University of Cincinnati

Date: 6/4/2020

I, **Brittany Green**, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Business Administration.

It is entitled:

Ultra-high Dimensional Semiparametric Longitudinal Data Analysis

Student's name: **Brittany Green**

This work and its defense approved by:

Committee chair: Peng Wang, Ph.D.

Committee chair: Yan Yu, Ph.D.

Committee member: Lenisa Chang, Ph.D.

36781

Ultra-high Dimensional Semiparametric Longitudinal Data Analysis

A dissertation submitted to the
Graduate School
of the University of Cincinnati
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

in the Department of Operations, Business Analytics, and Information Systems
of the College of Business

by

Brittany Green

M.I.S.E. University of Pittsburgh

August 2020

Committee Chairs: Yan Yu, Ph.D. and Peng Wang, Ph.D.

ABSTRACT

As ultra-high dimensional longitudinal data is becoming ever more apparent in fields such as public health, information systems, and bioinformatics, developing flexible methods with a sparse set of important variables is of high interest. In this setting, the dimension of the covariates can potentially grow exponentially with respect to the number of clusters. This dissertation research considers a flexible semiparametric approach, namely, partially linear single-index models, for ultra-high dimensional longitudinal data. Most importantly, we allow not only the partially linear covariates, but also the single-index covariates within the unknown flexible function estimated nonparametrically to be ultra-high dimensional. Using penalized generalized estimating equations, this approach can capture correlation within subjects, can perform simultaneous variable selection and estimation with a smoothly clipped absolute deviation penalty, and can capture nonlinearity and potentially some interactions among predictors. We establish asymptotic theory for the estimators including the oracle property in ultra-high dimension for both the partially linear and nonparametric components. An efficient algorithm is presented to handle the computational challenges, and we show the effectiveness of our method and algorithm via a simulation study and yeast cell cycle gene expression data. In addition, we develop an alternative solution methodology via the penalized quadratic inference function with partially linear single-index models for ultra-high dimensional longitudinal data. This methodology can improve the estimation efficiency when the working correlation structure is misspecified. Performance is demonstrated via a simulation study and analysis of a genomic dataset.

Copyright © 2020 by Brittany Green

All Rights Reserved

Dedicated to my wife and my parents.

ACKNOWLEDGMENTS

First and foremost, my deepest gratitude goes out to my advisor Professor Yan Yu. Your endless support, patience, and mentorship played an instrumental role in my success, without which this degree would not have been possible. You taught me perseverance, and gifted me excitement for future research. I am deeply honored and fortunate to work with you. I am also extremely grateful for my advisor Professor Peng Wang. I cannot thank you enough for your mentorship and the countless hours spent on research and guidance over the past 5 years. My deepest gratitude goes to Professor Heng Lian from City University of Hong Kong for his guidance. The theory in this work would be not possible without you. Also, I would like to thank my other committee member Lenisa Chang for her support during the dissertation process.

I also want to extend my appreciation to UC faculty, especially professors Craig Froehle, Sherah Daniel, and Suzanne Masterson for their guidance and encouragement throughout my time at UC. I would be remiss not to mention my fellow Ph.D. students in OBAIS, thank you for your friendship over all these years.

To my parents, thank you for being there for me over all these years and for believing in me. You are an integral part of this accomplishment. Finally, I would like to give my special and heartfelt thanks to my wife, Angela. Thank you for your endless encouragement and believing in me throughout this process. You have sacrificed so much for me, and I only hope I can return the love and support for you to achieve your dreams.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGMENTS	v
LIST OF TABLES	viii
LIST OF FIGURES	ix
1 Introduction	1
2 Model	6
2.1 Generalized Partially Linear Single-Index Models for Longitudinal Data	6
2.2 Generalized Estimating Equations and Large Sample Properties for Oracle Estimator	8
3 Penalized Generalized Estimating Equations for Ultra-high Dimensional Semi- parametric Longitudinal Data Analysis	10
4 Algorithm	11
4.1 An Efficient Algorithm	11
4.2 Tuning Parameter Selection	15
5 Numerical Study	17
5.1 Simulation	17
5.2 Additional Simulation Studies	21
5.3 An Application to Yeast Cell Cycle Data	30
6 Discussion	33
7 Penalized Quadratic Inference Function for Semiparametric Longitudinal Data Analysis in Ultra-high Dimensions	35
7.1 Quadratic Inference Function for Semiparametric Longitudinal Data Analysis	37
7.1.1 Quadratic Inference Function	39
7.1.2 Penalized Quadratic Inference Function for Ultra-high Di- mensional Data	41
7.2 Estimation Algorithm	42
7.2.1 Practical Implementation	46
7.3 Numerical Illustrations	49
7.4 An Application to an Advanced Renal Cancer Study	51
7.5 Discussion	53
8 Technical Proofs	54
8.1 Definitions	54
8.2 Assumptions	56
8.3 Proof of Theorem 1	58

8.4	Proof of Theorem 2	68
8.5	Proof of Theorem 3	72

LIST OF TABLES

1	Summary of Parameter Estimates for the Partially Linear Single-Index Model in Example 1 with $p_n = 250$ and $q_n = 250$	19
2	Summary of Estimation and Variable Selection Results from Various Methods in Example 1 with $p_n = 250$ and $q_n = 250$	20
3	Summary of Estimation and Variable Selection Results from Example 2 with Binary Response.	21
4	Summary of Parameter Estimates for Example 1 with Scenario 1 Design. . . .	24
5	Summary of Parameter Estimates for Example 1 with Scenario 2 Design. . . .	25
6	Summary of Parameter Estimates for Example 1 with a Complex Correlation Structure.	26
7	Summary of Parameter Estimates for the Partially Linear Single-Index Model with Unequal m_i with $p_n = 250$ and $q_n = 250$	27
8	Summary of Parameter Estimates for the Partially Linear Single-Index Model in Example 1 with Varying Number of Degrees and Knots with $p_n = 250$ and $q_n = 250$	27
9	Summary of Parameter Estimates for the Partially Linear Single-Index Model in Example 1 with $p_n = 500$ and $q_n = 500$	28
10	Summary of Parameter Estimates for the Partially Linear Single-Index Model in Example 1 with $p_n = 2500$ and $q_n = 2500$	28
11	Summary of Estimation and Variable Selection Results from Various Methods for Example 1 with $p_n = 2500$ and $q_n = 2500$	29
12	Summary of Estimation Results from the Yeast Cell Cycle Gene Expression Data.	33
13	Summary of Results for the Single-Index Model in Example 3 with Binary Response.	50
14	Summary of Results for CCI-779 Data.	52

LIST OF FIGURES

1	Flowchart diagram of the data integration steps for the yeast cell cycle gene expression data.	2
2	Fitted plot for the curve of $h(\cdot)$ from Example 1 with $p_n = 250$ and $q_n = 250$. .	19

1 Introduction

With recent advances, ultra-high dimensional longitudinal studies in bioinformatics and large scale health studies are increasingly common. One motivating example of a large scale gene expression study is the longitudinal yeast cell cycle gene expression data (Spellman et al. (1998); Luan and Li (2003)) in combination with transcription factors from ChIP data (Lee et al. (2002); Wang et al. (2007)). Transcription factors regulate the flow of genetic information by acting as binding agents from DNA to mRNA during the cell cycle process (Lee et al. (2002); Wang et al. (2007)). One main research question of interest is to identify important transcription factors from a large set of transcription factors that are associated with gene expression levels under a complex relationship. Within a biological process, investigating this association can give information into active transcriptional subnetworks anchored on the proximal promotor DNA from genome-wide mRNA profiles (Wang et al., 2007). As in Wang et al. (2012) and Inan and Wang (2017), we have $n = 297$ genes as the response $\mathbf{Y}$ and the binding probabilities of 96 transcription factors along with time t as the covariates $\mathbf{X}$. A detailed explanation of the data integration process is shown in the flowchart Figure 1 and is further discussed in Section 5.3.

Due to the complexity of gene transcriptional regulation, the relationship between the transcription factors and genes may often be quite complex in nature, and synergy or interactions among the transcription factors may also be present (Banerjee and Zhang, 2003; Das et al., 2004; Wang et al., 2007; Cheng and Li, 2008). For instance, it is noted that cooperative interaction exists between transcription factors MBP1 and SWI6 that form an important MluI cell cycle box binding factor (MBF) (Tsai et al. (2005); Banerjee and Zhang (2003)). Specifically, at the final portion of the G1 stage, cells can begin DNA replication and other processes, where the complex MBF is known to be a critical component involved in regulating this progression.

discussion of the general model setup in Section 2 and specifics for the yeast cell cycle data in Section 5.3.

Single-index models are known to be advantageous for nonparametric estimation and dimension reduction and have been largely studied for traditional cross-sectional data with fixed covariates (e.g. Carroll et al. (1997); Yu and Ruppert (2002); Yu et al. (2017)). They avoid the so-called “curse of dimensionality” by reducing the p_n -dimensional covariates $\mathbf{X}$ to a univariate “single-index” $\mathbf{X}\boldsymbol{\delta}_0$. The models are flexible by introducing a flexible univariate function through $h(\mathbf{X}\boldsymbol{\delta}_0)$. These models can also take into account some interactions among covariates $\mathbf{X}$. For example, a simple presence of a quadratic term in $h(\cdot)$ can naturally incorporate some two-way interaction into consideration.

Compared with the longitudinal single-index model (1), a longitudinal linear model in high-dimension $\mathbf{Y} = \mathbf{X}\boldsymbol{\delta} + \boldsymbol{\epsilon}$ (Wang et al. (2012)) may be too restrictive. It may mis-specify the relationship between the gene expression $\mathbf{Y}$ and transcription factors along with time t , which may lead to efficiency loss and reduce prediction accuracy. For example, interactions such as the cooperative interaction between MBP1 and SWI6 clearly cannot be captured in a typical linear model. Furthermore, modeling complex relationships can add even more challenges to such a challenging high dimensional problem.

Gene expression studies can even be in ultra-high dimension where the covariate dimension p_n may even increase exponentially with the cluster size n . Hence, identifying important variables under a flexible model is critical in estimation, inference, and prediction accuracy. As pointed out in Wang et al. (2012), the literature on variable selection for longitudinal data is limited. Ultra-high dimensionality and most importantly allowing the number of single-index covariates p_n within the nonparametric component $h(\cdot)$ to be ultra-high-dimensional add great flexibility in modeling as well as significant difficulty in computation and theory.

We adopt penalized generalized estimating equations (PGEE) (Wang et al. (2012) and references therein) along with the smoothly clipped absolute deviation penalty (SCAD) (Fan and Li, 2001) for variable selection in order to perform simultaneous estimation of the longitudinal single-index model (1) and potentially ultra-high dimensional variable selection for longitudinal data. The flexible function $h(\cdot)$ is modeled by B-splines non-parametrically, which is shown to be computationally stable. Along with a flexible semi-parametric model and simultaneous variable selection in ultra-high dimension, ignoring the dependence structure within subjects of longitudinal data analysis can potentially lead to less efficient estimates and potentially incorrect conclusions. We adopt the classical generalized estimating equation approach (GEE) (Zeger and Liang, 1986), which does not require knowing the correlation structure within subject’s observations and can yield a consistent estimator even if the working correlation structure is misspecified.

When we apply our proposed single-index model with PGEE and SCAD variable selection to the yeast cell cycle data at Stage G1, we find important transcription factors such as MBP1 and SWI6 are selected both under longitudinal single-index model (1) and the linear model. However, the important cooperative interaction between MBP1 and SWI6 as noted in the literature (Banerjee and Zhang, 2003; Das et al., 2004; Wang et al., 2007; Cheng and Li, 2008) cannot be captured under the linear model. This is also reflected in the better prediction error under longitudinal single-index model (1) as reported in Table 12 in Section 5.3. Another interesting finding is that the transcription factor YAP5, noted as an important factor in the previous literature (Banerjee and Zhang (2003) and references therein), is selected under single-index model (1) but not under the linear model.

We not only develop methodology and establish important theoretical properties under single-index model (1) for ultra-high dimensional longitudinal data, we further make two important methodological extensions to generalized partially linear single-index models to

allow discrete longitudinal responses; and to allow q_n ultra-high dimensional partially linear covariate terms $\mathbf{Z}$ to enter model (1) as a partially linear term $\mathbf{Z}\beta_0$. Our proposed approach of semiparametric penalized generalized estimating equations along with the SCAD penalty for the longitudinal single-index model (1) with ultra-high dimensionally covariates can also naturally allow discrete responses such as binary response as in the context of generalized modeling, where it is often difficult and intractable in many cases to specify the joint likelihood for correlated data. Particularly when the parameter dimension is diverging, high dimensional integration renders approximating the joint likelihood function computationally burdensome or impossible.

A key contribution of our approach is the ultra-high dimensional covariates $\mathbf{X}$ within the nonparametric portion $h(\mathbf{X}\delta_0)$, or equivalently, $h(\mathbf{X}_1\delta_1 + \dots + \mathbf{X}_{p_n}\delta_{p_n})$, of correlated data, where we need to select a potentially diverging number of important covariates. This diverging dimension of the covariates within the unknown, flexible function of correlated data creates opportunities and challenges due to high nonlinearity and the dimension of the covariates potentially increasing exponentially with the number of clusters n . Indeed, we allow both p_n and q_n to be ultra-high dimensional even at the order of $\exp(n^{1/2})$. We develop efficient algorithms to alleviate these difficulties through a key linear approximation of $h(\mathbf{X}\delta)$ along with penalized GEE for longitudinal data and implementing variable selection techniques. We establish the desirable theoretical properties including the oracle property (Fan and Li, 2001). To the best of our knowledge, this is the first work allowing ultra-high dimensional covariates in a nonparametric term for longitudinal data analysis.

The past literature on single-index models is extensive but mainly focuses on low to moderate fixed dimensions or on non-correlated data. For example, though the finding of nonlinear relationship is very interesting, the analysis in Carroll et al. (1997) is only conducted on exam 3 of Framingham Heart Study with one scalar response and four

fixed covariates, which is not applicable for general longitudinal data settings with high-dimensional covariates. Furthermore, as pointed out by Yu and Ruppert (2002), their computational algorithm via local linear approach may be unstable. For longitudinal data in fixed dimension, Ma et al. (2014), Chen et al. (2015), and Li et al. (2015) show promising results. Recent research in high-dimensional longitudinal data analysis with diverging number of covariates is mostly concerned with a linear model (Wang et al. (2012)). Lian et al. (2014) conduct a nice study by modeling nonlinear relationships in longitudinal data through partially linear additive models. However, only the partially linear portion is allowed to have a diverging number of covariates, while the nonparametric term is fixed as given a priori.

2 Model

2.1 Generalized Partially Linear Single-Index Models for Longitudinal Data

We observe longitudinal data with $i = 1, \dots, n$ subjects and $j = 1, \dots, m_i$ observations per subject. For subject i , we observe $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{im_i})^T$, $\mathbf{X}_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{im_i})^T$, $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \dots, \mathbf{Z}_{im_i})^T$ as the $m_i \times 1$ response vector, $m_i \times p_n$ matrix of single-index covariates, and $m_i \times q_n$ matrix of linear covariates respectively. This research focuses on a semiparametric approach, namely, generalized partially linear single-index models for longitudinal data, allowing the p_n -dimensional covariate vectors $\mathbf{X}_{ij}$ modeled nonparametrically and q_n -dimensional observed covariate vectors $\mathbf{Z}_{ij}$ modeled partially linearly to diverge and even be in ultra-high dimension. The observations are independent across different subjects but can be dependent within the same subject.

Generalized models are popular in modeling the relationship of the response variable from the general exponential family with a set of predictor variables (e.g. Carroll et al. (1997)). We model the conditional mean for longitudinal data semiparametrically by gen-

eralized partially linear single-index models

$$E(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i) := \boldsymbol{\mu}_i(\boldsymbol{\eta}_i) = g^{-1}(h(\mathbf{X}_i\boldsymbol{\delta}_0) + \mathbf{Z}_i\boldsymbol{\beta}_0), \quad i = 1, \dots, n, \quad (2)$$

where $g(\cdot)$ is the natural link function common to the generalized linear model as in Zeger and Liang (1986); $h(\cdot)$ is an unknown univariate function; $\boldsymbol{\delta}_0$ is a p_n -dimensional vector of single-index coefficients; and $\boldsymbol{\beta}_0$ is a q_n -dimensional vector of partially linear coefficients. The systematic component of the conditional mean is $\boldsymbol{\eta}_i = (\eta_{i1}, \dots, \eta_{im_i})^T = h(\mathbf{X}_i\boldsymbol{\delta}_0) + \mathbf{Z}_i\boldsymbol{\beta}_0$. Commonly the covariates are sparse in $\boldsymbol{\delta}_0 = (\delta_{01}, \dots, \delta_{0p_n})$ where many of them are zeros. Thus, there is a subset $\boldsymbol{\delta}_{0(1)} = (\delta_{01}, \dots, \delta_{0p_{1n}})$ such that $j = 1, \dots, p_{1n}$ are nonzero and the rest, $p_n - p_{1n}$, are zero without loss of generality of the ordering of the covariates, $\boldsymbol{\delta}_0 = (\boldsymbol{\delta}_{0(1)}^T, \mathbf{0}_{p_n - p_{1n}}^T)$. Similarly we can define $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_{0(1)}$ for the partially linear coefficients. Further, in order to model high-dimensional complex data, we allow both dimensions of important covariates p_{1n} and q_{1n} to diverge with the cluster size n . Note that the set-up of model (2) is quite general. When there is no partially-linear term and Gaussian identity link function is used, model (2) reduces to model (1) that is applied to the yeast cell cycle data.

For model identifiability, we commonly assume $\|\boldsymbol{\delta}_0\| = 1$ with the first component being positive (Yu and Ruppert, 2002). We estimate the unknown univariate function, $h(\cdot)$, via a polynomial spline approximation defined on an interval with potentially diverging support of $\mathbf{X}\boldsymbol{\delta}_0$. The support $[a, b]$ will be split into intervals based on $\tilde{K}$ interior knots, $(v_1, \dots, v_{\tilde{K}})$. The complete set of knots will be at $a = v_0 < v_1 < \dots < v_{\tilde{K}-1} < v_{\tilde{K}} < v_{\tilde{K}+1} = b$. Here $\tilde{K} \equiv \tilde{K}_n$ increases with the number of clusters n as discussed later for large sample properties. Polynomial splines of degree s guarantee that the approximated function on the intervals between knots are polynomials of order $s - 1$, and the approximated function will attain continuous differentiability up to $s - 2$ times. This holds if the approximated

function is of order $s \geq 2$ (Schumaker, 2007).

We define $\mathbf{B}(u)$ as the s -degree B-spline basis and $\boldsymbol{\psi} = (\psi_1, \dots, \psi_K)^\top$ are the $K \equiv K_n = 1 + s + \tilde{K}$ spline coefficients to estimate. Then, nonparametric estimation of the unknown function becomes $h(u) \approx \mathbf{B}(u)\boldsymbol{\psi} = \sum_{k=1}^K \mathbf{B}_k(u)\psi_k$, a linear combination of the B-spline bases. Overall, we are interested in estimating $\boldsymbol{\theta}_0 = (\boldsymbol{\psi}_0^\top, \boldsymbol{\delta}_0^\top, \boldsymbol{\beta}_0^\top)^\top$ where we approximate $h(\cdot)$ via $\mathbf{B}(\cdot)\boldsymbol{\psi}$ with $\boldsymbol{\eta}_i(\boldsymbol{\theta}_0) = \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0 + \mathbf{Z}_i\boldsymbol{\beta}_0$ and the conditional mean becomes $\boldsymbol{\mu}_i(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0)) = g^{-1}(\mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0 + \mathbf{Z}_i\boldsymbol{\beta}_0)$.

2.2 Generalized Estimating Equations and Large Sample Properties for Oracle Estimator

We examine first the problem when the true important single-index and partially linear variables are known a priori. This is referred to as the oracle estimator. We follow the conventional notation, using subscript 0 for true parameters, $\hat{\cdot}$ for an estimate, and subscript (1) for oracle. For example, $\hat{\boldsymbol{\theta}}_{(1)}$ denotes the oracle estimator of the true parameter $\boldsymbol{\theta}_{0(1)} = (\boldsymbol{\psi}_{0(1)}^\top, \boldsymbol{\delta}_{0(1)}^\top, \boldsymbol{\beta}_{0(1)}^\top)^\top$. Here $\boldsymbol{\delta}_{0(1)}$ are the p_{1n} -dimensional true single-index coefficients for the true important single-index covariates, $\mathbf{X}_{(1)i}$; and $\boldsymbol{\beta}_{0(1)}$ are the q_{1n} -dimensional true linear coefficients for the true partially-linear covariates, $\mathbf{Z}_{(1)i}$. For notational simplicity, we drop the subscript n throughout.

