Electronic Journal of Statistics
Vol. 6 (2012) 2235–2273
ISSN: 1935-7524
DOI: 10.1214/12-EJS744

A dimension reduction based approach for estimation and variable selection in partially linear single-index models with high-dimensional covariates

Jun Zhang

Shenzhen-Hong Kong Joint Research Centre for Applied Statistical Sciences
Shenzhen University
Shenzhen, China
e-mail: zhangjunstat@gmail.com

Tao Wang, Lixing Zhu[†]

Department of Mathematics
Hong Kong Baptist University
Hong Kong, China
e-mail: 10466029@life.hkbu.edu.hk; lzhu@math.hkbu.edu.hk

and

Hua Liang[‡]

Department of Biostatistics and Computational Biology
University of Rochester
Rochester, NY 14642, USA
e-mail: hliang@bst.rochester.edu

Abstract: In this paper, we formulate the partially linear single-index models as bi-index dimension reduction models for the purpose of identifying significant covariates in both the linear part and the single-index part through only one combined index in a dimension reduction approach. This is different from all existing dimension reduction methods in the literature, which in general identify two basis directions in the subspace spanned by the parameter vectors of interest, rather than the two parameter vectors themselves. This approach makes the identification and the subsequent estimation and variable selection easier than existing methods for multi-index models. When the number of parameters diverges with the sample size, we then adopt coordinate-independent sparse estimation procedure to select significant covariates and estimate the corresponding parameters. The resulting sparse dimension reduction estimators are shown to be consistent and asymptotically normal with the oracle property. Simulations are conducted to evaluate the performance of the proposed method, and a real data set is analysed for an illustration.

*This work was done when the first author visited the fourth author. Zhang's work was supported by the NSFC grant 11101157.

†Zhu's research was supported by a RGC grant from the Research Grants Council of Hong Kong, Hong Kong, China.

‡Corresponding author. Liang's research was partially supported by NSF grants DMS-1007167 and DMS-1207444.

AMS 2000 subject classifications: Primary 62G08; secondary 62G20, 62J02, 62F12.
Keywords and phrases: Coordinate-independent sparse estimation (CISE), cumulative slicing estimation, high-dimensional covariate, inverse regression, inverse regression, partially linear models, profile likelihood, sufficient dimension reduction.

Received January 2012.

1. Introduction

The single index model is an important generalization of the multiple linear regression model with an unknown link function. It has been widely studied and used to explore the complicated relation between the response and covariates of interest (Horowitz, 2009), and may reflect the interaction within covariates. To further effectively combine the interpretability of the multiple linear model and the flexibility of the single index model, their hybrid, partially linear single model (PLSIM), has been studied and applied for various complex data generated from biological and economic studies in the literature (Xia and Härdle, 2006; Yu and Ruppert, 2002). To the best of our knowledge, the first remarkable work on PLSIM was done by Carroll et al. (1997), who proposed a backfitting algorithm to estimate parameters of interest in a more general case; i.e., generalized PLSIM. Yu and Ruppert (2002) argued that the estimators proposed by Carroll et al. (1997) may be unstable, and suggested the penalized spline estimation procedure. Xia and Härdle (2006) applied the minimum average variance estimation (MAVE, Xia et al., 2002) to PLSIM and developed an effective algorithm. More recently, Wang et al. (2010) studied estimation in PLSIM with the additional assumptions imposed on model structure. Liang et al. (2010) proposed a profile least squares (PrLS) estimation procedure. However, when these methods are applied to deal with the case with diverging number of covariates, one may encounter some challenges. For example, MAVE may meet the sparseness problem as noted by Cui et al. (2011), and the PrLS estimation procedure is not easy to implement in high-dimensional settings because this method needs to minimize a high-dimensional nonlinear objective function. In this paper, we propose a method for estimation and variable selection in PLSIM when the dimensions of the covariates diverge. We integrate dimension reduction principle with a testing based variable selection approach.

There has been much work on the penalty based variable selection methods for semiparametric models with a diverging number of covariates. For example, Xie and Huang (2009) and Ni et al. (2009) studied variable selection for partially linear models (PLM), a special case of PLSIM, and established the selection consistency and the asymptotic normality for their estimators. They used respectively polynomial splines and smoothing splines to approximate the nonparametric function. Ravikumar et al. (2009) investigated high-dimensional nonparametric sparse additive models, developed a new class of algorithms for estimation and discussed the asymptotic properties of their estimators. Meier et al. (2009), Huang et al. (2010), and Li et al. (2012) studied

variable selection for high-dimensional nonparametric sparse additive models. Wang and Zhu (2011) derived almost necessary and sufficient conditions for the estimation consistency of parameter estimators for single-index models in "large p, small n" paradigms. See Fan and Li (2006) for a review on variable selection for high-dimensional data. Only Liang et al. (2010) carried out variable selection in the context of PLSIM using the smoothly clipped absolute deviation penalty (SCAD, Fan and Li, 2001) to simultaneously select significant covariates and estimate the corresponding parameters of interest. However, this method is limited to the case with the fixed dimension of covariates.

As an effective way to deal with the problem of "curse of dimensionality", dimension reduction techniques overcome this problem through identifying the subspace spanned by a few convex combinations of covariates, which can capture full information between response and covariates. This subspace is called central dimension reduction space (CS, Cook, 1998). The focus is therefore on the convex combinations, rather than the original covariate vector. If the convex combinations are all forms of mean regression functions, this subspace is called central mean subspace (Cook, 1998). For instance, the multiple linear model has only one convex combination of covariates to affect response. A rich list of literature includes Li (1991) for the sliced inverse regression (SIR), Cook and Weisberg (1991) for sliced average variance estimation (SAVE), Li (1992) for principal Hessian directions, Li and Wang (2007) for directional regression (DR), Wang and Xia (2008) for sliced regression, Zhu, Wang, Zhu and Ferré (2010) for discretization-expectation estimation, and Zhu, Zhu and Feng (2010) for simple cumulative slicing estimation (CUME). There has also been interest in investigating dimension reduction with a diverging number of covariates. As the first attempt in this direction, Zhu et al. (2006) revisited SIR, whereas Zhu and Zhu (2009b) suggested a weighted partial least squares method to cope with the highly correlated covariates in semiparametric regressions. It was known that these dimension reduction methods are usually unable to identify significant covariates that sometimes are of most interest, because these methods can identify only the central subspace or central mean subspace for general cases with more than one index. More recently, efforts have been made to incorporate dimension reduction into variable selection procedure. Important results of these efforts are the least squares approach for general multi-index models by Wu and Li (2011) with the SCAD penalty, the least squares formulation by Yin and Cook (2002), and coordinate-independent sparse estimation (CISE) by Chen et al. (2010), in which the authors introduced a coordinate-independent penalty to a least squares objective function formulated by Li (2007). The CISE is shown to produce sparse solution with the oracle property.

In this paper, we first formulate the PLSIM in a dimension reduction framework so that we can identify the direction of the nonzero coefficients. We then invoke the sufficient dimension reduction principle and incorporate a coordinate-independent penalty (Chen et al., 2010) to achieve a sparse dimension reduction. In theory, we justify that our method is capable to correctly identify significant covariates with probability approaching one. The selection helps us further derive asymptotically normally distributed estimators of the nonzero coefficients.

There is an interesting feature of the method that is of independent importance in dimension reduction. Note that in this model, there are two corresponding parameter vectors in the linear and single-index parts. When we formulate this model as a bi-index model in a dimension reduction framework that will be seen below, all existing dimension reduction approaches are to identify a CS (Cook, 1998) spanned by these two vectors. In other words, any basis vector in this space is a linear combination of them, and then in general, these two parameter vectors themselves cannot be identified. However, interestingly, we find that the partially linear single-index model has a particular dimension reduction framework. With it, we can identify the two parameter vectors of interest using only one basis vector in the CS rather than identifying the entire space. This is very different from all existing dimension reduction methods in the literature because for bi-index models we usually have to determine two basis vectors to identify the CS. This identification plays a key role in our procedure for variable selection.

We conduct Monte Carlo simulation experiments to examine the performance of the proposed procedures with moderate sample sizes, and compare the performance of the proposed methods based on two popular dimension reduction procedures: SIR and CUME. Our simulation results advocate our theoretical findings. The paper is organized as follows. In Section 2, we present the models and the basic framework. In Section 3, we describe the rationale of the proposed method, and present the asymptotic results including the selection and estimation consistency and the asymptotic distributions of the estimators. Simulation studies are reported in Section 4. In Section 5, we illustrate our proposed method through a real data set. All the technical proofs of the asymptotic results are postponed to the Appendix.

2. Models and dimension reduction framework

Let Y be the response variable and $(\mathbf{X}^\tau, \mathbf{Z}^\tau)^\tau$ be the vector of covariates in $R^p \times R^q$ whose relationship with Y following PLSIM can be described as

$$Y = \mathbf{X}^\tau \boldsymbol{\beta}_0 + g(\mathbf{Z}^\tau \boldsymbol{\theta}_0) + \varepsilon, \tag{2.1}$$

where $(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ is an unknown vector in $R^p \times R^q$ equipped with the Euclidean norm $\|\cdot\|_2$, ε is the error with mean zero and finite variance, and $g(\cdot)$ is an unknown univariate link function. For the sake of identifiability, we assume, without loss of generality, that $\boldsymbol{\theta}_0$ is unit and its first component is positive, i.e., the parameter space of $\boldsymbol{\theta}_0$ is $\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)^\tau, \|\boldsymbol{\theta}\|_2 = 1, \theta_1 > 0, \boldsymbol{\theta} \in R^q\}$. PLSIMs contain two important special cases. When $q = 1$, model (2.1) reduces to a partially linear model (PLM), for which there is much work in the literature, for example, Chen (1988); Engle et al. (1986); Heckman (1986), and Speckman (1988). Härdle et al. (2000) gave a comprehensive review for PLM. When $\boldsymbol{\beta}_0 = 0$, model (2.1) reduces to the single-index model. Ichimura (1993) proposed a semiparametric least squares estimation and Härdle et al. (1993) investigated the asymptotic normality of a kernel smoother based estimation. Naik and Tsai (2001) investigated the model selection. Wang and Yang (2009) proposed a regression spline based estimation method.

