Austral. J. Statist. 36(1), 1994, 75–93

RATES OF CONVERGENCE IN SEMI-PARAMETRIC MODELLING OF LONGITUDINAL DATA

R.A. MOYEED[1] AND P.J. DIGGLE[1]

University of Lancaster

Summary

We consider the problem of semi-parametric regression modelling when the data consist of a collection of short time series for which measurements within series are correlated. The objective is to estimate a regression function of the form $\mathrm{E}[Y(t) \mid x] = x'\beta + \mu(t)$, where $\mu(\cdot)$ is an arbitrary, smooth function of time t, and x is a vector of explanatory variables which may or may not vary with t. For the non-parametric part of the estimation we use a kernel estimator with fixed bandwidth h. When h is chosen without reference to the data we give exact expressions for the bias and variance of the estimators for β and $\mu(t)$, and an asymptotic analysis of the case in which the number of series tends to infinity whilst the number of measurements per series is held fixed. We also report the results of a small-scale simulation study to indicate the extent to which the theoretical results continue to hold when h is chosen by a data-based cross-validation method.

Key words: Autocorrelation; cross-validation; kernel regression; longitudinal data; semi-parametric regression; smoothing; time series.

1. Introduction

Longitudinal data consist of time-sequences of measurements made on each of a number of experimental units. A useful modelling framework for many longitudinal data problems is that the observed sequence of measurements on an experimental unit is sampled from a realisation of a continuous-time stochastic process $Y(t)$. Thus if $y(t)$ denotes the realisation of $Y(t)$, then the observed measurements are $y_j = y(t_j)$, where t_j $(j = 1, \ldots, m)$ are the times at which measurements are made. An important problem in many applications is to estimate the mean response, $\mathrm{E}[Y(t)]$, and the effect on the mean response of experimental treatments or other explanatory variables, whilst taking account of the covariance structure of $Y(t)$.

The linear modelling approach to this problem, in which we assume that $\mathrm{E}[Y(t_j)] = x_j'\beta$ for a suitable p-element vector x_j of explanatory variables measured at time t_j, is now very well established. Recent contributions include

Received April 1993; revised December 1993; accepted December 1993.

[1]Dept of Mathematics, Lancaster University, Lancaster LA1 4YF, UK.

Acknowledgements. This work was supported by the Science and Engineering Council under their Complex Stochastic Systems Initiative.

Pantula & Pollock (1985), Ware (1985), Jones (1987), Kenward (1987), Diggle (1988), Verbyla & Cullis (1990), Jones & Ackerson (1990), Jones & Boadi-Boteng (1991) and Munoz *et al.* (1992). The non-parametric approach, in which the mean response is assumed to be an arbitrary, smooth function of time, and explanatory variables are either absent altogether or are confined to treatment indicators, is considered in Hart & Wehrly (1986), Diggle & Hutchinson (1989), Altman (1990) and Hart (1991).

In this paper we consider the semi-parametric approach in which $\mathrm{E}[Y(t_j)] = x_j'\beta + \mu(t_j)$ where $\mu(t)$ is a smooth, but otherwise arbitrary, function of time. Zeger & Diggle (1994) considered this approach in the context of a specific application to the problem of estimating the time-trend in CD4 cell numbers amongst HIV sero-converters. The present paper gives a more general treatment of the properties of the estimators proposed by Zeger & Diggle (1994), and of a modified method of estimation which appears to have some theoretical advantages.

Treatments of semi-parametric modelling in a non-longitudinal context include Green *et al.* (1985), Heckman (1986), Rice (1986), Green (1987), Speckman (1988), Hastie & Tibshirani (1990) and Cuzick (1992).

In Section 2 we give the details of our assumed model. In Section 3 we review Zeger & Diggle's estimators for β and $\mu(t)$, and describe our proposed modifications. Section 4 gives our results from the asymptotic analysis of both methods of estimation, under the assumption that the number of measurements on each unit is fixed whilst the number of units tends to infinity. We also present numerical results which suggest that the asymptotic theory gives reliable pointers to the behaviour of the estimators in moderate samples. Section 5 is a brief summary of our conclusions.

2. The Model

We suppose that the data consist of n units, with m_i measurements on the ith unit. Let $\{y_{ij} : j = 1, \ldots, m_i; i = 1, \ldots, n\}$ denote the jth measurement on the ith unit, and t_{ij} the time at which the measurement y_{ij} is made. We consider the semi-parametric model

$$y_{ij} = x_{ij}'\beta + \mu(t_{ij}) + w_i(t_{ij}) + z_{ij} \tag{1}$$

where x_{ij} is a p-element vector of covariate values, β is an unknown p-element vector of regression parameters, $\mu(t)$ is an arbitrary smooth function of time t, the $w_i(t)$ $(i = 1, \ldots, n)$ are independent replicates of a zero mean stationary process with covariance function $\gamma(u) = \sigma_w^2\rho(u; \theta)$, where ρ is the correlation function, and z_{ij} are mutually independent measurement errors with mean zero and variance σ_z^2.

According to the model (1), the covariance structure of the complete set of measurements y_{ij} is given by

$$\operatorname{cov}(y_{ij}, y_{k\ell}) = \begin{cases} \sigma_z^2 + \sigma_w^2 & (i = k, j = \ell), \\ \sigma_w^2 \rho(t_{ij} - t_{k\ell}; \theta) & (i = k, j \neq \ell), \\ 0 & (i \neq k). \end{cases} \tag{2}$$

Because of the longitudinal nature of the data, it is necessary to introduce some additional notation. First, write $N = \sum_{i=1}^{n} m_i$ for the total number of observations. Quantities relating to individual units are written as $y_i = (y_{i1}, \ldots, y_{im_i})'$, $\mu_i = (\mu_{i1}, \ldots, \mu_{im_i})'$ and $X_i = (x_{i1}, \ldots, x_{im_i})'$. Note that μ_{ij} is shorthand for $\mu(t_{ij})$, and that X_i is an $m_i \times p$ matrix with jth row x'_{ij}. Let V_i be the $m_i \times m_i$ covariance matrix of y_i, with (j,k)th element $v_{jk} = \sigma_w^2 \rho(t_{ij} - t_{ik}; \theta) + \sigma_z^2 \delta_{j-k}$, where δ is the Kronecker delta. Finally, to represent quantities relating to the complete data set, we write $y = (y'_1, \ldots, y'_n)'$, $\mu = (\mu'_1, \ldots, \mu'_n)'$, $X = (X'_1, \ldots, X'_n)'$, the $N \times p$ design matrix for the parametric part of equation (1) and $V = \operatorname{diag}(V_1, \ldots, V_n)$, the $N \times N$ block diagonal covariance matrix of y.

3. Semi-parametric Regression

3.1. Estimation of μ and β

In matrix notation the model (1) for the full data set can be expressed as

$$y = X\beta + \mu + \epsilon \tag{3}$$

where $\mathrm{E}(\epsilon) = 0$ and $\operatorname{var}(\epsilon) = V$.