Using GEE (Zeger and Liang, 1986) and spline estimation of the unknown link function, our semiparametric generalized estimating equations are

$$\mathbf{S}(\boldsymbol{\theta}) := \sum_i \mathbf{U}_i(\boldsymbol{\theta}_{(1)})^\top \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_{(1)}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_{(1)}) (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_{(1)}))) = \mathbf{0}, \quad (3)$$

$$\text{where } \mathbf{U}_i^T(\boldsymbol{\theta}_{(1)}) = \begin{pmatrix} \mathbf{B}^T(\mathbf{X}_{(1)i}\boldsymbol{\delta}_{(1)}) \\ \mathbf{J}^T(\boldsymbol{\delta}_{(1)})\mathbf{X}_{(1)i}^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_{(1)i}\boldsymbol{\delta}_{(1)})\boldsymbol{\psi}_{(1)}) \\ \mathbf{Z}_{(1)i}^T \end{pmatrix}. \text{ Here } \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_{(1)})) \text{ and } \mathbf{A}_i(\boldsymbol{\theta}_{(1)})$$

are the spline-approximated marginal mean and variances of $\mathbf{Y}_i$ respectively. $\dot{\mathbf{B}}(\cdot)$ is the first derivative of the B-spline basis. We reparametrize the single-index parameter $\boldsymbol{\delta}_{(1)}$ for the identifiability constraint as in Yu et al. (2017). We treat $\boldsymbol{\delta}_{(1)}$ as a function of $p_1 - 1$ dimensional reparameterized single-index parameter $\boldsymbol{\delta}_{(1)}^{(-1)} = (\delta_2, \dots, \delta_{p_1})^T$, and accordingly define the Jacobian matrix $\mathbf{J}(\boldsymbol{\delta}_{(1)}) = \partial \boldsymbol{\delta}_{(1)} / \partial \boldsymbol{\delta}_{(1)}^{(-1)}$. $\mathbf{R}^{-1}(\gamma)$ is a pre-specified “working” correlation matrix. And γ will be estimated via the residual based moment method for a “working” correlation matrix. Typical working correlation matrices include independence, autocorrelation with lag 1 (AR(1)), or exchangeable. One of the main advantages of GEE is that we do not need to specify a full likelihood function and the “working” correlation matrix needs not be correct (see Wang et al. (2012) and references therein).

We establish the convergence rate and asymptotic normality of the oracle estimators of the semiparametric generalized estimating equations (3) with both diverging p_1 and q_1 , relying on assumptions that are consistent with the literature (Lian et al., 2014). Here $\hat{\boldsymbol{\xi}}_{(1)}$ are estimates of the reparametrized parameters $\boldsymbol{\xi}_{0(1)} = (\boldsymbol{\delta}_{0(1)}^{(-1)T}, \boldsymbol{\beta}_{0(1)}^T)^T$. Please refer to Section 8 for detailed definitions, assumptions, and technical proofs.

Theorem 2.1 (*Convergence rate*) Under assumptions (A1)-(A5) and that $K\sqrt{p_1} + p_1 + q_1 = o(\sqrt{n})$, $r_n\sqrt{K + p_1 + q_1}(\sqrt{K^3 p_1} + p_1 + q_1) = o(1)$, we have $\|\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{0(1)}\| = O_p(r_n)$, where $r_n = \sqrt{(K + p_1 + q_1)/n} + K^{-d}$.

Theorem 2.2 (*Asymptotic normality*) Under assumptions (A1)-(A6) and $\sqrt{nr_n^2}\sqrt{K + p_1 + q_1}(\sqrt{K^3 p_1} + p_1 + q_1) = o(1)$, $\sqrt{n(p_1 + q_1)(K + p_1 + q_1)}r_n K^{-d'} = o(1)$, then for any unit vector $\mathbf{a} \in \mathbf{R}^{p_1 + q_1 - 1}$, $\sqrt{n}\mathbf{a}^T \mathcal{C}_{n(1)}^{-1/2} \mathcal{D}_{n(1)}(\hat{\boldsymbol{\xi}}_{(1)} - \boldsymbol{\xi}_{0(1)}) \xrightarrow{d} N(0, 1)$.

3 Penalized Generalized Estimating Equations for Ultra-high Dimensional Semiparametric Longitudinal Data Analysis

In real applications, ultra-high-dimensional covariates may be present but usually only a subset of those are important for predicting the response. Hence, identifying important variables are critical in estimation, inference, and prediction accuracy. We propose the semiparametric penalized generalized estimating equations to perform simultaneous estimation and variable selection. With the penalty on the single-index and partially linear parameters, our single-index penalized generalized estimating equations are

$$\mathbf{S}^P(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{U}_i^T(\boldsymbol{\theta}) \mathbf{A}_i(\boldsymbol{\theta})^{1/2} \mathbf{R}^{-1} \mathbf{A}_i(\boldsymbol{\theta})^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\eta}_i(\boldsymbol{\theta}))) - n \mathbf{q}_\lambda(|(\boldsymbol{\delta}, \boldsymbol{\beta})|) \mathbf{sgn}(\boldsymbol{\delta}, \boldsymbol{\beta}). \quad (4)$$

We cannot guarantee an exact solution to $\mathbf{S}^P(\boldsymbol{\theta}) = \mathbf{0}$ as there are discontinuous points in $\mathbf{S}^P(\boldsymbol{\theta})$. Thus, we strive to achieve the asymptotic zero crossing (Johnson et al., 2008) which means a small change in the estimated parameters causes a sign change in the PGEE. Here $\mathbf{q}_\lambda(|(\boldsymbol{\delta}, \boldsymbol{\beta})|) = (\mathbf{0}_K, q_\lambda(|\delta_2|), \dots, q_\lambda(|\delta_{p_n}|), q_\lambda(|\beta_1|), \dots, q_\lambda(|\beta_{q_n}|))$ is a $K + p_n + q_n - 1$ vector of the first derivative of penalty functions, $q_\lambda(\cdot)$. $\mathbf{0}_K$ is a K dimensional vector of zeros, and $\mathbf{sgn}(\boldsymbol{\delta}, \boldsymbol{\beta})$ is similarly defined extracting their signs. Together, $\mathbf{q}_\lambda(|(\boldsymbol{\delta}, \boldsymbol{\beta})|) \mathbf{sgn}(\boldsymbol{\delta}, \boldsymbol{\beta})$ is the componentwise product.

We adopt the non-convex SCAD penalty to perform simultaneous variable selection and estimation. Here $\mathbf{q}_\lambda(|(\boldsymbol{\delta}, \boldsymbol{\beta})|)$ is the first derivative vector of the SCAD penalty, defined by $q_\lambda(\zeta) = \lambda \{I(\zeta \leq \lambda) + \frac{(a\lambda - \zeta)_+}{(a-1)\lambda} I(\zeta > \lambda)\}$, for $q_\lambda(0) = 0$ and $a > 2$. We use the SCAD penalty function as it can achieve the oracle property, that is, it performs as if the important variables were known in advance (Fan and Li, 2001). We use $a = 3.7$ as suggested by Fan and Li (2001). Other penalty functions such as LASSO (Tibshirani, 1996) may be used.

We consider the single-index penalized generalized estimating equations (4) in ultra-

high dimensions. We denote the $K + p + q - 1$ single-index PGEE by $\mathbf{S}^P$. Furthermore, $\mathbf{S}_{(1)}^P$ consists of the $K + p_1 + q_1 - 1$ estimating equations corresponding to the relevant predictors and $\mathbf{S}_{(2)}^P$ consists of the remaining $(p - p_1) + (q - q_1)$ equations in $\mathbf{S}^P$.

Theorem 3.1 *Under assumptions for Theorem 2.2, and that*

$$\sqrt{\log(p + q)/n} \ll \min\{\lambda_1, \lambda_2\} \leq \max\{\lambda_1, \lambda_2\} \ll \min\{|\delta_{0j}|, j \leq p_1, |\beta_{0j}|, j \leq q_1\},$$

there is $\check{\boldsymbol{\theta}} = (\check{\boldsymbol{\psi}}^T, \check{\boldsymbol{\delta}}^T, \check{\boldsymbol{\beta}}^T)^T$ that satisfies the following:

$$(i) \ P(\mathbf{S}_{(1)}^P(\check{\boldsymbol{\theta}}) = 0) \rightarrow 1$$

$$(ii) \ \|\mathbf{S}_{(2)}^P(\check{\boldsymbol{\theta}})\|_\infty = o_p(n \min\{\lambda_1, \lambda_2\})$$

$$(iii) \ \|\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(r_n) \text{ where } r_n = \sqrt{(K + p_1 + q_1)/n} + K^{-d} \text{ and } P(\check{\boldsymbol{\delta}}_{(2)} = 0, \check{\boldsymbol{\beta}}_{(2)} = 0) \rightarrow 1$$

$$(iv) \ \text{Let } \check{\boldsymbol{\xi}}_{(1)} = (\check{\boldsymbol{\delta}}_{(1)}^{(-1)T}, \check{\boldsymbol{\beta}}_{(1)}^T)^T. \text{ Then for any unit vector } \mathbf{a} \in \mathbf{R}^{p_1 + q_1 - 1},$$

$$\sqrt{n} \mathbf{a}^T \mathbf{C}_{n(1)}^{-1/2} \mathcal{D}_{n(1)}(\check{\boldsymbol{\xi}}_{(1)} - \boldsymbol{\xi}_{0(1)}) \xrightarrow{d} N(0, 1).$$

4 Algorithm

4.1 An Efficient Algorithm

We devise an efficient iterative algorithm to minimize the single-index penalized generalized estimating equations (4). The main challenge is due to the combination of nonlinearity and ultra-high dimensionality. That is, besides the potentially ultra-high dimensional number of partially linear covariates, the ultra-high dimensional single-index covariates lie within the nonlinear unknown flexible function estimated nonparametrically. Other challenges include the non-convex SCAD penalty functions.

The overall idea of the iterative algorithm is to estimate the spline coefficients and estimate and update the single-index and partially linear parameters iteratively. Note that when the single-index and partially linear coefficients, $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$, are known, this is a univariate longitudinal nonparametric problem (e.g., Hoover et al. (1998) using local methods and smoothing splines). In our implementation to estimate the spline coefficients $\boldsymbol{\psi}$, a linear GEE algorithm of Zeger and Liang (1986) can be simply adopted along with the polynomial B-spline basis. The main challenge lies in the second step given the spline coefficient estimates, we need to tackle the penalized ultra-high-dimensional *nonlinear* GEE problem to simultaneously estimate and select important variables.

We devise an efficient iterative algorithm to first estimate the spline coefficients, and secondly, to estimate and update the single-index and partially linear parameters iteratively. We focus on the more challenging second step of the algorithm. In this step, to tackle this main challenge, a key approximation we adopt is through a linear approximation of $h(\mathbf{X}_i\boldsymbol{\delta})$ around $\mathbf{X}_i\boldsymbol{\delta}_0$ as $h(\mathbf{X}_i\boldsymbol{\delta}) \cong h(\mathbf{X}_i\boldsymbol{\delta}_0) + \text{diag}(\dot{h}(\mathbf{X}_i\boldsymbol{\delta}_0))\mathbf{X}_i(\boldsymbol{\delta} - \boldsymbol{\delta}_0)$. Now that our problem becomes linear with respect to $\boldsymbol{\delta}$, we can use already existing linear methods such as a readily available R package, namely PGEE, from Wang et al. (2012) for our model.

To handle the non-convex SCAD penalty for variable selection, we adopt a similar approach as in Wang et al. (2012) for high-dimensional linear PGEE. As in Wang et al. (2012) and Inan and Wang (2017), for the linear PGEE step, we use the minorization-maximization algorithm (Hunter and Li, 2005) together with the Newton-Raphson method. We follow their approach to handle the non-convex SCAD penalty for variable selection with a quadratic approximation along with a perturbation of the penalty function. The quadratic approximation is based on the original local quadratic approximation idea in Fan and Li (2001) that given a small perturbation $\epsilon > 0$, the SCAD penalty can be

approximated by

$$q_\lambda(|\delta_j|) \cong q_\lambda(|\delta_j^{(0)}|) + \frac{\{\delta_j^2 - (\delta_j^{(0)})^2\}q'_\lambda(|\delta_j^{(0)}|+)}{2(\epsilon + |\delta_j^{(0)}|)}$$

for δ_j around $\delta_j^{(0)}$. As stated in Hunter and Li (2005), when $\delta_j^{(0)} = 0$, the denominator would have made above undefined without perturbation. Intuitively a small perturbation $\epsilon > 0$ is added. We follow Wang et al. (2012) and specify $\epsilon = 10^{-6}$. We refer readers to Hunter and Li (2005) for more details.

Next, we present the details for both steps of our efficient iterative algorithm for ultra-high-dimensional semiparametric longitudinal data analysis. We advocate using a readily available R package, PGEE (Inan and Wang, 2017), for the linear high-dimensional penalized longitudinal model (Wang et al., 2012). In particular, given the estimated spline coefficients $\hat{\psi}$ we break up $\mu_i(\eta_i(\theta)) = g^{-1}(\mathbf{B}(\mathbf{X}_i\hat{\delta})\hat{\psi} + \mathbf{X}_i^*(\delta - \hat{\delta}) + \mathbf{Z}_i\beta)$ into $g^{-1}(\mathbf{C}_i(\hat{\delta}, \hat{\psi}) + \mathbf{X}_i^*\delta + \mathbf{Z}_i\beta)$ where again $\mathbf{X}_i^* = \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i\hat{\delta})\hat{\psi})\mathbf{X}_i$, and $\mathbf{C}_i(\hat{\delta}, \hat{\psi}) = \mathbf{B}(\mathbf{X}_i\hat{\delta})\hat{\psi} - \mathbf{X}_i^*\hat{\delta}$ is a constant value from the previously estimated single-index and spline parameters. The iterative algorithm is the following:

Step 0: Initialize $\hat{\eta}^{(0)} = \begin{pmatrix} \hat{\delta}^{(0)} \\ \hat{\beta}^{(0)} \end{pmatrix}$

Step 1: Given $\hat{\eta}^{(k-1)} = \begin{pmatrix} \hat{\delta}^{(k-1)} \\ \hat{\beta}^{(k-1)} \end{pmatrix}$, estimate the spline coefficients, $\hat{\psi}^{(k-1)}$, by solving

$$\sum_{i=1}^n \mathbf{B}(\mathbf{X}_i\hat{\delta}^{(k-1)})\mathbf{A}_i\left(\psi, \hat{\delta}^{(k-1)}, \hat{\beta}^{(k-1)}\right)^{1/2} \hat{\mathbf{R}}^{-1}\mathbf{A}_i\left(\psi, \hat{\delta}^{(k-1)}, \hat{\beta}^{(k-1)}\right)^{-1/2} \cdot \\ (\mathbf{Y}_i - g^{-1}(\mathbf{B}(\mathbf{X}_i\hat{\delta}^{(k-1)})\psi + \mathbf{Z}_i\hat{\beta}^{(k-1)})) = \mathbf{0},$$

Step 2: Given the estimated spline coefficients $\hat{\psi}^{(k-1)}$, the k_{th} penalized estimator of $\hat{\eta}^{(k)}$

is determined by achieving the asymptotic zero crossing for $\hat{\boldsymbol{\eta}}^{(k)}$ via the penalized GEE

$$\mathbf{Q}(\boldsymbol{\eta}) = \sum_{i=1}^n \mathbf{V}_i(\hat{\boldsymbol{\delta}}^{(k-1)}, \hat{\boldsymbol{\psi}}^{(k-1)})^T \mathbf{A}_i \left(\hat{\boldsymbol{\psi}}^{(k-1)}, \boldsymbol{\delta}, \boldsymbol{\beta} \right)^{1/2} \hat{\mathbf{R}}^{-1} \mathbf{A}_i \left(\hat{\boldsymbol{\psi}}^{(k-1)}, \boldsymbol{\delta}, \boldsymbol{\beta} \right)^{-1/2} \cdot \\ (\mathbf{Y}_i - g^{-1}(\mathbf{C}_i(\hat{\boldsymbol{\delta}}^{(k-1)}, \hat{\boldsymbol{\psi}}^{(k-1)}) + \mathbf{X}_i^* \boldsymbol{\delta} + \mathbf{Z}_i \boldsymbol{\beta})) - n \mathbf{q}_\lambda(|(\boldsymbol{\delta}, \boldsymbol{\beta})|) \mathbf{sgn}(\boldsymbol{\delta}, \boldsymbol{\beta}),$$

with respect to $(\boldsymbol{\delta}, \boldsymbol{\beta})$ and with the identifiability constraints $\|\boldsymbol{\delta}\| = 1$ and $\delta_1 > 0$. $\mathbf{V}_i$ is similar to $\mathbf{U}_i$ previously except now the spline basis coefficients are determined in the previous step. Here $\mathbf{q}_\lambda(|(\boldsymbol{\delta}, \boldsymbol{\beta})|)$ is a vector of the first derivative of penalty functions and $\mathbf{sgn}(\boldsymbol{\delta}, \boldsymbol{\beta})$ is a vector extracting the sign of $(\boldsymbol{\delta}, \boldsymbol{\beta})$. Together, $\mathbf{q}_\lambda(|(\boldsymbol{\delta}, \boldsymbol{\beta})|) \mathbf{sgn}(\boldsymbol{\delta}, \boldsymbol{\beta})$ is the componentwise product.

Step 3: Repeat steps 1 and 2 until convergence.

Remark 4.1 *In many cases, the response is continuous and an identity link function is suitable for generalized estimating equations. In this case, we can further simplify the algorithm to promote better computational efficiency. Specifically, because $\boldsymbol{\mu}_i(\boldsymbol{\eta}_i(\boldsymbol{\theta})) = \mathbf{C}_i(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\psi}}) + \mathbf{X}_i^* \boldsymbol{\delta} + \mathbf{Z}_i \boldsymbol{\beta}$, we can include the constant term, $\mathbf{C}_i(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\psi}})$, with the continuous response by defining $\mathbf{Y}_i^* = \mathbf{Y}_i - \mathbf{C}_i(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\psi}})$. Then the generalized estimating equations can reduce to the following in terms of $\mathbf{X}_i^*$ and $\mathbf{Y}_i^*$: $\mathbf{S}(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{U}_i^T(\boldsymbol{\theta}) \mathbf{A}_i(\boldsymbol{\theta})^{1/2} \mathbf{R}^{-1} \mathbf{A}_i(\boldsymbol{\theta})^{-1/2} (\mathbf{Y}_i^* - \mathbf{X}_i^* \boldsymbol{\delta} - \mathbf{Z}_i \boldsymbol{\beta}) = \mathbf{0}$. The penalized GEE of Step 2 can further reduce to*

$$\mathbf{Q}(\boldsymbol{\eta}) = \sum_{i=1}^n \mathbf{V}_i(\hat{\boldsymbol{\delta}}^{(k-1)}, \hat{\boldsymbol{\psi}}^{(k-1)})^T \mathbf{A}_i \left(\hat{\boldsymbol{\psi}}^{(k-1)}, \boldsymbol{\delta}, \boldsymbol{\beta} \right)^{1/2} \hat{\mathbf{R}}^{-1} \mathbf{A}_i \left(\hat{\boldsymbol{\psi}}^{(k-1)}, \boldsymbol{\delta}, \boldsymbol{\beta} \right)^{-1/2} \cdot \\ (\mathbf{Y}_i^* - \mathbf{X}_i^* \boldsymbol{\delta} - \mathbf{Z}_i \boldsymbol{\beta}) - n \mathbf{q}_\lambda(|(\boldsymbol{\delta}, \boldsymbol{\beta})|) \mathbf{sgn}(\boldsymbol{\delta}, \boldsymbol{\beta}).$$

Hence, a simple modification of the linear high-dimensional PGEE package can now be directly adopted for the proposed semiparametric approach.

Additionally, well performing initial values for the algorithm are important as with

many similar types of optimization problems. For the lower dimensional case, classical linear GEE or penalized GEE can be used. For the ultra-high dimensional case, screening methods for longitudinal data such as the nonparametric independence screening and structure identification method from Cheng et al. (2014) or Xu et al. (2014) may be used to obtain initial values. In our numerical studies, starting values obtained from the sure independence screening (SIS) procedure in Fan and Lv (2008) work quite well. We advocate using a readily available R package, PGEE (Inan and Wang, 2017), for the linear high-dimensional penalized longitudinal model (Wang et al., 2012) and show promising results in our numerical study.

4.2 Tuning Parameter Selection

For the proposed iterative algorithm, tuning parameters include the degree and number of knots of B-spline basis for univariate smoothing, and more critically the penalty parameter λ for variable selection. As in Ma et al. (2014), we place the spline knots equally spaced in the support of the estimated index values. The estimated index value and the knots may change during the iterative estimation algorithm in practice. Quadratic and cubic splines are commonly used. For a smooth univariate function, a small number of knots are usually sufficient (Ma et al. 2014). In practice one may fix the number of knots when it is not very critical (e.g. Huang, Horowitz, and Wei 2010).

In order to choose the critical penalty parameter λ for variable selection, we use a high-dimensional Bayesian information criterion (HBIC) criterion for penalty parameter selection (Lian et al., 2014). We also allow separate penalization of the single-index parameters and the partially linear parameters through λ_1 and λ_2 respectively, which can give more flexibility in modeling. The high-dimensional Bayesian information criterion involves

minimizing

$$HBIC(\lambda_1, \lambda_2) = \log\left(\frac{1}{n} \sum_{i=1}^n (\mathbf{Y}_i - g^{-1}(\mathbf{B}(\mathbf{X}_i \hat{\boldsymbol{\delta}}_{\lambda_1}) \hat{\boldsymbol{\psi}}_{\lambda_1} + \mathbf{Z}_i \hat{\boldsymbol{\beta}}_{\lambda_2}))^2\right) + |M_\lambda| \frac{\log(n)}{2n} C_n.$$

Here $|M_\lambda|$ is the number of single-index and partially linear coefficients selected as important during the algorithm and $C_n = \log(\log(p_n + q_n))$. The log of average sum of squared error under working independence assumption is used as the loss function as in Lian et al. (2014).

