Biometrika (2002), **89**, 1, *pp.* 111–128

Printed in Great Britain

Varying-coefficient models and basis function approximations for the analysis of repeated measurements

By JIANHUA Z. HUANG

Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

jianhua@wharton.upenn.edu

COLIN O. WU

Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, Maryland 21218, U.S.A.

colin@mts.jhu.edu

and LAN ZHOU

Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

lzhou@cceb.upenn.edu

Summary

A global smoothing procedure is developed using basis function approximations for estimating the parameters of a varying-coefficient model with repeated measurements. Inference procedures based on a resampling subject bootstrap are proposed to construct confidence regions and to perform hypothesis testing. Conditional biases and variances of our estimators and their asymptotic consistency are developed explicitly. Finite sample properties of our procedures are investigated through a simulation study. Application of the proposed approach is demonstrated through an example in epidemiology. In contrast to the existing methods, this approach applies whether or not the covariates are time-invariant and does not require binning of the data when observations are sparse at distinct observation times.

Some key words: Basis function; Confidence band; Hypothesis testing; Least squares; Longitudinal data; Polynomial spline; Resampling subject bootstrap; Varying-coefficient model.

1. Introduction

In biomedical studies, subjects are often measured repeatedly over a given time period, so that the measurements within each subject are possibly correlated with each other but different subjects can be assumed to be independent. Data of this type are frequently referred to as longitudinal samples. Statistical analyses with longitudinal data have been a subject of intense investigation in the past two decades, using parametric models (Diggle et al., 1994; Davidian & Giltinan, 1995; Vonesh & Chinchilli, 1997) or a nonparametric or semiparametric approach (Hart & Wehrly, 1986; Müller, 1988; Rice & Silverman, 1991; Zeger & Diggle, 1994; Moyeed & Diggle, 1994; Lin & Carroll, 2000; Rice & Wu, 2001).

Let $\{Y(t), t \geqslant 0\}$ be a real-valued response process and let

$$\{X(t) = (X^{(0)}(t), \ldots, X^{(k)}(t))^{\mathrm{T}}, t \geqslant 0\}$$

be an R^{k+1}-valued covariate process for a group of independent subjects. The realisations of these processes for each subject are often obtained at a set of distinct and possibly irregularly spaced time points in a fixed interval $\mathcal{T}$ of finite length. Interest is often focused on evaluating the mean effects of t and $X(t)$ on the response $Y(t)$. A longitudinal sample from n randomly selected subjects is denoted by

$$\{(Y_{ij}, X_i(t_{ij}), t_{ij});\ i = 1, \ldots, n, j = 1, \ldots, n_i\},$$

where t_{ij} is the time when the jth measurement of the ith subject is made, n_i is the number of repeated mesaurements of the ith subject, and $Y_{ij} \equiv Y_i(t_{ij})$ and $X_i(t_{ij}) = (X_i^{(0)}(t_{ij}), \ldots, X_i^{(k)}(t_{ij}))^{\mathrm{T}}$ are the ith subject's observed outcome and covariates at t_{ij}. The total number of observations in the sample is $N = \sum_{i=1}^{n} n_i$. As in regression models where a baseline effect is desired, we set $X^{(0)}(t) \equiv 1$.

A useful nonparametric model for such a longitudinal sample is the varying-coefficient model

$$Y_{ij} = X_i^{\mathrm{T}}(t_{ij})\beta(t_{ij}) + \varepsilon_i(t_{ij}), \qquad (1{\cdot}1)$$

where, for all $t \geqslant 0$, $\beta(t) = (\beta_0(t), \ldots, \beta_k(t))^{\mathrm{T}}$ are smooth functions of t, $\varepsilon_i(t)$ is a realisation of a zero-mean stochastic process $\varepsilon(t)$, and $X_i(t_{ij})$ and ε_i are independent. Model (1·1) is a specific model of a class of functional linear models introduced by Ramsay & Silverman (1997, Ch. 9). This model contains as a special case the partially linear model studied by Zeger & Diggle (1994) and Moyeed & Diggle (1994), where only the intercept coefficient β_0 is allowed to be time-varying and $\beta_1, \ldots, \beta_k$ are all constants. It is also closely related to the varying-coefficient model for independent identically distributed data studied by Hastie & Tibshirani (1993).

To fit model (1·1), Hoover et al. (1998) used a class of smoothing splines and local polynomial estimators for the estimation of $\beta(t)$. However, because only one smoothing parameter is used, their local polynomial estimators may not be able to provide adequate smoothing for all the coefficient curves at the same time; also, their smoothing splines are extremely computationally intensive when there is a large number of distinct time points. To introduce different amounts of smoothing for different coefficient curves, Fan & Zhang (2000) suggested a two-step estimation method and Wu & Chiang (2000) and Chiang et al. (2001) proposed component-based kernel and smoothing spline estimators. When the observations at some distinct time points are sparse, Fan & Zhang (2000) suggested computing the raw estimates used in their two-step estimators by binning the data from adjacent time points. However, methods of bin selection and their effects have not been studied. The component-based methods, although adequate for time-invariant covariates, are not applicable when some of the covariates are time-dependent. None of these papers has investigated the testing of statistical hypotheses based on the above local smoothing methods.

Our approach, based on function approximation through basis expansions, applies to both time-invariant and time-dependent covariates when the observation times are either regularly or irregularly placed. The estimation method is a simple one-step procedure and no binning of data is needed when observations are sparse at distinct observation times. Different amounts of smoothing can be used for different individual coefficient curves. We approximate each $\beta_l(t)$ by a basis function expansion

$$\beta_l(t) \simeq \sum_{s=0}^{K_l} \gamma_{ls}^{*} B_{ls}(t),$$

where B_{ls} $(s=1, \ldots, K_l)$ is a set of basis functions, such as polynomial bases, Fourier bases or B-splines. The coefficients in the basis expansion, γ_{ls}^{*}, are estimated by least squares. The K_l play the role of smoothing parameters, and are selected using 'leave-one-subject-out' crossvalidation. Since the K_l $(l=0, \ldots, k)$ take only integer values, our crossvalidation is computationally less intensive than the crossvalidation required by Hoover et al. (1998). By taking $B_{l1}(t) \equiv 1$ to be the basis function in expanding $\beta_l(t)$ if $\beta_l(t)$ is known to be a constant, we ensure that our approach also suggests a noniterative solution to the partially linear model of Zeger & Diggle (1994) and Moyeed & Diggle (1994).

We describe our estimation procedures in § 2 and present the theoretical properties of our estimators in § 3. In § 4 we propose methods for constructing confidence bands and performing hypothesis testing using resampling subject bootstrap. Simulation results in § 5·1 illustrate finite sample performance of our estimation and inference procedures, and in § 5·2 we apply our procedures to a CD4 depletion sample from the Multicenter AIDS Cohort Study. Proofs of the main results are presented in an appendix.

2. Estimation method

2·1. *Basis approximation and least squares*

Suppose, for each $l=0, \ldots, k$, that there is a set of basis functions and constants $B_{ls}(t)$, γ_{ls}^{*}, for $s=1, \ldots, K_l$, such that

$$\beta_l(t) \simeq \sum_{s=1}^{K_l} \gamma_{ls}^{*} B_{ls}(t), \quad t \in \mathscr{T}.$$

Then we may approximate (1·1) by

$$Y_{ij} \simeq \sum_{l=0}^{k} \sum_{s=1}^{K_l} X_i^{(l)}(t_{ij}) \gamma_{ls}^{*} B_{ls}(t_{ij}) + \varepsilon_i(t_{ij}),$$

and the γ_{ls}^{*} can be estimated by minimising

$$l(\gamma) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} w_i \left\{ Y_{ij} - \sum_{l=0}^{k} \sum_{s=1}^{K_l} X_i^{(l)}(t_{ij}) B_{ls}(t_{ij}) \gamma_{ls} \right\}^2 \tag{2·1}$$

with respect to the γ_{ls}, where w_i is a nonnegative weight for the ith subject, $\gamma = (\gamma_0^{\mathrm{T}}, \ldots, \gamma_k^{\mathrm{T}})^{\mathrm{T}}$ and $\gamma_l = (\gamma_{l1}, \ldots, \gamma_{lK_l})^{\mathrm{T}}$. We require that $\sum_{i=1}^{n} n_i w_i = 1$. Assume that (2·1) can be uniquely minimised and denote its minimiser by $\hat{\gamma} = (\hat{\gamma}_0^{\mathrm{T}}, \ldots, \hat{\gamma}_k^{\mathrm{T}})^{\mathrm{T}}$, with $\hat{\gamma}_l = (\hat{\gamma}_{l1}, \ldots, \hat{\gamma}_{lK_l})$ for $l=0, \ldots, k$. Then it is natural to estimate $\beta_l(t)$ by

$$\hat{\beta}_l(t) = \sum_{s=1}^{K_l} \hat{\gamma}_{ls} B_{ls}(t),$$

to which we refer as the least squares basis estimator of $\beta_l(t)$.

