Semiparametric Penalized Quadratic Inference Functions for Longitudinal Data in Ultra-high Dimensions

Brittany Green[a], Heng Lian[b], Yan Yu[c], Tianhai Zu[c]

[a] *Department of Information Systems, Analytics, and Operations, University of Louisville, Louisville, Kentucky, U.S.A*
[b] *Department of Mathematics, City University of Hong Kong, Hong Kong*
[c] *Department of Operations, Business Analytics, & Information Systems,*
University of Cincinnati, Cincinnati, Ohio, U.S.A

Abstract

In many biomedical and health studies, multivariate data arise from repeated measurements on a sample of subjects over time. In order to analyze such longitudinal data, we need to consider the correlations from the same subject and it is inappropriate to use a simple multivariate model assuming independence structure. Motivated by a large scale longitudinal public health study that requires longitudinal data analysis with correlated multivariate discrete responses from repeated measurements and very high dimensional covariates, we adopt a flexible semi-parametric approach for simultaneous variable selection and estimation without the requirement of specifying the full likelihood. Specifically, we propose generalized partially linear single-index models using penalized quadratic inference functions for longitudinal data in ultra-high dimension. A key feature is that we allow the number of single-index covariates in the nonparametric term to diverge and even to be in ultra-high dimension. The penalized quadratic inference functions easily incorporate within-subject correlation and pursue efficient estimation, and the single-index models can incorporate nonlinearity and some interactions while avoiding the curse of dimensionality. In this challenging setting, we contribute both an efficient algorithm and new asymptotic theory for our proposed approach for diverging and even ultra-dimensional covariates and a discrete response in longitudinal data. We apply our method to investigate diabetes status within a continuing longitudinal public health study with very high-dimensional genetic variables and phenotype variables.

Keywords: Longitudinal data; Model selection; Multivariate correlated response; Partially linear model;
Single-index model.

1. Introduction

Numerous large scale public health research studies are longitudinal, where participants have repeated measurements taken over time. To analyze such kind of multivariate data that exhibit clear correlation among within-subject responses, we need to incorporate the correlation structure rather than assuming a simple multivariate regression model under independence.

One main goal of analyzing these types of longitudinal studies is to identify the genetic and phenotype factors related to a disease to provide insight into more effective treatment and disease prevention strategies. For example, researchers have discovered risk factors linked to various disease mechanisms (e.g., Meigs et al. [23]) using the Framingham data, an ongoing large scale multi-generational health study [6]. Within these types of large-scale longitudinal studies, while the true correlation among participants is usually difficult to uncover, incorporating within participant dependence can lead to more efficient estimation. Moreover, the disease of interest measured over time is sometimes a correlated discrete measure such as diabetes status. This non-normal correlated response creates a challenge in specifying the joint likelihood for longitudinal data.

In addition to the longitudinal nature of large scale public health studies, more recently, the genotype of participants is also collected. These genetic factors are very high dimensional as demonstrated in the Framingham data which collects a complex array of genetic data, including tens of thousands of single nucleotide polymorphisms (SNPs) from each participant. Notably, previous research has linked some genetic factors to disease. For instance, genomic studies have helped identify mechanisms of hypertension and diabetes [39]. In these very high dimensional settings, variable selection is imperative to identify the important risk factors, since usually only a few covariates relate to the response and including non-important variables lessens estimation efficiency and impedes inference. In addition, not only do these types of studies have high dimensional genetic data, recent research has found that genetic data interacts with phenotype variables. For example, Taylor et al. [31] shows that the genetic effects of hypertension are altered under phenotype factors such as BMI.

A motivating example of this paper to exemplify this complexity focuses on identifying factors related to diabetes status, a correlated discrete response, from the offspring cohort of the longitudinal Framingham data. Diabetes affects millions of people worldwide, and identifying genetic and phenotype factors that relate to diabetes can help inform preventative measures and further uncover biologic measures of diabetes [10]. In particular, one research question of interest is to determine which SNPs in high dimensions and phenotype factors relate to diabetes status among participants over time. While a traditional approach to this problem employs a linear model, due to the complexity of gene expression and interactions with phenotype factors, inflexible models with parametric assumptions may not account for the potential nonlinearity and synergy between genetic and phenotype data in high dimensional longitudinal data. To demonstrate, Figure 1 shows a clear nonlinear relationship under our proposed model between diabetes and the combination of SNPs and phenotype factors in the Framingham data.

To balance flexibility for accurate estimation and interpretability for this high dimensional set of risk factors, we consider a flexible semiparametric approach for repeated observations. Specifically, we adopt generalized partially linear single-index models (GPLSIM) [4] for longitudinal data in ultra-high dimension. Generalized partially linear single-index models achieve dimension reduction by reducing the high-dimensional predictors to a univariate index within a flexible function. Moreover, single-index models can capture some interactions among covariates as opposed to additive models. This is advantageous since SNPs and phenotypic risk factors do not relate to disease in isolation: the compound impact of the genotype-phenotype interaction has been shown to outperform the impact of using the conventional risk factors in isolation (e.g., Franks [10], Taylor et al. [31]).

Moreover, diabetes status, the outcome of interest measured during multiple waves of the Framingham data, is a correlated discrete response. This poses a challenge as the full joint likelihood can be intractable for correlated discrete data. To tackle this difficulty, we employ the penalized quadratic inference function (QIF) to account for within-subject correlation, perform model selection, and seek efficient estimation for diverging and potentially ultra-high dimensional longitudinal data. Previous research employing penalized generalized estimating equations (GEE) for diverging and ultra-high dimensional longitudinal data for linear and semiparametric models includes Wang et al. [37] and Green et al. [12]. However, generalized estimating equations are known to be less efficient and overfit the model compared to the quadratic inference function approach when the working correlation matrix is misspecified (e.g., Qu et al. [27],Cho and Qu [5]). In addition, the quadratic inference function is applicable in a variety of model setups and shows promising results such as in Qu and Li [26], Wang et al. [38], and Wang et al. [36].

Due to the many advantages of the quadratic inference function, a handful of works have proposed research employing partially linear single-index models using the quadratic inference function for longitudinal data (e.g., Bai et al. [2] and Lai et al. [18]). However, these works assume the dimension of both the single-index and partially linear covariates is fixed. Also in fixed finite dimension, Ma et al. [21] and Li et al. [19] incorporated variable selection for the partially linear single-index model employing the quadratic inference function for continuous longitudinal responses with the identity link function. In these studies, the real-data applications considered a fixed low-to-moderate dimensional set up: the application in Ma et al. [21] included a total of 11 covariates, and the application in Li et al. [19] included 13 covariates. In contrast, the Framingham data analyzed using our approach has 878 participants but involves over 50,000 SNP covariates and phenotype variables even within the nonparametric portion.

As opposed to the previous works, our approach allows the number of covariates in both the nonparametric and linear components of the generalized partially linear single-index model to diverge and even in ultra-high dimension. We also allow the number of important covariates to diverge. This is especially pertinent for our motivating example, since there are more than 50,000 genetic SNP variables. In particular, allowing flexible modeling and determining the sparse set of important covariates can lead to more accurate estimation. However, as a result of incorporating diverging covariates, we navigate additional challenges in both computation and theory when we allow ultra-high dimensional data within the nonlinear, unknown, flexible function with potentially diverging support of the single index. We establish asymptotic theory for ultra-dimensional covariates for model selection and estimation including the oracle property, which is much more challenging to establish than fixed-dimensional theory. In addition, while there are many previous approaches for estimating the coefficients of the generalized partially linear single-index model, efficient estimation becomes an even more challenging task when introducing ultra-high dimensional correlated data. This is because of the potentially high dimensional covariates in a nonlinear unknown function estimated nonparametrically together with the non-convex smoothly clipped absolute deviation (SCAD) penalty function as in Fan and Li [7], all within a longitudinal framework. Therefore, to select the sparse set of important covariates and to estimate the corresponding coefficients, we provide a computationally efficient iterative algorithm. This approach implements strategic approximations to reduce the computational burden and increase the effectiveness of the algorithm even for discrete correlated responses.

2. Quadratic Inference Function for Semiparametric Longitudinal Data Analysis

2.1. Model

For each subject i with $i = 1, \ldots, n$, assume correlation among its observations over time $t = 1, \ldots, T_i$, but independence from other subjects. We observe, for subject i, a correlated multivariate response vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{it}, \ldots, Y_{iT_i})$ over repeated measurements over time. We also observe at each time t, the $p_n \times 1$ dimensional single-index covariate vector $\mathbf{X}_{it} = (X_{it,1}, \ldots, X_{it,p_n})^{\mathrm{T}}$ and the $q_n \times 1$ dimensional linear covariate vector $\mathbf{Z}_{it} = (Z_{it,1}, \ldots, Z_{it,q_n})^{\mathrm{T}}$. Notably, we allow both the number of single-index and partially linear covariates, p_n and q_n, to diverge and potentially be in the exponential order. Also, the number of observations T_i can be different for each subject i for imbalanced observations.

We consider generalized partially linear single-index models for longitudinal data in ultra-high dimension to allow flexibility while avoiding the curse-of-dimensionality, that is,

$$E(Y_{it}|\mathbf{X}_{it}, \mathbf{Z}_{it}) = \mu_{it} = g^{-1}(\eta(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0) + \mathbf{Z}_{it}^{\mathrm{T}}\boldsymbol{\gamma}_0), \quad i = 1, \ldots, n; t = 1, \ldots, T_i. \tag{1}$$

Here $\eta(\cdot)$ is an unknown flexible function estimated non-parametrically; and $g(\cdot)$ is a link function of the exponential family. The coefficient vector for the single-index covariates is $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_n})^{\mathrm{T}}$ and the coefficient vector for the linear covariates is $\boldsymbol{\gamma}_0$ with $\boldsymbol{\gamma}_0 = (\gamma_{01}, \ldots, \gamma_{0q_n})^{\mathrm{T}}$. We assume a sparse set of important covariates, which is common in modern statistics literature [14]. In particular, there are a nonzero subset of p_{sn} coefficients from the total p_n coefficients and a subset of q_{sn} nonzero coefficients from the total q_n linear coefficients, where the rest of the coefficients are zero. For model identifiability, we assume $\left\|\boldsymbol{\beta}_0\right\| = 1$ with the first component positive [40].

We estimate the flexible univariate function of the conditional mean $\eta(\cdot)$ non-parametrically using polynomial splines. We first assume the support of the single-index $\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0$ is $[a, b]$. We note that the length of this support can be diverging due to the potentially ultra-high dimensional covariates, thus practically, we use the support of $\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}$ based on a given $\boldsymbol{\beta}$. We then divide this support based on H' interior knots to create subintervals $[c_k, c_{k+1})$, where $k \in \{0, \cdots, H'\}$ is determined by the partition, $a = c_0 < c_1 < \cdots < c_{H'} < c_{H'+1} = b$. Given we approximate $\eta(\cdot)$ with a degree $s \geq 2$ polynomial over each interval, a polynomial spline of order s is a $s - 1$ degree polynomial on each interval and globally $s-2$ times differentiable [29]. We consider the nonparametric estimation of the unknown flexible function $\eta(\cdot)$ with the single-index $u_{it} = \mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0$ as a linear combination of B-spline bases $\eta(u_{it}) \approx \mathbf{G}^{\mathrm{T}}(u_{it})\boldsymbol{\theta}$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_H)^{\mathrm{T}}$ are the basis coefficients of size $H \equiv H_n = 1 + s + H'$. Accordingly, since $\eta(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta})$ is estimated by $\mathbf{G}^{\mathrm{T}}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}$, the conditional mean $\mu_{it} = g^{-1}(\eta(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0) + \mathbf{Z}_{it}^{\mathrm{T}}\boldsymbol{\gamma}_0)$ now becomes $g^{-1}(\mathbf{G}^{\mathrm{T}}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0 + \mathbf{Z}_{it}^{\mathrm{T}}\boldsymbol{\gamma}_0)$. Therefore, in the rest of this paper, we focus on estimating the column coefficient vector $\boldsymbol{\alpha}_0 = (\boldsymbol{\theta}_0\ \boldsymbol{\beta}_0\ \boldsymbol{\gamma}_0)$, the spline, single-index, and the partial linear coefficients, respectively.

2.2. Quadratic Inference Function and Estimation

To account for the correlation within a subject's observations in longitudinal data, one may use the quadratic inference function (QIF) approach to estimate the spline basis, single-index, and partial linear coefficient vector, $\boldsymbol{\alpha}$, for the given covariates. The quadratic inference function from Qu et al. [27] replaces the inverse of the working correlation matrix with a linear combination of basis matrices, that is, $\mathbf{R}^{-1} \approx a_1\mathbf{M}_1 + \cdots + a_m\mathbf{M}_m$. Here $\mathbf{M}_1, \ldots, \mathbf{M}_m$ are predetermined, known symmetric matrices and $\mathbf{a} = (a_1, \ldots, a_m)^{\mathrm{T}}$ are constant coefficients. For most of the common correlation structures, there are available linear combinations of basis matrices to approximate $\mathbf{R}^{-1}$ as further discussed in Section 5.2.

To estimate $\boldsymbol{\alpha}$, Qu et al. [27] extend generalized estimating equations from Zeger and Liang [42] by adopting the generalized method of moments estimator from Hansen [13] to minimize the following quadratic inference function

$$Q_n(\boldsymbol{\alpha}) = \mathbf{g}_n^{\mathrm{T}}\mathbf{W}_n^{-1}\mathbf{g}_n, \tag{2}$$

where the extended score vector is

$$\mathbf{g}_n = \frac{1}{n}\sum_i \mathbf{g}_i(\boldsymbol{\alpha}) = \frac{1}{n}\left\{\begin{array}{c} \sum_i \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{A}_i(\boldsymbol{\alpha})^{1/2}\mathbf{M}_1\mathbf{A}_i(\boldsymbol{\alpha})^{-1/2}(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha})) \\ \sum_i \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{A}_i(\boldsymbol{\alpha})^{1/2}\mathbf{M}_2\mathbf{A}_i(\boldsymbol{\alpha})^{-1/2}(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha})) \\ \cdots \\ \sum_i \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{A}_i(\boldsymbol{\alpha})^{1/2}\mathbf{M}_m\mathbf{A}_i(\boldsymbol{\alpha})^{-1/2}(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha})) \end{array}\right\},$$

and $\mathbf{W}_n = \frac{1}{n} \sum_i \mathbf{g}_i(\boldsymbol{\alpha})\mathbf{g}_i(\boldsymbol{\alpha})^{\mathrm{T}}$. The resulting estimates are determined as $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} Q_n(\boldsymbol{\alpha})$, and $Q_n(\boldsymbol{\alpha})$ is known as the quadratic inference function [27]. For $\mathbf{g}_n$ in equation (2), we introduce matrix notation for subject i, where $\mathbf{Y}_i$ denotes the $T_i \times 1$ response vector, and $\boldsymbol{\mu}_i(\boldsymbol{\alpha}) = g^{-1}(\mathbf{G}(\mathbf{X}_i\boldsymbol{\beta})\boldsymbol{\theta} + \mathbf{Z}_i\boldsymbol{\gamma})$ is the spline approximated marginal mean, where $\mathbf{X}_i$ is the $T_i \times p_n$ dimensional single-index covariate matrix, $\mathbf{G}$ is the corresponding $T_i \times H$ spline basis, and $\mathbf{Z}_i$ is the $T_i \times q_n$ dimensional linear covariate matrix. Further, $\mathbf{A}_i(\boldsymbol{\alpha}) = diag\{\sigma^2_{i1}(\boldsymbol{\alpha}), \dots, \sigma^2_{iT_i}(\boldsymbol{\alpha})\}$ is a diagonal matrix of the variance of $\mathbf{Y}_i$. Here $\sigma^2_{it}(\boldsymbol{\alpha}) = \phi\dot{\mu}(h_{it})$ is the spline approximated marginal variance per subject i and observation t, where the systematic component $h_{it} = \mathbf{G}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}_{it}\boldsymbol{\beta})\boldsymbol{\theta} + \mathbf{Z}^{\mathrm{T}}_{it}\boldsymbol{\gamma}$, $\dot{\mu}(\cdot)$ is the first derivative with respect to h_{it}, and the scaling constant $\phi = 1$ as in Wang et al. [38]. $\mathbf{V}_i(\boldsymbol{\alpha})$ is defined as $\mathbf{V}_i(\boldsymbol{\alpha}) = \Big(\mathbf{G}(\mathbf{X}_i\boldsymbol{\beta}), \mathrm{diag}(\dot{\mathbf{G}}(\mathbf{X}_i\boldsymbol{\beta})\boldsymbol{\theta})\mathbf{X}_i\mathbf{J}(\boldsymbol{\beta}), \mathbf{Z}_i\Big)$, where $\dot{\mathbf{G}}(\cdot)$ is the first derivative of the spline basis. We define the Jacobian matrix to be $\mathbf{J}(\boldsymbol{\beta}) = \partial\boldsymbol{\beta}/\partial\boldsymbol{\beta}^{(-1)} = ((-\boldsymbol{\beta}^{(-1)}/(1 - \|\boldsymbol{\beta}^{(-1)}\|^2)^{1/2})^{\mathrm{T}}, I_{(p-1)\times(p-1)})^{\mathrm{T}}$, where we reparameterize $\boldsymbol{\beta}$ to be a function of $\boldsymbol{\beta}^{(-1)} = (\boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_p)$ using the "delete-one-component" method for identifiability as in Yu et al. [41] and Yu and Ruppert [40].