Note that there appears no universally agreed choice in high-dimensional BIC for C_n . Our choice follows that of Wang et al. (2009), whose consistency is proved only for moderate dimension. Even for the moderate dimension case as in Wang et al. (2009), this specific choice of C_n is based on the empirical evidence, and their consistency result works for many other choices of C_n . In Fan and Tang (2013), the authors consider more general loss and penalty with ultra-high dimensional case, and the choice of C_n would theoretically depend on a lot of factors including covariate distributions and error distribution. Showing consistency in our setting appears much more complicated. Thus we choose this particular C_n following Wang et al. (2009), also based on our empirical observation that this appears to work well. We acknowledge that there is an obvious gap between theory and practice in this aspect that deserves serious future research.

In order to find the minimum $HBIC(\lambda_1, \lambda_2)$ value, we can iteratively evaluate this measure with a grid of values for λ_1 and λ_2 . Alternatively, we can consecutively solve for one λ_i while the other is fixed and repeat for the other in order to reduce computational burden (Ruppert and Carroll, 2000). One may also use cross validation. That is, we may remove one fold at a time of all the data and use the rest of the folds as training data to fit the model. The penalty parameter can be chosen over a grid search to minimize the average prediction loss for all the testing folds. In our numerical studies, we find that

the high-dimensional Bayesian information criterion method of tuning parameter selection performs well.

For the number and placement of knots tuning parameters as in Ma et al. (2014), we place the spline knots equally spaced in the support of the estimated index values. The estimated index value and the knots may change during the iterative estimation algorithm in practice. Quadratic and cubic splines are commonly used. For a smooth univariate function, a small number of knots are usually sufficient (Ma et al. 2014). In practice one may fix the number of knots when it is not very critical (e.g. Huang, Horowitz, and Wei 2010). We find quadratic splines perform well in our simulation study and the Yeast Cell Cycle study. We have included a simulation study on varying number of knots and degrees for Example 1 in Table 8.

5 Numerical Study

We conduct Monte Carlo studies to assess the performance of the proposed estimation and variable selection approaches. In the Example 1 and 2 below, we use quadratic spline and three equally-spaced knots. In Section 5.2, we present additional results for higher-dimensional covariates, a different correlation level and a complex correlation structure, additional settings, and unequal m_i per subject etc. All simulations are based on 200 replications.

5.1 Simulation

Example 1. The correlated Gaussian response data are generated from the following model: $Y_{ij} = \sin((\mathbf{X}_{ij}^T \boldsymbol{\delta} - a)\pi)/(b - a) + \mathbf{Z}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij}$. The number of subjects is $n = 100$ and the number of observations per subject is $m = 10$. The true parameter vectors are $\boldsymbol{\delta}_0 = (1, 1, 1, 1, 1, 0, \dots, 0)^T / \sqrt{5}$ and $\boldsymbol{\beta}_0 = (1, 1, 1, 0, \dots, 0)$. The error terms ϵ_{ij} are drawn

from a multivariate normal distribution $N_{10}(0, \sigma^2 R(\rho))$ with mean 0 and an exchangeable correlation structure with $\rho = 0.6$ and error variance $\sigma^2 = 0.2$. The covariates $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are drawn from a multivariate normal distribution with mean 0 and an AR(1) correlation 0.5 and error variance 0.2. The constants $a = \sqrt{3}/2 - 1.645/\sqrt{12}$, $b = \sqrt{3}/2 + 1.645/\sqrt{12}$, and the number of covariates are $p_n = 250$ and $q_n = 250$.

As shown in Table 1, the parameter estimates for our proposed longitudinal single-index model are close to the true parameter values with small bias and standard error, using the exchangeable working correlation structure with SCAD penalty. On average it takes about 10 iterations for the single-index PGEE algorithm to converge. Figure 2 shows the fitted curve plot for $h(\cdot)$. We observe that the fitted curve virtually overlays the true function. Table 2 provides the estimation and variable selection results. We compare to the performance of the linear PGEE model (Wang et al., 2012); as well as single-index fit under independence working correlation structure; and single-index fit under LASSO penalty that we implement. The mean squared errors (MSE) for both the single-index and partially linear parameter estimates are lowest using our proposed method. In regards to the variable selection performance, in our limited study, our methods have better exact correct percentages (100%) for both the single-index portion and the partially linear portion in comparison to the linear model. In this limited study, the model with LASSO penalty tends to select more variables that are not actually important, as evidenced by the smaller true negatives (TN) in Table 2, which results in close to 0% correct percentage for the partially linear portion.

Table 1: Summary of Parameter Estimates for the Partially Linear Single-Index Model in Example 1 with $p_n = 250$ and $q_n = 250$. The true $\delta_0 = (1, 1, 1, 1, 1, 0, \dots, 0)^T / \sqrt{5}$ and the true $\beta_0 = (1, 1, 1, 0, \dots, 0)^T$. Example 1 has $n = 100$ subjects and $m = 10$ observations per subject, and the true error correlation structure is exchangeable with $\rho = 0.6$ and the error variance is 0.2. The sample mean, bias, and standard error are calculated over 200 simulations for single-index and partially linear parameter estimates.

	mean	bias	se
δ_1	0.4204	-0.0268	0.0120
δ_2	0.4554	0.0082	0.0121
δ_3	0.4476	0.0004	0.0108
δ_4	0.4337	-0.0135	0.0122
δ_5	0.4761	0.0289	0.0094
β_1	0.9972	-0.0028	0.0303
β_2	1.0128	0.0128	0.0298
β_3	0.9848	-0.0152	0.0312

Figure 2: Fitted plot for the curve of $h(\cdot)$ from Example 1 with $p_n = 250$ and $q_n = 250$. This figure is based on 200 simulation replications. The solid line is the true function. The dot dashed line is the average fitted curve and the dashed lines are the corresponding 2.5% and 97.5% quantiles.

Table 2: Summary of Estimation and Variable Selection Results from Example 1 with $p_n = 250$ and $q_n = 250$. Example 1 has $n = 100$ subjects and $m = 10$ observations per subject, and the true error correlation structure is exchangeable with $\rho = 0.6$ and the error variance is 0.2. The true single-index parameters $\boldsymbol{\delta}_0 = (1, 1, 1, 1, 1, 0, \dots, 0)^T / \sqrt{5}$ and the true partially linear parameters $\boldsymbol{\beta}_0 = (1, 1, 1, 0, \dots, 0)^T$. “Method” displays the models used in each comparison, where we implement our proposed longitudinal single-index models via semiparametric penalized GEE with SCAD penalty using exchangeable and independent correlation structure respectively; and with LASSO penalty using exchangeable correlation structure. Here the longitudinal linear model uses SCAD penalty as in the PGEE R package. Correct % is the percentage of 200 simulation replications where the exact true model is selected. “TNs” is the average of the true negatives. “FNs” is the average of the false negatives. “MSEp” is the mean squared error of the single-index parameter estimates or $\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\|^2$ averaged over simulation replications. “MSEq” is the same respective calculation for the partially linear parameter estimates.

Comparison A.

		Single-index Covariates				Partially Linear Covariates			
Method		Correct%	TNs	FNs	MSEp	Correct%	TNs	FNs	MSEq
Single-index	Single-index	100	245	0	0.0006	100	247	0	0.0026
	Linear	93	244.89	0	0.0206	86	246.79	0	0.0044

Comparison B.

		Single-index Covariates				Partially Linear Covariates			
Method		Correct%	TNs	FNs	MSEp	Correct%	TNs	FNs	MSEq
Single-index	Exchangeable	100	245	0	0.0006	100	247	0	0.0026
	Independence	75	244.72	0	0.0039	4	242.83	0	0.0320

Comparison C.

		Single-index Covariates				Partially Linear Covariates			
Method		Correct%	TNs	FNs	MSEp	Correct%	TNs	FNs	MSEq
Single-index	SCAD	100	245	0	0.0006	100	247	0	0.0026
	LASSO	53	244.39	0	0.0038	0	236.01	0	0.1975

Example 2. We consider the correlated binary responses from the marginal mean responses $\log(p_{ij}/(1 - p_{ij})) = \sin((\mathbf{X}_{ij}^T \boldsymbol{\delta} - a)\pi)/(b - a)$. The number of subjects is $n = 200$ and the number of observations per subject is $m = 10$. The true single-index vector $\boldsymbol{\delta}_0 = (1, 1, 1, 1, 0, \dots, 0)^T/\sqrt{4}$ and the number of single-index covariates $p_n = 100$. The covariates are sampled from an independent uniform distribution with mean 0 and variance 1. The correlated discrete binary outcome data is generated using the NORTA method (Touloumis, 2016). The latent exchangeable correlation structure with $\rho = 0.9$ is used.

As shown in Table 3, the variable selection results for the correct percentage are comparable to or better than those in the literature for binary responses, e.g. the longitudinal linear model (Wang et al., 2012). In addition, the false negatives (FNs) are zero and the number of true negatives (TNs) for all working correlation structures is close to the correct number of true negatives.

Table 3: Summary of Estimation and Variable Selection Results from Example 2 with Binary Response. The true single-index vector $\boldsymbol{\delta}_0 = (1, 1, 1, 1, 0, \dots, 0)^T/\sqrt{4}$ and the number of single-index covariates $p_n = 100$. Correct % is the percentage of 200 simulation replications where the exact true model is selected. “TNs” is the average of the true negatives, and “FNs” is the average of the false negatives. “MSEp” is the mean squared error of the single-index parameter estimates or $\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\|^2$ averaged over simulation replications.

	Correct%	TNs	FNs	MSEp
Independence	73	95.69	0	0.031
AR(1)	68	95.97	0	0.034
Exchangeable	72	95.67	0	0.032

5.2 Additional Simulation Studies

In this section, we conduct a variety of additional Monte Carlo studies to assess the performance of the proposed estimation and variable selection approaches using a continuous response. We compare the performance of our proposed method to the linear PGEE model (Wang et al., 2012); as well as the single-index fit under independent working correlation

structure; and single-index fit under LASSO penalty that we implement. All studies use the same design as in Example 1 unless otherwise noted. Here $p_n = 250$, $q_n = 250$, and $\rho = 0.6$ for Tables 4 - 8:

- two additional settings of non-constant true parameter values (Tables 4 & 5)
- a complex correlation structure (Table 6)
- unequal m_i per subject (Table 7)
- varying degree and knots (Table 8)
- additional estimation results with correlation coefficient $\rho = 0.2$ (Table 9)
- higher dimensional covariates $p_n = 2500$ and $q_n = 2500$ (Tables 10 & 11).

Also in Table 8 when $p_n = 2500, q_n = 2500$, we compare the estimation and variable selection performance to the

- linear PGEE model (Wang et al., 2012)
- single-index fit under independent working correlation structure
- single-index fit under LASSO penalty.

Over the study of various scenarios and comparisons for both the single-index and partially linear portions, we find that our proposed method provides effective variable selection and the parameter estimates are close to the true parameter values.

Because of the key approximation of $h(\mathbf{X}_i\boldsymbol{\delta})$ around $\mathbf{X}_i\boldsymbol{\delta}_0$, we essentially convert a complex nonlinear problem to a linear problem in step 2 of the algorithm. In terms of the computational complexity of the algorithm for the proposed longitudinal single-index model in practice, for our Example 1 with $p_n = 250$ and $q_n = 250$, on average it takes about

10 iterations for the single-index penalized GEE algorithm to converge at the convergence level 10^{-5} . This is equivalent to about 10 times of the computational time for the linear PGEE algorithm of Inan and Wang (2017). In particular, the average number of iterations is 10.56 with a standard deviation of 1.57 over 200 simulation replications. In the oracle scenario when no variable selection is involved, in this model setting, on average it takes about 8 iterations for the single-index GEE algorithm to converge.

We use mean, bias, and standard error (se) to assess parameter estimation performance of the single-index and partially linear parameter estimates. We assess the variable selection results using true negatives (TN), the average number of covariates the model selects as zero which are actually zero, and false negatives (FN) the average number of non-zero covariates that are incorrectly set to zero. We also use correct percentage (Correct %), the percentage of simulation replications with the exact true non-zero covariates correctly identified, to determine variable selection performance.

We present two additional settings of non-constant true parameter values. In Scenario 1, the true single-index parameter vector $\delta_0 = (0.1796, 0.1796, 0.3592, 0.5388, 0.7184, 0, \dots, 0)$ is the normalized vector of $(0.5, 0.5, 1, 1.5, 2, 0, \dots, 0)$. In Scenario 2, the true single-index parameter vector $\delta_0 = (0.2294, 0.2294, 0.4588, -0.4588, -0.6882, 0, \dots, 0)$ is the normalized vector of $(0.5, 0.5, 1, -1, -1.5, 0, \dots, 0)$. We keep the same setup as in Example 1 otherwise. As shown in Tables 4 and 5, the single-index and partially linear parameter estimates are close to the true parameter values.

Under the same design as Example 1, we also conduct a study under a more complex correlation structure. The correlation structure we use in this scenario is a mixture of an AR(1) correlation structure and an exchangeable correlation structure of $\rho = 0.6$ and error variance $\sigma^2 = 0.2$. Specifically, we simulate the errors with $Cov(e_{ij}, e_{ik}) = \sigma^2(\rho^{|j-k|}/2 +$

Table 4: Summary of Parameter Estimates for Example 1 with Scenario 1 Design. The true $\boldsymbol{\delta}_0 = (0.1796, 0.1796, 0.3592, 0.5388, 0.7184, 0, \dots, 0)$, which is the normalized vector of $(0.5, 0.5, 1, 1.5, 2, 0, \dots, 0)$ and the true $\boldsymbol{\beta}_0 = (1, 1, 1, 0, \dots, 0)^T$. The true correlation structure is exchangeable with $\rho = 0.6$ and the error variance is 0.2. The sample mean, bias, and standard error are calculated over 200 simulations for the single-index and partially linear parameter estimates. Here $p_n = 250$ and $q_n = 250$.

	mean	bias	se
δ_1	0.1815	0.0019	0.0130
δ_2	0.1654	-0.0142	0.0142
δ_3	0.3710	0.0118	0.0111
δ_4	0.5338	-0.0050	0.0132
δ_5	0.7186	0.0001	0.0113
β_1	1.0078	0.0078	0.0302
β_2	1.0017	0.0017	0.0291
β_3	0.9919	-0.0081	0.0311

$\rho/2$) for $i = 1, \dots, n$; $j, k = 1, \dots, m$; when $j \neq k$; and $Cov(e_{ij}, e_{ik}) = \sigma^2$ when $j = k$. Table 6 shows good parameter estimation results even under a misspecified working correlation matrix for an underlying complex correlation structure. This is consistent to the literature that GEE (Zeger and Liang, 1986) does not require knowing the correlation structure and can yield a consistent estimator even if the working correlation structure is misspecified.

Next we investigate unequal m_i per subject. In this case, we simulate 100 subjects with 10 observations per subject and 100 subjects with 5 observations per subject. The results in Table 7 show that the estimated parameters are close to the true parameters. The variable selection results have a 100% correct percentage. Our algorithm uses the high-dimensional linear PGEE R package (Inan and Wang (2017)) and classical linear GEE as a base, thus it is quite natural to incorporate unbalanced data under unequal m_i settings.

In addition, we implement a simulation study incorporating varying number of knots and degrees. We investigate the performance of our method with quadratic splines (degree=2) and cubic splines (degree=3) with 2, 3, and 5 number of knots of the B-spline basis

Table 5: Summary of Parameter Estimates for Example 1 with Scenario 2 Design. The true $\boldsymbol{\delta}_0 = (0.2294, 0.2294, 0.4588, -0.4588, -0.6882, 0, \dots, 0)$, which is the normalized vector of $(0.5, 0.5, 1, -1, -1.5, 0, \dots, 0)$ and the true $\boldsymbol{\beta}_0 = (1, 1, 1, 0, \dots, 0)^T$. The true correlation structure is exchangeable with $\rho = 0.6$ and the error variance is 0.2. The sample mean, bias, and standard error are calculated over 200 simulations for the single-index and partially linear parameter estimates. Here $p_n = 250$ and $q_n = 250$.

	mean	bias	se
δ_1	0.2168	-0.0126	0.0248
δ_2	0.2258	-0.0036	0.0281
δ_3	0.4568	-0.0020	0.0219
δ_4	-0.4599	-0.0011	0.0227
δ_5	-0.6922	-0.0040	0.0163
β_1	1.0046	0.0046	0.0386
β_2	0.9951	-0.0049	0.0411
β_3	0.9971	-0.0029	0.0388

for univariate spline approximation. As shown in Table 8, these combinations of degree and number of knots have a correct % of 100 and achieve the correct number of true negatives and false negatives. In our study, the choice of these specific tuning parameters does not seem to largely influence results.

The good performance of our algorithm in the numerical study section under a similar design as Example 1 is also confirmed in Table 9 below. In this case $p_n = 500$ and $q_n = 500$ and the covariates $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are drawn from a Uniform(0,1) distribution. The error terms are drawn from a multivariate normal distribution with mean 0 and AR(1) correlation structure with $\rho = 0.2$ and error variance $\sigma^2 = 0.2$.

Finally in Tables 10 & 11, we consider the same Example 1 design but with higher dimensional covariates where $p_n = 2500$ and $q_n = 2500$. As shown in Table 10 again the parameter estimates are close to the true values for both the single-index and partially linear parameter terms. Table 11 provides the estimation and variable selection results. We compare to the performance of the linear PGEE model (Wang et al., 2012); as well as single-index fit under independent working correlation structure; and single-index fit

Table 6: Summary of Parameter Estimates for Example 1 with a Complex Correlation Structure. The true $\boldsymbol{\delta}_0 = (1, 1, 1, 1, 1, 0, \dots, 0)^T / \sqrt{5}$ and the true $\boldsymbol{\beta}_0 = (1, 1, 1, 0, \dots, 0)^T$. The sample mean, bias, and standard error are calculated over 200 simulations for the single-index and partially linear parameter estimates. Here $p_n = 250$ and $q_n = 250$.

	mean	bias	se
δ_1	0.4205	-0.0267	0.0153
δ_2	0.4549	0.0077	0.0164
δ_3	0.4471	-0.0001	0.0136
δ_4	0.4332	-0.0140	0.0157
δ_5	0.4770	0.0298	0.0122
β_1	0.9949	-0.0051	0.0392
β_2	1.0137	0.0137	0.0375
β_3	0.9888	-0.0112	0.0352

under LASSO penalty that we implement. Our method has the lowest mean squared errors (MSE) of both the single-index and partially linear estimates. For variable selection, our proposed methods have higher correct percentage(%) than the other methods. In addition, our method achieves better results for the average number of true negatives (TN) most notably for the partially linear covariates. In this case, based on our simulation study, the model with the LASSO penalty has close to 0% correct percentage for the partially linear portion as it tends to select more variables that are not actually important, as evidenced by the smaller true negatives (TN) in Table 11.

Table 7: Summary of Parameter Estimates for the Partially Linear Single-Index Model with Unequal m_i with $p_n = 250$ and $q_n = 250$. The true $\delta_0 = (1, 1, 1, 1, 1, 0, \dots, 0)^T/\sqrt{5}$ and the true $\beta_0 = (1, 1, 1, 0, \dots, 0)^T$. This unequal m_i analysis has $n = n_1 + n_2$ subjects. $n_1 = 100$ subjects with $m = 10$ observations per subject and $n_2 = 100$ subjects with $m = 5$ observations per subject, and the true error correlation structure is exchangeable with $\rho = 0.6$ and the error variance is 0.2. The sample mean, bias, and standard error are calculated over 200 simulations for single-index and partially linear parameter estimates.

	mean	bias	se
δ_1	0.4251	-0.0221	0.0095
δ_2	0.4523	0.0051	0.0098
δ_3	0.4483	0.0011	0.0094
δ_4	0.4391	-0.0082	0.0112
δ_5	0.4695	0.0223	0.0082
β_1	1.0018	0.0018	0.0234
β_2	1.0040	0.0040	0.0260
β_3	1.0064	0.0064	0.0228

Table 8: Summary of Parameter Estimates for the Partially Linear Single-Index Model in Example 1 with Varying Number of Degrees and Knots with $p_n = 250$ and $q_n = 250$. The true $\delta_0 = (1, 1, 1, 1, 1, 0, \dots, 0)^T/\sqrt{5}$ and the true $\beta_0 = (1, 1, 1, 0, \dots, 0)^T$. Example 1 has $n = 100$ subjects and $m = 10$ observations per subject, and the true error correlation structure is exchangeable with $\rho = 0.6$ and the error variance is 0.2. The sample mean, bias, and standard error are calculated over 200 simulations for single-index and partially linear parameter estimates.