Write $\mathbf{T} = (\mathbf{X}^\tau, \mathbf{Z}^\tau)^\tau \in \mathbb{R}^{p_n+q_n}$. The dimensions of both $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$, say p_n and q_n respectively, may diverge with the sample size n. Note that model (2.1) can be broadly formulated as a sufficient dimension reduction (SDR) model (Zhu and Zhu, 2009a)

$$Y \perp\!\!\!\perp \mathbf{T} | \mathcal{S}^\tau \mathbf{T}, \tag{2.2}$$

with

$$\mathcal{S} = \begin{pmatrix} \boldsymbol{\beta}_0 & \mathbf{0}_{p_n \times 1} \\ \mathbf{0}_{q_n \times 1} & \boldsymbol{\theta}_0 \end{pmatrix},$$

where $\perp\!\!\!\perp$ indicates independence. That is, conditional on $\mathcal{S}^\tau \mathbf{T}$, Y and $\mathbf{T}$ are independent. $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ can be estimated with the help of SDR principle, whose major purpose is to seek a minimum CS subspace spanned by the columns of $\mathcal{S}$. So a SDR method does not provide estimators of $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$, instead two basis vectors in the subspace in general which cannot distinguish the covariates of the respective nonparametric and parametric components. Nevertheless, the two directions $\boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\|_2$ and $\boldsymbol{\theta}_0$ in our setting may be identifiable since the central subspace is two-dimensional and generated by $\mathcal{S}$. More specifically speaking, any vector in the central subspace is of form $(c_1\boldsymbol{\beta}_0^\tau, c_2\boldsymbol{\theta}_0^\tau)^\tau$. That means that the subvector consisting of the first p_n components is related only to $\boldsymbol{\beta}_0$, while the subvector consisting of the rest q_n components is related only to $\boldsymbol{\theta}_0$. Consequently, when we use a SDR method to identify the central subspace, we can use such a vector with some nonzero components in these two parts to respectively identify $\boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\|_2$ and $\boldsymbol{\theta}_0$. Moreover, such a subspace uniquely exists and contains all regression information of $Y|\mathbf{T}$ under the mild conditions (Cook, 1996a,b). Hence we proceed identifying $\boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\|_2$ and $\boldsymbol{\theta}_0$ as follows.

As shown by Li (2007), most of the commonly used SDR methods can be formulated as the following eigen-decomposition problem:

$$\Sigma^{-1/2}\mathcal{M}\Sigma^{-1/2}b = \lambda b,$$

where Σ is the covariance matrix of $\mathbf{T}$, λ is the eigenvalue and b is the associated eigenvector, $\mathcal{M}$ is a nonnegative definite method-specific symmetric kernel matrix. See Li (2007) for details on choices of $\mathcal{M}$ for various SDR methods. Let $\lambda_{\max}$ and u_0 be the largest eigenvalue and the associated eigenvector of $\Sigma^{-1/2}\mathcal{M}\Sigma^{-1/2}$. Note that if $\lambda_{\max}$ is nonzero, then $v_0 := \Sigma^{-1/2}u_0 \in \operatorname{span}(\mathcal{S})$ under some method-specific conditions on the marginal distribution of $\mathbf{T}$ such as the linearity condition (Li, 1991). This statement implies that there exists a vector $\varphi = (\varphi_1, \varphi_2)^\tau$ with φ_1 and φ_2 being nonzero such that $v_0 = \mathcal{S}\varphi$; that is, the first p_n elements of v_0 is proportional to $\boldsymbol{\beta}_0$, and the last q_n elements of v_0 to $\boldsymbol{\theta}_0$. Once v_0 is obtained, the estimates of the directions $\boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\|_2$ and $\boldsymbol{\theta}_0$ are obtained. In Appendix A.3, we discuss an identifiability assumption under which φ_1 and φ_2 can be nonzero. Hence, the first eigenvector obtained by a dimension reduction method can identify the two directions: $\boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\|_2$ and $\boldsymbol{\theta}_0$. Furthermore, selecting significant components of $\mathbf{X}$ is equivalent to identifying nonzero element of $\boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\|_2$. Thus, we achieve variable selection procedure, and

obtain estimated value of $\boldsymbol{\beta}_0$ by estimating the scalar $\|\boldsymbol{\beta}_0\|_2$ and the direction $\boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\|_2$.

Note that all the SDR methods aforementioned involve the whole original covariates $\mathbf{T}$. As a consequence, if $p_n \to \infty$ and $q_n \to \infty$, the estimated linear combination $v_0^\tau \mathbf{T}$ may be difficult to interpret and the significant covariates may be hard to identify because all insignificant covariates are also included in the estimated linear combination. To overcome this difficulty, we use the idea of CISE to penalize v_0 for obtaining a sparse estimator of v_0 as follows.

Let $\{(\mathbf{X}_i, \mathbf{Z}_i, Y_i); 1 \le i \le n\}$ be a sequence of independent and identically distributed (i.i.d.) samples from model (2.1). Denote by $\bar{\mathbf{T}}$ and Σ_n the sample mean and covariance matrix based on $(\mathbf{T}_1, \ldots, \mathbf{T}_n)$, which is defined similar to $\mathbf{T}$. Let $\tilde{u}_n$ be the following minimizer; that is,

$$\tilde{u}_n = \arg \min_{u \in \mathbb{R}^{(p_n+q_n)}} Q_n(u; \mathbf{G}_n, \Sigma_n) \quad \text{subject to} \quad u^\tau u = 1, \tag{2.3}$$

where $\mathbf{G}_n = \Sigma_n^{-1/2} \mathcal{M}_n \Sigma_n^{-1/2}$ and $Q_n(u; \mathbf{G}_n, \Sigma_n) = -u^\tau \mathbf{G}_n u + \rho_n(\Sigma_n^{-1/2} u)$ with $\rho_n(\Sigma_n^{-1/2} u) = \sum_{r=1}^{p_n+q_n} \alpha_r |[\Sigma_n^{-1/2} u]_{(r)}|$. Any dimension reduction based kernel matrix can be applied such as SIR or SAVE. In this paper, we choose $\mathcal{M}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{m}_n(Y_i) \mathbf{m}_n^\tau(Y_i)$, the sample version of the CUME based kernel matrix (Zhu, Zhu and Feng, 2010), where $\mathbf{m}_n(Y_i) = \frac{1}{n} \sum_{j=1}^n (\mathbf{T}_j - \bar{\mathbf{T}}) \mathbf{I}(Y_j \le Y_i)$, $[\Sigma_n^{-1/2} u]_{(r)}$ is the rth element of $\Sigma_n^{-1/2} u$, and $\{\alpha_r \ge 0, r = 1, \ldots, p_n + q_n\}$ are the penalty parameters. Then, the estimator of v_0 is defined as $\tilde{v}_n = \Sigma_n^{-1/2} \tilde{u}_n$.

We choose matrix $\mathcal{M}_n$ because it is easy to implement and avoids selecting other turning parameters in estimation such as the number of slices in SIR, SAVE and DR. A theoretical justification of CUME has been provided by Zhu, Zhu and Feng (2010) even when the dimensions p_n and q_n diverge with the sample size.

3. Estimation and main results

3.1. Estimation Procedure for $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$

We formulate the estimation procedure in following steps.

Step 1. Apply the CUME based kernel matrix for the CISE variable selection procedure (2.3) to obtain an estimator $\tilde{v}_n = (\tilde{v}_{n,I}^\tau, \tilde{v}_{n,II}^\tau)^\tau$, where $\tilde{v}_{n,I} = (\tilde{v}_{n,1}, \ldots, \tilde{v}_{n,p_n})^\tau$ and $\tilde{v}_{n,II} = (\tilde{v}_{n,p_n+1}, \ldots, \tilde{v}_{n,p_n+q_n})^\tau$.

Step 2. Check the first element of $\tilde{v}_{n,II}$, and define the estimator of $\boldsymbol{\theta}_0$ as $\hat{\boldsymbol{\theta}}_0 = \text{sign}(\tilde{v}_{n,p_n+1}) \tilde{v}_{n,II} / \|\tilde{v}_{n,II}\|_2$ to guarantee positiveness of the first element of $\hat{\boldsymbol{\theta}}_0$.

Step 3. Let $\hat{\Gamma}_i = \mathbf{X}_i^\tau \tilde{v}_{n,I}$, $\hat{\Lambda}_i = \mathbf{Z}_i^\tau \hat{\boldsymbol{\theta}}_0$. We then use the "synthesis" data $\{(\hat{\Gamma}_i, \hat{\Lambda}_i, Y_i); 1 \le i \le n\}$ to define an estimator $\hat{\boldsymbol{\kappa}}$ of the parameter $\boldsymbol{\kappa}$ in the following partially linear model:

$$Y_i \approx \boldsymbol{\kappa} \hat{\Gamma}_i + g(\hat{\Lambda}_i) + \varepsilon_i.$$

Step 4. Define an estimator of $\boldsymbol{\beta}_0$ as $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\kappa}} \tilde{v}_{n,I}$.

In Step 1, one may consider other SDR methods, such as SIR, SAVE or DR. See Zhu, Zhu and Feng (2010) for a discussion on advantages and disadvantages of these SDR methods. In Step 3, one can estimate the parameter $\boldsymbol{\kappa}$ with the commonly used partially linear model techniques such as the kernel method (Liang et al., 1999, 2004; Speckman, 1988) or spline method (Chen, 1988; Cuzick, 1992; Wahba, 1984). It is remarkable that the proposed procedure does not need any iteration, neither initial value. In contrast, spline method (Yu and Ruppert, 2002) and MAVE (Xia and Härdle, 2006) need to delicately choose initial values or iteration. Thus the proposed method is particularly computationally efficient compared to its competitors. The gain is substantial when p_n and q_n diverge. The proposed procedure still has appealing asymptotic properties (see Sections 3.2-3.4 for details). Moreover, our numerical studies suggest the good performance of our method.

It is noteworthy that if we study only estimation for model (2.1), we can still use the dimension reduction principle to obtain the estimator $\tilde{v}_n = (\tilde{v}_{n,I}^\tau, \tilde{v}_{n,II}^\tau)^\tau$, Steps 2 and 3 to obtain the estimators of $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$, which are consistent and asymptotically normal under mild conditions. This estimation method is of an independent interest in dimension reduction area, and provides an alternative way different from MAVE (Xia and Härdle, 2006) or profile likelihood based (Liang et al., 2010) methods, which need iteration for implementation.

