Received: 20 March 2019 | Accepted: 21 July 2020

DOI: 10.1111/biom.13348

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Ultra high-dimensional semiparametric longitudinal data analysis

Brittany Green[1] | **Heng Lian**[2] | **Yan Yu**[3] | **Tianhai Zu**[3]

[1] Department of Computer Information Systems, University of Louisville, Louisville, Kentucky

[2] Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong, China

[3] Department of Operations, Business Analytics, & Information Systems, University of Cincinnati, Cincinnati, Ohio

Correspondence
Yan Yu, Department of Operations, Business Analytics, and Information Systems, University of Cincinnati, Cincinnati, OH 45221.
Email: Yan.Yu@uc.edu

Abstract

As ultra high-dimensional longitudinal data are becoming ever more apparent in fields such as public health and bioinformatics, developing flexible methods with a sparse model is of high interest. In this setting, the dimension of the covariates can potentially grow exponentially as $\exp(n^{1/2})$ with respect to the number of clusters n. We consider 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, and we present an efficient algorithm to handle the computational challenges. We show the effectiveness of our method and algorithm via a simulation study and a yeast cell cycle gene expression data.

KEYWORDS

generalized estimating equations, oracle property, polynomial spline, SCAD, single-index model, variable selection

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

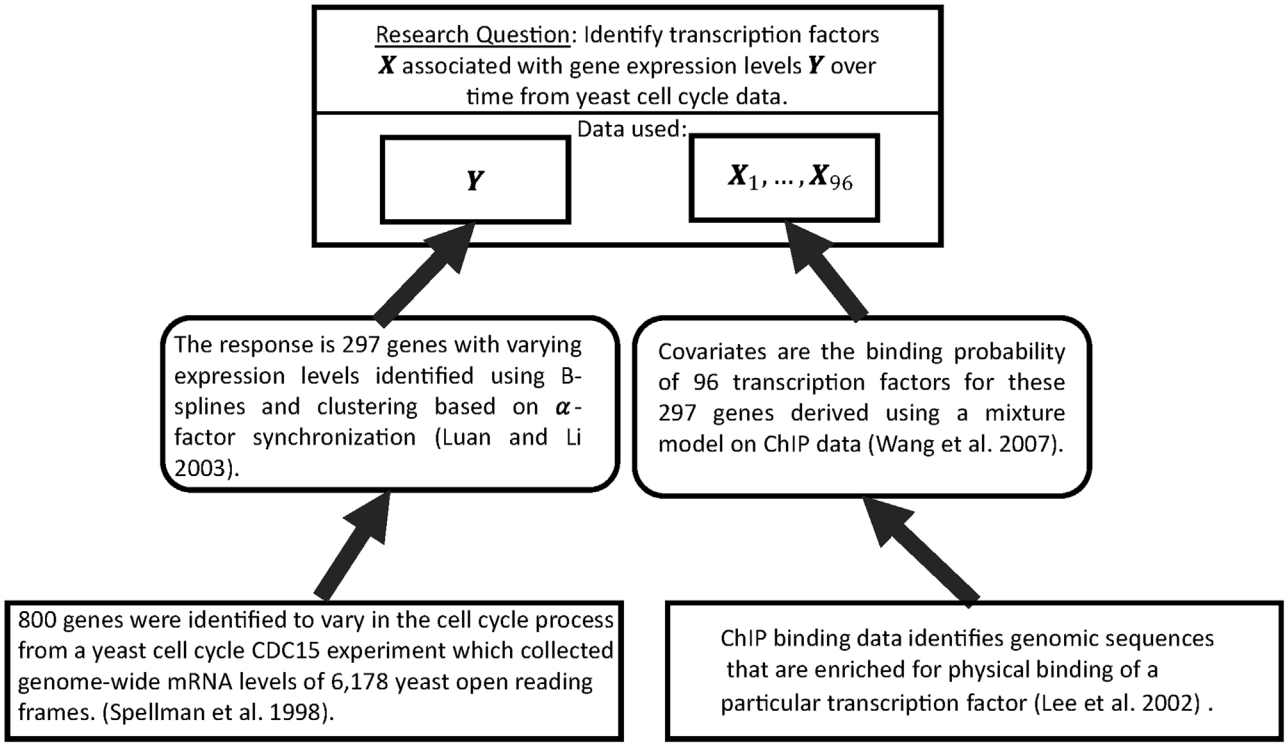

FIGURE 1 Flowchart diagram of the data integration steps for the yeast cell cycle gene expression data. The 297 genes identified are from an initial subset of 800 genes from Spellman *et al.* (1998) and a subsequent clustering model of Luan and Li (2003). The response is the log of the gene expression level. The covariates are the binding probabilities of 96 transcription factors determined using a mixture model from Wang *et al.* (2007) and ChIP data of Lee *et al.* (2002).

shown in the flowchart Figure 1 and is further discussed in Section 5.2.

Due to the complexity of gene transcriptional regulation, the relationship between the transciption 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) (Banerjee and Zhang, 2003; Tsai *et al.*, 2005). 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.

In order to offer flexible modeling and some interpretability in ultra high-dimensional longitudinal data, we consider a semiparametric approach, longitudinal single-index models

