{B_{rs}(t); s = 1, \dots, K_r\}$.

Using (4.19), the least squares estimators $\hat{\gamma}_{rs}$ of γ_{rs} can be obtained by minimizing

$$\ell(\gamma) = \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ w_i \left[Y_{ij} - \sum_{r=0}^k \sum_{s=1}^{K_r} \left(X_{ij}^{(r)} \gamma_{rs} B_{rs}(t_{ij}) \right) \right]^2 \right\}, \quad (4.20)$$

where $\gamma = (\gamma_0^T, \dots, \gamma_k^T)^T$ and $\gamma_r = (\gamma_{r1}, \dots, \gamma_{rK_r})^T$. If the minimizer of (4.20) uniquely exists, the basis function estimator of $\beta_r(t)$ is

$$\hat{\beta}_r^{BAS}(t) = \sum_{s=1}^{K_r} [\hat{\gamma}_{rs} B_{rs}(t)], \quad (4.21)$$

where K_r may depend on n and n_i , $i = 1, \dots, n$. Clearly, if K_r is finite and known and $\beta_r(t)$ belongs to the linear model spanned by $B_{rs}(t)$, $s = 1, \dots, K_r$, then (4.21) returns a parametric estimator of $\beta_r(t)$. On the other hand, if (4.18) holds with K_r unknown, a consistent nonparametric estimator produced by (4.21) may require K_r to be a function of n and n_i , $i = 1, \dots, n$, which may tend to infinity as n tends to infinity.

Depending on the underlying scientific nature of the data, many different bases may be used to approximate the components of $\beta(t)$. The most popular basis system in the

classical linear models is the polynomial basis $\{1, t, \dots, t^{K_r-1}\}$. A general class of bases that have certain numerical advantages over the above polynomial basis is the class of piecewise polynomials. Examples of piecewise polynomial bases include B-spline bases, such as linear, quadratic or cubic splines, or other types of truncated power series; see de Boor (1978) for further details of the explicit expressions of piecewise polynomials and their numerical properties. If $\beta_r(t)$ is believed to exhibit periodicity, Fourier series are often natural basis choices.

Huang, Wu and Zhou (2002) established the consistency of (4.21) and studied the practical performance of (4.21) with B-splines through an intensive simulation. Further asymptotic properties for the B-splines estimators of (4.21) are developed in Huang, Wu and Zhou (2004). In general, a B-spline estimator requires a smoothing parameter consisted of three aspects: degrees of the polynomials and number and location of the knots. Although generally desired, it is difficult, however, to simultaneously determine all three of these aspects from the data. Rice and Wu (2001) showed that the simple approach of using equally spaced knots often works well in practice, a finding also corroborated by the simulation of Huang, Wu and Zhou (2002).

3.7 A Cross-Validation Procedure

The most important factor that affects all of the above smoothing methods is the selection of appropriate smoothing parameters, such as the bandwidth, the positive penalty weight λ and the number and location of knots. It is of both theoretical and practical interest to select these values directly from the data.

Selecting data-driven smoothing parameters for nonparametric regression with independent identically distributed data has been a subject of intense investigation in the literature. Under the current context, a widely used method, suggested by Rice and Silverman (1991), is a cross-validation that deletes the entire repeated measurements of a subject, rather than an individual measurement, one at a time. Hart and Wehrly (1993) derived the consistency of this cross-validation for a simple nonparametric regression without the presence of covariates other than time. Without loss of generality, we denote ξ to be a vector of smoothing parameters, $\hat{\beta}(t; \xi)$ a smoothing estimator based on ξ and $\hat{\beta}^{(-i)}(t; \xi)$ an estimator computed using the same method as $\hat{\beta}(t; \xi)$ but with the i th subject's measurements deleted. The

cross-validation score for $\widehat{\beta}(t; \xi)$ is

$$\text{CV}(\xi) = \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ w_i \left[Y_{ij} - X_{ij}^T \widehat{\beta}^{(-i)}(t; \xi) \right]^2 \right\}, \quad (4.22)$$

which measures the predictive error of $\widehat{\beta}(t; \xi)$. The cross-validated smoothing parameter ξ_{cv} is then the minimizer of $\text{CV}(\xi)$, provided that the unique minimizer of $\text{CV}(\xi)$ exists.

The above cross-validation criterion is directly applicable to all the smoothing methods presented above, except the two-step smoothing of Section 2.4. For the estimators of Sections 2.2, 2.3 and 2.5 and B-splines with equally spaced knots, minimizing the corresponding cross-validation scores would either return a univariate bandwidth or a R^{k+1} -valued vector. An automatic search of the global minima usually requires a sophisticated optimization software. In practice, particularly when the smoothing parameter is multivariate, it is often reasonable to use a smoothing parameter whose cross-validation score is close to the global minima.

There are three intuitive reasons to use the cross-validation criterion (4.22). First, by deleting the subjects one at a time, it preserves the correlation structure of the data. Second, in contrast to alternatives such as the AIC, the BIC and the generalized cross-validation (e.g. Akaike, 1970, Schwarz, 1978, Shibata, 1981, and Wahba, 1990), (4.22) does not depend on the structure of the intra-subject correlations, hence, can be implemented in almost all the practical situations. Third, when the number of subjects is sufficiently large, minimizing (4.22) leads to a smoothing parameter that approximately minimizes the average squared error:

$$\text{ASE}(\widehat{\beta}(\cdot; \xi)) = \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ w_i \left[X_{ij}^T (\beta(t_{ij}) - \widehat{\beta}(t_{ij}; \xi)) \right]^2 \right\}. \quad (4.23)$$

The last assertion can be heuristically seen by the decomposition:

$$\begin{aligned} \text{CV}(\xi) &= \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ w_i \left[Y_{ij} - X_{ij}^T \beta(t_{ij}) \right]^2 \right\} \\ &+ 2 \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ w_i \left[Y_{ij} - X_{ij}^T \beta(t_{ij}) \right] \left[X_{ij}^T (\beta(t_{ij}) - \widehat{\beta}^{(-i)}(t_{ij}; \xi)) \right]^2 \right\} \\ &+ \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ w_i \left[X_{ij}^T (\beta(t_{ij}) - \widehat{\beta}^{(-i)}(t_{ij}; \xi)) \right]^2 \right\}. \end{aligned} \quad (4.24)$$

Here, (4.23) and the definition of $\widehat{\beta}^{(-i)}(t; \xi)$ imply that the third term at the right side of (4.24) is approximately the same as $\text{ASE}(\widehat{\beta}(\cdot; \xi))$. Because the first term at the right side of

(4.24) does not depend on the smoothing parameter and the second term is approximately zero, ξ_{cv} approximately minimizes $\text{ASE}(\widehat{\beta}(\cdot; \xi))$.

3.8 Asymptotic Inferences for Kernel Estimators

Confidence statements can be made either based on the asymptotic distributions of the estimators or through a bootstrap procedure. Currently, explicit expressions of asymptotic distributions have only been developed for the kernel estimators (4.6) and (4.16).

3.8.1 Pointwise Confidence Intervals

For both (4.6) and (4.16), their asymptotic distributions have been developed based on two important assumptions. First, the numbers of repeated measurements n_i are non-random and may or may not tend to infinity as n tending to infinity. Second, the time design points t_{ij} are random and independent identically distributed according to an unknown density function $f(\cdot)$. These assumptions are made for practical considerations as well as mathematical tractability.

Consider first the confidence procedures based on (4.6). Under the above assumptions and some additional mild regularity conditions, Wu, Chiang and Hoover (1998) showed that, if $w_i = 1/N$, $h = N^{-1/5}h_0$ and

$$\lim_{n \rightarrow \infty} N^{-6/5} \sum_{i=1}^n n_i^2 = \theta$$

for some constants $h_0 > 0$ and $0 \leq \theta < \infty$, $\widehat{\beta}^{LSK}(t)$ has an asymptotically multivariate normal distribution in the sense that

$$(Nh)^{1/2} \left[\widehat{\beta}^{LSK}(t) - \beta(t) \right] \longrightarrow \mathbf{N} (B(t), D^*(t)), \quad (4.25)$$

in distribution as $n \rightarrow \infty$. The bias, $B(t)$, and the variance-covariance matrix, $D^*(t)$, of (4.25) are

$$B(t) = [f(t)]^{-1} E_{XX^T}^{-1}(t) (b_0(t), \dots, b_k(t))^T \quad (4.26)$$

and

$$D^*(t) = [f(t)]^{-2} E_{XX^T}^{-1}(t) D(t) E_{XX^T}^1(t) \quad (4.27)$$

where $D(t)$ is a $(k+1) \times (k+1)$ matrix whose (l, r) th element is

$$\begin{aligned} D_{lr}(t) &= \sigma^2(t) E \left[X^{(l)}(t) X^{(r)}(t) \right] f(t) \left\{ \int [K(u)]^2 du \right\} \\ &\quad + \theta h_0 \rho_\epsilon(t) E \left[X^{(l)}(t) X^{(r)}(t) \right] [f(t)]^2, \end{aligned}$$

$\sigma^2(t) = E[\epsilon^2(t)]$, $\rho_\epsilon(t) = \lim_{a \rightarrow 0} E[\epsilon(t+a)\epsilon(t)]$ and

$$b_l(t) = h_0^{3/2} \sum_{c=0}^k \left\{ \left[\int u^2 K(u) du \right] \left\{ \beta'_c(t) \left[E \left[X^{(l)}(t) X^{(c)}(t) \right] \right]' f(t) \right. \right. \\ \left. \left. + \beta'_c(t) E \left[X^{(l)}(t) X^{(c)}(t) \right] f'(t) + (1/2) \beta''_c(t) E \left[X^{(l)}(t) X^{(c)}(t) \right] f(t) \right\} \right\}.$$

Then, there are lower and upper end points $L_\alpha(t)$ and $U_\alpha(t)$ given by

$$\left\{ A^T \widehat{\beta}^{LSK}(t) - (Nh)^{-1/2} A^T B(t) \right\} \pm Z_{\alpha/2} (Nh)^{-1/2} \left[A^T D^*(t) A \right]^{1/2}, \quad (4.28)$$

where $Z_{\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution, so that

$$\lim_{n \rightarrow \infty} P \left\{ L_\alpha(t) \leq A^T \beta(t) \leq U_\alpha(t) \right\} = 1 - \alpha. \quad (4.29)$$

Because $B(t)$ and $D^*(t)$ depend on unknown quantities, (4.28) is not implementable in practice. If $B(t)$ and $D^*(t)$ can be consistently estimated by $\widehat{B}(t)$ and $\widehat{D}^*(t)$, a pointwise $(1 - \alpha)$ confidence interval for $A^T \beta(t)$ can be approximated by $(\widehat{L}_\alpha(t), \widehat{U}_\alpha(t))$ with $\widehat{L}_\alpha(t)$ and $\widehat{U}_\alpha(t)$ being the lower and upper end points given by

$$\left\{ A^T \widehat{\beta}^{LSK}(t) - (Nh)^{-1/2} A^T \widehat{B}(t) \right\} \pm Z_{\alpha/2} (Nh)^{-1/2} \left[A^T \widehat{D}^*(t) A \right]^{1/2}. \quad (4.30)$$

Wu, Chiang and Hoover (1998) suggested to compute $\widehat{B}(t)$ and $\widehat{D}^*(t)$ by substituting $f(t)$, $\sigma^2(t)$, $\rho_\epsilon(t)$, $E[X^{(l)}(t)X^{(r)}(t)]$ and the required derivatives in (4.26) and (4.27) with their kernel estimators. Suppose that the kernel function $K(\cdot)$ is at least twice continuously differentiable in the interior of its support. These authors proposed to estimate $f(t)$, $\sigma^2(t)$, $\rho_\epsilon(t)$ and $E[X^{(l)}(t)X^{(r)}(t)]$ by

$$\widehat{f}(t) = (Nh)^{-1} \sum_{i=1}^n \sum_{j=1}^{n_i} K \left(\frac{t_{ij} - t}{h} \right), \\ \widehat{\sigma}^2(t) = \frac{1}{Nh \widehat{f}(t)} \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ \widehat{\epsilon}_i^2(t_{ij}) K \left(\frac{t_{ij} - t}{h} \right) \right\}, \\ \widehat{\rho}_\epsilon(t) = \frac{\sum_{i=1}^n \sum_{j_1 \neq j_2} \left\{ \widehat{\epsilon}_i(t_{ij_1}) \widehat{\epsilon}_i(t_{ij_2}) K \left(\frac{t_{ij_1} - t}{h} \right) K \left(\frac{t_{ij_2} - t}{h} \right) \right\}}{\sum_{i=1}^n \sum_{j_1 \neq j_2} \left\{ K \left(\frac{t_{ij_1} - t}{h} \right) K \left(\frac{t_{ij_2} - t}{h} \right) \right\}}$$

and

$$\widehat{E} \left[X^{(l)}(t) X^{(r)}(t) \right] = \frac{1}{Nh \widehat{f}(t)} \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ X_i^{(l)}(t_{ij}) X_i^{(r)}(t_{ij}) K \left(\frac{t_{ij} - t}{h} \right) \right\},$$

where $\widehat{\epsilon}_i(t_{ij}) = Y_{ij} - X_i^T(t_{ij}) \widehat{\beta}(t_{ij})$ are the residuals, and to estimate the first and second derivatives of $f(t)$, $\beta_l(t)$ and $E[X^{(l)}(t)X^{(r)}(t)]$ by the corresponding derivatives of $\widehat{f}(t)$,

$\hat{\beta}_l^{LSK}(t)$ and $\hat{E}[X^{(l)}(t)X^{(r)}(t)]$. Through an intensive simulation, these authors also suggested that the cross-validation bandwidth h_{cv} obtained from (4.22) may be used to compute all of the above estimators, although, in general, different bandwidths may be used for these estimators.

The above plug-in approach can also be extended to $\hat{\beta}_r^{COM}(t)$ of (4.16) when the covariates are time-invariant. Wu, Yu and Yuan (2000) have derived the explicit expressions of the bias, $B(\hat{\beta}_r^{COM}; t)$, and the standard deviation, $SD(\hat{\beta}_r^{COM}; t)$, of $\hat{\beta}_r^{COM}(t)$, and suggested to use the approximate $(1 - \alpha)$ confidence interval for $\beta_r(t)$ with end points

$$\left\{ \hat{\beta}_r^{COM}(t) - \hat{B}(\hat{\beta}_r^{COM}; t) \right\} \pm Z_{1-\alpha/2} \widehat{SD}(\hat{\beta}_r^{COM}; t),$$

where $\hat{B}(\hat{\beta}_r^{COM}; t)$ and $\widehat{SD}(\hat{\beta}_r^{COM}; t)$ are plug-in estimators of $B(\hat{\beta}_r^{COM}; t)$ and $SD(\hat{\beta}_r^{COM}; t)$. Because of the similarity it shares with $\hat{\beta}_l^{LSK}(t)$, we omit the details for this case.

The above asymptotic intervals differ from their counterparts with independent identically distributed data in the inclusion of intra-subject correlations in the variance term. When n_i are not negligible relative to n , θ in (4.27) may not be negligible, so that the contribution of the correlations may not be ignored. For the HIV/CD4 data (Section 1.2.2), the numbers of repeated measurements range from 1 to 14, while the number of subjects is 400. Asymptotic results that do not take the intra-subject correlations into account may not lead to adequate approximations. In this case, it is appropriate to estimate the correlations directly from the data. When the numbers of repeated measurements are negligible relative to the numbers of subjects, as in the ASGA data (Section 1.2.1), the contribution of the intra-subject correlation structures becomes negligible in the variances of the kernel estimators. The resulting confidence intervals are then similar to that with independent identically distributed samples.

3.8.2 Simultaneous Bands

In most applications, the main interest of inference lies in the overall confidence regions of $\beta_l(t)$ within a proper range of t values, rather than the confidence intervals at a particular time point. When the data are from independent identically distributed samples, simultaneous confidence regions for regression curves may be constructed using either extreme value theory of Gaussian processes (e.g. Eubank and Speckman, 1993) or variability bands bridged by pointwise intervals over a grid points (e.g. Knaf, Sacks and Ylvisaker, 1985,

Hall and Titterton, 1988, and Härdle and Marron, 1991). For longitudinal samples, analogous asymptotic theory of extreme values has not been developed. This leaves the latter approach to be the only practical simultaneous inferential tool in longitudinal analysis.

To construct a simultaneous band for $A^T\beta(t)$ over $t \in [a, b]$ based on the least squares kernel estimator $\hat{\beta}^{(LSK)}(t)$, we choose a positive integer M and partition $[a, b]$ into M equally spaced intervals with grid points $a = \xi_1 < \dots < \xi_{M+1} = b$, such that $\xi_{j+1} - \xi_j = (b - a)/M$ for $j = 1, \dots, M$. A set of approximate $(1 - \alpha)$ simultaneous confidence intervals for $A^T\beta(\xi_j)$, $j = 1, \dots, M + 1$, is then the collection of intervals $(\hat{l}_\alpha(\xi_j), \hat{u}_\alpha(\xi_j))$, $j = 1, \dots, M + 1$, which satisfies

$$\lim_{n \rightarrow \infty} P \left\{ \hat{l}_\alpha(\xi_j) \leq A^T\beta(\xi_j) \leq \hat{u}_\alpha(\xi_j) \text{ for all } j = 1, \dots, M + 1 \right\} \geq 1 - \alpha. \quad (4.31)$$

The Bonferroni adjustment suggests

$$(\hat{l}_\alpha(\xi_j), \hat{u}_\alpha(\xi_j)) = (\hat{L}_{\alpha/(M+1)}(\xi_j), \hat{U}_{\alpha/(M+1)}(\xi_j)), \quad (4.32)$$

where $(\hat{L}_\alpha(\xi_j), \hat{U}_\alpha(\xi_j))$ are defined in (4.30).

To establish a band that covers all the points between the grid points ξ_j , $j = 1, \dots, M + 1$, we first consider the interpolation of $A^T\beta(\xi_j)$ defined by

$$(A^T\beta)^{(I)}(t) = \left\{ \frac{M(\xi_{j+1} - t)}{b - a} \right\} [A^T\beta(\xi_j)] + \left\{ \frac{M(t - \xi_j)}{b - a} \right\} [A^T\beta(\xi_{j+1})], \quad (4.33)$$

for $t \in [\xi_j, \xi_{j+1}]$. A simultaneous band for $(A^T\beta)^{(I)}(t)$ over $t \in [a, b]$ is $(\hat{l}_\alpha^{(I)}(t), \hat{u}_\alpha^{(I)}(t))$, where $\hat{l}_\alpha^{(I)}(t)$ and $\hat{u}_\alpha^{(I)}(t)$ are the linear interpolations of $\hat{l}_\alpha(\xi_j)$ and $\hat{u}_\alpha(\xi_j)$, similarly defined as in (4.33). The gaps between the grid points are then bridged by the smoothness conditions of $A^T\beta(t)$. If $A^T\beta(t)$ satisfies

$$\sup_{t \in [a, b]} \left| (A^T\beta)'(t) \right| \leq c_1, \quad \text{for a known constant } c_1 > 0, \quad (4.34)$$

then it follows that

$$\left| A^T\beta(t) - (A^T\beta)^{(I)}(t) \right| \leq 2c_1 \left[\frac{M(\xi_{j+1} - t)(t - \xi_j)}{b - a} \right],$$

for all $t \in [\xi_j, \xi_{j+1}]$, and consequently

$$\left(\hat{l}_\alpha^{(I)}(t) - 2c_1 \left[\frac{M(\xi_{j+1} - t)(t - \xi_j)}{b - a} \right], \hat{u}_\alpha^{(I)}(t) + 2c_1 \left[\frac{M(\xi_{j+1} - t)(t - \xi_j)}{b - a} \right] \right) \quad (4.35)$$

is an approximate $(1 - \alpha)$ confidence band for $A^T\beta(t)$. If $A^T\beta(t)$ satisfies

$$\sup_{t \in [a, b]} \left| \left(A^T\beta \right)''(t) \right| \leq c_2, \quad \text{for a known constant } c_2 > 0, \quad (4.36)$$

then

$$\left| A^T\beta(t) - \left(A^T\beta \right)^{(I)}(t) \right| \leq \frac{c_2}{2} \left[\frac{M(\xi_{j+1} - t)(t - \xi_j)}{b - a} \right],$$

for all $t \in [\xi_j, \xi_{j+1}]$, and an approximate $(1 - \alpha)$ confidence band can be given by

$$\left(\hat{l}_\alpha^{(I)}(t) - \frac{c_2}{2} \left[\frac{M(\xi_{j+1} - t)(t - \xi_j)}{b - a} \right], \hat{u}_\alpha^{(I)}(t) + \frac{c_2}{2} \left[\frac{M(\xi_{j+1} - t)(t - \xi_j)}{b - a} \right] \right). \quad (4.37)$$

For smoothness conditions other than the ones considered in (4.34) and (4.36), the corresponding confidence bands may be similarly established. When the covariates are time-invariant, the same approach can be used to establish simultaneous confidence bands based on $\hat{\beta}^{COM}(t)$; see Wu, Yu and Yuan (2000) for details.

