Biometrika (2011), *pp.* 1–16
doi: 10.1093/biomet/asq080

Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation

BY YEHUA LI

Department of Statistics, University of Georgia, Athens, Georgia 30602, U.S.A.
yehuali@uga.edu

SUMMARY

For longitudinal data, when the within-subject covariance is misspecified, the semiparametric regression estimator may be inefficient. We propose a method that combines the efficient semiparametric estimator with nonparametric covariance estimation, and is robust against misspecification of covariance models. We show that kernel covariance estimation provides uniformly consistent estimators for the within-subject covariance matrices, and the semiparametric profile estimator with substituted nonparametric covariance is still semiparametrically efficient. The finite sample performance of the proposed estimator is illustrated by simulation. In an application to CD4 count data from an AIDS clinical trial, we extend the proposed method to a functional analysis of the covariance model.

Some key words: Covariance estimation; Functional data analysis; Generalized estimating equation; Kernel method; Partially linear model; Profile estimator.

1. INTRODUCTION

Semiparametric regression problems in longitudinal clustered data have received much attention recently (Zeger & Diggle, 1994; Lin & Carroll, 2001; Diggle et al., 2002). An important lesson from the literature is that, in order to obtain efficient estimators, one needs to take account of the within-cluster correlation in the correct way (Wang, 2003; Wang et al., 2005). Wang et al. (2005) showed that, in the semiparametric setting, the estimator for the parametric component in the model will achieve the semiparametric efficiency bound only if the within-cluster correlation matrix is specified correctly. Lin & Carroll (2006) generalized these methods to more general repeated measures problems.

There has been a vast volume of work on modelling covariance matrices in longitudinal data. Some recent work includes Wu & Pourahmadi (2003) and Huang et al. (2007). However, most of these methods assume that the observations are made on a regular time grid, and therefore are not suitable for longitudinal data collected at irregular and subject-specific times. More recently, Fan et al. (2007) and Fan & Wu (2008) proposed a semiparametric quasi maximum likelihood method to model and estimate longitudinal covariance functions. They studied the case in which the observation times are irregular on a continuous time interval. In their method, the variance function is modelled nonparametrically as a function in time, but the correlation is assumed to be a member of a known family of parametric correlation functions. The quasi maximum likelihood method provides a good trade-off between model flexibility and estimation efficiency, but it relies on correctly specifying the parametric model for the correlation function.

The next natural extension for these methods is to model the covariance function completely nonparametrically. Such ideas have become increasingly popular in functional data analysis,

where the covariance function is modelled as a smooth function and estimated by a kernel smoother. Some recent work on this topic includes Yao et al. (2005) and Hall et al. (2006).

It is an interesting question whether the semiparametric regression estimators can still achieve efficiency if the covariance function is modelled nonparametrically. In this paper we try to answer this question. We adopt a generalized partially linear model framework and show that when using a nonparametric kernel covariance estimator, the semiparametric estimator for the mean components remains asymptotically efficient. We refer to the proposed method as efficient semiparametric regression with nonparametric covariance estimation. We show by simulations that the proposed estimator has a comparable performance to using the correct covariance model and is more efficient than using a misspecified parametric covariance function.

2. THE MODEL AND DATA STRUCTURE

Suppose all longitudinal observations from different subjects or clusters are made on a fixed time interval $\mathcal{T} = [a, b]$. The data consist of n independent clusters, with the ith cluster having m_i observations. Let $T_i = (T_{i1}, \ldots, T_{im_i})^{\mathrm{T}}$ be the vector of random observation times in the ith cluster. Within the ith cluster, at time T_{ij}, we observe a response variable Y_{ij} and a p dimensional covariate vector X_{ij}. Let $Y_i = (Y_{i1}, \ldots, Y_{im_i})^{\mathrm{T}}$ and $X_i = (X_{i1}, \ldots, X_{im_i})^{\mathrm{T}}$. Suppose $2 \leqslant m_i \leqslant M$, for some finite number M, for all $i = 1, \ldots, n$.

Another approach is to model $Y_i(t)$ and $X_i(t)$ as random processes defined on continuous time $t \in \mathcal{T}$, with $X_i(t)$ being a multivariate process. Then, Y_i and X_i are observations on these processes at discrete times. We assume a generalized partially linear model (Wang et al., 2005)

$$\begin{aligned} &E\{Y_i(t) \mid X_i(s), s \in \mathcal{T}\} = E\{Y_i(t) \mid X_i(t)\} = \mu_i(t), \\ &g\{\mu_i(t)\} = X_i^{\mathrm{T}}(t)\beta + \theta(t), \quad t \in \mathcal{T}, \end{aligned} \tag{1}$$

where $g(\cdot)$ is a known monotone and differentiable link function, β is an unknown coefficient vector and $\theta(\cdot)$ is an unknown smooth function that represents the time effect. We also assume that the covariance of the response variable conditional on the covariates is a bivariate positive semidefinite function

$$\mathcal{R}(t_1, t_2) = \mathrm{cov}\{Y_i(t_1), Y_i(t_2) \mid X_i(s), s \in \mathcal{T}\}, \quad t_1, t_2 \in \mathcal{T}.$$

Fan et al. (2007) modelled the variance function and correlation function separately. Specifically, in their setting $\mathcal{R}(t_1, t_2) = \sigma(t_1)\sigma(t_2)\rho(t_1, t_2; \theta)$, where $\sigma^2(t)$ is the conditional variance of Y at time t modelled completely nonparametrically, $\rho(t_1, t_2; \theta)$ is a correlation function from a known family with only a few unspecified parameters, e.g., autoregressive moving average, ARMA, correlations.

We model $\mathcal{R}(t_1, t_2)$ as a bivariate nonparametric function, which is smooth except for the points on the diagonal line, $\{t_1 = t_2\}$, to allow possible nugget effects, see Diggle & Verbyla (1998), Yao et al. (2005) and Hall et al. (2006). To see this, let $\epsilon_i(t) = Y_i(t) - g^{-1}\{X_i(t)^{\mathrm{T}}\beta + \theta(t)\}$ be the error process, and by definition $\mathrm{cov}\{\epsilon_i(t_1), \epsilon_i(t_2)\} = \mathcal{R}(t_1, t_2)$. We assume that $\epsilon_i(t)$ can be decomposed into two independent components, $\epsilon_i(t) = \epsilon_{i0}(t) + \epsilon_{i1}(t)$, where $\epsilon_{i0}(\cdot)$ is a longitudinal process with smooth covariance function $\mathcal{R}_0(t_1, t_2)$; $\epsilon_{i1}(\cdot)$ is a white noise process usually caused by measurement errors. If $\sigma_1^2(t) = \mathrm{var}\{\epsilon_{i1}(t)\}$, then

$$\mathcal{R}(t_1, t_2) = \mathcal{R}_0(t_1, t_2) + \sigma_1^2(t_1) I(t_1 = t_2), \tag{2}$$

where $I(\cdot)$ is an indicator function. In (2), $\sigma_1^2(\cdot)$ is the nugget effect causing discontinuity in $\mathcal{R}(\cdot,\cdot)$. We assume that both $\mathcal{R}_0(\cdot,\cdot)$ and $\sigma_1^2(\cdot)$ are smooth functions. As a result, $\mathcal{R}(t_1,t_2)$ is a smooth surface except on the diagonal points where $t_1=t_2$, and it is also smooth along the diagonal direction. For time series data, without additional assumptions, some confounding will occur if both the mean and covariance functions are modelled nonparametrically. However, as illustrated by Yao et al. (2005) and Hall et al. (2006), this identifiability issue will not occur for longitudinal data, because of the independence between subjects.

Moreover, we assume that the observation times are random but not informative, i.e., $E(Y_{ij} \mid X_i, T_i) = \mu_{ij} = g^{-1}\{X_{ij}^{\mathrm{T}}\beta + \theta(T_{ij})\}$ and $\mathrm{cov}(Y_{ij}, Y_{ij'} \mid X_i, T_i) = \mathcal{R}(T_{ij}, T_{ij'})$; these assumptions are very common (Lin & Carroll, 2001; Wang et al., 2005).