To give an explicit expression for $\hat{\gamma}$ and $\hat{\beta}_l(t)$, we define

$$B(t)=\begin{pmatrix} B_{01}(t)\ldots B_{0K_0}(t) & 0\ldots 0 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots & \\ 0 \quad \ldots \quad 0 & 0\ldots 0 & B_{k1}(t) & \ldots & B_{kK_k}(t) \end{pmatrix},$$

$$U_i(t_{ij})=\{X_i^{\mathrm{T}}(t_{ij})B(t_{ij})\}^{\mathrm{T}}, \quad U_i=(U_i(t_{i1}),\ldots,U_i(t_{in_i}))^{\mathrm{T}}, \quad Y_i=(Y_{i1},\ldots,Y_{in_i})^{\mathrm{T}}.$$

Then (2·1) is equivalent to

$$l(\gamma)=\sum_{i=1}^{n}(Y_i-U_i\gamma)^{\mathrm{T}}W_i(Y_i-U_i\gamma), \tag{2·2}$$

where W_i is the $n_i\times n_i$ diagonal matrix $\mathrm{diag}(w_i,\ldots,w_i)$. Suppose that $\sum_{i=1}^{n}U_i^{\mathrm{T}}W_iU_i$ is invertible. The least squares estimator $\hat{\gamma}$ is uniquely defined by

$$\hat{\gamma}=\left(\sum_{i=1}^{n}U_i^{\mathrm{T}}W_iU_i\right)^{-1}\left(\sum_{i=1}^{n}U_i^{\mathrm{T}}W_iY_i\right). \tag{2·3}$$

With this matrix representation, our least squares basis estimator of $\beta(t)$ is

$$\hat{\beta}(t)=(\hat{\beta}_0(t),\ldots,\hat{\beta}_k(t))^{\mathrm{T}}=B(t)\hat{\gamma}. \tag{2·4}$$

Remark 1. The linear function spaces $\mathbb{G}_l$ spanned by the basis functions $\{B_{l1},\ldots,B_{lK_l}\}$ uniquely determine the basis estimators $\hat{\beta}_l$ $(0\leqslant l\leqslant k)$. Different sets of basis functions can be used to span the same space $\mathbb{G}_l$ and thus give the same estimator $\hat{\beta}_l$, although the corresponding $\hat{\gamma}$ may be different. For example, both the *B*-spline basis and the truncated power basis can be used to span a space of spline functions.

Remark 2. The choice of w_i in (2·1) may have a significant influence on the theoretical and practical properties of $\hat{\gamma}$ and $\hat{\beta}(t)$. The choice $w_i\equiv 1/N$ corresponds to an equal weight for each observation, while $w_i\equiv 1/(nn_i)$ corresponds to an equal weight for each subject. It is conceivable that an ideal choice of w_i may also depend on the intrasubject correlation structures of the data. However, because the actual correlation structures are usually completely unknown in practice, $w_i\equiv 1/N$ appears to be a practical choice if n_i $(i=1,\ldots,n)$ are relatively similar, while $w_i\equiv 1/(nn_i)$ may be appropriate otherwise. Some theoretical implications of the choices of the w_i are discussed later in Remark 4.

2·2. *Choosing a basis*

Any basis system for function approximation can be used. The Fourier basis may be desirable when the underlying functions exhibit periodicity, and polynomials are familiar choices which can provide good approximations to smooth functions. However, these bases may not be sensitive enough to exhibit certain local features without using a large K_l. In this respect, polynomial splines are often desirable. Ideally, a basis should be chosen to achieve an excellent approximation using a relatively small value of K_l. For some general guidance, see § 3.2.2 of Ramsay & Silverman (1997). We use *B*-spline bases in all our simulated and practical examples because they can exhibit local features and provide stable numerical solutions (de Boor, 1978, Ch. II).

2·3. *Selecting smoothing parameters*

In the spirit of articles such as Rice & Silverman (1991), Hart & Wehrly (1993) and Hoover et al. (1998), we choose the K_l by 'leave-one-subject-out' crossvalidation. Let $\hat{\gamma}^{(-i)}$

be the least squares estimator defined in (2·3) computed from the data with the measurements of the ith subject deleted, and let $\hat{\beta}^{(-i)}(t)$ be the estimator defined in (2·4) with $\hat{\gamma}$ replaced by $\hat{\gamma}^{(-i)}$. We define

$$\textsc{cv}(K)=\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i\{Y_{ij}-X_i^{\mathrm{T}}(t_{ij})\hat{\beta}^{(-i)}(t_{ij})\}^2 \tag{2·5}$$

to be the crossvalidation score for $K=(K_0, \ldots, K_k)$. The crossvalidated smoothing parameter K_{cv} is the minimiser of $\textsc{cv}(K)$. There are two main reasons for using this crossvalidation procedure. First, deletion of the entire measurements of the subject one at a time preserves the correlation in the data. Secondly, this approach does not require us to model the intrasubject correlation structure.

For an intuitive justification of K_{cv}, consider the average squared error

$$\textsc{ase}=\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i[X_i^{\mathrm{T}}(t_{ij})\{\beta(t_{ij})-\hat{\beta}(t_{ij})\}]^2 \tag{2·6}$$

and the decomposition

$$\begin{aligned}\textsc{cv}(K)=&\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i\{Y_{ij}-X_i^{\mathrm{T}}(t_{ij})\beta(t_{ij})\}^2\\ &+2\sum_{i=1}^{n}\sum_{j=1}^{n_i}[w_i\{Y_{ij}-X_i^{\mathrm{T}}(t_{ij})\beta(t_{ij})\}\{X_i^{\mathrm{T}}(t_{ij})\beta(t_{ij})-X_i^{\mathrm{T}}(t_{ij})\hat{\beta}^{(-i)}(t_{ij})\}]\\ &+\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i[X_i^{\mathrm{T}}(t_{ij})\{\beta(t_{ij})-\hat{\beta}^{(-i)}(t_{ij})\}]^2.\end{aligned} \tag{2·7}$$

The first term on the right-hand side of (2·7) does not depend on the smoothing parameters, and, because of the definition of $\hat{\beta}^{(-i)}(t)$, the expectation of the second term is zero. Thus, by minimising $\textsc{cv}(K)$, K_{cv} approximately minimises the third term on the right-hand side of (2·7), which is clearly an approximation of (2·6).

3. Statistical properties

3·1. *Conditional biases and variances of the basis estimators*

Let $\mathscr{X}=\{(X_i(t_{ij}), t_{ij});\, i=1, \ldots, n, j=1, \ldots, n_i\}$ be the set of the observed covariates. The conditional expectation $\tilde{\gamma}$ of $\hat{\gamma}$ given $\mathscr{X}$ is

$$\tilde{\gamma}=E(\hat{\gamma}\,|\,\mathscr{X})=\left(\sum_{i=1}^{n} U_i^{\mathrm{T}}W_iU_i\right)^{-1}\left(\sum_{i=1}^{n} U_i^{\mathrm{T}}W_i\tilde{Y}_i\right),$$

where $\tilde{Y}_i=E(Y_i\,|\,\mathscr{X})=(\tilde{Y}_{i1}, \ldots, \tilde{Y}_{in_i})$ with $\hat{Y}_{ij}=X_i^{\mathrm{T}}(t_{ij})\beta(t_{ij})$. The biases of $\hat{\beta}(t)$ and $\hat{\beta}_l(t)$ conditioning on $\mathscr{X}$ are

$$E\{\hat{\beta}(t)-\beta(t)\,|\,\mathscr{X}\}=B(t)\tilde{\gamma}-\beta(t), \quad E\{\hat{\beta}_l(t)-\beta_l(t)\,|\,\mathscr{X}\}=e_{l+1}^{\mathrm{T}}\{B(t)\tilde{\gamma}-\beta(t)\},$$

respectively, where e_{l+1} is the $(k\times 1)$ column vector with 1 as its $(l+1)$th element and zero elsewhere. If, for all $l=0, \ldots, k$, $\beta_l(t)$ belongs to the linear space spanned by $\{B_{l1}(t), \ldots, B_{lK_l}(t)\}$, we can write $\beta(t)=B(t)\gamma^*$ for some γ^* so that $\tilde{\gamma}=\gamma^*$ and the conditional bias $E\{\hat{\beta}(t)-\beta(t)\,|\,\mathscr{X}\}$ is 0. The conditional biases need not vanish in general, but one could make them asymptotically negligible as n tends to infinity by choosing a large K_l.

On the other hand, because the conditional variance-covariance matrix of Y_i is

$$V_i = \text{cov}(Y_i|\mathscr{X}) = \begin{pmatrix} C_\varepsilon(t_{i1}, t_{i1}) & \cdots & C_\varepsilon(t_{i1}, t_{in_i}) \\ \vdots & \vdots & \vdots \\ C_\varepsilon(t_{in_i}, t_{i1}) & \cdots & C_\varepsilon(t_{in_i}, t_{in_i}) \end{pmatrix},$$

where $C_\varepsilon(t_{ij}, t_{ij'}) = \text{cov}(\varepsilon_i(t_{ij}), \varepsilon_i(t_{ij'}))$ is the covariance between $\varepsilon_i(t_{ij})$ and $\varepsilon_i(t_{ij'})$, the conditional variance-covariance matrix of $\hat{\gamma}$ is

$$\text{cov}(\hat{\gamma}|\mathscr{X}) = \left(\sum_{i=1}^{n} U_i^{\mathrm{T}} W_i U_i\right)^{-1} \left(\sum_{i=1}^{n} U_i^{\mathrm{T}} W_i V_i W_i U_i\right) \left(\sum_{i=1}^{n} U_i^{\mathrm{T}} W_i U_i\right)^{-1}.$$

Then $\text{cov}\{\hat{\beta}(t)|\mathscr{X}\} = B(t)\,\text{cov}(\hat{\gamma}|\mathscr{X})B^{\mathrm{T}}(t)$ and $\text{var}\{\hat{\beta}_l(t)|\mathscr{X}\} = e_{l+1}^{\mathrm{T}}\,\text{cov}\{\hat{\beta}(t)|\mathscr{X}\}e_{l+1}$. If the $\varepsilon_i(t_{ij})$ are from a known Gaussian process and the conditional biases of the estimators are negligible, the above conditional variance-covariance matrices can be used for statistical inference. However, the intrasubject correlation structure $C_\varepsilon(t_{ij}, t_{ij'})$ is usually unknown in practice and needs to be estimated. Without the normality assumption on $\varepsilon_i(t)$, asymptotic inference is possible based on asymptotic distributional results.

