Bernoulli **20**(1), 2014, 141–163
DOI: 10.3150/12-BEJ479

Efficient semiparametric estimation in generalized partially linear additive models for longitudinal/clustered data

GUANG CHENG[1], LAN ZHOU[2,*] and JIANHUA Z. HUANG[2,**]

[1]*Purdue University, West Lafayette, IN 47907, USA. E-mail: chengg@purdue.edu*
[2]*Texas A&M University, College Station, TX 77843, USA.*
E-mail: [*]lzhou@stat.tamu.edu; [**]jianhua@stat.tamu.edu

We consider efficient estimation of the Euclidean parameters in a generalized partially linear additive models for longitudinal/clustered data when multiple covariates need to be modeled nonparametrically, and propose an estimation procedure based on a spline approximation of the nonparametric part of the model and the generalized estimating equations (GEE). Although the model in consideration is natural and useful in many practical applications, the literature on this model is very limited because of challenges in dealing with dependent data for nonparametric additive models. We show that the proposed estimators are consistent and asymptotically normal even if the covariance structure is misspecified. An explicit consistent estimate of the asymptotic variance is also provided. Moreover, we derive the semiparametric efficiency score and information bound under general moment conditions. By showing that our estimators achieve the semiparametric information bound, we effectively establish their efficiency in a stronger sense than what is typically considered for GEE. The derivation of our asymptotic results relies heavily on the empirical processes tools that we develop for the longitudinal/clustered data. Numerical results are used to illustrate the finite sample performance of the proposed estimators.

Keywords: GEE; link function; longitudinal data; partially linear additive models; polynomial splines

1. Introduction

The partially linear model has become a widely used semiparametric regression model because it provides a nice trade-off between model interpretability and flexibility. In a partially linear model, the mean of the outcome is assumed to depend on some covariates $\mathbf{X}$ parametrically and some other covariates $\mathbf{T}$ nonparametrically. Usually, the effects of $\mathbf{X}$ (e.g., treatment) are of major interest, while the effects of $\mathbf{T}$ (e.g., confounders) are nuisance parameters. Efficient estimation for partially linear models has been extensively studied and well understood for independent data; see, for example, Chen [3], Speckman [21], and Severini and Staniswalis [20]. The book of Härdle, Liang and Gao [8] provides a comprehensive review of the subject.

Efficient estimation of the Euclidean parameter (i.e., the parametric component) in the partially linear model for dependent data is by no means simple due to complication in data structure. Lin and Carroll [16,17] showed that, whether a natural application of the local polynomial kernel method can yield a semiparametric efficient estimator depends on whether the covariate modeled nonparametrically is a cluster-level covariate or not. Because the naive approach fails, Wang, Carroll and Lin [25] constructed a semiparametric efficient estimator by employing the iterative kernel method of Wang [24] that can effectively account for the within-cluster correlation. Alternatively, Zhang [27], Chen and Jin [4], and Huang, Zhang and Zhou [12] constructed semiparametric efficient estimators by extending the parametric generalized estimating equations (GEE) of Liang and Zeger [15]. He, Zhu and Fung [10] and He, Fung and Zhu [9] considered robust estimation, and Leng, Zhang and Pan [14] studied joint mean-covariance modeling for the partially linear model also by extending the GEE. In all these development, only one covariate is modeled nonparametrically.

In many practical situations, it is desirable to model multiple covariates nonparametrically. However, it is well known that multivariate nonparametric estimation is subject to the curse of dimensionality. A widely used approach for dimensionality reduction is to consider an additive model for the nonparametric part of the regression function in the partly linear model, which in turn results in the partially linear additive model. Although adapting this approach is a natural idea, there are major challenges for estimating the additive model for dependent data. Until only very recently, Carroll *et al.* [2] gave the first contribution on the partly linear additive model for longitudinal/clustered data, focusing on a simple setup of the problem, where there is the same number of observations per subject/cluster, and the identity link function is used.

The goal of the paper is to give a thorough treatment of the problem in the general setting that allows a monotonic link function and unequal number of observations among subjects/clusters. In this general setting, we derive the semiparametric efficient score and efficiency bound to obtain a benchmark for efficient estimation. In our derivation, we only assume the conditional moment restrictions instead of any distributional assumptions, for example, the multivariate Gaussian error assumption employed in Carroll *et al.* [2]. It turns out the definition of the efficient score involves solving a system of complex integral equations and there is no closed-form expression. This fact rules out the feasibility of constructing efficient estimators by plugging the estimated efficient influence function into their asymptotic linear expansions. We propose an estimation procedure that approximates the unknown functions by splines and uses the generalized estimating equations. To differentiate our procedure with the parametric GEE, we refer to it as the extended GEE. We show that the extended GEE estimators are semiparametric efficient if the covariance structure is correctly specified and they are still consistent and asymptotically normal even if the covariance structure is misspecified. In addition, by taking advantage of the spline approximation, we are able to give an explicit consistent estimate of the asymptotic variance without solving the system of integral equations that lead to the efficient scores. Having a closed-form expression for the asymptotic variance is an attractive feature of our method, in particular when there is no closed-form expression of the semiparametric efficiency bound. Another attractive feature of our method is the computational simplicity, there is no need to resort to the computationally more demanding backfitting type algorithm and numerical integration, as has been done in the previous work on the same model.

As a side remark, one highlight of our mathematical rigor is the careful derivation of the smoothness conditions on the least favorable directions from primitive conditions. This rather technical but important issue has not been well treated in the literature. To develop the asymptotic theory in this paper, we rely heavily on some new empirical process tools which we develop by extending existing results from the i.i.d. case to the longitudinal/clustered data.

The rest of the paper is organized as follows. Section 2 introduces the setup of the partially linear additive model and the formulation of the extended GEE estimator. Section 3 lists all regularity conditions, derives the semiparametric efficient score and the efficiency bound, and presents the asymptotic properties of the extended GEE estimators. Section 4 illustrates the finite sample performance of the GEE estimators using a simulation study and a real data. The proofs of some nonasymptotic results and the sketched proofs of the main asymptotic results are given in the Appendix. The supplementary file discusses the properties of the least favorable directions, presents the relevant empirical processes tools and the complete proofs of all asymptotic results.

Notation. For positive number sequences a_n and b_n, let $a_n \lesssim b_n$ mean that a_n/b_n is bounded, $a_n \asymp b_n$ mean that $a_n \lesssim b_n$ and $a_n \gtrsim b_n$, and $a_n \ll b_n$ mean that $\lim_n a_n/b_n = 0$. For two positive semidefinite matrices $\mathbf{A}$ and $\mathbf{B}$, let $\mathbf{A} \geq \mathbf{B}$ mean that $\mathbf{A} - \mathbf{B}$ is positive semidefinite. Define $x \vee y$ ($x \wedge y$) to be the maximum (minimum) value of x and y. For any matrix $\mathbf{V}$, denote $\lambda_V^{\max}$ ($\lambda_V^{\min}$) as the largest (smallest) eigenvalue of $\mathbf{V}$. Let $|\mathbf{V}|$ denote the Euclidean norm of the vector $\mathbf{V}$. Let $\|a\|_{L_2}$ denote the usual L_2 norm of a squared integrable function a, where the domain of integration and the dominating measure should be clear from the context.

2. The model setup

Suppose that the data consist of n clusters with the ith ($i = 1, \ldots, n$) cluster having m_i observations. In particular, for longitudinal data a cluster represents an individual subject. The data from different clusters are independent, but correlation may exist within a cluster. Let Y_{ij} and $(\mathbf{X}_{ij}, \mathbf{T}_{ij})$ be the response variable and covariates for the jth ($j = 1, \ldots, m_i$) observation in the ith cluster. Here $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijK})'$ is a $K \times 1$ vector and $\mathbf{T}_{ij} = (T_{ij1}, \ldots, T_{ijD})'$ is a $D \times 1$ vector. We consider the marginal model

$$\mu_{ij} = E(Y_{ij}|\mathbf{X}_{ij}, \mathbf{T}_{ij}), \tag{1}$$

and the marginal mean μ_{ij} depends on covariates $\mathbf{X}_{ij}$ and $\mathbf{T}_{ij}$ through a known monotonic and differentiable link function $\mu(\cdot)$:

$$\begin{aligned}\mu_{ij} &= \mu\bigl(\mathbf{X}'_{ij}\boldsymbol{\beta} + \theta_+(\mathbf{T}_{ij})\bigr)\\ &= \mu\bigl(\mathbf{X}'_{ij}\boldsymbol{\beta} + \theta_1(T_{ij1}) + \cdots + \theta_D(T_{ijD})\bigr),\end{aligned} \tag{2}$$

where $\boldsymbol{\beta}$ is a $K \times 1$ vector, and $\theta_+(\mathbf{t})$ is an additive function with D smooth additive component functions $\theta_d(t_d)$, $1 \leq d \leq D$. For identifiability, it is assumed that $\int_{\mathcal{T}_d} \theta_d(t_d)\,\mathrm{d}t_d = 0$, where $\mathcal{T}_d$ is the compact support of the covariate T_{ijd}. Applications of marginal models for longitudinal/clustered data are common in the literature (Diggle *et al.* [7]).

Denote

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{im_i} \end{pmatrix}, \qquad \boldsymbol{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \vdots \\ \mu_{im_i} \end{pmatrix}, \qquad \mathbf{X}_i = \begin{pmatrix} \mathbf{X}'_{i1} \\ \vdots \\ \mathbf{X}'_{im_i} \end{pmatrix}, \qquad \mathbf{T}_i = \begin{pmatrix} \mathbf{T}'_{i1} \\ \vdots \\ \mathbf{T}'_{im_i} \end{pmatrix},$$

$$\theta_+(\mathbf{T}_i) = \begin{pmatrix} \theta_+(\mathbf{T}_{i1}) \\ \vdots \\ \theta_+(\mathbf{T}_{im_i}) \end{pmatrix}, \qquad \mu\bigl(\mathbf{X}_i\boldsymbol{\beta} + \theta_+(\mathbf{T}_i)\bigr) = \begin{pmatrix} \mu\bigl(\mathbf{X}'_{i1}\boldsymbol{\beta} + \theta_+(\mathbf{T}_{i1})\bigr) \\ \vdots \\ \mu\bigl(\mathbf{X}'_{im_i}\boldsymbol{\beta} + \theta_+(\mathbf{T}_{im_i})\bigr) \end{pmatrix}.$$

Here and hereafter, we make the notational convention that application of a multivariate function to a matrix is understood as application to each row of the matrix, and similarly application of a univariate function to a vector is understood as application to each element of the vector. Using matrix notation, our model representation (1) and (2) can be written as

$$\boldsymbol{\mu}_i = E(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{T}_i) = \mu\big(\mathbf{X}_i\boldsymbol{\beta} + \theta_+(\mathbf{T}_i)\big). \tag{3}$$

Note that in our modeling framework no distributional assumptions are imposed on the data other than the moment conditions specified in (1) and (2). In particular, $\mathbf{X}_i$ and $\mathbf{T}_i$ are allowed to be dependent, as commonly seen for longitudinal/clustered data. Let $\boldsymbol{\Sigma}_i = \mathrm{var}(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{T}_i)$ be the true covariance matrix of $\mathbf{Y}_i$. Following the generalized estimating equations (GEE) approach of Liang and Zeger [15], we introduce a working covariance matrix $\mathbf{V}_i = \mathbf{V}_i(\mathbf{X}_i, \mathbf{T}_i)$ of $\mathbf{Y}_i$, which can depend on a nuisance finite-dimensional parameter vector τ distinct from $\boldsymbol{\beta}$. In the parametric setting, Liang and Zeger [15] showed that, consistency of the GEE estimator is guaranteed even when the covariance matrices are misspecified, and estimation efficiency will be achieved when the working covariance matrices coincide with the true covariance matrices, that is, when $\mathbf{V}_i(\tau^*) = \boldsymbol{\Sigma}_i$ for some τ^*. In this paper, we shall establish a similar result in a semiparametric context.

To estimate the functional parameters, we use basis approximations (e.g., Huang, Wu and Zhou [11]). We approximate each component function $\theta_d(t_d)$ of the additive function $\theta_+(\mathbf{t})$ in (2) by a basis expansion, that is,

$$\theta_d(t_d) \approx \sum_{q=1}^{Q_d} \gamma_{dq} B_{dq}(t_d) = \mathbf{B}_d'(t_d)\boldsymbol{\gamma}_d, \tag{4}$$

where $B_{dq}(\cdot), q = 1, \dots, Q_d$, is a system of basis functions, which is denoted as a vector $\mathbf{B}_d(\cdot) = (B_{d1}(\cdot), \dots, B_{dQ_d}(\cdot))'$, and $\boldsymbol{\gamma}_d = (\gamma_{d1}, \dots, \gamma_{dQ_d})'$ is a vector of coefficients. In principle, any basis system can be used, but B-splines are used in this paper for their good approximation properties. In fact, if $\theta_d(\cdot)$ is continuous, the spline approximation can be chosen to satisfy $\sup_t |\theta_d(t) - \mathbf{B}_d'(t)\boldsymbol{\gamma}_d| \to 0$ as $Q_d \to \infty$, and the rate of convergence can be characterized based on the smoothness of $\theta_d(\cdot)$; see de Boor [6].

It follows from (4) that

$$\theta_+(\mathbf{T}_{ij}) \approx \sum_{d=1}^{D}\sum_{q=1}^{Q_d} \gamma_{dq} B_{dq}(T_{ijd}) = \sum_{d=1}^{D} \mathbf{B}_d'(T_{ijd})\boldsymbol{\gamma}_d = \mathbf{Z}_{ij}'\boldsymbol{\gamma}, \tag{5}$$

where $\mathbf{Z}_{ij} = (\mathbf{B}_1'(T_{ij1}), \dots, \mathbf{B}_D'(T_{ijD}))'$, and $\boldsymbol{\gamma} = ({\boldsymbol{\gamma}_1}', \dots, {\boldsymbol{\gamma}_D}')'$. Denoting $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \dots, \mathbf{Z}_{im_i})'$, (3) and (5) together imply that

$$\boldsymbol{\mu}_i = E(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{T}_i) \approx \mu(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}). \tag{6}$$

Thus, the Euclidean parameters and functional parameters are estimated jointly by minimizing the following weighted least squares criterion

$$\sum_{i=1}^{n}\{\mathbf{Y}_i - \mu(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma})\}'\mathbf{V}_i^{-1}\{\mathbf{Y}_i - \mu(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma})\} \tag{7}$$

or, equivalently, by solving the estimating equations

$$\sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Delta}_i\mathbf{V}_i^{-1}\{\mathbf{Y}_i - \mu(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma})\} = 0 \tag{8}$$

and

$$\sum_{i=1}^{n}\mathbf{Z}_i'\boldsymbol{\Delta}_i\mathbf{V}_i^{-1}\{\mathbf{Y}_i - \mu(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma})\} = 0, \tag{9}$$

where $\boldsymbol{\Delta}_i$ is a diagonal matrix with the diagonal elements being the first derivative of $\mu(\cdot)$ evaluated at $\mathbf{X}_{ij}'\boldsymbol{\beta} + \mathbf{Z}_{ij}'\boldsymbol{\gamma}$, $j = 1, \ldots, m_i$. Denoting the minimizer of (7) as $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, then $\widehat{\boldsymbol{\beta}}$ estimates the parametric part of the model, and $\widehat{\theta}_1(\cdot) = \mathbf{B}_1'(\cdot)\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\theta}_D(\cdot) = \mathbf{B}_D'(\cdot)\widehat{\boldsymbol{\gamma}}_D$ estimate the nonparametric part of the model. We refer to these estimators the extended GEE estimators. In this paper, we shall show that, under regularity conditions, $\widehat{\boldsymbol{\beta}}$ is asymptotically normal and, if the correct covariance structure is specified, it is semiparametric efficient, and also show that $\widehat{\theta}_d(\cdot)$ is a consistent estimator of the true nonparametric function $\theta_d(\cdot)$, $d = 1, \ldots, D$.

When the link function $\mu(\cdot)$ is the identity function, the minimizer of the weighted least squares (7) or the solution to the estimating equations (8) and (9) has a closed-from expression:

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\gamma}}\end{pmatrix} = \left(\sum_{i=1}^{n}\mathbf{U}_i'\mathbf{V}_i^{-1}\mathbf{U}_i\right)^{-1}\sum_{i=1}^{n}\mathbf{U}_i'\mathbf{V}_i^{-1}\mathbf{Y}_i,$$

where $\mathbf{U}_i = (\mathbf{X}_i, \mathbf{Z}_i)$.

3. Theoretical studies of extended GEE estimators

3.1. Regularity conditions

We state the regularity conditions needed for the theoretical results in this paper. For the asymptotic analysis, we assume that the number of individuals/clusters goes to infinity while the number of observations per individual/cluster remains bounded.

C1. The random variables T_{ijd} are bounded, uniformly in $i = 1, \ldots, n$, $j = 1, \ldots, m_i$ and $d = 1, \ldots, D$. The joint distribution of any pair of T_{ijd} and $T_{ij'd'}$ has a density $f_{ijj'dd'}(t_{ijd}, t_{ij'd'})$ with respect to the Lebesgue measure. We assume that $f_{ijj'dd'}(\cdot,\cdot)$ is bounded away from 0 and infinity, uniformly in $i = 1, \ldots, n$, $j, j' = 1, \ldots, m_i$, and $d, d' = 1, \ldots, D$.

C2. The first covariate is constant 1, that is, $X_{ij1} \equiv 1$. The random variables X_{ijk} are bounded, uniformly in $i = 1, \dots, n$, $j = 1, \dots, m_i$ and $k = 2, \dots, K$. The eigenvalues of $E\{\mathbf{X}_{ij}\mathbf{X}'_{ij}|\mathbf{T}_{ij}\}$ are bounded away from 0, uniformly in $i = 1, \dots, n$, $j = 1, \dots, m_i$.
C3. The eigenvalues of true covariance matrices $\mathbf{\Sigma}_i$ are bounded away from 0 and infinity, uniformly in $i = 1, \dots, n$.
C4. The eigenvalues of the working covariance matrices $\mathbf{V}_i$ are bounded away from 0 and infinity, uniformly in $i = 1, \dots, n$.

Conditions similar to C1–C4 were used and discussed in Huang, Zhang and Zhou [12] when considering partially linear models with the identity link. Condition C1 is also used to ensure identifiability of the additive components, see Lemma 3.1 of Stone [22]. Condition C1 implies that the marginal density $f_{ijd}(\cdot)$ of T_{ijd} is bounded away from 0 on its support, uniformly in $i = 1, \dots, n$, $j = 1, \dots, m_i$, and $d = 1, \dots, D$. The condition on eigenvalues in C2 prevents the multicollinearity of the covariate vector $\mathbf{X}_{ij}$ and ensures the identifiability of $\boldsymbol{\beta}$. Since we assume that the cluster size (or the number of observations per subject) is bounded, we expect C3 is in general satisfied. Note that a zero eigenvalue of Σ_i indicates that there is a perfect linear relation among the residuals from subject i, which is unlikely to happen in reality.

Denote the true values of $\boldsymbol{\beta}$ and $\theta_+(t)$ by $\boldsymbol{\beta}_0$ and $\theta_{0,+}(t)$, respectively.

C5. (i) The link function μ is strictly monotone and has continuous second derivative; (ii) $\inf_s \mu'(s) = c_1 > 0$; (iii) μ' and μ'' are locally bounded around $\mathbf{x}^T\boldsymbol{\beta}_0 + \theta_{0,+}(\mathbf{t})$; (iv) $\mu(\pm v)$ increases slower than v^L as $v \to \infty$ for some $L > 0$.

Denote $e_{ij} = Y_{ij} - \mu_{ij}$ and $\mathbf{e}_i = (e_{i1}, \dots, e_{im_i})'$.

C6. The errors are uniformly sub-Gaussian, that is,

$$\max_{i=1,\dots,n} M_0^2 E\big\{\exp\big(|\mathbf{e}_i|^2/M_0^2\big) - 1|\mathbf{X}_i, \mathbf{T}_i\big\} \le \sigma_0^2 \qquad \forall n, \text{ a.s.} \tag{10}$$

for some fixed positive constants M_0 and σ_0.

Condition C5 on the link function is satisfied in all practical situations. The sub-Gaussian condition C6 relaxes the strict multivariate Gaussian error assumption, and is commonly used in the literature when applying the empirical process theory.

