Efficient Estimation in Marginal Partially Linear Models for Longitudinal/Clustered Data Using Splines

JIANHUA Z. HUANG

Department of Statistics, Texas A&M University

LIANGYUE ZHANG

Retail Financial Services, JPMorgan Chase

LAN ZHOU

Department of Statistics, Texas A&M University

ABSTRACT. We consider marginal semiparametric partially linear models for longitudinal/clustered data and propose an estimation procedure based on a spline approximation of the nonparametric part of the model and an extension of the parametric marginal generalized estimating equations (GEE). Our estimates of both parametric part and nonparametric part of the model have properties parallel to those of parametric GEE, that is, the estimates are efficient if the covariance structure is correctly specified and they are still consistent and asymptotically normal even if the covariance structure is misspecified. By showing that our estimate achieves the semiparametric information bound, we actually establish the efficiency of estimating the parametric part of the model in a stronger sense than what is typically considered for GEE. The semiparametric efficiency of our estimate is obtained by assuming only conditional moment restrictions instead of the strict multivariate Gaussian error assumption.

Key words: clustered data, generalized estimating equations (GEE), longitudinal data, marginal model, nonparametric regression, partially linear models, polynomial splines, semiparametric efficiency.

Running Heading: Partially Linear Models for Clustered Data

1 Introduction

There has been substantial recent interest in developing nonparametric and semiparametric regression methods for longitudinal/clustered data. An incomplete list of publications include Brumback & Rice (1998), Staniswalis & Lee (1998), Hoover *et al.* (1998), Wu *et al.* (1998), Fan & Zhang (2000), Lin & Carroll (2000), Chiang *et al.* (2001), Lin & Ying (2001), Martinussen & Scheike (2001), Wu & Zhang (2002), Wu & Liang (2004), Fan & Li (2004), Sun & Wu (2005), among many others. It is well-known that a fully nonparametric regression model suffers from the "curse of dimensionality", and some structure on the regression function is usually introduced to make the statistical analysis effective.

The partially linear model overcomes the curse of dimensionality by assuming some parametric structure. In a partially linear model, the mean of the outcome is assumed to depend on some covariates X parametrically and some other covariate T nonparametrically. This model is particularly appealing when the effects of X (e.g., treatment) are of major interest, while the effects of 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 (1988), Speckman (1988), and Severini & Staniswalis (1994). Martinussen *et al.* (2002) considered the partially linear model in the context of Cox's regression model for survival data. For a partially linear model for clustered data, how to effectively take into account within-cluster correlation has been a major concern. Lin & Carroll (2001a) showed that, when the parametric covariates are not independent of the nonparametric covariate, a natural application of the local polynomial kernel method can not properly account for the within-cluster correlation and therefore fails to yield a semiparametric efficient estimator. On the otherhand, the same kind of kernel method can account for within-cluster correlation and produce efficient estimator when the covariate modeled nonparametrically is a cluster-level covariate, such as a time-independent covariate in longitudinal data, or a family-level covariate in familial studies (Lin & Carroll, 2001b),.

The purpose of this paper is to propose a straightforward, unified method that is semiparametric efficient regardless of whether or not the covariate modeled nonparametrically is of cluster-level. Our method is based on a spline approximation of the nonparametric part of the model and an extension of the standard generalized estimating equations (GEE). It is shown that our estimates of both parametric part and nonparametric part of the model enjoy the same nice properties as parametric GEE, that is, the estimates are efficient if the covariance structure is correctly specified and they are still consistent and asymptotic normal even if the covariance structure is misspecified. It is also shown that our estimate of the parametric part achieves the semiparametric information bound and thus it is efficient in a stronger sense than what is typically considered for GEE. The semiparametric efficiency of our estimate is obtained by assuming only conditional moment restrictions and the strict multivariate Gaussian error assumption is not needed.

The proposed method can be viewed as an application of a general approach of functional modeling based on spline approximations (or basis expansion approximations, see Huang (2001). The attraction of this general methodology is that it explicitly converts a problem with an infinite-dimensional parameter to one with only a finite number of parameters. The close connection of this approach to the parametric approach has important implications. In particular, solutions to familiar parametric models immediately suggest practical solutions to problems with nonparametric components. Indeed, our proposed spline method can effectively take into account within-cluster correlation in estimation just as in the parametric GEE, while the traditional kernel method had difficulty in doing so. Our method is straightforward to implement; any software package for fitting the standard parametric GEE can be used for our purpose. It is also easy to extend the proposed estimation method to handle other semiparametric models such as partially linear additive models and partially linear varying coefficient models (see Section 6). Previous application of spline approximations in longitudinal data modeling includes He *et al.* (2002) and Huang *et al.* (2002, 2004). However, efficient estimation by accounting for within-subject correlation has not been considered in these work.

Wang *et al.* (2005) and Chen & Jin (2006) proposed methods for semiparametric efficient estimation of partially linear model for clustered data under the Multivariate Gaussian assumption. By employing the iterative kernel method of Wang (2003) that can effectively account for the within-cluster correlation, Wang *et al.* (2005) constructed an estimator that is semiparametric efficient for the model considered in this paper. The implementation of their method involves a computationally complex iterative backfitting algorithm and requires a good first approximation to the nonparametric function. Unlike the spline-based method, extension of the iterative kernel method is not straightforward to more general models such as a partially linear additive model, which models several covariates nonparametrically in an additive manner. The work of Chen & Jin (2006), which we are aware of during the revision of this paper, is closely related to ours. Chen & Jin proposed to use (non-smooth) piecewise polynomials instead of (smooth) spline approximations to approximate the nonparametric part of the model and then apply the idea of GEE. From the practical viewpoint, use of (smooth) spline approximations is favorable if a smooth fit of the nonparametric function is desired. Chen & Jin (2006) actually proposed a second stage kernel smoothing to get a smooth fit of the nonparametric function. Our work shows that this second stage smoothing is not necessary if a smooth spline approximation is used in GEE.

As a final note, it is interesting to point out that Wang *et al.* (2005) and Chen & Jin (2006) only established their asymptotic results under the assumption that data from different clusters/subjects are iid, although it is believed that the results should hold more generally. As a comparison, the iid assumption is not required when we establish our asymptotic results in this paper. This has important practical implications since the iid assumption has certainly ruled out many longitudinal data sets with unequal number of observations for different subjects. The iid assumption has been the focus in the econometrics literature on efficient estimation under conditional moment restrictions (Chamberlain, 1992; Ai & Chen, 2003), the departure from the iid assumption makes the current work beyond the scope of that literature.

The rest of the paper is organized as follows. In Section 2, we present the model and state the major assumptions. In Section 3, we describe the proposed estimation method. Section 4 states the theoretical results. Some numerical results are provided in Section 5. Section 6 gives concluding remarks. Finally, all technical proofs are contained in the Appendix.

2 The Model

Suppose that the data consist of n clusters with the ith ($i = 1, \dots, n$) cluster having m_i observations. In particular, a cluster represents an individual for longitudinal data. The data from different clusters are independent, but correlation may exist within a cluster. Let Y_{ij} and (X_{ij}, T_{ij}) be the response variable and covariates for the jth ($j = 1, \dots, m_i$) observation in the ith cluster. Here X_{ij}

is a $p \times 1$ vector and T_{ij} is a scalar that varies within each cluster. Denote

$$\underline{Y}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{im_i} \end{pmatrix}, \ \underline{X}_i = \begin{pmatrix} X_{i1}^t \\ \vdots \\ X_{im_i}^t \end{pmatrix}, \ \text{ and } \underline{T}_i = \begin{pmatrix} T_{i1} \\ \vdots \\ T_{im_i} \end{pmatrix}.$$

The basic assumption is that observations from different clusters are independent, and that

$$E(Y_{ij}|X_{ij}, T_{ij}, \underline{X}_i, \underline{T}_i) = E(Y_{ij}|X_{ij}, T_{ij}) = \mu_{ij}; \tag{1}$$

see below for a discussion of this assumption. The marginal mean μ_{ij} depends on X_{ij} and T_{ij} through a known monotonic and differentiable link function $g(\cdot)$:

$$g(\mu_{ij}) = X_{ij}^t \beta + \theta(T_{ij}), \tag{2}$$

where β is a $p \times 1$ vector and $\theta(\cdot)$ is an unknown smooth function. We thus model the effect of $X(p \times 1)$ parametrically and the effect of T nonparametrically. Denoting

$$\underline{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \vdots \\ \mu_{im_i} \end{pmatrix}, \ g(\underline{\mu}_i) = \begin{pmatrix} g(\mu_{i1}) \\ \vdots \\ g(\mu_{im_i}) \end{pmatrix}, \ \text{ and } \theta(\underline{T}_i) = \begin{pmatrix} \theta(T_{i1}) \\ \vdots \\ \theta(T_{im_i}) \end{pmatrix},$$

(2) can be written in matrix notation as $g(\underline{\mu}_i) = \underline{X}_i \beta + \theta(\underline{T}_i)$.

Applications of marginal models for longitudinal/clustered data are common, see for example Diggle *et al.* (2002). Model (2) differs from parametric marginal GEE models (Liang & Zeger, 1986) mainly through the presence of the nonparametric component $\theta(\cdot)$. It is motivated by the fact that the effect of the covariate T (e.g., time) may be complicated and might be better modeled nonparametrically (Zeger & Diggle, 1994). Note that in our model no distributional assumptions are imposed on the data other than the moment conditions specified in (1) and (2). In particular, X and T are allowed to be dependent, as commonly seen for longitudinal/clustered data.

The first equality in (1) is called full covariate conditional mean assumption (Diggle *et al.*, 2002, Chapter 12). This assumption was also used in Wang *et al.*, (2005). Pepe & Anderson (1994) concluded that using longitudinal data to estimate a cross-sectional model requires that either this condition is verified or that working independence GEE be used. This condition is satisfied when covariates are cluster-independent (or time-independent for longitudinal data), that is, $X_{ij} = X_i$, $T_{ij} = T_i$. It is also satisfied when $X_{ij} = X_i$ are baseline covariates and T_{ij} is strictly exogenous, such as observation time irrelevant to the outcome.

Let $\Sigma_i = \Sigma_i(\underline{X}_i, \underline{T}_i)$ and $V_i = V_i(\underline{X}_i, \underline{T}_i)$ be the true and assumed "working" covariances of $\underline{Y}_i$, where $\Sigma_i = \mathrm{var}(\underline{Y}_i | \underline{X}_i, \underline{T}_i)$. Throughout, we assume that V_i can depend on a nuisance finite-dimensional parameter vector τ, which is distinct from β. Semiparametric efficient estimator will be obtained when $\Sigma_i = V_i(\tau^*)$ for some τ^*.

3 The Estimation Method

Our estimation method is based on basis approximations (Huang *et al.*, 2002). The idea is to approximate the function $\theta(\cdot)$ in (2) by a basis expansion and then employ an extension of the standard GEE. Suppose that $\theta(\cdot)$ can be approximated well by a spline function so that

$$\theta(t) \approx \sum_{k=1}^{K_n} \gamma_k^* B_k(t) = B^t(t)\gamma^* \tag{3}$$

where $\{B_k(\cdot), k = 1, \ldots, K_n\}$ is a basis system of B-splines, $\gamma^* = (\gamma_1^*, \ldots, \gamma_{K_n}^*)^t$, and $B(t) = (B_1(t), \ldots, B_{K_n}(t))^t$. When $\theta(\cdot)$ is a smooth function, such a spline approximation always exists. In fact, if $\theta(\cdot)$ is continuous, the spline approximation can be chosen to satisfy $\sup_t |\theta(t) - B^t(t)\gamma^*| \to 0$ as $K_n \to \infty$ (de Boor, 2001). Let $Z_{ij} = B(T_{ij})$. Following (2) and (3), we have an approximated model

$$g(\mu_{ij}) \approx X_{ij}^t \beta + Z_{ij}^t \gamma^*. \tag{4}$$

In matrix notation, denoting $\underline{Z}_i = (Z_{i1}, \ldots, Z_{im_i})^t$, we have $g(\underline{\mu}_i) \approx \underline{X}_i\beta + \underline{Z}_i\gamma$.

Let $\mu(\cdot) = g^{-1}(\cdot)$ be the inverse of the link function. Let $\widehat{\beta}$ and $\widehat{\gamma}$ minimize

$$\sum_{i=1}^{n} \{\underline{Y}_i - \mu(\underline{X}_i\beta + \underline{Z}_i\gamma)\}^t V_i^{-1} \{\underline{Y}_i - \mu(\underline{X}_i\beta + \underline{Z}_i\gamma)\}$$

or equivalently, they solve the estimating equations

$$\sum_{i=1}^{n} \underline{X}_i^t \Delta_i V_i^{-1} \{\underline{Y}_i - \mu(\underline{X}_i\beta + \underline{Z}_i\gamma)\} = 0 \tag{5}$$

and

$$\sum_{i=1}^{n} \underline{Z}_i^t \Delta_i V_i^{-1} \{\underline{Y}_i - \mu(\underline{X}_i\beta + \underline{Z}_i\gamma)\} = 0, \tag{6}$$

where Δ_i is a diagonal matrix with the diagonal elements being the first derivative of $\mu(\cdot)$ evaluated at $X_{ij}^t\beta + Z_{ij}^t\gamma$, $j = 1, \dots, m_i$. Then $\widehat{\beta}$ estimates the parametric part of the model and $\widehat{\theta}(\cdot) = B^t(\cdot)\widehat{\gamma}$ estimates the nonparametric part of the model.

Our estimators can be viewed as an extension of GEE. If the approximation sign in (4) is replaced by a strict equality, the derivation of our estimator follows exactly the standard GEE. Since our estimator is motivated from an approximated model, the standard theory of the parametric GEE does not carry over. One of the main purpose of this paper is to provide justification of our extension of GEE in a semiparametric context. Such a justification requires carefully taking into account the consequence of use of function approximation. In particular, we show that $\widehat{\beta}$ is asymptotically normal and, if the correct covariance structure is specified, it is semiparametric efficient. We also show that $\widehat{\theta}(\cdot)$ is a consistent estimator of the true nonparametric function $\theta(\cdot)$. Our asymptotics assume that the number of individuals/clusters becomes large while the number of observations per individual/cluster remains bounded. We focus on the identity link case in our asymptotics.