Without loss of generality, denote $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^\tau, \boldsymbol{\beta}_{20}^\tau)^\tau$, $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{10}^\tau, \boldsymbol{\theta}_{20}^\tau)^\tau$, where $\boldsymbol{\beta}_{10}$ and $\boldsymbol{\theta}_{10}$ are p_0 and q_0 nonzero components of $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$, respectively, and $\boldsymbol{\beta}_{20}$ and $\boldsymbol{\theta}_{20}$ are two $(p_n - p_0)$- and $(q_n - q_0) \times 1$-zero vectors. Assume that p_0 and q_0 are fixed. Accordingly, $\mathbf{X}_0$ and $\mathbf{Z}_0$ are the first p_0 covariates of $\mathbf{X}$ and the first q_0 covariates of $\mathbf{Z}$. Furthermore, by a simple permutation, let the first $(p_0 + q_0)$ elements of the eigenvector v_0 correspond to the covariates $(\mathbf{X}_0^\tau, \mathbf{Z}_0^\tau)^\tau$, denoted as $v_{(0)}$. Thus, $v_0 = (v_{(0)}^\tau, v_{(1)}^\tau)^\tau$, where $v_{(0)} = (v_{(0),I}^\tau, v_{(0),II}^\tau)^\tau$ with $v_{(0),I}$ and $v_{(0),II}$ corresponding to $\mathbf{X}_0$ and $\mathbf{Z}_0$, while $v_{(1)}$ is a $(p_n + q_n - p_0 - q_0) \times 1$ zero vector. Let $\mathbf{T}_0 = (\mathbf{X}_0^\tau, \mathbf{Z}_0^\tau)^\tau$, $\mathbf{X}_0^\tau = (X_1, \ldots, X_{p_0})$, $\mathbf{Z}_0^\tau = (Z_1, \ldots, Z_{q_0})$, $\Sigma_{(0)} = \mathrm{Cov}(\mathbf{T}_0)$. Suppose $\dot{Y}$ is an independent copy of Y. Write $\mathbf{m}_{(0)}(y) = E\{\mathbf{T}_0 \mathbf{1}(Y \le y)\}$ and $\mathcal{M}_{(0)} = E\mathbf{m}_{(0)}(\dot{Y})\mathbf{m}_{(0)}^\tau(\dot{Y})$. Write $\mathbf{G}_{(0)} = \Sigma_{(0)}^{-1/2} \mathcal{M}_{(0)} \Sigma_{(0)}^{-1/2}$ and let $\lambda_1, \lambda_2, \ldots, \lambda_{p_0+q_0}$ be its eigenvalues ordered from the largest to the smallest, and $u_{(0)}^{(1)}, u_{(0)}^{(2)}, \ldots u_{(0)}^{(p_0+q_0)}$ be the corresponding eigenvectors. Theorem 1 of Zhu, Zhu and Feng (2010) shows that $\mathbf{G}_{(0)}$ has only two nonzero eigenvalues $\lambda_1 > \lambda_2 > 0$. That means $\lambda_m \equiv 0$ and $u_{(0)}^{(m)}$ are the eigenvectors corresponding to the 0 eigenvalue for $m \ge 3$.

With slight notation abuse, we redefine $\tilde{v}_n$ in the Algorithm as $\tilde{v}_n = (\tilde{v}_{n(\tilde{0})}^\tau, \tilde{v}_{n(\tilde{1})}^\tau)^\tau$, where $\tilde{v}_{n(\tilde{0})} = (\tilde{v}_{n(\tilde{0}),I}^\tau, \tilde{v}_{n(\tilde{0}),II}^\tau)^\tau$, $\tilde{v}_{n(\tilde{0}),I}$ and $\tilde{v}_{n(\tilde{0}),II}$ are the nonzero components of $\tilde{v}_{n,I}$ and $\tilde{v}_{n,II}$ respectively. Let $\mathbf{X}_I$, $\mathbf{Z}_I$ be the subset of the $\mathbf{X}$, $\mathbf{Z}$ with respect to $\tilde{v}_{n(\tilde{0}),I}$, $\tilde{v}_{n(\tilde{0}),II}$, and $\widetilde{p}_0$ and $\widetilde{q}_0$ be the lengths of $\tilde{v}_{n(\tilde{0}),I}$ and $\tilde{v}_{n(\tilde{0}),II}$ respectively. So $\widetilde{p}_0$ and $\widetilde{q}_0$ are estimates of p and q instead of constants. Analogously, define $\mathbf{T}_I = (\mathbf{X}_I^\tau, \mathbf{Z}_I^\tau)^\tau$, $\mathbf{T}_{iI} = (\mathbf{X}_{iI}^\tau, \mathbf{Z}_{iI}^\tau)^\tau$ for $i = 1, \ldots, n$.

In what follows, we write $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$ for any matrix or vector $\mathbf{A}$. $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ stand for the smallest and largest eigenvalues of $\mathbf{A}$ for any square

matrix $\mathbf{A}$. For any integer s, $\mathbf{0}_s$ and $\mathbf{I}_s$ denote the zero and identity matrices of size s.

3.2. Asymptotic property of $\tilde{v}_n$

We first present the asymptotic results for the eigenvector $\tilde{v}_n$.

Theorem 3.1. *Let $d_n = \max\{p_n, q_n\}$. Assume that Conditions (A1) and (A2) in the Appendix are satisfied, and furthermore $\sqrt{n} \max_{r \le p_0+q_0} \{\alpha_r\} \to 0$ and $d_n^3/n \to 0$, then the estimator $\tilde{v}_n$ satisfies*

$$\|\tilde{v}_n - v_0\|_2 = O_P(\sqrt{d_n^2/n}).$$

Remark 1. This theorem indicates that by properly choosing the penalty parameters $\{\alpha_r\}_{r=1}^{p_n+q_n}$, the estimator is still consistent when p_n and q_n diverge at a rate of $o(n^{1/3})$, which is the same as that in the context of variable selection for parametric (Fan and Peng, 2004) and semiparametric regressions (Zhu and Zhu, 2009a). Furthermore, we can observe that if p_n and q_n are fixed, we obtain a root-n estimator $\tilde{v}_n$. This conclusion coincides with Theorem 1 of Chen et al. (2010).

To investigate the oracle property of the estimator $\tilde{v}_n$, we define the following quantities. By replacing $\mathbf{T}_i$ with $\mathbf{T}_{iI}$, we define Σ_{nI}, $\bar{\mathbf{T}}_I$, $\mathbf{G}_{nI}$, and then $Q_{nI}(u; \mathbf{G}_{nI}, \Sigma_{nI})$, ρ_{nI}, $\mathcal{M}_{nI}$, $\mathbf{m}_{nI}(Y_i)$ in the same way as the corresponding quantities for (2.3). Write $\hat{v}_n^I = \Sigma_{nI}^{-1/2} \hat{u}_n^I$, where $\hat{u}_n^I$ is the following minimizer; i.e.,

$$\hat{u}_n^I = \arg \min_{u \in \mathbb{R}^{(\widetilde{p}_0+\widetilde{q}_0)\times 1}} Q_{nI}(u; \mathbf{G}_{nI}, \Sigma_{nI}) \quad \text{subject to} \quad u^\tau u = 1.$$

In the following theorem, we state the oracle property of $\tilde{v}_n$. Let $\mathcal{A}_n = \{j : \tilde{v}_{n,j} \neq 0\}$ and $\mathcal{A}_0 = \{1, 2, \ldots, p_0 + q_0\}$.

Theorem 3.2. *Under the conditions of Theorem 3.1, if $\sqrt{n} \max_{r > p_0+q_0} \{\alpha_r\}/d_n \to \infty$, then the estimator $\tilde{v}_n$ also satisfies*

(i) $P(\mathcal{A}_n = \mathcal{A}_0) \to 1$.
(ii) $\|\tilde{v}_{n(\tilde{0})} - \hat{v}_n^I\|_2^2 = o_P(1/n)$.

Remark 2. Theorem 3.2(i) indicates that the estimator $\tilde{v}_n$ can consistently select relevant covariates. That is, with probability approaching 1, the estimators of all zero elements of v_0 go to zero. Theorem 3.2(ii) is different from Theorem 2(ii) in Chen et al. (2010), in which the authors established the oracle property of the CISE procedure under the assumptions that the number of relevant covariates, q, is an unknown constant, while our $\widetilde{p}_0$ and $\widetilde{q}_0$ are both estimators of p_0 and q_0. Accordingly, $\tilde{v}_{n(\tilde{0})}$ and $\hat{v}_n^I$ are two estimators on the basis of the variable selection procedure. As a result, we can further use $\tilde{v}_{n(\tilde{0})}$ for estimating $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$, as required in **Steps** 2 and 3 in Section 3.1.

We further consider the asymptotic distribution of $\tilde{v}_{n(\tilde{0})}$, which is generally ignored in the literature of dimension reduction. Because $\Sigma_{(0)}$ is positive definite, it has an orthogonal decomposition such as $\Sigma_{(0)} = \mathbf{P}_{(0)}\Lambda_0\mathbf{P}_0^\tau$, where $\Lambda_0 = \operatorname{diag}\big(\lambda_1(\Sigma_{(0)}), \ldots, \lambda_{p_0+q_0}(\Sigma_{(0)})\big)$ consists of the eigenvalues of $\Sigma_{(0)}$, satisfying $\lambda_1(\Sigma_{(0)}) \geq \lambda_2(\Sigma_{(0)}) \geq \cdots \geq \lambda_{p_0+q_0}(\Sigma_{(0)}) > 0$, and the columns of $\mathbf{P}_{(0)}$ are the eigenvectors corresponding to $\Lambda_{(0)}$. Let $\mathbf{B}$ be a square matrix of size $p_0 + q_0$, whose (s,t)th element is equal to $-1/2\lambda_s^{-3/2}(\Sigma_{(0)})$ if $s = t$, and $\frac{\lambda_s^{-1/2}(\Sigma_{(0)}) - \lambda_t^{-1/2}(\Sigma_{(0)})}{\lambda_s(\Sigma_{(0)}) - \lambda_t(\Sigma_{(0)})}$ otherwise. Write