For $i = 1, \dots, n$, let $\mathbf{\Delta}_{i0}$ be a diagonal matrix with the jth diagonal element being the first derivative of $\mu(\cdot)$ evaluated at $\mathbf{X}'_{ij}\boldsymbol{\beta}_0 + \theta_{0,+}(\mathbf{T}_{ij})$, $j = 1, \dots, m_i$. Let $\mathbf{X}_{ik}$ denote the kth column of the matrix $\mathbf{X}_i$. For any additive function $\varphi_+(\mathbf{t}) = \varphi_1(t_1) + \dots + \varphi_D(t_D)$, $\mathbf{t} = (t_1, \dots, t_D)'$, denote $\varphi_+(\mathbf{T}_i) = (\varphi_+(\mathbf{T}_{i1}), \dots, \varphi_+(\mathbf{T}_{im_i}))'$. Let $\varphi^*_{k,+}(\cdot)$ be the additive function $\varphi_{k,+}(\cdot)$ that minimizes

$$\sum_{i=1}^n E\big[\big\{\mathbf{X}_{ik} - \varphi_{k,+}(\mathbf{T}_i)\big\}'\mathbf{\Delta}_{i0}\mathbf{V}_i^{-1}\mathbf{\Delta}_{i0}\big\{\mathbf{X}_{ik} - \varphi_{k,+}(\mathbf{T}_i)\big\}\big]. \tag{11}$$

Denote $\boldsymbol{\varphi}^*_+(\mathbf{T}_i) = (\varphi^*_{1,+}(\mathbf{T}_i), \dots, \varphi^*_{K,+}(\mathbf{T}_i))$ and define

$$\mathbf{I}_V \equiv \lim_n \frac{1}{n}\sum_{i=1}^n E\big[\big\{\mathbf{X}_i - \boldsymbol{\varphi}^*_+(\mathbf{T}_i)\big\}'\mathbf{\Delta}_{i0}\mathbf{V}_i^{-1}\mathbf{\Delta}_{i0}\big\{\mathbf{X}_i - \boldsymbol{\varphi}^*_+(\mathbf{T}_i)\big\}\big].$$

C7. The matrix $\mathbf{I}_V$ is positive definite.

Condition C7 is a positive information requirement that ensures the Euclidean parameter $\boldsymbol{\beta}$ can be root-n consistently estimated. When $\mathbf{V}_i$ is specified to be the true covariance matrix $\boldsymbol{\Sigma}_i$ for all i, $\varphi^*_{k,+}(\cdot)$ reduces to the least favorable direction $\psi^*_{k,+}(\cdot)$ in the definition of efficient score function and $\mathbf{I}_V$ reduces to the efficient information matrix $\mathbf{I}_{\text{eff}}$; see Section 3.2.

For $d = 1, \dots, D$, let $\mathbb{G}_d = \{\mathbf{B}'_d(t)\boldsymbol{\gamma}_d\}$ be a linear space of splines with degree r defined on the support $\mathcal{T}_d$ of T_{ijd}. Let $\mathbb{G}_+ = \mathbb{G}_1 + \cdots + \mathbb{G}_D$ be the additive spline space. We allow the dimension of $\mathbb{G}_d$, $1 \le d \le D$, and $\mathbb{G}_+$ to depend on n, but such dependence is suppressed in our notation to avoid clutter. For each spline space, we require that the knot sequence satisfies the quasi-uniform condition, that is, $\max_{j,j'}(u_{n,j+r+1} - u_{n,j})/(u_{n,j'+r+1} - u_{n,j'})$ is bounded uniformly in n for knots $\{u_{n,j}\}$. Let

$$\rho_n = \max\Big\{\inf_{g\in\mathbb{G}_+}\big\|g(\cdot) - \theta_{0,+}(\cdot)\big\|_\infty, \max_{1\le k\le K}\inf_{g\in\mathbb{G}_+}\big\|g(\cdot) - \varphi^*_{k,+}(\cdot)\big\|_\infty\Big\}$$

and $Q_n = \max\{Q_d = \dim(\mathbb{G}_d), 1 \le d \le D\}$.

C8. (i) $\lim_n Q_n^2 \log^4 n/n = 0$, (ii) $\lim_n n\rho_n^4 = 0$.

Condition C8(i) characterizes the growth rate of the dimension of the spline spaces relative to the sample size. Condition C8(ii) describes the requirement on the best rate of convergence that the functions $\theta_{0,+}(\cdot)$ and $\varphi^*_{k,+}(\cdot)$'s can be approximated by functions in the spline spaces. These requirements can be quantified by smoothness conditions on $\theta_{0,+}(\cdot)$ and $\varphi^*_{k,+}(\cdot)$'s, as follows. For $\alpha > 0$, write $\alpha = \alpha_0 + \alpha_1$, where α_0 is an integer and $0 < \alpha_1 \le 1$. We say a function is α-smooth, if its derivative of order α_0 satisfies a Hölder condition with exponent α_1. If all additive components of $\theta_{0,+}(\cdot)$ and $\varphi^*_{k,+}(\cdot)$'s are α-smooth, and the degree r of the splines satisfies $r \ge \alpha - 1$, then, by a standard result from approximation theory, $\rho_n \asymp Q_n^{-\alpha}$ for $\alpha > 1/2$ (Schumaker [19]). Condition C8 thus can be replaced by the following condition.

C8$'$. (i) $\lim_n Q_n^2 \log^4 n/n = 0$; (ii) additive components of $\theta_{0,+}(\cdot)$ and $\varphi^*_{k,+}(\cdot)$, $k = 1, 2, \dots, K$, are α-smooth for some $\alpha > 1/2$; (iii) $\lim_n Q_n^{4\alpha}/n = \infty$.

Since $\varphi^*_{k,+}$ is only implicitly defined, it is important to verify its smoothness requirement from primitive conditions. In the supplementary file (Cheng, Zhou and Huang [5]), that is, Section S.1, we shall show that $\varphi^*_{k,+}(\cdot)$ solves a system of integral equations and its smoothness is implied by smoothness requirements on the joint density of $\mathbf{X}_i$ and $\mathbf{T}_i$.

3.2. Semiparametric efficient score and efficiency bound

For estimating the Euclidean parameter in a semiparametric model, the efficiency bound provides a useful benchmark for the optimal asymptotic behaviors (e.g., Bickel *et al.* [1]). In this subsection, we give the semiparametric efficient score and efficient information matrix when the covariance structure is correctly specified. We do not make the normality assumption on the error distribution in the derivations.

The models studied in this paper have more than one nuisance function so that the efficient score function for $\boldsymbol{\beta}$, denoted as ℓ^*_β, is obtained by projecting onto a sum-space. In Lemma 1 below, we construct ℓ^*_β by the two-stage projection approach (Sasieni [18]). Recall that $\mathbf{e}_i = \mathbf{Y}_i - \mu(\mathbf{X}_i\boldsymbol{\beta} + \theta_+(\mathbf{T}_i))$, where $\theta_+(\mathbf{t}) = \theta_1(t_1) + \cdots + \theta_D(t_D)$. Write $f_i(\mathbf{x}_i, \mathbf{t}_i, \mathbf{y}_i - \mu(\mathbf{x}_i\boldsymbol{\beta} + \theta_+(\mathbf{t}_i)))$ as the joint density of $(\mathbf{X}_i, \mathbf{T}_i, \mathbf{Y}_i)$ for the ith cluster. We assume that $f_i(\cdot,\cdot,\cdot)$ is smooth, bounded and satisfies $\lim_{|e_{ij}|\to\infty} f_i(\cdot,\cdot,\mathbf{e}_i) = 0$ for all $j = 1, \ldots, m_i$.

Lemma 1. *The efficient score has the expression* $\ell^*_\beta = (\ell^*_{\beta,1}, \ldots, \ell^*_{\beta,K})'$ *with*

$$\ell^*_{\beta,k} = \sum_{i=1}^n \bigl(\mathbf{X}_{ik} - \psi^*_{k,+}(\mathbf{T}_i)\bigr)'\boldsymbol{\Delta}_{i0}\boldsymbol{\Sigma}_i^{-1}\bigl[\mathbf{Y}_i - \mu\bigl(\mathbf{X}_i\boldsymbol{\beta}_0 + \theta_{0,+}(\mathbf{T}_i)\bigr)\bigr], \tag{12}$$

where $\psi^*_{k,+}(\mathbf{t}) = \sum_{d=1}^D \psi^*_{kd}(t_d)$ *satisfies*

$$\sum_{i=1}^n E\bigl[\bigl(\mathbf{X}_{ik} - \psi^*_{k,+}(\mathbf{T})\bigr)'\boldsymbol{\Delta}_{i0}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Delta}_{i0}\psi_d(\mathbf{T}_{id})\bigr] = 0 \tag{13}$$

for any $\psi_d(t_d) \in L_2(\mathcal{T}_d)$, $d = 1, \ldots, D$.

The form of $\ell^*_{\beta,k}$ when $D = 1$ coincides with that derived in the partially linear models, for example, Lin and Carroll [16], under the strict multivariate Gaussian error assumption. In the supplementary material (Cheng, Zhou and Huang [5]), we shall see that $\psi^*_{kd}(t_d)$'s (or, more generally, $\varphi^*_{kd}(t_d)$'s) solve a Fredholm integral equation of the second kind (Kress [13]), and do not have a closed-form expression. In the same file, we also show that $\psi^*_{kd}(t_d)$'s (or, more generally, $\varphi^*_{kd}(t_d)$'s) are well defined and have nice properties such as boundedness and smoothness under reasonable assumptions on the joint density of $\mathbf{X}_i$ and $\mathbf{T}_i$. These properties are crucial for the feasibility to construct semiparametric efficient estimators but are not carefully studied in the literature.

The semiparametric efficient information matrix for $\boldsymbol{\beta}$ is

$$\begin{aligned}\mathbf{I}_{\text{eff}} &\equiv \lim_n \frac{1}{n} E\bigl(\ell^*_\beta \ell^{*\prime}_\beta\bigr) \\ &= \lim_n \frac{1}{n}\sum_{i=1}^n E\bigl[\bigl\{\mathbf{X}_i - \boldsymbol{\psi}^*_+(\mathbf{T}_i)\bigr\}'\boldsymbol{\Delta}_{i0}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Delta}_{i0}\bigl\{\mathbf{X}_i - \boldsymbol{\psi}^*_+(\mathbf{T}_i)\bigr\}\bigr],\end{aligned} \tag{14}$$

where $\boldsymbol{\psi}^*_+(\mathbf{T}_i) = (\psi^*_{1,+}(\mathbf{T}_i), \ldots, \psi^*_{K,+}(\mathbf{T}_i))$. The efficient information matrix $\mathbf{I}_{\text{eff}}$ here is the same as the quantity $\mathbf{I}_V$ in condition C7 when $\mathbf{V}_i = \boldsymbol{\Sigma}_i$. In the above result, different subjects/clusters need not have the same number of observations and thus $(\mathbf{X}_i, \mathbf{T}_i)$ may not be identically distributed. In the special case that $(\mathbf{X}_i, \mathbf{T}_i)$ are i.i.d., the efficient information can be simplified to

$$\mathbf{I}_{\text{eff}} = E\bigl[\bigl\{\mathbf{X}_i - \boldsymbol{\psi}^*_+(\mathbf{T}_i)\bigr\}'\boldsymbol{\Delta}_{i0}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Delta}_{i0}\bigl\{\mathbf{X}_i - \boldsymbol{\psi}^*_+(\mathbf{T}_i)\bigr\}\bigr],$$

where the kth component of $\boldsymbol{\psi}^*_+$ satisfies

$$E\bigl[\bigl\{\mathbf{X}_{ik} - \psi^*_{k,+}(\mathbf{T})\bigr\}'\boldsymbol{\Delta}_{i0}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Delta}_{i0}\psi_d(\mathbf{T}_{id})\bigr] = 0$$

for any $\psi_d(t_d) \in L_2(\mathcal{T}_d)$, $d = 1, \ldots, D$.

The function $\boldsymbol{\psi}_+^*(\mathbf{T}_i)$ involved in the efficient information matrix (14) actually corresponds to the least favorable direction (LFD) along $\theta_{0,+}(\mathbf{T}_i)$ in the least favorable submodel (LFS). To provide an intuitive interpretation, we assume for simplicity that $f_i(\mathbf{e}_i|\mathbf{x}_i, \mathbf{t}_i) \sim N(0, \boldsymbol{\Sigma}_i)$. Given the above distributional assumption, the parametric submodel (indexed by ε) passing through $(\boldsymbol{\beta}_0, \theta_{0,+})$ is constructed as

$$\varepsilon \mapsto -\frac{1}{2}\sum_{i=1}^{n}\big[\mathbf{y}_i - \boldsymbol{\mu}_i(\varepsilon)\big]'\boldsymbol{\Sigma}_i^{-1}\big[\mathbf{y}_i - \boldsymbol{\mu}_i(\varepsilon)\big], \tag{15}$$

where $\boldsymbol{\mu}_i(\varepsilon) = \mu\{\mathbf{x}_i(\boldsymbol{\beta}_0 + \varepsilon\mathbf{v}) + [\theta_{0,+}(\mathbf{t}_i) + \varepsilon h_+(\mathbf{t}_i)]\}$, for some vector $\mathbf{v} \in \mathbb{R}^K$ and perturbation direction $h_+(\cdot)$ around $\theta_{0,+}(\cdot)$. For any fixed $\mathbf{v}$, the information matrix for the parametric submodel (evaluated at $\varepsilon = 0$) is calculated as

$$\mathbf{I}_{\text{para}}(h_+) = \lim_n \frac{1}{n}\sum_{i=1}^{n} E\big[\big\{\mathbf{X}_i\mathbf{v} + h_+(\mathbf{T}_i)\big\}'\boldsymbol{\Delta}_{i0}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Delta}_{i0}\big\{\mathbf{X}_i\mathbf{v} + h_+(\mathbf{T}_i)\big\}\big].$$

The minimum $\mathbf{I}_{\text{para}}(h_+)$ over all possible perturbation directions is known as the semiparametric efficient information for $\mathbf{v}'\boldsymbol{\beta}$ (Bickel *et al.* [1]). The parametric submodel achieving the minimum is called the LFS and the associated direction is called LFD. By calculating the Fréchet derivative of the quadratic function $h_+ \mapsto \mathbf{I}_{\text{para}}(h_+)$ and considering (13), we can easily show that its minimum is achieved when $h_+ = -\boldsymbol{\psi}_+^*\mathbf{v}$. In view of the above discussion, the efficient information for $\boldsymbol{\beta}$ becomes the $\mathbf{I}_{\text{eff}}$ defined in (14).

Remark 1. Our derivation of the efficient score and efficient information matrix also applies when $\mathbf{T}$ is a cluster level covariate, that is, $T_{ijd} = T_{id}$ for $j = 1, \ldots, m_i$, $d = 1, \ldots, D$. Let $\widetilde{\mathbf{T}}_i = (T_{i1}, \ldots, T_{iD})'$. In this case, we only need to replace $\psi_{k,+}^*(\mathbf{T}_i)$ and $\psi_+^*(\mathbf{T}_i)$ by $\psi_{k,+}^*(\widetilde{\mathbf{T}}_i)\mathbf{1}$ and $\mathbf{1}(\psi_{1,+}^*(\widetilde{\mathbf{T}}_i), \ldots, \psi_{K,+}^*(\widetilde{\mathbf{T}}_i))$, where $\mathbf{1}$ is an m_i-vector of ones, and do similar changes for $\psi_{k,+}(\mathbf{T}_i)$ and $\psi_+(\mathbf{T}_i)$. It is interesting to note that, when $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{T}_i)$ are i.i.d., then $\psi_{k,+}^*(\cdot)$ has a closed form expression:

$$\psi_{k,+}^*(\mathbf{t}) = \frac{E(\mathbf{X}_{ik}'\boldsymbol{\Delta}_{i0}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Delta}_{i0}\mathbf{1}|\widetilde{\mathbf{T}}_i = \mathbf{t})}{E(\mathbf{1}'\boldsymbol{\Delta}_{i0}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Delta}_{i0}\mathbf{1}|\widetilde{\mathbf{T}}_i = \mathbf{t})}.$$

3.3. Asymptotic properties

In this subsection, we assume that the dimension of the Euclidean parameter, that is, K, is fixed. Define $f_0(\mathbf{x}, \mathbf{t}) = \mu(\mathbf{x}'\boldsymbol{\beta}_0 + \theta_{0,+}(\mathbf{t}))$. Define

$$\mathbb{F}_n = \big\{f(\mathbf{x}, \mathbf{t}) : f(\mathbf{x}, \mathbf{t}) = \mu\big(\mathbf{x}'\boldsymbol{\beta} + g(\mathbf{t})\big), \beta \in \mathbb{R}^K, g \in \mathbb{G}_+\big\}.$$

The extended GEE estimator can be written as

$$\arg\min_{f \in \mathbb{F}_n} \frac{1}{n}\sum_{i=1}^{n}\big\{\mathbf{Y}_i - f(\mathbf{X}_i, \mathbf{T}_i)\big\}'\mathbf{V}_i^{-1}\big\{\mathbf{Y}_i - f(\mathbf{X}_i, \mathbf{T}_i)\big\}.$$

The minimizer is $\widehat{f}_n(\mathbf{x},\mathbf{t}) = \mu(\mathbf{x}^T\widehat{\boldsymbol{\beta}}_V + \widehat{\theta}(\mathbf{t}))$ where $\widehat{\theta}(\mathbf{t}) = \mathbf{B}'(\mathbf{t})\widehat{\boldsymbol{\gamma}}$. The subscript of $\widehat{\boldsymbol{\beta}}_V$ denotes the dependence on the working covariance matrices.

According to condition C8 (or C8′), there is an additive spline function $\theta_n^*(\mathbf{t}) = \mathbf{B}'(\mathbf{t})\boldsymbol{\gamma}^* \in \mathbb{G}_+$ such that $\|\theta_n^* - \theta_{0,+}\|_\infty \lesssim \rho_n \to 0$. Then $f_n^*(\mathbf{x},\mathbf{t}) = \mu(\mathbf{x}'\boldsymbol{\beta}_0 + \theta_n^*(\mathbf{t}))$ is a spline-based approximation to the regression function. Define

$$\langle \xi_1, \xi_2\rangle_n = \frac{1}{n}\sum_i \xi_1'(\mathbf{X}_i,\mathbf{T}_i)\mathbf{V}_i^{-1}\xi_2(\mathbf{X}_i,\mathbf{T}_i)$$

and $\|\xi\|_n^2 = \langle \xi,\xi\rangle_n$.

Theorem 1 (Consistency). *The following results hold*:

$$\|\widehat{f}_n - f_n^*\|_n^2 = O_P\big(Q_n \log^2 n/n \vee \rho_n^2\big), \tag{16}$$

$$\|\widehat{f}_n - f_n^*\|_\infty = o_P(1), \tag{17}$$

$$\|\widehat{f}_n - f_0\|_\infty = o_P(1), \tag{18}$$

$$\widehat{\boldsymbol{\beta}}_V \xrightarrow{P} \boldsymbol{\beta}_0, \qquad \|\widehat{\theta} - \theta_{0,+}\|_\infty = o_P(1). \tag{19}$$

Theorem 1 says that the extended GEE estimators are consistent in estimating the parametric and nonparametric components of the model. Next we show that, our extended GEE estimator $\widehat{\boldsymbol{\beta}}$ is asymptotically normal even when the working covariance matrices $\mathbf{V}_i$'s are not necessarily the same as the true ones.

Denote $\mathbf{U}_i = (\mathbf{X}_i, \mathbf{Z}_i)$ as before. Let

$$\begin{aligned}\mathbf{H} &= \sum_{i=1}^n \mathbf{U}_i'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\boldsymbol{\Delta}_{i0}\mathbf{U}_i \equiv \begin{pmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{pmatrix}\\ &= \begin{pmatrix}\sum_{i=1}^n \mathbf{X}_i'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\boldsymbol{\Delta}_{i0}\mathbf{X}_i & \sum_{i=1}^n \mathbf{X}_i'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\boldsymbol{\Delta}_{i0}\mathbf{Z}_i\\ \sum_{i=1}^n \mathbf{Z}_i'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\boldsymbol{\Delta}_{i0}\mathbf{X}_i & \sum_{i=1}^n \mathbf{Z}_i'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\boldsymbol{\Delta}_{i0}\mathbf{Z}_i\end{pmatrix}.\end{aligned} \tag{20}$$

By the block matrix form of matrix inverse,

$$\begin{aligned}\begin{pmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{pmatrix}^{-1} &= \begin{pmatrix}\mathbf{H}^{11} & \mathbf{H}^{12}\\ \mathbf{H}^{21} & \mathbf{H}^{22}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{H}_{11\cdot 2}^{-1} & -\mathbf{H}_{11\cdot 2}^{-1}\mathbf{H}_{12}\mathbf{H}_{22}^{-1}\\ -\mathbf{H}_{22\cdot 1}^{-1}\mathbf{H}_{21}\mathbf{H}_{11}^{-1} & \mathbf{H}_{22\cdot 1}^{-1}\end{pmatrix},\end{aligned} \tag{21}$$

where $\mathbf{H}_{11\cdot 2} = \mathbf{H}_{11} - \mathbf{H}_{12}\mathbf{H}_{22}^{-1}\mathbf{H}_{21}$ and $\mathbf{H}_{22\cdot 1} = \mathbf{H}_{22} - \mathbf{H}_{21}\mathbf{H}_{11}^{-1}\mathbf{H}_{12}$. Define

$$\mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_V) \equiv \mathbf{H}^{11}\sum_{i=1}^{n}\big\{(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\boldsymbol{\Sigma}_i\mathbf{V}_i^{-1}\boldsymbol{\Delta}_{i0}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})\big\}\mathbf{H}^{11},$$

where the superscript $\triangle$ denotes the dependence on Δ_{i0}.

