Submitted to the Annals of Statistics

EFFICIENT ESTIMATION IN SEMIVARYING COEFFICIENT MODELS FOR LONGITUDINAL/CLUSTERED DATA

By Ming-Yen Cheng [¶,†], Toshio Honda[‖,*], and Jialiang Li[**,‡]

Hitotsubashi University[*], National Taiwan University[†], and National University of Singapore[‡]

In semivarying coefficient modeling of longitudinal/clustered data, of primary interest is usually the parametric component which involves unknown constant coefficients. First we study semiparametric efficiency bound for estimation of the constant coefficients in a general setup. It can be achieved by spline regression using the true within-subject covariance matrices, which are often unavailable in reality. Thus we propose an estimator when the covariance matrices are unknown and depend only on the index variable. To achieve this goal, we estimate the covariance matrices using residuals obtained from a preliminary estimation based on working independence and both spline and local linear regression. Then, using the covariance matrix estimates, we employ spline regression again to obtain our final estimator. It achieves the semiparametric efficiency bound under normality assumption and has the smallest asymptotic covariance matrix among a class of estimators even when normality is violated. Our theoretical results hold either when the number of within-subject observations diverges or when it is uniformly bounded. In addition, the local linear estimator of the nonparametric component is superior to the spline estimator in terms of numerical performance. The proposed method is compared with the working independence estimator and some existing method via simulations and application to a real data example.

[§]This research was partially supported by the Hitotsubashi International Fellow Program and a Taiwan Ministry of Education grant.

[¶]Corresponding author. Research was supported by the Ministry of Science and Technology grant 101-2118-M-002-001-MY3.

[‖]Research was supported by the JSPS Grant-in-Aids for Scientific Research (A) 24243031 and (C) 25400197.

[**]Research was supported by grants AcRF R-155-000-130-112 and NMRC/CBRG/0014/2012.

MSC 2010 subject classifications: Primary 62G08

Keywords and phrases: Covariance matrix estimation; local linear regression; semiparametric efficiency bound; spline functions.

1. Introduction. Suppose we have a scalar response Y, and two p-dimensional and q-dimensional covariate vectors $\boldsymbol{X}$ and $\boldsymbol{Z}$. Longitudinal data consist of $(Y_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, T_{ij}), i = 1, \ldots, n, j = 1, \ldots, m_i$, where Y_{ij}, $\boldsymbol{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$ and $\boldsymbol{Z}_{ij} = (Z_{ij1}, \ldots, Z_{ijq})^T$ are respectively the values of Y, $\boldsymbol{X}$ and $\boldsymbol{Z}$ of the ith subject at the jth observation time $T_{ij} \in [0, 1]$. Such kind of data are commonly acquired for various purposes, such as evidence based knowledge discovery and empirical study, in a wide range of subject areas. When the subjects are changed to clusters and the T_{ij}'s are observations on some index variable other than time, they are usually called clustered data. We assume that all the covariates are uniformly bounded for technical reasons. Besides, we let $Z_{ij1} \equiv 1$ and suppose $\boldsymbol{X}_{ij}$ has no constant element for all i and j.

For $i = 1, \ldots, n$, denote

$$\underline{\boldsymbol{X}}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{im_i})^T, \ \underline{\boldsymbol{Z}}_i = (\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{im_i})^T, \text{ and } \underline{T}_i = (T_{i1}, \ldots, T_{im_i})^T.$$

A popular model for longitudinal data analysis is the semivarying coefficient model, which is specified by

$$\begin{aligned} &\mathrm{E}(Y_{ij}|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, T_{ij}, \underline{\boldsymbol{X}}_i, \underline{\boldsymbol{Z}}_i, \underline{T}_i) \\ &\quad = \mathrm{E}(Y_{ij}|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, T_{ij}) \equiv \mu(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T\boldsymbol{g}(T_{ij})) = \mu_{ij}, \end{aligned} \tag{1.1}$$

where $\boldsymbol{A}^T$ stands for the transpose of a matrix $\boldsymbol{A}$. In model (1.1), $\mu(x)$ is a known strictly increasing smooth link function, $\boldsymbol{\beta}$ is an unknown regression coefficient vector, and $\boldsymbol{g}(t) = \big(g_1(t), \ldots, g_q(t)\big)^T$ is a vector of unknown smooth functions. Define

$$\underline{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im_i})^T = \underline{Y}_i - \underline{\mu}_i, \text{ and } \boldsymbol{\Sigma}_i = \mathrm{Var}(\underline{\epsilon}_i|\underline{\boldsymbol{X}}_i, \underline{\boldsymbol{Z}}_i, \underline{T}_i), \tag{1.2}$$

where $\underline{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$, $\underline{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})^T$, and $\boldsymbol{\Sigma}_i$ is an $m_i \times m_i$ positive definite matrix depending on $\underline{\boldsymbol{X}}_i$, $\underline{\boldsymbol{Z}}_i$, and $\underline{T}_i$, $i = 1, \ldots, n$. This is a standard marginal model in longitudinal data analysis [24].

Model (1.1) consists of a parametric component, which provides information on the constant impacts of some important covariates, and a nonparametric component which captures the dynamic impacts of the other covariates. In this way the model is able to reflect unknown nonlinear structures in the data while retaining similar interpretability as the classical linear models at the same time. There is an extensive literature on the variable selection, structure identification, estimation, and inference issues [6, 8, 12, 22, 25]. In particular, often of primary interest is to have access to the parametric component while the nonparametric component is viewed as the nuisance

part. In this regard, it is well known that assuming independence or some mis-specified working covariance structure yields less efficient estimation of the constant coefficients. Therefore, a substantial portion of the existing literature aimed at improving the efficiency via modeling and estimating the within-subject covariance structure [6, 7, 10, 18, 26, 27, 28], which is itself a challenging task.

In this article, we focus on the identity link function and make contributions to the efficient estimation problem for model (1.1) in three directions. First, we allow some of the m_i's to tend to infinity. As far as we know, this setup has not been treated before and the problem is nontrivial. Our results also hold when the m_i's are uniformly bounded and $\underline{\epsilon}_i$ satisfies the sub-Gaussian property. See the supplement [5] for the details. When all of the m_i's are diverging, that is, if we have densely observed data, it becomes a kind of functional data problem and is out of the scope of this paper. Second, we study explicit expression of the semiparametric efficiency bound for estimation of $\boldsymbol{\beta}$ and asymptotic normality of the generalized estimating equations (GEE) spline estimator under general covariance structures and error distributions. Using the true covariance matrices in the GEE estimation leads to optimality among all GEE estimators of the parametric component. Furthermore, it achieves the semiparametric efficiency bound when the errors are conditionally normal. Our results are in parallel to that for partially linear and partially linear additive models given by [13] and [4] respectively. Those models are among a rich variety of semiparametric ways of modeling longitudinal data, and they differ from semivarying coefficient models in that their nonparametric components admit more direct additive expressions. Partially linear (additive) models were also considered by [14, 15, 16, 17, 23], among which [14, 15, 16, 23] used kernel method and [17] used spline estimation.

Our third contribution is to deal with adaptive efficient estimation when the within-subject covariance matrices are estimated nonparametrically using the data at hand. Notice that [4] ignored this practical issue and did not consider estimation of the covariances, and [13] suggested using some parametric specification which can be estimated $\sqrt{n}$-consistently. We consider the case where the nonparametric within-subject covariance matrices depend only on the observation times but not on the other covariates. Such assumptions are reasonable because we do not assume that the observation times are regular across different subjects or they are dense. Indeed, with irregular and/or sparse observation times, estimating the covariances in a completely nonparametric way, by letting them to be dependent on all of the T_{ij}, $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_{ij}$ nonparametrically, is particularly problematic and even

unreliable as the curse-of-dimensionality problem arises. Our covariance estimator is constructed based on residuals yielded by an initial estimation. The final estimator of the true value of $\boldsymbol{\beta}$ is then given by plugging-in the covariance estimates to the GEE spline estimation. We show the asymptotic equivalence of our final estimator to the oracle efficient estimator which uses the true covariance matrices in the GEE spline estimation.

The above result is partly motivated by the study of [14] on efficient estimation in partially linear models under the same nonparametric covariance structure. However, the kernel profile method taken by [14] involves only local linear regression, thus, to achieve semiparametric efficiency it requires some complicated iterative backfitting calculation except for the identity link function [15, 16]. By comparison, our approach to estimating the parametric and nonparametric components in the mean function is different and much simpler. We ingeniously use both spline approximation and local linear estimation to avoid complicated calculation while allowing for the asymptotic equivalence property at the same time. To the best of our knowledge, there are no existing results for semivarying coefficient models, especially when some of the m_i's tend to infinity or when the $\mathbf{\Sigma}_i$s are estimated.

Our final estimator is some kind of feasible generalized least squares (FGLS) estimator since we replace the within-subject covariance matrices with their nonparametric estimates. Even if our assumption on the covariance matrices fails to hold, it still possesses the asymptotic normality under mild conditions and still makes use of some information of the covariance matrices. For example, if the covariances depend on some time-dependent covariates, to some extent such effects are still captured by our method. In this sense, compared with existing methods which use either parametrically estimated or some ad-hoc covariance matrices [7, 18, 21], our approach is more adaptive to the unknown covariance matrices. A promising cluster bootstrap inference method was proposed by [2]; it assumes some parametric within-cluster covariance structure, however. In the case where there is one observation for each subject/cluster, our assumption on the covariance matrices reduces to that of [20], which also suggested to improve the efficiency in a similar manner.

Our simulation study shows that numerically the proposed method outperforms the working independence approach and the quadratic inference functions (QIF) method by [18], and it behaves close to the oracle estimator which uses the true covariance matrices. Note that, while the QIF procedure is suitable when there is some kind of regularity and stationarity in the error process, our procedure adapts to both non-stationarity and irregularity. We also applied our method to the CD4 count dataset and identified some

interesting new effects not detected by the working independence approach.

After the semiparametric efficient estimation, we can estimate and make inference on the nonparametric component in the same way as in dealing with varying coefficient models, using the difference between the response and the estimated parametric part [25]. When p and q are both diverging and the model is sparse, [6] suggested a simultaneous variable selection and structure identification procedure and showed its consistency property. By combining the method with the proposed estimation procedure and by putting together the corresponding consistency and efficiency results, we have an efficient estimation procedure in this case.

The organization of this paper is as follows. In Section 2 we derive the semiparametric efficiency bound for the constant coefficient vector $\boldsymbol{\beta}$ and asymptotic normality of GEE spline estimators. In Section 3, we propose an efficient estimator of $\boldsymbol{\beta}$ when the errors have some general covariance structure and state its asymptotic equivalence to the oracle estimator which assumes the covariance matrices are known. Section 4 summarizes and discusses results of our simulation and empirical studies used to assess numerical performance of the proposed efficient estimator. Section 5 contains some technical assumptions and proof of the asymptotic equivalence. In the supplementary material [5] we give additional simulation results for estimation, proofs of the other theoretical results, some lemmas, and theoretical results when the m_i's are uniformly bounded.

2. Semiparametric efficiency bound for $\boldsymbol{\beta}$. In this section, $\boldsymbol{V}_i$ is a given $m_i \times m_i$ inverse weight matrix depending only on $\underline{\boldsymbol{X}}_i$, $\underline{\boldsymbol{Z}}_i$, and $\underline{T}_i$, $i = 1, \ldots, n$. We use a K_n-dimensional equispaced B-spline basis on $[0, 1]$, denoted by $\boldsymbol{B}(t)$, to approximate the function $\boldsymbol{g}(t)$. See [19] for the definition and properties of B-spline bases. We set $\boldsymbol{W}_{ij} = \boldsymbol{Z}_{ij} \otimes \boldsymbol{B}(T_{ij})$ and $\underline{\boldsymbol{W}}_i = (\boldsymbol{W}_{i1}, \ldots, \boldsymbol{W}_{im_i})^T$, where $\otimes$ is the Kronecker product, and we denote the true values of $\boldsymbol{\beta}$ and $\boldsymbol{g}(t)$ by $\boldsymbol{\beta}_0$ and $\boldsymbol{g}_0(t) = (g_{01}(t), \ldots, g_{0q}(t))^T$ respectively. Then we estimate $\boldsymbol{\beta}_0$ and $\boldsymbol{g}_0(t)$ by minimizing with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ simultaneously the following objective function:

$$\sum_{i=1}^{n} (\underline{Y}_i - \underline{\mu}(\underline{\boldsymbol{X}}_i\boldsymbol{\beta} + \underline{\boldsymbol{W}}_i\boldsymbol{\gamma}))^T \boldsymbol{V}_i^{-1} (\underline{Y}_i - \underline{\mu}(\underline{\boldsymbol{X}}_i\boldsymbol{\beta} + \underline{\boldsymbol{W}}_i\boldsymbol{\gamma})), \tag{2.1}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{qK_n}$ and the j th element of $\underline{\mu}(\underline{\boldsymbol{X}}_i\boldsymbol{\beta}+\underline{\boldsymbol{W}}_i\boldsymbol{\gamma})$ is $\mu(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}+\boldsymbol{W}_{ij}^T\boldsymbol{\gamma})$. Thus the generalized estimating equations are

$$\sum_{i=1}^{n} \underline{\boldsymbol{X}}_i^T \Delta_i \boldsymbol{V}_i^{-1}(\underline{Y}_i - \underline{\mu}(\underline{\boldsymbol{X}}_i\boldsymbol{\beta} + \underline{\boldsymbol{W}}_i\boldsymbol{\gamma})) = 0,$$

$$\text{and} \quad \sum_{i=1}^{n} \underline{\boldsymbol{W}}_i^T \Delta_i \boldsymbol{V}_i^{-1}(\underline{Y}_i - \underline{\mu}(\underline{\boldsymbol{X}}_i\boldsymbol{\beta} + \underline{\boldsymbol{W}}_i\boldsymbol{\gamma})) = 0, \tag{2.2}$$

where Δ_i is an $m_i \times m_i$ diagonal matrix defined by $\Delta_i = \text{diag}(\mu'(\boldsymbol{X}_{i1}^T\boldsymbol{\beta} + \boldsymbol{W}_{i1}^T\boldsymbol{\gamma}), \ldots, \mu'(\boldsymbol{X}_{im_i}^T\boldsymbol{\beta} + \boldsymbol{W}_{im_i}^T\boldsymbol{\gamma}))$. Denote the solution to (2.2) by $\widehat{\boldsymbol{\beta}}_{\boldsymbol{V}}$ and $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{V}} \equiv \big(\widehat{\gamma}_{1V}^T, \ldots, \widehat{\gamma}_{qV}^T\big)^T$. Then the GEE spline estimator with weight matrices $\boldsymbol{V}_i^{-1}$, $i = 1, \ldots, n$, for $\boldsymbol{\beta}_0$ is $\widehat{\boldsymbol{\beta}}_{\boldsymbol{V}}$ and that for $\boldsymbol{g}_0(t)$ is $\big(\widehat{\gamma}_{1V}^T\boldsymbol{B}(t), \ldots, \widehat{\gamma}_{qV}^T\boldsymbol{B}(t)\big)^T$.