A standard approach of estimating β and μ would be to adapt the existing estimators for the i.i.d. case (see Hastie & Tibshirani, 1990) by appropriately modifying them to take into consideration the correlation structure of the data, i.e. the sum of squared residuals is replaced by a quadratic form with weight matrix proportional to V^{-1}. Following Hastie & Tibshirani (1990), Zeger & Diggle (1994) have suggested a backfitting algorithm which computes the estimates $\hat{\beta}$ and $\hat{\mu}$ for the model (3) iteratively. They use a locally adaptive kernel smoother for estimating μ. Their procedure defines the estimators (henceforth referred to as the ZD estimators) as the solutions to the simultaneous equations

$$\hat{\mu}_{ZD} = K(y - X\hat{\beta}_{ZD}) \tag{4}$$

and

$$\hat{\beta}_{ZD} = (X'V^{-1}X)^{-1}X'V^{-1}(y - \hat{\mu}_{ZD}), \tag{5}$$

where K is the kernel smoother matrix. In fact, we can obtain a closed-form solution by substituting from (4) into (5) and re-arranging, to give

$$\hat{\beta}_{ZD} = \{X'V^{-1}(I - K)X\}^{-1}X'V^{-1}(I - K)y. \tag{6}$$

Kernel smoothing is a common method of estimating the mean function in any non-parametric regression model. The specific kernel estimator that is used in this article is the Nadaraya–Watson estimator defined by the weights

$$K_h(s,t_{ij}) = \kappa\left(\frac{s-t_{ij}}{h}\right)\Big/\sum_{k=1}^{n}\sum_{\ell=1}^{m_k}\kappa\left(\frac{s-t_{k\ell}}{h}\right), \tag{7}$$

where κ is the kernel function and h is called the bandwidth. We consider kernels with finite support, say on $[-1,1]$, with

$$\int_{-1}^{1}\kappa(t)t^j\,dt = \begin{cases} 1 & (j=0),\\ 0 & (1\le j\le \nu-1),\\ s_\kappa & (j=\nu), \end{cases}$$

where s_κ is assumed to be a non-zero constant. This defines a kernel of order ν. In practice, the most widely used kernels or the so-called 'standard kernels' are all of order 2.

It is possible to write down the exact expressions for the bias and variance of $\hat{\mu}_{ZD}$ and $\hat{\beta}_{ZD}$ for the model (3). For notational convenience substitute

$$R = X'V^{-1}(I-K) \quad \text{and} \quad W = \{X'V^{-1}(I-K)X\}^{-1} = (RX)^{-1}.$$

From (3) and (6) we obtain

$$\hat{\beta}_{ZD} = WRy = WR(\mu + X\beta + \epsilon) = \beta + WR(\mu+\epsilon).$$

Thus $\hat{\beta}_{ZD}$ has bias

$$\text{bias}(\hat{\beta}_{ZD}) = \text{E}(\hat{\beta}_{ZD}) - \beta = WR\mu \tag{8}$$

and variance

$$\text{var}(\hat{\beta}_{ZD}) = WRVR'W'. \tag{9}$$

As all N observations are used in the computation of the kernel estimate of μ at each point t_{ij}, we do not need to distinguish the measurements as coming from different individuals but can treat them as arising from a single process. Hence we simplify the notation by dropping one of the suffices and order the time points as $\{t_i,\, i=1,\ldots,N\}$. In order to look at the bias and variance of the mean function μ at any point t, we need to further define

$$k(t)' = \big(K_h(t,t_1),\ldots,K_h(t,t_N)\big)$$

and let $\hat{\mu}_0(t)$ denote the estimator of $\mu(t)$ that would be obtained by kernel smoothing if β were known exactly, i.e.

$$\hat{\mu}_0(t) = k(t)'(y - X\beta).$$

Noting that

$$\hat{\mu}_{ZD}(t) = \hat{\mu}_0(t) - k(t)'X(\hat{\beta}_{ZD} - \beta),$$

we have

$$\begin{aligned} \text{bias}[\hat{\mu}_{ZD}(t)] &= \mathrm{E}[\hat{\mu}_{ZD}(t)] - \mu(t) = \text{bias}[\hat{\mu}_0(t)] - k(t)'X\,\text{bias}(\hat{\beta}_{ZD}) \\ &= \{k(t)'\mu - \mu(t)\} - k(t)'XWR\mu \end{aligned} \tag{10}$$

and

$$\begin{aligned} \text{var}[\hat{\mu}_{ZD}(t)] &= \text{var}[\hat{\mu}_0(t)] - 2k(t)'X\,\text{cov}[\hat{\beta}_{ZD}, \hat{\mu}_0(t)] + k(t)'X\,\text{var}(\hat{\beta}_{ZD})X'k(t) \\ &= k(t)'Vk(t) - 2k(t)'XWRVk(t) + k(t)'XWRVR'W'X'k(t). \end{aligned} \tag{11}$$

3.2. Modified Estimator

One of the principal drawbacks of the above method is the bias of $\hat{\beta}_{ZD}$ which can dominate the variance asymptotically if x and t are correlated. We have also found that in certain situations $\text{var}(\hat{\beta}_{ZD})$ does not have the usual $\sqrt{n}$ rate of convergence and tends to behave oddly as the correlation ρ increases. Following Speckman (1988) we therefore consider another estimator of β (and hence of μ) which is closely related to $\hat{\beta}_{ZD}$. This estimator turns out to possess a lower order bias, which is achieved by undersmoothing the estimate of β without affecting the estimate of μ, by adjusting x for the dependence on t.

The modified estimate is computed by generalized least squares regression on the partial residuals

$$\tilde{X} = (I - K)X \qquad \text{and} \qquad \tilde{y} = (I - K)y$$

to obtain the estimator

$$\begin{aligned} \hat{\beta}_{\text{mod}} &= (\tilde{X}'V^{-1}\tilde{X})^{-1}\tilde{X}'V^{-1}\tilde{y} \\ &= \{X'(I-K)'V^{-1}(I-K)X\}^{-1}X'(I-K)'V^{-1}(I-K)y. \end{aligned} \tag{12}$$

The estimate of μ is defined in the same way as before with $\hat{\beta}_{ZD}$ replaced by $\hat{\beta}_{\text{mod}}$ in (4). Thus

$$\hat{\mu}_{\text{mod}} = K(y - X\hat{\beta}_{\text{mod}}). \tag{13}$$

Comparing (6) and (12) we notice that the difference in $\hat{\beta}_{ZD}$ and $\hat{\beta}_{\text{mod}}$ is achieved by replacing X' in (6) by $X'(I-K)'$ in (12). Note that the corresponding change to X is not made. Writing

$$\tilde{R} = X'(I-K)'V^{-1}(I-K)$$

and

$$\widetilde{W} = \{X'(I-K)'V^{-1}(I-K)X\}^{-1} = (\tilde{R}X)^{-1},$$

we find that

$$\text{bias}(\hat{\beta}_{\text{mod}}) = \widetilde{W}\tilde{R}\mu \tag{14}$$

and

$$\text{var}(\hat{\beta}_{\text{mod}}) = \widetilde{W}\tilde{R}V\tilde{R}'\widetilde{W}'. \tag{15}$$

The bias and variance of $\hat{\mu}_{\text{mod}}$ have expressions similar to those of $\hat{\mu}_{ZD}$, but with W and R being replaced by $\widetilde{W}$ and $\tilde{R}$ in equations (10) and (11).

