SCIENCE CHINA
Mathematics

• ARTICLES •

January 2011 Vol. 54 No. 1: 145–164
doi: 10.1007/s11425-010-4078-4

Joint semiparametric mean-covariance model in longitudinal study

MAO Jie & ZHU ZhongYi*

Department of Statistics, Fudan University, Shanghai 200433*, China*
Email: 071025012*@fudan.edu.cn, zhuzy@fudan.edu.cn*

Received October 19, 2009; accepted March 30, 2010

Abstract Semiparametric regression models and estimating covariance functions are very useful for longitudinal study. To heed the positive-definiteness constraint, we adopt the modified Cholesky decomposition approach to decompose the covariance structure. Then the covariance structure is fitted by a semiparametric model by imposing parametric within-subject correlation while allowing the nonparametric variation function. We estimate regression functions by using the local linear technique and propose generalized estimating equations for the mean and correlation parameter. Kernel estimators are developed for the estimation of the nonparametric variation function. Asymptotic normality of the the resulting estimators is established. Finally, the simulation study and the real data analysis are used to illustrate the proposed approach.

Keywords generalized estimating equation, kernel estimation, local linear regression, modified Cholesky decomposition, semiparametric varying-coefficient partially linear model

MSC(2000): 62F12, 62G08, 62G20

Citation: Mao J, Zhu Z Y. Joint semiparametric mean-covariance model in longitudinal study. Sci China Math, 2011, 54(1): 145–164, doi: 10.1007/s11425-010-4078-4

1 Introduction

Longitudinal data are often characterized by the dependence of repeated observations over time within the same subject [1]. Observations within the same subject tend to be correlated. So one key issue for longitudinal study involves the estimation of the covariance structure. Good estimation of the covariance structure improves the efficiency regression coefficients. There are several approaches for estimating a covariance structure. Diggle and Verbyla [2] provided a nonparametric estimator for the covariance structure without assuming stationarity, but their estimator is not guaranteed to be positive definite. To heed the positive-definiteness constraint, the modified Cholesky decomposition is considered by Pourahmadi [15, 16]. Wu and Pourahmadi [20] proposed nonparametric estimation of covariance matrices using a two-step estimation procedure. However, they only dealt with balanced data or nearly balanced longitudinal data. And longitudinal data are commonly collected at irregular and possibly subject-specific time points. To address these challenges, Ye and Pan [21] proposed an approach for joint modeling of the mean and covariance structures of longitudinal data within the framework of generalized estimation equations, but parametrically modeling the covariance structure. Parametric modeling may be restrictive for some applications and nonparametric modeling may be too flexible to make concise conclusions. To overcome these drawbacks, Fan et al. [4] and Fan and Wu [6] developed a semiparametric estimator for the

*Corresponding author

covariance structure by modeling the variance function nonparametrically and the correlation structure parametrically, but their methods involved constrained optimization.

In addition, in the analysis of longitudinal data, the partially linear model may not always be appropriate. For example, in a Multi-Center AIDS Cohort study [4], whether or not PreCD4 has a constant effect over time is unclear. Hence it is of potential interest in this study to allow a time-varying term. Then we study the semiparametric varying-coefficient partially linear model

$$y(t) = x'(t)\alpha(t) + z'(t)\beta + \varepsilon(t), \tag{1}$$

where $y(t)$ is the response variable, $x(t)$ and $z(t)$ are the covariate vectors at time t, $\alpha(t)$ comprises p unknown smooth functions, β is a q-dimensional unknown parameter vector and $E\{\varepsilon(t)|x(t), z(t)\} = 0$. It includes many useful models proposed in the literature. It is a useful extension of the commonly used partially linear model studied by Zeger and Diggle [22], Zhang et al. [23], He et al. [9, 10] and Fan and Li [5]. The partially linear model quantifies the time effect by allowing the intercept coefficient to vary over time but not the coefficients of the other covariate variables. It is also an extension of a useful varying-coefficient model [7, 8, 11]. Varying-coefficient models retain the flexibility of nonparametric models, but lose the explanatory power of parametric models. Therefore, model (1) has received increasing attention recently. It has been studied by Fan and Huang [3] in the case of i.i.d. observation and Martinussen and Scheike [13], Fan et al. [4] and Wang et al. [19] for longitudinal data.

In this article, we use the modified Cholesky decomposition to decompose the within-subject covariance function and then model the within-subject correlation in terms of covariate. We estimate regression functions using the local linear technique and propose generalized estimating equations for the mean and correlation parameter, and kernel estimators for variation. The estimation procedures introduced here are similar to those described by Ye and Pan [21] and Fan et al. [4]. However, this approach differs from the one proposed by Ye and Pan in that we study the semiparametric varying-coefficient partially linear model and fit the covariance structure by a semiparamtric model which allows a data analyst to easily incorporate prior information about the within-subject correlation structure; it is different from the approach developed by Fan et al. in that we adopt the modified Cholesky decomposition to decompose the covariance structure, then impose parametric within-subject correlation and allow the nonparametric variation function. The positive definiteness for the resulting estimate is therefore guaranteed. Moreover, the proposed approach does not need the true correlation structure and the normal distribution assumption. Hence the proposed method is robust and stable, which will be shown in the simulation study. However, if the correlation structure is correctly specified, Fan et al.'s method may be better than the proposed method.

The remainder of the paper is organized as follows. The estimation procedure for regression coefficients and covariance function is presented in Section 2. In Section 3 we discuss the asymptotic properties of the proposed procedures. Simulation study and real data analysis are given in Sections 4 and 5. Technical proofs appear in the Appendix.

2 Estimation of regression parameters and covariance

Assume that we have a sample of n subjects. For the ith subject, $i = 1, \ldots, n$, the response variable $y_i(t_{ij})$ and the covaraite vector $\{x_i(t_{ij}), z_i(t_{ij})\}$ are collected at time points $t = t_{ij}$, $j = 1, \ldots, J_i$, where J_i is the total number of observation on the i-th subject. Let $\varepsilon_i = (\varepsilon_i(t_{i1}), \ldots, \varepsilon_i(t_{iJ_i}))'$. Denote $\mu_{ij}(\beta, \alpha) = x_i'(t_{ij})\alpha(t_{ij}) + z_i'(t_{ij})\beta$ and $\mu_i(\beta, \alpha) = (\mu_{i1}(\beta, \alpha), \ldots, \mu_{iJ_i}(\beta, \alpha))'$. Write $r_{ij} = r_{ij}(\beta, \alpha) = y_i(t_{ij}) - \mu_{ij}(\beta, \alpha)$ and $r_i = (r_{i1}, \ldots, r_{iJ_i})'$. In addition, suppose $\mathrm{Cov}(\varepsilon_i|X_i) = \Sigma_i$ is the covariance matrix of y_i. Before going further, let $X_i = (x_i(t_{i1}), \ldots, x_i(t_{iJ_i}))'$, $X = (X_1', \ldots, X_n')'$, $Z_i = (z_i(t_{i1}), \ldots, z_i(t_{iJ_i}))'$, $Z = (Z_1', \ldots, Z_n')'$, $y_i = (y_i(t_{i1}), \ldots, y_i(t_{iJ_i}))'$, $y = (y_1', \ldots, y_n')'$.

In what follows, we first estimate $\alpha(t)$ by using the local linear regression technique, and then impose parametric within-subject correlation after decomposing the covariance structure by using the modified

Cholesky decomposition. We further construct estimating equations for the mean, generalized autoregressive parameters and develop a nonparametric estimator for innovation variances, respectively.

2.1 Local linear estimate of $\alpha(t)$

For a given β, let $y^*(t) = y(t) - z'(t)\beta$. Then the model (1) can be written as

$$y^*(t) = x'(t)\alpha(t) + \varepsilon(t).$$

This is a nonparametric regression problem. Thus we can use a nonparametric regression technique to estimate $\alpha(t)$. Here we use the local linear regression technique [4]. For any t in a neighborhood of t_0, it follows from Taylor's expansion that

$$\alpha_l(t) \approx \alpha_l(t_0) + \dot{\alpha}_l(t_0)(t - t_0) = a_l + b_l(t - t_0),$$

for $l = 1, \ldots, p$. Let $K(\cdot)$ be a kernel function and h be a bandwidth, $\dot{\alpha}(u) = d\alpha(u)/du$. We can find local parameters $(a_1, \ldots, a_p, b_1, \ldots, b_p)$ that minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\left[y_i^*(t_{ij}) - \sum_{l=1}^{p}\{a_l + b_l(t - t_0)x_{il}(t_{ij})\}\right]^2 K_h(t_{ij} - t_0), \tag{2}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$. The local linear estimate for $\alpha(t_0)$ is then simply $\widehat{\alpha}(t_0, \beta) = (\widehat{a}_1, \ldots, \widehat{a}_p)'$. By simple calculations, we have

$$\widehat{\alpha}(t_0, \beta) = (\mathbf{I}_p, \mathbf{0}_p)\Lambda^{-1}(t_0)\Psi(t_0)y^*, \tag{3}$$

where $\mathbf{I}_p$ is the $(p \times p)$ identity matrix and $\mathbf{0}_p$ is the $(p \times p)$ matrix with all entries being 0, $\Lambda(t_0) = \sum_{i=1}^{n}\Gamma_i'(t_0)W_i(t_0)\Gamma_i(t_0)$, $\Psi(t_0) = (\Gamma_1'(t_0)W_1(t_0), \ldots, \Gamma_n'(t_0)W_n(t_0))$,

$$\Gamma_i(t_0) = \begin{pmatrix} x_i'(t_{i1}) & (t_{i1} - t_0)x_i'(t_{i1}) \\ \vdots & \vdots \\ x_i'(t_{iJ_i}) & (t_{iJ_i} - t_0)x_i'(t_{iJ_i}) \end{pmatrix},$$

$W_i(t_0) = \mathrm{diag}(K_h(t_{i1} - t_0), \ldots, K_h(t_{iJ_i} - t_0))$, $y_i^* = (y_i^{*'}(t_{i1}), \ldots, y_i^{*'}(t_{iJ_i}))'$ and $y^* = (y_1^{*'}, \ldots, y_n^{*'})'$. With t_0 being replaced by t_{ij}, we can obtain

$$\widehat{\alpha}(t_{ij}, \beta) = (\mathbf{I}_p', \mathbf{0}_p')\Lambda^{-1}(t_{ij})\Psi(t_{ij})y^*.$$

Let

$$S_i = \begin{pmatrix} x_i'(t_{i1})(\mathbf{I}_p, \mathbf{0}_p)\Lambda^{-1}(t_{i1})\Psi(t_{i1}) \\ \vdots \\ x_i'(t_{iJ_i})(\mathbf{I}_p, \mathbf{0}_p)\Lambda^{-1}(t_{iJ_i})\Psi(t_{iJ_i}) \end{pmatrix} \tag{4}$$

and

$$\widehat{m}_i(\beta) = (x_i'(t_{i1})\widehat{\alpha}(t_{i1}, \beta), \ldots, x_i'(t_{iJ_i})\widehat{\alpha}(t_{iJ_i}, \beta))'.$$

Hence, the estimate $\widehat{m}_i(\beta)$ can be expressed as

$$\widehat{m}_i(\beta) = S_i y^*. \tag{5}$$

2.2 Estimation of β and Σ

In (5), we have estimated the nonparametric term. Then we replace the mean model with the following one:

$$\widehat{\mu}_{ij} = x'(t_{ij})\widehat{\alpha}(t_{ij}, \beta) + z'(t_{ij})\beta. \tag{6}$$

Thus the mean model only involves the unknown parameter β.

