JOURNAL OF NONPARAMETRIC STATISTICS
2019, VOL. 31, NO. 4, 867–886
https://doi.org/10.1080/10485252.2019.1651853

Efficient semiparametric regression for longitudinal data with regularised estimation of error covariance function

Shengji Jia[a], Chunming Zhang[a,b] and Hulin Wu[c]

[a]Department of Statistics, University of Wisconsin, Madison, WI, USA; [b]School of Mathematics and Statistics, Yunnan University, Kunming, People's Republic of China; [c]Department of Biostatistics, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA

ABSTRACT
Improving estimation efficiency for regression coefficients is an important issue in the analysis of longitudinal data, which involves estimating the covariance matrix of errors. But challenges arise in estimating the covariance matrix of longitudinal data collected at irregular or unbalanced time points. In this paper, we develop a regularisation method for estimating the covariance function and a stepwise procedure for estimating the parametric components efficiently in the varying-coefficient partially linear model. This procedure is also applicable to the varying-coefficient temporal mixed-effects model. Our method utilises the structure of the covariance function and thus has faster rates of convergence in estimating the covariance functions and outperforms the existing approaches in simulation studies. This procedure is easy to implement and its numerical performance is investigated using both simulated and real data.

ARTICLE HISTORY
Received 30 September 2017
Accepted 8 July 2019

KEYWORDS
Covariance function; local linear regression; method of regularisation; profile weighted least squares; semiparametric varying-coefficient partially linear model; Sobolov space; tensor product space

1. Introduction

There has been substantial recent interest in nonparametric and semiparametric methods for longitudinal or clustered data with dependence within subjects (or clusters) (see Lin and Carroll 2001; Diggle, Heagerty, Liang, and Zeger 2002; Wu and Zhang 2002b). Improving estimation efficiency is an important issue in the analysis of longitudinal data. In the nonparametric setting, Lin and Carroll (2001) recommended an approach which ignores the within-subject correlation completely and treats the data as if they were independent. However, Wang, Carroll, and Lin (2005) showed that, in the semiparametric setting, the estimator for parametric component in the model will achieve the semiparametric efficiency bound if the within-subject correlation structure is specified correctly. Thus estimating the covariance function is an important issue in the semiparametric model for longitudinal data.

Many authors have investigated the problem of within-subject correlation in longitudinal data. For example, Wu and Pourahmadi (2003) proposed nonparametric estimation

CONTACT Chunming Zhang cmzhang@stat.wisc.edu Department of Statistics, University of Wisconsin, Madison, WI 53706, USA; School of Mathematics and Statistics, Yunnan University, Kunming, Yunnan 650091, People's Republic of China

of large covariance matrices using a two step estimation procedure. But their method can only deal with balanced or nearly balanced longitudinal data. Challenges arise in estimating the covariance function if the data are collected at irregular or subject-specific time points. Wu and Zhang (see Wu and Zhang 2002a, 2006) proposed another method called local polynomial linear mixed-effects model (LLME) to analyse longitudinal data, which estimated the within-subject error directly instead of estimating the covariance function of the errors. Other methods for modelling the covariance function include the functional principal component analysis (FPCA) proposed by Yao, Müller, and Wang (2005).

In this paper, we consider a semiparametric varying-coefficient partially linear model

$$Y(t) = \boldsymbol{X}(t)^T\boldsymbol{\alpha}(t) + \boldsymbol{Z}(t)^T\boldsymbol{\beta} + \eta(t) + \zeta(t), \tag{1}$$

where $\boldsymbol{\alpha}(t)$ comprises p unknown smooth functions, $\boldsymbol{\beta}$ is a q-dimensional unknown parameter vector, and $\eta(t)$ captures the within-subject dependence with smooth covariance function $R_\eta(t_1, t_2)$, $\zeta(t)$ is just the measurement error with covariance function $\sigma_\zeta^2(t_1)\mathrm{I}(t_1 = t_2)$. All the temporal correlations are relegated to $\eta(t)$, so this decomposition is unique. Nonparametric models for longitudinal data can be viewed as special cases of model (1). Moreover, model (1) is an extension of the partially linear model and the time-varying-coefficient model.

We focus on estimating the covariance function $R_\eta(t_1, t_2)$ (defined in (5)) when observations are collected at irregular and possibly subject-specific time points. In this paper,

(i) A *varying-coefficient temporal mixed-effects model* is introduced as a good approximation of model (1). The within-subject correlated error $\eta(t)$ can be considered as a combination of some common random factors (not related to t) and some temporal functions.
(ii) A general framework of the regularisation method is applied to estimate the covariance function, which can be viewed as an extension of one-dimensional smoothing splines. This method is introduced through a careful characterisation of the function space (tensor product of Hilbert space) in which the covariance function $R_\eta(t_1, t_2)$ resides, and thus has faster rates of convergence compared to other methods which only assume $R_\eta(t_1, t_2)$ is a bivariate continuous function.
(iii) An explicit spectral decomposition of the estimated covariance function is established, and we can easily guarantee the estimated covariance function to be positive definite after truncating the negative eigenvalues. To improve the efficiency of estimating the regression coefficients, the weight matrix is chosen by the inverse of the adjusted covariance matrix in the weighted least squares method.
(iv) Our proposed method can be applied to both the sparse longitudinal data and the densely sampled functional data. Besides, our method also works quite well for the missing longitudinal/functional data, which can be considered as a special case of longitudinal/functional data collected at irregular time points.

There also has been a vast volume of work on modelling covariance functions in longitudinal data in literature. Fan, Huang, and Li (2007) and Fan and Wu (2008) proposed a quasi-maximum likelihood method to model covariance function of $\eta(t)$. In their method, the variance function $\mathrm{var}\{\eta(t)\}$ is modelled nonparametrically, but the correlation function $\mathrm{corr}\{\eta(t_1), \eta(t_2)\}$ is assumed to be a member of a known family of parametric correlation functions (e.g. an AR or ARMA correlation structure). The quasi-maximum likelihood method relies on correctly assuming the form of the correlation functions and can be misspecified easily. Li (2011) applied bivariate kernel smoothing techniques to estimate covariance functions. But the techniques of bivariate kernel smoothing do not utilise the structures of the covariance functions, and thus have slower rates of convergence compared with our method. Besides, the kernel covariance estimator is not guaranteed to be positive semidefinite, and an adjustment procedure is required by discretizing the kernel estimator on a dense grid, followed by taking the eigenvalue decomposition of the resulting covariance matrix. This discretizing procedure is quite subjective, since it depends heavily on the choice of dense grids.

The rest of the article is organised as follows. Section 2 describes the semiparametric varying-coefficient model for longitudinal data and decomposition of covariance function. A nonparametric estimation of the covariance function and an efficient estimation procedure for parameters based on profile least squares techniques is described in Section 3. Sampling properties of the proposed procedure are presented in Section 4. In Sections 5 and 6 the proposed method is illustrated via simulation studies and real data examples, respectively. All technical proofs are relegated to Appendix.

2. Model and covariance structure

Suppose all longitudinal observations from different subjects (or clusters) are made on a fixed time interval $\mathcal{T} \subset \mathbb{R}$, e.g. $\mathcal{T} = [0, 1]$. The data consist of n independent subjects. For the ith subject, $i = 1, \ldots, n$, the response variable $Y_i(t)$ and the covariates $\{\boldsymbol{X}_i(t), \boldsymbol{Z}_i(t)\}$ are collected at time points $t = t_{i,j}$, $j = 1, \ldots, J_i$, where J_i is the total number of observations for the ith subject. In this article, we consider a semiparametric varying-coefficient partially linear model

$$Y(t) = \boldsymbol{X}(t)^T \boldsymbol{\alpha}(t) + \boldsymbol{Z}(t)^T \boldsymbol{\beta} + \varepsilon(t)$$

i.e.

$$Y_i(t_{i,j}) = \boldsymbol{X}_i(t_{i,j})^T \boldsymbol{\alpha}(t_{i,j}) + \boldsymbol{Z}_i(t_{i,j})^T \boldsymbol{\beta} + \varepsilon_i(t_{i,j}), \quad i = 1, \ldots, n,\ j = 1, \ldots, J_i, \tag{2}$$

where $\boldsymbol{\alpha}(t)$ comprises p unknown smooth functions, $\boldsymbol{\beta}$ is a q-dimensional unknown parameter vector, and $\{\varepsilon_i(t) : i = 1, \ldots, n\}$ are i.i.d. error processes with $\mathrm{E}\{\varepsilon_i(t) \mid \boldsymbol{X}_i(t), \boldsymbol{Z}_i(t)\} = 0$. To consider the within-subject dependence, we assume that $\varepsilon_i(t)$ can be decomposed into two independent error processes:

$$\varepsilon_i(t) = \eta_i(t) + \zeta_i(t),$$

where $\{\eta_i(t) : i = 1, \ldots, n\}$ are i.i.d. mean zero error processes capturing the within-subject dependence or temporal correlation, and $\{\zeta_i(t) : i = 1, \ldots, n\}$ are the i.i.d. measurement error (see Yao et al. 2005; Hall, Müller, and Wang 2006). For $t_1 \in \mathcal{T}$ and $t_2 \in \mathcal{T}$, suppose

$$\begin{aligned} \operatorname{cov}\{\eta_i(t_1), \eta_i(t_2)\} &= R_\eta(t_1, t_2), \\ \operatorname{cov}\{\zeta_i(t_1), \zeta_i(t_2)\} &= \sigma_\zeta^2(t_1)\mathrm{I}(t_1 = t_2), \quad i = 1, \ldots, n, \end{aligned} \tag{3}$$

where $\mathrm{I}(\cdot)$ is an indicator function, $R_\eta(\cdot,\cdot)$ and $\sigma_\zeta^2(\cdot)$ are smooth functions. Then the covariance function $R(t_1,t_2)$ of $\varepsilon_i(t)$ is given by

$$R(t_1,t_2) \equiv \mathrm{cov}\{\varepsilon_i(t_1),\varepsilon_i(t_2)\} = R_\eta(t_1,t_2) + \sigma_\zeta^2(t_1)\mathrm{I}(t_1=t_2), \quad i=1,\ldots,n, \tag{4}$$

which is a smooth surface except on the diagonal points where $t_1=t_2$. In Section 3, we will estimate $R_\eta(\cdot,\cdot)$ and $\sigma_\zeta^2(\cdot)$ separately.

