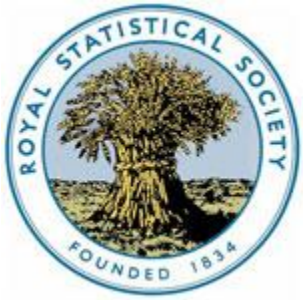

J. R. Statist. Soc. B (2000)
62, *Part* 2, *pp.* 303–322

Two-step estimation of functional linear models with applications to longitudinal data

Jianqing Fan and Jin-Ting Zhang

University of North Carolina, Chapel Hill, USA

[Received March 1998. Final revision September 1999]

Summary. 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 approach proposed. Simulation studies show that our two-step approach improves the kernel method proposed by Hoover and co-workers in several aspects such as accuracy, computational time and visual appeal of the estimators.

Keywords: Functional analysis of variance; Functional linear models; Local polynomial smoothing; Longitudinal data analysis

1. Introduction

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

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

where $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijd})^{\mathrm{T}}$ 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 cell count data set, the association is depicted in Fig. 1 in Section 4. To obtain such an association, some modelling 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 linear model

Address for correspondence: Jianqing Fan, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, USA.
E-mail: jfan@stat.unc.edu

 1369–7412/00/62303

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

where $\epsilon(t)$ is a zero-mean correlated stochastic process that cannot 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):

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

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

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) and Hand and Crowder (1996) and references therein) for longitudinal data by allowing coefficients to change over time. Although the traditional linear models provide useful tools for analysing longitudinal data, problems of 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 (1999) proposed 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), smoothing spline and kernel methods were studied whereas in Brumback and Rice (1998) the smoothing spline method was considered for functional analysis-of-variance (ANOVA) models which are special cases of functional linear models.

Although 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 must blindly invert a matrix of size 2000×2000, which takes a huge amount of central processor unit time and requires a large amount of random-access memory. 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 size of the matrix is approximately $nT \times nT$). This problem cannot easily be rescued by the back-fitting 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 neighbourhood and hence only part of the data is 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 quite 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) modified 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. Speaking simply, 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 by 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 cell count data in Section 4), although for a particular subject the measurements at some time points may be missing. Let t_j, $j = 1, \ldots, T$, 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 to obtain the raw estimates $b(t_j) = (b_1(t_j), \ldots, b_d(t_j))^{\mathrm{T}}$ for $\beta(t_j) = (\beta_1(t_j), \ldots, \beta_d(t_j))^{\mathrm{T}}$. This is the first step. Since the raw estimates are usually not smooth (see the examples given in Section 4), we must 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 $\{(t_j, b_r(t_j)), j = 1, 2, \ldots, T\}$. 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 neighbouring time points to improve the efficiency of the raw estimates. An extra benefit of our two-step procedure is that the smoothing step is one dimensional. This leads to several advantages. Firstly, for different components of the coefficient functions, different amounts of smoothing can be conducted. Secondly, a visualization of the raw estimates can assist us in choosing a sensible amount of smoothing. Thirdly, the smoothing step can be conducted with any existing smoothing technique. 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 with 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 the 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; 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 to use 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 approach proposed is applied to two longitudinal data sets, one of which 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 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 into our methodology when the sample size is large. Technical proofs are given in Appendix A.

2. Raw estimates

Let $\{t_j,\ j = 1, 2, \ldots, T\}$ be the distinct time points among $\{t_{ij},\ j = 1, 2, \ldots, T_i,\ i = 1, 2, \ldots, n\}$. 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

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

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

Assume that $\operatorname{rank}(\tilde{\mathbf{X}}_j) = d$ (see remark 1 for a discussion of the case $\operatorname{rank}(\tilde{\mathbf{X}}_j) < d$). Then standard least squares theory shows that $b(t_j) = (\tilde{\mathbf{X}}_j^{\mathrm{T}} \tilde{\mathbf{X}}_j)^{-1} \tilde{\mathbf{X}}_j^{\mathrm{T}} \tilde{\mathbf{Y}}_j$ is an estimator of $\beta(t_j)$ with

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

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

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

and

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

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

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

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 $P_j = \tilde{\mathbf{X}}_j(\tilde{\mathbf{X}}_j^{\mathrm{T}}\tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^{\mathrm{T}}$. It follows that

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

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

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

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

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

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

Remark 1. If $\mathrm{rank}(\tilde{\mathbf{X}}_j) < d$, we cannot obtain 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 raw estimates that are not missing. The second method is to increase the size of the neighbourhood. For instance, we can use all observations at time points t_{j-1}, t_j and t_{j+1} to fit model (1.2) with $t = t_j$. The third method is to impute some of the missing observations by obtaining information from the neighbouring time points. For example, we can use observations at time points t_{j-1} and t_{j+1} to impute the observations at t_j. 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 largely 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, because of their special structures, the functional ANOVA models should be handled with special care.