Next, we adopt the modified decomposition to decompose Σ_i as follows:

$$T_i \Sigma_i T_i' = D_i,$$

where the diagonal entries of T_i are 1 and the below-diagonal entries of T_i are the negatives of the coefficients of $\widetilde{y}_i(t_{ij}) = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{ijk}(y_i(t_{ik}) - \mu_{ik})$, the linear least-squares predictor of $y_i(t_{ij})$ based on its predecessors $y_i(t_{i(j-1)}), \ldots, y_i(t_{i1})$, and the diagonal entries of D_i are the prediction error variances $\sigma^2(t_{ij}) = \mathrm{Var}(\widetilde{\varepsilon}(t_{ij}))$, where $\widetilde{\varepsilon}(t_{ij}) = y_i(t_{ij}) - \widetilde{y}_i(t_{ij})$, for $0 \leqslant j \leqslant J_i, 1 \leqslant i \leqslant n$. Throughout this paper we refer to ϕ_{ijk} as generalized autoregressive parameters and $\sigma^2(t_{ij})$ as innovation variances. Since ϕ_{ijk} is unconstrained, we consider the model

$$\phi_{ijk} = z_{ijk}'\gamma, \tag{7}$$

where z_{ijk} is the $(d \times 1)$ vector of covariates and γ is the associated parameter. The covariates z_{ijk} may contain baseline covariates, polynomials in time and their interactions as well. When we use polynomials in time to model the generalized autoregressive parameters, the covariates may take the forms $z_{ijk} = (1, (t_{ij} - t_{ik}), \ldots, (t_{ij} - t_{ik})^{d-1})'$, provided that the within-subject correlation only depends on the elapsed time.

Similar to [21], we propose the following generalized estimating equations for β and γ respectively:

$$U_1(\beta) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i'}{\partial \beta}\right) \Sigma_i^{-1}(y_i - \mu_i) = 0, \tag{8}$$

$$U_2(\gamma) = \sum_{i=1}^{n} \left(\frac{\partial \widetilde{r}_i'}{\partial \gamma}\right) D_i^{-1}(r_i - \widetilde{r}_i) = 0, \tag{9}$$

where $\widetilde{r}_i$ is $(J_i \times 1)$ vector with jth components $\widetilde{r}_{ij} = E(r_{ij} | r_{i1}, \ldots, r_{i(j-1)}) = \sum_{k=1}^{j-1} \phi_{ijk} r_{ik}$. We can show that $D_i = \mathrm{diag}(\sigma_{i1}, \ldots, \sigma_{iJ_i})$ in $U_2(\gamma)$ is actually the covariance matrix of $r_i - \widetilde{r}_i$. In contrast to the mean equation $U_1(\beta)$, r_i in $U_2(\gamma)$ plays a role similar to that of the responses in the estimation of γ, viewed as working responses when modeling the generalized autoregressive parameters. So the idea behind equation (9) is in agreement with equation (8).

With μ_i being replaced by $\widehat{\mu}_i = (\widehat{\mu}_{i1}, \ldots, \widehat{\mu}_{iJ_i})$, and r_i and $\widetilde{r}_i$ by ξ_i and $\widetilde{\xi}_i$ respectively, equations (8) and (9) become

$$\widetilde{U}_1(\beta) = \sum_{i=1}^{n} \left(\frac{\partial \widehat{\mu}_i'}{\partial \beta}\right) \Sigma_i^{-1}(y_i - \widehat{\mu}_i) = 0, \tag{10}$$

$$\widetilde{U}_2(\gamma) = \sum_{i=1}^{n} \left(\frac{\partial \widetilde{\xi}_i'}{\partial \gamma}\right) D_i^{-1}(\xi_i - \widetilde{\xi}_i) = 0, \tag{11}$$

where $\partial \widehat{\mu}_i' / \partial \beta$ is a $(p \times J_i)$ matrix with jth column $z_i(t_{ij}) + Z'\Psi'(t_{ij})\Lambda^{-1}(t_{ij})(\mathbf{I}_p, \mathbf{0}_p)'$, ξ_i and $\widetilde{\xi}_i$ are $(J_i \times 1)$ vectors with jth components $\xi_{ij} = y_i(t_{ij}) - \widehat{\mu}_{ij}$ and $\widetilde{\xi}_{ij} = \sum_{k=1}^{j-1} \phi_{ijk}\xi_{ik}$, and $\partial \widetilde{\xi}_i' / \partial \gamma$ is a $(d \times J_i)$ matrix with jth column $\partial \widetilde{\xi}_{ij}' / \partial \gamma = \sum_{k=1}^{j-1} \xi_{ik} z_{ijk}$. In fact, the task behind equation (11) is equivalent to finding γ which minimizes

$$\sum_{i=1}^{n} \xi_i' T_i' D_i^{-1} T_i \xi_i.$$

In addition, note that

$$\sigma^2(t) = E[\widehat{\varepsilon}^2(t)|t].$$

A natural estimator for $\sigma^2(t)$ is the kernel estimator

$$\widehat{\sigma}^2(t) = \frac{\sum_{i=1}^n \sum_{j=1}^{J_i} (r_{ij} - \widetilde{r}_{ij})^2 K_{h_1}(t - t_{ij})}{\sum_{i=1}^n \sum_{j=1}^{J_i} K_{h_1}(t - t_{ij})}, \tag{12}$$

where $K_{h_1}(x) = h_1^{-1} K(x/h_1)$, $K(x)$ is a kernel density function and h_1 is a bandwidth. With r_{ij} being replaced by ξ_{ij}, (12) becomes

$$\widehat{\sigma}^2(t) = \frac{\sum_{i=1}^n \sum_{j=1}^{J_i} (\xi_{ij} - \widetilde{\xi}_{ij})^2 K_{h_1}(t - t_{ij})}{\sum_{i=1}^n \sum_{j=1}^{J_i} K_{h_1}(t - t_{ij})}. \tag{13}$$

Four remarks are worth mentioning. First, (10) is similar to the estimating equations in [14] and [21] except that it here includes the time-varying coefficient term. Second, instead of using the maximum-likelihood approach in [14], (10) and (11) are derived in the spirit of generalized estimating equations which do not require the normal distribution assumption. Third, (13) is different from the estimator for variance function in [4] in that it is the nonparametric estimator for innovation variance function. Finally, compared to the parametric model for covariance structure in [14] and [21], a semiparametric model is proposed to fit the covariance function, a parametric model for the within-subject correlation and a kernel estimator for variation.

By solving the equations (10) and (11), we can obtain the estimators of β and γ. Since functions $\widetilde{U}_1(\beta)$ and $\widetilde{U}_2(\gamma)$ are linear in β and γ respectively, the estimator of β and γ have the closed forms. Therefore, given Σ_i, the estimator of β can be updated by

$$\widehat{\beta} = \left\{ \sum_{i=1}^n (Z_i - S_i Z)' \Sigma_i^{-1} (Z_i - S_i Z) \right\}^{-1} \sum_{i=1}^n (Z_i - S_i Z)' \Sigma_i^{-1} (y_i - S_i y). \tag{14}$$

On the other hand, given β and D_i, the estimator of γ can be updated by

$$\widehat{\gamma} = \left\{ \sum_{i=1}^n Z'(i) D_i^{-1} Z(i) \right\}^{-1} \sum_{i=1}^n Z'(i) D_i^{-1} \widehat{r}_i, \tag{15}$$

where $Z(i) = (z(i,1), \ldots, z(i,J_i))'$ and $z(i,j) = \sum_{k=1}^{j-1} \widehat{r}_{ik} z_{ijk}$.

In summary, the algorithm below is used to calculate the parameter estimates.

Algorithm. **Step 1** Given an initial value $\gamma^{(0)}$ and $D_i^{(0)}$. $\Sigma_i^{(0)}$, the starting value of Σ_i, is obtained. Using (14), we can get $\beta^{(0)}$, the initial estimator of β.

Step 2 Use (15) to calculate the estimator $\gamma^{(1)}$ of γ and thus get $T_i^{(1)}$. And by (13), we can get $D_i^{(1)}$ and then $\Sigma_i^{(1)}$ using $\Sigma_i = T_i^{-1} D_i T_i^{-1'}$.

Step 3 Update the value $\beta^{(0)}$ by

$$\beta^{(1)} = \left\{ \sum_{i=1}^n (Z_i - S_i Z)' \Sigma_i^{(1)-1} (Z_i - S_i Z) \right\}^{-1} \sum_{i=1}^n (Z_i - S_i Z)' \Sigma_i^{(1)-1} (y_i - S_i y).$$

Step 4 Repeat Steps 2 and 3 until the convergence of the parameter estimates.

A convenient initial value for γ and D_i is $\gamma^{(0)} = 0$ and $D_i^{(0)} = I_{J_i \times J_i}$, where $I_{J_i \times J_i}$ is the $(J_i \times J_i)$ identity matrix. In other words, $I_{J_i \times J_i}$ is chosen as the starting value for the covariance matrix Σ_i. We have demonstrated in the simulation that the above algorithm has fast convergence. Thus its computational cost is much lower than constrained optimization. Finally, $\widehat{\alpha}(\cdot; \widehat{\beta})$ can be obtained by $\widehat{\alpha}(\cdot; \beta)$ with β being replaced by $\widehat{\beta}$.

2.3 Bandwidths selection

A question has not been addressed in the implementation of the foregoing procedure. That is how to select the bandwidths. Similar to [4], we use the MCV (multifold CV) method to choose the bandwidth of $\widehat{\alpha}(t)$. Specifically, we partition the data into Q groups, namely, $d_1, \ldots, d_Q$, each of which has approximately the same number of subjects. The cross-validation score is defined as the sum of residual squares,

$$\mathrm{CV}(h) = \sum_{k=1}^{Q} \sum_{i \in d_k} \sum_{j=1}^{J_i} \{y_i(t_{ij}) - \widehat{y}_{(d_k)}(t_{ij})\}^2,$$

where $\widehat{y}_{(d_k)}(t_{ij})$ is the fitted value for ith subject at observed time t_{ij} with data in d_k deleted, using working independence correlation. Then, we have the optimal bandwidth

$$h_{opt} = \arg\min_h \mathrm{CV}(h).$$

In addition, as $\widehat{\varepsilon}(t_{ij})$, $j = 1, \ldots, J_i$ and $i = 1, \ldots, n$, are independent, so we may choose an optimal bandwidth for $\widehat{\sigma}^2(t)$ using various existing bandwidth selectors for independent data. Here we adopt the plug-in bandwidth selector, details of which are given in [17].