3·2. *Large-sample properties*

We establish in this section the consistency and convergence rates of $\hat{\beta}(t)$, based on the following technical assumptions.

Assumption 1. The observation time points follow a random design in the sense that t_{ij}, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$, are chosen independently from an unknown distribution $F(.)$ with a density $f(.)$ on the finite interval $\mathscr{T}$. The density function $f(t)$ is uniformly bounded away from 0 and infinity; that is, there are positive constants M_1 and M_2 such that $M_1 \leqslant f(t) \leqslant M_2$ for all $t \in \mathscr{T}$.

Assumption 2. Let $E_{XX^{\mathrm{T}}}(t) = E\{X(t)X^{\mathrm{T}}(t)\}$ and let $\lambda_0(t) \leqslant \ldots \leqslant \lambda_k(t)$ be the eigenvalues of $E_{XX^{\mathrm{T}}}(t)$. Then $\lambda_r(t)$ $(r = 0, \ldots, k)$ are uniformly bounded away from 0 and infinity for all $t \in \mathscr{T}$.

Assumption 3. The range of the covariates $X^{(l)}(t)$ is bounded in the sense that there is a positive constant M_3 such that $|X^{(l)}(t)| \leqslant M_3$ for all $t \in \mathscr{T}$ and $l = 0, \ldots, k$.

Assumption 4. There is a positive constant M_4 such that $E\{\varepsilon(t)^2\} \leqslant M_4$ for all $t \in \mathscr{T}$.

We first introduce a distance measure to assess the performance of our estimators. Let $\|a\|_{L_2} = [\int_{\mathscr{T}} \{a(t)\}^2\, dt]^{\frac{1}{2}}$ be the L_2 norm of any square integrable real-valued function $a(t)$ on $\mathscr{T}$ and let $\|A\|_{L_2} = \{\sum_{l=0}^{k} \|a_l\|_{L_2}^2\}^{\frac{1}{2}}$ be the L_2 norm of $A(t) = (a_0(t), \ldots, a_k(t))^{\mathrm{T}}$, where $a_l(t)$ are real-valued functions on $\mathscr{T}$. Define the integrated squared error of $\hat{\beta}_l(t)$ as

$$\textsc{ise}(\hat{\beta}_l) = \|\hat{\beta}_l - \beta_l\|_{L_2}^2 = \int_{\mathscr{T}} \{\hat{\beta}_l(t) - \beta_l(t)\}^2\, dt$$

and the integrated squared error of $\hat{\beta}(t) = (\hat{\beta}_0(t), \ldots, \hat{\beta}_k(t))^{\mathrm{T}}$ as

$$\textsc{ise}(\hat{\beta}) = \sum_{l=0}^{k} \{\textsc{ise}(\hat{\beta}_l)\}.$$

We say that $\hat{\beta}(.)$ is a consistent estimator of $\beta(.)$ if $\lim_{n\to\infty} \textsc{ise}(\hat{\beta}) = 0$ holds in probability, or equivalently $\lim_{n\to\infty} \textsc{ise}(\hat{\beta}_l) = 0$ holds in probability for $l = 0, \ldots, k$.

Let $\tilde{\beta}(t)=E\{\hat{\beta}(t)|\mathscr{X}\}$ and $\tilde{\beta}_l(t)$ $(l=0,\ldots,k)$ be the $(l+1)$th element of $\tilde{\beta}(t)$. By the Cauchy–Schwarz inequality, $\hat{\beta}_l(.)$ is a consistent estimator of $\beta_l(.)$, in the sense that $\lim_{n\to\infty}\textsc{ise}(\hat{\beta}_l)=0$ in probability, if and only if both $\|\hat{\beta}_l-\tilde{\beta}_l\|_{L_2}$ and $\|\tilde{\beta}_l-\beta_l\|_{L_2}$ tend to zero in probability. Thus, the consistency of $\hat{\beta}(.)$ holds if and only if $\|\hat{\beta}_l-\tilde{\beta}_l\|_{L_2}$ and $\|\tilde{\beta}_l-\beta_l\|_{L_2}$ tend to zero in probability for all $l=0,\ldots,k$. Since we approximate $\beta_l(t)$ by functions in a linear space, the asymptotic derivations of $\textsc{ise}(\hat{\beta})$ depend on some L_∞ distances between $\beta_l(t)$ and the chosen linear space. Specifically, let $\mathbb{G}_l$ be the linear space spanned by $\{B_{l1}(t),\ldots,B_{lK_l}(t)\}$ and let $D(\beta_l,\mathbb{G}_l)=\inf_{g\in\mathbb{G}_l}\sup_{t\in\mathscr{T}}|\beta_l(t)-g(t)|$ be the L_∞ distance between $\beta_l(.)$ and $\mathbb{G}_l$. Then the asymptotic properties of $\textsc{ise}(\hat{\beta})$ depend on $\rho_n=\sum_{l=0}^{k}D(\beta_l,\mathbb{G}_l)$,

$$A_{n,l}=\sup_{g\in\mathbb{G}_l,\,\|g\|_{L_2}\neq 0}\frac{\sup_{t\in\mathscr{T}}|g(t)|}{\|g\|_{L_2}},\quad A_n=\max_{0\leqslant l\leqslant k}A_{n,l}.$$

Examples of ρ_n and A_n for the commonly used bases, such as the polynomials, splines and trigonometric bases, can be found in Huang (1998, § 2.2).

Define $K_n=\max_{0\leqslant l\leqslant k}K_l$, which may or may not tend to infinity as n tends to infinity. The next theorem, whose proof is given in the Appendix, shows the consistency and the convergence rates of $\hat{\beta}(.)$.

Theorem 1. *If Assumptions* 1–4 *are satisfied*, $\lim_{n\to\infty}\rho_n=0$ *and*

$$\lim_{n\to\infty}\left[A_n^2K_n\max\left\{\max_{1\leqslant i\leqslant n}(n_iw_i),\ \sum_{i=1}^{n}n_i^2w_i^2\right\}\right]=0, \tag{3·1}$$

then $\hat{\beta}(.)$ *uniquely exists with probability tending to one and is a consistent estimator of* $\beta(.)$. *Moreover*,

(a) $\|\hat{\beta}-\tilde{\beta}\|_{L_2}^2=O_p(K_n\sum_{i=1}^{n}n_i^2w_i^2)$,
(b) $\|\tilde{\beta}-\beta\|_{L_2}=O_p(\rho_n)$,
(c) $\textsc{ise}(\hat{\beta})=O_p\{K_n\sum_{i=1}^{n}(n_i^2w_i^2)+\rho_n^2\}$.

Note that Theorem 1 gives the consistency of $\hat{\beta}(.)$ for general basis choices, including polynomials, splines and trigonometric bases. The convergence rates, however, may be improved when a particular type of basis is used. For an interesting special case, Theorem 2, whose proof is given in the Appendix, gives the improved convergence rates for a class of spline estimators. In this theorem, we assume that each $\mathbb{G}_l$ is a space of polynomial splines on $\mathscr{T}$ with a fixed degree and the knots have bounded mesh ratio; that is, the ratios of the differences between consecutive knots are bounded away from zero and infinity uniformly in n.

Theorem 2. *Suppose that* $\hat{\beta}(t)$ *is defined as in* (2·4) *with a spline basis. If the conditions of Theorem* 1 *are satisfied, then*

(a) $\|\hat{\beta}-\tilde{\beta}\|_{L_2}^2=O_p[\sum_{i=1}^{n}n_i^2w_i^2\{(K_n/n_i)+1\}]$,
(b) $\|\tilde{\beta}-\beta\|_{L_2}=O_p(\rho_n)$,
(c) $\textsc{ise}(\hat{\beta})=O_p[\sum_{i=1}^{n}n_i^2w_i^2\{K_n/n_i)+1\}+\rho_n^2]$.

Remark 3. Different choices of w_i generally lead to different convergence rates of the estimators. For the general situation in Theorem 1, we have

$$\sum_{i=1}^{n}K_nn_i^2w_i^2=\begin{cases}K_n/n, & \text{if } w_i=1/(nn_i),\\ K_n\sum_{i=1}^{n}n_i^2/N^2, & \text{if } w_i=1/N.\end{cases}$$

As shown in Hoover et al. (1998),

$$\lim_{n\to\infty} \sum_{i=1}^{n} n_i^2/N^2 = 0$$

if and only if $\lim_{n\to\infty} \max_{1\leqslant i\leqslant n}(n_i/N)=0$. Thus, as with local smoothing methods, the $w_i=1/N$ weight may lead to inconsistent estimators $\hat{\beta}(.)$, while $w_i=1/(nn_i)$ leads to consistent $\hat{\beta}(.)$ for all choices of n_i.