3.1 Identity Link

For general link, our extended GEE estimates can be computed using an iterative algorithm. For the identity link case, both $\widehat{\beta}$ and $\widehat{\theta}(\cdot)$ have a closed form expression. This facilitates a finite sample analysis. Denote $\underline{U}_i = (\underline{X}_i, \underline{Z}_i)$. It is easily seen that

$$\begin{pmatrix} \widehat{\beta} \\ \widehat{\gamma} \end{pmatrix} = \Big(\sum_{i=1}^n \underline{U}_i^t V_i^{-1} \underline{U}_i \Big)^{-1} \sum_{i=1}^n \underline{U}_i^t V_i^{-1} \underline{Y}_i. \tag{7}$$

Let $\mathbb{X}$ and $\mathbb{T}$ denote respectively the collections of all X_{ij}'s and all T_{ij}'s. Let $\delta(t) = \theta(t) - B^t(t)\gamma^*$. Then

$$E\Big(\begin{pmatrix} \widehat{\beta} \\ \widehat{\gamma} \end{pmatrix} \Big| \mathbb{X}, \mathbb{T} \Big) = \begin{pmatrix} \beta \\ \gamma^* \end{pmatrix} + \Big(\sum_{i=1}^n \underline{U}_i^t V_i^{-1} \underline{U}_i \Big)^{-1} \sum_{i=1}^n \underline{U}_i^t V_i^{-1} \delta(\underline{T}_i)$$

and

$$\mathrm{var}\Big(\begin{pmatrix} \widehat{\beta} \\ \widehat{\gamma} \end{pmatrix} \Big| \mathbb{X}, \mathbb{T} \Big) = \Big(\sum_{i=1}^n \underline{U}_i^t V_i^{-1} \underline{U}_i \Big)^{-1} \Big(\sum_{i=1}^n \underline{U}_i^t V_i^{-1} \Sigma_i V_i^{-1} \underline{U}_i \Big) \Big(\sum_{i=1}^n \underline{U}_i^t V_i^{-1} \underline{U}_i \Big)^{-1}. \tag{8}$$

The conditional mean and variance of $\widehat{\theta}(t) = B^t(t)\widehat{\gamma}$ follow easily from the above equations. Note when $V_i = \Sigma_i$ for all i, the right-hand side of (8) reduces to $(\sum_{i=1}^n \underline{U}_i^t V_i^{-1} \underline{U}_i)^{-1}$.

Let $\mathcal{T}$ denote the common support of T_{ij}.

Theorem 1. *Both* $\mathrm{var}(\widehat{\beta}|\mathbb{X},\mathbb{T})$ *and* $\mathrm{var}(\widehat{\theta}(t)|\mathbb{X},\mathbb{T})$, $t \in \mathcal{T}$, *are minimized when* $V_i = \Sigma_i$ *for all* i.

This theorem says that in the class of spline-based extended GEE estimators, the most efficient estimator is the one corresponding to the correctly specified covariance structure. This is a finite sample result that holds for any fixed n and any configuration of m_i. No asymptotic analysis is involved.

We now give some expressions that are useful for further analysis. Let

$$\sum_{i=1}^{n} \underline{U}_i^t V_i^{-1} \underline{U}_i = \begin{pmatrix} \sum_{i=1}^{n} \underline{X}_i^t V_i^{-1} \underline{X}_i & \sum_{i=1}^{n} \underline{X}_i^t V_i^{-1} \underline{Z}_i \\ \sum_{i=1}^{n} \underline{Z}_i^t V_i^{-1} \underline{X}_i & \sum_{i=1}^{n} \underline{Z}_i^t V_i^{-1} \underline{Z}_i \end{pmatrix} \triangleq \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}.$$

It follows from well-known block matrix forms of matrix inverse that

$$\begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}^{-1} = \begin{pmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{pmatrix} = \begin{pmatrix} H_{11\cdot 2}^{-1} & -H_{11\cdot 2}^{-1} H_{12} H_{22}^{-1} \\ -H_{22\cdot 1}^{-1} H_{21} H_{11}^{-1} & H_{22\cdot 1}^{-1} \end{pmatrix}, \tag{9}$$

where $H_{11\cdot 2} = H_{11} - H_{12} H_{22}^{-1} H_{21}$ and $H_{11\cdot 2} = H_{22} - H_{21} H_{11}^{-1} H_{12}$. Consequently,

$$\widehat{\beta} = H^{11} \Big\{ \sum_{i=1}^{n} \underline{X}_i^t V_i^{-1} \underline{Y}_i - H_{12} H_{22}^{-1} \sum_{i=1}^{n} \underline{Z}_i^t V_i^{-1} \underline{Y}_i \Big\}, \tag{10}$$

and the conditional mean and covariance matrix of $\widehat{\beta}$ are given by

$$E(\widehat{\beta}|\mathbb{X},\mathbb{T}) = \beta + H^{11} \Big\{ \sum_{i=1}^{n} \underline{X}_i^t V_i^{-1} \delta(\underline{T}_i) - H_{12} H_{22}^{-1} \sum_{i=1}^{n} \underline{Z}_i^t V_i^{-1} \delta(\underline{T}_i) \Big\} \tag{11}$$

and

$$\mathrm{var}(\widehat{\beta}|\mathbb{X},\mathbb{T}) = H^{11} \sum_{i=1}^{n} \{ (\underline{X}_i - \underline{Z}_i H_{22}^{-1} H_{21})^t V_i^{-1} \Sigma_i V_i^{-1} (\underline{X}_i - \underline{Z}_i H_{22}^{-1} H_{21}) \} H^{11}. \tag{12}$$

In particular, when the working covariance matrix equals the true covariance matrix (i.e., $V_i = \Sigma_i$), $\mathrm{var}(\widehat{\beta}|\mathbb{X},\mathbb{T}) = H^{11}$. Using the same technique, we obtain that

$$\mathrm{var}(\widehat{\gamma}|\mathbb{X},\mathbb{T}) = H^{22} \sum_{i=1}^{n} \{ (\underline{Z}_i - \underline{X}_i H_{11}^{-1} H_{12})^t V_i^{-1} \Sigma_i V_i^{-1} (\underline{Z}_i - \underline{X}_i H_{11}^{-1} H_{12}) \} H^{22}. \tag{13}$$

The closed form expression of $\mathrm{var}(\widehat{\theta}(t)|\mathbb{X},\mathbb{T})$ is then given by $\mathrm{var}(\widehat{\theta}(t)|\mathbb{X},\mathbb{T}) = B^t(t)\mathrm{var}(\widehat{\gamma}|\mathbb{X},\mathbb{T})B(t)$.

4 Asymptotic Results

This section studies the asymptotic behavior for the proposed estimator. We present our results for the identity link case only. Similar results hold for general link case, but with much more involved technical arguments (Zhang, 2004). The general link function is allowed when we derive the semiparametric efficient score and information bound in Section 4.4.

4.1 Assumptions for the Asymptotic Results

We assume the following conditions hold.

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

(C2) The random variables X_{ij} are bounded, uniformly in $i = 1, \ldots, n$, $j = 1, \ldots, m_i$. The eigenvalues of $E\{(1, X_{ij}^t)^t(1, X_{ij}^t)|T_{ij}\}$ are bounded away from 0, uniformly in $i = 1, \ldots, n$, $j = 1, \ldots, m_i$.

(C3) The eigenvalues of $\Sigma_i = \mathrm{var}(\underline{Y}_i|\underline{X}_i, \underline{T}_i)$ are bounded away from 0 and infinity, uniformly in $i = 1, \ldots, n$.

(C4) The eigenvalues of the working covariance matrices V_i are bounded away from 0 and infinity, uniformly in $i = 1, \ldots, n$.

Some remarks of the above assumptions are in order. Condition (C1) is a weak regularity condition that is commonly used in the literature. The condition on eigenvalues in (C2) is essentially a requirement that the vector $(1, X_{ij}^t)^t$ is not multicolinear. The boundedness conditions on the covariates in (C2) also implies that the eigenvalues of $E\{(1, X_{ij}^t)^t(1, X_{ij}^t)\}$ are bounded away from infinity, uniformly in $i = 1, \ldots, n$, $j = 1, \ldots, m_i$. It follows from (C3) that $\mathrm{var}(Y_{ij}|\underline{X}_i, \underline{T}_i)$ is bounded uniformly in $i = 1, \cdots, n$, $j = 1, \cdots, m_i$.

To make β estimable at the $\sqrt{n}$ rate, we need a condition that ensures X and T are not functional related. To introduce such a condition, we need some notations. Let $\underline{X}_{ik}$ denote the k-th column of the matrix $\underline{X}_i$ and denote $\varphi(\underline{T}_i) = (\varphi(T_{i1}), \ldots, \varphi(T_{im_i}))^t$. Let $\varphi_k^*(\cdot)$ be the function

$\varphi(\cdot)$ that minimizes

$$\sum_i E\big[\{\underline{X}_{ik} - \varphi(\underline{T}_i)\}^t V_i^{-1}\{\underline{X}_{ik} - \varphi(\underline{T}_i)\}\big].$$

Denote $\underline{\varphi}^*(\underline{T}_i) = (\varphi_1^*(\underline{T}_i), \ldots, \varphi_p^*(\underline{T}_i))$ and define

$$I_V \triangleq \lim_n \frac{1}{n}\sum_i E\big[\{\underline{X}_i - \underline{\varphi}^*(\underline{T}_i)\}^t V_i^{-1}\{\underline{X}_i - \underline{\varphi}^*(\underline{T}_i)\}\big].$$

When the working covariance matrices are specified to be the true covariance matrices, that is, $V_i = \Sigma_i$ for all i, $\varphi_k^*(\cdot)$ reduces to the function $\psi_k^*(\cdot)$ in the definition of the efficient score and I_V reduces to the information bound (Section 4.4). We introduce $\varphi_k^*(\cdot)$ and I_V here in order to deal with general working covariance matrices.

(C5) The matrix I_V is positive definite.

To gain some insight on this condition, consider the case that data from different clusters are iid. If the matrix I_V is singular and Condition (C4) holds, then $X_{ijk}, k = 1, \ldots, p$, and T_{ij} are functionally related; specifically, there is a non-zero vector $(c_1, \ldots, c_p)^t$ such that $\sum_k c_k\{X_{ijk} - \varphi_k^*(T_{ij})\} = 0$, a.s., for all j, and in particular, when $p = 1$, $X_{ij1} = \varphi_1^*(T_{ij})$.

We next introduce some assumptions on the properties of the spline space. Let $\mathbb{G}_n = \{\sum_{k=1}^{K_n} \gamma_k B_k(t)\}$ be a space of splines with degree d defined on the common support $\mathcal{T}$ of T_{ij}. It is assumed that the knot sequence $\{u_{n,j}\}$ satisfies that $\max_{j,j'}(u_{n,j+d+1} - u_{n,j})/(u_{n,j'+d+1} - u_{n,j'})$ is bounded uniformly in n. Let

$$\rho_n = \max\left\{\inf_{g\in\mathbb{G}_n} \|g(\cdot) - \theta(\cdot)\|_\infty, \max_{1\le k\le p}\inf_{g\in\mathbb{G}_n} \|g(\cdot) - \varphi_k^*(\cdot)\|_\infty\right\}.$$

(C6) (i) $\lim_n K_n^2 \log n/n = 0$, (ii) $\lim_n n\rho_n^4 = 0$, and (iii) $\lim_n K_n\rho_n^2 = 0$.

Assumption (C6)(i) characterizes the rate of growth of the dimension of the spline spaces relative to the sample size. Assumptions (C6) (ii) (iii) describe the requirement on the best rate of convergence that the functions $\theta(\cdot)$ and $\varphi_k^*(\cdot)$ can be approximated by functions in the spline spaces. These requirements can be quantified by smoothness conditions on $\theta(\cdot)$ and $\varphi_k^*(\cdot)$, as usually used in the literature of nonparametric or semiparametric estimation.

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. Suppose that $\theta(\cdot)$ and $\varphi_k^*(\cdot)$, $k = 1, \ldots, p$, are α-smooth. (The smoothness of $\varphi_k^*(\cdot)$, $k = 1, \ldots, p$, is implied by smoothness requirement on the joint density of $\underline{X}_i$ and $\underline{T}_i$. This is shown in some detail for the special case when $V_i = \Sigma_i$ in Remark 2 of Section 4.4. Results for general working covariance matrices follow from similar argument.) Suppose also that the degree d of the splines satisfies $d \geq \alpha - 1$. Then, by a standard result from approximation theory, $\rho_n \asymp K_n^{-\alpha}$ (Schumaker, 1981). Condition (C6) thus can be replaced by the following condition.

(C6)$'$ (i) $\theta(\cdot)$ and $\varphi_k^*(\cdot)$, $k = 1, \ldots, p$, are α-smooth, (ii) $\lim_n K_n^2 \log n/n = 0$, (iii) $\lim_n K_n^{4\alpha}/n = \infty$, and (iv) $\alpha > 1/2$.

4.2 Estimation of the Parametric Part

Denote by β_0 the true value of β. Let $\widehat{\beta}_V$ denote the estimator $\widehat{\beta}$ corresponding to working covariance matrices V_i. Let $R(\widehat{\beta}_V) = \text{var}(\widehat{\beta}_V | \mathbb{X}, \mathbb{T})$ denote the conditional covariance matrix of $\widehat{\beta}_V$. See (12) for a closed form expression of $R(\widehat{\beta}_V)$. The following result gives the asymptotic distribution of $\widehat{\beta}_V$ for general working covariance matrices.

Theorem 2. *Let* $\mathbb{I}$ *denote the* $p \times p$ *identity matrix. Then*

$$\{R(\widehat{\beta}_V)\}^{-1/2}(\widehat{\beta}_V - \beta_0) \to \text{Normal}(0, \mathbb{I}).$$

The usual robust or sandwich variance estimate corresponds to replacing Σ_i by $(\underline{y}_i - \underline{X}_i\widehat{\beta} - \underline{Z}_i\widehat{\gamma})(\underline{y}_i - \underline{X}_i\widehat{\beta} - \underline{Z}_i\widehat{\gamma})^t$ in the expression of $R(\widehat{\beta}_V)$.

According to Theorem 2, our extended GEE estimator is still consistent and is asymptotically normal when the working covariance matrix V is misspecified. It follows from Theorem 1 that $R(\widehat{\beta}_V) \geq R(\widehat{\beta}_\Sigma)$, which means that $\widehat{\beta}_\Sigma$ 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 (Liang & Zeger, 1986). However, we shall show in Section 4.5 that $\widehat{\beta}_\Sigma$ is optimal in a much stronger sense, that is, it is the most efficient among all regular estimators (see Bickel *et al.*, 1993, for the precise definition of regular estimators).