Notably, the quadratic inference function does not need to estimate the coefficients $\mathbf{a} = (a_1, \dots, a_m)^{\mathrm{T}}$. This may be especially beneficial in a semiparametric high dimensional setting with likely even more nuisance parameters than a moderate dimensional setting [5, 36]. Moreover, Qu et al. [27] show the resulting estimators from minimizing this quadratic inference function are the most efficient estimators given the same class of estimating functions, which includes generalized estimating equations. This is advantageous since mis-specified working correlation structures may cause efficiency loss, yet the true correlation structure is not often known in practice. In addition, the QIF only requires the first two moments of the response distribution alleviating the difficulty of specifying the full joint likelihood for correlated discrete responses [27].

3. Penalized Quadratic Inference Function for Ultra-high Dimensional Data

Variable selection is an essential task, since over selecting variables can negatively impact estimation efficiency and inference for a sparse set of important variables. Nevertheless, identifying a sparse set of important covariates in high dimensions is difficult [8]. In our motivating example there are more than 50,000 SNPs and the correlated discrete response creates a further challenge to specify the joint likelihood. Thus, for concurrent variable selection and estimation with diverging and even potentially ultra-high dimensional covariates along with correlated discrete responses, we adopt the penalized quadratic inference function for the generalized partially linear single-index model in (1). We define our penalized quadratic inference function to minimize

$$Q_P(\boldsymbol{\alpha}) = Q_n(\boldsymbol{\alpha}) + \sum_{j=1}^{p_n} q_{\lambda_p}(|\beta_j|) + \sum_{k=1}^{q_n} q_{\lambda_q}(|\gamma_k|). \tag{3}$$

Here $Q_n(\boldsymbol{\alpha})$ refers to the QIF equation (2), $q_{\lambda_p}(|\beta_j|)$ with $j \in \{1, \dots, p_n\}$ is the penalty function for each single-index coefficient with corresponding tuning parameter λ_p. Similarly, $q_{\lambda_q}(|\gamma_k|)$ with $k \in \{1, \dots, q_n\}$ is the penalty function for each linear coefficient with corresponding tuning parameter λ_q.

While other penalty functions can be used, we implement the smoothly clipped absolute deviation (SCAD) penalty. The first derivative of the SCAD penalty function is defined as $\dot{q}_\lambda(\zeta) = \lambda\{I(\zeta \leq \lambda) + \frac{(a\lambda-\zeta)_+}{(a-1)\lambda} I(\zeta > \lambda)\}$, for $q_\lambda(0) = 0$ and $a > 2$ for a given regularization parameter λ. As suggested in Fan and Li [7], we set $a = 3.7$.

4. Asymptotic Properties

In our proposed semiparametric approach for longitudinal data, the total number of covariates can be ultra-high dimensional in both the nonparametric and partially linear portions. Additionally, the true important covariates, p_{sn} and q_{sn}, can be diverging. Especially for longitudinal data, few existing approaches incorporate diverging or even ultra-high dimensional covariates with nonlinearity for discrete responses. In this challenging setting, we establish important theoretical properties for the estimators of both the partially linear and, more importantly, the nonparametric single-index components, not only in moderately high dimension but even in ultra-high dimension.

We first establish asymptotic theory, namely, convergence rate and asymptotic normality, in the oracle case (i.e., when the exact true covariates are given ahead of time). In this case, we use subscript (s) to denote the oracle. That is, we let $\mathbf{X}_{(s)i}$ be the true $T_i \times p_{sn}$ dimensional single-index covariate matrix, and $\mathbf{Z}_{(s)i}$ be the true $T_i \times q_{sn}$ dimensional linear covariate matrix. The true non-zero p_{sn}-dimensional single-index parameter vector is $\boldsymbol{\beta}_{0(s)} = \{\beta_{01}, \dots, \beta_{0p_{sn}}\}^{\mathrm{T}}$, and the true non-zero q_{sn}-dimensional partially linear vector is $\boldsymbol{\gamma}_{0(s)} = \{\gamma_{01}, \dots, \gamma_{0q_{sn}}\}^{\mathrm{T}}$. The true non-zero spline parameters corresponding to the single-index $\mathbf{X}_{(s)i}\boldsymbol{\beta}_{0(s)}$ are $\boldsymbol{\theta}_{0(s)}$. Then for the oracle estimators, we let $\hat{\boldsymbol{\alpha}}_{(s)}$ refer to the oracle estimator for the true important parameters $\boldsymbol{\alpha}_{0(s)} = (\boldsymbol{\theta}_{0(s)}\ \boldsymbol{\beta}_{0(s)}\ \boldsymbol{\gamma}_{0(s)})$, and $\hat{\boldsymbol{\zeta}}_{(s)}$ refer to the estimate for $\boldsymbol{\zeta}_{0(s)} = \left(\boldsymbol{\beta}_{0(s)}^{(-1)}\ \boldsymbol{\gamma}_{0(s)}\right)$. Also, we let $\mathbf{V}_{0(s)i} = \left(\mathbf{G}\left(\mathbf{X}_{(s)i}\boldsymbol{\beta}_{0(s)}\right), \text{diag}\left\{\dot{\boldsymbol{\eta}}\left(\mathbf{X}_{(s)i}\boldsymbol{\beta}_{0(s)}\right)\right\}\mathbf{X}_{(s)i}\mathbf{J}\left(\boldsymbol{\beta}_{0(s)}\right), \mathbf{Z}_{(s)i}\right)$ with $\mathbf{J}(\boldsymbol{\beta}_{0(s)}) = \partial\boldsymbol{\beta}_{0(s)}/\partial\boldsymbol{\beta}_{0(s)}^{(-1)}$, and we define the true conditional mean of the response as $\boldsymbol{\mu}_{0(s)i} = g^{-1}(\boldsymbol{\eta}(\mathbf{X}_{(s)i}\boldsymbol{\beta}_{0(s)}) + \mathbf{Z}_{(s)i}\boldsymbol{\gamma}_{0(s)})$.

We use $\dot{f}, \ddot{f}$ to denote the first and second derivative of functions. In the theoretical study, we assume that $T_i \equiv T$ for simplicity, and we assume T is fixed. For two T-dimensional vectors $\mathbf{a}$ and $\mathbf{b}, \mathbf{a} \odot \mathbf{b}$ denotes the Hadamard product taken component-wise resulting in another T-dimensional vector. For a matrix $\mathbf{A}$ with T rows, $\mathbf{A} \odot \mathbf{a}$ is the matrix of the same size as $\mathbf{A}$ resulting from applying the Hadamard product to each column of $\mathbf{A}$. We assume in (A2) below that η is smooth. Under that assumption, there exists a vector $\boldsymbol{\theta}_0$ such that $\|\mathbf{G}(\cdot)\boldsymbol{\theta}_0 - \eta(\cdot)\|_\infty \le CH_n^{-d}$.

We rely on the following assumptions.

(A1) $\sup_{1\le i\le n, 1\le t\le T}\left\|\mathbf{X}_{(s)it}\right\| = O_p\left(\sqrt{p_{sn}}\right), \sup_{1\le i\le n, 1\le t\le T}\left\|\mathbf{Z}_{(s)it}\right\| = O_p\left(\sqrt{q_{sn}}\right)$, and $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ are i.i.d.

(A2) $\eta \in \mathcal{H}^d(M)$ for some $d \ge 2$ and a constant $M > 0$, where $\mathcal{H}^d$ contains all functions η such that $\left|\eta^{(d_1)}(x) - \eta^{(d_1)}(y)\right| \le M|x-y|^{d-d_1}, d_1$ is the largest integer strictly smaller than d and $\eta^{(d_1)}$ is the d_1-th order derivative of η. We also assume η is a bounded function.

(A3) $E\left[\left\|\mathbf{Y}_i - \boldsymbol{\mu}_{0(s)i}\right\|^{2+\delta}\right] < \infty$ for some $\delta > 0$.

(A4) There exists positive constants c_1 and c_2 such that $c_1 \le \lambda_{\min}\left(n^{-1}\sum_i \mathbf{V}_{0(s)i}^{\mathrm{T}}\mathbf{V}_{0(s)i}\right) \le$ $\lambda_{\max}\left(n^{-1}\sum_i \mathbf{V}_{0(s)i}^{\mathrm{T}}\mathbf{V}_{0(s)i}\right) \le c_2$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues of a matrix, respectively.

(A5) On the set $\left\{\boldsymbol{\alpha}_{(s)} : \left\|\boldsymbol{\alpha}_{(s)} - \boldsymbol{\alpha}_{0(s)}\right\| \le Cr_n\right\}$, where C is a positive constant, $\dot{\boldsymbol{\mu}}_i(\boldsymbol{\alpha})$, is uniformly bounded away from 0 and ∞, and $\ddot{\boldsymbol{\mu}}_i\left(\boldsymbol{\alpha}_{0(s)}\right)$ is uniformly bounded.

(A6) $\mathbf{M}_1 = \mathbf{I}, \mathbf{M}_2, \ldots, \mathbf{M}_m$ are linear independent non-negative definition matrices with bounded eigenvalues.

Further, for use in the proof of asymptotic normality, we define the following projection. Let $\mathcal{M}_t = \left\{g : Eg^2\left(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0\right) < \infty\right\}$. For any random vector $\mathbf{a} \in \mathbf{R}^T$ that is a function of $(\mathbf{X}_i, \mathbf{Z}_i)$, we define $E_{\mathcal{M}}[\mathbf{a}] = \mathbf{g}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right) = \left(g_1\left(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_0\right), \ldots, g_T\left(\mathbf{X}_{iT}^{\mathrm{T}}\boldsymbol{\beta}_0\right)\right)^{\mathrm{T}}$, where $\mathbf{g} = (g_1, \ldots, g_T)^{\mathrm{T}}$ is the minimizer of

$$E\left[\left(\mathbf{a} - \mathbf{g}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)^{\mathrm{T}}\boldsymbol{\Omega}\left(\mathbf{a} - \mathbf{g}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\right] \tag{4}$$

over $g_j \in \mathcal{M}_j, j = 1, \ldots, T$, where

$$\boldsymbol{\Omega} = \mathbf{F}_i\left(E\left[\mathbf{F}_i^{\mathrm{T}}\mathbf{R}\mathbf{F}_i\right]\right)^{-1}\mathbf{F}_i^{\mathrm{T}}$$

$\mathbf{F}_i = \left(\mathbf{A}_{0i}^{1/2}\mathbf{M}_1\mathbf{A}_{0i}^{1/2}\mathbf{V}_{0i}, \ldots, \mathbf{A}_{0i}^{1/2}\mathbf{M}_m\mathbf{A}_{0i}^{1/2}\mathbf{V}_{0i}\right), \mathbf{R} = \mathbf{A}_{0i}^{-1/2}E\left[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\mathrm{T}}\right]\mathbf{A}_{0i}^{-1/2}$, where $\mathbf{R}$ is the true correlation matrix of the error term.

Interpreting Ω as a weight matrix (in the non-longitudinal case it is just a 1×1 weight), the above can indeed be regarded as a projection. Using projections in the proof of asymptotic normality in the parametric part is an important technique in various semiparametric models [43]. This definition of projection can be extended to the case when $\mathbf{a}$ is a random matrix with T columns such that the projection is obtained row by row.

Using the defined projection, we write $E_{\mathcal{M}}\left[\mathbf{Z}_{(s)i}\right] = \left\{f_{tj}\left(\mathbf{X}_{(s)i}\boldsymbol{\beta}_{0(s)}\right)\right\}_{1\le t\le T, 1\le j\le q_{sn}}$ and $E_{\mathcal{M}}\left[\text{diag}\left(\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\mathbf{X}_i\right] = \left\{m_{tj}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right\}_{1\le t\le T, 1\le j\le p_{sn}}$. Define

$$\mathbf{U}_{0(s)i}^{\mathrm{T}} = \begin{pmatrix} \mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_{0(s)}\right)\mathbf{X}_{(s)i}^{\mathrm{T}}\,\text{diag}\left(\dot{\eta}\left(\mathbf{X}_{(s)i}\boldsymbol{\beta}_{0(s)}\right)\right) - E_{\mathcal{M}}\left[\mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_{0(s)}\right)\mathbf{X}_{(s)i}^{\mathrm{T}}\,\text{diag}\left(\dot{\eta}\left(\mathbf{X}_{(s)i}\boldsymbol{\beta}_{0(s)}\right)\right)\right] \\ \mathbf{Z}_{(s)i}^{\mathrm{T}} - E_{\mathcal{M}}\left[\mathbf{Z}_{(s)i}^{\mathrm{T}}\right] \end{pmatrix}$$

(A7) $f_{tj}, m_{tj} \in \mathcal{H}^{d'}$ for some $d' \ge 1$.

(A8) There exist positive constants c_3 and c_4 such that $c_3 \le \lambda_{\min}\left(n^{-1}\sum_i \mathbf{U}_{0(s)i}^{\mathrm{T}}\mathbf{U}_{0(s)i}\right) \le$ $\lambda_{\max}\left(n^{-1}\sum_i \mathbf{U}_{0(s)i}^{\mathrm{T}}\mathbf{U}_{0(s)i}\right) \le c_4$.

(A9) $\boldsymbol{\epsilon}_i$ is a sub-Gaussian random vector.

Theorem 1. *(Convergence rate of oracle estimator.) Under assumptions (A1)-(A6) and that* $\left(H_n^3 + H_n^2 p_{sn} + q_{sn}\right)r_n^2 \to 0$, *we have*

$$\left\|\widehat{\boldsymbol{\alpha}}_{(s)} - \boldsymbol{\alpha}_{0(s)}\right\| = O_p\left(r_n\right),$$

where $r_n = \sqrt{\left(H_n + p_{sn} + q_{sn}\right)/n} + H_n^{-d}$.

Theorem 2. *(Asymptotic normality of oracle estimator.) Under assumptions (A1)-(A8) and that* $n\left(H_n^3 + H_n^2 p_{sn} + q_{sn}\right)^{1/2} r_n^3 \to 0$*, then for any unit vector* $\mathbf{a} \in \mathbf{R}^{p_{sn}+q_{sn}-1}$*,*

$$\sqrt{n}\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{(s)}^{-1/2}\left(\widehat{\boldsymbol{\zeta}}_{(s)} - \boldsymbol{\zeta}_{0(s)}\right) \xrightarrow{d} N(0,1),$$

where $\boldsymbol{\Sigma}_{(s)} = E\left[\mathbf{U}_{0(s)}^{\mathrm{T}}\mathbf{F}_{0(s)}\right]\left(E\left[\mathbf{F}_{0(s)}^{\mathrm{T}}\mathbf{R}\mathbf{F}_{0(s)}\right]\right)^{-1}E\left[\mathbf{F}_{0(s)}^{\mathrm{T}}\mathbf{U}_{0(s)}\right]$*.*

Next, we establish asymptotic properties when the single-index covariates and the partially linear covariates are ultra-high dimensional for our semiparametric penalized quadratic inference function estimators.