Theorem 2 (Asymptotic normality). *The extended GEE estimator $\widehat{\boldsymbol{\beta}}_V$ is asymptotically linear, that is*,

$$\widehat{\boldsymbol{\beta}}_V = \boldsymbol{\beta}_0 + \mathbf{H}^{11}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\mathbf{e}_i + o_P\left(\frac{1}{\sqrt{n}}\right). \tag{22}$$

Consequently,

$$\big\{\mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_V)\big\}^{-1/2}(\widehat{\boldsymbol{\beta}}_V - \boldsymbol{\beta}_0) \xrightarrow{d} \text{Normal}(0, \mathbf{Id}), \tag{23}$$

where $\mathbf{Id}$ *denotes the* $K \times K$ *identity matrix.*

When applying the asymptotic normality result for asymptotic inference, the variance $\mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_V)$ can be estimated by replacing $\boldsymbol{\Sigma}_i$ with $(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_V - \mathbf{Z}_i\widehat{\boldsymbol{\gamma}})(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_V - \mathbf{Z}_i\widehat{\boldsymbol{\gamma}})'$, and substituting parameter estimates in $\boldsymbol{\Delta}_{i0}$. The resulting estimator of variance is referred to as the Sandwich estimator.

Theorem 3. $\mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_V) \geq \mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_\Sigma)$.

Theorem 3 says that $\widehat{\boldsymbol{\beta}}_\Sigma$ is the most efficient in the class of extended GEE estimators with general working covariance matrices. Such a result is in parallel to that for standard parametric GEE estimators (Liang and Zeger [15]). This theorem is a consequence of the generalized Cauchy–Schwarz inequality and can be proved using exactly the same argument as Theorem 1 of Huang, Zhang and Zhou [12].

When the covariance matrices are correctly specified, the extended GEE estimators are efficient in a stronger sense than just described. Next, we show that the extended GEE estimator of $\boldsymbol{\beta}$ is the most efficient one among all regular estimators (see Bickel *et al.* [1] for the precise definition of regular estimators). In other words, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_\Sigma$ achieves the semiparametric efficiency bound, that is, the inverse of the efficient information matrix.

Corollary 1. *The estimator $\widehat{\boldsymbol{\beta}}_\Sigma$ is asymptotically normal and semiparametric efficient*, *that is*,

$$(n\mathbf{I}_{\text{eff}})^{1/2}(\widehat{\boldsymbol{\beta}}_\Sigma - \boldsymbol{\beta}_0) \xrightarrow{d} \text{Normal}(0, \mathbf{Id}). \tag{24}$$

In the below, we sketch the proof of Corollary 1 and postpone the details to the Appendix. Fixing $\mathbf{V}_i = \boldsymbol{\Sigma}_i$ in the definition of $\mathbf{H}$ as given in (20), we see that $\mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_\Sigma)$ can be written as

$$\mathbf{H}^{11}\sum_{i=1}^{n}\big\{(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})'\boldsymbol{\Delta}_{i0}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Delta}_{i0}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})\big\}\mathbf{H}^{11}.$$

Using the block matrix inversion formula (21) and examining the (1, 1)-block of the identity $\mathbf{H}^{-1} = \mathbf{H}^{-1}\mathbf{H}\mathbf{H}^{-1}$, we obtain that $\mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_{\Sigma}) = \mathbf{H}^{11}$. Denote $\widehat{\mathbf{I}}_n^{-1} = n\mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_{\Sigma})$. It is easily seen using (21) that

$$\widehat{\mathbf{I}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i' \boldsymbol{\Delta}_{i0} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Delta}_{i0} \mathbf{X}_i - \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i' \boldsymbol{\Delta}_{i0} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Delta}_{i0} \mathbf{Z}_i \left(\sum_{i=1}^{n} \mathbf{Z}_i' \boldsymbol{\Delta}_{i0} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Delta}_{i0} \mathbf{Z}_i\right)^{-1} \sum_{i=1}^{n} \mathbf{Z}_i' \boldsymbol{\Delta}_{i0} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Delta}_{i0} \mathbf{X}_i.$$

The asymptotic normality result in Theorem 2 can be rewritten as

$$(n\widehat{\mathbf{I}}_n)^{1/2}(\widehat{\boldsymbol{\beta}}_{\Sigma} - \boldsymbol{\beta}_0) \xrightarrow{d} \text{Normal}(0, \mathbf{Id}).$$

Thus, Corollary 1 follows from Theorem 2 and the result that $\widehat{\mathbf{I}}_n \to \mathbf{I}_{\text{eff}}$. The matrix $\widehat{\mathbf{I}}_n$ can be interpreted as a spline-based consistent estimate of the efficient information matrix.

Remark 2. When $\mathbf{T}$ is a cluster-level covariate, that is, $\mathbf{T}_{ij} = \widetilde{\mathbf{T}}_i = (T_{i1}, \ldots, T_{iD})'$, $j = 1, \ldots, m_i$, Theorems 1, 2 and Corollary 1 still hold. In that case, We can simplify C1 to the following condition.

C1′. The random variables T_{id} are bounded, uniformly in $i = 1, \ldots, n$, and $d = 1, \ldots, D$. The joint distribution of any pair of T_{id} and $T_{id'}$ has a density $f_{idd'}(t_{id}, t_{id'})$ with respect to the Lebesgue measure. We assume that $f_{idd'}(\cdot, \cdot)$ is bounded away from 0 and infinity, uniformly in $i = 1, \ldots, n$, and $d, d' = 1, \ldots, d$.

Remark 3. Our asymptotic result on estimation of the Euclidean parameter is quite insensitive to the choice of the number of terms Q_d in the basis expansion which plays the role of a smoothing parameter. Specifically, suppose that additive components of $\theta_{0,+}(\cdot)$ and $\varphi_{k,+}^{*}(\cdot)$, $k = 1, \ldots, K$, all have bounded second derivatives, that is, condition C8′ is satisfied with $\alpha = 2$. Then the requirement on Q_d reduces to $n^{1/8} \ll Q_d \ll n^{1/2}/\log^2 n$, a wide range for choosing Q_d. Thus, the precise determination of Q_d is not of particular concern when applying our asymptotic results. This insensitivity of smoothing parameter is also confirmed by our simulation study. In practice, it is advisable to use the usual data driven methods such as delete-cluster(subject)-out cross-validation to select Q_d and then check the sensitivity of the results (Huang, Zhang and Zhou [12]).

Remark 4. For simplicity, we assume in our asymptotic analysis that the working correlation parameter vector τ in $\mathbf{V}_i$ is known. It can be estimated via the method of moments using a quadratic function of Y_i's, just as in the application of the standard parametric GEEs (Liang and Zeger [15]). Similar to the parametric case, as long as such an estimate of τ converges in probability to some $\tau^{\dagger}$ at $\sqrt{n}$ rate, there is no asymptotic effect on $\widehat{\boldsymbol{\beta}}$ due to the estimation of τ; see Huang, Zhang and Zhou [12], Remark 1.

Remark 5. Our method does not require the assumption of normal error distribution. However, because it is essentially a least squares method, it is not robust to outliers. To achieve robustness to outlying observations, it is recommended to use the M-estimator type method as considered in He, Fung and Zhu [9].

4. Numerical results

4.1. Simulation

We conducted simulation studies to evaluate the finite sample performance of the proposed method. When the number of observations are the same per subject/cluster and the identity link function is used, our method performs comparably to the method of Carroll *et al.* [2] (see supplementary materials). In this section, we focus on simulation setups that cannot be handled by the existing method of Carroll *et al.* [2]. We generated data from the model

$$E(Y_{ij}|X_{ij}, Z_{ij1}, Z_{ij2}) = g\big\{\beta_0 + X_{ij}\beta_1 + f_1(Z_{ij1}) + f_2(Z_{ij2})\big\}, \qquad j = 1, \ldots, n_i, i = 1, \ldots, n,$$

where g is a link function which will be specified below, $\beta_0 = 0$, $\beta_1 = 0.5$, $f_1(t) = \sin\{2\pi(t - 0.5)\}$, and $f_2(t) = t - 0.5 + \sin\{2\pi(t - 0.5)\}$. The covariates Z_{ij1} and Z_{ij2} were generated from independent $\text{Normal}(0.5, 0.25)$ random variables but truncated to the unit interval [0, 1]. The covariate X_{ij} was generated as $X_{ij} = 3(1 - 2Z_{ij1})(1 - 2Z_{ij2}) + u_{ij}$ where u_{ij} were independently drawn from $\text{Normal}(0, 0.25)$. We obtained different simulation setups by varying the observational time distribution, the correlation structure, the parameters of the correlation function, the data distribution, and the number of subjects. We present results for five different setups, the details of which are given below.

For each simulation setup, 400 simulation runs were conducted and summary statistics of the results were calculated. For each simulated data set, the proposed generalized GEE estimator was calculated using a working independence (WI), an exchangeable (EX) correlation, or an autoregressive correlation structure. The correlation parameter ρ was estimated using the method of moments. Cubic splines were used with the number of knots chosen from the range 1–7 by the five-fold delete-subject-out cross-validation. The bias, variance, and the mean squared errors of Euclidean parameters were calculated for each scenario based on the 400 runs. The mean integrated squared errors (MISE), calculated using 100 grip points over [0, 1], for estimating $f_1(\cdot)$ and $f_2(\cdot)$, were also computed.

Setup 1. The longitudinal responses are from multivariate normal distribution with the autoregressive correlation structure and the identity link function. For each subject, six observational times are evenly placed between 0 and 1. The results are summarized in Table 1.

Setup 2. The same as setup 1, except that the log link is used. The results are summarized in Table 2.

Setup 3. This setup is the same as setup 1, except that the exchangeable correlation structure is used and the observational time distribution is different. For each subject, ten observational times are first evenly placed between 0 and 1. Then 40% of the observations are removed from each dataset and thus different subjects may have different number of observations and the observational times may be irregularly placed. The results are summarized in Table 3.

Table 1. Summary of simulation results for setup 1, based on 400 replications. The generalized GEE estimators using a working independence (WI), an exchangeable (EX) correlation structure, and an autoregressive (AR) structure are compared. The true correlation structure is the autoregressive with the lag-one correlation being ρ. Each entry of the table equals the original value multiplied by 10^5

		$\beta_0 = 0$			$\beta_1 = 0.5$			$f_1(\cdot)$	$f_2(\cdot)$
ρ	Method	Bias	SD	MSE	Bias	SD	MSE	MISE(f_1)	MISE(f_2)
					$n = 100$				
0.2	WI	−27	391	391	−176	60	60	1428	1330
	EX	14	381	381	−173	58	58	1359	1307
	AR	−27	373	373	−180	57	57	1311	1279
0.5	WI	−102	586	586	−169	59	60	1452	1322
	EX	44	524	524	−168	48	49	1245	1116
	AR	−42	474	474	−152	40	40	990	920
0.8	WI	−194	924	925	−100	60	60	1448	1358
	EX	−13	787	787	−133	29	29	747	662
	AR	−96	686	686	−93	16	16	463	461
					$n = 200$				
0.2	WI	−239	181	182	−35	26	26	689	709
	EX	−273	180	181	−47	25	25	669	698
	AR	−238	175	176	−32	26	26	656	664
0.5	WI	−261	270	271	4	27	27	676	712
	EX	−281	258	259	−38	23	23	569	604
	AR	−208	241	242	−18	19	19	482	493
0.8	WI	−183	448	449	62	30	30	677	723
	EX	−243	400	401	−20	13	13	338	361
	AR	−162	369	369	−7	8	8	224	245

Setup 4. It is the same as setup 3, except that the log link is used. The results are summarized in Table 4.

Setup 5. This setup is the same as setup 4, except that the Poisson distribution is used as the marginal distribution. All regression parameters in the general setup, the Euclidean and the functional, are halved for appropriate scaling of the response variable. The results are summarized in Table 5.

We have the following observations from the simulation results: for both Euclidean parameters, the estimator accounting for the correlation is more efficient (and sometimes significantly so) than the estimator using working independence correlation structure, even when the correlation structure is misspecified. Using the correct correlation structure usually produces the most efficient estimation. Efficiency gain gets bigger when the correlation parameter ρ gets larger. The variance is usually a dominating factor when comparing the MSEs between the two estimators. We have also observed that the sandwich estimated SEs work reasonably well; the averages of

Table 2. Summary of simulation results for setup 2, based on 400 replications. The generalized GEE estimators using a working independence (WI), an exchangeable (EX) correlation structure, and an autoregressive (AR) structure are compared. The true correlation structure is the autoregressive with the lag-one correlation being ρ. Each entry of the table equals the original value multiplied by 10^5

		$\beta_0 = 0$			$\beta_1 = 0.5$			$f_1(\cdot)$	$f_2(\cdot)$
ρ	Method	Bias	SD	MSE	Bias	SD	MSE	MISE(f_1)	MISE(f_2)
				$n = 100$					
0.2	WI	−2294	970	1022	44	30	30	1407	2280
	EX	−2223	962	1011	77	29	29	1397	2195
	AR	−2265	951	1003	64	29	29	1374	2104
0.5	WI	−2164	1137	1183	62	33	33	1356	2198
	EX	−1928	1007	1045	117	26	26	1146	1846
	AR	−1711	799	828	89	23	23	948	1394
0.8	WI	−2361	1727	1783	85	37	37	1378	2234
	EX	−2091	1017	1061	140	17	17	742	1303
	AR	−1911	722	758	116	12	12	518	824
				$n = 200$					
0.2	WI	−1387	388	407	88	16	16	601	1010
	EX	−1498	390	412	99	16	16	582	1024
	AR	−1497	384	407	98	15	15	565	985
0.5	WI	−1433	499	519	84	17	17	618	1068
	EX	−1532	435	458	108	14	14	534	830
	AR	−1525	387	410	97	12	12	436	660
0.8	WI	−1410	712	732	76	18	18	623	1095
	EX	−1192	332	346	81	9	9	302	482
	AR	−1433	277	298	88	5	5	219	323

the sandwich estimated SEs are close to the Monte Carlo sample standard deviations (numbers not shown to save space). For the functional parameters $f_1(\cdot)$ and $f_2(\cdot)$, the spline estimator accounting for the correlation is more efficient and the most efficient when the working correlation is the same as the true correlation structure. We also examined the Normal Q–Q plots of the Euclidean parameter estimates and observed that the distributions of the estimates are close to normal. These empirical results agree nicely with our theoretical results.

4.2. The longitudinal CD4 cell count data

To illustrate our method on a real data set, we considered the longitudinal CD4 cell count data among HIV seroconverters previously analyzed by Zeger and Diggle [26]. This data set contains 2376 observations of CD4+ cell counts on 369 men infected with the HIV virus. See Zeger and Diggle [26] for more detailed description of the data. We fit a partially linear additive model

Table 3. Summary of simulation results for setup 3, based on 400 replications. The generalized GEE estimators using a working independence (WI) and an exchangeable (EX) correlation structure are compared. The true correlation structure is the exchangeable with parameter ρ. Each entry of the table equals the original value multiplied by 10^5

		$\beta_0 = 0$			$\beta_1 = 0.5$			$f_1(\cdot)$	$f_2(\cdot)$
ρ	Method	Bias	SD	MSE	Bias	SD	MSE	MISE(f_1)	MISE(f_2)
				$n = 100$					
0	WI	−129	337	337	−125	61	61	1426	1412
	EX	−109	336	336	−116	61	61	1416	1410
0.2	WI	−96	527	527	−56	61	61	1445	1423
	EX	−161	511	511	−66	55	55	1297	1347
0.5	WI	−125	797	798	14	62	62	1515	1399
	EX	−216	735	735	−39	37	37	924	962
0.8	WI	−23	1054	1054	58	62	62	1552	1362
	EX	−164	914	914	−27	15	15	455	464
				$n = 200$					
0	WI	48	149	149	70	29	29	780	649
	EX	54	149	149	74	29	29	782	659
0.2	WI	−99	253	253	39	29	29	798	677
	EX	−21	237	237	37	25	25	693	609
0.5	WI	−192	403	404	−3	31	31	768	690
	EX	−64	354	354	15	16	16	470	432
0.8	WI	−240	564	565	−60	32	32	718	702
	EX	−96	466	466	6	7	7	236	227

using the log link with the CD4 counts as the response, covariates entering the model linearly including smoking status measured by packs of cigarettes, drug use (yes, 1; no 0), number of sex partners, and depression status measures by the CESD scale (large values indicating more depression symptoms), and the effects of age and time since seroconversion being modeled nonparametrically. We would like to remark that the partially linear additive model here provides a good balance of model interpretability and flexibility. Age and time are of continuous type and thus their effects are naturally modeled nonparametrically. Other variables are of discrete type and are not suitable for a nonparametric model.

Table 6 gives the estimates of the Euclidean parameters using both the WI and EX correlation structures. Cubic splines were used for fitting the additive functions and reported results correspond to the number of knots selected by the five-fold delete-subject-out cross-validation from the range of 0–10. The selected numbers of knots are 8 for time and 4 for age when using the WI structure and 8 for time and 3 for age when using the EX structure. The estimates of the

Table 4. Summary of simulation results for setup 4, based on 400 replications. The generalized GEE estimators using a working independence (WI) and an exchangeable (EX) correlation structure are compared. The true correlation structure is the exchangeable with parameter ρ. Each entry of the table equals the original value multiplied by 10^5

		$\beta_0 = 0$			$\beta_1 = 0.5$			$f_1(\cdot)$	$f_2(\cdot)$
ρ	Method	Bias	SD	MSE	Bias	SD	MSE	MISE(f_1)	MISE(f_2)
				$n = 100$					
0	WI	−2451	756	816	−218	27	28	1313	2318
	EX	−2500	773	835	−229	27	28	1329	2314
0.2	WI	−2581	1054	1120	−176	30	31	1461	2184
	EX	−2440	974	1034	−164	26	26	1240	2012
0.5	WI	−2455	1514	1574	−88	36	36	1574	2287
	EX	−1970	918	956	−102	19	19	851	1482
0.8	WI	−2520	2029	2093	3	41	41	1806	2342
	EX	−2547	1036	1101	24	10	10	629	1025
				$n = 200$					
0	WI	−883	329	336	114	14	14	653	826
	EX	−866	329	337	116	14	14	655	823
0.2	WI	−1090	475	487	84	14	14	728	844
	EX	−903	390	398	84	13	13	631	734
0.5	WI	−1310	718	736	44	16	16	779	932
	EX	−951	344	353	67	9	9	421	538
0.8	WI	−1533	966	989	13	18	18	744	1086
	EX	−1245	285	301	65	4	4	209	381

Euclidean parameters using the EX structure have smaller SE than those using the WI structure, suggesting that the EX structure produces more efficient estimates for this data set.

Appendix

A.1. Proof of Lemma 1 (derivation of the efficient score)

Let $\dot{\ell}_\beta$ denote the ordinary score for $\boldsymbol{\beta}$ when only $\boldsymbol{\beta}$ is unknown. Let $\mathcal{P}_f$ and $\mathcal{P}_\theta$ be the models with only $\{f_i, i = 1, \ldots, n\}$ and $\theta_+(\cdot)$ unknown, respectively, and let $\dot{\mathcal{P}}_f$ and $\dot{\mathcal{P}}_\theta$ be the corresponding tangent spaces. Following the discussions in Section 3.4 of Bickel *et al.* [1] (see also Appendix A6 of Huang, Zhang and Zhou [12]), we have

$$\ell^*_\beta = \Pi\big[\dot{\ell}_\beta | \dot{\mathcal{P}}_f^\perp\big] - \Pi\big[\Pi\big(\dot{\ell}_\beta | \dot{\mathcal{P}}_f^\perp\big) | \Pi\big[\dot{\mathcal{P}}_\theta | \dot{\mathcal{P}}_f^\perp\big]\big], \tag{25}$$

Table 5. Summary of simulation results for setup 5, based on 400 replications. The generalized GEE estimators using a working independence (WI) and an exchangeable (EX) correlation structure are compared. The true correlation structure is the exchangeable with parameter ρ. Each entry of the table equals the original value multiplied by 10^5

		$\beta_0 = 0$			$\beta_1 = 0.5$			$f_1(\cdot)$	$f_2(\cdot)$
ρ	Method	Bias	SD	MSE	Bias	SD	MSE	MISE(f_1)	MISE(f_2)
					$n = 100$				
0	WI	−2967	366	454	−379	113	114	1376	1446
	EX	−2983	368	457	−353	112	113	1368	1439
0.2	WI	−2557	738	803	−456	120	122	1394	1385
	EX	−2998	777	867	−367	98	99	1031	1110
0.5	WI	−1952	1101	1140	221	126	126	1446	1484
	EX	−2272	1339	1390	215	70	70	506	628
0.8	WI	−1979	1344	1383	369	126	127	1464	1567
	EX	−2349	1651	1706	506	71	74	411	545
					$n = 200$				
0	WI	−1563	190	214	−214	51	52	685	780
	EX	−1586	191	216	−208	51	52	691	781
0.2	WI	−1015	405	415	−195	54	55	637	771
	EX	−1355	402	421	−154	45	45	516	589
0.5	WI	−1301	599	616	143	55	55	742	777
	EX	−1751	634	665	218	30	30	256	300
0.8	WI	−1381	636	655	341	52	53	768	802
	EX	−1942	662	699	434	30	32	224	281

Table 6. Estimates of the Euclidean parameters in the CD4 cell counts study using the spline-based estimates. Working correlation structures used are working independence (WI) and exchangeable (EX). The standard errors (SE) are calculated using the sandwich formula

	WI		EX	
Parameter	Estimate	SE	Estimate	SE
Smoking	0.0786	0.0119	0.0619	0.0111
Drug	0.0485	0.0421	0.0134	0.0294
Sex partners	−0.0056	0.0043	0.0017	0.0035
Depression	−0.0025	0.0014	−0.0031	0.0013

where $\Pi[\cdot|\cdot]$ denote the projection operator, and $\dot{\mathcal{P}}^{\perp}$ denote the orthogonal complement of $\dot{\mathcal{P}}$. Lemma A.4 in Huang, Zhang and Zhou [12] directly implies that

$$\Pi\big[\dot{\ell}_{\beta}|\dot{\mathcal{P}}_f^{\perp}\big] = \sum_{i=1}^{n} \mathbf{X}_i' \boldsymbol{\Delta}_{i0} \boldsymbol{\Sigma}_i^{-1} \big[\mathbf{Y}_i - \mu\big(\mathbf{X}_i \boldsymbol{\beta}_0 + \theta_{0,+}(\mathbf{T}_i)\big)\big]. \tag{26}$$

Similarly, by constructing parametric submodels for each $\theta_k(\cdot)$ and slightly adapting the same Lemma, we have

$$\Pi\big[\dot{\mathcal{P}}_{\theta}|\dot{\mathcal{P}}_f^{\perp}\big] = \sum_{i=1}^{n} \left(\sum_{d=1}^{D} \psi_d(\mathbf{T}_{id})\right)' \boldsymbol{\Delta}_{i0} \boldsymbol{\Sigma}_i^{-1} \big[\mathbf{Y}_i - \mu\big(\mathbf{X}_i \boldsymbol{\beta}_0 + \theta_{0,+}(\mathbf{T}_i)\big)\big], \tag{27}$$

where $\psi_d(\mathbf{T}_{id}) = (\psi_d(T_{i1d}), \ldots, \psi_{im_id}(T_{im_id}))'$, for $\psi_d(\cdot) \in L_2(\mathcal{T}_d)$. Combination of (25)–(27) gives (12).