$$\aleph_{\Sigma_{(0)}^{-1/2}}(\mathbf{T}_0) = \mathbf{P}_{(0)}\Big(\mathbf{B} \odot \Big[\mathbf{P}_{(0)}^\tau\Big\{\big(\mathbf{T}_0 - E\mathbf{T}_0\big)^{\otimes 2} - E\mathbf{T}_0\big(\mathbf{T}_0 - E\mathbf{T}_0\big)^\tau\Big\}\mathbf{P}_{(0)}\Big]\Big)\,\mathbf{P}_{(0)}^\tau,$$

where $\odot$ is the Hadamard product operator. Furthermore, write

$$\begin{aligned}\aleph_{\mathcal{M}_{(0)}}(\mathbf{T}_0, Y) \;=\; & 2\Big\{E\big[\mathbf{T}_0\mathbf{1}(Y \leq \dot{Y})\mathbf{m}_{(0)}^\tau(Y)\big|\,(\mathbf{T}_0, Y)\,\big] + E\big[\mathbf{m}_{(0)}(\dot{Y})\mathbf{T}_0^\tau \\ & \mathbf{1}(Y \leq \dot{Y})\big|\,(\mathbf{T}_0, Y)\,\big] + \mathbf{m}_{(0)}(Y)\mathbf{m}_{(0)}^\tau(Y) - 3\mathcal{M}_{(0)}\Big\}.\end{aligned}$$

$$\begin{aligned}\mathbf{\Phi}_0 = \Sigma_{(0)}^{-1/2}\Big\{ & \Sigma_{(0)}^{1/2}\aleph_{\Sigma_{(0)}^{-1/2}}(\mathbf{T}_0)\,\mathcal{M}_{(0)} + \mathcal{M}_{(0)}\aleph_{\Sigma_{(0)}^{-1/2}}(\mathbf{T}_0)\,\Sigma_{(0)}^{1/2} \\ & + \aleph_{\mathcal{M}_{(0)}}(\mathbf{T}_0, Y)\Big\}\Sigma_{(0)}^{-1/2}.\end{aligned}$$

We now present the asymptotic distribution of $\tilde{v}_{n(\tilde{0})}$.

Theorem 3.3. *Under the conditions of Theorem 3.2, the estimator $\tilde{v}_{n(\tilde{0})}$ is asymptotically normally distributed with covariance matrix $\mathbf{\Omega}_0$, where*

$$\mathbf{\Omega}_0 = \operatorname{Var}\Bigg(\Sigma_{(0)}^{-1/2}\sum_{m=2}^{p_0+q_0}\frac{u_0^{(m)}u_0^{(m)\tau}\mathbf{\Phi}_0 u_0^{(m)}}{\lambda_1 - \lambda_m I(m \leq 2)} + \aleph_{\Sigma_{(0)}^{-1/2}}(\mathbf{T}_0)\,\Sigma_{(0)}^{1/2}v_{(0)}\Bigg) \tag{3.1}$$

with $I(D)$ being the indicator function of the set D.

3.3. Asymptotic distribution of the estimator of $\boldsymbol{\theta}_{10}$

Write

$$\tilde{\boldsymbol{\pi}}_1 = (\mathbf{0}_{\widetilde{p}_0}, \mathbf{I}_{\widetilde{q}_0}), \boldsymbol{\pi}_1 = (\mathbf{0}_{p_0}, \mathbf{I}_{q_0}) \;\text{ and }\; \mathbf{J}_{\theta_{10}} = \frac{1}{\|\boldsymbol{\pi}_1 v_{(0)}\|_2}\big(\mathbf{I}_{q_0} - \boldsymbol{\theta}_{10}\boldsymbol{\theta}_{10}^\tau\big).$$

Note that $\tilde{v}_{n(\tilde{0})} = (\tilde{v}_{n(\tilde{0}),I}^\tau, \tilde{v}_{n(\tilde{0}),II}^\tau)^\tau$, and $\tilde{v}_{n(\tilde{0}),II}$ is an estimator of $v_{(0),II}$. Recall that the first element of $\boldsymbol{\theta}_{10}$ is assumed to be positive. Then the estimator of $\boldsymbol{\theta}_{10}$ can be defined as

$$\hat{\boldsymbol{\theta}}_{10} = \operatorname{sign}(\tilde{v}_{n(\tilde{0}),\widetilde{p}_0+1})\frac{\tilde{v}_{n(\tilde{0}),II}}{\|\tilde{v}_{n(\tilde{0}),II}\|_2} = \operatorname{sign}(\tilde{v}_{n(\tilde{0}),\widetilde{p}_0+1})\frac{\tilde{\boldsymbol{\pi}}_1\tilde{v}_{n(\tilde{0})}}{\|\tilde{\boldsymbol{\pi}}_1\tilde{v}_{n(\tilde{0})}\|_2}.$$

Theorem 3.4. *Under the conditions of Theorem 3.2, the estimator $\hat{\boldsymbol{\theta}}_{10}$ is asymptotically normally distributed with mean $\boldsymbol{\theta}_{10}$ and variance $\mathbf{J}_{\theta_{10}}\boldsymbol{\pi}_1\mathbf{\Omega}_0\boldsymbol{\pi}_1^\tau\mathbf{J}_{\theta_{10}}$.*

3.4. Asymptotic distributions of the estimators of $\boldsymbol{\kappa}$ and $\boldsymbol{\beta}_{10}$

We first state the estimation procedure for $\boldsymbol{\kappa}$ and its asymptotic distribution. Write $r_Y(t;\varsigma) = E(Y|\varsigma^\tau \mathbf{Z}_I = t)$. $r_{\mathbf{X}_I}(t;\varsigma)$ is a $\widetilde{p}_0$-vector whose elements are $r_{X_l}(t;\varsigma)$, where $r_{X_l}(t;\varsigma) = E(X_l|\varsigma^\tau \mathbf{Z}_I = t)$ for $l \in \{j : \tilde{v}_{n(\tilde{0}),j} \neq 0, 1 \le j \le \widetilde{p}_0\}$, and the local linear estimators of these elements are respectively denoted as $\hat{r}_Y(t;\varsigma) = \hat{E}(Y|\varsigma^\tau \mathbf{Z}_I = t)$, $\hat{r}_{X_l}(t;\varsigma) = \hat{E}(X_l|\varsigma^\tau \mathbf{Z}_I = t)$, $\hat{r}_{\mathbf{X}_I}(t;\varsigma)$ is a $\widetilde{p}_0$-vector whose elements are $\hat{r}_{X_l}(t;\varsigma)$, that is,

$$\hat{r}_Y(t;\varsigma) = \frac{\sum_{i=1}^n \psi_i(t,\varsigma) Y_i}{\sum_{i=1}^n \psi_i(t,\varsigma)}, \qquad \hat{r}_{X_l}(t;\varsigma) = \frac{\sum_{i=1}^n \psi_i(t,\varsigma) X_{il}}{\sum_{i=1}^n \psi_i(t,\varsigma)}, \quad \text{for } l \in \{j : \tilde{v}_{n(\tilde{0}),j} \neq 0, 1 \le j \le \widetilde{p}_0\},$$

where $\psi_i(t,\varsigma) = K_h(\varsigma^\tau \mathbf{Z}_{iI} - t)\big[V_{n,2}(t,\varsigma) - (\varsigma^\tau \mathbf{Z}_{iI} - t)V_{n,1}(t,\varsigma)\big]$ for $i = 1,\ldots,n$, $V_{n,j}(t,\varsigma) = \sum_{i=1}^n K_h(\varsigma^\tau \mathbf{Z}_{iI} - t)(\varsigma^\tau \mathbf{Z}_{iI} - t)^j$ for $j = 1, 2$, $K_h(\cdot) = h^{-1}K(\cdot/h)$ with the kernel function $K(\cdot)$ satisfying the conditions in the Appendix, and h being a bandwidth.

In the following, denote $\breve{\mathbf{X}}_0 = \mathbf{X}_0 - E(\mathbf{X}_0|\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_0)$, $\breve{\mathbf{Z}}_0 = \mathbf{Z}_0 - E(\mathbf{Z}_0|\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_0)$, and $\Sigma_{\breve{\mathbf{X}}_0} = \mathrm{Cov}(\breve{\mathbf{X}}_0)$. Furthermore, let $\boldsymbol{\pi}_2 = (\mathbf{I}_{p_0}, \mathbf{0}_{q_0})$, $\tilde{\boldsymbol{\pi}}_2 = (\mathbf{I}_{\widetilde{p}_0}, \mathbf{0}_{\widetilde{q}_0})$, and

$$\mathbf{W} = \begin{pmatrix} \sigma_\varepsilon^2 \boldsymbol{\beta}_{10}^\tau \Sigma_{\breve{\mathbf{X}}_0} \boldsymbol{\beta}_{10}, & \mathbf{W}_{12} \\ \mathbf{W}_{12}^\tau, & \boldsymbol{\beta}_{10}^\tau \Sigma_{\breve{\mathbf{X}}_0} \boldsymbol{\pi}_2 \boldsymbol{\Omega} \boldsymbol{\pi}_2^\tau \Sigma_{\breve{\mathbf{X}}_0} \boldsymbol{\beta}_{10} \end{pmatrix} \tag{3.2}$$

with

$$\mathbf{W}_{12} = \boldsymbol{\beta}_{10}^\tau E\Bigg\{ \varepsilon \breve{\mathbf{X}}_0 \Bigg(\sum_{m=2}^{p_0+q_0} \frac{u_0^{(m)\tau} u_0^{(m)\tau} \boldsymbol{\Phi} u_0^{(m)}}{\lambda_1 - \lambda_m I(m \le 2)} \Sigma_{(0)}^{-1/2} + v_{(0)}^\tau \Sigma_{(0)}^{1/2} \aleph_{\Sigma_{(0)}^{-1/2}}(\mathbf{T}_0) \Bigg) \Bigg\} \boldsymbol{\pi}_2^\tau \Sigma_{\breve{\mathbf{X}}_0} \boldsymbol{\beta}_{10}.$$