3. The estimation procedure

3·1. *Stage* 1: *the initial estimator*

We first apply the working independence estimator of Lin & Carroll (2001), assuming that the within-cluster covariance matrices are identities. For a given value β, let $\hat{\alpha} = \hat{\alpha}(t,\beta) = \{\hat{\alpha}_0(t,\beta), \hat{\alpha}_1(t,\beta)\}^{\mathrm{T}}$ be the solution of the local estimating equation

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i} T_{ij}(t)\mu^{(1)}(X_{ij},t)K_{h_1}(T_{ij}-t)\{Y_{ij}-\mu(X_{ij},t)\} = 0,$$

where $T_{ij}(t) = \{1, (T_{ij}-t)/h\}^{\mathrm{T}}$, $\mu(\cdot)$ and $\mu^{(1)}(\cdot)$ are $g^{-1}(\cdot)$ and its first derivative evaluated at $X_{ij}^{\mathrm{T}}\beta + T_{ij}^{\mathrm{T}}(t)\alpha$, $K(\cdot)$ is a kernel function, $K_h(t) = h^{-1}K(t/h)$, and h_1 is the bandwidth used in Stage 1. Then the kernel estimator for $\theta(t)$ is given by $\hat{\theta}(t,\beta) = \hat{\alpha}_0(t,\beta)$.

We then proceed to estimate β by solving the profile estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{\partial\mu\{X_{ij}^{\mathrm{T}}\beta + \hat{\theta}(t,\beta)\}}{\partial\beta}[Y_{ij} - \mu\{X_{ij}^{\mathrm{T}}\beta + \hat{\theta}(T_{ij},\beta)\}] = 0. \tag{3}$$

The detailed algorithm for solving (3) is given in Lin & Carroll (2001).

3·2. *Stage* 2: *nonparametric covariance estimation*

Let $\hat{\epsilon}_{ij} = Y_{ij} - \mu\{X_{ij}^{\mathrm{T}}\hat{\beta} + \hat{\theta}(T_{ij},\hat{\beta})\}$, which is an estimate of $\epsilon_{ij} = Y_{ij} - \mu_{ij}$. Suppose $\mathcal{R}$ has a decomposition as in (2); we first estimate the smooth part $\mathcal{R}_0$ using a bivariate local linear smoother. Let $\hat{\mathcal{R}}_0(t_1,t_2) = \hat{\alpha}_0$, where $(\hat{\alpha}_0,\hat{\alpha}_1,\hat{\alpha}_2)$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{j'\neq j}\{\hat{\epsilon}_{ij}\,\hat{\epsilon}_{ij'} - \alpha_0 - \alpha_1(T_{ij}-t_1) - a_2(T_{ij'}-t_2)\}^2 K_{h_2}(T_{ij}-t_1)K_{h_2}(T_{ij'}-t_2). \tag{4}$$

Define

$$N_R = \sum_{i=1}^{n} m_i(m_i-1),$$

$$S_{pq} = \frac{1}{N_R}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{j'\neq j}\left(\frac{T_{ij}-t_1}{h_2}\right)^p\left(\frac{T_{ij'}-t_2}{h_2}\right)^q K_{h_2}(T_{ij}-t_1)K_{h_2}(T_{ij'}-t_2),$$

$$R_{pq} = \frac{1}{N_R} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{j' \neq j} \hat{\epsilon}_{ij} \hat{\epsilon}_{ik} \left(\frac{T_{ij} - t_1}{h_2} \right)^p \left(\frac{T_{ij'} - t_2}{h_2} \right)^q K_{h_2}(T_{ij} - t_1) K_{h_2}(T_{ij'} - t_2).$$

Then the following solution for (4) is given in Hall et al. (2006):

$$\hat{\mathcal{R}}_0(s,t) = (\mathcal{A}_1 R_{00} - \mathcal{A}_2 R_{10} - \mathcal{A}_3 R_{01}) \mathcal{B}^{-1}, \tag{5}$$

where $\mathcal{A}_1 = S_{20}S_{02} - S_{11}^2$, $\mathcal{A}_2 = S_{10}S_{02} - S_{01}S_{11}$, $\mathcal{A}_3 = S_{01}S_{20} - S_{10}S_{11}$, $\mathcal{B} = \mathcal{A}_1 S_{00} - \mathcal{A}_2 S_{10} - \mathcal{A}_3 S_{01}$.

The diagonal values on $\mathcal{R}(\cdot,\cdot)$ require a special treatment. Let $\sigma^2(t) = \mathcal{R}_0(t,t) + \sigma_1^2(t)$, then it can be estimated by a one-dimensional local linear smoother. We put $\hat{\sigma}^2(t) = \hat{\alpha}_0$, where $(\hat{\alpha}_0, \hat{\alpha}_1)$ minimizes

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \{\hat{\epsilon}_{ij}^2 - \alpha_0 - \alpha_1 (T_{ij} - t)\}^2 K_{h_3}(T_{ij} - t). \tag{6}$$

The covariance function is estimated by

$$\hat{\mathcal{R}}(s,t) = \hat{\mathcal{R}}_0(s,t) I(s \neq t) + \hat{\sigma}^2(t) I(s = t). \tag{7}$$

3·3. *Stage* 3: *the refined estimator*

First, we interpolate the estimated covariance function at the observation times to estimate the within-cluster covariance matrices. Let Σ_i and $\hat{\Sigma}_i$ be the true and estimated covariance matrix within the ith cluster, with $\hat{\Sigma}_i = \{\hat{\mathcal{R}}(T_{ij}, T_{ij'})\}_{j,j'=1}^{m_i}$. Following Wang et al. (2005), define $G_{ij}(t)$ to be an $m_i \times 2$ matrix with $\{1, (T_{ij} - t)/h_4\}$ on the jth row, and 0 in every other entry.

Algorithm 1. The refined estimator is obtained by iterating between the following two steps:

Step 1. Let $\widetilde{\theta}(\cdot)$ be the current estimator for $\theta(\cdot)$. Given β, update $\hat{\theta}(t)$ by $\hat{\alpha}_0(t,\beta)$, where $\hat{\alpha} = \{\hat{\alpha}_0(t,\beta), \hat{\alpha}_1(t,\beta)\}$ solves the local estimating equation

$$\sum_{i=1}^{n} \sum_{j=1}^{m_i} K_{h_4}(T_{ij} - t) \mu_{ij}^{(1)}(\beta, \hat{\alpha}) G_{ij}^{\mathrm{T}}(t) \hat{\Sigma}_i^{-1} [Y_i - \mu^*\{t, X_i, T_i, \beta, \hat{\alpha}, \widetilde{\theta}(T_i, \beta)\}] = 0, \tag{8}$$

where $\mu^*\{t, X_i, T_i, \beta, \hat{\alpha}, \widetilde{\theta}(T_i, \beta)\}]$ is a m_i dimensional vector with the ℓth entry given by

$$\mu[X_{i\ell}^{\mathrm{T}} \beta + I(\ell = j)\{\hat{\alpha}_0 + \hat{\alpha}_1 (T_{ij} - t)/h_4\} + I(\ell \neq j) \widetilde{\theta}(T_{i\ell}, \beta)].$$

Step 2. Update $\hat{\beta}$ by solving the estimating equation

$$\sum_{i=1}^{n} \frac{\partial \mu\{X_i \beta + \hat{\theta}(T_i, \beta)\}^{\mathrm{T}}}{\partial \beta} \hat{\Sigma}_i^{-1} [Y_i - \mu\{X_i \beta + \hat{\theta}(T_i, \beta)\}] = 0, \tag{9}$$

where $\hat{\theta}(T_i, \beta) = \{\hat{\theta}(T_{i1}, \beta), \ldots, \hat{\theta}(T_{im_i}, \beta)\}^{\mathrm{T}}$, and likewise $\mu(\cdot)$ is the vector pooling mean values within the same cluster.

Additional implementation details of Algorithm 1 will be postponed to §5, after discussion of the asymptotic properties of the proposed method.

4. Asymptotic theory

Let β_0 and $\theta_0(\cdot)$ be the true parameters in (1). For simplicity, we assume $m_i = m$ for all i, with $m \geqslant 2$. Let $f_1(\cdot)$ be the density for T_{ij}, and $f_2(\cdot, \cdot)$ be the joint density for any $(T_{ij}, T_{ij'})$, $j \neq j'$. We first show that the local linear variance and covariance estimators given by (4)–(7) are uniformly consistent.