$$\mathbf{Y} = h(\mathbf{X}\boldsymbol{\delta}_0) + \epsilon, \tag{1}$$

where $\mathbf{Y}$ is a longitudinal response variable and $\mathbf{X}$ are p_n-dimensional covariates; ϵ is an error term $E(\epsilon|\mathbf{X}) = 0$; and $h(\cdot)$ is a flexible univariate function. More details are presented in subsequent sections.

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 (eg, 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 interactions.

Compared with the longitudinal single-index model (1), a longitudinal linear model in high-dimension $\mathbf{Y} = \mathbf{X}\boldsymbol{\delta} + \epsilon$ (Wang *et al.*, 2012) may be too restrictive. It may misspecify 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 a sparse set of 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) (eg, Wang *et al.*, 2012) 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 nonparametrically, which is shown to be computationally stable. Along with a flexible semiparametric 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 4 in Section 5.2. 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.

In this paper, 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 $\mathbf{Z}\boldsymbol{\beta}_0$. Our proposed approach of semiparametric penalized generalized estimating equations along with the SCAD penalty for the longitudinal single-index model (1) with ultra high-dimensional 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 $h(\mathbf{X}\boldsymbol{\delta}_0)$, or equivalently, $h(\mathbf{X}_1\delta_1 + \cdots + \mathbf{X}_{p_n}\delta_{p_n})$, 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 $h(\mathbf{X}\boldsymbol{\delta})$ 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 noncorrelated data. For example, though the finding discovery of a nonlinear relationship is very interesting, the analysis in Carroll *et al.* (1997) is only conducted on exam 3 of the 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 a 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 a 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 $\mathbf{Y}_i = (Y_{i1}, \dots Y_{im_i})^{\mathrm{T}}$, $\mathbf{X}_i = (\mathbf{X}_{i1}, \dots \mathbf{X}_{im_i})^{\mathrm{T}}$, $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \dots \mathbf{Z}_{im_i})^{\mathrm{T}}$ 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 paper 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. 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 (eg, Carroll *et al.*, 1997). We model the conditional mean for longitudinal data semiparametrically by generalized partially linear single-index models