The estimator $\hat{\boldsymbol{\kappa}}$ can be obtained through the local linear smoothing in **Step** 3; that is,

$$\hat{\boldsymbol{\kappa}} = \frac{\frac{1}{n}\sum_{i=1}^n \big\{Y_i - \hat{r}_Y(\hat{\boldsymbol{\theta}}_{10}^\tau \mathbf{Z}_{iI}; \hat{\boldsymbol{\theta}}_{10})\big\}\big\{\mathbf{X}_{iI} - \hat{r}_{\mathbf{X}_I}(\hat{\boldsymbol{\theta}}_{10}^\tau \mathbf{Z}_{iI}; \hat{\boldsymbol{\theta}}_{10})\big\}^\tau (\tilde{\boldsymbol{\pi}}_2 \tilde{v}_{n(\tilde{0})})}{\frac{1}{n}\sum_{i=1}^n \Big[\big\{\mathbf{X}_{iI} - \hat{r}_{\mathbf{X}_I}(\hat{\boldsymbol{\theta}}_{10}^\tau \mathbf{Z}_{iI}; \hat{\boldsymbol{\theta}}_{10})\big\}^\tau (\tilde{\boldsymbol{\pi}}_2 \tilde{v}_{n(\tilde{0})})\Big]^2}. \tag{3.3}$$

Theorem 3.5. *In addition to the conditions of Theorem 3.2, assume that Conditions (A3)-(A5) are satisfied. The estimator $\hat{\boldsymbol{\kappa}}$ is asymptotically normal with variance*

$$\sigma_{\hat{\boldsymbol{\kappa}}}^2 = \frac{\boldsymbol{\kappa}^2}{(\boldsymbol{\beta}_{10}^\tau \Sigma_{\breve{\mathbf{X}}_0} \boldsymbol{\beta}_{10})^2} (1, -\boldsymbol{\kappa}) \mathbf{W} (1, -\boldsymbol{\kappa})^\tau. \tag{3.4}$$

Finally we define an estimator of $\boldsymbol{\beta}_{10}$ as $\hat{\boldsymbol{\beta}}_{10} = \hat{\boldsymbol{\kappa}} \times (\tilde{v}_{n(\tilde{0}),I}) = \hat{\boldsymbol{\kappa}} \times (\tilde{\boldsymbol{\pi}}_2 \tilde{v}_{n(\tilde{0})})$, and present its asymptotic distribution in the following theorem.

Theorem 3.6. *Under the conditions of Theorem 3.5, the estimator* $\hat{\boldsymbol{\beta}}_{10}$ *is asymptotically normal with covariance matrix:*

$$\Sigma_{\beta_{10}} = \mathrm{Var}\Bigg\{ \frac{\boldsymbol{\beta}_{10}^{\tau} \breve{\mathbf{X}}_0 \varepsilon \boldsymbol{\beta}_{10}}{\boldsymbol{\beta}_{10}^{\tau} \Sigma_{\breve{\mathbf{X}}_0} \boldsymbol{\beta}_{10}} - \boldsymbol{\kappa} \left(\frac{\boldsymbol{\beta}_{10} \boldsymbol{\beta}_{10}^{\tau} \Sigma_{\breve{\mathbf{X}}_0}}{\boldsymbol{\beta}_{10}^{\tau} \Sigma_{\breve{\mathbf{X}}_0} \boldsymbol{\beta}_{10}} - \mathbf{I}_{p_0} \right) \boldsymbol{\pi}_2 \Bigg(\Sigma_{(0)}^{-1/2} \sum_{m=2}^{p_0+q_0} \frac{u_0^{(m)} u_0^{(m)\tau} \boldsymbol{\Phi}_0 u_0^{(m)}}{\lambda_1 - \lambda_m I(m \le 2)} + \aleph_{\Sigma_{(0)}^{-1/2}} (\mathbf{T}_0) \, \Sigma_{(0)}^{1/2} v_{(0)} \Bigg) \Bigg\}. \quad (3.5)$$

4. Simulation studies

In this section, we report simulation results to evaluate the performance of the proposed procedure. Two dimension reduction methods, SIR and CUME have been adopted for a comparison. The number of slices for the SIR method was chosen as 5. The experiments were repeated 500 times, each consisting of $n = 150$ samples from the following two models:

$$Y = \mathbf{X}^{\tau} \boldsymbol{\beta}_0 + \exp(\mathbf{Z}^{\tau} \boldsymbol{\theta}_0) + \varepsilon, \quad (4.1)$$

$$Y = \mathbf{X}^{\tau} \boldsymbol{\beta}_0 + 3 \sin(\mathbf{Z}^{\tau} \boldsymbol{\theta}_0) + \varepsilon. \quad (4.2)$$

The first model has a monotonic link function for the single-index part and the second link function is of high frequency. The dimensions of $\mathbf{X}$ and $\mathbf{Z}$ are $(10, 10)$, $(30, 20)$, $(20, 30)$, $(50, 30)$ and $(30, 50)$, respectively. The following three cases were considered:

- Case 1. $(\mathbf{X}^{\tau}, \mathbf{Z}^{\tau})^{\tau}$ follows normal distribution $N(0_{(p_n+q_n)\times 1}, \mathbf{I}_{p_n+q_n})$, and ε follows $N(0, 0.1^2)$;
- Case 2. $(\mathbf{X}^{\tau}, \mathbf{Z}^{\tau})^{\tau}$ follows normal distribution $N(0_{(p_n+q_n)\times 1}, \Sigma)$, where $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = 0.5^{|i-j|}$, and ε is the same as in Case 1;
- Case 3. $(\mathbf{X}^{\tau}, \mathbf{Z}^{\tau})^{\tau}$ are the same as in Case 2, while ε is generated from $N\big(0, 0.1^2 \times (|\mathbf{X}_1| + |\mathbf{Z}_1|)\big)$, correlated with $(\mathbf{X}^{\tau}, \mathbf{Z}^{\tau})^{\tau}$. Here $\mathbf{X}_1$, $\mathbf{Z}_1$ are the first elements of $\mathbf{X}$ and $\mathbf{Z}$, respectively.

To estimate the parameter $\boldsymbol{\kappa}$ in Step 3, we used the local linear smoother as mentioned in Section 3.3 to obtain nonparametric estimators $\hat{E}(Y|\hat{\Lambda})$ and $\hat{E}(X|\hat{\Lambda})$ with the Epanechnikov kernel function $K(t) = 3/4(1 - t^2)I(|t| \le 1)$. For selecting bandwidth h, the cross-validation criterion was applied (Fan and Gijbels, 1996, page 149). Following Chen et al. (2010), let

$$\alpha_r = \alpha_0 \Big| [\Sigma_n^{-1/2} \hat{u}_n]_{(r)} \Big|^{-\varpi},$$

where $\hat{u}_n = \arg \min_{u \in \mathbb{R}^{(p_n+q_n)}} (-u^{\tau} \mathbf{G}_n u)$ subject to $u^{\tau} u = 1$, and $\mathbf{G}_n$ was defined in (2.3). That is, $\hat{u}_n$ is the first eigenvector of $\mathbf{G}_n$ with respect to it largest nonzero eigenvalue, and $[\Sigma_n^{-1/2} \hat{u}_n]_{(r)}$ is the r-th component of $\Sigma_n^{-1/2} \hat{u}_n$. α_0 and

ϖ are positive tuning parameters that were selected by minimizing the following BIC-type criterion (Chen et al., 2010):

$$f(\alpha_0, \varpi) = -\tilde{v}^{\tau}_{n(\alpha_0,\varpi)} \mathcal{M}_n \tilde{v}_{n(\alpha_0,\varpi)} + \frac{\log n}{n}(N_{(\alpha_0,\varpi)} - 1), \tag{4.3}$$

where $\tilde{v}_{n(\alpha_0,\varpi)}$ denotes the estimator of v_0 through (2.3) for a given pair (α_0, ϖ), $N_{(\alpha_0,\varpi)}$ stands for the number of the non-zero elements of $\tilde{v}_{n(\alpha_0,\varpi)}$, $\log n/n$ is the BIC-type factor, and $\mathcal{M}_n$ is the sample version of either the CUME kernel matrix defined in (2.3) or the SIR kernel matrix: $\mathrm{Cov}[E\{\mathbf{T} - E(\mathbf{T}|\mathbf{Y})\}]$ when these two methods are applied. This minimization can be easily solved by a two-dimensional grid search. To simplify this minimization, Chen et al. (2010) fixed $\varpi = 0.5$ in their simulation. But our numerical experience suggests that the data-driven strategy performs better with a slight increase of computational burden.

To measure the selection and estimation accuracy, we define ω_{u,β_0}, ω_{c,β_0} and ω_{o,β_0} as the proportions of underfitted, correctly fitted and overfitted models. In the case of overfitted, the labeled "1", "2" and "≥ 3" are the proportions of models including 1, 2 and more than 2 insignificant covariates. Denote by Medse_{β_0} the median of the squared error $\|\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\|_2^2$, "$\mathrm{C}_{\beta_0}$" and "$\mathrm{IN}_{\beta_0}$" the average number of the zero coefficients that were correctly set to be zero, and the average number of the non-zero coefficients that were incorrectly set to be zero, respectively. In the same way, define the quantities ω_{u,θ_0}, ω_{c,θ_0}, ω_{o,θ_0}, $\mathrm{Medse}_{\theta_0}$, "$\mathrm{C}_{\theta_0}$", and "$\mathrm{IN}_{\theta_0}$". Tables 1-4 report the values of these quantities under various configurations when the true parameters are chosen as $\boldsymbol{\beta}_0 = (3, 1.5, 0.5, 0, \ldots, 0)^{\tau}$ and $\boldsymbol{\theta}_0 = (1/\sqrt{2}, 1/\sqrt{2}, 0, \ldots, 0)$. Overall, the SIR and CUME based procedures successfully distinguish significant and insignificant covariates in the three cases. That is, the values of "C_{β_0}" and "IN_{β_0}" are respectively close to the true values $(p_n - 3)$ and 0, and the values of "C_{θ_0}" and "IN_{θ_0}" close to the true values $(q_n - 2)$ and 0. For the linear components of $\mathbf{X}$, the proportion of which the model is correctly fitted (column ω_{c,β_0}) is above 70% in all the three cases even when the number of the covariates increases to 50. The average proportions of which the model is correctly fitted for the SIR and CUME based methods are 99.23% and 99.52%, respectively. The proportions of which the model is underfitted (column ω_{u,β_0}) and overfitted (columns under ω_{c,β_0}) are about 20% and 10%, respectively. In the overfitted case, the proportion of models including 1 insignificant covariate dominates the ones including 2 or more insignificant covariates. The latter is nearly 0% in most situations. This indicates that our method most likely selects model that is very close to the true one. Compared with the SIR-based procedure, the CUME-based procedure performs better regarding model complexity with slightly higher proportions of correctly selected model in most situations. However, it also more often inclines to underfitting. A similar but better pattern can be observed from the results for the single-index components. For instance, the proportions of correctly fitted models are all about 80%, and the values of ω_{u,θ_0}, ω_{c,θ_0}, ω_{o,θ_0} are smaller, larger, and smaller than the corresponding values of ω_{u,β_0}, ω_{c,β_0}, ω_{o,β_0}, respectively.