Proposition 1. *If the assumptions in the Appendix hold, as $n \to \infty$, then*

$$\sup_{s,t\in[a,b]} |\hat{\mathcal{R}}_0(s,t) - \mathcal{R}_0(s,t)| = O_p[h_2^2 + \{\log(n)/(nh_2^2)\}^{1/2} + h_1^2 + \{\log(n)/(nh_1)\}^{1/2}],$$

$$\sup_{t\in[a,b]} |\hat{\sigma}^2(t) - \sigma^2(t)| = O_p[h_3^2 + \{\log(n)/(nh_3)\}^{1/2} + h_1^2 + \{\log(n)/(nh_1)\}^{1/2}].$$

The proof of Proposition 1 is given in the Appendix. Next, we study the asymptotic properties of the refined estimators $\hat{\theta}(\cdot)$ and $\hat{\beta}$ in Stage 3. The following notation is similar to that in Wang et al. (2005). Let Δ be a m-dimensional diagonal matrix, with the jth diagonal element being $\mu^{(1)}\{X_j^{\mathrm{T}}\beta + \theta(T_j)\}$. The function $\varphi_{\mathrm{eff}}(t)$ is the solution of the integral equation

$$\sum_{j=1}^{m}\sum_{\ell=1}^{m} E[\Delta_{jj}\sigma^{j\ell}\Delta_{\ell\ell}\{X_\ell - \varphi(T_\ell)\} \mid T_j = t] f_1(t) = 0, \tag{10}$$

where $\sigma^{j\ell}$ is the (j, ℓ)th element of Σ^{-1} and Δ_{jj} is the jth diagonal element of Δ. As described in Wang et al. (2005), equation (10) is a Fredholm integral equation of the second kind.

Proposition 2. *If all assumptions in the Appendix hold, as $n \to \infty$, we have the following convergence in distribution:*

$$n^{1/2}(\hat{\beta} - \beta_0) \longrightarrow N(0, \widetilde{A}^{-1}), \tag{11}$$

$$(nh_4)^{1/2}\{\hat{\theta}(t) - \theta_0(t) - h_4^2 B_*(t)\} \longrightarrow N\{0, V_\theta(t)\},$$

where $\widetilde{A} = E(\widetilde{X}^{\mathrm{T}}\Delta\Sigma^{-1}\Delta\widetilde{X})$, $\widetilde{X} = X - \varphi_{\mathrm{eff}}(T)$ and $B_*(t)$ and $V_\theta(t)$ are given in (A3) *and* (A4).

The key in our derivation is to show that the estimated covariance function is uniformly consistent and therefore, when it is interpolated at the subject-specific observation times and inserted into the profile local estimating equations, the estimation errors in the covariance only introduce an asymptotically negligible error into the final estimator.

An easier way to appreciate these results is to revisit the theory that Wang et al. (2005) developed for general working covariance functions. The working covariance function in our method is estimated from the data and is asymptotically consistent for the true covariance. As a result, the resulting estimator is asymptotically efficient. By comparing the asymptotic variance in (11) and that in Proposition 2 in Wang et al. (2005), we can see that the proposed estimator can still reach the semiparametric information bound.

One interesting question regarding our algorithm is whether we should iterate between Stages 2 and 3, i.e., re-estimate the covariance function when better estimates of the residuals are available. According to our theoretical derivation, there is no gain of efficiency in such iterations, since we need only a uniformly consistent covariance estimator for the Stage 3 estimator to work. From our experience in the simulation study, the working independence estimator in Stage 1 provides very good initial estimates for both β and $\theta(\cdot)$, and the covariance estimator based on the residuals from Stage 1 is fairly accurate. Therefore, we do not recommend such iterations.

The assumption of m being fixed for all subjects was for ease of exposition, and for easy comparison with the semiparametric information bound developed in Wang et al. (2005). For the case of varying m_i among the subjects, we can model m_i as independent variables with an identical distribution, as in Yao et al. (2005). Essentially, the same asymptotic results as in Proposition 2 can be developed: we just redefine integral equation (10) by taking another expectation over m, and in the definition of $\widetilde{A}$ the expectation is taken over m as well. Similar modifications are needed for the expressions of $B_*(t)$ and $V_\theta(t)$.

5. IMPLEMENTATION ISSUES

5·1. *Calculation of the profile estimator and sandwich formula*

The calculation of the semiparametric profile estimator in Stage 3 is not trivial. The kernel estimator in (8) is referred to by Lin et al. (2004) as the seemingly unrelated kernel estimator. Under the linear link function, an equivalent noniterative procedure is given by their Proposition 1. In this case, for a given β, the kernel estimator (8) can be written as

$$\hat{\theta}(t, \beta) = \mathcal{S}^{\mathrm{T}}(t)(Y - X\beta), \tag{12}$$

where Y and X are the response vector and design matrix obtained by pooling all subjects together, $\mathcal{S}^{\mathrm{T}}(t)$ is given in Lin et al. (2004). Now,

$$\frac{\partial}{\partial \beta}\hat{\theta}(t, \beta) = -\mathcal{S}(t)X,$$

for all β and (9) can easily be solved.

Motivated by (9), a sandwich formula for estimating the covariance of $\hat{\beta}$ is

$$\begin{aligned}\hat{\mathrm{var}}(\hat{\beta}) = &\left(\sum_{i=1}^{n} \widetilde{X}_i^{\mathrm{T}} \Delta_i \hat{\Sigma}_i^{-1} \Delta_i \widetilde{X}_i\right)^{-1} \left\{\sum_{i=1}^{n} \widetilde{X}_i^{\mathrm{T}} \Delta_i \hat{\Sigma}_i^{-1} (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^{\mathrm{T}} \hat{\Sigma}_i^{-1} \Delta_i \widetilde{X}_i\right\} \\ &\times \left(\sum_{i=1}^{n} \widetilde{X}_i^{\mathrm{T}} \Delta_i \hat{\Sigma}_i^{-1} \Delta_i \widetilde{X}_i\right)^{-1},\end{aligned} \tag{13}$$

where $\widetilde{X}_i = X_i + (\partial/\partial\beta)\hat{\theta}(T_i, \beta)$, $\Delta_i = \mathrm{diag}\{\mu_{ij}^{(1)}\}$, both evaluated at the estimated values $\{\hat{\beta}, \hat{\theta}(\cdot)\}$, and $\hat{\mu}_i$ is the estimated mean vector within cluster i.

Since $\hat{\theta}$ is a linear smoother, and by (12), we have $\mathrm{var}\{\hat{\theta}(t, \beta_0)\} = \mathcal{S}^{\mathrm{T}}(t)\mathrm{cov}(Y)\mathcal{S}(t)$. Notice that $\mathrm{cov}(Y)$ is a block diagonal matrix, with the ith block on the diagonal being the covariance matrix within the ith cluster. Therefore, a sandwich-type variance estimator is

$$\hat{\mathrm{var}}\{\hat{\theta}(t)\} = \mathcal{S}^{\mathrm{T}}(t)\hat{\Sigma}_{\mathrm{Sand}}\mathcal{S}(t), \tag{14}$$

where $\hat{\Sigma}_{\mathrm{Sand}} = \mathrm{diag}(\hat{\Sigma}_{i,\mathrm{Sand}})_{i=1}^{n}$, with $\hat{\Sigma}_{i,\mathrm{Sand}} = (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^{\mathrm{T}}$.

Similar sandwich formulae as (13) and (14) for the working independence estimator were provided in Lin & Carroll (2001). In our case, these formulae take the $\hat{\Sigma}_i$s as given, and therefore ignore the extra variability in $\hat{\beta}$ and $\hat{\theta}$ caused by substituting for the estimated covariance matrices. On the other hand, our theory says that this extra variability is of higher order. Therefore, it seems still reasonable to use these formulae. Our simulation results in §6 also confirm that these sandwich formulae perform acceptably for finite samples.

5·2. *The adjusted covariance function*

As noted in § 4, the key to the success of the proposed method is that the kernel covariance estimator provides uniformly consistent estimation of the covariance function. When the sample is large enough, estimation error in the covariance function is negligible. On the other hand, as noted in Hall et al. (1994) and Li et al. (2007), the kernel covariance estimator is not guaranteed to be positive semidefinite, and therefore some adjustment is needed to enforce the condition. This is particularly important when the sample size is relatively small, and the adjustment procedure would help by further regularizing the covariance estimator.

A commonly used spectral decomposition of the covariance functions for longitudinal data is (Yao et al., 2005; Hall et al., 2006)

$$\mathcal{R}_0(s,t) = \sum_{k=1}^{\infty} \omega_k \psi_k(s)\psi_k(t),$$

where $\omega_1 \geqslant \omega_2 \geqslant \cdots \geqslant 0$ are the eigenvalues of the covariance function and $\psi_k(t)$ are the corresponding eigenfunctions, with $\int_{\mathcal{T}} \psi_k(t)\psi_{k'}(t)dt = I(k = k')$.