A.2. Proof sketch for Theorem 1 (consistency)

Let $\epsilon_n = (Q_n/n)^{1/2} \log n \vee \rho_n$. To show (16), it suffices to show that $P(\|\widehat{f}_n - f_n^*\|_n > \epsilon_n) \to 0$ as $n \to \infty$. Applying the peeling device (see the proof of Theorem 9.1 of van de Geer [23]), we can bound the above probability by the sum of $2C_0 \exp(-n\epsilon_n^2/(256C_0^2))$ and $P(\|y - f_n^*\|_n > \sigma)$ for some positive constant C_0. Considering condition C8 and choosing some proper σ related to ρ_n, we complete the proof of (16). As for (17), we have that

$$\|\widehat{f}_n - f_n^*\|_{\infty} \lesssim \big\|\mathbf{x}'\widehat{\boldsymbol{\beta}}_V + \widehat{\theta} - \big(\mathbf{x}'\boldsymbol{\beta}_0 + \theta_n^*\big)\big\|_{\infty} \lesssim Q_n^{1/2} \big\|\mathbf{x}'\widehat{\boldsymbol{\beta}}_V + \widehat{\theta} - \big(\mathbf{x}'\boldsymbol{\beta}_0 + \theta_n^*\big)\big\|$$

by Condition C5(iii) and Lemma S.2 in the supplementary note that

$$\big\|\mathbf{x}'\beta + g(\mathbf{t})\big\|_{\infty} \lesssim {Q_n}^{1/2} \big\|\mathbf{x}'\beta + g(\mathbf{t})\big\| \qquad \text{for } g \in \mathbb{G}_+. \tag{28}$$

It then follows by condition C5(ii) and (16) that

$$Q_n^{1/2} \big\|\mathbf{x}'\widehat{\boldsymbol{\beta}}_V + \widehat{\theta} - \big(\mathbf{x}'\boldsymbol{\beta}_0 + \theta_n^*\big)\big\| \lesssim Q_n^{1/2} \mathrm{O}_P\big\{(Q_n/n)^{1/2} \log n + \rho_n\big\} = \mathrm{o}_P(1)$$

since $(Q_n \log n)^2/n \to 0$ and $Q_n \rho_n^2 \to 0$ by condition C8 and the fact that $\rho_n \asymp Q_n^{-\alpha}$ for $\alpha > 1/2$. We thus obtain (17). Due to condition C5(iii), it follows that $\|f_n^* - f_0\|_{\infty} = \mathrm{O}(\|\theta_n^* - \theta_{0,+}\|_{\infty}) = \mathrm{O}(\rho_n)$ by Taylor's theorem. Combining this with (17), we obtain (18). From the proof of (17), we have that

$$\big\|\mathbf{x}'\widehat{\boldsymbol{\beta}}_V + \widehat{\theta} - \big(\mathbf{x}'\boldsymbol{\beta}_0 + \theta_n^*\big)\big\| = \mathrm{O}_P\big\{(Q_n/n)^{1/2} \log n + \rho_n\big\}.$$

Considering Lemma 3.1 of Stone [22], we obtain that $\|\mathbf{x}'(\widehat{\boldsymbol{\beta}}_V - \boldsymbol{\beta}_0)\|^2 = \mathrm{o}_P(1)$, which together with the no-multicollinearity condition C2 implies $\widehat{\boldsymbol{\beta}}_V \xrightarrow{P} \boldsymbol{\beta}_0$. By (28), we also obtain

$$\big\|\widehat{\theta} - \theta_n^*\big\|_{\infty} = Q_n^{1/2} \mathrm{O}_P\big\{(Q_n/n)^{1/2} \log n + \rho_n\big\} = \mathrm{o}_P(1).$$

Since $\|\theta_n^* - \theta_{0,+}\|_\infty = \mathrm{O}(\rho_n) = \mathrm{o}(1)$, application of the triangle inequality yields $\|\widehat{\theta} - \theta_{0,+}\|_\infty = \mathrm{o}_P(1)$, the last conclusion.

A.3. Proof sketch for Theorem 2 (asymptotic normality)

Note that $\widehat{\boldsymbol{\beta}}_V \in \mathbb{R}^K$ and $\widehat{\boldsymbol{\gamma}} \in \mathbb{R}^{Q_n}$ solve the estimating equations

$$\sum_{i=1}^n \mathbf{U}_i' \widehat{\boldsymbol{\Delta}}_i \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \mu(\mathbf{X}_i \widehat{\boldsymbol{\beta}}_V + \mathbf{Z}_i \widehat{\boldsymbol{\gamma}})\} = 0 \tag{29}$$

with $\mathbf{U}_i = (\mathbf{X}_i, \mathbf{Z}_i)$, and $\widehat{\boldsymbol{\Delta}}_i$ is a diagonal matrix with the diagonal elements being the first derivative of $\mu(\cdot)$ evaluated at $X_{ij}' \widehat{\boldsymbol{\beta}}_V + Z_{ij}' \widehat{\boldsymbol{\gamma}}$, $j = 1, \ldots, m_i$. Using the Taylor expansion, we have that

$$\mu(\mathbf{X}_i \widehat{\boldsymbol{\beta}}_V + \mathbf{Z}_i \widehat{\boldsymbol{\gamma}}) \approx \mu\big(\mathbf{X}_i \boldsymbol{\beta}_0 + \theta_0(\mathbf{T}_i)\big) + \boldsymbol{\Delta}_{i0}\{\mathbf{X}_i(\widehat{\boldsymbol{\beta}}_V - \boldsymbol{\beta}_0) + \mathbf{Z}_i \widehat{\boldsymbol{\gamma}} - \theta_0(\mathbf{T}_i)\}. \tag{30}$$

Recall that $\boldsymbol{\gamma}^*$ is assumed to satisfy $\rho_n = \|\theta_{0,+} - \mathbf{B}' \boldsymbol{\gamma}^*\|_\infty \to 0$. Substituting (30) into (29) yields

$$0 = \sum_{i=1}^n \mathbf{U}_i' (\widetilde{\mathbf{J}}_1 + \widetilde{\mathbf{J}}_2) - \sum_{i=1}^n \mathbf{U}_i' \boldsymbol{\Delta}_{i0} \mathbf{V}_i^{-1} \boldsymbol{\Delta}_{i0} \mathbf{U}_i \begin{pmatrix} \widehat{\boldsymbol{\beta}}_V - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^* \end{pmatrix}, \tag{31}$$

where

$$\widetilde{\mathbf{J}}_1 = (\widehat{\boldsymbol{\Delta}}_i - \boldsymbol{\Delta}_{i0}) \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \mu(\mathbf{X}_i \widehat{\boldsymbol{\beta}}_V + \mathbf{Z}_i \widehat{\boldsymbol{\gamma}})\}$$

and

$$\widetilde{\mathbf{J}}_2 = \boldsymbol{\Delta}_{i0} \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \mu\big(\mathbf{X}_i \boldsymbol{\beta}_0 + \theta_0(\mathbf{T}_i)\big) - \boldsymbol{\Delta}_{i0}\big(\mathbf{Z}_i \boldsymbol{\gamma}^* - \theta_0(\mathbf{T}_i)\big)\}.$$

Recalling (20) and using (21), we obtain from (31) that

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_V &= \boldsymbol{\beta}_0 + \mathbf{H}^{11} \sum_{i=1}^n \big(\mathbf{X}_i - \mathbf{Z}_i \mathbf{H}_{22}^{-1} \mathbf{H}_{21}\big)' (\widetilde{\mathbf{J}}_1 + \widetilde{\mathbf{J}}_2) \\
&= \boldsymbol{\beta}_0 + \mathbf{H}^{11} \sum_{i=1}^n \big(\mathbf{X}_i - \mathbf{Z}_i \mathbf{H}_{22}^{-1} \mathbf{H}_{21}\big)' \boldsymbol{\Delta}_{i0} \mathbf{V}_i^{-1} \mathbf{e}_i + \pi_n,
\end{aligned}$$

where the error term π_n has an explicit form and can be shown to be $\mathrm{o}_P(n^{-1/2})$ (the proof of this part relies heavily on the empirical process theory and is very lengthy). By the asymptotic linear expansion (22), we have

$$\begin{aligned}
&\{\mathbf{R}^\triangle(\widehat{\boldsymbol{\beta}}_V)\}^{-1/2} (\widehat{\boldsymbol{\beta}}_V - \boldsymbol{\beta}_0) \\
&\quad = \{\mathbf{R}^\triangle(\widehat{\boldsymbol{\beta}}_V)\}^{-1/2} \left(\mathbf{H}^{11} \sum_{i=1}^n \big(\mathbf{X}_i - \mathbf{Z}_i \mathbf{H}_{22}^{-1} \mathbf{H}_{21}\big)' \boldsymbol{\Delta}_{i0} \mathbf{V}_i^{-1} \mathbf{e}_i \right) + \mathrm{o}_P(1).
\end{aligned}$$

Then by applying the central limit theorem to the above equation and using the fact that

$$\mathrm{var}\left(\mathbf{H}^{11}\sum_{i=1}^{n}\big(\mathbf{X}_i-\mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21}\big)'\mathbf{\Delta}_{i0}\mathbf{V}_i^{-1}\mathbf{e}_i\,\Big|\{\mathbf{X}_i,\mathbf{T}_i\}_{i=1}^{n}\right)=\mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_V),$$

we complete the whole proof of (23).

A.4. Proof of Corollary 1

We only need to show that $\widehat{\mathbf{I}}_n \to \mathbf{I}_{\mathrm{eff}}$. Fix $\mathbf{V}_i = \mathbf{\Sigma}_i$ in the definitions of $\langle\xi_1,\xi_2\rangle_n^{\triangle}$ and $\langle\xi_1,\xi_2\rangle^{\triangle}$. Let $\widehat{\psi}_{k,n} = \arg\min_{\psi\in\mathbb{G}_+}\|x_k-\psi\|_n^{\triangle}$. From (21), we see that $\widehat{\mathbf{I}}_n = (\mathbf{H}_{11}-\mathbf{H}_{12}\mathbf{H}_{22}^{-1}\mathbf{H}_{21})/n$. Thus, the (k,k')th element of $\widehat{\mathbf{I}}_n$ is $\langle x_k-\widehat{\psi}_{k,n}, x_{k'}-\widehat{\psi}_{k',n}\rangle_n^{\triangle}$. On the other hand, by (13) and (14), the (k,k')th element of $\mathbf{I}_{\mathrm{eff}}$ is the limit of $\langle x_k-\psi_k^*, x_{k'}-\psi_{k'}^*\rangle_n^{\triangle}$, where $\psi_k^* = \psi_{k,+}^* = \arg\min_{L_{2,+}}\|x_k-\psi\|^{\triangle}$. Hence, it suffices to show that

$$\big\|\widehat{\psi}_{k,n}-\psi_k^*\big\|_n^{\triangle} = \mathrm{o}_P(1), \qquad k=1,2,\ldots,K, \tag{32}$$

because, if this is true, then by the triangle inequality,

$$\begin{aligned}\widehat{\mathbf{I}}_n\big(k,k'\big) &= \langle x_k-\widehat{\psi}_{k,n}, x_{k'}-\widehat{\psi}_{k',n}\rangle_n^{\triangle}\\ &= \big\langle x_k-\psi_k^*, x_{k'}-\psi_{k'}^*\big\rangle_n^{\triangle} + \mathrm{o}_P(1) = \mathbf{I}_{\mathrm{eff}}\big(k,k'\big)+\mathrm{o}_P(1).\end{aligned}$$

To show (32), we use $\psi_{k,n}^* = \Pi_n^{\triangle}x_k$ as a bridge. Notice that

$$\big\|\widehat{\psi}_{k,n}-\psi_k^*\big\|_n^{\triangle} \le \big\|\psi_{k,n}^*-\psi_k^*\big\|_n^{\triangle} + \big\|\widehat{\psi}_{k,n}-\psi_{k,n}^*\big\|_n^{\triangle}.$$

We inspect separately the sizes of the two terms on the right-hand side of the above inequality. First note that $\psi_{k,n}^* = \Pi_n^{\triangle}\psi_k^*$ since $\mathbb{G}_+\subset L_{2,+}$. Thus, $\|\psi_{k,n}^*-\psi_k^*\|^{\triangle} = \inf_{g\in\mathbb{G}_+}\|g-\psi_k^*\|^{\triangle} \asymp \inf_{g\in\mathbb{G}_+}\|g-\psi_k^*\|_{L_2} = \mathrm{O}(\rho_n) = \mathrm{o}(1)$, using Lemma S.2 in the supplementary note. Since $E(\{\|\psi_{k,n}^*-\psi_k^*\|_n^{\triangle}\}^2) = \{\|\psi_{k,n}^*-\psi_k^*\|^{\triangle}\}^2$, we have that $\|\psi_{k,n}^*-\psi_k^*\|_n^{\triangle} = \mathrm{o}_P(1)$. On the other hand, since $\psi_{k,n}^* = \Pi_n^{\triangle}x_k$ and $\widehat{\psi}_{k,n} = \widehat{\Pi}_n^{\triangle}x_k$, we have $\{\|\widehat{\psi}_{k,n}-\psi_{k,n}^*\|^{\triangle}\}^2 = \{\|x_k-\widehat{\psi}_{k,n}\|^{\triangle}\}^2 - \{\|x_k-\psi_{k,n}^*\|^{\triangle}\}^2$ and $\{\|x_k-\widehat{\psi}_{k,n}\|_n^{\triangle}\}^2 \le \{\|x_k-\psi_{k,n}^*\|_n^{\triangle}\}^2$. These two relations and Lemma S.3 in the supplementary note imply that $\|\widehat{\psi}_{k,n}-\psi_{k,n}^*\|^{\triangle} = \mathrm{o}_P(1)$, which in turn by the same lemma implies $\|\widehat{\psi}_{k,n}-\psi_{k,n}^*\|_n^{\triangle} = \mathrm{o}_P(1)$. As a consequence, $\|\widehat{\psi}_{k,n}-\psi_k^*\|_n^{\triangle} = \mathrm{o}_P(1)$, which is exactly (32). The proof is complete.

Acknowledgements

G. Cheng supported by NSF Grant DMS-09-06497 and NSF CAREER Award DMS-1151692. L. Zhou supported in part by NSF Grant DMS-09-07170. J. Z. Huang supported in part by NSF Grants DMS-06-06580, DMS-09-07170, NCI (CA57030), and Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

Supplementary Material

Supplement to "Efficient semiparametric estimation in generalized partially linear additive models for longitudinal/clustered data" (DOI: 10.3150/12-BEJ479SUPP; .pdf). The supplementary file (Cheng, Zhou and Huang [5]) includes the properties of the least favorable directions and the complete proofs of Theorems 1 and 2 together with some empirical processes results for the clustered/longitudinal data. The results of a simulation study that compares our method with that by Carroll *et al.* [2] are also included.

References

[1] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. *Johns Hopkins Series in the Mathematical Sciences*. Baltimore, MD: Johns Hopkins Univ. Press. MR1245941

[2] Carroll, R.J., Maity, A., Mammen, E. and Yu, K. (2009). Efficient semiparametric marginal estimation for partially linear additive model for longitudinal/clustered data. *Statistics in BioSciences* **1** 10–31.

[3] Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann*. *Statist*. **16** 136–146. MR0924861

[4] Chen, K. and Jin, Z. (2006). Partial linear regression models for clustered data. *J*. *Amer*. *Statist*. *Assoc*. **101** 195–204. MR2268038

[5] 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.

[6] de Boor, C. (2001). *A Practical Guide to Splines*, revised ed. *Applied Mathematical Sciences* **27**. New York: Springer. MR1900298

[7] Diggle, P.J., Heagerty, P.J., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford: Oxford Univ. Press. MR2049007

[8] Härdle, W., Liang, H. and Gao, J. (2000). *Partially Linear Models*. New York: Springer.

[9] He, X., Fung, W.K. and Zhu, Z. (2005). Robust estimation in generalized partial linear models for clustered data. *J*. *Amer*. *Statist*. *Assoc*. **100** 1176–1184. MR2236433

[10] He, X., Zhu, Z.Y. and Fung, W.K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89** 579–590. MR1929164

[11] Huang, J.Z., Wu, C.O. and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89** 111–128. MR1888349

[12] Huang, J.Z., Zhang, L. and Zhou, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scand*. *J*. *Stat*. **34** 451–477. MR2368793

[13] Kress, R. (1999). *Linear Integral Equations*, 2nd ed. *Applied Mathematical Sciences* **82**. New York: Springer. MR1723850

[14] Leng, C., Zhang, W. and Pan, J. (2010). Semiparametric mean-covariance regression analysis for longitudinal data. *J*. *Amer*. *Statist*. *Assoc*. **105** 181–193. With supplementary material available online. MR2656048

[15] Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

[16] Lin, X. and Carroll, R.J. (2001). Semiparametric regression for clustered data. *Biometrika* **88** 1179–1185. MR1872228

[17] Lin, X. and Carroll, R.J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J*. *Amer*. *Statist*. *Assoc*. **96** 1045–1056. MR1947252

[18] Sasieni, P. (1992). Nonorthogonal projections and their application to calculating the information in a partly linear Cox model. *Scand*. *J*. *Stat*. **19** 215–233. MR1183198

[19] Schumaker, L.L. (1981). *Spline Functions*: *Basic Theory*. New York: Wiley. MR0606200

[20] Severini, T.A. and Staniswalis, J.G. (1994). Quasi-likelihood estimation in semiparametric models. *J*. *Amer*. *Statist*. *Assoc*. **89** 501–511. MR1294076

[21] Speckman, P. (1988). Kernel smoothing in partial linear models. *J*. *R*. *Stat*. *Soc*. *Ser*. *B Stat*. *Methodol*. **50** 413–436. MR0970977

[22] Stone, C.J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann*. *Statist*. **22** 118–171.

[23] van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge: Cambridge Univ. Press.

[24] Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **90** 43–52. MR1966549

[25] Wang, N., Carroll, R.J. and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *J*. *Amer*. *Statist*. *Assoc*. **100** 147–157. MR2156825

[26] Zeger, S.L. and Diggle, P.J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50** 689–699.

[27] Zhang, L. (2004). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. Ph.D. thesis, Univ. Pennsylvania.

Received October 2011 and revised September 2012

Supplementary materials to

Efficient Semiparametric Estimation in Generalized Partially Linear Additive Models for Longitudinal/Clustered Data

Guang Cheng[*], **Lan Zhou**[†] **and Jianhua Z. Huang**[‡].
Purdue University and Texas A&M University

In this supplementary note, we discuss the properties of the least favorable directions, and provide the complete proofs of estimation consistency and asymptotic normality together with some empirical process results for clustered/longitudinal data. We also include some additional numerical results. The note is organized as follows. Section S.1 discusses the properties of the LFD under primitive conditions. Section S.2 defines the inner products and norms to be used in the note. Section S.3 presents the empirical processes tools we develop for the clustered/longitudinal data. Section S.4 contains the estimation consistency proof. Section S.5 collects the lemmas needed for the asymptotic normality proof which is given in Section S.6. Section S.7 reports the results of a simulation study that compares our method with that by Carroll *et al.* (2009).