3.9 Bootstrap Variability Bands

The above asymptotic inferences subject to two restrictions which, to some degree, limit their applications in longitudinal analysis. First, because the asymptotic distributions have so far only been developed for the two kernel type estimators, $\hat{\beta}^{LSK}(t)$ and $\hat{\beta}^{COM}(t)$, confidence procedures for other estimators are still not available. Given that smoothing methods such as splines and local polynomials have exhibited a number of theoretical and practical advantages over the kernel methods, particularly at the boundary of the support of t , inferential procedures based on these smoothing methods are in demand. Second, because the plug-in estimators require the estimation of the design densities, covariance functions and the other quantities appeared in the bias and variance terms of the estimators, the procedure is usually computationally intensive and may introduce additional errors in its coverage probabilities.

A more appealing inferential procedure that has been suggested in the literature is the “resampling-subject” bootstrap. This approach has broader appeal in longitudinal analysis since it resamples the subjects of the original data, which may preserve the intra-subject correlations. Although its theoretical properties have not been well-understood, practical performances of this “resampling-subject” bootstrap have been investigated by a number of simulation studies. Let $\hat{\beta}(t) = (\hat{\beta}_0(t), \dots, \hat{\beta}_k(t))^T$ be an estimator of $\beta(t)$ constructed based on any of the previously mentioned smoothing method. An approximate $(1 - \alpha)$ pointwise percentile interval for $A^T E[\hat{\beta}(t)]$ can be constructed by the following steps:

1. Randomly draw n subjects with replacement from the original dataset and denote the resulting bootstrap sample to be $\{(Y_{ij}^*, t_{ij}^*, X_{ij}^*); i = 1, \dots, n, j = 1, \dots, n_i\}$.
2. Compute the bootstrap estimator $\widehat{\beta}^{boot}(t)$, hence $A^T \widehat{\beta}^{boot}(t)$, based on the above bootstrap sample and the smoothing method specified for $\widehat{\beta}(t)$.
3. Repeating the above two steps B times, so that B bootstrap estimators $A^T \widehat{\beta}^{boot}(t)$ are obtained.
4. Calculate $L_{\alpha}^{boot}(t)$ and $U_{\alpha}^{boot}(t)$, the lower and upper $[100 \times (\alpha/2)]$ th percentiles, respectively, of the B bootstrap estimators $A^T \widehat{\beta}^{boot}(t)$. The approximate $(1 - \alpha)$ bootstrap interval is then $(L_{\alpha}^{boot}(t), U_{\alpha}^{boot}(t))$.

When $A^T E[\widehat{\beta}(t)]$ satisfies the smoothness conditions (4.34) or (4.36), simultaneous confidence bands for $A^T E[\widehat{\beta}(t)]$ can be constructed using (4.35) and (4.37) with (4.32) replaced by $(L_{\alpha/(M+1)}^{boot}(\xi_j), U_{\alpha/(M+1)}^{boot}(\xi_j))$.

The main advantages of this bootstrap are its generality and simplicity. It is not limited to kernel type estimators and does not depend on the correlations and designs of the data. Despite its potential, several related theoretical and practical issues have still yet to be resolved. Because the biases of the estimators have not been adjusted, the resulting intervals or bands may not always have desirable coverage probabilities for $A^T \beta(t)$. If a consistent estimator of the bias is also available, improved confidence regions for $A^T \beta(t)$ may be obtained by adjusting the bias appeared in $(L_{\alpha/(M+1)}^{boot}(\xi_j), U_{\alpha/(M+1)}^{boot}(\xi_j))$. Currently, consistent bias estimators can only be obtained on a case-by-case basis, and no general procedure is available. A natural alternative to the percentile end points used in Step 4 is to consider normal approximated intervals with end points $A^T \widehat{\beta}(t) \pm z_{(1-\alpha/2)} \widehat{se}^{boot}(t)$, where $\widehat{se}^{boot}(t)$ is the sample standard error of the B bootstrap estimators $A^T \widehat{\beta}^{boot}(t)$. Asymptotic properties for both the percentile and the normal approximation bootstrap procedures have not been investigated.

3.10 Application to Alabama Fetal Growth Study

Normal fetal growth is naturally thought to influence infant survival and proper child development. Our objective is to investigate the effects of maternal risk factors and maternal anthropometric measurements on the patterns of fetal growth. Although the outcomes measured by fetal abdominal circumference, biparietal diameter and femur length are all

time-dependent, the covariates of interest may be either time-dependent or time-invariant. A typical time-dependent covariate is the maternal placental thickness measured by ultrasound at each visit. On the other hand, mother's height, weight and body mass index measured at the beginning of pregnancy, are time-invariant. Other variables, such as maternal habits of cigarette smoking and alcohol consumption, may be either time-dependent or time-invariant depending on how these variables are defined. A simple way to define time-invariant maternal smoking and drinking status is to categorize the mothers as smokers (ever smoked cigarettes during the pregnancy) versus non-smokers (never smoked cigarettes during the pregnancy) and non-drinkers/light-drinkers (consumed one beer/one glass of wine or less per day in average during the pregnancy) versus heavy-drinkers (consumed more than one beer or one glass of wine per day in average during the pregnancy). As in most self-reported questionnaires, the data contain the average numbers of cigarettes smoked and the average amount of alcohol consumed per day per subject. These actual cigarette and alcohol consumptions are clearly time-dependent as some of the participating subjects change their behaviors during the study. Depending on the specific scientific questions, both smoking and drinking categories and the actual consumptions could be considered in the analysis.

For the purpose of illustration, the analysis present here focuses on the effects of maternal smoking/drinking categories and placental thickness on the growth of fetal abdominal circumference. Other covariate and outcome measurements can be similarly investigated, provided that the models have clear and meaningful biological interpretations. Although the general trend of Figure 1 shows an upward growth pattern, it hardly provides any clue on the relationship between fetal growth and the covariates of interest. A nonparametric analysis with (1.2) seems a natural start.

Let $Y(t)$ and $X^{(1)}(t)$ be the fetal abdominal circumference and placental thickness, respectively, at t weeks of gestation; $X^{(2)}$ and $X^{(3)}$ be the mother's drinking and smoking categories defined by

$$X^{(2)} = \begin{cases} 1 & \text{if she is a non-drinker/light-drinker,} \\ 0 & \text{if she is a heavy-drinker,} \end{cases} \quad X^{(3)} = \begin{cases} 1 & \text{if she is a smoker,} \\ 0 & \text{otherwise;} \end{cases}$$

and $X^{(4)}$ be the mother's height (in centimeters) at the beginning of the pregnancy.

In view that proper placental development may also be affected by drinking and smoking, we first consider the effects of the time-invariant covariate vector $X = (1, X^{(2)}, X^{(3)}, X^{(4)})^T$.

Although we can fit (1.2) directly with $(Y(t), t, X)$ and describe the covariate effects by $\beta(t) = (\beta_0(t), \beta_2(t), \beta_3(t), \beta_4(t))^T$, a better biological interpretation can be obtained if $X^{(4)}$ were replaced by its centered version $X^{(*4)} = X^{(4)} - E[X^{(4)}]$, so that the covariate effects are characterized by $\beta^*(t) = (\beta_0^*(t), \beta_2(t), \beta_3(t), \beta_4(t))^T$. For the latter case, the baseline coefficient curve $\beta_0^*(t)$ represents the mean abdominal circumference at t weeks of gestation for a non-smoking and non-drinking/light-drinking mother whose height is at average, while, for the former, $\beta_0(t)$ itself does not have a biological interpretation.

To fit model (1.2) with $(Y(t), t, X^{(*)})$, $X^{(*)} = (1, X^{(2)}, X^{(3)}, X^{(*4)})^T$, we computed $X_i^{(*4)}$, $i = 1, \dots, 1475$, by subtracting the sample average of $\{X_j^{(4)}; j = 1, \dots, 1475\}$ from $X_i^{(4)}$. Figure 2 shows the estimated coefficient curves, including the baseline growth curve and the covariate effects characterized by alcohol consumption, cigarette smoking and mother's height, and their corresponding 95% simultaneous confidence bands. These coefficient curves were computed using the componentwise estimators of (4.16) with the Epanechnikov kernel, the cross-validated bandwidths described in (4.22) and $w_i = 1/(nn_i)$. It is worthwhile noting that in this data set the numbers of repeated measurements, most of which are around 4, are much smaller compared with the number of subjects $n = 1475$. Thus, asymptotic results obtained by assuming n tending to infinity and n_i remaining finite are expected to give adequate approximations. For kernel smoothing estimators, this means that both $w_i = 1/(nn_i)$ and $w_i = 1/N$ lead to very similar estimates, and the inter-subject correlations can be ignored in the asymptotic variances of the estimators. Thus, no covariance estimators are needed in the construction of asymptotically approximate confidence bands. Based on the same kernel and bandwidths used in the coefficient curve estimates, the simultaneous confidence bands were computed using the asymptotic approximation (4.35) and the Bonferroni adjustment with $M = 40$ and $c_1 = 5$. These graphs suggest an upward linear baseline curve $\beta_0^*(t)$ and undetectable effects from alcohol consumption, cigarette smoking and mother's height. However, because the confidence bands used here tend to be conservative, they may not be sensitive enough to detect small influences of the covariates. The curve estimates and their corresponding confidence bands can also be computed using the least squares kernel method of (4.6). These results are omitted from the presentation, because they are similar to the ones shown in Figure 2.

The above nonparametric results, i.e. graphs shown in Figure 2, suggest that the relationship between fetal abdominal circumference $Y(t)$, gestational age t , alcohol consump-

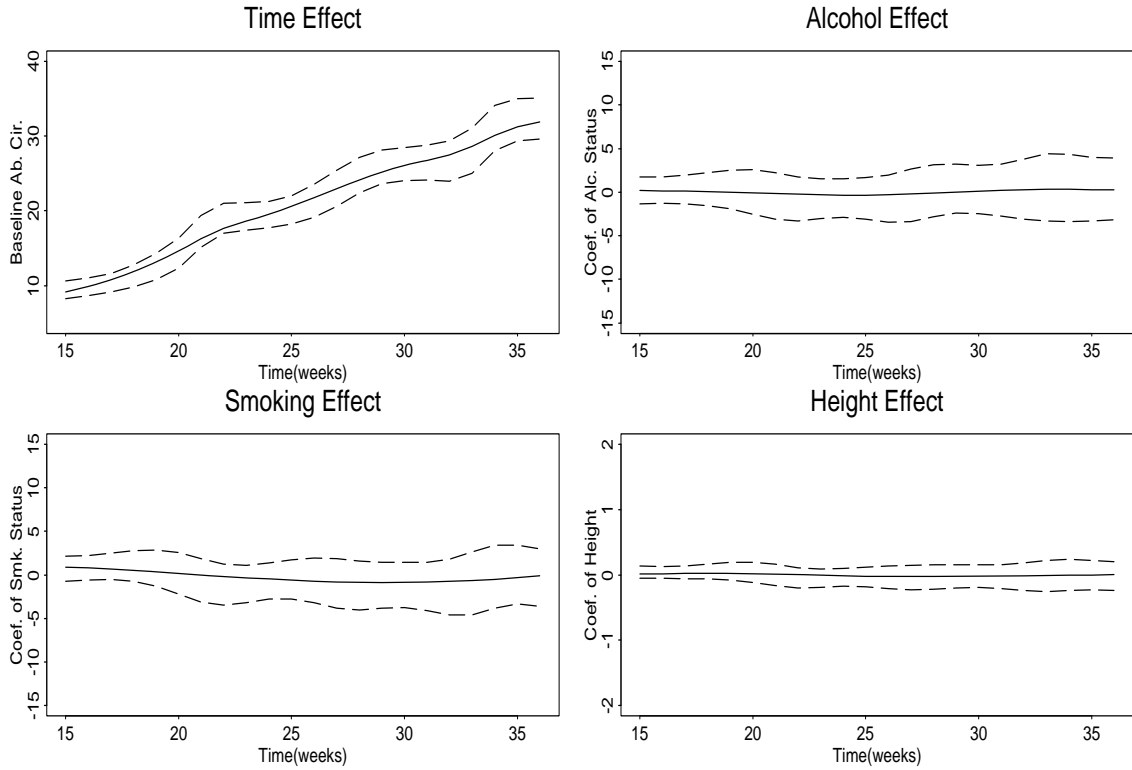


Figure 2: Solid lines: componentwise kernel estimates of the coefficient curves (covariate effects) computed using the Epanechnikov kernel, the cross-validated bandwidths and $w_i = 1/(nn_i)$. Dashed lines: the 95% Bonferroni-type confidence bands.

tion $X^{(2)}$, cigarette smoking $X^{(3)}$ and centered maternal height $X^{(*4)}$ can be reasonably described by the linear model

$$Y(t) = \beta_{00} + \beta_{01}t + \beta_2X^{(2)} + \beta_3X^{(3)} + \beta_4X^{(*4)} + \epsilon(t)$$

with unknown parameters $(\beta_{00}, \beta_{01}, \beta_2, \beta_3, \beta_4)$ and a mean zero error process $\epsilon(t)$. This model can be fitted using the *Mixed-Effects Procedure* in S-plus (Bates and Pinherio, 1999). Table 1 shows the parameter estimates and the corresponding standard errors computed from the above linear model and the S-plus procedure. The results from this linear model suggested clearly non-significant effects for alcohol consumption and maternal height and a very weak, but slightly positive, effect for cigarette smoking. The weak smoking effect shown in this linear analysis is likely caused by the random variations of the data, rather than any substantial association between fetal size and smoking. These results generally agree with the findings obtained from the above nonparametric analysis.

Table 1: Parameter estimates and their standard errors computed using the *Mixed-Effects Procedure* in S-plus.

	Parameter Estimate	Standard Error	Z-ratio
β_{00}	-6.5496	0.0614	-106.5880
β_{01}	1.0645	0.0021	496.1262
β_2	0.0026	0.0551	0.0478
β_3	0.1009	0.0516	1.9555
β_4	0.0007	0.0035	0.1996

When placental thickness $X^{(1)}(t)$ is added to the model, smoothing has to be carried out with time-dependent covariates. In order to obtain a meaningful biological interpretation for the baseline coefficient curve, we use the centered covariate $X^{(*1)}(t) = X^{(1)}(t) - E[X^{(1)}(t)]$, the difference between a subject's placental thickness at time t and the conditional mean at t . To avoid starting with a model that has too many covariates, we consider first fitting (1.2) with $(t, X^{(*1)}(t), X^{(*4)})$ as the covariate vector. The top panel of Figure 3 shows the estimated coefficient curve for $X^{(*1)}(t)$ computed using the kernel method of (4.11) with the standard Gaussian kernel, the cross-validated bandwidths and $w_i = 1/N$. This estimate appears to be undersmoothed, as it can not be explained by a clear biological interpretation. An alternative, perhaps biologically more transparent, estimated coefficient curve of $X^{(*1)}(t)$, shown in the bottom panel of Figure 3, is computed using the same method except with bandwidth vector $(\gamma_1, h_0, h_1, h_4) = (1.5, 1.0, 2.0, 1.0)$. This bandwidth vector was chosen because its cross-validation score was very close to that of the cross-validated bandwidths. Bootstrap percentile intervals are used to demonstrate the variability of the estimates, while inferences based on asymptotic approximations are still not yet available for this type of estimators.

Figure 3 suggests, at least qualitatively, some positive association between placental thickness and fetal abdominal circumference. The estimated coefficient curve for the centered maternal height $X^{(*4)}(t)$ stays constantly close to zero, suggesting a non-significant effect for the maternal height. The estimated baseline coefficient curve is also very close to the one presented in Figure 2. Hence, these curves are omitted from the presentation. Also omitted are the analysis with the mother's drinking and smoking status, $X^{(2)}$ and $X^{(3)}$, added to the model, as their effects are very similar to the ones shown in Figure 2.

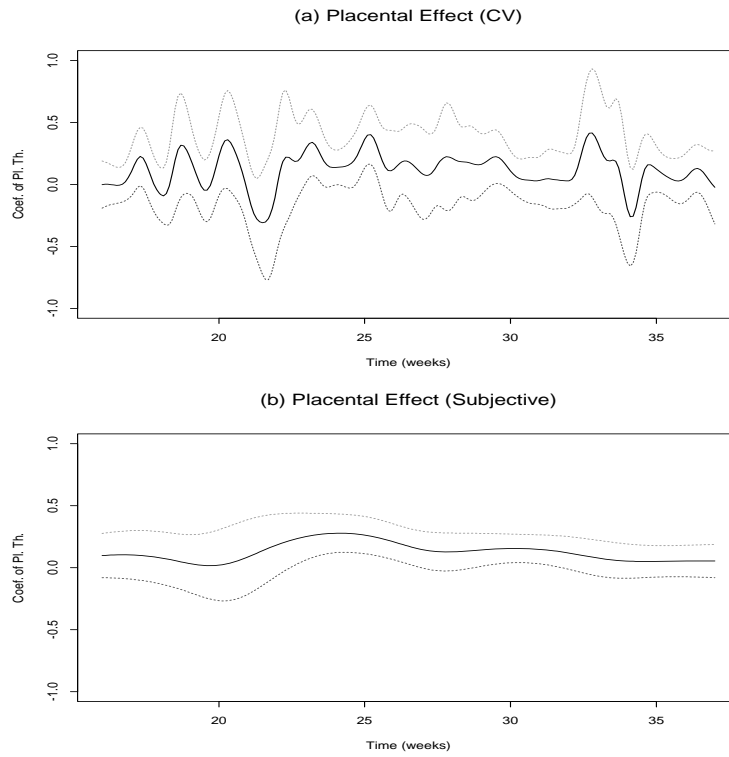


Figure 3: Solid lines: estimated coefficient curve (covariate effect) for placental thickness, computed using (4.11) with the standard Gaussian kernel, $w_i = 1/N$, cross-validated bandwidths (top panel) and bandwidth vector $(\gamma_1, h_0, h_1, h_4) = (1.5, 1.0, 2.0, 1.0)$ (bottom panel). Dashed lines: the 95% pointwise intervals computed using the “resampling-subject” bootstrap percentiles.

3.11 Application to BMACS CD4/HIV Study

Let t_{ij} denote the i th subject's time length (in years) for his j th measurement since HIV infection. Our objective is to evaluate the effects of two factors, the pre-HIV infection CD4 percent $X^{(1)}$ and the smoking status $X^{(2)}(t)$, on the post-HIV infection depletion of CD4 percent $Y(t)$ over time. The first covariate $X^{(1)}$ does not depend on the time since HIV infection. The second covariate $X^{(2)}(t)$ equals 1 if the subject is classified as a smoker at time t and zero otherwise. Because some of the subjects change their smoking habits during the study, $X^{(2)}(t)$ is a time-dependent variable. Owing to the lack of an existing parametric or semiparametric model that is known to describe the scientific relevance between these variables, it is reasonable to consider an initial analysis with the nonparametric model (1.2).

The same rationale used in the analysis of the ASGA study suggests that, in terms of biological interpretability, the center variable $X^{(*1)} = X^{(1)} - E[X^{(1)}]$ is more preferable than its uncentered version $X^{(1)}$ in the model (1.2). However, because $X^{(2)}(t)$ is a time-dependent binary variable, it is unnecessary to be centered. Thus, with $X_i^{(*1)}$ estimated by subtracting the corresponding sample mean from $X_i^{(1)}$, the model (1.2) can be fitted with the data $\{(Y_{ij}, t_{ij}, X_{ij}^*); i = 1, \dots, 400, j = 1, \dots, n_i\}$. The baseline coefficient curve $\beta_0^*(t)$ represents the mean CD4 percent at t years after the infection for those who are non-smokers at time t and have average level of CD4 percent before the infection. The effects $\beta_1^*(t)$ and $\beta_2(t)$ of $X^{(*1)}$ and $X^{(2)}(t)$, respectively, can be interpreted the usual way.