4. Asymptotic Behaviour

4.1. Assumptions

In order to gain general insight into the behaviour of these estimators it is very helpful to carry out some sort of asymptotic analysis. For the asymptotics to work we need $N \to \infty$, which can result if either or both of n and m_i go to infinity. In many practical situations (e.g. Zeger & Diggle, 1994), it is natural as well as easier to include a large number (n) of experimental units and to take measurements on each of them at a small number (m_i) of time points, which would be expected to differ from one unit to another. If we let $n \to \infty$ and keep m_i finite, then our asymptotic theory requires that the units do not have a common set of measurement times. Of course, care should be taken in interpreting an asymptotic analysis in the case of small samples, or when some of the underlying assumptions are suspect, as may be the case in particular applications. Before we can state and prove the theorems, we need to discuss some features of the model and other issues connected with the asymptotic investigation.

Throughout, we give equal importance to the estimation of both μ and β in our semi-parametric setup without treating either as a nuisance parameter. Our present analysis proceeds bearing this in mind. First we consider the non-parametric regression problem of estimating μ in equation (3) when the regression parameter vector β is known, with other conditions remaining unchanged. Recently, Hart & Wehrly (1986), Altman (1990) and Hart (1991) have addressed the kernel smoothing of data with serially correlated errors. Summarizing the results from these papers and from our own study, we may conclude that the order of magnitude of bias($\hat{\mu}_0$) is unaffected by the correlation structure. Irrespective of whether the errors are i.i.d. or serially correlated, the leading term in the bias is

$$\text{bias}(\hat{\mu}_0) = O(h^{\nu}) \tag{16}$$

assuming that the kernel is of order ν and the function μ has at least ν derivatives with $\mu^{(\nu)}$ being bounded and continuous on [0,1]. On the other hand, the correlation structure of the errors can affect the order of magnitude of the variance of $\hat{\mu}$. If observations sufficiently apart are virtually uncorrelated and the sum of the correlations is finite, it can be shown that var($\hat{\mu}$) is larger than in the i.i.d. case, and increases as the correlation increases. However, the variances are essentially of the same order, $O\big((Nh)^{-1}\big)$. In contrast, with a continuous correlation function, e.g. if the data are longitudinal in nature, the order of the variance differs from the above case and is given by

$$\operatorname{var}(\hat{\mu}_0) = O(n^{-1}). \tag{17}$$

It is evident from (2) that V, the covariance matrix of y, can be written as

$$V = \sigma_z^2 I + \sigma_w^2 V_\rho = \sigma^2 I + c(V_\rho - I) \tag{18}$$

where V_ρ is a block diagonal correlation matrix with n blocks. The ith block is an $m_i \times m_i$ matrix with (j,k)th element $v_{ijk} = \rho(t_{ij} - t_{ik}; \theta)$, $\sigma^2 = (\sigma_z^2 + \sigma_w^2)$ and $c = \sigma_w^2/\sigma^2$. We assume that the correlation function ρ is even with $\rho(0) = 1$ and $0 \le \rho(u) \le 1$ for all $u \in [-1,1]$.

In the following we use the matrix spectral norm, $\|A\|$ = modulus of the largest singular value of A. Also, if $x = (x_1, \dots, x_n)' \in \mathcal{R}^n$, then we define $\|x\|^2 = \sum x_i^2$. Throughout, $\operatorname{tr}(A)$ denotes the trace of a matrix A. Although we do not refer to the matrix Euclidean norm $\|A\|_E = \big(\operatorname{tr}(A'A)\big)^{1/2}$ explicitly anywhere we use the well known result that $\|A\| \le \|A\|_E$. We also define $m = \max_i(m_i) < \infty$.

Let $\lambda_{\rho,\max}$ and $\lambda_{\rho,\min}$ be the largest and the smallest eigenvalues of the matrix V_ρ respectively. It follows from equation (18) and our assumptions on $\rho(u)$ above that $0 \le \lambda_{\rho,\min} \le 1$ and $1 \le \lambda_{\rho,\max} \le m$ and hence both $\|V\|$ and $\|V^{-1}\|$ are $O(1)$. We do not need any specific parameterisation of $\rho(u)$ for the theorems and lemmas of the next section to hold.

However, if we do assume that

$$\rho(u;\theta) = \exp(-\theta u^d)$$

for some positive constant d, we have some useful results. Under this model, $\lambda_{\rho,\min}$ is a decreasing function and $\lambda_{\rho,\max}$ is an increasing function of the lag-one correlation $\alpha = \exp(-\theta)$. When $\alpha = 0$, $V_\rho = I$ and, consequently, $\lambda_{\rho,\min} = \lambda_{\rho,\max} = 1$. On the other hand, if $\alpha = 1$, we have $\lambda_{\rho,\min} = 0$ and $\lambda_{\rho,\max} = m$. Denote $f_{1,\rho} = 1 - \lambda_{\rho,\min}$ and $f_{2,\rho} = \lambda_{\rho,\max}/m$. Hence both $f_{1,\rho}$ and $f_{2,\rho}$ are now increasing functions of α with $0 \le f_{1,\rho} \le 1$ and $1/m \le f_{2,\rho} \le 1$. Now, referring to equation (18), the spectral decomposition of V gives

$$\|V\| = O\big(\sigma^2(1 - c + f_{2,\rho} mc)\big) \tag{19}$$

and

$$\|V^{-1}\| = O\Big(\frac{1}{\sigma^2}(1 - f_{1,\rho}c)^{-1}\Big). \tag{20}$$

We henceforth let $x'_{ij} = x(_{ij1}, \dots, x_{ijp})$ in model (1) and for $k = 1, \dots, p$ write $X = \{x_{ijk}\}$. We suppose that the variables x_{ijk} and t are related via the regression model

$$x_{ijk} = g_k(t_{ij}) + \eta_{ijk} \qquad (i = 1, \dots, n;\ j = 1, \dots, m_i;\ k = 1, \dots, p), \tag{21}$$

where $g_k(t)$ are smooth functions with ν continuous derivatives and η_{ijk} are independent mean zero random variables independent of w_i and z_{ij}. We may suppose that $(\eta_{i1},\ldots,\eta_{ip})$ $(i=1,\ldots,N)$ are independent identically distributed p-dimensional random vectors with mean zero and finite variance-covariance matrix. The model (21) is a very general one and is intended to mimic a range of practical situations including the case $g_k(t)=0$, i.e. x and t are uncorrelated, as well as cases in which x variables show a positive association with time. We also look at the case when $x_{ijk}=g_{ki}$ for all $j=1,\ldots,m_i$, a situation which arises if x is independent of t. Although of practical importance, the latter case is not amenable to a precise asymptotic analysis like the other cases covered by (21). For this case we report here only the findings based on a simulation study. In line with the notation used to express model (3) we may write (21) as

$$X = g + \eta, \tag{22}$$

where the kth column of g is $g_k = \big(g_k(t_1),\ldots,g_k(t_N)\big)'$ and the kth column of η is $\eta_k = (\eta_{1k},\ldots,\eta_{Nk})'$.