Degree	Knots	Correct%	TNs	FNs	MSEp	Correct%	TNs	FNs	MSEq
3	2	100	245	0	0.0010	100	247	0	0.0027
3	3	100	245	0	0.0006	100	247	0	0.0032
3	5	100	245	0	0.0006	100	247	0	0.0032
2	2	100	245	0	0.0027	100	247	0	0.0046
2	3	100	245	0	0.0006	100	247	0	0.0026
2	5	100	245	0	0.0006	100	247	0	0.0032

Table 9: Summary of Parameter Estimates for the Partially Linear Single-Index Model in Example 1 with $p_n = 500$ and $q_n = 500$. The true $\delta_0 = (1, 1, 1, 1, 1, 0, \dots, 0)^T/\sqrt{5}$ and the true $\beta_0 = (1, 1, 1, 0, \dots, 0)^T$. Example 1 has $n = 100$ subjects and $m = 10$ observations per subject, and the true error correlation structure is exchangeable with $\rho = 0.2$ and the error variance is 0.2. The sample mean, bias, and standard error are calculated over 200 simulations for single-index and partially linear parameter estimates.

	mean	bias	se
δ_1	0.4491	0.0019	0.0182
δ_2	0.4489	0.0017	0.0183
δ_3	0.4422	-0.0049	0.0182
δ_4	0.4475	0.0003	0.0178
δ_5	0.4462	-0.0010	0.0185
β_1	0.9990	-0.0010	0.0488
β_2	1.0037	0.0037	0.0541
β_3	0.9988	-0.0012	0.0530

Table 10: Summary of Parameter Estimates for the Partially Linear Single-Index Model in Example 1 with $p_n = 2500$ and $q_n = 2500$. The true $\delta_0 = (1, 1, 1, 1, 1, 0, \dots, 0)^T/\sqrt{5}$ and the true $\beta_0 = (1, 1, 1, 0, \dots, 0)^T$. Example 1 has $n = 100$ subjects and $m = 10$ observations per subject, and the true error correlation structure is exchangeable with $\rho = 0.6$ and the error variance is 0.2. The sample mean, bias, and standard error are calculated over 200 simulations for single-index and partially linear parameter estimates.

	mean	bias	se
δ_1	0.4534	0.0062	0.0167
δ_2	0.4474	0.0001	0.0274
δ_3	0.4448	-0.0025	0.0166
δ_4	0.4464	-0.0008	0.0159
δ_5	0.4415	-0.0057	0.0266
β_1	0.9994	-0.0006	0.0332
β_2	0.9923	-0.0077	0.0318
β_3	0.9988	-0.0012	0.0270

5.3 An Application to Yeast Cell Cycle Data

Investigating the influences of gene expression during the cell cycle process can give information into how the cell cycle affects biological processes and insights into cell cycle regulation. The yeast cell cycle data is from the CDC15 experiment which collected genome-wide mRNA levels of 6,178 yeast open reading frames for 18 time points or 119 minutes divided into 7 minute intervals (Spellman et al. (1998)), where certain genes are found to change periodically throughout the yeast cell cycle process, but the influences of these changes were unknown. Transcription factors (TFs) (Lee et al. (2002) and Wang et al. (2007)) are a critical part of the cell cycle process, where transcription factors have been shown to influence gene expression by regulating the flow of genetic information from DNA to mRNA by acting as binding agents to certain DNA sequences during the cell cycle process. We are interested in selecting which transcription factors from a large set of candidates are associated with yeast gene expression levels while simultaneously allowing flexible modeling and incorporating some interactions among these TFs.

Flowchart 1 shows the data integration process. Spellman et al. (1998) found that 800 genes are changed periodically by unknown influences. Then Luan and Li (2003) further identified a subset of $n = 297$ genes regulated during the cell cycle based on α -factor synchronization experiments. As in Wang et al. (2012), the response $\mathbf{Y}_i$ is the log-transformed gene expression level at time point $j = 1, \dots, m$ for gene $i = 1, 2, \dots, n$. We use 96 transcription factors $\mathbf{X}_{ik}$, which are the matching scores for each transcription factor k in relation to gene i . The matching score is the binding probability of each transcription factor on the promoter region of a gene. These binding probabilities were computed via a mixture modeling approach from Wang et al. (2007) utilizing the ChIP dataset of Lee et al. (2002). Specifically, first using the ChIP binding experiment, a significance test for no enrichment of gene i from transcription factor k is performed using the standardized

enrichment variables from the experiment. In order to link this information to identify specific enriched genes, the p-values from this hypothesis test are mapped to a normal Z score. These then create a distribution having a mixture of 2 groups: an unenriched group and an enriched group. The final binding probability is the probability that a latent variable is in the enriched group or the unenriched group given the data which can be estimated using the EM algorithm. Please see Section 2.1 of Wang et al. (2007) for more details.

We apply our proposed longitudinal single-index model along with PGEE and SCAD penalty to investigate the relationship between the longitudinal response gene expression $\mathbf{Y}$ and 96 transcription factors along with time t , that is, $\mathbf{Y}_i = h(\sum_{k=0}^{96} \delta_k \mathbf{X}_{ik}) + \epsilon_i$. Here Gaussian identity link function is used for the continuous response gene expression. For notational simplicity, X_{i0} corresponds to time t , and the following $\mathbf{X}_{ik}$ with $k = 1, \dots, 96$ covariates are transcription factors. The cell cycle process consists of cell growth, replication of DNA, separation of chromosomes, and then division of a cell. Commonly, this process is divided into 5 stages. We will focus on the G1 or “GAP 1” stage of the cell cycle process in order to benchmark with Inan and Wang (2017) and Wang et al. (2012). As in Inan and Wang (2017) and Wang et al. (2012), transcription factors are standardized to have mean 0 and standard deviation of 1 and no penalty is added on time t . When $h(\cdot)$ is an identity function, the single-index model reduces to the linear PGEE model of Wang et al. (2012), i.e., $\mathbf{Y}_i = \alpha_0 + \sum_{k=0}^{96} \delta_k \mathbf{X}_{ik} + \epsilon_i$. Note that complex relationship and potential interactions among transcription factors and time cannot be captured through this linear model.

Consistent to the previous research, we find that important transcription factors such as MBP1 and SWI6 are selected both under our longitudinal single-index model (1) and the linear longitudinal model of Wang et al. (2012) during the G1 stage. This is not

surprising as the covariate MBP1 is known to be highly active in the G1 phase (Cheng and Li, 2008). In addition, SWI6 along with MBP1 form the complex MBF which is involved in the regulation of genes in the G1 stage as discussed in Cheng and Li (2008); Bähler (2005); Koch and Nasmyth (1994). Specifically, at the final portion of the G1 stage cells can begin DNA replication and other processes. The complex MBF is a critical component involved in regulating this progression by activating expression of genes in late G1 phase (Bähler, 2005). We find that ten transcription factors, ABF1, FKH1, FKH2, GAT3, GCR2, MBP1, NDD1, STB1, SWI4, and SWI6, are commonly selected under the single-index model and the linear model. The single-index model has selected a total of 25, 23, 24 TFs while the linear model has selected 16, 12, 11 TFs under the independence, AR(1), and exchangeable correlation structure respectively.

Note that the important cooperative interaction between MBP1 and SWI6 (Banerjee and Zhang, 2003; Das et al., 2004; Wang et al., 2007; Cheng and Li, 2008) cannot be captured under the linear model of Wang et al. (2012). This is also reflected in the better prediction error under the longitudinal single-index model. In particular, Table 12 reports the model mean squared errors (MSE) and model prediction errors (PE) under the selected longitudinal single-index model and linear model. For this particular application, the single-index model gives smaller prediction error compared with the linear model. The single-index model under independence structure gives smallest prediction error with largest model size. We choose the tuning parameter λ for the SCAD penalty via grid search, where the best λ corresponds to the lowest HBIC value as discussed in Section 4.2. The estimation of flexible univariate function $h(\cdot)$ uses quadratic spline basis with 1 knot that yields good results. The prediction error PE is computed through a 10-fold cross validation. The corresponding standard errors of model MSE and model PE are computed through bootstrap sampling by subject of the chosen models for 200 replications.

Table 12: Summary of Estimation Results from the Yeast Cell Cycle Gene Expression Data. “Model Type” displays the model used. “Single-index Model” refers to our proposed high-dimensional longitudinal single-index model using penalized GEE, and “Linear Model” refers to the longitudinal linear PGEE model. “MSE” is the model mean squared error, and “PE” refers to the mean prediction error resulting from a 10-fold cross validation of the data. The standard errors “se” are based on 200 bootstrap samples.

	Model Type	MSE(se)	PE(se)
Independence	Single-index Model	0.2279(0.0133)	0.2316(0.0132)
	Linear Model	0.2576(0.0144)	0.2855(0.0143)
AR-1	Single-index Model	0.2312(0.0130)	0.2353(0.0134)
	Linear Model	0.2648(0.0141)	0.2879(0.0156)
Exchangeable	Single-index Model	0.2305(0.0131)	0.2339(0.0134)
	Linear Model	0.2677(0.0144)	0.2633(0.0159)

One very interesting finding is that the transcription factor YAP5, noted as an important factor in the previous literature (e.g. Banerjee and Zhang (2003)), is constantly selected under the single-index model but not at all under the linear model. In addition, TFs such as MET31 as noted in (Tsai et al. (2005)) and GCR1 among 21 TFs found in Song et al. (2014) are also selected under the single-index model but not under the linear model. On the other hand, TFs such as MSN4, PHD1, RGM1, RLM1, SMP1, SRD1 are selected under the linear model but not under the single-index model. Among them, MSN4 was also noted as an important factor in Tsai et al. (2005). To the best of our knowledge, the rest 5 TFs have not been commonly noted in the previous literature for Stage G1.

6 Discussion

In this essay, we examine the partially linear single-index models using penalized generalized estimating equations in order to potentially capture nonlinearity, maintain interpretability, and reduce dimensionality of ultra-high longitudinal data. Most significantly,

we allow ultra-high dimensional predictors both within the nonparametric component along with the linear portion. Computationally, we invoke a linear approximation on the nonparametric portion as well as institute an approximation method to handle the non-convex SCAD penalty. This efficient algorithm simultaneously performs variable selection and estimation. We contribute to the theory in establishing asymptotic properties for the estimators including the oracle property in ultra-high dimension for both the linear and more importantly the nonparametric single-index components.

We acknowledge that automatically selecting which variables to include in the unknown flexible function as opposed to the linear portion of the partially linear single-index model is an important yet challenging problem to solve. We wish to explore further in future research.

7 Penalized Quadratic Inference Function for Semiparametric Longitudinal Data Analysis in Ultra-high Dimensions

Numerous ultra-high dimensional public health and genetics research studies are longitudinal, where participants have measurements taken at multiple times. In many of these studies, not only are the covariates high dimensional with a few sparse important variables, requiring accurate variable selection, but also the correlated response can be discrete, creating a challenge in modeling the joint likelihood for correlated data. Further, as demonstrated in recent research such as in Wang et al. (2014) and Dogan et al. (2018), inflexible models with parametric assumptions may not account for both linear and nonlinear effects of covariates in high dimensional longitudinal data.

In longitudinal studies, incorporating the within-subject dependence in statistical modeling can lead to more efficient estimation and accuracy. Moreover, the outcome of interest can be a correlated discrete response. This poses a challenge as the full joint likelihood can be intractable for correlated discrete data. To alleviate this difficulty of specifying the full joint likelihood for correlated data, Zeger and Liang (1986) proposed generalized estimating equations to account for within-subject correlation. Notably, the resulting coefficient estimates from the generalized estimating equations method are consistent even if the working correlation matrix is misspecified (Zeger and Liang, 1986). However, under a misspecified working correlation matrix, the resulting coefficient estimates from the generalized estimating equations approach may be inefficient, as demonstrated in Qu et al. (2000).

To pursue estimation efficiency, Qu et al. (2000) developed the quadratic inference function for longitudinal studies. Unlike the generalized estimating equations approach, the quadratic inference function need not estimate the nuisance parameters from the inverse of the working correlation matrix. This may be especially beneficial in a high dimensional

setting with likely even more nuisance parameters (Cho and Qu, 2013). Further, the quadratic inference function is applicable in a variety of model setups and shows promising results such as in Qu and Li (2006), Wang et al. (2012), and Wang et al. (2014).

In this section, we employ the penalized quadratic inference function to account for within-subject correlation, perform model selection, and pursue efficient estimation for high dimensional and potentially ultra-high dimensional longitudinal data. Along with the penalized quadratic inference function, we use the generalized partially linear single-index model to balance flexibility for accurate estimation and interpretability. We provide methodology for this approach and propose an estimation algorithm for computational expediency and effectiveness. We note that since this is a stand-alone section, the notation is intentionally different from the previous sections.

Due to the many advantages of both the quadratic inference function and partially linear single-index models, a handful of works have proposed research employing both procedures together, for example in Bai et al. (2009), Ma et al. (2014), Lai et al. (2013), and Li et al. (2015). However, all of these works assume the dimension of both the single-index and partially linear covariates is fixed. Without variable selection, Bai et al. (2009) developed a procedure for single-index models by combining penalized splines and the quadratic inference function method with discrete and continuous responses. Also, Lai et al. (2013) implemented a bias-corrected quadratic inference function method for partially linear single-index models estimated via local linear smoothing. A few approaches incorporated variable selection for the partially linear single-index model employing the quadratic inference function. Ma et al. (2014) used the penalized quadratic inference function for variable selection and estimation and established asymptotic theory. They also provided a two-step procedure employing the profile principle to estimate parameters. Li et al. (2015) implemented variable selection through the penalized bias-corrected version

of the quadratic inference function along with theoretical justifications for the estimators with a continuous response.

Considering a growing number of longitudinal studies are high dimensional, it is especially pertinent to develop new approaches where the covariates are allowed to diverge. Importantly, 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 grow even in the exponential order, $\exp(n^{1/2})$, with the number of clusters. As a result, we navigate additional challenges in computation due to allowing ultra-high dimensional data within the highly-nonlinear, unknown, flexible function with potential diverging support of the single-index.

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 highly, nonlinear unknown function estimated nonparametrically, together with the non-convex smoothly clipped absolute deviation (SCAD) penalty function, all within a longitudinal framework. To address these challenges, we provide a computationally efficient iterative algorithm to select the sparse set of important single-index and linear covariates and to estimate the corresponding coefficients. Our approach implements strategic approximations to reduce the computational burden and increase the effectiveness of the algorithm.

7.1 Quadratic Inference Function for Semiparametric Longitudinal Data Analysis

For each subject i with $i = 1, \dots, n$, assume the longitudinal response $Y_{i,t}$ has correlation among its observations, $t = 1, \dots, T_i$, but that it is independent of other subjects. We observe the $T_i \times 1$ response vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT_i})^T$, the $T_i \times p_n$ dimensional single-

index covariate vector $\mathbf{X}_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{iT_i})^T$, and the $T_i \times q_n$ dimensional linear covariate vector $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \dots, \mathbf{Z}_{iT_i})^T$ over T_i total time points. Notably, we allow the single-index and partially linear covariates to diverge and potentially be on the order of $\exp(n^{1/2})$. Also, the observations T_i can be different for each subject i . We reiterate that since this is a stand-alone section, the notation is intentionally different from the previous sections.

To allow flexibility but retain some interpretability, we consider the generalized partially linear single-index model

$$E(\mathbf{Y}_i | \mathbf{X}_i, \mathbf{Z}_i) := g^{-1}(\eta(\mathbf{X}_i \boldsymbol{\beta}_0) + \mathbf{Z}_i \boldsymbol{\gamma}_0), \quad i = 1, \dots, n, \quad (5)$$

as described in Carroll et al. (1997) and Yu et al. (2017). Here $\eta(\cdot)$ is an unknown flexible function estimated non-parametrically, and $g^{-1}(\cdot)$ is the link function of the exponential family. The coefficients for the single-index covariates are $\boldsymbol{\beta}_0 = \{\beta_{01}, \dots, \beta_{0p_n}\}$ and the coefficients for the linear covariates are $\boldsymbol{\gamma}_0$ with $\boldsymbol{\gamma}_0 = \{\gamma_{01}, \dots, \gamma_{0q_n}\}$. In this high dimensional space, we assume a sparse set of important covariates. Specifically, we assume there exists a nonzero subset of p_1 coefficients from the total p_n coefficients and a subset of q_1 nonzero coefficients from the total q_n linear coefficients, where the rest of the coefficients are zero. Also, we assume $\|\boldsymbol{\beta}_0\| = 1$ with the first component positive for model identifiability (Yu and Ruppert, 2002).

We estimate the flexible univariate function of the conditional mean $\eta(\cdot)$ non-parametrically using polynomial splines, a global smoothing method more computationally efficient than local approaches (Wang et al., 2011). We first assume the support of the single-index $\mathbf{X}_i \boldsymbol{\beta}_0$ 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 $\mathbf{X}_i \boldsymbol{\beta}$ 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 $k \in \{0, \dots, H'\}$ is determined by the partition, $a = c_0 < c_1 < \dots < c_{H'} < c_{H'+1} = b$.

Given we approximate $\eta(\cdot)$ with a degree $d \geq 2$ polynomial over each interval, a polynomial spline of order d is a function that provides a $d - 1$ degree polynomial or less and global $d - 2$ continuous derivatives on $[a, b]$ (Schumaker, 2007). We consider the nonparametric estimation of the unknown flexible function $\eta(\cdot)$ with the single-index $\mathbf{u}_i = \mathbf{X}_i\boldsymbol{\beta}$ as a linear combination of B-spline bases

$$\eta(\mathbf{u}_i) \approx \mathbf{G}(\mathbf{u}_i)\boldsymbol{\theta} = \mathbf{G}_1(\mathbf{u}_i)\theta_1 + \mathbf{G}_2(\mathbf{u}_i)\theta_2 \cdots + \mathbf{G}_H(\mathbf{u}_i)\theta_H,$$

and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_H)^\top$ are the basis coefficients of size $H \equiv H_n = 1 + d + H'$. Accordingly, since $\eta(\mathbf{X}_i\boldsymbol{\beta})$ is estimated by $\mathbf{G}(\mathbf{X}_i\boldsymbol{\beta})\boldsymbol{\theta}$, the conditional mean $E(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i) := g^{-1}(\eta(\mathbf{X}_i\boldsymbol{\beta}_0) + \mathbf{Z}_i\boldsymbol{\gamma}_0)$ now becomes $g^{-1}(\mathbf{G}(\mathbf{X}_i\boldsymbol{\beta}_0)\boldsymbol{\theta}_0 + \mathbf{Z}_i\boldsymbol{\gamma}_0)$. Therefore, in the rest of this paper, we focus on estimating $\boldsymbol{\alpha}_0 = (\boldsymbol{\theta}_0^\top, \boldsymbol{\beta}_0^\top, \boldsymbol{\gamma}_0^\top)^\top$: the spline, single-index, and the linear coefficients, respectively.

7.1.1 Quadratic Inference Function

In longitudinal studies, ignoring correlation among a participant's repeated observations is well-known to decrease estimation efficiency and produce biased estimates. To account for the correlation within a subject's observations, one may use the generalized estimating equations (GEE) approach to estimate the spline basis, single-index, and linear coefficients, $\boldsymbol{\alpha} = (\boldsymbol{\theta}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ by $\sum_i \mathbf{V}_i(\boldsymbol{\alpha})^\top \mathbf{A}_i^{1/2}(\boldsymbol{\alpha}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\alpha}) (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha})) = \mathbf{0}$ from Zeger and Liang (1986). Here $\boldsymbol{\mu}_i(\boldsymbol{\alpha})$ is the conditional mean. $\mathbf{A}_i(\boldsymbol{\alpha})$ is a diagonal matrix of the variance of $\mathbf{Y}_i$. $\mathbf{V}_i^\top(\boldsymbol{\alpha})$ is defined as

$$\mathbf{V}_i^\top(\boldsymbol{\alpha}_0) = \begin{pmatrix} \mathbf{G}^\top(\mathbf{X}_i\boldsymbol{\beta}_0) \\ \mathbf{J}^\top(\boldsymbol{\beta}_0)\mathbf{X}_i^\top \text{diag}(\dot{\mathbf{G}}(\mathbf{X}_i\boldsymbol{\beta}_0)\boldsymbol{\theta}_0) \\ \mathbf{Z}_i^\top \end{pmatrix},$$

with $\dot{\mathbf{G}}(\cdot)$ as the first derivative of the spline basis. We define the Jacobian matrix to be

$$\mathbf{J}(\boldsymbol{\beta}_0) = \frac{\partial \boldsymbol{\beta}_0}{\partial \boldsymbol{\beta}_0^{(-1)}} = \begin{pmatrix} -\frac{\boldsymbol{\beta}_0^{(-1)}}{(1 - \|\boldsymbol{\beta}_0^{(-1)}\|^2)^{1/2}} \\ I_{(p_n-1) \times (p_n-1)} \end{pmatrix},$$

where we reparameterize $\boldsymbol{\beta}_0$ to be a function of $\boldsymbol{\beta}_0^{(-1)} = (\boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_{p_1})$ using the “delete-one-component” method where p_1 is the number of true important covariates for identifiability as in Yu et al. (2017) and Yu and Ruppert (2002). Moreover, to estimate $\boldsymbol{\alpha}$, the GEE must estimate the nuisance parameters of the working correlation matrix $\mathbf{R}^{-1}$ (Zeger and Liang, 1986). This may be detrimental, especially in high dimensions where there tends to be more parameters inside of the potentially highly nonlinear, unknown function $\eta(\cdot)$ (Qu et al., 2000; Wang et al., 2014).

To avoid estimating any nuisance parameters of the working correlation matrix, we employ a competing method, namely, the quadratic inference function to estimate $\boldsymbol{\alpha} = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$. The quadratic inference function from Qu et al. (2000) replaces the inverse of the working correlation matrix in the GEE with a linear combination of basis matrices, that is, $\mathbf{R}^{-1} \approx a_1 \mathbf{M}_1 + \dots + a_m \mathbf{M}_m$. Here $\mathbf{M}_1, \dots, \mathbf{M}_m$ are predetermined, known symmetric matrices and $\mathbf{a} = (a_1, \dots, a_m)^T$ are constant coefficients. There are available linear combinations of basis matrices to approximate $\mathbf{R}^{-1}$ for most of the common correlation structures as further discussed in Section 7.2.1.