Besides the difference in covariate centering, there is another important difference in the estimation and inferences between this and the previous example. The numbers of repeated measurements in this data set can not be simply ignored compared with the number of subjects. Thus, at least for the known case of kernel estimation, the asymptotic approximations assuming n tending to infinity and n_i remaining bounded may not lead to adequate estimators of the variances, although both $w_i = 1/(nn_i)$ and $w_i = 1/N$ seem to be reasonable weight choices. Because the correlation structure of the data is completely unknown and difficult to be estimated accurately, Wu, Chiang and Hoover (1998) suggested that it is appropriate in this case to obtain conservative Bonferroni-type bands with the covariance $\rho_\epsilon(t)$ in (4.28) replaced by the variance $\sigma^2(t)$, an upper bound for $|\rho_\epsilon(t)|$. The graphs in Figure 4 show the individuals' depletion of CD4 percent over time, the estimated coefficient curves and their corresponding conservative Bonferroni-type 95% asymptotic confidence bands. The estimated coefficient curves were computed using (4.6) with Epanechnikov ker-

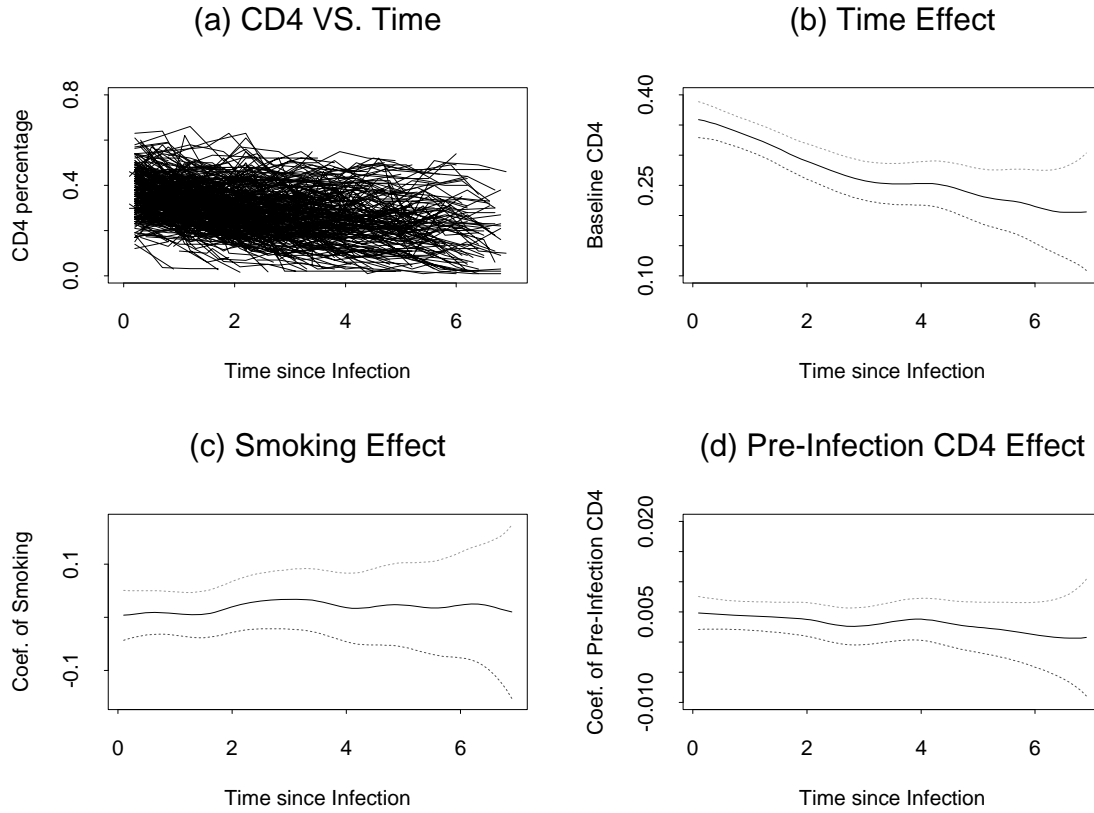


Figure 4: (a) Individuals' CD4 percent versus time (in years) since HIV infection. (b)-(d) Estimated baseline CD4 percent, coefficient curve for smoking and coefficient curve for pre-HIV infection CD4 percent (solid curves) and their corresponding 95% simultaneous confidence bands (dotted curves).

nel, the cross-validated bandwidth and the $w_i = 1/N$ weight. The confidence bands were computed using (4.30) and (4.35) with $M = 138$, $c_1 = 3$ and $\rho_\epsilon(t)$ replaced by $\sigma^2(t)$. The same kernel and bandwidth used in computing (4.6) were also used in computing all the plug-in kernel estimators required in (4.30).

Figure 4(b) shows a declining baseline CD4 percent curve over time since HIV infection, which coincides with the basic trend suggested by the plot shown in Figure 4(a). The simultaneous band for the coefficient curve of the pre-infection CD4 percent stays positive at least for the first four years after HIV infection, suggesting strongly the benefit of high pre-infection CD4 level for the initial period since the infection. However, the positive effect of the pre-infection CD4 percent on the post-infection CD4 percent appears tapering

down at the later stage of the infection. Although the estimated curve in Figure 4(c) stays positive throughout the seven-year time range considered in this data set, the confidence band obtained for this curve does not show any significant positive association between cigarette smoking and post-infection CD4 level. This may be either caused by the weak association between these two variables or the conservative nature of our confidence bands. Clearly, our findings here only provide some exploratory insights on the data. Biomedical implications and parametric models that provide additional meaningful descriptions of the biological mechanisms have to be further developed and independently confirmed by other studies. Nevertheless, the usefulness of nonparametric regression, particularly the varying-coefficient models, in the initial exploration of longitudinal data is transparent, as was shown in this and the previous examples.

3.12 Discussion and Further Remarks

This section has presented a series of parametric, semiparametric and nonparametric models and their estimation and inferential methods for the analysis of longitudinal data. These methods have a wide range of applications in biomedical studies. Theory and methods for parametric models, particularly the linear models, have been extensively studied in the literature. Estimation and inferences based on parametric models can be easily implemented using existing statistical software packages, such as SAS and S-plus. Methods based on semiparametric and nonparametric models, on the other hand, represent the most current progress in this active research field.

The nonparametric estimation and inferential methods introduced here are all based on the general framework of varying-coefficient models. These methods have the advantage of being flexible while applicable to large longitudinal studies. Smoothing methods for these models have been developed using local polynomials and splines, each has its own advantages and disadvantages in practice. Generally speaking, the componentwise smoothing methods are flexible and computationally feasible when the covariates are time-invariant, while methods based on ordinary and penalized least squares and basis approximations can be applied to models with both time-dependent and time-invariant covariates. Pointwise and simultaneous confidence bands for the coefficient curves can be constructed using either asymptotic approximations or the “resampling-subject” bootstrap. The asymptotic confidence procedures have only been developed for the kernel methods. The “resampling-

subject” bootstrap may in principle be used with any smoothing estimators. However, despite the usefulness of this bootstrap shown by a number of simulation studies, its theoretical properties have not been investigated. The approach of two-step smoothing appears to be useful to overcome some of the drawbacks of the ordinary least squares. But, in order for this approach to be useful in an unbalanced longitudinal study, further research is needed to establish specific methods for calculating the raw estimates and the asymptotic properties of the final estimators. Finally, a practical consideration is the use of the uniform weight $w_i = 1/N$ versus the uniform subject weight $w_i = 1/(nn_i)$. Although none of these weight uniformly dominates the other in all the longitudinal designs and an ideal weight may depend on the unknown correlation structure and how fast n_i , $i = 1, \dots, n$, tending to infinity relative to n , simulation studies that have been reported in the literature so far suggest that both weight choices are appropriate when all the subjects have approximately the same numbers of repeated measurements, while the $w_i = 1/(nn_i)$ weight is usually preferred when the numbers of repeated measurements differ from each other significantly.

There are a number of topics that warrant further investigation. First and foremost, although estimation and confidence tools are important in longitudinal analyses, methods that are enormously useful in biomedical studies are testing procedures that can evaluate the statistical evidence for different hypotheses. Such procedures distinguish a parametric submodel that explains a given scientific hypothesis from the general nonparametric model. The main task of decision making is to determine the distributions of the appropriate test statistics. Another practically important problem is to improve the confidence procedures. The procedures presented in this article are known to be conservative, which often hinders their usefulness in practice. Further work needs to focus on reducing the widths of the bands while maintaining satisfactory coverage probabilities. Finally, in view that the varying-coefficient models are still inadequate for a number of longitudinal settings, there is a need to further extend these models. A useful extension is to consider regression models where the outcome variable depends on the history as well as the current values of the covariates. All the estimation and inference methods will have to be redeveloped for this extension.

4 Nonparametric Models for Distribution Functions

4.1 Motivation and Justification

The regression methods discussed in Section 2 and Section 3 are generally based on modeling the conditional mean and covariance structures of the response variables given a set of covariates, which could be either time-varying or time-invariant. Since the mean and covariance structures could be either parametric or nonparametric, the conditional-mean based regression models are undoubtedly the primary tools in a longitudinal analysis, and constitute the majority of techniques developed in the literature. Although popular in practice, this class of methods could be inadequate when the conditional means and covariances are ill-suited for answering the scientific questions being investigated. Such scenarios could arise when the scientific objectives are defined by the outcome variables through their conditional distribution functions which can not be adequately approximated by the normal distributions.

The NGHS data of Example 3, Section 1.3, is a typical example where many relevant scientific questions would better answered by evaluating the conditional distribution functions. As described in Section 1.3, an important objective of the NGHS is to evaluate the effects of age, race and obesity on several cardiovascular risk factors and the temporal trends of cardiovascular health status determined by the corresponding risk factors, such as blood pressure (BP) and hypertension, during adolescence. Since the conditional distributions of the cardiovascular risk factors observed in the NGHS are unknown *a priori* and usually non-Gaussian, statistical inferences for the effects of age, race and obesity on the conditional means of these risk factors may not have meaningful clinical interpretations, because statistical effects on the conditional means may not have direct implications on the distributions of cardiovascular health status for the population of interest. A more meaningful approach is to investigate the statistical effects of age, race and obesity on the distributions of the risk factors, so that various options of clinical interventions may be explored to reduce the chance of developing undesired health status.

4.2 Unstructured Conditional-Distribution Models

Let $F_t[y|\mathbf{X}(t)] = P[Y(t) \leq y|\mathbf{X}(t), t]$ be the conditional cumulative distribution function (CDF) of $Y(t)$ given $\{t, \mathbf{X}(t)\}$. If there is no structure imposed on the relationship between $\{t, \mathbf{X}(t)\}$ and $F_t[y|\mathbf{X}(t)]$, the estimation of $F_t[y|\mathbf{X}(t)]$ based on the longitudinal sample

$\{(Y_{ij}, t_{ij}, \mathbf{X}_{ij}) : i = 1, \dots, n, j = 1, \dots, n_i\}$ may be carried out by extending the kernel methods of Hall, Wolff and Yao (1999) to the current sampling framework. This extension may be straightforward when the number of covariates involved in $\mathbf{X}(t)$ is small, for example, $K = 0$ or 1 . In particular, when $K = 0$, the univariate local logistic method and adjusted Nadaraya-Watson Estimator described in Section 3 of Hall, Wolff and Yao (1999) may be directly applied. When $K \geq 1$, the multivariate estimators described in Section 3.2 of Hall, Wolff and Yao (1999) have to be considered. However, as noted in Hall, Wolff and Yao (1999), when the number of covariates K is large, the multivariate kernel smoothing methods may be computationally unstable due to the well-known problem of “curse of dimensionality” (Fan and Gijbels, 2006, p.264). In addition, statistical results and inferences obtained from a completely unstructured nonparametric estimator for $F_t[y|\mathbf{X}(t)]$ could be difficult to interpret in practical situations. These potential drawbacks, namely computational instability and difficulty in interpretations, often render the unstructured smoothing estimation of $F_t[y|\mathbf{X}(t)]$ impractical. The rest of this section describes the estimation and inference based on a class of structural nonparametric models for $F_t[y|\mathbf{X}(t)]$, the time-varying transformation models.

4.3 Time-Varying Transformation Models

Transformation models with time-to-event data have been studied extensively in survival analysis. Methods of estimation and inference with various right-censored time-to-event data may be found in Cheng, Wei and Ying (1995, 1997), Lu and Ying (2004), Lu and Tsiatis (2006), and Zeng and Lin (2006), among others. In contrast to the conditional-mean based regression models, the transformation models provide a class of functional structures for the conditional CDFs, which has been shown to be an effective dimension-reduction strategy to approximate the conditional distribution functions.

By extending the varying-coefficient approach to the transformation models, Wu, Tian and Yu (2010) suggests that $F_t[y|\mathbf{X}(t)]$ can be modeled by the time-varying linear transformation models of the form:

$$g\{S_t[y|\mathbf{X}(t)]\} = h(y, t) + \mathbf{X}^T(t) \beta(t), \quad (4.1)$$

where $g(\cdot)$ is a known decreasing link function, $S_t[y|\mathbf{X}(t)] = 1 - F_t[y|\mathbf{X}(t)]$, $h(\cdot, \cdot)$ is a unknown baseline function strictly increasing in y , $\beta(t) = (\beta_0(t), \dots, \beta_K(t))^T$, and $\beta_k(t)$ are smooth functions on the time range \mathcal{T} . For each fixed t within the time range, the

model (4.1) is a semiparametric linear transformation model with a nonparametric baseline function $h(y, t)$ and the linear coefficients $(\beta_0(t), \dots, \beta_K(t))^T$. The covariate effect $\beta_k(t)$ at any given t represents the change of $g\{S_t[y|\mathbf{X}(t)]\}$ for all $y \in R$ associated with a unit increase of $X^{(k)}(t)$. Since the effect of y on $F_t[y|\mathbf{X}(t)]$ is summarized in $h(y, t)$, the covariate effects of $X^{(k)}(t)$, $k = 0, \dots, K$, on $F_t[y|\mathbf{X}(t)]$ are constant for all possible values of y when t is fixed. Thus, by imposing a functional linear structure, the model complexity of (4.1) is greatly reduced compared with that of the unstructured nonparametric conditional-distribution models.

It is assumed throughout this section that the form of the link function $g(\cdot)$ is known and chosen by the investigators depending on the goals of the specific analysis. Well-known special cases include the proportional hazard model $g\{S_t[y|\mathbf{X}(t)]\} = \log\{-\log[S_t(y|\mathbf{X}(t))]\}$ and the proportional odds model $g\{S_t[y|\mathbf{X}(t)]\} = -\log\{S_t[y|\mathbf{X}(t)]/F_t[y|\mathbf{X}(t)]\}$. In practice, the fitness of (4.1) may be graphically evaluated by first dividing the time range into small time bins and then examining the linearity of the plots of $g\{S_t[y|\mathbf{X}(t)]\}$ versus $X^{(k)}(t)$ for $k = 0, \dots, K$ and t within all the time bins. Estimation and prediction based on the model (4.1) with unknown link function $g(\cdot)$ deserve substantial investigation, but these results have not been well-established in the present literature. To simplify the notation, each of the n subjects is observed at a randomly selected subset of $J > 1$ distinct design time points $\mathbf{t} = (t_{(1)}, \dots, t_{(J)})^T$. Since not all the subjects are observed at every $t_{(j)}$, \mathcal{S}_j denotes the set of subjects whose observations are available at time $t_{(j)}$, $\mathcal{Z} = \{Y_i(t_{(j)}), \mathbf{X}_i(t_{(j)}), t_{(j)}; i \in \mathcal{S}_j, j = 1, \dots, J\}$ the longitudinal sample of $\{Y(t), \mathbf{X}(t), t \in \mathcal{T}\}$, and $\mathcal{D} = \{\mathbf{X}_i(t_{(j)}), t_{(j)}; i \in \mathcal{S}_j, j = 1, \dots, J\}$ the set of observed covariates. Here $Y_i(t_{(j)})$ and $\mathbf{X}_i(t_{(j)}) = (X_{i0}(t_{(j)}), \dots, X_{iK}(t_{(j)}))^T$ are the outcome and covariate vector, respectively, at $t_{(j)}$ for the i th subject when $i \in \mathcal{S}_j$. Let $m_j = \#\{i \in \mathcal{S}_j\}$ be the number of subjects in \mathcal{S}_j , and $m_{j_1 j_2} = \#\{i \in \mathcal{S}_{j_1} \cap \mathcal{S}_{j_2}\}$ the number of subjects in both \mathcal{S}_{j_1} and \mathcal{S}_{j_2} when $j_1 \neq j_2$. Clearly $m_{j_1 j_2} \leq \min(m_{j_1}, m_{j_2})$.

The main results of this section are focused on the estimation and inference of the time-varying covariate effects $\beta(t)$ in (4.1) and the practical interpretations of the statistical results. In addition to the estimation of $\beta(t)$, nonparametric smoothing estimation and prediction of $h(y, t)$, $F_t[y|\mathbf{X}(t)]$ and their functions, such as the conditional quantiles, also have important applications in biomedical studies. For example, appropriate nonparametric predictors of $F_t[y|\mathbf{x}]$ for any given $\mathbf{X}(t) = \mathbf{x}$ may be used to identify subgroups of the

population who may have excessive risks at different time periods. But, methods and theory for the nonparametric estimation and prediction of $h(y, t)$, $F_t[y|\mathbf{X}(t)]$ and their functions based on (4.1) are currently still under development, hence, are not included in the present discussion. In survival analysis with the linear transformation models, most estimation methods are developed for time-to-event data with random censoring, for example, Cheng, Wei and Ying (1995, 1997). In longitudinal studies, however, repeatedly measured outcome variables and covariates are usually not censored, although censoring remains a theoretical possibility. Thus, in the estimation of $\beta(t)$ in (4.1), none of the variables in the dataset are censored.

4.4 Two-Step Smoothing Methods

As discussed above, when t is fixed, the model (4.1) reduces to the linear transformation model of Cheng, Wei and Ying (1995), so that $\beta(t)$ may be estimated by the estimating equations developed in their paper. When t changes within \mathcal{T} , a useful smoothing method for the estimation of the coefficient curves in (4.1) is to first compute a set of raw estimate the coefficient curves at time design points \mathbf{t} , and then compute the smoothing estimates of the coefficient curves at $t \in \mathcal{T}$ based on the raw estimates at \mathbf{t} .

4.4.1 Raw Estimates of Coefficients

The coefficients $\beta(t_{(j)})$ of (4.1) can be estimated by adapting the estimating equations of Cheng, Wei and Ying (1995) to the observations at $t_{(j)}$. Denote

$$\epsilon_{i(j)} = g\left\{S_{t_{(j)}}\left[Y_i(t_{(j)})\middle|\mathbf{X}_i(t_{(j)})\right]\right\}.$$

It can be verified from

$$P\left[\epsilon_{i(j)} \leq u \middle| \mathbf{X}_i(t_{(j)}), t_{(j)}\right] = P\left\{S_{t_{(j)}}\left[Y_i(t_{(j)})\middle|\mathbf{X}_i(t_{(j)})\right] \geq g^{-1}(u) \middle| \mathbf{X}_i(t_{(j)}), t_{(j)}\right\}$$

and the similar derivation used in the equation (1.4) of Cheng, Ying and Wei (1995), that (4.1) is equivalent to

$$h\left[Y_i(t_{(j)}), t_{(j)}\right] = -\mathbf{X}_i^T(t_{(j)})\beta(t_{(j)}) + \epsilon_{i(j)}, \quad (4.2)$$

where $\epsilon_{i(j)}$ are random errors with distribution function $G(\cdot) = 1 - g^{-1}(\cdot)$. Let

$$Z_{i_1, i_2}(t_{(j)}) = 1_{[Y_{i_1}(t_{(j)}) \geq Y_{i_2}(t_{(j)})]}, \quad \mathbf{X}_{i_1, i_2}(t_{(j)}) = \mathbf{X}_{i_1}(t_{(j)}) - \mathbf{X}_{i_2}(t_{(j)})$$

and $\xi(s) = \int_{-\infty}^{\infty} \{1 - G(t + s)\} dG(t)$. It follows from (4.2) that

$$\begin{aligned} E \left[Z_{i_1, i_2}(t_{(j)}) \middle| \mathbf{X}_{i_1}, \mathbf{X}_{i_2}, t_{(j)} \right] &= P \left\{ h \left[Y_{i_1}(t_{(j)}), t_{(j)} \right] \geq h \left[Y_{i_2}(t_{(j)}), t_{(j)} \right] \middle| \mathbf{X}_{i_1}, \mathbf{X}_{i_2}, t_{(j)} \right\} \\ &= P \left[\epsilon_{i_1, (j)} - \epsilon_{i_2, (j)} \geq \mathbf{X}_{i_1, i_2}^T(t_{(j)}) \beta(t_{(j)}) \right] \\ &= \xi \left[\mathbf{X}_{i_1, i_2}^T(t_{(j)}) \beta(t_{(j)}) \right]. \end{aligned} \quad (4.3)$$

A raw estimator for $\beta(t_{(j)})$ is a solution $\tilde{\beta}(t_{(j)})$ to the estimating equation

$$\sum_{i_1 \neq i_2 \in \mathcal{S}_j} U_{i_1 i_2} \left[\tilde{\beta}(t_{(j)}) \right] = 0, \quad (4.4)$$

where, with $w(\cdot)$ being a known weight function,

$$U_{i_1 i_2} \left[\beta(t_{(j)}) \right] = w \left[\mathbf{X}_{i_1, i_2}^T(t_{(j)}) \beta(t_{(j)}) \right] \mathbf{X}_{i_1, i_2}(t_{(j)}) \left\{ Z_{i_1, i_2}(t_{(j)}) - \xi \left[\mathbf{X}_{i_1, i_2}^T(t_{(j)}) \beta(t_{(j)}) \right] \right\}.$$

It has been shown in Cheng, Wei and Ying (1995, Section 2 and Appendix 1) that, if the weights $w(\cdot)$ are positive, their estimating equation has asymptotically a unique solution, and when $w(\cdot) = 1$ and the matrix $\sum \sum Z_{ij} Z_{ij}^T$ of their equation (2.1) is positive definite, their estimating equation has a unique solution. For each fixed $t_{(j)}$, the estimating equation (4.4) is identical to the estimating equation (2.2) of Cheng, Wei and Ying (1995) without censoring, so that the conclusions of uniqueness and asymptotically uniqueness of the solutions also hold for the estimating equation (4.4). Although the explicit expressions for the finite sample mean and variance of $\tilde{\beta}(t_{(j)})$ are not yet available the asymptotic properties of $\tilde{\beta}(t_{(j)})$ for a fixed time point $t_{(j)}$ are the same as properties developed in Cheng, Wei and Ying (1995). When different time points in \mathbf{t} are involved, the potential intra-subject correlations of the data imply that our raw estimators at different time points are potentially correlated. Large sample approximations of the mean, covariance and variance of $\tilde{\beta}(t_{(j)})$ are presented later in this section.