It follows immediately from the properties of kernel smoothing (assuming $N\to\infty$, $h\to 0$ and $Nh\to\infty$) that

(a) $\operatorname{tr}(K_h'K_h) = O(h^{-1})$,
(b) $\|(I-K_h)\mu\| = O(N^{1/2}h^{\nu})$,
(c) $\|(I-K_h)g_k\| = O(N^{1/2}h^{\nu})$.

From the assumptions on η_k following equations (21) and (22), we also have

(d) $\|A\eta_k\|^2 = O\big(\operatorname{tr}(A'A)\big)$, for any matrix A,
(e) $a'\eta_k = O(\|a\|)$, for any vector $a\in\mathcal{R}^N$,
(f) $N^{-1}\eta'V^{-1}\eta = \Sigma_V = O\big((1-f_{1,\rho}c)^{-1}\big)$.

Assumption (f) follows from (20) and holds in probability. Note that when we say $N\to\infty$ we mean $n\to\infty$ with m being always finite. Also, in the remainder of the paper we make the dependence of K on h implicit and simply write K instead of K_h. Finally, for our theorems to be valid we need to assume that t has an asymptotic density $p(t)$ on [0,1].

4.2. Theorems

The first two theorems are on the ZD estimators, while Theorems 3 and 4 relate to the modified estimators. All the results hold under the conditions set out in Section 4.1. Although we do not state this explicitly at the beginning of each theorem or lemma, it is understood to be applicable in each case.

Lemma 1. *If $n\to\infty$, $h\to 0$ and $nh\to\infty$, then $N^{-1}X'V^{-1}(I-K)X\to\Sigma_V$.*

Theorem 1. *If $n\to\infty$ and $h\to 0$, then*

$$\mathrm{E}(\hat\beta_{ZD})-\beta = \Sigma_V^{-1}\|N^{-1/2}V^{-1}g\|O(h^{\nu}) + O(N^{-1/2}h^{\nu}) \tag{23}$$

and if $\nu=2$, then

$$\operatorname{var}(\hat\beta_{ZD}) = N^{-1}\Sigma_V^{-1}+N^{-2}\Sigma_V^{-1}g'V^{-1}(I-K)V(I-K)'V^{-1}g\Sigma_V^{-1}+o(N^{-1}). \tag{24}$$

The results of Theorem 1 demonstrate that the first term of the bias dominates if g is non-vanishing. We can also show that the bias is $O(h^\nu)$ and is not affected by the correlation structure. Equation (20) and assumption (f) imply that $O(\Sigma_V^{-1})O(V^{-1}) \approx 1$, and since $\|N^{-1/2}g\| = O(1)$ it follows that $\text{bias}(\hat\beta_{ZD}) = O(h^\nu)$. If $g = 0$, the bias is given by the second term of equation (23) and is $O(N^{-1/2}h^\nu) = o(1)$.

Equation (24) shows that $N^{-2}\Sigma_V^{-1}g'V^{-1}(I-K)V(I-K)'V^{-1}g\Sigma_V^{-1}$ can dominate $\text{var}(\hat\beta_{ZD})$ unless it has the same order as the first term. However, it can be shown that the term $N^{-1}g'V^{-1}(I-K)V(I-K)'V^{-1}g$ is $O(1)$ if $\nu = 2$ and so $\text{var}(\hat\beta_{ZD}) = O(1/N)$. By assumption (f), the term $\Sigma_V^{-1} = O(1 - f_{1,\rho}c)$ and clearly this is a decreasing function of α. So, if $g = 0$, the second term is zero and the variance will be a decreasing function of the correlation. However, if g is non-vanishing then the second term can dominate the first one, e.g. if g has a large signal to noise ratio. In that case $\text{var}(\hat\beta_{ZD})$ may not exhibit a fixed pattern as α varies. However, since $g'V^{-1}(I-K)V(I-K)'V^{-1}g$ is an increasing function of the correlation and if this term is dominating, the most likely scenario is that even if there is an initial fall in the value of $\text{var}(\hat\beta_{ZD})$, it will eventually start to rise as α becomes moderate to large.

Theorem 2. *If $n \to \infty$ and $h \to 0$, then*

$$\text{bias}[\hat\mu_{ZD}(t)] = \text{bias}[\hat\mu_0(t)] + k(t)'X\,\text{bias}(\hat\beta_{ZD}) = O(h^\nu) \tag{25}$$

and when $\nu = 2$,

$$\text{var}[\hat\mu_{ZD}(t)] = \text{var}[\hat\mu_0(t)][1 + o(1)] = \max\big(O((Nh)^{-1}), O(n^{-1})\big). \tag{26}$$

The following lemma and the theorems relate to the modified estimators.

Lemma 2. *If $n \to \infty$, $h \to 0$ and $nh \to \infty$, then $N^{-1}\tilde X'V^{-1}\tilde X \to \Sigma_V$.*

Before stating Theorem 3, we need a more precise statement about the bias of a kernel estimator as shown in assumption (c) by specifying the leading term of the bias,

$$(I-K)g_k = h^\nu a(t)g_k^{(\nu)} + o(h^\nu) \tag{27}$$

where $a(t)$ is some bounded function and $g_k^{(\nu)}$ is the νth derivative of g_k.

Theorem 3. *If $n \to \infty$ and $h \to 0$, then*

$$\text{E}(\hat\beta_{\text{mod}}) - \beta = \Sigma_V^{-1}\|N^{-1/2}V^{-1}g^{(\nu)}\|O(h^{2\nu}) + O\big(h^\nu(Nh)^{-1/2}\big). \tag{28}$$

In addition, if $nh^2 \to \infty$ and $\nu = 2$, then

$$\text{var}(\hat\beta_{\text{mod}}) = N^{-1}\Sigma_V^{-1} + o(N^{-1}). \tag{29}$$

For a reason similar to the one given in Theorem 1, we see that bias$(\hat{\beta}_{\text{mod}}) = O(h^{2\nu})$ and is not affected by the correlation structure to any significant degree. If g vanishes, bias$(\hat{\beta}_{\text{mod}})$ reduces to $O\big(h^{\nu}(Nh)^{-1/2}\big) = o(1)$. Equation (29) shows that the asymptotic var$(\hat{\beta}_{\text{mod}})$ does not involve g. It is $O(N^{-1})$ and goes down with decreasing correlation because Σ_V^{-1} is a decreasing function of α, as noted in our discussion following Theorem 1.

Theorem 4. *If $n \to \infty$ and $h \to 0$, then*

$$\text{bias}[\hat{\mu}_{\text{mod}}(t)] = \text{bias}[\hat{\mu}_0(t)][1 + o(1)] = O(h^{\nu}). \tag{30}$$

If in addition, $nh^2 \to \infty$ and $\nu = 2$, then

$$\text{var}[\hat{\mu}_{\text{mod}}(t)] = \text{var}[\hat{\mu}_0(t)][1 + o(1)] = \max\big(O\big((Nh)^{-1}\big), O(n^{-1})\big). \tag{31}$$

It is evident from Theorems 2 and 4 that asymptotically both $\hat{\mu}_{ZD}$ and $\hat{\mu}_{\text{mod}}$ perform almost equally well, although bias$[\hat{\mu}_{\text{mod}}(t)]$ could be slightly lower than bias$(\hat{\mu}_{ZD})$. However, both Theorems 1 and 3 tell us that in an asymptotic sense $\hat{\beta}_{\text{mod}}$ fares better than $\hat{\beta}_{ZD}$ both in terms of the bias and the variance. There is a marked improvement in the bias of the modified estimator and in both the cases the variance is asymptotically of the same order. However, whereas var$(\hat{\beta}_{\text{mod}})$ is a decreasing function of the correlation, this is not so in general for var$(\hat{\beta}_{ZD})$.