TABLE 1
Simulating results for model (4.1) when $\boldsymbol{\beta}_0 = (3, 1.5, 0.5, 0, \ldots, 0)^\tau$. *The performance of* $\widehat{\boldsymbol{\beta}}$. *The true* C_{β_0} *value is equal to* $(p_n - 3)$

				ω_{o,β_0} (%)				No of zeros	
(p_n, q_n)	Method	ω_{u,β_0} (%)	ω_{c,β_0} (%)	"1 (%)"	"2" (%)	"≥ 3"(%)	Medse_{β_0}	C_{β_0}	IN_{β_0}
Case 1									
(10, 10)	SIR	13.00	77.20	9.80	0.40	0.00	0.1620	6.862	0.130
(10, 10)	CUME	10.60	81.80	7.60	0.00	0.00	0.1617	6.906	0.106
(30, 20)	SIR	15.80	70.40	13.80	0.00	0.00	0.2108	26.784	0.158
(30, 20)	CUME	16.00	77.40	6.20	0.40	0.00	0.2110	26.870	0.160
(20, 30)	SIR	10.80	77.40	11.80	0.00	0.00	0.2038	16.842	0.108
(20, 30)	CUME	15.00	77.80	6.80	0.40	0.00	0.2084	16.888	0.150
(50, 30)	SIR	22.80	66.40	10.40	0.40	0.00	0.2380	46.758	0.228
(50, 30)	CUME	21.40	68.20	10.40	0.00	0.00	0.2356	46.814	0.214
(30, 50)	SIR	22.80	70.00	7.20	0.00	0.00	0.2479	26.872	0.228
(30, 50)	CUME	22.40	72.20	5.40	0.00	0.00	0.2511	26.888	0.224
Case 2									
(10, 10)	SIR	12.80	77.60	9.40	0.40	0.00	0.1719	6.866	0.126
(10, 10)	CUME	20.40	70.80	8.60	0.20	0.00	0.1955	6.860	0.204
(30, 20)	SIR	22.60	69.40	7.80	0.20	0.00	0.2330	26.840	0.226
(30, 20)	CUME	26.40	71.40	2.20	0.00	0.00	0.2504	26.934	0.264
(20, 30)	SIR	19.40	74.20	6.40	0.00	0.00	0.2241	16.884	0.194
(20, 30)	CUME	24.80	71.00	4.20	0.00	0.00	0.2540	16.928	0.248
(50, 30)	SIR	22.20	69.40	8.20	0.20	0.00	0.2444	46.840	0.222
(50, 30)	CUME	26.20	71.40	2.40	0.00	0.00	0.2576	46.948	0.262
(30, 50)	SIR	21.60	72.40	5.60	0.40	0.00	0.2260	26.884	0.216
(30, 50)	CUME	27.60	70.00	2.40	0.00	0.00	0.2547	26.954	0.276
Case 3									
(10, 10)	SIR	17.40	73.80	8.20	0.60	0.00	0.1937	6.844	0.174
(10, 10)	CUME	24.40	72.00	3.60	0.20	0.00	0.2125	6.938	0.244
(30, 20)	SIR	19.40	69.60	10.80	0.20	0.00	0.2130	26.826	0.194
(30, 20)	CUME	26.20	71.00	2.80	0.00	0.00	0.2340	26.938	0.262
(20, 30)	SIR	17.00	74.60	8.20	0.20	0.00	0.2071	16.862	0.170
(20, 30)	CUME	25.00	71.00	4.00	0.00	0.00	0.2411	16.922	0.250
(50, 30)	SIR	17.60	69.80	12.20	0.20	0.20	0.2241	46.792	0.176
(50, 30)	CUME	23.80	72.40	3.80	0.00	0.00	0.2359	46.912	0.238
(30, 50)	SIR	18.40	72.60	8.80	0.20	0.00	0.2090	26.858	0.184
(30, 50)	CUME	27.40	70.20	2.40	0.00	0.00	0.2604	26.936	0.274

It is worth mentioning that the smaller value of $\boldsymbol{\beta}_0$ increases the chance of choosing underfitted model. This may be common in variable selection procedure in that the smaller parameters are hard to detect and easily to be penalized to zero. To confirm this observation, we increased the third element of $\boldsymbol{\beta}_0$ to 1.5 but keep $\boldsymbol{\theta}_0$ the same as before. We run additional simulations and report the results for the linear components when $(p_n, q_n) = (30, 50), (50, 30)$ in Tables 5 and 6, which indicate that, for the linear components $\boldsymbol{\beta}_0$, the proportions of underfitted models and overfitted models decrease, while the proportion of correctly fitted models increases and the estimation accuracy of $\boldsymbol{\beta}_0$ also gets improved.

5. Real Data Analysis

Now we illustrate the proposed method by analyzing a real dataset from an economic growth study. The data include 59 potential covariates that describe economic, political, social, and geographical characteristics of the countries from 1960-1992. Sala-I-Martin (1997) analyzed the data using a linear regression model and found that 22 out of the 59 variables appear to be "significant". As

TABLE 2
Simulating results for model (4.1) when $\boldsymbol{\beta}_0 = (3, 1.5, 0.5, 0, \ldots, 0)^\tau$. *The performance of* $\widehat{\boldsymbol{\theta}}$. *The true* C_{θ_0} *value is equal to* $(q_n - 2)$

				ω_{o,θ_0} (%)				No of zeros	
(p_n, q_n)	Method	ω_{u,θ_0} (%)	ω_{c,θ_0} (%)	"1 (%)"	"2" (%)	"≥ 3"(%)	Medse_{θ_0}	C_{θ_0}	IN_{θ_0}
Case 1									
(10, 10)	SIR	0.20	91.00	8.60	0.20	0.00	0.0276	7.910	0.002
(10, 10)	CUME	0.00	91.00	8.80	0.20	0.00	0.0127	7.908	0.000
(30, 20)	SIR	0.00	83.20	16.60	0.20	0.00	0.0149	17.830	0.000
(30, 20)	CUME	0.20	89.00	10.60	0.20	0.00	0.0141	17.890	0.002
(20, 30)	SIR	0.20	89.80	10.00	0.00	0.00	0.0296	27.900	0.002
(20, 30)	CUME	0.00	83.20	16.60	0.20	0.00	0.0146	27.830	0.000
(50, 30)	SIR	0.40	85.20	14.00	0.40	0.00	0.0169	27.852	0.004
(50, 30)	CUME	0.60	87.20	12.00	0.20	0.00	0.0139	27.874	0.006
(30, 50)	SIR	0.40	76.00	23.00	0.60	0.00	0.0153	47.758	0.004
(30, 50)	CUME	0.20	83.40	16.00	0.40	0.00	0.0165	47.832	0.002
Case 2									
(10, 10)	SIR	0.00	93.80	5.60	0.60	0.00	0.0264	7.932	0.000
(10, 10)	CUME	0.60	93.40	5.80	0.20	0.00	0.0233	7.936	0.006
(30, 20)	SIR	0.00	91.40	8.40	0.00	0.00	0.0284	17.914	0.000
(30, 20)	CUME	2.00	93.60	4.40	0.00	0.00	0.0339	17.956	0.020
(20, 30)	SIR	0.20	90.20	9.60	0.00	0.00	0.0240	27.904	0.002
(20, 30)	CUME	1.20	94.60	4.20	0.00	0.00	0.0404	27.958	0.012
(50, 30)	SIR	0.20	90.40	9.40	0.00	0.00	0.0313	27.906	0.002
(50, 30)	CUME	1.20	96.00	2.80	0.00	0.00	0.0307	27.972	0.012
(30, 50)	SIR	0.40	86.40	13.20	0.00	0.00	0.0304	47.866	0.004
(30, 50)	CUME	1.60	92.80	5.60	0.00	0.00	0.0359	47.944	0.016
Case 3									
(10, 10)	SIR	0.02	92.20	7.60	0.00	0.00	0.0298	7.924	0.002
(10, 10)	CUME	1.80	93.40	4.80	0.00	0.00	0.0331	7.952	0.018
(30, 20)	SIR	0.00	90.80	9.20	0.00	0.00	0.0250	17.908	0.000
(30, 20)	CUME	1.60	94.00	4.40	0.00	0.00	0.0341	17.954	0.016
(20, 30)	SIR	0.40	86.60	12.60	0.40	0.00	0.0304	27.864	0.004
(20, 30)	CUME	0.40	95.40	4.20	0.00	0.00	0.0358	27.958	0.004
(50, 30)	SIR	0.80	88.60	10.40	0.20	0.00	0.0297	27.892	0.008
(50, 30)	CUME	1.60	93.00	5.20	0.20	0.00	0.0366	27.944	0.016
(30, 50)	SIR	0.60	83.00	16.20	0.20	0.00	0.0338	47.834	0.006
(30, 50)	CUME	1.60	93.60	4.60	0.20	0.00	0.0442	47.950	0.016

a consequence, he had to fit 30,856 regressions per variable or a total of nearly 2 million regressions, which poses a computational challenge. Another concern is whether the linear regression is proper, since other investigators found some nonlinear structure between the covariates and the response (economic growth gamma). As an illustrative purpose, we used model (2.1) and the proposed procedure to examine the relationship between the response variable and 17 continuous covariates, which are listed in Table 7. We first fitted the response Y and each covariate with local linear smoothing and obtained a 95% pointwise confidence band, and also fitted a linear regression. If the linear straight line was encompassed in the confidence band, we classified that covariate as a linear component, and a single index one otherwise. As a result, we suggested "h60", "abslatit", "urb60", "lforce60" as single-index components. We

TABLE 3
Simulating results for model (4.2) when $\boldsymbol{\beta}_0 = (3, 1.5, 0.5, 0, \ldots, 0)^\tau$. *The performance of* $\widehat{\boldsymbol{\beta}}$. *The true* C_{β_0} *value is equal to* $(p_n - 3)$