It is worthwhile to note that the assumption of having J distinct design time points $\mathbf{t} = (t_{(1)}, \dots, t_{(J)})^T$ is a mathematical simplification for the purpose of simplifying the theoretical discussion. In practical biomedical samples, the design time points are usually not prespecified, and \mathbf{t} may be chosen from a practical round-off of the time points based on biological or clinical justifications. When the number of distinct time points is large and there are very few subjects observed at t_j , the raw estimates $\tilde{\beta}(t_j)$ may not exist at some time points $t_{(j)}$. In such situations, a practical approach is to group the observed time points into small time bins with $t_{(j)}$ being the center of the j th time bin, so that the raw estimates

can be computed at each bin. Fan and Zhang (2000) considered this binning approach for a two-step estimation with conditional-mean based varying-coefficient models. When a binning method is used, we require the bin sizes to be small so that the raw estimates are undersmoothed relative to the smoothing parameters computed in the smoothing step. Practical effects of various bin choices deserves have not been systematically investigated. We assume throughout this section that \mathbf{t} already contains the centers of the properly chosen time bins, so that no further binning is necessary. A practical implication of this assumption is that there are sufficient numbers of subjects having observations at all the time points in \mathbf{t} .

4.4.2 Smoothing Estimates of Coefficient Curves

Based on the raw estimates at time points $\mathbf{t} = \{t_{(1)}, \dots, t_{(J)}\}$, a smoothing estimator of $\beta(t)$ can be constructed for all $t \in \mathcal{T}$. To see why the smoothing step is necessary in addition to the raw estimates, it is important to note that the raw estimates are only for the coefficients at $\{t_{(1)}, \dots, t_{(J)}\}$ and the smoothing step computes the curve estimates for all t within \mathcal{T} . In addition, the raw estimates often have large variations over different time design points and such “spiky” estimates generally do not have meaningful biological interpretations. Through the smoothing step, the resultant estimators have reduced variations by sharing information from the adjacent time points.

Applying the least squares kernel approach of Section 3.2 to the raw estimates $\tilde{\beta}(t_{(j)})$ for $j = 1, \dots, J$, a kernel estimator of $\beta_d(t)$ for $0 \leq d \leq K$ can be obtained by

$$\hat{\beta}_d(t) = \frac{\sum_{j=1}^J \tilde{\beta}_d(t_{(j)}) K_h(t - t_{(j)})}{\sum_{j=1}^J K_h(t - t_{(j)})}, \quad (4.5)$$

where $K_h(s) = (1/h)K(s/h)$ and $K(\cdot)$ is a non-negative kernel function. Given the known drawbacks of potentially having large boundary biases for kernel estimators (Fan and Gijbels, 1996), a more preferable smoothing approach for the estimation of $\beta_d(t)$ is through the local polynomial method. More generally, if $\beta_d(t)$ is $(Q + 1)$ times continuously differentiable with respect to t , a local polynomial estimator of the q th derivatives of $\beta_d^{(q)}(t)$ is given by

$$\hat{\beta}_d^{(q)}(t) = \sum_{j=1}^J w_{q, Q+1}(t_{(j)}, t) \tilde{\beta}_d(t_{(j)}), \quad (4.6)$$

where $w_{q, Q+1}(t_{(j)}, t)$ is determined by the smoothing method.

When $q = 0$, $\widehat{\beta}_d(t) = \widehat{\beta}_d^{(0)}(t)$ are local polynomial estimators of $\beta_d(t)$, and specific choices of $w_{q,Q+1}(t_{(j)}, t)$ in (4.6) determine the smoothness and statistical properties of the smoothing estimators. Let $C_j = (1, t_{(j)} - t, \dots, (t_{(j)} - t)^Q)^T$, $C = (C_1, \dots, C_J)^T$, $W_j = K\{(t_{(j)} - t)/h\}$ with $K(\cdot)$ being a non-negative kernel function, and $W = \text{diag}(W_1, \dots, W_J)$. The weight function for the q th order local polynomial estimator of (4.6) is

$$w_{q,Q+1}(t_{(j)}, t; h) = q! e_{q+1,Q+1}^T (C^T W C)^{-1} C_j W_j, \quad j = 1, \dots, J, \quad (4.7)$$

where $e_{q+1,Q+1} = (0, \dots, 0, 1, 0, \dots, 0)^T$ with 1 at its $(q+1)$ th place. This expression, (4.7) is the same as Equation (3.5) of Fan and Zhang (2000).

Coefficient curves in the conditional-mean models may also be estimated by a “one-step smoothing method”, such as Hoover et al. (1998), Lin and Carroll (2000) and Wu and Chiang (2000). Since the estimation of (4.1) is generally obtained through rank-based methods, similar one-step smoothing methods are not yet available for the current setting, because there currently lacks a rank-based smoothing method which does not depend on the initial raw estimation. The two-step smoothing approach is considered in Wu, Tian and Yu (2010) because it is computationally simple. This two-step approach is also capable of automatically adjusting different smoothing needs for different coefficient curves. A potential drawback, however, is that the two-step smoothing approach often requires that there are sufficient observations available at each time design point $t_{(j)}$.

In practice, it is often important to estimate and predict the conditional quantiles and cumulative distributions of the response variable $Y(t)$ at given covariates $\{\mathbf{X}(t), t\}$. Under this circumstance, this implies that there is a need to construct a nonparametric estimator of $h(y, t)$ that is monotone increasing or “order-preserving” in y for all $t \in \mathcal{T}$. In principle, $h(y, t)$ can be estimated by smoothing some raw estimators obtained at the distinct time points $\{t_{(j)}; j = 1, \dots, J\}$, and a set of raw estimators can be constructed using the approach described in Cheng, Wei and Ying (1997). However, as demonstrated in Hall and Müller (2003), the local polynomial method does not automatically lead to “order-preserving” smoothing estimators except for the simple case of kernel estimators with non-negative kernels. The construction of “order-preserving” nonparametric estimators for the conditional quantiles and cumulative distribution functions based on (4.1) require different methods and theory from the ones used in this section.

4.4.3 Bandwidth Choices

The choices of bandwidth h in (4.5) and (4.7) are crucial for obtaining an appropriate smoothing estimator. In practice, subjective bandwidths may be chosen by examining the fitted curves and evaluating the specific clinical settings. Although subjectively chosen bandwidths may lead to scientifically meaningful estimators, data-driven bandwidths are often required as a useful alternative bandwidth choice in practical studies.

For conditional-mean based regression models, a popular choice for selecting bandwidths with longitudinal data is the “deleting-subject” cross-validation (CV) approach, which deletes the entire observations of a subject one at a time, e.g., Hoover et al. (1998) and Wu and Chiang (2000). Extending the CV approach to our setting, we consider here two CV methods for selecting the bandwidths in (4.5) and (4.6) based on the data. The first approach is a direct extension of the “deleting-subject” CV bandwidth, which minimizes the cross-validation score

$$CV_Z(\mathbf{h}) = \sum_{j=1}^J \sum_{i_1 \neq i_2 \in \mathcal{S}_j} \left[Z_{i_1, i_2}(t_{(j)}) - \xi \left\{ \mathbf{X}_{i_1, i_2}^T(t_{(j)}) \widehat{\beta}_{-(i_1, i_2)}(t_{(j)}; \mathbf{h}) \right\} \right]^2, \quad (4.8)$$

where $\widehat{\beta}_{-(i_1, i_2)}(t_{(j)}; \mathbf{h})$ is the two-step local polynomial or kernel estimator computed with all the observations from the subject pair (i_1, i_2) deleted and $\mathbf{h} = (h_0, \dots, h_D)^T$ is the vector of bandwidths for $(\widehat{\beta}_0(t), \dots, \widehat{\beta}_D(t))^T$. In practice, it can be computationally intensive when (4.8) involves a large sample size n . An alternative approach to speed up the computation is to replace (4.8) with a “ M -fold” cross-validation score, which is calculated by deleting a block of subjects each time. To do this, we randomly divide the subjects into M blocks, $\{b(m); m = 1, \dots, M\}$, and compute \mathbf{h}_{CV_Z} which minimizes

$$CV_Z^{(M)}(\mathbf{h}) = \sum_{j=1}^J \sum_{m=1}^M \sum_{i_1 \neq i_2 \in \mathcal{S}_j; (i_1, i_2) \in b(m)} \left[Z_{i_1, i_2}(t_{(j)}) - \xi \left\{ \mathbf{X}_{i_1, i_2}^T(t_{(j)}) \widehat{\beta}_{-b(m)}(t_{(j)}; \mathbf{h}) \right\} \right]^2, \quad (4.9)$$

where $\widehat{\beta}_{-b(m)}(t_{(j)}; \mathbf{h})$ is the two-step kernel or local polynomial estimator computed with the entire block $b(m)$ deleted.

In both (4.8) and (4.9), $CV_Z(\mathbf{h})$ and $CV_Z^{(M)}(\mathbf{h})$ are minimized over $\mathbf{h} = (h_0, \dots, h_K)^T$, which could still be computationally intensive when K is large. Another approach, which relies on a component-wise approach, is to find a CV bandwidth, $h_{d, CV}$, which minimizes

$$CV_{\beta_d}(h_d) = \sum_{j=1}^J \left\{ \tilde{\beta}_d(t_{(j)}) - \widehat{\beta}_{d, -(j)}(t_{(j)}; h_d) \right\}^2, \quad (4.10)$$

for each d with $0 \leq d \leq K$, where $\widehat{\beta}_{d,-(j)}(t_{(j)}, h_d)$ is the smoothing estimator computed with the raw estimate $\widetilde{\beta}_d(t_{(j)})$ at time point $t_{(j)}$ deleted. It is straightforward to see from (4.10) that, by minimizing the CV scores for each d , substantial amount of computation is saved in (4.10) compared with (4.8) or (4.9). A clear implication of (4.10) is that, by deleting raw estimates at each time design point $t_{(j)}$, it ignores the potential intra-subject correlations of the data, and consequently the theoretical properties of (4.10) are potentially different from the CV bandwidths obtained from (4.8) and (4.9). Theoretical properties of these CV procedures have not been systematically investigated.

4.5 Inferences

Asymptotic distributions for the smoothing estimators of the conditional-distribution based regression models under the current setting have not been systematically developed. Given the lack of a reliable asymptotic result, asymptotically approximated statistical inferences for the smoothing estimators of this section have not been established. Two difficult issues for establishing adequate statistical inferences are: (a) correcting the potential biases of the smoothing estimators; (b) quantifying the standard errors of the smoothing estimators. In practice, the estimation biases are often difficult to estimate. However, it has been shown in the literature (e.g., Fan and Zhang, 2000; Huang, Wu and Zhou, 2002) that a “ $\pm Z_{1-\alpha/2}$ standard error band”, which ignores the bias, can often be used to approximate a $[100 \times (1 - \alpha)]\%$ pointwise confidence interval for a smoothing estimator. If the variances of $\widehat{\beta}_d^{(q)}(t)$ given \mathcal{D} can be consistently estimated, say, by $\widehat{\text{var}}\{\widehat{\beta}_d^{(q)}(t)|\mathcal{D}\}$, a “ $\pm Z_{1-\alpha/2}$ standard error band” for $\beta_d^{(q)}(t)$ can be expressed as

$$\widehat{\beta}_d^{(q)}(t) \pm Z_{1-\alpha/2} \left[\widehat{\text{var}}\{\widehat{\beta}_d^{(q)}(t)|\mathcal{D}\} \right]^{1/2}. \quad (4.11)$$

To obtain an appropriate estimate $\widehat{\text{var}}\{\widehat{\beta}_d^{(q)}(t)|\mathcal{D}\}$, a practical approach is to use a “resampling-subject” bootstrap in which a bootstrap sample is generated by randomly resampling n “bootstrap subjects” from the original sample with replacement. Let $\widehat{\beta}_{d,b}^{(q)}(t)$ be the smoothing estimator of $\beta_d^{(q)}(t)$ computed from the b th bootstrap sample. The conditional variance estimator $\widehat{\text{var}}\{\widehat{\beta}_d^{(q)}(t)|\mathcal{D}\}$ are then computed by the sample variances of $\{\widehat{\beta}_{d,b}^{(q)}(t); b = 1, \dots, B\}$.

An alternative approach for constructing an $[100 \times (1 - \alpha)]\%$ confidence intervals for $\beta_d^{(q)}(t)$ without using the normal approximation as in (4.11) is to use the bootstrap quantile

intervals

$$\left(L_{\alpha/2}, U_{\alpha/2}\right), \quad (4.12)$$

where $L_{\alpha/2}$ and $U_{\alpha/2}$ are the lower and upper $[100 \times (\alpha/2)]$ th percentiles of $\{\widehat{\beta}_{d,b}^{(q)}(t); b = 1, \dots, B\}$. Both (4.11) and (4.12) may be used as approximate inference tools in practice. When the sample size is large, such as the examples of Section 1.3, the approximate confidence intervals constructed by (4.11) and (4.12) are similar and may be used interchangeably.

4.6 Applications to the NGHS Data

Applying the conditional-distribution based regression approach to the NGHS data, the age-specific covariate effects of race, height and BMI on the distribution functions of systolic blood pressure (SBP) and diastolic blood pressure (DBP) can be evaluated using the time-varying transformation model (4.1). Before fitting the SBP and DBP data with either a conditional-mean based regression model or a conditional-distribution based regression model, Wu, Tian and Yu (2010) reported in a series of preliminary evaluations for normality, including the Shapiro-Wilk tests, the Kolmogorov-Smirnov tests and visual inspections of the quantile-quantile plots, that the conditional distributions of SBP and DBP for this population of girls given age, race, height (in *dm*) and BMI (in *kg/m²*) were clearly not normal. In particular, many of these conditional distributions were skewed and were rejected for normality by the goodness-of-fit tests at 5% significance level.

The preliminary findings reported in Wu, Tian and Yu (2010) suggest that the existing results obtained in Daniels et al. (1998) or Thompson et al. (2007) may not give an adequate description of the covariate effects on the overall conditional distributions of SBP and DBP. In particular, Daniels et al. (1998) used the conditional-mean regression models, which may not be adequate for describing the effects of these covariates on the probabilities of unhealthy levels of cardiovascular risks, and the conclusions of Thompson et al. (2007) depend on their specific threshold choices for the outcome variables, which may not hold if other threshold values were used. Thus, there are two major advantages for using (4.1) over the conditional-mean based regression analyses. First, it has a flexible and parsimonious structure to summarize the age-specific effects of these covariates on the overall distributions of the outcome variables. Second, statistical inferences for the age-dependent coefficient curves have the same biological interpretations as in Cheng, Wei and Ying (1995) and do

not depend on the normality assumption of the outcome variables or the threshold values for defining unhealthy risk levels.

Because the repeated measurements are obtained for girls between 9 and 19 years of age, the analysis presented here is confined to the age range $\mathcal{T} = [9, 19)$. For a girl at age $t \in \mathcal{T}$, the outcome variable, $Y(t)$, is either the girl's SBP or DBP value at age t , and the covariate vector is $\mathbf{X}(t) = (Race, Height(t), BMI(t))^T$, where $Race = 0$ or 1 if the girl is Caucasian or African-American, and $Height(t)$ and $BMI(t)$ are the girl's height and body mass index, respectively, at age t . Since the visit time for this study is not regularly spaced, it is clinically meaningful to round up the age to one tenth of a year, so that the age range can be grouped into equally spaced bins $[9.0, 9.1), \dots, [18.9, 19.0)$ with time design points $\mathbf{t} = \{9.0, 9.1, 9.2, \dots, 18.9\}$. The i th girl's age, covariates and outcome are denoted by $\{t_j, \mathbf{X}_i(t_j), Y_i(t_j)\}$ if her actual age at the visit falls into the bin $[t_j, t_{j+1})$. This binning scheme is based on the clinical definition of age and the assumption that $Height(t)$, $BMI(t)$ and $Y(t)$ are smooth functions of age t .

In the exploratory analysis reported in Wu, Tian and Yu (2010), the odds-ratios plots of SBP with a range of threshold values under various strata of age, race, height and BMI showed that the proportional odds models gave reasonable approximations to the relationships between the conditional distribution functions of SBP and the covariates $\mathbf{X}(t)$. Thus, for this analysis, the following time-varying proportional odds model is considered:

$$\begin{aligned} & \log \left\{ \frac{P[Y(t) > y | t, Race, Height, BMI]}{P[Y(t) \leq y | t, Race, Height, BMI]} \right\} \\ & = -h(y, t) + \beta_0(t) \times Race + \beta_1(t) \times Height(t) + \beta_2(t) \times BMI(t), \end{aligned} \quad (4.13)$$

where a positive (negative) value for $\beta_0(t)$ suggests that African-American girls tend to have higher (lower) SBP or DBP values than Caucasian girls at age t , and $\beta_1(t)$ and $\beta_2(t)$ represent the changes of the log-odds of $Y(t) > y$ associated with a unit increase of $Height(t)$ and $BMI(t)$, respectively, at age t .

Following the estimation procedure of Section 4.4, the raw estimates are first computed at the time design points $\mathbf{t} = \{9.0, 9.1, 9.2, \dots, 18.9\}$ with the $w(\cdot) = 1$ weight, and then the smoothing estimates $\hat{\beta}(t) = (\hat{\beta}_0(t), \hat{\beta}_1(t), \hat{\beta}_2(t))^T$ are computed using the local linear method with the Epanechnikov kernel and the cross-validated bandwidths chosen from (4.9) and (4.10). For the M -block CV bandwidths, ten subject blocks, $M = 10$, are used to minimize the CV score (4.9). Figure 5 shows the two-step local linear estimators of

$\beta(t) = (\beta_0(t), \beta_1(t), \beta_2(t))^T$ computed from the CV bandwidths based on (4.10) and the “ ± 1.96 standard error bands” computed from 500 bootstrap repetitions. In order to ease the computational burden, the CV bandwidths obtained from the original sample are used for the bootstrap standard error bands. The covariate effects shown in Figure 5 are similar for both SBP and DBP. The contribution of race $\beta_0(t)$ is close to zero and slightly increases with age, suggesting that African-American girls tend to have slightly higher odds of SBP and DBP than Caucasian girls at later years. The positive estimates of $\beta_1(t)$ and $\beta_2(t)$ suggest that both height and BMI contribute positively to the odds of SBP and DBP. The effects $Height(t)$ and $BMI(t)$ on the conditional distribution of SBP appear to decrease as the girls are getting older. The $BMI(t)$ seems to have different effects on the conditional distribution of SBP and the conditional distribution of DBP. For the conditional distribution of SBP, the effects of $BMI(t)$, $\beta_2(t)$, appears decreasing linearly as t increases. For the conditional distribution of DBP, the effects of $BMI(t)$, $\beta_2(t)$, appears to be a nonlinear function of the girl’s age t .

4.7 Discussion and Potential Extensions

The time-varying linear transformation models discussed in this section use a conditional-distribution based regression approach for modeling the entire conditional distribution functions and evaluating the covariate effects on these distributions. This approach is conceptually superior to the conditional-mean based models when the scientific objective depends on the conditional distribution functions. As illustrated in the application to the NGHS blood pressure data, the practical advantages of modeling the conditional distribution functions may be generalized to other typical biomedical studies. Similar to the conditional-mean based varying-coefficient models, the time-varying transformation models effectively reduce the problem of “curse-of-dimensionality” and at the same time retain a high degree of model flexibility. From the computational point of view, the two-step estimation procedure is conceptually simple and can be carried out by combining the existing rank-based estimation procedures in survival analysis and the smoothing procedures in nonparametric curve estimation.

