Two-Step Estimation of Functional Linear Models with Applications to Longitudinal Data

Jianqing Fan *and* Jin-Ting Zhang
Department of Statistics
UNC-Chapel Hill, NC 27599-3260

September 19, 1999

Abstract

Functional linear models are useful in longitudinal data analysis. They include many classical and recently proposed statistical models for longitudinal data and other functional data. Recently, smoothing spline and kernel methods have been proposed for estimating their coefficient functions nonparametrically but these methods are either intensive in computation or inefficient in performance. To overcome these drawbacks, in this paper, a simple and powerful two-step alternative is proposed. In particular, the implementation of the proposed approach via local polynomial smoothing is discussed. Methods for estimating standard deviations of estimated coefficient functions are also proposed. Some asymptotic results for the local polynomial estimators are established. Two longitudinal data sets, one of which involves time-dependent covariates, are used to demonstrate the proposed approach. Simulation studies show that our two-step approach improves the kernel method proposed in Hoover, *et al* (1998) in several aspects such as accuracy, computation time and visual appealingness of the estimators.

Key Words And Phrases: Functional linear models, functional ANOVA, local polynomial smoothing, longitudinal data analysis.

Short title : Functional linear models

1 Introduction

Longitudinal data arise frequently in many scientific studies. See Jones (1993), Diggle, Liang and Zeger (1994) and Hand and Crowder (1996) for many interesting examples. Take the CD4 data presented in Section 4 as an example. The CD4 cell percentage of each subject along with some important covariates was measured over a period of time in order to monitor AIDS progression. Let $\{t_{ij}, j = 1, \cdots, T_i\}$ be the times over which the measurements of the i^{th} subject took place. Let Y_{ij} be the observed response (such as the CD4 percentage) and $\mathbf{X}_{ij}$ be the observed covariates (such as Age, Smoking status and PreCD4 level, among others) for the i^{th} subject at time t_{ij}. This results in data of the form

$$(t_{ij}, \mathbf{X}_{ij}, y_{ij}), \quad j = 1, 2, \cdots, T_i; \quad i = 1, 2, \cdots, n, \tag{1.1}$$

where $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \cdots, X_{ijd})^T$ are the d covariate variables measured at time t_{ij}. Of interest is to study the association between the covariates and the responses and to examine how the association varies with time. For the CD4 data set, the association is depicted in Figure 1 in Section 4. To obtain such an association, some modeling between the covariates and the response is needed.

A simple and useful model for studying the association between the covariates $\mathbf{X}(t)$ and response $Y(t)$ is the following linear model:

$$Y(t) = \mathbf{X}(t)^T \beta(t) + \varepsilon(t), \tag{1.2}$$

where $\varepsilon(t)$ is a zero mean correlated stochastic process that can not be explained by the covariates. By letting $X_1(t) \equiv 1$, model (1.2) allows a time-varying intercept term. The repeated measurements (1.1) are regarded as a random sample from model (1.2):

$$Y_i(t_{ij}) = \mathbf{X}_i(t_{ij})^T \beta(t_{ij}) + \varepsilon_i(t_{ij}), \tag{1.3}$$

where $Y_i(t_{ij}) = Y_{ij}$ and $\mathbf{X}_i(t_{ij}) = \mathbf{X}_{ij}$ and $\varepsilon_i(t)$ is a zero mean stochastic process with covariance function $\gamma(s,t) = \text{cov}(\varepsilon_i(s), \varepsilon_i(t))$.

Model (1.2) includes many useful models proposed in the literature. It is a useful extension of commonly-used linear models (Lindsey 1993, Jones 1993, Diggle, *et al* 1994, Hand and Crowder 1996 and references therein) for longitudinal data by allowing coefficients to change over time. While the traditional linear models provide useful tools for analyzing longitudinal data, problems on the adequacy of model fitting often arise. Model (1.2) is also an extension of a useful semiparametric model studied by Zeger and Diggle (1994) and Moyeed and Diggle (1994). The semiparametric model quantifies the time effect by allowing the intercept coefficient to vary over time but not the coefficients of the other covariate variables. In the specific case where there is only an intercept

covariate $X_1(t) \equiv 1$ (namely, no real covariates are of interest) in model (1.2), the model is called a mean function model in Zhang (1999). The mean function model has been extensively studied by Hart and Wehrly (1986, 1993) and Rice and Silverman (1991) respectively in the contexts of repeated measurements and functional data under slightly different formulations. There, a cross-validation procedure removing one subject each time is suggested for bandwidth selection.

Model (1.2) is a specific model of a class of functional linear models introduced by Ramsay and Silverman (1997) in a somewhat different context. It is closely related to the varying-coefficient models (for cross sectional data rather than functional data) proposed in Cleveland, *et al* (1991). For the varying-coefficient models, smoothing spline and kernel methods are proposed in Hastie and Tibshirani (1993). Fan and Zhang (1997) propose a two-step procedure to overcome inflexibility of the traditional spline and kernel methods. Some of these methods can also be adopted in the context of functional linear models. Examples are provided by Ramsay and Silverman(1997), Hoover, *et al* (1998) and Brumback and Rice (1998). In Hoover, *et al* (1998), the smoothing spline and kernel methods are studied while in Brumback and Rice (1998), the smoothing spline method is considered for functional ANOVA models which are special cases of functional linear models.

While the spline method has better performance than the kernel method due to its introduction of multiple smoothing parameters (Hoover, *et al*, 1998), its computation is very intensive even for a longitudinal data set of moderate size (Brumback and Rice, 1998), not to mention the difficulty of selecting the multiple smoothing parameters which involves high dimensional optimization problems. This is particularly the case when functional ANOVA is considered. Taking the nested functional ANOVA model as an example, the number of coefficient functions in model (1.2) can grow extremely fast. For the progesterone data discussed in Section 4.2, there are 91 coefficient functions. Estimating these 91 coefficient functions imposes quite a challenge to the spline method. According to Brumback and Rice (1998), one has to blindly invert a matrix of size 2000×2000, which takes a lot of CPU and requires large amount of RAM. The size of this matrix grows very fast either as the number of subjects n or the number of distinct time points T increases [the matrix size is approximately $(nT) \times (nT)$]. This problem can not easily be rescued by the backfitting algorithm of Hastie and Tibshirani (1993), since there are 91 functions to iterate. This makes the spline method very expensive to compute. It also poses an interesting challenge to statisticians to choose appropriately 91 smoothing parameters.

Compared with the spline method, the kernel method is less intensive since its calculation is indeed conducted around a neighborhood and hence only part of the data are actually involved. However, since the kernel method involves only one smoothing parameter, it often undersmooths some of the underlying coefficient functions when these coefficient functions admit different degrees of smoothness (Hoover, *et al* 1998). Moreover, the kernel method is still pretty intensive in computation. This is especially the case when the cross-validation method of removing one subject

each time is employed to select the smoothing parameter. There are many possible approaches for overcoming these disadvantages of the spline and kernel methods. For instance, Wu and Chiang (1998) modify the kernel method by allowing different smoothing parameters for different coefficient functions although their approach is applicable only when the covariates are all time-independent. Some other ideas, different from the conventional spline and kernel methods, are outlined in Fan and Zhang (1998).

To overcome the disadvantages of the existing approaches for functional linear models, in this paper, an alternative approach—a two-step procedure, is proposed. Simply speaking, we first calculate the raw estimates of the coefficient functions via fitting a standard linear model and then smooth the raw estimates to obtain the smooth estimates of the coefficient functions using one of the existing smoothing techniques. Compared with the spline and the kernel methods proposed in Hoover, *et al* (1998) and Brumback and Rice (1998), our new procedure has many nice properties. It is simple to understand, easy to implement, fast to calculate and effective in performance.

Our new procedure is motivated by a special structure of many longitudinal data sets: measurements are collected at the same scheduled time points for all subjects or can be viewed as so (see the CD4 data in Section 4) although for a particular subject, the measurements at some time points may be missing. Let $t_j, j = 1, \cdots, T$ be the distinct time points where data were collected. Since there are a number of observations (not necessarily n) collected at time t_j, it is possible that for this fixed t_j, we use the data collected there (or around t_j to increase the sample size if needed) to fit the linear model (1.2) and obtain the raw estimates $b(t_j) = (b_1(t_j), \cdots, b_d(t_j))^T$ for $\beta(t_j) = (\beta_1(t_j), \cdots, \beta_d(t_j))^T$. This is the first step. Since the raw estimates are usually not smooth (see examples given in Section 4), we have to smooth them to obtain the smooth estimates for the coefficient functions. Thus, in the second step, for each given component r, a smoothing technique is applied to the data $\{(t_j, b_r(t_j)), j = 1, 2, \cdots, T\}$. This smoothing step is crucial since it gives smooth estimates for the underlying smooth coefficient functions and moreover it allows us to pool information from neighboring time points to improve the efficiency of the raw estimates. An extra benefit of our two-step procedure is that the smoothing step is actually one-dimensional. This leads to several advantages. Firstly, for different components of the coefficient functions, different amounts of smoothing can be conducted. Secondly, visualization of the raw estimates can assist us in choosing a sensible amount of smoothing. Thirdly, the smoothing step can be conducted with any one of existing smoothing techniques. Finally, the existing well-developed smoothing parameter selectors such as the bandwidth selector proposed by Ruppert, *et al* (1995) can be employed easily in the smoothing step when a local linear fit is employed.