				ω_{o,β_0} (%)				No of zeros	
(p_n, q_n)	Method	ω_{u,β_0} (%)	ω_{c,β_0} (%)	"1 (%)"	"2" (%)	"$\geq$ 3"(%)	Medse_{β_0}	C_{β_0}	IN_{β_0}
Case 1									
(10, 10)	SIR	19.20	79.20	1.60	0.00	0.00	0.1947	6.960	0.192
(10, 10)	CUME	10.40	85.80	3.80	0.00	0.00	0.1796	6.940	0.104
(30, 20)	SIR	19.60	79.00	1.40	0.00	0.00	0.2478	26.920	0.196
(30, 20)	CUME	14.60	79.00	6.40	0.00	0.00	0.2157	26.858	0.146
(20, 30)	SIR	21.60	76.20	2.20	0.00	0.00	0.2368	16.936	0.216
(20, 30)	CUME	23.20	75.60	1.20	0.00	0.00	0.2206	16.956	0.232
(50, 30)	SIR	21.60	73.60	4.80	0.00	0.00	0.2605	46.840	0.216
(50, 30)	CUME	26.80	71.60	1.60	0.00	0.00	0.2607	46.882	0.268
(30, 50)	SIR	23.20	74.00	2.80	0.00	0.00	0.2506	26.900	0.232
(30, 50)	CUME	23.60	74.80	1.40	0.20	0.00	0.2372	26.956	0.236
Case 2									
(10, 10)	SIR	13.40	77.60	9.00	0.00	0.00	0.1598	6.874	0.134
(10, 10)	CUME	21.00	75.20	3.80	0.00	0.00	0.1921	6.928	0.210
(30, 20)	SIR	19.80	71.00	9.20	0.00	0.00	0.2237	26.830	0.198
(30, 20)	CUME	18.00	75.20	6.80	0.00	0.00	0.1995	26.894	0.180
(20, 30)	SIR	17.40	74.20	8.40	0.00	0.00	0.1987	16.878	0.174
(20, 30)	CUME	23.60	73.20	3.20	0.00	0.00	0.2134	16.942	0.236
(50, 30)	SIR	21.40	71.20	7.20	0.20	0.00	0.2171	46.844	0.214
(50, 30)	CUME	26.60	72.20	1.20	0.00	0.00	0.2507	46.918	0.266
(30, 50)	SIR	15.80	76.60	7.60	0.00	0.00	0.2115	26.862	0.158
(30, 50)	CUME	28.00	69.00	3.00	0.00	0.00	0.2384	26.946	0.280
Case 3									
(10, 10)	SIR	14.20	77.60	8.20	0.00	0.00	0.1886	6.876	0.142
(10, 10)	CUME	22.40	74.40	3.20	0.00	0.00	0.1901	6.940	0.224
(30, 20)	SIR	19.40	73.20	7.40	0.00	0.00	0.2113	26.870	0.194
(30, 20)	CUME	24.00	72.00	4.00	0.00	0.00	0.2341	26.932	0.240
(20, 30)	SIR	15.20	75.80	8.80	0.20	0.00	0.1984	16.866	0.152
(20, 30)	CUME	27.00	70.00	3.00	0.00	0.00	0.2353	16.932	0.270
(50, 30)	SIR	15.60	71.40	13.00	0.00	0.00	0.2113	46.810	0.156
(50, 30)	CUME	26.20	70.60	3.20	0.00	0.00	0.2417	46.934	0.262
(30, 50)	SIR	17.60	76.20	6.20	0.00	0.00	0.2147	26.874	0.176
(30, 50)	CUME	26.60	69.80	3.60	0.00	0.00	0.2593	26.934	0.266

then applied our procedure to estimate and select nonzero elements of $(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$. The final estimated values of $(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ and the standard errors based on 1000 bootstrap resamples are reported in Table 7, which show that the SIR-based and CUME-based procedures select out the variables X_1, X_2, X_4 and X_6, and the SIR-based procedure selects two more variable X_{11} and X_{12}. The procedures also distinguish two single-index variables: Z_2 and Z_4. We estimated the nonparametric function $g(\cdot)$ by using the estimated values $(\hat{\boldsymbol{\beta}}_{0,\text{SIR}}, \hat{\boldsymbol{\theta}}_{0,\text{SIR}})$, and $(\hat{\boldsymbol{\beta}}_{0,\text{CUME}}, \hat{\boldsymbol{\theta}}_{0,\text{CUME}})$ and show the estimated curves of $g(\cdot)$ in Figure 1, which show a similar pattern but difference in magnitude.

As a referee suggested, we use the additive model and adaptive COSSO method proposed by Lin and Zhang (2006) to select significant component. The selected covariates are listed in Table 7, from which we can see that 5-fold CV adaptive COSSO tends to select more covariates than our two dimension-reduction based methods. All the unimportant covariates identified by the adaptive COSSO are also identified as unimportant covariates by the CUME dimension reduction based method. Moreover, the leave-one-out prediction error by

TABLE 4
Simulating results for model (4.2) when $\boldsymbol{\beta}_0 = (3, 1.5, 0.5, 0, \ldots, 0)^\tau$. *The performance of* $\widehat{\boldsymbol{\theta}}$. *The true* C_{θ_0} *value is equal to* $(q_n - 2)$

				ω_{o,θ_0} (%)				No of zeros	
(p_n, q_n)	Method	ω_{u,θ_0} (%)	ω_{c,θ_0} (%)	"1 (%)"	"2" (%)	"$\geq$ 3"(%)	Medse_{θ_0}	C_{θ_0}	IN_{θ_0}
Case 1									
(10, 10)	SIR	0.00	94.40	5.60	0.00	0.00	0.0061	7.944	0.000
(10, 10)	CUME	0.00	94.20	5.80	0.00	0.00	0.0041	7.942	0.000
(30, 20)	SIR	0.00	92.60	7.40	0.00	0.00	0.0052	17.926	0.000
(30, 20)	CUME	0.00	89.60	10.40	0.00	0.00	0.0057	17.896	0.000
(20, 30)	SIR	0.00	88.60	11.40	0.00	0.00	0.0057	27.886	0.000
(20, 30)	CUME	0.00	90.40	9.60	0.00	0.00	0.0052	27.904	0.000
(50, 30)	SIR	0.00	93.80	6.20	0.00	0.00	0.0064	27.938	0.000
(50, 30)	CUME	0.00	94.40	5.60	0.00	0.00	0.0052	27.944	0.000
(30, 50)	SIR	0.00	83.00	17.00	0.00	0.00	0.0061	47.830	0.000
(30, 50)	CUME	0.00	91.20	8.80	0.00	0.00	0.0043	47.912	0.000
Case 2									
(10, 10)	SIR	0.20	93.00	6.60	0.20	0.00	0.0267	7.930	0.002
(10, 10)	CUME	1.20	93.20	5.60	0.00	0.00	0.0285	7.944	0.012
(30, 20)	SIR	1.00	90.60	8.40	0.00	0.00	0.0238	17.916	0.010
(30, 20)	CUME	0.40	91.40	8.20	0.00	0.00	0.0319	17.918	0.004
(20, 30)	SIR	1.20	87.20	11.40	0.20	0.00	0.0279	27.876	0.012
(20, 30)	CUME	0.80	93.60	5.60	0.00	0.00	0.0304	27.944	0.008
(50, 30)	SIR	0.00	91.00	9.00	0.00	0.00	0.0372	27.910	0.000
(50, 30)	CUME	1.40	94.60	4.00	0.00	0.00	0.0330	27.960	0.014
(30, 50)	SIR	0.80	87.60	11.60	0.00	0.00	0.0301	47.880	0.008
(30, 50)	CUME	0.40	94.60	5.00	0.00	0.00	0.0324	47.950	0.004
Case 3									
(10, 10)	SIR	1.00	92.00	6.60	0.40	0.00	0.0221	7.924	0.010
(10, 10)	CUME	1.00	96.20	2.80	0.00	0.00	0.0271	7.970	0.010
(30, 20)	SIR	0.40	91.60	8.00	0.00	0.00	0.0275	17.920	0.004
(30, 20)	CUME	1.60	95.20	3.20	0.00	0.00	0.0404	17.968	0.016
(20, 30)	SIR	0.20	89.00	10.60	0.20	0.00	0.0269	27.890	0.002
(20, 30)	CUME	0.60	95.20	4.00	0.20	0.00	0.0345	27.956	0.006
(50, 30)	SIR	0.00	90.20	9.60	0.20	0.00	0.0313	27.900	0.000
(50, 30)	CUME	0.60	95.40	4.00	0.00	0.00	0.0401	27.960	0.006
(30, 50)	SIR	0.20	85.60	14.20	0.00	0.00	0.0327	47.858	0.002
(30, 50)	CUME	0.60	94.40	5.00	0.00	0.00	0.0341	47.948	0.006

the adaptive COSSO procedure is 5.3597×10^{-4}, while the leave-one-out prediction error by the dimension-reduction based method are 2.0303×10^{-4} with SIR, and 1.7052×10^{-4} with CUME. Consequently, the CUME based dimension reduction procedure has the smallest prediction error.

6. Discussion

We have proposed a dimension reduction based procedure for estimation and variable selection in PLSIM when the dimensions of the covariates diverge with the sample size. The procedure naturally inherits the advantages of sufficient dimension reduction and PLM, and avoids computational complexity and limitations in the existing estimation and variable selection methods for PLSIM. However, the corresponding theory for the procedure is subject to the assumption $d_n^3/n \to 0$. The difficulty mainly comes from estimating the covariance matrix. Like most dimension reduction methods, our method is limited to continuous covariates. Further investigations for the discrete covariates would be of great value.