S.1. Properties of the least favorable directions

In this subsection, we show that the least favorable directions (or, more generally, $\varphi^*_{kd}(t_d)$'s) are well defined and have nice properties such as boundedness and smoothness under reasonable assumptions on the joint density of $\mathbf{X}_i$ and $\mathbf{T}_i$. These properties are crucial for the feasibility to construct semiparametric efficient estimators but are not carefully studied in the literature. We present our results for the *generalized* efficient score functions, which are defined using the working covariance matrices. Recall that $\varphi^*_{k,+}(\mathbf{t}) = \varphi^*_{k1}(t_1) + \cdots + \varphi^*_{kD}(t_D)$ is the additive function $\varphi_{k,+}(\cdot)$ that minimizes

$$\sum_{i=1}^{n} E\big[\{\boldsymbol{X}_{ik} - \varphi_{k,+}(\mathbf{T}_i)\}'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\boldsymbol{\Delta}_{i0}\{\boldsymbol{X}_{ik} - \varphi_{k,+}(\mathbf{T}_i)\}\big].$$

[*]Corresponding author. Purdue University, West Lafayette, IN 47907, Email: chengg@purdue.edu. Research sponsored by NSF (DMS-0906497) and NSF CAREER Award (DMS-1151692).

[†]Texas A&M University, College Station, TX 77843, Email: lzhou@stat.tamu.edu. Research partly sponsored by NSF (DMS-0907170)

[‡]Texas A&M University, College Station, TX 77843, Email: jianhua@stat.tamu.edu. Research partly sponsored by NSF (DMS-0606580, DMS-0907170), NCI (CA57030), and Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

Since $\varphi^*_{k,+}(\cdot)$ can be viewed as an orthogonal projection, it always exists and is uniquely defined under the condition that $\mu(\cdot)$ is strictly monotone and $\mathbf{V}_i^{-1}$ is positive definite, see the alternating projection theorem (Bickel *et al.*, 1993, Theorem A.4.2).

In general, $\varphi^*_{kd}(\cdot)$'s do not have a closed-form expression. We now show that they are the solution of a Fredholm integral equation of the second kind (Kress, 1999). The stationary equations of the minimization problem are, for $k = 1, 2 \dots, K$,

$$E\left[\sum_{i=1}^{n}\sum_{j,j'=1}^{m_i}\{X_{ijk} - \sum_{d'=1}^{D}\varphi^*_{kd'}(T_{ijd'})\}\Delta_{ijj}v_i^{jj'}\Delta_{ij'j'}\varphi_d(T_{ij'd})\right] = 0, \tag{S.1}$$

for any $\varphi_d \in L_2(\mathcal{T}_d)$, $d = 1, \dots, D$, where Δ_{ijj} is the j-th diagonal element of the diagonal matrix $\mathbf{\Delta}_{i0}$, and $v_i^{jj'}$ is the Recall that f_{ijd} and $f_{ijj'dd'}$ are the density function for T_{ijd} and the joint density function for $(T_{ijd}, T_{ij'd'})$, respectively. By considering conditional expectations and writing out the conditional expectations as integrals, we see that $\varphi^*_{kd}(\cdot)$'s solve the following system of Fredholm integral equations of the second kind (Kress, 1999):

$$\begin{aligned}\varphi^*_{kd}(t) = q_d(t) - \int_{\mathcal{T}_d} H_{dd}(t,s)\varphi^*_{kd}(s)ds \\ - \sum_{d'\neq d}\int_{\mathcal{T}_{d'}}[G_{dd'}(t,s) + H_{dd'}(t,s)]\varphi^*_{kd'}(s)ds,\end{aligned} \tag{S.2}$$

where

$$q_d(t) = \frac{a_d(t)}{b_d(t)}, \quad H_{dd'}(t,s) = \frac{A_{dd'}(t,s)}{b_d(t)}, \quad G_{dd'}(t,s) = \frac{B_{dd'}(t,s)}{b_d(t)},$$

with

$$\begin{aligned}a_d(t) &= \sum_i\sum_j\sum_{j'} E(\Delta_{ijj'}v_i^{jj'}\Delta_{ij'j'}X_{ijk}|T_{ij'd} = t)f_{ij'd}(t),\\ b_d(t) &= \sum_i\sum_j E(\Delta^2_{ijj}v_i^{jj}|T_{ijd} = t)f_{ijd}(t),\\ A_{dd'}(t,s) &= \sum_i\sum_j\sum_{j'\neq j} E(\Delta_{ijj}v_i^{jj'}\Delta_{ij'j'}|T_{ijd} = t, T_{ij'd'} = s)f_{ijj'd'd}(t,s),\\ B_{dd'}(t,s) &= \sum_i\sum_j E(\Delta^2_{ijj}v_i^{jj}|T_{ijd} = t, T_{ijd'} = s)f_{ijjdd'}(t,s),\end{aligned}$$

for any fixed $d = 1, 2, \dots, D$ and $k = 1, 2, \dots, K$.

By condition C1, C2, C4, C5(iii), and the boundedness of $\theta_{0,+}$, it can be concluded that $A_{dd'}(t,s)/n$ and $a_d(t)/n$ are bounded, and $b_d(t)/n$ is bounded below by some positive constant that is independent of n. Consequently, $H_{dd'}(t,s)$ and $q_d(t)$ are both bounded. Using the fact that $\varphi^*_{kd} \in L_2(\mathcal{T}_d)$, and applying the Cauchy–Schwarz inequality for the integrals in (S.2), we obtain that φ^*_{kd} is bounded.

We next show that the smoothness of φ_{kd}^* follows from the smoothness assumptions on the joint density of $\mathbf{X}_i$ and $\mathbf{T}_i$. This result is useful to verify Condition C8$'$. To prove this result, we prepare with a lemma.

Let $\lceil\alpha\rceil$ denote the smallest integer bigger than or equal to α and $m_\alpha = \lceil\alpha\rceil - 1$. We say a function $a(\mathbf{s}, t)$ supported on compacta is *uniformly α-smooth in t relative to* $\mathbf{s}$ if the m-th derivative of $a(\mathbf{s}, t)$ w.r.t. t, denoted as $D_t^m a(\mathbf{s}, t)$, for $m = 0, \ldots, m_\alpha$, are all bounded, and

$$\sup_{\mathbf{s}} \sup_{t_1 \neq t_2} \frac{|D_t^{m_\alpha} a(\mathbf{s}, t_2) - D_t^{m_\alpha} a(\mathbf{s}, t_1)|}{|t_2 - t_1|^{\alpha'}} < \infty,$$

with $\alpha' = \alpha - m_\alpha$.

Lemma S.1. *(i) If $a(\mathbf{s}, t) = a(\mathbf{s}_1, \mathbf{s}_2, t)$ is uniformly α-smooth in t relative to $\mathbf{s}_1$ and $\mathbf{s}_2$, then $\int_{\mathcal{S}_1} a(\mathbf{s}_1, \mathbf{s}_2, t)\, d\mathbf{s}_1$ is uniformly α-smooth in t relative to $\mathbf{s}_2$.*

(ii) If $a(\mathbf{s}, t)$ and $b(\mathbf{s}, t)$ are both uniformly α-smooth in t relative to $\mathbf{s}$, then $c(\mathbf{s}, t) \equiv a(\mathbf{s}, t) b(\mathbf{s}, t)$ is uniformly α-smooth in t relative to $\mathbf{s}$.

(iii) If $a(\mathbf{s}, t)$ is uniformly α-smooth in t relative to $\mathbf{s}$ and $f(\cdot) \in C^{\lceil\alpha\rceil}$, then $f(a(\mathbf{s}, t))$ is uniformly α-smooth in t relative to $\mathbf{s}$.

Proof: (i) Note that $D_t^m a(\mathbf{s}_1, \mathbf{s}_2, t)$ is bounded for $0 \leq m \leq m_\alpha$, by the dominated convergence theorem, we can take derivative inside the integral to obtain

$$D_t^m \Big(\int_{\mathcal{S}_1} a(\mathbf{s}_1, \mathbf{s}_2, t)\, d\mathbf{s}_1 \Big) = \int_{\mathcal{S}_1} D_t^m a(\mathbf{s}_1, \mathbf{s}_2, t)\, d\mathbf{s}_1,$$

which implies that $D_t^m (\int_{\mathcal{S}_1} a(\mathbf{s}_1, \mathbf{s}_2, t)\, d\mathbf{s}_1)$ is bounded for $0 \leq m \leq m_\alpha$. Using this and the fact that

$$\begin{aligned} & \frac{\big| D_t^{m_\alpha} \big(\int_{\mathcal{S}_1} a(\mathbf{s}_1, \mathbf{s}_2, t_2)\, d\mathbf{s}_1 \big) - D_t^{m_\alpha} \big(\int_{\mathcal{S}_1} a(\mathbf{s}_1, \mathbf{s}_2, t_1)\, d\mathbf{s}_1 \big) \big|}{|t_2 - t_1|^{\alpha'}} \\ & \leq \int_{\mathcal{S}_1} \sup_{\mathbf{s}_1, \mathbf{s}_2} \sup_{t_1 \neq t_2} \frac{|D_t^{m_\alpha} a(\mathbf{s}_1, \mathbf{s}_2, t_2) - D_t^{m_\alpha} a(\mathbf{s}_1, \mathbf{s}_2, t_1)|}{|t_2 - t_1|^{\alpha'}}\, d\mathbf{s}_1 < \infty, \end{aligned}$$

for all $\mathbf{s}_2$ and $t_1 \neq t_2$, we conclude that $\int_{\mathcal{S}_1} a(\mathbf{s}_1, \mathbf{s}_2, t)\, d\mathbf{s}_1$ is indeed uniformly α-smooth in t relative to $\mathbf{s}_2$.

(ii) The result is true because

$$D_t^m c = \sum_{i+j=m} D_t^i a D_t^j b$$

is bounded for $0 \leq m \leq m_\alpha$. Also we note that for $i < m_\alpha$,

$$\frac{|D_t^i a(\mathbf{s}, t_2) - D_t^i a(\mathbf{s}, t_1)|}{|t_2 - t_1|^{\alpha'}} = \frac{|\int_{t_1}^{t_2} D_t^{i+1} a(\mathbf{s}, t)\, dt|}{|t_2 - t_1|^{\alpha'}}.$$

It can then be easily verified that

$$\sup_{\mathbf{s}} \sup_{t_1 \neq t_2} \frac{|D_t^{m_\alpha} c(\mathbf{s}, t_2) - D_t^{m_\alpha} c(\mathbf{s}, t_1)|}{|t_2 - t_1|^{\alpha'}} < \infty.$$

(iii) When $0 < \alpha \leq 1$, the result follows from the observation that

$$\frac{f(a(\mathbf{s}, t_2)) - f(a(\mathbf{s}, t_1))}{|t_2 - t_1|^\alpha} = \frac{f(a(\mathbf{s}, t_2)) - f(a(\mathbf{s}, t_1))}{|a(\mathbf{s}, t_2) - a(\mathbf{s}, t_1)|} \cdot \frac{|a(\mathbf{s}, t_2) - a(\mathbf{s}, t_1)|}{|t_2 - t_1|^\alpha}.$$

Using the chain rule, the above observation and part (ii) of the lemma, the desired result can be obtained by induction for general α. □

Note that φ_{kd}^*'s satisfy the system of integral equations (S.2). The quantities involved in the integral equations have certain properties under the regularity conditions listed in Section 3.1 Under Condition C5(ii), $\mu(\cdot)$ is strictly monotone. Condition C4 states that $\mathbf{V}_i^{-1}$ is positive definite. It follows that $b_d(t)$ is strictly positive. According to Lemma S.1, we have that $q_d(t)$, $\int_{\mathcal{T}_{d'}} H_{dd'}(t,s)\varphi_{kd'}^*(s)ds$ and $\int_{\mathcal{T}_{d'}} G_{rd'}(t,s)\varphi_{kd'}^*(s)ds$ are α-smooth in t if we assume that $f_{X,T}(\mathbf{x}_i, \mathbf{t}_i)$ and elements of $\mathbf{V}_i^{-1} = (v_i^{kl})$ are all uniformly α-smooth in t_{ijd} relative to the rest of the variates, and also that $\mu'(\cdot) \in C^{\lceil \alpha \rceil}$. Therefore, it follows from (S.2) that φ_{kd}^* is α-smooth.

S.2. Inner products and norms

Our theoretical development extends the geometric arguments used in Huang (1998, 2003), and Huang, Wu and Zhou (2004). This subsection gives the definition of some relevant inner products and norms in our context and presents some useful lemmas. We first introduce two inner products. For $\xi_1, \xi_2 \in L_2(d\mathbf{x} \times d\mathbf{t})$, define the empirical inner product as

$$\langle \xi_1, \xi_2 \rangle_n^\triangle = \frac{1}{n} \sum_i \xi_1'(\mathbf{X}_i, \mathbf{T}_i) \mathbf{\Delta}_{i0} \mathbf{V}_i^{-1} \mathbf{\Delta}_{i0} \xi_2(\mathbf{X}_i, \mathbf{T}_i)$$

and the theoretical inner product as $\langle \xi_1, \xi_2 \rangle^\triangle = E[\langle \xi_1, \xi_2 \rangle_n^\triangle]$. Corresponding norms are denoted as $\|\cdot\|_n^\triangle$ and $\|\cdot\|^\triangle$. Note that the superscript $\triangle$ reflects the dependence on the Δ_{i0}. When the link function is the identity function, two Δ_{i0}'s become the identity matrix and thus $\langle \cdot, \cdot \rangle_n^\triangle$ reduced to $\langle \cdot, \cdot \rangle_n$ as previously defined in Section 3.3 of the main paper.

LEMMA S.2. *(i) There are constants $M_1, M_2 > 0$ such that $M_1\|g\|_{L_2} \leq \|g\|^\triangle \leq M_2\|g\|_{L_2}$ for all $g \in \mathbb{G}_+$. (ii) There is a constant M_3 such that $\|g\|_\infty \leq M_3 Q_n^{1/2} \|g\|^\triangle$ for all $g \in \mathbb{G}_+$. (iii) There is a constant M_4 such that $\|\mathbf{x}'\beta + g(\mathbf{t})\|_\infty \leq M_4 {Q_n}^{1/2} \|\mathbf{x}'\beta + g(\mathbf{t})\|$ for $g \in \mathbb{G}_+$.*

Let $x_k(\cdot)$ denote the coordinate mapping that maps x to its k-th component so that $x_k(X_{ij}) = X_{ijk}$ for $k = 1, \ldots, p$.

LEMMA S.3. *Suppose* $\lim_n Q_n^2 \log n/n = 0$. *Then, for* $k = 1, \ldots, p$,

$$\sup_{g\in\mathbb{G}_+}\left|\frac{[\|x_k - g\|_n^{\triangle}]^2}{[\|x_k - g\|^{\triangle}]^2} - 1\right| = o_P(1), \tag{S.3}$$

$$\sup_{g\in\mathbb{G}_+}\left|\frac{[\|g\|_n^{\triangle}]^2}{[\|g\|^{\triangle}]^2} - 1\right| = o_P(1). \tag{S.4}$$

LEMMA S.4. *Let* $\{h_n\}$ *be a sequence of functions on* $\mathcal{T}$ *such that* $\|h_n\|_\infty \le M$ *for* $n \ge 1$. *Then,*

$$\sup_{g\in\mathbb{G}_+}\frac{|\langle h_n, g\rangle_n^{\triangle} - \langle h_n, g\rangle^{\triangle}|}{\|g\|^{\triangle}} = O_P\left(\sqrt{\frac{Q_n}{n}}\right)\|h_n\|^{\triangle}.$$

These lemmas are slight extensions of existing results. Lemma S.2 follows from Conditions C1 and C5 (ii)–(iii), Lemma 3.1 of Stone (1994), and a property of the spline functions as presented in Appendix A2 of Huang, Wu and Zhou (2004). Lemma S.3 can be proved as Lemma A2 of Huang, Wu and Zhou (2004). Lemma S.4 follows from the same argument as the proof of Lemma 11 in Huang (1998).

S.3. Empirical process results for clustered/longitudinal data

To facilitate our asymptotic analysis, we provide in this section some empirical process results for clustered/longitudinal data, which are extensions of existing results for i.i.d. data (e.g., Chapters 8 & 9 of van de Geer, 2000).

LEMMA S.5. *Let* $\boldsymbol{W} = (W_1, ..., W_m)'$ *be a zero mean random vector that satisfies the sub-Gaussian condition:*

$$M_0^2\{E\exp(|\boldsymbol{W}|^2/M_0^2) - 1\} \le \sigma_0^2, \tag{S.5}$$

Let $\mathbf{V}$ *be a symmetric positive definite matrix. Then, for all* $\boldsymbol{\beta} \in \mathbb{R}^m$,

$$E\exp(\boldsymbol{\beta}'\mathbf{V}^{-1}\boldsymbol{W}) \le \exp\{2(M_0^2 + \sigma_0^2)|\lambda_{\mathbf{V}^{-1}}^{\max}|^2|\boldsymbol{\beta}|^2\}. \tag{S.6}$$

Proof. The proof is a slight modification of the proof of Lemma 8.1 on Page 126 of van de Geer (2000). Take $c = 1 + \sigma_0^2/M_0^2$. By (S.5) and Chebyshev's inequality, we have that, for all $t > 0$

$$P(|\boldsymbol{W}| > t) \le c\exp(-t^2/M_0^2).$$

We then have that, for $k = 2, 3, \ldots$,

$$E|\boldsymbol{W}|^k = \int_0^\infty k\,t^{k-1}P(|\boldsymbol{W}| > t)\,dt \le c\int_0^\infty k\,t^{k-1}\exp(-t^2/M_0^2)\,dt.$$

Do the substitution $u = (t/M_0)^2$ in the above integral, then $t = M_0\sqrt{u}$, $dt = M_0 u^{-1/2}\,du/2$, and it follows that

$$E|\boldsymbol{W}|^k \le \frac{ckM_0^k}{2}\int_0^\infty u^{k/2-1}e^{-u}\,du = cM_0^k\Gamma\Big(\frac{k}{2}+1\Big).$$

Note that $E\boldsymbol{\beta}'\mathbf{V}^{-1}\boldsymbol{W} = 0$ and $\boldsymbol{\beta}'\mathbf{V}^{-1}\boldsymbol{W} \le \lambda^{\max}_{\mathbf{V}^{-1}}|\boldsymbol{\beta}||\boldsymbol{W}|$. Hence,

$$\begin{aligned}
E\exp\{\boldsymbol{\beta}'\mathbf{V}^{-1}\boldsymbol{W}\} &\le 1 + \sum_{k=2}^\infty \frac{(\lambda^{\max}_{\mathbf{V}^{-1}})^k|\boldsymbol{\beta}|^k E|\boldsymbol{W}|^k}{k!} \\
&\le 1 + c\sum_{k=2}^\infty \frac{(M_0\lambda^{\max}_{\mathbf{V}^{-1}}|\boldsymbol{\beta}|)^k}{k!}\Gamma\Big(\frac{k}{2}+1\Big).
\end{aligned}$$

The fact $\{\Gamma(k/2+1)\}^2 \le \Gamma(k+1) = k!$, $k = 2, 3, \ldots$, implies that the above summation is bounded by

$$\sum_{k=2}^\infty \frac{(M_0|\boldsymbol{\beta}|\lambda^{\max}_{\mathbf{V}^{-1}})^k}{\Gamma\big(\frac{k}{2}+1\big)} = \sum_{k=1}^\infty \frac{\big\{M_0^2(\lambda^{\max}_{\mathbf{V}^{-1}})^2|\boldsymbol{\beta}|^2\big\}^k}{\Gamma(k+1)} + \sum_{k=1}^\infty \frac{\big\{M_0^2(\lambda^{\max}_{\mathbf{V}^{-1}})^2|\boldsymbol{\beta}|^2\big\}^{k+\frac{1}{2}}}{\Gamma\big(k+\frac{3}{2}\big)}.$$