2.1. *Nested functional analysis of variance*

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 levels of

nesting. The motivation for our study comes from an analysis of the progesterone curves measured over 21 conceptive and 70 non-conceptive women's menstrual cycles (top level nesting, namely group effects). A woman in the non-conceptive group can have as many as five 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 $k = 1, 2, \ldots, K_{ij}$ (the number of cycles of subject j in group i), $j = 1, 2, \ldots, J_i$ and $i = 1, 2, \ldots, I$. 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

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

Model (2.6) is a special case of model (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 $\alpha_i(t_l)$ $(i = 1, 2, \ldots, I)$ are given by

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

$$\operatorname{var}\{\hat{\alpha}_i(t_l)\} = \gamma(t_l, t_l) \Big/ \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} \delta_{ijkl},$$

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

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

$$\operatorname{var}\{\hat{\alpha}_i(t_l) + \hat{\beta}_{ij}(t_l)\} = \gamma(t_l, t_l) \Big/ \sum_{k=1}^{K_{ij}} \delta_{ijkl},$$

if $\Sigma_{k=1}^{K_{ij}} \delta_{ijkl} > 0$; 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 raw estimates that are not missing 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 $\operatorname{var}\{\hat{\alpha}_i(t_l)\}$ and the average of the variances of those estimates $\hat{\alpha}_i(t_l) + \hat{\beta}_{ij}(t_l)$ that are not missing. These ideas can also be employed to impute the missing observations.

2.2. *Crossed functional analysis of variance*

We discuss only a two-way crossed functional ANOVA model. Multiple-way crossed functional ANOVA models can be dealt with similarly. A two-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 $i = 1, 2, \ldots, I_b$, $j = 1, 2, \ldots, J_\tau$. 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 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:

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

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

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

$$\operatorname{var}\{\hat{\mu}(t_l)\} = \gamma(t_l, t_l) \Big/ \sum_{i=1}^{I_b} \sum_{j=1}^{J_\tau} \delta_{ijl},$$

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

3. Refining the raw estimates

There are several reasons for us to refine the raw estimates obtained in the last section. Firstly, the raw estimates are generally not smooth. Secondly, they are inefficient since they have not used the information from the neighbouring time points and hence their efficiency can be improved. Thirdly, there may be some missing raw estimates because of insufficient 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 non-design 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 $\{(t_j, b_r(t_j)), j = 1, 2, \ldots, T\}$ 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 that $\beta_r(t)$ is $(p + 1)$-times continuously differentiable and we wish to estimate its qth derivative for some $0 \leqslant q < p + 1$. 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

$$E\{\widehat{\beta_r^{(q)}}(t)|\mathcal{D}\} = \sum_{j=1}^{T} w_r(t_j, t)\,\beta_r(t_j), \tag{3.2}$$

$$\text{var}\{\widehat{\beta_r^{(q)}}(t)|\mathcal{D}\} = \sum_{j=1}^{T}\sum_{k=1}^{T} w_r(t_j, t)\,w_r(t_k, t)\,\text{cov}\{b_r(t_j), b_r(t_k)|\mathcal{D}\}. \tag{3.3}$$

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

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

which is also called a 95% pointwise confidence interval by some 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 $C_j = (1, t_j - t, \ldots, (t_j - t)^p)^{\mathrm{T}}, j = 1, 2, \ldots, T$, and $K_h(t) = K(t/h)/h$ be a kernel function with a bandwidth h. Then

$$w_{q,p+1}(t_j, t) = q!\, e_{q+1,p+1}^{\mathrm{T}}(C^{\mathrm{T}}WC)^{-1}C_jW_j, \qquad j = 1, 2, \ldots, T, \tag{3.5}$$