Remark 1. For simplicity, we assume in our asymptotic analysis that the working correlation parameter vector τ in V is known. It can be estimated via the method of moments using a quadratic function of Y's, just as in the application of the standard parametric GEEs (Liang & Zeger, 1986). Similar to the parametric case, as long as such an estimate of τ converges in probability to some τ^* at a $\sqrt{n}$ rate, then there is no asymptotic effect on our estimator of β due to estimation of

τ, i.e., Theorem 2 still holds. In fact, $\{R(\widehat{\beta}_{\widehat{V}})\}^{-1/2}(\widehat{\beta}_{\widehat{V}} - \beta_0) \to \text{Normal}(0, \mathbb{I})$, where $\widehat{V}$ is the estimated covariance matrix. The proof of this result is technically much involved and will be given in Appendix A.7.

4.3 Estimation of the Nonparametric Part

Our method automatically outputs a spline estimate of the nonparametric part of the model. In this subsection we study the asymptotic property of this estimate. Denote by $\theta_0(t)$ the true value of $\theta(t)$. We say that $\widehat{\theta}(\cdot)$ is a consistent estimator of $\theta(\cdot)$ if $\lim_{n\to\infty} \|\widehat{\theta}(\cdot) - \theta_0(\cdot)\|_{L_2} = 0$ hold in probability.

Theorem 3. *(i) (Global rate of convergence.)* $\|\widehat{\theta}(\cdot) - \theta_0(\cdot)\|_{L_2}^2 = O_P(\rho_n^2 + K_n/n)$. *(ii) (Asymptotic normality.) For* $t \in \mathcal{T}$, $\{\widehat{\theta}(t) - E(\widehat{\theta}(t)|\mathbb{X}, \mathbb{T})\}/\{\text{var}(\widehat{\theta}(t)|\mathbb{X}, \mathbb{T})\}^{1/2} \to N(0, 1)$ *in distribution. (iii) (Sup norm of the bias.)* $\sup_{t\in\mathcal{T}} |E(\widehat{\theta}(t)|\mathbb{X}, \mathbb{T}) - \theta_0(t)| = O_P(\rho_n)$.

The part (i) of this theorem implies that $\widehat{\theta}(\cdot)$ is consistent. The rate of convergence given here is the same as those obtained for spline regression for independent data (Huang, 1998). If $\theta(\cdot)$ is α-smooth for $\alpha > 1/2$, then $\rho_n \asymp K_n^{-\alpha}$ (Schumaker, 1981). Choosing $K_n \asymp n^{1/(1+2\alpha)}$, we obtain that $\|\widehat{\theta}(\cdot) - \theta(\cdot)\|_{L_2}^2 = O_P(n^{-2\alpha/(1+2\alpha)})$, which is the optimal rate of convergence given by Stone (1982). In particular, when $\alpha = 2$ (the commonly used second derivative assumption), the choice of $K_n \asymp n^{1/5}$ gives the the optimal rate of convergence $n^{-4/5}$.

Part (ii) of the theorem states that our spline estimate of nonparametric part is also asymptotic normal. According to Theorem 1, the asymptotic variance is smallest if the working covariance structure equals to the true structure. Part (iii) of the theorem says that the sup norm of the bias is bounded by the best possible approximation rate of $\theta_0(t)$ using splines.

4.4 Semiparametric Efficient Score and Information Bound

For semiparametric problems, efficient score and information bound provide useful benchmark for optimal asymptotic behavior of regular estimators (see Bickel *et al.*, 1993). Lin & Carroll (2001a) and Wang *et al.* (2005) derived the semiparametric efficient score for the model in consideration under the normality assumption. In this subsection we give derivation without the normality assumption.

Let $\|a\|_{L_2} = [\int_{\mathcal{T}} \{a(t)\}^2 dt]^{1/2}$ denote the L_2 norm of any square integrable function $a(t)$ on the common support $\mathcal{T}$ of T_{ij}. We denote $a \in L_2(\mathcal{T})$ if $\|a\|_{L_2} < \infty$. Recall $\underline{X}_i = (X_{i1}, \ldots, X_{im_i})^t$ and denote $\psi(\underline{T}_i) = (\psi(T_{i1}), \ldots, \psi(T_{im_i}))^t$ for $\psi(\cdot) \in L_2(\mathcal{T})$. We show in Appendix A.6 that for our semiparametric partially linear model the efficient score for β is

$$\ell^*_\beta = \sum_i \{\underline{X}_i - \underline{\psi}^*(\underline{T}_i)\}^t \Delta_i \Sigma_i^{-1} \big[\underline{Y}_i - \mu\{\underline{X}_i\beta_0 + \theta_0(\underline{T}_i)\}\big], \tag{14}$$

where β_0 and $\theta_0(\cdot)$ are the true values of β and $\theta(\cdot)$, and $\underline{\psi}^*(\underline{T}_i) = (\psi^*_1(\underline{T}_i), \ldots, \psi^*_p(\underline{T}_i))$. The function $\psi^*_k(\cdot) \in L_2(\mathcal{T})$, $k = 1, \ldots, p$, satisfies

$$\sum_i E\big[\{\underline{X}_{ik} - \psi^*_k(\underline{T}_i)\}^t \Delta_i \Sigma_i^{-1} \Delta_i \psi(\underline{T}_i)\big] = 0, \qquad \psi(\cdot) \in L_2(\mathcal{T}). \tag{15}$$

The semiparametric information bound for β is

$$nI \triangleq E(\ell^*_\beta \ell^{*t}_\beta) = \sum_i E\big[\{\underline{X}_i - \underline{\psi}^*(\underline{T}_i)\}^t \Delta_i \Sigma_i^{-1} \Delta_i \{\underline{X}_i - \underline{\psi}^*(\underline{T}_i)\}\big]. \tag{16}$$

The efficient score and information bound here are derived without the normal distribution assumption, but they coincide with those obtained by Lin & Carroll (2001a) under the multivariate normality assumption.

Denote $f_{ij}(\cdot)$ the marginal density function of T_{ij}, and $f_{ilj}(\cdot,\cdot)$ the joint density function of (T_{il}, T_{ij}). Write $\Sigma_i^{-1} = (\sigma_i^{jl})$. According to (15), ψ^*_k solves the following Fredholm integral equation of the second kind (Kress, 1989)

$$\psi^*_k(t) = q(t) + \int H(t,s)\psi^*_k(s)\, ds, \tag{17}$$

where

$$H(t,s) = \frac{\sum_i \sum_j \sum_{l \neq j} E(\Delta_{ijj}\sigma_i^{jl}\Delta_{ill} | T_{il} = s, T_{ij} = t) f_{ijl}(s,t)}{\sum_i \sum_j E(\sigma_i^{jj}\Delta_{ijj}^2 | T_{ij} = t) f_{ij}(t)}$$

and

$$q(t) = \frac{\sum_i \sum_j \sum_l E(\Delta_{ijj}\sigma_i^{jl}\Delta_{ill} X_{ijk} | T_{ij} = t) f_{ij}(t)}{\sum_i \sum_j E(\sigma_i^{jj}\Delta_{ijj}^2 | T_{ij} = t) f_{ij}(t)}.$$

In general the $\psi^*_k(\cdot)$ above does not have a close form expression. However, under the condition that $\mu(\cdot)$ is strictly monotone and Σ_i^{-1} is positive definite, it always exists and is unique. This follows easily from the fact that $\psi_k^*(\cdot)$ can be viewed as an orthogonal projection relative to an appropriate inner product. Specifically, for $\xi_1, \xi_2 \in L_2(dx \times dt)$, define an inner product as

$$\langle \xi_1, \xi_2 \rangle = E\left\{ \frac{1}{n} \sum_i \xi_1^t(\underline{X}_i, \underline{T}_i) \Delta_i \Sigma_i^{-1} \Delta_i \xi_2(\underline{X}_i, \underline{T}_i) \right\} \tag{18}$$

and the corresponding norm is denoted as $\| \cdot \|$. For $k = 1, \ldots, p$, let $x_k(\cdot)$ denote the coordinate mapping that maps x to its k-th component so that $x_k(X_{ij}) = X_{ijk}$. Then (15) implies that $\psi_k^*(\cdot)$ is the orthogonal projection of $x_k(\cdot)$ onto $L_2(\mathcal{T})$. We also have that $\psi_k^* = \arg\min_{\psi \in L_2(\mathcal{T})} \|x_k - \psi\|$.

Our derivation of efficient score and information bound also applies when T is a cluster-level covariate, that is, $T_{ij} = T_i$, $j = 1, \ldots, m_i$. We only need replace $\psi_k^*(\underline{T}_i)$ and $\underline{\psi}^*(\underline{T}_i)$ in (14)-(16) respectively by $\psi_k^*(T_i)\mathbf{1}$ and $\mathbf{1}(\psi_1^*(T_i), \ldots, \psi_p^*(T_i))$, where $\mathbf{1}$ is a m_i-vector of ones, and do similar changes for $\psi_k(\underline{T}_i)$ and $\psi(\underline{T}_i)$. It is interesting to note that in this particular case, if $(\underline{Y}_i, \underline{X}_i, T_i)$ are i.i.d., then $\psi_k^*(\cdot)$ has a closed form expression. Indeed,

$$\psi_k^*(t) = \frac{E(X_{ik}^t \Delta_i \Sigma_i^{-1} \Delta_i \mathbf{1} | T_i = t)}{E(\mathbf{1} \Delta_i \Sigma_i^{-1} \Delta_i \mathbf{1} | T_i = t)}, \qquad k = 1, \ldots, p.$$

The resulting efficient score functions coincide with that obtained in Lin & Carroll (2001b) under multivariate normality assumption.

Remark 2. In obtaining asymptotic results, it is necessary to require ψ_k^* to be smooth functions. The smoothness of ψ_k^* follows from the smoothness assumptions on the joint density $f_{X,T}(\underline{x}_i, \underline{t}_i)$ of $\underline{X}_i$ and $\underline{T}_i$. Note that the denominator in the definitions of $H(t,s)$ and $q(t)$ are strictly positive under our assumption that $\mu(\cdot)$ is strictly monotone and Σ_i^{-1} is positive definite. Suppose that $f_{X,T}(\underline{x}_i, \underline{t}_i)$ and elements of $\Sigma_i^{-1} = (\sigma_i^{jl})$ are α-smooth as functions of t_{ij}, $j = 1, \ldots, m_i$. Suppose also that $\mu^{(1)}(\cdot) \in C^{\lceil \alpha \rceil}$, where $\lceil \alpha \rceil$ denote the smallest integer bigger than or equal to α. Then, $q(t)$ and $\int H(t,s) \psi_k^*(s)\, ds$ are α-smooth in t. Therefore, it follows from (17) that ψ_k^* is α-smooth.

4.5 Achieving the Semiparametric Efficiency Bound

In this section, we show that our estimate of the parametric part of the partially linear model achieves the semiparametric efficiency bound when the covariance structure is correctly specified. For simplicity of presentation, we assume that parameter τ in the covariance matrix specification is known. The results in this section remain hold when τ can be $\sqrt{n}$-consistently estimated (see Remark 1 in Section 4.3).

Theorem 2 shows that $\{R(\widehat{\beta}_V)\}^{-1/2}(\widehat{\beta}_V - \beta_0) \to \mathrm{Normal}(0, \mathbb{I})$ for general working covariance matrices. When the working covariance matrices are specified to be the true covariance matrices, we have $R(\widehat{\beta}_\Sigma) = (n\widehat{I}_n)^{-1}$, where

$$n\widehat{I}_n = \left(\sum_{i=1}^n \underline{X}_i^t \Sigma_i^{-1} \underline{X}_i\right) - \left(\sum_{i=1}^n \underline{X}_i^t \Sigma_i^{-1} \underline{Z}_i\right)\left(\sum_{i=1}^n \underline{Z}_i^t \Sigma_i^{-1} \underline{Z}_i\right)^{-1}\left(\sum_{i=1}^n \underline{Z}_i^t \Sigma_i^{-1} \underline{X}_i\right). \tag{19}$$

To facilitate our discussion, let us introduce two inner products. For $\xi_1, \xi_2 \in L_2(dx \times dt)$, define the empirical inner product as

$$\langle \xi_1, \xi_2 \rangle_n = \frac{1}{n} \sum_i \xi_1^t(\underline{X}_i, \underline{T}_i) V_i^{-1} \xi_2(\underline{X}_i, \underline{T}_i)$$

and the theoretical inner product as $\langle \xi_1, \xi_2 \rangle = E[\langle \xi_1, \xi_2 \rangle_n]$. Corresponding norms are denoted as $\| \cdot \|_n$ and $\| \cdot \|$. When $V_i = \Sigma_i$ for all i, the theoretical inner product here is the same as that introduced in (18) with the identity link. We use the same notation for the two versions of the theoretical inner product for notational simplicity. This should not cause confusion since the definition of theoretical inner product in (18) is only used when we discuss efficient scores.

It is easy to see that φ_k^* defined in Section 4.1 satisfies $\varphi_k^* = \arg\min_{\varphi \in L_2} \|x_k - \varphi\|$ (as in the previous section). Define $\widehat{\varphi}_{k,n} = \arg\min_{\varphi \in \mathbb{G}_n} \|x_k - \varphi\|_n$. The next result says that $\widehat{\varphi}_{k,n}$ provides a consistent estimate of φ_k^*.

Theorem 4. *$\|\widehat{\varphi}_{k,n} - \varphi_k^*\|_n^2 = o_P(1)$, $k = 1, \ldots, p$.*

This result has an important implication. Since φ_k^* becomes the function ψ_k^* in the definition of efficient score (14), Theorem 4 suggests us to estimate the efficient score by replacing ψ_k^* in (14) with

$$\widehat{\psi}_{k,n} = \mathrm{argmin}_{\psi \in \mathbb{G}_n} \frac{1}{n} \sum_i \{\underline{X}_{ik} - \psi(\underline{T}_i)\}^t \Sigma_i^{-1} \{\underline{X}_{ik} - \psi(\underline{T}_i)\}.$$

It also suggests an estimate of the efficient information as explained below.