Our procedure is also easy to implement using existing software. For each fixed t_j, model (1.2) is a standard linear model with independent error structure. All statistical software containing least squares procedures can be used to obtain the raw estimates. In the second step, all popular

smoothing techniques such as spline (Wahba, 1990, Green and Silverman, 1994), kernel (Gasser and Müller, 1979, Wand and Jones, 1995) and local polynomial (Fan, 1992, Ruppert and Wand, 1994, and Fan and Gijbels, 1996) can be employed. The codes for many of them can be found in SAS, S-plus, and Matlab, among others. Thus little programming effort is needed for using our procedure.

Further, our procedure is fast to compute. This can be seen in our simulation studies conducted in Section 5. The main reasons are as follows. In the first step, the calculation just focuses on a particular point and hence the data involved are very few compared with the whole data set. In the second step, the calculation is performed just for several one-dimensional smoothing problems. This is of course very fast compared with the multi-dimensional smoothing techniques used by Hoover, *et al* (1998).

The paper is organized as follows. Section 2 discusses how to obtain the raw estimates of the coefficient functions and their variances. In particular, the approaches for how to deal with the raw estimates of two kinds of functional ANOVA models are presented in detail. In Section 3, we describe how to refine the raw estimates via smoothing. Then in Section 4, the proposed approach is applied to two longitudinal datasets, one of which actually involves a time-dependent covariate. This is quite different from Hoover, *et al* (1998), Brumback and Rice (1998), and Wu and Chiang (1998) since their examples actually involve no time-dependent covariates. These applications show that our methodology is indeed useful and powerful. To compare our method with the kernel method proposed in Hoover, *et al* (1998), extensive simulation studies with models involving time-dependent covariates are conducted in Section 5. In Section 6, some asymptotic results for the local polynomial estimators in the current context are established. They provide useful insights to our methodology when the sample size is large. Technical proofs are given in the Appendix.

2 Raw Estimates

Let $\{t_j, j = 1, 2, \cdots, T\}$ be the distinct time points among $\{t_{ij}, j = 1, 2, \cdots, T_i, i = 1, 2, \cdots, n\}$. For each given time t_j, let N_j be the collection of the subject indices of all y_{ij} observed at t_j. Collect all $\mathbf{X}_{ij}$ and y_{ij} whose subject indices are in N_j and form the design matrix $\tilde{\mathbf{X}}_j$ and the response vector $\tilde{\mathbf{Y}}_j$ respectively. Then from model (1.2), the data collected at time t_j follow the linear model

$$\tilde{\mathbf{Y}}_j = \tilde{\mathbf{X}}_j \beta(t_j) + \tilde{e}_j, \tag{2.1}$$

where $\tilde{e}_j$ is defined similarly to $\tilde{\mathbf{Y}}_j$ and $\tilde{\mathbf{X}}_j$. Note that

$$\mathrm{E}(\tilde{e}_j) = 0, \qquad \mathrm{cov}(\tilde{e}_j) = \gamma(t_j, t_j) I_{n_j},$$

where n_j denotes the number of subjects observed at time t_j, namely, n_j is the number of the elements in N_j. Clearly model (2.1) is a standard linear model.

Assume $\text{Rank}(\tilde{\mathbf{X}}_j) = d$ (see Remark 2.1 for discussions on the case $\text{Rank}(\tilde{\mathbf{X}}_j) < d$). Then the standard least-squares theory shows that $b(t_j) = (\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^T\tilde{\mathbf{Y}}_j$ is an estimator of $\beta(t_j)$ with

$$\text{E}(b(t_j)) = \beta(t_j), \qquad \text{cov}(b(t_j)) = \gamma(t_j, t_j)(\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j)^{-1}.$$

For $r = 1, 2, \cdots, d$, let $b_r(t_j)$ be the r-th component of $b(t_j)$. Then

$$b_r(t_j) = e_{r,d}^T(\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^T\tilde{\mathbf{Y}}_j, \qquad \text{E}(b_r(t_j)|\mathcal{D}) = \beta_r(t_j), \tag{2.2}$$

and

$$\text{cov}(b_r(t_j), b_r(t_k)|\mathcal{D}) = \gamma(t_j, t_k)e_{r,d}^T(\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^T M_{jk}\tilde{\mathbf{X}}_k(\tilde{\mathbf{X}}_k^T\tilde{\mathbf{X}}_k)^{-1}e_{r,d}, \tag{2.3}$$

where here and throughout $\mathcal{D} = \{(\mathbf{X}_{ij}, t_j), j = 1, 2, \cdots, T; i = 1, 2, \cdots, n\}$ and $e_{r,d}$ stands for a d-dimensional unit vector with one at its r^{th} entry. If the α^{th} entry of $\tilde{\mathbf{Y}}_j$ and the β^{th} entry of $\tilde{\mathbf{Y}}_k$ come from the same subject, the $(\alpha, \beta)^{th}$ entry of M_{jk} takes value 1 and otherwise 0. It is worthwhile to notice that M_{jj} is an identity matrix which results in a simpler expression for the variance of $b_r(t_j)$:

$$\text{Var}(b_r(t_j)) = \gamma(t_j, t_j)e_{r,d}^T(\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j)^{-1}e_{r,d}. \tag{2.4}$$

To estimate the covariance of $b_r(t_j)$ and $b_r(t_k)$, we need to estimate $\gamma(t_j, t_k)$. Let $\hat{\tilde{e}}_j = (I_{n_j} - P_j)\tilde{\mathbf{Y}}_j$ denote the residuals from the least-squares fit where $P_j = \tilde{\mathbf{X}}_j(\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^T$. It follows that

$$\text{Etr}\{\hat{\tilde{e}}_j\hat{\tilde{e}}_k^T\} = \text{tr}\{(I_{n_k} - P_k)M_{jk}^T(I_{n_j} - P_j)^T\}\gamma(t_j, t_k).$$

If $\text{tr}\{(I_{n_k} - P_k)M_{jk}^T(I_{n_j} - P_j)^T\} \neq 0$, then a natural estimator for $\gamma(t_j, t_k)$ is given by

$$\hat{\gamma}(t_j, t_k) = \text{tr}\{\hat{\tilde{e}}_j\hat{\tilde{e}}_k^T\}/\text{tr}\{(I_{n_k} - P_k)M_{jk}^T(I_{n_j} - P_j)^T\}. \tag{2.5}$$

In particular, when $j = k$ and $n_j > d$, we have

$$\hat{\gamma}(t_j, t_j) = \hat{\tilde{e}}_j^T\hat{\tilde{e}}_j/(n_j - d).$$

An estimator for $\text{cov}(b_r(t_j), b_r(t_k)|\mathcal{D})$ can be obtained via replacing $\gamma(t_j, t_k)$ by $\hat{\gamma}(t_j, t_k)$ in (2.3).

Remark 2.1 If $\text{Rank}(\tilde{\mathbf{X}}_j) < d$, we can not get a raw estimate for $\beta(t_j)$. There are four methods to handle this situation. The first method is to leave it missing. If there are only a few such time points, we can estimate the corresponding missing values by smoothing the unmissing raw estimates. The second method is to increase the size of neighborhood. For instance, we can use all observations at time points t_{j-1}, t_j and t_{j+1} to fit the model (1.2) with $t = t_j$. The third method is to impute some of missing observations via getting information from the neighboring

time points. For example, one can use observations at time points t_{j-1} and t_{j+1} to impute the observations at t_j. Note that as long as $\beta(t)$ is smooth and the time window is small, the biases created by the second and third methods are negligible. The fourth method is via using a binning technique. This is particularly the case when the data are heavily missing or the scheduled time points are not the same for all subjects. Examples of using binning techniques can be found in Fan and Marron (1994).

We now turn to discuss a class of special functional linear models–functional ANOVA models whose covariates are time-invariant. By introducing some dummy covariates, these models can be written in the form of model (1.2). However, due to their special structures, the functional ANOVA models should be handled with special care.

2.1 Nested Functional ANOVA

We here consider only a two-level nested functional ANOVA model for simplicity of presentation. The basic ideas can be extended easily to general cases of multiple-level of nesting. The motivation of our study comes from an analysis of the progesterone curves measured over 21 conceptive and 70 nonconceptive women's menstrual cycles (top level nesting, namely, group effects). A woman in the nonconceptive group can have as many as 5 cycles of data for analysis (second level of nesting, namely, subject effects). See Brumback and Rice (1998) and Section 4.2 for more details.

A two-level nested functional ANOVA is of the form:

$$y_{ijk}(t) = \alpha_i(t) + \beta_{ij}(t) + e_{ijk}(t), \tag{2.6}$$

where $k = 1, 2, \cdots, K_{ij}$ (number of cycles of subject j in group i); $j = 1, 2, \cdots, J_i$ and $i = 1, 2, \cdots, I$. The coefficient functions $\alpha_i(t)$ and $\beta_{ij}(t)$ are assumed to be smooth; they are the first and second level effects respectively. The terms $e_{ijk}(t)$ are the error processes with mean function 0 and common covariance function $\gamma(s, t)$. To make model (2.6) identifiable, the second level effects should satisfy some identifiability conditions, say,

$$\sum_{j=1}^{J_i} \beta_{ij}(t) = 0, i = 1, 2, \cdots, I. \tag{2.7}$$

Note that model (2.6) is a special case of (1.2).