are the local polynomial weights for estimating the qth derivative of an underlying function where $C = (C_1, C_2, \ldots, C_T)^{\mathrm{T}}$ and $W = \text{diag}(W_1, \ldots, W_T)$ with $W_j = K_h(t_j - t)$. In particular, the local linear weights are given by $w_{0,2}(t_j, t), j = 1, 2, \ldots, T$. 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 for $\text{var}\{b_r(t_j)\}$ in equation (2.4), we have $a^2(t_j) = e_{r,d}^{\mathrm{T}}(\tilde{\mathbf{X}}_j^{\mathrm{T}}\tilde{\mathbf{X}}_j)^{-1}e_{r,d}$ and $\sigma^2(t_j) = \gamma(t_j, t_j)$. Thus, the data $\{(t_j, b_r(t_j)), j = 1, 2, \ldots, T\}$ are heteroscedastic. Note that $\sigma^2(t)$ may vary slowly if we assume that it is smooth. However, $a^2(t)$ may change dramatically owing to different numbers of data points observed at different times. This knowledge can be incorporated in the construction of the local polynomial weights $w_{q,p+1}(t_j, t), j = 1, 2, \ldots, T$, 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 obtained similarly.

4. Applications to longitudinal data

4.1. CD4 cell percentage in human immunodeficiency virus seroconverters

The human immunodeficiency virus (HIV) destroys CD4 cells (T-lymphocytes, a vital component of the immune system) so the number or percentage of the CD4 cells in the blood of a human body will change after the human subject has been infected with HIV. Thus the CD4 cell level marks the progression of disease 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 whereas 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 takes binary values 1 or 0, according to whether a subject is a smoker or a non-smoker. 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

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

where $Y(t)$ is the percentage of CD4 cells at time t. In the data, the time point t_{ij} indicates the time (in years) when the ith subject paid his jth visit after HIV infection. All subjects were scheduled to pay their visits twice a year but the definite 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 stage of 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 0. As a result, the intercept function $\beta_0(t)$ can be interpreted as the base-line CD4 percentage curve for a non-smoker 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.

Fig. 1 depicts the fitted coefficient functions (full curves) with ±2 pointwise standard error bands (broken curves). The circles indicate the raw estimates of the coefficient functions at

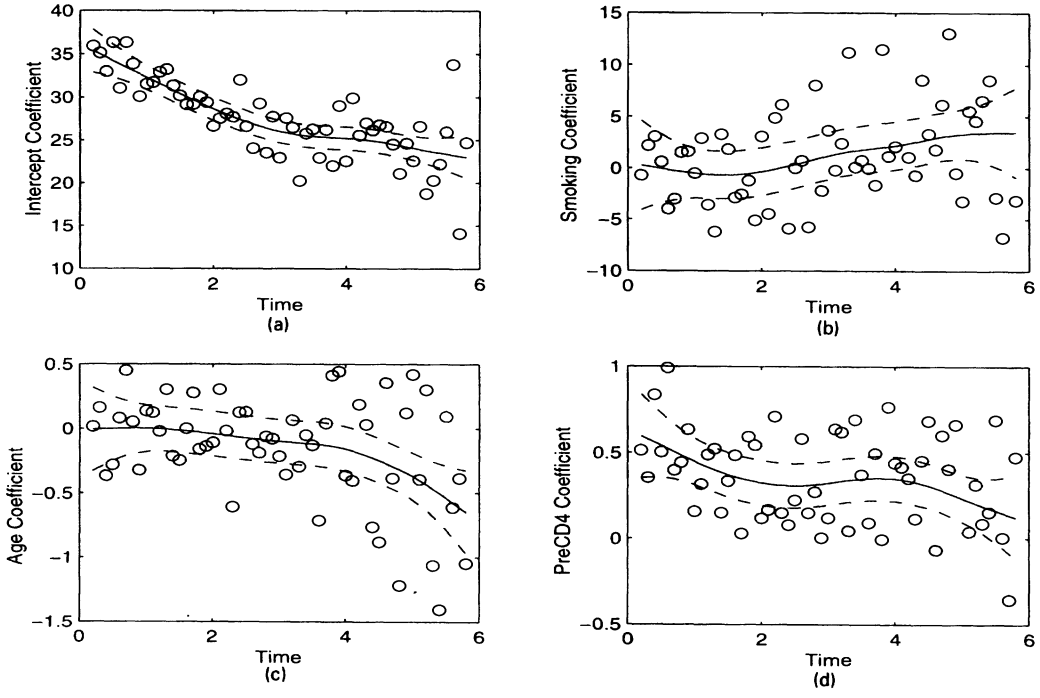

Fig. 1. Estimated coefficient curves for the base-line CD4 cell percentage and the effects of covariates (a) intercept, (b) Smoking, (c) Age and (d) PreCD4 on the percentage of CD4 cells: ———, smoothed effects; – – –, ±2 pointwise standard error bands; ○, raw estimates

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 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 cell coefficient functions are about the same, indicating that they admit a similar amount of smoothness.