Hereafter we focus on the identity link function and present the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{V}}$ in Proposition 1 under general error distributions as specified in Assumption A6 given in Section 5. We allow some of the m_i's to diverge in a way like $\sum_{i=1}^n m_i^5 = O(n)$ and $\max_{1\le i\le n} m_i = O(n^{1/8})$. See Assumptions A1 and A2 for the specific conditions on the m_i's. We refer to the supplement [5] for the results for general link functions when the m_i's are uniformly bounded and the $\underline{\epsilon}_i$'s satisfy the sub-Gaussian property.

First, we introduce some function spaces, inner products and projections. Let L_2 denote the space of square integrable functions on $[0, 1]$ and recall $\boldsymbol{B}(t)$ is the equispaced B-spline basis on $[0, 1]$. We define two function spaces:

$$\boldsymbol{G} = \{(g_1, \ldots, g_q)^T \,|\, g_j \in L_2,\, j = 1, \ldots, q\},$$
$$\text{and} \quad \boldsymbol{G}_B = \{(\boldsymbol{B}^T\boldsymbol{\gamma}_1, \ldots, \boldsymbol{B}^T\boldsymbol{\gamma}_q)^T \,|\, \boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_q^T)^T \in \mathbb{R}^{qK_n}\}.$$

Note that $\boldsymbol{G}_B \subset \boldsymbol{G}$. Next, let v_1 and v_2 be two stochastic processes each taking scalar values at T_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$. Then we define two inner products of v_1 and v_2 by $\langle v_1, v_2\rangle_n^V = \frac{1}{n}\sum_{i=1}^n \underline{v}_{1i}^T\boldsymbol{V}_i^{-1}\underline{v}_{2i}$ and $\langle v_1, v_2\rangle^V = \text{E}\{\langle v_1, v_2\rangle_n^V\}$, where $\underline{v}_{1i}$ and $\underline{v}_{2i}$ are defined in the same way as $\underline{T}_i$, and we define the associated norms by $\|v\|_n^V = (\langle v, v\rangle_n^V)^{1/2}$ and $\|v\|^V = (\langle v, v\rangle^V)^{1/2}$. The projections, with respect to $\|\cdot\|^V$, of the kth element of $\boldsymbol{X}$ onto $\boldsymbol{Z}^T\boldsymbol{G}$ and $\boldsymbol{Z}^T\boldsymbol{G}_B$ are given by

$$\Pi_{\boldsymbol{V}}X_k = \underset{\boldsymbol{g}\in\boldsymbol{G}}{\text{argmin}}\, \|X_k - \boldsymbol{Z}^T\boldsymbol{g}\|^V \text{ and } \Pi_{\boldsymbol{V}n}X_k = \underset{\boldsymbol{g}\in\boldsymbol{G}_B}{\text{argmin}}\, \|X_k - \boldsymbol{Z}^T\boldsymbol{g}\|^V, \tag{2.3}$$

where $\|X_k - \boldsymbol{Z}^T\boldsymbol{g}\|^V = \frac{1}{n}\text{E}\Big\{\sum_{i=1}^n(\underline{X}_{ik} - \underline{(\boldsymbol{Z}^T\boldsymbol{g})}_i)^T\boldsymbol{V}_i^{-1}(\underline{X}_{ik} - \underline{(\boldsymbol{Z}^T\boldsymbol{g})}_i)\Big\}$, with $\underline{X}_{ik} = (X_{i1k}, \ldots, X_{im_ik})^T$ and $\underline{(\boldsymbol{Z}^T\boldsymbol{g})}_i = (\boldsymbol{Z}_{i1}^T\boldsymbol{g}(T_{i1}), \ldots, \boldsymbol{Z}_{im_i}^T\boldsymbol{g}(T_{im_i}))$. Hereafter we write $\boldsymbol{\varphi}_{\boldsymbol{V}k}^* = \Pi_{\boldsymbol{V}}X_k \in \overline{\boldsymbol{G}}$ and $\overline{\boldsymbol{\varphi}}_{\boldsymbol{V}k} = \Pi_{\boldsymbol{V}n}X_k \in \boldsymbol{G}_B$.

Assumption S

(i) The projections $\boldsymbol{\varphi}^*_{\boldsymbol{V}k}(t)$, $k = 1, \ldots, p$, and the varying coefficient function $\boldsymbol{g}_0$ are twice continuously differentiable on $[0, 1]$, and they and their second order derivatives are uniformly bounded in n.

(ii) We take $K_n = \lfloor c_K n^{1/5} \rfloor$ for some positive constant c_K, where $\lfloor x \rfloor$ is the largest integer no greater than x.

Assumption S(i) is a mild and standard assumption for semiparametric models. We consider the existence and smoothness properties of $\boldsymbol{\varphi}^*_{\boldsymbol{V}k}(t)$ in Section 5. Recall that all the covariates are assumed to be uniformly bounded. Since the relevant functions are assumed to be at least twice continuously differentiable, we recommend quadratic or cubic spline approximation. Then the order of K_n specified in Assumption S(ii) is optimal. If the smoothness of different functions varies, we refer to [1] for the convergence rate interfere phenomenon.

The following matrices are necessary in order to present asymptotic normality of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{V}}$:

$$\boldsymbol{H} = \begin{pmatrix} \sum_{i=1}^n \underline{\boldsymbol{X}}_i^T \boldsymbol{V}_i^{-1} \underline{\boldsymbol{X}}_i & \sum_{i=1}^n \underline{\boldsymbol{X}}_i^T \boldsymbol{V}_i^{-1} \underline{\boldsymbol{W}}_i \\ \sum_{i=1}^n \underline{\boldsymbol{W}}_i^T \boldsymbol{V}_i^{-1} \underline{\boldsymbol{X}}_i & \sum_{i=1}^n \underline{\boldsymbol{W}}_i^T \boldsymbol{V}_i^{-1} \underline{\boldsymbol{W}}_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} \end{pmatrix}, \tag{2.4}$$

$$\boldsymbol{H}_{11\cdot 2} = \boldsymbol{H}_{11} - \boldsymbol{H}_{12}\boldsymbol{H}_{22}^{-1}\boldsymbol{H}_{21}\,, \quad \text{and} \quad \boldsymbol{H}^{11} = (\boldsymbol{H}_{11\cdot 2})^{-1}\,.$$

Let $\Omega_{\boldsymbol{V}n}$ be a $p \times p$ matrix whose (k, l)th element is

$$\begin{aligned} &\langle X_k - \boldsymbol{Z}^T \boldsymbol{\varphi}^*_{\boldsymbol{V}k}, X_l - \boldsymbol{Z}^T \boldsymbol{\varphi}^*_{\boldsymbol{V}l} \rangle^V \\ &\quad = \frac{1}{n} \sum_{i=1}^n \mathrm{E}\Big\{ (\underline{X}_{ik} - \underline{(\boldsymbol{Z}^T \boldsymbol{\varphi}^*_{\boldsymbol{V}k})}_i)^T \boldsymbol{V}_i^{-1} (\underline{X}_{il} - \underline{(\boldsymbol{Z}^T \boldsymbol{\varphi}^*_{\boldsymbol{V}l})}_i) \Big\}. \end{aligned}$$

Note that $n^{-1}\boldsymbol{H}_{11\cdot 2}$ is an estimate of $\Omega_{\boldsymbol{V}n}$. We assume that there exists a $p \times p$ positive definite matrix $\Omega_{\boldsymbol{V}}$ such that

$$\lim_{n\to\infty} \Omega_{\boldsymbol{V}n} = \Omega_{\boldsymbol{V}}. \tag{2.5}$$

Now we are ready to state the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{V}}$ under general error distributions as specified in Assumption A6 given in Section 5. Its proof is given in the supplement [5]. We denote the normal distribution with mean η and covariance Ω by $\mathrm{N}(\eta, \Omega)$, and by "$\xrightarrow{d}$" we mean convergence in distribution. Let $\boldsymbol{I}_l$ be the l-dimensional identity matrix.

PROPOSITION 1. *(Asymptotic normality of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{V}}$) Under Assumption S, (2.5), and Assumptions A1-6 given in Section 5, we have*

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{V}} = \boldsymbol{\beta}_0 + \boldsymbol{H}^{11} \sum_{i=1}^{n} (\underline{\boldsymbol{X}}_i - \underline{\boldsymbol{W}}_i \boldsymbol{H}_{22}^{-1} \boldsymbol{H}_{21})^T \boldsymbol{V}_i^{-1} \underline{\epsilon}_i + o_p\Big(\frac{1}{\sqrt{n}}\Big).$$

We also have

$$\Gamma_{\boldsymbol{V}}^{-1/2} (\widehat{\boldsymbol{\beta}}_{\boldsymbol{V}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{I}_p),$$

where $\Gamma_{\boldsymbol{V}}$ is given by

$$\boldsymbol{H}^{11} \sum_{i=1}^{n} \Big\{ (\underline{\boldsymbol{X}}_i - \underline{\boldsymbol{W}}_i \boldsymbol{H}_{22}^{-1} \boldsymbol{H}_{21})^T \boldsymbol{V}_i^{-1} \boldsymbol{\Sigma}_i \boldsymbol{V}_i^{-1} (\underline{\boldsymbol{X}}_i - \underline{\boldsymbol{W}}_i \boldsymbol{H}_{22}^{-1} \boldsymbol{H}_{21}) \Big\} \boldsymbol{H}^{11}. \tag{2.6}$$

Under (2.5), $\widehat{\boldsymbol{\beta}}_{\boldsymbol{V}}$ is $\sqrt{n}$-consistent for $\boldsymbol{\beta}_0$. We can estimate its asymptotic covariance $\Gamma_{\boldsymbol{V}}$ given in (2.6) by replacing the $\boldsymbol{\Sigma}_i$'s with some estimates based on $\widehat{\boldsymbol{\beta}}_{\boldsymbol{V}}$ and $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{V}}$. For example, we can replace $\boldsymbol{\Sigma}_i$ with $\widetilde{\underline{\epsilon}}_i \widetilde{\underline{\epsilon}}_i^T$ where

$$\widetilde{\underline{\epsilon}}_i = \underline{Y}_i - \underline{\boldsymbol{X}}_i^T \widehat{\boldsymbol{\beta}}_{\boldsymbol{V}} - \underline{\boldsymbol{W}}_i^T \widehat{\boldsymbol{\gamma}}_{\boldsymbol{V}}.$$

However, this approach may be too crude and it does not make use of the common information on the covariance structure contained in different subjects. Alternatively, we can estimate the $\boldsymbol{\Sigma}_i$'s by applying smoothing techniques to some residuals based on some assumption on the covariance structure. We investigate this problem in Section 3.

Next, Proposition 2 gives the semiparametric efficiency bound for estimation of $\boldsymbol{\beta}_0$. It can be proved in almost the same way as in Section 4.4 of [13] and Lemma 1 of [4] and the proof is omitted. We denote the semiparametric efficient score function of $\boldsymbol{\beta}$ by

$$\boldsymbol{l}_{\boldsymbol{\beta}}^* = (l_{\boldsymbol{\beta}1}^*, \ldots, l_{\boldsymbol{\beta}p}^*)^T.$$

Its expression is given in Proposition 2. Then we denote $\boldsymbol{\varphi}_{\boldsymbol{\Sigma}k}^*(t)$ by $\boldsymbol{\varphi}_{eff,k}^*(t)$ when $\boldsymbol{V}_i = \boldsymbol{\Sigma}_i$ in (2.1).

PROPOSITION 2. *(Semiparametric efficiency bound) Under the same assumptions as in Proposition 1, we have*

$$l_{\boldsymbol{\beta}k}^* = \sum_{i=1}^{n} (\underline{X}_{ik} - \underline{(\boldsymbol{Z}^T \boldsymbol{\varphi}_{eff,k}^*)}_i)^T \boldsymbol{\Sigma}_i^{-1} \{\underline{Y}_i - \underline{\boldsymbol{X}}_i^T \boldsymbol{\beta}_0 - \underline{(\boldsymbol{Z}^T \boldsymbol{g}_0)}_i\},$$

and the semiparametric efficient information matrix for $\boldsymbol{\beta}$ is given by

$$\lim_{n\to\infty} \frac{1}{n} \mathrm{E}\{\boldsymbol{l}_{\boldsymbol{\beta}}^* (\boldsymbol{l}_{\boldsymbol{\beta}}^*)^T\} = \Omega_{\boldsymbol{\Sigma}} \text{ with } \boldsymbol{V}_i = \boldsymbol{\Sigma}_i \text{ in (2.5)}.$$

Proposition 3 gives the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}$, the so called oracle estimator, which uses the true covariance structure in the GEE spline regression. It also asserts that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}$ achieves the semiparametric efficiency bound derived from Proposition 2. The proof is given in the supplement [5].