It is easily seen that the efficient information I defined in (16) has (k, k')-entry $\langle x_k - \psi_k^*, x_{k'} - \psi_{k'}^* \rangle$ where $V_i = \Sigma_i$ for all i in the definition of theoretical inner product. Some calculation reveals that $\widehat{I}_n$ defined in (19) has (k, k')-entry $\langle x_k - \widehat{\psi}_{k,n}, x_{k'} - \widehat{\psi}_{k',n} \rangle_n$, where $V_i = \Sigma_i$ for all i in the definition of the empirical inner product. Theorem 4 together with the triangle inequality implies that

$$\widehat{I}_n(k, k') = \langle x_k - \psi_k^*, x_{k'} - \psi_{k'}^* \rangle_n + o_P(1) = I(k, k') + o_P(1),$$

that is, $\widehat{I}_n$ is a consistent estimate of the efficient information I for β. In light of this result, $\widehat{I}_n$ will be called the estimated information. Note the consistency result given here holds when Σ_i is replaced by a general covariance matrix V_i in the definition of I and $\widehat{I}_n$.

Let $\widehat{\beta}_\Sigma$ denote $\widehat{\beta}$ when $V_i = \Sigma_i$, $i = 1, \ldots, n$. The above discussion together with Theorems 2 immediately imply the following result.

Corollary 1. *The estimator $\widehat{\beta}_\Sigma$ is asymptotically normal and achieves the semiparametric efficiency bound, that is,*

$$(nI)^{1/2}(\widehat{\beta}_\Sigma - \beta_0) \to \text{Normal}(0, \mathbb{I}).$$

The efficient information I can be consistently estimate by $\widehat{I}_n$.

When $(\underline{Y}_,\underline{X}_i, \underline{T}_i)$ are i.i.d., $I = E[\{\underline{X}_i - \varphi^*(\underline{T}_i)\}^t \Delta_i \Sigma_i^{-1} \Delta_i \{\underline{X}_i - \varphi^*(\underline{T}_i)\}]$ does not depend on n. The result in Corollary 1 can also be written as $n^{1/2}(\widehat{\beta}_\Sigma - \beta_0) \to \text{Normal}(0, I^{-1})$.

4.6 Discussion

Choice of K_n. Our asymptotic results are quite insensitive to the choice of the number of terms K_n in the basis expansion, which plays the role of a smoothing parameter. Specifically, suppose that $\theta(\cdot)$ and $\varphi_k^*(\cdot)$, $k = 1, \ldots, p$, have bounded second derivatives (corresponding to $\alpha = 2$ in Condition (C6)$'$). Let $a_n \ll b_n$ (or $b_n \gg a_n$) means $\lim_n a_n/b_n = 0$. Then the requirement on K_n reduces to $n^{1/8} \ll K_n \ll \{n/(\log n)\}^{1/2}$. This is a wide range, including in particular $K_n \asymp n^{1/5}$ that corresponds to the optimal rate of convergence $n^{-4/5}$. It follows that undersmoothing (i.e. $K_n \gg n^{1/5}$) is not necessary for semiparametric efficient estimation using splines, where undersmoothing means that one uses more knots than what is needed to achieve the optimal rate of convergence. The result also suggests that precise determination of K_n is not of particular concern when applying our asymptotic results. It is advisable to use the usual data driven methods such as delete-subject-out cross-validation (see, for example, Section 2.3 of Huang *et al.*, 2004) to select K_n and then check the sensitivity of the results.

Smoothness conditions. Our asymptotic results indicate that semiparametric efficient estimator of β exists under smoothness conditions weaker than the usual bounded second derivative condition used in the literature (Lin & Carroll 2001ab, Wang *et al.* 2005). Indeed, suppose $\theta(\cdot)$

and $\varphi_k^*(\cdot)$, $k=1,\ldots,p$, are α-smooth for $\alpha > 1/2$. Condition (C6) reduces to $n^{1/(4\alpha)} \ll K_n \ll \{n/(\log n)\}^{1/2}$ and $\alpha > 1/2$. Therefore, as long as $\alpha > 1/2$, there is an K_n to fulfill the requirements in (C6) and Corollary 1 shows the existence of semiparametric efficient estimator of β for such a choice of K_n.

Cluster-level covariates. When T_i is a cluster-level covariate, that is, $T_{ij} = T_i$, $j = 1,\ldots,m_i$, Theorems 2–4 and Corollary 1 still hold. We only need replace Condition (C1) by the following condition.

(C1)$'$ The random variables T_i are bounded, uniformly in $i = 1,\ldots,n$. The distribution of T_i has a density $f_i(\cdot)$ with respect to the Lebesgue measure. We assume that $f_i(\cdot)$ is bounded away from 0 and ∞ on its support, uniformly in $i = 1,\ldots,n$.

Specification of covariance structure. Specifying the correct covariance structure for longitudinal data is a difficult task. This difficulty is exactly the motivation of using working covariance structure in GEE. Our results are in parallel to those for parametric GEE: If the covariance structure is correctly specified up to an $\sqrt{n}$-consistent estimable parameter, we get the most efficient estimate, otherwise we can only get an asymptotic normal estimator with correct inference using the sandwich estimate of variance. Techniques for specifying parametric covariance structure have been studied extensively in the literature (Diggle *et al.*, 2002); nonparametric techniques have just started to emerging (Wu & Pourahmadi, 2003, Huang *et al.*, 2006, 2007).

5 Numerical Results

5.1 Simulation

In this section, we report some results from a simulation study to show the finite sample performance of the proposed method. For the first example, we simulated 250 data sets from the model

$$Y_{ij} = X_{ij1}\beta_1 + X_{ij2}\beta_2 + \theta(T_{ij}) + \epsilon_{ij}, \qquad j = 1,\ldots,4,\ i = 1,\ldots,100,$$

where $\beta_1 = \beta_2 = 1$ and $\theta(t) = \sin(8t)$. The covariates T_{ij} were generated as uniform $[0,1]$ random variables and $X_{ij1} = T_{ij} + \delta_{ij}$ with δ_{ij} being $N(0,1)$ random variables. Mimicking a treatment indicator, the covariates X_{ij2} were generated as Bernoulli random variables with probability of success 0.5. The errors ϵ_{ij} follow a multivariate normal distribution with mean 0, marginal variance

1 and exchangeable correlation $\rho = 0.5$. For each simulated data set, the spline-based estimators were calculated with both a working independence (WI) and an exchangeable (EX) correlation structure. The exchangeable correlation parameter ρ was estimated using the method of moments. Cubic splines were used with the number of knots chosen from the range 0–10 by the five-fold delete-subject-out cross-validation.

The simulation results for comparing two estimators are summarized in Table 1. For both regression parameters, the estimator accounting for the correlation properly is more efficient than the estimator using working independence correlation structure. The average MSEs (mean squared errors) of the latter is more than 1.5 times that of the former for both $\widehat{\beta}_1$ and $\widehat{\beta}_2$. The variance is a dominating factor when comparing the MSEs between the two estimators. The difference of bias between the two estimators is negligible; the paired t-test gives a p-value 0.7726 for $\widehat{\beta}_1$ and 0.7730 for $\widehat{\beta}_2$. We also observed that the sandwich estimated SEs work reasonably well; the averages of the sandwich estimated SEs are close to the sample standard derivations. The mean integrated squared error (MISE), calculated using 100 grip points over $[0, 1]$, for estimating $\theta(\cdot)$ is also given in Table 1. The spline method using the exchangeable covariance structure has smaller MISE than that using working independence. Normal Q-Q plots of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ also indicate that the distributions of the estimates are close to normal. These empirical results agree nicely with the theory.

Place Table 1 around here.

Table 1: *Summary of simulation results for comparing the estimators using a working independence (WI) and an exchangeable (EX) correlation structure, based on 250 replications. Cubic splines are used with the number of knots chosen from the range 0–10 by the five-fold delete-subject-out cross-validation. Each entry equals the original value multiplied by 10.*

		$\beta_1 = 1$			$\beta_2 = 1$		$\theta(\cdot)$
Method	Bias	SD	MSE	Bias	SD	MSE	MISE
WI	-.0370	.4941	.0319	-.0090	.9937	.1089	.3114
EX	.0443	.3907	.0181	.0033	.7876	.0677	.2276

To understand how insensitive the two estimators, corresponding to a working independence and an exchangeable correlation structure, are to the choice of the number of knots, we did the following. Both estimators were computed for the number of knots being in the range of 1–10, for each of the 250 simulated data sets. For both of β_0 and β_1, we calculated the correlation coefficient of the parameter estimates corresponding to each pair of choices of the number of knots. The pairwise correlations are all above 0.98, showing strong evidence that the estimates are not sensitive to the choice of the number of knots. The reason for using the range 1–10 here is because, any number of knots within this range was selected by the five-fold delete-subject-out cross-validation for certain data set and 0 was never selected for any data set.

In the second example, we use exactly the same setup as used in Wang *et al.* (2005). This setup differs from the setup in our first example only in the way how X_{ij1} and T_{ij} are generated, $\rho = 0.6$ and $\theta(t) = \sin(2t)$; the rest are all the same. The covariates T_{ij} and X_{ij1} are generated as sums of independent uniform $[-1, 1]$ random variables and a common uniform $[0, 1]$ random variables. The two covariates so generated are time-varying and they are also correlated with each other. Summary of the results is given in Table 2. The results for the kernel method are adapted from Wang *et al.* (2005). The spline method is implemented in the same way as in the first example. It is clear that the results using splines are comparable to those using kernels, which is not surprising since the two methods should give asymptotically equivalent results, except that for estimation of β_2, the kernel method with the correctly specified covariance structure does not improve much over the kernel method with working independence structure. We think this departure from the theory is a finite sample phenomenon and may be related to the actual implementation of the iterative kernel method, which in turn could be improved by fine tuning of the computational algorithm.

Place Table 2 around here.

5.2 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 (1994). This data set contains 2376 observations of CD4+ cell counts on 369 men infected with the HIV virus. See Zeger and Diggle (1994) for more detailed description of the data. Following Wang *et al.* (2005), we fit a partially linear model with the square root-transform CD4 counts as response, covariates entering

Table 2: *Summary of simulation results for the example in Wang et al. (2005) from 250 replications. Each entry equals the original value multiplied by 10. Working independence (WI) and the true exchangeable covariance (EX) structures are used in combination with Wang et al.'s iterative kernel (kernel) and proposed spline methods (spline).*

	$\beta_1 = 1$			$\beta_2 = 1$		
Method	Bias	SD	ave MSE	Bias	SD	ave MSE
WI (kernel)	.0732	.8564	.0739	.0135	1.6486	.2718
EX (kernel)	.0118	.5675	.0322	.0107	1.6324	.2665
WI (spline)	.0413	.8506	.0722	-.0974	1.0322	.1071
EX (spline)	-.0173	.5852	.0341	-.0786	.6982	.0492

the model linearly including age, 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 effect of time since seroconversion being modeled nonparametrically.

We used a working covariance structure called by Zeger and Diggle (1994) as "random intercept plus serial correlation and measurement error" (we shall abbreviate it as RSM). Such a structure can be obtained by fitting a saturated model to the data and carefully inspecting the variogram of the residuals. See Chapter 5 of Diggle *et al.* (2002) for detailed description of how to specify a suitable parametric model of covariance structure for longitudinal data. The working covariance matrices we used have a form of $\tau^2 I + \nu^2 J + \omega^2 H$, where I is an identity matrix, J is a matrix of 1's and $H(j,k) = \exp(-\alpha|T_{ij} - T_{ik}|)$. We used the covariance parameters by Wang *et al.* (2005), $(\widehat{\tau}^2, \widehat{\nu}^2, \widehat{\omega}^2, \widehat{\alpha}^2) = (11.32, 3.26, 22.15, .23)$, which were obtained by leaving out residuals in the boundary and coupling a least squares method in variogram analysis and a moment variance estimation.

Place Table 3 around here.

Table 3 gives the regression coefficient estimates of the parametric covariates using both WI and RSM covariance structures. The results for the kernel method are adapted from Wang *et*

Table 3: *Regression coefficients in the CD4 cell counts study in HIV seroconverters using the kernel and spline estimates. Working covariance structure used are working independence (WI) and "random intercept plus serial correlation plus measurement error" (RSM).*

	Kernel				Spline			
	WI		RSM		WI		RSM	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Age	.014	.035	.008	.032	.015	.034	.008	.033
Smoking	.984	.182	.579	.139	.971	.182	.629	.140
Drug	1.049	.526	.584	.335	1.102	.528	.717	.334
Sex partners	-.054	.059	.078	.039	-.069	.059	.025	.038
Depression	-.033	.021	-.046	.014	-.033	.021	-.041	.014

al. (2006). Cubic splines were used for the proposed method and reported results correspond to the number of knots selected by the five-fold delete-subject-out cross-validation from the range of 0–20. For a given working covariance structure, the kernel and spline methods give comparable results. That estimates using the RSM structure have smaller SE than those using the WI structure is in accordance with our theory. Note some fairly large numerical differences between estimates using different working covariance structures. Such discrepancy may occur because of randomness as explained in Wang *et al.* (2005). The selected numbers of knots are 9 and 15 for WI and RSM working covariance structure respectively. The spline fit of the nonparametric function using the selected number of knots is noisier than the kernel fit reported in Wang *et al.* (2005), but the regression coefficient estimates of the parametric covariates are not sensitive to choice of the number of knots of splines.

6 Concluding Remarks

This paper can be viewed as an application of the general approach of functional modeling based on spline approximations. Naive application of the traditional local polynomial kernel approach has difficulty in effectively dealing with the intrinsic structure of the partially linear model for longitudinal/clustered data and has failed to construct methods that can account for within-class correlation structure and produce a semiparametric efficient estimator. We have shown that the spline method, as a global smoothing method, provides a simple solution to the problem, due to its close connection to the parametric approach. This solution has attractive statistical and numerical properties. It can deliver a semiparametric efficient estimator of the parametric part and a consistent (smooth) estimator of the nonparametric part of the model. It is also more straightforward to implement than the computationally complex iterative kernel method.