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

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 $\boldsymbol{\eta}_i = (\eta_{i1}, \dots, \eta_{im_i})^{\mathrm{T}} = h(\mathbf{X}_i\boldsymbol{\delta}_0) + \mathbf{Z}_i\boldsymbol{\beta}_0$. 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 $\mathbf{B}(u)\boldsymbol{\psi}$ based on $\tilde{K}$ interior knots. Here $\mathbf{B}(u)$ is an s-degree B-spline basis and $\boldsymbol{\psi} = (\psi_1, \dots, \psi_K)^{\mathrm{T}}$ are the $K \equiv K_n = 1 + s + \tilde{K}$ spline coefficients to estimate (see Web Appendix for details). 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 $\boldsymbol{\theta}_0 = (\boldsymbol{\psi}_0^{\mathrm{T}}, \boldsymbol{\delta}_0^{\mathrm{T}}, \boldsymbol{\beta}_0^{\mathrm{T}})^{\mathrm{T}}$, 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, $\hat{}$ for an estimate, and subscript (1) for oracle. For example, $\hat{\boldsymbol{\theta}}_{(1)}$ denotes the oracle estimator of the true parameter $\boldsymbol{\theta}_{0(1)} = (\boldsymbol{\psi}_{0(1)}^{\mathrm{T}}, \boldsymbol{\delta}_{0(1)}^{\mathrm{T}}, \boldsymbol{\beta}_{0(1)}^{\mathrm{T}})^{\mathrm{T}}$. 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 the paper.

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

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

where

$$\mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\theta}_{(1)}) = \begin{pmatrix} \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{(1)i}\boldsymbol{\delta}_{(1)}) \\ \mathbf{J}^{\mathrm{T}}(\boldsymbol{\delta}_{(1)})\mathbf{X}_{(1)i}^{\mathrm{T}}\mathrm{diag}(\dot{\mathbf{B}}(\mathbf{X}_{(1)i}\boldsymbol{\delta}_{(1)})\boldsymbol{\psi}_{(1)}) \\ \mathbf{Z}_{(1)i}^{\mathrm{T}} \end{pmatrix}.$$

Here $\boldsymbol{\mu}(\boldsymbol{\eta}_i(\boldsymbol{\theta}_{(1)}))$ 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 $\boldsymbol{\delta}_{(1)}^{(-1)} = (\delta_2, \dots, \delta_{p_1})^{\mathrm{T}}$, and accordingly define the Jacobian matrix $\mathbf{J}(\boldsymbol{\delta}_{(1)}) = \partial\boldsymbol{\delta}_{(1)}/\partial\boldsymbol{\delta}_{(1)}^{(-1)}$. $\mathbf{R}^{-1}(\boldsymbol{\gamma})$ is a pre-specified "working" correlation matrix. And $\boldsymbol{\gamma}$ 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 need 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 $\widehat{\boldsymbol{\xi}}_{(1)}$ are estimates of the reparametrized parameters $\boldsymbol{\xi}_{0(1)} = (\boldsymbol{\delta}_{0(1)}^{(-1)\mathrm{T}}, \boldsymbol{\beta}_{0(1)}^{\mathrm{T}})^{\mathrm{T}}$. Please refer to Web Appedix A for detailed definitions, assumptions, and technical proofs.

Theorem 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* $\|\widehat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{0(1)}\| = O_p(r_n)$, *where* $r_n = \sqrt{(K + p_1 + q_1)/n} + K^{-d}$.

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

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

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

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 (see Remark in Web Appendix). Here $\mathbf{q}_\lambda(|\boldsymbol{\delta}, \boldsymbol{\beta}|) = (\mathbf{0}_K, q_\lambda(|\delta_2|), \ldots, q_\lambda(|\delta_{p_n}|), q_\lambda(|\beta_1|), \ldots, q_\lambda(|\beta_{q_n}|))$ 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 nonconvex 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 $q_\lambda(\zeta) = \lambda\{I(\zeta \le \lambda) + \frac{(a\lambda - \zeta)_+}{(a-1)\lambda} I(\zeta > \lambda)\}$, 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. *Under assumptions for Theorem 2, and that*

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

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

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

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

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

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

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

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 nonconvex 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 (eg, 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.

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 nonconvex SCAD penalty for variable selection, we adopt a similar approach as in Wang *et al.* (2012) for high-dimensional linear PGEE, where the minorization-maximization algorithm method is used (see Hunter and Li, 2005 and web appendix for details). We use this method together with the Newton-Raphson method to get the estimates similar to Inan and Wang (2017).