PROPOSITION 3. *(Oracle efficient estimator) If we take $\boldsymbol{V}_i = \boldsymbol{\Sigma}_i$ in (2.2) then, under the same assumptions as in Proposition 1, we have*

$$\sqrt{n}\,\Omega_{\boldsymbol{\Sigma}}^{1/2}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{I}_p).$$

In practice, usually the $\boldsymbol{\Sigma}_i$'s are unknown and we have no direct access to the semiparametric efficient score function or the oracle estimator. In the next section we study nonparametric estimation of the covariances so as to improve the efficiency.

3. Efficient estimation. The semiparametric efficiency bound of $\boldsymbol{\beta}$ given in Proposition 2 indicates that knowledge, or at least estimation, of the $\boldsymbol{\Sigma}_i$'s is necessary in order to construct a semiparametric efficient estimator. On the other hand, as discussed in the Introduction, when the $\boldsymbol{\Sigma}_i$'s are unknown it is almost impossible to estimate them in a fully nonparametric way. Fortunately, for longitudinal or clustered data sets, it is reasonable to make some assumptions such as

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}(\underline{T}_i),\ i = 1, \ldots, n, \tag{3.1}$$

where the (j, j)th element of $\boldsymbol{\Sigma}_i$ is given by $\sigma^2(T_{ij})$ and the (j, j')th element is given by $\sigma(T_{ij}, T_{ij'})$ when $j \neq j'$, for some smooth functions $\sigma^2(t)$ and $\sigma(s, t)$. Based on (3.1), in Section 3.1 we construct nonparametric estimates of the covariances and then use them to derive an FGLS procedure to improve the efficiency, and we show in Section 3.2 its asymptotic equivalence to the oracle estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}$. We also discuss estimation of the nonparametric component.

3.1. *Methodology.* A preliminary estimation of $\boldsymbol{\beta}_0$ and $\boldsymbol{g}_0$ is necessary before we can estimate the covariances. For simplicity and robustness, we utilize working independence in the GEE spline estimation. As noted following Proposition 1 we could then use the resultant residuals to estimate the covariance matrices directly. However it is intuitively better to further make use of the covariance structure (3.1) by applying some nonparametric smoothing techniques to the residuals. In addition, alternative to the spline estimator, we could apply smoothing techniques to the pseudo responses $\underline{Y}_i - \underline{\boldsymbol{X}}_i^T \widehat{\boldsymbol{\beta}}_{\boldsymbol{V}}$ to obtain another estimator of $\boldsymbol{g}_0$. We take this latter approach for technical and numerical reasons given in Remark 1. After the preliminary

estimation, for each $i = 1, \ldots, n$, we estimate $\boldsymbol{\Sigma}_i$ by applying local linear regression and denote the resultant estimate by $\widehat{\boldsymbol{\Sigma}}_i$. Our final estimator of $\boldsymbol{\beta}_0$ is then obtained by taking $\boldsymbol{V}_i = \widehat{\boldsymbol{\Sigma}}_i$, $i = 1, \ldots, n$, in the GEE spline estimation. Note that in the trivial case where m_i is fixed for all i and the T_{ij}'s are equi-spaced, we can estimate $\boldsymbol{\Sigma}_i$ without using any smoothing techniques.

Let K be a given kernel function. Our estimation procedure is formally specified as follows:

Step 1. Estimate $\boldsymbol{\beta}_0$ by the GEE spline method given in Section 2 with $\boldsymbol{V}_i = \boldsymbol{I}_{m_i}$, $i = 1, \ldots, n$, and denote the resultant working independence estimate by $\widehat{\boldsymbol{\beta}}_I$.

Step 2. Estimate $\boldsymbol{g}_0(t)$ by applying local linear regression to $\{Y_{ij} - \boldsymbol{X}_{ij}^T \widehat{\boldsymbol{\beta}}_I,\ i = 1, \ldots, n, j = 1, \ldots, m_i\}$, using bandwidth h_1. We denote the resultant estimate by $\widehat{\boldsymbol{g}}(t)$, which is written as

(3.2)
$$\widehat{\boldsymbol{g}}(t) = D_q (A_{1n}(t))^{-1} \frac{1}{N_1 h_1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \boldsymbol{Z}_{ij} \otimes \begin{pmatrix} 1 \\ \frac{T_{ij}-t}{h_1} \end{pmatrix} K\Big(\frac{T_{ij}-t}{h_1}\Big)(Y_{ij} - \boldsymbol{X}_{ij}^T \widehat{\boldsymbol{\beta}}_I),$$

where $N_1 = \sum_{i=1}^{n} m_i$, $D_q = \boldsymbol{I}_q \otimes (1 \quad 0)$, and

$$A_{1n}(t) = \frac{1}{N_1 h_1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (\boldsymbol{Z}_{ij} \boldsymbol{Z}_{ij}^T) \otimes \begin{pmatrix} 1 & \frac{T_{ij}-t}{h_1} \\ \frac{T_{ij}-t}{h_1} & (\frac{T_{ij}-t}{h_1})^2 \end{pmatrix} K\Big(\frac{T_{ij}-t}{h_1}\Big).$$

Step 3. Calculate the residuals, denoted as $\widehat{\epsilon}_{ij}$, given by

$$\widehat{\epsilon}_{ij} = Y_{ij} - \boldsymbol{X}_{ij}^T \widehat{\boldsymbol{\beta}}_I - \boldsymbol{Z}_{ij}^T \widehat{\boldsymbol{g}}(T_{ij}),\ i = 1, \ldots, n, j = 1, \ldots, m_i.$$

Step 4. Estimate the variance function $\sigma^2(t)$ by applying to the squared residuals local linear regression with bandwidth h_2. Denote the resultant estimate by $\widehat{\sigma^2}(t)$; it can be expressed as

$$\widehat{\sigma^2}(t) = (1\,0)(A_{2n}(t))^{-1} \frac{1}{N_1 h_2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \begin{pmatrix} 1 \\ \frac{T_{ij}-t}{h_2} \end{pmatrix} K\big(\frac{T_{ij}-t}{h_2}\big)(\widehat{\epsilon}_{ij})^2, \tag{3.3}$$

where $A_{2n}(t) = \frac{1}{N_1 h_2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \begin{pmatrix} 1 & \frac{T_{ij}-t}{h_2} \\ \frac{T_{ij}-t}{h_2} & (\frac{T_{ij}-t}{h_2})^2 \end{pmatrix} K\big(\frac{T_{ij}-t}{h_2}\big)$.

Step 5. Estimate the covariance function $\sigma(s,t)$ by applying to $\{\widehat{\epsilon}_{ij}\widehat{\epsilon}_{ij'},\ j \neq j', i = 1, \ldots, n\}$ local linear regression with bandwidth h_3. We denote

the resultant estimate by $\widehat{\sigma}(s,t)$; it has the following expression:

$$\widehat{\sigma}(s,t) = (1\,0\,0)(A_{3n}(s,t))^{-1} \times \frac{1}{N_2 h_3^2}\sum_{i=1}^{n}\sum_{j\neq j'}\begin{pmatrix}1\\ \frac{T_{ij}-s}{h_3}\\ \frac{T_{ij'}-t}{h_3}\end{pmatrix} K(\frac{T_{ij}-s}{h_3})K(\frac{T_{ij'}-t}{h_3})\widehat{\epsilon}_{ij}\widehat{\epsilon}_{ij'}, \tag{3.4}$$

where $N_2 = \sum_{i=1}^{n} m_i(m_i-1)$ and

$$A_{3n}(s,t) = \frac{1}{N_2 h_3^2}\sum_{i}^{n}\sum_{j\neq j'}\begin{pmatrix}1\\ \frac{T_{ij}-s}{h_3}\\ \frac{T_{ij'}-t}{h_3}\end{pmatrix}\begin{pmatrix}1 & \frac{T_{ij}-s}{h_3} & \frac{T_{ij'}-t}{h_3}\end{pmatrix} K(\frac{T_{ij}-s}{h_3})K(\frac{T_{ij'}-t}{h_3}).$$

Step 6. Calculate $\widehat{\boldsymbol{\Sigma}}_i$ by combining the results from steps 4 and 5 by letting

$$\widehat{\boldsymbol{\Sigma}}_i(j,j') = \widehat{\sigma}(T_{ij},T_{ij'})I(j\neq j') + \widehat{\sigma^2}(T_{ij})I(j=j'),$$

and then estimate $\boldsymbol{\beta}_0$ with $\boldsymbol{V}_i = \widehat{\boldsymbol{\Sigma}}_i$ in the GEE (2.2). Denote the resultant estimate of $\boldsymbol{\beta}_0$ by $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}}}$.

Step 7. Update the nonparametric estimator of $\boldsymbol{g}_0(t)$ given in Step 2 by replacing $Y_{ij} - \boldsymbol{X}_{ij}^T\widehat{\boldsymbol{\beta}}_I$ with $Y_{ij} - \boldsymbol{X}_{ij}^T\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}}}$, $i = 1,\ldots,n, j = 1,\ldots,m_i$. Denote the resultant estimator by $\widehat{\boldsymbol{g}}_U(t)$. Alternatively, we can estimate $\boldsymbol{g}_0(t)$ with splines, by replacing $\boldsymbol{\beta}$ with $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}}}$ and taking $\boldsymbol{V}_i = \widehat{\boldsymbol{\Sigma}}_i$ in the GEE (2.2). Denote the resultant estimator by $\widehat{\boldsymbol{g}}_S(t)$.

In general the covariance function estimate $\widehat{\sigma}(s,t)$ given by step 5 may not be positive semidefinite. We can modify it by truncating the eigenfunctions in its spectral decomposition that have eigenvalues not exceeding some nonnegative constant λ_L. Then we have positive definite covariance estimates if we replace $\widehat{\sigma}(s,t)$ with this modified version in step 6.

REMARK 1. *When we calculate $\widehat{\boldsymbol{\beta}}_I$ in step 1, we also have $\widehat{\boldsymbol{\gamma}}_I$ and get the set of residuals $\{\widetilde{\epsilon}_{ij} = Y_{ij} - \boldsymbol{X}_{ij}^T\widehat{\boldsymbol{\beta}}_I - \boldsymbol{W}_{ij}^T\widehat{\boldsymbol{\gamma}}_I\}$. Then we could omit steps 2 and 3 of our procedure by exploiting this set of residuals when we estimate $\boldsymbol{\Sigma}_i$ in steps 4-6. However, our simulation results summarized in Section 4 indicate that this simplified approach is inferior to the proposed one. Intuitively speaking, to achieve the semiparametric efficiency in the GEE spline estimation of $\boldsymbol{\beta}_0$, to some extent the accompanying estimation of $\boldsymbol{g}_0(t)$ requires undersmoothing and thus it often exhibits spurious wiggling patterns. Besides, it is difficult to justify theoretically this simplified approach as the local property of spline estimators seems to be intractable.*

3.2. *Asymptotic results.* First we establish the asymptotic equivalence between the data-driven estimator $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}}}$ and the oracle estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}$ by exploiting some desirable properties of $\widehat{\boldsymbol{\Sigma}}_i$. First, we specify our assumptions on the smoothness of $\boldsymbol{g}_0(t)$, $\sigma^2(t)$ and $\sigma(s,t)$. We need Assumption B given below, which is more restrictive than usual, in order to evaluate the difference between $\widehat{\boldsymbol{\Sigma}}_i^{-1}$ and $\boldsymbol{\Sigma}_i^{-1}$.

Assumption B.

(i) Assumption (3.1) holds.
(ii) The true varying coefficient function $\boldsymbol{g}_0(t)$ is three times continuously differentiable on $[0,1]$.
(iii) The variance function $\sigma^2(t)$ is three times continuously differentiable on $[0,1]$.
(iv) The covariance function $\sigma(s,t)$ is three times continuously differentiable on $[0,1]^2$.

In the following we collect our assumptions on the kernel function K and the three bandwidths used in the construction of the proposed estimator. Assumption H(i) on K is a standard one. When Assumption B holds, our assumptions on the bandwidths h_1, h_2 and h_3 are not restrictive. For example, the optimal order of h_1 and h_2 is $n^{-1/5}$ which falls into the specified range. A larger order is recommended only for h_3 due to the two-dimensional smoothing in step 5. However, since the effective number of observations used in step 5 of the procedure is N_2 we anticipate that bandwidth choice will not seriously affect the performance of our final estimator.

Assumption H.

(i) The kernel function K is some continuously differentiable symmetric density function with a compact support.
(ii) The bandwidths h_1, h_2 and h_3 satisfy $h_1 = c_1 n^{-a_h}$ for some $1/6 < a_h \leq 1/4$, $h_2 = c_2 n^{-b_h}$ for some $1/6 < b_h \leq 1/4$ and $h_3 = c_3 n^{-c_h}$ for some $1/6 < c_h < 1/4$, where c_1, c_2 and c_3 are some positive constants.

The asymptotic expression of $\widehat{\boldsymbol{\Sigma}}_i$ is given in Proposition 4, which is verified in the supplementary material [5]. Note that we need more elaborate representations than those used by [14] since we deal with a $(p+qK_n)$-dimensional linear regression model. Note also that the functions B_j, $j = 1, \ldots, 4$, that appear in Proposition 4 are implicitly defined in the proof of the proposition and only their boundedness property is needed in the proof of Theorem 1.