By induction and the fact that $\Gamma(\ell+1) = \ell\Gamma(\ell), \ell > 0$, we can verify that $\Gamma\Big(k+\dfrac{3}{2}\Big) \ge \Gamma(k+1), k = 1, 2, \ldots$. In addition, note that $c < c^k$ for $k = 2, 3\ldots$. Consequently,

$$\begin{aligned}
E\exp(\boldsymbol{\beta}'\mathbf{V}^{-1}\boldsymbol{W}) &\le 1 + \sum_{k=1}^\infty \frac{\big\{cM_0^2(\lambda^{\max}_{\mathbf{V}^{-1}})^2|\boldsymbol{\beta}|^2\big\}^k}{\Gamma(k+1)}\big[1 + \big\{cM_0^2(\lambda^{\max}_{\mathbf{V}^{-1}})^2|\boldsymbol{\beta}|^2\big\}^{1/2}\big] \\
&\le \exp\{cM_0^2(\lambda^{\max}_{\mathbf{V}^{-1}})^2|\boldsymbol{\beta}|^2\}\left(1 + \sum_{k=1}^\infty \frac{\big\{cM_0^2(\lambda^{\max}_{\mathbf{V}^{-1}})^2|\boldsymbol{\beta}|^2\big\}^k}{\Gamma(k+1)}\right) \\
&\le \exp\{2cM_0^2(\lambda^{\max}_{\mathbf{V}^{-1}})^2|\boldsymbol{\beta}|^2\} = \exp\big\{2(M_0^2+\sigma_0^2)|\lambda^{\max}_{\mathbf{V}^{-1}}|^2|\boldsymbol{\beta}|^2\big\},
\end{aligned}$$

where the second inequality follows from $A^{1/2} \le e^A$ for any $A > 0$. □

LEMMA S.6. *Let $\boldsymbol{W}_1 \in \mathbb{R}^{m_1}, \ldots, \boldsymbol{W}_n \in \mathbb{R}^{m_n}$ be independent zero mean random vectors satisfying the uniform sub-Gaussian condition:*

$$\max_{i=1,\ldots,n} M_0^2\big\{E\exp(|\boldsymbol{W}_i|^2/M_0^2) - 1\big\} \le \sigma_0^2. \tag{S.7}$$

Let $\mathbf{V}_i^{-1}$ be a symmetric $m_i \times m_i$ positive definite matrix, $i = 1, ..., n$. Then, for any $\boldsymbol{\gamma}_1 \in \mathbb{R}^{m_1}, \ldots, \boldsymbol{\gamma}_n \in \mathbb{R}^{m_n}$, and $a > 0$,

$$P\left(\Big|\sum_{i=1}^n \boldsymbol{\gamma}_i'\mathbf{V}_i^{-1}\boldsymbol{W}_i\Big| \ge a\right) \le 2\exp\Big\{-\frac{a^2}{8(M_0^2+\sigma_0^2)\sum_{i=1}^n(\lambda^{\max}_{\mathbf{V}_i^{-1}})^2|\boldsymbol{\gamma}_i|^2}\Big\} \tag{S.8}$$

Proof. The proof is almost exactly the same as Lemma 8.2 on Page 127 of van de Geer (2000). We still present it here for completeness. By Lemma S.5, for all $\beta \in \mathbb{R}$,

$$E\exp\Big(\beta\sum_{i=1}^{n}\boldsymbol{\gamma}_i'\mathbf{V}_i^{-1}\boldsymbol{W}_i\Big) = \prod_{i=1}^{n} E\exp(\beta\boldsymbol{\gamma}_i'\mathbf{V}_i^{-1}\boldsymbol{W}_i) \le \exp\Big\{2(M_0^2+\sigma_0^2)\beta^2\sum_{i=1}^{n}(\lambda_{\mathbf{V}_i^{-1}}^{\max})^2|\boldsymbol{\gamma}_i|^2\Big\}.$$

This together with the Chebyshev's inequality implies that, for $\beta > 0$,

$$P\Big(\sum_{i=1}^{n}\boldsymbol{\gamma}_i'\mathbf{V}_i^{-1}\boldsymbol{W}_i \ge a\Big) \le \exp\Big\{2(M_0^2+\sigma_0^2)\beta^2\sum_{i=1}^{n}(\lambda_{\mathbf{V}_i^{-1}}^{\max})^2|\boldsymbol{\gamma}_i|^2 - \beta a\Big\}.$$

Take

$$\beta = \frac{a}{4(M_0^2+\sigma_0^2)\sum_{i=1}^{n}(\lambda_{\mathbf{V}_i^{-1}}^{\max})^2|\boldsymbol{\gamma}_i|^2},$$

in the above inequality to obtain that

$$P\Big(\sum_{i=1}^{n}\boldsymbol{\gamma}_i'\mathbf{V}_i^{-1}\boldsymbol{W}_i \ge a\Big) \le \exp\Big\{-\frac{a^2}{8(M_0^2+\sigma_0^2)\sum_{i=1}^{n}(\lambda_{\mathbf{V}_i^{-1}}^{\max})^2|\boldsymbol{\gamma}_i|^2}\Big\}.$$

Because the same inequality holds for $-\sum_{i=1}^{n}\boldsymbol{\gamma}_i'\mathbf{V}_i^{-1}\boldsymbol{W}_i$, the desired result follows. □

Recall that $\mathbb{F}_n = \{f(\mathbf{x},\mathbf{t}) : f(\mathbf{x},\mathbf{t}) = \mu(\mathbf{x}'\boldsymbol{\beta}+g(\mathbf{t})),\ \beta \in \mathbb{R}^K, g \in \mathbb{G}_+\}$. Recall the definition of the norm $\|\cdot\|_n$ given above the statement of Theorem 1. For $R > 0$, denote $\mathbb{F}_n(R) = \{f \in \mathbb{F}_n : \|f - f_n^*\|_n \le R\}$. For $f \in \mathbb{F}_n(R)$, $f(\mathbf{X}_i,\mathbf{T}_i)$ denotes a vector whose elements are $f(\boldsymbol{X}_{ij},\boldsymbol{T}_{ij})$, $j = 1,\dots,m_i$. Let $N(r,\mathbb{F}_n(R),\|\cdot\|_n)$ be the minimal number of balls of radius r that is needed to cover $\mathbb{F}_n(R)$ under the norm $\|\cdot\|_n$. Correspondingly, $H(r,\mathbb{F}_n(R),\|\cdot\|_n) = \log N(r,\mathbb{F}_n(R),\|\cdot\|_n)$ is the entropy number.

Lemma S.7. *Suppose that $\boldsymbol{e}_i$ satisfies (10) in the main paper. For some constant C_0 depending only on M_0 and σ_0 specified in (10) of the main paper, and for $\delta > 0$ and $\sigma > 0$ satisfying $\delta/\sigma < R$, and*

$$n^{1/2}\delta \ge 2C_0\Big\{\int_{\delta/8\sigma}^{R} H^{1/2}(u,\mathbb{F}_n(R),\|\cdot\|_n)du \bigvee R\Big\}, \tag{S.9}$$

we have that

$$P\Big[\Big\{\sup_{f\in\mathbb{F}_n(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}\{f(\mathbf{X}_i,\mathbf{T}_i)\}'\mathbf{V}_i^{-1}\boldsymbol{e}_i\Big| \ge \delta\Big\} \cap \{\|\boldsymbol{e}\|_n \le \sigma\}\Big|\{\mathbf{X}_i,\mathbf{T}_i\}_{i=1}^{n}\Big] \le C_0\exp\Big(-\frac{n\delta^2}{4C_0^2R^2}\Big) \quad a.s., \tag{S.10}$$

where $\boldsymbol{e} = (\boldsymbol{e}_1',\dots,\boldsymbol{e}_n')'$ and $\|\boldsymbol{e}\|_n^2 = (1/n)\sum_{i=1}^{n}\boldsymbol{e}_i'\mathbf{V}_i^{-1}\boldsymbol{e}_i$.

Proof. This is a generalization of Corollary 8.3 of van de Geer (2000) to cluster/longitudinal data. The proof is similar as that of Lemma 3.2 on Page 29 of van de Geer (2000) and included here for completeness. Denote $C^2 = 8\, r_v(M_0^2+\sigma_0^2)$, where $r_v = \max_{1\le i\le n}\{(\lambda^{\max}_{\mathbf{V}_i^{-1}})^2/\lambda^{\min}_{\mathbf{V}_i^{-1}}\}$. According to condition (C4), r_v is bounded. By the fact that

$$\sum_{i=1}^n |\lambda^{\max}_{\mathbf{V}_i^{-1}}\boldsymbol{\gamma}_i|^2 \le r_v \sum_{i=1}^n \boldsymbol{\gamma}_i'\mathbf{V}_i^{-1}\boldsymbol{\gamma}_i := n r_v \|\boldsymbol{\gamma}\|_n^2$$

and Lemma S.6, we have

$$P\left(\frac{1}{n}\left|\sum_{i=1}^n \boldsymbol{\gamma}_i'\mathbf{V}_i^{-1}\boldsymbol{e}_i\right| \ge a\right) \le 2\exp\left\{-\frac{na^2}{C^2\|\boldsymbol{\gamma}\|_n^2}\right\}. \tag{S.11}$$

To apply the chaining technique, we index the functions in $\mathbb{F}_n(R)$ by a parameter $\phi\in\Phi$, i.e. $\mathbb{F}_n(R) = \{f_\phi\}_{\phi\in\Phi}$, without loss of generality. Let $N_s = N(2^{-s}R, \mathbb{F}_n(R), \|\cdot\|_n)$ and $\{f_j^s\}_{j=1}^{N_s}$ be the minimal $(2^{-s}R)$-covering set of $\mathbb{F}_n(R)$. Then, for each ϕ, there exists a $f_\phi^s \in \{f_1^s,\ldots,f_{N_s}^s\}$ such that $\|f_\phi - f_\phi^s\| \le 2^{-s}R$. Let

$$S = \min\{s\ge 1 : 2^{-s}R \le \delta/2\sigma\}. \tag{S.12}$$

Note that $S\ge 1$ since we assume $\delta/\sigma < R$. On the set $\{\|\boldsymbol{e}\|_n \le \sigma\}$, we apply the Cauchy–Schwarz inequality to obtain that

$$\left|\frac{1}{n}\sum_{i=1}^n \{f_\phi(\mathbf{X}_i,\mathbf{T}_i) - f_\phi^S(\mathbf{X}_i,\mathbf{T}_i)\}'\mathbf{V}_i^{-1}\boldsymbol{e}_i\right| \le \sigma\|f_\phi - f_\phi^S\|_n \le \frac{\delta}{2},$$

where the second inequality uses (S.12). Then it suffices to prove an analog of (S.10) for the $(2^{-S}R)$-covering net $\{f_j^S\}_{j=1}^{N_S}$,

$$P\left(\max_{j=1,\ldots,N_S}\left|\frac{1}{n}\sum_{i=1}^n \{f_j^S(\mathbf{X}_i,\mathbf{T}_i)\}'\mathbf{V}_i^{-1}\boldsymbol{e}_i\right| \ge \frac{\delta}{2}\middle|\{\mathbf{X}_i,\mathbf{T}_i\}_{i=1}^n\right).$$

Now apply the chaining technique. Let $f_\phi^0 = 0$. We have

$$f_\phi^S = \sum_{s=1}^S (f_\phi^s - f_\phi^{s-1}).$$

By the triangle inequality,

$$\begin{aligned}\|f_\phi^s - f_\phi^{s-1}\|_n &\le \|f_\phi^s - f_\phi\|_n + \|f_\phi - f_\phi^{s-1}\|_n\\ &\le 2^{-s}R + 2^{-s+1}R = (2^{-s})3R.\end{aligned} \tag{S.13}$$

Let η_s be positive numbers satisfying $\sum_{s=1}^{S}\eta_s \le 1$. Then

$$\begin{aligned}
&P\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n}\boldsymbol{e}_i'\mathbf{V}_i^{-1}\{f_\phi^s(\mathbf{X}_i,\mathbf{T}_i)-f_\phi^{s-1}(\mathbf{X}_i,\mathbf{T}_i)\}\right|>\frac{\delta}{2}\middle|\{\mathbf{X}_i,\mathbf{T}_i\}_{i=1}^n\right)\\
&\le \sum_{s=1}^{S}P\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{e}_i'\mathbf{V}_i^{-1}\{f_\phi^s(\mathbf{X}_i,\mathbf{T}_i)-f_\phi^{s-1}(\mathbf{X}_i,\mathbf{T}_i)\}\right|>\frac{\delta\eta_s}{2}\middle|\{\mathbf{X}_i,\mathbf{T}_i\}_{i=1}^n\right) \qquad \text{(S.14)}\\
&\le \sum_{s=1}^{S}2\exp\left\{2H(2^{-s}R,\mathbb{F}_n(R),\|\cdot\|_n)-\frac{n\delta^2\eta_s^2}{36C^2 2^{-2s}R^2}\right\},
\end{aligned}$$

with the last inequality being obtained by the fact that

$$\begin{aligned}
&\left\{\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{e}_i'\mathbf{V}_i^{-1}\{f_\phi^s(\mathbf{X}_i,\mathbf{T}_i)-f_\phi^{s-1}(\mathbf{X}_i,\mathbf{T}_i)\}\right|>\frac{\delta\eta_s}{2}\right\}\\
&\subset \bigcup_{1\le j\le N_s,\,1\le l\le N_{s-1}}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{e}_i'\mathbf{V}_i^{-1}\{f_j^s(\mathbf{X}_i,\mathbf{T}_i)-f_l^{s-1}(\mathbf{X}_i,\mathbf{T}_i)\}\right|>\frac{\delta\eta_s}{2}\right\}
\end{aligned}$$

and (S.11) and (S.13).

Set

$$\eta_s=\frac{12C2^{-s}RH^{1/2}(2^{-s}R,\mathbb{F}_n(R),\|\cdot\|_n)}{n^{1/2}\delta}\bigvee\frac{2^{-s}s^{1/2}}{8}.$$

We next verify that $\sum_{s=1}^{S}\eta_s\le 1$. By taking C_0 sufficiently large in (S.9), we have that

$$\frac{n^{1/2}\delta}{2}\ge 12C\sum_{s=1}^{S}2^{-s}RH^{1/2}(2^{-s}R,\mathbb{F}_n(R),\|\cdot\|_n) \qquad \text{(S.15)}$$

Moreover,

$$\begin{aligned}
\sum_{s=1}^{S}2^{-s}s^{1/2}&\le 1+\int_1^\infty 2^{-x}x^{1/2}\,dx\\
&\le 1+\int_0^\infty 2^{-x}x^{1/2}\,dx = 1+\left(\frac{\pi}{\log 2}\right)^{1/2}\le 4.
\end{aligned} \qquad \text{(S.16)}$$

Hence, by (S.15) and (S.16),

$$\sum_{s=1}^{S}\eta_s\le\sum_{s=1}^{S}\frac{12C2^{-s}RH^{1/2}(2^{-s}R,\mathbb{F}_n(R),\|\cdot\|_n)}{n^{1/2}\delta}+\sum_{s=1}^{S}\frac{2^{-s}s^{1/2}}{8}\le\frac{1}{2}+\frac{1}{2}=1.$$

The definition of η_s implies that

$$H(2^{-s}R,\mathbb{F}_n(R),\|\cdot\|_n)\le\frac{n\delta^2\eta_s^2}{144C^2 2^{-2s}R^2},$$

which is used to bound the right side of (S.14) so that

$$\sum_{s=1}^{S} 2\exp\Big\{2H(2^{-s}R,\mathbb{F}_n(R),\|\cdot\|_n) - \frac{n\delta^2\eta_s^2}{36C^2 2^{-2s}R^2}\Big\} \le \sum_{s=1}^{S} 2\exp\Big\{-\frac{n\delta^2\eta_s^2}{72C^2 2^{-2s}R^2}\Big\}.$$

Next, invoke that $\eta_s \ge 2^{-s}s^{1/2}/8$ to get

$$\begin{aligned}
&\sum_{s=1}^{S} 2\exp\Big\{-\frac{n\delta^2\eta_s^2}{72C^2 2^{-2s}R^2}\Big\} \\
&\le \sum_{s=1}^{S} 2\exp\Big\{-\frac{n\delta^2 s}{4608C^2R^2}\Big\} \le \sum_{s=1}^{\infty} 2\exp\Big\{-\frac{n\delta^2 s}{4608C^2R^2}\Big\} \\
&= 2\exp\Big\{-\frac{n\delta^2}{4608C^2R^2}\Big\}\Big[1-\exp\Big\{-\frac{n\delta^2}{4608C^2R^2}\Big\}\Big]^{-1} \\
&\le 4\exp\Big\{-\frac{n\delta^2}{4608C^2R^2}\Big\},
\end{aligned}$$

where in the last inequality, we used (S.9) with a sufficiently large choice of C_0 to ensure that

$$\frac{n^{1/2}\delta}{2} \ge (1152\log 2)^{1/2}CR \qquad \text{or} \qquad \frac{n\delta^2}{4608C^2R^2} \ge \log 2. \tag{S.17}$$

Recall that $C^2 = 8r_v(M_0^2+\sigma_0^2)$. Hence we have proved (S.10) by setting $C_0^2 = 16 \vee 36846\, r_v(M_0^2+\sigma_0^2)/4$. □

Lemma S.8. *Suppose that $\mu(\pm v)$ increases slower than v^L as $v \to \infty$ for some $L > 0$. For some constant M independent of n,*

$$N(r,\mathbb{F}_n(R),\|\cdot\|_n) \le \Big(\frac{MR^LQ_n^{L/2}}{r}\Big)^{2(DQ_n+K+1)} \quad a.s.. \tag{S.18}$$

Proof. We use $\mathbb{F}_n^{\text{id}}$ and $\mathbb{F}_n^{\text{id}}(R)$ to denote the special case of $\mathbb{F}$ and $\mathbb{F}_n(R)$ when $\mu(\cdot)$ is the identity link. By Example 3.7.4d on Page 40 of van de Geer (2000), we know that $\mathbb{F}_n^{\text{id}}$ is a VC subgraph class with VC index $V(\mathbb{F}_n^{\text{id}}) \le DQ_n+K+2$. As a subset of $\mathbb{F}_n^{\text{id}}$, $\mathbb{F}_n^{\text{id}}(R)$ is also a VC subgraph class with VC index $V(\mathbb{F}_n^{\text{xid}}(R)) \le DQ_n+K+2$. Since μ is a monotone function, by Lemma 2.6.18(viii) of van der Vaart and Wellner (1996), we have that $V(\mathbb{F}_n(R)) \le V(\mathbb{F}_n^{\text{id}}(R)) \le DQ_n+K+2$. By the monotonicity of μ, Lemmas S.2 and S.3, and the assumption on the growth rate of $\mu(\pm v)$ as $v\to\infty$, we have that

$$\begin{aligned}
\|\mu(f)\|_\infty &\le |\mu(\|f\|_\infty)| + |\mu(-\|f\|_\infty)| \\
&\le |\mu(M_3\sqrt{Q_n}R)| + |\mu(-M_3\sqrt{Q_n}R)| \lesssim R^LQ_n^{L/2},
\end{aligned}$$

for any $\mu(f) \in \mathbb{F}_n(R)$, as $n \to \infty$. Applying Theorem 3.11 on Page of 40 of van de Geer (2000) with $\delta = rR^{-L}Q_n^{-L/2}$, $p = 2$, and $Q = \|\cdot\|_n$, we obtain that

$$\begin{aligned} N(r, \mathbb{F}_n(R), \|\cdot\|_n) &\le CV(\mathbb{F}_n(R))(16e)^{V(\mathbb{F}_n(R))}\Big(\frac{R^L Q_n^{L/2}}{r}\Big)^{2\{V(\mathbb{F}_n(R))-1\}} \\ &\le C(DQ_n + K + 2)(16e)^{DQ_n+K+2}\Big(\frac{R^L Q_n^{L/2}}{r}\Big)^{2(DQ_n+K+1)} \\ &\le \Big(\frac{MR^L Q_n^{L/2}}{r}\Big)^{2(DQ_n+K+1)} \end{aligned}$$

for some absolute constant M. □

LEMMA S.9. *Let*

$$J_\delta = \int_{\delta^2/128\sigma}^{\delta} H^{1/2}(u, \mathbb{F}_n(\delta), \|\cdot\|_n)du \bigvee \delta, \qquad 0 < \delta < 128\sigma.$$

We have that

$$n^{1/2}\delta^2 \ge 32C_0 J_\delta, \tag{S.19}$$

for $\delta_n \le \delta < 128\sigma$ and some δ_n satisfying $(Q_n/n)^{1/2} \le \delta_n \le (Q_n/n)^{1/2}\log n$.