Theorem 3. *(Asymptotic normality of PQIF estimator.) Under the same assumptions for Theorem 2 and (A9) and* $\left(\sqrt{\frac{(H_n^3+H_n^2 p_{sn}+q_{sn})log p_n}{n}} + \sqrt{H_n + p_{sn} + q_{sn}}\right) r_n << \lambda_p << \min_{j\le p_{sn}}\left|\beta_{0j}\right|$, $\sqrt{H_n + p_{sn} + q_{sn}}\left(1 + \sqrt{\frac{log q_n}{n}}\right) r_n << \lambda_q << \min_{j\le q_{sn}}\left|\gamma_{0j}\right|$*, there is an* r_n*-consistent local minimizer* $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ *such that for any unit vector* $\mathbf{a}$*,*

(i)

$$\sqrt{n}\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{(s)}^{-1/2}\left(\widehat{\boldsymbol{\zeta}}_{(s)} - \boldsymbol{\zeta}_{0(s)}\right) \xrightarrow{d} N(0,1),$$

where $\boldsymbol{\Sigma}_{(s)} = E\left[\mathbf{U}_{0(s)}^{\mathrm{T}}\mathbf{F}_{0(s)}\right]\left(E\left[\mathbf{F}_{0(s)}^{\mathrm{T}}\mathbf{R}\mathbf{F}_{0(s)}\right]\right)^{-1}E\left[\mathbf{F}_{0(s)}^{\mathrm{T}}\mathbf{U}_{0(s)}\right]$*.*

(ii) $\widehat{\beta}_{p_{sn}+1} = \cdots = \widehat{\beta}_{p_n} = \widehat{\gamma}_{q_{sn}+1} = \cdots = \widehat{\gamma}_{q_n} = 0$ *with probability approaching one.*

One key contribution we make is to establish the above challenging theoretical properties for ultra-high dimensional correlated discrete response data. Here we allow not only ultra-high dimensional partially linear covariates but importantly also the single-index covariates within the nonlinear flexible unknown function to be ultra-high dimensional. We select important variables that are potentially diverging in the generalized partially linear single-index model framework with PQIF functions. We relegate detailed technical proofs along with four lemmas to Section 9. We hope similar techniques can be adopted for other highly nonlinear ultra-high dimensional modeling.

5. Algorithm and Detailed Implementation

5.1. Algorithm

Estimation of the spline, single-index, and partial linear coefficients of penalized quadratic inference function (3) involves computational challenges in high dimensions, including the non-convex penalty function and the potential ultra-high dimensional covariates in both the nonparametric and linear portions of the generalized partially linear single-index model. We avoid the prohibitive computational burden of estimating all parameters in one step by implementing an iterative algorithm. Moreover, we apply two strategic approximations: a linear approximation of the unknown flexible function and a local quadratic approximation of the non-convex SCAD penalty.

We first focus on easing the computational burden of the potentially ultra-high dimensional covariates embedded inside the likely nonlinear, unknown function estimated nonparametrically. This nonlinear optimization over a high-dimensional space becomes a computationally demanding task. To alleviate this, we invoke a linear approximation of $\eta(\cdot)$ to convert this penalized nonlinear problem into a penalized linear problem. Then we can use existing linear algorithms, which are more computationally advantageous than nonlinear estimation especially in high dimensions. Specifically, we apply the first order Taylor series approximation of $\eta(\mathbf{X}_i\boldsymbol{\beta})$ at the point $\mathbf{X}_i\boldsymbol{\beta}_0$ as $\eta(\mathbf{X}_i\boldsymbol{\beta}_0) \cong \eta(\mathbf{X}_i\boldsymbol{\beta}_0) + diag(\dot{\eta}(\mathbf{X}_i\boldsymbol{\beta}_0))\mathbf{X}_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, where $\eta(\cdot)$ is the first derivative of the function $\dot{\eta}(\cdot)$. Notably, the parameters to be estimated $\boldsymbol{\beta}$ are now outside of the unknown, flexible function. We can now use the linear penalized quadratic inference function method as in Cho and Qu [5] to estimate the parameters.

Still, even though the problem is now linear, minimization of the linear penalized quadratic inference function is difficult due to the non-convex SCAD penalty. Following Cho and Qu [5], we approximate the linear penalized quadratic inference function by implementing a local quadratic approximation of the SCAD penalty and then minimize this approximated quadratic function using the Newton-Rhapson algorithm. Specifically, on iteration k of the iterative algorithm, we let the column coefficient vector $\boldsymbol{\zeta}^{(k)} = (\boldsymbol{\beta}^{(k)}\ \boldsymbol{\gamma}^{(k)})$. Then as in Fan and Li [7], the local quadratic approximation of the SCAD penalty function for iteration k and coefficient $\zeta_j^{(k)}$ is $q_\lambda(|\zeta_j|) \cong q_\lambda(|\zeta_j^{(k)}|) + \frac{1}{2}\frac{\dot{q}_\lambda(|\zeta_j^{(k)}|)}{|\zeta_j^{(k)}|}(\zeta_j^2 - \zeta_j^{(k)2})$ with $\zeta_j \approx \zeta_j^{(k)}$ and $\zeta_j^{(k)} \neq 0$. See Cho and Qu [5] for the full estimation algorithm for the linear penalized quadratic inference function problem.

Altogether, our two-step iterative algorithm consists of, in step one, estimating the spline coefficients $\boldsymbol{\theta}$ via a quadratic inference function in conjunction with a polynomial B-spline basis. In this step, the single-index and linear coefficient estimates $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}$ are given from the previous iteration or as initial values during the first iteration of the algorithm. In the second step, given the estimated spline coefficients $\hat{\boldsymbol{\theta}}$, we estimate the linear and single-index coefficients $\boldsymbol{\beta}, \boldsymbol{\gamma}$ via the linear approximation of $\eta(\cdot)$, which converts the nonlinear optimization to a linear penalized quadratic inference function method from Cho and Qu [5]. In this step, the conditional mean with $\eta(\cdot)$ estimated by polynomial splines $g^{-1}(\mathbf{G}(\mathbf{X}_i\boldsymbol{\beta})\hat{\boldsymbol{\theta}} + \mathbf{Z}_i\boldsymbol{\gamma})$ becomes $g^{-1}(\mathbf{G}(\mathbf{X}_i\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\theta}} + \text{diag}(\dot{\mathbf{G}}(\mathbf{X}_i\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\theta}})\mathbf{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathbf{Z}_i\boldsymbol{\gamma})$ upon implementation of the linear approximation of $\eta(\mathbf{X}_i\boldsymbol{\beta})$ at the point $\mathbf{X}_i\hat{\boldsymbol{\beta}}$. We continue this process of first estimating the spline coefficients and then estimating the linear and single-index coefficients until convergence.

A detailed description of the iterative algorithm is the following:

Step 0: Initialize $\hat{\boldsymbol{\zeta}}^{(0)} = \begin{pmatrix} \hat{\boldsymbol{\beta}}^{(0)} \\ \hat{\boldsymbol{\gamma}}^{(0)} \end{pmatrix}$. See Section 5.2 for details on obtaining initial values under various scenarios.

Step 1: Given $\hat{\boldsymbol{\zeta}}^{(k-1)} = \begin{pmatrix} \hat{\boldsymbol{\beta}}^{(k-1)} \\ \hat{\boldsymbol{\gamma}}^{(k-1)} \end{pmatrix}$, estimate the spline coefficients, $\hat{\boldsymbol{\theta}}^{(k-1)}$, by minimizing the quadratic inference function $\mathbf{g}_n^{\mathrm{T}}\mathbf{W}_n^{-1}\mathbf{g}_n$, where $\mathbf{g}_n = \frac{1}{n}\sum_i \mathbf{g}_i(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}^{(k-1)}, \hat{\boldsymbol{\gamma}}^{(k-1)})$.

Step 2: Given the estimated spline coefficients $\hat{\boldsymbol{\theta}}^{(k-1)}$, the k_{th} penalized estimator of $\hat{\boldsymbol{\zeta}}^{(k)} = (\hat{\boldsymbol{\beta}}^{(k)}\ \hat{\boldsymbol{\gamma}}^{(k)})$ is determined by minimizing the penalized quadratic inference function

$$Q_n(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma}) + \sum_{j=1}^{p_n} q_{\lambda_p}(|\beta_j|) + \sum_{k=1}^{q_n} q_{\lambda_q}(|\gamma_k|), \tag{5}$$

where $Q_n(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbf{g}_n^{\mathrm{T}}\mathbf{W}_n^{-1}\mathbf{g}_n$ and $\mathbf{g}_n = \frac{1}{n}\sum_i \mathbf{g}_i(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. We also assume the identifiability constraints $||\boldsymbol{\beta}|| = 1$ and $\beta_1 > 0$.

Using the approach from Cho and Qu [5], we estimate the single-index and partially linear parameters $\boldsymbol{\beta}, \boldsymbol{\gamma}$ by taking a quadratic approximation of the penalized quadratic inference function with the SCAD penalty and invoking the Newton-Rhapson algorithm. See Cho and Qu [5] for further implementation details. We then repeat steps 1 and 2 until convergence. In practice, we observe that the algorithm converges in around 3 steps.

5.2. Practical Implementation Information for Algorithm

In this section, we describe the practical implementation aspects that contribute to the performance of our approach and give suggestions for desirable performance. We must decide on an appropriate linear combination of basis matrices to use for the quadratic inference function, the degree, number, and placement of knots for the B-spline basis for univariate smoothing, and the penalty parameter for variable selection. Further implementation decisions include initial value determination and screening method selection to prohibit extensive computational burden when incorporating ultra-high dimensional covariates. We provide additional studies in the Web Appendix.

Choice of Basis Matrices for Quadratic Inference Function

To approximate the inverse of the working correlation structure for the quadratic inference function, one must select a linear combination of basis matrices $\mathbf{M}_i$ in (2) and (3). As noted in Qu et al. [27], if the correlation structure is exchangeable then $\mathbf{R}^{-1}$ can be approximated by $a_1\mathbf{I} + a_2\mathbf{M}_2$, where $\mathbf{M}_2$ has 1 on the off-diagonal and 0s on the diagonal. Here $a_1 = -(m-2)\rho + 1/l_1$ and $a_2 = \rho/l_1$ with $l_1 = (m-1)\rho^2 - (n-2)\rho - 1$. For the AR(1) case, $\mathbf{R}^{-1}$ can be approximated by $a_1\mathbf{I} + a_2\mathbf{M}_2 + a_3\mathbf{M}_3$, where $\mathbf{M}_2$ has 0 everywhere except the two main off-diagonals have 1s, and $\mathbf{M}_3$ has 0 everywhere except at $(1, 1)$ and (T, T), which are 1s where T is the largest time point. $a_1 = (1+\rho^2)/l_2$, $a_2 = -\rho/l_2$, and $a_3 = -\rho^2/l_2$ with $l_2 = 1 - \rho^2$. When no previous information exists about the possible correlation structure, one may implement the approach from Zhou and Qu [44] to consistently select the correlation structure by employing informative basis matrices in a sufficient class of the true structure. More information on further working correlation structures can be found in [26, 27, 44].

Initial Values and Screening

In ultra-high dimensions, the computational time of most approaches tends to be prohibitive for practical usage [8]. For this reason, screening is commonly used in relevant literature to rapidly reduce the covariate dimension in practice (e.g., Cai et al. [3] and Fang et al. [9]). We echo Fan and Lv [8] and incorporate sure independence screening to first efficiently reduce dimensionality to benefit step 2 of our iterative algorithm: the linear penalized quadratic inference function.

On the other hand, in moderately high dimensions, the estimates from the linear penalized quadratic inference function can be obtained in a satisfactory amount of time for practical usage without implementing a screening approach. In the relevant literature, initial values with good properties are often used in moderately high dimensions.

For example, Wang et al. [35] use linear GEE initials. One may also use sure independence screening estimates at the original dimension as initial values [8]. For our approach in moderately high dimensions, we may use linear quadratic inference function estimates as initial values.

Tuning Parameters for Penalization

As with many optimization problems, tuning parameter selection is vital for desirable variable selection performance. One must find a suitable value for the tuning parameter of the penalty function λ_p for the single-index parameters and λ_q for the linear parameters of the model. Given that the role of the quadratic inference function is analogous to negative twice the log likelihood, we can use the high dimensional Bayesian information criterion (HBIC) for tuning parameter selection as in Wang et al. [34]. The model selection criteria is

$$HBIC(\lambda_p, \lambda_q) = Q_n(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}_{\lambda_p}, \hat{\boldsymbol{\gamma}}_{\lambda_q}) + d_\lambda \frac{log(n)}{2n} C_n, \tag{6}$$

where d_λ is the number of important nonzero coefficients from both the single-index and partially linear portions of the fitted model. C_n is considered as $log(log(p_n + q_n))$ and is based on empirical evidence as in the literature Wang et al. [34]. We find the minimum HBIC using grid search over separate λ_p and λ_q values, which provides less restrictions in modeling.

Spline Smoothing Tuning Parameters

For the spline basis, one must select the number, degree, and placement of knots. Selection criteria for the number of interior knots for the B-spline basis is based on a variety of approaches. As explained in Ma et al. [21], the number of knots can be chosen by a BIC type criteria focusing on consistency or AIC and cross-validation approaches when efficiency is of interest [15]. In Ruppert and Carroll [28] and Yu and Ruppert [40], the number of knots implemented depends on the characteristics and shape of the function to be estimated. In particular, monotonicity and discontinuity dictate the number of knots that are needed in an empirical analysis.

Placement of knots is usually based on equally spaced knots or knot placement at the quantiles of the support of the estimated single-index. We note that during the iterative algorithm the estimated index values and consequently the knot placement may change in practice. Further, for best performance, a knot must be near any discontinuities. We find equally spaced knots with 2 interior knots perform well in our simulation studies. Huang et al. [16] claim that when the number of knots does not greatly influence performance, one can fix the amount of knots. Regarding the choice of degree, in practice, usually quadratic and cubic splines are implemented [21, 41].

6. Simulation Studies

We demonstrate the estimation and variable selection performance of our generalized partially linear single-index model using the penalized quadratic inference function through simulation studies. We mainly focus our investigation on a correlated binary response example in very high dimension. We also examine our approach under imbalanced data and a complex correlation structure. We present additional settings and a continuous response example in the Web Appendix. For both the single-index and linear covariates, "Correct%" is the percentage of simulation replications that select only the relevant variables, "TN" represents the true negatives, the number of true zero coefficients that the model sets to zero, and "FN" represents the false negatives, the number of true nonzero coefficients that are set falsely to zero. For the estimation results, "MSEp" is the average mean squared error of the single-index coefficient estimates $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ and "MSEq" is the same for the partial linear coefficients $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|^2$.

Binary Response Example.

We consider the correlated binary responses from the marginal mean

$$log\left(\frac{p_{it}}{1 - p_{it}}\right) = \sin\frac{((\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta} - a)\pi)}{b - a} + \mathbf{Z}_{it}^{\mathrm{T}}\boldsymbol{\gamma}.$$

There are $n = 400$ subjects with $T_i = T = 10$ time points for all subjects, and $p_n + q_n = 500$ and $p_n + q_n = 5000$ with $p_n = q_n$. The coefficient vector is $\boldsymbol{\beta}_0 = (1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}/\sqrt{3}$ and $\boldsymbol{\gamma}_0 = (1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$, and a and b are constants equal to $\sqrt{3}/2 - 1.645/\sqrt{12}$ and $\sqrt{3}/2 + 1.645/\sqrt{12}$ respectively. The covariates $\mathbf{X}_{it}$ are sampled from an independent uniform distribution with minimum at 0 and maximum at 0.5, and the covariates $\mathbf{Z}_{it}$ are sampled from an independent normal distribution with mean at 0 and standard deviation of 0.5. We use the method described in Macke et al. [22] to simulate correlated binary data with exchangeable correlation structure with $\rho = 0.2$. We evaluate the performance using independence, AR(1), and exchangeable working correlation matrices.

Table 1 reports the estimation and variable selection results, showing that our GPLSIM PQIF model selects the true important variables with a high percentage over 200 simulation replications. On average, the model also correctly accounts for the number of true negatives of the covariates. In particular, when $p_n + q_n = 500$, the number of true negatives is very close to the true amount of 247 for the single-index covariates, and the number of true negatives is close to the true amount of 247 for the linear covariates in all scenarios. Similarly, in the $p_n + q_n = 5000$ case, the number of true negatives is close to the true amount of 2497 for the single-index case, and the number of true negatives is also close to the true amount of 2497 for the linear covariates. Moreover, the number of true important single-index or linear variables under selected is low, since the number of false negatives in the $p_n + q_n = 500$ case is 0, and it is below 5% in all instances of the $p_n + q_n = 5000$ case. Here the model selection performance under various correlation structures is quite similar. However, the performance is better in almost all cases under the true exchangeable correlation structure.