4.3. Numerical Results

In order to get a clearer picture of the results of the preceding section we present here a numerical study of the bias and variance of the ZD and of the modified estimators. Our objective is to show what the asymptotic results actually mean in practice and how well they hold for relatively small sample sizes.

We report here some numerical calculations based on a sample size of $n = 50$ with $m_i = m = 10$ measurements on each unit. For the sake of simplicity, we included only one covariate, $x(t) = 5t^2 + \eta$, where $\eta \sim N(0, 0.5)$. The mean function was $\mu(t) = \cos(\pi t)$ for $t \in [0,1]$. The kernel used in the study was the quartic kernel $w(s) = (15/16)(1-s^2)^2 I(|s| \leq 1)$. For convenience we do not use a boundary-corrected kernel. It is expected that the conclusions of our simulations as regards the relative comparison of the two methods would not be changed by using a kernel modified at the boundaries. However, in applications we would strongly recommend the use of boundary-corrected kernels to reduce the bias in small samples. We assumed the covariance function to be $\gamma(u) = \sigma_w^2 \exp(-\theta|u|)$, with $\sigma_w^2 = 0.4$. Also, we fixed the variance of the measurement errors at $\sigma_z^2 = 0.1$. To look at the effect of various amounts of correlation, we varied $\alpha = \exp(-\theta)$ from 0 to 1 at steps of 0.1. For this study we simulated measurement times t as independent random samples from $U(0,1)$, and $x(t)$ as independent random samples from $N(5t^2, 0.5)$. Finally, we decided to take $h = 0.15$, a value that seemed to give the smallest mean-square error when averaged across all values of α. In fact, the optimum value of h varied very little with α.

TABLE 1

Bias and variance of ZD estimators $\hat{\mu}_{ZD}$ and $\hat{\beta}_{ZD}$ when x is correlated with t

$e^{-\theta}$	ISB	IVAR	MISE	bias($\hat{\beta}_{ZD}$)	var($\hat{\beta}_{ZD}$)
0.0	0.005891	0.013280	0.019171	−0.030261	0.001606
0.1	0.000659	0.015585	0.016243	−0.006619	0.000839
0.2	0.000657	0.015583	0.016241	−0.006605	0.000786
0.3	0.000687	0.015676	0.016363	−0.006897	0.000765
0.4	0.000742	0.015852	0.016594	−0.007406	0.000763
0.5	0.000828	0.016140	0.016968	−0.008160	0.000781
0.6	0.000962	0.016607	0.017569	−0.009250	0.000828
0.7	0.001184	0.017406	0.018590	−0.010878	0.000928
0.8	0.001596	0.018927	0.020523	−0.013487	0.001147
0.9	0.002531	0.022460	0.024991	−0.018269	0.001705
1.0	0.005776	0.035075	0.040851	−0.029922	0.003875

TABLE 2

Bias and variance of modified estimators $\hat{\mu}_{\mathrm{mod}}$ and $\hat{\beta}_{\mathrm{mod}}$ when x is correlated with t

$e^{-\theta}$	ISB	IVAR	MISE	bias($\hat{\beta}_{\mathrm{mod}}$)	var($\hat{\beta}_{\mathrm{mod}}$)
0.0	0.000256	0.014837	0.015092	−0.001197	0.001899
0.1	0.000211	0.014595	0.014806	−0.000124	0.000758
0.2	0.000211	0.014472	0.014683	−0.000116	0.000687
0.3	0.000211	0.014382	0.014593	−0.000120	0.000640
0.4	0.000211	0.014305	0.014517	−0.000134	0.000603
0.5	0.000212	0.014235	0.014447	−0.000156	0.000571
0.6	0.000213	0.014166	0.014380	−0.000191	0.000543
0.7	0.000215	0.014098	0.014313	−0.000243	0.000516
0.8	0.000219	0.014026	0.014244	−0.000326	0.000489
0.9	0.000224	0.013946	0.014170	−0.000474	0.000460
1.0	0.000239	0.013845	0.014083	−0.000818	0.000423

Table 1 refers to the ZD estimators with columns 2 and 3 showing the integrated squared bias (ISB) and the integrated variance (IVAR) of $\hat{\mu}_{ZD}$ as a function of α. Columns 4 and 5 display bias($\hat{\beta}_{ZD}$) and var($\hat{\beta}_{ZD}$) respectively. Table 2 shows the same for the modified estimators. We summarize the results as follows:

1. bias($\hat{\mu}_{\mathrm{mod}}$) remains relatively constant as α varies, whereas bias($\hat{\mu}_{ZD}$) appears to decrease and then increase with increasing α. The integrated squared bias is consistently smaller for $\hat{\mu}_{\mathrm{mod}}$ than for $\hat{\mu}_{ZD}$.
2. Both var($\hat{\mu}_{\mathrm{mod}}$) and var($\hat{\mu}_{ZD}$) change very little with correlation once α becomes positive.
3. bias($\hat{\beta}_{\mathrm{mod}}$) is substantially lower than bias($\hat{\beta}_{ZD}$). Neither is affected by correlation to any significant effect.
4. As expected, var($\hat{\beta}_{\mathrm{mod}}$) falls as α increases, whereas var($\hat{\beta}_{ZD}$) decreases at

TABLE 3

Bias and variance of ZD estimators $\hat{\mu}_{ZD}$ and $\hat{\beta}_{ZD}$ when x is a constant

$e^{-\theta}$	ISB	IVAR	MISE	bias($\hat{\beta}_{ZD}$)	var($\hat{\beta}_{ZD}$)
0.0	0.000216	0.009843	0.010059	−0.004923	0.017387
0.1	0.000208	0.035428	0.035636	−0.001465	0.089005
0.2	0.000207	0.038442	0.038648	−0.000311	0.100352
0.3	0.000206	0.040586	0.040792	0.000330	0.108675
0.4	0.000206	0.042291	0.042498	0.000642	0.115413
0.5	0.000206	0.043699	0.043906	0.000664	0.121051
0.6	0.000206	0.044872	0.045079	0.000378	0.125814
0.7	0.000207	0.045846	0.046053	−0.000271	0.129782
0.8	0.000208	0.046659	0.046053	−0.001370	0.133086
0.9	0.000211	0.047397	0.047608	−0.003011	0.136003
1.0	0.000216	0.048332	0.048548	−0.004923	0.139457

TABLE 4

Bias and variance of modified estimators $\hat{\mu}_{\text{mod}}$ and $\hat{\beta}_{\text{mod}}$ when x is a constant

$e^{-\theta}$	ISB	IVAR	MISE	bias($\hat{\beta}_{\text{mod}}$)	var($\hat{\beta}_{\text{mod}}$)
0.0	0.000207	0.009808	0.010015	0.000040	0.017358
0.1	0.000206	0.031114	0.031321	0.000526	0.078016
0.2	0.000206	0.035054	0.035260	0.000951	0.092008
0.3	0.000206	0.038215	0.038421	0.001687	0.103390
0.4	0.000207	0.041066	0.041273	0.002819	0.113768
0.5	0.000209	0.043785	0.043995	0.004493	0.123739
0.6	0.000215	0.046478	0.046694	0.006961	0.133675
0.7	0.000231	0.049241	0.049472	0.010685	0.143953
0.8	0.000270	0.052217	0.052487	0.016640	0.155094
0.9	0.000388	0.055748	0.056136	0.027346	0.168419
1.0	0.000908	0.061440	0.062348	0.052691	0.190002

first and then rises at an accelerating rate. The latter behaviour is due to the large signal-to-noise ratio of $g(t)$.