Let δ_{ijkl} be 1 if $y_{ijk}(t_l)$ is observed and 0 otherwise. Then, the raw estimates (2.2) and their variances for the first level effects $\alpha_i(t_l)$ $(i = 1, 2, \cdots, I)$ are given by

$$\hat{\alpha}_i(t_l) = \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} y_{ijk}(t_l)\delta_{ijkl} / \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} \delta_{ijkl}, \quad \mathrm{Var}\{\hat{\alpha}_i(t_l)\} = \gamma(t_l, t_l) / \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} \delta_{ijkl},$$

if $\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}}\delta_{ijkl} > 0$; otherwise, $\hat{\alpha}_i(t_l)$ and its variance are left as missing. The raw estimates and their variances for the sum $\alpha_i(t_l)+\beta_{ij}(t_l)(j=1,2,\cdots,J_i; i=1,2,\cdots,I)$ are given by

$$\hat{\alpha}_i(t_l)+\hat{\beta}_{ij}(t_l) = \sum_{k=1}^{K_{ij}} y_{ijk}(t_l)\delta_{ijkl} / \sum_{k=1}^{K_{ij}}\delta_{ijkl}, \quad \mathrm{Var}\{\hat{\alpha}_i(t_l)+\hat{\beta}_{ij}(t_l)\} = \gamma(t_l,t_l) / \sum_{k=1}^{K_{ij}}\delta_{ijkl},$$

if $\sum_{k=1}^{K_{ij}}\delta_{ijkl} > 0$; otherwise, leave them missing. Obviously these raw estimates and their variances are consistent with the least-squares estimators.

If only a few raw estimates are missing, they can be estimated by using unmissing raw estimates via smoothing, say. Otherwise, we can use the upper level effects as substitutes. For example, if $\hat{\alpha}_i(t_l)+\hat{\beta}_{ij}(t_l)$ is missing, it can be estimated by $\hat{\alpha}_i(t_l)$ via setting $\hat{\beta}_{ij}(t_l)=0$. The corresponding variance is assumed to be the sum of $\mathrm{Var}(\hat{\alpha}_i(t_l))$ and the average of the variances of those unmissing estimates $\hat{\alpha}_i(t_l)+\hat{\beta}_{ij}(t_l)$. These ideas can also be employed to impute the missing observations.

2.2 Crossed Functional ANOVA

We discuss only a two-way crossed functional ANOVA model. Multiple-way crossed functional ANOVA models can similarly be dealt with. A 2-way crossed functional ANOVA model is of the form:

$$y_{ij}(t) = \mu(t) + b_i(t) + \tau_j(t) + e_{ij}(t), \tag{2.8}$$

where $i=1,2,\cdots,I_b; j=1,2,\cdots,J_\tau$. The function $\mu(t)$ is the grand mean function, $b_i(t)$ the block effect at level i and $\tau_j(t)$ the treatment effect at level j. In the expression (2.8), the functions $e_{ij}(t)$ are error processes with mean function 0 and common covariance function $\gamma(s,t)$. To make model (2.8) identifiable, we impose the following conditions for the block and treatment effects:

$$\sum_{i=1}^{I_b} b_i(t) = 0, \qquad \sum_{j=1}^{J_\tau}\tau_j(t) = 0. \tag{2.9}$$

Let $\delta_{ijl} = 1$ if $y_{ij}(t_l)$ is observed and 0 otherwise. The approaches for calculating the raw estimates and their variances of the grand means, the block and the treatment effects are similar to those in the nested functional ANOVA models. For example, we compute the raw estimates and their variances of the grand means by

$$\hat{\mu}(t_l) = \sum_{i=1}^{I_b}\sum_{j=1}^{J_\tau} y_{ij}(t_l)\delta_{ijl} / \sum_{i=1}^{I_b}\sum_{j=1}^{J_\tau}\delta_{ijl}, \quad \mathrm{Var}\{\hat{\mu}(t_l)\} = \gamma(t_l,t_l) / \sum_{i=1}^{I_b}\sum_{j=1}^{J_\tau}\delta_{ijl},$$

if $\sum_{i=1}^{I_b}\sum_{j=1}^{J_\tau}\delta_{ijl} > 0$; otherwise, we leave them missing.

3 Refining the Raw Estimates

There are several reasons for us to refine the raw estimates obtained in last section. Firstly, the raw estimates are generally not smooth. Secondly, they are inefficient since they haven't used the information from the neighboring time points and hence their efficiency can be improved. Thirdly, there may be some missing raw estimates due to insufficient amount of data around some time points and it is desirable to impute them. Finally, we may also want to estimate the values of the coefficient curves at nondesign points.

A natural way to refine the raw estimates is to smooth them over time. We now describe briefly how to smooth the raw estimates $\{(t_j, b_r(t_j)), j = 1, 2, \cdots, T\}$ for obtaining the smooth coefficient function $\hat{\beta}_r(t)$ via one of the existing smoothing techniques. Most of the existing smoothing techniques are linear in the responses. Suppose $\beta_r(t)$ is $(p+1)$-times continuously differentiable and we wish to estimate its q-th derivative for some $0 \leq q < p+1$. Then a typical linear estimator is given by

$$\widehat{\beta_r^{(q)}}(t) = \sum_{j=1}^{T} w_r(t_j, t) b_r(t_j), \tag{3.1}$$

where the weights $w_r(t_j, t)$ can be constructed by various smoothing techniques such as spline, kernel or local linear regression.

Simple calculation shows that

$$\begin{aligned} \mathrm{E}(\widehat{\beta_r^{(q)}}(t)|\mathcal{D}) &= \sum_{j=1}^{T} w_r(t_j, t)\beta_r(t_j), & (3.2) \\ \mathrm{Var}(\widehat{\beta_r^{(q)}}(t)|\mathcal{D}) &= \sum_{j=1}^{T}\sum_{k=1}^{T} w_r(t_j, t) w_r(t_k, t) \mathrm{cov}(b_r(t_j), b_r(t_k)|\mathcal{D}). & (3.3) \end{aligned}$$

By the discussions given in Section 2, $\mathrm{cov}(b_r(t_j), b_r(t_l)|\mathcal{D})$ can be estimated by using (2.3) and (2.5). Then the ± 2 standard error bands can be constructed by

$$\widehat{\beta_r^{(q)}}(t) \pm 2\{\widehat{\mathrm{Var}}(\widehat{\beta_r^{(q)}}(t)|\mathcal{D})\}^{1/2}, \tag{3.4}$$

which is also called a 95% pointwise confidence interval by some authors on the ground that the bias term is also ignored in constructing confidence intervals for parametric models since these parametric models hold at best approximately.

We now turn to local polynomial fitting. Let $C_j = (1, t_j - t, \cdots, (t_j - t)^p)^T, j = 1, 2, \cdots, T$ and $K_h(t) = K(t/h)/h$ be a kernel function with a bandwidth h. Then

$$w_{q,p+1}(t_j, t) = q! e_{q+1,p+1}^T (C^T W C)^{-1} C_j W_j, \quad j = 1, 2, \cdots, T, \tag{3.5}$$

are the local polynomial weights for estimating the q-th derivative of an underlying function where $C = (C_1, C_2, \cdots, C_T)^T$ and $W = \text{diag}(W_1, \cdots, W_T)$ with $W_j = K_h(t_j - t)$. In particular, the local linear weights are given by $w_{0,2}(t_j, t), j = 1, 2, \cdots, T$. See Fan and Gijbels (1996) for details.

The variances of the raw estimates obtained in Section 2 often take the form $a^2(t)\sigma^2(t)$ where $a^2(t)$ is a known function taking positive values. For example, in the expression of $\text{Var}(b_r(t_j))$ in (2.4), we have $a^2(t_j) = e_{r,d}^T(\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j)^{-1}e_{r,d}$ and $\sigma^2(t_j) = \gamma(t_j, t_j)$. Thus, the data $\{(t_j, b_r(t_j)), j = 1, 2, \cdots, T\}$ are heteroscedastic. Note that $\sigma^2(t)$ may vary slowly if we assume it is smooth. However, $a^2(t)$ may change dramatically due to different numbers of data points observed at different times. This knowledge can be incorporated in the construction of the local polynomial weights $w_{q,p+1}(t_j, t), j = 1, 2, \cdots, T$ so that the refined estimates can be improved further. For example, the local polynomial fit can be more effective if the kernel weight $K_h(t_j - t)$ is replaced by $K_h(t_j - t)/a^2(t_j)$. The standard errors for the weighted local polynomial fit can be similarly obtained.