Given the practical implications of conditional-distribution based regression models, further research based on this approach is warranted. First, notice that, in the application to the NGHS blood pressure data, the fitness of the model (4.1) is examined in an *ad hoc* fashion

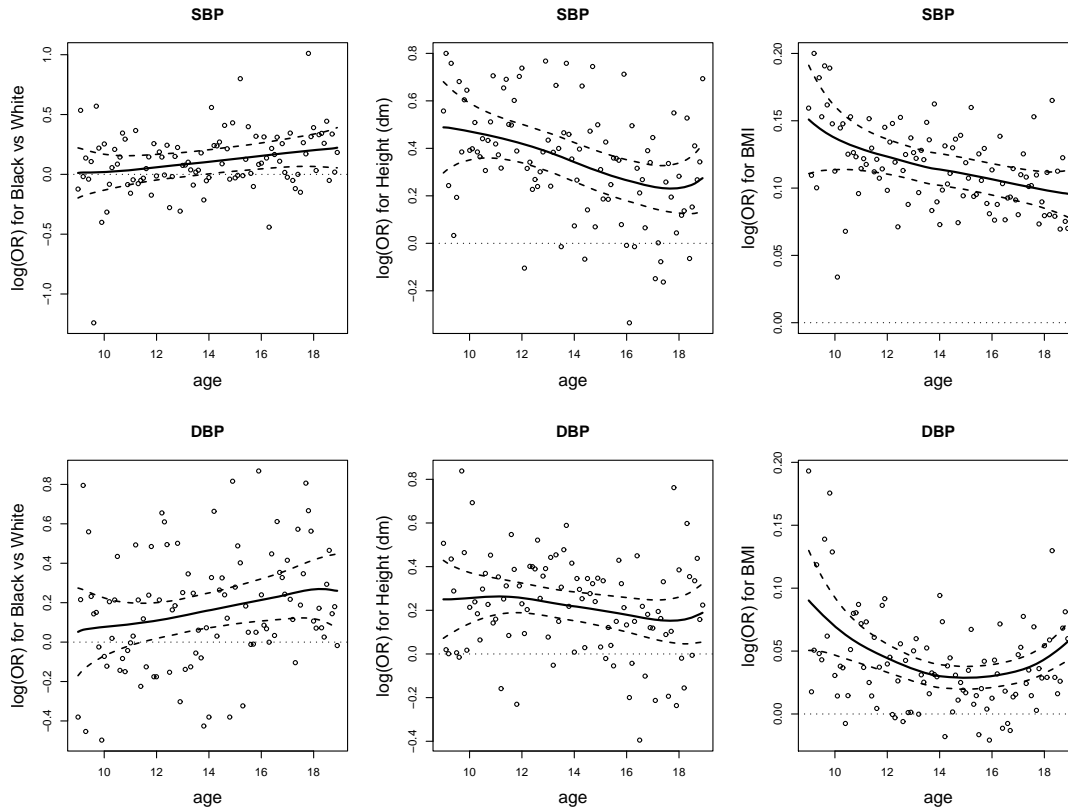


Figure 5: Top panels: systolic blood pressure (SBP). Bottom panels: diastolic blood pressure (DBP). Each row shows the raw estimates, the two-step local linear estimates (solid lines) computed with the CV bandwidths from (4.10), and their ± 1.96 bootstrap standard error bands (dashed lines) for $\beta_0(t)$, $\beta_1(t)$ and $\beta_2(t)$ of (4.1).

by a series of exploratory plots. More rigorously, appropriate goodness-of-fit tests should be developed to formally test the adequacy of the time-varying models under appropriate statistical hypotheses. Second, the model (4.1) considered in this section involves only a univariate outcome variable. In many situations the scientific objective is better described through the joint distributions of a multivariate outcome variable. Extensions to conditional distributions with multivariate outcome variables warrant further investigation. Third, the two-step procedure discussed in this section is limited to the estimation of the covariate effects over time. Smoothing methods for the estimation and prediction of conditional distributions and conditional quantiles have useful applications in longitudinal studies and should be developed as well.

5 Regression Methods for Outcome-Adaptive Covariates

It is well-known in the literature (for example, Pepe and Anderson, 1994) that when the values of a covariate depend on the outcome values at the previous time points, the wide range of conditional-mean based regression methods of Section 2 and Section 3 may lead to unsatisfactory results. This type of covariates, whose values may depend on the previous outcome values, are referred herein as “outcome-adaptive” covariates. As a special case of longitudinal analysis with “outcome-adaptive” covariates, this section presents the modeling and estimation approaches for longitudinal data with the presence of a concomitant intervention. After a brief introduction of data with a concomitant intervention, the rest of this section summarizes the similarities and differences between the modeling approaches of Wu, Tian and Bang (2008) and Wu, Tian and Jiang (2011) and their corresponding estimation and inference procedures.

5.1 Outcome-Adaptive Covariates

For longitudinal clinical trials with randomly assigned study treatments, longitudinal effects of the study treatments are modeled through a time-invariant categorical covariate vector, while other factors of interest, such as age, gender, ethnicity and disease risk factors, can be modeled through either time-invariant or time-dependent covariates. In many longitudinal studies, some time-dependent covariates are “outcome-adaptive” in the sense that their potential values at a time point may depend on the values or time-trends of the outcome variable prior to that time point.

A typical scenario which involves “outcome-adaptive” covariates is the presence of concomitant interventions in longitudinal clinical trials. Unlike the study treatments of a clinical trial which are randomly assigned to the study subjects, concomitant interventions are not randomly assigned, because they are initiated, usually due to ethical reasons, to study subjects who exhibit less satisfactory trends in their health outcomes. The scenario of concomitant interventions bears some similarities to longitudinal studies with informative missing data. In the case of informative missing data, the study subjects with undesirable outcome time-trends tend to drop out from the study earlier than those with more desirable outcome time-trends. The only difference is that, in studies with concomitant interventions, the outcomes of the study subjects continue to be observed after the start of the concomitant interventions. In a randomized longitudinal clinical trial with randomly assigned study treatments, study subject who have taken a concomitant intervention in addition to their assigned study treatments may generally have different disease pathology from those who do not need the concomitant intervention. Thus, in addition to the primary objective of evaluating the effects of the study treatments, an important secondary objective is to evaluate the effects of the additional concomitant interventions on the outcome variables of the study populations.

The Enhancing Recovery in Coronary Heart Disease (ENRICHD) Study described in Example 4 is a typical example which involves a concomitant intervention in addition to the randomly assigned treatment regimens. In this randomized clinical trial for evaluating the efficacy of a six-month cognitive behavior therapy (CBT) versus usual cardiovascular care (UC), the Beck Depression Inventory (BDI) scores for patients in the CBT arm were repeatedly measured at weekly visits during the treatment and four yearly follow-up visits, while BDI scores for patients in the UC arm were only measured at baseline, the six-month visit and the yearly follow-up visits. By the study design (ENRICHD, 2001), pharmacotherapy with antidepressants was allowed as a concomitant intervention in both the CBT and the UC arms if a patient had high baseline BDI scores or nondecreasing BDI trends five weeks after enrollment or antidepressants were requested by the patient or the primary-care physicians. Although Taylor et al. (2005) reported that pharmacotherapy improved survival among 1834 depressed ENRICHD patients, their results, however, did not address the question of whether pharmacotherapy was beneficial for lowering the patients’ depression severity.

Using the repeatedly measured BDI scores in a subsample of 91 ENRICHHD patients in the CBT arm who received pharmacotherapy within the treatment period, Wu, Tian and Bang (2008) showed that the naïve mixed-effects models gave misleading results for the pharmacotherapy effects on the BDI trends over time, and proposed a varying-coefficient mixed-effects model to reduce the potential bias associated with the estimated pharmacotherapy effects. A main drawback of Wu, Tian and Bang (2008) is the potential loss of information because their regression model can not be applied to patients who have already received pharmacotherapy at baseline or have not received pharmacotherapy during the study.

As a generalization of the varying-coefficient mixed-effects model, Wu, Tian and Jiang (2011) proposed a comprehensive regression method for evaluating the concomitant intervention effects which is capable to incorporate information from all the study subjects in a longitudinal study. Using the framework of shared-parameter models in Follmann and Wu (1995), the approach of Wu, Tian and Jiang (2011) describes the covariate effects on the response variable through a change-point mixed-effects model, and incorporates the random coefficients and the intervention starting time (change-point time) through a series of joint distributions. Patients who have received a concomitant intervention at baseline or have not received any concomitant intervention during the study period are treated as censored. A likelihood-based method is established for statistical estimation and inferences, and its computation is implemented through a two-stage iteration procedure. Applying their procedures to the ENRICHHD pharmacotherapy data, the results of Wu, Tian and Jiang (2011) suggest that their proposed method leads to adequate estimates when a concomitant intervention is present, while the naïve mixed-effects model is likely misspecified under such situations.

5.2 Structure for Data with One Concomitant Intervention

Following the notation of Section 1.2, let \mathcal{T}_0 and \mathcal{T}_1 be the beginning and ending times of a study, and n be the total number of randomly selected subjects. The i th subject has n_i visits and observations $(T_{ij}, Y_{ij}, \mathbf{X}_i)$ at the j th visit, where T_{ij} , the study time, is the time elapsed from the beginning of the study to the j th visit, \mathbf{X}_i is a time-invariant covariate vector, and Y_{ij} is the real-valued outcome variable. For simplicity, we assume throughout the section that the study involves only one concomitant intervention, and the i th subject can change

from “without concomitant intervention” to “concomitant intervention” only once during the study with S_i being the concomitant intervention starting time or change-point time. Let $\lambda_{ij} = 0$ or 1 , if $T_{ij} \leq S_i$ or $T_{ij} > S_i$, respectively, be the concomitant intervention indicator for the i th subject. Since not every subject has a change-point time during the study, the i th subject’s change-point time is observed if $T_{i1} \leq S_i \leq T_{in_i}$. If $S_i < T_{i1}$ or $S_i > T_{in_i}$, the subject’s change-point time is left or right censored, respectively. The indicator variable for censoring $\delta_i^{(c)}$ is defined by $\delta_i^{(c)} = 0$ if $T_{i1} \leq S_i \leq T_{in_i}$, 1 if $S_i > T_{in_i}$, and 2 if $S_i < T_{i1}$. The observed change-point times are $\{\mathcal{S}_i^{(c)} = (S_i^{(c)}, \delta_i^{(c)}); i = 1, \dots, n\}$, where $S_i^{(c)} = S_i$ if $\delta_i^{(c)} = 0$, T_{in_i} if $\delta_i^{(c)} = 1$, and T_{i1} if $\delta_i^{(c)} = 2$.

5.3 Model Formulations and Interpretations

5.3.1 N ave Mixed-Effects Change-Point Models

Since a concomitant intervention is not randomly assigned, it is understood in practice that the effects of a concomitant intervention can not be properly evaluated by directly comparing the outcome values between the subjects who received the intervention and those who did not receive the intervention. This fact has been noted in ENRICHD (2003) and Taylor et al (2005) for the ENRICHD Study. A better approach that has been suggested by Wu, Tian and Bang (2008) and Wu, Tian and Jiang (2011) is to use a change-point model, which assumes that the outcome variable Y_{ij} follows different trajectories before and after the concomitant intervention, so that the effects of the concomitant intervention can be evaluated through the differences of the mean trajectories.

Let $\mu_0(T_{ij}, \mathbf{X}_i; \mathbf{a}_i)$ be the i th subject’s trajectory before the concomitant intervention, which is parameterized by the subject-specific parameter $\mathbf{a}_i^T = (a_{i1}, \dots, a_{id_0})^T$, $d_0 \geq 1$, and let $\mu_1(T_{ij}, \mathbf{X}_i, R_{ij}; \mathbf{b}_i)$ be the change of the trajectory after the concomitant intervention, which is parameterized by the subject-specific parameter $\mathbf{b}_i = (b_{i1}, \dots, b_{id_1})^T$, $d_1 \geq 1$, and may depend on the “intervention duration time” $R_{ij} = T_{ij} - S_i$ as well as T_{ij} and \mathbf{X}_i . The usual mixed-effects model framework (Davadian and Giltinan, 1995; Verbeke and Molenberghs, 2000; Diggle et al., 2002) suggests that a n ave mixed-effects model for evaluating the pre- and post-intervention trajectories can be expressed as

$$\begin{cases} Y_{ij} = \mu_0(T_{ij}, \mathbf{X}_i; \mathbf{a}_i) + \lambda_{ij} \mu_1(T_{ij}, \mathbf{X}_i, R_{ij}; \mathbf{b}_i) + \epsilon_{ij}, \\ (\mathbf{a}_i^T, \mathbf{b}_i^T)^T \sim \text{joint distribution } G(\cdot), \end{cases} \quad (5.1)$$

where $(\mathbf{a}_i^T, \mathbf{b}_i^T)^T$ has unknown mean $(\alpha^T, \beta^T)^T$ and covariance matrix Σ , ϵ_{ij} are mean zero

random errors with $cov(\epsilon_{ij_1}, \epsilon_{ij_2}) = \sigma_{ij_1j_2}$, and $\epsilon_{i_1j_1}$ and $\epsilon_{i_2j_2}$ are independent if $i_1 \neq i_2$. For mathematical convenience, the joint distribution of \mathbf{a}_i and \mathbf{b}_i may be assumed to be multivariate Gaussian $\mathcal{N}\{(\alpha^T, \beta^T)^T, \Sigma\}$. Under (5.1), a positive (or negative) value for $\mu_1(T_{ij}, \mathbf{X}_i, R_{ij}; \mathbf{b}_i)$ would suggest that the intervention tends to increase (or decrease) the mean of Y_{ij} given $(T_{ij}, \mathbf{X}_i, R_{ij})$.

5.3.2 An Example for the Biases of Näive Mixed-Effects Models

Because the “self-selectiveness” of the intervention is ignored, Wu, Tian and Bang (2008) shows that (5.1) can be a misspecified model even if $\mu_0(\cdot; \mathbf{a}_i)$, $\mu_1(\cdot; \mathbf{b}_i)$, ϵ_{ij} and $G(\cdot)$ are correctly specified. Sometimes misleading conclusions may occur even under simple situations where (5.1) appears to have natural interpretations. The following simple synthetic example illustrates the potential erroneous conclusions which may result from the standard estimation procedures of a naïve mixed-effects model.

Suppose that a longitudinal study has $n = 24$ independent subjects, and, for $1 \leq i \leq 24$, $j = 1, \dots, n_i$, $n_i = 10$ and $T_{ij} = j$. Let S_i be the change-point time for the i th subject changing from “no concomitant intervention” to “with concomitant intervention”, and let λ_{ij} be the corresponding “concomitant intervention indicator” such that $\lambda_{ij} = 0$ if $S_i < T_{ij}$, and $\lambda_{ij} = 1$ if $S_i \geq T_{ij}$. For the first 12 subjects, i.e., $1 \leq i \leq 12$, their change-point time is $S_i = 2$, and their outcomes are generated by $Y_{ij} = 20 + e_{ij}$ for $\lambda_{ij} = 0$ and $1 \leq j \leq 2$, and $Y_{ij} = 19 + e_{ij}$ for $\lambda_{ij} = 1$ and $3 \leq j \leq 10$, where e_{ij} are independent identically distributed with the $N(0, 3)$ distribution. For the remaining 12 subjects, i.e., $13 \leq i \leq 24$, their change-point time is $S_i = 8$, and their outcomes are generated by $Y_{ij} = 19 + e_{ij}$ for $\lambda_{ij} = 0$ and $1 \leq j \leq 8$, and $Y_{ij} = 17 + e_{ij}$ for $\lambda_{ij} = 1$ and $9 \leq j \leq 10$, where e_{ij} are independent identically distributed with the $N(0, 3)$ distribution. Suppose that it is known that Y_{ij} does not depend on T_{ij} for all $1 \leq i \leq 24$ and $1 \leq j \leq 10$, T_{ij} . Then, a special case of the naïve mixed-effects change-point model of (5.1) which may be considered for the current situation is

$$\begin{cases} Y_{ij} = a_i + b_i \lambda_{ij} + \epsilon_{ij}, \\ (a_i, b_i)^T \sim \text{joint distribution } G(\cdot), \end{cases} \quad (5.2)$$

where, for all $1 \leq i \leq 24$ and $1 \leq j \leq 10$, ϵ_{ij} are independent identically distributed with the $N(0, 3)$ distribution, $E(a_i) = \alpha$ and $E(b_i) = \beta$ are the fixed-effects, and $G(\cdot)$ is an unknown distribution. Since, for $1 \leq i \leq 12$, the true effect of the concomitant intervention is $b_i = -1$, and, for $13 \leq i \leq 24$, the true effect of the concomitant intervention is $b_i = -2$,

the true fixed mean effect for the population is $\beta = -1.5$, which is unknown and needs to be estimated from the data.

Given a sample $\{(Y_{ij}, T_{ij}, S_i); i = 1, \dots, 24, j = 1, \dots, 10\}$ generated from the above specification, the mean and covariance parameters of (5.1) can be estimated using a number of standard statistical analysis software packages, such as SAS and R. Although the exact distribution function $G(\cdot)$ is generally unknown, adequate estimates of the mean parameters $\{\alpha, \beta\}$ may often be obtained in practice by assuming $G(\cdot)$ to be a multivariate normal distribution with a suitable correlation structure. Since the correlation structures of the data are generally unknown, three LME procedures in the R statistical package can be used to estimate $\{\alpha, \beta\}$: LME with working independent correlation structure (LMEWI), LME with random intercept (LMERI), and LME with random intercept and slope (LMERIS). Similarly, parameter estimates can also be obtained using the generalized estimation equation procedure with three correlation structures: GEE with working independent correlation structure (GEEWI), GEE with exchangeable correlation structure (GEEEC), and GEE with unstructured correlation structure (GEEUC). Further details for the parameter estimations with linear mixed-effects models can be in Verbeke and Molenberghs (2000).

To examine whether the naïve mixed-effects change-point model (5.2) can lead to appropriate estimates for the mean concomitant intervention effect β , 10,000 independent samples of $\{(Y_{ij}, T_{ij}, S_i); i = 1, \dots, 24, j = 1, \dots, 10\}$ were generated in a simulation study. The estimators $\hat{\beta}$ and their standard errors were computed from each of the simulated samples using each of the estimation procedures, namely LMEWI, LMERI, LMERIS, GEEWI, GEEEC and GEEUC, in SAS. Table 2 summarizes the averages of the estimators and their standard errors (SE) and the empirical coverage probabilities of the corresponding 95% confidence intervals covering the true parameter $\beta = -1.5$ computed from the 10,000 simulated samples and the naïve mixed-effects model (5.2). These results suggest that all these LME and GEE procedures with different correlation structure assumptions give similar estimates for β , which are around -0.6 and -0.7 and far from the true value of $\beta = -1.5$, with comparable standard errors. Consequently, the 95% confidence intervals shown in Table 2 have low empirical coverage probabilities, which suggests that the model (5.2) leads to excessive biases and inadequate estimates for β .

Are there simple approaches which can lead to better estimates of the mean concomitant intervention effect β ? Obviously, for the simple setup considered here, one can consider

Table 2: Averages of the parameter estimates and their standard errors (SE) and the empirical coverage probabilities of their corresponding 95% confidence intervals (CI) covering the true parameter $\beta = -1.5$ computed from 10,000 simulated samples with the naïve mixed-effects change-point model (5.2)

R Procedure with Correlation Structure	Estimate	SE	Empirical Coverage Probability of 95% CI
LME Working Independence	-0.598	0.395	37.8%
LME Random Intercept	-0.731	0.403	51.0%
LME Random Intercept & Slope	-0.788	0.431	60.4%
GEE Working Independence	-0.598	0.385	36.2%
GEE Exchangeable	-0.709	0.385	46.7%
GEE Unstructured	-0.752	0.572	43.9%

an “individual fitting” estimator of β based on (5.2), which is given by

$$\hat{\beta}_{ind} = (1/n) \sum_{i=1}^n \left\{ \frac{\sum_{j=1}^{n_i} (Y_{ij} 1_{[\delta_{ij}=1]})}{\sum_{j=1}^{n_i} 1_{[\delta_{ij}=1]}} - \frac{\sum_{j=1}^{n_i} (Y_{ij} 1_{[\delta_{ij}=0]})}{\sum_{j=1}^{n_i} 1_{[\delta_{ij}=0]}} \right\}. \quad (5.3)$$

A simple application of (5.3) to our simulated samples led to estimates of β very close to its true value of -1.5. However, it is known in the literature that “individual fitting” estimation methods based on the repeated measurements separately from each subject, such as (5.3), are generally less efficient than the well-known procedures, such as MLEs, REML estimates and GEEs (e.g., Verbeke and Molenberghs, 2000; Diggle et al., 2002). In addition, it is also difficult to generalize the “individual fitting” approaches to regression models with more complicated terms and patterns than the simple cases exhibited in (5.2).