In terms of estimation, the MSEs in Table 1 are small for all parameter estimates. Further, Table 2 indicates that all parameter estimates are close to the true parameter values. As is the case with variable selection for this example, the parameter estimation performance under various correlation structures is quite similar. However, it is better in almost all cases under the true exchangeable correlation structure.

Complex Correlation Structure. We further investigate a more complex correlation structure, which is a mixture of an AR(1) correlation structure and an exchangeable correlation structure with $\rho = 0.5$ and $p_n = 250$ and $q_n = 250$. In particular, the correlation structure simulated is $Corr(Y_{it}, Y_{ik}) = (\rho^{|t-k|}/2 + \rho/2)$ with $t, k = 1, \cdots, T$ when $t \neq k$; and $Corr(Y_{it}, Y_{ik}) = 1$ when $t = k$ for each subject i. Table 3 indicates that our proposed model, labeled as "GPLSIM SCAD", yields good selection results for an underlying complex correlation structure even when the working correlation matrix is misspecified. This is consistent with similar literature, e.g. Qu et al. [27], that QIF can yield a consistent estimator under a misspecified working correlation matrix.

We further examine the estimation and variable selection results with a popular LASSO penalty [32] in comparison to SCAD penalty. Table 3 indicates when using the LASSO penalty for all correlation structures, the GPLSIM PQIF model tends to over-select variables that are not important in the single-index portion. In addition, the parameter estimates for the SCAD penalty are closer to the true parameter values compared to those using LASSO penalty. This is similarly observed for the linear PQIF model, where for all correlation structures the LASSO penalty appears to perform worse in two areas: non-important covariates are selected in the single-index portion and all parameter estimates are farther from the true values.

Unequal T_i per Subject. Next we investigate unequal T_i per subject for imbalanced data. We simulate 400 subjects with 10 observations per subject with $p_n + q_n = 500$, and 400 subjects with 5 observations per subject for $p_n + q_n = 500$. The remaining set is the same as in the binary response example with a simulated exchangeable correlation structure and $\rho = 0.5$. The results in Table 4 indicate that the estimated parameters are close to the true parameters. The variable selection results have a high correct percentage, with a range of 97-100% for the single-index parameters and near 100% for the linear parameters. Our algorithm uses the high-dimensional linear penalized QIF from Cho and Qu [5] and linear QIF from Qu et al. [27] as a base, thus incorporating imbalanced data is natural.

7. An Application to Diabetes Analysis

Diabetes is a widespread disease associated with various health complications including but not limited to an increased risk of stroke and vision loss [33]. Due to the known impact of both phenotype and high dimensional genotype variables on diabetes [10], identifying important genetic and phenotype risk factors may allow early diagnosis and more effective treatment. To investigate this relationship between risk factors and diabetes status using our proposed model, we use the ongoing Framingham Heart Study data [6]. This study is a continuing longitudinal study of cardiovascular disease, and researchers have also used this data to investigate various diseases such as diabetes (e.g., Meigs et al. [23]). For more in-depth details on the phenotype and genotype variables we used, please visit the Framingham study page at `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v32.p13`.

To investigate factors related to diabetes status, we use a subset of 878 participants among the offspring cohort of the Framingham data measured over four exams. To avoid a large gap between exam 1 and exam 2, we use a subset that covers four waves of exams to ensure the diabetes indicators and diabetes-related quantitative traits are comparable. According to the Framingham Data dictionary, the participants' diabetes status variable, the correlated binary response $\mathbf{Y}_i$, is derived by an algorithm considering blood glucose test results, treatment status, and other information per the protocol of vr_diab_ex09_1_1002s. We include ten phenotype covariates that are potentially linked to diabetes from previous research: age, systolic blood pressure (SBP), total cholesterol (TC), smocking status (SMK), cigarette per day (CPD), body mass index (BMI), weight (WGT), Ventricular rate

(VENT_RT), Triglyceride(TRIG), and HDL cholesterol. We also include high dimensional genetic data from participants as covariates in our analysis. Participants were genotyped on the Affymetrix 500K creating 500,568 single nucleotide polymorphisms (SNPs). After removing SNPs that are vastly missing or possessing zero variance, quality control filters yielded 53,722 SNPs for modeling. For computational efficiency, we also follow the common practice and employ a combined screening method as in [11] and [17].

We apply our proposed generalized partial linear single-index model with PQIF approach to this subset of Framingham data to investigate the relationship between the longitudinal response between phenotypes and SNPs. The logit link function is used for the binary response, diabetes status. The only binary phenotype variable is smoking status, which naturally goes into the partial linear term when fitting the model while the rest of the phenotype variables and SNPs are embedded in the single-index term. Specifically, for participant i over 4 waves, $log(p_{it}/(1-p_{it})) = \eta(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}) + Z_{it}\gamma$, where all phenotype and genotype variables except smoking status enter the single-index term, and smoking status (Z_{it}) enters partially linearly. We find the best penalty parameter using HBIC, and we use quadratic degree with 3 equally spaced knots. See Section 5.2 for more information on tuning parameters and setup.

Figure 1 indicates a clear nonlinear relationship between the flexible function and the single-index made up of genetic and phenotype risk factors that were selected as important using our approach. A linear model is likely to mis-specify the relationship. For comparison, we also report the linear SCAD penalized quadratic inference function along with our proposed generalized partially linear single-index model using the SCAD penalized quadratic inference function. For the linear model, the same covariates are included. We use the BIQIF from Cho and Qu [5] to determine the best performing penalty parameter for the linear model.

Table 6 reports the 3-fold cross-validation area under the curve (AUC) and the out-of-sample model AUC with a 70/30 training and testing split of the data. Under the exchangeable structure, our proposed GPLSIM model clearly outperforms the linear penalized QIF model in terms of higher AUC for both cross validation and out-of-sample testing. Under the correlation structure of AR(1) and independence, although with less margin, our proposed partially linear single-index model using penalized QIF still consistently outperforms the linear penalized QIF model.

Table 6 also reports the number of phenotype variables selected and the number of genotype variables selected by the model. Under the exchangeable structure, 5 phenotype variables have been selected. They are age, SBP, WGT, TRIG and HDL. All of them have been identified as risk factors for diabetes in the previous literature and have been used as key components in predictive models of incidence of diabetes mellitus such as [30] and [20]. Among the large amount of genotype variables, the proposed model selected six SNPs, three of them have been confirmed by literature: rs4506565, rs10946398 and rs5018648 (Diabetes Genetics Initiative (2007); [1, 24, 25]).

8. Conclusion

In public health, researchers are increasingly conducting large-scale longitudinal studies to investigate the relationship between disease and genetic and phenotype factors. These studies can provide insight into more effective treatment and disease prevention strategies. To incorporate correlation with a discrete response and to account for the complexity and synergy among genetic factors and phenotype variables, we propose an approach via penalized quadratic inference functions for generalized partially linear single-index models. Specifically, the quadratic inference functions can yield efficient estimation when the working correlation structure is misspecified, and the generalized partially linear single-index models are flexible and can incorporate some interactions. We allow genetics factors that are diverging and even in the exponential order with the number of participants not only in the linear portion of the model, but importantly also in the nonlinear portion estimated non-parametrically. We establish theoretical results such as asymptotic normality and the oracle property in ultra-high dimension. Moreover, we develop an efficient estimation algorithm for computational expediency. We employ our approach to investigate diabetes status for an ongoing longitudinal public health study with genetics factors in very high dimensions.

Table 1: Summary of Variable Selection Results for the Generalized Partially linear Single-Index Model for the Binary Response Example. The total numbers of covariates are $p_n + q_n = 500$ and $p_n + q_n = 5000$ with $p_n = q_n$ calculated over 200 simulations. “Correct%” is the percentage of times the true important variables are selected over the iterations. “TN” is the average of the true negatives over the iterations, and “FN” is the average of the false negatives over the iterations. “MSEp” and “MSEq” are the average mean squared errors for the single-index parameters and partially linear parameters over all simulation iterations.

		Single-index Covariates				Partially Linear Covariates			
	Structure	Correct%	TNs	FNs	MSEp	Correct%	TNs	FNs	MSEq
$\boldsymbol{p_n + q_n = 500}$									
	Independence	86	246.81	0	0.0240	100	247	0	0.0180
	AR(1)	92	246.92	0	0.0120	100	247	0	0.0160
	Exchangeable	99	246.98	0	0.0080	100	247	0	0.0130
$\boldsymbol{p_n + q_n = 5000}$									
	Independence	80	2496.91	0.12	0.0586	93	2496.93	0	0.0216
	AR(1)	79	2496.94	0.15	0.0658	95	2496.95	0	0.0195
	Exchangeable	97	2497.00	0.03	0.0179	98	2496.98	0	0.0156

Table 2: Summary of Parameter Estimates for the Generalized Partially linear Single-Index Model for the Binary Response Example. The total numbers of covariates are $p_n + q_n = 500$ and $p_n + q_n = 5000$ with $p_n = q_n$. The sample mean, bias, and standard error are calculated over 200 simulations for single-index and partially linear parameter estimates.

		Independence			AR(1)			Exchangeable		
	par.	mean	bias	se	mean	bias	se	mean	bias	se
$\boldsymbol{p_n + q_n = 500}$										
	β_1	0.5671	-0.0103	0.0590	0.5739	-0.0035	0.0544	0.5757	-0.0017	0.0484
	β_2	0.5659	-0.0114	0.0630	0.5731	-0.0042	0.0519	0.577	-0.0004	0.0504
	β_3	0.5780	0.0006	0.0602	0.5746	-0.0027	0.0557	0.5722	-0.0052	0.0520
	γ_1	1.0024	0.0024	0.0778	1.0156	0.0156	0.0734	1.0171	0.0171	0.0625
	γ_2	0.9997	-0.0003	0.0772	1.0123	0.0123	0.0703	1.0168	0.0168	0.0654
	γ_3	0.9938	-0.0062	0.0812	1.0098	0.0098	0.0709	1.0108	0.0108	0.0659
$\boldsymbol{p_n + q_n = 5000}$										
	β_1	0.5855	0.0081	0.0746	0.5626	-0.0148	0.1359	0.5778	0.0005	0.0551
	β_2	0.5241	-0.0533	0.1694	0.5374	-0.0400	0.1729	0.5556	-0.0217	0.1098
	β_3	0.5718	-0.0056	0.1216	0.5751	-0.0022	0.1189	0.5831	0.0057	0.0497
	γ_1	1.0022	0.0022	0.0833	1.0144	0.0144	0.0798	1.0165	0.0165	0.0761
	γ_2	1.0125	0.0125	0.0786	1.0190	0.0190	0.0717	1.0228	0.0228	0.0625
	γ_3	1.0017	0.0017	0.0811	1.0052	0.0052	0.0808	1.0091	0.0091	0.0681

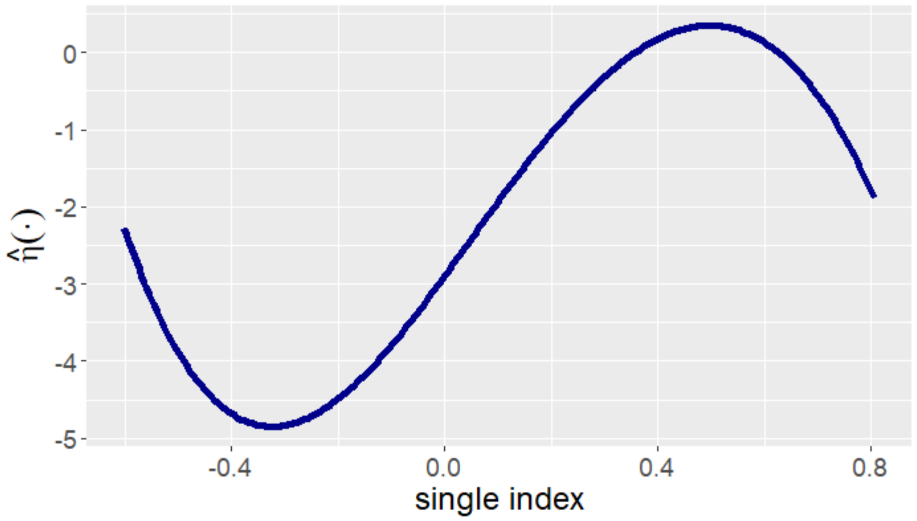

Figure 1: Curve Estimates for Real Data Application to Diabetes Analysis. We apply penalized quadratic inference function approach to generalized partial linear single-index model for longitudinal data in high dimension. Here the response variable of interest is diabetes status, a binary correlated discrete response, and all continuous phenotype and genotype variables are embedded in the single-index term.

Table 3: Summary of Estimation and Variable Selection Results for the Generalized Partially linear Single-Index Model for the Binary Response Example with Complex Correlation with Different Penalties for $p_n = 250$ and $q_n = 250$. The true correlation structure is $Corr(Y_{it}, Y_{ik}) = (\rho^{|t-k|}/2 + \rho/2)$ with $t, k = 1, \cdots, T$ when $t \neq k$, $\rho = 0.5$, and $Corr(Y_{it}, Y_{ik}) = 1$ when $t = k$ for each subject i. The “GPLSIM SCAD” and “GPLSIM LASSO” refer to our proposed generalized partially linear single-index model using the penalized quadratic inference function with SCAD penalty and LASSO penalty respectively. Similarly, The “Linear PQIF SCAD” and “Linear PQIF LASSO‘ refer to linear penalized quadratic inference function with SCAD penalty and LASSO penalty respectively. The remaining settings are the same as in the binary response example. “TN” is the average of the true negatives over the iterations, and “FN” is the average of the false negatives over the iterations. “MSEp” and “MSEq” are the average mean squared errors for the single-index parameters and partially linear parameters over all simulation iterations.

		Single-index Covariates				Partially Linear Covariates			
Model	Structure	Correct%	TNs	FNs	MSEp	Correct%	TNs	FNs	MSEq
GPLSIM SCAD	Independence	93	246.93	0	0.0169	100	247	0	0.0143
	AR(1)	88	246.88	0	0.0142	99	246.99	0	0.0106
	Exchangeable	96	246.96	0	0.0127	100	247	0	0.0123
GPLSIM LASSO	Independence	87	246.74	0.13	0.0929	86	246.85	0	0.1298
	AR(1)	79	246.33	0.18	0.1805	100	247	0	0.0976
	Exchangeable	89	246.71	0.09	0.0868	99	247	0	0.0840
Linear PQIF SCAD	Independence	8	245	0	0.5370	78	246.85	0	0.0177
	AR(1)	32	246.10	0.78	0.5488	90	247	0	0.0100
	Exchangeable	28	246.41	0.98	0.5246	88	246.80	0	0.0138
Linear PQIF LASSO	Independence	2	246.09	1.37	0.7846	29	245.81	0	0.0652
	AR(1)	22	245.98	0.42	0.3213	52	246.29	0	0.0784
	Exchangeable	55	246.68	0.34	0.2427	50	246.30	0	0.0654

Table 4: Summary of Estimation and Variable Selection Results for the Binary Response Example with Unequal T_i with $p_n = 250$ and $q_n = 250$. This analysis with unequal numbers of observations per subject T_i has $n = n_1 + n_2$ subjects. In particular, $n_1 = 400$ subjects with $T = 10$ observations per subject and $n_2 = 400$ subjects with $T = 5$ observations per subject. The remaining settings are the same as in the binary response example. “Correct%” is the percentage of times the true important variables are selected over the iterations. “TN” is the average of the true negatives over the iterations, and “FN” is the average of the false negatives over the iterations.

	Single-index Covariates				Partially Linear Covariates			
Structure	Correct%	TNs	FNs	MSEp	Correct%	TNs	FNs	MSEq
Independence	97	246.96	0	0.0175	100	247	0	0.0174
AR(1)	100	247	0	0.0089	100	247	0	0.0142
Exchangeable	100	247	0	0.0092	100	247	0	0.0168

Table 5: Summary of Parameter Estimates for the Continuous Response Example. The total numbers of covariates are $p_n + q_n = 5000$ with $p_n = q_n$. The sample mean, bias, and standard error are calculated over 200 simulations for the single-index and partially linear parameter estimates.