4 Applications to Longitudinal Data

4.1 CD4 Cell Percentage in HIV Seroconverters

Human immune-deficiency virus(HIV) destroys CD4 cells (T-lymphocytes, a vital component of the immune system) so that the number or percentage of the CD4 cells in the blood of a human body will change after the human subject is infected with HIV. Thus the CD4 cell level marks the disease progression of a subject. To use the CD4 marker effectively in studies of new therapies or for monitoring individual subjects, it is important to build some statistical models for the CD4 cell counts or percentage. For CD4 cell counts, Lange, *et al* (1992) proposed some Bayesian models while Zeger and Diggle (1994) employed a semiparametric model. For further related references, see Lange, *et al* (1992).

The data set came from the Multi-Center AIDS Cohort Study. It contains the HIV status of 283 homosexual men who were infected with HIV during the follow-up period between 1984 and 1991. See Kaslow, *et al* (1987) for the related design, methods and medical implications of this study. The response variable is the CD4 cell percentage of a subject at distinct time points after HIV infection. We took three covariates for this study. The first one takes binary values 1 or 0, according to whether a subject is a smoker or nonsmoker. The second covariate is the age of a subject at the time when the measurement was collected and hence it is time-dependent. The third covariate is the CD4 cell percentage level before HIV infection. Our model can be written as follows:

$$Y(t) = \beta_0(t) + \beta_1(t)\ \text{Smoking} + \beta_2(t)\ \text{Age}(t) + \beta_3(t)\ \text{PreCD4} + e(t), \tag{4.1}$$

where $Y(t)$ is the % of CD4 cells at time t. In the data, the time point t_{ij} indicates the time (in years) when the i^{th} subject paid his j^{th} visit after HIV infection. All subjects were scheduled to pay their visits twice a year but the concrete time points for different subjects are not the same. The aim of this study is to assess the effects of cigarette smoking, age at the disease progression and pre-HIV infection CD4 cell percentage on the CD4 cell percentage depletion over time.

For a clear interpretation of the coefficient functions, we centralized the variables Age(t) and PreCD4 so that their sample means are zero. As a result, the intercept function $\beta_0(t)$ can be interpreted as the baseline CD4 percentage curve for a nonsmoker with average pre-infection CD4 percentage and average age. See Wu and Chiang(1998) for a detailed account of other advantages of such a normalization.

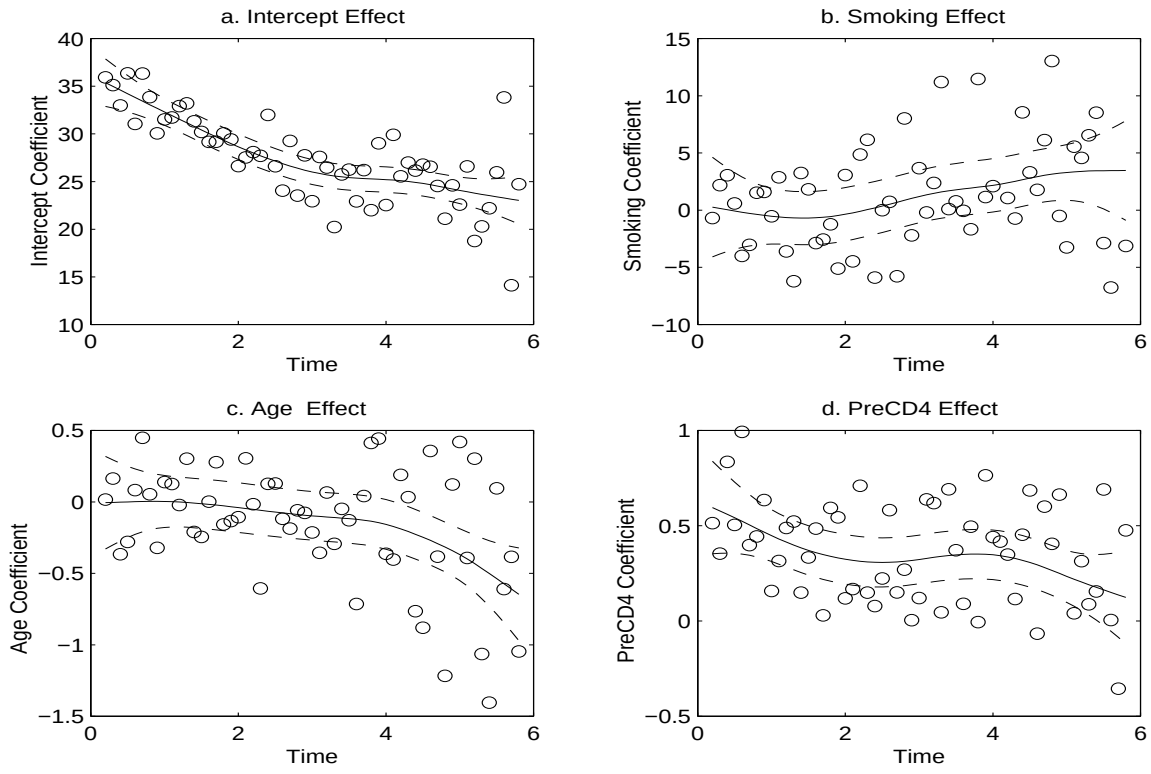

Figure 1: Estimated coefficient curves for the baseline CD4 percentage and the effects of Smoking, Age and PreCD4 on the percentage of CD4 cells. Solid curves—smoothed effects; dashed curves —±2 pointwise standard error bands; circles—raw estimates.

Figure 1 depicts the fitted coefficient functions (solid curves) with ±2 pointwise standard error bands (dashed curves). The circles indicate the raw estimates of the coefficient functions at the possible visiting time points. There are some outliers in the raw estimates (off the scale of the plots) and they were deleted before the smoothing was performed. As an example, here the fitted coefficient functions are obtained via smoothing the raw estimates of each coefficient function respectively by a cubic smoothing spline fit (Green and Silverman, 1994) with smoothing parameters

chosen by cross-validation. It is worthwhile to mention that the smoothing parameters selected by cross-validation for all CD4 coefficient functions are about the same, indicating that they admit similar amount of smoothness.

The fitted intercept function (baseline CD4 percentage curve) is displayed in Figure 1 (a). It has a quick drop during the first three years and a slower drop afterwards. The fitted smoking coefficient function is displayed in Figure 1 (b). It seems that $\beta_1(t) \geq 0$ for most of the time. This may suggest that the smoking population has higher CD4 percentage if we hold other covariates fixed. The suggestion, however, may not be so convincing since the estimated standard error bands cover 0 most of the time. The age effect in general decreases over time and is more pronounced as time evolves, as shown in Figure 1(c). The estimated standard error bands suggest that the age effect is probably near zero within the first four years but not afterwards. The effect of the pre-HIV CD4 cell percentage seems generally decreasing with time, and far from zero since the estimated standard error bands do not cover 0 except near the end of the study.

4.2 Progesterone Data Analysis

The data used here are a sample of urinary metabolite progesterone curves (Munro , *et al*, 1991) measured over 21 conceptive and 70 nonconceptive women menstrual cycles. A woman in the nonconceptive group can be measured up to 5 menstrual cycles while she contributes only one cycle if she is in the conceptive group. The data have been aligned and truncated around the day of ovulation so that the data curves have the same design points. Due to various reasons, not all measurements in a menstrual cycle are available, and this results in some missing responses in some cycles. This curve data set has been carefully studied in Brumback and Rice (1998) as an interesting illustration of their smoothing spline models for the analysis of nested samples of curves. Unlike for the CD4 data example presented in the previous subsection, where a smoothing spline fit is used in the smoothing step, as an example, here the raw estimates are smoothed by local linear regression with the Gaussian kernel, and the bandwidths are selected by the data-driven method of Ruppert, *et al* (1995). Since the covariance function of the raw estimates is about n^{-1} of that of a subject [see (2.3) and (2.4)], the dependence of the raw estimates has little effect on the bandwidth selection.

Figures 2 (a) and (b) depict the fitted coefficient curves of the nonconceptive and conceptive group effects (solid curves) and ± 2 pointwise standard error bands (dashed curves). Their raw estimates are indicated by the circles which clearly show the shapes of the underlying group effect curves. While these two group effect curves progress similarly during 8 days before and after the day of the ovulation, they show different tendencies from the eighth day after the ovulation: the progesterone curve for the nonconceptive group decreases rapidly while the progesterone curve for