By replacing the inverse of the correlation matrix, $\mathbf{R}^{-1} \approx a_1 \mathbf{M}_1 + \dots + a_m \mathbf{M}_m$, the GEE can be approximated by components of a linear combination of the extended score vector

$$\mathbf{g}_n = \frac{1}{n} \sum_i \mathbf{g}_i(\boldsymbol{\alpha}) = \frac{1}{n} \begin{pmatrix} \sum_i \mathbf{V}_i^T(\boldsymbol{\alpha}) \mathbf{A}_i(\boldsymbol{\alpha})^{1/2} \mathbf{M}_1 \mathbf{A}_i(\boldsymbol{\alpha})^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha})) \\ \sum_i \mathbf{V}_i^T(\boldsymbol{\alpha}) \mathbf{A}_i(\boldsymbol{\alpha})^{1/2} \mathbf{M}_2 \mathbf{A}_i(\boldsymbol{\alpha})^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha})) \\ \dots \\ \sum_i \mathbf{V}_i^T(\boldsymbol{\alpha}) \mathbf{A}_i(\boldsymbol{\alpha})^{1/2} \mathbf{M}_m \mathbf{A}_i(\boldsymbol{\alpha})^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha})) \end{pmatrix}.$$

Then Qu et al. (2000) adopt the generalized method of moments estimator from Hansen (1982) to minimize the following quadratic inference function

$$Q_n(\boldsymbol{\alpha}) = \mathbf{g}_n^T \mathbf{W}^{-1} \mathbf{g}_n, \quad (6)$$

where $\mathbf{W}^{-1} = Cov(\mathbf{g}_n)$ is estimated consistently by $\frac{1}{n} \sum_i \mathbf{g}_i(\boldsymbol{\alpha}) \mathbf{g}_i(\boldsymbol{\alpha})^T$. $Q_n(\boldsymbol{\alpha})$ is known as the quadratic inference function (Qu et al., 2000), and the resulting estimates are determined as $\hat{\boldsymbol{\alpha}} = \arg \min_{\boldsymbol{\alpha}} Q_n(\boldsymbol{\alpha})$. Notably, the quadratic inference function does not need to estimate the coefficients $\mathbf{a} = (a_1, \dots, a_m)$. Moreover, Qu et al. (2000) show the resulting estimators from minimizing this quadratic inference function are the most efficient estimators given the same class of estimating functions, which includes the GEE. This is advantageous since misspecified working correlation structures cause efficiency loss, yet the true correlation structure is not often known in practice.

7.1.2 Penalized Quadratic Inference Function for Ultra-high Dimensional Data

Variable selection becomes an essential task for high and ultra-high dimensional data with a sparse set of important variables, since over selecting variables can negatively impact estimation efficiency and inference for the important variables. Nevertheless, identifying a sparse set of important covariates in high dimensions proves difficult (Fan and Lv, 2008). We adopt the penalized quadratic inference function for the generalized partially linear

single-index model in (5) with diverging and even potentially ultra-high dimensional covariates for concurrent variable selection and estimation. We define our penalized quadratic function to minimize

$$Q_P(\boldsymbol{\alpha}) = Q_n(\boldsymbol{\alpha}) + \sum_{l=1}^{p_n} q_{\lambda_p}(|\beta_l|) + \sum_{k=1}^{q_n} q_{\lambda_q}(|\gamma_k|). \quad (7)$$

Here $Q_n(\boldsymbol{\alpha})$ refers to equation (6), $q_{\lambda_p}(|\beta_l|)$ with $l \in 1, \dots, p_n$ 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 (Fan and Li, 2001). The first derivative of the SCAD penalty function is defined as $q'_\lambda(\zeta) = \lambda\{I(\zeta \leq \lambda) + \frac{(a\lambda - \zeta)_+}{(a-1)\lambda}I(\zeta > \lambda)\}$, for $q_\lambda(0) = 0$ and $a > 2$ for a given regularization parameter λ . As suggested in Fan and Li (2001), we set $a = 3.7$.

7.2 Estimation Algorithm

Estimation of the spline, single-index, and linear coefficients of (7) involves computational challenges in high dimensions, including the non-convex penalty function and the potentially 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. To this end, 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 $\eta(\mathbf{X}_i\boldsymbol{\beta}_0) \cong \eta(\mathbf{X}_i\boldsymbol{\beta}_0) + \text{diag}(\dot{\eta}(\mathbf{X}_i\boldsymbol{\beta}_0))\mathbf{X}_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, where $\dot{\eta}(\cdot)$ is the first derivative of the function $\eta(\cdot)$. 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 (2013) 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 (2013), 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-Raphson algorithm. Specifically, on iteration k of the iterative algorithm, we let $\boldsymbol{\zeta}^{(k)} = (\boldsymbol{\beta}^{(k)\text{T}}, \boldsymbol{\gamma}^{(k)\text{T}})^{\text{T}}$. Then as in Fan and Li (2001), the local quadratic approximation of the SCAD penalty function for iteration k and coefficient $\zeta_j^{(k)}$ is $q_\lambda(|\zeta_j|) \cong q_\lambda(|\zeta_j^{(k)}|) + \frac{1}{2} \frac{q'_\lambda(|\zeta_j^{(k)}|)}{|\zeta_j^{(k)}|} (\zeta_j^2 - \zeta_j^{(k)2})$ with $\zeta_j \approx \zeta_j^{(k)}$ and $\zeta_j^{(k)} \neq 0$. See Cho and Qu (2013) 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 non-

linear optimization to a linear penalized quadratic inference function method from Cho and Qu (2013). 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 7.2.1 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 $\mathbf{g}_n^T \mathbf{W}^{-1} \mathbf{g}_n$. Here $\mathbf{g}_n = \frac{1}{n} \sum_i \mathbf{g}_i(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}^{(k-1)}, \hat{\boldsymbol{\gamma}}^{(k-1)})$ is

$$\frac{1}{n} \left\{ \begin{array}{c} \sum_{i=1}^n \mathbf{G}(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(k-1)})\mathbf{A}_i\left(\boldsymbol{\theta}, \hat{\boldsymbol{\zeta}}^{(k-1)}\right)^{1/2} \mathbf{M}_1\mathbf{A}_i\left(\boldsymbol{\theta}, \hat{\boldsymbol{\zeta}}^{(k-1)}\right)^{-1/2} (\mathbf{Y}_i - g^{-1}(\mathbf{G}(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(k-1)})\boldsymbol{\theta} + \mathbf{Z}_i\hat{\boldsymbol{\gamma}}^{(k-1)})) \\ \sum_{i=1}^n \mathbf{G}(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(k-1)})\mathbf{A}_i\left(\boldsymbol{\theta}, \hat{\boldsymbol{\zeta}}^{(k-1)}\right)^{1/2} \mathbf{M}_2\mathbf{A}_i\left(\boldsymbol{\theta}, \hat{\boldsymbol{\zeta}}^{(k-1)}\right)^{-1/2} (\mathbf{Y}_i - g^{-1}(\mathbf{G}(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(k-1)})\boldsymbol{\theta} + \mathbf{Z}_i\hat{\boldsymbol{\gamma}}^{(k-1)})) \\ \dots \\ \sum_{i=1}^n \mathbf{G}(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(k-1)})\mathbf{A}_i\left(\boldsymbol{\theta}, \hat{\boldsymbol{\zeta}}^{(k-1)}\right)^{1/2} \mathbf{M}_m\mathbf{A}_i\left(\boldsymbol{\theta}, \hat{\boldsymbol{\zeta}}^{(k-1)}\right)^{-1/2} (\mathbf{Y}_i - g^{-1}(\mathbf{G}(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(k-1)})\boldsymbol{\theta} + \mathbf{Z}_i\hat{\boldsymbol{\gamma}}^{(k-1)})) \end{array} \right\}$$

Here $\mathbf{M}_1, \dots, \mathbf{M}_m$ are basis matrices used to approximate a specific working correlation structure as discussed further in Section 7.2.1, and $\mathbf{W}^{-1} = \text{Cov}(\mathbf{g}_n)$ is estimated consistently by $\frac{1}{n} \sum_i \mathbf{g}_i(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}^{(k-1)}, \hat{\boldsymbol{\gamma}}^{(k-1)})\mathbf{g}_i(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}^{(k-1)}, \hat{\boldsymbol{\gamma}}^{(k-1)})^T$.

Step 2: Given the estimated spline coefficients $\hat{\boldsymbol{\theta}}^{(k-1)}$, the k_{th} penalized estimator of

$\hat{\boldsymbol{\zeta}}^{(k)} = (\hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)})^T$ is determined by minimizing the penalized quadratic inference function

$$Q_n(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma}) + \sum_{l=1}^{p_n} q_{\lambda_p}(|\beta_l|) + \sum_{k=1}^{q_n} q_{\lambda_q}(|\gamma_k|), \quad (8)$$

where $Q_n(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbf{g}_n^T \mathbf{W}^{-1} \mathbf{g}_n$ and $\mathbf{g}_n = \frac{1}{n} \sum_i \mathbf{g}_i(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is

$$\frac{1}{n} \left\{ \begin{array}{c} \sum_{i=1}^n \mathbf{U}_i^T \mathbf{A}_i \left(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma} \right)^{1/2} \mathbf{M}_1 \mathbf{A}_i \left(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma} \right)^{-1/2} (\mathbf{Y}_i - g^{-1}(\mathbf{C}_i + \mathbf{X}_i^* \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\gamma})) \\ \sum_{i=1}^n \mathbf{U}_i^T \mathbf{A}_i \left(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma} \right)^{1/2} \mathbf{M}_2 \mathbf{A}_i \left(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma} \right)^{-1/2} (\mathbf{Y}_i - g^{-1}(\mathbf{C}_i + \mathbf{X}_i^* \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\gamma})) \\ \dots \\ \sum_{i=1}^n \mathbf{U}_i^T \mathbf{A}_i \left(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma} \right)^{1/2} \mathbf{M}_m \mathbf{A}_i \left(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma} \right)^{-1/2} (\mathbf{Y}_i - g^{-1}(\mathbf{C}_i + \mathbf{X}_i^* \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\gamma})) \end{array} \right\}.$$

For brevity, we let $\mathbf{X}_i^* = \text{diag}(\dot{\mathbf{G}}(\mathbf{X}_i \hat{\boldsymbol{\beta}}) \hat{\boldsymbol{\theta}}) \mathbf{X}_i$, and $\mathbf{C}_i = \mathbf{G}(\mathbf{X}_i \hat{\boldsymbol{\beta}}) \hat{\boldsymbol{\theta}} - \mathbf{X}_i^* \hat{\boldsymbol{\beta}}$. These are derived from the linear approximation to $\eta(\cdot)$, since the conditional mean with this approximation, $g^{-1}(\mathbf{G}(\mathbf{X}_i \hat{\boldsymbol{\beta}}) \hat{\boldsymbol{\theta}} + \text{diag}(\dot{\mathbf{G}}(\mathbf{X}_i \hat{\boldsymbol{\beta}}) \hat{\boldsymbol{\theta}}) \mathbf{X}_i (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathbf{Z}_i \boldsymbol{\gamma})$, can be rearranged into $g^{-1}(\mathbf{C}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) + \mathbf{X}_i^* \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\gamma})$. We also assume the identifiability constraints $\|\boldsymbol{\beta}\| = 1$ and $\beta_1 > 0$. $\mathbf{U}_i^T = \mathbf{U}_i^T(\hat{\boldsymbol{\beta}}^{(k-1)}, \hat{\boldsymbol{\theta}}^{(k-1)})$ is similar to $\mathbf{V}_i^T(\boldsymbol{\alpha})$ except now the spline basis coefficients are determined in the previous step. $\mathbf{W}^{-1} = \text{Cov}(\mathbf{g}_n)$ is estimated by $\frac{1}{n} \sum_i \mathbf{g}_i(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \mathbf{g}_i(\hat{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\beta}, \boldsymbol{\gamma})^T$. As above, $\mathbf{M}_1, \dots, \mathbf{M}_m$ are basis matrices used to approximate a specific working correlation structure as discussed further in Section 7.2.1.

Using the approach from Cho and Qu (2013), we estimate the single-index and partially linear parameters $\boldsymbol{\beta}, \boldsymbol{\gamma}$ by taking a quadratic approximation of (8) and invoking the Newton-Rhaphson algorithm. See Cho and Qu (2013) 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.

In the next subsection, we provide recommendations for implementing this algorithm

such as choosing tuning parameters for the penalty function, selecting basis matrices for quadratic inference function, and obtaining initial values for step 0.

7.2.1 Practical Implementation

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 with ultra-high dimensional covariates.

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 (6) and (7). As noted in Qu et al. (2000), 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 (m, m) , which are 1s. $a_1 = (1 + \rho^2)/l_2$, $a_2 = -\rho/l_2$, 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 (2012) 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 (Qu et al., 2000; Qu and Li, 2006; Zhou and Qu, 2012).

Initial Values and Screening

In ultra-high dimensions, the computational time of most approaches tends to be prohibitive for practical usage (Fan and Lv, 2008). For this reason, screening is commonly used in relevant literature to rapidly reduce the covariate dimension in practice (Brdic et al., 2011; Cai et al., 2015; Hui et al., 2015; Fang et al., 2015). We echo Fan and Lv (2008) 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. (2011) uses linear GEE initials. One may also use sure Independence screening estimates at the original dimension as initial values (Fan and Lv, 2008). 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. (2009). The model for selection is

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

where h_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. (2009). 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. (2014), 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 (He et al., 2002). In Ruppert and Carroll (2000) and Yu and Ruppert (2002), 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. (2010) 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 (Ma et al., 2014; Yu et al., 2017). In this study, we use quadratic splines for our simulation analysis and for the CCI-779 study.

7.3 Numerical Illustrations

In this section, we demonstrate the variable selection performance of our generalized partially linear single-index model using the penalized quadratic inference function. We evaluate the performance of each example over 100 simulation replications. 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.

Example 3.

We consider the correlated binary responses from the marginal mean

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

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 $\boldsymbol{\beta}_0 = (1, 1, 1, 1, 1, 0, \dots, 0)^T / \sqrt{5}$ and $\boldsymbol{\gamma}_0 = (1, 1, 1, 0, \dots, 0)$, 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 mean 0 and variance 1, and the covariates $\mathbf{Z}_{it}$ are sampled from an independent standard normal distribution. We use the method described in Macke et al. (2009) 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.

In Table 13, we report the variable selection results. This table indicates that our approach selects the true important variables with a high percentage of simulation replications. On average, the model also correctly accounts for the number of true negatives

Table 13: Summary of Results for the Single-Index Model in Example 3 with Binary Response. The total covariates are $p_n + q_n = 500$ and $p_n + q_n = 5000$ with $p_n = q_n$ calculated over 100 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.

	Single-index Covariates			Linear Covariates		
	Correct%	TN	FN	Correct%	TN	FN
$p_n + q_n = 500$						
Independence	91	244.95	0	1	247	0
AR(1)	93	244.92	0	1	247	0
Exchangeable	98	244.98	0	1	247	0
$p_n + q_n = 5000$						
Independence	84	2494.88	0.05	1	2497	0
AR(1)	89	2494.93	0.04	1	2497	0
Exchangeable	95	2494.98	0.03	99	2497	0.01

for all dimensions of covariates. In particular, when $p_n + q_n = 500$, the number of true negatives are very close to the true amount of 245 for the single-index covariates, and the number of true negatives are 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 are close to the true amount of 2495 for the single-index case, and the number of true negatives are also close to the true amount of 2497 for the linear covariates. Moreover, the number of true important single-index or linear variable 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. In this example, the model selection performance with various correlation structures is quite similar. However, it is better in almost all cases under the true exchangeable correlation structure.

7.4 *An Application to an Advanced Renal Cancer Study*

In this section, we use our generalized single-index model using the penalized quadratic inference function to analyze a longitudinal pharmacokinetics study of CCI-779 from Boni et al. (2005). CCI-779 is a known renal cancer inhibitor and tumor suppressor, and CCI-779's pharmacogenomic profile may be predicted by gene transcripts in blood mononuclear cells. One aim of this study is to determine which specific gene transcripts are associated with the concentration of CCI-779, potentially leading to more effective treatment for renal cancer patients. The outcome of the data, $\mathbf{Y}_i$, is the CCI-779 cumulative AUC of 33 patients measured at baseline, 8, and 16 weeks following the start of CCI-779 treatment. Also, the data is unbalanced, and there are 54 observations total. The covariates, $\mathbf{X}_{it}$, consist of 12,626 gene transcripts in blood mononuclear cells that are potentially associated with the concentration of CCI-779.

For such a high-dimensional problem, we use sure independence screening from Fan and Lv (2008) to reduce dimensionality initially. We then use our longitudinal single-index model and the penalized quadratic inference function to explore the association between CCI-779 cumulative AUC and the covariates using the model $\mathbf{Y}_i = \eta(\sum_t \beta_t \mathbf{X}_{it}) + \epsilon_i$. The response, CCI-779 cumulative AUC, is continuous, and we use the Gaussian identity link function. Here $\mathbf{X}_{i0}$ is time and the remaining covariates are gene transcripts. We compare the performance of our longitudinal single-index penalized quadratic inference function approach to a longitudinal linear penalized quadratic inference function method $\mathbf{Y}_i = \sum_t \beta_t \mathbf{X}_{it} + \epsilon_i$.

For the single-index model, we choose the tuning parameter λ for the SCAD penalty for the single-index model and the linear model via grid search, where the best λ corresponds to the lowest HBIC value. We use a quadratic spline basis with 1 knot to estimate the flexible univariate function $\eta(\cdot)$. The average prediction error is computed using 5-fold

cross validation performance from our method fitted on the full data.

Table 14 provides the in-sample mean squared error and the out-of-sample mean prediction error over 5 folds using our single-index model with the penalized quadratic inference function approach and the linear penalized quadratic inference function approach. We find that the single-index model has lower in-sample mean squared error and out-of-sample mean prediction error than the linear penalized quadratic inference function. This is probably because a linear model cannot capture complexity and synergy among covariates as the single-index model can. We show results from the independence case only as the performance of other working correlation structures is similar or worse because of challenges from such a small sample size for both our single-index model with the penalized quadratic inference function and the linear penalized quadratic inference function method.

Table 14: Summary of Results for CCI-779 Data. “MSE” is the model mean squared error, and “PE” is the mean prediction error using 5-fold cross validation. “Single-index Model” refers to our proposed single-index model using the penalized quadratic inference function, and “Linear Model” refers to the linear model using the penalized quadratic inference function.

	MSE	PE
Single-index Model	0.0183	0.0240
Linear Model	0.0661	0.1379

7.5 Discussion

Motivated by growing areas of high dimensional longitudinal data analysis with a correlated discrete response, we adopt a flexible semiparametric approach for simultaneous variable selection and estimation without the requirement of specifying the full likelihood. Moreover, using the penalized quadratic inference function, the resulting regression estimates pursue estimation efficiency and variable selection accuracy. A key challenge is that we allow the number of single-index covariates in the nonparametric term to diverge and to even be in ultra-high dimension. Especially for longitudinal data, few existing approaches incorporate diverging or ultra-high dimensional covariates with nonlinearity for potential discrete responses. Future work includes establishing new asymptotic theory for our generalized partially linear single-index model using the penalized quadratic inference function for diverging and even ultra-dimensional covariates allowing discrete responses in longitudinal data.

8 Technical Proofs

In this section, we provide detailed definitions for our generalized partially linear single-index longitudinal model using penalized generalized estimating equations and establish necessary definitions and assumptions. We then present assumptions and technical proofs for all theorems and Lemma 1. We prove Theorem 1 that establishes the convergence rate for the generalized partially linear single-index models assuming known but potentially diverging number of important variables of longitudinal data via GEE. We then prove Theorem 2 of the asymptotic normality of the oracle estimators. Further, we prove Theorem 3 that establishes the large sample properties for the penalized GEE estimators under ultra-high dimension for both single-index covariates in the nonparametric component and the partially linear component.

8.1 Definitions

First we provide a more detailed explanation of the approach we use to estimate the unknown univariate function, $h(\cdot)$. We use polynomial splines to estimate $h(\cdot)$ defined on an interval with potentially diverging support of $\mathbf{X}\boldsymbol{\delta}_0$. The support $[a, b]$ will be split into intervals based on $\tilde{K}$ interior knots, $(v_1, \dots, v_{\tilde{K}})$. The complete set of knots will be at $a = v_0 < v_1 < \dots < v_{\tilde{K}-1} < v_{\tilde{K}} < v_{\tilde{K}+1} = b$. Here $\tilde{K} \equiv \tilde{K}_n$ increases with the number of clusters n as discussed later for large sample properties. Polynomial splines of degree s guarantee that the approximated function on the intervals between knots are polynomials of order $s - 1$, and the approximated function will attain continuous differentiability up to $s - 2$ times. This holds if the approximated function is of order $s \geq 2$ (Schumaker, 2007). We use the B-spline basis, $\mathbf{B}(u)\boldsymbol{\psi}$, to estimate the unknown function.

Next, we introduce definitions for the assumptions and technical proofs. $\hat{\boldsymbol{\theta}}_{0(1)}$ denotes the oracle estimator of the true parameter $\boldsymbol{\theta}_{0(1)} = (\boldsymbol{\psi}_{0(1)}^T, \boldsymbol{\delta}_{0(1)}^T, \boldsymbol{\beta}_{0(1)}^T)^T$. $\boldsymbol{\theta}_{0(1)} =$

$(\boldsymbol{\psi}_{0(1)}^T, \boldsymbol{\delta}_{0(1)}^T, \boldsymbol{\beta}_{0(1)}^T)^T$ where $\boldsymbol{\psi}_{0(1)}$ are the spline coefficients corresponding to the true single-index $\mathbf{X}_{(1)i}\boldsymbol{\delta}_{0(1)}$. Here $\boldsymbol{\delta}_{0(1)}$ are the true single-index coefficients, and $\boldsymbol{\beta}_{0(1)}$ are the true linear coefficients. For notational simplicity, we drop the subscript n . p_{1n} and q_{1n} are the number of important nonzero covariates for the true important single-index covariates, $\mathbf{X}_{(1)i}$, and partially-linear covariates, $\mathbf{Z}_{(1)i}$, respectively.