The fitted intercept function (base-line CD4 cell percentage curve) is displayed in Fig. 1(a). It has a quick drop during the first 3 years and a slower drop afterwards. The fitted smoking coefficient function is displayed in Fig. 1(b). It seems that $\beta_1(t) \geqslant 0$ for most of the time. This may suggest that the smoking population has a higher CD4 cell percentage if we hold the 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 Fig. 1(c). The estimated standard error bands suggest that the age effect is probably near zero within the first 4 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 non-conceptive womens' menstrual cycles. A woman in the non-conceptive group can be measured up to five menstrual cycles whereas 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. For various reasons, not all measurements in a menstrual cycle are available, and this results in some missing responses in some cycles. This 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 Section 4.1, 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 equations (2.3) and (2.4)), the dependence of the raw estimates has a limited effect on the selection of the bandwidth.

Figs 2(a) and 2(b) depict the fitted coefficient curves of the non-conceptive and conceptive group effects (full curves) and ± 2 pointwise standard error bands (broken curves). Their raw estimates are indicated by the circles, which clearly show the shapes of the underlying group effect curves. Whereas these two group effect curves progress similarly for 8 days before and after the day of the ovulation, they show different tendencies from the eighth day after ovulation: the progesterone curve for the non-conceptive group decreases rapidly whereas the progesterone curve for 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 Fig. 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. Fig. 2(d) presents the

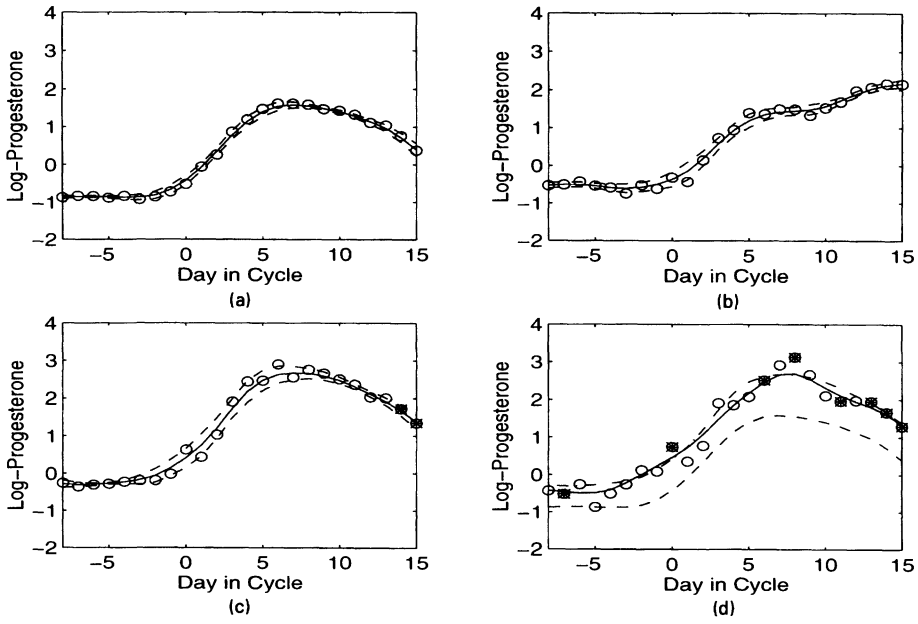

Fig. 2. Estimated coefficient curves for the progesterone data: (a) non-conceptive effect; (b) conceptive effect; (c) non-conceptive plus subject 11 effect (———, smoothed effects; – – –, ±2 standard error bands; ○, raw estimates; ∗, imputed raw estimates); (d) non-conceptive plus subject 11 plus cycle 1 effect (———, smoothed effect for cycle 1; – – –, smoothed non-conceptive effects; · — · —, smoothed non-conceptive plus subject 11 effects; ○, raw estimates; ∗, imputed raw estimates)

smoothed effect (full 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 Fig. 2(d). The non-conceptive effect curve and the non-conceptive 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 because of the intensive computation of the spline approach, not to mention the difficulty in choosing multiple smoothing parameters.