PROPOSITION 4. *(Representations of the covariance estimators) Under the assumptions in Proposition 1 with* $\boldsymbol{V}_i = \boldsymbol{I}_{m_i}$, *and Assumptions B and H,*

we have the following representations of $\widehat{\sigma^2}(t)$ *and* $\widehat{\sigma}(s,t)$*. Uniformly in* t*,*

$$\widehat{\sigma^2}(t) - \sigma^2(t) = B_1(t)h_2^2 + B_2(t)E_1(t) + O_p(h_1^3 + h_2^3) + O_p\Big(\frac{\log n}{nh_1} + \frac{\log n}{nh_2}\Big)$$

where uniformly in t

$$E_1(t) = \frac{1}{N_1 h_2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\begin{pmatrix}1\\ \frac{T_{ij}-t}{h_2}\end{pmatrix} K\Big(\frac{T_{ij}-t}{h_2}\Big)(\epsilon_{ij}^2 - \sigma^2(T_{ij})) = O_p\Big(\sqrt{\frac{\log n}{nh_2}}\Big),$$

and $B_1(t)$ *and* $B_2(t)$ *are bounded functions. Uniformly in* s *and* t $(s \neq t)$*,*

$$\widehat{\sigma}(s,t) - \sigma(s,t) = B_3(s,t)h_2^2 + B_4(s,t)E_2(s,t) + O_p(h_1^3 + h_3^3) + O_p\Big(\frac{\log n}{nh_1} + \frac{\log n}{nh_3^2}\Big),$$

where

$$\begin{aligned} E_2(s,t) &= \frac{1}{N_2 h_3^2}\sum_{i=1}^{n}\sum_{j\neq j'}\begin{pmatrix}1\\ \frac{T_{ij}-s}{h_3}\\ \frac{T_{ij'}-t}{h_3}\end{pmatrix} K\Big(\frac{T_{ij}-s}{h_3}\Big)K\Big(\frac{T_{ij'}-t}{h_3}\Big)(\epsilon_{ij}\epsilon_{ij'} - \sigma(T_{ij}, T_{ij'})) \\ &= O_p\Big(\sqrt{\frac{\log n}{nh_3^2}}\Big) \quad \textit{uniformly in } s \textit{ and } t, \end{aligned}$$

and $B_3(s,t)$ *and* $B_4(s,t)$ *are bounded functions.*

We state in Theorem 1 the desirable equivalence property of $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}}}$ to the oracle estimator. The proof uses Proposition 4; it is tedious and technical and thus is postponed to Section 5.4. We have not yet obtained a similar result for general link functions even when the m_i's are uniformly bounded, and that is a future research topic.

THEOREM 1. *Under the assumptions in Proposition 4, we have*

$$\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} + o_p(n^{-1/2}).$$

Suppose (3.1) fails to hold, but $\mathrm{Var}(\underline{\epsilon}_i \,|\, \underline{T}_i)$ still can be represented by some functions $\sigma^2(t)$ and $\sigma(s,t)$. Then Proposition 1 and Theorem 1 continue to hold $\boldsymbol{\Sigma_i} = \mathrm{Var}(\underline{\epsilon}_i \,|\, \underline{\boldsymbol{X}}_i, \underline{\boldsymbol{Z}}_i, \underline{T}_i)$ is replaced by $\mathrm{Var}(\underline{\epsilon}_i \,|\, \underline{T}_i)$. We are still exploiting the information on $\mathrm{Var}(\underline{\epsilon}_i \,|\, \underline{T}_i)$.

Besides, we can replace the three times continuously differentiability with the twice continuously differentiability and the Hölder continuity of the second derivatives of order α_1, α_2, and α_3 in assumptions B(ii), B(iii), and

B(iv), respectively. In this case, the bandwidths in steps 2, 4, and 5 of our method have to satisfy the condition $\sqrt{n}(h_1^{2+\alpha_1} + h_2^{2+\alpha_2} + h_3^{2+\alpha_3}) \to 0$. Note that α_3 must be positive because step 5 of our procedure requires two-dimensional smoothing. Then we can prove similar results when $0 \le \alpha_1 < 1$, $0 \le \alpha_2 < 1$, and $0 < \alpha_3 < 1$. Specifically, the $O_p(h_j^3)$ terms in Proposition 4 will be replaced by $O_p(h_j^{2+\alpha_j})$, $j = 1, 2, 3$.

REMARK 2. *In Proposition 2, no assumptions on the structure of the $\mathbf{\Sigma}_i$'s or the conditional normality of the $\underline{\epsilon}_i$'s is imposed. However, as mentioned before it is difficult to estimate the $\mathbf{\Sigma}_i$'s in a fully nonparametric way and thus we impose assumption (3.1). On the other hand, when (3.1) holds, we should use this information in calculating the semiparametric efficient score function. Unfortunately, under general errors this task seems intractable and we have no results in this regard. Nevertheless, when (3.1) and some regularity conditions hold, we come up with some remedies to improve the efficiency, as compared to using some working covariance structure. Indeed, $\widehat{\boldsymbol{\beta}}_{\widehat{\mathbf{\Sigma}}}$ has the smallest asymptotic variance among all $\widehat{\boldsymbol{\beta}}_{\mathbf{V}}$ in this case, based on Propositions 1-3, Theorem 1, and the fact that it is an FGLS estimator. Furthermore, it is semiparametric efficient when $\underline{\epsilon}_i$ is normally distributed conditionally on $\underline{\boldsymbol{X}}_i$, $\underline{\boldsymbol{Z}}_i$ and $\underline{T}_i$, as discussed in A.1 of [23].*

Suppose we use cubic splines in the final spline estimator given in Step 7. Then, under the assumptions in Proposition 4 and assume the minimum eigenvalue of $\boldsymbol{H}_{22.1} = \boldsymbol{H}_{22} - \boldsymbol{H}_{21}\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{21}$ is bounded below by Cn/K_n for some positive constant C, we can show the following asymptotic normality:

$$\sqrt{n/K_n}\mathbf{\Psi}(t)^{-1/2}\big(\widehat{\boldsymbol{g}}_S(t) - \boldsymbol{g}_0(t)\big) \xrightarrow{d} \mathrm{N}\big(0, \boldsymbol{I}_q\big),$$

where $\mathbf{\Psi}(t) = \lim_{n\to\infty} nK_n^{-1}(\boldsymbol{I}_q \otimes \boldsymbol{B}(t)^T)\boldsymbol{H}_{22.1}^{-1}(\boldsymbol{I}_q \otimes \boldsymbol{B}(t))$. As for the updated local linear estimator given in Step 7, let $\mu_2 = \int u^2 K(u)du$ and $\nu_0 = \int K(u)^2 du$, and suppose the assumptions in Proposition 4 hold and $h_1 = Cn^{-1/5}$, then we have the following asymptotic normality:

$$\sqrt{N_1 h_1}\big(\widehat{\boldsymbol{g}}_U(t) - \boldsymbol{g}_0(t) - \frac{h_1^2}{2}\mu_2 \boldsymbol{g}_0''(t)\big) \xrightarrow{d} \mathrm{N}\big(0, \nu_0 \mathbf{\Psi}_U(t)\big)$$

where $\mathbf{\Psi}_U(t) = \mathbf{\Lambda}_1^{-1}\mathbf{\Lambda}_2\mathbf{\Lambda}_1^{-1}$, $\mathbf{\Lambda}_1 = \lim\limits_{n\to\infty} \dfrac{1}{N_1}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m_i} \mathrm{E}(\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^T | T_{ij} = t) f_{ij}(t)$, $\mathbf{\Lambda}_2 = \lim\limits_{n\to\infty} \dfrac{1}{N_1}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m_i} \mathrm{E}(\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^T | T_{ij} = t) f_{ij}(t) \mathrm{E}(\epsilon_{ij}^2 | T_{ij} = t)$, and $f_{ij}(t)$ denotes the density of T_{ij}.

4. Numerical studies.

4.1. *Simulation study.* In our simulation study summarized in this section, the data were generated from the following model:

$$Y_{ij} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_0 + \boldsymbol{Z}_{ij}^T\boldsymbol{g}_0(T_{ij}) + \epsilon_i(T_{ij}),\ j = 1, \ldots, m_i,\ i = 1, \ldots, n,$$

with the first component of $\boldsymbol{Z}_{ij}$ being taken as 1. The number of observation time points in the ith subject was set as $m_i = m_0 + binomial(m_r, 0.65)$. Then the observation time points T_{ij} were uniformly distributed over the interval $[(j-1)/(m_0+m_r), j/(m_0+m_r)]$, $j = 1, \cdots, m_i$. We note that when $m_i = m_0 + m_r$, the subject is observed at all follow-up time points; when $m_i < m_0 + m_r$, the subject may be lost to follow up. This setup is intended to model real and more complicated scenarios that often happen in practice. We set $m_0 = 6$ and $m_r = 6$. We generated the other $(p+q-1)-$dimensional covariates from a multivariate Gaussian distribution, and we considered the following coefficients settings:

$p = 4$, $q = 4$, $\boldsymbol{\beta}_0 = (5, 5, -5, -5)^T$ and
$\boldsymbol{g}_0(t) = \big(3.5\sin(2\pi t), 5(1-t)^2, 3.5(\exp(-(3t-1)^2)+\exp(-(4t-3)^2))-1.5, 3.5t^{1/2}\big)^T$.

The random error process $\epsilon_i(t)$ was simulated from an ARMA$(1,1)$ Gaussian process with mean zero and covariance function $\mathrm{cov}(\epsilon_i(s), \epsilon_i(t)) = \omega\rho^{|s-t|}$. We set $\omega = 4.95$ and considered $\rho = 0.4$ or 0.8.

We considered two types of working covariance structure: working independence covariances and the proposed covariance estimates. For the sake of comparison, we also considered using the true covariances and using the covariance estimator with the crude raw residuals obtained from Step 1.

Throughout the numerical studies, following [9], we used cubic splines and took the spline dimension K_n as $K_n = \lfloor 2n^{1/5} \rfloor$. For the efficient estimator, h_1 and h_2 were selected via the commonly used leave-one-subject-out cross-validation, and the bandwidth h_3 was set as $h_3 = 2h_1$. We report in Table 1 the average estimation bias and estimated standard error (SE) obtained from 200 repetitions. The empirical standard errors are very close to the estimated standard errors and thus are omitted. In general, the efficient estimator could yield smaller estimation bias and variance, compared to the naive estimator assuming working independence. In particular, the standard error for the efficient estimator is only $20 \sim 50\%$ of that of the working independence estimator, indicating a remarkable reduction. In addition, we note that the efficient estimator has very similar performance to that of the oracle estimator. Regarding the crude estimator, as it is based on a simplified

residual construction it produces relatively less accurate covariance estimation. Thus, its estimation bias and standard error are respectively larger than that for the efficient estimator.

TABLE 1

Estimation results of 200 simulations. "Independent" corresponds to $\boldsymbol{V}_i = \boldsymbol{I}_{m_i}$; "Efficient" refers to using $\boldsymbol{V}_i = \widehat{\boldsymbol{\Sigma}}_i$; "Oracle" refers to using the true Σ_i as V_i; "Crude" refers to using residuals directly from Step 1 to estimate the covariances.

			Independent		Efficient		Oracle		Crude		Quadratic	
n	ρ		bias	SE	bias	SE	bias	SE	bias	SE	bias	SE
100	0.4	β_1	.0214	.0726	.0128	.0366	.0133	.0245	.0165	.0425	.0154	.0421
		β_2	-.0218	.0727	-.0186	.0362	-.0146	.0251	-.0165	.0442	.0102	.0425
		β_3	-.0309	.0718	-.0126	.0364	-.0147	.0245	-.0127	.0435	.0095	.0455
		β_4	.0199	.0736	.0145	.0369	.0132	.0246	.0210	.0438	-.0113	.0398
200	0.4	β_1	-.0072	.0525	-.0082	.0247	-.0028	.0176	-.0122	.0337	.0049	.0302
		β_2	.0088	.0528	.0136	.0226	.0034	.0174	.0115	.0356	.0089	.0345
		β_3	-.0071	.0526	.0075	.0256	.0112	.0174	-.0146	.0354	-.0076	.0312
		β_4	.0094	.0525	.0124	.0272	.0132	.0178	-.0204	.0355	-.0075	.0305
100	0.8	β_1	.0257	.0723	.0245	.0334	-.0070	.0109	.0347	.033	.0112	.0378
		β_2	-.0179	.0731	-.0122	.0328	-.0112	.0106	.0436	.0332	-.0109	.0344
		β_3	.0388	.0729	-.0257	.0335	.0214	.0107	.0279	.0332	-.0179	.0394
		β_4	-.0193	.0735	.0447	.0334	-.0122	.0108	-.0345	.0326	.0184	.0404
200	0.8	β_1	.0173	.0497	.0149	.0194	.0057	.0089	.0144	.0248	.0089	.0250
		β_2	.0169	.0512	-.0146	.0196	-.0010	.0092	-.0167	.0242	-.0064	.0248
		β_3	-.0364	.0499	.0145	.0190	.0058	.0090	.0135	.0232	-.0053	.0212
		β_4	.0289	.0496	-.0139	.0182	-.0035	.0089	-.0222	.0238	.0083	.0196

There are also other existing methods based on estimating equations. We specifically considered the one based on quadratic inference function (QIF) [18] in which, to incorporate the longitudinal dependence, the correlation matrix is approximated using a matrix expansion. We used the same basis matrices as recommended by [18], i.e., the first order basis matrix with 0 on the diagonal and 1 off-diagonal, which is suitable for unequal cluster sizes and irregular time points. Any negative eigenvalue was set to zero whenever it occurred. From Table 1, we notice that this approach is more efficient than the estimator assuming working independence but is less efficient than our proposed method. The QIF approach indirectly models the correlations using some matrix approximation while our method directly models the covariances. The actual covariance dependence may differ from the pattern suggested by the basis matrices in the quadratic inference function. When that happens the estimation results using QIF method may be less satisfactory than our nonparametric approach. Therefore our method may incorporate a more accurate covariance structure in the estimation and thus achieve better efficiency. Besides, the covariance of the estimating function depends on the unknown parameters, and is estimated and integrated in the QIF. This may decrease the stability in solving the optimization problem.