We follow the classical linear GEE set-up (Zeger and Liang, 1986) and assume that the marginal density of Y_{ij} is from a canonical exponential family. We can write the marginal variance per subject and time point as $\sigma_{ij}^2 = \phi \dot{\mu}(\eta_{ij})$, where $\dot{\mu}$ denotes the first derivative of μ as a function of η_{ij} ; and ϕ is a scaling constant for dispersion (also see Wang et al. (2012)). Commonly a simple dispersion parameter $\phi = 1$ may be adopted, then the marginal variance becomes $\sigma_{ij}^2 = \dot{\mu}(\eta_{ij})$. If overdispersion or underdispersion is suspected to exist, a dispersion parameter can be added in this marginal variance function. Then the spline approximated version is $\sigma_{ij}^2(\boldsymbol{\theta}_{0(1)}) = \dot{\mu}(\eta_{ij}(\boldsymbol{\theta}_{0(1)}))$. Let $\mathbf{A}_i(\boldsymbol{\theta}_{0(1)}) = \text{diag}(\sigma_{i1}^2(\boldsymbol{\theta}_{0(1)}), \dots, \sigma_{im_i}^2(\boldsymbol{\theta}_{0(1)}))$ be the spline-approximated marginal variances of $\mathbf{Y}_i$; and $\boldsymbol{\mu}_i(\boldsymbol{\eta}_i(\boldsymbol{\theta}_{0(1)}))$ be the spline-approximated marginal mean with $\boldsymbol{\eta}_i(\boldsymbol{\theta}_{0(1)}) = \mathbf{B}(\mathbf{X}_{(1)i}\boldsymbol{\delta}_0) + \mathbf{Z}_{(1)i}\boldsymbol{\delta}_{0(1)}$. To handle the identifiability constraint for the single-index parameter $\boldsymbol{\delta}_{(1)}$, we treat $\boldsymbol{\delta}_{(1)}$ as a function of $\boldsymbol{\delta}_{(1)}^{(-1)} = (\delta_2, \dots, \delta_{p_1})^T$ where $\boldsymbol{\delta}_{(1)}^{(-1)}$ is the $p_1 - 1$ dimensional reparameterized single-index parameter. The Jacobian matrix is

$$\mathbf{J}(\boldsymbol{\delta}_{(1)}) = \frac{\partial \boldsymbol{\delta}_{(1)}}{\partial \boldsymbol{\delta}_{(1)}^{(-1)}} = \begin{pmatrix} -\frac{\boldsymbol{\delta}_{(1)}^{(-1)}}{(1 - \|\boldsymbol{\delta}_{(1)}^{(-1)}\|^2)^{1/2}} \\ \mathbf{I}_{(p_1-1) \times (p_1-1)} \end{pmatrix}.$$

We denote the true values of $\boldsymbol{\delta}_{(1)}$ and $\boldsymbol{\beta}_{(1)}$ by $\boldsymbol{\delta}_{0(1)}$ and $\boldsymbol{\beta}_{0(1)}$, respectively, and we denote $\mathbf{J}(\boldsymbol{\delta}_{0(1)})$ by $\mathbf{J}_{0(1)}$.

8.2 Assumptions

Given the previous definitions, we examine the convergence rate and asymptotic normality of the oracle estimators for the generalized estimating equations model with both diverging p_1 for the nonparametric component and diverging q_1 for the partially linear component. We use $\dot{f}, \ddot{f}, \dddot{f}$ to denote derivatives of functions from first order to third order, while for general d -th order derivative we will also use superscript (d) to denote it. We assume that $m_i \equiv m$ to simplify the proof. In practice unequal m_i can be easily dealt with. For two m -dimensional vectors $\mathbf{a}, \mathbf{b}$, $\mathbf{a} \odot \mathbf{b}$ is also an m -dimensional vector with componentwise product (Hadamard product). For a matrix $\mathbf{A}$ with m rows and an m -vector $\mathbf{b}$, $\mathbf{A} \odot \mathbf{b}$ is a matrix of the same size as $\mathbf{A}$ resulting from applying Hadamard product of $\mathbf{b}$ with each column of $\mathbf{A}$.

Under assumption (A2) below that h is d -smooth, there is a vector $\boldsymbol{\psi}_0$ with $\|\mathbf{B}(\cdot)\boldsymbol{\psi}_0 - h(\cdot)\|_\infty \leq CK^{-d}$ (De Boor, 2001). Also define $\boldsymbol{\eta}_{0(1)i} = h(\mathbf{X}_{i(1)}\boldsymbol{\delta}_{0(1)}) + \mathbf{Z}_{i(1)}\boldsymbol{\beta}_{0(1)}$ and the true conditional mean of the response is then $\boldsymbol{\mu}(\boldsymbol{\eta}_{0(1)i})$.

We rely on the following assumptions that are consistent with the literature (Lian et al., 2014).

$$(A1) \quad \sup_{1 \leq i \leq n, 1 \leq j \leq m} \|\mathbf{X}_{(1)ij}\| = O_p(\sqrt{p_1}), \sup_{1 \leq i \leq n, 1 \leq j \leq m} \|\mathbf{Z}_{(1)ij}\| = O_p(\sqrt{q_1}).$$

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

$$(A3) \quad E[\|\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_{0(1)i})\|^{2+\delta}] < \infty \text{ for some } \delta > 0.$$

$$(A4) \quad \text{There exists positive constants } c_1 \text{ and } c_2 \text{ such that } c_1 \leq \lambda_{\min}(n^{-1} \sum_i \mathbf{U}_{0(1)i}^T \mathbf{U}_{0(1)i}) \leq \lambda_{\max}(n^{-1} \sum_i \mathbf{U}_{0(1)i}^T \mathbf{U}_{0(1)i}) \leq c_2, \text{ where } \lambda_{\min} \text{ and } \lambda_{\max} \text{ denotes the smallest and largest}$$

eigenvalue of a matrix, respectively.

(A5) Let $C = \{\boldsymbol{\theta}_{(1)} : \|\boldsymbol{\theta}_{(1)} - \boldsymbol{\theta}_{0(1)}\| \leq Cr_n\}$, where C is a positive constant. Then $\dot{\mu}(\eta_{ij}(\boldsymbol{\theta}_{0(1)}))$, $1 \leq i \leq n, 1 \leq j \leq m$, are uniformly bounded away from 0 and ∞ on C ; $\ddot{\mu}(\eta_{ij}(\boldsymbol{\theta}_{0(1)}))$ and $\ddot{\mu}(\eta_{ij}(\boldsymbol{\theta}_{0(1)}))$, $1 \leq i \leq n, 1 \leq j \leq m$, are uniformly bounded on C .

Define $\boldsymbol{\xi}_{0(1)} = (\boldsymbol{\delta}_{0(1)}^{(-1)\text{T}}, \boldsymbol{\beta}_{0(1)}^{\text{T}})^{\text{T}}$ and $\widehat{\boldsymbol{\xi}}_{(1)} = (\widehat{\boldsymbol{\delta}}_{(1)}^{(-1)\text{T}}, \widehat{\boldsymbol{\beta}}_{(1)}^{\text{T}})^{\text{T}}$. Further, we define the following projection. Let $\mathcal{M}_j = \{g : Eg^2(\mathbf{X}_{ij}^{\text{T}}\boldsymbol{\delta}_0) < \infty\}$. For any random variable $\boldsymbol{\Omega} \in \mathbf{R}^m$ which is a function of $(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\epsilon}_i)$, we define $E_{\mathcal{M}}[\boldsymbol{\Omega}] = (g_1(\mathbf{X}_{i1}^{\text{T}}\boldsymbol{\delta}_0), \dots, g_m(\mathbf{X}_{im}^{\text{T}}\boldsymbol{\delta}_0))^{\text{T}}$ where $g_1, \dots, g_m$ is the minimizer of

$$E[(\boldsymbol{\Omega} - \mathbf{g})^{\text{T}} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} (\boldsymbol{\Omega} - \mathbf{g})]$$

over $g_1 \in \mathcal{M}_1, \dots, g_m \in \mathcal{M}_m$ where $\mathbf{g} = (g_1(\mathbf{X}_{ij}^{\text{T}}\boldsymbol{\delta}_0), \dots, g_m(\mathbf{X}_{ij}^{\text{T}}\boldsymbol{\delta}_0))^{\text{T}}$. This definition can be extended to the case when $\boldsymbol{\Omega}$ is a random matrix with m rows such that the projection is obtained column by column.

Using the defined projection, we write $E_{\mathcal{M}}[\mathbf{Z}_{(1)i}] = (f_1(\mathbf{X}_{(1)i}\boldsymbol{\delta}_{0(1)}), \dots, f_q(\mathbf{X}_{(1)i}\boldsymbol{\delta}_{0(1)}))$ with $f_j(\mathbf{X}_{(1)i}\boldsymbol{\delta}_{0(1)}) = (f_{j1}(\mathbf{X}_{i1}^{\text{T}}\boldsymbol{\delta}_0), \dots, f_{jm}(\mathbf{X}_{im}^{\text{T}}\boldsymbol{\delta}_0))^{\text{T}}$, and write

$$E_{\mathcal{M}}[\text{diag}(h(\mathbf{X}_i\boldsymbol{\delta}_0))\mathbf{X}_i] = (g_1(\mathbf{X}_i\boldsymbol{\delta}_0), \dots, g_p(\mathbf{X}_i\boldsymbol{\delta}_0))$$

with $g_j(\mathbf{X}_i\boldsymbol{\delta}_0) = (g_{j1}(\mathbf{X}_{i1}^{\text{T}}\boldsymbol{\delta}_0), \dots, g_{jm}(\mathbf{X}_{im}^{\text{T}}\boldsymbol{\delta}_0))^{\text{T}}$. Define

$$\mathbf{W}_{0(1)i} = \begin{pmatrix} \mathbf{Z}_{(1)i}^{\text{T}} - E_{\mathcal{M}}[\mathbf{Z}_{(1)i}^{\text{T}}] \\ \mathbf{J}^{\text{T}}(\boldsymbol{\delta}_{0(1)})\mathbf{X}_{(1)i}^{\text{T}}\text{diag}(\dot{h}(\mathbf{X}_{(1)i}\boldsymbol{\delta}_{0(1)})) - E_{\mathcal{M}}[\mathbf{J}^{\text{T}}(\boldsymbol{\delta}_{0(1)})\mathbf{X}_{(1)i}^{\text{T}}\text{diag}(\dot{h}(\mathbf{X}_{(1)i}\boldsymbol{\delta}_{0(1)}))] \end{pmatrix}.$$

(A6) $f_{jl}, g_{jl} \in \mathcal{H}^{d'}$ for some $d' > 1$.

Further for Theorems 2 and 3, we define $\mathcal{C}_{n(1)} = \frac{1}{n} \sum_i \mathbf{W}_{0(1)i} \mathbf{A}_{0(1)i}^{1/2} \mathbf{R}^{-1} \mathbf{R}_0 \mathbf{R}^{-1} \mathbf{A}_{0(1)i}^{1/2} \mathbf{W}_{0(1)i}$, $\mathcal{D}_{n(1)} = \frac{1}{n} \sum_i \mathbf{W}_{0(1)i} \mathbf{A}_{0(1)i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0(1)i}^{1/2} \mathbf{W}_{0(1)i}$, where $\mathbf{R}_0$ is the true correlation matrix.

8.3 Proof of Theorem 1

We define

$$\bar{\mathbf{S}}_0(\boldsymbol{\theta}) = \sum_i \mathbf{U}_i(\boldsymbol{\theta}_{0(1)})^T \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_{0(1)}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_{0(1)}) (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_{(1)}))),$$

and

$$\mathbf{S}_0(\boldsymbol{\theta}) = \sum_i \mathbf{U}_{0(1)i}^T \mathbf{A}_{0(1)i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0(1)i}^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_{(1)}))),$$

where $\mathbf{A}_{0(1)i} = \text{diag}(\dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0(1)i}))$ and

$$\mathbf{U}_{0(1)i}^T = \begin{pmatrix} \mathbf{B}^T(\mathbf{X}_{(1)i} \boldsymbol{\delta}_{0(1)}) \\ \mathbf{J}_{0(1)}^T \mathbf{X}_{(1)i}^T \text{diag}(\dot{h}(\mathbf{X}_{(1)i} \boldsymbol{\delta}_{0(1)})) \\ \mathbf{Z}_{(1)i}^T \end{pmatrix}.$$

Note the difference between $\bar{\mathbf{S}}_0(\boldsymbol{\theta})$ and $\mathbf{S}_0(\boldsymbol{\theta})$ is that we use spline approximation for the true function h in several places for the former.

In proving Theorem 1, we rely on Lemma 1 below.

Lemma 8.1

$$\begin{aligned} \sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq Cr_n} \|\mathbf{S}(\boldsymbol{\theta}) - \bar{\mathbf{S}}_0(\boldsymbol{\theta})\| &= O_p(\sqrt{n}(\sqrt{K^3 p_1 + p_1 + q_1} r_n \\ &\quad + n r_n^2 \sqrt{K + p_1 + q_1} (\sqrt{K^3 p_1 + p_1 + q_1}))) \end{aligned}$$

and

$$\begin{aligned} \sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq Cr_n} \|\bar{\mathbf{S}}_0(\boldsymbol{\theta}) - \mathbf{S}_0(\boldsymbol{\theta})\| &= O_p(\sqrt{n}K^{-d}(K\sqrt{p_1} + \sqrt{q_1}) \\ &\quad + nK^{-d}(K\sqrt{p_1} + \sqrt{q_1})\sqrt{K + p_1 + q_1}r_n). \end{aligned}$$

For notational simplicity, subscript (1) is suppressed when referring to important covariates, corresponding true parameters, and oracle estimators for all proofs for Lemma 1, Theorems 1 and 2.

Let $\boldsymbol{\theta}^{(-1)} = (\boldsymbol{\psi}^\top, \boldsymbol{\delta}^{(-1)\top}, \boldsymbol{\beta}^\top)^\top$. As in Wang (2011), we only need to show that for any $a > 0$, there exists $L > 0$ such that for all n sufficiently large,

$$P\left(\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq Lr_n} (\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)})^\top \mathbf{S}(\boldsymbol{\theta}) < 0\right) \geq 1 - a. \quad (9)$$

First, we have, by Lemma 8.1,

$$(\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)})^\top (\mathbf{S}(\boldsymbol{\theta}) - \mathbf{S}_0(\boldsymbol{\theta})) = O_p(\sqrt{n}(\sqrt{K^3 p_1} + p_1 + q_1)r_n^2) = o_p(nr_n^2).$$

Second, using that

$$\begin{aligned} &\boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta})) - \boldsymbol{\mu}(\boldsymbol{\eta}_{0i}) \\ &= \text{diag}(\dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i}))(\boldsymbol{\eta}_i(\boldsymbol{\theta}) - \boldsymbol{\eta}_{0i}) + O_p((K + p_1 + q_1)r_n^2) \\ &= \text{diag}(\dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i}))\mathbf{U}_{0i}(\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)}) + \text{diag}(\dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i}))(\mathbf{U}_i(\boldsymbol{\theta}^*) - \mathbf{U}_{0i})(\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)}) \\ &\quad + \text{diag}(\dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i}))(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_{0i}) + O_p((K + p_1 + q_1)r_n^2) \\ &= \text{diag}(\dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i}))\mathbf{U}_{0i}(\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)}) + \text{diag}(\dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i}))(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_{0i}) \\ &\quad + O_p((\sqrt{K^3 p_1} + p_1)r_n^2) + O_p((K + p_1 + q_1)r_n^2), \end{aligned}$$

where $\boldsymbol{\theta}^*$ lies between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$, we have

$$\begin{aligned}
& (\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)})^T \mathbf{S}_0(\boldsymbol{\theta}) \\
&= (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \sum_i \mathbf{U}_{0i}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta}))) \\
&= (\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)})^T \sum_i \mathbf{U}_{0i}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} \boldsymbol{\epsilon}_i \\
&\quad - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \sum_i \mathbf{U}_{0i}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} (\boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta})) - \boldsymbol{\mu}(\boldsymbol{\eta}_{0i})) \\
&= (\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)})^T \sum_i \mathbf{U}_{0i}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} \boldsymbol{\epsilon}_i \\
&\quad - (\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)})^T \sum_i \mathbf{U}_{0i}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} \mathbf{U}_{0i} (\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)}) \\
&\quad - (\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)})^T \sum_i \mathbf{U}_{0i}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} (\boldsymbol{\eta}_i(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_{0i}) \\
&\quad + O_p(n(\sqrt{K^3 p_1 + p_1}) r_n^2) + O_p(n(K + p_1 + q_1) r_n^2).
\end{aligned}$$

The sum in the first term above has mean zero and thus the first term is $O_p(\sqrt{n}(Lr_n)\sqrt{K + p_1 + q_1})$.

The second term is bounded below by $Cn(Lr_n)^2$ with probability approaching one by assumption (A4). The third term can be bounded as

$$\begin{aligned}
& \left| (\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)})^T \sum_i \mathbf{U}_{0i}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} (\boldsymbol{\eta}_i(\boldsymbol{\theta}_0) - \boldsymbol{\eta}_{0i}) \right| \\
&\leq |(\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)})^T \sum_i \mathbf{U}_{0i}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} \mathbf{U}_{0i} (\boldsymbol{\theta}^{(-1)} - \boldsymbol{\theta}_0^{(-1)})|^{1/2} \\
&\quad \times \left| (\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 - h(\mathbf{X}_i \boldsymbol{\delta}_0))^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} (\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 - h(\mathbf{X}_i \boldsymbol{\delta}_0)) \right|^{1/2} = O_p(nr_n K^{-d}),
\end{aligned}$$

which is dominated by the second term when L is large enough. This proves (9).

Proof of Lemma 1. The proofs involve messy calculations to an extent that is quite challenging to manage. Our strategy is to organize the expressions below in various decompositions into a sum of first order terms that is dominant and other terms that are of

smaller order whose exact expressions are no longer important. For example, we have

$$\begin{aligned}
& \eta_{ij}(\boldsymbol{\theta}) - \eta_{ij}(\boldsymbol{\theta}_0) \\
&= \mathbf{B}^T(\mathbf{X}_{ij}^T \boldsymbol{\delta}) \boldsymbol{\psi} - \mathbf{B}^T(\mathbf{X}_{ij}^T \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 + \mathbf{Z}_{ij}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
&= \mathbf{B}^T(\mathbf{X}_{ij}^T \boldsymbol{\delta}_0)(\boldsymbol{\psi} - \boldsymbol{\psi}_0) + \dot{\mathbf{B}}^T(\mathbf{X}_{ij}^T \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 \mathbf{X}_{ij}^T(\boldsymbol{\delta} - \boldsymbol{\delta}_0) + \mathbf{Z}_{ij}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
&\quad + \dot{\mathbf{B}}^T(\mathbf{X}_{ij}^T \boldsymbol{\delta}^*)(\boldsymbol{\psi} - \boldsymbol{\psi}_0) \mathbf{X}_{ij}^T(\boldsymbol{\delta} - \boldsymbol{\delta}_0) + \ddot{\mathbf{B}}^T(\mathbf{X}_{ij}^T \boldsymbol{\delta}^*) \boldsymbol{\psi}_0 (\mathbf{X}_{ij}^T(\boldsymbol{\delta} - \boldsymbol{\delta}_0))^2 \\
&= (\mathbf{B}^T(\mathbf{X}_{ij}^T \boldsymbol{\delta}_0), \dot{\mathbf{B}}^T(\mathbf{X}_{ij}^T \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 \mathbf{X}_{ij}^T, \mathbf{Z}_{ij}^T)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + O_p(\sqrt{K^3 p_1 + p_1} r_n^2),
\end{aligned}$$

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

$$\boldsymbol{\eta}_i(\boldsymbol{\theta}) - \boldsymbol{\eta}_i(\boldsymbol{\theta}_0) = (\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 \mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + O_p(\sqrt{K^3 p_1 + p_1} r_n^2),$$

and $\|\boldsymbol{\eta}_i(\boldsymbol{\theta}) - \boldsymbol{\eta}_i(\boldsymbol{\theta}_0)\| = O_p(\sqrt{K + p_1 + q_1} r_n)$.