Remark 4. When specific smoothness conditions are given, more precise convergence rates can be deduced by determining the size of $D(\beta_l, \mathbb{G}_l)$, the discrepancy between $\beta_l(.)$ and the linear space $\mathbb{G}_l$. For example, when $\beta_l(t)$ has bounded second derivatives and $\mathbb{G}_l$ is a space of cubic splines with K_n interior knots on $\mathcal{T}$, we have $D(\beta_l, \mathbb{G}_l)=O(K_n^{-2})$ (Schumaker, 1981, Theorem 6.27) and, by Theorem 1, $\text{ISE}(\hat{\beta})=O_p(K_n/n+K_n^{-4})$. For the special choice of $K_n=O(n^{1/5})$, this reduces to $\text{ISE}(\hat{\beta})=O_p(n^{-4/5})$, which is the optimal convergence rate for nonparametric regression with independent, identically distributed data under the same smoothness conditions; see for example Stone (1982).

4. Some inference procedures based on a resampling subject bootstrap

4·1. *Pointwise confidence intervals*

Let $\{(Y_{ij}^*, X_i^*(t_{ij}^*), t_{ij}^*);\ 1\leqslant i\leqslant n,\ 1\leqslant j\leqslant n_i\}$ be a bootstrap sample obtained by sampling n subjects with replacement from the original data, and let $\hat{\gamma}^*$ and $\hat{\beta}^*(t)$ be the estimators computed as in (2·3) and (2·4), respectively, using the bootstrap sample. With B independent replications, we obtain B bootstrap estimators $\hat{\gamma}^*$ and $\hat{\beta}^*(t)$. A $(1-\alpha)$ confidence interval for $E\{\hat{\beta}_l(t)\}$ based on bootstrap percentiles can be given by

$$(L_{l,\alpha/2}(t),\ U_{l,\alpha/2}(t)), \tag{4·1}$$

where $L_{l,\alpha/2}(t)$ and $U_{l,\alpha/2}(t)$ are the $\{100\times(\alpha/2)\}$th and $\{100\times(1-\alpha/2)\}$th percentiles of the bootstrap estimators of $\beta_l(t)$. A $(1-\alpha)$ interval for $E\{\hat{\beta}_l(t)\}$ based on a normal approximation is

$$\hat{\beta}_l(t) \pm z_{1-\alpha/2}\hat{s}_{l,B}(t), \tag{4·2}$$

where z_p is the 100pth percentile of the standard Gaussian distribution and $\hat{s}_{l,B}(t)$ is the sample standard error of $\hat{\beta}_l(t)$ computed from the B bootstrap estimators of $\beta_l(t)$. Since the bias of $\hat{\beta}_l(t)$ has not been adjusted, (4·1) and (4·2) do not generally lead to adequate confidence intervals for $\beta_l(t)$ unless the bias of $\hat{\beta}_l(t)$ is negligible relative to its variance. In practice, one may either estimate the bias or make it negligible by selecting a large K_l, for example, in the computation of $\hat{\beta}_l(t)$. Alternatively, one can simply admit that $E\{\hat{\beta}_l(t)\}$ is the estimable part of $\beta_l(t)$ and treat $E\{\hat{\beta}_l(t)\}$ as the parameter of interest. This is a sensible approach since $E\{\hat{\beta}(t)\}$, as a good approximation of $\beta_l(t)$, is expected to capture the main feature of $\beta_l(t)$; see § 3.5 of Hart (1997) for a similar argument in the context of kernel smoothing.

4·2. *Simultaneous confidence bands*

We present here a simple approach that extends the above pointwise confidence intervals to simultaneous bands for $E\{\hat{\beta}_l(t)\}$ over a given subinterval $[a, b]$ of $\mathcal{T}$. Partitioning $[a, b]$ into $M+1$ equally spaced grid points $a=\xi_1<\ldots<\xi_{M+1}=b$ for some integer

$M \geqslant 1$, we get a set of approximate $(1-\alpha)$ simultaneous confidence intervals $(l_{l,\alpha}(\xi_r), u_{l,\alpha}(\xi_r))$ for $E\{\hat{\beta}_l(\xi_r)\}$, such that

$$\lim_{n\to\infty} \text{pr}[l_{l,\alpha}(\xi_r) \leqslant E\{\hat{\beta}_l(\xi_r)\} \leqslant u_{l,\alpha}(\xi_r), \text{ for all } r=1,\ldots,M+1] \geqslant 1-\alpha.$$

In particular, if we use Bonferroni adjustment, $(l_{l,\alpha}(\xi_r), u_{l,\alpha}(\xi_r))$ may be given by

$$(L_{l,\alpha/\{2(M+1)\}}(\xi_r), U_{l,\alpha/\{2(M+1)\}}(\xi_r)) \quad \text{or} \quad (\hat{\beta}_l(\xi_r) \pm z_{1-\alpha/\{2(M+1)\}}\hat{s}_{l,B}(\xi_r)). \tag{4·3}$$

Let $E^{(I)}\{\hat{\beta}_l(t)\}$ be the linear interpolation of $E\{\hat{\beta}_l(\xi_r)\}$ and $E\{\hat{\beta}_l(\xi_{r+1})\}$, for $\xi_r \leqslant t \leqslant \xi_{r+1}$, such that

$$E^{(I)}\{\hat{\beta}_l(t)\} = M\left(\frac{\xi_{r+1}-t}{b-a}\right)E\{\hat{\beta}_l(\xi_r)\} + M\left(\frac{t-\xi_r}{b-a}\right)E\{\hat{\beta}_l(\xi_{r+1})\}.$$

Then $(l_{l,\alpha}^{(I)}(t), u_{l,\alpha}^{(I)}(t))$ is an approximate $(1-\alpha)$ confidence band for $E^{(I)}\{\hat{\beta}_l(t)\}$ in the sense that

$$\lim_{n\to\infty} \text{pr}[l_{l,\alpha}^{(I)}(t) \leqslant E^{(I)}\{\hat{\beta}_l(t)\} \leqslant u_{l,\alpha}^{(I)}(t), \text{ for all } t \in [a,b]] \geqslant 1-\alpha,$$

where $l_{l,\alpha}^{(I)}(t)$ and $u_{l,\alpha}^{(I)}(t)$ are linear interpolations of $l_{l,\alpha}(\xi_r)$ and $u_{l,\alpha}(\xi_r)$.

To construct the bands for $E\{\hat{\beta}_l(t)\}$, we assume that either

$$\sup_{t\in[a,b]} |[E\{\hat{\beta}_l(t)\}]'| \leqslant c_1, \tag{4·4}$$

for a known constant $c_1 > 0$, or

$$\sup_{t\in[a,b]} |[E\{\hat{\beta}_l(t)\}]''| \leqslant c_2, \tag{4·5}$$

for a known constant $c_2 > 0$. Direct calculation shows that, for $\xi_r \leqslant t \leqslant \xi_{r+1}$,

$$|E\{\hat{\beta}_l(t)\} - E^{(I)}\{\hat{\beta}_l(t)\}| \leqslant \begin{cases} 2c_1 M\{(\xi_{r+1}-t)(t-\xi_r)/(b-a)\}, & \text{if (4·4) holds,} \\ \tfrac{1}{2}c_2(\xi_{r+1}-t)(t-\xi_r), & \text{if (4·5) holds.} \end{cases}$$

If we adjust the bands for $E^{(I)}\{\hat{\beta}_l(t)\}$, our approximate $(1-\alpha)$ confidence bands for $E\{\hat{\beta}_l(t)\}$ are

$$\left(l_{l,\alpha}^{(I)}(t) - 2c_1 M\left\{\frac{(\xi_{r+1}-t)(t-\xi_r)}{b-a}\right\}, u_{l,\alpha}^{(I)}(t) + 2c_1 M\left\{\frac{(\xi_{r+1}-t)(t-\xi_r)}{b-a}\right\}\right) \tag{4·6}$$

or

$$\left(l_{l,\alpha}^{(I)}(t) - \frac{1}{2}c_2(\xi_{r+1}-t)(t-\xi_r), u_{l,\alpha}^{(I)}(t) + \frac{1}{2}c_2(\xi_{r+1}-t)(t-\xi_r)\right), \tag{4·7}$$

when (4·4) or (4·5) holds, respectively.

Remark 5. The Bonferroni adjustment (4·3), although very simple, often leads to conservative bands. For refinements, one may use the inclusion-exclusion identities to calculate $(l_{l,\alpha}(\xi_r), u_{l,\alpha}(\xi_r))$ with more accurate coverage probabilities; see for example Naiman & Wynn (1997). These refinements, however, usually involve extensive computations, and may not be practical for large longitudinal studies. A related issue is the choice of M. Although some heuristic suggestions for choosing M for the simple case of kernel regression with independent identically distributed samples have been provided by Hall & Titterington (1988), optimal choices of M under the current situation are not available.