Comparing the underlying mechanism for generating $\{(Y_{ij}, T_{ij}, S_i); i = 1, \dots, 24, j = 1, \dots, 10\}$ with the model (5.2), one potential flaw of using (5.2) with its LME and GEE estimation procedures is that the model does not take the potential relationship between the change-point time S_i and the values of Y_{ij} before the change-point time, i.e., the values of Y_{ij} when $\delta_{ij} = 0$. Indeed, under the real data generating mechanism, subjects with subject-specific mean value of Y_{ij} at $j = 1$ to be 20 have change-point time at $S_i = 2$, while subjects with subject-specific mean value of Y_{ij} at $j = 1$ to be 19 have change-point time at $S_i = 8$. Although this fact is unknown at the estimation stage, the possibility of the potential relationship between S_i and Y_{ij} when δ_{ij} is not allowed in the MLEs, REML estimates and GEEs based on the naïve model (5.2). To see whether the potential bias for

the estimation of β could be reduced by incorporating the change-point time S_i into the model, we consider here the following simple generation of (5.2):

$$\begin{cases} Y_{ij} = a_{0i} + a_{1i}S_i + b_i\lambda_{ij} + \epsilon_{ij}, \\ (a_{0i}, a_{1i}, b_i)^T \sim N((\alpha_0, \alpha_1, \beta)^T, \Gamma), \end{cases} \quad (5.4)$$

where $(\alpha_0, \alpha_1, \beta)^T$ is the vector of mean parameters and Γ is the unknown covariance matrix. The structures of Γ generally do not have major influences on the estimation of $(\alpha_0, \alpha_1, \beta)^T$, and many commonly used parametric structures may be used when implementing the estimation procedures (Diggle et al., 2002). It is important to note that the interpretation of β is the same in both (5.2) and (5.4). Strictly speaking, the model (5.4) is not correct for the underlying data generating mechanism, since Y_{ij} does not depend on S_i through a simple linear model. Although (5.4) is at best a rough approximation of the true underlying data generating mechanism, the main intent here is to evaluate whether the bias for the estimation of the concomitant intervention effect can be reduced by incorporating S_i into the model. Details on the justifications of (5.4) and interpretations of its parameters are given in Section 5.4, where it is described as a special case of the varying-coefficient mixed-effects models of Wu, Tian and Bang (2008).

Table 3: Averages of the parameter estimates and their standard errors (SE) and the empirical coverage probabilities of their corresponding 95% confidence intervals (CI) covering the true parameter $\beta = -1.5$ computed from 10,000 simulated samples with the mixed-effects model (5.4)

R Procedure with Correlation Structure	Estimate	SE	Empirical Coverage Probability of 95% CI
LME Working Independence	-1.498	0.485	94.7%
LME Random Intercept	-1.498	0.483	94.6%
LME Random Intercept & Slopes	-1.498	0.496	95.2%

Table 3 summarizes the averages of the estimators and their standard errors (SE) and the empirical coverage probabilities of the corresponding 95% confidence intervals covering the true parameter $\beta = -1.5$ computed from the 10,000 simulated samples, the mixed-effects model (5.4) and the same LME procedures as the ones used in Table 2. The mean estimates for β in Table 3 are very close to the true value of $\beta = -1.5$, with comparable standard errors in both Table 2 and Table 3. The 95% confidence intervals shown in Table

3 have empirical coverage probabilities which are very close to the nominal level of 95% and much higher than the ones shown in Table 2. Clearly, (5.4) leads to much smaller bias for the estimation of β than (5.2).

5.3.3 General Shared-Parameter Models

The data structure of Section 5.2 is a special case of outcome-adaptive covariates, which involves only one concomitant intervention and each study subject has at most one change-point from “without concomitant intervention” to “concomitant intervention”. This simple structure of outcome-adaptiveness suggests that a natural extension for the mixed-effects approach of (5.1) is to incorporate the initiation of the concomitant intervention into the regression model. This extension of (5.1) can be achieved by allowing the intervention starting time S_i to be correlated with the pre-intervention random coefficients \mathbf{a}_i or more generally $\{\mathbf{a}_i, \mathbf{b}_i\}$. Let $\mu_0(\cdot; \mathbf{a}_i)$ and $[\mu_0(\cdot; \mathbf{a}_i) + \mu_1(\cdot; \mathbf{b}_i)]$ be the subject-specific response curves before and after the start of the concomitant intervention, respectively. With $\mu_1(\cdot; \mathbf{b}_i)$ being interpreted as the concomitant intervention effect, a shared-parameter change-point model for the given dataset $\{Y_{ij}, T_{ij}, \mathbf{X}_i, S_i\}$ is

$$\begin{cases} Y_{ij} = \mu_0(T_{ij}, \mathbf{X}_i; \mathbf{a}_i) + \delta_{ij} \mu_1(T_{ij}, \mathbf{X}_i, R_{ij}; \mathbf{b}_i) + \epsilon_{ij}, \\ (\mathbf{a}_i^T, \mathbf{b}_i^T, S_i)^T \sim \text{Joint Distribution}, \end{cases} \quad (5.5)$$

where $R_{ij} = T_{ij} - S_i$, ϵ_{ij} are mean zero errors with $\text{cov}(\epsilon_{ij_1}, \epsilon_{ij_2}) = \sigma_{ij_1 j_2}$, $\epsilon_{i_1 j_1}$ and $\epsilon_{i_2 j_2}$ are independent if $i_1 \neq i_2$, and, conditioning on $\{\mathbf{a}_i, \mathbf{b}_i\}$, S_i and $\{T_{ij}, \mathbf{X}_i\}$ are independent. In addition, we assume that $\{\mathbf{a}_i, \mathbf{b}_i\}$ and $\{T_{ij}, \mathbf{X}_i\}$ are independent. Using the matrix representation $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$ and $\mathbf{T}_i = (T_{i1}, \dots, T_{in_i})^T$, the joint likelihood function of $(\mathbf{Y}_i^T, S_i)^T$ given $\{\mathbf{T}_i, \mathbf{X}_i\}$ based on (5.5) is

$$f(\mathbf{Y}_i, S_i | \mathbf{T}_i, \mathbf{X}_i) = \int f(\mathbf{Y}_i | \mathbf{T}_i, \mathbf{X}_i, S_i, \mathbf{a}_i, \mathbf{b}_i) f(S_i | \mathbf{a}_i, \mathbf{b}_i) dH(\mathbf{a}_i, \mathbf{b}_i), \quad (5.6)$$

where $f(\cdot | \cdot)$ denotes the conditional density and $H(\cdot, \cdot)$ is the joint distribution function of $\{\mathbf{a}_i, \mathbf{b}_i\}$. The extra $f(S_i | \mathbf{a}_i, \mathbf{b}_i)$ in the integrand distinguishes (5.6) from the usual likelihood functions for the mixed-effects models (Verbeke and Molenberghs, 2000, p24).

In (5.5), the parameters $\{\mathbf{a}_i, \mathbf{b}_i\}$ are associated with both the response curves of Y_{ij} and the distribution of S_i . The shared parameters approach was proposed in Follmann and Wu (1995) for the purpose of modeling the behaviors of informative missing data. However, in (5.5), the subjects are still being observed after the change-point time. The correlation

between S_i and \mathbf{a}_i suggests that the i th subject's change-point time is affected by the pre-intervention response curve $\mu_0(\cdot)$, and the correlation between S_i and \mathbf{b}_i suggests that S_i may also influence the response curve $\mu_1(\cdot)$, which characterizes the intervention effects.

5.4 Varying-Coefficient Mixed-Effects Models - A Special Case

5.4.1 Formulation of Varying-Coefficient Mixed-Effects Models

The joint likelihood function (5.6) requires a known distribution function for $\{\mathbf{a}_i, \mathbf{b}_i, S_i\}$, and maximizing this likelihood can be computationally intensive in practice. When all the subjects have observed change-point times within the study period, that is, $\mathcal{T}_0 < S_i < \mathcal{T}_1$ for all $1 \leq i \leq n$, a simpler regression method, which does not depend on the distribution function of S_i , may be considered. Since S_i are observed for all $1 \leq i \leq n$, we can consider the conditional distribution

$$f(\mathbf{Y}_i | S_i, T_i, \mathbf{X}_i) = \int f(\mathbf{Y}_i | T_i, \mathbf{X}_i, S_i, \mathbf{a}_i, \mathbf{b}_i) dG(\mathbf{a}_i, \mathbf{b}_i | S_i), \quad (5.7)$$

and rewrite (5.5) as a varying-coefficient model using the conditional distribution of $\{\mathbf{a}_i, \mathbf{b}_i\}$ given S_i . Although $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are allowed to take general parametric or nonparametric forms, this approach is illustrated here assuming that $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are linear functions of the form $\mu_0(T_{ij}, \mathbf{X}_i; \mathbf{a}_i) = \mathbf{Z}_{ij}^T \mathbf{a}_i$ and $\mu_1(T_{ij}, \mathbf{X}_i, S_i; \mathbf{b}_i) = \mathbf{W}_{ij}^T \mathbf{b}_i$, where $\mathbf{Z}_{ij} = (Z_{ij0}, \dots, Z_{ijD_1})^T$ is generated by $\{(T_{ij}, \mathbf{X}_i); 1 \leq j \leq n_i, \delta_{ij} = 0\}$, and $\mathbf{W}_{ij} = (W_{ij0}, \dots, W_{ijD_2})^T$ is generated by $\{(T_{ij}, \mathbf{X}_i, S_i); 1 \leq j \leq n_i, \delta_{ij} = 1\}$.

Let $\alpha(S_i) = E(\mathbf{a}_i | S_i)$, $\beta(S_i) = E(\mathbf{b}_i | S_i)$, $\mathbf{a}_i^* = \mathbf{a}_i - \alpha(S_i)$ and $\mathbf{b}_i^* = \mathbf{b}_i - \beta(S_i)$. A varying-coefficient mixed-effects model for the data $\{Y_{ij}, T_{ij}, S_i; 1 \leq i \leq n, 1 \leq j \leq n_i\}$ is

$$\begin{cases} Y_{ij} = \mathbf{Z}_{ij}^T [\alpha(S_i) + \mathbf{a}_i^*] + \delta_{ij} \mathbf{W}_{ij}^T [\beta(S_i) + \mathbf{b}_i^*] + \epsilon_{ij}, \\ \left(\mathbf{a}_i^{*T}, \mathbf{b}_i^{*T} \right)^T \Big| S_i \sim G(\cdot | S_i) \end{cases} \quad (5.8)$$

where, for $S_i = s$, $G(\cdot | s)$ is a distribution function with mean zero and covariance matrix $cov[(\mathbf{a}_i^{*T}, \mathbf{b}_i^{*T})^T | s] = \mathbf{C}(s)$. The population-mean parameters of interest are $\alpha(s)$ and $\beta(s)$, which, in this case, are both smooth functions of s . When $S_i = s$, the mean concomitant intervention effect is $\beta(s)$. The special choice of $\beta(s) = 0$ for all $s \in (\mathcal{T}_0, \mathcal{T}_1)$ implies that the concomitant intervention has no population-mean effect on the time-trend curve of Y_{ij} .

A number of interesting special cases of (5.8) may be considered in real applications by specifying the forms of $\alpha(\cdot)$, $\beta(\cdot)$ and $G(\cdot | \cdot)$. An obvious choice for $G(\cdot | S_i)$ is the multivariate normal distribution with mean zero and covariance matrix $\mathbf{C} = cov[(\mathbf{a}_i^{*T}, \mathbf{b}_i^{*T})^T | s]$, which is

assumed to be time-invariant for simplicity. Extension to time-dependent covariances can be made by modeling $\mathbf{C}(s)$. Since the main objective is to evaluate the population-mean effects of the concomitant intervention and explicit forms of $G(\cdot|S_i)$ are often unknown, using appropriate models for $\alpha(s)$ and $\beta(s)$ is often more important than using a suitable model for $\mathbf{C}(s)$. Linear models for $\alpha(s)$ and $\beta(s)$ can be expressed as $\alpha(s; \gamma) = (\alpha_0(s; \gamma_0), \dots, \alpha_{D_1}(s; \gamma_{D_1}))^T$ and $\beta(s; \tau) = (\beta_0(s; \tau_0), \dots, \beta_{D_2}(s; \tau_{D_2}))^T$, where

$$\alpha_d(s; \gamma) = \sum_{l=0}^{L_d} \gamma_{dl} \mathcal{T}_{dl}(s) \quad \text{and} \quad \beta_d(s; \tau) = \sum_{m=0}^{M_d} \tau_{dm} \mathcal{T}_{dm}^*(s) \quad (5.9)$$

where $\{L_d, M_d\}$ are fixed, and $\{\mathcal{T}_{dl}(s), \mathcal{T}_{dm}^*(s)\}$ are known transformations of s . The choice of $\mathcal{T}_{dl}(s) = s^l$ and $\mathcal{T}_{dm}^*(s) = s^m$ leads to the global polynomials $\alpha_d(s; \gamma)$ and $\beta_d(s; \tau)$.

When the parametric forms of $\alpha(s)$ and $\beta(s)$ are unknown, nonparametric analysis can be performed by approximating $\alpha(s)$ and $\beta(s)$ with basis expansions. If $\{\mathcal{B}_{d_1}(s) = (\mathcal{B}_{d_1 0}(s), \dots, \mathcal{B}_{d_1 \mathcal{L}_{d_1}}(s))^T; 0 \leq d_1 \leq D_1\}$ and $\{\mathcal{B}_{d_2}^*(s) = (\mathcal{B}_{d_2 0}^*(s), \dots, \mathcal{B}_{d_2 \mathcal{M}_{d_2}}(s))^T; 0 \leq d_2 \leq D_2\}$ are two sets of pre-specified basis functions, their basis approximations for $\alpha(s)$ and $\beta(s)$ are given by

$$\alpha_d(s; \gamma) \approx \sum_{l=0}^{\mathcal{L}_d} \gamma_{dl} \mathcal{B}_{dl}(s) \quad \text{and} \quad \beta_d(s; \tau) \approx \sum_{m=0}^{\mathcal{M}_d} \tau_{dm} \mathcal{B}_{dm}^*(s), \quad (5.10)$$

where \mathcal{L}_d and \mathcal{M}_d may tend to infinity as $n \rightarrow \infty$. Common choices of basis functions include truncated polynomial bases, Fourier bases or B-splines. Currently, only B-splines with fixed knot sequences have been investigated for the model (5.8) in the literature (Wu, Tian and Bang, 2008) because of the superior numerical stability of B-spline approximatoins. An alternative smoothing approach is to approximate $\alpha(s)$ and $\beta(s)$ by smoothing splines (Lin and Zhang, 1999; Chiang, Rice and Wu, 2001). But nonparametric estimation and inference with smoothing splines in (5.8) have not been studied, since the explicit expressions and statistical properties of smoothing spline estimators are very different from B-splines.

5.4.2 Least-Squares Estimation

Likelihood-based estimates of $\alpha(s)$ and $\beta(s)$ for (5.8) can not be computed if the explicit forms of $G(\cdot|S_i)$ and the distribution of ϵ_{ij} are unknown. When $\alpha(s) = \alpha(s; \gamma)$ and $\beta(s) = \beta(s; \tau)$ are parametrized by Euclidean valued parameters γ and τ , respectively, a practical approach, which does not require the explicit distribution functions, is to compute the

weighted least-squares estimators $\hat{\gamma}_{LS}$ and $\hat{\tau}_{LS}$ which minimize

$$\begin{aligned} \ell(\gamma, \tau) = & \sum_{i=1}^n \left\{ \left[\mathbf{Y}_i - \left(\mathbf{Z}_i^T \alpha(S_i; \gamma) + (\delta \mathbf{W})_i^T \beta(S_i; \tau) \right) \right]^T \right. \\ & \left. \times \Lambda_i \left[\mathbf{Y}_i - \left(\mathbf{Z}_i^T \alpha(S_i; \gamma) + (\delta \mathbf{W})_i^T \beta(S_i; \tau) \right) \right] \right\}, \end{aligned} \quad (5.11)$$

where $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \dots, \mathbf{Z}_{in_i})^T$, $(\delta \mathbf{W})_i = (\delta_{i1} \mathbf{W}_{i1}, \dots, \delta_{in_i} \mathbf{W}_{in_i})^T$, and Λ_i are pre-specified symmetric nonsingular $n_i \times n_i$ weight matrices. Explicit expressions of the weighted least-squares estimators are

$$\begin{pmatrix} \hat{\gamma}_{LS}(\mathcal{T}) \\ \hat{\tau}_{LS}(\mathcal{T}) \end{pmatrix} = \left\{ \sum_{i=1}^n [\mathcal{W}_i \mathcal{T}_i]^T \Lambda_i [\mathcal{W}_i \mathcal{T}_i] \right\}^{-1} \left\{ \sum_{i=1}^n [\mathcal{W}_i \mathcal{T}_i]^T \Lambda_i \mathbf{Y}_i \right\}, \quad (5.12)$$

where $\sum_{i=1}^n [\mathcal{W}_i \mathcal{T}_i]^T \Lambda_i [\mathcal{W}_i \mathcal{T}_i]$ is nonsingular, and the j th row of \mathcal{W}_i is $(\mathbf{Z}_{ij}^T, \delta_{ij} \mathbf{W}_{ij}^T)$.

For nonparametric functions of $\alpha(s)$ and $\beta(s)$ with basis approximations (5.10), the least-squares based nonparametric estimators of $\alpha(s)$ and $\beta(s)$ are computed by substituting the right-hand side terms of (5.10) into (5.11), which lead to

$$(\tilde{\alpha}_{LS}^T(s; \mathcal{B}), \tilde{\beta}_{LS}^T(s; \mathcal{B}))^T = \mathcal{B}(s) \left(\tilde{\gamma}_{LS}^T(\mathcal{B}), \tilde{\tau}_{LS}^T(\mathcal{B}) \right)^T, \quad (5.13)$$

where $\{\tilde{\gamma}_{LS}(\mathcal{B}), \tilde{\tau}_{LS}(\mathcal{B})\}$ are given in (5.12) with $\mathcal{T}(s)$ replaced by $\mathcal{B}(s)$.

When $\Lambda_i = \mathbf{V}_i^{-1}$ and the distribution functions are assumed to be normal, (5.12) and (5.13) are the same as the maximum likelihood estimators or their approximated versions based on B-splines. When \mathbf{V}_i are unknown, as often the case in practice, subjective choices for Λ_i may be used. One potential ‘‘plug-in’’ approach is to estimate \mathbf{V}_i from the data and compute the estimates by substituting Λ_i with the estimates of \mathbf{V}_i^{-1} . But, in practice, \mathbf{V}_i is often difficult to estimate, it is unclear whether such ‘‘plug-in’’ estimators have superior statistical properties than the estimators with subjective Λ_i choices.

5.5 Estimation with Shared-Parameter Models

5.5.1 Linear and Additive Shared-Parameter Models

For many situations where concomitant interventions are involved, the change-point time S_i , which depends on the pre-intervention trend of Y_{ij} , may not be observed for all the study subjects, since some subjects may have the concomitant intervention change-points before or after the study period. In such situations, where S_i is referred in Section 5.2 as double censored, a natural strategy is to further simplify (5.2) into a sufficiently general model that

is practically useful. Suppose that the concomitant intervention effects only depend on the pre-intervention trends \mathbf{a}_i of the outcome variable but not on the change-point time S_i . A useful special case of (5.2) is

$$\begin{cases} Y_{ij} = \mu_0(T_{ij}, \mathbf{X}_i; \mathbf{a}_i) + \lambda_{ij} \mu_1(T_{ij}, \mathbf{X}_i, R_{ij}; \mathbf{b}_i) + \epsilon_{ij}, \\ \mathbf{a}_i \sim F_a(\cdot), S_i | \mathbf{a}_i \sim F_s(\cdot | \mathbf{a}_i), \mathbf{b}_i | \mathbf{a}_i \sim F_b(\cdot | \mathbf{a}_i), \end{cases} \quad (5.14)$$

where $F_a(\cdot)$ is the cumulative distribution function (CDF) of \mathbf{a}_i , $F_s(\cdot | \mathbf{a}_i)$ and $F_b(\cdot | \mathbf{a}_i)$ are the conditional CDF's of S_i and \mathbf{b}_i , respectively, given \mathbf{a}_i , and \mathbf{b}_i and S_i are independent given \mathbf{a}_i . In contrast to the varying-coefficient model (5.8), where the conditional means of \mathbf{a}_i and \mathbf{b}_i given S_i are used, (5.14) incorporates S_i through $F_s(\cdot | \mathbf{a}_i)$. By modeling the conditional distribution of S_i given \mathbf{a}_i , (5.14) allows S_i to be left or right censored. For simplicity, (5.14) assumes that \mathbf{b}_i does not depend on S_i , although further generalizations may allow the distribution of \mathbf{b}_i to depend on (S_i, \mathbf{a}_i) .