We next considered the situation where m_i might diverge for some subjects i. We randomly selected $n_0 = \mathsf{C}n^{3/8}$ subjects such that their observation points are $\mathsf{B}n^{1/8}m_i$ equally spaced on $[0, 1]$ and we let the ramaining

$n - n_0$ subjects to have m_i observations, where m_i was generated in the same way as described above. All the other model settings are identical to that in the previous simulation studies. For different values of B and C, we obtained the results given in Table 2. We notice that all the considered estimators improve with relatively smaller biases and smaller standard errors as compared with the respective bounded m_i case. The efficient estimator still performs much better than the independent estimator in all cases. We do not report results for the QIF method by [18] here, as it is not tailored for the case of diverging m_i and becomes relatively unstable in this case.

TABLE 2

Estimation results of 200 simulations. "Independent" corresponds to $\boldsymbol{V}_i = \boldsymbol{I}_{m_i}$; "Efficient" refers to using $\boldsymbol{V}_i = \widehat{\boldsymbol{\Sigma}}_i$; "Oracle" refers to using the true Σ_i as V_i. B adjusts the diverging m_i and C controls the proportion of cases with diverging m_i.

$\mathsf{B}=1.5, \mathsf{C}=4$			Independent		Efficient		Oracle	
n	ρ		bias	SE	bias	SE	bias	SE
100	0.4	β_1	.0182	.0707	.0087	.0361	-.0017	.0204
		β_2	-.0186	.0717	-.0172	.0329	-.0055	.0205
		β_3	-.0236	.0702	.0041	.0336	-.0056	.0205
		β_4	.0100	.0702	-.0034	.0346	.0008	.0205
200	0.4	β_1	-.0130	.0517	-.0157	.0228	-.0037	.0153
		β_2	.0146	.0516	.0177	.0227	.0028	.0151
		β_3	-.0151	.0512	.0041	.0224	.0011	.0152
		β_4	-.0076	.0517	.0065	.0229	.0038	.0153
100	0.8	β_1	.0181	.0683	-.0175	.0213	.0028	.0102
		β_2	-.0111	.0682	-.0147	.0203	.0028	.0102
		β_3	-.0030	.0674	-.0105	.0199	-.0015	.0100
		β_4	.0260	.0675	.0125	.0208	.0028	.0101
200	0.8	β_1	-.0017	.0499	-.0024	.0132	.0014	.0076
		β_2	-.0005	.0496	.0006	.0129	-.0001	.0076
		β_3	.0045	.0499	.0041	.0133	.0004	.0076
		β_4	-.0052	.0496	-.0059	.0130	-.0009	.0075
$\mathsf{B}=1.5, \mathsf{C}=4$			Independent		Efficient		Oracle	
n	ρ		bias	SE	bias	SE	bias	SE
100	0.4	β_1	.0105	.0710	.0039	.0315	-.0026	.0174
		β_2	-.0180	.0715	-.0095	.0313	-.0046	.0174
		β_3	-.0122	.0730	-.0104	.0323	.0010	.0176
		β_4	.0141	.0707	.0105	.0317	.0034	.0174
200	0.4	β_1	-.0085	.0510	-.0060	.0223	-.0036	.0134
		β_2	-.0066	.0513	-.0062	.0225	-.0018	.0135
		β_3	.0094	.0510	-.0015	.0225	-.0016	.0136
		β_4	.0062	.0514	.0001	.0224	.0006	.0137
100	0.8	β_1	-.0154	.0703	.0042	.0212	-.0040	.0087
		β_2	-.0152	.0690	.0028	.0215	.0001	.0087
		β_3	.0129	.0677	.0044	.0208	-.0002	.0092
		β_4	-.0076	.0699	-.0032	.0215	.0008	.0088
200	0.8	β_1	-.0141	.0489	.0111	.0157	-.0001	.0067
		β_2	-.0136	.0490	-.0145	.0147	-.0003	.0069
		β_3	.0058	.0491	.0016	.0142	-.0001	.0069
		β_4	.0071	.0483	.0041	.0150	-.0001	.0072

We also conducted additional simulations to examine performance of estimation of the nonparametric coefficients and estimation accuracy of parametric coefficients using modified approaches. For space consideration, we report the results in the supplement [5].

4.2. *Real data example.* We now present an application of our method to the CD4 count data from the AIDS Clinical Trial Group 193A Study [11]. The data came from a randomized, double-blind study of AIDS patients with CD4 counts of $\leq$ 50 cells/mm3. The patients were randomized to one of four treatments with roughly equal group sizes; each consisted of a daily regimen of 600 mg of zidovudine. Treatment 1 is zidovudine alternating monthly with 400 mg didanosine; Treatment 2 is zidovudine plus 225 mg of zalcitabine; Treatment 3 is zidovudine plus 400 mg of didanosine; Treatment 4 is a triple therapy consisting of zidovudine plus 400 mg of didanosine plus 400 mg of nevirapine. Measurements of CD4 counts were scheduled to be collected at baseline and at eight week intervals during the 40 weeks of follow-up. However, the real observation times were unbalanced due to mistimed measurements, skipped visits and dropouts. The number of measurements of CD4 counts during the 40 weeks of follow-up varied from 1 to 9, with a median of 4. The response variable was taken as the log-transformed CD4 counts, Y =log(CD4 counts + 1). There was also gender and baseline age information about each patient. A total of 1309 patients were enrolled in the study. We eliminated the 122 patients who dropped out immediately after the baseline measurement.

We considered the following available covariates: treatments 2, 3 and 4 (coded by three indicator variables for treatment groups 2, 3 and 4, respectively), age (years), sex (coded as 1 for male and 0 for female), and interaction effects between these covariates. Using the group SCAD structure identification procedure of Cheng et al. (2014), we found that the coefficients for treatment 3, treatment 4 and the interaction between treatment 2 and sex are varying, and the coefficients given in Table 3 are constants. The group SCAD procedure also suggested that we remove all the other interaction effects. The estimated varying intercept (i.e. effect of treatment 1) and the varying coefficients are displayed in Figure 1 along with 95% confidence intervals. The curves in the figures are updated local linear estimates without using the covariance function estimates. We used cross validation to select the bandwidth. The constant coefficient estimates and their estimated standard errors are provided in Table 3. To facilitate a comparison, we reported the results using the estimators assuming working independence and the efficient estimator proposed in this paper. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ and

$\underline{\boldsymbol{U}}_i = (\underline{\boldsymbol{X}}_i, \underline{\boldsymbol{W}}_i)$. In practice, the variances for the efficient parameter estimates were obtained from the first p diagonal elements of the following matrix: $\left(\sum_{i=1}^n \underline{\boldsymbol{U}}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \underline{\boldsymbol{U}}_i\right)^{-1}$, and for the working independence parameter estimates the variances were obtained from the first p diagonal elements of the following matrix: $\left(\sum_{i=1}^n \underline{\boldsymbol{U}}_i^T \underline{\boldsymbol{U}}_i\right)^{-1} \sum_{i=1}^n \underline{\boldsymbol{U}}_i^T \widehat{\boldsymbol{\Sigma}}_i \underline{\boldsymbol{U}}_i \left(\sum_{i=1}^n \underline{\boldsymbol{U}}_i^T \underline{\boldsymbol{U}}_i\right)^{-1}$.

TABLE 3

Estimation results for CD4 count data. "Independent" corresponds to using $\boldsymbol{V}_i = \boldsymbol{I}_{m_i}$*; "Efficient" refers to using* $\boldsymbol{V}_i = \widehat{\boldsymbol{\Sigma}}_i$*; "Quadratic" refers to the QIF based method.*

	Independent		Efficient		Quadratic	
Covariates	Coefficients	SE	Coefficients	SE	Coefficients	SE
treatment 2	.3614	.2257	.4038	.2027	.3532	.1318
age	.0946	.0274	.0818	.0245	.0882	.0171
sex	.1704	.1768	.2246	.1587	.1187	.1034
treatment 3:sex	-.2922	.2472	-.2908	.2209	-.2625	.2485
treatment 4:sex	-.5321	.2416	-.5653	.2146	-.5580	.1574

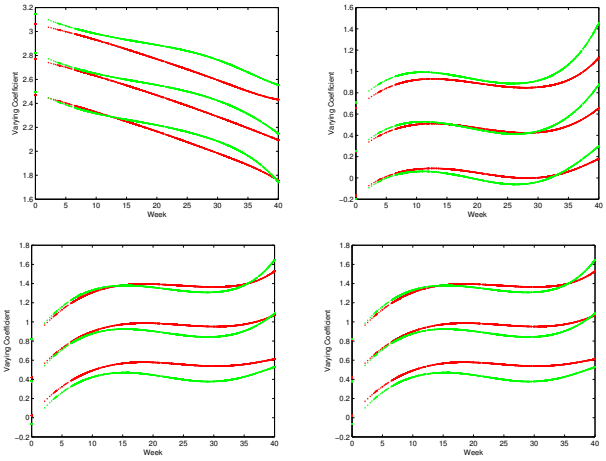

FIG 1. *Estimated varying-coefficients along with 95% confidence intervals for the intercept (upper left), treatment 3 (upper right), treatment 4 (lower left), and interaction between treatment 2 and sex (lower right). The red curves are efficient estimators while the green curves are estimators obtained under working independence.*

From Table 3, we note that the estimated constant coefficients for treatment 2, age, and the interaction between treatment 4 and sex are all quite significant. The constant coefficient estimates for sex are not significant but are still kept in the model since we include the interactions between treatments and sex. The efficient estimates for all the constant and varying coef-

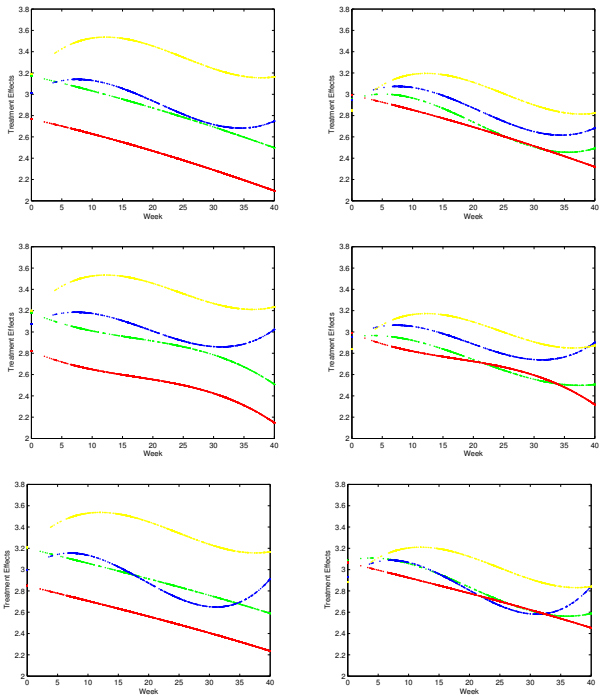

FIG 2. *Estimated treatment effects for the four treatment groups. The panels in the top, middle and bottom rows are respectively the proposed efficient estimates, the estimates assuming independence and the estimates based on the QIF method. The panels in the left and right columns are respectively for the females and the males. Red, green, blue and yellow curves are for treatment groups 1, 2, 3 and 4, respectively.*

ficients have smaller standard errors than the respective estimates assuming working independence. In fact, the Wald test statistic for the coefficient of treatment 2 is $.3614/.2257 = 1.60 < 1.96$ under the working independence, failing to declare a significant difference. On the other hand, the Wald test statistic for the same coefficient is $.4038/.2027 = 1.99 > 1.96$ from the efficient estimation, leading to a significant treatment difference. Other than these, because the sample size in this study was rather large, the two types of estimates for all the constant and varying coefficients appear to be very similar. For the sake of comparison, we also present the estimation results for these regression coefficients from the estimating equation methods based on the QIF method [18]. The conclusions on the estimation significance and effect direction remain the same as for the efficient estimation while the

magnitude of the estimated coefficients slightly differs. For this particular dataset, sometimes the QIF estimator seems to have smaller standard error than the efficient estimator. An explanation is that it choses a covariance structure like compound symmetry in the matrix basis, thus it will be more efficient than our estimator when this structure is plausible (which is possibly the case here). Otherwise, it is generally not as good when the covariance structure is mis-specified.

In general, the CD4 count tends to increase with age in the fitted model. Our estimation results suggest that there exist interaction effects between treatment and sex. Specifically, for the females (sex=0), subjects receiving treatments 2, 3 and 4 tend to have increasingly higher CD4 counts than those under treatment 1. The effect for treatment 2 (as compared with treatment 1) is estimated as a constant and is significant, while those for the other two treatment groups are varying (the upper right and the lower left panels in Figure 1) with even greater positive differences from treatment 1. For the males (sex=1), subjects receiving treatments 2, 3 and 4 also tend to have higher mean CD4 counts than those receiving treatment 1. The interaction between treatment 2 and sex is varying over time (the lower right panel in Figure 1) while those for treatments 3 and 4 are constant. The effects of treatments 3 and 4 are significantly different from that of treatment 1, judging from Table 3. Also, we notice that the differences between treatments seem to be greater between the females than between the males.