There are many different methods to analyse the within-subject error $\eta_i(t)$. For example, Wu and Zhang (2002a) used the local polynomial method to decompose $\eta_i(t)$. Here let us consider functional principal component analysis model for $\eta_i(t)$. Let $\lambda_{(1)} \geq \lambda_{(2)} \geq \cdots \geq 0$ be ordered values of the eigenvalues of the linear operator determined by $R_\eta(\cdot,\cdot)$ with $\sum_{k=1}^\infty \lambda_{(k)} < \infty$, and the $\psi_k(\cdot)$'s be the corresponding orthonormal eigenfunctions or principal components, see Hall et al. (2006). Then, $R_\eta(\cdot,\cdot)$ admits the spectral decomposition:

$$R_\eta(t_1,t_2) = \sum_{k=1}^\infty \lambda_{(k)}\psi_k(t_1)\psi_k(t_2), \tag{5}$$

and $\eta_i(t)$ admits the Karhunen-Loeve expansion as follows,

$$\eta_i(t) = \sum_{k=1}^\infty \xi_{i,k}\psi_k(t),$$

where $\xi_{i,k} = \int_{t\in\mathcal{T}} \eta_i(t)\psi_k(t)\,\mathrm{d}t$ are uncorrelated random variables with $\mathrm{E}(\xi_{i,k})=0$ and $\mathrm{E}(\xi_{i,k}\xi_{i,j}) = \lambda_{(k)}\mathrm{I}(j=k)$. If $\lambda_{(k)} \approx 0$ for $k \geq L+1$, then model (2) can be approximated by

$$Y_i(t_{i,j}) \approx \boldsymbol{X}_i(t_{i,j})^T\boldsymbol{\alpha}(t_{i,j}) + \boldsymbol{Z}_i(t_{i,j})^T\boldsymbol{\beta} + \sum_{k=1}^L \xi_{i,k}\psi_k(t_{i,j}) + \zeta_i(t_{i,j}). \tag{6}$$

Model (6) can be regarded as a *varying-coefficient temporal mixed-effects model*, since $\xi_{i,k}$ are random variables and $\psi_k(t)$ are 'unknown' but fixed basis functions.

As for the estimation of $\sigma_\zeta^2(\cdot)$ and $R_\eta(\cdot,\cdot)$, we assume that $\sigma_\zeta^2(\cdot)$ is a smooth function so that smoothing techniques such as the local linear regression can be applied to estimate the variance function $\sigma_\zeta^2(\cdot)$. By the covariance function decomposition (5), we assume that $R_\eta(\cdot,\cdot)$ resides in a tensor product of Hilbert space $\mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$, which is the closure of the following linear space

$$\mathrm{span}\{f(s)g(t) : f(\cdot), g(\cdot) \in \mathcal{W}_2^2(\mathcal{T})\}, \tag{7}$$

where $\mathcal{W}_2^2(\mathcal{T}) = \{f : f', f'' \text{ are absolutely continuous}, f'' \in L_2(\mathcal{T})\}$ is a Sobolev space endowed with the squared norm $\int_\mathcal{T}(f'')^2$. Because $\mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$ is dense in the continuous bivariate function space, we can find an element in $\mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$ that approximates any continuous bivariate function very well.

3. Proposed methodology for estimation

In practice, estimation of $\{\boldsymbol{\alpha}(t), \boldsymbol{\beta}\}$ must be done in multiple steps. Their initial estimates are constructed by ignoring within-subject correlation. With the initial estimates of $\{\boldsymbol{\alpha}(t), \boldsymbol{\beta}\}$, we can estimate $R(\cdot,\cdot)$ based on residuals. Finally, we can estimate $\{\boldsymbol{\alpha}(t), \boldsymbol{\beta}\}$ more efficiently by using the estimate of $R(\cdot,\cdot)$. In this section, we propose the efficient estimates for $\{\boldsymbol{\alpha}(t), \boldsymbol{\beta}\}$ using profile weighted least squares techniques.

3.1. Step 1: initial estimator

For a given $\boldsymbol{\beta}$, model (2) reduced to a varying-coefficient model:

$$Y_i(t_{i,j}) - \boldsymbol{Z}_i(t_{i,j})^T\boldsymbol{\beta} = \boldsymbol{X}_i(t_{i,j})^T\boldsymbol{\alpha}(t_{i,j}) + \eta_i(t_{i,j}) + \zeta_i(t_{i,j}). \tag{8}$$

Ignoring the within-subject correlation or $\eta_i(t)$, we use the profile local linear regression to get initial estimates of $\{\boldsymbol{\alpha}(t), \boldsymbol{\beta}\}$, see Fan and Huang (2005). For any t in the neighbourhood of t_0, the lth component $\alpha_l(t)$ of $\boldsymbol{\alpha}(t)$, admits Taylor's expansion as follows:

$$\begin{aligned}\alpha_l(t) &\approx \alpha_l(t_0) + \alpha_l'(t_0)(t - t_0)\\ &\equiv a_l + b_l(t - t_0), \quad \text{for } l = 1, \dots, p.\end{aligned}$$

Let $K(\cdot)$ be a kernel function and h_1 be a bandwidth. Thus we can find the local linear estimator $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(t_0)$ of $\boldsymbol{\alpha}(t_0)$, where $\boldsymbol{\alpha}(t)$ is the true varying function in model (8). Let $(\widehat{a}_1, \dots, \widehat{a}_p, \widehat{b}_1, \dots, \widehat{b}_p)$ be the minimiser of

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\left[Y_i(t_{i,j}) - \boldsymbol{Z}_i(t_{i,j})^T\boldsymbol{\beta} - \sum_{l=1}^{p}\{a_l + b_l(t_{i,j} - t_0)\}X_{il}(t_{i,j})\right]^2 K_{h_1}(t_{i,j} - t_0), \tag{9}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, and $\boldsymbol{X}_i(t) = (X_{i1}(t), \dots, X_{ip}(t))^T$. Then $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(t_0) = (\widehat{a}_1, \dots, \widehat{a}_p)^T$. Note that the profile least squares estimator of $(\boldsymbol{\alpha}(t), \boldsymbol{\beta})$ has a closed form using the following matrix notation. Let

$$\begin{aligned}
\boldsymbol{Y} &= (\boldsymbol{Y}_1^T, \dots, \boldsymbol{Y}_n^T)^T, \quad \boldsymbol{Y}_i = (Y_i(t_{i,1}), \dots, Y_i(t_{i,J_i}))^T,\\
\mathbf{Z} &= (\mathbf{Z}_1^T, \dots, \mathbf{Z}_n^T)^T, \quad \mathbf{Z}_i = (\boldsymbol{Z}_i(t_{i,1}), \dots, \boldsymbol{Z}_i(t_{i,J_i}))^T,\\
\boldsymbol{m} &= (\boldsymbol{m}_1^T, \dots, \boldsymbol{m}_n^T)^T, \quad \boldsymbol{m}_i = (\boldsymbol{X}_i(t_{i,1})^T\boldsymbol{\alpha}(t_{i,1}), \dots, \boldsymbol{X}_i(t_{i,J_i})^T\boldsymbol{\alpha}(t_{i,J_i}))^T.
\end{aligned}$$

Then model (8) can be written as

$$\boldsymbol{Y} - \mathbf{Z}\boldsymbol{\beta} = \boldsymbol{m} + \boldsymbol{\eta} + \boldsymbol{\zeta}, \tag{10}$$

where $\boldsymbol{\eta} = (\eta_1(t_{1,1}), \dots, \eta_n(t_{n,J_n}))^T$ and $\boldsymbol{\zeta} = (\zeta_1(t_{1,1}), \dots, \zeta_n(t_{n,J_n}))^T$. Since the estimator $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(\cdot)$ is linear in $\boldsymbol{Y} - \mathbf{Z}\boldsymbol{\beta}$, given $\boldsymbol{\beta}$, the estimator of $\boldsymbol{m}$ is of the form $\widehat{\boldsymbol{m}} = \mathbf{S}(\boldsymbol{Y} - \mathbf{Z}\boldsymbol{\beta})$, where $\mathbf{S}$ is a smoothing matrix of the local linear smoother, see Fan and Gijbels (1996). Substituting $\widehat{\boldsymbol{m}}$ into model (10) results in the linear model,

$$(\mathbf{I} - \mathbf{S})\boldsymbol{Y} \approx (\mathbf{I} - \mathbf{S})\mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\zeta},$$

where $\mathbf{I}$ is an identity matrix. So an initial estimator for $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}^{\text{ini}} = \{\mathbf{Z}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\mathbf{Z}\}^{-1}\mathbf{Z}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\boldsymbol{Y}. \tag{11}$$

Then the profile least squares estimator for the nonparametric component $\boldsymbol{\alpha}(\cdot)$ is just $\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}^{\text{ini}}}(\cdot)$.