The leave-one-out cross-validation method is used to select the bandwith 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 cell data set. The covariates of the CD4 cell data set are kept fixed and the true coefficient curves are taken as the full curves presented in Fig. 1. Following Wu and Chiang (1998), we shall sample the errors ϵ_{ij} from the Gaussian process with zero mean and covariance function

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

This is a decaying exponential stationary covariance function, indicating that the correlation will be decreasing with time lag. The variance factor 16 is chosen differently from the 0.0625 given by Wu and Chiang (1998) since the standard deviation of the CD4 cell 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 cell data. For each subject, about 12 time points are randomly selected from the set $\{t_j = 0.1j,\ j = 1, \ldots, 60\}$ to make the simulated data sufficiently similar to the original CD4 cell data. The observed data are then the sum of the errors and the expected values at various time points, i.e.

$$Y_{ij} = X_{ij}^{\mathrm{T}} \beta(t_{ij}) + \epsilon_{ij}, \qquad j = 1, 2, \ldots, T_i, \quad i = 1, 2, \ldots, n,$$

with β being the fitted coefficient functions presented in Fig. 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

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

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

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

or its unweighted average squared error UASE, defined similarly to WASE but with no weights in equation (5.1).

The box plots of the MADE, WASE and UASE ratios two-step/kernel are presented in panels 1, 2 and 3 of Fig. 3(a) respectively. It seems that both methods perform quite 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 1/30–1/50 of that for the kernel method.

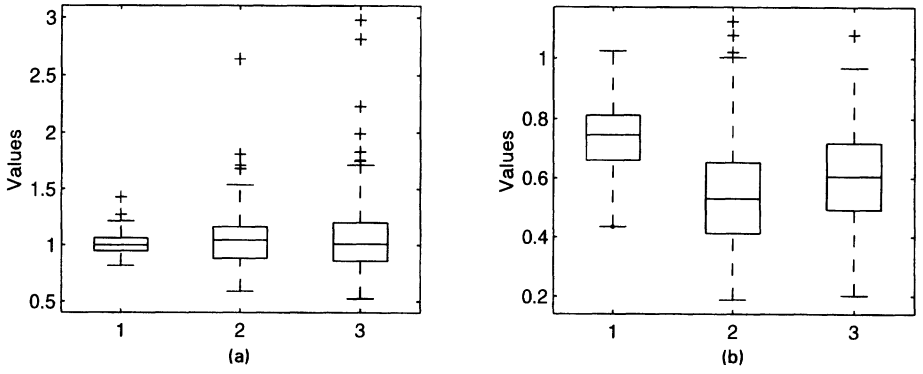

Fig. 3. Comparison of the two-step method with the kernel method: (a) box plots for the ratios two-step/kernel of MADEs (panel 1), WASEs (panel 2) and UASEs (panel 3) for model 1; (b) as for (a) but for model 2

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. An examination of the resulting plots (which are omitted here for brevity) reveals that the kernel method generally undersmooths some or all of the true coefficient curves. This 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 cell coefficient functions are not sufficiently challenging 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

$$\begin{aligned}\beta_0(t) &= 15 + 8.7\ \sin(2\pi t),\\ \beta_1(t) &= 4 - 17(t - \tfrac{1}{2})^2,\\ \beta_2(t) &= 1 + 11.2t,\\ \beta_3(t) &= 1 + 2t^2 + 11.3(1-t)^3.\end{aligned}$$

They represent four different types of curve. The four covariates are chosen as follows. First, we let $X_0(t) \equiv 1$. We then let $X_1(t)$ be a binomial random variable with probability of success $p = 0.6$ 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 0 and conditional variance

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

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

$$\operatorname{cov}(\epsilon_{i_1 j_1}, \epsilon_{i_2 j_2}) = \begin{cases} 5.27\ \exp(-0.5|t_{i_1 j_1} - t_{i_2 j_2}|), & \text{if } i_1 = i_2,\\ 0, & \text{if } i_1 \neq i_2.\end{cases}$$

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 equispaced over [0, 1]. For each subject, we let 60% of the 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 in our first model simulation, we sampled 201 data sets from this model, calculated their MADEs, WASEs and UASEs for both the two-step and the kernel methods, and then presented their ratio box plots in Fig. 3(b). We can see that the two-step method has a much better performance using all three accuracy measures. An 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 1/30–1/50 of that for the kernel method.