The estimated effects of the four treatment groups are plotted in Figure 2 for the efficient estimator, the working independence estimator and the QIF estimator. Note that treatment effects given by the efficient estimator rarely cross each other, giving nice interpretation and ordering of the different treatments, whereas this is not the case for those given by the QIF or the working independence estimator. Previous authors identified a similar pattern on the order of magnitude of the time-varying treatment effects [14]. However, they ignored the interactions between the treatments and sex. Our findings suggest the treatment effect curves might be rather different between the males and the females.

5. Proofs of the main results.

5.1. *Additional assumptions and definitions.* We denote the Euclidean norm of a vector a by $|a|$. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ stand for the minimum and maximum eigenvalues of a symmetric matrix A, respectively. Besides, C, C_1, C_2, ... are generic positive constants whose values may vary from line to line. Recall that the density function of T_{ij} is denoted by by $f_{ij}(t)$, $i = 1, \ldots, n$ and $j = 1, \cdots, m_i$. Also, we denote the joint density func-

tion of T_{ij} and $T_{ij'}$ $(j \neq j')$ by $f_{ijj'}(s,t)$. In Assumptions A1 and A2, we consider sparse and irregular observation times. Note that we carry out two-dimensional smoothing in step 5 and there are three bandwidths involved in our method. Therefore we impose these restrictive assumptions to avoid complicated assumptions involving m_i, $m_{\max}$, and the bandwidths simultaneously. Roughly speaking, these assumptions imply we should have $\sum_{i=1}^{n} m_i^5 = O(n)$.

Assumption A1. For some positive constant C_{A1}, we have $m_{\max} \equiv \max_{1\leq i\leq n} m_i = O(n^{1/8})$ and $\sum_{i=1}^{n} m_i < C_{A1} n$.

Assumption A2. The joint density functions $f_{ij}(t)$ and $f_{ijj'}(s,t)$ are uniformly bounded and we have for some positive constant C_{A2},

$$\frac{1}{C_{A2}} < \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i} f_{ij}(t) \leq \frac{1}{n}\sum_{i=1}^{n} m_i^4 \sum_{j=1}^{m_i} f_{ij}(t) < C_{A2} \text{ on } [0,1], \quad \text{and}$$

$$\frac{1}{C_{A2}} < \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq j'} f_{ijj'}(s,t) \leq \frac{1}{n}\sum_{i=1}^{n} m_i^3 \sum_{j\neq j'} f_{ijj'}(s,t) < C_{A2} \text{ on } [0,1]^2.$$

Assumption A3. For some positive constants C_{A3} and C_{A4}, we have

$$C_{A3}\boldsymbol{I}_{p+q} \leq \mathrm{E}\left\{ \begin{pmatrix} \boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^T & \boldsymbol{X}_{ij}\boldsymbol{Z}_{ij}^T \\ \boldsymbol{Z}_{ij}\boldsymbol{X}_{ij}^T & \boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^T \end{pmatrix} \,\middle|\, T_{ij} \right\} \leq C_{A4}\boldsymbol{I}_{p+q}, \text{ uniformly in } i \text{ and } j.$$

Assumption A4. For some positive constants C_{A5} and C_{A6}, we have $C_{A5} \leq \lambda_{\min}(\boldsymbol{\Sigma}_i) \leq \lambda_{\max}(\boldsymbol{\Sigma}_i) \leq C_{A6} m_i$, uniformly in i.

Assumption A5. For some positive constants C_{A7} and C_{A8}, we have $C_{A7} \leq \lambda_{\min}(\boldsymbol{V}_i) \leq \lambda_{\max}(\boldsymbol{V}_i) \leq C_{A8} m_i$, uniformly in i.

Assumption A6. For some positive constants C_{A9} and C_{A10}, we have $\mathrm{E}\{\exp(C_{A9}|\epsilon_{ij}|) \,|\, \underline{\boldsymbol{X}}_i, \underline{\boldsymbol{Z}}_i, \underline{T}_i\} < C_{A10}$, uniformly in i and j.

Assumption A3 is a standard one and is necessary for identification of the constant coefficients and the varying coefficient functions. When $\underline{\epsilon}_i$ consists of some stochastic process and i.i.d. errors, we have $\boldsymbol{\Sigma_i} = \Xi(\underline{T}_i) + \eta^2 \boldsymbol{I}_{m_i}$, where $\Xi(\underline{T}_i)$ is positive definite. Hence we impose Assumptions A4 and A5 on $\boldsymbol{V}_i$ and $\boldsymbol{\Sigma}_i$, respectively. In [4], it is assumed that $\underline{\epsilon}_i$ has the sub-Gaussian property in order to deal with general link functions. The sub-Gaussian assumption prevents m_i from tending to infinity. Assumption A6, which is less restrictive, is enough for the identity link function since we do not need to employ any results from the empirical process theory in this case.

For $\boldsymbol{g} = (g_1, \ldots, g_q)^T \in \boldsymbol{G}$, we define the sup and L_2 norms by $\|\boldsymbol{g}\|_{G,\infty} = \sum_{j=1}^{q} \sup_{t\in[0,1]} |g_j(t)|$ and $\|\boldsymbol{g}\|_{G,2}^2 = \sum_{j=1}^{q} \int_0^1 g_j^2(t)dt$. Assumptions A2 and

A3 imply there are positive constants C_1 and C_2 such that

$$C_1\|\boldsymbol{g}\|_{G,2} \le \|\boldsymbol{Z}^T\boldsymbol{g}\|^V \le C_2\|\boldsymbol{g}\|_{G,2} \tag{5.1}$$

for any $\boldsymbol{g} \in \boldsymbol{G}$. The details are given in Lemma 1. In (2.3), we define two kinds of projections of X_k. We define another one here:

$$\widehat{\boldsymbol{\varphi}}_{\boldsymbol{V}k} = \widehat{\Pi}_{\boldsymbol{V}n}X_k = \operatorname*{argmin}_{\boldsymbol{g}\in\boldsymbol{G}_B} \|X_k - \boldsymbol{Z}^T\boldsymbol{g}\|_n^V. \tag{5.2}$$

5.2. *Spline approximation and projections.* Recall we assume all the relevant functions are at least twice continuously differentiable and they and their second order derivatives are uniformly bounded. Hence the sup norm of approximation errors by spline functions is bounded from above by $C_{approx}K_n^{-2}$, where C_{approx} depends on the relevant functions. See Corollary 6.26 of [19].

Note that $\langle\cdot,\cdot\rangle^V$ and $\|\cdot\|^V$ are defined on $\{v \mid \sum_{i,j} \mathrm{E}(v_{ij}^2) < \infty\}$ and that $\{\boldsymbol{Z}^T\boldsymbol{g}\}$ is a closed linear subspace due to (5.1). Therefore the projections $\boldsymbol{\varphi}^*_{\boldsymbol{V}k} = (\varphi^*_{\boldsymbol{V}k1}, \ldots, \varphi^*_{\boldsymbol{V}kq})^T$, $k = 1, \ldots, p$, exist uniquely. Next, we set $\boldsymbol{V}_i^{-1} = (v_i^{j_1j_2})$. Note that $\boldsymbol{\varphi}^*_{\boldsymbol{V}k} = \Pi_{\boldsymbol{V}}X_k$ defined in (2.3) satisfies

$$\langle X_k - \boldsymbol{Z}^T\Pi_{\boldsymbol{V}}X_k, \boldsymbol{Z}^T\boldsymbol{g}\rangle^V = 0 \quad \forall \boldsymbol{g} \in \boldsymbol{G}\,.$$

By representing the above equality explicitly, we can derive the following integral equations for $\boldsymbol{\varphi}^*_{\boldsymbol{V}k}(t)$. For $d_1 = 1, \ldots, q$,

$$\sum_{d_2=1}^{q} a_{d_2}^{(d_1)}(t)\varphi^*_{\boldsymbol{V}kd_2}(t) = b^{(d_1)}(t) + \int_0^1 \sum_{d_2=1}^{q} c_{d_2}^{(d_1)}(s,t)\varphi^*_{\boldsymbol{V}kd_2}(s)ds, \tag{5.3}$$

where

$$a_{d_2}^{(d_1)}(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} \mathrm{E}\{Z_{ijd_2}v_i^{jj}Z_{ijd_1} \,|\, T_{ij} = t\}f_{ij}(t),$$

$$b^{(d_1)}(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{1\le j_1,j_2\le m_i} \mathrm{E}\{X_{ij_1k}v_i^{j_1j_2}Z_{ij_2d_1} \,|\, T_{ij_2} = t\}f_{ij_2}(t),$$

$$c_{d_2}^{(d_1)}(s,t) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j_1\ne j_2} \mathrm{E}\{Z_{ij_1d_2}v_i^{j_1j_2}Z_{ij_2d_1} \,|\, T_{ij_1} = s, T_{ij_2} = t\}f_{ij_1j_2}(s,t).$$

Let $\boldsymbol{A}(t)$ be the $q \times q$ matrix whose (d_1, d_2)th element is $a_{d_2}^{(d_1)}(t)$. Assumptions A2 and A3 imply that $|\boldsymbol{A}(t)| \ne 0$ on $[0,1]$ and we set $\psi^*_{\boldsymbol{V}kd_1}(t) = \sum_{d_2=1}^{q} a_{d_2}^{(d_1)}(t)\varphi^*_{\boldsymbol{V}kd_2}(t)$. Then (5.3) reduces to (S.2) of [3] and the same argument there applies. Therefore $\boldsymbol{\varphi}^*_{\boldsymbol{V}k}(t)$ has the required smoothness properties under similar regularity conditions.

5.3. *Remarks on the proofs of Propositions 1–3.* We can proceed as in [13] (and [3]) by replacing Z_{ij}, $\underline{Z}_i$, and $\varphi_k^*(t)$ in [13] (and $\boldsymbol{Z}_{ij}$, $\boldsymbol{Z}_i$, and $\varphi_k^*(\boldsymbol{t})$ in [3]) with $\boldsymbol{W}_{ij}$, $\underline{\boldsymbol{W}}_i$, and $\boldsymbol{Z}^T\boldsymbol{\varphi}^*_{\boldsymbol{V}k}(t)$, respectively. They used several lemmas in their proofs. We reorganize the corresponding lemmas in our setup into Lemma 1 given in the following. Its proof and outlines of the proofs of Propositions 1-3 are given in the supplement [5].

LEMMA 1. *Assume that Assumptions A1-5 hold.*

(i) There are positive constants C_1 and C_2 such that for any $\boldsymbol{g} \in \boldsymbol{G}$, $C_1\|\boldsymbol{g}\|_{G,2} \le \|\boldsymbol{Z}^T\boldsymbol{g}\|^V \le C_2\|\boldsymbol{g}\|_{G,2}$.

(ii) There are positive constants C_3 and C_4 such that for any $\boldsymbol{g} \in \boldsymbol{G}_B$, $\|\boldsymbol{g}\|^2_{G,\infty} \le C_3K_n\|\boldsymbol{g}\|^2_{G,2} \le C_4K_n(\|\boldsymbol{Z}^T\boldsymbol{g}\|^V)^2$.

(iii) There is a positive constant C_5 such that for any $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{g} \in \boldsymbol{G}_B$, $\|\boldsymbol{X}^T\boldsymbol{\beta}+\boldsymbol{Z}^T\boldsymbol{g}\|_\infty \le C_5K_n^{1/2}\|\boldsymbol{X}^T\boldsymbol{\beta}+\boldsymbol{Z}^T\boldsymbol{g}\|^V$, where $\|v\|_\infty = \max_{i,j}|v_{ij}|$. Besides, for some positive constant C_6, $\|v\|^V \le C_6\|v\|_\infty$.

(iv)

$$\sup_{\boldsymbol{g}_1,\boldsymbol{g}_2\in\boldsymbol{G}_B}\Big|\frac{\langle\boldsymbol{Z}^T\boldsymbol{g}_1,\boldsymbol{Z}^T\boldsymbol{g}_2\rangle_n^V - \langle\boldsymbol{Z}^T\boldsymbol{g}_1,\boldsymbol{Z}^T\boldsymbol{g}_2\rangle^V}{\|\boldsymbol{Z}^T\boldsymbol{g}_1\|^V\|\boldsymbol{Z}^T\boldsymbol{g}_2\|^V}\Big| = O_p(K_n\sqrt{\log n/n}).$$

(v) For any positive constant M, we have $\langle X_j - \boldsymbol{Z}^T\boldsymbol{g}_j, X_k - \boldsymbol{Z}^T\boldsymbol{g}_k\rangle_n^V - \langle X_j - \boldsymbol{Z}^T\boldsymbol{g}_j, X_k - \boldsymbol{Z}^T\boldsymbol{g}_k\rangle^V = o_p(1)$ uniformly in $\boldsymbol{g}_j \in \boldsymbol{G}_B$ and $\boldsymbol{g}_k \in \boldsymbol{G}_B$ satisfying $\|\boldsymbol{g}_j\|_{G,2} \le M$ and $\|\boldsymbol{g}_k\|_{G,2} \le M$.