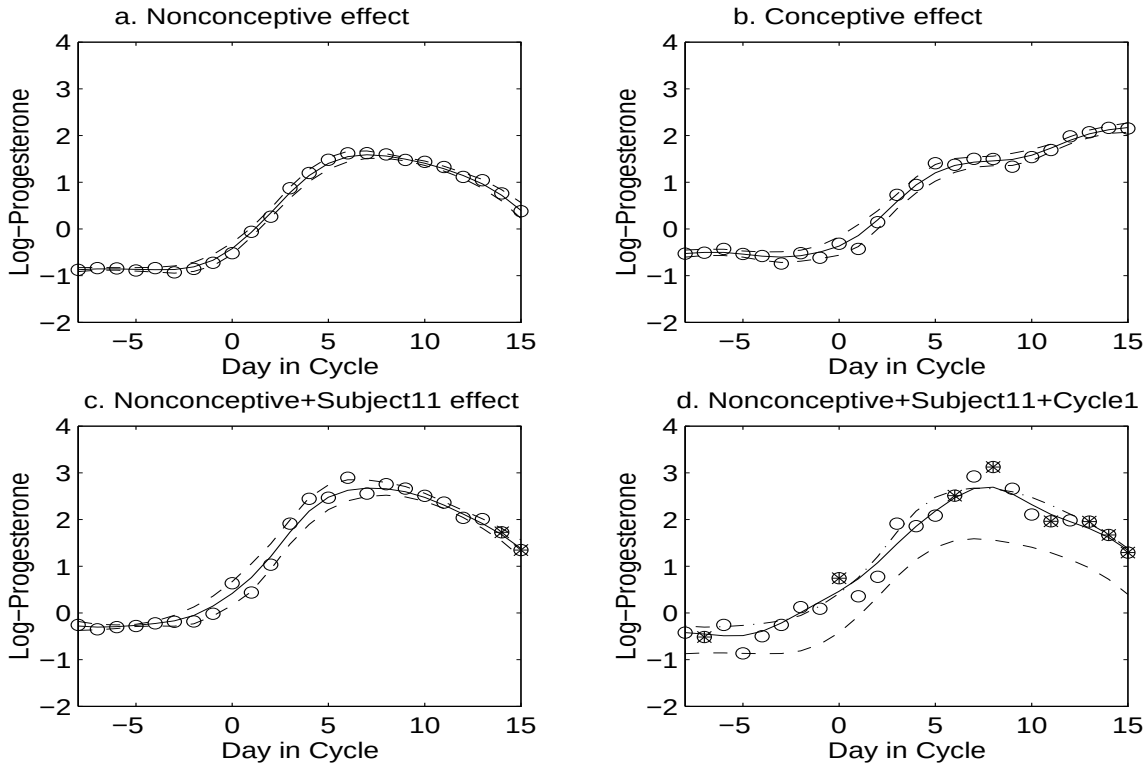

Figure 2: Estimated coefficient curves for the progesterone data. In (a)-(c), solid curves—smoothed effects; dashed curves—±2 standard error bands; circles—raw estimates; stars—imputed raw estimates. In (d), dashed curve—smoothed nonconceptive effects; dotdashed curve—smoothed nonconceptive+ Subject 11 effects; solid curve—smoothed effect for Cycle 1.

the conceptive group increases steadily. This can possibly be applied for self-administered assays of detecting fertile periods and early pregnancy. In this nested functional ANOVA model, there are 91 estimated coefficient functions. We only selectively report some of them for illustration. The subject effect curve for subject 11 is presented in Figure 2 (c). It is noticed that the standard error bands here are substantially wider than those for the group effects since we now use only the data within Subject 11. Figure 2 (*d*) presents the smoothed effect (solid curve) for Cycle 1 of Subject 11. The raw estimates here for the cycle effect are actually the observations or imputed values, indicated respectively by circles or stars in the figure. The nonconceptive effect curve and the nonconceptive plus the effect curve of Subject 11 are superimposed there for comparison.

5 Simulation Studies

The aim of this section is to compare the performance of our two-step procedure with that of the kernel method proposed in Hoover, *et al* (1998) via simulation studies. Although the spline approach of Hoover, *et al* (1998) is a nice one to compete with, we opt for not doing so due to the

intensive computation of the spline approach, not to mention the difficulty in choosing multiple smoothing parameters.

The leaving-one-subject-out cross-validation method is used to select the bandwidth for the kernel method of Hoover, *et al* (1998). For our two-step procedure, the bandwidth selector proposed by Ruppert, *et al* (1995) will be employed since a local linear fit is used in the smoothing step.

In this simulation study, two models will be explored. The first model tries to mimic the CD4 data set. The covariates of the CD4 data set are kept fixed and the true coefficient curves are taken as the solid ones presented in Figure 1. Following Wu and Chiang (1998), we shall sample the errors ε_{ij} from the Gaussian process with zero mean and covariance function:

$$\mathrm{cov}(\varepsilon_{i_1 j_1}, \varepsilon_{i_2 j_2}) = \begin{cases} 16\exp(-|t_{i_1 j_1} - t_{i_2 j_2}|), & \text{if } i_1 = i_2, \\ 0, & \text{if } i_1 \neq i_2. \end{cases}$$

This is a decayed exponential stationary covariance function, indicating the correlation will be decreasing with time. The variance factor 16 is chosen differently from the one .0625 given by Wu and Chiang (1998) since the standard deviation of the CD4 data for each subject is about 4. The scheduled distinct time points for a simulated data set are chosen similarly to those in the original CD4 data. For each subject, about 12 time points are randomly selected from the set $\{t_j = 0.1j, j = 1, \cdots, 60\}$ to make the simulated data sufficiently similar to the original CD4 data. The observed data are then the sum of the errors and the expected values at various time points. That is,

$$Y_{ij} = X_{ij}^T \beta(t_{ij}) + \varepsilon_{ij}, \quad j = 1, 2, \cdots, T_i; \quad i = 1, 2, \cdots, n,$$

with β being the fitted coefficient functions presented in Figure 1.

We sampled 201 data sets from this model and fitted them respectively by the two-step method and the kernel method. The performance of a fit is measured by its Mean Absolute Deviation Error (MADE) from the true curves, defined as

$$\mathrm{MADE} = (4T)^{-1} \sum_{j=1}^{T} \sum_{r=0}^{3} |\beta_r(t_j) - \hat{\beta}_r(t_j)| / \mathrm{range}(\beta_r),$$

where $\mathrm{range}(\beta_r)$ is the range of the function $\beta_r(t)$. The weights are introduced to account for the different scales of the coefficient functions. Traditionally, the performance of a fit may also be measured by its Weighted Average Squared Error (WASE), defined as

$$\mathrm{WASE} = (4T)^{-1} \sum_{j=1}^{T} \sum_{r=0}^{3} (\beta_r(t_j) - \hat{\beta}_r(t_j))^2 / \mathrm{range}^2(\beta_r), \tag{5.1}$$

or its Unweighted Average Squared Error(UASE), defined similarly to WASE but with no weights in the equation (5.1).

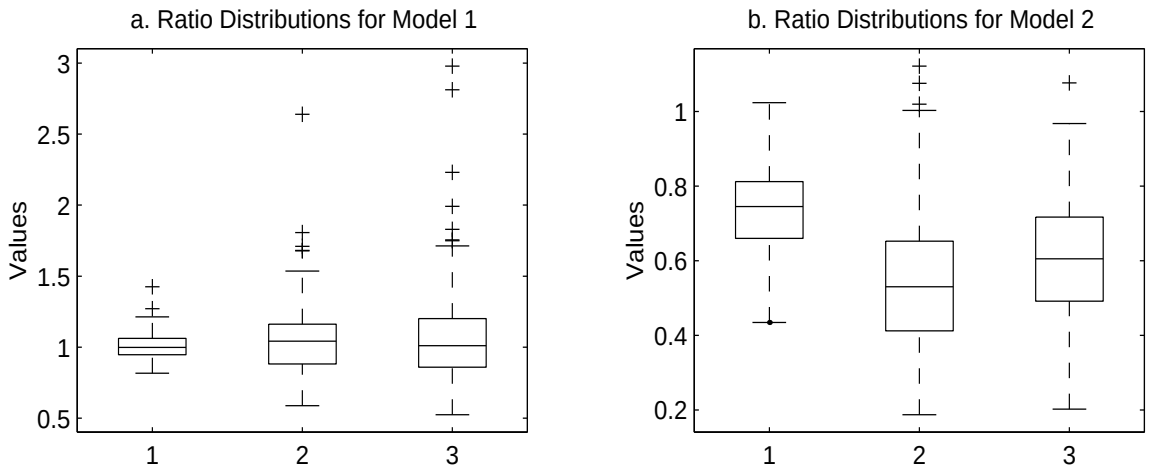

Figure 3: Comparison of the two-step method with the kernel method. (a) Boxplots for the ratios (two-step/kernel) of MADEs (panel 1), WASEs (panel 2) and UASEs (panel 3) for Model 1. (b) Same caption but now for Model 2.

The boxplots of the MADE, WASE, UASE ratios (two-step/kernel) are presented in panels 1, 2 and 3 of Figure 3 (a) respectively. It seems that both methods perform pretty comparably for all three measures since the underlying functions admit similar degrees of smoothness. Hence, the advantages of the two-step estimator do not show up in this simple situation. However, the computation time of the two-step method is only about $1/30 \sim 1/50$ of that for the kernel method.

Let us compare the median performance of both methods. The median performance is indicated by a fitted coefficient curve whose MADE, say, attains the median value among 201 simulations. Since the simulated data sets, for which the two-step method and the kernel method achieve the median performance, are not necessarily the same, we compare the coefficient curves with median performance of one method with those coefficient curves fitted from the same data set using the other method. Examination of the resulting plots (which are omitted here for space saving) reveals that the kernel method generally undersmooths some or all of the true coefficient curves. This fact has also been noticed by Hoover, *et al* (1998).