	Independence			AR(1)			Exchangeable		
par.	mean	bias	se	mean	bias	se	mean	bias	se
β_1	0.4440	-0.0032	0.0241	0.4457	-0.0015	0.0194	0.4459	-0.0013	0.0150
β_2	0.4499	0.0027	0.0200	0.4516	0.0044	0.0139	0.4492	0.0020	0.0136
β_3	0.4474	0.0002	0.0206	0.4478	0.0006	0.0163	0.4482	0.0009	0.0134
β_4	0.4458	-0.0014	0.0217	0.4448	-0.0024	0.0177	0.4473	0.0000	0.0155
β_5	0.4464	-0.0008	0.0204	0.4445	-0.0027	0.0151	0.4444	-0.0028	0.0144
γ_1	0.9944	-0.0056	0.0386	0.9967	-0.0033	0.0283	0.9980	-0.0020	0.0226
γ_2	1.0000	0.0000	0.0410	1.0026	0.0026	0.0308	1.0026	0.0026	0.0243
γ_3	0.9991	-0.0009	0.0456	1.0007	0.0007	0.0337	1.0049	0.0049	0.0270

Table 6: Summary of Results for Real Data Application to Diabetes Analysis. "CV AUC" and "OOS AUC" are cross validation model area under the curve with 3-fold cross validation and out-of-sample model area under the curve using 70/30 training testing split data set. "PQIF-GPLSIM" refers to our proposed generalized partial linear single-index model using penalized QIF, and "PQIF-linear" refers to the linear penalized QIF model. The numbers of selected phenotype and genotype variables are reported with full data.

		OOS AUC	CV AUC	# of phenotype selected	# of gene selected
PQIF-GPLSIM	Exchangeable	0.809	0.819	5	6
	AR(1)	0.801	0.794	6	7
	Independence	0.802	0.799	4	9
PQIF-Linear	Exchangeable	0.786	0.787	4	7
	AR(1)	0.773	0.781	5	7
	Independence	0.781	0.783	6	9

9. Technical Proofs

We provide detailed proofs and supporting lemmas for the asymptotic properties of estimators in ultra-high dimension for our proposed generalized partially linear single-index model using the penalized quadratic inference function. In Section 9.1, we first prove Theorem 1, the convergence rate under the oracle setting. In Section 9.2, we provide the proof of Theorem 2, asymptotic normality, in the oracle setting. In Section 9.3, we prove Theorem 3 which determines asymptotic properties for the penalized quadratic inference function estimator when both the single-index and linear variables can diverge and even be in ultra-high dimension. In both the oracle and the ultra-high dimensional settings, the important variables can diverge for both the partially linear portion and the single-index portion. For simplicity of notation, we drop all n subscript in this section.

9.1. Proof of Convergence Rate for Oracle Estimators

Let $\mathbf{V}_i(\boldsymbol{\alpha}) = \left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}\right), \mathrm{diag}\left\{\dot{\mathbf{G}}\left(\mathbf{X}_i\boldsymbol{\beta}\right)\boldsymbol{\theta}\right\}\mathbf{X}_i\mathbf{J}(\boldsymbol{\beta}), \mathbf{Z}_i\right), \mathbf{V}_{0i} = \left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \mathrm{diag}\left\{\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right\}\mathbf{X}_i\mathbf{J}\left(\boldsymbol{\beta}_0\right), \mathbf{Z}_i\right)$, $\mathbf{K}_{i\ell}(\boldsymbol{\alpha}) = \mathbf{V}_i^{\mathrm{T}}(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}(\boldsymbol{\alpha})\mathbf{M}_\ell\mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}), \mathbf{K}_{0i\ell} = \mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}, \mathbf{K}_i(\boldsymbol{\alpha}) = \left(\mathbf{K}_{i1}^{\mathrm{T}}(\boldsymbol{\alpha}), \ldots, \mathbf{K}_{im}^{\mathrm{T}}(\boldsymbol{\alpha})\right)^{\mathrm{T}}$, and $\mathbf{K}_{0i} = \left(\mathbf{K}_{0i1}^{\mathrm{T}}, \ldots, \mathbf{K}_{0im}^{\mathrm{T}}\right)^{\mathrm{T}}$. Define $\mathbf{g}_i(\boldsymbol{\alpha}) = \mathbf{K}_i(\boldsymbol{\alpha})\left(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right), \mathbf{g}_{0i}(\boldsymbol{\alpha}) = \mathbf{K}_{0i}\left(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right)$, $\mathbf{g}_{0i} = \mathbf{K}_{0i}\boldsymbol{\epsilon}_i$ with $\boldsymbol{\epsilon}_i = \mathbf{Y}_i - \boldsymbol{\mu}_{0i}, \overline{\mathbf{g}}_n(\boldsymbol{\alpha}) = \frac{1}{n}\sum_i \mathbf{g}_i(\boldsymbol{\alpha}), \overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_i \mathbf{g}_{0i}(\boldsymbol{\alpha}), \overline{\mathbf{g}}_{0n} = \frac{1}{n}\sum_i \mathbf{g}_{0i}, \mathbf{W}_n(\boldsymbol{\alpha}) = \frac{1}{n}\sum_i \mathbf{g}_{0i}(\boldsymbol{\alpha})\mathbf{g}_{0i}^{\mathrm{T}}(\boldsymbol{\alpha}), \mathbf{W}_{0n} = \frac{1}{n}\sum_i \mathbf{g}_{0i}\mathbf{g}_{0i}^{\mathrm{T}}$, $Q_n(\boldsymbol{\alpha}) = \overline{\mathbf{g}}_n(\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{W}_n^{-1}(\boldsymbol{\alpha})\overline{\mathbf{g}}_n(\boldsymbol{\alpha})$, and $Q_{0n}(\boldsymbol{\alpha}) = \overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{W}_{0n}^{-1}\overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha})$.

Let $r_n = \sqrt{\frac{H+p_s+q_s}{n}} + H^{-d}$. To get the convergence rate, we need to show that for L sufficiently large, with probability approaching one as $n \to \infty$,

$$\inf_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|=Lr_n} Q_n(\boldsymbol{\alpha}) - Q_n\left(\boldsymbol{\alpha}_0\right) \geq Cr_n^2. \tag{7}$$

The above will be implied by (8) and (9) below,

$$\sup_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|\leq Cr_n} \left|Q_n(\boldsymbol{\alpha}) - Q_{0n}(\boldsymbol{\alpha})\right| = o_p\left(r_n^2\right), \tag{8}$$

and for L sufficiently large,

$$\inf_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|=Lr_n} Q_{0n}(\boldsymbol{\alpha}) - Q_{0n}\left(\boldsymbol{\alpha}_0\right) \geq CL^2r_n^2. \tag{9}$$

As a first step we prove (8).

$$\begin{aligned}
&Q_n(\boldsymbol{\alpha}) - Q_{0n}(\boldsymbol{\alpha})\\
=&\overline{\mathbf{g}}_n(\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{W}_n^{-1}(\boldsymbol{\alpha})\overline{\mathbf{g}}_n(\boldsymbol{\alpha}) - \overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha})\mathbf{W}_{0n}^{-1}\overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha})\\
=&\left(\overline{\mathbf{g}}_n(\boldsymbol{\alpha}) - \overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha})\right)^{\mathrm{T}}\mathbf{W}_n^{-1}(\boldsymbol{\alpha})\overline{\mathbf{g}}_n(\boldsymbol{\alpha}) + \overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha})^{\mathrm{T}}\left(\mathbf{W}_n^{-1}(\boldsymbol{\alpha}) - \mathbf{W}_{0n}^{-1}\right)\overline{\mathbf{g}}_n(\boldsymbol{\alpha})\\
&+ \overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{W}_{0n}^{-1}\left(\overline{\mathbf{g}}_n(\boldsymbol{\alpha}) - \overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha})\right) \qquad (10)\\
=&O_p\left(\sqrt{\frac{H^3 + H^2p_s + q_s}{n}}r_n^2 + \sqrt{H^3 + H^2p_s + q_s r_n^3}\right)\\
=&o_p\left(r_n^2\right),
\end{aligned}$$

using (a), (c), (d), and (e) from Lemma 1 results below.

In Step 2, we prove (9).

$$\begin{aligned}
&Q_{0n}(\boldsymbol{\alpha}) - Q_{0n}\left(\boldsymbol{\alpha}_0\right)\\
=&\left(\overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha}) - \overline{\mathbf{g}}_{0n}\left(\boldsymbol{\alpha}_0\right)\right)^{\mathrm{T}}\mathbf{W}_{0n}\left(\overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha}) - \overline{\mathbf{g}}_{0n}\left(\boldsymbol{\alpha}_0\right)\right) + 2\overline{\mathbf{g}}_{0n}\left(\boldsymbol{\alpha}_0\right)^{\mathrm{T}}\mathbf{W}_{0n}\left(\overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha}) - \overline{\mathbf{g}}_{0n}\left(\boldsymbol{\alpha}_0\right)\right)\\
\geq&CL^2r_n^2,
\end{aligned}$$

using (a), (b), and (e) from Lemma 1 results below.

Lemma 1. *We establish the following properties:*

(a) $\sup_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|\leq Cr_n} \left\|\overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha}) - \overline{\mathbf{g}}_{0n}\left(\boldsymbol{\alpha}_0\right)\right\| = O_p\left(r_n\right).$

(b) $\inf_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|=Lr_n} \left\|\overline{\mathbf{g}}_{0n}(\boldsymbol{\alpha}) - \overline{\mathbf{g}}_{0n}\left(\boldsymbol{\alpha}_0\right)\right\| \geq CLr_n.$

(c) $\sup_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|\leq Cr_n} \left\|\bar{\mathbf{g}}_n(\boldsymbol{\alpha}) - \bar{\mathbf{g}}_{0n}(\boldsymbol{\alpha})\right\| = O_p\left(\sqrt{\frac{H^3+H^2p_s+q_s}{n}}r_n\right)$.

(d) $\sup_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|\leq Cr_n} \left\|\mathbf{W}_n^{-1}(\boldsymbol{\alpha}) - \mathbf{W}_{0n}^{-1}\right\| = O_p\left(\sqrt{H^3 + H^2 p_s + q_s}r_n\right)$.

(e) $\left\|\bar{\mathbf{g}}_{0n}\left(\boldsymbol{\alpha}_0\right)\right\| = O_p\left(r_n\right)$.

Proof of Lemma 1. We first prove (e), where

$$\begin{aligned}&\sum_i \mathbf{K}_{0i}\left(\mathbf{Y}_i - \boldsymbol{\mu}_i\left(\boldsymbol{\alpha}_0\right)\right)\\ =&\sum_i \mathbf{K}_{0i}\boldsymbol{\epsilon}_i + \sum_i \mathbf{K}_{0i}\left(\boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i\left(\boldsymbol{\alpha}_0\right)\right)\\ =&O_p\left(\sqrt{n\left(H+p_s+q_s\right)} + nH^{-d}\right) = O_p\left(nr_n\right).\end{aligned}$$

Here we used that for any unit vector $\mathbf{a}$ of appropriate dimension,

$$\|\mathbf{a}^{\mathrm{T}}\sum_i^{\mathrm{T}} \mathbf{K}_i\left(\boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i\left(\boldsymbol{\alpha}_0\right)\right)\| \leq \left(\sum_i \left|\mathbf{a}^{\mathrm{T}}\mathbf{K}_i\mathbf{K}_i^{\mathrm{T}}\mathbf{a}\right|^2\right)^{1/2}\left(\sum_i \|\boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i\left(\boldsymbol{\alpha}_0\right)\|^2\right)^{1/2} = O_p\left(nH^{-d}\right).$$

For (c), the proof is based on repeated application of Taylor's expansion, but the diverging dimension makes it quite messy and therefore hard to keep track of the higher order terms. Let $\delta_{it} = \mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0\right)\boldsymbol{\theta}_0 - \eta\left(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0\right)$ and $\boldsymbol{\delta}_i = \left(\delta_{i1},\ldots,\delta_{iT}\right)^{\mathrm{T}}$. Then employing Taylor's expansion,

$$\begin{aligned}&h_{it}(\boldsymbol{\alpha}) - h_{0it}\\ =&\mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)\boldsymbol{\theta} - \mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0\right)\boldsymbol{\theta}_0 + \mathbf{Z}_{it}^{\mathrm{T}}\left(\boldsymbol{\gamma} - \gamma_0\right) + \delta_{it}\\ =&\mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right) + \dot{\mathbf{G}}^{\mathrm{T}}\left(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0\right)\boldsymbol{\theta}_0\mathbf{X}_{it}^{\mathrm{T}}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right) + \mathbf{Z}_{it}^{\mathrm{T}}\left(\boldsymbol{\gamma} - \gamma_0\right)\\ &+ \dot{\mathbf{G}}^{\mathrm{T}}\left(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}^*\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right)\mathbf{X}_{it}^{\mathrm{T}}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right) + \ddot{\mathbf{G}}^{\mathrm{T}}\left(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}^*\right)\boldsymbol{\theta}_0\left(\mathbf{X}_{it}^{\mathrm{T}}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)\right)^2 + \delta_{it}\\ =&\left(\mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0\right), \dot{\eta}\left(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0\right)\mathbf{X}_{it}^{\mathrm{T}}, \mathbf{Z}_{it}^{\mathrm{T}}\right)\left(\alpha - \alpha_0\right) + \delta_{it} + O_p\left(\left(\sqrt{H^3p_s} + p_s\right)r_n^2\right),\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ in the following quantities, where superscript $*$ always indicates such values that arise from Taylor's expansion, or

$$\begin{aligned}\mathbf{h}_i(\boldsymbol{\alpha}) - \mathbf{h}_{0i} &= \left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{X}_i, \mathbf{Z}_i\right)\left(\alpha - \alpha_0\right) + \boldsymbol{\delta}_i + O_p\left(\left(\sqrt{H^3p_s} + p_s\right)r_n^2\right)\\ &= O_p\left(\sqrt{H + p_s + q_s}r_n\right).\end{aligned} \tag{11}$$

We also have

$$\sum_i \|\mathbf{h}_i(\boldsymbol{\alpha}) - \mathbf{h}_i\left(\boldsymbol{\alpha}_0\right)\|^2 = O_p\left(nr_n^2 + nH^2\min\left\{H, p_s\right\}r_n^4\right).$$

Similarly, we have by Taylor's expansion

$$\begin{aligned}&\mathbf{A}_i^{1/2}(\boldsymbol{\alpha}) - \mathbf{A}_{0i}^{1/2}\\ =&\frac{1}{2}\mathbf{A}_{0i}^{-1/2}\operatorname{diag}\left(\ddot{\boldsymbol{\mu}}_{0i}\right)\operatorname{diag}\left(\mathbf{h}_i(\boldsymbol{\alpha}) - \mathbf{h}_{0i}\right) + O_p\left(\|\mathbf{h}_i(\boldsymbol{\alpha}) - \mathbf{h}_{0i}\|^2\right)\\ =&\frac{1}{2}\mathbf{A}_{0i}^{-1/2}\operatorname{diag}\left(\ddot{\boldsymbol{\mu}}_{0i}\right)\operatorname{diag}\left\{\left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{X}_i, \mathbf{Z}_i\right)\left(\alpha - \alpha_0\right) + \boldsymbol{\delta}_i\right\}\\ &+ O_p\left(\left(\sqrt{H^3p_s} + p_s\right)r_n^2 + \left(H + p_s + q_s\right)r_n^2\right),\end{aligned}$$

$$
\begin{aligned}
&\sum_i \left\|\mathbf{A}_i^{1/2}(\boldsymbol{\alpha}) - \mathbf{A}_{0i}^{1/2}\right\|^2 \\
=&\sum_i \left\|\frac{1}{2}\mathbf{A}_{0i}^{-1/2}\operatorname{diag}\left(\ddot{\boldsymbol{\mu}}_{0i}\right)\operatorname{diag}\left\{\left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{X}_i, \mathbf{Z}_i\right)\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\right) + \boldsymbol{\delta}_i\right\}\right\|^2 + O_p\left(nH^2\min\left\{H, p_s\right\}r_n^4\right) \\
=&O_p\left(nr_n^2 + nH^2\min\left\{H, p_s\right\}r_n^4\right),
\end{aligned}
$$

$$
\begin{aligned}
&\mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}) - \mathbf{A}_{0i}^{-1/2} \\
=&-\frac{1}{2}\mathbf{A}_{0i}^{-3/2}\operatorname{diag}\left(\ddot{\boldsymbol{\mu}}_{0i}\right)\operatorname{diag}\left(\mathbf{h}_i(\boldsymbol{\alpha}) - \mathbf{h}_{0i}\right) + O_p\left(\|\mathbf{h}_i(\boldsymbol{\alpha}) - \mathbf{h}_{0i}\|^2\right) \\
=&-\frac{1}{2}\mathbf{A}_{0i}^{-3/2}\operatorname{diag}\left(\ddot{\boldsymbol{\mu}}_{0i}\right)\operatorname{diag}\left\{\left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{X}_i, \mathbf{Z}_i\right)\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\right)\right\} \\
&+ O_p\left(\left(\sqrt{H^3p_s} + p_s\right)r_n^2 + \left(H + p_s + q_s\right)r_n^2\right),
\end{aligned}
$$

$$
\begin{aligned}
&\sum_i \left\|\mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}) - \mathbf{A}_{0i}^{-1/2}\right\|^2 \\
=&O_p\left(nr_n^2 + nH^2\min\left\{H, p_s\right\}r_n^4\right) \\
&\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}\right) - \mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right) \\
=&\dot{\mathbf{G}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\odot\left(\mathbf{X}_i\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right) + O_p\left(H^{5/2}p_sr_n^2\right),
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_i \left\|\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}\right) - \mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right\|^2 \\
=&\sum_i \left\|\dot{\mathbf{G}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\odot\left(\mathbf{X}_i\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right)\right\|^2 + O_p\left(nH^5p_sr_n^4\right) \\
=&O_p\left(nH^3r_n^2 + nH^5p_sr_n^4\right).
\end{aligned}
$$