3 Asymptotic properties

In this section we study the sampling properties of estimates. Here we consider the data as a random sample from the population process $\{y(t), x(t), z(t)\}, t \in [0, T]$. Assume that J_i, $i = 1, \ldots, n$, are i.i.d. random variables with $0 < E(J_i) < \infty$. Given J_i, t_{ij}, $j = 1, \ldots, J_i$, are independent and identically distributed according to a density $f(t)$.

Set $G(t) = Ex(t)x'(t)$, $\Psi(t) = Ex(t)z'(t)$ and write

$$\widetilde{X}_i = (\Psi(t_{i1})G^{-1}(t_{i1})x_i(t_{i1}), \ldots, \Psi(t_{iJ_i})G^{-1}(t_{iJ_i})x_i(t_{iJ_i}))'.$$

Let γ_0 and $\sigma_0^2(t)$ be the true values of γ and innovation variance at time t when the model (7) is correct. Denote the estimates of γ and $\sigma^2(t)$ by $\widehat{\gamma}$ and $\widehat{\sigma}^2(t)$ at time t. Write $D_{0i} = \mathrm{diag}(\sigma_0^2(t_{i1}), \ldots, \sigma_0^2(t_{iJ_i}))$, $\widehat{D}_i = \mathrm{diag}(\widehat{\sigma}^2(t_{i1}), \ldots, \widehat{\sigma}^2(t_{iJ_i}))$, $\widehat{\Sigma}_i^{-1} = T_i'(\widehat{\gamma})\widehat{D}_i^{-1}T_i(\widehat{\gamma})$ and $\Sigma_i^{-1}(\gamma) = T_i'(\gamma)D_{0i}^{-1}T_i(\gamma)$, where $i = 1, \ldots, n$. Then, we have the following results.

Theorem 1. *Under the regularity conditions* 1–5 *in the appendix, if the following matrices A and B exist, and there exists a matrix Σ_i such that $\widehat{\Sigma}_i \xrightarrow{P} \Sigma_i$, then, as $n \to \infty$,*

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} N(0, A^{-1}BA^{-1}),$$

where

$$A = E\{(Z_1 - \widetilde{X}_1)\Sigma_1^{-1}(Z_1 - \widetilde{X}_1)\},$$
$$B = E\{(Z_1 - \widetilde{X}_1)'\Sigma_1^{-1}Cov(\varepsilon_1|X_1, Z_1)\Sigma_1^{-1}(Z_1 - \widetilde{X}_1)\}.$$

and

$$\widetilde{X}_1 = (\Psi(t_{11})G^{-1}(t_{11})x_1(t_{11}), \ldots, \Psi(t_{1J_1})G^{-1}(t_{1J_1})x_1(t_{1J_1}))'.$$

The technical condition that $\widehat{\Sigma}_i \xrightarrow{P} \Sigma_i$ seems strong. However, it is reasonable under the regularity conditions 1–10 and additional condition that model (7) is correct. As mentioned in Section 2, the estimation of β and $\alpha(t)$ depends on the estimation of γ and $\sigma^2(t)$. On the other hand, the estimation of γ and $\sigma^2(t)$ in turn depends on the estimation of β and $\alpha(t)$. As shown in the proof of Theorems 1 and 2 in the Appendix, the initial estimates of β and $\alpha(t)$ are consistent and then the initial estimates of γ and $\sigma^2(t)$ are also consistent which is presented in the proof of Theorems 3 and 4. Repeat this process at convergence of the estimates, and then we can get $\widehat{\gamma} \xrightarrow{P} \gamma_0$ and $\widehat{D}_i \xrightarrow{P} D_{i0}$. In addition, it follows from the additional condition that $\mathrm{Cov}(\varepsilon_1|X_1, Z_1) = \Sigma_1(\gamma_0)$. In this case

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} N(0, B_0^{-1}),$$

where $B_0 = E\{(Z_1 - \widetilde{X}_1)'\mathrm{Cov}(\varepsilon_1|X_1, Z_1)^{-1}(Z_1 - \widetilde{X}_1)\}$.

To give the asymptotic result of estimator $\widehat{\alpha}(t)$, let

$$\upsilon_i = \int v^i K(v)dv, \quad \kappa_i = \int v^i K^2(v)dv \quad \text{and} \quad \ddot{\alpha}(u) = d^2\frac{\alpha(u)}{du^2}.$$

Theorem 2. *Suppose that conditions of Theorem* 1 *hold. Then the following results hold: if* $n \to \infty$, $nh^5 = O(1)$, *then*

$$\sqrt{nh}\Big(\widehat{\alpha}(t) - \alpha(t) - \frac{1}{2}\upsilon_2 h^2 \ddot{\alpha}(t)\Big) \xrightarrow{\mathcal{L}} N\Big(0, \frac{\kappa_0}{f(t)E(J_1)}\widetilde{\sigma}^2(t)G^{-1}(t)\Big),$$

where $\widetilde{\sigma}^2(t) = E[\varepsilon^2(t)|t]$.

From Theorem 2, the efficiency of $\widehat{\beta}$ does not play an important role in the asymptotic bias and variance of $\widehat{\alpha}(t)$. This implies that we may choose a bandwidth by using one of the existing bandwidth selectors used for independent data. This result is identical to that obtained in [4].

Theorem 3. *Assume that model* (7) *is correct and the conditions of Theorems* 1 *and* 2 *hold. Further suppose that* $\sup_{t\in[0,T]} |\widehat{\sigma}^2(t) - \sigma_0^2(t)| = O_P(h)$. *Under the regularity conditions* 6–8 *in the appendix, then, as* $n \to \infty$, $\widehat{\gamma}$ *is asymptotically normally distributed as follows*:

$$\sqrt{n}(\widehat{\gamma} - \gamma_0) \xrightarrow{\mathcal{L}} N(0, I^{-1}(\gamma_0)V(\gamma_0)I^{-1}(\gamma_0)),$$

where $I(\gamma) = (I_{ij}(\gamma))_{(q+d)\times(q+d)}$, $V(\gamma) = (v_{ij})_{(q+d)\times(q+d)}$,

$$I_{ij}(\gamma) = E\bigg[\mathrm{tr}\bigg\{\Sigma_1^{-1}(\gamma)\bigg(\frac{\partial\Sigma_1(\gamma)}{\partial\gamma_i}\Sigma_1^{-1}(\gamma)\frac{\partial\Sigma_1(\gamma)}{\partial\gamma_j} + \frac{\partial\Sigma_1(\gamma)}{\partial\gamma_j}\Sigma_1^{-1}(\gamma)\frac{\partial\Sigma_1(\gamma)}{\partial\gamma_i}\bigg)\bigg\}\bigg]$$

and

$$v_{ij} = E\bigg\{\bigg(\varepsilon_1'\frac{\partial\Sigma_1^{-1}(\gamma)}{\partial\gamma_i}\varepsilon_1\varepsilon_1'\frac{\partial\Sigma_1^{-1}(\gamma)}{\partial\gamma_j}\varepsilon_1\bigg) - \mathrm{tr}\bigg(\frac{\partial\Sigma_1^{-1}(\gamma)}{\partial\gamma_i}\Sigma_1(\gamma)\bigg)\mathrm{tr}\bigg(\frac{\partial\Sigma_1^{-1}(\gamma)}{\partial\gamma_j}\Sigma_1(\gamma)\bigg)\bigg\}.$$

Theorem 4. *Suppose that conditions of Theorems* 1–3 *hold. Under conditions* 9 *and* 10 *in the Appendix, if* $c < nh_1^5 < C$, *and* $c < h/h_1 < C$ *for some positive constants* c *and* C, *then, as* $n \to \infty$,

$$\sqrt{nh_1}(\widehat{\sigma}^2(t) - \sigma_0^2(t) - b(t)) \xrightarrow{\mathcal{L}} N(0, v(t)),$$

where

$$b(t) = \frac{h_1^2}{2}\bigg\{\ddot{\sigma}_0^2(t) + \frac{2\dot{\sigma}_0^2(t)\dot{f}(t)}{f(t)}\upsilon_2\bigg\}$$

and

$$v(t) = \frac{\mathrm{Var}(\widetilde{\varepsilon}^2(t))\kappa_0}{f(t)E(J_1)}.$$

Note that the asymptotic bias and variance of $\widehat{\sigma}^2(t)$ depend on the efficiency of $\widehat{\gamma}$ and further on the efficiency of $\widehat{\beta}$ and $\widehat{\alpha}(t)$. This can be explained by noting the definition of $\widetilde{\varepsilon}(t_{ij})$.

In Theorem 1, the asymptotic covariance matrix of $\widehat{\beta}$ can be estimated by $\widehat{A}^{-1}\widehat{B}\widehat{A}^{-1}$, where

$$\widehat{A} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - S_iZ)'\widehat{\Sigma}_i^{-1}(Z_i - S_iZ),$$

$$\widehat{B} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - S_i Z)'\widehat{\Sigma}_i^{-1}\widehat{r}_i\widehat{r}_i'\widehat{\Sigma}_i^{-1}(Z_i - S_i Z),$$
$$\widehat{r}_i = y_i - X_i\widehat{\beta} - \widehat{\alpha}_i(\widehat{\beta}), \quad \widehat{\Sigma}_i^{-1} = T_i'(\widehat{\gamma})\widehat{D}_i^{-1}T_i(\widehat{\gamma})$$

and $\widehat{D}_i = \mathrm{diag}(\widehat{\sigma}^2(t_{i1}), \ldots, \widehat{\sigma}^2(t_{iJ_i}))$. In Theorem 3, the asymptotic covariance matrix of $\widehat{\gamma}$ can be estimated by $\widehat{I}^{-1}\widehat{V}\widehat{I}^{-1}$, where $\widehat{I} = (\widehat{I}_{ij})$ and $\widehat{V} = (\widehat{v}_{ij})$. More explicitly, $\widehat{I}_{ij}$ and $\widehat{v}_{ij}$ are given by

$$\widehat{I}_{ij} = \frac{1}{n}\sum_{l=1}^{n}\mathrm{tr}\left[\widehat{\Sigma}_l^{-1}\left(\frac{\partial\widehat{\Sigma}_l}{\partial\gamma_i}\widehat{\Sigma}_l^{-1}\frac{\partial\widehat{\Sigma}_l}{\partial\gamma_j} + \frac{\partial\widehat{\Sigma}_l}{\partial\gamma_j}\widehat{\Sigma}_l^{-1}\frac{\partial\widehat{\Sigma}_l}{\partial\gamma_i}\right)\right]$$

and

$$\widehat{v}_{ij} = \frac{4}{n}\sum_{l=1}^{n}\left\{\left(\widehat{r}_l'\frac{\partial\widehat{\Sigma}_l^{-1}}{\partial\gamma_i}\widehat{r}_l\widehat{r}_l'\frac{\partial\widehat{\Sigma}_l^{-1}}{\partial\gamma_j}\widehat{r}_l\right) - \mathrm{tr}\left(\frac{\partial\widehat{\Sigma}_l^{-1}}{\partial\gamma_i}\widehat{\Sigma}_l\right)\mathrm{tr}\left(\frac{\partial\widehat{\Sigma}_l^{-1}}{\partial\gamma_j}\widehat{\Sigma}_l\right)\right\}.$$

The proofs of Theorem 1–4 are given in the Appendix.