3.2. Step 2: covariance estimator

After we get the initial estimators $\widehat{\boldsymbol{\beta}}^{\text{ini}}$ and $\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}^{\text{ini}}}(t)$ in Step 1, the residuals are

$$\widehat{\varepsilon}_i(t_{i,j}) = Y_i(t_{i,j}) - \boldsymbol{X}_i(t_{i,j})^T \widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}^{\text{ini}}}(t_{i,j}) - \boldsymbol{Z}_i(t_{i,j})^T \widehat{\boldsymbol{\beta}}^{\text{ini}}, \quad i = 1, \ldots, n,\ j = 1, \ldots, J_i. \tag{12}$$

Then we will derive the nonparametric estimator of $R(\cdot,\cdot)$ based on $\widehat{\varepsilon}_i(t_{i,j})$. Since there are too many parameters in $R_\eta(\cdot,\cdot) \in \mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$, similar to the idea of smoothing spline, a penalty for over-parametrization is imposed to regularise the covariance function. Let $\widehat{R}_\eta(s,t) \in \mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$ be the minimiser of

$$\frac{1}{\sum_{i=1}^n J_i(J_i-1)} \sum_{i=1}^n \sum\sum_{1 \le j_1 \ne j_2 \le J_i} \left\{\widehat{\varepsilon}_i(t_{i,j_1})\widehat{\varepsilon}_i(t_{i,j_2}) - R_\eta(t_{i,j_1}, t_{i,j_2})\right\}^2 + \lambda_n P(R_\eta) \tag{13}$$

where $\lambda_n \ge 0$ is a tuning parameter, and $P(R_\eta)$ is a penalty function for $R_\eta(s,t) = \sum_{j \ge 1} a_j f_j(s) g_j(t)$ defined in (7).

The diagonal element of $R(\cdot,\cdot)$ requires a special treatment since it involves both $R_\eta(t,t)$ and $\sigma_\zeta^2(t)$. Denote $\sigma^2(t) \equiv R(t,t)$, which can be estimated by an one-dimensional local linear smoother. Let $(\widehat{\gamma}_0, \widehat{\gamma}_1)$ be the minimiser of

$$\sum_{i=1}^n \sum_{j=1}^{J_i} \{\widehat{\varepsilon}_i^2(t_{i,j}) - \gamma_0 - \gamma_1(t_{i,j} - t)\}^2 K_{h_2}(t_{i,j} - t), \tag{14}$$

where h_2 is a new bandwidth which can be different from h_1 in Step 1. Define $\widehat{\sigma}^2(t) = \max\{\widehat{\gamma}_0, 0\}$, where we take the maximum to avoid a negative $\widehat{\gamma}_0$. According to the definition of $R(s,t)$ in (4), $R(s,t)$ can also be written as

$$R(s,t) = R_\eta(s,t)\mathrm{I}(s \ne t) + \sigma^2(t)\mathrm{I}(s = t),$$

that is, $R(s,t)$ is $R_\eta(s,t)$ when $s \ne t$ and $\sigma^2(t)$ when $s = t$. So the estimator $\widehat{R}(\cdot,\cdot)$ of the covariance function is a combination of $\widehat{R}_\eta(s,t)$ and $\widehat{\sigma}^2(t)$:

$$\widehat{R}(s,t) = \widehat{R}_\eta(s,t)\mathrm{I}(s \ne t) + \widehat{\sigma}^2(t)\mathrm{I}(s = t). \tag{15}$$

By the definition of $\sigma^2(t)$, $\sigma^2(t) = R_\eta(t,t) + \sigma_\zeta^2(t) > R_\eta(t,t)$. Since $\widehat{\sigma}^2(t)$ is a consistent estimator of $\sigma^2(t)$, it can be shown that with probability tending to 1, $\widehat{\sigma}^2(t) > \widehat{R}_\eta(t,t)$.

3.3. Step 3: refined estimator

First, let Σ_i and $\widehat{\Sigma}_i$ be the true and estimated covariance matrix within the ith subject, i.e.

$$\Sigma_i = \left[R(t_{i,j}, t_{i,k})\right]_{j,k=1}^{J_i}, \quad \widehat{\Sigma}_i = \left[\widehat{R}(t_{i,j}, t_{i,k})\right]_{j,k=1}^{J_i}. \tag{16}$$

Since we ignore the within-subject correlation or $\eta_i(t)$ in Step 1, the initial least squares estimator $\widehat{\boldsymbol{\beta}}^{\text{ini}}$ in (11) is not efficient. To improve the efficiency for estimating $\boldsymbol{\beta}$, we use the profile weighted least squares estimator as follows:

$$\{\mathbf{Z}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{Z}\}^{-1}\mathbf{Z}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\boldsymbol{Y}, \tag{17}$$

where $\mathbf{W}$ is a weight matrix, called a working covariance matrix. Then the initial estimate $\widehat{\boldsymbol{\beta}}^{\text{ini}}$ in (11) is just a special case of (17) with $\mathbf{W}$ being an identity matrix. As usual, misspecification of the working covariance matrix does not affect the consistency of the resulting estimate, but does affect the efficiency. Fan et al. (2007) has shown that the most efficient estimator for $\boldsymbol{\beta}$ among the profile weighted least squares estimates given in (17) is the one that uses the inverse of the true variance-covariance matrix of errors as the weight matrix, that is, $\mathbf{W} = \operatorname{diag}(\Sigma_1^{-1}, \ldots, \Sigma_n^{-1})$. Because Σ_i's are unknown, we can use the estimators $\widehat{\Sigma}_i$ in (16), so the final refined estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \{\mathbf{Z}^T(\mathbf{I}-\mathbf{S})^T\widehat{\mathbf{W}}(\mathbf{I}-\mathbf{S})\mathbf{Z}\}^{-1}\mathbf{Z}^T(\mathbf{I}-\mathbf{S})^T\widehat{\mathbf{W}}(\mathbf{I}-\mathbf{S})\boldsymbol{Y}, \tag{18}$$

where $\widehat{\mathbf{W}} = \operatorname{diag}(\widehat{\Sigma}_1^{-1}, \ldots, \widehat{\Sigma}_n^{-1})$. The profile least squares estimator for the nonparametric component is simply $\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}}(\cdot)$.

3.4. Adjusted covariance function estimator

It can be shown that $\widehat{R}_\eta$ in Step 2 is a consistent estimator of R_η, but is not guaranteed to be positive semidefinite, and therefore some adjustment is needed to enforce the positive semidefinite condition, which is particularly important when the sample size is relatively small. The idea is to take a spectral decomposition of $\widehat{R}_\eta$ and truncate the negative components.

Computing the eigenvalues and eigenfunctions of a symmetric bivariate function is generally nontrivial. Typically this is done by discretizing the covariance function estimation and approximating its eigenfunctions by the respective eigenvectors (see Hall, Müller, and Yao 2008). However, discretizing the covariance function is quite subjective since it depends heavily on the choice of dense grids. Fortunately, if the choice of the penalty function $P(R_\eta)$ for $R_\eta(s,t) = \sum_{j\geq 1} a_j f_j(s) g_j(t)$ follows Cai and Yuan (2010),

$$P(R_\eta) = \|R_\eta\|^2_{\mathcal{W}_2^2(\mathcal{T})\otimes\mathcal{W}_2^2(\mathcal{T})}, \tag{19}$$

then

$$\begin{aligned}
P(R_\eta) &= \iint_{t\in\mathcal{T},s\in\mathcal{T}} \left\{\frac{\partial^4 R_\eta(s,t)}{\partial^2 s\partial^2 t}\right\}^2 \mathrm{d}s\,\mathrm{d}t \\
&= \iint_{t\in\mathcal{T},s\in\mathcal{T}} \left\{\sum_{j\geq 1} a_j f_j''(s) g_j''(t)\right\}^2 \mathrm{d}s\,\mathrm{d}t \\
&= \sum\sum_{i,j\geq 1} a_i a_j \int_{\mathcal{T}} \{f_i''(s) f_j''(s)\}\,\mathrm{d}s \int_{\mathcal{T}} \{g_i''(t) g_j''(t)\}\,\mathrm{d}t,
\end{aligned}$$

and the minimiser $\widehat{R}_\eta$ of (13) must have the following form

$$\widehat{R}_\eta(s,t) = \sum_{i=1}^{n} \boldsymbol{H}_i(s)^T \widehat{\mathbf{A}}_i \boldsymbol{H}_i(t) = \sum_{i=1}^{n} \sum_{j,k=1}^{J_i} \widehat{\mathbf{A}}_i(j,k) H(s,t_{i,j}) H(t,t_{i,k}), \tag{20}$$

where $\widehat{\mathbf{A}}_i$ is a $J_i \times J_i$ symmetric matrix, $\widehat{\mathbf{A}}_i(j,k)$ is the element in the jth row and kth column of $\widehat{\mathbf{A}}_i$, and

$$\boldsymbol{H}_i(s) = \left(H(s,t_{i,1}), \ldots, H(s,t_{i,J_i})\right)^T \in \mathbb{R}^{J_i},$$

where $H(s,t) = \frac{1}{4}B_2(s)B_2(t) - \frac{1}{24}B_4(|s-t|)$ and B_r is the rth Bernoulli polynomial, see Milton and Irene (1972). Thanks to the representation (20), the eigenvalues and eigenfunctions of $\widehat{R}_\eta(s,t) = \sum_{i=1}^n \boldsymbol{H}_i(s)^T \widehat{\mathbf{A}}_i \boldsymbol{H}_i(t)$ can actually be computed explicitly. Let $\boldsymbol{h}(\cdot) = (\boldsymbol{H}_1(\cdot)^T, \ldots, \boldsymbol{H}_n(\cdot)^T)^T$, and

$$\widehat{\mathbf{A}} = \begin{pmatrix} \widehat{\mathbf{A}}_1 & 0 & \ldots & 0 \\ 0 & \widehat{\mathbf{A}}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \widehat{\mathbf{A}}_n \end{pmatrix}_{N \times N}$$

where $\widehat{\mathbf{A}}_i$'s are defined in (20) and $N = J_1 + \cdots + J_n$ is the total number of observations.