Our spline-based method can be extended easily to deal with generalizations of partially linear models. For example, the marginal mean can be modeled as

$$g(\mu_{ij}) = X_{ij}^t\beta + \theta_1(T_{ij1}) + \cdots + \theta_q(T_{ijq}),$$

which yields a partially linear additive model. Such a model allows more covariates to be modeled nonparametrically. Alternatively, a partially linear time-varying coefficient model specifies the marginal mean of a longitudinal outcome as

$$g(\mu_{ij}) = X_{ij}^t\beta + X_{ij,p+1}\theta_1(T_{ij}) + \cdots + X_{ij,p+q}\theta_q(T_{ij}).$$

This is a special case of the time-varying coefficient model of Hoover *et al.* (1998) which restricts some coefficients to be not time varying. Such restrictions make the relevant coefficients estimable at the parametric $\sqrt{n}$ rate. An intuitively appealing approach to fit these models would be to approximate the nonparametric functions in the model by splines, and then construct estimating equations as in this paper. Of course, rigorous theoretical justification of this approach requires further work.

Splines are typically defined as piecewise polynomials with global smoothness requirements. For example, cubic splines are piecewise cubic polynomials that globally have continuous second derivatives. Non-smooth piecewise polynomial approximations have been applied for nonparametric and semiparametric models for longitudinal data by Carroll *et al.* (2003) and Chen & Jin (2006). It would be interesting to point out that non-smooth piecewise polynomials without global smoothness requirement can also be viewed as splines by allowing multiple knots at the same location and they have B-spline representations (de Boor, 2001). Therefore, the theoretical results developed in

Huang *et al.* (2004) and in this paper are general enough to cover methods based on non-smooth piecewise polynomial approximations. Development of such general theory is important because use of (smooth) spline approximations is favorable from the practical viewpoint. Methods employing piecewise polynomial approximations usually require a second stage kernel smoothing to get a smooth fit of the nonparametric function. Application of spline approximations can yield a smooth nonparametric fit automatically in a one-stage procedure.

Acknowledgement

Jianhua Z. Huang's work is partially supported by grants from the US National Science Foundation. We would like to thank Dylan Small and Chuck Stone for helpful comments.

References

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* **71**, 1795–1843.

de Boor, C. (2001). *A practical guide to splines*, Revised Edition. Springer, New York.

Bickel, P. J. & Doksum, K. A. (2001). *Mathematical statistics: Basic ideas and selected topics*, 2nd Ed. Prentice Hall, Upper Saddle River, Jew Jersey.

Bickel, P. J., Klaassen, A. J., Ritov, Y. & Wellner, J. A. (1993). *Efficient and adaptive inference in semiparametric models.* Johns Hopkins University Press, Baltimore.

Brumback, B. A. & Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J. Amer. Statist. Assoc.* **93**, 961–976.

Carroll, R. J., Hall, P., Apanasovich, T. V. & Lin, X. (2004). Histospline method in nonparametric regression models with application to clustered/longitudinal data. *Statist. Sinica* **14**, 633–658.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics* **34**, 305–334.

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

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

Chiang, C. T., Rice, J. A. & Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Amer. Statist. Assoc.* **96**, 605–619.

Diggle, P. J., Heagerty, P., Liang, K.-Y. & Zeger, S. L. (2002). *Analysis of longitudinal data.*, 2nd Ed. Oxford University Press, Oxford, England.

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

Fan, J. & Zhang, J.-T. (2000). Functional linear models for longitudinal data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62**, 303–322.

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

Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822.

Huang, J. Z. (1998). Projection estimation in multiple regression with applications to functional ANOVA models. *Ann. Statist.* **26**, 242–272.

Huang, J. Z. (2001). Concave extended linear modeling: A theoretical synthesis. *Statist. Sinica* **11**, 173–197.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31**, 1600–1635.

Huang, J. Z., Wu, C. O. & Zhou, L. (2002). Varying coefficient models and basis function approximation for the analysis of repeated measurements. *Biometrika* **89**, 111–128.

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

Huang, J. Z., Liu, N., Pourahmadi, M., & Liu, L. (2006). Covariance selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85–98.

Huang, J. Z., Liu, L., & Liu, N. (2007). Estimation of large covariance matrices of longitudinal data with basis function approximations. *J. Comput. Graph. statist.*, to appear.

Kress, R. (1999). *Linear integral equations,* 2nd Ed. Springer, New York.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lin, Z. & Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Amer. Statist. Assoc.* **95**, 520–534.

Lin, X. & Carroll, R. J. (2001a). Semiparametric regression for clustered data using generalized estimating equations. *J. Amer. Statist. Assoc.* **96**, 1045–1056.

Lin, X. & Carroll, R. J. (2001b). Semiparametric regression for clustered data. *Biometrika* **88**, 1179–1185.

Lin, D. Y. & Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc.*, **96**, 103–112.

Martinussen, T. & Scheike, T. H. (2001). Sampling adjusted analysis of dynamic additive regression models for longitudinal data. *Scand. J. Statist.* **28**, 303–323.

Martinussen, T., Scheike, T. H. & Skovgaard, IB M. (2002). Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. *Scand. J. Statist.* **29**, 57–74.

Pepe, M. S. & Anderson, G.A. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communication in Statistics – Simulation* **23**, 939–51.

Schumaker, L. L. (1981). *Spline functions: Basic theory*, New York: Wiley.

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

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

Staniswalis, J. G. & Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc.* **93**, 1403–1418.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1348–1360.

Sun, Y. & Wu, H. (2005). Semiparametric time-varying coefficients regression model for Longitudinal Data. *Scand. J. Statist.* **32**, 21-47.

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

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

Wu, C. O., Chiang, C. -T. & Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Amer. Statist. Assoc.*, **93**, 1388–1402.

Wu, H. & Liang, H. (2004). Backfitting random varying-coefficient models with time-dependent smoothing covariates. *Scand. J. Statist.* **31**, 3–19.

Wu, H. & Zhang, J.T. (2002). Local polynomial mixed-effects models for longitudinal data. *J. Amer. Statist. Assoc.* **97**, 883-897.

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

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

Zhang, L. (2004). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. Ph.D. Thesis, Department of Statistics, University of Pennsylvania.

Jianhua Z. Huang, Department of Statistics, 447 Blocker Building, Texas A&M University, College Station, TX 77843-3143, U.S.A.
Email: jianhua@stat.tamu.edu

Appendix

A.1 Proof of Theorem 1

The generalized Cauchy–Schwarz inequality states that, if $((\Sigma_{11}, \Sigma_{12})^t, (\Sigma_{21}, \Sigma_{22})^t)^t$ is symmetric positive semidefinite, then $\Sigma_{11} \geq \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ (see Section B.10.2.2 of Bickel and Doksum, 2001). Now

$$\begin{pmatrix} \sum_{i=1}^n U_i^t V_i^{-1}\Sigma_i V_i^{-1} U_i & \sum_{i=1}^n U_i^t V_i^{-1} U_i \\ \sum_{i=1}^n U_i^t V_i^{-1} U_i & \sum_{i=1}^n U_i^t \Sigma_i^{-1} U_i \end{pmatrix} = \sum_{i=1}^n U_i^t \begin{pmatrix} V_i^{-1} \\ \Sigma_i^{-1} \end{pmatrix} \Sigma_i \left(V_i^{-1} \Sigma_i^{-1} \right) U_i \geq 0.$$

Thus, an application of the generalized Cauchy–Schwarz inequality yields

$$\sum_{i=1}^n U_i^t V_i^{-1}\Sigma_i V_i^{-1} U_i \geq \left(\sum_{i=1}^n U_i^t V_i^{-1} U_i\right)\left(\sum_{i=1}^n U_i^t \Sigma_i^{-1} U_i\right)^{-1}\left(\sum_{i=1}^n U_i^t V_i^{-1} U_i\right),$$

which in turn yields

$$\Big(\sum_{i=1}^{n} U_i^t V_i^{-1} U_i\Big)^{-1} \sum_{i=1}^{n} U_i^t V_i^{-1} \Sigma_i V_i^{-1} U_i \Big(\sum_{i=1}^{n} U_i^t V_i^{-1} U_i\Big)^{-1} \geq \Big(\sum_{i=1}^{n} U_i^t \Sigma_i^{-1} U_i\Big).$$

The left hand side of the above equation corresponds to $\mathrm{var}((\widehat{\beta}, \widehat{\gamma})^T | \mathbb{X}, \mathbb{T}))$ (see (8)), and the right hand side corresponds to the special case of the same quantity when $V_i = \Sigma_i$ for all i. The desired results follows.

A.2 Some Usfeul Lemmas

We shall employ the machinery developed in Huang (1998, 2003) and Huang *et al.* (2004) to prove our asymptotic results. Following these papers, it is convenient to take a geometric viewpoint. The empirical and theoretical inner products introduced in Section 4.5 play an important role in our technical arguments. Let $\widehat{\Pi}_n$, Π_n denote respectively the projection onto $\mathbb{G}_n$ relative to the empirical and the theoretical inner products (referred to as the empirical and theoretical projections).

Lemma A.1. *There are constants $M_1, M_2 > 0$ such that $M_1 \|g\|_{L_2} \leq \|g\| \leq M_2 \|g\|_{L_2}$ for all $g \in \mathbb{G}_n$.*

Let $x_0 \equiv 0$. For $k = 1, \ldots, p$, let $x_k(\cdot)$ denote the coordinate mapping that maps x to its k-th component so that $x_k(X_{ij}) = X_{ijk}$.

Lemma A.2. *Suppose $\lim_n K_n^2 \log n / n = 0$. Then*

$$\sup_{g \in \mathbb{G}_n} \left| \frac{\|x_k - g\|_n^2}{\|x_k - g\|^2} - 1 \right| = O_P\left(\sqrt{\frac{K_n^2 \log n}{n}}\right), \qquad k = 1, \ldots, p.$$

Lemma A.3.

$$\sup_{g \in \mathbb{G}_n} \frac{|\langle h_n, g\rangle_n - \langle h_n, g\rangle|}{\|g\|} = O_P\left(\sqrt{\frac{K_n}{n}}\right) \|h_n\|.$$

Lemma A.1 follows easily from Conditions (C1) and (C4). Lemma A.2 can be proved as Lemma A.2 of Huang *et al.* (2004). Lemma A.3 follows from the same argument as in the proof of Lemma 11 in Huang (1998), but using the B-spline basis in the argument.

Notation. For positive numbers a_n and b_n for $n \geq 1$, let $a_n \lesssim b_n$ (or $b_n \gtrsim a_n$) mean that a_n / b_n is bounded and let $a_n \asymp b_n$ mean that $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

A.3 Proof of Theorem 2

We first show, by using the Cramér-Wold device and checking the Linderberg condition, that $\{R(\widehat{\beta}_V)\}^{1/2}\{\widehat{\beta} - E(\widehat{\beta}|\mathbb{X},\mathbb{T})\} \to N(0,\mathbb{I})$. By (10),

$$\widehat{\beta} - E(\widehat{\beta}|\mathbb{X},\mathbb{T}) = H^{11}\bigg(\sum_{i=1}^{n} \underline{X}_i^t V_i^{-1}\underline{e}_i - H_{12}H_{22}^{-1}\sum_{i=1}^{n} \underline{Z}_i^t V_i^{-1}\underline{e}_i\bigg),$$

where $\underline{e}_i = \underline{Y}_i - \underline{X}^t\beta - \theta(\underline{T}_i)$. Thus, for $c \in \mathbb{R}^p$ with $|c| = 1$, we can write

$$c^t\{\widehat{\beta} - E(\widehat{\beta}|\mathbb{X},\mathbb{T})\} = \sum_{i=1}^{n} a_i\epsilon_i,$$

where,

$$a_i^2 = c^t H^{11}(\underline{X}_i - \underline{Z}_i H_{22}^{-1}H_{21})^t V_i^{-1}\Sigma_i V_i^{-1}(\underline{X}_i - \underline{Z}_i H_{22}^{-1}H_{21})H^{11}c,$$

and conditioning on $(\mathbb{X},\mathbb{T})$, ϵ_i are independent with mean 0 and variance 1. It follows from Theorem 4 that $\widehat{I}_n \to^P I$, where, abusing notation slightly, we replaced V_i by Σ_i in the definition of I and $\widehat{I}_n$. Therefore $(nH^{11})^{-1} = \widehat{I}_n = I + o_P(1)$ and thus the eigenvalues of H^{11} are of order $1/n$. Then

$$\max_{1\le i\le m_i} a_i^2 \le |c|^2 \max_{1\le i\le n}\max_{1\le j\le m_i} |X_{ij} - H_{12}H_{22}^{-1}Z_{ij}|^2 O\Big(\frac{1}{n^2}\Big).$$

Note that $\|g\|_\infty \lesssim K_n^{1/2}\|g\|$ for $g \in \mathbb{G}$. The k-th component of $H_{12}H_{22}^{-1}Z_{ij}$, which is $(\widehat{\Pi}_n x_k)(T_{ij})$, is less than

$$\|\widehat{\Pi}_n x_k\|_\infty \le K_n^{1/2}\|\widehat{\Pi}_n x_k\| = K_n^{1/2}\|\widehat{\Pi}_n x_k\|_n(1 + o_P(1)) = O_P(K_n^{1/2}).$$

Thus, $\max_{1\le i\le m_i} a_i^2 = O_P(K_n/n^2)$. On the other hand, $\sum_{i=1}^n a_i^2 = \mathrm{var}(c^t\widehat{\beta}|\mathbb{X},\mathbb{T}) \gtrsim c^t H^{11}c \gtrsim 1/n$. Hence, $\max_{1\le i\le n} a_i^2/\sum_{i=1}^n a_i^2 = O_P(K_n/n) = o_P(1)$. Then, it follows by checking the Lindergerg condition that, $\sum_{i=1}^n a_i\epsilon_i/\{\sum_{i=1}^n a_i^2\}^{1/2} \to N(0,1)$, which is the desired result.

We next show that $|E(\widehat{\beta}|\mathbb{X},\mathbb{T}) - \beta| = o_P(1/\sqrt{n})$. It follows from (11) that

$$\begin{aligned} E(\widehat{\beta}|\mathbb{X},\mathbb{T}) - \beta &= H^{11}\bigg[\sum_{i=1}^{n} \underline{X}_i^t V_i^{-1}\bigg\{\delta(\underline{T}_i) - \underline{Z}_i H_{22}^{-1}\sum_{i=1}^{n} \underline{Z}_i^t V_i^{-1}\delta(\underline{T}_i)\bigg\}\bigg] \\ &= H^{11}\bigg[\sum_{i=1}^{n} \underline{X}_i^t V_i^{-1}\Big\{\delta(\underline{T}_i) - (\widehat{\Pi}_n\delta)(\underline{T}_i)\Big\}\bigg] \triangleq nH^{11}S, \end{aligned}$$

where $S = (S_1,\cdots,S_p)^t$ and $S_k = \langle x_k, \delta - \widehat{\Pi}_n\delta\rangle_n = \langle x_k - \widetilde{\varphi}_{k,n}, \delta - \widehat{\Pi}_n\delta\rangle_n$ for any $\widetilde{\varphi}_{k,n} \in \mathbb{G}_n$. Let $\varphi_k^* = \arg\min_{\varphi\in L_2}\|x_k - \varphi\|$. In the following, we pick $\widetilde{\varphi}_{k,n}$ such that it satisfies $\|\widetilde{\varphi}_{k,n} - \varphi_k^*\|_\infty \lesssim \rho_n$.