4 Simulation study

In this section, we are going to investigate four issues concerning the proposed estimators via Monte Carlo simulations. First, how does the performance of the proposed estimator $\widehat{\beta}$ compare with that obtained by quasi-likelihood (QL) and minimum generalized variance method (MGV)? Second, how does the proposed estimator $\widehat{\beta}$ perform when the noise does not follow the normal distribution? Third, how well are the finite-sample performances of the proposed estimate $\widehat{\alpha}(t)$, $\widehat{\gamma}$ and $\sigma^2(t)$? Finally, how does the proposed method compare with the parametric model (18) to be specified later? Details of QL and MGV are given in [4].

4.1 Simulation models

In each simulation run, a simple random sample of 50 subjects is generated according to the model

$$y(t_{ij}) = x'(t_{ij})\alpha(t_{ij}) + z'(t_{ij})\beta + \varepsilon(t_{ij}) \quad (j = 1, \ldots, J_i, i = 1, \ldots, n). \tag{16}$$

To generate the observation times, each individual has a set of 'scheduled' time points $\{0, 1, \ldots, 12\}$, and each scheduled time, except time 0, has a probability 20% being skipped. The actual observation time is a random perturbation of the scheduled time: a uniform (0, 1) random variable is added to non-skipped scheduled time to obtain the actual observation time.

In the simulation $\varepsilon(t)$ is generated from a Gaussian process with mean 0 and four covariance structures which are presented as follows:

• Σ_1: variance function $\sigma^2(t) = 0.5\exp(t/12)$ and ARMA(1, 1) correlation structure $(\mathrm{Corr}(\varepsilon(s), \varepsilon(t)) = \eta\rho^{|t-s|}, s \neq t)$ with $(\eta, \rho) = (0.85, 0.6)$.

• Σ_2: the variance function is as in Σ_1 and exchangeable correlation structure with correlation $\rho = 0.5$.

• Σ_3: the same variance function as in Σ_1, but unstructured correlation structure which is described as follows: $\mathrm{Corr}(y_{ij}, y_{i(j-k)}) = 0.7$, $0 < j-k \leqslant 3$; $\mathrm{Corr}(y_{ij}, y_{i(j-k)}) = 0.5$, $3 < j-k \leqslant 6$; $\mathrm{Corr}(y_{ij}, y_{i(j-k)}) = 0.3$, $6 < j-k \leqslant 9$; $\mathrm{Corr}(y_{ij}, y_{i(j-k)}) = 0.2$, $9 < j-k \leqslant 12$, where j is the actual order of t_{ij} in the time sequence when there is no skipped observations.

• Σ_4: proposed working structure with innovation variance function $\sigma^2(t) = 0.5\exp(t/12)$ and $\gamma = (0.6817, -0.3410, 0.0515, -0.0024)'$.

We let $\alpha(t)$ and β be two-dimensional in our simulation. We further set $x_1(t) = 1$ to include an intercept term. The covariates are chosen as follows: for a given t, $(x_2(t), z_1(t))'$ follows the bivariate normal distribution with mean 0, variance 1, and correlation 0.5, and $z_2(t)$ follows Bernoulli-distributed random variable with success probability 0.5 and independent of $x_2(t)$ and $z_1(t)$. In this simulation, set $\beta = (1, 2)'$, $\alpha_1(t) = \sqrt{t/12}$ and $\alpha_2(t) = \sin(2\pi t/12)$.

Our following simulation results are all based on 1000 independent repetitions.

4.2 Performance of $\widehat{\beta}$

To demonstrate the efficiency improvement, we also estimate β using working independence correlation structure and true correlation structure in which the parameters are set to be the true value. The estimate using the true covariance is the most efficient estimate and serves as a benchmark.

The kernel function involved in the local linear modeling is taken to be Epanechnikov kernel $K(t) = 0.75(1-t^2)_+$. After tuning, the bandwidth of $\widehat{\alpha}(t)$ is chosen to be 1.5. Since the efficiency of $\widehat{\beta}$, $\widehat{\alpha}(t)$ and $\widehat{\gamma}$ affect the asymptotic bias and variance of $\widehat{\sigma}^2(t)$, we need to choose a bandwidth for $\widehat{\sigma}^2(t)$ in each simulation. Here we directly use the plug-in bandwidth selector to choose a bandwidth. In addition, cubic polynomial in lag and time is used to model autoregressive coefficients, namely, $z_{ijk} = (1, (t_{ij}-t_{ik}), (t_{ij}-t_{ik})^2, (t_{ij}-t_{ik})^3)'$.

Here we compare the proposed method with Fan et al.'s method [4] in terms of relative efficiency of the estimators of the mean parameters β. The relative efficiency of $\widehat{\beta}_k$ $(k=1,2)$ is the ratio of the MSE of Fan et al.'s estimator $\widehat{\beta}_k^{FHL}$ [4] to the MSE of the proposed estimator $\widehat{\beta}_k^{MZ}$:

$$e(\widehat{\beta}_k) = \mathrm{MSE}(\widehat{\beta}_k^{FHL})/\mathrm{MSE}(\widehat{\beta}_k^{MZ}).$$

Table 1 summarizes the results obtained based on Σ_1–Σ_4. In the table, 'Bias' represents the sample average over 1000 estimates subtracting the true value of β, 'SD' represents the sample standard deviation over 1000 estimates and 'MSE' represents the sample mean squared error over 1000 estimates. Table 1 shows that the standard deviation of the estimator of β is significantly reduced by incorporating the covariance matrix. When the true covariance structure is Σ_1, the values for $e(\widehat{\beta}_k)$ are all in the range 0.92–0.93 if Fan et al.'s method uses ARMA(1, 1) for correlation structure, namely, the working correlation

Table 1 Performance of $\widehat{\beta}$

		$\widehat{\beta}_1$			$\widehat{\beta}_2$		
	Method	Bias	SD	MSE	Bias	SD	MSE
Σ_1	True	−0.1260	3.4479	0.1189	−0.0989	6.3652	0.4049
	Independence	−0.2637	4.6919	0.2200	0.1409	8.1190	0.6585
	QL (ARMA(1, 1))	−0.1869	3.5035	0.1230	−0.0698	6.4229	0.4122
	MGV (ARMA(1, 1))	−0.1335	3.5822	0.1284	−0.1242	6.4989	0.4221
	QL (AR(1))	−0.1871	3.8580	0.1488	−0.2152	6.9245	0.4795
	MGV (AR(1))	−0.1960	3.9554	0.1564	−0.2402	7.1467	0.5018
	Proposed method	−0.1949	3.6613	0.1343	−0.0311	6.6607	0.4432
Σ_2	True	0.0568	3.4067	0.1160	0.2361	5.8950	0.3477
	Independence	0.0807	5.0586	0.2557	0.2620	8.3835	0.7023
	MGV (ARMA(1, 1))	−0.0595	4.3273	0.1871	0.2605	7.5440	0.5692
	MGV (AR(1))	−0.0485	5.4377	0.2954	0.1500	10.1262	1.0246
	Proposed method	0.0546	3.5574	0.1265	0.2437	6.2938	0.3963
Σ_3	True	−0.0427	2.3384	0.0546	0.0853	4.1746	0.1742
	Independence	0.1313	4.7873	0.2291	−0.2124	8.3542	0.6977
	MGV (ARMA(1, 1))	−0.0284	3.3913	0.1149	−0.1409	5.8721	0.3447
	MGV (AR(1))	−0.0749	4.1627	0.1732	−0.0896	7.3054	0.5332
	Proposed method	0.0141	2.9796	0.0887	0.0075	5.1402	0.2640
Σ_4	True	−0.2511	4.2853	0.1841	−0.3017	7.6823	0.5905
	Independence	−0.1340	5.4154	0.2932	−0.3037	9.4405	0.8913
	MGV (ARMA(1, 1))	−0.3561	4.7670	0.2283	−0.1176	8.6292	0.7440
	MGV (AR(1))	0.5078	5.9994	0.3621	−0.1177	10.3940	1.0794
	Proposed method	−0.2957	4.3391	0.1890	−0.3240	7.7709	0.6043

* Values in the columns of Bias, SD and MSE are multiplied by a factor of 100.

structure is correctly specified. That means that if the correlation structure is correctly specified, Fan et al.'s method is better than the proposed method. However, if Fan et al.'s method uses AR(1) for the correlation structure in this case, the values for $e(\widehat{\beta}_k)$ are all in the range 1.08–1.16. As Fan et al.'s method only uses ARMA(1, 1) and AR(1) for correlation, so when the true covariance matrix is Σ_2, Σ_3 or Σ_4, Fan et al.'s working correlation is misspecified. The values for $e(\widehat{\beta}_k)$ are all in the range 1.44–2.59 for the case of Σ_2, in the range 1.30–2.02 for the case of Σ_3 and in the range 1.21–1.91 for the case of Σ_4. Obviously, the proposed method yields better estimates for β than those obtained using MGV method if Fan et al.'s working correlation is misspecified. This further confirms that the proposed method provides robustness to misspecification of the covariance structure. In addition, we can find from Table 1 that the efficiency loss by using the proposed method for Σ_1 is smaller than the efficiency loss for Σ_4 by using MGV approach, which implies that the proposed method is more stable than MGV approach.

Next we evaluate the performance of $\widehat{\beta}$ when the noise follows marginal double exponential (DE) distribution with mean 0 and covariance matrix Σ_3 or Σ_4. All other components are simulated the same way as in the previous case. Simulation results over 1000 repetitions are reported in Table 2. From Table 2, we find that $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are very close to their corresponding true values. This confirms that the proposed approach does not require normal distribution assumption.