An adjustment procedure has been proposed and theoretically justified by Hall et al. (2008) to transform $\hat{\mathcal{R}}_0$ into a valid covariance function. The idea is to take a spectral decomposition of $\hat{\mathcal{R}}_0$ and truncate the negative components. Letting $\hat{\omega}_k$ and $\hat{\psi}_k(\cdot)$ $(k = 1, 2, \ldots)$, be the eigenvalues and eigenfunctions of $\hat{\mathcal{R}}_0$, and $K_n = \max\{k; \hat{\omega}_k > 0\}$, then the adjusted estimator for $\mathcal{R}$ is

$$\widetilde{\mathcal{R}}_0(s,t) = \sum_{k=1}^{K_n} \hat{\omega}_k \hat{\psi}_k(s)\hat{\psi}_k(t), \quad \widetilde{\mathcal{R}}(s,t) = \widetilde{\mathcal{R}}_0(s,t)I(s \neq t) + \hat{\sigma}^2(t)I(s = t). \tag{15}$$

This adjustment procedure is carried out by discretizing $\hat{\mathcal{R}}_0(\cdot,\cdot)$ on a dense grid in $\mathcal{T}^2$, and then taking the eigenvalue decomposition of the resulting covariance matrix. By doing so, we obtain $\hat{\omega}_k$ and a discrete version of $\hat{\psi}_k(\cdot)$. To get the adjusted estimator of the subject-specific covariance matrix $\widetilde{\Sigma}_i$, we need to interpolate the $\hat{\psi}_k(\cdot)$s and $\hat{\sigma}^2(\cdot)$ on T_i and construct $\widetilde{\Sigma}_i$ according to (15). Our experience is that this procedure can effectively regularize the covariance matrices and stabilize the final estimator.

6. Simulation studies

6·1. *Simulation* 1

We set $\mathcal{T} = [0, 1]$ and let the observation times T_{ij} be independent variables with uniform distribution on $\mathcal{T}$. Let the sample size n be 200, and each subject has $m = 5$ observations. For each subject, we observe two covariates: X_1 is time dependent with $X_{1,ij} = T_{ij} + U_{ij}$, where $U_{ij} \sim \mathrm{Un}[-1, 1]$; X_2 is a binary time independent variable with $\mathrm{pr}(X_2 = 1) = 0{\cdot}5$. The response is generated from the model $Y_{ij} = \beta_1 X_{1,ij} + \beta_2 X_{2,ij} + \theta_0(T_{ij}) + \epsilon_{ij}$, where $\beta_1 = \beta_2 = 1$, $\theta_0(t) = \sin(2\pi t)$, and ϵ_{ij} is generated from a mixed effect model

$$\epsilon_{ij} = \xi_{0,ij} + \sum_{k=1}^{3} \xi_{k,i}\phi_k(T_{ij}),$$

where $\xi_{0,ij}, \xi_{k,i} \sim N(0, 0{\cdot}3)$ are independent random effects for all i, j and k, $\phi_1(t) = t^2 + 0{\cdot}5$, $\phi_2(t) = \sin(3\pi t)$, $\phi_3(t) = \cos(3\pi t)$.

Table 1. *Summary of the simulation results. The columns are the empirical bias* $\times$ 100*, standard error* $\times$ 100 *and the mean of the sandwich standard error estimator* $\times$100 *for* β_1 *and* β_2.

		β_1			β_2		
		Bias	SE	SWSE	Bias	SE	SWSE
Simulation 1	WI	0·1	4·5	4·9	0·5	8·5	8·3
	ESPR-ARMA	−0·1	4·2	4·2	0·3	8·2	7·9
	ESPR-NPC	−0·1	3·7	3·9	0·4	7·8	7·1
	ESPR-TC	−0·1	3·6	3·8	0·4	7·7	7·2
Simulation 2	WI	0·6	6·1	5·9	−0·7	12·1	12·3
	ESPR-ARMA	0·2	4·1	4·0	−0·9	11·9	12·1
	ESPR-NPC	0·2	4·3	4·1	−0·7	12·0	12·0

SE, standard error; SWSE, sandwich standard error; WI, working independence; ESPR-ARMA, efficient semiparametric regression with autoregressive moving-average correlation; ESPR-NPC, efficient semiparametric regression with nonparametric covariance estimation; ESPR-TC, efficient semiparametric regression with the true covariance.

In this model, ξ_1, ξ_2 and ξ_3 represent random effects of unidentified sources of variation, and ξ_0 is the measurement error which introduces the nugget effect. The covariance function $\mathcal{R}_0(s,t) = 0.3\sum_{k=1}^{3}\phi_k(s)\phi_k(t)$ has a periodic pattern, due to the periodicity in ϕ_2 and ϕ_3. Such a pattern cannot be modelled with any parametric covariance functions, but can be well estimated by the nonparametric covariance estimator.

We repeat the simulation 200 times. For each simulated dataset, we fit both the working independence estimator and our proposed method. For comparison, we also show the results for two more estimators: the efficient semiparametric regression estimator using the true covariance and the efficient estimator using a misspecified ARMA(1,1) covariance model, which is a widely used parametric model and allows for nugget effects. For the ARMA(1,1) method, we assume the covariance is of the form $\mathcal{R}(s,t) = \sigma(s)\sigma(t)\rho(s,t;\gamma,\nu)$, where $\sigma^2(t)$ is modelled nonparametrically, and the correlation function is of the parametric form

$$\rho(s,t;\gamma,\nu) = \gamma\exp(-|s-t|/\nu)I(s \neq t) + I(s=t), \tag{16}$$

with $\gamma \in [0,1]$ and $\nu \in (0,\infty)$. The parameters in the ARMA model are estimated using the quasi maximum likelihood method (Fan & Wu, 2008).

For a fair comparison, we use the same bandwidth when estimating $\theta(\cdot)$ for all four estimators. As the simulation results in Wang et al. (2005) show, neither the working independence estimator nor the semiparametric efficient estimator is sensitive to the choice of bandwidth. In the following, we show results using $h = 0{\cdot}1$ for $\hat{\theta}(\cdot)$, but our conclusions are not sensitive to this choice. To choose the bandwidths in Stage 2 of our estimator, we use leave one subject out crossvalidation (Rice & Silverman, 1991) in a pilot dataset, and then fix the values at the crossvalidated choice. For the results presented below, we use $h_2 = 0{\cdot}12$ and $h_3 = 0{\cdot}15$.

The results for estimating β are presented in the top part of Table 1. All four estimators have much smaller biases than the standard error, which is in line with the theory that they are asymptotically unbiased. Define relative efficiency between two estimators $\hat{\beta}_{m1}$ and $\hat{\beta}_{m2}$ to be $\mathrm{Eff}(\hat{\beta}_{m1},\hat{\beta}_{m2}) = \mathrm{var}(\hat{\beta}_{m2})/\mathrm{var}(\hat{\beta}_{m1})$. Then our proposed method has (27%, 12%) efficiency gain over the estimator assuming ARMA covariance for (β_1,β_2), and has (44%, 20%) efficiency gain over the working independence estimator. We can also see that the sandwich standard error estimator gives good approximations to the true standard errors for all cases.

For estimation of the nonparametric part, the pointwise mean squared error of $\hat{\theta}(t)$ by our method is almost uniformly lower than those of the methods assuming working independence or ARMA covariance, and is almost the same as using the true covariance.

6·2. *Simulation* 2

We now further illustrate the robustness of the proposed method. The data are generated with the same setup as in Simulation 1, except that the true correlation function is set to be the ARMA model given in (16). We set the marginal variance to be $\sigma^2 = 1{\cdot}2$, and set $\gamma = 0{\cdot}75$ and $\nu = 1$ for the correlation function (16). We apply the semiparametric regression methods assuming working independence, ARMA covariance and nonparametric covariance to the simulated data and repeat the simulation 200 times. The bandwidths used here are the same as in Simulation 1. The results for estimating β are summarized in Table 1.

Again, all estimators considered are asymptotically unbiased, which is confirmed from the numerical results: the biases are much smaller than the standard errors in all cases. In comparison to the standard error, the method assuming the ARMA covariance model outperforms the others. This is expected because the correlation model is correctly specified. The proposed estimator with nonparametric covariance estimation, on the other hand, performs reasonably well: the standard errors of our estimator are very close to those assuming the correct covariance model, and much smaller than those of the working independence estimator.