Further specifications of $F_a(\cdot)$, $F_s(\cdot | \mathbf{a}_i)$ and $F_b(\cdot | \mathbf{a}_i)$ may be considered in practice to balance the computational feasibility and flexibility of the model. A useful and mathematically tractable specification for (5.14) is to assume that S_i and \mathbf{b}_i depend on \mathbf{a}_i only through their conditional means, which are linear functions of \mathbf{a}_i . A linear shared-parameter model for (5.14) is

$$\begin{cases} Y_{ij} = \mu_0(T_{ij}, \mathbf{X}_i; \mathbf{a}_i) + \lambda_{ij} \mu_1(T_{ij}, \mathbf{X}_i, R_{ij}; \mathbf{b}_i) + \epsilon_{ij}, \\ \mathbf{a}_i = \alpha + e_i^{(a)}, S_i = \gamma^T (1, \mathbf{a}_i^T)^T + e_i^{(s)}, \mathbf{b}_i = \beta^T (1, \mathbf{a}_i^T)^T + e_i^{(b)}, \end{cases} \quad (5.15)$$

where $\alpha = (\alpha_1, \dots, \alpha_{d_0})^T$, $\alpha_d \in \mathbf{R}$, $\beta = (\beta_1^T, \dots, \beta_{d_1}^T)^T$, $\beta_l = (\beta_{l0}, \dots, \beta_{ld_0})^T$, $\beta_{ld} \in \mathbf{R}$, $\gamma = (\gamma_0, \dots, \gamma_{d_0})^T$, and $\{\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{in_i})^T, e_i^{(a)}, e_i^{(b)}, e_i^{(s)}\}$ are independent mean zero random errors with covariance matrices $\{\mathbf{V}_y, \mathbf{V}_a, \mathbf{V}_b, \sigma_s^2\}$, respectively. The unknown parameters in (5.15) are the mean components $\theta = (\alpha^T, \beta_1^T, \dots, \beta_{d_1}^T, \gamma^T)^T$ and the covariance structures $\mathbf{V} = \{\mathbf{V}_y, \mathbf{V}_a, \mathbf{V}_b, \sigma_s^2\}$.

When the relationship between S_i and \mathbf{a}_i in (5.14) are unknown, a nonparametric model for $\{S_i, \mathbf{a}_i\}$ is $S_i = \mu^{(s)}(\mathbf{a}_i) + \epsilon_i^{(s)}$, where $\mu^{(s)}(\mathbf{a}_i) = E(S_i | \mathbf{a}_i)$ is a smooth function of \mathbf{a}_i . Since unstructured estimation of $\mu^{(s)}(\mathbf{a}_i)$ could be difficult when \mathbf{a}_i is a high dimensional vector, a simple additive approach is to replace the relationship between S_i and \mathbf{a}_i in (5.15) with

$$S_i = \sum_{d=0}^{d_0} \mu_d^{(s)}(a_{id}) + \epsilon_i^{(s)}, \quad (5.16)$$

where $\mu_d^{(s)}(a_{id})$ are smooth functions of a_{id} , so that an additive shared-parameter model for

(5.14) is

$$\begin{cases} Y_{ij} = \mu_0(T_{ij}, \mathbf{X}_i; \mathbf{a}_i) + \lambda_{ij} \mu_1(T_{ij}, \mathbf{X}_i, R_{ij}; \mathbf{b}_i) + \epsilon_{ij}, \\ \mathbf{a}_i = \alpha + e_i^{(a)}, S_i = \sum_{d=0}^{d_0} \mu_d^{(s)}(a_{id}) + \epsilon_i^{(s)}, \mathbf{b}_i = \beta^T (1, \mathbf{a}_i^T)^T + e_i^{(b)}. \end{cases} \quad (5.17)$$

Further generalizations of (5.17) are theoretically possible but at the expense of computational complexity.

5.5.2 Maximum Likelihood Estimation Methods

If the distribution functions are explicitly specified with a known parametric form, the parameters in (5.14) can be estimated by a maximum likelihood (ML). Denote by $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$, $\mathbf{T}_i = (T_{i1}, \dots, T_{in_i})^T$, $\mathcal{D}_i = (\mathbf{T}_i, \mathbf{X}_i)$, and $f_y(\cdot)$, $f_b(\cdot)$, $f_s(\cdot)$ and $f_a(\cdot)$ the densities of Y_{ij} , \mathbf{b}_i , S_i and \mathbf{a}_i . The joint density of $(\mathbf{b}_i, S_i, \mathbf{a}_i)$ can be expressed as

$$f(\mathbf{b}_i, S_i, \mathbf{a}_i) = f_b(\mathbf{b}_i | \mathbf{a}_i) f_s(S_i | \mathbf{a}_i) f_a(\mathbf{a}_i).$$

Then the conditional density of (\mathbf{Y}_i, S_i) given $\mathcal{D}_i = (\mathbf{T}_i, \mathbf{X}_i)$ can be derived by integrating over \mathbf{a}_i and \mathbf{b}_i and is given by

$$f_{(y,s)}(\mathbf{Y}_i, S_i | \mathcal{D}_i) = \iint f_y(\mathbf{Y}_i | \mathcal{D}_i, S_i, \mathbf{a}_i, \mathbf{b}_i) f_b(\mathbf{b}_i | \mathbf{a}_i) f_s(S_i | \mathbf{a}_i) f_a(\mathbf{a}_i) d\mathbf{a}_i d\mathbf{b}_i. \quad (5.18)$$

Since the observed change-point time is the double censored version $\{S_i^{(c)}, \delta_i^{(c)}\}$, the conditional density function (5.18) may not be directly used in estimation, and one has to consider the following conditional density of $\mathcal{S}_i^{(c)} = (S_i^{(c)}, \delta_i^{(c)})$ given \mathbf{a}_i :

$$f_s(\mathcal{S}_i^{(c)} | \mathbf{a}_i) = \begin{cases} f_s(S_i | \mathbf{a}_i), & \text{if } \delta_i^{(c)} = 0, \\ 1 - F_s(T_{in_i} | \mathbf{a}_i), & \text{if } \delta_i^{(c)} = 1, \\ F_s(T_{i1} | \mathbf{a}_i), & \text{if } \delta_i^{(c)} = 2. \end{cases} \quad (5.19)$$

By (5.18) and (5.19), the log-likelihood function for $(\mathbf{Y}_i, \mathcal{S}_i^{(c)})$ conditioning on \mathcal{D}_i , $i = 1, \dots, n$, is

$$L_c = \frac{1}{n} \sum_{i: \delta_i^{(c)}=0} \log f_{(y,s)}(\mathbf{Y}_i, S_i | \mathcal{D}_i) + \frac{1}{n} \sum_{l=1,2} \sum_{i: \delta_i^{(c)}=l} \log f_{(y,l)}(\mathbf{Y}_i | \mathcal{D}_i) \quad (5.20)$$

where $f_{(y,s)}(\cdot | \cdot)$ is given above, $f_{(y,1)}(\cdot | \mathcal{D}_i, \mathbf{a}_i)$ and $f_{(y,2)}(\cdot | \mathcal{D}_i, \mathbf{a}_i, \mathbf{b}_i)$ are the densities of \mathbf{Y}_i given $\{\mathcal{D}_i, \mathbf{a}_i, \delta_i^{(c)} = 1\}$ and $\{\mathcal{D}_i, \mathbf{a}_i, \mathbf{b}_i, \delta_i^{(c)} = 2\}$, respectively,

$$f_{(y,1)}(\mathbf{Y}_i | \mathcal{D}_i) = \int f_{(y,1)}(\mathbf{Y}_i | \mathcal{D}_i, \mathbf{a}_i) \{1 - F_s(T_{in_i} | \mathbf{a}_i)\} f_a(\mathbf{a}_i) d\mathbf{a}_i$$

and

$$f_{(y,2)}(\mathbf{Y}_i|\mathcal{D}_i) = \iint f_{(y,2)}(\mathbf{Y}_i|\mathcal{D}_i, \mathbf{a}_i, \mathbf{b}_i) F_s(T_{i1}|\mathbf{a}_i) f_b(\mathbf{b}_i|\mathbf{a}_i) f_a(\mathbf{a}_i) d\mathbf{a}_i d\mathbf{b}_i.$$

If the parametric family for $f_{(y,s)}(\cdot|\cdot)$ is denoted by $\{f_{(y,s)}(\cdot|\cdot|\mathcal{D}_i); \phi = (\theta, \mathbf{V})\}$, the ML estimators for $\phi = (\theta, \mathbf{V})$ can be obtained by maximizing the log-likelihood function (5.20).

5.5.3 Approximate Maximum Likelihood Estimation Methods

Estimation for the additive shared-parameter model (5.17) can be achieved by maximizing an approximate likelihood function for (5.17). Under some mild smoothness conditions on $\mu_d^{(s)}(\cdot)$ (e.g., Huang, Wu and Zhou, 2004), $\mu_d^{(s)}(\cdot)$ can be approximated by the B-spline expansion

$$\mu_d^{(s)}(a_{id}) \approx \sum_{p=1}^{P_d} \gamma_p^{(d)} B_p^{(d)}(a_{id}) = \left(\gamma^{(d)}\right)^T \mathbf{B}^{(d)}(a_{id}) \quad (5.21)$$

where, for some P_d , $\{B_p^{(d)}(\cdot); 1 \leq p \leq P_d\}$ is a spline basis,

$$\mathbf{B}^{(d)}(a_{id}) = \left(B_1^{(d)}(a_{id}), \dots, B_{P_d}^{(d)}(a_{id})\right)^T$$

and $\gamma^{(d)} = (\gamma_1^{(d)}, \dots, \gamma_{P_d}^{(d)})^T$ is a set of real-valued coefficients. It follows from (5.16) that

$$S_i \approx \sum_{d=0}^{d_0} (\gamma^{(d)})^T \mathbf{B}^{(d)}(a_{id}) + \epsilon_i^{(s)}. \quad (5.22)$$

By substituting $\sum_{d=0}^{d_0} \mu_d^{(s)}(a_{id})$ of (5.16) with $\sum_{d=0}^{d_0} (\gamma^{(d)})^T \mathbf{B}^{(d)}(a_{id})$, the parameters are

$$\theta = \left(\alpha^T, \beta_1^T, \dots, \beta_{d_1}^T, \gamma^T\right)^T, \quad \gamma = \left((\gamma^{(0)})^T, \dots, (\gamma^{(d_0)})^T\right)^T \quad \text{and} \quad \mathbf{V} = \left\{\mathbf{V}_y, \mathbf{V}_a, \mathbf{V}_b, \sigma_s^2\right\}.$$

Let $f_s^*(\cdot; \gamma, \sigma_s|\mathbf{a}_i)$ be the conditional density of $\sum_{d=0}^{d_0} \{(\gamma^{(d)})^T \mathbf{B}^{(d)}(a_{id})\} + \epsilon_i^{(s)}$ given \mathbf{a}_i . If the distributions of ϵ_{ij} , $e_i^{(a)}$, $e_i^{(b)}$ and $e_i^{(s)}$, which all have zero means, are parameterized by the vector of variance parameters \mathbf{V} , the density $f_s(S_i|\mathbf{a}_i)$ can be approximated by $f_s^*(S_i; \gamma, \sigma_s|\mathbf{a}_i)$, and the parameters of (5.17) can be obtained by maximizing the following approximate log-likelihood function for $(\mathbf{Y}_i, \mathcal{S}_i^{(c)})$ given \mathcal{D}_i ,

$$L_c^*(\phi) = \frac{1}{n} \sum_{i:\delta_i^{(c)}=0} \log f_{(y,s)}^*(\mathbf{Y}_i, S_i; \phi|\mathcal{D}_i) + \frac{1}{n} \sum_{l=1,2} \sum_{i:\delta_i^{(c)}=l} \log f_{(y,l)}^*(\mathbf{Y}_i; \phi|\mathcal{D}_i) \quad (5.23)$$

where $\phi = (\theta, \mathbf{V})$, $f_{(y,s)}^*(\cdot|\mathcal{D}_i)$, $f_{(y,k)}^*(\cdot|\mathcal{D}_i)$, $k = 1, 2$, are given in (5.20) with $f_s(S_i|\mathbf{a}_i)$ replaced by $f_s^*(S_i; \gamma, \sigma_s|\mathbf{a}_i)$. If $L_c^*(\phi)$ satisfies the regularity conditions for MLE's, the approximate MLE $\hat{\phi}$ satisfy $L_c^*(\hat{\phi}) = \max_{\phi} L_c^*(\phi)$.

5.5.4 A Two-Stage Estimation Procedure

The likelihood functions (5.20) and (5.23) involve nonlinear terms of the parameters. A global maximization of (5.20) or (5.23) over θ and \mathbf{V} simultaneously could be computationally unfeasible in practice. In order to alleviate the computational burden, Wu, Tian and Jiang (2011) suggests to use the following two-stage procedure, which combines REMLE with the Newton-Raphson algorithm:

- (S1) Assume that $\{\epsilon_{ij}, \mathbf{a}_i, \mathbf{b}_i, S_i\}$ of (5.15) or (5.17) are independent random variables with covariance matrices $\mathbf{V} = \{\mathbf{V}_y, \mathbf{V}_a, \mathbf{V}_b, \sigma_s^2\}$, that is, the naïve mixed-effects model (5.1) holds. Compute $\hat{\mathbf{V}}$ of \mathbf{V} using the REMLE procedure.
- (S2) Substitute \mathbf{V} with $\hat{\mathbf{V}}$, and maximize $L_c(\theta, \hat{\mathbf{V}})$ with respect to θ using the Newton-Raphson procedure. The maximizer $\hat{\theta} = \arg \max_{\theta} L_c(\theta, \hat{\mathbf{V}})$ is the approximate ML estimator for θ .

From the expressions of $f_{(y,s)}(\mathbf{Y}_i, S_i | \mathcal{D}_i)$ and $f_{(y,k)}(\mathbf{Y}_i | \mathcal{D}_i)$ for $k = 1, 2$, it is easy to see that the Newton-Raphson algorithm for maximizing $L_c(\theta, \hat{\mathbf{V}})$ at stage (S2) involves multi-dimensional integrations over the functions of \mathbf{a}_i , \mathbf{b}_i and S_i with respect to the joint distributions of \mathbf{a}_i and \mathbf{b}_i . All the necessary quantities involved in the Newton-Raphson algorithm, including the log-likelihood functions, and their gradients and Hessian matrices, can be computed using Monte Carlo simulations, in which case large Monte-Carlo samples are required to compute the gradient and the Hessian matrix in each iteration, so that a complete Newton-Raphson algorithm can be costly to implement. If a suitable initial estimator is available, computation of the algorithm can be significantly reduced by a “one-step” Newton-Raphson procedure (Bickel, 1975). In Wu, Tian and Jiang (2011), the authors suggest to use the estimators computed from the REMLE procedure as a natural candidate for the initial estimator $\hat{\theta}_0$ and to compute the initial estimators of γ by fitting the regression model $S_i = \gamma^T(1, \tilde{\mathbf{a}}_i^{pred}) + \epsilon_i^{(s)}$ using the subjects with S_i observed (i.e., $\delta_i^{(c)} = 0$), where $\tilde{\mathbf{a}}_i^{pred}$ is the predicted value for \mathbf{a}_i .

5.6 Bootstrap Confidence Intervals

In theory, approximate inferences for the parameter vector ϕ can be constructed using the asymptotic distribution of the ML estimator $\hat{\phi}$, when n is large and the model (5.15) follows a known parametric family. Under suitable regularity conditions (Serfling, 1980, Ch. 4),

the asymptotic normality of the MLE's implies that $\hat{\phi}$ has approximately the multivariate normal distribution $\mathcal{N}(\phi, Var(\phi))$, so that an approximate $[100 \times (1 - \alpha)]$ th confidence interval for $\ell(\phi)$, a linear combination of ϕ , is $\ell(\hat{\phi}) \pm Z_{\alpha/2}[\hat{Var}\{\ell(\phi)\}]^{1/2}$, where $\hat{Var}\{\ell(\phi)\}$ is the variance estimator and $Z_{\alpha/2}$ is the $[100 \times (1 - \alpha/2)]$ th percentile of the standard normal distribution. For the additive model (5.17), where nonparametric components are present, asymptotic distributions of the approximate MLE in (5.23) have not yet been developed. As a practical alternative, a bootstrap procedure is to generate bootstrap samples by resampling the subjects with replacement and compute the corresponding bootstrap estimators. The estimates obtained from the original sample are natural choices for the initial estimates for the bootstrap samples. A $[100 \times (1 - \alpha)]$ th confidence interval based on percentiles is given by the corresponding lower and upper $[100 \times (\alpha/2)]$ th percentiles ($L_{\alpha/2}, U_{\alpha/2}$) of the bootstrap estimators. This bootstrap approach has been used in Wu, Tian and Bang (2008) and Wu, Tian and Jiang (2011). Alternatively, one can also compute $\hat{Var}\{\ell(\phi)\}$ from the bootstrap samples, and use the approximate confidence interval $\ell(\hat{\phi}) \pm Z_{\alpha/2}[\hat{Var}\{\ell(\phi)\}]^{1/2}$.

5.7 Applications to the ENRICHD Pharmacotherapy Data

5.7.1 Application to Subjects with Observed Change-Points

A brief summary of the ENRICHD study has been described in Example 4 of Section 1.3. The objective here is to evaluate the additional effects of pharmacotherapy (antidepressants) on the trends of depression severity measured by BDI scores for patients who received pharmacotherapy during the six-month cognitive behavior therapy (CBT) treatment period. Pharmacotherapy with antidepressants is a concomitant intervention in this trial because the decision of using antidepressants and its starting time was made by the patients or their primary care physicians.

Using the regression framework of Section 5.4, this analysis includes 91 patients with a total of 1,446 observations in the CBT arm who received pharmacotherapy during the treatment period and had clear records of their pharmacotherapy starting time. Data from patients in the usual care (UC) arm are excluded, since the pharmacotherapy starting time and repeated BDI scores were not accurately recorded for these patients. Excluded are also data from patients in the CBT arm who did not have pharmacotherapy starting time accurately recorded during the six month CBT treatment period. Among the 91 patients analyzed here, 43 of them started pharmacotherapy at baseline and 48 started

pharmacotherapy between 7 and 172 days. The number of visits for these patients ranges from 5 to 36 and has the median of 16.

Following Section 5.2, Y_{ij} , T_{ij} , S_i , $R_{ij} = T_{ij} - S_i$ and $\delta_{ij} = 1_{[T_{ij} \geq S_i]}$ are the i th subject's BDI score, trial time (in months), starting time of pharmacotherapy, time from initiation of pharmacotherapy, and pharmacotherapy indicator, respectively, at the j th visit. A simple case of the naïve mixed-effects models for evaluating the trends of BDI score over T_{ij} is the linear mixed-effects model

$$Y_{ij} = a_{0i} + a_{1i}T_{ij} + b_{0i}\delta_{ij} + b_{1i}\delta_{ij}R_{ij} + \epsilon_{ij}, \quad (5.24)$$

where $E(a_{0i}, a_{1i}, b_{0i}, b_{1i})^T = (\alpha_0, \alpha_1, \beta_0, \beta_1)^T$. When $\delta_{ij} = 1$ and $R_{ij} = r$, $(\beta_0 + \beta_1 r)$ describes the mean pharmacotherapy effect at r months since the start of pharmacotherapy.

Since (5.24) ignores the possible correlation between S_i and the pre-pharmacotherapy depression trends, which may lead to potential bias, its varying-coefficient generalization is

$$Y_{ij} = \alpha_0(S_i) + \alpha_1(S_i)T_{ij} + \beta_0\delta_{ij} + \beta_1\delta_{ij}R_{ij} + e_{ij}, \quad (5.25)$$

where $e_{ij} = a_{i0}^* + a_{i1}^*T_{ij} + b_{i0}^*\delta_{ij} + b_{i1}^*\delta_{ij}R_{ij} + \epsilon_{ij}$, $\alpha_0(S_i) = \gamma_{00} + \gamma_{01}S_i$ and $\alpha_1(S_i) = \gamma_{10} + \gamma_{11}S_i$. The mean pre-pharmacotherapy BDI trend in (5.25) is associated with S_i through intercept $\alpha_0(S_i)$ and slope $\alpha_1(S_i)$. The mean pharmacotherapy effect at r months after the start of pharmacotherapy is $\beta_0 + \beta_1 r$. A negative (positive) value for $\beta_0 + \beta_1 r$ corresponds to a beneficial (harmful) effect for reducing depression.

For mathematical simplicity, (5.25) assumes that $\beta_0(S_i) \equiv \beta_0$ and $\beta_1(S_i) \equiv \beta_1$, so that the effects of pharmacotherapy only depend on how long the antidepressant has been used. Under this assumption, β_0 and β_1 have the same interpretations in both the naïve linear mixed-effects model (Näive LME) (5.24) and the shared-parameter linear mixed-effects model (SP-LME) (5.25), although the mean BDI scores have different pre-pharmacotherapy time trends.

Table 4 summarizes the estimates for β_0 and β_1 and their corresponding standard errors, 95% CIs and p-values obtained by the REML procedure with unstructured correlations. The negative estimates for (β_0, β_1) suggest that the beneficial effect of pharmacotherapy for this patient population is detected under both the Näive LME and SP-LME models, when only the patients who had pharmacotherapy change-point time within the CBT period are included in the analysis.

Table 4: Parameter estimates for β_0 and β_1 and their standard errors (SE), 95% confidence intervals (CIs) and p-values were obtained by restricted maximum likelihood with unstructured correlations for the naïve linear mixed-effects model (Näive LME) (5.24) and the shared-parameter linear mixed-effects model (SP-LME) (5.25).