Now we test the accuracy of the proposed standard error formula for the simulated data from the normal distribution and double exponential distribution respectively. Table 3 depicts the simulations results for Σ_3 and Σ_4. In Table 3, 'SD' represents the sample standard deviation over 1000 estimates, 'SE' represents the average of 1000 estimated standard errors using the proposed standard error formula and 'Std' represents the standard deviation of these 1000 standard errors. Table 3 shows that the proposed standard error formula performs well for both correctly specified and misspecified cases.

4.3 Performance of $\widehat{\alpha}(t)$, $\widehat{\gamma}$ and $\widehat{\sigma}^2(t)$

We evaluate the performance of $\widehat{\alpha}(t)$. It can be assessed by the square root of the average squared errors (RASE),

$$\mathrm{RASE}(\alpha_j(\cdot)) = n_{\mathrm{grid}}^{-1}\left[\sum_{k=1}^{n_{\mathrm{grid}}}(\widehat{\alpha}_j(t_k)-\alpha_j(t_k))^2\right]^{1/2}, \tag{17}$$

where $\{t_k, k=1,\dots,n_{\mathrm{grid}}\}$ are the grid points at which $\alpha_j(\cdot)$ is estimated and $j=1,\dots,p$. Table 5 displays the mean and standard deviation (in parentheses) of RASE for Σ_3. Results for other cases are

Table 2 Performance of the proposed $\widehat{\beta}$ for double exponential distribution with covariance structure Σ_3 or Σ_4

	$\widehat{\beta}_1$			$\widehat{\beta}_2$		
	Bias	SD	MSE	Bias	SD	MSE
Σ_3	−0.0912	2.8420	0.0808	−0.2720	4.7992	0.2308
Σ_4	−0.0166	4.4605	0.1988	0.2950	7.3054	0.5340

* Values in the columns of Bias, SD and MSE are multiplied by a factor of 100.

Table 3 Standard errors

	$\widehat{\beta}_1$		$\widehat{\beta}_2$	
Noise	SD	SE (Std)	SD	SE (Std)
True covariance matrix: Σ_3				
N	0.0298	0.0286 (0.0034)	0.0514	0.0496 (0.0055)
DE	0.0284	0.0276 (0.0037)	0.0480	0.0481 (0.0057)
True covariance matrix: Σ_4				
N	0.0434	0.0415 (0.0050)	0.0777	0.0710 (0.0077)
DE	0.0434	0.0396 (0.0056)	0.0731	0.0685 (0.0084)

* N, normal; DE, double exponential.

Table 4 RASE for estimating $\alpha(\cdot)$

	N		DE	
WC	RASE ($\widehat{\alpha}_1(\cdot)$)	RASE ($\widehat{\alpha}_2(\cdot)$)	RASE ($\widehat{\alpha}_1(\cdot)$)	RASE ($\widehat{\alpha}_2(\cdot)$)
True	0.1279 (0.0516)	0.1146 (0.0299)	0.1278 (0.0514)	0.1137 (0.0313)
Independence	0.1308 (0.0524)	0.1155 (0.0305)	0.1292 (0.0521)	0.1149 (0.0321)
MGV (ARMA(1, 1))	0.1369 (0.0510)	0.1193 (0.0311)	0.1377 (0.0507)	0.1173 (0.0324)
MGV (AR(1))	0.1501 (0.0557)	0.1254 (0.0336)	0.1507 (0.0561)	0.1230 (0.0360)
PM	0.1380 (0.0511)	0.1150 (0.0302)	0.1275 (0.0503)	0.1134 (0.3140)

* N, normal; DE, double exponential; WC, working covariance.

Table 5 Estimates of γ

Noise		γ_1	γ_2	γ_3	γ_4
N	Estimate	0.6540	−0.3241	0.0486	−0.0023
	SD	0.0884	0.0682	0.0140	0.0008
	SE	0.0846	0.0649	0.0133	0.0008
	Std	0.0050	0.0044	0.0010	0.0001
DE	Estimate	0.6451	−0.3199	0.0479	−0.0022
	SD	0.0957	0.0729	0.0149	0.0009
	SE	0.0880	0.0659	0.0133	0.0008
	Std	0.0109	0.0083	0.0018	0.0001

* N, normal; DE, double exponential.

similar. From Table 5, we find that the result using the proposed method is slightly better than the other four working structures. This result is consistent with Fan et al. [4] and the theoretical result of Theorem 2 that the weight matrix does not affect the asymptotic bias and variance of $\alpha(t)$.

Next, we investigate the performance of the proposed $\widehat{\gamma}$. Our $\widehat{\gamma}$ is based on the estimation of other components for the simulated data from the normal distribution and double exponential distribution with covariance matrix Σ_4 respectively. Table 4 reports the mean of $\widehat{\gamma}$ in the first row and standard deviations in the second row. The third row gives SE of $\widehat{\gamma}$ with Std in the last row. From Table 4, we can see that $\widehat{\gamma}$ are all significant and close to their corresponding true ones. This agrees with our consistency result. Table 4 also indicates that the standard deviation of $\widehat{\gamma}$ is very close to the asymptotic standard errors using our asymptotic normality formulas. This means that the proposed standard error formula of $\widehat{\gamma}$ performs well.

Similar to (17), we can define the square root of average squared errors for $\widehat{\sigma}^2(t)$. Here we only consider the result for Σ_4. The mean and standard deviation (in parenthesis) of RASE are 0.0854 (0.0535) and 0.1202 (0.0778) for the simulated data from the normal distribution and double exponential distribution respectively. This shows that the estimator $\widehat{\sigma}^2(t)$ performs well.

4.4 Comparison with the parametric model

In this subsection, we demonstrate the flexibility of the proposed method by comparing its performance with the following parametric model for the covariance matrix

$$\phi_{ijk} = z'_{ijk}\gamma, \quad \log \sigma^2_{ij} = z'_{ij}\lambda, \tag{18}$$

which was considered by Pourahmadi [15, 16] and Ye and Pan [21]. To make comparison, we generate 1,000 datasets from (16) with the following:

- Case I: all components are the same as those specified in Section 4.1 except that the true covariance structure is the parametric model (18) with $\gamma = (0.6873, -0.3430, 0.0516, -0.0024)'$ and $\lambda = (-0.6851, -0.0251, 0.0153, -0.0007)'$.
- Case II: all components are the same as the case for Σ_4 but innovation variance function $0.05 + 16(T/12)^2$.

Table 6 Comparison with the parametric model

	$\widehat{\beta}_1$			$\widehat{\beta}_2$		
Method	Bias	SD	MSE	Bias	SD	MSE
Case I						
Parametric model	−0.1848	3.7312	0.1394	−0.0740	6.8830	0.4733
Proposed method	−0.1934	3.7710	0.1424	−0.0487	6.9514	0.4828
Case II						
Parametric model	−0.0138	4.3088	0.1853	0.2724	8.5648	0.7328
Proposed method	−0.0974	3.9104	0.1527	0.1754	7.4463	0.5537

* Values in the columns of Bias, SD and MSE are multiplied by a factor of 100.

To illustrate the flexibility of the proposed method, we fit data generated under the setting of Case I and Case II using the parametric model (18) and the proposed method. Simulation results are summarized in Table 6, in which the notation is the same as that given in Table 1. From Table 6, we can find that the proposed method is comparable to parametric model in Case I, but is much better in Case II. Thus the proposed method offers a good balance between flexibility and estimation efficiency.

5 Real data analysis

5.1 Background of the study

In this section, we apply the proposed method to the actual longitudinal data. The data are a subset of data from the Multi-Center AIDS Cohort study. They comprise the human immunodeficiency virus (HIV) status of 283 homosexual men who were infected with HIV during the following-up period of 1984–1991. The data set has been analyzed by Fan et al. [4] using the semiparametric varying-coefficient partially linear model. Details have been given in [4].

For the ith subject, let x_1 be PreCD4, z_1 be smoking status (1 for a smoker and 0 for a nonsmoker) and z_2 be age (x_1 and z_2 are standardized variables with mean 0 and standard deviation 1). We consider the response y as the CD4 cell percentage of a subject at distinct time points after HIV infection. Similar to [4], we fit the data by the following model

$$y(t_{ij}) = \alpha_1(t_{ij}) + \alpha_2(t_{ij})x_1(t) + \beta_1 z_1(t_{ij}) + \beta_2 z_2(t_{ij}) + \varepsilon(t_{ij}),$$

where $\alpha_1(t)$ and $\alpha_2(t)$ are varying-coefficient terms.

5.2 Bandwidths and model selection

According to Fan et al. [4], the optimal bandwidth is $h = 21.8052$ and h_1 is chosen by using the plug-in bandwidth selector. In addition, our method involves the choice of d. To minimize notation, we use Poly(d) as the shorthand for polynomial models of degree d for autoregressive coefficients. As shown in Pourahmadi [16], we here choose BIC, defined as

$$\mathrm{BIC} = -\frac{1}{n}\widehat{\mathrm{QL}}_{\max} + d\frac{\log n}{n},$$

where QL is the log quasi-likelihood of response, and $\widehat{\mathrm{QL}}_{\max}$ is the maximized log quasi-likelihood for the models with the specified degree d. Smaller values of BIC are associated with better-fitting models. Judging from the BIC values in Table 7, the Ploy(4) is clearly the model of choice for autoregressive coefficients, namely, $z_{ijk} = (1, (t_{ij} - t_{ik}), (t_{ij} - t_{ik})^2, (t_{ij} - t_{ik})^3)'$.

5.3 Estimation

Since the covariance estimator does not affect the asymptotic bias and the variance of $\alpha(t)$, the resulting

Table 7 Values of BIC for several models

Model	BIC	Model	BIC
Ploy(2)	30.3547	Ploy(5)	29.9575
Ploy(3)	29.9606	Ploy(6)	29.9559
Ploy(4)	29.9394	Ploy(7)	29.9519

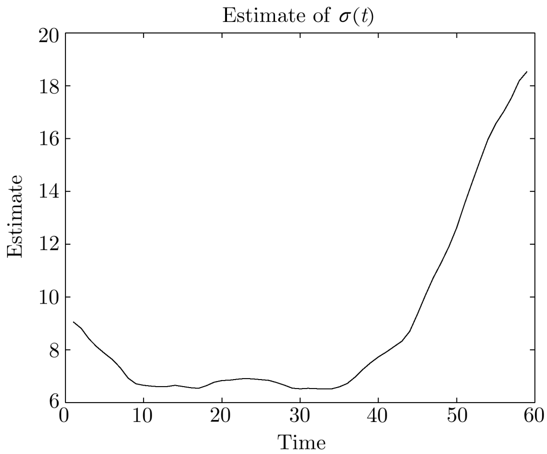

Figure 1 Plot of the estimate of $\sigma(t)$

estimate of $\alpha(t)$ is similar to [4]. Thus the corresponding plots of $\alpha(t)$ are not displayed here. The resulting estimate of $\sigma(t)$ is depicted in Figure 1, from which we can see that $\sigma(t)$ seems to be constant in the first three years and then increases markedly as time increases. This shows that the CD4 percentage becomes harder to be predicted as time evolves. Note that the proposed estimate remains relatively low and stable during the first three years, whereas the estimate of variance function in [4] remains constant only during the first and half year. This implies that the proposed estimate is correspondingly more stable and easier to be predicted.