It is worth noting that, under the ARMA correlation model, the covariance surface $\mathcal{R}_0(s, t) = \sigma^2\gamma \exp(-|s - t|/\nu)$ is not differentiable on the diagonal line, $\{s = t\}$. Therefore, our assumption in the asymptotic theory that $\mathcal{R}_0$ is twice differentiable is violated. The good performance of our method in this situation shows that it is quite robust against the smoothness assumption. In fact, the smoothness assumption in the asymptotic theory is to guarantee the best possible convergence rate for the kernel covariance estimator. When such assumptions are mildly violated, the kernel covariance estimator is still uniformly consistent, but converge at a slower rate. This phenomenon was also observed in Li et al. (2007).

7. Application to CD4 data from AIDS clinical trial

7·1. *Data structure*

We now present an application of our method to the CD4 count data from the AIDS Clinical Trial Group 193A Study (Henry et al., 1998). The data are from a randomized, double-blind study of AIDS patients with CD4 counts of $\leqslant 50$ cells/mm^3. The patients were randomized to one of four treatments; each consisted of a daily regimen of 600 mg of zidovudine. Treatment 1 is zidovudine alternating monthly with 400 mg didanosine; Treatment 2 is zidovudine plus 2·25 mg of zalcitabine; Treatment 3 is zidovudine plus 400 mg of didanosine; Treatment 4 is a triple therapy consisting of zidovudine plus 400 mg of didanosine plus 400 mg of nevirapine.

Measurements of CD4 counts were scheduled to be collected at baseline and at eight week intervals during the 40 weeks of follow-up. However, the real observation times were unbalanced due to mistimed measurements, skipped visits and dropouts. The number of measurements of CD4 counts during the 40 weeks of follow-up varied from 1 to 9, with a median of 4. The response variable was the log-transformed CD4 counts, $Y = \log(\text{CD4 counts} + 1)$. There was also gender and baseline age information about each patient.

A total of 1309 patients were enrolled in the study. We eliminate the 122 patients who dropped out immediately after the baseline measurement. The remaining patients are mostly male, with 1044 males and only 143 females. We therefore concentrate on the male group. In Fig. 1, we

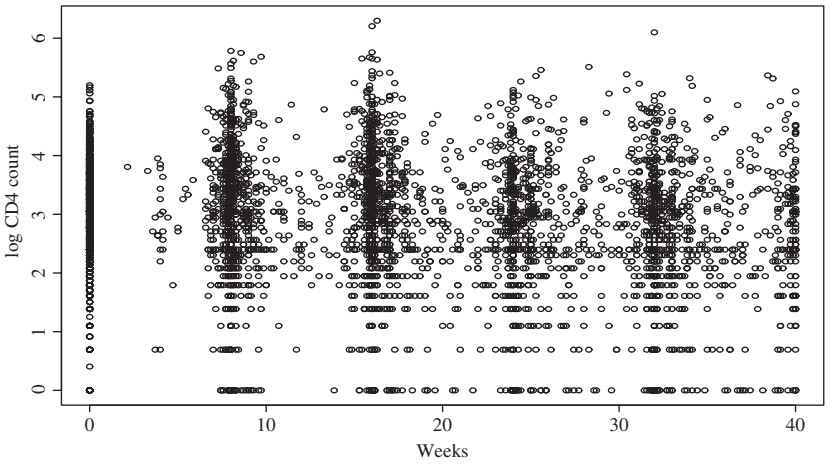

Fig. 1. Log CD4 count versus observation time in weeks.

show the scatter plot of Y versus observation times for all 1044 male patients. We have roughly the same number of patients, about 260, for each of the four treatment groups.

7·2. *A functional analysis of the covariance model*

We adopt the following model

$$Y_{k,ij} = X_{k,ij}\beta + \theta_k(T_{k,ij}) + \epsilon_{k,ij}, \quad i = 1, \ldots, n_k;\ j = 1, \ldots, m_i;\ k = 1, \ldots, 4, \tag{17}$$

where $Y_{k,ij}$ is the log CD4 count for the jth visit of the ith subject within the kth treatment group and $X_{k,ij}$ is the baseline age of the subject, which is time invariant. The time effects, $\theta_k(T)$, are modelled nonparametrically and are different in the different treatment groups.

In (17), the data within each treatment are modelled with a partially linear model as in (1), but we assume that all subjects have the same age effect β. In other words, there is no interaction between treatment and baseline age. Since the main effect of the kth treatment is given by the function θ_k, this model can be considered as a functional analysis of variance model as in Brumback & Rice (1998). Since we also have a covariate, baseline age, the model can be considered as a functional analysis of covariance model. We assume that each treatment group has the same within-subject covariance function, which is a common assumption in analysis of variance. In other words, we assume $\mathrm{cov}(\epsilon_{k,ij}, \epsilon_{k,ij'}) = \mathcal{R}(T_{k,ij}, T_{k,ij'})$ for all k, where $\mathcal{R}(\cdot, \cdot)$ is modelled nonparametrically with the structure as in (2).

The profile method described in § 3 can be easily extended to this new model. Given β, $\theta_k(\cdot)$ can be estimated by local linear smoothers $\hat{\theta}_k(t, \beta) = \mathcal{S}_k(t)(Y_k - X_k\beta)$, where Y_k and X_k are the response vector and design matrix within the kth treatment group. In Stage 1, the $\mathcal{S}_k$s are the working independence kernel smoothers, and in Stage 3 the $\mathcal{S}_k$s are the seemingly unrelated kernel smoothers.

To estimate β, we need to solve an estimating equation that pools all treatment groups together

$$\sum_{k=1}^{4}\sum_{i=1}^{n_k}\{X_{k,i} + \frac{\partial}{\partial\beta}\hat{\theta}_k(T_{k,i}, \beta)\}V_{k,i}^{-1}\{Y_{k,i} - X_{k,i}\beta - \hat{\theta}_k(T_{k,i}, \beta)\} = 0,$$

where the working covariance $V_{k,i}$ is identity in Stage 1 and is the kernel covariance estimator in Stage 3.

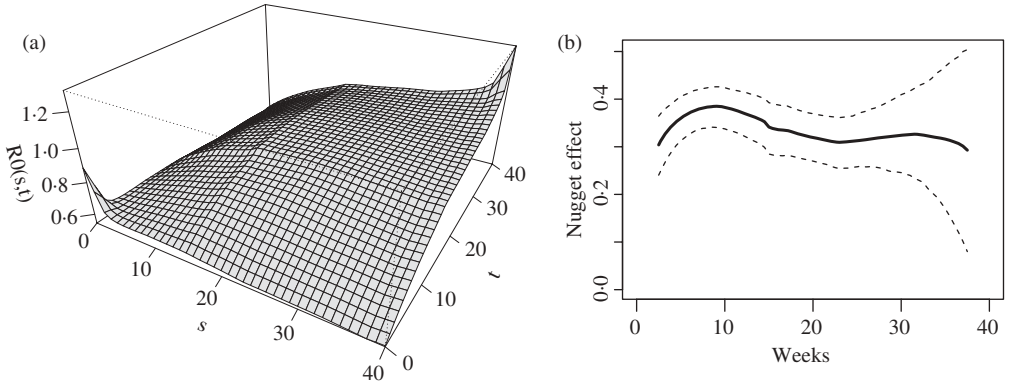

Fig. 2. Estimated covariance function and nugget effect for the CD4 count data. Panel (a) is $\hat{\mathcal{R}}_0(s,t)$. In panel (b), the solid curve is $\hat{\sigma}_1^2(t)$, and the two dotted curves are the 95% pointwise confidence intervals.

7·3. *Estimation results*

We first perform a preliminary analysis using a working independence profile kernel estimator, then estimate the covariance function based on the residuals pooled from all four treatment groups. In Fig. 2(a), we show the estimated covariance function $\hat{\mathcal{R}}_0(s,t)$. Although the covariance estimator shows some boundary effects, overall the covariance surface is quite smooth. In addition, the range of within subject covariance is rather long, i.e., all observations within the same subject are strongly correlated. In Fig. 2(b), we show the estimated nugget effect function, $\hat{\sigma}_1^2(t)$, and its 95% pointwise confidence intervals. The pointwise confidence intervals are obtained by a stratified bootstrap procedure, where a bootstrap sample consists of a random sample of n_k subjects from the kth treatment group, for $k = 1, \ldots, 4$. The estimation procedures in Stages 1 and 2 are performed to each bootstrap sample to get the bootstrap version $\hat{\sigma}_{1,B}^2(t)$. The bootstrap procedure is repeated 1000 times and the confidence intervals are taken as the pointwise 2·5 and 97·5 percentiles of $\hat{\sigma}_{1,B}^2(t)$. From the plot, we can see that the nugget effect is roughly constant over time.