Model	Parameter	Estimate	SE	95% CI	p-value
Näive LME	β_0	-3.410	0.994	(-5.399, -1.422)	0.0013
	β_1	-1.584	0.521	(-2.626, -0.542)	0.0039
SP-LME	β_0	-4.302	1.041	(-6.385, -2.220)	0.0001
	β_1	-2.062	0.773	(-3.608, -0.516)	0.0105

5.7.2 Application to Subjects with Censored Change-Points

The previous analysis uses only a sub-sample of 91 depressed patients in the ENRICHD CBT arm who had their change-point time S_i observed during the CBT treatment period. Thus, the conclusion of the beneficial effects of antidepressants for lowering the BDI scores ignores the information from patients who did not start pharmacotherapy during the CBT period. Using the shared-parameter models, the analysis here is based on 557 depressed patients who had their exact dates of antidepressant starting time recorded and attended 5 or more CBT sessions during the six-month treatment period. For practical considerations, patients in the UC arm, patients whose starting dates of antidepressant use were not recorded, and patients who had poor adherence to the required weekly CBT sessions (attended less than 5 sessions) are excluded from the analysis. Because antidepressant use for each patient was individually monitored and recorded as accurate as possibly by study psychiatrists (Taylor et al., 2005, page 794), it is reasonable to assume that the missing records on antidepressant starting dates were missing at random. The longitudinal sample then includes 11 patients who used antidepressants before baseline, 92 patients who started antidepressant during the treatment period, and 454 patients who did not use antidepressants before and during the treatment period. The number of visits for these patients ranges from 5 to 36 and has a median of 12.

With a slight modification of the notation in Section 5.7.1, Y_{ij} , T_{ij} , $S_i^{(c)}$, and $R_{ij} = T_{ij} - S_i^{(c)}$ denote the i th patient’s BDI score, trial time (months), starting time (months) of antidepressant use, and antidepressant duration time (months), respectively, at the j th visit. For all $1 \leq i \leq n$, the observed $(S_i^{(c)}, \delta_i^{(c)})$ is $(S_i^{(c)} = S_i, \delta_i^{(c)} = 0)$ if the i th patient used an-

tidepressants within the CBT period, ($S_i^{(c)} = T_{in_i}, \delta_i^{(c)} = 1$) if the patient did not use antidepressant within the CBT period, and ($S_i = 0, \delta_i^{(c)} = 2$) if the patient used antidepressants before baseline. When $\lambda_{ij} = 1_{[S_i < T_{ij}]}$ and the linear models $\mu_0(T_{ij}; a_{i0}, a_{i1}) = a_{i0} + a_{i1}T_{ij}$ and $\mu_1(R_{ij}; b_{i0}, b_{i1}) = b_{i0} + b_{i1}R_{ij}$ are used, (a_{i0}, a_{i1}) represents the intercept and slope of the i th subject's BDI trajectory before antidepressant use, and (b_{i0}, b_{i1}) is the intercept and slope of the change of the subject's BDI trajectory after antidepressant use. A series of preliminary analyses described in Wu, Tian and Jiang (2011, Appendix D.1 of the Web-Supplementary Materials) suggest that the above linear models for $\mu_0(T_{ij}; a_{i0}, a_{i1})$ and $\mu_1(R_{ij}; b_{i0}, b_{i1})$ can be used as a parsimonious approximation to the BDI time trends for this study.

Similar to (5.24), the REMLE procedure is applied to the following naïve mixed-effects model to estimate the unknown population-mean parameters $\alpha_0, \alpha_1, \beta_0$ and β_1 :

$$\begin{cases} Y_{ij} = a_{i0} + a_{i1}T_{ij} + \lambda_{ij}(b_{i0} + b_{i1}R_{ij}) + \epsilon_{ij}, \\ (a_{i0}, a_{i1}, b_{i0}, b_{i1})^T \sim \mathcal{N}((\alpha_0, \alpha_1, \beta_0, \beta_1)^T, \Sigma), \end{cases} \quad (5.26)$$

where $(\alpha_0, \alpha_1, \beta_0, \beta_1)^T$ is the unknown mean vector and Σ is the unstructured covariance matrix for the multivariate normal distribution $\mathcal{N}(\cdot, \cdot)$. The population-mean concomitant intervention effects are β_0 and β_1 , which are the mean intercept and slope for the “correction term” after antidepressant use, and zero value of $\beta_0 + \beta_1 R_{ij}$ indicates ignorable antidepressant effect on the mean BDI scores.

To account for the possible link between pharmacotherapy change-point time and the BDI trend before the use of antidepressants, a shared-parameter model that directly generalizes (5.26) is

$$\begin{cases} Y_{ij} = a_{i0} + a_{i1}T_{ij} + \lambda_{ij}(b_{i0} + b_{i1}R_{ij}) + \epsilon_{ij}, & (a_{i0}, a_{i1})^T = (\alpha_0, \alpha_1)^T + \epsilon_i^{(a)}, \\ b_{i0} = \beta_0 + \epsilon_{i0}^{(b)}, & b_{i1} = \beta_1 + \epsilon_{i1}^{(b)}, & S_i = \gamma_0 + \gamma_1 a_{i0} + \epsilon_i^{(s)}, \end{cases} \quad (5.27)$$

where $\epsilon_i^{(a)}$ and $\epsilon_i^{(b)} = (\epsilon_{i0}^{(b)}, \epsilon_{i1}^{(b)})^T$ are mean zero bivariate normal random vectors with unstructured covariance matrices $\Sigma^{(a)}$ and $\Sigma^{(b)}$, respectively, $\epsilon_i^{(s)}$ is a mean zero normal random variable with variance σ_s^2 , and $\epsilon_i^{(a)}$, $\epsilon_i^{(b)}$ and $\epsilon_i^{(s)}$ are independent. The interpretations of the population-mean parameters $\alpha_0, \alpha_1, \beta_0$ and β_1 in (5.27) are the same as their counterparts specified in (5.26).

Table 5 shows the estimates of $\alpha_0, \alpha_1, \beta_0$ and β_1 , and their standard errors and 95% confidence intervals (CI) computed using REMLE with unstructured covariance matrix Σ under the naïve linear mixed-effects model (Naïve-LMEM) (5.26) and the two-stage ML

Table 5: Parameter estimates, standard errors (SE) and 95% confidence intervals (CI) computed for the ENRICHD pharmacotherapy data based on the naïve linear mixed-effects model (Näive-LMEM) (5.26), the linear shared-parameter model (LSPM)(5.27).

Effect	Näive-LMEM			LSPM		
Parameter	Estimate	SE	95% CI	Estimate	SE	95% CI
α_0	14.454	0.312	(13.842, 15.066)	15.867	0.375	(15.216, 16.771)
α_1	-1.887	0.067	(-2.018, -1.756)	-1.816	0.070	(-1.989, -1.714)
β_0	3.579	0.825	(1.962, 5.196)	-6.646	0.915	(-8.239, -4.990)
β_1	0.036	0.227	(-0.409, 0.481)	-0.453	0.284	(-0.962, 0.122)
γ_0	—	—	—	15.162	1.134	(13.497, 17.652)
γ_1	—	—	—	0.480	0.050	(-0.593, -0.392)

procedure with ten Newton-Raphson iterations under the linear shared-parameter model (LSPM) (5.27). Under Näive-LMEM, the negative estimate $\tilde{\alpha}_1 = -1.887$ suggests that the mean BDI score for these patients tends to decrease over the trial time since the start of the CBT sessions, while the positive estimate $\tilde{\beta}_0 = 3.579$ and its 95% CI seem to suggest that the use of antidepressants increase the patients' mean BDI scores. Since the “self-selectiveness” of the antidepressant use as a concomitant intervention is not considered in (5.26), the positive estimate of β_0 under this model does not reflect the real effect of pharmacotherapy on depression severity. On the other hand, under the LSPM (5.27), the negative estimates $\hat{\beta}_0 = -6.646$ and $\hat{\beta}_1 = -0.453$ suggest that pharmacotherapy has on average a beneficial effect for lowering a patient's depression severity. The 95% CI's for γ_0 and γ_1 obtained from (5.27) suggest that (5.26) is likely a misspecified model for this dataset.

5.8 Discussion and Further Remarks

The main results of this section show that, in general, outcome-adaptive covariates, such as concomitant interventions, should not be treated as usual time-dependent covariates in naïve mixed-effects models. For the simple case of a concomitant intervention, a shared-parameter model may be considered to reduce the estimation bias and correct the “self-selectiveness” of the concomitant intervention. The methods presented here have a narrow focus on a single concomitant intervention in a longitudinal clinical trial. Concomitant interventions may commonly appear in other settings, such as in an epidemiological study where study subjects may take antihypertensive medication during the study when their

blood pressure levels either exhibit some undesirable trends or stay in an intolerable range. In the ENRICHHD pharmacotherapy data, pharmacotherapy as a concomitant was initiated under a vague guideline and a linear shared-parameter model appears to be a reasonable choice. However, this model may not be suitable when the intervention selection mechanism is changed, and in some situations the entire shared-parameter approach may have to be re-evaluated.

As a special case of the shared-parameter models, a varying-coefficient mixed-effects model may be considered mainly because it has a simple and clear biological interpretation for the simple situation where there is only one concomitant intervention and the change-point time is observed for all subjects in the study. Compared with the shared-parameter models, the least-squares based estimation method for the varying-coefficient mixed-effects models does not require the known parametric forms of the distribution functions. The shared-parameter models, on the other hand, may be applied to concomitant interventions with double censored change-point time, but their estimation requires computationally intensive ML and approximate ML algorithms.

Future research in this area may be pursued with several potentially worthy extensions. First, subjects in longitudinal studies may have single or multiple concomitant interventions which can be turned on or off at different time points. In such situations, more general shared-parameter models may be needed to accommodate the possibility of multiple interventions and/or multiple change-points. Second, all the shared-parameter models studied in this section rely on linear functions to describe the time-trends before and after the intervention, but it is possible that linear response curves are inadequate for certain disease outcomes. Models with nonlinear response curves can be justified in practice and should be investigated. Third, the estimation approach of this section depends on the classical frequentist's framework for the B-spline methods. In a different context, Fahrmeir and Lang (2000) demonstrated a promising Bayesian inference procedure for generalized additive mixed models based on Markov random field priors. Similar Bayesian estimation and inference approaches for the shared-parameter models of this section may lead to computationally simpler estimation and inference procedures. Fourth, large sample properties, such as convergence rates and asymptotic distributions, of the ML and approximate ML estimators are still not well-understood and should be systematically developed to provide theoretical justifications for these estimators. Finally, since it may not be always clear

whether an intervention is a concomitant intervention, a model diagnostic method for evaluating the appropriateness of a shared-parameter model would be a valuable tool to be developed.

References

- [1] AKAIKE, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203–217.
- [2] ALTMAN, N. S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.* **85**, 749–759.
- [3] BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–972.
- [4] BATES, D. M. and PINHEIRO, J. C. (1999) *Mixed Effects Models in S*. Springer-Verlag, New York.
- [5] BICKEL, P. J. (1975). One-step Huber estimates in linear models. *Journal of the American Statistical Association* **34**, 584-653.
- [6] CHENG, S. C., WEI, L. J. and YING, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–845.
- [7] CHENG, S. C., WEI, L. J. and YING, Z. (1997). Predicting survival probabilities with semiparametric transformation models. *Journal of the American Statistical Association* **92**, 227–235.
- [8] CHIANG, C.-T., RICE, J. A. and WU, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variable. *J. Amer. Statist. Assoc.* **96**(454):605-619.
- [9] DANIELS, S. R., MCMAHON, R. P., OBARZANEK, E., WACLAWIW, M. A., SIMILO, S. L., BIRO, F. M, SCHREIBER, G. B., KIMM, S. Y. S., MORRISON, J. A. and BARTON, B. A. (1998). Longitudinal correlates of change in blood pressure in adolescent girls. *Hypertension* **31**, 97–103.

- [10] DAVIDIAN, M. and GILTINAN, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*. London; New York: Chapman Hall.
- [11] DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York.
- [12] DIGGLE, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics* **44**, 959-971.
- [13] DIGGLE, P. J., HEAGERTY, P., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data.*, 2nd ed. Oxford: Oxford University Press, England.
- [14] EUBANK, R. L. and SPECKMAN, P. L. (1993). Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.* **88**, 1287-1301.
- [15] ENRICHHD INVESTIGATORS. (2001). Enhancing recovery in coronary heart disease patients (ENRICHHD): study int ervention rationale and design *Psychosomatic Medicine* **63**, 747-755.
- [16] ENRICHHD INVESTIGATORS. (2003). Enhancing recovery in coronary heart disease patients (ENRICHHD): the effects of treating depression and low perceived social support on clinical events after myocardial infarction. *Journal of the American Medical Association* **289**, 3106-3116.
- [17] FAHRMEIR, L. and LANG, S. (2000). Bayesian inference for generalised additive mixed models based on Markov random field priors. *Applied Statistics*, **50**, 201-220.
- [18] FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modeling and Its Applications*. London: Chapman and Hall.
- [19] FAN, J. and MARRON, J. S. (1994). Fast implementations of nonparametric curves estimators. *J. Comput. Graph. Statist.*, **3**, 35-56.
- [20] FAN, J. Q. and ZHANG, J.-T. (2000). Functional linear models for longitudinal data. *J. Roy. Statist. Soc. B* **62**, 303-322.
- [21] FOLLMANN, D. and WU, M. C. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51**, 151-168.

- [22] HALL, P., and MÜLLER, H.G. (2003). Order-preserving Nonparametric regression, with applications to conditional distributions and quantile function estimation. *J. Amer. Statist. Assoc.* **98**, 598–608.
- [23] HALL, P. and TITTERINGTON, D. M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Multi. Anal.* **27**, 228-254.
- [24] HALL, P., WOLFF, R. C. L. and YAO, Q. (1999). Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.* **94**, 154-163.
- [25] HÄRDLE, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, U.K.
- [26] HÄRDLE, W. and MARRON, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19**, 778-796.
- [27] HART, T. D. (1991). Kernel regression estimation with time series errors. *J. Roy. Statist. Soc., Ser. B* **53**, 173-187.
- [28] HART, T. D. and WEHRLY, T. E. (1986). Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.* **81**, 1080-1088.
- [29] HARVILLE, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383-385.
- [30] HASTIE, T. J. and TIBSHIRANI, R. J. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. B* **55**, 757-796.
- [31] HOOVER, D. R., RICE, J. A., WU, C. O. and YANG, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809-822.
- [32] HUANG, J., WU, C. O. and ZHOU, L. (2002). Varying coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89**, 111-128.
- [33] HUANG, J. Z., WU, C. O. and ZHOU, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* **14**, 763-788.

- [34] JONES, R. H. and ACKERSON, L. M. (1990). Serial correlation in unequally spaced longitudinal data. *Biometrika* **77**, 721-731.
- [35] JONES, R. H. and BOADI-BOTENG, F. (1991). Unequally spaced longitudinal data with serial correlation. *Biometrics* **47**, 161-175.
- [36] KASLOW, R. A., OSTROW, D. G., DETELS, R., PHAIR, J. P., POLK, B. F. and RINALDO, C. R. (1987). The Multicenter AIDS Cohort Study: Rationale, Organization and Selected Characteristics of the Participants. *American Journal of Epidemiology* **126**, 310-318.
- [37] KIMM, S. Y., BARTON, B. A., OBARZANEK, E., MCMAHON, R. P., SABRY, Z. I., WACLAWIW, M. A., et al. (2001). Racial divergence in adiposity during adolescence: the NHLBI Growth and Health Study. *Pediatrics* 2001;107:E34.
- [38] KIMM, S. Y., BARTON, B. A., OBARZANEK, E., MCMAHON, R. P., KRONBERG, S. S., WACLAWIW, M. A., et al. (2002). Obesity development during adolescence in a biracial cohort: the NHLBI Growth and Health Study. *Pediatrics* 2002;110:E54.
- [39] KIMM, S. Y., GLYNN, N. W., KRISKA, A. M., FITZGERALD, S. L., AARON, D. J., SIMILO, S. L., et al. (2000). Longitudinal changes in physical activity in a biracial cohort during adolescence. *Med Sci Sports Exerc* 2000;**32**:1445-54.
- [40] KNAFL, G., SACKS, J. and YLVIKAKER, D. (1985). Confidence bands for regression functions. *J. Amer. Statist. Assoc.* **80**, 683-691.
- [41] LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.
- [42] LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.
- [43] LIN, X. and CARROLL, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Statist. Assoc.* **95**, 520-534.
- [44] LIN, X. and ZHANG, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B*, **61**(2), 381-400.

- [45] LU, W. and TSIATIS, A. A. (2006). Semiparametric transformation models for the case-cohort study. *Biometrika* **93**, 207–214.
- [46] LU, W. and YING, Z. (2004). On semiparametric transformation cure models. *Biometrika* **91**, 331–343.
- [47] MOLENBERGHS, G. and VERBEKE, G. (2005). *Models for Discrete Longitudinal Data*. Springer: New York, NY.
- [48] MOYEED, R. A. and DIGGLE, P. J. (1994). Rates of convergence in semiparametric modeling of longitudinal data. *Austral. J. Statist.* **36**, 75–93.
- [49] MÜLLER, H.-G. (1988). Nonparametric Regression Analysis of Longitudinal Data. *Lecture Notes in Statistics*, **46**. Springer-Verlag, Berlin.
- [50] NATIONAL HIGH BLOOD PRESSURE EDUCATION PROGRAM WORKING GROUP ON HIGH BLOOD PRESSURE IN CHILDREN AND ADOLESCENTS (NHBPEP Working Group) (2004). The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents. *Pediatrics* **114**, 555–576.
- [51] NATIONAL HEART, LUNG, AND BLOOD INSTITUTE GROWTH AND HEALTH RESEARCH GROUP (NGHSRG) (1992). Obesity and cardiovascular disease risk factors in black and white girls: the NHLBI Growth and Health Study. *Am J. Public Health* **82** 1613–1620.
- [52] OBARZANEK, E., WU, C. O., CUTLER, J. A., KAVEY, R. W., PEARSON, G. D. and DANIELS, S. R. (2010). Prevalence and incidence of hypertension in adolescent girls. *The Journal of Pediatrics* **157**(3), 461–467.
- [53] PANTULA, S. G. and POLLOCK, K. H. (1985). Nested analysis of variance with auto-correlated errors. *Biometrics*, **41**, 909–920.
- [54] PATTERSON, H. D. and THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.
- [55] PEPE, M. S. and ANDERSON, G. (1994). A cautionary note on inference of marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics: Simulation and Computation*, **23**, 939–951.

- [56] RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53**, 233-243.
- [57] RICE, J. A. and WU, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253-259.
- [58] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.
- [59] SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley & Sons.
- [60] SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.
- [61] TAYLOR, C. B., YOUNGBLOOD, M. E., CATELLIER, D., VEITH, R. C., CARNEY, R. M., BURG, M. M., KAUFMANN, P., SHUSTER, J., MELLMAN, T., BLUMENTHAL, J. A., KRISHNAN, R. and JAFFE, A. S. (2005). Effects of antidepressant medication on morbidity and mortality in depressed patients after myocardial infarction. *Archives of General Psychiatry*, **62**, 792–298.
- [62] THOMPSON, D. R., OBARZANEK, E., FRANKO, D. L., BARTON, B. A., MORRISON, J., BIRO, F. M., DANIELS, S. R. and STRIEGEL-MOORE, R. H. (2007). Childhood overweight and cardiovascular disease risk factors: The National Heart, Lung, and Blood Institute Growth and Health Study. *Journal of Pediatrics* **150**, 18–25.
- [63] VERBEKE, G. and MOLENBERGHS, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.
- [64] VONESH, E. F. and CHINCHILLI, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Marcel Dekker, New York.
- [65] WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.
- [66] WARE, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician*, **39**, 95-101.
- [67] WU, C. O. and CHIANG, C.-T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica* **10**, 433-456.

- [68] WU, C. O., CHIANG, C.-T. and HOOVER, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Amer. Statist. Assoc.* **93**, 1388-1402.
- [69] WU, C. O., TIAN, X. and BANG, H. (2008). A varying-coefficient model for the evaluation of time-varying concomitant intervention effects in longitudinal studies. *Statistics in Medicine* **27**, 3042–3056.
- [70] WU, C. O., TIAN, X. and JIANG, W. (2011). A shared parameter model for the estimation of longitudinal concomitant intervention effects. *Biostatistics* **12**(4):737–749.
- [71] WU, C. O., TIAN, X. and YU, J. (2010). Nonparametric estimation for time-varying transformation models with longitudinal data. *Journal of Nonparametric Statistics* **22**, 133–147.
- [72] WU, C. O., YU, K. F. and CHIANG, C.-T. (2000). A two-step smoothing method for varying-coefficient models with repeated measurements. *Ann. Inst. Statist. Math.* **52**, 519-543.
- [73] WU, C. O., YU, K. F. and YUAN, V. W. S. (2000). Large sample properties and confidence bands for component-wise varying-coefficient regression with longitudinal dependent variable. *Commun. Statist.–Theory Meth.* **29**, 1017-1037.
- [74] ZEGER, S. L. and DIGGLE, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689-699.
- [75] ZEGER, S. L., LIANG, K.-Y. and ALBERT, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049-1060.
- [76] ZENG, D. and LIN, D. Y. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika* **93**, 627–640.