Using the identity

$$
\begin{aligned}
&A_1B_1C_1 - A_0B_0C_0 \\
=&\left(A_1 - A_0\right)B_0C_0 + A_0\left(B_1 - B_0\right)C_0 + A_0B_0\left(C_1 - C_0\right) \\
&+A_0\left(B_1 - B_0\right)\left(C_1 - C_0\right) + \left(A_1 - A_0\right)B_0\left(C_1 - C_0\right) + \left(A_1 - A_0\right)\left(B_1 - B_0\right)C_0 \\
&+\left(A_1 - A_0\right)\left(B_1 - B_0\right)\left(C_1 - C_0\right),
\end{aligned}
$$

we have

$$
\begin{aligned}
&\mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_i\boldsymbol{\beta}\right)\mathbf{A}_i^{1/2}(\boldsymbol{\alpha})\mathbf{M}_\ell\mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}) - \mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2} \\
=&\dot{\mathbf{G}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\odot\left(\mathbf{X}_i\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2} \\
&+\frac{1}{2}\mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{A}_{0i}^{-1/2}\operatorname{diag}\left(\ddot{\boldsymbol{\mu}}_{0i}\right)\operatorname{diag}\left\{\left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{X}_i, \mathbf{Z}_i\right)\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\right) + \boldsymbol{\delta}_i\right\}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2} \\
&-\frac{1}{2}\mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-3/2}\operatorname{diag}\left(\ddot{\boldsymbol{\mu}}_{0i}\right)\operatorname{diag}\left\{\left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{X}_i, \mathbf{Z}_i\right)\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\right) + \boldsymbol{\delta}_i\right\} \\
&+ O_p\left(\sqrt{H^3p_s\left(H + p_s + q_s\right)}r_n^2 + \sqrt{H}\left(H + p_s + q_s\right)r_n^2\right),
\end{aligned}
$$

and

$$
\sum_i \left\|\mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_i\boldsymbol{\beta}\right)\mathbf{A}_i^{1/2}(\boldsymbol{\alpha})\mathbf{M}_\ell\mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}) - \mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\right\|^2 = O_p\left(nH^3r_n^2\right).
$$

Using $\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha}) = \boldsymbol{\epsilon}_i + \left(\boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right), \boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i(\boldsymbol{\alpha}) = O_p\left(\sqrt{H + p_s + q_s}r_n\right)$, and $\sum_i \left\|\boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right\|^2 = O_p\left(nr_n^2\right)$, we obtain

$$\begin{aligned}
&\sum_i \left(\mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_i\boldsymbol{\beta}\right)\mathbf{A}_i^{1/2}(\boldsymbol{\alpha})\mathbf{M}_\ell\mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}) - \mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\right)\left(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right)\\
=&\sum_i \left(\mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_i\boldsymbol{\beta}\right)\mathbf{A}_i^{1/2}(\boldsymbol{\alpha})\mathbf{M}_\ell\mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}) - \mathbf{G}^{\mathrm{T}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\right)\left(\boldsymbol{\epsilon}_i + \boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right)\\
=&O_p\left(\sqrt{nH^3}r_n\right).
\end{aligned}$$

Other components of $\mathbf{g}_i(\boldsymbol{\alpha}) - \mathbf{g}_{0i}(\boldsymbol{\alpha})$ can be similarly dealt with. More specifically, we have

$$\begin{aligned}
&\mathbf{J}^{\mathrm{T}}(\boldsymbol{\beta})\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\mathbf{G}}\left(\mathbf{X}_i\boldsymbol{\beta}\right)\boldsymbol{\theta}\right) - \mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_0\right)\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\\
=&\frac{\partial\mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_0\right)}{\partial\boldsymbol{\beta}}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\\
&+ \mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_0\right)\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\ddot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\odot\left(\mathbf{X}_i\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)\right)\right)\\
&+ \mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_0\right)\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\mathbf{G}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right) + \dot{\boldsymbol{\delta}}_i\right)\\
&+ O_p\left(\sqrt{p_s\left(H^3 + p_s\right)}r_n^2\right)\\
=&O_p\Biggl(\biggl(\sqrt{p_s\left(H^3p_s + p_s\right)}r_n\biggr),
\end{aligned}$$

where $\dot{\boldsymbol{\delta}}_i = \dot{\mathbf{G}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\boldsymbol{\theta}_0 - \dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)$.

$$\begin{aligned}
&\mathbf{J}^{\mathrm{T}}(\boldsymbol{\beta})\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\mathbf{G}}\left(\mathbf{X}_i\boldsymbol{\beta}\right)\boldsymbol{\theta}\right)\mathbf{A}_i^{1/2}(\boldsymbol{\alpha})\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}) - \mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_0\right)\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\\
=&\frac{\partial\mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_0\right)}{\partial\boldsymbol{\beta}}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\\
&+ \mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_0\right)\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\ddot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\odot\left(\mathbf{X}_i\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)\right)\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\\
&+ \mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_0\right)\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\mathbf{G}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right) + \dot{\boldsymbol{\delta}}_i\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\\
&+ \frac{1}{2}\mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_0\right)\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\mathbf{A}_{0i}^{-1/2}\operatorname{diag}\left(\ddot{\boldsymbol{\mu}}_{0i}\right)\\
&\operatorname{diag}\left\{\left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{X}_i, \mathbf{Z}_i\right)\left(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\right) + \boldsymbol{\delta}_i\right\}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\\
&- \frac{1}{2}\mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_0\right)\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-3/2}\operatorname{diag}\left(\ddot{\boldsymbol{\mu}}_{0i}\right)\\
&\operatorname{diag}\left\{\left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{X}_i, \mathbf{Z}_i\right)\left(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\right) + \boldsymbol{\delta}_i\right\} + O_p\biggl(\sqrt{\left(H^3p_s + p_s^2\right)\left(H + p_s + q_s\right)}r_n^2\\
&+ \sqrt{p_s}\left(H + p_s + q_s\right)r_n^2\biggr)\\
=&O_p\left(\sqrt{H^3p_s + p_s^2}r_n + \sqrt{p_s\left(H + p_s + q_s\right)}r_n\right),
\end{aligned}$$

and

$$\begin{aligned}
&\sum_i\left\|\mathbf{J}^{\mathrm{T}}(\boldsymbol{\beta})\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\mathbf{G}}\left(\mathbf{X}_i\boldsymbol{\beta}\right)\boldsymbol{\theta}\right)\mathbf{A}_i^{1/2}(\boldsymbol{\alpha})\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}) - \mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_0\right)\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\right\|^2\\
=&O_p\left(nH^2p_sr_n^2\right),
\end{aligned}$$

which in turn implies, using $\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha}) = \boldsymbol{\epsilon}_i + \left(\boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right)$,

$$\begin{aligned}
&\textstyle\sum_i\left(\mathbf{J}^{\mathrm{T}}(\boldsymbol{\beta})\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\mathbf{G}}\left(\mathbf{X}_i\boldsymbol{\beta}\right)\boldsymbol{\theta}\right)\mathbf{A}_i^{1/2}(\boldsymbol{\alpha})\mathbf{M}_\ell\mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}) - \mathbf{J}^{\mathrm{T}}\left(\boldsymbol{\beta}_0\right)\mathbf{X}_i^{\mathrm{T}}\operatorname{diag}\left(\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\mathbf{A}_{0i}^{1/2}\right.\\
&\left.\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\right)\left(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right) = O_p\left(nH^2p_sr_n^2\right).
\end{aligned}$$

We can similarly show

$$\sum_i\left(\mathbf{Z}_i^{\mathrm{T}}\mathbf{A}_i^{1/2}(\boldsymbol{\alpha})\mathbf{M}_\ell\mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}) - \mathbf{Z}_i^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\right)\left(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right) = O_p\left(nq_sr_n^2\right),$$

since $\sum_i \left\|\mathbf{Z}_i^{\mathrm{T}}\mathbf{A}_i^{1/2}(\boldsymbol{\alpha})\mathbf{M}_\ell\mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}) - \mathbf{Z}_i^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\right\|^2 = O_p\left(nq_s r_n^2\right)$.

Thus, we finally get

$$\sup_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|\le Cr_n}\left\|\sum_i \mathbf{g}_i(\boldsymbol{\alpha}) - \sum_i \mathbf{g}_{0i}(\boldsymbol{\alpha})\right\| = O_p\left(\sqrt{n\left(H^3 + H^2 p_s + q_s\right)}r_n\right).$$

To prove (a) and (b), for any unit vector $\mathbf{a}$,

$$\begin{aligned}&\mathbf{a}^{\mathrm{T}}\sum_i \left(\mathbf{g}_{0i}(\boldsymbol{\alpha}) - \mathbf{g}_{0i}\left(\boldsymbol{\alpha}_0\right)\right)\\ =&\mathbf{a}^{\mathrm{T}}\sum_i \mathbf{K}_{0i}\left(\boldsymbol{\mu}_i\left(\boldsymbol{\alpha}_0\right) - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right)\\ \le&\left(\sum_i \left|\mathbf{a}^{\mathrm{T}}\mathbf{K}_{0i}\mathbf{K}_{0i}\mathbf{a}\right|\right)^{1/2}\left(\sum_i \left\|\boldsymbol{\mu}_i\left(\boldsymbol{\alpha}_0\right) - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right\|^2\right)^{1/2} = O_p\left(nr_n\right).\end{aligned}$$

And for the lower bound, we similarly have for $\mathbf{a} = \boldsymbol{\alpha}_0 - \boldsymbol{\alpha}$, and $\ell \in \{1,\ldots,m\}$,

$$\begin{aligned}&\mathbf{a}^{\mathrm{T}}\sum_i \left(\mathbf{g}_{0i\ell}(\boldsymbol{\alpha}) - \mathbf{g}_{0i\ell}\left(\boldsymbol{\alpha}_0\right)\right)\\ =&\mathbf{a}^{\mathrm{T}}\sum_i \mathbf{K}_{0i\ell}\left(\boldsymbol{\mu}_i\left(\boldsymbol{\alpha}_0\right) - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right)\\ =&\sum_i \mathbf{a}^{\mathrm{T}}\mathbf{K}_{0i\ell}\operatorname{diag}\left(\dot{\boldsymbol{\mu}}_{0i}\right)\mathbf{V}_{0i}\left(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}\right)\\ &+\sum_i \mathbf{a}^{\mathrm{T}}\mathbf{K}_{0i\ell}\left(\boldsymbol{\mu}_i\left(\boldsymbol{\alpha}_0\right) - \boldsymbol{\mu}_i(\boldsymbol{\alpha}) - \operatorname{diag}\left(\dot{\boldsymbol{\mu}}_{0i}\right)\mathbf{V}_{0i}\left(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}\right)\right)\\ =&\sum_i \left(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}\right)^{\mathrm{T}}\mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{1/2}\mathbf{V}_{0i}\left(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}\right) + O_p\left(nH^2\left(H + p_s\right)r_n^4\right).\end{aligned}$$

Thus,

$$\begin{aligned}&\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|\left\|\sum_i \left(\mathbf{g}_{0i\ell}(\boldsymbol{\alpha}) - \mathbf{g}_{0i\ell}\left(\boldsymbol{\alpha}_0\right)\right)\right\|\\ \ge&\sum_i \left(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}\right)^{\mathrm{T}}\mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{1/2}\mathbf{V}_{0i}\left(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}\right) - O_p\left(nH^2\left(H + p_s\right)r_n^4\right)\\ =&Cn\left\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\right\|^2,\end{aligned}$$

which implies (b).

To prove (d), start with

$$\begin{aligned}
&\mathbf{W}_n(\boldsymbol{\alpha}) - \mathbf{W}_{0n}\\
=&\frac{1}{n}\sum_i \mathbf{g}_i(\boldsymbol{\alpha})\mathbf{g}_i^{\mathrm{T}}(\boldsymbol{\alpha}) - \frac{1}{n}\sum_i \mathbf{g}_{0i}\mathbf{g}_{0i}^{\mathrm{T}}\\
=&\frac{1}{n}\sum_i \left(\mathbf{g}_i(\boldsymbol{\alpha}) - \mathbf{g}_{0i}\right)\mathbf{g}_{0i}^{\mathrm{T}} + \frac{1}{n}\sum_i \mathbf{g}_{0i}\left(\mathbf{g}_i(\boldsymbol{\alpha}) - \mathbf{g}_{0i}\right)^{\mathrm{T}} + \frac{1}{n}\sum_i \left(\mathbf{g}_i(\boldsymbol{\alpha}) - \mathbf{g}_{0i}\right)\left(\mathbf{g}_i(\boldsymbol{\alpha}) - \mathbf{g}_{0i}\right)^{\mathrm{T}}
\end{aligned}$$

$$\begin{aligned}
&\left\|\sum_i \left(\mathbf{g}_{0i}(\boldsymbol{\alpha}) - \mathbf{g}_{0i}\right)\mathbf{g}_{0i}^{\mathrm{T}}\right\|^2\\
=&\left\|\sum_i \mathbf{K}_{0i}\left(\boldsymbol{\mu}_i(\boldsymbol{\alpha}) - \boldsymbol{\mu}_{0i}\right)\boldsymbol{\epsilon}_i^{\mathrm{T}}\mathbf{K}_{0i}^{\mathrm{T}}\right\|^2\\
=&O_p\left(\sum_i \left\|\mathbf{K}_{0i}^{\mathrm{T}}\mathbf{K}_{0i}\left(\boldsymbol{\mu}_i(\boldsymbol{\alpha}) - \boldsymbol{\mu}_{0i}\right)\right\|^2\right)\\
=&O_p\left(n\left(H + p_s + q_s\right)^2 r_n^2\right).
\end{aligned}$$

$$\begin{aligned}
&\left\|\sum_i \left(\mathbf{g}_i(\boldsymbol{\alpha}) - \mathbf{g}_{0i}(\boldsymbol{\alpha})\right)\mathbf{g}_{0i}^{\mathrm{T}}\right\|^2\\
\le&\left\|\sum_i \left(\mathbf{K}_i(\boldsymbol{\alpha}) - \mathbf{K}_{0i}\right)\left(\boldsymbol{\epsilon}_i + \boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right)\boldsymbol{\epsilon}_i^{\mathrm{T}}\mathbf{K}_{0i}^{\mathrm{T}}\right\|^2\\
=&O_p\left(\left\|\sum_i \left(\mathbf{K}_i(\boldsymbol{\alpha}) - \mathbf{K}_{0i}\right)\mathbf{K}_{0i}^{\mathrm{T}}\right\|^2 + \sum_i \left\|\mathbf{K}_{0i}^{\mathrm{T}}\left(\mathbf{K}_i(\boldsymbol{\alpha}) - \mathbf{K}_{0i}\right)\left(\boldsymbol{\mu}_i(\boldsymbol{\alpha}) - \boldsymbol{\mu}_{0i}\right)\right\|^2\right)\\
=&O_p\left(n^2\left(H^3 + H^2 p_s + q_s\right) r_n^2\right),
\end{aligned}$$

using that $\sum_i \|\mathbf{K}_i(\boldsymbol{\alpha}) - \mathbf{K}_{0i}\|^2 = O_p\left(n\left(H^3 + H^2 p_s + q_s\right) r_n^2\right)$. Then

$$\begin{aligned}
&\sum_i \left(\mathbf{g}_i(\boldsymbol{\alpha}) - \mathbf{g}_{0i}\right)\left(\mathbf{g}_i(\boldsymbol{\alpha}) - \mathbf{g}_{0i}\right)^{\mathrm{T}}\\
=&\sum_i \left(\left(\mathbf{K}_i(\boldsymbol{\alpha}) - \mathbf{K}_{0i}\right)\left(\boldsymbol{\epsilon}_i + \boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right) + \mathbf{K}_{0i}\left(\boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i(\boldsymbol{\alpha})\right)\right)^{\otimes 2}\\
=&O_p\left(\sum_i \|\mathbf{K}_i(\boldsymbol{\alpha}) - \mathbf{K}_{0i}\|^2 + \sum_i \|\mathbf{K}_i(\boldsymbol{\alpha}) - \mathbf{K}_{0i}\|^2\left\|\boldsymbol{\mu}_i(\boldsymbol{\alpha}) - \boldsymbol{\mu}_{0i}\right\|^2 + \sum_i \left\|\boldsymbol{\mu}_i(\boldsymbol{\alpha}) - \boldsymbol{\mu}_{0i}\right\|^2\right)\\
=&O_p\left(n\left(H^3 + H^2 p_s + q_s\right) r_n^2\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\mathbf{W}_n(\boldsymbol{\alpha}) - \mathbf{W}_{0n}\| &= O_p\left(\frac{H + p_s + q_s}{\sqrt{n}} r_n + \sqrt{H^3 + H^2 p_s + q_s} r_n + \left(H^3 + H^2 p_s + q_s\right) r_n^2\right)\\
&= O_p\left(\sqrt{H^3 + H^2 p_s + q_s} r_n\right).
\end{aligned}$$