The coefficient functions in the first model simulation admit quite similar amounts of smoothness (the selected bandwidths or smoothing parameters are close to each other, as observed in Section 4.1). This explains why the two-step method and the kernel methods perform similarly for this simulation study. The CD4 coefficient functions are not challenging enough for the two-step method. In the second simulation model, we test both methods by using somewhat more inhomogeneous functions.

The second model of our simulation study is designed as follows. Four true coefficient functions

are chosen as:

$$\beta_0(t) = 15 + 8.7\sin(2\pi t), \quad \beta_1(t) = 4 - 17(t - 1/2)^2,$$
$$\beta_2(t) = 1 + 11.2t, \qquad\qquad \beta_3(t) = 1 + 2t^2 + 11.3(1 - t)^3.$$

They represent four different types of curves. The four covariates are chosen as follows. First of all, we let $X_0(t) \equiv 1$. We then let $X_1(t)$ be a binomial random variable with probability of success $p = .6$ and let $X_2(t)$ be a uniform random variable over the time-dependent interval $[t/4, 1 + 3t/4]$. Finally we let $X_3(t)$, when conditioning on $X_2(t)$, be a normal random variable with mean zero and conditional variance

$$\mathrm{var}(X_3(t)|X_2(t)) = (1 + X_2(t))/(2 + X_2(t)).$$

As in the first simulation study, the errors are sampled (independently from the covariates) from a stationary Gaussian process with zero mean and a decayed exponential covariance function:

$$\mathrm{cov}(\varepsilon_{i_1 j_1}, \varepsilon_{i_2 j_2}) = \begin{cases} 5.27\exp(-.5|t_{i_1 j_1} - t_{i_2 j_2}|), & \text{if } i_1 = i_2, \\ 0, & \text{if } i_1 \neq i_2. \end{cases}$$

Note that the correlation is larger for the present simulation study.

Without loss of generality, we let the time interval be $[0, 1]$. We also chose $N = 100$ subjects and $T = 45$ time points. These T time points are equi-spaced over $[0, 1]$. For each subject, we let 60% of data be randomly missing so that unequal numbers of observations for subjects are obtained. The expected number of data points for a simulation data set is 1800.

As we did in our first model simulation, we sampled 201 data sets from the above model, calculated their MADEs, WASEs, UASEs for both the two-step and the kernel methods, and then presented their ratio boxplots in Figure 3 (b). We can see that the two-step method has much better performance using all three accuracy measures. Examination of the median performance reveals the same conclusion as that for Model 1 and the computation time for the two-step method is about $1/30 \sim 1/50$ of that for the kernel method.

6 Asymptotic results

We first impose some conditions on the covariance structure of $\varepsilon_i(t)$ in model (1.3). We assume that the error $\varepsilon_i(t)$ consists of two parts: trajectory (subject) effect $v_i(t)$ and measurement error process $e_i(t)$ so that

$$\varepsilon_i(t) = v_i(t) + e_i(t). \tag{6.1}$$

This formulation is a generalization of that in Section 5.6 of Diggle, *et al* (1994). The trajectory process $\{v_i(t)\}$ is assumed to be continuous with covariance function $\gamma_0(s, t)$ and the noise process

$\{e_i(t)\}$ is assumed to be uncorrelated with the variance function $\sigma^2(t)$. Thus, the covariance function of $\{\varepsilon_i(t)\}$ is

$$\gamma(s,t) = \gamma_0(s,t) + \sigma^2(t)1_{\{s=t\}}.$$

As in Zeger and Diggle (1994), the covariance function $\gamma(s,t)$ is not necessarily continuous around the diagonal elements.

Local polynomial fitting technique is used for smoothing the raw estimates because of their good sampling properties (Fan and Gijbels, 1996). To get some further insight on the refined estimates, some asymptotic results will be derived for estimation at an interior point in the support of the design density. The treatments for boundary points are along the same lines and are omitted here. The local polynomial estimator of the q-th derivative of $\beta_r(t)$ based on the raw estimates $b_r(t_j), j = 1, 2, \cdots, T$ is as follows:

$$\widehat{\beta_r^{(q)}}(t) = \sum_{j=1}^{T} w_{q,p+1}(t_j, t) b_r(t_j), \quad q = 0, 1, 2, \cdots, p, \tag{6.2}$$

where $w_{q,p+1}$ is given in (3.5). Let $K_{q,p+1}$ be the equivalent kernel of the local polynomial fit (see Fan and Gijbels, 1996), defined by

$$K_{q,p+1}(t) = e_{q+1,p+1}^T S^{-1}(1, t, \cdots, t^p)^T K(t), \tag{6.3}$$

with $S = (s_{ij})_{i,j=0,1,\cdots,p}$ and $s_{ij} = \int K(u)u^{i+j}du$.

We first derive the asymptotic bias. Since $\mathrm{E}\{b_r(t_j)|\mathcal{D}\} = \beta_r(t_j), j = 1, 2, \cdots, T$, the correlation within subjects doesn't affect the bias structure of the estimator. This leads to the following theorem. The technical conditions and the proofs of the theorems are given in the Appendix.

Theorem 6.1 *Suppose Condition A_1 in the Appendix holds. Then when $h \to 0$ and $Th \to \infty$ as $T \to \infty$,*

$$Bias(\widehat{\beta_r^{(q)}}(t)|\mathcal{D}) = \frac{q!\beta_r^{(p+1)}(t)h^{p-q+1}}{(p+1)!} B_{p+1}(K_{q,p+1})(1 + o_p(1)),$$

where $B_{p+1}(K) = \int K(u)u^{p+1}du$.

It is more involved to derive the asymptotic variance of the estimator (6.2). The main difficulty is that the variance-covariance structure of the raw estimates $b_r(t_j), j = 1, 2, \cdots, T$ is very complicated. Let n_j, n_k and n_{jk} be the numbers of elements in N_j, N_k and $N_j \cap N_k$ respectively. Set $\Omega_l = \mathrm{E}(\mathbf{X}_{1l}\mathbf{X}_{1l}^T), l = j, k$ and $\Omega_{jk} = \mathrm{E}(\mathbf{X}_{1j}\mathbf{X}_{1k}^T)$ for all $j, k = 1, 2, \cdots, T$. Then by the Law of Large Numbers and Condition A_{61}, we deduce from (2.3) that

$$\mathrm{cov}(b_r(t_j), b_r(t_k)|\mathcal{D}) = \gamma(t_j, t_k)\frac{n_{jk}}{n_j n_k} e_{r,d}^T \Omega_j^{-1}\Omega_{jk}\Omega_k^{-1} e_{r,d}(1 + o_p(1)) \tag{6.4}$$

when n_j, n_k and n_{jk} are large. In particular,

$$\mathrm{Var}(b_r(t_j)|\mathcal{D}) = \gamma(t_j,t_j)e_{r,d}^T\Omega_j^{-1}e_{r,d}/n_j(1+o_p(1)).$$

If the covariates $\mathbf{X}_{ij}$ satisfy Condition A_{62}, i.e., they are time-invariant as those in the progesterone data, then $\Omega_j = \Omega_k = \Omega_{jk} = \Omega_1$ for all j and k. In this case, the expression (6.4) can be simplified as

$$\mathrm{cov}(b_r(t_j), b_r(t_k)|\mathcal{D}) = \omega^{rr}\gamma(t_j,t_k)\frac{n_{jk}}{n_j n_k}(1+o_p(1)), \tag{6.5}$$

where $\omega^{rr} = e_{r,d}^T\Omega_1^{-1}e_{r,d}$, the $(r,r)^{th}$ entry of Ω_1^{-1}.

We now derive the asymptotic variance for two specific situations: n_{jk} is either small or large. Let $I_t = \{j : |t_j - t| \le h\}$ be the indices of the local neighborhood. In some situations, n_{jk} may be much smaller than n_j or n_k for all $j \neq k$, $j,k \in I_t$ and $n_j, j \in I_t$ are about the same proportion of n. In other words, we have $n_{jk}^2/(n_j n_k) \approx 0$ and $n_j \approx cn$ for some constant $0 < c < 1$ for $j \neq k, (j,k \in I_t)$. These situations approximately satisfy the conditions of the following theorem.

Theorem 6.2 *Under Conditions $A_1 - A_3, A_5$ and A_{62} in the Appendix, if $\gamma(t,t)$ is continuous for all t and*

$$n_{jk}/(n_j n_k) = \begin{cases} o\{1/(nTh^{2q+1})\}, & j \neq k, \\ 1/(cn) + o\{1/(nTh^{2q+1})\}, & j = k, \end{cases}$$

holds uniformly for all $j,k \in I_t$ for some constant $0 < c < 1$, then when $h \to 0$ and $nTh^{2q+1} \to \infty$ as $nT \to \infty$,

$$Var(\widehat{\beta_r^{(q)}}(t)|\mathcal{D}) = \frac{q!^2\omega^{rr}\gamma(t,t)}{cnTh^{2q+1}f(t)}V(K_{q,p+1})(1+o_p(1)), \tag{6.6}$$

where $V(K) = \int K^2(u)du$.