Similarly, we have by Taylor's expansion

$$\begin{aligned}
& \mathbf{A}_i^{1/2}(\boldsymbol{\theta}) - \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0) \\
&= 0.5 \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \text{diag}(\boldsymbol{\eta}_i(\boldsymbol{\theta}) - \boldsymbol{\eta}_i(\boldsymbol{\theta}_0)) + O_p(\|\boldsymbol{\eta}_i(\boldsymbol{\theta}) - \boldsymbol{\eta}_i(\boldsymbol{\theta}_0)\|^2) \\
&= 0.5 \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \text{diag}\{(\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 \mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} \\
&\quad + O_p((\sqrt{K^3 p_1 + p_1} r_n^2 + (K + p_1 + q_1) r_n^2),
\end{aligned}$$

$$\begin{aligned}
& \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}) - \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \\
= & -0.5\mathbf{A}_i^{-3/2}(\boldsymbol{\theta}_0)\text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0)))\text{diag}(\boldsymbol{\eta}_i(\boldsymbol{\theta}) - \boldsymbol{\eta}_i(\boldsymbol{\theta}_0)) + O_p(\|\boldsymbol{\eta}_i(\boldsymbol{\theta}) - \boldsymbol{\eta}_i(\boldsymbol{\theta}_0)\|^2) \\
= & -0.5\mathbf{A}_i^{-3/2}(\boldsymbol{\theta}_0)\text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0)))\text{diag}\{(\mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\psi_0\mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} \\
& + O_p((\sqrt{K^3p_1 + p_1})r_n^2 + (K + p_1 + q_1)r_n^2),
\end{aligned}$$

$$\begin{aligned}
& \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}) - \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0) \\
= & \dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0) \odot (\mathbf{X}_i(\boldsymbol{\delta} - \boldsymbol{\delta}_0)) + O_p(K^{5/2}p_1r_n^2),
\end{aligned}$$

where $\odot$ denotes the componentwise (Hadamard) product and

$$\begin{aligned}
\mathbf{A}_i^{1/2}(\boldsymbol{\theta}) - \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0) &= O_p(\sqrt{K + p_1 + q_1}r_n), \\
\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}) - \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) &= O_p(\sqrt{K + p_1 + q_1}r_n), \\
\mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}) - \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0) &= O_p(K^{3/2}p_1^{1/2}r_n).
\end{aligned}$$

Using the identity

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

we have

$$\begin{aligned}
& \mathbf{B}^T(\mathbf{X}_i\boldsymbol{\delta})\mathbf{A}_i^{1/2}(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}) - \mathbf{B}^T(\mathbf{X}_i\boldsymbol{\delta}_0)\mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0)\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \\
= & \dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0) \odot (\mathbf{X}_i(\boldsymbol{\delta} - \boldsymbol{\delta}_0))\mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0)\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \\
& + 0.5\mathbf{B}^T(\mathbf{X}_i\boldsymbol{\delta}_0)\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0)\text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0)))\text{diag}\{(\mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0\mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} \\
& \mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) - 0.5\mathbf{B}^T(\mathbf{X}_i\boldsymbol{\delta}_0)\mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0)\mathbf{R}^{-1}\mathbf{A}_i^{-3/2}(\boldsymbol{\theta}_0)\text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \\
& \text{diag}\{(\mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0\mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} \\
& + O_p(\sqrt{K^3p_1(K+p_1+q_1)}r_n^2 + \sqrt{K}(K+p_1+q_1)r_n^2),
\end{aligned}$$

and

$$\begin{aligned}
& \mathbf{B}^T(\mathbf{X}_i\boldsymbol{\delta})\mathbf{A}_i^{1/2}(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}) - \mathbf{B}^T(\mathbf{X}_i\boldsymbol{\delta}_0)\mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0)\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \\
= & O_p(\sqrt{K^3p_1}r_n + \sqrt{K(K+p_1+q_1)}r_n).
\end{aligned}$$

Using $\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta})) = \boldsymbol{\epsilon}_i + (\boldsymbol{\mu}(\boldsymbol{\eta}_{0i}) - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta})))$ and $\boldsymbol{\mu}(\boldsymbol{\eta}_{0i}) - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta})) = O_p(\sqrt{K+p_1+q_1}r_n)$,

we obtain

$$\begin{aligned}
& \sum_i (\mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta}) \mathbf{A}_i^{1/2}(\boldsymbol{\theta}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}) - \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta}_0) \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0)) (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta}))) \\
= & \sum_i \dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}_0) \odot (\mathbf{X}_i(\boldsymbol{\delta} - \boldsymbol{\delta}_0)) \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \boldsymbol{\epsilon}_i \\
& + \sum_i 0.5 \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta}_0) \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \text{diag}\{(\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 \mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} \\
& \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \boldsymbol{\epsilon}_i \\
& - \sum_i 0.5 \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta}_0) \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0) \mathbf{R}^{-1} \mathbf{A}_i^{-3/2}(\boldsymbol{\theta}_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \\
& \text{diag}\{(\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 \mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} \boldsymbol{\epsilon}_i \\
& + \sum_i (\mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta}) \mathbf{A}_i^{1/2}(\boldsymbol{\theta}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}) - \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta}_0) \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0)) (\boldsymbol{\mu}(\boldsymbol{\eta}_0) - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta}))) \\
& + O_p(n\sqrt{K^3 p_1(K + p_1 + q_1)} r_n^2 + n\sqrt{K}(K + p_1 + q_1) r_n^2) \\
= & \sum_i \dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}_0) \odot (\mathbf{X}_i(\boldsymbol{\delta} - \boldsymbol{\delta}_0)) \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \boldsymbol{\epsilon}_i \\
& + \sum_i 0.5 \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta}_0) \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \text{diag}\{(\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 \mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} \\
& \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \boldsymbol{\epsilon}_i \\
& - \sum_i 0.5 \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta}_0) \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0) \mathbf{R}^{-1} \mathbf{A}_i^{-3/2}(\boldsymbol{\theta}_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \text{diag}\{(\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 \mathbf{X}_i, \mathbf{Z}_i) \\
& (\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} \boldsymbol{\epsilon}_i \\
& + O_p(n\sqrt{K^3 p_1(K + p_1 + q_1)} r_n^2 + n\sqrt{K}(K + p_1 + q_1) r_n^2).
\end{aligned}$$

By direct variance calculations, the first three terms above are

$$O_p(\sqrt{nK^3 p_1} r_n), O_p(\sqrt{nK(K + p_1 + q_1)} r_n), O_p(\sqrt{nK(K + p_1 + q_1)} r_n).$$

Other components of $\mathbf{U}_i(\boldsymbol{\theta})$ can be similarly dealt with. More specifically, we have

$$\begin{aligned}
& \mathbf{J}^T(\boldsymbol{\delta})\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta})\boldsymbol{\psi}) - \mathbf{J}^T(\boldsymbol{\delta}_0)\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0) \\
&= \frac{\partial \mathbf{J}^T(\boldsymbol{\delta}_0)}{\partial \boldsymbol{\delta}}(\boldsymbol{\delta} - \boldsymbol{\delta}_0)\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0) \\
&\quad + \mathbf{J}^T(\boldsymbol{\delta}_0)\mathbf{X}_i^T \text{diag}(\ddot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0 \odot (\mathbf{X}_i(\boldsymbol{\delta} - \boldsymbol{\delta}_0))) \\
&\quad + \mathbf{J}^T(\boldsymbol{\delta}_0)\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)(\boldsymbol{\psi} - \boldsymbol{\psi}_0)) \\
&\quad + O_p(\sqrt{p_1(K^3 + p_1)}r_n^2) \\
&= O_p((\sqrt{p_1(K^3 p_1 + p_1)}r_n).
\end{aligned}$$

$$\begin{aligned}
& \mathbf{J}^T(\boldsymbol{\delta})\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta})\boldsymbol{\psi})\mathbf{A}_i^{1/2}(\boldsymbol{\theta})\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}) - \mathbf{J}^T(\boldsymbol{\delta}_0)\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0) \\
& \quad \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0)\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \\
&= \frac{\partial \mathbf{J}^T(\boldsymbol{\delta}_0)}{\partial \boldsymbol{\delta}}(\boldsymbol{\delta} - \boldsymbol{\delta}_0)\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0)\mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0)\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \\
&\quad + \mathbf{J}^T(\boldsymbol{\delta}_0)\mathbf{X}_i^T \text{diag}(\ddot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0 \odot (\mathbf{X}_i(\boldsymbol{\delta} - \boldsymbol{\delta}_0)))\mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0)\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \\
&\quad + \mathbf{J}^T(\boldsymbol{\delta}_0)\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)(\boldsymbol{\psi} - \boldsymbol{\psi}_0))\mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0)\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \\
&\quad + 0.5\mathbf{J}^T(\boldsymbol{\delta}_0)\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0)\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0)\text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \\
&\quad \text{diag}\{(\mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0\mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\}\mathbf{R}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \\
&\quad - 0.5\mathbf{J}^T(\boldsymbol{\delta}_0)\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0)\mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0)\mathbf{R}^{-1}\mathbf{A}_i^{-3/2}(\boldsymbol{\theta}_0)\text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \\
&\quad \text{diag}\{(\mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0\mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} + O_p(\sqrt{(K^3 p_1 + p_1^2)(K + p_1 + q_1)}r_n^2 \\
&\quad + \sqrt{p_1}(K + p_1 + q_1)r_n^2) \\
&= O_p(\sqrt{K^3 p_1 + p_1^2}r_n + \sqrt{p_1}(K + p_1 + q_1)r_n).
\end{aligned}$$

Using $\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta})) = \boldsymbol{\epsilon}_i + (\boldsymbol{\mu}(\boldsymbol{\eta}_{0i}) - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta})))$ and $\boldsymbol{\mu}(\boldsymbol{\eta}_{0i}) - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta})) = O_p(\sqrt{K + p_1 + q_1}r_n)$,

$$\begin{aligned}
& \sum_i (\mathbf{J}^T(\delta) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta) \psi) \mathbf{A}_i^{1/2}(\theta) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\theta) - \mathbf{J}^T(\delta_0) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0) \mathbf{A}_i^{1/2}(\theta_0)) \\
& \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\theta_0)) (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\theta))) \\
= & \frac{\partial \mathbf{J}^T(\delta_0)}{\partial \delta} (\delta - \delta_0) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0) \mathbf{A}_i^{1/2}(\theta_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\theta_0) \boldsymbol{\epsilon}_i \\
& + \mathbf{J}^T(\delta_0) \mathbf{X}_i^T \text{diag}(\ddot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0 \odot (\mathbf{X}_i(\delta - \delta_0))) \mathbf{A}_i^{1/2}(\theta_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\theta_0) \boldsymbol{\epsilon}_i \\
& + \mathbf{J}^T(\delta_0) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta_0) (\psi - \psi_0)) \mathbf{A}_i^{1/2}(\theta_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\theta_0) \boldsymbol{\epsilon}_i \\
& + 0.5 \mathbf{J}^T(\delta_0) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0) \mathbf{A}_i^{-1/2}(\theta_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\theta_0))) \\
& \text{diag}\{(\mathbf{B}(\mathbf{X}_i \delta_0), \dot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0 \mathbf{X}_i, \mathbf{Z}_i)(\theta - \theta_0)\} \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\theta_0) \boldsymbol{\epsilon}_i \\
& - 0.5 \mathbf{J}^T(\delta_0) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0) \mathbf{A}_i^{1/2}(\theta_0) \mathbf{R}^{-1} \mathbf{A}_i^{-3/2}(\theta_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\theta_0))) \\
& \text{diag}\{(\mathbf{B}(\mathbf{X}_i \delta_0), \dot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0 \mathbf{X}_i, \mathbf{Z}_i)(\theta - \theta_0)\} \boldsymbol{\epsilon}_i \\
& + \sum_i (\mathbf{J}^T(\delta) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta) \psi) \mathbf{A}_i^{1/2}(\theta) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\theta) - \mathbf{J}^T(\delta_0) \mathbf{X}_i^T \\
& \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0) \mathbf{A}_i^{1/2}(\theta_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\theta_0)) (\boldsymbol{\mu}(\boldsymbol{\eta}_0) - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\theta))) \boldsymbol{\epsilon}_i \\
& + O_p(n \sqrt{(K^3 p_1 + p_1^2)(K + p_1 + q_1)} r_n^2 + n \sqrt{p_1} (K + p_1 + q_1) r_n^2) \\
= & \frac{\partial \mathbf{J}^T(\delta_0)}{\partial \delta} (\delta - \delta_0) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0) \mathbf{A}_i^{1/2}(\theta_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\theta_0) \boldsymbol{\epsilon}_i \\
& + \mathbf{J}^T(\delta_0) \mathbf{X}_i^T \text{diag}(\ddot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0 \odot (\mathbf{X}_i(\delta - \delta_0))) \mathbf{A}_i^{1/2}(\theta_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\theta_0) \boldsymbol{\epsilon}_i \\
& + \mathbf{J}^T(\delta_0) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta_0) (\psi - \psi_0)) \mathbf{A}_i^{1/2}(\theta_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\theta_0) \boldsymbol{\epsilon}_i \\
& + 0.5 \mathbf{J}^T(\delta_0) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0) \mathbf{A}_i^{-1/2}(\theta_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\theta_0))) \\
& \text{diag}\{(\mathbf{B}(\mathbf{X}_i \delta_0), \dot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0 \mathbf{X}_i, \mathbf{Z}_i)(\theta - \theta_0)\} \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\theta_0) \boldsymbol{\epsilon}_i \\
& - 0.5 \mathbf{J}^T(\delta_0) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0) \mathbf{A}_i^{1/2}(\theta_0) \mathbf{R}^{-1} \mathbf{A}_i^{-3/2}(\theta_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\theta_0))) \\
& \text{diag}\{(\mathbf{B}(\mathbf{X}_i \delta_0), \dot{\mathbf{B}}(\mathbf{X}_i \delta_0) \psi_0 \mathbf{X}_i, \mathbf{Z}_i)(\theta - \theta_0)\} \boldsymbol{\epsilon}_i \\
& + O_p(n \sqrt{(K^3 p_1 + p_1^2)(K + p_1 + q_1)} r_n^2 + n \sqrt{p_1} (K + p_1 + q_1) r_n^2).
\end{aligned}$$

By direct variance calculations, the first five terms above are of order

$$O_p(\sqrt{np_1}r_n), O_p(\sqrt{np_1^2}r_n), O_p(\sqrt{nK^3p_1}r_n), O_p(\sqrt{np_1(K+p_1+q_1)}r_n), O_p(\sqrt{np_1(K+p_1+q_1)}r_n).$$

Finally, we have

$$\begin{aligned} & \sum_i (\mathbf{Z}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\theta}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}) - \mathbf{Z}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0)) (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta}))) \\ = & 0.5 \mathbf{Z}_i^T \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \text{diag}\{(\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 \mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \boldsymbol{\epsilon}_i \\ & - 0.5 \mathbf{Z}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0) \mathbf{R}^{-1} \mathbf{A}_i^{-3/2}(\boldsymbol{\theta}_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \text{diag}\{(\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 \mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} \boldsymbol{\epsilon}_i \\ & + \sum_i (\mathbf{Z}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\theta}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}) - \mathbf{Z}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0)) (\boldsymbol{\mu}(\boldsymbol{\eta}_0) - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta}))) \\ & + O_p(n\sqrt{q_1}(K+p_1+q_1)r_n^2) \\ = & 0.5 \mathbf{Z}_i^T \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \text{diag}\{(\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 \mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\boldsymbol{\theta}_0) \boldsymbol{\epsilon}_i \\ & - 0.5 \mathbf{Z}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\theta}_0) \mathbf{R}^{-1} \mathbf{A}_i^{-3/2}(\boldsymbol{\theta}_0) \text{diag}(\ddot{\boldsymbol{\mu}}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_0))) \text{diag}\{(\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0), \dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 \mathbf{X}_i, \mathbf{Z}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} \boldsymbol{\epsilon}_i \\ & + O_p(n\sqrt{q_1}(K+p_1+q_1)r_n^2), \end{aligned}$$

with the first two terms are of order $O_p(\sqrt{nq_1(K+p_1+q_1)}r_n)$.

Thus we finally get

$$\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq Cr_n} \|\mathbf{S}(\boldsymbol{\theta}) - \bar{\mathbf{S}}_0(\boldsymbol{\theta})\| = O_p(\sqrt{n}(\sqrt{K^3p_1+p_1+q_1}r_n + nr_n^2\sqrt{K+p_1+q_1}(\sqrt{K^3p_1+p_1+q_1}))).$$

The second bound is proved similarly.

8.4 Proof of Theorem 2

Since $\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\beta}}$ is the zero of

$$\sum_i \begin{pmatrix} \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta}) \\ \mathbf{J}^T(\boldsymbol{\delta}) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}) \widehat{\boldsymbol{\psi}}) \\ \mathbf{Z}_i^T \end{pmatrix} \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\psi}}, \boldsymbol{\delta}, \boldsymbol{\beta}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\widehat{\boldsymbol{\psi}}, \boldsymbol{\delta}, \boldsymbol{\beta}) (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\psi}}, \boldsymbol{\delta}, \boldsymbol{\beta}))) = \mathbf{0},$$

it is the zero of

$$\mathbf{Q}(\boldsymbol{\xi}) := \sum_i \begin{pmatrix} \mathbf{Z}_i^T - \gamma_1 \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta}) \\ \mathbf{J}^T(\boldsymbol{\delta}) [\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}) \widehat{\boldsymbol{\psi}}) - \gamma_2 \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta})] \end{pmatrix} \mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\psi}}, \boldsymbol{\delta}, \boldsymbol{\beta}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\widehat{\boldsymbol{\psi}}, \boldsymbol{\delta}, \boldsymbol{\beta}) (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\psi}}, \boldsymbol{\delta}, \boldsymbol{\beta}))).$$

The above is true for any $\gamma_1 \in \mathbf{R}^{q_1 \times K}, \gamma_2 \in \mathbf{R}^{(p_1-1) \times K}$ but in order for our proof to work we should choose γ_1, γ_2 that satisfies $\|E_{\mathcal{M}}[\mathbf{Z}_i^T] - \gamma_1 \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta}_0)\| = O_p(\sqrt{q_1} K^{-d'})$, $\|E_{\mathcal{M}}[\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}) \widehat{\boldsymbol{\psi}})] - \gamma_2 \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta})\| = O_p(\sqrt{p_1} K^{-d'})$ which is possible by assumption (A6). As proved in Lemma 8.1, the above is close to

$$\sum_i \begin{pmatrix} \mathbf{Z}_i^T - E_{\mathcal{M}}[\mathbf{Z}_i^T] \\ \mathbf{J}^T(\boldsymbol{\delta}_0) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{h}}(\mathbf{X}_i \boldsymbol{\delta}_0)) - E_{\mathcal{M}}[\mathbf{J}^T(\boldsymbol{\delta}_0) \mathbf{X}_i^T \text{diag}(\dot{\mathbf{h}}(\mathbf{X}_i \boldsymbol{\delta}_0))] \end{pmatrix} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\widehat{\boldsymbol{\psi}}, \boldsymbol{\delta}, \boldsymbol{\beta}))),$$

with the difference $o_p(\sqrt{n})$. We define $\mathbf{W}_i(\boldsymbol{\theta})$ and $\mathbf{W}_{0i}$ by

$$\mathbf{W}_i(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{Z}_i^T - \gamma_1 \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta}) \\ \mathbf{J}^T(\boldsymbol{\delta}) [\mathbf{X}_i^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \boldsymbol{\delta}) \widehat{\boldsymbol{\psi}}) - \gamma_2 \mathbf{B}^T(\mathbf{X}_i \boldsymbol{\delta})] \end{pmatrix}$$

and

$$\mathbf{W}_{0i} = \begin{pmatrix} \mathbf{Z}_i^T - E_{\mathcal{M}}[\mathbf{Z}_i^T] \\ \mathbf{J}^T(\boldsymbol{\delta}_0) \mathbf{X}_i^T \text{diag}(\dot{h}(\mathbf{X}_i \boldsymbol{\delta}_0)) - E_{\mathcal{M}}[\mathbf{J}^T(\boldsymbol{\delta}_0) \mathbf{X}_i^T \text{diag}(\dot{h}(\mathbf{X}_i \boldsymbol{\delta}_0))] \end{pmatrix}.$$

That is, we have

$$\sup_{\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\| \leq Cr_n} \|\mathbf{Q}(\boldsymbol{\xi}) - \sum_i \mathbf{W}_{0i}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\hat{\boldsymbol{\psi}}, \boldsymbol{\delta}, \boldsymbol{\beta})))\| = o_p(\sqrt{n}). \quad (10)$$