The above findings are consistent with the asymptotic properties established in Section 4.2.

In Tables 3 and 4 we present calculations for the case when the covariate x is constant across time t for each of the n units with x varying only from unit to unit. Our experiment is exactly the same as before with the exception that now $x_{ij} = g_i$, a random constant. Tables 3 and 4 refer to the ZD and the modified estimators respectively. These results show that the performances of both the estimators are very similar when the correlation is low, but that the ZD procedure gives better results as the correlation increases. The most worrying feature of these results is that the order of magnitude of var($\hat{\beta}$) appears to be much greater than N^{-1} and increases as the correlation increases. One plausible

TABLE 5

Effect on the MISE, bias and variance of $\hat\beta$ when h is chosen by a data-based cross-validation method

$e^{-\theta}$	MISE		bias($\hat\beta$)		var($\hat\beta$)	
	$\hat h_{\text{opt}}$	h_{opt}	$\hat h_{\text{opt}}$	h_{opt}	$\hat h_{\text{opt}}$	h_{opt}
0.0	0.035241	0.017788	−0.058510	−0.020028	0.001932	0.001723
0.1	0.016296	0.015793	−0.013639	−0.010720	0.000873	0.000795
0.5	0.018141	0.016725	−0.017433	−0.011815	0.000810	0.000739
0.9	0.034068	0.024991	−0.039227	−0.018269	0.001759	0.001705
1.0	0.059861	0.040731	−0.061221	−0.026311	0.003851	0.004011

explanation is that the term $X\beta$ in model (3) gets confounded with $\mu(t)$, because x is constant. This results in a rather poor estimate of β. A large value for the variance is indicative of this, although the bias tends to mask the effect. By analysing individual sets of simulated data, we have seen that, in any single case, the apparent bias of $\hat\beta$ can be quite large but may take either a positive or a negative value, resulting in a relatively small true bias.

Finally, we carried out a small scale simulation study of the behaviour of the estimators when the bandwidth parameter h was chosen from the data using the cross-validation method of Rice & Silverman (1991) and adopted by Zeger & Diggle (1994) for analysing the data on CD4 cell numbers. The idea is to choose an h that minimizes the mean integrated (average) squared error $\mathrm{E}\left[\sum_{ij}\{\hat\mu(t_{ij} - \mu(t_{ij})\}^2\right]$. We call this bandwidth h_{opt}. The cross-validation method of Rice–Silverman when applied here would result in selecting an $h = \hat h_{\text{opt}}$ which minimizes

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{y_{ij} - x_{ij}'\beta - \hat\mu^{(i)}(t_{ij})\}^2,$$

where $\hat\mu^{(i)}(t_{ij})$ is the kernel estimate of $(y_{ij} - x_{ij}'\beta)$ obtained by leaving out the complete vector of observations of the ith unit. As β is unknown we would substitute $\beta = \hat\beta^{[k]}$ at the kth iteration of the backfitting algorithm suggested by Zeger & Diggle (1994) for estimating $\hat\mu_{ZD}$ and $\hat\beta_{ZD}$.

Again with exactly the same set up as described for Tables 1 and 2, we simulated 100 data sets and estimated $\hat h_{\text{opt}}$ in each case. Table 5 shows the average MISE, bias($\hat\beta$) and var($\hat\beta$) as attained by $\hat h_{\text{opt}}$ across these 100 samples along with their true values when $h = h_{\text{opt}}$ for five different values of the correlation. The method performs quite satisfactorily when the correlation is small to moderate. As the correlation approaches 1, the true MISE and bias($\hat\beta$) differ considerably from the estimated ones. When the correlation is exactly zero, the method does not seem to do well either. In this case, as all observations are uncorrelated, it would be more efficient to use the standard cross-validation score constructed by leaving out individual data points rather than whole vectors of observations.

5. Discussion

Our main objective in this paper has been to provide a deeper understanding of the heuristic procedures proposed by Zeger & Diggle (1994) for semi-parametric regression modelling of longitudinal data. To achieve this we have established both exact and asymptotic results for the bias and variance of the ZD estimators, and have demonstrated by explicit calculation of some special cases that the asymptotic results give reliable indications of how well the estimators perform in finite samples.

Our results assume that the non-parametric part of the regression estimation is carried out using a fixed-bin-width kernel, whereas Zeger & Diggle used a variable-bin-width kernel. The latter is generally preferable for smoothing data observed at an irregular set of time points (see, for example, Silverman, 1984), but is less amenable to theoretical analysis. The practical differences between the two methods are small when the density of observation times is approximately constant over the time-interval of interest, as is the case in our numerical examples.

We have also considered the behaviour of a modified estimation procedure. The modified estimator for β makes explicit adjustments for the dependence on t of both the response variable and the explanatory variables, by working with $\tilde{y} = (I - K)y$ and $\tilde{X} = (I - K)X$ in place of y and X. The resulting estimator, $\hat{\beta}_{\text{mod}}$, achieves a lower order of bias than $\hat{\beta}_{ZD}$ without having to undersmooth for the estimation of μ, provided that X is correlated with t. When all of the explanatory variables are constant over time, this consideration is irrelevant, and indeed our results confirm that in this case the ZD estimator has the better performance. As defined, both the modified estimators and the non-iterative form of the ZD estimators involve multiplication of N by N matrices. However, in most applications it is sufficient to compute $\hat{\mu}(t)$ for a set of $N_0 \ll N$ values of t chosen to span the full range of t, thereby reducing the required matrix multiplications to order N_0 by N.

One issue which we have addressed only briefly is the data-based selection of a value for the kernel band-width, h. Our simulations suggest that the Rice & Silverman prescription leads to only a very slight drop in performance, relative to the theoretically optimal choice of h, when the correlation is moderate, but a substantial drop at the two extremes of high and low correlation.