We compare the semiparametric regression methods with three covariance structures: working independence, ARMA(1,1) covariance and the nonparametric covariance in Fig. 2. The parameters in the ARMA covariance function are estimated by the quasi maximum likelihood method. For β, the estimate, standard error and p-value are: $(1{\cdot}111, 0{\cdot}365, 0{\cdot}233) \times 10^{-2}$ for working independence; $(1{\cdot}109, 0{\cdot}102, 0{\cdot}000) \times 10^{-2}$ for ARMA (1,1); and $(1{\cdot}108, 0{\cdot}098, 0{\cdot}000) \times 10^{-2}$ for nonparametric covariance. The three estimators give essentially the same values for $\hat{\beta}$. The method using nonparametric covariance has the smallest standard error, indicating some efficiency gain over the other two estimators. Baseline age has a significant effect on CD4 count.

Unlike many other semiparametric regression problems, the nonparametric functions $\theta_k(\cdot)$ are of greater interest to us in this dataset since they are the treatment effects. In Fig. 3, we show the estimated θ functions for the four treatment groups, obtained from the proposed method. In each panel, the two dashed curves are the pointwise standard error bands, which are given by $\hat{\theta}_k(t) \pm 2 \times SE\{\hat{\theta}_k(t)\}$. At baseline, all treatment groups have almost the same mean CD4 count. As time moves on, differences between the treatment groups start to emerge. Treatment 1 does not seem to be effective, since the CD4 count still drops dramatically after the treatment. The other three treatments all show some positive effects, as they can delay the drop of CD4 counts. Treatment 4 seems most promising, since it can even increase CD4 counts in the short term after starting the treatment.

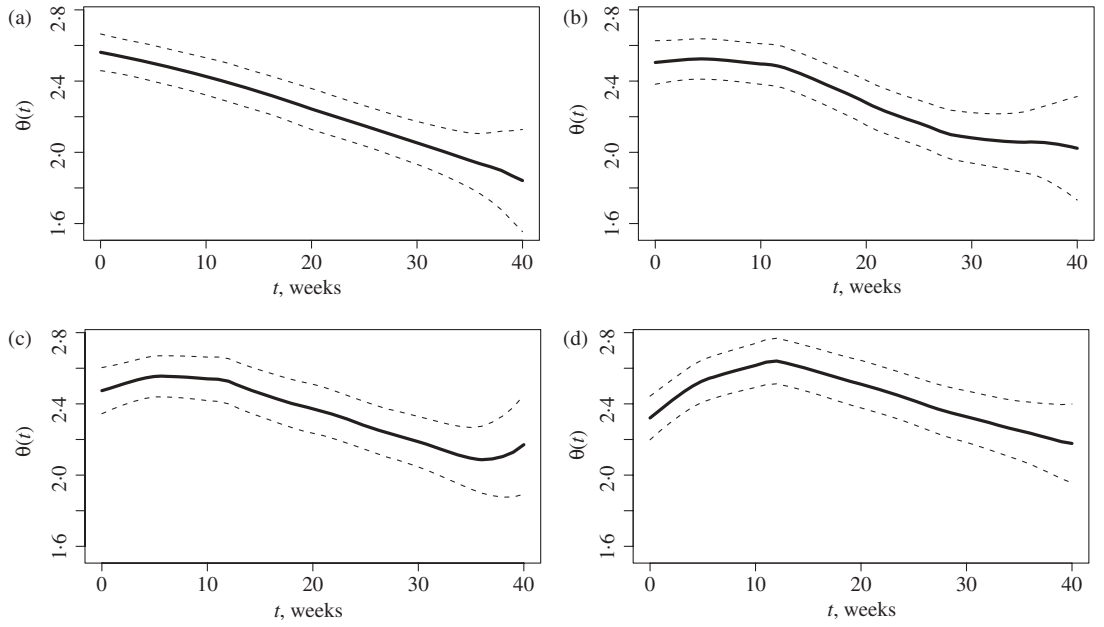

Fig. 3. Estimated time effect $\theta_k(t)$ of the CD4 count data. In each panel, the two dashed curves are $\hat{\theta}_k(t) \pm 2 \times SE\{\hat{\theta}_k(t)\}$. (a) Treatment 1, (b) treatment 2, (c) treatment 3 and (d) treatment 4.

Similar results are obtained from using the working independence and ARMA covariance structures. For these methods, the estimated $\tilde{\theta}(\cdot)$ are similar, but the bands are wider than those from our proposed method.

7·4. *Test for treatment effects*

To test the hypothesis, $H_0 : \theta_1(t) = \cdots = \theta_4(t)$, we assume that $\epsilon_{k,ij}$ are Gaussian errors and perform a generalized likelihood ratio test (Fan & Jiang, 2005). The estimated likelihood under the full model is

$$\hat{\ell}_{\text{full}} = -\frac{1}{2} \sum_{k=1}^{4} \sum_{i=1}^{n_k} \log(|\hat{\Sigma}_{k,i}|) + \{Y_{k,i} - X_{k,i}\hat{\beta} - \hat{\theta}_k(T_{k,i})\}^{\mathrm{T}} \hat{\Sigma}_{k,i}^{-1} \{Y_{k,i} - X_{k,i}\hat{\beta} - \hat{\theta}_k(T_{k,i})\},$$

where $\hat{\Sigma}_{k,i}$ is the nonparametric covariance estimator interpolated at the subject-specific times. The estimated likelihood under the reduced model, denoted as $\hat{\ell}_{\text{red}}$, is defined similarly under the assumption that the θ_ks are the same. The generalized likelihood ratio test statistic is defined as $\lambda_n(H_0) = \hat{\ell}_{\text{full}} - \hat{\ell}_{\text{red}}$. To assess the distribution of $\lambda_n(H_0)$ under the null hypothesis, we adopt a stratified conditional bootstrap procedure:

(i) To generate data satisfying the null hypothesis, take the residuals within each treatment group, $Y^*_{k,ij} = Y_{k,ij} - \hat{\theta}_k(T_{k,ij})$ for all k, i and j.
(ii) Sample with replacement n_k subjects within the kth treatment group and include the residuals, Y^*, within the resampled subjects into the bootstrap sample.
(iii) Calculate the generalized likelihood ratio test statistic, $\lambda^*_n(H_0)$, in the bootstrap sample, using the same bandwidths as for the real data.
(iv) Repeat steps (ii) and (iii) a large number of times and use the empirical distribution of $\lambda^*_n(H_0)$ to approximate the null distribution of λ_n.

The p-value of the test is estimated by the empirical frequency of λ_n^* that are greater than the observed test statistic λ_n. With a bootstrap sample size of 1000, we find the p-value for the null hypothesis to be 0·018. Therefore, we conclude that the four treatments have different effects.

Acknowledgement

This research was supported by a grant from the U.S. National Science Foundation. The author thanks the editor, the associate editor and two referees, whose helpful comments and suggestions greatly improved the paper. He also thanks Nicole Lazar for proofreading the paper.

Appendix

Proofs

The assumptions for our theoretical results.

Assumption A1. Assume that $\theta(\cdot)$ is a smooth function, and $\theta^{(2)}(\cdot)$ exists and is continuous on $[a, b]$.

Assumption A2. Assume that $\mathcal{R}(t_1, t_2) = \mathcal{R}_0(t_1, t_2) + \sigma_1^2(t_1) I(t_1 = t_2)$, where $\mathcal{R}_0$ is twice continuously differentiable on $\mathcal{T}^2$, and $\sigma^2(t) = \mathcal{R}_0(t, t) + \sigma_1^2(t)$ is twice continuously differentiable in t. Denote $\mathcal{R}_0^{(\ell_1,\ell_2)}$ as the partial derivatives, and they exist for $\ell_1, \ell_2 \geqslant 0$ and $\ell_1 + \ell_2 \leqslant 2$.

Assumption A3. Assume that f_1 has a compact support on $[a, b]$, and $0 < m_T \leqslant f_1(t) \leqslant M_T$ for $t \in [a, b]$. Both f_1 and f_2 are twice continuously differentiable.

Assumption A4. The kernel function $K(\cdot)$ is a symmetric continuously differentiable probability density function on $[-1, 1]$. Denote $\sigma_K^2 = \int K(t) t^2 dt$, $\nu_K = \int K^2(t) dt$.