Consider the decomposition

$$S_k = \langle x_k - \varphi_k^*, \delta - \Pi_n \delta \rangle_n + \langle x_k - \varphi_k^*, \Pi_n \delta - \widehat{\Pi}_n \delta \rangle_n + \langle \varphi_k^* - \widetilde{\varphi}_{k,n}, \delta - \widehat{\Pi}_n \delta \rangle_n \triangleq S_{1k} + S_{2k} + S_{3k}.$$

Because φ_k^* satisfies $\langle x_k - \varphi_k^*, \varphi \rangle = 0$ for $\varphi \in L_2$, $E(S_{1k}) = 0$. Note that $\|\delta - \Pi_n \delta\|_\infty \leq \|\delta\|_\infty + K_n^{1/2} \|\Pi_n \delta\| \lesssim K_n^{1/2} \rho_n$. Thus, since the eigenvalues of V_i are bounded away from 0 and infinity, we obtain that

$$\text{var}(S_{1k}) = \frac{1}{n^2} \sum_{i=1}^n E\Big[\{\underline{X}_{i,k} - \varphi_k^*(\underline{T}_i)\}^t V_i^{-1} \{\delta(\underline{T}_i) - (\Pi_n \delta)(\underline{T}_i)\} \Big]^2 \lesssim \frac{K_n \rho_n^2}{n} \|x_k - \varphi_k^*\|^2,$$

which together with $E(S_{1k}) = 0$ implies that $|S_{1k}| = O_P((K_n \rho_n^2/n)^{1/2})$. Note that $\|\delta - \Pi_n \delta\| = O(\rho_n)$. It follows from Lemmas A.2 and A.3 that

$$\|\widehat{\Pi}_n \delta - \Pi_n \delta\|_n = \sup_{g \in \mathbb{G}_n} \frac{|\langle \delta - \Pi_n \delta, g \rangle_n - \langle \delta - \Pi_n \delta, g \rangle|}{\|g\|_n} = O_P\bigg(\sqrt{\frac{K_n \rho_n^2}{n}} \bigg),$$

and therefore, $|S_{2k}| = O_P((K_n \rho_n^2/n)^{1/2})$. Moreover,

$$|S_{3k}| \leq \|\varphi_k^* - \widetilde{\varphi}_{k,n}\|_n \|\delta - \widehat{\Pi}_n \delta\|_n \leq \|\varphi_k^* - \widetilde{\varphi}_{k,n}\|_n \|\delta\|_n = O_P(\rho_n^2).$$

Note that $nH^{11} = O_P(1)$. Put all the above things together, we obtain that $|E(\widehat{\beta}|\mathbb{X}, \mathbb{T}) - \beta| = o_P(1/\sqrt{n})$ provided $n\rho_n^4 \to 0$ and $K_n \rho_n^2 \to 0$.

A.4 Proof of Theorems 3

The theorem can be proved using similar arguments as the proof of Theorems 2–4 of Huang *et al.* (2004). The details are omitted.

A.5 Proof of Theorem 4

Define $\varphi_{k,n}^* = \arg\min_{\varphi \in \mathbb{G}_n} \|x_k - \varphi\|$. We have that

$$\|\widehat{\varphi}_{k,n} - \varphi_k^*\|_n \leq \|\varphi_{k,n}^* - \varphi_k^*\|_n + \|\widehat{\varphi}_{k,n} - \varphi_{k,n}^*\|_n.$$

We inspect separately the sizes of the two terms on the right side of the above inequality. First note that $\varphi_{k,n}^* = \Pi_n \varphi_k^*$. Thus $\|\varphi_{k,n}^* - \varphi_k^*\| = \inf_{g \in \mathbb{G}} \|g - \varphi_k^*\| \asymp \inf_{g \in \mathbb{G}} \|g - \varphi_k^*\|_{L_2} = O(\rho_n) = o(1)$. Since $E(\|\varphi_{k,n}^* - \varphi_k^*\|_n) = \|\varphi_{k,n}^* - \varphi_k^*\|$, we have that $\|\varphi_{k,n}^* - \varphi_k^*\|_n = O_P(\rho_n) = o_P(1)$. On the other hand, since $\varphi^*_{k,n} = \Pi_n x_k$ and $\widehat{\varphi}_{k,n} = \widehat{\Pi}_n x_k$, we have $\|\widehat{\varphi}_{k,n} - \varphi^*_{k,n}\|^2 = \|x_k - \widehat{\varphi}_{k,n}\|^2 - \|x_k - \varphi^*_{k,n}\|^2$ and $\|x_k - \widehat{\varphi}_{k,n}\|_n^2 \le \|x_k - \varphi^*_{k,n}\|_n^2$. These two relations and Lemma A.2 imply that

$$\|\widehat{\varphi}_{k,n} - \varphi^*_{k,n}\|^2 = o_P(\|x_k - \widehat{\varphi}_{k,n}\|^2) + o_P(\|x_k - \varphi^*_{k,n}\|^2). \tag{A.1}$$

It follows from the triangle inequality and the fact that $\|x_k - \varphi^*_{k,n}\|$ is bounded that

$$\|x_k - \widehat{\varphi}_{k,n}\| \le \|x_k - \varphi^*_{k,n}\| + \|\widehat{\varphi}_{k,n} - \varphi^*_{k,n}\| \lesssim 1 + \|\widehat{\varphi}_{k,n} - \varphi^*_{k,n}\|. \tag{A.2}$$

By (A.1) and (A.2), we obtain that $\|\widehat{\varphi}_{k,n} - \varphi^*_{k,n}\|^2 = o_P(1)$, which by Lemma A.2 implies $\|\widehat{\varphi}_{k,n} - \varphi^*_{k,n}\|_n^2 = o_P(1)$. As a consequence, $\|\widehat{\varphi}_{k,n} - \varphi^*_k\|_n^2 = o_P(1)$. The proof is complete.

A.6 Semiparametric Efficient Score

In this section we derive the semiparametric efficient score and information bound following Bickel *et al.* (1993). As a preparation for getting results for the model specified by (1) and (2), we first give a result on a simplified model. Let Y_{ij} and X_{ij} be the response variable and covariate vector for the jth ($j = 1, \ldots, m_i$) observation in the ith cluster ($i = 1, \ldots, n$). Denote

$$\underline{Y}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{im_i} \end{pmatrix}, \quad \underline{X}_i = \begin{pmatrix} X_{i1}^t \\ \vdots \\ X_{im_i}^t \end{pmatrix}, \quad g_\nu(\underline{X}_i) = \begin{pmatrix} g_\nu(X_{i1}) \\ \vdots \\ g_\nu(X_{im_i}) \end{pmatrix}, \quad \underline{e}_i = \begin{pmatrix} e_{i1} \\ \vdots \\ e_{im_i} \end{pmatrix}.$$

Consider the model

$$Y_{ij} = g_\nu(X_{ij}) + e_{ij}, \qquad i = 1, \ldots, n, j = 1, \ldots, m_i, \tag{A.3}$$

where $\{g_\nu(\cdot)\}$ are known functions indexed by $\nu \in \mathbb{R}^k$ and

$$E(\underline{e}_i | \underline{X}_i) = 0, \qquad i = 1, \ldots, n. \tag{A.4}$$

The data from different clusters are assumed to be independent. Let $f_i(\underline{x}_i, \underline{y}_i - g_\nu(\underline{x}_i))$ be the joint density of $(\underline{X}_i, \underline{Y}_i)$. It is assumed that $f_i(\cdot, \cdot)$ is smooth, bounded and satisfies $\lim_{e_{ij} \to \pm\infty} f_i(\underline{x}_i, \underline{e}_i) = 0$.

Lemma A.4. *For the model specified by (A.3) and (A.4), the contribution of cluster i to the efficient score of ν is given by*

$$\ell^*_{\nu, f_i} = \{\dot{g}_\nu(\underline{X}_i)\}^t \{E(\underline{e}_i \underline{e}_i^t | \underline{X}_i)\}^{-1} \underline{e}_i, \tag{A.5}$$

where

$$\dot{g}_\nu(\underline{X}_i) = \begin{pmatrix} \frac{\partial g_\nu(X_{i1})}{\partial \nu_1} & \cdots & \frac{\partial g_\nu(X_{i1})}{\partial \nu_k} \\ \cdots & \cdots & \cdots \\ \frac{\partial g_\nu(X_{im_i})}{\partial \nu_1} & \cdots & \frac{\partial g_\nu(X_{im_i})}{\partial \nu_k} \end{pmatrix}$$

The efficient score for ν *is* $\sum_i \ell^*_{\nu,f_i}$.

Proof. We only need calculate the contribution to the efficient score by each cluster. Since data from different clusters are independent, the (overall) efficient score is the summation of individual clusters' contributions.

Let ${}_{f_i}\dot{\mathcal{P}}_{\nu,f_i}$ denote the tangent space at (ν, f_i) when ν is fixed. Let

$$L_2^0(P_{\nu,f_i}) = \left\{ b(\underline{x}_i, \underline{e}_i) : \int b^2(\underline{x}_i, \underline{e}_i) f_i(\underline{x}_i, \underline{e}_i)\, d\underline{x}_i\, d\underline{e}_i < \infty \text{ and } \int b(\underline{x}_i, \underline{e}_i) f_i(\underline{x}_i, \underline{e}_i)\, d\underline{x}_i\, d\underline{e}_i = 0 \right\}$$

and

$$\mathcal{Q}_i = \{c^t \underline{e}_i h(x_{i1}, \ldots, x_{im_i}) : c \in \mathbb{R}^{m_i}, c^t \underline{e}_i h \in L_2^0(P_{\nu,f_i})\}.$$

We claim that

$$_{f_i}\dot{\mathcal{P}}_{\nu,f_i} = \mathcal{Q}_i^\perp \subset L_2^0(P_{\nu,f_i}). \tag{A.6}$$

To see this, we first show that ${}_{f_i}\dot{\mathcal{P}}_{\nu,f_i} \subset \mathcal{Q}_i^\perp$ or, equivalently, that any score function $a_i(\underline{x}_i, \underline{e}_i) \in {}_{f_i}\dot{\mathcal{P}}_{\nu,f_i}$ satisfies

$$\int a_i(\underline{x}_i, \underline{e}_i) f_i(\underline{x}_i, \underline{e}_i)\, d\underline{x}_i\, d\underline{e}_i = 0 \tag{A.7}$$

and

$$\int \underline{e}_i h(\underline{x}_i) a_i(\underline{x}_i, \underline{e}_i) f_i(\underline{x}_i, \underline{e}_i)\, d\underline{x}_i d\underline{e}_i = 0, \qquad \underline{e}_i h(\underline{x}_i) \in L_2^0(P_{\nu,f_i}). \tag{A.8}$$

Note that, for any one-dimensional submodel $\{f_{i,\eta}\}$ with $f_{i,0} = f_i$ and having score function a_i at $\eta = 0$, as $\eta \to 0$,

$$\frac{f_{i,\eta}^{1/2} - f_i^{1/2}}{\eta} \to \frac{1}{2} a_i f_i^{1/2} \quad \text{in } L_2(d\underline{x}_i, d\underline{e}_i).$$

Multiplying both sides by $f_{i,\eta}^{1/2} + f_i^{1/2}$ and using the boundedness of $f_{i,\eta}$, we have

$$\frac{f_{i,\eta} - f_i}{\eta} \to a_i f_i \quad \text{in } L_2(d\underline{x}_i, d\underline{e}_i).$$

Thus, since $f_{i,\eta}$ are densities, we have

$$0 \equiv \int \frac{f_{i,\eta} - f_i}{\eta}\, d\underline{x}_i\, d\underline{e}_i \to \int a_i(\underline{x}_i, \underline{e}_i) f_i(\underline{x}_i, \underline{e}_i)\, d\underline{x}_i\, d\underline{e}_i,$$

which gives (A.7). On the other hand, for fixed ν, according to (A.4), any perturbation $f_{i,\eta}$ of f_i within the model satisfies

$$\int \underline{e}_i f_{i,\eta}(\underline{x}_i, \underline{e}_i)\, d\underline{e}_i = 0. \tag{A.9}$$

Hence, for any $\underline{e}_i h(\underline{x}_i) \in L_0^2(P_{\nu,f_i})$ we have

$$0 \equiv \int \underline{e}_i h(\underline{x}_i) \frac{f_{i,\eta} - f_i}{\eta}\, d\underline{x}_i\, d\underline{e}_i \to \int \underline{e}_i h(\underline{x}_i) a_i(\underline{x}_i, \underline{e}_i) f_i(\underline{x}_i, \underline{e}_i)\, d\underline{x}_i\, d\underline{e}_i,$$

which implies (A.8). Next we construct a dense subset of $\mathcal{Q}_i^{\perp}$ which is contained in ${}_{f_i}\dot{\mathcal{P}}_{\nu,f_i}$. For any bounded and smooth $a_i \in \mathcal{Q}_i^{\perp}$, consider the submodel $\{f_{i,\eta} = (1+\eta a_i) f_i, |\eta| < \varepsilon\}$ for some sufficiently small $\varepsilon > 0$. It is easily seen that $f_{i,\eta}$ are bounded densities satisfying (A.9) with a_i as the score function at $\eta = 0$. As a consequence, a_i is in the tangent space ${}_{f_i}\dot{\mathcal{P}}_{\nu,f_i}$. This construction together with (A.7) and (A.8) yields (A.6).