Appendix: Proofs of Theorems

Proof of Lemma 1. Using equation (22), we can write

$$\begin{aligned} &N^{-1}X'V^{-1}(I-K)X \\ &= N^{-1}\big(g'V^{-1}(I-K)g + g'V^{-1}(I-K)\eta + \eta'V^{-1}(I-K)g + \eta'V^{-1}(I-K)\eta\big). \end{aligned}$$

We now show that all the terms except the last one tend to zero. Since p is finite, it follows from equation (20) and assumption (c)

$$\begin{aligned}N^{-1}\|g'V^{-1}(I-K)g\| &\le N^{-1}\|g\|\,\|V^{-1}\|\,\|(I-K)g\| \\ &= N^{-1}O(N^{1/2})O(1)O(N^{1/2}h^{\nu}) = O(h^{\nu}) = o(1).\end{aligned}$$

From equation (20) and assumptions (c) and (e), we have

$$\begin{aligned}N^{-1}\|\eta'V^{-1}(I-K)g\| &= N^{-1}O\big(\|V^{-1}(I-K)g\|\big) \\ &\le N^{-1}O\big(\|V^{-1}\|\,\|(I-K)g\|\big) \\ &= N^{-1}O(N^{1/2}h^{\nu}) = O(N^{-1/2}h^{\nu}) = o(1).\end{aligned}$$

Similarly,

$$\begin{aligned}N^{-1}\|g'V^{-1}(I-K)\eta\| &= N^{-1}O\big(\|g'V^{-1}(I-K)\|\big) \\ &= N^{-1}O(N^{1/2}h^{-1/2}) = O\big((Nh)^{-1/2}\big) = o(1).\end{aligned}$$

Finally, assumptions (a), (d) and (f) give

$$\begin{aligned}N^{-1}\eta'V^{-1}(I-K)\eta &= N^{-1}(\eta'V^{-1}\eta - \eta'V^{-1}K\eta) \\ &= \Sigma_V + N^{-1}O(N^{1/2})O(h^{-1/2}) \\ &= \Sigma_V + O(N^{-1/2}h^{-1/2}) = \Sigma_V + o(1)\end{aligned}$$

and the proof is complete.

Proof of Theorem 1. From equation (8) we have

$$\operatorname{bias}(\hat\beta_{ZD}) = \big(X'V^{-1}(I-K)X\big)^{-1}X'V^{-1}(I-K)\mu.$$

It follows from Lemma 1 that

$$\big(X'V^{-1}(I-K)X\big)^{-1} \to N^{-1}\Sigma_V^{-1}.$$

Now consider

$$N^{-1}X'V^{-1}(I-K)\mu = N^{-1}\big(g'V^{-1}(I-K)\mu + \eta'V^{-1}(I-K)\mu\big).$$

The order of the first term is

$$\begin{aligned}N^{-1}\|g'V^{-1}(I-K)\mu\| &\le N^{-1}\|V^{-1}g\|\,\|(I-K)\mu\| \\ &= N^{-1}\|V^{-1}g\|O(N^{1/2}h^{\nu}) = \|N^{-1/2}V^{-1}g\|O(h^{\nu})\end{aligned}$$

by assumption (b). The order of the second is

$$N^{-1}\|\eta'V^{-1}(I-K)\mu\| \le N^{-1}O\big(\|V^{-1}(I-K)\mu\|\big) \le O(N^{-1/2}h^{\nu}) = o(1)$$

by assumption (b) and equation (20).

To prove equation (24) we refer to equation (9) and consider the term RVR'. Put $Q = V^{-1}(I-K)V(I-K)'V^{-1}$ and using equation (22) write

$$RVR' = X'QX = g'Qg + g'Q\eta + \eta'Qg + \eta'Q\eta.$$

Following assumptions (a), (d), (f) and the fact that $\|K'V^{-1}\eta\| = O(h^{-1/2})$, we have

$$\begin{aligned} N^{-1}\eta'Q\eta &= N^{-1}(\eta'V^{-1}\eta - \eta'V^{-1}K\eta - \eta'K'V^{-1}\eta + \eta'V^{-1}KVK'V^{-1}\eta) \\ &= \Sigma_V + O(N^{-1/2}h^{-1/2}) + O(N^{-1}h^{-1}) \\ &= \Sigma_V + o(1). \end{aligned}$$

Just as above, the order of $g'Q\eta$ or of $\eta'Qg$ is given by

$$\begin{aligned} N^{-1}\|g'Q\eta\| &\le N^{-1}\big(\|g'(I-K)'V^{-1}\eta\| + \|g'V^{-1}K\eta\| + \|g'V^{-1}KVK'V^{-1}\eta\|\big) \\ &= O(N^{-1/2}h^{\nu}) + O(N^{-1/2}) + O(N^{-1/2}) = o(1). \end{aligned}$$

This follows from the fact that for second order kernels, i.e. when $\nu = 2$, both $\|g'V^{-1}K\|$ and $\|g'V^{-1}KVK'\|$ are $O(N^{-1/2})$. However, for higher order kernels this is not true in general and for any ν other than 2 we would need to impose further conditions to achieve an order of $o(1)$ for $N^{-1}g'Q\eta$. Now, the term $N^{-1}g'Qg$ can be shown to be of $O(1)$ and, therefore, does not vanish as $N \to \infty$. Hence,

$$N^{-1}RVR' = \Sigma_V + N^{-1}g'Qg + o(1)$$

and using Lemma 1 once more, we have

$$\begin{aligned} \operatorname{var}(\hat\beta_{ZD}) &= N^{-1}\Sigma_V^{-1}\big(\Sigma_V + N^{-1}g'Qg + o(1)\big)\Sigma_V^{-1} \\ &= N^{-1}\Sigma_V^{-1} + N^{-2}\Sigma_V^{-1}g'Qg\Sigma_V^{-1} + o(N^{-1}). \end{aligned}$$

Proof of Theorem 2. Definition of $k(t)$, equation (7) and assumptions (c) and (e) imply that $\|k(t)'X\| = \|k(t)'g + k(t)'\eta\| = O\big(g(t)\big) = O(1)$. It now follows at once from equations (10), (16) and (23) that bias $\big(\hat\mu_{ZD}(t)\big) = O(h^{\nu})$. Like bias $\big(\hat\mu_0(t)\big)$ or bias$(\hat\beta_{ZD})$, bias $\big(\hat\mu_{ZD}(t)\big)$ is not affected by serial correlation. The order of the bias also does not change if g vanishes.

We see from equation (11) that the variance of $\hat\mu$ is the sum of three terms. Since $\operatorname{var}(\hat\beta_{ZD}) = O(N^{-1})$ and $\|k(t)'X\| = O(1)$ as noted in the preceding paragraph,

$$\|k(t)'X \operatorname{var}(\hat\beta_{ZD})X'k(t)\| = O(N^{-1}).$$

To find the order of the term involving cov $\big(\hat\beta_{ZD}, \hat\mu_0(t)\big)$, consider

$$\begin{aligned} \|RVk(t)\| &= \|X'V^{-1}(I-K)Vk(t)\| \\ &\le \|g'V^{-1}(I-K)Vk(t)\| + \|\eta'V^{-1}(I-K)Vk(t)\| \\ &= O(1) + o(1) \end{aligned}$$

and thus by Lemma 1, cov $\left(\hat{\beta}_{ZD}, \hat{\mu}_0(t)\right) = O(N^{-1})$. Finally, using equation (18) we can write

$$\begin{aligned} k(t)'Vk(t) &= \sigma^2 k(t)'k(t) + ck(t)'(V_\rho - I)k(t) \\ &= O\big((Nh)^{-1}\big) + O(n^{-1}) \end{aligned}$$

by equation (17) and assumption (a). Now, if the correlation $\rho = 0$, $k(t)'Vk(t) = O\big((Nh)^{-1}\big)$ and as soon as ρ becomes positive $k(t)'Vk(t) = O\big((Nh)^{-1}\big) + O(n^{-1})$ and the value of mh determines which of the terms is dominant.