Additionally, we echo findings in Fan and Lv (2008) that stress the computational speed and accuracy benefits of sure independence screening (SIS) to first efficiently reduce dimensionality, especially in the ultra high-dimensional case when the computational performance without screening is not desirable for practical usage. In our simulation studies, we incorporate screening and find SIS works well. Other variants of SIS (eg, Cheng *et al.*, 2014) may also be used. For the moderately high-dimensional case, one may adopt well-performing initial values such as estimates from classical linear GEE or penalized GEE. We include the detailed algorithm in the Web Appendix.

4.2 | Tuning parameter selection

For the proposed iterative algorithm, tuning parameters include the degree and number of knots of the 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 (eg, Huang *et al.*, 2010).

In order to choose the critical penalty parameter λ for variable selection, we use a high-dimensional BIC (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 HBIC involves minimizing

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

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). We provide more discussion about the tuning parameter selection in the Web Appendix.

5 | NUMERICAL STUDY

We conduct Monte Carlo studies to assess the performance of the proposed estimation and variable selection approaches. In the Web Appendix, 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. In the numerical study below, we use quadratic splines and three equally-spaced knots. All simulations are based on 200 replications.

5.1 | Simulation

Example 1. The correlated Gaussian response data are generated from the following model: $Y_{ij} = \sin((\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\delta} - a)\pi)/(b - a) + \mathbf{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{ij}$. The number of subjects is $n = 100$ and the number of observations

TABLE 1 Summary of parameter estimates for the partially linear single-index model in Example 1 with $p_n = 250$ and $q_n = 250$

	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

Note. The true $\boldsymbol{\delta}_0 = (1,1,1,1,1,0,\ldots,0)^{\mathrm{T}}/\sqrt{5}$ and the true $\boldsymbol{\beta}_0 = (1,1,1,0,\ldots,0)^{\mathrm{T}}$. 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.

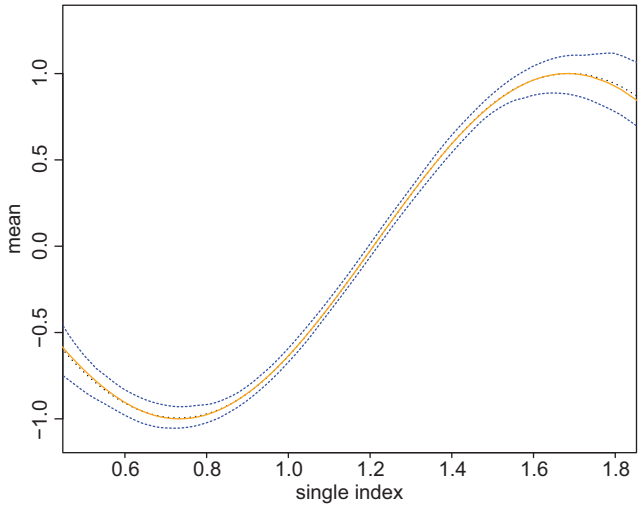

FIGURE 2 Fitted plot for the curve of $h(\cdot)$ from Example 1 with $p_n = 250$ and $q_n = 250$. This figure is based on 200 simulation replications. The solid line is the true function. The dot dashed line is the average fitted curve and the dashed lines are the corresponding 2.5% and 97.5% quantiles. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

per subject is $m = 10$. The true parameter vectors are $\boldsymbol{\delta}_0 = (1,1,1,1,1,0,\ldots,0)^{\mathrm{T}}/\sqrt{5}$ and $\boldsymbol{\beta}_0 = (1,1,1,0,\ldots,0)$. 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 $h(\cdot)$. (This figure appears in color in the electronic version of this article, and any mention of color refers to that version.) 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.

Example 2. We consider the correlated binary responses from the marginal mean responses $log(p_{ij}/(1-p_{ij})) = \sin((\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\delta} - a)\pi)/(b-a)$. The number of subjects is $n = 200$ and the number of observations per subject is $m = 10$. The true single-index vector $\boldsymbol{\delta}_0 = (1,1,1,1,0,\ldots,0)^{\mathrm{T}}/\sqrt{4}$ 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 are 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, for example, 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.