□

9.2. Proof of Asymptotic Normality for Oracle Estimators

Let $\mathbf{N}_{0i} = \left(\operatorname{diag}\left(\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\mathbf{X}_i\mathbf{J}\left(\boldsymbol{\beta}_0\right), \mathbf{Z}_i\right)$ and define

$$\begin{aligned}\mathbf{P} = \arg\min_{\mathbf{Q}}(\mathbf{N}-\mathbf{GQ})^{\mathrm{T}} &\begin{pmatrix} \mathbf{A}_{01}^{1/2}\mathbf{M}_1\mathbf{A}_{01}^{1/2}\mathbf{V}_{01} & \cdots & \mathbf{A}_{01}^{1/2}\mathbf{M}_m\mathbf{A}_{01}^{1/2}\mathbf{V}_{01} \\ \vdots & \vdots & \vdots \\ \mathbf{A}_{0n}^{1/2}\mathbf{M}_1\mathbf{A}_{0n}^{1/2}\mathbf{V}_{0n} & \cdots & \mathbf{A}_{0n}^{1/2}\mathbf{M}_m\mathbf{A}_{0n}^{1/2}\mathbf{V}_{0n} \end{pmatrix}\mathbf{W}_{0n}^{-1} \\ &\begin{pmatrix} \mathbf{V}_{01}^{\mathrm{T}}\mathbf{A}_{01}^{1/2}\mathbf{M}_1\mathbf{A}_{01}^{1/2} & \cdots & \mathbf{V}_{0n}^{\mathrm{T}}\mathbf{A}_{0n}^{1/2}\mathbf{M}_1\mathbf{A}_{0n}^{1/2} \\ \vdots & \vdots & \vdots \\ \mathbf{V}_{01}^{\mathrm{T}}\mathbf{A}_{01}^{1/2}\mathbf{M}_m\mathbf{A}_{01}^{1/2} & \cdots & \mathbf{V}_{0n}^{\mathrm{T}}\mathbf{A}_{0n}^{1/2}\mathbf{M}_m\mathbf{A}_{0n}^{1/2} \end{pmatrix}(\mathbf{N}-\mathbf{GQ}).\end{aligned} \tag{12}$$

Note (12) is the empirical version of (4). We write $\boldsymbol{\zeta} = \left(\boldsymbol{\beta}^{(-1)\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}\right)^{\mathrm{T}}$ for the parameters of the parametric portion. Using the reparametrization $\boldsymbol{\theta}^* = \boldsymbol{\theta} + \mathbf{P}\boldsymbol{\zeta}$, there is a 1-1 mapping between $(\boldsymbol{\theta}^*, \boldsymbol{\zeta})$ and $(\boldsymbol{\theta}, \boldsymbol{\zeta})$. Thus the problem of minimizing $Q_n(\boldsymbol{\theta}, \boldsymbol{\zeta})$ over $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ is equivalent to minimizing over $(\boldsymbol{\theta}^*, \boldsymbol{\zeta})$. We will show in Lemma 2 that $\|\mathbf{P}\|_{op}$ is bounded despite its diverging dimension. This means that a r_n -consistent estimator $(\widehat{\boldsymbol{\theta}, \boldsymbol{\zeta}})$ is equivalent to a r_n-consistent estimator $\left(\widehat{\boldsymbol{\theta}}^*, \widehat{\boldsymbol{\zeta}}\right)$. In the following, we always regard the parameters as $(\boldsymbol{\theta}^*, \boldsymbol{\zeta})$, and we simply write $Q_n\left(\boldsymbol{\theta}^*, \boldsymbol{\zeta}\right)$ for the QIF objective we are minimizing when using such a reparametrization and do the same for other quantities that depend on the parameters $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{\zeta})$. Fixing $\boldsymbol{\theta}^*$ at $\widehat{\boldsymbol{\theta}}^* = \widehat{\boldsymbol{\theta}} + \mathbf{P}\widehat{\boldsymbol{\zeta}}$, then obviously $\widehat{\boldsymbol{\zeta}}$ minimizes $Q_n\left(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\zeta}\right)$.

Let $\mathbf{U}_i = \mathbf{N}_{0i} - \mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{P}$, where this can be interpreted as orthogonalized predictors for the parametric part. Define $Q_{0n}(\boldsymbol{\zeta}) = \overline{\mathbf{g}}_{0n}(\boldsymbol{\zeta})^{\mathrm{T}}\mathbf{W}_{0n}\overline{\mathbf{g}}_{0n}(\boldsymbol{\zeta})$, with $\overline{\mathbf{g}}_{0n}(\boldsymbol{\zeta}) = \frac{1}{n}\sum_i \mathbf{g}_{0i}(\boldsymbol{\zeta}), \mathbf{g}_{0i}(\boldsymbol{\zeta}) = \left(\mathbf{g}_{0i1}^{\mathrm{T}}(\boldsymbol{\zeta}), \ldots, \mathbf{g}_{0im}^{\mathrm{T}}(\boldsymbol{\zeta})\right)^{\mathrm{T}}$, $\mathbf{g}_{0i\ell}(\boldsymbol{\zeta}) = \mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}\left(\boldsymbol{\epsilon}_i - \mathbf{A}_{0i}\mathbf{G}\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0^*\right) - \mathbf{A}_{0i}\mathbf{U}_i\left(\zeta - \zeta_0\right)\right)$.

Let $\widetilde{\zeta}$ be the minimizer of $Q_{0n}(\zeta)$. We first establish the asymptotic normality of $\widetilde{\zeta}$. Obviously, $Q_{0n}(\zeta)$ is a quadratic function of $\boldsymbol{\zeta}$ with a close-form minimizer

$$\widetilde{\boldsymbol{\zeta}} = \boldsymbol{\zeta}_0 + \left(\mathbf{S}_{0n}^{\mathrm{T}}\mathbf{W}_{0n}^{-1}\mathbf{S}_{0n}\right)^{-1}\mathbf{S}_{0n}^{\mathrm{T}}\mathbf{W}_{0n}^{-1}\frac{1}{n}\sum_i \mathbf{K}_{0i}\left(\boldsymbol{\epsilon}_i - \mathbf{A}_{0i}\mathbf{G}\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)\left(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0^*\right)\right),$$

where $\mathbf{S}_{0n} = \frac{1}{n}\sum_i \mathbf{K}_{0i}\mathbf{A}_{0i}\mathbf{U}_i$. To establish the asymptotic normality of $\widetilde{\boldsymbol{\zeta}}$, we need to show

$$\mathbf{S}_{0n}^{\mathrm{T}}\mathbf{W}_{0n}^{-1}\frac{1}{n}\sum_i \mathbf{K}_{0i}\mathbf{A}_{0i}\mathbf{G}\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right)\left(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0^*\right) = o_p\left(n^{-1/2}\right).$$

We note that $\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0^*$ has a nonparametric rate and a naive bound would fail to show the $o_p\left(n^{-1/2}\right)$ rate above. However, it turns out the above is exactly zero due to the definition of $\mathbf{P}$. In fact, the first order condition of the optimization problem (12) is just

$$\mathbf{S}_{0n}^{\mathrm{T}}\mathbf{W}_{0n}^{-1}\frac{1}{n}\sum_i \mathbf{K}_{0i}\mathbf{A}_{0i}\mathbf{G}\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0\right) = \mathbf{0}.$$

To show $\widehat{\boldsymbol{\zeta}}$ has the same asymptotic distribution as $\widetilde{\boldsymbol{\zeta}}$, we need to establish

$$\sup_{\|\boldsymbol{\zeta}-\boldsymbol{\zeta}_0\|\leq Cr_n}\left|Q_n\left(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\zeta}\right) - Q_{0n}(\boldsymbol{\zeta})\right| = o_p(1/n), \tag{13}$$

and

$$Q_{0n}(\boldsymbol{\zeta}) - Q_{0n}(\widetilde{\boldsymbol{\zeta}}) \geq C\|\boldsymbol{\zeta} - \widetilde{\boldsymbol{\zeta}}\|^2. \tag{14}$$

Indeed, if (13) and (14) hold, we will have for any $\epsilon > 0$,

$$\inf_{\|\boldsymbol{\zeta}-\widetilde{\boldsymbol{\zeta}}\|=\epsilon/\sqrt{n}} Q_n\left(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\zeta}\right) - Q_n\left(\widehat{\boldsymbol{\theta}}^*, \widetilde{\boldsymbol{\zeta}}\right) \geq C\|\boldsymbol{\zeta} - \widetilde{\boldsymbol{\zeta}}\|^2 - o_p(1/n) > 0.$$

Since $\widehat{\boldsymbol{\zeta}}$ minimizes $Q_n\left(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\zeta}\right)$, the above implies $\|\widehat{\boldsymbol{\zeta}} - \widetilde{\boldsymbol{\zeta}}\|_\infty \leq \|\widehat{\boldsymbol{\zeta}} - \widetilde{\boldsymbol{\zeta}}\| = o_p\left(n^{-1/2}\right)$, and thus $\widehat{\boldsymbol{\zeta}}$ has the same asymptotic distribution as $\widetilde{\boldsymbol{\zeta}}$, which finishes the proof.

Note that (13) is already shown in (10), where the more stringent assumption for Theorem 2 makes the rate

$o_p(1/n)$ instead of $o_p\left(r_n^2\right)$, and (14) is shown in Lemma 3.

Lemma 2. *The operator norm (largest singular value) of* $\mathbf{P}$ *defined in (12) is bounded.*

Proof of Lemma 2. We have the closed-form

$$\mathbf{P} = \left\{\left(\frac{1}{n}\sum_i \mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)^{\mathrm{T}} \mathbf{A}_{0i}\mathbf{K}_{0i}^{\mathrm{T}}\right)\mathbf{W}_{0n}^{-1}\left(\frac{1}{n}\sum_i \mathbf{K}_{0i}\mathbf{A}_{0i}\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\right\}^{-1}$$
$$\left(\frac{1}{n}\sum_i \mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)^{\mathrm{T}} \mathbf{A}_{0i}\mathbf{K}_{0i}^{\mathrm{T}}\right)\mathbf{W}_{0n}^{-1}\left(\frac{1}{n}\sum_i \mathbf{K}_{0i}\mathbf{A}_{0i}\mathbf{N}_{0i}\right).$$

The sample averages that appear above (note $\mathbf{W}_{0n}$ is also a sample average) are obviously converging to their population counterparts, and we thus only consider the population quantities.

First, $E\left[\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)^{\mathrm{T}}\mathbf{A}_{0i}\mathbf{K}_{0i\ell}\right]$ and $E\left[\mathbf{K}_{0i\ell}\mathbf{A}_{0i}\mathbf{N}_{0i}\right]$ are both submatrices of $E\left[\mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{1/2}\mathbf{V}_{0i}\right]$, and thus their operator norms are bounded. Second, the quantity that we take the inverse of is a principal submatrix

$$\sum_\ell E\left[\mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{1/2}\mathbf{V}_{0i}\right]W_{0n}^{-1}E\left[\mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{1/2}\mathbf{V}_{0i}\right],$$

whose eigenvalues are bounded away from zero and infinity, and thus

$$\left\{\left(\frac{1}{n}\sum_i \mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)^{\mathrm{T}} \mathbf{A}_{0i}\mathbf{K}_{0i}^{\mathrm{T}}\right)\mathbf{W}_{0n}^{-1}\left(\frac{1}{n}\sum_i \mathbf{K}_{0i}\mathbf{A}_{0i}\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\right\}^{-1}$$

also has bounded eigenvalues. □

The next lemma proves (14).

Lemma 3. $Q_{0n}(\boldsymbol{\zeta}) - Q_{0n}(\widetilde{\boldsymbol{\zeta}}) \geq C\|\boldsymbol{\zeta} - \widetilde{\boldsymbol{\zeta}}\|^2$.

Proof of Lemma 3. Using that $Q_{0n}(\boldsymbol{\zeta})$ is a quadratic form with minimizer $\widetilde{\boldsymbol{\zeta}}$, we have

$$Q_{0n}(\boldsymbol{\zeta}) - Q_{0n}(\widetilde{\boldsymbol{\zeta}}) \geq \lambda_{\min}(\mathbf{D})\left\|\boldsymbol{\zeta} - \widetilde{\boldsymbol{\zeta}}\right\|^2,$$

where $\lambda_{\min}(\mathbf{D})$ denotes the minimum eigenvalue of the matrix

$$\mathbf{D} = \left(\frac{1}{n}\sum_i \begin{pmatrix} \mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_1\mathbf{A}_{0i}^{1/2}\mathbf{U}_i \\ \vdots \\ \mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{1/2}\mathbf{U}_i \end{pmatrix}\right)^T \mathbf{W}_{0n}^{-1}\left(\frac{1}{n}\sum_i \begin{pmatrix} \mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_1\mathbf{A}_{0i}^{1/2}\mathbf{U}_i \\ \vdots \\ \mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{1/2}\mathbf{U}_i \end{pmatrix}\right),$$

which is a principal submatrix of

$$\mathbf{D}' := \left(\tfrac{1}{n}\textstyle\sum_i \begin{pmatrix} \mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_1\mathbf{A}_{0i}^{1/2}\left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \mathbf{U}_i\right) \\ \vdots \\ \mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{1/2}\left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \mathbf{U}_i\right) \end{pmatrix}\right)^{\mathrm{T}} \mathbf{W}_{0n}^{-1}$$
$$\cdot\left(\tfrac{1}{n}\textstyle\sum_i \begin{pmatrix} \mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_1\mathbf{A}_{0i}^{1/2}\left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \mathbf{U}_i\right) \\ \vdots \\ \mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{1/2}\left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \mathbf{U}_i\right) \end{pmatrix}\right).$$

Noting $\mathbf{U}_i = \mathbf{N}_{0i} - \mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{P}$, we have

$$\left(\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right), \mathbf{U}_i\right) = \mathbf{V}_{0i}\begin{pmatrix} \mathbf{I} & -\mathbf{P} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}.$$

Since $\|\mathbf{P}\|_{op}$ is bounded, both

$$\begin{pmatrix} \mathbf{I} & -\mathbf{P} \\ 0 & \mathbf{I} \end{pmatrix}$$

and its inverse

$$\begin{pmatrix} \mathbf{I} & \mathbf{P} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$$

have bounded eigenvalues. Furthermore, since $\mathbf{W}_{0n}^{-1}$ has eigenvalues bounded away from zero, and the sample average in the definition of $\mathbf{D}'$ can be approximated by its population counterpart, we only need to show that

$$\sum_{\ell} E\left[\mathbf{V}_{0i}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\mathbf{M}_{\ell}\mathbf{A}_{0i}^{1/2}\mathbf{V}_{0i}\right]^{\otimes 2}$$

has eigenvalues bounded away from zero, which is true by assumption. □

9.3. Proof of Asymptotic Properties of PQIF Estimators in Ultra-High Dimension

For convergence rate, we only need to show

$$\inf_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|=Lr_n} Q_n(\boldsymbol{\alpha}) + \sum_{j=1}^{p} q_{\lambda_p}\left(\left|\beta_j\right|\right) + \sum_{k=1}^{q} q_{\lambda_q}\left(|\gamma_k|\right) > Q_n\left(\boldsymbol{\alpha}_0\right) + \sum_{j=1}^{p} q_{\lambda_p}\left(\left|\beta_{0j}\right|\right) + \sum_{k=1}^{q} q_{\lambda_q}\left(|\gamma_{0k}|\right). \tag{15}$$

We already see in (7) that $\inf_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|=Lr_n} Q_n(\boldsymbol{\alpha}) > Q_n\left(\boldsymbol{\alpha}_0\right)$. We will show when $\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\| \le Lr_n, q_{\lambda_p}\left(\left|\beta_j\right|\right) \ge q_{\lambda_p}\left(\left|\beta_{0j}\right|\right), j = 1, \dots, p$ (similarly we can show $q_{\lambda_q}\left(|\gamma_k|\right) \ge q_{\lambda_q}\left(|\gamma_{0k}|\right), k = 1, \dots, q\big)$, which immediately implies (15). Indeed, when $j > p_s, q_{\lambda_p}\left(\left|\beta_j\right|\right) \ge 0 = q_{\lambda_p}\left(\left|\beta_{0j}\right|\right)$. On the other hand, when $j \le p_s$, since $\left|\beta_{0j}\right| \ge C\lambda_p$ and $\left|\widehat{\beta}_j - \beta_{0j}\right| \le \|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\| = o\left(\lambda_p\right)$, both $\left|\beta_{0j}\right|$ and $\left|\widehat{\beta}_{0j}\right|$ are large enough to be in the region of the domain of q_{λ_p} that is nonzero by the specific expression of the SCAD penalty, and thus $q_{\lambda_p}\left(\left|\beta_j\right|\right) = q_{\lambda_p}\left(\left|\beta_{0j}\right|\right)$.