Assume that $\widehat{\mathbf{A}} = \widehat{\mathbf{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{U}}^T$ is the eigenvalue decomposition of $\widehat{\mathbf{A}}$, where $\widehat{\boldsymbol{\Lambda}} = \mathrm{diag}\{\widehat{\lambda}_{(1)}, \widehat{\lambda}_{(2)}, \ldots, \widehat{\lambda}_{(N)}\}$ is the diagonal matrix of the decreasing eigenvalues $\widehat{\lambda}_{(1)} \geq \widehat{\lambda}_{(2)} \geq \ldots \geq \widehat{\lambda}_{(N)}$, and $\widehat{\mathbf{U}} \equiv (\widehat{\boldsymbol{u}}_1, \widehat{\boldsymbol{u}}_2, \ldots, \widehat{\boldsymbol{u}}_N)$ is the matrix of the corresponding eigenvectors. Then $\widehat{R}_\eta(s,t)$ in (20) admits the following spectral decomposition:

$$\begin{aligned} \widehat{R}_\eta(s,t) &= \boldsymbol{h}(s)^T \widehat{\mathbf{A}} \boldsymbol{h}(t) \\ &= \boldsymbol{h}(s)^T \left(\widehat{\mathbf{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{U}}^T\right) \boldsymbol{h}(t) \\ &= \boldsymbol{h}(s)^T \left\{ \sum_{k=1}^{N} \widehat{\lambda}_{(k)} \widehat{\boldsymbol{u}}_k \widehat{\boldsymbol{u}}_k^T \right\} \boldsymbol{h}(t) \\ &\equiv \sum_{k=1}^{N} \widehat{\lambda}_{(k)} \widehat{\psi}_k(s) \widehat{\psi}_k(t) \end{aligned}$$

where $\widehat{\psi}_k(\cdot) = \widehat{\boldsymbol{u}}_k^T \boldsymbol{h}(\cdot)$ is the estimator of $\psi_k(\cdot)$ in (6). Since $\widehat{\lambda}_{(k)}$'s are not necessarily positive, we can first truncate the negative eigenvalues. Then the adjusted estimators for R_η and R are defined as

$$\widetilde{R}_\eta(s,t) = \sum_{k=1}^{N} \widehat{\lambda}_{(k)} \mathrm{I}(\widehat{\lambda}_{(k)} > \tau) \widehat{\psi}_k(s) \widehat{\psi}_k(t), \quad s,t \in \mathcal{T}, \tag{21}$$

$$\widetilde{R}(s,t) = \widetilde{R}_\eta(s,t)\mathrm{I}(s \neq t) + \max\{\widetilde{R}_\eta(t,t), \widehat{\sigma}^2(t)\}\mathrm{I}(s = t), \tag{22}$$

where $\tau \geq 0$ is a predetermined threshold for the eigenvalues (e.g. 0.01) or a percentage (e.g. 1 percent) of the sum of all the positive eigenvalues. It will be shown in the next section that $\widetilde{R}(s,t)$ is positive definite. Thus when we want to take the inverse of $\widehat{\Sigma}_i$ in (16) to estimate $\boldsymbol{\beta}$, it is better to replace $\widehat{R}(t_{i,j}, t_{i,k})$ by $\widetilde{R}(t_{i,j}, t_{i,k})$ in the expression of $\widehat{\Sigma}_i$ in (16) and $\widehat{\mathbf{W}}$ in (18).

4. Theoretical results

In this section we investigate sampling properties of the covariance function estimator as $n \to \infty$. The proposed estimation procedures are applicable for various formulations for collecting longitudinal data. To facilitate the presentation, we assume that $\{J_i : i = 1, \ldots, n\}$ are independent and identically distributed random variables with $0 < \mathrm{E}(J_i) < \infty$, and $\{t_{i,1}, \ldots, t_{i,J_i} \mid J_i\}_{i=1}^n$ are independent and identically distributed on $\mathcal{T}$ with a density $f_T(t)$, see Fan et al. (2007). In this section and Sections 5–6, the penalty function $P(R_\eta)$ takes the special form (19).

First, we show that the residuals $\widehat{\varepsilon}_i(t_{i,j})$ in (12) are uniformly consistent estimators for the true errors $\varepsilon_i(t_{i,j})$.

Proposition 4.1: *Assume regularity conditions* C1–C5 *in the Appendix. If* $\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{X}(t)^T\}$ *is positive definite for each* $t \in \mathcal{T}$, *then*

$$\sup_{i,j} |\widehat{\varepsilon}_i(t_{i,j}) - \varepsilon_i(t_{i,j})| = O_{\mathrm{P}}(h_1^2 + \{-\log(h_1)/(nh_1)\}^{1/2}).$$

From the proof of Proposition 4.1 and the definition of $\widehat{\boldsymbol{\beta}}$ in (18), it is easy to derive the consistency of the proposed estimators of $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\beta}$ in Corollary 4.1.

Corollary 4.1: *Under the conditions of Proposition* 4.1, *we have*

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_{\mathrm{P}}(n^{-1/2});$$
$$\sup_{t\in\mathcal{T}} \left|\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}}(t) - \boldsymbol{\alpha}(t)\right| = O_{\mathrm{P}}(h_1^2 + \{-\log(h_1)/(nh_1)\}^{1/2}).$$

Next, define $\overline{\sigma}^2(t) = \overline{\gamma}_0$, where $(\overline{\gamma}_0, \overline{\gamma}_1)$ minimise

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\{\varepsilon_i^2(t_{i,j}) - \gamma_0 - \gamma_1(t_{i,j} - t)\}^2 K_{h_2}(t_{i,j} - t) \tag{23}$$

and $\overline{R}_\eta \in \mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$ be the minimiser of

$$\frac{1}{\sum_{i=1}^n J_i(J_i - 1)} \sum_{i=1}^{n} \sum\sum_{1\le j_1 \ne j_2 \le J_i} \left\{\varepsilon_i(t_{i,j_1})\varepsilon_i(t_{i,j_2}) - R_\eta(t_{i,j_1}, t_{i,j_2})\right\}^2 + \lambda_n P(R_\eta). \tag{24}$$

So (14) and (13) are data versions of (23) and (24) after we replace the unobserved $\varepsilon_i(t_{i,j})$ by residuals $\widehat{\varepsilon}_i(t_{i,j})$. $\overline{\sigma}^2$ and $\overline{R}_\eta$ are called pseudo-estimators for σ^2 and R_η. The next proposition showed the consistency of $\overline{\sigma}^2$ and $\overline{R}_\eta$.

Proposition 4.2: (i) *Under the regularity conditions* C1–C7 *in the Appendix*,

$$\sup_{t\in\mathcal{T}} \left|\overline{\sigma}^2(t) - \sigma^2(t)\right| = O_{\mathrm{P}}(h_2^2 + \{-\log(h_2)/(nh_2)\}^{1/2}). \tag{25}$$

(ii) *Under the regularity conditions* C1–C7 *in the Appendix*,

$$\|\overline{R}_\eta - R_\eta\|_{L_2} = O_{\mathrm{P}}\big(\{\log(n)/n\}^{2/5}\big), \tag{26}$$

if the tuning parameter $\lambda_n \asymp \{\log(n)/n\}^{2/5}$. *Here* $\|R_\eta(\cdot,\cdot)\|_{L_2} = \{\iint_{s,t\in\mathcal{T}} R_\eta^2(s,t)\,\mathrm{d}s\,\mathrm{d}t\}^{1/2}$ *denotes the integrated squared norm of a bivariate function*, *see* Bosq (2000).

Now let's consider the rates of convergence for the estimators $\overline{\sigma}^2$ and $\overline{R}_\eta$. From (25), when $h_2 \asymp \{\log(n)/n\}^{1/5}$, we have $\sup_{t\in\mathcal{T}} |\overline{\sigma}^2 - \sigma^2| = O_{\mathrm{P}}(\{\log(n)/n\}^{2/5})$. Thus $\overline{\sigma}^2$ has the same optimal rates of convergence as $\overline{R}_\eta$ in (26). On the other hand, if we use two-dimensional smoothing techniques to estimate $R_\eta(\cdot,\cdot)$, the optimal L_2-convergence rate is $n^{-1/3}$, which is much larger than the convergence rate $\{\log(n)/n\}^{2/5}$ in Proposition 4.2. This is because the bivariate continuous functions space with continuous second derivatives $C^{(2)}(\mathcal{T})$ is much larger than the tensor product Hilbert space $\mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$.

Finally, we show in the next proposition that the adjusted covariance function estimator defined in Section 3.4 is positive definite.

Proposition 4.3: *The adjusted covariance function estimator* $\widetilde{R}(s,t)$ *in (22) is positive definite almost surely.*

5. Simulation study

5.1. Simulation 1

In this section, we investigate the finite sample properties of the estimators proposed in Sections 3 through Monte Carlo simulations. Suppose the data are generated from the following model:

$$Y_i(t_{i,j}) = \boldsymbol{X}_i(t_{i,j})^T\boldsymbol{\alpha}(t_{i,j}) + \boldsymbol{Z}_i(t_{i,j})^T\boldsymbol{\beta} + \eta_i(t_{i,j}) + \zeta_i(t_{i,j}), \quad i = 1,\ldots,n,\ j = 1,\ldots,J_i. \tag{27}$$

We set the sample size n be 200, and $\{J_i : 1 \le i \le n\}$ be independent discrete uniform random variables on $\{5,6,7\}$. Let $\mathcal{T} = [0,1]$ and given J_i, the observation times $\{t_{i,j} : 1 \le i \le n,\ 1 \le j \le J_i\}$ be independent variables with uniform distribution on $\mathcal{T}$. We let the coefficients of $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\beta}$ be two dimensional in our simulation, and further set $X_1(t) \equiv 1$ to include an intercept term. We generate the covariates in the following way: For a given t, $(X_2(t), Z_1(t))^T$ follows a bivariate normal distribution with mean 0, variance 1, and correlation 0.5, and $Z_2(t)$ is a Bernoulli distributed random variable with success probability 0.5 and independent of $X_2(t)$ and $Z_1(t)$. In this simulation we set $\boldsymbol{\beta} = (1,2)^T$, $\alpha_1(t) = \sqrt{t}$, and $\alpha_2(t) = \sin(2\pi t)$. For $i = 1,\ldots,n$, the within-subject errors $\eta_i(t)$ are generated from a temporal mixed-effects model:

$$\eta_i(t) = \sum_{k=1}^{L} \xi_{i,k}\psi_k(t), \tag{28}$$

where $\xi_{i,k}$'s are independent standard normal random variables. Thus the covariance function is $R_\eta(s,t) = \sum_{k=1}^{L} \psi_k(s)\psi_k(t)$. We set $L = 1$, and $\psi_1(t) = \cos(\pi t)$. Finally we assume the measurement errors $\zeta_i(t_{i,j})$ follow $N(0, (\sqrt{0.1})^2)$ and are independent of $\eta_i(t_{i,j})$.