6. Asymptotic results

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

$$\epsilon_i(t) = \upsilon_i(t) + e_i(t). \tag{6.1}$$

This formulation is a generalization of that in section 5.6 of Diggle *et al.* (1994). The trajectory process $\{\upsilon_i(t)\}$ 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 $\{\epsilon_i(t)\}$ is

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

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

A local polynomial fitting technique is used for smoothing the raw estimates because of their good sampling properties (Fan and Gijbels, 1996). To obtain some further insight into 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 qth derivative of $\beta_r(t)$ based on the raw estimates $b_r(t_j)$, $j = 1, 2, \ldots, T$, is

$$\widehat{\beta_r^{(q)}}(t) = \sum_{j=1}^{T} w_{q,p+1}(t_j, t)\, b_r(t_1), \qquad q = 0, 1, 2, \ldots, p, \tag{6.2}$$

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

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

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

We first derive the asymptotic bias. Since $E\{b_r(t_j)|\mathcal{D}\} = \beta_r(t_j)$, $j = 1, 2, \ldots, T$, the correlation within subjects does not 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 Appendix A.

Theorem 1. Suppose that condition (a) in Appendix A holds. Then, when $h \to 0$ and $Th \to \infty$ as $T \to \infty$,

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

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

It is more involved to derive the asymptotic variance of estimator (6.2). The main difficulty is that the variance–covariance structure of the raw estimates $b_r(t_j)$, $j = 1, 2, \ldots, T$, 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 $\Omega_l = E(\mathbf{X}_{1l}\mathbf{X}_{1l}^{\mathrm{T}})$, $l = j, k$, and $\Omega_{jk} = E(\mathbf{X}_{1j}\mathbf{X}_{1k}^{\mathrm{T}})$ for all $j, k = 1, 2, \ldots, T$. Then by the law of large numbers and condition (f) in Appendix A we deduce from equation (2.3) that

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

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

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

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

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

where $\omega^{rr} = e_{r,d}^{\mathrm{T}} \Omega_1^{-1} e_{r,d}$, 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 $I_t = \{j\colon |t_j - t| \leqslant h\}$ be the indices of the local neighbourhood. 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 2. Under conditions (a)–(c), (e) and (g) in Appendix A, 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 $nT \to \infty$,

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

where $V(K) = \int K^2(u)\, \mathrm{d}u$.

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

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

Theorem 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 $j,\, k = 1, 2, \ldots, n$. Let $\gamma_{\alpha,\beta}(s,\, t)$ denote $\partial^{\alpha+\beta}\gamma_0(s,\, t)/\partial s^\alpha \partial t^\beta$ for any integers $\alpha,\, \beta = 0, 1, \ldots, p+1$.

Theorem 3. Suppose that conditions (a)–(e) and (g) in Appendix A hold. Assume that $n_{jk}/n_j n_k = 1/n + o(1/n)$ holds uniformly for all $j,\, k = 1, 2, \ldots, T$. Then when $h \to 0$ and $nTh^{2q+1} \to \infty$ as $n,\, T \to \infty$,

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

When the underlying process $v(t)$ defined in equation (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, $\gamma_{q,q}(t,\,t) = 0$ and $\gamma_{q,p+1}(t,\,t) = 0$ for all $q = 1,\,2,\,\ldots,\,p$. 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 1. Under the conditions of theorem 3, if the trajectory process $v(t)$ is stationary, then for all $q = 1,\,2,\,\ldots,\,p$ we have

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

and

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

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 distinct time points T, such an effect is very small and can be ignored.

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 note that for the raw estimates of the functional ANOVA models the corresponding $\omega^{rr} = 1$.

Acknowledgements

We are very grateful to the Joint Editor and two referees for helpful comments and suggestions which made it possible for our manuscript to be improved substantially. We also owe much to Professor W. Lasley and Professor B. Brumback for making the hormone data available to us and to Professor Colin O. Wu and Professor Donald Hoover and their project supported by the National Institute on Drug Abuse grant R01 DA10184-01 for providing us Multi-Center AIDS Cohort Study public use data set release PO4 (1984–1991). Thanks also go to Professor Colin O. Wu for his helpful comments which have greatly improved the presentation of this paper. Fan's research was partially supported by National Science Foundation grant DMS-9504414 and National Security Agency grant 96-1-0015.