4·3. *Hypothesis testing*

It is often of interest to test whether one or several coefficient functions are time-varying or are identically zero. We propose here a goodness-of-fit test based on the comparison of the weighted residual sum of squares from weighted least squares fits under both the null hypothesis and the alternative. Consider for example testing the null hypothesis that none of the coefficient functions, except the baseline curve, is time-varying:

$$H_0 : \beta_l(t) = \beta_l^0,$$

for all $t \in \mathscr{T}$ and all $1 \leqslant l \leqslant k$, where β_l^0 are unknown constants, versus the general alternative that one or more of the coefficient functions $\beta_l(t)$ $(l = 1, \ldots, k)$ are time-varying. If we approximate the baseline $\beta_0(t)$ by $\sum_{s=1}^{K_0} \{\gamma_{0s} B_{0s}(t)\}$, the weighted residual sum of squares under H_0 is

$$\text{RSS}_0 = \sum_{i=1}^{n} \sum_{j=1}^{n_i} w_i \left\{ Y_{ij} - \sum_{s=1}^{K_0} X_i^{(0)}(t_{ij}) B_{0s}(t_{ij}) \hat{\gamma}_{0s}^0 - \sum_{l=1}^{k} \sum_{s=1}^{K_l} X_i^{(l)}(t_{ij}) \hat{\beta}_l^0 \right\}^2,$$

where $\hat{\gamma}_{0s}^0$ and $\hat{\beta}_l^0$ minimise the weighted residual sum of squares among all choices of γ_{0s} and β_l^0. Under the general alternative that all the coefficient functions are allowed to be time-varying, we use the least squares method of § 2·1, so that the corresponding weighted residual sum of squares is

$$\text{RSS}_1 = \sum_{i=1}^{n} \sum_{j=1}^{n_i} w_i \left\{ Y_{ij} - \sum_{l=0}^{k} \sum_{s=1}^{K_l} X_i^{(l)}(t_{ij}) B_{ls}(t_{ij}) \hat{\gamma}_{lk} \right\}^2.$$

THEOREM 3. *Suppose that the conditions of Theorem* 1 *are satisfied*, $\inf_{t \in \mathscr{T}} \sigma^2(t) > 0$, $\sup_{t \in \mathscr{T}} E\{\varepsilon^4(t)\} < \infty$, *and* $T_n = (\text{RSS}_0 - \text{RSS}_1)/\text{RSS}_1$. *Under* H_0, $T_n \to 0$ *in probability as* $n \to \infty$. *Otherwise, if* $\inf_{c \in R} \|\beta_l - c\|_{L_2} > 0$ *for some* $l = 1, \ldots, k$, *then there exists a* $\delta > 0$ *such that, with probability tending to one*, $T_n > \delta$.

The proof of this theorem is in the Appendix. This result motivates the use of a test procedure that rejects H_0 when T_n is larger than an appropriate critical value. We use a resampling subject bootstrap to evaluate the null distribution of T_n. To be specific, let

$$\hat{\varepsilon}_{ij} = Y_{ij} - \sum_{l=0}^{k} \sum_{s=1}^{K_l} X_i^{(l)}(t_{ij}) B_{ls}(t_{ij}) \hat{\gamma}_{lk}$$

be the residuals of (2·4) and let $\{Y_{ij}^p;\ i = 1, \ldots, n, j = 1, \ldots, n_i\}$,

$$Y_{ij}^p = \sum_{s=1}^{K_0} X_i^{(0)}(t_{ij}) B_{0s}(t_{ij}) \hat{\gamma}_{0s}^0 + \sum_{l=1}^{k} \sum_{s=1}^{K_l} X_i^{(l)}(t_{ij}) \hat{\beta}_l^0 + \hat{\varepsilon}_{ij},$$

be a set of pseudo-responses under the null hypothesis. The following bootstrap procedure can be used to evaluate the null distribution of T_n and the p-values of the test.

Step 1. Resample n subjects with replacement from

$$\{(Y_{ij}^p, X_i(t_{ij}), t_{ij});\ i = 1, \ldots, n, j = 1, \ldots, n_i\}$$

to obtain the bootstrap sample $\{(Y_{ij}^{p*}, X_i^*(t_{ij}^*), t_{ij}^*);\ i = 1, \ldots, n, j = 1, \ldots, n_i^*\}$.

Step 2. Repeat the above sampling procedure B times.

Step 3. From each bootstrap sample, calculate the test statistic T_n^* and derive the empirical distribution of T_n^* based on the B independent bootstrap samples.

Step 4. Reject the null hypothesis H_0 at the significance level α when the observed test statistic T_n is greater than or equal to the $\{100 \times (1-\alpha)\}$th percentile of the empirical distribution of T_n^*. The p-value of the test is the empirical probability of $\{T_n^* \geqslant T_n\}$.

For ease of computation, the same basis system is used for each $\beta_l(t)$ in the calculations of T_n^* and T_n in our implementation in § 5. In principle, however, different basis approximations may be used for different bootstrap replications at additional computational cost.

The above testing procedure can be modified in a straightforward way to test other null hypotheses such as that one or a subset of coefficient functions are constants or are identically zero; see § 5·2 for some examples. Theorem 3 can be adapted easily to the general situation. The proposed testing procedure may also be used with other estimation methods, such as local polynomial fitting and smoothing splines.

5. Numerical results

5·1. *Monte Carlo simulation*

In each simulation run, a simple random sample of 200 subjects is generated according to the model

$$Y_{ij} = \beta_0(t_{ij}) + \sum_{k=1}^{3} X_i^{(k)}(t_{ij})\beta_k(t_{ij}) + \varepsilon_i(t_{ij}) \quad (j = 1, \ldots, n_i, i = 1, \ldots, 200).$$

The covariates are chosen as follows: $X_i^{(1)}(t)$ is a uniform random variable over the time-dependent interval $[t/10, 2 + t/10]$; $X_i^{(2)}(t)$, when conditional on $X_i^{(1)}(t)$, is a normal random variable with mean zero and conditional variance $\{1 + X_i^{(1)}(t)\}/\{2 + X_i^{(1)}(t)\}$; and $X_i^{(3)}(t)$, which is independent of $X_i^{(1)}(t)$ and $X_i^{(2)}(t)$, is a Bernoulli random variable with probability of success 0·6. The coefficient curves are given by

$$\beta_0(t) = 15 + 20 \sin\left(\frac{t\pi}{60}\right), \quad \beta_1(t) = 2 - 3 \cos\left\{\frac{(t-25)\pi}{15}\right\},$$
$$\beta_2(t) = 6 - 0{\cdot}2t, \quad \beta_3(t) = -4 + \frac{(20-t)^2}{2000}. \tag{5·1}$$

To generate the observation times, each individual has a set of 'scheduled' time points $\{0, 1, \ldots, 30\}$, and each scheduled time, except time 0, has a probability 60% of being skipped. The actual observation time is a random perturbation of the scheduled time: a $\mathrm{Un}(-0{\cdot}5, 0{\cdot}5)$ random deviate is added to the non-skipped scheduled time to obtain the actual observation time. This leads to unequal numbers of repeated measurements n_i and different observed time points t_{ij} per subject. The random errors $\varepsilon_i(t_{ij})$ are independent from the covariates and are given by $\varepsilon_i(r_{ij}) = Z_i(t_{ij}) + E_{ij}$, where the $Z_i(t_{ij})$ are generated from a stationary Gaussian process with zero mean and a decayed exponential covariance function

$$\mathrm{cov}(Z_{i_1}(t_{i_1 j_1}), Z_{i_2}(t_{i_2 j_2})) = \begin{cases} 4 \exp(-|t_{i_1 j_1} - t_{i_2 j_2}|), & \text{if } i_1 = i_2, \\ 0, & \text{if } i_1 \neq i_2; \end{cases}$$

the E_{ij} are independent measurement errors with the $N(0, 4)$ distribution.

For each simulated dataset, we computed the least squares basis estimators using cubic splines with equally spaced knots. The number of knots were chosen by the leave-one-subject-out crossvalidation described in § 2·3. We allowed different numbers of knots for different coefficient curves. Define the mean absolute deviation of errors by

$$\text{MADE} = \sum_{l=0}^{3} \sum_{r=1}^{n_{gr}} n_{gr}^{-1}\{|\hat{\beta}_l(t_r) - \beta_l(t_r)|/\text{range}(\beta_l)\},$$

where $\{t_r; r = 1, \ldots, n_{gr}\}$ is a set of equally spaced grid points in the support of t_{ij}; $n_{gr} = 301$ in our implementation. In Fig. 1, we present fitted varying coefficient curves corresponding to the nine deciles of the MADE values from 200 simulation runs.

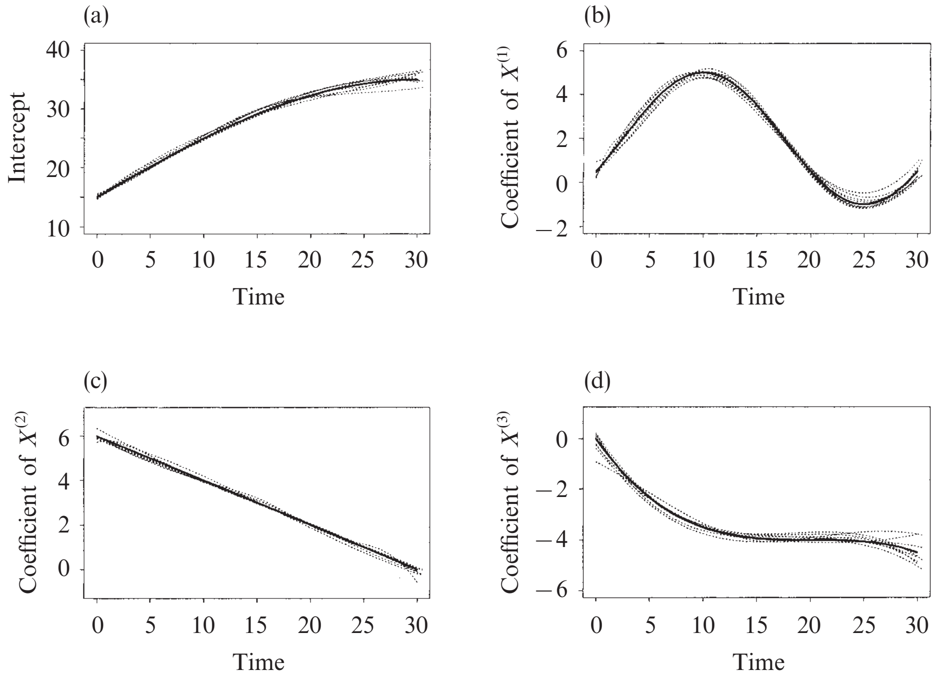

Fig. 1. Monte Carlo simulation. Fitted varying-coefficient curves, dotted lines, corresponding to the nine deciles of the MADE values in 200 simulation runs. The solid lines are the true coefficient functions.