For fixed f_i, the ordinary score for ν is

$$\dot{\ell}_{\nu,f_i}(\underline{x}_i, \underline{e}_i) = -\sum_{j=1}^{m_i} \frac{(\partial_{e_{ij}} f_i) \dot{g}_\nu(x_{ij})}{f_i}.$$

Using the same argument as in the previous paragraph, we can show that $\dot{\ell}_{\nu,f_i} \in L_2^0(P_{\nu,f_i})$. By Theorem 3.4.1 of Bickel *et al.* (1993), the efficient score for ν is

$$\ell^*_{\nu,f_i}(\underline{x}_i, \underline{e}_i) = \dot{\ell}_{\nu,f_i}(\underline{x}_i, \underline{e}_i) - \Pi\{\dot{\ell}_{\nu,f_i}(\underline{x}_i, \underline{e}_i) \mid {}_{f_i}\dot{\mathcal{P}}_{\nu,f_i}\} = \Pi\{\dot{\ell}_{\nu,f_i}(\underline{x}_i, \underline{e}_i) \mid \mathcal{Q}_i\}. \tag{A.10}$$

For any $b(\underline{x}_i, \underline{e}_i) \in L_2^0(P_{\nu,f_i})$,

$$\Pi(b \mid \mathcal{Q}_i) = \sum_{j=1}^{m_i} e_{ij} h_{j0}(\underline{x}_i) = \underline{e}_i^t \underline{h}_0(\underline{x}_i),$$

where, h_0 satisfies

$$E\{b(\underline{X}_i, \underline{e}_i) \underline{e}_i^t \underline{h}(\underline{X}_i)\} \equiv E\{\underline{h}_0^t(\underline{X}_i) \underline{e}_i \underline{e}_i^t \underline{h}(\underline{X}_i)\}, \qquad \underline{e}_i^t \underline{h}(\underline{x}_i) \in L_2^0(P_{\nu,f_i}),$$

or equivalently,

$$\underline{h}_0(\underline{X}_i) = \{E(\underline{e}_i \underline{e}_i^t | \underline{X}_i)\}^{-1} E\{b(\underline{X}_i, \underline{e}_i) \underline{e}_i | \underline{X}_i\}.$$

Let $b = -\sum_{j=1}^{m_i} (\partial_{e_{ij}} f_i)\{\partial_{\nu_s} g_\nu(x_{ij})\}/f_i$, the s-th column of $\dot{\ell}_{\nu,f_i}(\underline{x}_i, \underline{e}_i)$. Then,

$$E\{b(\underline{X}_i, \underline{e}_i) \underline{e}_i | \underline{X}_i\} = -\frac{\sum_{j=1}^{m_i} \partial_{\nu_s} g_\nu(x_{ij}) \int (\partial_{e_{ij}} f_i) \underline{e}_i\, d\underline{e}_i}{\int f_i(\underline{x}_i, \underline{e}_i)\, d\underline{e}_i}.$$

It follows from integration by parts that

$$\int (\partial_{e_{ij}} f_i)\underline{e}_i \, d\underline{e}_i = -\int f_i(\underline{x}_i, \underline{e}_i)\, d\underline{e}_i \begin{pmatrix} 0 \\ \cdots \\ 1 \\ \cdots \\ 0 \end{pmatrix}. \qquad \leftarrow j\text{-th position}$$

Hence

$$E\{b(\underline{X}_i, \underline{e}_i)\underline{e}_i | \underline{X}_i\} = \begin{pmatrix} \partial_{\nu_s} g_\nu(x_{i1}) \\ \cdots \\ \partial_{\nu_s} g_\nu(x_{im_i}) \end{pmatrix}.$$

Plugging into (A.10) yields (A.5). This completes the proof of Lemma A.4.

Now let $\mathcal{P}$ denote the model specified by (1) and (2). To derive the efficient score and information bound for this model, consider the following submodels:

$$\begin{aligned} &\mathcal{P}_1 : \text{Model } \mathcal{P} \text{ with only } \beta \text{ unknown}, \\ &\mathcal{P}_2 : \text{Model } \mathcal{P} \text{ with only } \theta(\cdot) \text{ unknown}, \\ &\mathcal{P}_3 : \text{Model } \mathcal{P} \text{ with both } \beta \text{ and } \theta(\cdot) \text{ known}. \end{aligned}$$

Denote by $\dot{\mathcal{P}}_1$, $\dot{\mathcal{P}}_2$, and $\dot{\mathcal{P}}_3$ the tangent spaces corresponding to these submodels. Let $\dot{\ell}_\beta$ be the ordinary score for β in Model $\mathcal{P}_1$. The efficient score for β in Model $\mathcal{P}$ is given by

$$\begin{aligned} \ell_\beta^* = \dot{\ell}_\beta - \Pi(\dot{\ell}_\beta | \dot{\mathcal{P}}_2 + \dot{\mathcal{P}}_3) &= \dot{\ell}_\beta - \Pi(\dot{\ell}_\beta | \dot{\mathcal{P}}_3) - \Pi(\dot{\ell}_\beta | \Pi_{\dot{\mathcal{P}}_3^\perp} \dot{\mathcal{P}}_2) \\ &= \Pi(\dot{\ell}_\beta | \dot{\mathcal{P}}_3^\perp) - \Pi\big\{\Pi(\dot{\ell}_\beta | \dot{\mathcal{P}}_3^\perp) \big| \Pi_{\dot{\mathcal{P}}_3^\perp} \dot{\mathcal{P}}_2\big\} \end{aligned}$$

(see Section 3.4 of Bickel *et al.*, 1993). Let $\mathcal{P}_{13}$ be Model $\mathcal{P}$ with $\theta(\cdot)$ known. Then $\Pi(\dot{\ell}_\beta | \dot{\mathcal{P}}_3^\perp)$ is the efficient score for β in Model $\mathcal{P}_{13}$. According to Lemma A.4,

$$\Pi(\dot{\ell}_\beta | \dot{\mathcal{P}}_3^\perp) = \sum_{i=1}^n \underline{X}_i^t \Delta_i \Sigma_i^{-1} [\underline{Y}_i - \mu\{\underline{X}_i \beta_0 + \theta_0(\underline{T}_i)\}],$$

where $\Sigma(\underline{X}_i, \underline{T}_i) = \text{var}(\underline{Y}_i | \underline{X}_i, \underline{T}_i)$ is the conditional covariance matrix. By considering parametric submodels of $\mathcal{P}_2$ with $\theta(\cdot)$ replaced by $\theta(\eta, \cdot)$ and applying Lemma A.4, we can show that

$$\Pi_{\dot{\mathcal{P}}_3^\perp} \dot{\mathcal{P}}_2 = \left\{ \sum_{i=1}^n \psi^t(\underline{T}_i) \Delta_i \Sigma_i^{-1} [\underline{Y}_i - \mu\{\underline{X}_i \beta_0 + \theta_0(\underline{T}_i)\}], \quad \psi(\cdot) \in L_2(\mathcal{T}) \right\}.$$

Therefore,

$$\ell_\beta^* = \sum_{i=1}^n \{\underline{X}_i - \psi^*(\underline{T}_i)\}^t \Delta_i \Sigma_i^{-1}[\underline{Y}_i - \mu\{\underline{X}_i\beta_0 + \theta_0(\underline{T}_i)\}],$$

where $\psi^*(\cdot)$ satisfies

$$\sum_{i=1}^n E\big[\{\underline{X}_i - \psi^*(\underline{T}_i)\}^t \Delta_i \Sigma_i^{-1} \Delta_i \psi(\underline{T}_i)\}\big] = 0, \quad \psi(\cdot) \in L_2(\mathcal{T}).$$

A.7 Proof of the Result in Remark 1

Assume that the working covariance matrices V_i depend on an unknown parameter τ that lies in a compact subset of a finite-dimensional space. Assume that $\widehat{\tau} - \tau^* = O_P(1/\sqrt{n})$ for some τ^* and the elements of $V_i^{-1}(\cdot)$ have bounded second derivatives. We show that Theorem 2 still holds. For notational convenience, suppose that τ is one-dimensional from now on. (The argument below goes through with some notational complications when τ is more than one-dimensional.)

Denote the first and second derivatives of V_i^{-1} with respect to τ as $DV_i^{-1}(\cdot)$ and $D^2V_i^{-1}(\cdot)$. Denote $\widehat{V}_i^{-1} = V_i^{-1}(\widehat{\tau})$ and $V_i^{*-1} = V_i^{-1}(\tau^*)$. Let $\widehat{\Pi}_{n,\widehat{\tau}}$ and $\widehat{\Pi}_{n,\tau^*}$ be the empirical projections onto $\mathbb{G}_n$ when the inverse working covariance matrix is respectively $\widehat{V}_i^{-1}$ and V_i^{*-1}. Similarly, let $< \cdot,\cdot >_{n,\widehat{\tau}}$ and $< \cdot,\cdot >_{\widehat{\tau}}$ be the empirical and theoretical inner products when the inverse working correlation matrix is $\widehat{V}_i^{-1}$, and $< \cdot,\cdot >_{n,\tau^*}$ and $< \cdot,\cdot >_{\tau^*}$ be the empirical and theoretical inner products when the inverse working correlation matrix is V_i^{*-1}.

We first establish some useful lemmas.

For a fixed τ, define an inner product between two longitudinal observation $y^{(1)}$ and $y^{(2)}$ as

$$< y^{(1)}, y^{(2)} >_{n,\tau} = \frac{1}{n}\sum_{i=1}^n \underline{y}_i^{(1)t} V_i^{-1}(\tau) \underline{y}_i^{(2)}.$$

The induced norm is denoted as $\|y\|_{n,\tau}$.

Lemma A.5. *There is an absolute constant C independent of n such that, for any longitudinal observation y,*

$$\big|\|y\|_{n,\widehat{\tau}}^2 - \|y\|_{n,\tau^*}^2\big| \le C\|y\|_{n,\tau^*}^2 |\widehat{\tau} - \tau^*|. \tag{A.11}$$

Moreover,

$$\big\|\widehat{\Pi}_{n,\tau^*}y - \widehat{\Pi}_{n,\widehat{\tau}}y\big\|_{n,\tau^*}^2 \le \frac{2C|\widehat{\tau} - \tau^*|}{1 - C|\widehat{\tau} - \tau^*|}\big\|y - \widehat{\Pi}_{n,\tau^*}y\big\|_{n,\tau^*}^2 \le \frac{2C|\widehat{\tau} - \tau^*|}{1 - C|\widehat{\tau} - \tau^*|}\|y\|_{n,\tau^*}^2.$$

Proof. By Taylor's theorem and applying the Cauchy–Schwarz inequality twice, we obtain that

$$\big|\|y\|_{n,\widehat{\tau}} - \|y\|_{n,\tau^*}\big| = \frac{1}{n}\Big|\sum_{i=1}^{n} \underline{y}_i^t V_i^{-1}(\widehat{\tau})\underline{y}_i - \sum_{i=1}^{n} \underline{y}_i^t V_i^{-1}(\tau^*)\underline{y}_i\Big| = \frac{|\widehat{\tau}-\tau^*|}{n}\Big|\sum_{i=1}^{n} \underline{y}_i^t DV_i^{-1}(\tilde{\tau})\underline{y}_i\Big|,$$

where $\tilde{\tau}$ is between τ^* and $\widehat{\tau}$. Let λ_1 denote the smallest absolute value among the eigenvalues of $V_i^{-1}(\tau^*)$, $i = 1, \ldots, n$, and let λ_2 denote the largest absolute value among the eigenvalues of $DV_i^{-1}(\tilde{\tau})$ for $\tilde{\tau}$ between τ^* and $\widehat{\tau}$, $i = 1, \cdots, n$. By our assumptions, λ_2/λ_1 is bounded by some constant C that is independent of n. The first result follows. To obtain the second result, using the property of orthogonal projection and repeatedly applying (A.11), we have

$$\big\|y - \widehat{\Pi}_{n,\widehat{\tau}}y\big\|_{n,\tau^*}^2 \le \frac{\big\|y - \widehat{\Pi}_{n,\widehat{\tau}}y\big\|_{n,\widehat{\tau}}^2}{1 - C|\widehat{\tau}-\tau^*|} \le \frac{\big\|y - \widehat{\Pi}_{n,\tau^*}y\big\|_{n,\widehat{\tau}}^2}{1 - C|\widehat{\tau}-\tau^*|} \le \frac{1 + C|\widehat{\tau}-\tau^*|}{1 - C|\widehat{\tau}-\tau^*|}\big\|y - \widehat{\Pi}_{n,\tau^*}y\big\|_{n,\tau^*}^2.$$

Therefore,

$$\big\|\widehat{\Pi}_{n,\tau^*}y - \widehat{\Pi}_{n,\widehat{\tau}}y\big\|_{n,\tau^*}^2 = \big\|y - \widehat{\Pi}_{n,\widehat{\tau}}y\big\|_{n,\tau^*}^2 - \big\|y - \widehat{\Pi}_{n,\tau^*}y\big\|_{n,\tau^*}^2 \le \frac{2C|\widehat{\tau}-\tau^*|}{1 - C|\widehat{\tau}-\tau^*|}\big\|y - \widehat{\Pi}_{n,\tau^*}y\big\|_{n,\tau^*}^2$$

as desired.

The following is a restatement of Lemma 6.1 of Huang (2003).

Lemma A.6. *Let A, B be symmetric positive definite matrices. Define*

$$\epsilon_n = \sup_w \Big|\frac{w^tAw}{w^tBw} - 1\Big|.$$

We have

$$\sup_{u,v}\Bigg|\frac{u^tA^{-1}v - u^tB^{-1}v}{\sqrt{u^tA^{-1}u}\sqrt{v^tB^{-1}v}}\Bigg| \le \frac{\epsilon_n^2}{1-\epsilon_n} + 2\frac{\epsilon_n}{\sqrt{1-\epsilon_n}}.$$

Lemma A.7.

$$\big[nH^{11}(\widehat{\tau})\big]_{ij} - \big[nH^{11}(\tau^*)\big]_{ij} = O_P\Big(\frac{1}{\sqrt{n}}\Big) \qquad i,j = 1, \cdots, p.$$

Proof. Applying Lemma A.6 to $A = H(\widehat{\tau})$ and $B = H(\tau^*)$ and noting that $\epsilon_n = O_P(1/\sqrt{n})$ by Lemma A.5, we obtain that

$$\sup_{u,v}\Bigg|\frac{u^tH^{-1}(\widehat{\tau})v - u^tH^{-1}(\tau^*)v}{\sqrt{u^tH^{-1}(\tau^*)u}\sqrt{v^tH^{-1}(\tau^*)v}}\Bigg| = O_P\Big(\frac{1}{\sqrt{n}}\Big).$$

This is in particular true when H^{-1} is replaced by H^{11}, a submatrix of H^{-1}. From the proof of Theorem 2, the eigenvalues of $H^{11}(\tau^*)$ are of order $O_P(1/n)$. Combining these observations gives the result.