Assumption A5. Assume that $E|\epsilon_{ij}|^{4+\delta_0} < \infty$ for some $\delta_0 > 0$.

Assumption A6. Assume that $h_1 \sim n^{-\nu_1}$, $1/5 \leqslant \nu_1 \leqslant 1/3$, as $n \to \infty$ (Lin & Carroll, 2001).

Assumption A7. Assume that $h_\ell \to 0$ as $n \to \infty$ for $\ell = 2, 3$, $nh_2^2/\log(n) \to \infty$, $nh_3/\log(n) \to \infty$.

Assumption A8. Assume that $h_4 \to 0$ as $n \to \infty$, such that $nh_4^8 \to 0$ and $nh_4/\log(1/h_4) \to \infty$ (Wang et al., 2005).

Proof of Proposition 1

The following asymptotic results regarding the working independence estimator were given by Lin & Carroll (2001) and Wang et al. (2005).

Lemma A1. *If $\hat{\beta}_{WI}$ and $\hat{\theta}_{WI}(t, \hat{\beta}_{WI})$ are the working independence estimators with bandwidth satisfying Assumption A6, then*

$$\hat{\beta}_{WI} - \beta_0 = O_p(n^{-1/2}), \quad \sup_{t \in \mathcal{T}} |\hat{\theta}_{WI}(t, \hat{\beta}_{WI}) - \theta_0(t)| = O_p[h_1^2 + \{\log(n)/(nh_1)\}^{1/2}].$$

Proof. By standard calculation,

$$(\hat{\mathcal{R}}_0 - \mathcal{R}_0)(s, t) = (\mathcal{A}_1 R_{00}^* - \mathcal{A}_2 R_{10}^* - \mathcal{A}_3 R_{01}^*)\mathcal{B}^{-1}, \tag{A1}$$

where $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$ and $\mathcal{B}$ are defined as in (5), and for $p, q \geqslant 0$,

$$R^*_{pq} = \frac{1}{N_R}\sum_i\sum_j\sum_{k\neq j}\{\hat\epsilon_{ij}\hat\epsilon_{ik} - \mathcal{R}_0(s,t) - \mathcal{R}_0^{(1,0)}(s,t)(T_{ij}-s) - \mathcal{R}_0^{(0,1)}(s,t)(T_{ik}-t)\}$$
$$\times\left(\frac{T_{ij}-s}{h_2}\right)^p\left(\frac{T_{ik}-t}{h_2}\right)^q K_{h_2}(T_{ij}-s)K_{h_2}(T_{ik}-t).$$

Denoting $\epsilon^*_{i,jk} = \epsilon_{ij}\epsilon_{ik} - \mathcal{R}_0(T_{ij}, T_{ik})$, we have $E(\epsilon_{i,jk}) = 0$. Then $R^*_{00} = (R^*_{00,a} + R^*_{00,v} + R^*_{00,b}) \times \{1 + o_p(1)\}$, where

$$R^*_{00,a} = \frac{1}{N_R}\sum_i\sum_j\sum_{k\neq j}(\hat\epsilon_{ij}\hat\epsilon_{ik} - \epsilon_{ij}\epsilon_{ik})K_{h_2}(T_{ij}-s)K_{h_2}(T_{ik}-t),$$

$$R^*_{00,v} = \frac{1}{N_R}\sum_i\sum_j\sum_{k\neq j}\epsilon^*_{i,jk}K_{h_2}(T_{ij}-s)K_{h_2}(T_{ik}-t),$$

$$R^*_{00,b} = \frac{1}{N_R}\sum_i\sum_j\sum_{k\neq j}\left\{\mathcal{R}_0(T_{ij},T_{ik}) - \mathcal{R}_0(s,t) - \mathcal{R}_0^{(1,0)}(s,t)(T_{ij}-s) - \mathcal{R}_0^{(0,1)}(s,t)(T_{ik}-t)\right\}$$
$$\times K_{h_2}(T_{ij}-s)K_{h_2}(T_{ik}-t).$$

By simple algebra, $R^*_{00,a} = N_R^{-1}\sum_i\sum_j\sum_{k\neq j}\{\epsilon_{ij}(\hat\epsilon_{ik}-\epsilon_{ik}) + (\hat\epsilon_{ij}-\epsilon_{ij})\epsilon_{ik} + (\hat\epsilon_{ij}-\epsilon_{ij})(\hat\epsilon_{ik}-\epsilon_{ik})\}$ $K_{h_2}(T_{ij}-s)K_{h_2}(T_{ik}-t)$. By Lemma A1, $\hat\epsilon_{ij}-\epsilon_{ij}$ is bounded by an $O_p[\{\log(n)/(nh_1)\}^{1/2} + h_1^2]$ term uniformly for all i and j. If we decompose $R^*_{00,a}$ into three terms according to the three terms in the brace, it can be easily seen that each term, and thus $R^*_{00,a}$ itself, is of order $O_p[\{\log(n)/(nh_1)\}^{1/2} + h_1^2]$ for all $s, t \in \mathcal{T}$.

The variance part $R^*_{00,v}$ is the two-dimensional kernel smoother applied to zero-mean variables $\epsilon^*_{i,jk}$. By classic uniform convergence rates for the kernel smoother (Mack & Silverman, 1982; Masry, 1995), we have $R^*_{00,v} = O_p[\{\log(n)/(nh_2^2)\}^{1/2}]$ uniformly for all (s,t). It can also be shown that the bias part is $R^*_{00,b} = 2^{-1}\sigma_K^2h_2^2f_2(s,t)\{\mathcal{R}_0^{(2,0)} + \mathcal{R}_0^{(0,2)}\}(s,t) + o_p[\{\log(n)/(nh_2^2)\}^{1/2} + h_2^2]$ uniformly for all (s,t).

Based on similar calculations, R^*_{10} and R^*_{01} have order $O_p[h_2^3 + \log(n)/(nh_2^2) + h_1^2 + \{\log(n)/(nh_1)\}^{1/2}]$ uniformly for all (s,t), and

$$S_{00} = f_2(s,t) + O_p[\{\log(n)/(nh_2^2)\}^{1/2} + h_2^2],$$
$$S_{01} = f_2^{(0,1)}(s,t)\sigma_K^2h_2 + O_p[\{\log(n)/(nh_2^2)\}^{1/2} + h_2^3],$$
$$S_{10} = f_2^{(1,0)}(s,t)\sigma_K^2h_2 + O_p[\{\log(n)/(nh_2^2)\}^{1/2} + h_2^3],$$
$$S_{02} = f_2(s,t)\sigma_K^2 + O_p[\{\log(n)/(nh_2^2)\}^{1/2} + h_2^2],$$
$$S_{20} = f_2(s,t)\sigma_K^2 + O_p[\{\log(n)/(nh_2^2)\}^{1/2} + h_2^2],$$
$$S_{11} = f_2^{(1,1)}(s,t)\sigma_K^4h_2^2 + O_p[\{\log(n)/(nh_2^2)\}^{1/2} + h_2^4].$$

The convergence rate for $\hat{\mathcal{R}}_0$ is obtained by inserting these rates back into (A1).

The convergence rate for $\hat\sigma^2(\cdot)$ is that of a one-dimensional local linear smoother, which can be obtained by similar derivations. □

Proof of Proposition 2

LEMMA A2. *Let A_n be a sequence of random matrices converging to an invertible matrix A. Then*

$$A_n^{-1} = A^{-1} - A^{-1}(A_n - A)A^{-1} + O_p(\|A_n - A\|^2),$$

where $\|A\| = \mathrm{tr}(A^{\mathrm{T}}A)^{1/2}$.

Proof. The proof is carried out by proving the profile estimating equations (8) and (9) are asymptotically equivalent to the estimating equations (3) and (4) in Wang et al. (2005) when the true covariance matrices are used. In other words, the error incurred by substituting for the estimated covariance matrix is asymptotically negligible.

Let $\delta_n = h_2^2 + h_3^2 + \{\log(n)/(nh_2^2)\}^{1/2} + \{\log(n)/(nh_3)\}^{-1/2}$, then by Proposition 1 $\hat{\mathcal{R}}(s,t) - \mathcal{R}(s,t) = O_p(\delta_n)$ uniformly for all (s,t). By Lemma A2, $\hat{\Sigma}_i^{-1} - \Sigma_i^{-1} = O_p(\delta_n)$ uniformly for all i. Let $\sigma_i^{j\ell}$ and $\hat{\sigma}_i^{j\ell}$ be the (j,ℓ)th element of Σ_i and $\hat{\Sigma}_i$, then $\hat{\sigma}_i^{j\ell} - \sigma_i^{j\ell} = O_p(\delta_n)$ uniformly.