To demonstrate the power of the test, we considered the null hypothesis that $\beta_1(t)$ was time-invariant versus the general alternative that $\beta_1(t)$ was time-varying. The same data-generating mechanism as above was used to create the longitudinal datasets under both the null and alternative hypotheses. Under both hypotheses, $\beta_0(t)$, $\beta_2(t)$ and $\beta_3(t)$ were as defined in (5·1), and we used a sequence of alternative models indexed by λ:

$$\beta_1(t; \lambda) = c + \lambda\{\beta_1(t) - c\} \quad (0 \leqslant \lambda \leqslant 1),$$

where $\beta_1(t)$ is defined in (5·1) and $c = \int_0^{30} \beta_1(t)\, dt/30$.

For each λ in $\{0, 0{\cdot}025, 0{\cdot}050, \ldots, 1{\cdot}00\}$, we generated 200 independent longitudinal samples, each with 200 independent subjects, and applied to each simulated sample the bootstrap test of § 4·3 based on 200 bootstrap repetitions and cubic spline estimators with 5 equally spaced knots. Figure 2 shows the resulting empirical power function as a function of λ. As expected, the power is a monotone increasing function of λ, which is close to 0·05 when λ is close to 0, and is close to 1 at $\lambda = 0{\cdot}3$.

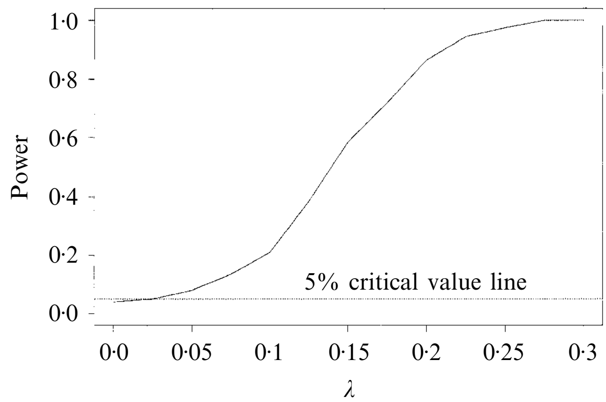

Fig. 2. Monte Carlo simulation. Plot of power curve against λ for the goodness-of-fit test.

5·2. *Application to* AIDS *data*

As a subset from the Multicenter AIDS Cohort Study, the data include the repeated measurements of physical examinations, laboratory results and CD4 cell counts and percentages of 283 homosexual men who became HIV-positive between 1984 and 1991. All individuals were scheduled to have their measurements made at semi-annual visits, but, because many individuals missed some of their scheduled visits and the HIV infections happened randomly during the study, there are unequal numbers of repeated measurements and different measurement times per individual. Further details about the design, methods and medical implications of the study can be found in Kaslow et al. (1987).

The aims of our statistical analysis are to describe the trend of mean CD4 percentage depletion over time and to evaluate the effects of cigarette smoking, pre-HIV infection CD4 percentage and age at HIV infection on the mean CD4 percentage after the infection. Denote by t_{ij} the time in years of the jth measurement of the ith individual after HIV infection, by Y_{ij} the ith individual's CD4 percentage at time t_{ij} and by $X_i^{(1)}$ the ith individual's smoking status; $X_i^{(1)}$ is 1 or 0 if the ith individual ever or never smoked cigarettes, respectively, after the HIV infection. In order to get a clear biological interpretation, we define $X_i^{(2)}$ to be the ith individual's centred age at HIV infection, obtained by subtracting the sample average age at infection from the ith individual's age at infection. The ith individual's centred pre-infection CD4 percentage, denoted by $X_i^{(3)}$, is computed by subtracting the average pre-infection CD4 percentage of the sample from the ith individual's actual pre-infection CD4 percentage. These covariates, except the time, are time-invariant. The varying-coefficient model for Y_{ij}, t_{ij} and $X_i = (1, X_i^{(1)}, X_i^{(2)}, X_i^{(3)})^{\mathrm{T}}$ is

$$Y_{ij} = \beta_0(t_{ij}) + X_i^{(1)}\beta_1(t_{ij}) + X_i^{(2)}\beta_2(t_{ij}) + X_i^{(3)}\beta_3(t_{ij}) + \varepsilon_{ij}, \tag{5·2}$$

where $\beta_0(t)$, the baseline CD4 percentage, represents the mean CD4 percentage t years after the infection for a non-smoker with average pre-infection CD4 percentage and average age at HIV infection, and $\beta_1(t)$, $\beta_2(t)$ and $\beta_3(t)$ describe the time-varying effects for cigarette smoking, age at HIV infection and pre-infection CD4 percentage, respectively, on the post-infection CD4 percentage at time t.

In two prior analyses of the same dataset, Wu & Chiang (2000) and Fan & Zhang (2000) considered the nonparametric estimation of $\beta_l(t)$ $(l = 0, 1, 2, 3)$ using local smoothing methods. We analysed the data using the estimation and inference procedures of §§ 2 and 4 with cubic splines and equally spaced knots. From the crossvalidation procedure of § 2·3, the numbers of interior knots for $\hat{\beta}_0(.)$, $\hat{\beta}_1(.)$, $\hat{\beta}_2(.)$ and $\hat{\beta}_3(.)$ were chosen to be 0, 5, 1 and 3, respectively. Figure 3 shows the fitted coefficient functions, solid curves, and their 95% pointwise confidence intervals, dotted curves. These intervals were computed using the percentile procedure (4·1) with 1000 bootstrap replications. These results have the following implications: the baseline CD4 percentage of the population decreases with time, but at a rate that appears to be gradually slowing down; cigarette smoking and age of HIV infection do not show any significant effect on the post-infection CD4 percentage; and pre-infection CD4 percentage appears to be positively associated with high post-infection CD4 percentage. These findings basically agree with those discovered by the local smoothing methods of Wu & Chiang (2000) and Fan & Zhang (2000), although, possibly because of the random variation of the data, some of the estimated curves vary slightly from method to method.

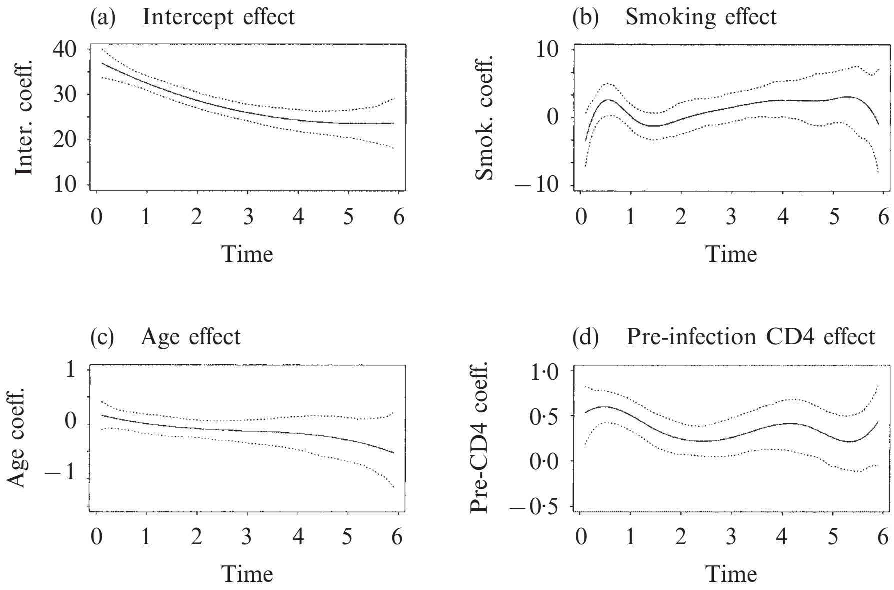

Fig. 3. Application to AIDS data. Estimated coefficient curves for the baseline CD4 percentage and the effects of smoking, age and pre-infection CD4 percentage on the percentage of CD4 cells. Solid curves, estimated effects; dotted curves, 95% bootstrap pointwise confidence intervals.

To test some specific models that may have scientific relevance, we considered four null hypotheses for the coefficient curves: H_{01}, smoking has no effect; H_{02}, age has no effect; H_{03}, the baseline CD4 percentage curve is time-invariant; and H_{04}, pre-infection CD4 effect is time-invariant. We used the bootstrap procedure of § 4·3 to test each of the above null hypotheses separately against its corresponding general alternative: for testing H_{01}, or H_{02}, we constructed RSS_0 by assuming that $\beta_1(t)=0$, or $\beta_2(t)=0$, and that other coefficients are time-varying in model (5·2), and computed RSS_1 by assuming all coefficients to be time-varying. For testing H_{03}, or H_{04}, we constructed RSS_0 by assuming that $\beta_0(t)$ is constant, or $\beta_3(t)$ is constant, and other coefficients are time-varying in model (5·2), and computed RSS_1 by assuming all coefficients are time-varying. The time-varying coefficients were then approximated by cubic splines with five equally spaced knots and fitted using the least squares method of § 2·1 with $w_i = 1/(nn_i)$. The number of bootstrap replicates was chosen to be $B = 1000$. The observed test statistics and their p-values are summarised in Table 1. At the 0·05 significance level, there is insufficient evidence to reject H_{01}, H_{02} and H_{04} and convincing evidence for the rejection of H_{03}. However, with a p-value of 0·059, the decision about H_{04} is borderline; we increased B from 1000 to 5000 in the test of H_{04}, and obtained a p-value of 0·061.