For comparison, in each simulated dataset, we fit the model using four different estimators of the covariance function:

Method I. Working independence (WI) estimator (see Lin and Carroll 2001).

Method II. Our proposed method.

Method III. True covariance function.

Method IV. Misspecified ARMA(1, 1) method. We assume the covariance function $R(s,t)$ for $\varepsilon_i(t) = \eta_i(t) + \zeta_i(t)$ is of the form

$$R(s,t) = R_\eta(s,t)\mathrm{I}(s \neq t) + \sigma^2(t)\mathrm{I}(s = t), \tag{29}$$

where $R_\eta(s,t) = \sigma(s)\sigma(t)\gamma\rho^{|s-t|}$, has the misspecified ARMA(1, 1) structure. The parameters $\gamma \in [0,1]$ and $\rho \in [0,1]$ in the Method IV are estimated using the quasi-maximum likelihood method (QL), see Fan et al. (2007). For a fair comparison, we use the same bandwidth $h_1 = 0.1$ when estimating $\boldsymbol{\alpha}(\cdot)$ for all three estimators. The bandwidth h_2 in (14) is selected by the plug-in method (Fan and Gijbels 1996). Throughout the simulations and the real data examples in the next section, we use the Epanechnikov kernel, and the tuning parameter λ_n in (13) is selected automatically by the package 'ssfcov', which is based on the tuning parameter selection method in smoothing splines.

Table 1 summarises of the results over 200 simulations. We assess the performance of different approaches by calculating the bias and standard errors of 200 estimates. In the table, 'Bias' represents the median of the 200 estimates subtracting the true value, 'SD' represents the median absolute deviation of the 200 estimates divided by a factor of 0.6745, and 'RE' of a current estimator represents the relative efficiency between the oracle estimator (Method III) and the current estimator, which is defined as SD^2(oracle estimator)/SD^2(current estimator). Intuitively, if the relative efficiency of a method is larger, the SD of the coefficient using this method is smaller, so this method will be better (i.e. more efficient).

From Table 1, all parameter 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 terms of the efficiency, theoretically, the oracle estimator (Method III) using the true covariance function should be the best, and our proposed approach (Method II) should be much better than that using both working independence structure which ignored $\eta_i(t)$ (Method I) and the misspecified ARMA(1, 1) correlation structure for $\eta_i(t)$ (Method IV). The results in Table 1 agreed with this conjecture. For example, our proposed method has 30% efficiency gain over the estimator assuming working independence (Method I) for $\widehat{\beta}_1$, while the Method IV has only 21% efficiency gain over the estimator assuming working independence (Method I) for $\widehat{\beta}_1$.

Table 1. *(Simulation 1)* Compare the performance of $\widehat{\boldsymbol{\beta}}$ using different methods (with $n = 200$).

	$\widehat{\beta}_1$			$\widehat{\beta}_2$		
Method	Bias	SD	RE	Bias	SD	RE
Method I	−.0021	.0255	.33	−.0002	.0378	.22
our Method II	.0013	.0184	.63	−.0033	.0294	.37
Method III	.0005	.0146	1.0	−.0008	.0179	1.0
Method IV	−.0031	.0198	.54	−.0034	.0298	.36

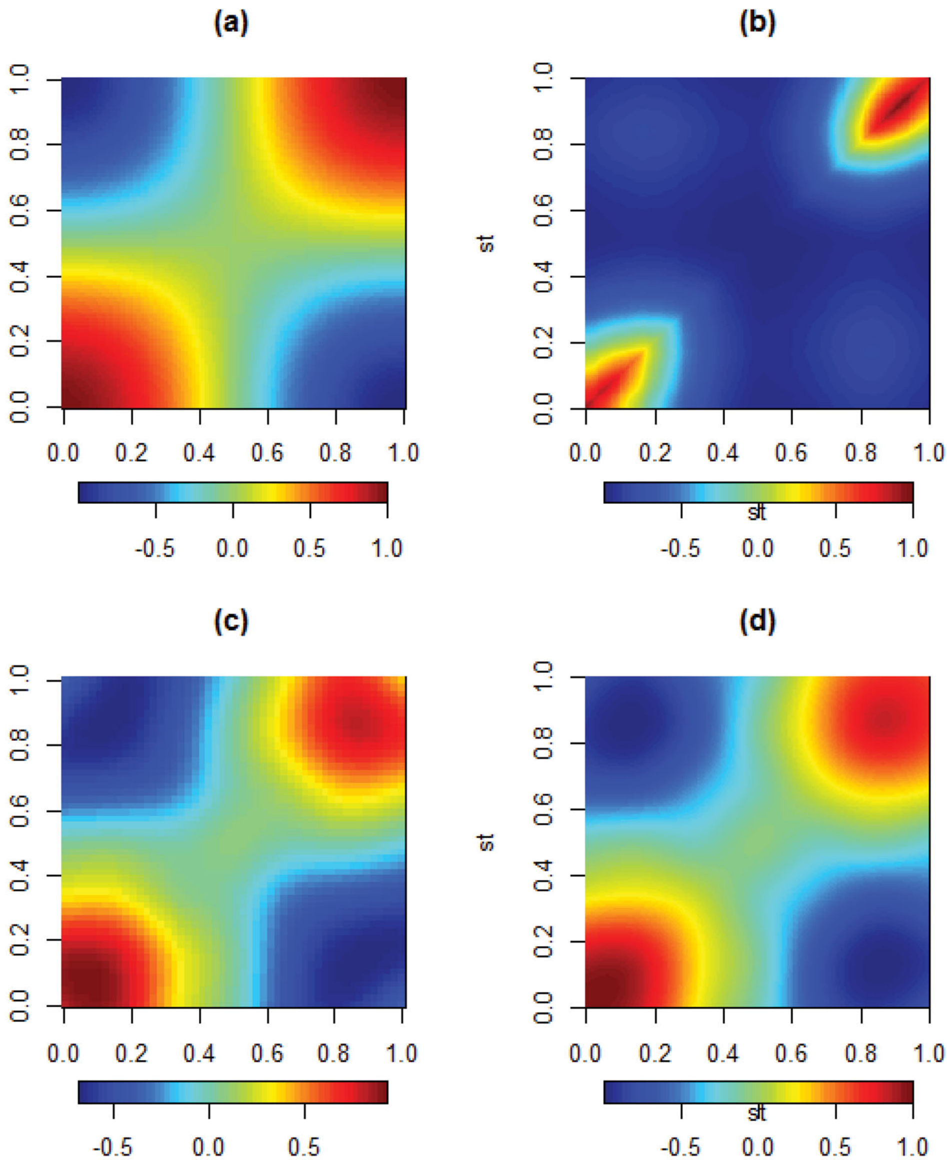

Figure 1. *(Simulation 1)* Panel (a): true covariance function $R_\eta(\cdot,\cdot)$; panel (b): estimated covariance function $\widehat{R}_\eta(\cdot,\cdot)$ using ARMA(1, 1) model for $\eta_i(t)$ (Method IV); panel (c): estimated covariance function $\widehat{R}_\eta(\cdot,\cdot)$ using our proposed method (Method II); panel (d): adjusted covariance function estimator $\widetilde{R}_\eta(\cdot,\cdot)$ based on (21) for one simulated data (with $n = 200$).

Figure 1 shows the advantage of our proposed method. From the plot, the true covariance function $R_\eta(s,t)$ for $\eta_i(t)$ in Figure 1(a) is very complicated and cannot be estimated by any parametric model such as AR(1) or ARMA(1, 1) model. Figure 1(b) shows the parametric ARMA(1, 1) estimator of $R_\eta(s,t)$ for $\eta_i(t)$ and Figure 1(c) shows the nonparametric estimator $\widehat{R}_\eta(s,t)$ using our proposed method for one simulated data. Figure 1(d) shows the adjusted estimator $\widetilde{R}_\eta(s,t)$ in (21), which is very close to $\widehat{R}_\eta(s,t)$ in Figure 1(c) but is

Table 2. *(Simulation 2)* Compare the performance of $\widehat{\boldsymbol{\beta}}$ using different methods (with $n = 100$).

	$\widehat{\beta}_1$			$\widehat{\beta}_2$		
Method	Bias	SD	RE	Bias	SD	RE
Method I	.0055	.0517	0.45	.0043	.0819	0.60
our Method II	.0052	.0415	0.70	.0005	.0742	0.74
Method III	−.0013	.0348	1.0	.0050	.0637	1.0
Method IV	−.0008	.0341	1.04	.0057	.0625	1.04

positive definite. From the plot, our proposed method captures the structure of covariance function very well, and the estimators $\widehat{R}_\eta(s,t)$ and $\widetilde{R}_\eta(s,t)$ are obviously consistent.