Next we consider variable selection consistency. Suppose, by way of contradiction, that $\widehat{\beta}_{j^*} \ne 0$ for some $j^* \in \{p_s + 1, \dots, p\}$, and components of $\widehat{\gamma}$ can be similarly dealt with. Define $\check{\boldsymbol{\beta}}$ such that its j^*-component is zero while other components are equal to those of $\widehat{\boldsymbol{\beta}}$. We will show that

$$Q_n(\check{\boldsymbol{\alpha}}) + \sum_{j=1}^{p} q_{\lambda_p}\left(\left|\check{\beta}_j\right|\right) + \sum_{k=1}^{q} q_{\lambda_q}\left(\left|\check{\gamma}_j\right|\right) < Q_n(\widehat{\boldsymbol{\alpha}}) + \sum_{j=1}^{p} q_{\lambda_p}\left(\left|\widehat{\beta}_j\right|\right) + \sum_{k=1}^{q} q_{\lambda_q}\left(\left|\widehat{\gamma}_k\right|\right), \tag{16}$$

which leads to a contradiction. In fact, in Lemma 4, we show $Q_n(\check{\boldsymbol{\alpha}}) - Q_n\left(\boldsymbol{\alpha}_0\right) = O_p\left(\lambda_p\right)\|\check{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}\|$. Furthermore, by the definition of $\check{\boldsymbol{\beta}}$ which only differs from $\widehat{\boldsymbol{\beta}}$ in the j^*-th component, we have

$$\sum_{j=1}^{p} q_{\lambda_p}\left(\left|\check{\beta}_{0j}\right|\right) - \sum_{j=1}^{p} q_{\lambda_p}\left(\left|\widehat{\beta}_j\right|\right) = -q_{\lambda_p}\left(\left|\widehat{\beta}_{j^*}\right|\right) = -\lambda_p\left|\widehat{\beta}_{j^*}\right|,$$

where the last step is due to $\left|\widehat{\beta}_{j^*}\right| \le \left\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right\| = o_p\left(\lambda_p\right)$ implying $\left|\widehat{\beta}_{j^*}\right|$ is in the region of the domain of $q_{\lambda_p}(.)$ that is a linear function by the specific expression of the SCAD penalty. This finishes the proof of (16).

Lemma 4. *Uniformly for $j^* = p_s + 1, \dots, p$, where $\check{\boldsymbol{\alpha}}$ below implicitly depends on j^*,*

$$Q_n(\widehat{\boldsymbol{\alpha}}) - Q_n(\check{\boldsymbol{\alpha}}) = O_p\left(\sqrt{\frac{(H^3 + H^2 p_s + q_s)\, logp}{n}} + \sqrt{H + p_s + q_s}\right) r_n \left|\widehat{\beta}_{j^*}\right|.$$

Proof of Lemma 4. The proof is based on Taylor's expansion largely the same as in Lemma 1. We only briefly present some of the calculations for illustration. We decompose

$$\begin{aligned} & Q_n(\widehat{\boldsymbol{\alpha}}) - Q_n(\check{\boldsymbol{\alpha}}) \\ &= \left(\overline{\mathbf{g}}_n(\widehat{\boldsymbol{\alpha}}) - \overline{\mathbf{g}}_n(\check{\boldsymbol{\alpha}})\right)^{\mathrm{T}} \mathbf{W}_{0n}^{-1}\overline{\mathbf{g}}_{0n}\left(\boldsymbol{\alpha}_0\right) + \overline{\mathbf{g}}_{0n}\left(\boldsymbol{\alpha}_0\right)^{\mathrm{T}}\left(\mathbf{W}_n^{-1}(\boldsymbol{\alpha}) - \mathbf{W}_n^{-1}(\check{\boldsymbol{\alpha}})\right)\overline{\mathbf{g}}_{0n}\left(\boldsymbol{\alpha}_0\right) \\ &\quad + \overline{\mathbf{g}}_{0n}\left(\boldsymbol{\alpha}_0\right)^{\mathrm{T}}\mathbf{W}_{0n}^{-1}\left(\overline{\mathbf{g}}_n(\widehat{\boldsymbol{\alpha}}) - \overline{\mathbf{g}}_n(\check{\boldsymbol{\alpha}})\right) + \cdots, \end{aligned} \tag{17}$$

where we omitted the higher order terms. Consider the first term as an example. $\left\|\overline{\mathbf{g}}_{0n}\left(\boldsymbol{\alpha}_0\right)\right\| = O_p\left(r_n\right)$ using (e) of Lemma 1. For $\overline{\mathbf{g}}_n(\widehat{\boldsymbol{\alpha}}) - \overline{\mathbf{g}}_n(\check{\boldsymbol{\alpha}})$, similar to the calculations in Lemma 1 we can get that, for example, the main terms of

$$\begin{aligned}&\textstyle\sum_i \mathbf{G}\left(\mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\alpha}})\mathbf{M}_\ell\mathbf{A}_i^{-1/2}(\widehat{\boldsymbol{\alpha}})\left(\mathbf{Y}_i-\boldsymbol{\mu}_i(\widehat{\boldsymbol{\alpha}})\right)\\ &-\textstyle\sum_i \mathbf{G}\left(\mathbf{X}_i\check{\boldsymbol{\beta}}\right)\mathbf{A}_i^{1/2}(\check{\boldsymbol{\alpha}})\mathbf{M}_\ell\mathbf{A}_i^{-1/2}(\check{\boldsymbol{\alpha}})\left(\mathbf{Y}_i-\boldsymbol{\mu}_i(\check{\boldsymbol{\alpha}})\right)\end{aligned}$$

are

$$\begin{aligned}&\sum_i\left\{\dot{\mathbf{G}}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{-1/2}+\frac{1}{2}\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{A}_{0i}^{1/2}\operatorname{diag}\left(\ddot{\boldsymbol{\mu}}_{0i}\odot\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\right.\\ &\left.-\frac{1}{2}\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{A}_{0i}^{-3/2}\operatorname{diag}\left(\ddot{\boldsymbol{\mu}}_{0i}\odot\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\right\}\left(\mathbf{X}_{i(j^*)}\odot\boldsymbol{\epsilon}_i\right)\widehat{\beta}_{j^*}\\ &+\sum_i\mathbf{G}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\mathbf{A}_{0i}^{1/2}\mathbf{M}_\ell\mathbf{A}_{0i}^{1/2}\operatorname{diag}\left(\dot{\eta}\left(\mathbf{X}_i\boldsymbol{\beta}_0\right)\right)\mathbf{X}_{i(j^*)}\widehat{\beta}_{j^*},\end{aligned}\tag{18}$$

where the T-dimensional vector $\mathbf{X}_{i(j^*)}$ is the j^* -th column of $\mathbf{X}_i$. The first term of (18) has mean zero, and is of order $O_p\left(\sqrt{nH^3\log p}\left|\hat{\beta}_{j^*}\right|\right)$, where the logarithmic term comes from applying Bernstein's inequality to get uniform bound over j^*. The second term in (18) is more easily derived to be of order $n\sqrt{H}\left|\widehat{\beta}_{j^*}\right|$. This and similar bounds would give

$$\left\|\bar{\mathbf{g}}_n(\widehat{\boldsymbol{\alpha}})-\bar{\mathbf{g}}_n(\tilde{\boldsymbol{\alpha}})\right\|=O_p\left(\sqrt{\frac{(H^3+H^2p_s+q_s)\log p}{n}}+\sqrt{H+p_s+q_s}\right)\left|\widehat{\beta}_{j^*}\right|.$$

Then the first term in (17) would be of order $O_p\left(\sqrt{\frac{(H^3+H^2p_s+q_s)\log p}{n}}+\sqrt{H+p_s+q_s}\right)r_n\left|\widehat{\beta}_{j^*}\right|$. The third term in (17) is easily seen to be of the same order, while the second term in (17) can be shown to be of smaller order as is also the case in (10).

References

[1] An, P., M. Feitosa, S. Ketkar, A. Adelman, S. Lin, I. Borecki, and M. Province (2009, December). Epistatic interactions of CDKN2B-TCF7L2 for risk of type 2 diabetes and of CDKN2B-JAZF1 for triglyceride/high-density lipoprotein ratio longitudinal change: evidence from the Framingham Heart Study. *BMC proceedings 3 Suppl 7*, S71.

[2] Bai, Y., W. K. Fung, and Z. Y. Zhu (2009). Penalized quadratic inference functions for single-index models with longitudinal data. *Journal of Multivariate Analysis 100*(1), 152–161.

[3] Cai, L., H. Wu, D. Li, K. Zhou, and F. Zou (2015). Type 2 diabetes biomarkers of human gut microbiota selected via iterative sure independent screening method. *PloS one 10*(10), 1–15.

[4] Carroll, R. J., J. Fan, I. Gijbels, and M. P. Wand (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association 92*(438), 477–489.

[5] Cho, H. and A. Qu (2013). Model selection for correlated data with diverging number of parameters. *Statistica Sinica 23*(2), 901–927.

[6] Dawber, T. R., G. F. Meadors, and F. E. Moore Jr (1951). Epidemiological approaches to heart disease: the Framingham Study. *American Journal of Public Health and the Nations Health 41*(3), 279–286.

[7] Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

[8] Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B 70*(5), 849–911.

[9] Fang, Y., Y. Qin, N. Zhang, J. Wang, H. Wang, and X. Zheng (2015). DISIS: prediction of drug response through an iterative sure independence screening. *PloS one 10*(3), 1–13.

[10] Franks, P. W. (2011). Gene× environment interactions in type 2 diabetes. *Current diabetes reports 11*(6), 552.

[11] Gaulton, K. J., T. Ferreira, Y. Lee, A. Raimondo, R. Mägi, M. E. Reschen, A. Mahajan, et al. (2015, December). Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nature Genetics 47*(12), 1415–1425.

[12] Green, B., H. Lian, Y. Yu, and T. Zu (2020). Ultra high-dimensional semiparametric longitudinal data analysis. *Biometrics*.

[13] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054.

[14] Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the LASSO and generalizations*. CRC press.

[15] He, X., Z.-Y. Zhu, and W.-K. Fung (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika 89*(3), 579–590.

[16] Huang, J., J. L. Horowitz, and F. Wei (2010). Variable selection in nonparametric additive models. *Annals of statistics 38*(4), 2282.

[17] Karlsson Linnér, R., P. Biroli, E. Kong, et al. (2019, February). Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics 51*(2), 245–257. Number: 2 Publisher: Nature Publishing Group.

[18] Lai, P., G. Li, and H. Lian (2013). Quadratic inference functions for partially linear single-index models with longitudinal data. *Journal of Multivariate Analysis 118*, 115–127.

[19] Li, G., P. Lai, and H. Lian (2015). Variable selection and estimation for partially linear single-index models with longitudinal data. *Statistics and Computing 25*(3), 579–593.

[20] Lindström, J. and J. Tuomilehto (2003, March). The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk. *Diabetes Care 26*(3), 725–731. Publisher: American Diabetes Association Section: Epidemiology/Health Services/Psychosocial Research.

[21] Ma, S., H. Liang, and C.-L. Tsai (2014). Partially linear single index models for repeated measurements. *Journal of Multivariate Analysis 130*, 354–375.

[22] Macke, J. H., P. Berens, A. S. Ecker, A. S. Tolias, and M. Bethge (2009). Generating spike trains with specified correlation coefficients. *Neural computation 21*(2), 397–423.

[23] Meigs, J. B., A. K. Manning, C. S. Fox, J. C. Florez, C. Liu, L. A. Cupples, and J. Dupuis (2007). Genome-wide association with diabetes-related traits in the framingham heart study. *BMC medical genetics 8*(1), 1–10.

[24] Perry, J. R. B., M. I. McCarthy, A. T. Hattersley, E. Zeggini, Wellcome Trust Case Control Consortium, M. N. Weedon, and T. M. Frayling (2009, June). Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. *Diabetes 58*(6), 1463–1467.

[25] Prasad, R. B. and L. Groop (2015, March). Genetics of Type 2 Diabetes—Pitfalls and Possibilities. *Genes 6*(1), 87–123.

[26] Qu, A. and R. Li (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics 62*(2), 379–391.

[27] Qu, A., B. G. Lindsay, and B. Li (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika 87*(4), 823–836.

[28] Ruppert, D. and R. J. Carroll (2000). Theory & methods: Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal of Statistics 42*(2), 205–223.

[29] Schumaker, L. (2007). *Spline functions: basic theory*. Cambridge University Press.

[30] Stern, M. P., K. Williams, and S. M. Haffner (2002, April). Identification of persons at high risk for type 2 diabetes mellitus: do we need the oral glucose tolerance test? *Annals of Internal Medicine 136*(8), 575–581.

[31] Taylor, J. Y., Y. V. Sun, S. C. Hunt, and S. L. Kardia (2010). Gene-environment interaction for hypertension among African American women across generations. *Biological Research for Nursing 12*(2), 149–155.

[32] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

[33] Vos, T., C. Allen, M. Arora, R. M. Barber, Z. A. Bhutta, A. Brown, A. Carter, D. C. Casey, F. J. Charlson, A. Z. Chen, et al. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. *The lancet 388*(10053), 1545–1602.

[34] Wang, H., B. Li, and C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B 71*(3), 671–683.

[35] Wang, L., X. Liu, H. Liang, and R. J. Carroll (2011). Estimation and variable selection for generalized additive partial linear models. *Annals of statistics 39*(4), 1827.

[36] Wang, L., L. Xue, A. Qu, H. Liang, et al. (2014). Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *The Annals of Statistics 42*(2), 592–624.

[37] Wang, L., J. Zhou, and A. Qu (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics 68*(2), 353–360.

[38] Wang, P., G.-f. Tsai, and A. Qu (2012). Conditional inference functions for mixed-effects models with unspecified random-effects distribution. *Journal of the American Statistical Association 107*(498), 725–736.

[39] Young, A., I. Ferrier, S. Ball, M. Mohiuddin, C. Todhunter, J. Mansfield, et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*.

[40] Yu, Y. and D. Ruppert (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association 97*(460), 1042–1054.

[41] Yu, Y., C. Wu, and Y. Zhang (2017). Penalised spline estimation for generalised partially linear single-index models. *Statistics and Computing 27*(2), 571–582.

[42] Zeger, S. L. and K.-Y. Liang (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics 42*(1), 121–130.

[43] Zhang, Y., H. Lian, and Y. Yu (2017). Estimation and variable selection for quantile partially linear single-index models. *Journal of Multivariate Analysis 162*, 215–234.

[44] Zhou, J. and A. Qu (2012). Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistical Association 107*(498), 701–710.