Next, we estimate β. The resulting estimators of β are shown in Table 8 with standard errors (SE) in the parentheses. From Table 8, we can see that age and smoking status are found to have no significant effect on CD4 cell percentage, whereas the p-value for the proposed $\widehat{\beta}_1$ is smaller than Fan et al. [4]'s.

Table 9 presents the estimates for γ. The numbers in parenthesis are the corresponding standard errors. Table 9 shows that all estimators are significant. Thus the correlation of the data can be explained by the proposed covariance structure.

6 Conclusion

In this article, we used the modified Cholesky decomposition to decompose the within-subject covariance function and then imposed the parametric within-subject correlation while allowing the nonparametric

Table 8 Estimates of β

	Independence	Proposed method	Fan et al. [4]	
			QL	MGV
$\widehat{\beta}_1$	0.8726 (1.1545)	1.6991 (1.1575)	0.6848 (0.9972)	0.6328 (1.0864)
$\widehat{\beta}_2$	−0.5143 (0.6110)	0.2824 (0.5202)	0.0556 (0.4718)	−0.3658 (0.5488)

Table 9 Estimates of γ

	γ_1	γ_2	γ_3	γ_4
Estimate	0.8243	−0.0679	0.0018	−0.0000
SE	0.1152	0.0157	0.0006	0.0000

variation function. We have further proposed generalized estimating equations for the mean and correlation parameter, and kernel estimator for variation. The asymptotic normality is established.

Semiparametric modeling for covariance structure keeps the flexibility of the nonparametric modeling and maintains the explanatory power of parametric modeling. The resulting estimate of the covariance function can be guaranteed to be positive definite. Moreover, the proposed method does not require the prior information about the correlation structure and the normal distribution assumption. This can provide robustness to misspecification of the covariance structure, as we expect it often will be.

A more general case is that the marginal density of $y_i(t_{ij})$ can be written in the form

$$f(y_i(t_{ij})) = \exp[\{y_i(t_{ij})\theta_{ij} - a(\theta_{ij}) + b(y_i(t_{ij}))\}\phi], \tag{19}$$

and $g(\mu_{ij}) = x_i(t_{ij})'\beta + \alpha(t_{ij})$. It is of interest to apply the proposed method to the model (19). Though the proposed method performs well for a continuous distribution under some moment conditions, its applicability to the data from model (19) requires more attention because this model includes discrete distributions.

Acknowledgements This work was partly supported by National Natural Science Foundation of China (Grant Nos. 10931002, 10911120386). The authors cordially thank the referees for their careful reading and helpful comments.

References

1 Diggle P T, Heagerty P J, Liang K, et al. Analysis of Longitudinal Data, 2nd ed. Oxford Statistical Science Series. Oxford: Oxford University Press, 2002

2 Diggle P, Verbyla A. Nonparametric estimation of covariance structure in longitudinal data. Biometrics, 1998, 54: 401–415

3 Fan J, Huang T. Profile likelihood inferences on semiparametric varying coefficient partially linear models. Bernoulli, 2005, 11: 1031–1059

4 Fan J, Huang T, Li R. Analysis of longitudinal data with semiparametric estimation of covariance function. J Amer Statist Assoc, 2007, 100: 632–641

5 Fan J, Li R. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. J Amer Statist Assoc, 2004, 99: 710–723

6 Fan J, Wu Y. Semiparametric estimation of covariance matrices for longitudinal data. J Amer Statist Assoc, 2008, 103: 1520–1533

7 Fan J, Zhang W. Statistical estimation in varying-coefficient models. J Amer Statist Assoc, 1999, 27: 1491–1518

8 Fan J, Zhang W. Two-step estimation of functional linear models with application to longitudinal data. J Roy Statist Soc Ser B, 2000, 62: 303–322

9 He X, Fung W K, Zhu Z Y. Robust estimation in generalized partial linear models for clustered data. J Amer Statist Assoc, 2005, 100: 1176–1184

10 He X, Zhu Z Y, Fung W K. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. Biometrika, 2002, 89: 579–590

11 Huang J Z, Wu C O, Zhou L. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika, 2002, 89: 111–128

12 Liang K Y, Zeger S L. Longitudinal data analysis using generalized linear model. Biometrika, 1986, 73: 13–22

13 Martinussen T, Scheike T H. A semiparametric additive regression model for longitudinal data. Biometrika, 1999, 86: 691–702

14 Pan J, Mackenzie G. Model selection for joint mean-covariance structures in longitudinal studies. Biometrika, 2003, 90: 239–244

15 Pourahmadi M. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. Biometrika, 1999, 86: 677–690

16 Pourahmadi M. Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. Biometrika, 2000, 87: 425–435

17 Ruppert D, Shealther S J, Wand M P. An effect bandwidth selector for local least squares regression. J Amer Statist Assoc, 1995, 90: 1257–1270

18 Sun Y, Zhang W, Tong H. Estimation of the covariance matrix of random effects in longitudinal studies. Ann Statist, 2007, 35: 2795–2814

19 Wang H, Zhu Z, Zhou J. Quantile regression in partially linear varying coefficient models. Ann Statist, 2009, 37: 3841–3866

20 Wu W B, Pourahmadi M. Nonparametric estimation of large covariance matrices of longitudinal data. Biometrika, 2003, 90: 831–844

21 Ye H, Pan J. Modelling of covariance structure in generalised estimation equations for longitudinal data. Biometrika, 2006, 93: 927–941

22 Zeger S L, Diggle P J. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. Biometrics, 1994, 50: 689–699

23 Zhang D W, Lin X H, Raz J, et al. Semiparametric stochastic mixed models for longitudinal data. J Amer Statist Assoc, 1998, 93: 710–719

Appendix

In this Appendix, the proofs of Theorems 1–4 will be given. To facilitate the presentation, let $\mathcal{D} = \{(t_i, X_i, Z_i) : i = 1, \ldots, n\}$.

The following regularity conditions are imposed:

1 The density function $f(\cdot)$ of t has a continuous derivative and is bounded away from 0, where $t \in [0, T]$ and T is an constant. The kernel function $K(\cdot)$ is a symmetric density function with a compact support. In addition, $\upsilon_2 = O(1)$, $\kappa_0 = O(1)$, $\kappa_1 = O(1)$ and $\kappa_2 = O(1)$;

2 $nh^8 \to 0, nh^2/(\log n)^3 \to \infty$;

3 $Ex(t)x'(t)$ and $Ex(t)z'(t)$ are Lipschitz-continuous;

4 J_i has a finite moment-generating function, $E\|x(t)\|^4 + E\|z(t)\|^2 < \infty$;

5 $\alpha_0(t)$ has a continuous second derivative;

6 The parameter space of γ, Θ, is a compact subset of R^d, and the true parameter value γ_0 is in the interior of the parameter space Θ. The covarite z_{ijk} is bounded, meaning that all the elements of the vectors are bounded;

7 For any $\gamma \in \Theta$, it holds with probability one that matrices $\partial\Sigma_i^{-1}(\gamma)/\partial\gamma_k$ and $\partial^2\Sigma_i^{-1}(\gamma)/\partial\gamma_k\partial\gamma_l$ are bounded, where $i = 1, \ldots, n$ and $k, l = 1, \ldots, d$;

8 $E\varepsilon^4(t) < \infty$;

9 $E\widetilde{\varepsilon}^4(t) < \infty$;

10 $\sigma_0^2(t)$ has a continuous second derivative.

Lemma 1. *Under the conditions of Theorem* 1, *we have that*

$$(x'(t), \mathbf{0})\left\{\sum_{i=1}^{n}\Gamma_i'(t)W_i(t)\Gamma_i(t)\right\}^{-1}\sum_{i=1}^{n}\Gamma_i'(t)W_i(t)Z_i = x'(t)G(t)^{-1}\Psi(t)\{1 + O_P(c_n)\}$$

holds uniformly in $t \in [0, T]$, *where* $c_n = h^2 + \{-\log h/(nh)\}^{1/2}$.

This follows immediately from the result obtained by Fan and Huang [3].

Proof of Theorem 1. The proof is very similar to [3] and [4] and below we only briefly present it. For the convenience of presentation, we let $\mathbf{m}_{0i} = (x_i'(t_{i1})\alpha_0(t_{i1}), \ldots, x_i'(t_{iJ_i})\alpha_0(t_{iJ_i}))'$, $\mathbf{m}_0 = (\mathbf{m}_{01}', \ldots, \mathbf{m}_{0n}')'$ and $\epsilon = (\varepsilon_1, \ldots, \varepsilon_n)$. By the orthogonality between the mean part and covariance part and the condition in Theorem 1, a routine calculation reveals that

$$\widehat{\beta} - \beta_0 = \left\{\frac{1}{n}\sum_{i=1}^{n}(Z_i - S_iZ)'\Sigma_i^{-1}(Z_i - S_iZ)\right\}^{-1}(\mathcal{K}_{n1} + \mathcal{K}_{n2})(1 + o_P(1)), \tag{A.20}$$

where

$$\mathcal{K}_{n1} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - S_iZ)'\Sigma_i^{-1}\{\mathbf{m}_{0i} - S_i\mathbf{m}_0\}$$

and

$$\mathcal{K}_{n2} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - S_iZ)'\Sigma_i^{-1}(\varepsilon_i - S_i\epsilon).$$

By Lemma 1 and the assumption at the beginning of Section 4, it is easy to show that

$$\frac{1}{n}\sum_{i=1}^{n}(Z_i - S_iZ)'\Sigma_i^{-1}(Z_i - S_iZ) \to A, \tag{A.21}$$

where A is as shown in Theorem 1.

Similar to [3] and [4], it follows from Lemma 1 that

$$\mathcal{K}_{n1} = o_P(n^{-1/2}). \tag{A.22}$$

By using the similar argument as before and applying the central limit theorem, we have

$$\sqrt{n}\mathcal{K}_{n2} \xrightarrow{\mathcal{L}} N(0, B), \tag{A.23}$$

where B is as given in Theorem 1. Combining (A.20) and (A.21) with (A.22) and (A.23) gives

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, A^{-1}BA^{-1}).$$

This completes the proof.

Proof of Theorem 2. The proof of Theorem 2 can be completed by methods similar to Fan et al. [4].