5.2. Simulation 2

In simulation 2, we study the robustness of our proposed method. The data are generated with the same setup as in the previous simulation, except that $\eta_i(t)$ is a Gaussian process with the covariance function $R_\eta(s,t)$ in (29). We set the sample size n be 100 and the marginal variance $\sigma^2(t) = 1$, and set $\rho = 0.35$ and $\gamma = 0.75$ in (29). We apply the semiparametric regression methods assuming working independence (Method I), nonparametric covariance (Method II), true covariance (Method III) and ARMA$(1,1)$ covariance structure for $\eta_i(t)$ (Method IV) to the simulated data and repeat the simulation 200 times. The selection methods of h_1, h_2 and λ_n are the same as in the previous simulation. The results for estimating $\boldsymbol{\beta}$ are summarised in Table 2.

From Table 2, again all parameter estimators are asymptotically unbiased since the biases are much smaller than the standard errors. When we compare the efficiencies of the estimates, theoretically, since the true correlation structure for $\eta_i(t)$ is ARMA$(1,1)$ model, the efficiency of the estimator using ARMA$(1,1)$ correlation structure (Method IV) should be close to the oracle estimator (Method III) using the true covariance function, and both of them should be more efficient than our proposed approach (Method II). The estimator using working independence structure (Method I) should be the least efficient. Again the results in Table 2 agreed with this conjecture. Our proposed estimator (Method II), on the other hand, performs reasonably well: the standard errors of our estimators are much smaller than those of the working independence estimator (Method I), for example, our proposed method has 25% efficiency gain over the estimator assuming working independence (WI) for $\widehat{\beta}_1$.

6. Real data example

6.1. Case 1: multi-centre AIDS cohort data

We now present an application of our proposed method to the Multi-Center AIDS Cohort study. The dataset comprises the information of 283 subjects who were infected with human immunodeficiency virus (HIV) during the study in year 1984–1991. A total of $N = 1817$ observations were made in this study, with between 1 and 14 observations per subject. This dataset was also analysed by Fan et al. (2007) and Huang, Wu, and Zhou (2002). Our target is to describe the trend in mean CD4 (cluster of differentiation 4) percentage depletion over time and to evaluate the effects of smoking, pre-HIV infection CD4 percentage, and age at infection on the mean CD4 percentage.

We take the response Y to be CD4 cell percentage, X_1 to be the standardised variable for PreCD4, Z_1 to be the smoking status (1 for a smoker and 0 for a nonsmoker) and Z_2 to be the standardised variable for age. The observation time is divided by 6 so that the rescaled observation time t is in between 0 and 1. Now consider a semiparametric varying-coefficient partially linear model

$$Y(t) = \alpha_1(t) + \alpha_2(t)X_1(t) + \beta_1 Z_1(t) + \beta_2 Z_2(t) + \eta(t) + \zeta(t). \tag{30}$$

We apply a multifold cross-validation method to select a bandwidth h_1 for $\boldsymbol{\alpha}(t)$. After partitioning the data into 14 groups, we fit model (30) for the data excluding the kth-group for each $k = 1, \ldots, 14$. For the computational issue, we minimise the cross-validation (CV) score on a rough grid $h_1 \in \{0.5\kappa^b : b = 0, \ldots, 12\}$, with $\kappa = 0.8$. The resulting optimal bandwidth is $h_1^{\text{opt}} = 0.054$. We can estimate $\alpha_1(t)$ and $\alpha_2(t)$ more precisely by choosing different bandwidths 0.054 and 0.081 for $\alpha_1(t)$ and $\alpha_2(t)$ respectively to avoid the under-smoothness of $\alpha_2(t)$, see Fan and Zhang (1999). As for the estimation of $\sigma^2(t)$, this is a one-dimensional kernel regression of the squared residuals. In this application, we directly use the plug-in bandwidth selector (Fan and Gijbels 1996) and choose the bandwidth $h_2^{\text{opt}} = 0.080$.

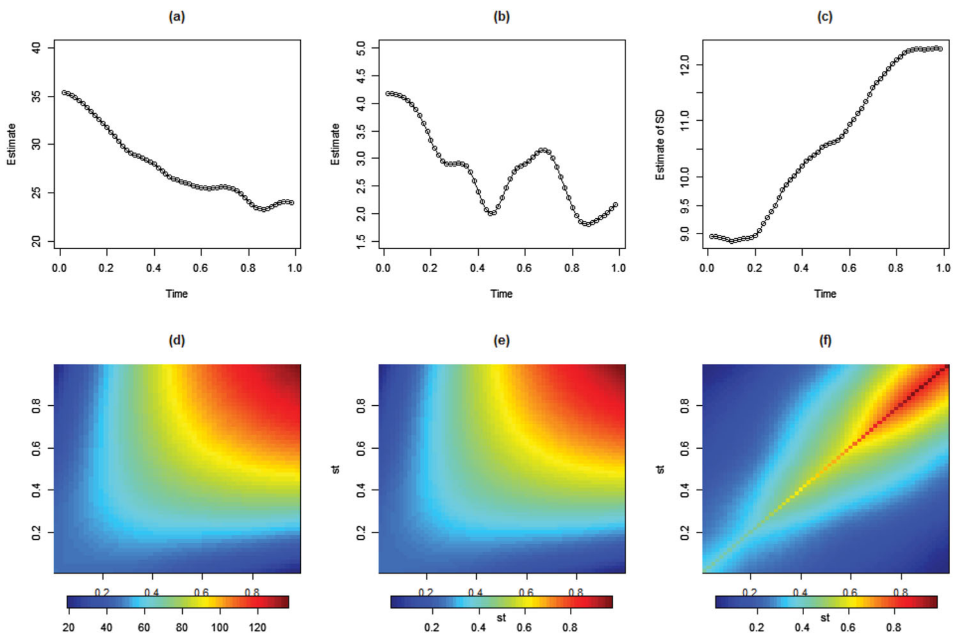

Figure 2. (*Real data case 1*) Panel (a): estimate of $\alpha_1(t)$; panel (b): estimate of $\alpha_2(t)$; panel (c): estimate of $\sigma(t)$. In panels (a)–(c), circles ($\circ$) represent the estimates of the functions at observation times, with lines ($-$) connecting them. Panel (d): estimated covariance function $\widehat{R}_\eta(\cdot,\cdot)$ using our proposed method for $\eta_i(t)$; panel (e): adjusted covariance function estimator $\widetilde{R}_\eta(\cdot,\cdot)$ based on (21); panel (f): estimated covariance function $\widehat{R}_\eta(\cdot,\cdot)$ using ARMA(1, 1) model for $\eta_i(t)$.

Table 3. *(Real data case 1)* Compare estimates of $\boldsymbol{\beta}$ using different methods.

	Method I	Method II	Method IV
$\widehat{\beta}_1$	0.686	0.563	0.734
$\widehat{\beta}_2$	−0.529	0.096	0.045

The resulting estimate of $\boldsymbol{\alpha}(t)$ is depicted in Figures 2(a,b). The intercept function decreases with time, implying an overall trend of CD4 cell percentage is decreasing over time. The trend for $\alpha_2(t)$ implies that the impact of PreCD4 on CD4 cell percentage decreases gradually during the first 3 years after infection and then increases a bit. The resulting estimate of $\sigma(t)$ in Figure 2(c), indicates that $\sigma(t)$ seems to be increasing as time increases. This means predicting the CD4 percentage becomes harder and harder over time. Figure 2(d–f) show the estimates $\widehat{R}_\eta(\cdot,\cdot)$, $\widetilde{R}_\eta(\cdot,\cdot)$ using our proposed method and the estimate $\widehat{R}_\eta(\cdot,\cdot)$ based on ARMA$(1,1)$ structure for $\eta_i(t)$, which characterises the within-subject dependence. From the plot, our estimate of the covariance function $R_\eta(\cdot,\cdot)$ is quite different from that using the ARMA$(1,1)$ structure for $\eta_i(t)$, so the within-subject correlation is misspecified if we just use quasi-maximum likelihood method (QL).

Table 3 shows the estimates of β_1 and β_2 with three different covariance structures: working independence (Method I), our proposed method (Method II) and ARMA$(1,1)$ covariance structure for $\eta_i(t)$ by Fan et al. (2007) (Method IV). The estimates $(\widehat{\beta}_1,\widehat{\beta}_2)$ using Method II and Method IV are quite different, so the ARMA$(1,1)$ covariance structure for $\eta_i(t)$ is misspecified. Finally, the effects of age on the mean CD4 percentage is negative if we assume working independence (Method I) but is positive if we use more efficient estimation method.

6.2. *Case 2: progesterone data*

We now apply the proposed methods to the longitudinal progesterone data. Progesterone, which is a reproductive hormone, is responsible for normal fertility and menstrual cycling. A longitudinal hormone study on progesterone collected urine samples from 34 healthy women (control group) in a menstrual cycle on alternative days, see Sowers et al. (1998). A total of 492 observations were made in this study, with between 11 and 28 observations per subject. The observation time is divided by 30 so that the rescaled observation time t is in between 0 and 1.