We write

$$\begin{aligned} & \mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\hat{\boldsymbol{\psi}}, \boldsymbol{\delta}, \boldsymbol{\beta})) \\ = & \boldsymbol{\epsilon}_i - \text{diag}\{\dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i})\} [\text{diag}\{\dot{h}(\mathbf{X}_i \boldsymbol{\delta}_0)\} \mathbf{X}_i \mathbf{J}_0 (\boldsymbol{\delta}^{(-1)} - \boldsymbol{\delta}_0^{(-1)}) - \mathbf{Z}_i (\boldsymbol{\beta} - \boldsymbol{\beta}_0)] - \mathbf{r}_i, \end{aligned}$$

where $\mathbf{r}_i = (r_{i1}, \dots, r_{im})^T$ with

$$\begin{aligned}
& \mathbf{r}_i \\
&= \boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\psi}}, \boldsymbol{\delta}, \boldsymbol{\beta})) - \boldsymbol{\mu}(\boldsymbol{\eta}_{0i}) - \text{diag}\{\dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i})\}[\text{diag}\{\dot{h}(\mathbf{X}_i\boldsymbol{\delta}_0)\}\mathbf{X}_i\mathbf{J}_0(\boldsymbol{\delta}^{(-1)} - \boldsymbol{\delta}_0^{(-1)}) - \mathbf{Z}_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] \\
&= (\boldsymbol{\mu}(\boldsymbol{\eta}_i^*) - \dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i})) \odot (\boldsymbol{\eta}_i - \boldsymbol{\eta}_{0i}) + \boldsymbol{\mu}(\boldsymbol{\eta}_{0i}) \odot (\boldsymbol{\eta}_i - \boldsymbol{\eta}_{0i}) - \text{diag}\{\dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i})\}[\text{diag}\{\dot{h}(\mathbf{X}_i\boldsymbol{\delta}_0)\} \\
&\quad \mathbf{X}_i\mathbf{J}_0(\boldsymbol{\delta}^{(-1)} - \boldsymbol{\delta}_0^{(-1)}) - \mathbf{Z}_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] \\
&= \dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i}) \odot [\boldsymbol{\eta}_i - \boldsymbol{\eta}_{0i} - \text{diag}\{\dot{h}(\mathbf{X}_i\boldsymbol{\delta}_0)\}\mathbf{X}_i\mathbf{J}_0(\boldsymbol{\delta}^{(-1)} - \boldsymbol{\delta}_0^{(-1)}) - \mathbf{Z}_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] \\
&\quad + O_p((K + p_1 + q_1)r_n^2) \\
&= \dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i}) \odot [\mathbf{B}(\mathbf{X}_i\boldsymbol{\delta})\widehat{\boldsymbol{\psi}} - \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0 + \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0 - h(\mathbf{X}_i\boldsymbol{\delta}_0) - \dot{h}(\mathbf{X}_i\boldsymbol{\delta}_0)\mathbf{X}_i\mathbf{J}_0(\boldsymbol{\delta}^{(-1)} - \boldsymbol{\delta}_0^{(-1)})] \\
&\quad + O_p((K + p_1 + q_1)r_n^2) \\
&= \dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i}) \odot [(\mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}) - \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0))\boldsymbol{\psi}_0 + \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta})(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) + \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0 - h(\mathbf{X}_i\boldsymbol{\delta}_0) \\
&\quad - \dot{h}(\mathbf{X}_i\boldsymbol{\delta}_0)\mathbf{X}_i\mathbf{J}_0(\boldsymbol{\delta}^{(-1)} - \boldsymbol{\delta}_0^{(-1)})] + O_p((K + p_1 + q_1)r_n^2) \\
&= \dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i}) \odot [(\dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}^*)\boldsymbol{\psi}_0\mathbf{X}_i(\boldsymbol{\delta} - \boldsymbol{\delta}_0) + \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0)(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) + (\mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}) - \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0))(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \\
&\quad + \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0 - h(\mathbf{X}_i\boldsymbol{\delta}_0) - \dot{h}(\mathbf{X}_i\boldsymbol{\delta}_0)\mathbf{X}_i\mathbf{J}_0(\boldsymbol{\delta}^{(-1)} - \boldsymbol{\delta}_0^{(-1)})] + O_p((K + p_1 + q_1)r_n^2) \\
&= \dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i}) \odot [\mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0)(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) + \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0 - h(\mathbf{X}_i\boldsymbol{\delta}_0) + O_p((\sqrt{K^3p_1} + p_1)r_n^2) \\
&\quad + ((\dot{\mathbf{B}}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0 - \dot{h}(\mathbf{X}_i\boldsymbol{\delta}_0))\mathbf{X}_i\mathbf{J}_0(\boldsymbol{\delta}^{(-1)} - \boldsymbol{\delta}_0^{(-1)})] + O_p((K + p_1 + q_1)r_n^2) \\
&= \dot{\boldsymbol{\mu}}(\boldsymbol{\eta}_{0i}) \odot [\mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0)(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) + \mathbf{B}(\mathbf{X}_i\boldsymbol{\delta}_0)\boldsymbol{\psi}_0 - h(\mathbf{X}_i\boldsymbol{\delta}_0)] + O_p((\sqrt{K^3p_1} + p_1 + q_1)r_n^2).
\end{aligned}$$

Then we have

$$\begin{aligned}
& \sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}(\hat{\boldsymbol{\psi}}, \boldsymbol{\delta}, \boldsymbol{\beta}))) \\
= & \sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} \boldsymbol{\epsilon}_i \\
& - \sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} [\text{diag}\{\dot{h}(\mathbf{X}_i \boldsymbol{\delta}_0)\} \mathbf{X}_i \mathbf{J}_0(\boldsymbol{\delta}^{(-1)} - \boldsymbol{\delta}_0^{(-1)}) - \mathbf{Z}_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] \\
& - \sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} \mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0)(\boldsymbol{\psi} - \boldsymbol{\psi}_0) \\
& - \sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} [\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 - h(\mathbf{X}_i \boldsymbol{\delta}_0)] \\
& + O_p(n\sqrt{p_1 + q_1}(\sqrt{K^3 p_1} + p_1 + q_1)r_n^2). \tag{11}
\end{aligned}$$

To get a usable bound for the 3rd and 4th terms above, the critical observation is that $\sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} \mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0)$ and $\sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} [\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 - h(\mathbf{X}_i \boldsymbol{\delta}_0)]$ both have mean zero, due to the definition of the projection in $\mathbf{W}_{0i}$ (the projection means $\sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} g(\mathbf{X}_i \boldsymbol{\delta}_0)$ has mean zero for any square integrable function g), thus we can bound the two terms above by a simple variance calculation to get

$$\begin{aligned}
& \sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}(\hat{\boldsymbol{\psi}}, \boldsymbol{\delta}, \boldsymbol{\beta}))) \\
= & \sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} \boldsymbol{\epsilon}_i \\
& - \sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} [\text{diag}\{\dot{h}(\mathbf{X}_i \boldsymbol{\delta}_0)\} \mathbf{X}_i \mathbf{J}_0(\boldsymbol{\delta}^{(-1)} - \boldsymbol{\delta}_0^{(-1)}) - \mathbf{Z}_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] \\
& + O_p(\sqrt{n(p_1 + q_1)K}r_n) + O_p(n\sqrt{p_1 + q_1}(\sqrt{K^3 p_1} + p_1 + q_1)r_n^2).
\end{aligned}$$

Let $\tilde{\boldsymbol{\xi}} = (\tilde{\boldsymbol{\delta}}^{(-1)\text{T}}, \tilde{\boldsymbol{\beta}}^{\text{T}})^{\text{T}}$ be the zero of the first two terms above (we will denote the sum of

the two terms by $\tilde{\mathbf{Q}}(\xi)$). That is, $\tilde{\mathbf{Q}}(\tilde{\xi}) = \mathbf{0}$. More explicitly, with $\xi_0 = (\delta_0^{(-1)\top}, \beta_0^\top)^\top$,

$$\tilde{\xi} - \xi_0 = \left(\sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} \mathbf{W}_{0i}^\top \right)^{-1} \sum_i \mathbf{W}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} \epsilon_i.$$

The asymptotic normality of $\tilde{\xi}$ easily attains by the central limit theorem. For $\hat{\xi}$, by 10 we have $\tilde{\mathbf{Q}}(\hat{\xi}) = o_p(\sqrt{n})$. On the other hand, for any ξ with $\|\xi - \tilde{\xi}\| = c/\sqrt{n}$ for some small $c > 0$, the linearity of $\tilde{\mathbf{Q}}$ implies that

$$\begin{aligned} & \|\tilde{\mathbf{Q}}(\xi)\| \\ = & \|\tilde{\mathbf{Q}}(\xi) - \tilde{\mathbf{Q}}(\tilde{\xi})\| \\ = & \left\| \sum_i \mathbf{J}_{0i} \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} [\text{diag}\{\dot{h}(\mathbf{X}_i \delta_0)\} \mathbf{X}_i \mathbf{J}_0 (\delta^{(-1)} - \tilde{\delta}^{(-1)}) - \mathbf{Z}_i (\beta - \tilde{\beta})] \right\| \end{aligned}$$

is bounded below by a constant multiple of $c\sqrt{n}$ with probability approaching one. Thus $\|\hat{\xi} - \tilde{\xi}\| = o_p(1/\sqrt{n})$ which means $\hat{\xi}$ and $\tilde{\xi}$ have the same asymptotic normal distribution.

8.5 Proof of Theorem 3

Suppose $\hat{\theta} = (\hat{\psi}^\top, \hat{\delta}^\top, \hat{\beta}^\top)^\top$ is the estimator obtained when the irrelevant predictors are removed. Let $\check{\beta} = (\hat{\beta}^\top, 0^\top)^\top$, $\check{\delta} = (\hat{\delta}^\top, 0^\top)^\top$ and $\check{\psi} = \hat{\psi}$. Obviously, $\check{\theta} = (\check{\psi}^\top, \check{\delta}^\top, \check{\beta}^\top)^\top$ satisfies

$$\sum_i \mathbf{U}_{(1)i}^\top(\check{\theta}) \mathbf{A}_i^{1/2}(\check{\theta}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\check{\theta}) (\mathbf{Y}_i - \mu(\eta_i(\check{\theta}))) = 0,$$

where

$$\mathbf{U}_{(1)i}^\top(\theta) = \begin{pmatrix} \mathbf{B}^\top(\mathbf{X}_i \delta) \\ \mathbf{J}_{(1)}^\top(\delta) \mathbf{X}_{(1)i}^\top \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \delta) \psi) \\ \mathbf{Z}_{(1)i}^\top \end{pmatrix},$$

$\mathbf{X}_{(1)i}$ ($\mathbf{Z}_{(1)i}$) contains the first p_1 (q_1) rows of $\mathbf{X}_i$ ($\mathbf{Z}_i$) and $\mathbf{J}_{(1)}(\boldsymbol{\delta})$ is the upper-right $p_1 \times (p_1 - 1)$ submatrix of $\mathbf{J}(\boldsymbol{\delta})$. Combined with $q_{\lambda_1}(|\check{\delta}_j|) = 0, j > p_1$ and $q_{\lambda_2}(|\check{\beta}_j|) = 0, j > q_1$ since $|\check{\delta}_j| \gg \lambda_1, |\check{\beta}_j| \gg \lambda_2$, we proved (i).

For (ii), we only need to bound

$$\sum_i \mathbf{Z}_{ij}^T \mathbf{A}_i^{1/2}(\check{\boldsymbol{\theta}}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\check{\boldsymbol{\theta}}) (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\check{\boldsymbol{\theta}}))), j > q_1$$

and

$$\mathbf{X}_{ij}^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \check{\boldsymbol{\delta}}) \check{\boldsymbol{\psi}}) \mathbf{A}_i^{1/2}(\check{\boldsymbol{\theta}}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\check{\boldsymbol{\theta}}) (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\check{\boldsymbol{\theta}}))), j > p_1.$$

We focus on the second expression above since the two are similar. Obviously, its value does not change if $\mathbf{X}_{ij}^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \check{\boldsymbol{\delta}}) \check{\boldsymbol{\psi}})$ is replaced by $\mathbf{X}_{ij}^T \text{diag}(\dot{\mathbf{B}}(\mathbf{X}_i \check{\boldsymbol{\delta}}) \check{\boldsymbol{\psi}}) - \gamma_{2j}^T \mathbf{B}^T(\mathbf{X}_i^T \check{\boldsymbol{\delta}})$, and we denote the latter by $\mathbf{V}_{ij}^T(\check{\boldsymbol{\theta}})$. As in (10), the difference between

$$\mathbf{V}_{ij}^T(\check{\boldsymbol{\theta}}) \mathbf{A}_i^{1/2}(\check{\boldsymbol{\theta}}) \mathbf{R}^{-1} \mathbf{A}_i^{-1/2}(\check{\boldsymbol{\theta}}) (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\check{\boldsymbol{\theta}})))$$

and

$$\mathbf{V}_{0ij}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\check{\boldsymbol{\theta}})))$$

is $o_p(\sqrt{n})$ where $\mathbf{V}_{0ij}^T = \mathbf{X}_{ij}^T \text{diag}(\dot{h}(\mathbf{X}_i \boldsymbol{\delta}_0)) - E_{\mathcal{M}}[\mathbf{X}_{ij}^T \text{diag}(\dot{h}(\mathbf{X}_i \boldsymbol{\delta}_0))]$. Then we have, as in

(11),

$$\begin{aligned}
& \mathbf{V}_{0ij}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\eta}_i(\check{\boldsymbol{\theta}}))) \\
&= \sum_i \mathbf{V}_{0ij}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{-1/2} \boldsymbol{\epsilon}_i \\
&\quad - \sum_i \mathbf{V}_{0ij}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} [\text{diag}\{\dot{h}(\mathbf{X}_i \boldsymbol{\delta}_0)\} \mathbf{X}_i \mathbf{J}_0(\boldsymbol{\delta}^{(-1)} - \boldsymbol{\delta}_0^{(-1)}) - \mathbf{Z}_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] \\
&\quad - \sum_i \mathbf{V}_{0ij}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} \mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0)(\boldsymbol{\psi} - \boldsymbol{\psi}_0) \\
&\quad - \sum_i \mathbf{V}_{0ij}^T \mathbf{A}_{0i}^{1/2} \mathbf{R}^{-1} \mathbf{A}_{0i}^{1/2} [\mathbf{B}(\mathbf{X}_i \boldsymbol{\delta}_0) \boldsymbol{\psi}_0 - h(\mathbf{X}_i \boldsymbol{\delta}_0)] \\
&\quad + O_p(n\sqrt{p_1 + q_1}(\sqrt{K^3 p_1} + p_1 + q_1)r_n^2) \\
&= O_p(\sqrt{n\log(p+q)}) + O_p(n\sqrt{p_1 + q_1 + K r_n}) + O_p(n\sqrt{p_1 + q_1}(\sqrt{K^3 p_1} + p_1 + q_1)r_n^2).
\end{aligned}$$

In the last line the bound $O_p(\sqrt{n\log(p+q)})$ is the result of applying Bernstein's inequality for the first term together with Boole's inequality to make the bound uniform over different j , which is the reason for the extra logarithmic term to emerge. By our assumption this is of order $O_p(n\lambda_1)$, which proves (ii). The penalty terms are obviously zero here. Properties (iii) and (iv) of Theorem 3 are trivial based on the definition of $\check{\boldsymbol{\theta}}$.

Remark 8.1 *It is easy to see the bound (ii) implies that $\check{\boldsymbol{\theta}}$ has a zero crossing property: for $a > 0$ sufficiently small,*

$$\lim_{a \rightarrow 0, a \neq 0} \mathbf{S}_j^P(\check{\boldsymbol{\psi}}, \check{\boldsymbol{\delta}} + a\mathbf{e}_j, \check{\boldsymbol{\beta}}) \mathbf{S}_j^P(\check{\boldsymbol{\psi}}, \check{\boldsymbol{\delta}} - a\mathbf{e}_j, \check{\boldsymbol{\beta}}) < 0, j = p_1 + 1, \dots, p,$$

with probability approaching one, here $\mathbf{S}_j^P$ is the estimating equation in $\mathbf{S}^P$ corresponding to $\mathbf{X}_{ij}$. Similar property holds for coordinate-wise perturbation of $\check{\boldsymbol{\beta}}$.

The reason the displayed above is true is that since $\mathbf{S}_j(\check{\boldsymbol{\psi}}, \check{\boldsymbol{\delta}}, \check{\boldsymbol{\beta}}) = o_p(n\lambda_1)$, by continuity, as long as a is small enough, $\mathbf{S}_j(\check{\boldsymbol{\psi}}, \check{\boldsymbol{\delta}} + a\mathbf{e}_j, \check{\boldsymbol{\beta}}) \mathbf{S}_j(\check{\boldsymbol{\psi}}, \check{\boldsymbol{\delta}} - a\mathbf{e}_j, \check{\boldsymbol{\beta}}) = o_p(n\lambda_1)$. For the penalty,

$nq_{\lambda_1}(a)\text{sgn}(a) = n\lambda_1$ and $nq_{\lambda_1}(-a)\text{sgn}(-a) = -n\lambda_1$ for $a > 0$, which are the dominant terms and thus $\lim_{a \rightarrow 0, a \neq 0} \mathbf{S}_j^P(\check{\psi}, \check{\delta} + a\mathbf{e}_j, \check{\beta})\mathbf{S}_j^P(\check{\psi}, \check{\delta} - a\mathbf{e}_j, \check{\beta}) < 0$.

BIBLIOGRAPHY

- Bähler, J. (2005). Cell-cycle control of gene expression in budding and fission yeast. *Annu. Rev. Genet.* **39**, 69–94.
- Bai, Y., Fung, W. K., and Zhu, Z. Y. (2009). Penalized quadratic inference functions for single-index models with longitudinal data. *Journal of Multivariate Analysis* **100**, 152–161.
- Banerjee, N. and Zhang, M. Q. (2003). Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Research* **31**, 7024–7031.
- Boni, J. P., Leister, C., Bender, G., Fitzpatrick, V., Twine, N., Stover, J., Dorner, A., Immermann, F., and Burczynski, M. E. (2005). Population pharmacokinetics of CCI-779: correlations to safety and pharmacogenomic responses in patients with advanced renal cancer. *Clinical Pharmacology & Therapeutics* **77**, 76–89.
- Bradic, J., Fan, J., and Wang, W. (2011). Penalized composite quasi-likelihood for ultra-high dimensional variable selection. *Journal of the Royal Statistical Society: Series B* **73**, 325–349.
- Cai, L., Wu, H., Li, D., Zhou, K., and Zou, F. (2015). Type 2 diabetes biomarkers of human gut microbiota selected via iterative sure independent screening method. *PloS one* **10**, 1–15.
- Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association* **92**, 477–489.
- Chen, J., Li, D., Liang, H., and Wang, S. (2015). Semiparametric GEE analysis in partially linear single-index models for longitudinal data. *The Annals of Statistics* **43**, 1682–1715.

- Cheng, C. and Li, L. M. (2008). Systematic identification of cell cycle regulated transcription factors from microarray time series data. *BMC Genomics* **9**, 116.
- Cheng, M.-Y., Honda, T., Li, J., and Peng, H. (2014). Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *The Annals of Statistics* **42**, 1819–1849.
- Cho, H. and Qu, A. (2013). Model selection for correlated data with diverging number of parameters. *Statistica Sinica* **23**, 901–927.
- Das, D., Banerjee, N., and Zhang, M. Q. (2004). Interacting models of cooperative gene regulation. *Proceedings of the National Academy of Sciences* **101**, 16234–16239.
- De Boor, C. (2001). *A Practical Guide to Splines; Rev. Ed.* Applied mathematical sciences. Springer.
- Dogan, M. V., Grumbach, I. M., Michaelson, J. J., and Philibert, R. A. (2018). Integrated genetic and epigenetic prediction of coronary heart disease in the framingham heart study. *PloS one* **13**,.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.
- Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B* **70**, 849–911.
- Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B* **75**, 531–552.
- Fang, Y., Qin, Y., Zhang, N., Wang, J., Wang, H., and Zheng, X. (2015). DISIS: prediction of drug response through an iterative sure independence screening. *PloS one* **10**, 1–13.

- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* pages 1029–1054.
- He, X., Zhu, Z.-Y., and Fung, W.-K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89**, 579–590.
- Hoover, D., Rice, J., Wu, C., and Yang, L. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822.
- Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics* **38**, 2282.
- Hui, F. K., Warton, D. I., and Foster, S. D. (2015). Tuning parameter selection for the adaptive lasso using ERIC. *Journal of the American Statistical Association* **110**, 262–269.
- Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics* **33**, 1617–1642.
- Inan, G. and Wang, L. (2017). PGEE: An R package for analysis of longitudinal data with high-dimensional covariates. *R Journal* **9**, 393–402.
- Johnson, B. A., Lin, D., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103**, 672–680.
- Koch, C. and Nasmyth, K. (1994). Cell cycle regulated transcription in yeast. *Current Opinion in Cell Biology* **6**, 451–459.
- Lai, P., Li, G., and Lian, H. (2013). Quadratic inference functions for partially linear

- single-index models with longitudinal data. *Journal of Multivariate Analysis* **118**, 115–127.
- Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002). Transcriptional regulatory networks in *saccharomyces cerevisiae*. *Science* **298**, 799–804.
- Li, G., Lai, P., and Lian, H. (2015). Variable selection and estimation for partially linear single-index models with longitudinal data. *Statistics and Computing* **25**, 579–593.
- Lian, H., Liang, H., and Wang, L. (2014). Generalized additive partial linear models for clustered data with diverging number of covariates using GEE. *Statistica Sinica* **24**, 173–196.
- Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics* **19**, 474–482.
- Ma, S., Liang, H., and Tsai, C.-L. (2014). Partially linear single index models for repeated measurements. *Journal of Multivariate Analysis* **130**, 354–375.
- Macke, J. H., Berens, P., Ecker, A. S., Tolias, A. S., and Bethge, M. (2009). Generating spike trains with specified correlation coefficients. *Neural computation* **21**, 397–423.
- Qu, A. and Li, R. (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics* **62**, 379–391.
- Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823–836.
- Ruppert, D. and Carroll, R. J. (2000). Theory & methods: Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal of Statistics* **42**, 205–223.

- Schumaker, L. (2007). *Spline functions: basic theory*. Cambridge University Press.
- Song, R., Yi, F., and Zou, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica* **24**, 1735–1752.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., Futcher, B., et al. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273–3297.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58**, 267–288.
- Touloumis, A. (2016). Simulating correlated binary and multinomial responses under marginal model specification: The SimCorMultRes package. *The R Journal* **8**, 79–91.
- Tsai, H.-K., Lu, H. H.-S., and Li, W.-H. (2005). Statistical methods for identifying yeast cell cycle transcription factors. *Proceedings of the National Academy of Sciences* **102**, 13532–13537.
- Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B* **71**, 671–683.
- Wang, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics* **39**, 389–417.
- Wang, L., Chen, G., and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23**, 1486–1494.

- Wang, L., Liu, X., Liang, H., and Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *The Annals of Statistics* **39**, 1827.
- Wang, L., Xue, L., Qu, A., Liang, H., 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**, 592–624.
- Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68**, 353–360.
- Wang, P., Tsai, G.-f., and Qu, A. (2012). Conditional inference functions for mixed-effects models with unspecified random-effects distribution. *Journal of the American Statistical Association* **107**, 725–736.
- Xu, P., Zhu, L., and Li, Y. (2014). Ultrahigh dimensional time course feature selection. *Biometrics* **70**, 356–365.
- Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* **97**, 1042–1054.
- Yu, Y., Wu, C., and Zhang, Y. (2017). Penalised spline estimation for generalised partially linear single-index models. *Statistics and Computing* **27**, 571–582.
- Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.
- Zhou, J. and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistical Association* **107**, 701–710.