Proof. By Lemma S.8 we have that

$$J_\delta \le \int_{\delta^2/128\sigma}^{\delta} \Big\{2(DQ_n + K + 1)\log\Big(\frac{M\delta^L Q_n^{L/2}}{u}\Big)\Big\}^{1/2} du \bigvee \delta.$$

Since $\delta < 128\sigma$ and the integrand is a decreasing function of u, we have

$$J_\delta \le \psi_\delta = \begin{cases} \delta\big\{2(DQ_n + K + 1)\log\big(128M\sigma\delta^{L-2}Q_n^{L/2}\big)\big\}^{1/2} & \text{if } 0 < L < 2; \\ \delta\big\{2(DQ_n + K + 1)\log\big(M(128\sigma)^{L-1}Q_n^{L/2}\big)\big\}^{1/2} & \text{if } L \ge 2. \end{cases}$$

Note that $\delta^{-2}\psi_\delta$ is a decreasing function of δ when $0 < \delta < 128\sigma$. Thus,

$$n^{1/2}\delta^2 \ge 32C_0\psi_\delta,$$

for all $\delta_n < \delta < 128\sigma$ with δ_n satisfies

$$n^{1/2}\delta_n^2 = 32C_0\psi_{\delta_n}.$$

It is easy to verify that, for $\delta_{n1} = (Q_n/n)^{1/2}\log n$,

$$\delta_{n1}^{-2}\psi_{\delta_{n1}} \asymp \begin{cases} \frac{n^{1/2}}{\log n}\Big[\log\big\{128M\sigma\big(n/Q_n\big)^{1-L/2}(\log n)^{L/2-1}Q_n^{L/2}\big\}\Big]^{1/2} & \text{if } 0 < L < 2; \\ \frac{n^{1/2}}{\log n}\Big[\log\big\{M(128\sigma)^{L-1}Q_n^{L/2}\big\}\Big]^{1/2} & \text{if } L \ge 2 \end{cases}$$

Thus, $\delta_{n1}^{-2}\psi_{\delta_{n1}} \le n^{1/2}/(32C_0)$ for n large. This implies that $\delta_n \le \delta_{n1} = (Q_n/n)^{1/2}\log n$, because $\delta^{-2}\psi_\delta$ is decreasing in δ. Similarly it can be checked that, for $\delta_{n2} = (Q_n/n)^{1/2}$, $\delta_{n2}^{-2}\psi_{\delta_{n2}} \ge n^{1/2}/(32C_0)$ for n large, which implies that $\delta_n \ge \delta_{n2} = (Q_n/n)^{1/2}$. □

S.4. Proof of Theorem 1

Denote $\boldsymbol{W}_i = \boldsymbol{Y}_i - f_n^*(\mathbf{X}_i, \mathbf{T}_i)$. Then

$$\boldsymbol{W}_i = \{\boldsymbol{Y}_i - f_0(\mathbf{X}_i, \mathbf{T}_i)\} + \{f_0(\mathbf{X}_i, \mathbf{T}_i) - f_n^*(\mathbf{X}_i, \mathbf{T}_i)\} \equiv \boldsymbol{e}_i + \boldsymbol{\xi}_i.$$

In the following, $\boldsymbol{Y}_i$, $\boldsymbol{W}_i$, $\boldsymbol{e}_i$ and $\boldsymbol{\xi}_i$ will be viewed as evaluation of some functions, denoted as y, w, e and ξ respectively. For example, $y = y(\mathbf{x}, \mathbf{t})$ is a function that satisfies $y(\boldsymbol{X}_{ij}, \boldsymbol{T}_{ij}) = Y_{ij}$, and $w = w(\mathbf{x}, \mathbf{t})$ is a function that satisfies $w(\boldsymbol{X}_{ij}, \boldsymbol{T}_{ij}) = W_{ij}$. This functional viewpoint helps simplify our presentation. In particular, it facilitates the use of inner products and norms for representing some complicated expressions.

Recall the definition of the norm $\|\cdot\|_n$ given before the statement of Theorem 1. Denote the corresponding inner product as $\langle \cdot, \cdot \rangle_n$. Since $w = y - f_n^*$ and

$$\|y - f_n^* + f_n^* - \widehat{f}_n\|_n^2 = \|y - \widehat{f}_n\|_n^2 \le \|y - f_n^*\|_n^2,$$

we have that

$$\|\widehat{f}_n - f_n^*\|_n^2 \le 2\langle w, \widehat{f}_n - f_n^* \rangle_n. \tag{S.20}$$

Let σ be a positive constant to be determined later. It follows from (S.20) and the Cauchy–Schwarz inequality $|\langle \varepsilon_1, \varepsilon_2 \rangle_n| \le \|\varepsilon_1\|_n \|\varepsilon_2\|_n$ that $\|\widehat{f}_n - f_n^*\|_n \le 2\sigma$ on $\{\|w\|_n \le \sigma\}$. Let $\delta > 0$ satisfy $\delta \le 2\sigma$. Define $S = \max\{s = 0, 1, \cdots : 2^s\delta < 2\sigma\}$. Applying the peeling device (see the proof of Theorem 9.1 of van de Geer, 2000), we obtain that

$$\begin{aligned}
&P\big(\{\|\widehat{f}_n - f_n^*\|_n > \delta\} \cap \{\|w\|_n \le \sigma\} \big| \{\mathbf{X}_i, \mathbf{T}_i\}_{i=1}^n\big) \\
&\le \sum_{s=0}^{S} P\big(\{2^{s+1}\delta \ge \|\widehat{f}_n - f_n^*\|_n > 2^s\delta\} \cap \{\|w\|_n \le \sigma\} \big| \{\mathbf{X}_i, \mathbf{T}_i\}_{i=1}^n\big) \\
&\le \sum_{s=0}^{S} P\left(\left\{\sup_{f \in \mathbb{F}_n(2^{s+1}\delta)} 2\langle w, f - f_n^* \rangle_n > 2^{2s}\delta^2\right\} \cap \{\|w\|_n \le \sigma\} \middle| \{\mathbf{X}_i, \mathbf{T}_i\}_{i=1}^n\right),
\end{aligned}$$

where $\mathbb{F}_n(R) = \{f \in \mathbb{F}_n : \|f - f_n^*\|_n \le R\}$, as defined earlier. Note that $\|\xi\|_n < c\rho_n$. If we set $\delta > 8c\rho_n$, then we have

$$\big|\langle \xi, f - f_n^* \rangle_n\big| \le c\rho_n \cdot 2^{s+1}\delta \le 2^{2s-2}\delta^2, \qquad f \in \mathbb{F}_n(2^{s+1}\delta).$$

Notice the fact that $w = e + \xi$. The triangle inequality suggests that $\|e\|_n \le 2\sigma$ if $\|w\|_n \le \sigma$ and $\sigma > c\rho_n$. Combining the above discussion, we have that

$$\begin{aligned}
&P\big(\{\|\widehat{f}_n - f_n^*\|_n > \delta\} \cap \{\|w\|_n \le \sigma\} \big| \{\mathbf{X}_i, \mathbf{T}_i\}_{i=1}^n\big) \\
&\le \sum_{s=0}^{S} P\Bigg(\left\{\sup_{f \in \mathbb{F}_n(2^{s+1}\delta)} \big|\langle e, f - f_n^* \rangle_n\big| > 2^{2s-2}\delta^2\right\} \\
&\qquad\qquad \cap \{\|e\|_n \le 2\sigma\} \Bigg| \{\mathbf{X}_i, \mathbf{T}_i\}_{i=1}^n\Bigg).
\end{aligned} \tag{S.21}$$

Note by Lemma S.9 that

$$n^{1/2}2^{2s-2}\delta^2 \geq 2C_0\Big\{\int_{2^{2s-2}\delta^2/8\sigma}^{2^{s+1}\delta} H^{1/2}(u, \mathbb{F}_n(2^{s+1}\delta), \|\cdot\|_n)du \bigvee 2^{s+1}\delta\Big\},$$

provided that $2^{s+1}\delta > \delta_n$ (to ensure this, assume WLOG $\delta > \delta_n$). Now we apply Lemma S.7. First, (S.9) is satisfied when δ and R there are set to be $2^{2s-2}\delta^2$ and $2^{s+1}\delta$, respectively. Then, by (S.10), we have

$$P\bigg(\Big\{\sup_{f\in\mathbb{F}_n(2^{s+1}\delta)} \big|\langle e, f - f_n^*\rangle_n\big| > 2^{2s-2}\delta^2\Big\} \cap \{\|e\|_n \leq 2\sigma\} \bigg| \{\mathbf{X}_i, \mathbf{T}_i\}_{i=1}^n\bigg)$$
$$\leq C_0 \exp\Big(-\frac{n2^{2s}\delta^2}{256C_0^2}\Big).$$

It then follows from (S.21) that for $\delta \geq \delta_n + 8c\rho_n$,

$$P\big(\{\|\widehat{f}_n - f_n^*\|_n > \delta\} \cap \{\|w\|_n \leq \sigma\}\big|\{\mathbf{X}_i, \mathbf{T}_i\}_{i=1}^n\big)$$
$$\leq \sum_{s=0}^{S} C_0 \exp\Big(-\frac{n2^{2s}\delta^2}{256C_0^2}\Big) \leq 2C_0 \exp\Big(-\frac{n\delta^2}{256C_0^2}\Big);$$

here, we use the inequality

$$\sum_{s=0}^{\infty} \exp(-a2^{2s}) \leq \sum_{k=1}^{\infty} \exp(-ak) = \frac{e^{-a}}{1-e^{-a}} \leq 2e^{-a},$$

provided that $e^{-a} < 1/2$. Therefore,

$$P\big(\|\widehat{f}_n - f_n^*\|_n > \delta\big|\{\mathbf{X}_i, \mathbf{T}_i\}_{i=1}^n\big) \leq 2C_0 \exp\Big(-\frac{n\delta^2}{256C_0^2}\Big)$$
$$+ P(\|w\|_n > \sigma|\{\mathbf{X}_i, \mathbf{T}_i\}_{i=1}^n).$$

Note that we previously require $\delta \geq \delta_n + 8c\rho_n$. Then it follows that

$$P\big(\|\widehat{f}_n - f_n^*\|_n > \delta_n + 8c\rho_n\big)$$
$$= E\big\{P\big(\|\widehat{f}_n - f_n^*\|_n > \delta_n + 8c\rho_n\big|\{\mathbf{X}_i, \mathbf{T}_i\}_{i=1}^n\big)\big\}$$
$$\leq 2C_0 \exp\Big(-\frac{n\delta^2}{256C_0^2}\Big) + P\big(\|w\|_n > \sigma\big).$$

Let $c_v = \max_i\{\lambda^{\max}_{\mathbf{V}_i^{-1}}\}$, then $c_v < \infty$ by Condition C4. Set $\sigma = \|e\|_n/2 + c\max_n \rho_n + (2c_v)^{1/2}\sigma_0$ (σ_0 is defined in Condition C6). By the Chebyshev inequality,

$$P(\|w\|_n > \sigma) \leq P(\|e\|_n > (2c_v)^{1/2}2\sigma_0) \leq \exp\Big(-\frac{8nc_v\sigma_0^2}{M_0^2}\Big)E\left(\exp\Big(\frac{n\|e\|_n^2}{M_0^2}\Big)\right).$$

According to Condition C6,

$$E\left\{\exp\Big(\frac{n\|e\|_n^2}{M_0^2}\Big)\right\} \le E\left\{\exp\Big(\sum_{i=1}^n \frac{c_v|\boldsymbol{e}_i|^2}{M_0^2}\Big)\right\} \le \Big(1+\frac{\sigma_0^2}{M_0^2}\Big)^{nc_v} \le \exp\Big(\frac{nc_v\sigma_0^2}{M_0^2}\Big).$$

Hence

$$P(\|w\|_n > \sigma) \le \exp\Big(-\frac{7nc_v\sigma_0^2}{M_0^2}\Big) \to 0.$$

Consequently,

$$P\big(\|\widehat{f}_n - f_n^*\|_n > \delta_n + 8c\rho_n\big) \to 0, \quad \text{as } n \to \infty. \tag{S.22}$$

Since $\delta_n \le (Q_n/n)^{1/2}\log n$, as required in application of Lemma S.9, the first conclusion follows.

By Condition C5(iii) and Lemma S.2, we have that

$$\begin{aligned}\|\widehat{f}_n - f_n^*\|_\infty &\lesssim \|\mathbf{x}'\widehat{\boldsymbol{\beta}}_V + \widehat{\theta} - (\mathbf{x}'\boldsymbol{\beta}_0 + \theta_n^*)\|_\infty \\ &\lesssim Q_n^{1/2}\|\mathbf{x}'\widehat{\boldsymbol{\beta}}_V + \widehat{\theta} - (\mathbf{x}'\boldsymbol{\beta}_0 + \theta_n^*)\|.\end{aligned}$$

Note that $(Q_n \log n)^2/n \to 0$. Moreover, $Q_n\rho_n^2 \to 0$ by Condition C8 and the fact that $\rho_n \asymp Q_n^{-\alpha}$ for $\alpha > 1/2$. It then follows by Condition C5(ii) and the first conclusion that

$$\begin{aligned}&Q_n^{1/2}\|\mathbf{x}'\widehat{\boldsymbol{\beta}}_V + \widehat{\theta} - (\mathbf{x}'\boldsymbol{\beta}_0 + \theta_n^*)\| \\ &\quad \lesssim Q_n^{1/2}\|\widehat{f}_n - f_n^*\| = Q_n^{1/2}O_P\{(Q_n/n)^{1/2}\log n + \rho_n\} = o_P(1).\end{aligned}$$

We thus obtain the second conclusion.

By Taylor's theorem,

$$\begin{aligned}&f_n^*(\mathbf{x},t) - f_0(\mathbf{x},t) \\ &\quad = \int_0^1 \mu'[\mathbf{x}^T\beta_0 + \theta_{0,+}(t) + u\{\theta_n^*(t) - \theta_{0,+}(t)\}]\,du\,\{\theta_n^*(t) - \theta_{0,+}(t)\}.\end{aligned} \tag{S.23}$$

Due to Condition C5(iii), it follows that $\|f_n^* - f_0\|_\infty = O(\|\theta_n^* - \theta_{0,+}\|_\infty) = O(\rho_n)$. Combining this with the second conclusion gives $\|\widehat{f}_n - f_0\|_\infty = o_P(1)$, which is the third conclusion.

From the proof of the second conclusion, we have that

$$\|\mathbf{x}'\widehat{\boldsymbol{\beta}}_V + \widehat{\theta} - (\mathbf{x}'\boldsymbol{\beta}_0 + \theta_n^*)\| = O_P\{(Q_n/n)^{1/2}\log n + \rho_n\}.$$

By Lemma 3.1 of Stone (1994),

$$\|\mathbf{x}'(\widehat{\boldsymbol{\beta}}_V - \boldsymbol{\beta}_0)\|^2 + \|\widehat{\theta} - \theta_n^*\|^2 \lesssim \|\mathbf{x}'\widehat{\boldsymbol{\beta}}_V + \widehat{\theta}_n - (\mathbf{x}'\boldsymbol{\beta}_0 + \theta_n^*)\|^2.$$

Combining these two results, we obtain that $\|\mathbf{x}'(\widehat{\boldsymbol{\beta}}_V - \boldsymbol{\beta}_0)\|^2 = o_P(1)$, which together with the no-multicollinearity condition C2 implies $\widehat{\boldsymbol{\beta}}_V \xrightarrow{P} \boldsymbol{\beta}_0$. We also obtain that $\|\widehat{\theta} - \theta_n^*\| = O_P\{(Q_n/n)^{1/2}\log n + \rho_n\}$. Thus, by Lemma S.2,

$$\|\widehat{\theta} - \theta_n^*\|_\infty = Q_n^{1/2} O_P\{(Q_n/n)^{1/2}\log n + \rho_n\} = o_P(1).$$

Since $\|\theta_n^* - \theta_{0,+}\|_\infty = O(\rho_n) = o(1)$, application of the triangle inequality yields $\|\widehat{\theta} - \theta_{0,+}\|_\infty = o_P(1)$, the last conclusion.

The proof of the theorem is complete.

S.5. Some Lemmas for Theorem 2

The following Lemmas are needed for the asymptotic normality result Theorem 2.

LEMMA S.10. *There exists a positive constant M_5 that is independent of n such that*

$$|\mu'(\boldsymbol{X}_{ij}'\widehat{\boldsymbol{\beta}}_V + \boldsymbol{Z}_{ij}'\widehat{\gamma}) - \mu'(\boldsymbol{X}_{ij}'\boldsymbol{\beta}_0 + \theta_{0,+}(\boldsymbol{T}_{ij})| \le M_5|\widehat{d}_{ij}|,$$

where $\widehat{d}_{ij}$ is the jth element of $\widehat{\boldsymbol{d}}_i = \mathbf{X}_i\widehat{\boldsymbol{\beta}}_V + \mathbf{Z}_i\widehat{\gamma} - \mathbf{X}_i\boldsymbol{\beta}_0 - \theta_{0,+}(\mathbf{T}_i)$.

Proof. This is an immediate consequence of the intermediate value theorem, Theorem 1, and Condition C5(iii). □

Let $\widehat{\Pi}_n^\triangle$, $\Pi_n^\triangle$ denote respectively the projection onto $\mathbb{G}_+$ relative to $\langle\cdot,\cdot\rangle_n^\triangle$ and $\langle\cdot,\cdot\rangle^\triangle$. Let X_{ijk} denote the k-th component of $\boldsymbol{X}_{ij}$ and $\mathbf{H}_{21,k}$ denote the k-th column of $\mathbf{H}_{21}$ where $\mathbf{H}_{21}$ was defined in (20) of the main paper.

LEMMA S.11. *There exists a positive constant M_6 that is independent of n such that,*

$$\max_{1\le i\le n}\max_{1\le j\le m_i} |X_{ijk} - \boldsymbol{Z}_{ij}'\mathbf{H}_{22}^{-1}\mathbf{H}_{21,k}| \le M_6, \qquad k = 1,\ldots,K, \tag{S.24}$$

on Ω_n^c with Ω_n satisfying $P(\Omega_n) \to 0$.

Proof. Using the above defined inner product, we see that the least favorable direction $\varphi_{k,+}^* = \arg\min_{\varphi\in L_{2,+}} \|x_k - \varphi\|^\triangle$, where $L_{2,+}$ denote the space of square integrable additive functions. To avoid clutter, we drop the subscript $+$ in the rest of the proof of this lemma. Define $\varphi_{k,n}^* = \Pi_n^\triangle x_k$ and $\widehat{\varphi}_{k,n} = \widehat{\Pi}_n^\triangle x_k$. Recall that $x_k(\cdot)$ denotes the coordinate mapping that maps x to its k-th component. We know that $(x_k - \widehat{\Pi}_n^\triangle x_k)$ maps X_{ijk} to $(X_{ijk} - \boldsymbol{Z}_{ij}'\mathbf{H}_{22}^{-1}\mathbf{H}_{21,k})$. Thus, in order to show (S.24), we only need to show the boundedness of $\widehat{\varphi}_{k,n}$. We use the boundedness of φ_k^* and $\varphi_{k,n}^*$ as a bridge.

In subsection S.1, we have already shown the the boundedness of φ_k^*. Now we show the the boundedness of $\varphi_{k,n}^*$. Let $g_n^* \in \mathbb{G}_+$ be the minimizer of $\|\varphi_k^* - g\|_\infty$. Then we have $\|\varphi_k^* - g_n^*\|_\infty \lesssim \rho_n$. Since $\varphi_{k,n}^* - g_n^* = \Pi_n^\triangle(\varphi_k^* - g_n^*)$, we have

$$\|\varphi_{k,n}^* - g_n^*\|^\triangle \le \|\varphi_k^* - g_n^*\|^\triangle \lesssim \rho_n.$$

It follows by applying Lemma S.2(ii) that $\|\varphi^*_{k,n} - g^*_n\|_\infty \lesssim Q_n^{1/2}\|\varphi^*_{k,n} - g^*_n\|^\triangle \lesssim Q_n^{1/2}\rho_n$ Hence,

$$\begin{aligned}\|\varphi^*_{k,n}\|_\infty &\le \|\varphi^*_{k,n} - g^*_n\|_\infty + \|g^*_n - \varphi^*_k\|_\infty + \|\varphi^*_k\|_\infty \\ &\lesssim Q_n^{1/2}\rho_n + \rho_n + 1 \lesssim 1.\end{aligned} \tag{S.25}$$

Next we show the boundedness of $\widehat{\varphi}_{k,n}$. By the definition of $\widehat{\varphi}_{k,n}$ and $\varphi^*_{k,n}$ and the Cauchy–Schwarz inequality,

$$\begin{aligned}\|\widehat{\varphi}_{k,n} - \varphi^*_{k,n}\|_n^\triangle &= \sup_{g\in\mathbb{G}_+} \frac{|\langle \widehat{\varphi}_{k,n} - \varphi^*_{k,n}, g\rangle_n^\triangle|}{\|g\|_n^\triangle} = \sup_{g\in\mathbb{G}_+} \frac{|\langle x_k - \varphi^*_{k,n}, g\rangle_n^\triangle|}{\|g\|_n^\triangle} \\ &= \sup_{g\in\mathbb{G}_+} \frac{|\langle x_k - \varphi^*_{k,n}, g\rangle_n^\triangle - \langle x_k - \varphi^*_{k,n}, g\rangle^\triangle|}{\|g\|_n^\triangle}.\end{aligned}$$

By Lemma S.3, Lemma S.4 and the fact that $\|x_k - \varphi^*_{k,n}\|_\infty$ is bounded, i.e. (S.25), we have

$$\begin{aligned}&\sup_{g\in\mathbb{G}_+} \frac{|\langle x_k - \varphi^*_{k,n}, g\rangle_n^\triangle - \langle x_k - \varphi^*_{k,n}, g\rangle^\triangle|}{\|g\|_n^\triangle} \\ &\qquad = O_P\Big(\sqrt{\frac{Q_n}{n}}\Big)\|x_k - \varphi^*_{k,n}\|_n^\triangle = O_P\Big(\sqrt{\frac{Q_n}{n}}\Big).\end{aligned}$$

Therefore, $\|\widehat{\varphi}_{k,n} - \varphi^*_{k,n}\|_n^\triangle = O_P((Q_n/n)^{1/2})$. Applying Lemma S.3, we have that

$$\|\widehat{\varphi}_{k,n} - \varphi^*_{k,n}\|^\triangle = (1 + o_P(1))\|\widehat{\varphi}_{k,n} - \varphi^*_{k,n}\|_n^\triangle = O_P\Big(\sqrt{\frac{Q_n}{n}}\Big),$$

which implies that

$$\|\widehat{\varphi}_{k,n} - \varphi^*_{k,n}\|_\infty \le \sqrt{Q_n}\|\widehat{\varphi}_{k,n} - \varphi^*_{k,n}\|^\triangle = O_P\Big(\sqrt{\frac{Q_n^2}{n}}\Big) = o_P(1).$$

Combining the triangular inequality with (S.25), we obtain that $\|\widehat{\varphi}_{k,n}\|_\infty$ is bounded with probability tending to one. □

LEMMA S.12. $n\mathbf{H}^{11} = O_P(1)$.