Appendix A

A.1. Preliminaries

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

(a) The time points $t_1, t_2, \ldots, t_T$ 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.
(b) The noise variance $\sigma^2(t)$ is continuous in the support of f.
(c) The coefficient function $\beta_r(t)$ is $p+1$ times continuously differentiable for some p.
(d) The covariance function $\gamma_0(s, t)$ of the underlying trajectory process $v(t)$ (see equation (6.1)) is $p+1$ times continuously differentiable for both s and t for some p.
(e) The kernel function K is a bounded symmetric probability density function with a bounded support $[-1, 1]$, say.
(f) For a fixed $j \in \{1, 2, \ldots, T\}$, the covariates $\mathbf{X}_{ij}$, $i = 1, 2, \ldots, n$, are independently and identically distributed as $\mathbf{X}_{1j} = (X_{1j1}, \ldots, X_{1jd})^T$ with $\Omega_j = E(\mathbf{X}_{1j}\mathbf{X}_{1j}^T)$ positive definite.
(g) The covariates $\mathbf{X}_{ij}$ satisfy condition (f) and they are time invariant, i.e. $\mathbf{X}_{ij} = \mathbf{X}_{i1}$ for all $j = 1, 2, \ldots, T$.

Conditions (a)–(e) 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 (f) says that, for a fixed time point, the covariates for different subjects are independently and identically distributed. Condition (g) holds for many longitudinal data sets. One of the data sets presented in Section 4 is a typical example.

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

Lemma 1. Suppose that conditions (a) and (e) hold. If $h \to 0$ and $Th \to \infty$ as $T \to \infty$, then

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

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

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

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

Moreover, by lemma 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 1 and 2 respectively.

Lemma 3. Suppose that conditions (a), (b), (d) and (e) hold. If $h \to 0$ and $Th \to \infty$ as $T \to \infty$, then

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

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

Proof. Clearly,

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

By lemma 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 (d), 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 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 that $\gamma_{q,p+1}(t,\, t) = \gamma_{p+1,\,q}(t,\, t)$. The assertion then follows.

A.2. Proofs

A.2.1. *Proof of theorem 1*

Suppose that the conditions imposed for theorem 1 hold. By equation (3.2), lemmas 1 and 2, and Taylor expansion, we have

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

Theorem 1 follows.

A.2.2. *Proof of theorem 2*

By the assumptions and equation (3.3), we have

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

The last equality follows from lemmas 2 and 3. Theorem 2 follows.

A.2.3. Proof of theorem 3

Suppose that the conditions given for theorem 3 hold. Then we have

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

Theorem 3 then follows from lemma 3.

References

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

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

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

Fan, J. (1992) Design-adaptive nonparametric regression. *J. Am. Statist. Ass.*, **87**, 998–1004.

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

Fan, J. and Marron, J. S. (1994) Fast implementations of nonparametric curve estimators. *J. Comput. 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 B. Brumback and J. A. Rice). *J. Am. Statist. Ass.*, **93**, 980–983.

Fan, J. and Zhang, W. (1999) Statistical estimation in varying-coefficient models. *Ann. Statist.*, to be published.

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

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

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

Hart, J. D. and Wehrly, T. E. (1986) Kernel regression estimation using repeated measurements data. *J. Am. Statist. Ass.*, **81**, 1080–1088.

——— (1993) Consistency of cross-validation when the data are curves. *Stoch. Process. Applic.*, **45**, 351–361.

Hastie, T. and Tibshirani, R. (1993) Varying-coefficient models (with discussion). *J. R. Statist. Soc.* B, **55**, 757–796.

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.

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

Kaslow, R. A., Ostraw, 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. *Am. J. Epidem.*, **126**, 310–318.

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. Am. Statist. Ass.*, **87**, 615–632.

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

Moyeed, R. A. and Diggle, P. J. (1994) Rates of covergence in semi-parametric modeling of longitudinal data. *Aust. 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 metabolites as measured by enzyme immunoassay and radioimmunoassay. *Clin. Chem.*, **37**, 838–844.

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

Rice, J. A. and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B, **53**, 233–243.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995) An effective bandwidth selector for local least squares regression. *J. Am. Statist. Ass.*, **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. *Regl Conf. Ser. Appl. Math.*, **59**.

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

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

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.

Zhang, J. T. (1999) Smoothed functional data analysis. *Dissertation*. University of North Carolina, Chapel Hill.