As the proofs of Lemma 2 and Theorems 3 and 4 are more or less similar to those given above, we omit repetitions and only provide those aspects of the proof which are new or which yield a different result from the theorems on the ZD estimators.

Proof of Lemma 2. The proof is similar to that of Lemma 1.

Proof of Theorem 3. From equation (14) we have

$$\operatorname{bias}(\hat{\beta}_{\mathrm{mod}}) = (\tilde{X}'V^{-1}\tilde{X})^{-1}\tilde{X}'V^{-1}\tilde{\mu},$$

where $\tilde{\mu} = (I-K)\mu$. Writing

$$\tilde{g} = (I-K)g \qquad \text{and} \qquad \tilde{\eta} = (I-K)\eta,$$

we note that

$$N^{-1}\tilde{X}'V^{-1}\tilde{\mu} = N^{-1}\tilde{g}'V^{-1}\tilde{\mu} + N^{-1}\tilde{\eta}'V^{-1}\tilde{\mu}.$$

By applying assumption (b) and equation (27) we get

$$\begin{aligned} N^{-1}\|\tilde{g}'V^{-1}\tilde{\mu}\| &\le N^{-1}O(h^\nu)\|V^{-1}g^{(\nu)}\|O(N^{1/2}h^\nu) \\ &= \|N^{-1/2}V^{-1}g^{(\nu)}\|O(h^{2\nu}). \end{aligned}$$

Next we observe that

$$\begin{aligned} N^{-1}\|\tilde{\eta}'V^{-1}\tilde{\mu}\| &\le N^{-1}(\|\eta'V^{-1}\tilde{\mu}\| + \|\eta'K'V^{-1}\tilde{\mu}\|) \\ &\le O(N^{-1/2}h^\nu) + O\big(h^\nu(Nh)^{-1/2}\big). \end{aligned}$$

As $O\big(h^\nu(Nh)^{-1/2}\big) > O(N^{-1/2}h^\nu)$, equation (28) follows from Lemma 2 and the results obtained above.

To prove equation (29) we proceed along the same lines as in Theorem 1. Write

$$\tilde{R}V\tilde{R}' = \tilde{X}'Q\tilde{X} = \tilde{g}'Q\tilde{g} + \tilde{g}'Q\tilde{\eta} + \tilde{\eta}'Q\tilde{g} + \tilde{\eta}'Q\tilde{\eta}.$$

As before we can show that

$$N^{-1}\tilde{\eta}'Q\tilde{\eta} = \Sigma_V + o(1)$$

provided $nh^2 \to \infty$. Similarly,

$$N^{-1}\|\tilde{g}'Q\tilde{\eta}\| \le N^{-1/2}O(h^{\nu-3/2}) = o(1).$$

Here in contrast to the term $N^{-1}g'Qg$ in Theorem 1, we can show that $N^{-1}\tilde{g}'Q\tilde{g}$ has $o(1)$. From assumption (c) and noting that $\|V\| = O(1)$, $\|V^{-1}\| = O(1)$, $\|K\| = O(h^{-1/2})$ and $\|Q\| = O(h^{-1})$, we have

$$\begin{aligned} N^{-1}\|\tilde{g}'Q\tilde{g}\| &\le N^{-1}\|\tilde{g}\|^2\|Q\| \\ &= N^{-1}O(Nh^{2\nu})O(h^{-1}) = O(h^{2\nu-1}) = o(1). \end{aligned}$$

Hence,

$$N^{-1}\tilde{R}V\tilde{R}' = \Sigma_V + o(1)$$

and thus

$$\text{var}(\hat{\beta}_{\text{mod}}) = (\tilde{X}'V^{-1}\tilde{X})^{-1}\tilde{R}V\tilde{R}'(\tilde{X}'V^{-1}\tilde{X})^{-1} = N^{-1}\Sigma_V^{-1} + o(N^{-1})$$

by Lemma 2.

Proof of Theorem 4. Same as in Theorem 2.

References

ALTMAN, N.S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.* **85**, 749–759.

CUZICK, J. (1992). Semiparametric additive regression. *J. Roy. Statist. Soc. Ser. B* **54**, 831–843.

DIGGLE, P.J. (1988). An approach to the analysis of repeated measurements. *Biometrics* **44**, 959–971.

— & HUTCHINSON, M.F. (1989). On spline smoothing with autocorrelated errors. *Austral. J. Statist.* **31**, 166–182.

GREEN, P.J. (1987). Penalized likelihood for general semi-parametric regression models. *Internat. Statist. Rev.* **55**, 245–259.

—, JENNISON, C. & SEHEULT, A.H. (1985). Analysis of field experiments by least squares smoothing. *J. Roy. Statist. Soc. Ser. B* **47**, 299–315.

HART, J.D. (1991). Kernel regression estimation with time series errors. *J. Roy. Statist. Soc. Ser. B* **53**, 173–187.

— & WEHRLY, T.E. (1986). Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.* **81**, 1080–1088.

HASTIE, T.J. & TIBSHIRANI, R.J. (1990). *Generalized Additive Models.* London: Chapman and Hall.

HECKMAN, N.E. (1986). Spline smoothing in a partly linear model. *J. Roy. Statist. Soc. Ser. B* **48**, 244–248.

JONES, R.H. (1987). Serial correlation in unbalanced mixed models. *Bull. Internat. Statist. Inst.* **46**, 105–122.

— & ACKERSON, L.M. (1990). Serial correlation in unequally spaced longitudinal data. *Biometrika* **77**, 721–731.

— & BOADI-BOTENG, F. (1991). Unequally spaced longitudinal data with serial correlation. *Biometrics* **47**, 161–175.

KENWARD, M.G. (1987). A method for comparing profiles of repeated measurements. *Appl. Statist.* **36**, 296–308.

MUNOZ, A., CAREY, V. SCHOUTEN, J.P., SEGAL, M. & ROSNER, B. (1992). A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics* **48**, 733–742.

PANTULA, S.G. & POLLOCK, K.H. (1985). Nested analysis of variance with autocorrelated errors. *Biometrics* **41**, 909–920.

RICE, J. (1986). Convergence rates for partially splined models. *Statist. Probab. Lett.* **4**, 203–208.

— & SILVERMAN, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53**, 233–243.

SILVERMAN, B.W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.* **12**, 898–916.

SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50**, 413–436.

VERBYLA, A.P. & CULLIS, B.R. (1990). Modelling in repeated measures experiments. *Appl. Statist.* **39**, 341–356.

WARE, J.H. (1985). Linear models for the analysis of longitudinal studies. *Amer. Statist.* **39**, 95–101.

ZEGER, S.L. & DIGGLE, P.J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**. To appear.