Proof. The same argument as in the proof of Theorem 2 of Huang, Zhang and Zhou (2007) gives this result. □

S.6. Proof of Theorem 2

Note that $\widehat{\boldsymbol{\beta}}_V \in \mathbb{R}^K$ and $\widehat{\boldsymbol{\gamma}} \in \mathbb{R}^{Q_n}$ solve the estimating equations

$$\begin{cases}\sum_{i=1}^n \mathbf{X}_i'\widehat{\boldsymbol{\Delta}}_i\mathbf{V}_i^{-1}\{\boldsymbol{Y}_i - \mu(\mathbf{X}_i\widehat{\boldsymbol{\beta}}_V + \mathbf{Z}_i\widehat{\boldsymbol{\gamma}})\} = 0,\\ \sum_{i=1}^n \mathbf{Z}_i'\widehat{\boldsymbol{\Delta}}_i\mathbf{V}_i^{-1}\{\boldsymbol{Y}_i - \mu(\mathbf{X}_i\widehat{\boldsymbol{\beta}}_V + \mathbf{Z}_i\widehat{\boldsymbol{\gamma}})\} = 0,\end{cases}$$

or, equivalently

$$\sum_{i=1}^{n} \mathbf{U}_i' \widehat{\mathbf{\Delta}}_i \mathbf{V}_i^{-1} \{\boldsymbol{Y}_i - \mu(\mathbf{X}_i \widehat{\boldsymbol{\beta}}_V + \mathbf{Z}_i \widehat{\boldsymbol{\gamma}})\} = 0, \tag{S.26}$$

with $\mathbf{U}_i = (\mathbf{X}_i, \mathbf{Z}_i)$, and $\widehat{\mathbf{\Delta}}_i$ is a diagonal matrix with the diagonal elements being the first derivative of $\mu(\cdot)$ evaluated at $X_{ij}' \widehat{\boldsymbol{\beta}}_V + Z_{ij}' \widehat{\boldsymbol{\gamma}}$, $j = 1, \ldots, m_i$. Recall that $\boldsymbol{\gamma}^*$ is assumed to satisfy $\rho_n = \|\theta_{0,+} - \boldsymbol{B}' \boldsymbol{\gamma}^*\|_\infty \to 0$. Then the boundedness of $\mathbf{x}' \boldsymbol{\beta}_0 + \theta_0(\mathbf{t})$ implies the same property of $\mathbf{x}' \boldsymbol{\beta}_0 + \boldsymbol{B}'(t) \boldsymbol{\gamma}^*$. By (17) in the main paper and Condition C5(ii), $\mathbf{x}' \widehat{\boldsymbol{\beta}}_V + \boldsymbol{B}'(t) \widehat{\boldsymbol{\gamma}}$ is also bounded.

Using the Taylor expansion, we have that

$$\begin{aligned} \mu(\mathbf{X}_i \widehat{\boldsymbol{\beta}}_V + \mathbf{Z}_i \widehat{\boldsymbol{\gamma}}) = \mu\big(\mathbf{X}_i \boldsymbol{\beta}_0 + \theta_0(\boldsymbol{T}_i)\big) \\ + \mathbf{\Delta}_{i0} \big\{\mathbf{X}_i (\widehat{\boldsymbol{\beta}}_V - \boldsymbol{\beta}_0) + \mathbf{Z}_i \widehat{\boldsymbol{\gamma}} - \theta_0(\boldsymbol{T}_i)\big\} + \widehat{\boldsymbol{r}}_i \end{aligned} \tag{S.27}$$

where $\widehat{\boldsymbol{r}}_i$ is a remainder term satisfying

$$\widehat{\boldsymbol{r}}_i \lesssim \widehat{\boldsymbol{d}}_i' \widehat{\boldsymbol{d}}_i \tag{S.28}$$

with

$$\widehat{\boldsymbol{d}}_i = (\widehat{d}_{i1}, \ldots, \widehat{d}_{im_i})' = \mathbf{X}_i \widehat{\boldsymbol{\beta}}_V + \mathbf{Z}_i \widehat{\boldsymbol{\gamma}} - \mathbf{X}_i \boldsymbol{\beta}_0 - \theta_0(\boldsymbol{T}_i), \qquad i = 1, \cdots, n.$$

Note that the second line of (S.27) can be rewritten as

$$\mathbf{\Delta}_{i0} \mathbf{U}_i \begin{pmatrix} \widehat{\boldsymbol{\beta}}_V - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^* \end{pmatrix} + \mathbf{\Delta}_{i0} \big(\mathbf{Z}_i \boldsymbol{\gamma}^* - \theta_0(\boldsymbol{T}_i)\big) + \widehat{\boldsymbol{r}}_i. \tag{S.29}$$

Define

$$\widetilde{\boldsymbol{J}}_1 = (\widehat{\mathbf{\Delta}}_i - \mathbf{\Delta}_{i0}) \mathbf{V}_i^{-1} \big\{\boldsymbol{Y}_i - \mu(\mathbf{X}_i \widehat{\boldsymbol{\beta}}_V + \mathbf{Z}_i \widehat{\boldsymbol{\gamma}})\big\}$$

and

$$\boldsymbol{J}_2 = \mathbf{\Delta}_{i0} \mathbf{V}_i^{-1} \big\{\boldsymbol{Y}_i - \mu\big(\mathbf{X}_i \boldsymbol{\beta}_0 + \theta_0(\boldsymbol{T}_i)\big) - \widehat{\boldsymbol{r}}_i - \mathbf{\Delta}_{i0} \big(\mathbf{Z}_i \boldsymbol{\gamma}^* - \theta_0(\boldsymbol{T}_i)\big)\big\}.$$

Substituting (S.27) into (S.26) yields

$$0 = \sum_{i=1}^{n} \mathbf{U}_i' (\boldsymbol{J}_1 + \boldsymbol{J}_2) - \sum_{i=1}^{n} \mathbf{U}_i' \mathbf{\Delta}_{i0} \mathbf{V}_i^{-1} \mathbf{\Delta}_{i0} \mathbf{U}_i \begin{pmatrix} \widehat{\boldsymbol{\beta}}_V - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^* \end{pmatrix}. \tag{S.30}$$

Recalling (20) and using (21) in the main paper, we obtain from (S.30) that

$$\widehat{\boldsymbol{\beta}}_V = \boldsymbol{\beta}_0 + \mathbf{H}^{11} \sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{Z}_i \mathbf{H}_{22}^{-1} \mathbf{H}_{21})' (\widetilde{\boldsymbol{J}}_1 + \boldsymbol{J}_2). \tag{S.31}$$

Recall that $\boldsymbol{e}_i = \boldsymbol{Y}_i - \mu(\mathbf{X}_i\boldsymbol{\beta}_0 + \theta_0(\boldsymbol{T}_i))$ and define $\boldsymbol{\zeta}_i = \mu(\mathbf{X}_i\boldsymbol{\beta}_0 + \theta_0(\boldsymbol{T}_i)) - \mu(\mathbf{X}_i\widehat{\boldsymbol{\beta}}_V + \mathbf{Z}_i\widehat{\boldsymbol{\gamma}})$. Define

$$\boldsymbol{I}_1 = \mathbf{H}^{11}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})'(\widehat{\boldsymbol{\Delta}}_i - \boldsymbol{\Delta}_{i0})\mathbf{V}_i^{-1}\boldsymbol{e}_i$$
$$\boldsymbol{I}_2 = \mathbf{H}^{11}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})'(\widehat{\boldsymbol{\Delta}}_i - \boldsymbol{\Delta}_{i0})\mathbf{V}_i^{-1}\boldsymbol{\zeta}_i$$
$$\boldsymbol{I}_3 = \mathbf{H}^{11}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\boldsymbol{e}_i$$
$$\boldsymbol{I}_4 = \mathbf{H}^{11}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\widehat{\boldsymbol{r}}_i$$
$$\boldsymbol{I}_5 = \mathbf{H}^{11}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\boldsymbol{\Delta}_{i0}\{\mathbf{Z}_i\boldsymbol{\gamma}^* - \theta_0(\boldsymbol{T}_i)\}$$

Then, (S.31) can be written as

$$\widehat{\boldsymbol{\beta}}_V = \boldsymbol{\beta}_0 + \boldsymbol{I}_1 + \boldsymbol{I}_2 + \boldsymbol{I}_3 - \boldsymbol{I}_4 - \boldsymbol{I}_5. \tag{S.32}$$

To prove the asymptotic linear expansion (22) in the main paper, we need to show that

$$|\boldsymbol{I}_1| + |\boldsymbol{I}_2| + |\boldsymbol{I}_4| + |\boldsymbol{I}_5| = o_P\Big(n^{-1/2}\Big). \tag{S.33}$$

Write $\boldsymbol{I}_1 = n\mathbf{H}^{11}\boldsymbol{I}_1^o$, where

$$\boldsymbol{I}_1^o = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})'(\widehat{\boldsymbol{\Delta}}_i - \boldsymbol{\Delta}_{i0})\mathbf{V}_i^{-1}\boldsymbol{e}_i.$$

Set

$$h_k(\mathbf{x},\mathbf{t};\boldsymbol{\beta},\boldsymbol{\gamma}) = \{\mu'(\mathbf{x}'\boldsymbol{\beta} + \boldsymbol{B}'(\mathbf{t})\boldsymbol{\gamma}) - \mu'(\mathbf{x}'\boldsymbol{\beta}_0 + \theta_{0,+}(\mathbf{t}))\}\{x_k - \boldsymbol{B}(\mathbf{t})'\mathbf{H}_{22}^{-1}\mathbf{H}_{21,k}\},$$

where x_k is the k-th element of $\mathbf{x}$ and $\mathbf{H}_{21,k}$ is the k-th column of $\mathbf{H}_{21}$. Then the k-th element of $\boldsymbol{I}_1^o$ is

$$[\boldsymbol{I}_1^o]_k = \frac{1}{n}\sum_{i=1}^{n}h_k'(\mathbf{X}_i,\mathbf{T}_i;\widehat{\boldsymbol{\beta}}_V,\widehat{\boldsymbol{\gamma}})\mathbf{V}_i^{-1}\boldsymbol{e}_i$$

for $k = 1,2,\ldots,K$. Note that $\widehat{f}(\mathbf{x},\mathbf{t}) = \mu(\mathbf{x}'\widehat{\boldsymbol{\beta}}_V + B'(\mathbf{t})\widehat{\boldsymbol{\gamma}})$ and $f_0(\mathbf{x},\mathbf{t}) = \mu(\mathbf{x}'\boldsymbol{\beta}_0 + \theta_{0,+}(\mathbf{t}))$. For notational simplicity, we write $h_k(\cdot)$ as $h(\cdot)$ hereafter. By Lemmas S.10 and S.11, Condition C5, and (S.22), the following holds on Ω_n^c with $P(\Omega_n) \to 0$,

$$\|h(\cdot;\widehat{\boldsymbol{\beta}}_V,\widehat{\boldsymbol{\gamma}})\|_n \le c_1\|\widehat{f}_n - f_0\|_n \le c_1\|\widehat{f}_n - f_n^*\|_n + c_1\|f_n^* - f_0\|_n \le c_1(\delta_n + \rho_n) \equiv \tau_n.$$

Recall that $\mathbb{F}_n^{\text{id}} = \{f(\mathbf{x}, \mathbf{t}) = \mathbf{x}'\boldsymbol{\beta} + g(\mathbf{t}) : \boldsymbol{\beta} \in \mathbb{R}^K, g \in \mathbb{G}_+\}$. Let

$$\mathbb{H}_n = \{h(\mathbf{x}, \mathbf{t}; f) : h(\mathbf{x}, \mathbf{t}; f) = [\mu'(f(\mathbf{x}, \mathbf{t})) - \mu'(f_0(\mathbf{x}, \mathbf{t}))][x_k - \boldsymbol{B}(\mathbf{t})'\mathbf{H}_{22}^{-1}\mathbf{H}_{21,k}], f \in \mathbb{F}_n^{\text{id}}\}$$

and $\mathbb{H}_n(R) = \{h \in \mathbb{H}_n : \|h\|_n \leq R\}$. For $f_1, f_2 \in \mathbb{F}_n^{\text{id}}$, let $h_1(\cdot) = h(\cdot; f_1)$ and $h_2(\cdot) = h(\cdot; f_2) \in \mathbb{H}_n$. It follows from Lemmas S.10 and S.11 that $\|h_1 - h_2\|_n \leq c_2\|f_1 - f_2\|_n$ for some constant $c_2 > 0$. Thus,

$$N(2c_2 u, \mathbb{H}_n(\tau_n), \|\cdot\|_n) \leq N(u, \mathbb{F}_n^{\text{id}}(\tau_n), \|\cdot\|_n) \leq \left(\frac{4\tau_n + u}{u}\right)^{Q_n+K}$$

by Lemma 2.5 on Page 20 of van de Geer (2000). We have that, on Ω_n^c,

$$|[\boldsymbol{I}_1^o]_k| \leq \sup_{h \in \mathbb{H}_n(\tau_n)} n^{-1}\left|\sum_{i=1}^n h(\mathbf{X}_i, \boldsymbol{T}_i)'\mathbf{V}_i^{-1}\boldsymbol{e}_i\right| \equiv \sup_{h \in \mathbb{H}_n(\tau_n)} |\widetilde{\boldsymbol{I}}_1|.$$

Note that for some large σ, $\|e\|_n < \sigma$ with probability tending to one. Now we apply Lemma S.7 with $R = \tau_n$ and $\sigma = 2\|e\|_n$. Note that (S.9) holds for $\delta = c_3\tau_n(Q_n \log n/n)^{1/2}$. We have by (S.10) that, on Ω_n^c,

$$P\left(\left\{\sup_{h \in \mathbb{H}_n(\tau_n)} |\widetilde{\boldsymbol{I}}_1| \geq c_3\tau_n\left(\frac{Q_n \log n}{n}\right)^{1/2}\right\} \cap \{\|e\|_n \leq \sigma\} \,\middle|\, \{\mathbf{X}_i, \mathbf{T}_i\}_{i=1}^n\right) \leq C_0 \exp\left(-\frac{c_3^2 Q_n \log n}{4C_0^2}\right).$$

Therefore,

$$P\left(|[\boldsymbol{I}_1^o]_k| \geq c_3\tau_n\left(\frac{Q_n \log n}{n}\right)^{1/2}\right) \leq C_0 \exp\left(-\frac{c_3^2 Q_n \log n}{4C_0^2}\right) + P(\Omega_n) \to 0 \quad \text{as } n \to \infty.$$

This together with Lemma S.12 and Condition C8 implies that

$$\boldsymbol{I}_1 = \left(n\mathbf{H}^{11}\right)\boldsymbol{I}_1^o = O_P(1) O_P\left(c_3\tau_n\left(\frac{Q_n \log n}{n}\right)^{1/2}\right) = o_P\left(n^{-1/2}\right).$$

By the Cauchy–Schwarz inequality, Condition C4, Lemmas S.10, S.11 and S.12, and then by Theorem 1 and Condition C8,

$$|\boldsymbol{I}_2| = O_P(1)\left(\frac{1}{n}\sum_{i=1}^n \sum_{j=1}^{m_i} \widehat{d}_{ij}^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n \sum_{j=1}^{m_i} \zeta_{ij}^2\right)^{1/2}$$

$$= O_P\Big(\sqrt{\frac{Q_n}{n}}\log n + \rho_n\Big)^2 = o_P\Big(n^{-1/2}\Big).$$

It follows from Lemma S.11, Conditions C4 and C5(ii), inequality (S.28), and then from Theorem 1 that

$$|\boldsymbol{I}_4| = O_P(1)\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\widehat{d}_{ij}^2 = O_P\Big(\sqrt{\frac{Q_n}{n}}\log n + \rho_n\Big)^2 = o_P\Big(n^{-1/2}\Big).$$

We now give an upper bound for $\boldsymbol{I}_5$. We show that the same argument as in dealing with the bias term for the identity link case (Huang, Zhang and Zhou, 2007) can be used for our purposes. Denote $\delta(\mathbf{t}) = \boldsymbol{B}'(\mathbf{t})\boldsymbol{\gamma}^* - \theta_0(\mathbf{t})$ and $\boldsymbol{\delta}_i = \delta(\mathbf{T}_i)$. We have that

$$\boldsymbol{I}_5 = \mathbf{H}^{11}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\boldsymbol{\Delta}_{i0}\boldsymbol{\delta}_i \equiv n\mathbf{H}^{11}\boldsymbol{S},$$

where $\boldsymbol{S} = (S_1, \cdots, S_K)'$ and

$$S_k = \langle x_k - \widehat{\Pi}_n^{\triangle}x_k, \delta\rangle_n^{\triangle} = \langle x_k - \widehat{\Pi}_n^{\triangle}x_k, \delta - \widehat{\Pi}_n^{\triangle}\delta\rangle_n^{\triangle} = \langle x_k, \delta - \widehat{\Pi}_n^{\triangle}\delta\rangle_n^{\triangle}.$$

Observe that $\|\delta(\cdot)\|_\infty \le \rho_n$. It follows from exactly the same argument as in the proof of Theorem 2 of Huang, Zhang and Zhou (2007) that $|S_k| = o_P(n^{-1/2})$. By Lemma S.12, $n\mathbf{H}^{11} = O_P(1)$ and as a consequence, $\boldsymbol{I}_5 = o_P(n^{-1/2})$. The proof of (22) in the main paper is complete.

By Lemma S.11, Lemma S.12, Conditions C3, C4, and C5(iii), we have

$$\mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_V) = O_P\Big(\frac{1}{n}\Big).$$

Then, we have, by the asymptotic linear expansion proven above,

$$\begin{aligned}&\{\mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_V)\}^{-1/2}(\widehat{\boldsymbol{\beta}}_V - \boldsymbol{\beta}_0)\\ &\qquad = \{\mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_V)\}^{-1/2}\Big(\mathbf{H}^{11}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\boldsymbol{e}_i\Big) + o_P(1).\end{aligned}$$

Then by applying the central limit theorem to the above equation and using the fact that

$$\mathrm{var}\Big(\mathbf{H}^{11}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{H}_{22}^{-1}\mathbf{H}_{21})'\boldsymbol{\Delta}_{i0}\mathbf{V}_i^{-1}\boldsymbol{e}_i\Big|\{\mathbf{X}_i, \mathbf{T}_i\}_{i=1}^n\Big) = \mathbf{R}^{\triangle}(\widehat{\boldsymbol{\beta}}_V),$$

we obtain the second conclusion. This completes the whole proof.

S.7. More Simulation Results

We have ran a simulation study to compare our method with that by Carroll *et al.* (2009) under exactly the same setups used in their paper. The results are reported in Table S1. We found that our spline method performs comparably to the local polynomial methods of Carroll *et al.* (2009) in these simple settings.

TABLE S1

MSEs $\times 10^4$ of the estimates for the Euclidean parameter based on 500 samples of size 200 on the exact same setups as Carroll et al. (2009). The methods "Local constant", "Local linear", and our spline method are labeled as LC, LL, and SP, respectively. Results for LC, LL are adapted from Carroll, et al. (2009), Table 1.

Setup	Accounting for correlation			Working independence		
	LC	LL	SP	LC	LL	SP
1	2	2	1	12	12	7
2	9	9	4	13	13	7
3	15	16	6	16	16	6
4	4	4	3	14	14	8
5	2	3	1	14	14	6
6	6	6	12	62	63	29
7	31	31	27	67	67	28

References

BICKEL, P. J., KLAASSEN, A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

CARROLL, R. J., MAITY, A., MAMMEN, E. and YU, K. (2009). Efficient semiparametric marginal Estimation for partially linear additive model for longitudinal/clustered data. *Statistics in BioSciences* **1** 10–31.

HUANG, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *Annals of Statistics* **26** 242–272.

HUANG, J. Z. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics* **31** 1600–1635.

HUANG, J. Z., WU, C. O. and ZHOU, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* **14** 763–788.

HUANG, J. Z., ZHANG, L. and ZHOU, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scandinavian Journal of Statistics* **34** 451–477.

KRESS, R. (1999). *Linear Integral Equations*, Second ed. Springer, New York.

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics* **22** 118–171.

VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.