Lemma A.8.

$$\sup_{u,v}\left|\frac{u^tH_{22}^{-1}(\widehat{\tau})v-u^tH_{22}^{-1}(\tau^*)v}{\sqrt{u^tH_{22}^{-1}(\tau^*)u}\sqrt{v^tH_{22}^{-1}(\tau^*)v}}\right|=O_P\Big(\frac{1}{\sqrt{n}}\Big). \tag{A.12}$$

Proof. The desired result follows from Lemma A.6 by noting that H_{22} is a submatrix of H.

Lemma A.9. *For any two longitudinal samples $y^{(1)}$ and $y^{(2)}$,*

$$\sup_{y^{(1)},y^{(2)}}\frac{|<y^{(1)},y^{(2)}>_{n,\widehat{\tau}}-<y^{(1)},y^{(2)}>_{n,\tau^*}|}{\|y^{(1)}\|_{n,\tau^*}\|y^{(2)}\|_{n,\tau^*}}=O_P\Big(\frac{1}{\sqrt{n}}\Big).$$

Proof. Note that (A.12) holds with H_{22}^{-1} replaced by V_i^{-1}. The desired result follows by applying the Cauchy–Schwarz inequality.

Lemma A.10. *Eigenvalues for $H_{22}^{-1}(\tau^*)$ are of order $O_P(K_n/n)$.*

Proof. For any vector $c=(c_1,\cdots,c_{K_n})^t$, denote $g(t)=\sum_{l=1}^{K_n}c_lB_l(t)$, and then $g(\underline{T}_i)=\underline{Z}_ic$. It follows from Lemma A.2 that, except on a set with probability tending to zero,

$$c^t\frac{1}{n}H_{22}(\tau^*)c=c^t\frac{1}{n}\sum_{i=1}^n\underline{Z}_i^tV_i^{-1}(\tau^*)\underline{Z}_ic=<g,g>_{n,\tau^*}\le 2\|g\|_{\tau^*}^2\asymp\|g\|_{L^2}^2\asymp\frac{|c|^2}{K_n}.$$

Hence, the eigenvalues of $H_{22}(\tau^*)/n$ are of order $O_P(1/K_n)$, which in turn implies that the eigenvalues for $H_{22}^{-1}(\tau^*)$ are of order $O_P(K_n/n)$.

Define

$$\widehat{\beta}_{\widehat{V}}=H^{11}\Big\{\sum_{i=1}^n\underline{X}_i^t\widehat{V}_i^{-1}\underline{Y}_i-H_{12}H_{22}^{-1}\sum_{i=1}^n\underline{Z}_i^t\widehat{V}_i^{-1}\underline{Y}_i\Big\}$$

and

$$A\widehat{\beta}_{\widehat{V}}=\beta+H^{11}\Big\{\sum_{i=1}^n\underline{X}_i^t\widehat{V}_i^{-1}\delta(\underline{T}_i)-H_{12}H_{22}^{-1}\sum_{i=1}^n\underline{Z}_i^t\widehat{V}_i^{-1}\delta(\underline{T}_i)\Big\}.$$

Then

$$\begin{aligned}\widehat{\beta}_{\widehat{V}}-A\widehat{\beta}_{\widehat{V}}&=H^{11}\Big\{\sum_{i=1}^n\underline{X}_i^t\widehat{V}_i^{-1}\underline{e}_i-H_{12}H_{22}^{-1}\sum_{i=1}^n\underline{Z}_i^t\widehat{V}_i^{-1}\underline{e}_i\Big\}\\&=H^{11}\sum_{i=1}^n\underline{X}_i^t\widehat{V}_i^{-1}\Big\{\underline{e}_i-(\widehat{\Pi}_{n,\widehat{\tau}}\underline{e})(\underline{T}_i)\Big\}.\end{aligned}$$

Define $\widehat{\beta}_{V^*}$ and $A\widehat{\beta}_{V^*}$ similarly by replacing $\widehat{V}$ with V^* in the above expressions. Consider the decomposition

$$\widehat{\beta}_{\widehat{V}}-\beta_0=\widehat{\beta}_{V^*}-\beta_0+\{(\widehat{\beta}_{\widehat{V}}-A\widehat{\beta}_{\widehat{V}})-(\widehat{\beta}_{V^*}-A\widehat{\beta}_{V^*})\}+(A\widehat{\beta}_{\widehat{V}}-A\widehat{\beta}_{V^*}).$$

Careful analysis can show that (see Zhang, 2004, for details)

$$A\widehat{\beta}_{\widehat{V}} - A\widehat{\beta}_{V^*} = o_P\Big(\frac{1}{\sqrt{n}}\Big) \tag{A.13}$$

and that

$$(\widehat{\beta}_{\widehat{V}} - A\widehat{\beta}_{\widehat{V}}) - (\widehat{\beta}_{V^*} - A\widehat{\beta}_{V^*}) = o_P\Big(\frac{1}{\sqrt{n}}\Big). \tag{A.14}$$

(A.13) and (A.14) together with Theorem 2 give us $\{R(\widehat{\beta}_{V^*})\}^{-1/2}(\widehat{\beta}_{\widehat{V}} - \beta_0) \to \mathrm{Normal}(0, \mathbb{I})$. We shall prove that

$$\{R(\widehat{\beta}_{\widehat{V}})\}^{-1/2} = \{R(\widehat{\beta}_{V^*})\}^{-1/2} + o_P(\sqrt{n}). \tag{A.15}$$

Therefore, $\{R(\widehat{\beta}_{\widehat{V}})\}^{-1/2}(\widehat{\beta}_{\widehat{V}} - \beta_0) \to \mathrm{Normal}(0, \mathbb{I})$.

Proof of (*A*.15). For a fixed τ, define a new inner product between two longitudinal observation y_1 and y_2 as

$$< y^{(1)}, y^{(2)} >^1_{n,\tau} = \frac{1}{n}\sum_{i=1}^{n} \underline{y}_i^{(1)t} V_i^{-1}(\tau)\Sigma_i V_i^{-1}(\tau)\underline{y}_i^{(2)}.$$

The induced norm is denoted as $\|y\|^1_{n,\tau}$.

Lemma A.11. *The norms* $\|\cdot\|^1_{n,\widehat{\tau}}, \|\cdot\|^1_{n,\tau^*}, \|\cdot\|_{n,\widehat{\tau}}, \|\cdot\|_{n,\tau^*}$ *are all equivalent in the sense that the ratio of any two norms are bounded away from 0 and infinity uniformly for all longitudinal observations.*

Proof. The result follows from the observation that the eigenvalues for V_i^{-1} and Σ_i are bounded away from 0 and infinity uniformly across $i = 1, \cdots, n$.

Lemma A.12. *For any two longitudinal samples* $y^{(1)}$ *and* $y^{(2)}$,

$$\sup_{y^{(1)},y^{(2)}} \frac{|< y^{(1)}, y^{(2)} >^1_{n,\widehat{\tau}} - < y^{(1)}, y^{(2)} >^1_{n,\tau^*}|}{\|y^{(1)}\|^1_{n,\tau^*}\|y^{(2)}\|_{n,\tau^*}} = O_P\Big(\frac{1}{\sqrt{n}}\Big).$$

Lemma A.13. *There is an absolute constant* C *independent of* n *such that, for any longitudinal observation* y,

$$\big|(\|y\|^1_{n,\widehat{\tau}})^2 - (\|y\|^1_{n,\tau^*})^2\big| \le C(\|y\|^1_{n,\tau^*})^2|\widehat{\tau} - \tau^*|.$$

The proofs for the above two lemmas are exactly the same to the proofs of Lemma A.5 and Lemma A.9. We only need replace the V_i^{-1} by $V_i^{-1}\Sigma_i V_i^{-1}$ in the proofs.

Write

$$\widehat{W}=\frac{1}{n}\sum_{i=1}^{n}\{\underline{X}_i-\underline{Z}_iH_{22}^{-1}(\widehat{\tau})H_{21}(\widehat{\tau})\}^tV_i^{-1}(\widehat{\tau})\Sigma_iV_i^{-1}(\widehat{\tau})\{\underline{X}_i-\underline{Z}_iH_{22}^{-1}(\widehat{\tau})H_{21}(\widehat{\tau})\}$$

and

$$W^*=\frac{1}{n}\sum_{i=1}^{n}\{\underline{X}_i-\underline{Z}_iH_{22}^{-1}(\tau^*)H_{21}(\tau^*)\}^tV_i^{-1}(\tau^*)\Sigma_iV_i^{-1}(\tau^*)\{\underline{X}_i-\underline{Z}_iH_{22}^{-1}(\tau^*)H_{21}(\tau^*)\}.$$

The difference between the (k,l)-th $(k,l=1,\cdots,p)$ entries of $\widehat{W}$ and W^* can be written as

$$\begin{aligned}[\widehat{W}]_{kl}-[W^*]_{kl} &= (< x_k-\widehat{\Pi}_{n,\widehat{\tau}}x_k, x_l-\widehat{\Pi}_{n,\widehat{\tau}}x_l >^1_{n,\widehat{\tau}} - < x_k-\widehat{\Pi}_{n,\widehat{\tau}}x_k, x_l-\widehat{\Pi}_{n,\widehat{\tau}}x_l >^1_{n,\tau^*})\\ &\quad + (< \widehat{\Pi}_{n,\tau^*}x_k-\widehat{\Pi}_{n,\widehat{\tau}}x_k, x_l-\widehat{\Pi}_{n,\widehat{\tau}}x_l >^1_{n,\tau^*})\\ &\quad + (< x_k-\widehat{\Pi}_{n,\tau^*}x_k, \widehat{\Pi}_{n,\widehat{\tau}}x_l-\widehat{\Pi}_{n,\tau^*}x_l >^1_{n,\tau^*})\\ &\triangleq I_8+I_9+I_{10}.\end{aligned}$$

It follows from Lemmas A.11 and A.12 that,

$$\begin{aligned}I_8 &= \|x_k-\widehat{\Pi}_{n,\widehat{\tau}}x_k\|^1_{n,\tau^*}\|x_l-\widehat{\Pi}_{n,\widehat{\tau}}x_l\|^1_{n,\tau^*}O_P\Big(\frac{1}{\sqrt{n}}\Big)\\ &= \|x_k-\widehat{\Pi}_{n,\widehat{\tau}}x_k\|_{n,\widehat{\tau}}\|x_l-\widehat{\Pi}_{n,\widehat{\tau}}x_l\|_{n,\widehat{\tau}}O_P\Big(\frac{1}{\sqrt{n}}\Big)=O_P\Big(\frac{1}{\sqrt{n}}\Big).\end{aligned}$$

By Lemma A.11 and the second conclusion of Lemma A.5,

$$I_9\le\|\widehat{\Pi}_{n,\tau^*}x_k-\widehat{\Pi}_{n,\widehat{\tau}}x_k\|_{n,\tau^*}\|x_l-\widehat{\Pi}_{n,\widehat{\tau}}x_l\|_{n,\widehat{\tau}}=o_P\Big(\frac{1}{n^{1/4}}\Big)\|x_l\|_{n,\widehat{\tau}}\|x_k\|_{n,\tau^*}=o_P\Big(\frac{1}{n^{1/4}}\Big).$$

Arguing similarly we obtain that $I_{10}=o_P(n^{-1/4})$. Hence, $\widehat{W}-W^*=o_P(n^{-1/4})$. Therefore, by Lemma A.7, we have

$$\begin{aligned}&nR(\widehat{\beta}_{\widehat{V}})-nR(\widehat{\beta}_{V^*})\\ &=nH^{11}(\widehat{\tau})(\widehat{W}/n)\{nH^{11}(\widehat{\tau})\}-nH^{11}(\tau^*)(W^*/n)\{nH^{11}(\tau^*)\}=o_P(n^{-1/4})=o_P(1).\end{aligned}$$

Consequently, $\{R(\widehat{\beta}_{\widehat{V}})\}^{-1/2}=\{R(\widehat{\beta}_{V^*})\}^{-1/2}+o_P(\sqrt{n})$. The proof of (A.15) is complete.

Table 1: *Summary of simulation results for comparing the estimators using a working independence (WI) and an exchangeable (EX) correlation structure, based on 250 replications. Cubic splines are used with the number of knots chosen from the range 0–10 by the five-fold delete-subject-out cross-validation. Each entry equals the original value multiplied by 10.*

		$\beta_1 = 1$			$\beta_2 = 1$		$\theta(\cdot)$
Method	Bias	SD	MSE	Bias	SD	MSE	MISE
WI	-.0370	.4941	.0319	-.0090	.9937	.1089	.3114
EX	.0443	.3907	.0181	.0033	.7876	.0677	.2276

Table 2: *Summary of simulation results for the example in Wang et al. (2005) from 250 replications. Each entry equals the original value multiplied by 10. Working independence (WI) and the true exchangeable covariance (EX) structures are used in combination with Wang et al.'s iterative kernel (kernel) and proposed spline methods (spline).*

	$\beta_1 = 1$			$\beta_2 = 1$		
Method	Bias	SD	ave MSE	Bias	SD	ave MSE
WI (kernel)	.0732	.8564	.0739	.0135	1.6486	.2718
EX (kernel)	.0118	.5675	.0322	.0107	1.6324	.2665
WI (spline)	.0413	.8506	.0722	-.0974	1.0322	.1071
EX (spline)	-.0173	.5852	.0341	-.0786	.6982	.0492

Table 3: *Regression coefficients in the CD4 cell counts study in HIV seroconverters using the kernel and spline estimates. Working covariance structure used are working independence (WI) and "random intercept plus serial correlation plus measurement error" (RSM).*

	Kernel				Spline			
	WI		RSM		WI		RSM	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Age	.014	.035	.008	.032	.015	.034	.008	.033
Smoking	.984	.182	.579	.139	.971	.182	.629	.140
Drug	1.049	.526	.584	.335	1.102	.528	.717	.334
Sex partners	-.054	.059	.078	.039	-.069	.059	.025	.038
Depression	-.033	.021	-.046	.014	-.033	.021	-.041	.014