For a given β, local equation estimating (8) is

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{h_4}(T_{ij}-t)\mu_{ij}^{(1)}(\beta,\hat{\alpha})G_{ij}^{\mathrm{T}}(t)\hat{\Sigma}_i^{-1}[Y_i - \mu^*\{t, X_i, T_i, \beta, \hat{\alpha}, \widetilde{\theta}(T_i,\beta)\}] \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{h_4}(T_{ij}-t)\mu_{ij}^{(1)}(\beta,\hat{\alpha})G_{ij}^{\mathrm{T}}(t)\Sigma_i^{-1}[Y_i - \mu^*\{t, X_i, T_i, \beta, \hat{\alpha}, \widetilde{\theta}(T_i,\beta)\}] \\
&\qquad - \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{h_4}(T_{ij}-t)\mu_{ij}^{(1)}(\beta,\hat{\alpha})G_{ij}^{\mathrm{T}}(t)\Sigma_i^{-1}(\hat{\Sigma}_i - \Sigma_i)\Sigma_i^{-1}\right. \\
&\qquad\quad \left.\times\,[Y_i - \mu^*\{t, X_i, T_i, \beta, \hat{\alpha}, \widetilde{\theta}(T_i,\beta)\}]\right) \times \{1 + o_p(1)\}. \qquad \text{(A2)}
\end{aligned}$$

The dominant term in (A2) is the same as in the local estimating equation (3) in Wang et al. (2005), which is of order $O_p(h_4^2 + \{\log(n)/(nh_4)\}^{1/2})$. The second term in (A2) is a higher order $o_p[h_4^2 + \{\log(n)/(nh_4)\}^{1/2}]$ term. Therefore, for a given β, the nonparametric estimator $\hat{\theta}(t,\beta)$ with inserted covariance estimator is asymptotically equivalent to that using the true covariance. By differentiation with respect to β on both sides of (A2) and following similar lines as in the proof of Proposition 1 in Wang et al. (2005), we can show that $\varphi(t) = \partial\hat{\theta}(t,\beta)/\partial\beta$ is asymptotically equivalent to that found by inserting the true covariance, and therefore $\varphi(t) = \varphi_{\mathrm{eff}}(t) \times \{1 + o_p(1)\}$, where $\varphi_{\mathrm{eff}}(t)$ is the solution of the integration equation (10). Therefore, the profile estimating function in (9) is equivalent to

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{\partial\mu\{X_i\beta + \hat{\theta}(T_i,\beta)\}^{\mathrm{T}}}{\partial\beta}\Sigma_i^{-1}[Y_i - \mu\{X_i\beta + \hat{\theta}(T_i,\beta)\}] \\
&\quad - \left(\sum_{i=1}^{n} \frac{\partial\mu\{X_i\beta + \hat{\theta}(T_i,\beta)\}^{\mathrm{T}}}{\partial\beta}\Sigma_i^{-1}(\hat{\Sigma}_i - \Sigma_i)\Sigma_i^{-1}[Y_i - \mu\{X_i\beta + \hat{\theta}(T_i,\beta)\}]\right) \times \{1 + o_p(1)\}.
\end{aligned}$$

Again the first term in the expression above is the same as equation (4) in Wang et al. (2005), and the second term is asymptotically negligible. Therefore, (9) is asymptotically equivalent to (4) in Wang et al. when the true covariance is inserted. The asymptotic distribution of $\hat{\beta}$ follows directly from Proposition 2 in Wang et al. (2005), with working covariance V equal to the true covariance Σ.

Following Wang et al. (2005), $\hat{\theta}(t)$ has the asymptotic expansion

$$\begin{aligned}
\hat{\theta}(t) - \theta_0(t) = &\frac{1}{2}h_4^2 B_*(t) + \frac{1}{nW_2(t)}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{ij}^{(1)}K_{h_4}(T_{ij}-t)\left\{\sum_{\ell=1}^{m}\sigma_i^{j\ell}(Y_{i\ell} - \mu_{i\ell})\right\} \\
&+ o_p\{\log(n)/(nh_4)^{1/2}\}, \qquad \text{(A3)}
\end{aligned}$$

where $W_2(t) = \sum_{j=1}^{m} E\{\Delta_{jj}^2\sigma^{jj} \mid T_j = t\}f_1(t)$, $B_*(t) = b_*(t) \times \{1 + o(1)\}$ and $b_*(t)$ satisfies the integral equation $b_*(t) = \theta^{(2)}(t) - W_2^{-1}(t)\sum_j\sum_{\ell\neq j} E\{\Delta_{jj}\sigma^{j\ell}\Delta_{\ell\ell}b_*(T_\ell) \mid T_j = t\}f_1(t)$. By routine calculation,

one can show that the asymptotic variance for $\hat{\theta}(t)$ is given by $(nh_4)^{-1}V_\theta(t)$, where

$$V_\theta(t) = \nu_K \left[\sum_{j=1}^{m} E\{\Delta_{jj}^2 \sigma^{jj} \mid T_j = t\} f_1(t) \right]^{-1}. \tag{A4}$$

This asymptotic variance is the same as (13) in Wang (2003). The asymptotic normality of $\hat{\theta}$ follows from the central limit theorem. □

REFERENCES

BRUMBACK, B. A. & RICE, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Am. Statist. Assoc.* **93**, 961–76.

DIGGLE, P. J. & VERBYLA, A. P. (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics* **54**, 401–15.

DIGGLE, P. J., HEAGERTY, P., LIANG, K.-Y., & ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. New York: Oxford University Press.

FAN, J. & JIANG, J. (2005). Nonparametric inferences for additive models. *J. Am. Statist. Assoc.* **100**, 890–907.

FAN, J. & WU, Y. (2008). Semiparametric estimation of covariance matrixes for longitudinal data. *J. Am. Statist. Assoc.* **103**, 1520–33.

FAN, J., HUANG, T. & LI, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Am. Statist. Assoc.* **102**, 632–41.

HALL, P., FISHER, N. I. & HOFFMAN, B. (1994). On the nonparametric estimation of covariance functions. *Ann. Statist.* **22**, 2115–34.

HALL, P., MÜLLER, H. G. & WANG, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34**, 1493–517.

HALL, P., MÜLLER, H. G. & YAO, F. (2008). Modelling sparse generalized longitudinal observations with latent Gaussian processes. *J. R. Statist. Soc.* B. **70**, 703–23.

HENRY, K., ERICE, A., TIERNEY, C., BALFOUR, H. H. JR, FISCHL, M. A., KMACK, A., LIOU, S. H., KENTON, A., HIRSCH, M. S., PHAIR, J., MARTINEZ, A. & KAHN J. O. (1998). A randomized, controlled, double-blind study comparing the survival benefit of four different reverse transcriptase inhibitor therapies (three-drug, two-drug, and alternating drug) for the treatment of advanced AIDS. *J. Acq. Immune Defic. Synd. Hum. Retrovir.*, **19**, 339–49.

HUANG, J. Z., LIU, L., & LIU, N. (2007). Estimation of large covariance matrices of longitudinal data with basis function approximations. *J. Comp. Graph. Statist.* **16**, 189–209.

LI, Y., WANG, N., HONG, M., TURNER, N., LUPTON, J. & CARROLL, R. J. (2007). Nonparametric estimation of correlation functions in spatial and longitudinal data, with application to colon carcinogenesis experiments. *Ann. Statist.* **35**, 1608–43.

LIN, X. & CARROLL, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Am. Statist. Assoc.* **96**, 1045–56.

LIN, X. & CARROLL, R. J. (2006). Semiparametric estimation in general repeated measures problems. *J. R. Statist. Soc.* B **68**, 69–88.

LIN, X., WANG, N., WELSH, A. H. & CARROLL, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. *Biometrika*, **91**, 177–93.

MACK, Y. P. & SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahr. verw. Geb.* **61**, 405–15.

MASRY, E. (1995). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.* **17**, 571–599.

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.

WANG, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, **90**, 43–52.

WANG, N., CARROLL, R. J. & LIN, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Am. Statist. Assoc.*, **100**, 147–57.

WU, W. B., & POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, **90**, 831–44.

YAO, F., MÜLLER, H. G. & WANG, J. L. (2005). Functional data analysis for sparse longitudinal data. *J. Am. Statist. Assoc.*, **100**, 577–90.

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 September* 2009. *Revised November* 2010]