Proof of Theorem 3. By simple calculation, it can be seen that

$$\sqrt{n}(\widehat{\gamma} - \gamma_0) = \left\{\frac{1}{n}\sum_{i=1}^{n} Z'(i)\widehat{D}_i^{-1}Z(i)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z'(i)\widehat{D}_i^{-1}(\widehat{r}_i - Z(i)\gamma_0)\right\}.$$

Then a routine calculation reveals that

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z'(i)\widehat{D}_i^{-1}(\widehat{r}_i - Z(i)\gamma_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J_i} z'(i,j)\widehat{\sigma}^{-2}(t_{ij})(\widehat{r}_{ij} - z(i,j)\gamma_0)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J_i} z'(i,j)(\widehat{\sigma}^{-2}(t_{ij}) - \sigma_0^{-2}(t_{ij}))(\widehat{r}_{ij} - z(i,j)\gamma_0)\\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J_i} z'(i,j)\sigma_0^{-2}(t_{ij})(\widehat{r}_{ij} - z(i,j)\gamma_0).\end{aligned}$$

By boundedness of other components and the condition in Theorem 3, we have

$$\sqrt{n}(\widehat{\gamma} - \gamma_0) = \left\{\frac{1}{n}\sum_{i=1}^{n} Z'(i)D_i^{-1}Z(i)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z'(i)D_i^{-1}(\widehat{r}_i - Z(i)\gamma_0)\right\}(1 + o_P(1)).$$

First we are ready to prove

$$\frac{1}{n}\sum_{i=1}^{n} Z'(i)D_i^{-1}Z(i) \xrightarrow{P} I(\gamma_0). \tag{A.24}$$

Note

$$\frac{1}{n}\sum_{i=1}^{n} Z'(i)D_i^{-1}Z(i) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \mathrm{tr}(\widehat{R}_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma\partial\gamma'}, \tag{A.25}$$

the (s,t)th entry of which can be decomposed as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \mathrm{tr}(\widehat{R}_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma_s\partial\gamma_t} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \mathrm{tr}(R_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma_s\partial\gamma_t} + 2Q_{n1} + Q_{n2}, \tag{A.26}$$

where $s, t = 1, \ldots, d$, $R_i = \varepsilon_i\varepsilon_i'$, $\widehat{R}_i = \widehat{r}_i\widehat{r}_i'$, $\widehat{r}_i = (\widehat{r}_{i1}, \ldots, \widehat{r}_{iJ_i})'$ and $\widehat{r}_{ij} = y_i(t_{ij}) - x_i'(t_{ij})\widehat{\alpha}(t_{ij}, \widehat{\beta}) - z_i'(t_{ij})\widehat{\beta}$,

$$Q_{n1} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i'\frac{\partial^2\Sigma_i^{-1}(\gamma_0)}{\partial\gamma_s\partial\gamma_t}(X_i(\widehat{\alpha}(t_i, \widehat{\beta}) - \alpha_0(t_i)) + Z_i(\widehat{\beta} - \beta_0))$$

and

$$Q_{n2} = \frac{1}{n}\sum_{i=1}^{n}(X_i(\widehat{\alpha}(t_i,\widehat{\beta}) - \alpha_0(t_i)) + Z_i(\widehat{\beta}-\beta_0))' \frac{\partial^2 \Sigma_i^{-1}(\gamma_0)}{\partial\gamma_s\partial\gamma_t} \times (X_i(\widehat{\alpha}(t_i,\widehat{\beta}) - \alpha_0(t_i)) + Z_i(\widehat{\beta}-\beta_0)).$$

It is obvious from Theorems 1 and 2 that $Q_{n1} = o_P(1)$ and $Q_{n2} = o_P(1)$. Thus we can get

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \mathrm{tr}(\widehat{R}_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma\partial\gamma'} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \mathrm{tr}(R_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma\partial\gamma'} + o_P(1). \tag{A.27}$$

Note that $\partial^2 tr\{R_i\Sigma_i^{-1}(\gamma_0)\}/\partial\gamma\partial\gamma'$ is a $(d\times d)$ vector with (s,t)th element $\partial^2 tr\{\Sigma_i^{-1}(\gamma_0)R_i\}/\partial\gamma_s\partial\gamma_t$, the expectation of which, after some tedious calculation, equals

$$E\left[\mathrm{tr}\left\{\Sigma_1^{-1}(\gamma_0)\left(\frac{\partial\Sigma_1(\gamma_0)}{\partial\gamma_s}\Sigma_1^{-1}(\gamma_0)\frac{\partial\Sigma_1(\gamma_0)}{\partial\gamma_t} + \frac{\partial\Sigma_1(\gamma_0)}{\partial\gamma_t}\Sigma_1^{-1}(\gamma_0)\frac{\partial\Sigma_1(\gamma_0)}{\partial\gamma_s}\right)\right\}\right]. \tag{A.28}$$

In addition, $\partial^2\{\Sigma_i^{-1}(\gamma_0)\}/\partial\gamma_s\partial\gamma_t$ is a symmetric matrix. This together with condition 7 implies that $\partial^2\{\Sigma_i^{-1}(\gamma_0)\}/\partial\gamma_s\partial\gamma_t$ has finite eigenvalues. Let δ_{mst} be the maximum value among all eigenvalues. Therefore the absolute value of the (s,t)th entry of $\partial^2\mathrm{tr}\{R_i\Sigma_i^{-1}(\gamma_0)\}/\partial\gamma\partial\gamma'$ is bounded by $|\delta_{mst}|\mathrm{tr}(R_i)$, whose expectation exists and is finite. Combining this and (A.27)–(A.28) with conditions 6–7 and by the law of large numbers yields

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \mathrm{tr}(\widehat{R}_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma\partial\gamma'} \xrightarrow{P} I(\gamma_0).$$

This along with (A.25) establishes (A.24).

To get the desired asymptotic normality result, we also need to show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z'(i)D_i^{-1}(\widehat{r}_i - Z(i)\gamma_0) \xrightarrow{\mathcal{L}} N(0, V(\gamma_0)).$$

Next, we prove the above result. Note that

$$\frac{1}{n}\sum_{i=1}^{n} Z'(i)D_i^{-1}(\widehat{r}_i - Z(i)\gamma_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathrm{tr}(\widehat{R}_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma}.$$

Similarly, we can obtain

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathrm{tr}(\widehat{R}_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma_s} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathrm{tr}(R_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma_s} + 2\widetilde{Q}_{n1} + \widetilde{Q}_{n2}, \tag{A.29}$$

where $s = 1,\dots,d$,

$$\widetilde{Q}_{n1} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i'\frac{\partial\Sigma_i^{-1}(\gamma_0)}{\partial\gamma_s}(X_i(\widehat{\alpha}(t_i,\widehat{\beta}) - \alpha_0(t_i)) + Z_i(\widehat{\beta}-\beta_0))$$

and

$$\widetilde{Q}_{n2} = \frac{1}{n}\sum_{i=1}^{n}(X_i(\widehat{\alpha}(t_i,\widehat{\beta}) - \alpha_0(t_i)) + Z_i(\widehat{\beta}-\beta_0))' \frac{\partial \Sigma_i^{-1}(\gamma_0)}{\partial\gamma_s} \times (X_i(\widehat{\alpha}(t_i,\widehat{\beta}) - \alpha_0(t_i)) + Z_i(\widehat{\beta}-\beta_0)).$$

Now we consider the term $\widetilde{Q}_{n1}$. It follows from tedious calculation and the proof of Theorem 2 that

$$\widetilde{Q}_{n1} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{J_i} w_{ijk}(x_i'(t_{ij})(\widehat{\alpha}(t_{ij},\widehat{\beta}) - \alpha_0(t_{ij})) + z_i'(t_{ij})(\widehat{\beta}-\beta_0))\varepsilon_{ik}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{J_i} w_{ijk}(x_i'(t_{ij}),\mathbf{0}_p')\bigg\{\sum_{r=1}^{n}\Gamma_r'(t_{ij})W_r(t_{ij})\Gamma_r(t_{ij})\bigg\}^{-1}\sum_{r=1}^{n}\Gamma_r'(t_{ij})W_r(t_{ij})\varepsilon_r\varepsilon_{ik}\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i'\frac{\partial\Sigma_i^{-1}(\gamma_0)}{\partial\gamma_s}(Z_i-S_iZ)(\widehat{\beta}-\beta_0)+o_p(n^{-1/2})\\
&=\widetilde{Q}_{n11}+\widetilde{Q}_{n12}+o_P(n^{-1/2}),
\end{aligned}
$$

where w_{ijk} is the (j,k)th entry of the matrix $\partial\Sigma_i^{-1}(\gamma_0)/\partial\gamma_s$. First, we consider the term $\widetilde{Q}_{n11}$. It follows from straightforward calculation that

$$
\begin{aligned}
(x_i'(t_{ij}),&\mathbf{0}_p')\bigg\{\sum_{r=1}^{n}\Gamma_r'(t_{ij})W_r(t_{ij})\Gamma_r(t_{ij})\bigg\}^{-1}\sum_{r=1}^{n}\Gamma_r'(t_{ij})W_r(t_{ij})\varepsilon_r\varepsilon_{ik}\\
&=\frac{x_i'(t_{ij})G^{-1}(t)}{nhE(J_1)f(t_{ij})}\sum_{r=1}^{n}\sum_{l=1}^{J_r}K_h(t_{rl}-t_{ij})x_r(t_{rl})\varepsilon_{rl}\varepsilon_{ik}(1+O_P(c_n)).
\end{aligned}
$$

By boundedness of the kernel function, independent of random errors between subjects and condition 8, we have

$$E(\widetilde{Q}_{n11}^2|\mathcal{D})=O_P\bigg(\frac{(\log n)^6}{(nh)^2}\bigg),$$

that is

$$\widetilde{Q}_{n11}=o_P(n^{-1/2}). \tag{A.30}$$

Similarly, it follows from Theorem 1 that

$$\widetilde{Q}_{n12}=o_P(n^{-1/2}), \tag{A.31}$$

which together with (A.30) gives

$$\widetilde{Q}_{n1}=o_P(n^{-1/2}). \tag{A.32}$$

In addition, it follows from the results of Theorems 1 and 2 that

$$\widetilde{Q}_{n2}=o_P(n^{-1/2}), \tag{A.33}$$

which along with (A.29) and (A.32) results in

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathrm{tr}(\widehat{R}_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma}=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathrm{tr}(R_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma}+o_P(n^{-1/2}).$$

And then we only need to prove

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\mathrm{tr}(R_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma}\xrightarrow{\mathcal{L}}N(0,V(\gamma_0)). \tag{A.34}$$

By straightforward calculation, we have

$$\frac{\partial\mathrm{tr}(R_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma}=Z(i0)'D_i^{-1}(\varepsilon_i-\widetilde{\varepsilon}_i),$$

where $\widetilde{\varepsilon}_i$ is the $(J_i\times 1)$ vector with jth components $\widetilde{\varepsilon}_{ij}=E(\varepsilon_{ij}|\varepsilon_{i1},\ldots,\varepsilon_{i(j-1)})=\sum_{k=1}^{j-1}\phi_{ijk}\varepsilon_{ik}$ and $Z(i0)$ is the $(J_i\times d)$ matrix with jth column $\sum_{k=1}^{j-1}\varepsilon_{ik}z_{ijk}$. Thus we have

$$E\bigg(\frac{\partial\mathrm{tr}(R_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma}\bigg)=0. \tag{A.35}$$

This together with condition 7 and by the central limit theorem proves (A.34). Using Slutsky's theorem, we can get

$$\sqrt{n}(\widehat{\gamma}-\gamma_0) \xrightarrow{\mathcal{L}} N\big(0, I^{-1}(\gamma_0)V(\gamma_0)I^{-1}(\gamma_0)\big).$$

This completes the proof of Theorem 3.