5.2 | 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

TABLE 2 Summary of estimation and variable selection results from Example 1 with $p_n = 250$ and $q_n = 250$

Comparison A.	Single-index covariates				Partially linear covariates			
Method	**Correct%**	**TNs**	**FNs**	**MSEp**	**Correct%**	**TNs**	**FNs**	**MSEq**
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
Single-index 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
Single-index LASSO	53	244.39	0	0.0038	0	236.01	0	0.1975

Note. 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 $\boldsymbol{\delta}_0 = (1, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}/\sqrt{5}$ and the true partially linear parameters $\boldsymbol{\beta}_0 = (1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$. "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.

TABLE 3 Summary of estimation and variable selection results from Example 2 with binary response

	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

Note. The true single-index vector $\boldsymbol{\delta}_0 = (1, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}/\sqrt{4}$ 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.

cycle affects biological processes and insights into cell cycle regulation. The yeast cell cycle data are from the CDC15 experiment that collected genome-wide mRNA levels of 6178 yeast open reading frames for 18 time points or 119 min divided into 7 min 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 $j = 1, \ldots, m$ for gene $i = 1, 2, \ldots, n$. 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 two 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 that can be estimated using the EM

TABLE 4 Summary of estimation results from the yeast cell cycle gene expression data

	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)

Note. "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.

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 the 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 $k = 1, \ldots, 96$ 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 five 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), TFs 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), that is, $\mathbf{Y}_i = \alpha_0 + \sum_{k=0}^{96} \delta_k \mathbf{X}_{ik} + \epsilon_i$. Note that complex relationship and potential interactions among TFs and time cannot be captured through this linear model.

Consistent to the previous research, we find that important TFs 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 that is involved in the regulation of genes in the the G1 stage as discussed in Cheng and Li (2008), Bähler (2005), and 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 the G1 phase (Bähler, 2005). We find that 10 TFs, 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 4 reports the 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 the flexible univariate function $h(\cdot)$ uses the quadratic spline basis with one knot that yields good results. The 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.

One very interesting finding is that the transcription factor YAP5, noted as an important factor in the previous literature (eg, 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, and 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 remaining five TFs have not been commonly noted in the previous literature for Stage G1.

6 | DISCUSSION

In this paper, 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 nonconvex 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. We also provide discussion on assessing variable uncertainty in the Web Appendix.

ACKNOWLEDGMENTS

We are grateful to the editor, the associate editor, and three referees for their insightful comments, which have led to significant improvement of our paper.

DATA AVAILABILITY STATEMENT

The Yeast Cell Cycle data that support the findings of this paper are openly available in the PGEE R package from Inan and Wang (2017). The PGEE package can be accessed at https://CRAN.R-project.org/package=PGEE.

ORCID

Yan Yu https://orcid.org/0000-0002-2859-3093

REFERENCES

Bähler, J. (2005) Cell-cycle control of gene expression in budding and fission yeast. *Annual Review of Genetics*, 39, 69–94.

Banerjee, N. and Zhang, M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Research*, 31, 7024–7031.

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.

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.

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.

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. and Wei, F. (2010) Variable selection in nonparametric additive models. *Annals of statistics*, 38(4), 2282–2313.

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. *The 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.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., 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.

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., 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 (Methodological)*, 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, L., Chen, G. and Li, H. (2007) Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23, 1486–1494.

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

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.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 1-5 are available with this paper at the Biometrics website on Wiley Online Library. In addition, the R codes and ReadMe file for analyzing the yeast cell cycle data and the R codes for a simulation example are available at the Biometrics website on Wiley Online Library.

How to cite this article: Green B, Lian H, Yu Y, Zu T. Ultra high-dimensional semiparametric longitudinal data analysis. *Biometrics*. 2021;77:903–913. https://doi.org/10.1111/biom.13348