Similar to the procedure by Zhang, Lin, Raz, and Sowers (1998), a logarithmic transformation is applied on the progesterone level to make the data homoscedastic. We take the response to be the difference between the jth log-transformed progesterone level measured at rescaled time $t_{i,j}$ and the individual's average log-transformed progesterone level. For the ith subject, let X_i and Z_i denote age and body mass index, both of which are standardised to have mean 0 and standard deviation 1. We consider the semiparametric model

$$Y_i(t_{i,j}) = \alpha_1(t_{i,j}) + \beta_1 X_i(t_{i,j}) + \beta_2 Z_i(t_{i,j}) + \eta_i(t_{i,j}) + \zeta_i(t_{i,j}). \tag{31}$$

We apply a multifold cross-validation method to select a bandwidth h_1 for $\alpha_1(t)$. We partition the data into 17 groups, and each group contains 2 subjects. We fit model (31) for the

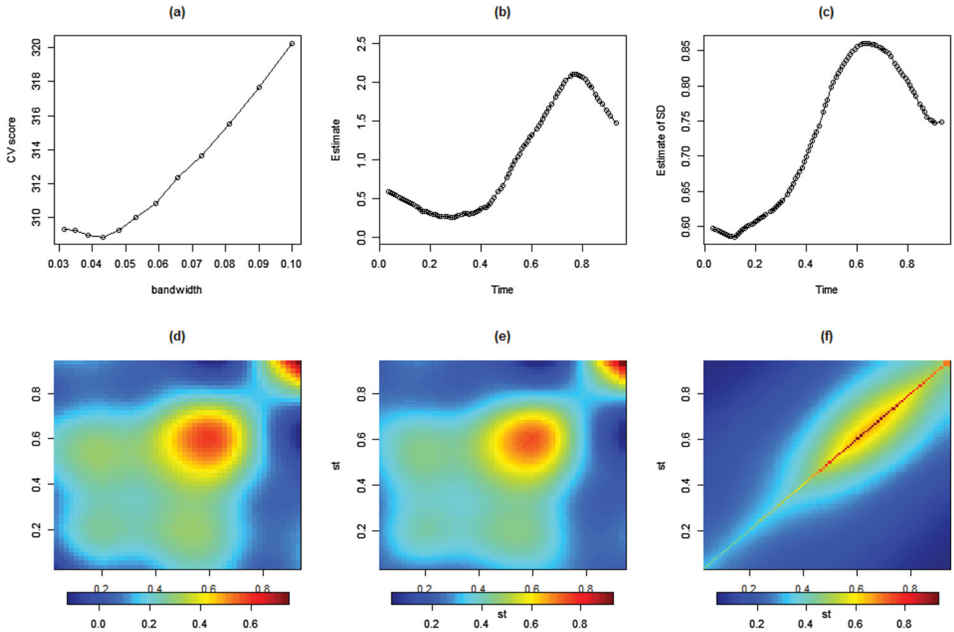

Figure 3. *(Real data case 2)* Panel (a): cross-validation score; panel (b): estimate of $\alpha_1(t)$; panel (c): estimate of $\sigma(t)$; panel (d): estimated covariance function $\widehat{R}_\eta(\cdot,\cdot)$ using our proposed method for $\eta_i(t)$; panel (e): adjusted covariance function estimator $\widetilde{R}_\eta(\cdot,\cdot)$ based on (21); panel (f): estimated covariance function $\widehat{R}_\eta(\cdot,\cdot)$ using ARMA(1, 1) model for $\eta_i(t)$.

data excluding the kth-group for each $k = 1, \ldots, 17$. For the computational issue, we minimise the cross-validation (CV) score on a rough grid $h_1 \in \{0.1\kappa^b : b = 0, \ldots, 12\}$, here we choose $\kappa = 0.9$. From Figure 3(a), the resulting optimal bandwidth is $h_1^{\text{opt}} = 0.043$. As for the estimation of $\sigma^2(t)$, this is a one-dimensional kernel regression of the squared residuals. In this application we directly use the plug-in bandwidth selector (Fan and Gijbels 1996) and choose the bandwidth $h_2^{\text{opt}} = 0.096$.

The resulting estimate of $\alpha_1(t)$ is depicted in Figures 3(b). The shape of intercept function implies an overall trend of progesterone level over time. The resulting estimate of $\sigma(t)$ in Figure 3(c), indicates that $\sigma(t)$ seems to be increasing as time increases from 0 to 0.6. This means predicting the progesterone level becomes harder and harder over time. Figure 3(d–f) show the estimates $\widehat{R}_\eta(\cdot,\cdot)$, $\widetilde{R}_\eta(\cdot,\cdot)$ using our proposed method and the estimate $\widehat{R}_\eta(\cdot,\cdot)$ based on ARMA(1, 1) structure for $\eta_i(t)$, which characterises the within-subject dependence. From the plot, our estimate of the covariance function is quite complicated and different from the ARMA(1, 1) or AR(1) covariance structure for $\eta_i(t)$, so the within-subject correlation is misspecified if we use quasi-maximum likelihood method (QL).

Table 4 shows the estimates of β_1 and β_2 with four covariance structures: working independence (Method I), our proposed method (Method II), ARMA(1, 1) covariance structure for $\eta_i(t)$ (Method IV), and Mixed model by Zhang et al. (1998), which is a special case of (6). The values of $(\widehat{\beta_1}, \widehat{\beta_2})$ between Method II and Method IV are quite different, showing that the ARMA(1, 1) structure for $\eta_i(t)$ is misspecified. Finally, the effects of age

Table 4. *(Real data case 2)* Compare estimates of $\boldsymbol{\beta}$ using different methods.

	Method I	Method II	Method IV	Mixed model (Zhang et al. 1998)
$\widehat{\beta}_1$	0.082	−0.009	0.068	0.925
$\widehat{\beta}_2$	−0.117	−0.187	−0.099	−2.913

on the mean progesterone level is negative if we use our proposed method (Method II), but is positive if we assume other three methods.

7. Conclusion

In this article we proposed a class of nonparametric models for the covariance function of longitudinal data at irregular or subject-specific time points. We further developed an estimation procedure for $\sigma^2(t) = R(t,t)$ using local linear regression, estimation procedure for $R_\eta(t_1, t_2)$ using regularisation approach, and estimation procedure for regression coefficients $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\beta}$ using profile weighted least squares. We also showed that the varying-coefficient temporal mixed-effects model is a good approximation of the semiparametric varying-coefficient partially linear model.

Although we just focused on analysing the longitudinal data at irregular or subject-specific time points, our proposed method can also be applied to equally-spaced or regular time points. In this balanced case, directly estimating the 'covariance matrix' of the errors may be better than estimating the 'covariance function' of the errors, but our method continues to work reasonably well and will not lose much efficiency.

Several issues are desirable for future research. First, in the presence of outliers, one should consider a robust method to estimate $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\beta}$ instead of using profile weighted least squares. Second, in the simulations and real data examples we just checked the plots of the estimated covariance functions and argued that the covariance function for $\eta_i(t)$ cannot be estimated by any parametric model such as AR(1) or ARMA(1, 1) model. It is better to develop a new procedure to test whether the covariance structure for $\eta_i(t)$ has a parametric form such as ARMA(1, 1) model. This research topic is beyond the scope of this article and further research is needed.

Acknowledgements

The authors thank the two referees for insightful comments and suggestions, and Yongsu Lee for the helps in R programming techniques.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

C. Zhang's research is supported by the U.S. NSF Division of Mathematical Sciences [grant number DMS– 1712418], Wisconsin Alumni Research Foundation, and National Natural Science Foundation of China [grant number 11690014]. Hulin Wu's reserach is supported by grants NIAID [grant number RO1 AI087135], and CPRIT [grant number RP170668].

References

Bosq, D. (2000), *Linear Processes in Function Spaces: Theory and Applications*, New York: Springer.

Cai, T.T., and Yuan, M. (2010), 'Nonparametric Covariance Function Estimation for Functional and Longitudinal Data'. Technical Report.

Diggle, P.J., Heagerty, P., Liang, K.Y., and Zeger, S.L. (2002), *Analysis of Longitudinal Data* (2nd ed.), New York: Oxford University Press.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Application*, London: Chapman & Hall.

Fan, J., and Huang, T. (2005), 'Profile Likelihood Inferences on Semiparametric Varying Coefficient Partially Linear Models', *Bernoulli*, 11, 1031–1059.

Fan, J., Huang, T., and Li, R. (2007), 'Analysis of Longitudinal Data with Semiparametric Estimation of Covariance Function', *Journal of the American Statistical Association*, 102, 632–641.

Fan, J., and Wu, Y. (2008), 'Semiparametric Estimation of Covariance Matrixes for Longitudinal Data', *Journal of the American Statistical Association*, 103, 1520–1533.

Fan, J, and Zhang, W.Y. (1999), 'Statistical Estimation in Varying Coefficient Models', *The Annals of Statistics*, 27, 1491–1518.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications* (Vol. II, 2nd ed.), New York: John Wiley and Sons.

Hall, P., Müller, H.G., and Wang, J.L. (2006), 'Properties of Principal Component Methods for Functional and Longitudinal Data Analysis', *The Annals of Statistics*, 34, 1493–1517.

Hall, P., Müller, H.G., and Yao, F. (2008), 'Modelling Sparse Generalized Longitudinal Observations with Latent Gaussian Processes', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 703–723.

Huang, J.Z., Wu, C.O., and Zhou, L. (2002), 'Varying-coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements', *Biometrika*, 89, 111–128.

Li, Y. (2011), 'Efficient Semiparametric Regression for Longitudinal Data with Nonparametric Covariance Estimation', *Biometrika*, 98, 355–370.

Lin, X., and Carroll, R.J. (2001), 'Semiparametric Regression for Clustered Data Using Generalized Estimating Equations', *Journal of the American Statistical Association*, 96, 1045–1056.

Mack, Y.P., and Silverman, B.W. (1982), 'Weak and Strong Uniform Consistency of Kernel Regression Estimates', *Zeitschrift Für Wahrscheinlichkeitstheorie Und Verwandte Gebiete*, 61, 405–415.

Milton, A., and Irene, A.S. (1972), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York: Dover.

Sowers, M., Randolph, J.F., Crutchfield, M., Jannausch, M.L., Shapiro, B., Zhang, B., and La Pietra, M. (1998), 'Urinary Ovarian and Gonadotropin Hormone Levels in Premenopausal Women With Low Bone Mass', *Journal of Bone and Mineral Research*, 13, 1191–1202.