Proof of Theorem 4. Let $e_{ij} = x_i'(t_{ij})(\widehat{\alpha}(t_{ij},\widehat{\beta}) - \alpha_0(t_{ij})) + z_i'(t_{ij})(\widehat{\beta}-\beta_0)$ and $e_i = (e_{i1},\dots,e_{iJ_i})'$. Then, $\widehat{r}_i = \varepsilon_i - e_i$.

Decompose $\widehat{\sigma}^2(t)$ as

$$\begin{aligned}
\widehat{\sigma}^2(t) &= \frac{\sum_{i=1}^n\sum_{j=1}^{J_i}(\widehat{r}_{ij} - \sum_{k=1}^{j-1}\widehat{r}_{ik}z_{ijk}'\widehat{\gamma})^2 K_{h_1}(t-t_{ij})}{\sum_{i=1}^n\sum_{j=1}^{J_i}K_{h_1}(t-t_{ij})}\\
&= \frac{\sum_{i=1}^n\sum_{j=1}^{J_i}(\widehat{r}_{ij} - \sum_{k=1}^{j-1}\widehat{r}_{ik}z_{ijk}'\gamma_0 - \sum_{k=1}^{j-1}\widehat{r}_{ik}z_{ijk}'(\widehat{\gamma}-\gamma_0))^2 K_{h_1}(t-t_{ij})}{\sum_{i=1}^n\sum_{j=1}^{J_i}K_{h_1}(t-t_{ij})}\\
&= \frac{\sum_{i=1}^n \widehat{r}_i'T_i'(\gamma_0)W_{i1}T_i(\gamma_0)\widehat{r}_i}{\sum_{i=1}^n\sum_{j=1}^{J_i}K_{h_1}(t-t_{ij})} + \frac{\sum_{i=1}^n\sum_{j=1}^{J_i}(\sum_{k=1}^{j-1}\widehat{r}_{ik}z_{ijk}'(\widehat{\gamma}-\gamma_0))^2 K_{h_1}(t-t_{ij})}{\sum_{i=1}^n\sum_{j=1}^{J_i}K_{h_1}(t-t_{ij})}\\
&\quad - 2\frac{\sum_{i=1}^n\sum_{j=1}^{J_i}(\widehat{r}_{ij} - \sum_{k=1}^{j-1}\widehat{r}_{ik}z_{ijk}'\gamma_0)(\sum_{k=1}^{j-1}\widehat{r}_{ik}z_{ijk}'(\widehat{\gamma}-\gamma_0))K_{h_1}(t-t_{ij})}{\sum_{i=1}^n\sum_{j=1}^{J_i}K_{h_1}(t-t_{ij})}\\
&= L_1 + L_3 - 2L_2.
\end{aligned}$$

First, we consider the term L_2. It follows from straightforward but tedious calculation and Theorems 1–3 that

$$\begin{aligned}
L_2 &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{J_i}\left(\varepsilon_{ij} - \sum_{k=1}^{j-1}\varepsilon_{ik}z_{ijk}'\gamma_0 - e_{ij} + \sum_{k=1}^{j-1}e_{ik}z_{ijk}'\gamma_0\right)\left(\sum_{k=1}^{j-1}(\varepsilon_{ik}-e_{ik})z_{ijk}'(\widehat{\gamma}-\gamma_0)\right)K_{h_1}(t-t_{ij})\\
&= \mathrm{tr}\Bigg[\Bigg(\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{J_i}K_{h_1}(t-t_{ij})(\varepsilon_{ij}-\widetilde{\varepsilon}_{ij})\Bigg\{\sum_{k=1}^{j-1}\varepsilon_{ik}\Bigg(\frac{1}{n}\sum_{i=1}^n\frac{\partial\mathrm{tr}(\widehat{R}_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma}\Bigg)z_{ijk}'\Bigg\}\Bigg)\\
&\quad\times\Bigg\{\frac{1}{n}\sum_{i=1}^n Z'(i)D_i^{-1}Z(i)\Bigg\}^{-1}\Bigg](1+o_P(1)) + o_P(n^{-1/2})\\
&= \mathrm{tr}\Bigg[\Bigg(\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{J_i}K_{h_1}(t-t_{ij})(\varepsilon_{ij}-\widetilde{\varepsilon}_{ij})\Bigg\{\sum_{k=1}^{j-1}\varepsilon_{ik}\Bigg(\frac{1}{n}\sum_{i=1}^n\frac{\partial\mathrm{tr}(R_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma}\Bigg)z_{ijk}'\Bigg\}\Bigg)\\
&\quad\times\Bigg\{\frac{1}{n}\sum_{i=1}^n Z'(i)D_i^{-1}Z(i)\Bigg\}^{-1}\Bigg](1+o_P(1)) + o_P(n^{-1/2}).
\end{aligned}$$

Then it is sufficient to consider the term

$$\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{J_i}K_{h_1}(t-t_{ij})(\varepsilon_{ij}-\widetilde{\varepsilon}_{ij})\Bigg\{\sum_{k=1}^{j-1}\varepsilon_{ik}\Bigg(\frac{1}{n}\sum_{i=1}^n\frac{\partial\mathrm{tr}(R_i\Sigma_i^{-1}(\gamma_0))}{\partial\gamma}\Bigg)z_{ijk}'\Bigg\},$$

the (s,t)th entry of which can be written as

$$\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{J_i}K_{h_1}(t-t_{ij})(\varepsilon_{ij}-\widetilde{\varepsilon}_{ij})\Bigg\{\sum_{k=1}^{j-1}\varepsilon_{ik}\Bigg(\frac{1}{n}\sum_{i=1}^n\varepsilon_i'\frac{\partial\Sigma_i^{-1}(\gamma_0)}{\partial\gamma_s}\varepsilon_i\Bigg)z_{ijkt}\Bigg\},$$

where z_{ijkt} is the tth entry of z_{ijk}. Recall that $\widetilde{\varepsilon}_{ij} = E(\varepsilon_{ij}|\varepsilon_{i1},\dots,\varepsilon_{i(j-1)})$. Using similar techniques to consider $\widetilde{Q}_{n11}$ in the proof of Theorem 3, following tedious calculation, it follows that $\sqrt{nh_1}L_2 = o_P(1)$. It is clear from Theorem 3 that $\sqrt{nh_1}L_3 = o_P(1)$.

Now we consider L_1. Decompose L_1 as

$$L_1 = L_{11} + L_{12} + L_{13},$$

where

$$L_{11} = \frac{\sum_{i=1}^n \varepsilon_i' T_i'(\gamma_0) W_{i1} T_i(\gamma_0) \varepsilon_i}{\sum_{i=1}^n \sum_{j=1}^{J_i} K_{h_1}(t - t_{ij})},$$
$$L_{12} = \frac{\sum_{i=1}^n \varepsilon_i' T_i'(\gamma_0) W_{i1} T_i(\gamma_0) e_i}{\sum_{i=1}^n \sum_{j=1}^{J_i} K_{h_1}(t - t_{ij})},$$
$$L_{13} = \frac{\sum_{i=1}^n e_i' T_i'(\gamma_0) W_{i1} T_i(\gamma_0) e_i}{\sum_{i=1}^n \sum_{j=1}^{J_i} K_{h_1}(t - t_{ij})},$$

and

$$W_{i1} = \mathrm{diag}(K_{h_1}(t - t_{ij}), \ldots, K_{h_1}(t - t_{iJ_i})).$$

By straightforward calculation, we have

$$\begin{aligned} L_{12} &= \frac{1}{nh} \sum_{i=1}^n \sum_{j=1}^{J_i} \sum_{k=1}^{J_i} v_{ijk}(x_i'(t_{ij}), \mathbf{0}_p') \left\{ \sum_{r=1}^n \Gamma_r'(t_{ij}) W_r(t_{ij}) \Gamma_r(t_{ij}) \right\}^{-1} \\ &\quad \times \sum_{r=1}^n \Gamma_r'(t_{ij}) W_r(t_{ij}) \varepsilon_r \varepsilon_{ik} + \frac{1}{nh} \sum_{i=1}^n \varepsilon_i' T_i'(\gamma_0) W_{i1} T_i(\gamma_0)(Z_i - \widetilde{X}_i)(\widehat{\beta} - \beta_0) \\ &\quad + o_P(n^{-1/2}) \\ &= L_{121} + L_{122} + o_P(n^{-1/2}), \end{aligned}$$

By similar argument to the proof of (A.29) in Theorem 3 and boundness of kernel functions, we can show that $\sqrt{nh_1} L_{121} = o_P(1)$ and $\sqrt{nh_1} L_{122} = o_P(1)$. Using Theorems 1 and 2, we can easily get $\sqrt{nh_1} L_{13} = o_P(1)$.

It remains to show that Q_1 is asymptotically normal. Since $\widetilde{\varepsilon}_i = T(\gamma_0)\varepsilon_i$, we have

$$L_1 = \frac{\sum_{i=1}^n \sum_{j=1}^{J_i} \widetilde{\varepsilon}_i^2(t_{ij}) K_{h_1}(t - t_{ij})}{\sum_{i=1}^n \sum_{j=1}^{J_i} K_{h_1}(t - t_{ij})}.$$

Similar to Fan et al. [4], it follows from $E(\widetilde{\varepsilon}^2(t)) = \sigma_0^2(t)$ that

$$\sqrt{nh_1}(L_1 - \sigma_0^2(t) - b(t)) \xrightarrow{\mathcal{L}} N(0, v(t)).$$

Using Slutsky's theorem,

$$\sqrt{nh_1}(\widehat{\sigma}^2(t) - \sigma_0^2(t) - b(t)) \xrightarrow{\mathcal{L}} N(0, v(t)).$$

This completes the proof.