It follows that the corresponding asymptotic conditional mean square error (MSE) of $\widehat{\beta_r^{(q)}}(t)$ is given by

$$\begin{aligned} \mathrm{MSE}(\widehat{\beta_r^{(q)}}(t)|\mathcal{D}) &= \{\frac{q!\beta_r^{(p+1)}(t)}{(p+1)!}B_{p+1}(K_{q,p+1})\}^2 h^{2(p-q+1)} \\ &+ \frac{q!^2\omega^{rr}\gamma(t,t)}{cnTh^{2q+1}f(t)}V(K_{q,p+1}) + o_p\{h^{2(p-q+1)} + (nTh^{2q+1})^{-1}\}. \end{aligned}$$

Theorem 6.2 implies that when the sampling is taken very carefully, the correlation influence can be ignored. In this case, the optimal bandwidth is $O\{(nT)^{-1/(2p+3)}\}$, the same as that for uncorrelated data.

In some other situations, n_j, n_k and n_{jk} are about the same as n. A longitudinal data set with no missing values provides an extreme example where $n_{jk} = n_j = n$ for all $j,k = 1,2,\cdots,n$. Let $\gamma_{\alpha,\beta}(s,t)$ denote $\partial^{\alpha+\beta}\gamma_0(s,t)/\partial s^\alpha \partial t^\beta$ for any integers $\alpha, \beta = 0,1,\cdots,p+1$.

Theorem 6.3 *Suppose Conditions $A_1 - A_5$ and A_{62} hold. Assume that $n_{jk}/(n_j n_k) = 1/n + o\{1/n\}$ holds uniformly for all $j, k = 1, 2, \cdots, T$. Then when $h \to 0$ and $nTh^{2q+1} \to \infty$ as $n, T \to \infty$,*

$$\begin{aligned} Var(\widehat{\beta_r^{(q)}}(t)|\mathcal{D}) &= \frac{\omega^{rr}}{n}\{\gamma_{q,q}(t,t) + \frac{2q!\gamma_{q,p+1}(t,t)h^{p-q+1}}{(p+1)!}B_{p+1}(K_{q,p+1})\} \\ &+ \frac{q!^2\sigma^2(t)\omega^{rr}}{nTh^{2q+1}f(t)}V(K_{q,p+1}) + o_p\{n^{-1}h^{p-q+1} + (nTh^{2q+1})^{-1}\}. \end{aligned}$$

When the underlying process $v(t)$ defined in (6.1) is stationary, which is assumed in Section 5.6 of Diggle, *et al* (1994), $\gamma_0(t,t) = \gamma_0(0,0)$ is a constant. Thus, $\gamma_{q,q}(t,t) = 0, \gamma_{q,p+1}(t,t) = 0$ for all $q = 1, 2, \cdots, p$. It follows that the local polynomial derivative estimator will be consistent under milder conditions. For example, we do not need $n \to \infty$. However, if $\gamma_{q,q}(t,t) \neq 0$, the local polynomial estimators in this case are consistent only when $n \to \infty$.

Corollary 6.1 *Under the conditions of Theorem 6.3, if the trajectory process $v(t)$ is stationary, then for all $q = 1, 2, \cdots, p$, we have*

$$Var(\widehat{\beta_r^{(q)}}(t)\mathcal{D}) = \frac{q!^2\sigma^2(t)\omega^{rr}}{nTh^{2q+1}f(t)}V(K_{q,p+1}) + o_p\{(nTh^{2q+1})^{-1}\}, \tag{6.7}$$

and

$$\begin{aligned} MSE(\widehat{\beta_r^{(q)}}(t)|\mathcal{D}) &= \{\frac{q!\beta_r^{(p+1)}(t)}{(p+1)!}B_{p+1}(K_{q,p+1})\}^2 h^{2(p-q+1)} \\ &+ \frac{q!^2\omega^{rr}\sigma^2(t)}{nTh^{2q+1}f(t)}V(K_{q,p+1}) + o_p\{h^{2(p-q+1)} + (nTh^{2q+1})^{-1}\}. \end{aligned}$$

If $\gamma_{q,p+1}(t,t) \neq 0$, then the correlation within a subject will affect the choice of the bandwidth. However, when the subject number n is much larger than the number of the distinct time points T, such an effect is very small and can be ignored.

The similar asymptotic results can be established for both the nested functional ANOVA and the crossed functional ANOVA models since they are special cases of functional linear models. In all asymptotic results, we need only to notice that for the raw estimates of the functional ANOVA models, the corresponding $\omega^{rr} = 1$.

Appendix

A Preliminaries

In this Appendix, we outline the proofs for some asymptotic results given in Section 6. For convenience, we collect technical conditions as follows.

Conditions

A_1. The time points $t_1, t_2, \cdots, t_T$ are a random sample from the probability density f and t is a continuous point of f in the interior of the support of f.

A_2. The noise variance $\sigma^2(t)$ is continuous in the support of f.

A_3. The coefficient function $\beta_r(t)$ is $(p+1)$-times continuously differentiable for some p.

A_4. The covariance function $\gamma_0(s,t)$ of the underlying trajectory process $v(t)$ [see (6.1)] is $(p+1)$-times continuously differentiable for both s and t for some p.

A_5. The kernel function K is a bounded symmetric probability density function with a bounded support $[-1,1]$, say.

A_{61}. For a fixed $j \in \{1,2,\cdots,T\}$, the covariates $\mathbf{X}_{ij}, i = 1,2,\cdots,n$ are independently and identically distributed as $\mathbf{X}_{1j} = (X_{1j1},\cdots,X_{1jd})^T$ with $\Omega_j = \mathrm{E}(\mathbf{X}_{1j}\mathbf{X}_{1j}^T)$ positive definite.

A_{62}. The covariates $\mathbf{X}_{ij}$ satisfy A_{61} and they are time-invariant. That is $\mathbf{X}_{ij} = \mathbf{X}_{i1}$ for all $j = 1,2,\cdots,T$.

Conditions $A_1 - A_5$ are just some regularity conditions for the asymptotic results and are not the weakest possible conditions. They are imposed for convenience of the technical proofs. Condition A_{61} says that for a fixed time point, the covariates for different subjects are independently and identically distributed. Condition A_{62} holds for many longitudinal data sets. One of the datasets presented in Section 4 is a typical example.

Before we proceed to prove the results, we list the following three lemmas on the properties of the local polynomial weights $w_{q,p+1}$ given in (3.5). See Fan and Gijbels (1996), page 64 for a proof of Lemma A.1.

Lemma A.1 *Suppose Conditions A_1 and A_5 hold. If $h \to 0$ and $Th \to \infty$ as $T \to \infty$, then*

$$w_{q,p+1}(t_j,t) = \frac{q!}{Th^{q+1}f(t)} K_{q,p+1}\left(\frac{t_j - t}{h}\right)(1 + o_p(1)), j = 1,2,\cdots,T, \tag{A.1}$$

where $K_{q,p+1}$ is the equivalent kernel defined by (6.3)

Lemma A.2 *Under the conditions given in the Lemma A.1, we have*

$$\sum_{j=1}^{T} w_{q,p+1}(t_j,t)(t_j - t)^k = q!1_{\{k=q\}}, \quad k = 0,1,2,\cdots,p. \tag{A.2}$$

Moreover, by Lemma A.1, we have

$$\sum_{j=1}^{T} w_{q,p+1}(t_j,t)(t_j - t)^{p+1} = q!h^{p-q+1}B_{p+1}(K_{q,p+1})(1 + o_p(1)), \tag{A.3}$$

$$\sum_{j=1}^{T} w_{q,p+1}^2(t_j,t) = \frac{q!^2}{Th^{2q+1}f(t)} V(K_{q,p+1})(1 + o_p(1)), \tag{A.4}$$

where B_{p+1} and V are given in Theorems 5.1 and 5.2, respectively.

Lemma A.3 *Suppose Conditions A_1, A_2, A_4 and A_5 hold. If $h \to 0$ and $Th \to \infty$ as $T \to \infty$, then*

$$\begin{aligned}\sum_{j=1}^{T}\sum_{k=1}^{T} w_{q,p+1}(t_j,t)w_{q,p+1}(t_k,t)\gamma(t_j,t_k) &= \{\gamma_{q,q}(t,t) + \frac{2q!\gamma_{q,p+1}(t,t)h^{p-q+1}}{(p+1)!}B_{p+1}(K_{q,p+1})\} \\ &+ \frac{q!^2\sigma^2(t)}{Th^{2q+1}f(t)}V(K_{q,p+1}) + o_p\{h^{p-q+1} + (Th^{2q+1})^{-1}\},\end{aligned}$$

where $\gamma(s,t) = \gamma_0(s,t) + \sigma^2(t)1_{\{s=t\}}$.