Wang, N., Carroll, R.J., and Lin, X. (2005), 'Efficient Semiparametric Marginal Estimation for Longitudinal/Clustered Data', *Journal of the American Statistical Association*, 100, 147–157.

Wu, W.B., and Pourahmadi, M. (2003), 'Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data', *Biometrika*, 90, 831–844.

Wu, H.L., and Zhang, J.T. (2002a), 'Local Polynomial Mixed-effects Models for Longitudinal Data', *Journal of the American Statistical Association*, 97, 883–897.

Wu, H.L., and Zhang, J.T. (2002b), 'The Study of Long-term HIV Dynamics Using Semiparametric Nonlinear Mixed-effects Models', *Statistics in Medicine*, 21, 3655–3675.

Wu, H.L., and Zhang, J.T. (2006), *Nonparametric Regression Methods for Longitudinal Data: Mixed-effects Modeling Approaches*, New York: Wiley.

Yao, F., Müller, H.G., and Wang, J.L. (2005), 'Functional Data Analysis for Sparse Longitudinal Data', *Journal of the American Statistical Association*, 100, 577–590.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998), 'Semiparametric Stochastic Mixed Models for Longitudinal Data', *Journal of the American Statistical Association*, 93, 710–719.

Appendix. Conditions and proofs of main results

The following technical conditions are imposed. They are not the weakest possible conditions, but they are imposed to facilitate the proofs. For notational convenience, given a vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^T$, define $|\boldsymbol{\alpha}| = (|\alpha_1|, \dots, |\alpha_p|)^T$.

C1. The density function $f_T(\cdot)$ is Lipschitz-continuous and bounded away from 0 and infinity. The function $K(\cdot)$ is a symmetric density function with a compact support.
C2. $nh_1^8 \to 0$ and $nh_1^2/\{\log(n)\}^3 \to \infty$ as $n \to \infty$.
C3. $\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{X}(t)^T\}$ and $\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{Z}(t)^T\}$ are Lipschitz-continuous.
C4 $\boldsymbol{\alpha}(t)$ has a continuous second derivative.
C5. J_i has a finite moment-generating function in some neighbourhood of the origin. In addition, $\mathrm{E}\{\|\boldsymbol{X}(t)\|^4\} + \mathrm{E}\{\|\boldsymbol{Z}(t)\|^2\} < \infty$.
C6. $\sigma_\zeta^2(\cdot)$ has a continuous second derivative, and $\mathrm{E}\{\varepsilon(t)^{4+\delta_0}\} < \infty$ for some $\delta_0 > 0$.
C7. $h_2\{\log(n)\}^2 \to 0$ and $nh_2/\log(n) \to \infty$ as $n \to \infty$.

Proof of Proposition 4.1.: First, By Fan et al. (2007, Theorem 1 and result (A.1), (A.2)), we have

$$\widehat{\boldsymbol{\beta}}^{\mathrm{ini}} - \boldsymbol{\beta}^\circ = O_{\mathrm{P}}(n^{-1/2}) \tag{A1}$$

$$\sup_{t\in\mathcal{T}} \left|\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(t) - \boldsymbol{\alpha}^\circ(t) + \left[\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{X}(t)^T\}\right]^{-1}\left[\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{Z}(t)^T\}\right](\boldsymbol{\beta} - \boldsymbol{\beta}^\circ)\right| = O_{\mathrm{P}}(c_n), \tag{A2}$$

where $c_n = h_1^2 + \{-\log(h_1)/(nh_1)\}^{1/2}$, $\boldsymbol{\beta}^\circ$ and $\boldsymbol{\alpha}^\circ(t)$ are true parametric and nonparametric components. Since $\mathcal{T}$ is compact, together with condition C3 and (A1),

$$\sup_{t\in\mathcal{T}} \left|\left[\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{X}(t)^T\}\right]^{-1}\left[\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{Z}(t)^T\}\right](\widehat{\boldsymbol{\beta}}^{\mathrm{ini}} - \boldsymbol{\beta}^\circ)\right| = O_{\mathrm{P}}(n^{-1/2}). \tag{A3}$$

Utilizing triangle inequality and (A2), (A3) yields

$$\sup_{t\in\mathcal{T}} \left|\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}^{\mathrm{ini}}}(t) - \boldsymbol{\alpha}^\circ(t)\right| = O_{\mathrm{P}}(c_n).$$

Finally, by the definition of $\widehat{\varepsilon}(t)$ in (12) and $\varepsilon(t)$, we have

$$\sup_{t\in\mathcal{T}} \left|\widehat{\varepsilon}(t) - \varepsilon(t)\right| = \sup_{t\in\mathcal{T}} \left|\boldsymbol{X}(t)^T\left\{\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}^{\mathrm{ini}}}(t) - \boldsymbol{\alpha}^\circ(t)\right\} + \boldsymbol{Z}(t)^T\left(\widehat{\boldsymbol{\beta}}^{\mathrm{ini}} - \boldsymbol{\beta}^\circ\right)\right| = O_{\mathrm{P}}(c_n).$$

This completes the proof. ∎

Proof of Proposition 4.2.: Without loss of generality, suppose $\mathcal{T} = [0, 1]$ and $f_T(\cdot)$ is uniform density on $[0, 1]$. For case (i), by Condition C5, suppose $\mathrm{E}(e^{tJ_i}) \leq C$ for some $t > 0$. By Markov's inequality, $\mathrm{P}(J_i \geq a) = \mathrm{P}(e^{tJ_i} \geq e^{ta}) \leq \mathrm{E}(e^{tJ_i})/e^{ta} \leq Ce^{-ta}$, then

$$\begin{aligned}
&\mathrm{P}\left(\max_{1\leq i\leq n} J_i \geq d\log(n)\right) \\
&= 1 - \mathrm{P}\left(\max_{1\leq i\leq n} J_i < d\log(n)\right) \\
&= 1 - \mathrm{P}(J_i < d\log(n))^n \\
&= 1 - \left[1 - \mathrm{P}(J_i \geq d\log(n))\right]^n \\
&\leq 1 - \left(1 - Ce^{-td\log(n)}\right)^n \\
&= 1 - \left(1 - Cn^{-td}\right)^n \to 0,
\end{aligned}$$

when d is large enough such that $td > 1$. So $\max_{1\le i\le n} J_i = O_P(\log(n))$. Next let $\{T_i^{(1)}, \ldots, T_i^{(J_i)}\}$ be order statistics of $\{t_{i,1}, \ldots, t_{i,J_i}\}$. According to statement by Feller (1971, p.42), for a given J_i,

$$P(T_i^{(2)} - T_i^{(1)} > 2h_2, \ldots, T_i^{(J_i)} - T_i^{(J_i-1)} > 2h_2) = \left[1 - 2h_2(J_i - 1)\right]_+^{J_i} \tag{A4}$$

where $[g]_+$ is the positive part of g. Since $\max_{1\le i\le n} J_i = O_P(\log(n))$, with probability tending to 1,

$$\left[1 - 2h_2(J_i - 1)\right]_+^{J_i} \ge \left[1 - 2dh_2 \log(n)\right]_+^{2d\log(n)}.$$

By condition C7, we have $[1 - 2dh_2 \log(n)]_+^{2d\log(n)} \to 1$, so it is unlikely for each individual to have more than two observations in the same neighbourhood $[t - h_2, t + h_2]$. Thus in what follows, the $\varepsilon_i^2(t_{i,j})$'s can be treated as independent similar to the proof of Theorem 1 in Fan and Wu (2008). By classic uniform convergence rates for the kernel smoother (Mack and Silverman 1982), we have

$$\sup_{t\in\mathcal{T}} \left|\overline{\sigma}^2(t) - \sigma^2(t)\right| = O_P(h_2^2 + \{-\log(h_2)/(nh_2)\}^{1/2}).$$

For case (ii), after we take $\alpha = 2$ in Theorem 4 according to Cai and Yuan (2010), we have

$$\lim_{D\to\infty} \limsup_{n\to\infty} P\left(\|\overline{R}_\eta - R_\eta\|_{L_2}^2 > D\left(\{\log(n)/(nm)\}^{4/5} + n^{-1}\right)\right) = 0$$

that means

$$\|\overline{R}_\eta - R_\eta\|_{L_2}^2 = O_P\left(\{\log(n)/(nm)\}^{4/5} + n^{-1}\right), \tag{A5}$$

where $m = E(J_i)$ is the expected number of observations within each subject. Since condition C5 implies that the first moment $m = E(J_i)$ is finite, (A5) becomes $\|\overline{R}_\eta - R_\eta\|_{L_2}^2 = O_P(\{\log(n)/n\}^{4/5})$, which is exactly (26). ■

Proof of Proposition 4.3.: First, note that $\tau \ge 0$, it is easy to show that $\widetilde{R}_\eta(s,t) = \sum_{k=1}^N \widehat{\lambda}_{(k)} I(\widehat{\lambda}_{(k)} > \tau)\widehat{\psi}_k(s)\widehat{\psi}_k(t)$ defined in (21) is positive definite since all the eigenvalues of $\widetilde{R}_\eta(s,t)$ are positive. Secondly, note that

$$\begin{aligned}\widetilde{R}(s,t) &= \widetilde{R}_\eta(s,t)I(s \ne t) + \max\{\widetilde{R}_\eta(t,t), \widehat{\sigma}^2(t)\}I(s = t)\\ &= \widetilde{R}_\eta(s,t) + \max\{0, \widehat{\sigma}^2(t) - \widetilde{R}_\eta(t,t)\}I(s = t).\end{aligned}$$

Obviously the second term, $\max\{0, \widehat{\sigma}^2(t) - \widetilde{R}_\eta(t,t)\}I(s = t)$, is positive semidefinite since it is diagonal and all the diagonal entries are nonnegative. So $\widetilde{R}(s,t)$ is the summation of a positive definite and a positive semidefinite covariance function, which is of course positive definite. ■