(vi) For any process δ_n taking scalar values at T_{ij} such that $\|\delta_n\|_\infty$ is uniformly bounded in n and $\{\delta_{n,ij}\}_{j=1}^{m_i}$ are mutually independent in i,

$$\sup_{\boldsymbol{g}\in\boldsymbol{G}_B}\Big|\frac{\langle\delta_n,\boldsymbol{Z}^T\boldsymbol{g}\rangle_n^V - \langle\delta_n,\boldsymbol{Z}^T\boldsymbol{g}\rangle^V}{\|\boldsymbol{Z}^T\boldsymbol{g}\|^V}\Big| = O_p(\sqrt{K_n/n})\|\delta_n\|_\infty.$$

(vii) We also suppose Assumption S holds. Then for $k = 1,\dots,p$, $\|\widehat{\boldsymbol{\varphi}}_{\boldsymbol{V}k}\|_\infty = O_p(1)$, $\|\boldsymbol{Z}^T(\boldsymbol{\varphi}^*_{\boldsymbol{V}k} - \widehat{\boldsymbol{\varphi}}_{\boldsymbol{V}k})\|_n^V = o_p(1)$, and $\|\boldsymbol{Z}^T(\boldsymbol{\varphi}^*_{\boldsymbol{V}k} - \widehat{\boldsymbol{\varphi}}_{\boldsymbol{V}k})\|^V = o_p(1)$.

5.4. *Proof of Theorem 1.* Since we consider the identity link function, we have explicit expressions of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} - \boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}}} - \boldsymbol{\beta}_0$:

$$\begin{aligned}(5.4)\quad \widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} - \boldsymbol{\beta}_0 =& \boldsymbol{H}^{11}\sum_{i=1}^n(\underline{\boldsymbol{X}}_i - \underline{\boldsymbol{W}}_i\boldsymbol{H}_{22}^{-1}\boldsymbol{H}_{21})^T\boldsymbol{\Sigma}_i^{-1}\underline{\epsilon}_i \\ &- \boldsymbol{H}^{11}\sum_{i=1}^n(\underline{\boldsymbol{X}}_i - \underline{\boldsymbol{W}}_i\boldsymbol{H}_{22}^{-1}\boldsymbol{H}_{21})^T\boldsymbol{\Sigma}_i^{-1}(\underline{\boldsymbol{W}}_i\boldsymbol{\gamma}^* - \underline{(\boldsymbol{Z}^T\boldsymbol{g}_0)}_i) \\ =& I_1 - I_2 \quad \text{(say)},\end{aligned}$$

$$
\begin{aligned}
(5.5)\quad \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}}} - \boldsymbol{\beta}_0 =& \widehat{\boldsymbol{H}}^{11} \sum_{i=1}^{n} (\underline{\boldsymbol{X}}_i - \underline{\boldsymbol{W}}_i \widehat{\boldsymbol{H}}_{22}^{-1} \widehat{\boldsymbol{H}}_{21})^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \underline{\epsilon}_i \\
&- \widehat{\boldsymbol{H}}^{11} \sum_{i=1}^{n} (\underline{\boldsymbol{X}}_i - \underline{\boldsymbol{W}}_i \widehat{\boldsymbol{H}}_{22}^{-1} \widehat{\boldsymbol{H}}_{21})^T \widehat{\boldsymbol{\Sigma}}_i^{-1} (\underline{\boldsymbol{W}}_i \boldsymbol{\gamma}^* - \underline{(\boldsymbol{Z}^T \boldsymbol{g}_0)}_i) \\
=& \widehat{I}_1 - \widehat{I}_2 \quad \text{(say)},
\end{aligned}
$$

where $\widehat{\boldsymbol{H}}^{11}$, $\widehat{\boldsymbol{H}}_{22}$ and $\widehat{\boldsymbol{H}}_{21}$ are defined as in (2.4) with $\boldsymbol{V}_i = \widehat{\boldsymbol{\Sigma}}_i$, $i = 1, \ldots, n$, and $\boldsymbol{\gamma}^* = (\boldsymbol{\gamma}_1^{*T}, \ldots, \boldsymbol{\gamma}_q^{*T})^T$ satisfies $|\boldsymbol{B}^T(t)\boldsymbol{\gamma}_j^* - g_{0j}(t)| \le C_g K_n^{-2}$, $j = 1, \ldots, q$, for some positive constant C_g depending on $\boldsymbol{g}_0(t)$. Proposition 4 and Assumption A4 imply that with probability tending to 1, $C_1 \boldsymbol{I}_{m_i} \le \widehat{\boldsymbol{\Sigma}}_i \le C_2 m_i \boldsymbol{I}_{m_i}$ uniformly in i for some positive constants C_1 and C_2. As for $\widehat{\boldsymbol{\Sigma}}_i^{-1}$,

$$
\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_i^{-1} - \boldsymbol{\Sigma}_i^{-1} &= \widehat{\boldsymbol{\Sigma}}_i^{-1} (\boldsymbol{\Sigma}_i - \widehat{\boldsymbol{\Sigma}}_i) \boldsymbol{\Sigma}_i^{-1} \\
&= \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\Sigma}_i - \widehat{\boldsymbol{\Sigma}}_i) \boldsymbol{\Sigma}_i^{-1} + \widehat{\boldsymbol{\Sigma}}_i^{-1} (\boldsymbol{\Sigma}_i - \widehat{\boldsymbol{\Sigma}}_i) \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\Sigma}_i - \widehat{\boldsymbol{\Sigma}}_i) \boldsymbol{\Sigma}_i^{-1}.
\end{aligned}
$$

It follows from Proposition 4, Assumption A4, and the above identity that

$$
(5.6)\quad \widehat{\boldsymbol{\Sigma}}_i^{-1} - \boldsymbol{\Sigma}_i^{-1} = \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\Sigma}_i - \widehat{\boldsymbol{\Sigma}}_i) \boldsymbol{\Sigma}_i^{-1} + m_i^2 O_p\Big(h_2^4 + h_3^4 + \frac{\log n}{n h_2} + \frac{\log n}{n h_3^2} \Big).
$$

The last term in the right-hand side of (5.6) is in the sense of eigenvalue evaluation. By using Assumption A4 and Proposition 4, we get an expression of each element of $\boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\Sigma}_i - \widehat{\boldsymbol{\Sigma}}_i) \boldsymbol{\Sigma}_i^{-1}$. This expression, along with the assumptions for Theorem 1 and the local property of the B-spline basis, will be employed in the proofs of the following lemmas. These lemmas are needed in order to evaluate $\widehat{I}_1 - I_1$ and their proofs are given in the supplement [5].

LEMMA 2. *Assume the same conditions as in Theorem 1. Let $h_{12,kl}$ and $\widehat{h}_{12,kl}$ be the (k, l) element of $\boldsymbol{H}_{12}$ and $\widehat{\boldsymbol{H}}_{12}$, respectively. Then we have uniformly in k and l,*

$$
\begin{aligned}
&\frac{1}{n} h_{12,kl} = O_p(K_n^{-1}), \ \frac{1}{n}(h_{12,kl} - \widehat{h}_{12,kl}) = K_n^{-1} O_p\Big(h_2^2 + h_3^2 + \sqrt{\frac{\log n}{n h_2}} + \sqrt{\frac{\log n}{n h_3^2}} \Big), \\
&\Big\{ \sum_{l=1}^{qK_n} (n^{-1} h_{12,kl})^2 \Big\}^{1/2} = O_p(K_n^{-1/2}), \\
&\Big[\sum_{l=1}^{qK_n} \{ n^{-1} (h_{12,kl} - \widehat{h}_{12,kl}) \}^2 \Big]^{1/2} = K_n^{-1/2} O_p\Big(h_2^2 + h_3^2 + \sqrt{\frac{\log n}{n h_2}} + \sqrt{\frac{\log n}{n h_3^2}} \Big).
\end{aligned}
$$

LEMMA 3. *Assume the same conditions as in Theorem 1. Then, with probability tending to 1,* $C_1K_n^{-1} \le \lambda_{\min}(n^{-1}\boldsymbol{H}_{22}) \le \lambda_{\max}(n^{-1}\boldsymbol{H}_{22}) \le C_2K_n^{-1}$ *for some positive constants* C_1 *and* C_2*. We also have*

$$\begin{aligned}&\max\big\{|\lambda_{\min}(n^{-1}(\widehat{\boldsymbol{H}}_{22}-\boldsymbol{H}_{22}))|, |\lambda_{\max}(n^{-1}(\widehat{\boldsymbol{H}}_{22}-\boldsymbol{H}_{22}))|\big\}\\ &= K_n^{-1}O_p\Big(h_2^2+h_3^2+\sqrt{\log n/(nh_2)}+\sqrt{\log n/(nh_3^2)}\Big).\end{aligned}$$

Hence we have $\max\big\{|\lambda_{\min}(n^{-1}\widehat{\boldsymbol{H}}_{22})|, |\lambda_{\max}(n^{-1}\widehat{\boldsymbol{H}}_{22})|\big\} = O_p(K_n^{-1})$ *and* $\max\big\{|\lambda_{\min}((n^{-1}\widehat{\boldsymbol{H}}_{22})^{-1}-(n^{-1}\boldsymbol{H}_{22})^{-1})|, |\lambda_{\max}((n^{-1}\widehat{\boldsymbol{H}}_{22})^{-1}-(n^{-1}\boldsymbol{H}_{22})^{-1})|\big\}$ *is also bounded from above by* $K_nO_p\Big(h_2^2+h_3^2+\sqrt{\log n/(nh_2)}+\sqrt{\log n/(nh_3^2)}\Big)$.

LEMMA 4. *Under the same conditions as in Theorem 1, we have* $\frac{1}{n}\widehat{\boldsymbol{H}}_{11} = \frac{1}{n}\boldsymbol{H}_{11}+o_p(1)$ *and* $\frac{1}{n}\widehat{\boldsymbol{H}}_{12}\big(\frac{1}{n}\widehat{\boldsymbol{H}}_{22}\big)^{-1}\frac{1}{n}\widehat{\boldsymbol{H}}_{21} = \frac{1}{n}\boldsymbol{H}_{12}\big(\frac{1}{n}\boldsymbol{H}_{22}\big)^{-1}\frac{1}{n}\boldsymbol{H}_{21}+o_p(1)$, *where* $o_p(1)$ *means both componentwise and in the meaning of eigenvalue evaluation. Hence we have* $n\widehat{\boldsymbol{H}}^{11} = n\boldsymbol{H}^{11}+o_p(1)$.

LEMMA 5. *Assume the same conditions as in Theorem 1. Then we have for some positive constants* C_1 *and* C_2, $\frac{C_1}{K_n}\boldsymbol{I}_{qK_n} \le \mathrm{cov}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \underline{\boldsymbol{W}}_i^T\boldsymbol{\Sigma}_i^{-1}\underline{\epsilon}_i\Big) \le \frac{C_2}{K_n}\boldsymbol{I}_{qK_n}$. *In addition we have*

$$\begin{aligned}&\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \underline{\boldsymbol{W}}_i^T(\widehat{\boldsymbol{\Sigma}}_i^{-1}-\boldsymbol{\Sigma}_i^{-1})\underline{\epsilon}_i\Big|\\ &= \sqrt{\frac{n}{K_n}}O_p\Big(\frac{\log n}{nh_1}+\frac{\log n}{nh_2}+\frac{\log n}{nh_3^2}\Big)+\sqrt{\frac{n}{K_n}}O_p(h_1^3+h_2^3+h_3^3)\\ &\quad +O_p(h_2^2+h_3^2)+O_p\Big(\frac{1}{\sqrt{nh_2}}+\frac{1}{\sqrt{nh_3^2}}+\frac{1}{\sqrt{nK_nh_2}}+\frac{1}{\sqrt{nK_nh_3^2}}\Big).\end{aligned}$$

LEMMA 6. *Assume the same conditions as in Theorem 1. Then we have for some positive constants* C_1 *and* C_2, $C_1\boldsymbol{I}_p \le \mathrm{cov}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \underline{\boldsymbol{X}}_i^T\boldsymbol{\Sigma}_i^{-1}\underline{\epsilon}_i\Big) \le C_2\boldsymbol{I}_p$. *In addition we have*

$$\begin{aligned}&\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \underline{\boldsymbol{X}}_i^T(\widehat{\boldsymbol{\Sigma}}_i^{-1}-\boldsymbol{\Sigma}_i^{-1})\underline{\epsilon}_i\Big|\\ &= \sqrt{n}O_p\Big(\frac{\log n}{nh_1}+\frac{\log n}{nh_2}+\frac{\log n}{nh_3^2}\Big)+\sqrt{n}O_p(h_1^3+h_2^3+h_3^3)\\ &\quad +O_p(h_2^2+h_3^2)+O_p\Big(1/(\sqrt{n}h_2)+1/(\sqrt{n}h_3^2)\Big).\end{aligned}$$

Now we prove that $\widehat{I}_1 - I_1 = o_p(n^{-1/2})$. Write

$$I_1 = \boldsymbol{H}^{11}\sum_{i=1}^{n}\underline{\boldsymbol{X}}_i^T\boldsymbol{\Sigma}_i^{-1}\underline{\epsilon}_i - \boldsymbol{H}^{11}\boldsymbol{H}_{12}\boldsymbol{H}_{22}^{-1}\sum_{i=1}^{n}\underline{\boldsymbol{W}}_i^T\boldsymbol{\Sigma}_i^{-1}\underline{\epsilon}_i = \boldsymbol{H}^{11}(I_{11} - I_{12}) \text{ (say)}.$$

We define $\widehat{I}_{11}$ and $\widehat{I}_{12}$ similarly. From Proposition 1 and Lemma 4, we have only to prove

$$\frac{1}{\sqrt{n}}(\widehat{I}_{11} - I_{11}) = o_p(1) \quad \text{and} \quad \frac{1}{\sqrt{n}}(\widehat{I}_{12} - I_{12}) = o_p(1). \tag{5.7}$$