Table 1. *Application to* AIDS *data. Summary of goodness-of-fit tests*

Null hypothesis	Value of test statistic	p-value
Smoking has no effect	0·0125	0·176
Age has no effect	0·0102	0·301
Baseline effect is constant	0·1103	0·000
Pre-CD4 effect is contant	0·0118	0·059

Acknowledgement

Partial support for this research was provided by the National Institute on Drug Abuse, the National Science Foundation, the Acheson Duncan Fund of the Johns Hopkins University and the University of Pennsylvania Research Foundation. The authors would like to thank the editor and the reviewers for insightful comments and suggestions and Dr Joseph B. Margolick for providing the data for § 5.

Appendix

Proofs

Proof of Theorem 1. We assume, without loss of generality, that $B_{ls}(t)$, for $s=1, \ldots, K_l$, is an orthonormal basis for the linear space $\mathbb{G}_l$, for $l=1, \ldots, k$, with inner product $\langle f_1, f_2 \rangle = \int_{\mathcal{T}} f_1(t) f_2(t)\, dt$. Then for any $g_l \in \mathbb{G}_l$ there is a unique representation $g_l(t)=\sum_{s=1}^{K_l} \gamma_{ls} B_{ls}(t)$, so that the L_2-norm of $g(t)=(g_0(t), \ldots, g_k(t))^{\mathrm{T}}$ is $\|g\|_{L_2}=\{\sum_{l=0}^{k} \|g_l\|_{L_2}^2\}^{1/2}=(\sum_{l=0}^{k} \sum_{s=1}^{K_l} \gamma_{ls}^2)^{1/2}$.

Following the notation of Huang (1998, p. 246), we write $a_n \asymp b_n$ if both a_n and b_n are positive and a_n/b_n and b_n/a_n are bounded for all n. Let T be the random variable of time with distribution $F(.)$ and density $f(.)$. The proof is derived from the following series of lemmas.

Lemma A1. *If* (3·1) *is satisfied, then*

$$\sup_{g_l \in \mathbb{G}_l,\, l=0,\ldots,k} \left| \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} w_i \{\sum_{l=0}^{k} X_i^{(l)}(t_{ij}) g_l(t_{ij})\}^2}{E\{\sum_{l=0}^{k} X^{(l)}(T) g_l(T)\}^2} - 1 \right| = o_p(1).$$

Proof. The lemma can be proved using arguments similar to those in the proof of Lemma 10 of Huang (1998). Details are omitted.

Lemma A2. *Suppose that* (3·1) *holds and* $U=(U_1^{\mathrm{T}}, \ldots, U_n^{\mathrm{T}})^{\mathrm{T}}$ *with* U_i *defined as in* (2·2). *There is an interval* $[M_1, M_2]$ *with* $0<M_1<M_2$ *such that, as* $n \rightarrow \infty$,

$$\operatorname{pr}\{\textit{all the eigenvalues of } U^{\mathrm{T}} W U \textit{ fall in } [M_1, M_2]\} \rightarrow 1.$$

Consequently, with probability tending to 1, $U^{\mathrm{T}} W U=\sum_{i=1}^{n}(U_i^{\mathrm{T}} W_i U_i)$ *is invertible and* $\hat{\beta}(.)$ *exists uniquely.*

Proof. By Lemma A1, the following holds with probability tending to one as $n \rightarrow \infty$:

$$\gamma^{\mathrm{T}} U^{\mathrm{T}} W U \gamma=\sum_{i=1}^{n} \sum_{j=1}^{n_i}\left[w_i\left\{\sum_{l=0}^{k} X_i^{(l)}(t_{ij}) g_l(t_{ij})\right\}^2\right] \asymp E\left[\left\{\sum_{l=0}^{k} X^{(l)}(T) g_l(T)\right\}^2\right],$$

where $g_l=\sum_{s=1}^{K_l}\gamma_{ls}B_{ls}$, for $l=0,\ldots,k$, and γ is the vector with entries γ_{ls}, for $s=1,\ldots,K_l$ and $l=0,\ldots,k$. Using conditional expectations and Assumptions 1 and 2, we have that

$$E\left[\left\{\sum_{l=0}^{k}X^{(l)}(T)g_l(T)\right\}^2\right]=\int_{\mathscr{T}}g^{\mathrm{T}}(t)E_{XX^{\mathrm{T}}}(t)g(t)f(t)\,dt\asymp\int_{\mathscr{T}}g^{\mathrm{T}}(t)g(t)\,dt=\sum_{l=0}^{k}\|g_l\|_{L_2}^2$$

holds uniformly for all $g_l\in\mathbb{G}_l$ $(l=0,\ldots,k)$. Thus, $\gamma^{\mathrm{T}}U^{\mathrm{T}}WU\gamma\asymp\gamma^{\mathrm{T}}\gamma$ holds uniformly for all γ. The conclusion of the lemma follows.

LEMMA A3. *If* (3·1) *holds*, *then*

$$\|\hat{\beta}-\tilde{\beta}\|_{L_2}^2=O_p\left(K_n\sum_{i=1}^{n}n_i^2w_i^2\right),\quad\left(\sum_{l=0}^{k}\|\tilde{\beta}_l-\beta_l\|_{L_2}^2\right)^{\frac{1}{2}}=O_p(\rho_n).$$

Proof. By direct calculation, we have $\|\hat{\beta}-\tilde{\beta}\|_{L_2}^2=\sum_{l=0}^{k}\sum_{s=1}^{K_l}|\hat{\gamma}_{ls}-\tilde{\gamma}_{ls}|^2$,

$$\hat{\gamma}-\tilde{\gamma}=\left(\sum_{i=1}^{n}U_i^{\mathrm{T}}W_iU_i\right)^{-1}\sum_{i=1}^{n}U_i^{\mathrm{T}}W_i\varepsilon_i=(U^{\mathrm{T}}WU)^{-1}U^{\mathrm{T}}W\varepsilon$$

and, by Lemma A2, with probability tending to 1, $|(U^{\mathrm{T}}WU)^{-1}U^{\mathrm{T}}W\varepsilon|^2\asymp\varepsilon^{\mathrm{T}}WUU^{\mathrm{T}}W\varepsilon$. Using the Cauchy–Schwarz inequality and Assumptions 3 and 4, we have that

$$E(|U_i^{\mathrm{T}}W_i\varepsilon_i|^2)=E\left[\sum_{l=0}^{k}\sum_{s=1}^{K_l}w_i^2\left\{\sum_{j=1}^{n_i}X_i^{(l)}(t_{ij})B_{ls}(t_{ij})\varepsilon_{ij}\right\}^2\right]=O(K_nn_i^2w_i^2)\qquad(\mathrm{A}{\cdot}1)$$

and, consequently,

$$E(\varepsilon^{\mathrm{T}}WUU^{\mathrm{T}}W\varepsilon)=\sum_{i=1}^{n}E(\varepsilon_i^{\mathrm{T}}W_iU_iU_i^{\mathrm{T}}W_i\varepsilon_i)=O\left(K_n\sum_{i=1}^{n}n_i^2w_i^2\right).$$

The Markov inequality then implies that

$$|(U^{\mathrm{T}}WU)^{-1}U^{\mathrm{T}}W\varepsilon|^2=O_P\left(K_n\sum_{i=1}^{n}n_i^2w_i^2\right).$$

This proves the first assertion of the lemma.

To prove the second assertion, we consider $g^*(t)=(g_0^*(t),\ldots,g_k^*(t))^{\mathrm{T}}$ and $g_l^*\in\mathbb{G}_l$, such that $\sup_{t\in\mathscr{T}}|g_l^*(t)-\beta_l(t)|=D(\beta_l,\mathbb{G}_l)$. Since $|\tilde{\beta}_l(t)-\beta_l(t)|\leqslant|\tilde{\beta}_l(t)-g_l^*(t)|+|g_l^*(t)-\beta_l(t)|$, it suffices to show that $(\sum_{l=0}^{k}\|\tilde{\beta}_l-g_l^*\|_{L_2}^2)^{1/2}=O_p(\rho_n)$. There is a $\gamma^*=(\gamma_0^{*\mathrm{T}},\ldots,\gamma_k^{*\mathrm{T}})^{\mathrm{T}}$ with $\gamma_l^*=(\gamma_{l1}^*,\ldots,\gamma_{lK_l}^*)^{\mathrm{T}}$ such that $g^*(t)=B(t)\gamma^*$. Note that $\tilde{\beta}(t)=B(t)\tilde{\gamma}$. It follows from Lemma A2 that

$$\sum_{l=0}^{k}\|\tilde{\beta}_l-g_l^*\|_{L_2}^2=\sum_{l=0}^{k}\sum_{s=1}^{K_l}|\tilde{\gamma}_{ls}-\gamma_{ls}^*|^2\asymp(\tilde{\gamma}-\gamma^*)^{\mathrm{T}}\left(\sum_{i=1}^{n}U_i^{\mathrm{T}}W_iU_i\right)(\tilde{\gamma}-\gamma^*).$$

Set $\tilde{Y}_i=(\tilde{Y}_{i1},\ldots,\tilde{Y}_{in_i})$ with $\tilde{Y}_{ij}=X_i(t_{ij})^{\mathrm{T}}\beta(t_{ij})$ for $j=1,\ldots,n_i$ and $i=1,\ldots,n$. Since

$$\sum_{i=1}^{n}\{U_i^{\mathrm{T}}W_i(\tilde{Y}_i-U_i\tilde{\gamma})\}=0,$$

we have that

$$\sum_{i=1}^{n}w_i|U_i\tilde{\gamma}-U_i\gamma^*|^2\leqslant\sum_{i=1}^{n}w_i|\tilde{Y}_i-U_i\gamma^*|^2$$

and, by Assumption 3, $|X_i^{\mathrm{T}}(t_{ij})\{\beta(t_{ij})-B(t_{ij})\gamma^*\}|=O(\rho_n)$. Thus,

$$(\tilde{\gamma}-\gamma^*)^{\mathrm{T}}\left(\sum_{i=1}^{n}U_i^{\mathrm{T}}W_iU_i\right)(\tilde{\gamma}-\gamma^*)\leqslant\sum_{i=1}^{n}\sum_{j=1}^{n_i}w_i\rho_n^2=\rho_n^2.$$

The assertion of the lemma then follows from the above computations.