Proof Clearly,

$$\begin{aligned}\sum_{j=1}^{T}\sum_{k=1}^{T} w_{q,p+1}(t_j,t)w_{q,p+1}(t_k,t)\gamma(t_j,t_k) &= \sum_{j=1}^{T}\sum_{k=1}^{T} w_{q,p+1}(t_j,t)w_{q,p+1}(t_k,t)\gamma_0(t_j,t_k) \\ &+ \sum_{j=1}^{T} w_{q,p+1}^2(t_j,t)\sigma^2(t_j)\}.\end{aligned}$$

By Lemma A.2, we obtain that

$$\sum_{j=1}^{T} w_{q,p+1}^2(t_j,t)\sigma^2(t_j) = \frac{q!^2\sigma^2(t)}{Th^{2q+1}f(t)}V(K_{q,p+1})(1+o_p(1)).$$

Under Condition A_4, the Taylor expansion of $\gamma_0(t_j,t_k)$ at (t,t) is given by

$$\gamma_0(t_j,t_k) = \sum_{\alpha=0}^{p+1}\sum_{\beta=0}^{p+1}\gamma_{\alpha,\beta}(t,t)\frac{(t_j-t)^\alpha}{\alpha!}\frac{(t_k-t)^\beta}{\beta!} + o\{(t_j-t)^{p+1}(t_k-t)^{p+1}\}.$$

By Lemma A.2 again, we have

$$\sum_{j=1}^{T}\sum_{k=1}^{T} w_{q,p+1}(t_j,t)w_{q,p+1}(t_k,t)\sum_{\alpha=0}^{p}\sum_{\beta=0}^{p}\gamma_{\alpha,\beta}(t,t)\frac{(t_j-t)^\alpha}{\alpha!}\frac{(t_k-t)^\beta}{\beta!} = \gamma_{q,q}(t,t),$$

$$\sum_{j=1}^{T}\sum_{k=1}^{T} w_{q,p+1}(t_j,t)w_{q,p+1}(t_k,t)\mathcal{A} = \frac{q!\gamma_{q,p+1}(t,t)h^{p-q+1}}{(p+1)!}B_{p+1}(K_{q,p+1})(1+o_p(1)),$$

where $\mathcal{A} = \sum_{\alpha=0}^{p}\gamma_{\alpha,p+1}(t,t)\frac{(t_j-t)^\alpha}{\alpha!}\frac{(t_k-t)^{p+1}}{(p+1)!}$ and

$$\sum_{j=1}^{T}\sum_{k=1}^{T} w_{q,p+1}(t_j,t)w_{q,p+1}(t_k,t)\mathcal{B} = \frac{q!\gamma_{p+1,q}(t,t)h^{p-q+1}}{(p+1)!}B_{p+1}(K_{q,p+1})(1+o_p(1)),$$

where $\mathcal{B} = \sum_{\beta=0}^{p}\gamma_{p+1,\beta}(t,t)\frac{(t_j-t)^{p+1}}{(p+1)!}\frac{(t_k-t)^\beta}{\beta!}$. Since $\gamma_0(s,t) = \gamma_0(t,s)$, we have $\gamma_{q,p+1}(t,t) = \gamma_{p+1,q}(t,t)$. The assertion then follows.

B Proofs

Proof of Theorem 6.1 Suppose the conditions imposed for this theorem hold. By (3.2), Lemmas A.1 and A.2, and the Taylor expansion, we have

$$\begin{aligned}
\mathrm{E}(\widehat{\beta_r^{(q)}}(t)|\mathcal{D}) &= \sum_{j=1}^{T} w_{q,p+1}(t_j,t)\beta_r(t_j) \\
&= \sum_{j=1}^{T} w_{q,p+1}(t_j,t)\{\sum_{k=0}^{p+1}\beta_r^{(k)}(t)\frac{(t_j-t)^k}{k!} + o((t_j-t)^{p+1})\} \\
&= \beta_r^{(q)}(t) + \frac{q!\beta_r^{(p+1)}(t)h^{p-q+1}}{(p+1)!}B_{p+1}(K_{q,p+1})(1+o_p(1)).
\end{aligned}$$

Theorem 6.1 follows.

Proof of Theorem 6.2 By the assumptions and (3.3), we have

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta_r^{(q)}}(t)|\mathcal{D}) &= \sum_{j=1}^{T}\sum_{k=1}^{T} w(t_j,t)w(t_k,t)\mathrm{cov}(b_r(t_j),b_r(t_k)|\mathcal{D}) \\
&= \omega^{rr}\sum_{j=1}^{T}\sum_{k=1}^{T} w_{q,p+1}(t_j,t)w_{q,p+1}(t_k,t)\gamma(t_j,t_k)n_{jk}/(n_jn_k)(1+o_p(1)) \\
&= \omega^{rr}/(cn)\sum_{j=1}^{T} w_{q,p+1}^2(t_j,t)\gamma(t_j,t_j)(1+o_p(1)) \\
&+ o\{1/(nTh^{2q+1})\}\omega^{rr}\{\sum_{j=1}^{T}\sum_{k=1}^{T} w_{q,p+1}(t_j,t)w_{q,p+1}(t_k,t)\gamma(t_j,t_k)\} \\
&= \frac{\omega^{rr}q!^2\gamma(t,t)}{cnTh^{2q+1}f(t)}V(K_{q,p+1})(1+o_p(1)).
\end{aligned}$$

The last equality follows from Lemmas A.2 and A.3. Theorem 6.2 follows.

Proof of Theorem 6.3 Suppose the conditions given for this theorem hold. Then we have

$$\mathrm{Var}(\widehat{\beta_r^{(q)}}(t)|\mathcal{D}) = \omega^{rr}\{1/n + o(1/n)\}\sum_{j=1}^{T}\sum_{k=1}^{T} w_{q,p+1}(t_j,t)w_{q,p+1}(t_k,t)\gamma(t_j,t_k)(1+o_p(1)).$$

Theorem 6.3 then follows from Lemma A.3.

Acknowledgments. We are very grateful to the Joint Editor Dr. Chris Jones and two referees for helpful comments and suggestions which made it possible for our manuscript to be improved substantially. We also owe a lot to Professors W. Lasley and B. Brumback for making the hormone data available to us and to Professors Colin O. Wu and Donald Hoover and their project supported by the National Institute on Drug Abuse grant R01 DA10184-01 for providing us MACS Public Use Data Set Release PO4 (1984-1991). Thanks also go to Dr. Colin O. Wu for his helpful comments which have greatly improved the presentation of this paper. Fan's research was partially supported by Grant DMS-9504414 and NSA Grant 96-1-0015.

References

Brumback, B. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J. Amer. Statist. Assoc.,* **93**, 961-994.

Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991). Local regression models. In *Statistical Models in S* (Chambers, J. M. and Hastie, T. J., eds), 309–376. Wadsworth & Brooks, Pacific Grove.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford University Press, Oxford, England.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.,* **87**, 998-1004.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and its Applications.* Chapman and Hall, London.

Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators. *J. Computat. Graph. Statist.*, **3**, 35-56.

Fan, J. and Zhang, J. T. (1998). Comments on "Smoothing spline models for the analysis of nested and crossed samples of curves" by Brumback and Rice. *J. Amer. Statist. Assoc.,* **93**, 980-983.

Fan, J. and Zhang, W. (1997). Statistical estimation in varying-coefficient models. Manuscript.

Gasser, T. and Müller, H. G. (1979). Kernel estimation of regression functions. In: *Smoothing Techniques for Curve Estimations*, eds. Gasser and Rosenblatt. Springer-Verlag, Heidelberg.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

Hand, D. and Crowder, M. (1996). *Practical Longitudinal Data Analysis*, Chapman and Hall, London.

Hart, J. D. and Wehrly, T. E. (1986). Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.* **81**, 1080-88.

Hart, J. D. and Wehrly, T. E. (1993). Consistency of cross-validation when the data are curves. *Stochastic Process and their Applications*, **45**, 351-361.

Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. B* **55**, 757-96.

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*, to appear.

Jones, R. M. (1993). *Longitudinal Data with Serial Correlation: a state-space approach.* Chapman and Hall, London.

Kaslow, R. A., Ostraw, D. G., Detels, R. *et al* (1987). The multicenter AIDS Cohort Study: rationale, organization and selected characteristics of the participants. *American Journal of Epidemiology*, **126**, 310-18.

Lange, N., Carlin, B. P. and Gelfand, A. E. (1992). Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers. *J. Amer. Statist. Assoc.,* **87**, 615-632.

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

Lindsey, J. K. (1993). *Models for Repeated Measurements.* Oxford University Press, Oxford, England.

Moyeed, R. A. and Diggle, P. J.(1994). Rates of convergence in semi-parametric modeling of longitudinal data. *Austr. J. Statist.* **36**, 75-93.

Munro, C., Stabenfeldt, G., Cragun, J., Addiego, L., Overstreet, J., and Lasley, B. (1991). Relationship of serum estradiol and progesterone concentrations to the excretion profiles of their major urinary methabolites as measured by enzyme immunoassay and radioimmunoassay. *Clinical Chemistry*, **37**, 838-844.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*, Springer-Verlag, Berlin.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when data are curves. *J. Roy. Statist. Soc. B*, **53**, 233-43.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.,* **90**, 1257-1270.

Ruppert, D. and Wand, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist.,* **22**, 1346-1370.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59.

Wand, M.P. and Jones, M.C.(1995). *Kernel Smoothing.* Chapman and Hall, London.

Wu, C. O. and Chiang, C. T. (1998). Kernel smoothing on varying coefficient models with longitudinal dependent variable. Manuscript.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689-99.

Zhang, J. T. (1999). *Smoothed Functional Data Analysis.* Dissertation, UNC-Chapel Hill.