The former result in (5.7) can be handled in the same way as the latter and we consider only the latter. Write

$$\begin{aligned}\frac{1}{\sqrt{n}}(\widehat{I}_{12} - I_{12}) &= \frac{1}{n}\widehat{\boldsymbol{H}}_{12}\big(\frac{1}{n}\widehat{\boldsymbol{H}}_{22}\big)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{\boldsymbol{W}}_i^T(\widehat{\boldsymbol{\Sigma}}_i^{-1} - \boldsymbol{\Sigma}_i^{-1})\underline{\epsilon}_i \\ &\quad + \frac{1}{n}\widehat{\boldsymbol{H}}_{12}\big\{\big(\frac{1}{n}\widehat{\boldsymbol{H}}_{22}\big)^{-1} - \big(\frac{1}{n}\boldsymbol{H}_{22}\big)^{-1}\big\}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{\boldsymbol{W}}_i^T\boldsymbol{\Sigma}_i^{-1}\underline{\epsilon}_i \\ &\quad + \big(\frac{1}{n}\widehat{\boldsymbol{H}}_{12} - \frac{1}{n}\boldsymbol{H}_{12}\big)\big(\frac{1}{n}\boldsymbol{H}_{22}\big)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{\boldsymbol{W}}_i^T\boldsymbol{\Sigma}_i^{-1}\underline{\epsilon}_i \\ &= DI_{12}^{(1)} + DI_{12}^{(2)} + DI_{12}^{(3)} \quad \text{(say)}.\end{aligned}$$

Lemmas 2, 3, and 5 imply

$$\begin{aligned}DI_{12}^{(1)} =& \sqrt{n}O_p\Big(\frac{\log n}{nh_1} + \frac{\log n}{nh_2} + \frac{\log n}{nh_3^2}\Big) + \sqrt{n}O_p(h_1^3 + h_2^3 + h_3^3) \\ &+ \sqrt{K_n}O_p\Big(\frac{1}{\sqrt{nh_2}} + \frac{1}{\sqrt{nh_3^2}} + \frac{1}{\sqrt{nK_nh_2}} + \frac{1}{\sqrt{nK_nh_3^2}}\Big) \\ &+ \sqrt{K_n}O_p(h_2^2 + h_3^2) = o_p(1), \\ DI_{12}^{(j)} =& \sqrt{K_n}O_p\Big(h_2^2 + h_3^2 + \sqrt{\log n/(nh_2)} + \sqrt{\log n/(nh_3^2)}\Big) = o_p(1),\ j = 2, 3.\end{aligned}$$

Hence we have established

$$\widehat{I}_1 - I_1 = o_p(n^{-1/2}). \tag{5.8}$$

Next we deal with $\widehat{I}_2 - I_2$ and two more lemmas are necessary.

LEMMA 7. *Under the same conditions as in Theorem 1,*

$$\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{\boldsymbol{W}}_i^T\boldsymbol{\Sigma}_i^{-1}(\underline{\boldsymbol{W}}_i\boldsymbol{\gamma}^*-\underline{(\boldsymbol{Z}^T\boldsymbol{g}_0)}_i)\Big| = O_p(\sqrt{n}K_n^{-5/2}), \quad \textit{and}$$

$$\begin{aligned}&\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{\boldsymbol{W}}_i^T(\widehat{\boldsymbol{\Sigma}}_i^{-1}-\boldsymbol{\Sigma}_i^{-1})(\underline{\boldsymbol{W}}_i\boldsymbol{\gamma}^*-\underline{(\boldsymbol{Z}^T\boldsymbol{g}_0)}_i)\Big|\\ &\quad= \sqrt{n}K_n^{-5/2}O_p\Big(h_2^2+h_3^2+\sqrt{\log n/(nh_2)}+\sqrt{\log n/(nh_3^2)}\Big).\end{aligned}$$

LEMMA 8. *Under the same conditions as in Theorem 1,*

$$\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{\boldsymbol{X}}_i^T\boldsymbol{\Sigma}_i^{-1}(\underline{\boldsymbol{W}}_i\boldsymbol{\gamma}^*-\underline{(\boldsymbol{Z}^T\boldsymbol{g}_0)}_i)\Big| = O_p(\sqrt{n}K_n^{-2}) \quad \textit{and}$$

$$\begin{aligned}&\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{\boldsymbol{X}}_i^T(\widehat{\boldsymbol{\Sigma}}_i^{-1}-\boldsymbol{\Sigma}_i^{-1})(\underline{\boldsymbol{W}}_i\boldsymbol{\gamma}^*-\underline{(\boldsymbol{Z}^T\boldsymbol{g}_0)}_i)\Big|\\ &\quad= \sqrt{n}K_n^{-2}O_p\Big(h_2^2+h_3^2+\sqrt{\log n/(nh_2)}+\sqrt{\log n/(nh_3^2)}\Big).\end{aligned}$$

Now we can show that $\widehat{I}_2 - I_2 = o_p(n^{-1/2})$. Write

$$\begin{aligned}I_2 =&\boldsymbol{H}^{11}\sum_{i=1}^{n}\underline{\boldsymbol{X}}_i^T\boldsymbol{\Sigma}_i^{-1}(\underline{\boldsymbol{W}}_i\boldsymbol{\gamma}^*-\underline{(\boldsymbol{Z}^T\boldsymbol{g}_0)}_i)-\boldsymbol{H}^{11}\boldsymbol{H}_{12}\boldsymbol{H}_{22}^{-1}\sum_{i=1}^{n}\underline{\boldsymbol{W}}_i^T\boldsymbol{\Sigma}_i^{-1}(\underline{\boldsymbol{W}}_i\boldsymbol{\gamma}^*-\underline{(\boldsymbol{Z}^T\boldsymbol{g}_0)}_i)\\ =&\boldsymbol{H}^{11}(I_{21}-I_{22})\ (\text{say}).\end{aligned}$$

We define $\widehat{I}_{21}$ and $\widehat{I}_{22}$ similarly and write $\widehat{I}_2 = \widehat{\boldsymbol{H}}^{11}(\widehat{I}_{21}-\widehat{I}_{22})$. From Proposition 1 and Lemma 4, we have only to prove $\frac{1}{\sqrt{n}}(\widehat{I}_{21}-I_{21}) = o_p(1)$ and $\frac{1}{\sqrt{n}}(\widehat{I}_{22}-I_{22}) = o_p(1)$. The former result in the above can be handled in the same way as the latter and we consider only the latter. Write

$$\begin{aligned}\frac{1}{\sqrt{n}}(\widehat{I}_{22}-I_{22}) &= \frac{1}{n}\widehat{\boldsymbol{H}}_{12}\Big(\frac{1}{n}\widehat{\boldsymbol{H}}_{22}\Big)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{\boldsymbol{W}}_i^T(\widehat{\boldsymbol{\Sigma}}_i^{-1}-\boldsymbol{\Sigma}_i^{-1})(\underline{\boldsymbol{W}}_i\boldsymbol{\gamma}^*-\underline{(\boldsymbol{Z}^T\boldsymbol{g}_0)}_i)\\ &\quad+\frac{1}{n}\widehat{\boldsymbol{H}}_{12}\Big\{\Big(\frac{1}{n}\widehat{\boldsymbol{H}}_{22}\Big)^{-1}-\Big(\frac{1}{n}\boldsymbol{H}_{22}\Big)^{-1}\Big\}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{\boldsymbol{W}}_i^T\boldsymbol{\Sigma}_i^{-1}(\underline{\boldsymbol{W}}_i\boldsymbol{\gamma}^*-\underline{(\boldsymbol{Z}^T\boldsymbol{g}_0)}_i)\\ &\quad+\Big(\frac{1}{n}\widehat{\boldsymbol{H}}_{12}-\frac{1}{n}\boldsymbol{H}_{12}\Big)\Big(\frac{1}{n}\boldsymbol{H}_{22}\Big)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{\boldsymbol{W}}_i^T\boldsymbol{\Sigma}_i^{-1}(\underline{\boldsymbol{W}}_i\boldsymbol{\gamma}^*-\underline{(\boldsymbol{Z}^T\boldsymbol{g}_0)}_i).\\ &= DI_{22}^{(1)}+DI_{22}^{(2)}+DI_{22}^{(3)} \quad (\text{say})\end{aligned}$$

Lemmas 2, 3, and 7 imply, for $j = 1, 2, 3$,

$$DI_{22}^{(j)} = \sqrt{n}K_n^{-2}O_p\Big(h_2^2 + h_3^2 + \sqrt{\log n/(nh_2)} + \sqrt{\log n/(nh_3^2)}\Big) = o_p(1).$$

Hence we have established $\widehat{I}_2 - I_2 = o_p(n^{-1/2})$. The desired result follows from (5.4), (5.5), (5.8) and the above result.

Acknowledgements. The authors thank the associate editor and three referees for their thoughtful and constructive comments on a previous submission, which led to significant improvement of this paper.

SUPPLEMENTARY MATERIAL

Supplement A: Some technical material
(doi: xx.xxxx/xx-AOSxxxxSUPP). Estimation of the nonparametric component, additional simulation results, proofs of the propositions and lemmas, and theory for the case of uniformly bounded m_i and general link function.

References.

[1] Cheng, G. and Wang, X. (2011). Semiparametric additive transformation model under current status data. *Electronic J. Statist.* **5** 1735–1764.
[2] Cheng, G., Yu, Z. and Huang, J. Z. (2013). The cluster bootstrap consistency in generalized estimating equations. *J. Multivariate Anal.* **115** 33–47.
[3] Cheng, G., Zhou, L. and Huang, J. Z. (2014). Supplement to "Efficient semiparametric estimation in generalized partially linear additive models for longitudinal/clustered data". *doi: 10.3150/12-BEJ479SUPP.*
[4] Cheng, G., Zhou, L. and Huang, J. Z. (2014). Efficient semiparametric estimation in generalized partially linear additive models for longitudinal/clustered data. *Bernoulli* **20** 141-163.
[5] Cheng, M. Y., Honda, T. and Li, J. (2015). Supplement to "Efficient estimation in semivarying coefficient models for longitudinal/clustered data". *doi: xx.xxxx/xx-AOSxxxxSUPP.*
[6] Cheng, M. Y., Honda, T., Li, J. and Peng, H. (2014). Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *Ann. Stat.* **42** 1819-1849.
[7] Fan, J., Huang, T. and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Amer. Statist. Assoc.* **102** 632–641.
[8] Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99** 710 – 723.
[9] Fan, J., Ma, Y. and Dai, W. (2014). Nonparametric independence screening in sparse ultra- high dimensional varying coefficient models. *J. Amer. Statist. Assoc.* **109** 1270–1284.
[10] Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrixes for longitudinal data. *J. Amer. Statist. Assoc.* **103** 1520–1533.
[11] Henry, K., Erice, A., Tierney, C., Balfour, H. H. J., Fischl, M. A., Kmack, A., Liou, S. H., Kenton, A., Hirsch, M. S., Phair, J., Martinez, A.,

Kahn, J. O. and for the AIDS Clinical Trial Group 193A Study Team (1998). A randomized, controlled, double-blind study comparing the survival benefit of four different reverse transcriptase inhibitor therapies (three-drug, two-drug, and alternating drug) for the treatment of advanced AIDS. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **19** 339-349.

[12] Huang, J. Z., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14** 763–788.

[13] Huang, J. Z., Zhang, L. and Zhou, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scand. J. Statist.* **34** 451–477.

[14] Li, Y. (2011). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika* **98** 355–370.

[15] Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 69-88.

[16] Lin, X., Wang, N., Welsh, A. H. and Carroll, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. *Biometrika* **91** 177-193.

[17] Ma, S. (2012). Two-step spline estimating equations for generalized additive partially linear models with large cluster sizes. *Ann. Stat.* **40** 2943–2872.

[18] Qu, A. and Li, R. (2006). Quadratic Inference Functions for Varying-Coefficient Models with Longitudinal Data. *Biometrics* **62** 379–391.

[19] Schumaker, L. L. (2007). *Spline Functions: Basic Theory, 3rd ed.* Cambridge University Press, Cambridge.

[20] Shen, S. L., Cui, J. L., Mei, C. L. and Wang, C. W. (2014). Estimation and inference of semi-varying coefficient models with heteroscedastic errors. *J. Multivariate Anal.* **124** 70–93.

[21] Tian, R., Xue, L. and Liu, C. (2014). Penalized quadratic inference functions for semiparametric varying coefficient partially linear models with longitudinal data. *J. Multivariate Anal.* **132** 94–110.

[22] Wang, L. and Qu, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 177–190.

[23] Wang, N., Carroll, R. J. and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Amer. Statist. Assoc.* **100** 147–157.

[24] Wu, H. and Zhang, J. T. (2006). *Nonparametric Regression Methods for Longitudinal Data: mixed-effects modeling approaches.* Wiley, New York.

[25] Xia, Y., Zhang, W. and Tong, H. (2004). Efficient estimation for semivarying-coefficient models. *Biometrika* **91** 661–681.

[26] Yao, W. and Li, R. (2013). New local estimation procedure for a non-parametric regression function for longitudinal data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **75** 123–138.

[27] Zhang, W., Fan, J. and Sun, Y. (2009). A semiparametric model for cluster data. *Ann. Stat.* **37** 2377-2408.

[28] Zhou, J. and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *J. Amer. Statist. Assoc.* **107** 701–710.

M.-Y. Cheng
Department of Mathematics
National Taiwan University
Taipei 106, Taiwan
E-mail: cheng@math.ntu.edu.tw

T. Honda
Graduate School of Economics
Hitotsubashi University
Kunitachi, Tokyo 186-8601, Japan
E-mail: t.honda@r.hit-u.ac.jp

J. Li
Department of Statistics & Applied Probability
National University of Singapore
Singapore 117546
E-mail: stalj@nus.edu.sg