Theorem 1 is then a direct consequence of Lemma A3 and the triangle inequality. □

Proof of Theorem 2. This theorem can be proved along the same lines as Theorem 1, but we need to use the special properties of spline functions. For $l=1,\ldots,k$, set $B_{ls}=K_l^{1/2}N_{ls}$ $(s=1,\ldots,K_l)$, where N_{ls} are B-splines as defined in de Boor (1978, Ch. IX). The B-splines N_{ls} are nonnegative functions satisfying $\sum_{s=1}^{K_l} N_{ls}(t)=1$ for $t\in\mathscr{T}$ and $\int_{\mathscr{T}} N_{ls}(t)\,dt \leqslant M/K_l$ for some constant M. Moreover, there are positive constants M_1 and M_2 such that

$$\frac{M_1}{K_l}\sum_{s=1}^{K_l}\gamma_{ls}^2 \leqslant \int_{\mathscr{T}}\left\{\sum_{s=1}^{K_l}\gamma_{ls}N_{ls}(t)\right\}^2 dt \leqslant \frac{M_2}{K_l}\sum_{s=1}^{K_l}\gamma_{ls}^2,$$

for $\gamma_{ls}\in R$ and $s=1,\ldots,K_l$. If we use these properties of B-splines, (A·1) can be strengthened to

$$E(|U_i^{\mathrm{T}}W_i\varepsilon_i|^2)=E\left[\sum_{l=0}^{k}\sum_{s=1}^{K_l} w_i^2\left\{\sum_{j=1}^{n_i} X_i^{(l)}(t_{ij})B_{ls}(t_{ij})\varepsilon_{ij}\right\}^2\right]\leqslant w_i^2\sum_{l=0}^{k}\left\{n_i+(n_i^2-n_i)\frac{1}{K_l}\right\}K_l.$$

The rest of the proof is similar to that of Theorem 1 and thus is omitted. □

Proof of Theorem 3. Write $\hat\beta^0(t)=(\hat\beta_0^0(t),\hat\beta_1^0,\ldots,\hat\beta_k^0)^{\mathrm{T}}$, where $\hat\beta_0^0(t)=\sum_{s=1}^{K_0}\{\hat\gamma_{0s}^0B_{0s}(t)\}$. It can be shown by direct calculation and Lemma A1 that, with probability tending to one as $n\to\infty$,

$$\mathrm{RSS}_0-\mathrm{RSS}_1=\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i[X_i^{\mathrm{T}}(t_{ij})\{\hat\beta(t_{ij})-\hat\beta^0(t_{ij})\}]^2 \asymp \|\hat\beta-\hat\beta^0\|_{L_2}^2.$$

Then, under H_0, $\|\hat\beta-\hat\beta^0\|_{L_2}\leqslant\|\hat\beta-\beta\|_{L_2}+\|\hat\beta^0-\beta\|_{L_2}\to 0$, in probability as $n\to\infty$. On the other hand, because $\|\hat\beta-\hat\beta^0\|_{L_2}\geqslant\|\hat\beta^0-\beta\|_{L_2}-\|\hat\beta-\beta\|_{L_2}$, we have that, as $n\to\infty$,

$$\|\hat\beta-\hat\beta^0\|_{L_2}\geqslant\sum_{l=1}^{k}\|\hat\beta_l^0-\beta_l\|_{L_2}-o_p(1)\geqslant\sum_{l=1}^{k}\inf_{c\in R}\|\beta_l-c\|_{L_2}-o_p(1).$$

It remains to show that, with probability tending to 1, RSS_1 is bounded away from zero and infinity. By the definition of RSS_1 and $\tilde\beta(t)=E[\hat\beta(t)|\mathscr{X}]$, we have that

$$\mathrm{RSS}_1=\sum_{i=1}^{n}\sum_{j=1}^{n_i}[Y_{ij}-X_i^{\mathrm{T}}(t_{ij})\beta(t_{ij})+X_i^{\mathrm{T}}(t_{ij})\{\beta(t_{ij})-\tilde\beta(t_{ij})\}+X_i^{\mathrm{T}}(t_{ij})\{\tilde\beta(t_{ij})-\hat\beta(t_{ij})\}]^2.$$

It follows from the proof of Theorem 1 that

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i[X_i^{\mathrm{T}}(t_{ij})\{\beta(t_{ij})-\tilde\beta(t_{ij})\}]^2=o_p(1),\quad \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i[X_i^{\mathrm{T}}(t_{ij})\{\tilde\beta(t_{ij})-\hat\beta(t_{ij})\}]^2=o_p(1).$$

Thus, it suffices to show that, with probability tending to 1, $\sum_{i=1}^{n}\sum_{j=1}^{n_i}w_i\varepsilon_i^2(t_{ij})$ is bounded away from zero and infinity. By $\sup_{t\in\mathscr{T}}E\{\varepsilon^4(t)\}<\infty$, there is a constant $c>0$ such that

$$\mathrm{var}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n_i}w_i\varepsilon_i^2(t_{ij})\right\}\leqslant\sum_{i=1}^{n}w_i^2n_i\sum_{j=1}^{n_i}E\{\varepsilon_i^4(t_{ij})\}\leqslant\sum_{i=1}^{n}n_i^2w_i^2c\to 0.$$

The Chebyshev inequality then implies that, as $n\to\infty$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}w_i\varepsilon_i^2(t_{ij})-E\left\{\sum_{i=1}^{n}\sum_{j=1}^{n_i}w_i\varepsilon_i^2(t_{ij})\right\}\to 0,$$

in probability. Since $\sum_{i=1}^{n}(n_iw_i)=1$ and $E\{\varepsilon_i^2(t_{ij})\}$ is bounded away from zero and infinity, the result follows. □

References

Boor, C. de (1978). *A Practical Guide to Splines*. New York: Springer.

Chiang, C. T., Rice, J. A. & Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Am. Statist. Assoc.* **96**, 605–19.

DAVIDIAN, M. & GILTINAN, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. London: Chapman and Hall.

DIGGLE, P. J., LIANG, K.-Y. & ZEGER, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

FAN, J. & ZHANG, J.-T. (2000). Functional linear models for longitudinal data. *J. R. Statist. Soc.* B **62**, 303–22.

HALL, P. & TITTERINGTON, D. M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Mult. Anal.* **27**, 228–54.

HART, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer-Verlag.

HART, J. D. & WEHRLY, T. E. (1986). Kernel regression estimation using repeated measurements data. *J. Am. Statist. Assoc.* **81**, 1080–8.

HART, J. D. & WEHRLY, T. E. (1993). Consistency of cross-validation when the data are curves. *Stoch. Proces. Appl.* **45**, 351–61.

HASTIE, T. J. & TIBSHIRANI, R. J. (1993). Varying-coefficient models. *J. R. Statist. Soc.* B **55**, 757–76.

HOOVER, D. R., RICE, J. A., WU, C. O. & YANG, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–22.

HUANG, J. Z. (1998). Projection estimation in multiple regression with applications to functional ANOVA models. *Ann. Statist.* **26**, 242–72.

KASLOW, R. A., OSTROW, D. G., DETELS, R., PHAIR, J. P., POLK, B. F. & RINALDO, C. R. (1987). The Multicenter AIDS Cohort Study: rationale, organization and selected characteristics of the participants. *Am. J. Epidem.* **126**, 310–8.

LIN, X. & CARROLL, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Statist. Assoc.* **95**, 520–34.

MOYEED, R. A. & DIGGLE, P. J. (1994). Rates of convergence in semiparametric modelling of longitudinal data. *Aust. J. Statist.* **36**, 75–93.

MÜLLER, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*, Lecture Notes in Statistics, **46**. Berlin: Springer-Verlag.

NAIMAN, D. & WYNN, H. (1997). Abstract tubes, improved inclusion-exclusion identities and inequalities and importance sampling. *Ann. Statist.* **36**, 75–93.

RAMSAY, J. O. & SILVERMAN, B. W. (1997). *Functional Data Analysis*. New York: Springer-Verlag.

RICE, J. A. & SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B **53**, 233–43.

RICE, J. A. & WU, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–9.

SCHUMAKER, L. (1981). *Spline Functions: Basic Theory*. New York: John Wiley & Sons.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1348–60.

VONESH, E. F. & CHINCHILLI, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker.

WU, C. O. & CHIANG, C.-T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statist. Sinica* **10**, 433–56.

ZEGER, S. L. & DIGGLE, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–99.

[*Received June* 2000. *Revised March* 2001]