TABLE 5
Simulating results for model (4.1) when $\boldsymbol{\beta}_0 = (3, 1.5, 1.5, 0, \ldots, 0)^\tau$. *The performance of* $\widehat{\boldsymbol{\beta}}$. *The true* C_{β_0} *value is equal to* $(p_n - 3)$

				ω_{o,β_0} (%)				No of zeros	
(p_n, q_n)	Method	ω_{u,β_0} (%)	ω_{c,β_0} (%)	"1 (%)"	"2" (%)	"≥ 3"(%)	Medse_{β_0}	C_{β_0}	IN_{β_0}
Case 1									
(50, 30)	SIR	0.00	91.00	9.00	0.00	0.00	0.1346	46.910	0.000
(50, 30)	CUME	0.00	91.00	9.00	0.00	0.00	0.1747	46.910	0.000
(30, 50)	SIR	0.00	96.00	4.00	0.00	0.00	0.1323	26.960	0.000
(30, 50)	CUME	0.00	93.00	7.00	0.00	0.00	0.1840	26.930	0.000
Case 2									
(50, 30)	SIR	0.00	98.00	2.00	0.00	0.00	0.1550	46.980	0.000
(50, 30)	CUME	0.00	94.00	6.00	0.00	0.00	0.2098	46.940	0.000
(30, 50)	SIR	0.00	98.00	2.00	0.00	0.00	0.2115	26.980	0.000
(30, 50)	CUME	0.00	99.00	1.00	0.00	0.00	0.1991	26.990	0.000
Case 3									
(50, 30)	SIR	0.00	96.00	4.00	0.00	0.00	0.2129	46.960	0.000
(50, 30)	CUME	0.00	91.00	9.00	0.00	0.00	0.2224	46.910	0.000
(30, 50)	SIR	0.00	99.00	1.00	0.00	0.00	0.2041	26.990	0.000
(30, 50)	CUME	0.00	96.00	4.00	0.00	0.00	0.2213	26.960	0.000
				ω_{o,θ_0} (%)				No of zeros	
(p_n, q_n)	Method	ω_{u,θ_0} (%)	ω_{c,θ_0} (%)	"1 (%)"	"2" (%)	"≥ 3"(%)	Medse_{θ_0}	C_{θ_0}	IN_{θ_0}
Case 1									
(50, 30)	SIR	1.00	97.00	2.00	0.00	0.00	0.0228	27.980	0.020
(50, 30)	CUME	10.00	86.00	4.00	0.00	0.00	0.0354	27.960	0.100
(30, 50)	SIR	4.00	90.00	6.00	0.00	0.00	0.0198	47.940	0.040
(30, 50)	CUME	8.00	86.00	7.00	0.00	0.00	0.0360	47.940	0.080
Case 2									
(50, 30)	SIR	6.00	94.00	0.00	0.00	0.00	0.0336	28.000	0.100
(50, 30)	CUME	1.00	95.00	4.00	0.00	0.00	0.0496	27.960	0.010
(30, 50)	SIR	12.00	86.00	2.00	0.00	0.00	0.0347	47.980	0.200
(30, 50)	CUME	6.00	91.00	3.00	0.00	0.00	0.0449	47.970	0.060
Case 3									
(50, 30)	SIR	6.00	94.00	0.00	0.00	0.00	0.0493	28.000	0.080
(50, 30)	CUME	6.00	93.00	1.00	0.00	0.00	0.0575	27.990	0.060
(30, 50)	SIR	0.00	98.00	2.00	0.00	0.00	0.0498	47.980	0.000
(30, 50)	CUME	4.00	94.00	2.00	0.00	0.00	0.0530	47.980	0.040

Our estimation procedure needs to estimate covariance Σ or inverse covariance matrix $\Sigma^{-1/2}$. For high-dimensional settings like $p >> n$, it is always assumed that covariance matrices is sparsity; that is, many entries of the off-diagonal elements are zero and the number of nonzero off-diagonal entries grows slowly. Under the sparsity condition, regularization and thresholding procedures have been proposed to construct estimators of Σ and Σ^{-1} (Bickel and Levina, 2008a,b; Cai and Liu, 2011; Lam and Fan, 2009). However, there is little dimension reduction literature on such a setting because there are additional challenges for estimating the dimension reduction kernel matrix $\mathcal{M}$, besides estimating the covariance Σ. For example, for the SIR dimension reduction method, one needs to estimate $\mathcal{M}_{\text{sir}} = \text{Cov}\{E(X^\tau, Z^\tau)^\tau | Y)\}$ as well. But the usual assumptions like off-diagonal elements being zero may be inappropriately to impose on $\mathcal{M}_{\text{sir}}$ directly. To the best of our acknowledge, only Zhu et al. (2006) recently investigated estimation of $\mathcal{M}_{\text{sir}}$ when the dimension is divergent but smaller than the sample size. How to handle the settings like $p >> n$ needs much more efforts and warrants further study.

TABLE 6
Simulating results for model (4.2) when $\boldsymbol{\beta}_0 = (3, 1.5, 1.5, 0, \ldots, 0)^\tau$. *The performance of* $\widehat{\boldsymbol{\beta}}$. *The true* C_{β_0} *value is equal to* $(p_n - 3)$

				ω_{o,β_0} (%)				No of zeros	
(p_n, q_n)	Method	ω_{u,β_0} (%)	ω_{c,β_0} (%)	"1 (%)"	"2" (%)	"$\geq$ 3"(%)	Medse_{β_0}	C_{β_0}	IN_{β_0}
Case 1									
(50, 30)	SIR	0.00	93.00	7.00	0.00	0.00	0.1497	46.930	0.000
(50, 30)	CUME	0.20	89.40	10.40	0.00	0.00	0.1844	46.896	0.002
(30, 50)	SIR	0.00	99.00	1.00	0.00	0.00	0.1810	26.990	0.000
(30, 50)	CUME	0.40	93.40	6.00	0.00	0.00	0.1928	26.940	0.004
Case 2									
(50, 30)	SIR	0.00	98.00	2.00	0.00	0.00	0.1651	46.980	0.000
(50, 30)	CUME	1.00	95.00	4.00	0.00	0.00	0.1949	46.960	0.010
(30, 50)	SIR	0.00	99.00	1.00	0.00	0.00	0.1810	26.990	0.000
(30, 50)	CUME	0.04	95.20	4.40	0.00	0.00	0.1856	26.956	0.004
Case 3									
(50, 30)	SIR	0.00	100.00	0.00	0.00	0.00	0.1723	47.000	0.000
(50, 30)	CUME	0.00	100.00	0.00	0.00	0.00	0.2113	47.000	0.000
(30, 50)	SIR	1.00	97.00	2.00	0.00	0.00	0.2049	26.980	0.010
(30, 50)	CUME	1.00	97.00	2.00	0.00	0.00	0.1861	26.980	0.010
				ω_{o,θ_0} (%)				No of zeros	
(p_n, q_n)	Method	ω_{u,θ_0} (%)	ω_{c,θ_0} (%)	"1 (%)"	"2" (%)	"$\geq$ 3"(%)	Medse_{θ_0}	C_{θ_0}	IN_{θ_0}
Case 1									
(50, 30)	SIR	0.00	95.00	5.00	0.00	0.00	0.0102	27.950	0.000
(50, 30)	CUME	0.60	93.60	5.80	0.00	0.00	0.0097	27.942	0.006
(30, 50)	SIR	6.00	92.00	2.00	0.00	0.00	0.0117	47.980	0.080
(30, 50)	CUME	0.20	90.20	9.60	0.00	0.00	0.0083	47.904	0.002
Case 2									
(50, 30)	SIR	4.00	95.00	1.00	0.00	0.00	0.0348	27.990	0.060
(50, 30)	CUME	2.00	97.00	1.00	0.00	0.00	0.0404	27.990	0.020
(30, 50)	SIR	6.00	92.00	2.00	0.00	0.00	0.0470	47.980	0.080
(30, 50)	CUME	2.40	95.60	2.00	0.00	0.00	0.0533	47.976	0.024
Case 3									
(50, 30)	SIR	9.00	89.00	2.00	0.00	0.00	0.0343	27.980	0.100
(50, 30)	CUME	3.00	97.00	0.00	0.00	0.00	0.0514	28.000	0.040
(30, 50)	SIR	14.00	86.00	0.00	0.00	0.00	0.0516	48.000	0.190
(30, 50)	CUME	9.00	91.00	0.00	0.00	0.00	0.0527	47.990	0.100

It should be worth pointing out that the procedure developed in this paper applies for fixed p_0 and q_0. Relax this case to infinite p_0 andq_0 would enhance the applicability in real data analysis. But substantial efforts seem to need because the current method relies on the asymptotic property for estimation of Σ, which is valid only for fixed p_0 and q_0. To overcome this challenge, alternative approach may be needed and is worth further studying.

We thank a referee for raising the question on which covariates for the linear part and which for the single index part. There is little literature on linear versus nonlinear forms for additive regression models (Zhang et al., 2011). Whether their procedure works for PLSIM needs additional efforts and beyond the scope of this paper. Currently, we use a guideline as follows. The effects of all the continuous covariates are put in the single-index part and those of the discrete covariates in the linear part. If the estimation results show that some of the continuous covariate effects can be relocated in the linear part, then a new model can be fitted with those continuous covariate effects moved to the linear part. The procedure is iterated several times if needed.

TABLE 7
Results for real data analysis. The estimated values and standard errors (SE) of $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$. "AC" stands for "Adaptive COSSO"

Linear component				
Variable $\mathbf{X}$		$\hat{\boldsymbol{\beta}}_{0,\mathrm{SIR}}$ (SE, $\times 10^{-3}$)	$\hat{\boldsymbol{\beta}}_{0,\mathrm{CUME}}$ (SE, $\times 10^{-3}$)	AC
$\mathbf{X}_1$	primary school enrollment rate in 1960	0.0128 (0.2235)	0.0049 (0.1459)	$\surd$
$\mathbf{X}_2$	area index	-0.0014 (0.0238)	-0.0026 (0.0247)	$\surd$
$\mathbf{X}_3$	average rate of growth of population between 1960 and 1990	N/A	N/A	
$\mathbf{X}_4$	number of years on open economy	0.0196 (0.1975)	0.0155 (0.149)	$\surd$
$\mathbf{X}_5$	number of revolutions and coups	N/A	N/A	$\surd$
$\mathbf{X}_6$	political rights	-0.0013 (0.0286)	-0.0018 (0.0272)	$\surd$
$\mathbf{X}_7$	index of civil liberties	N/A	N/A	$\surd$
$\mathbf{X}_8$	fraction of primary exports in total exports in 1970	N/A	N/A	$\surd$
$\mathbf{X}_9$	work index in 1960	N/A	N/A	
$\mathbf{X}_{10}$	fraction Catholic	N/A	N/A	$\surd$
$\mathbf{X}_{11}$	fraction Muslim	0.0037 (0.1534)	N/A	
$\mathbf{X}_{12}$	fraction Protestant	-0.0008 (0.2156)	N/A	$\surd$
$\mathbf{X}_{13}$	fraction GDP in mining	N/A	N/A	$\surd$
Single-index component				
Variable $\mathbf{Z}$		$\hat{\boldsymbol{\theta}}_{0,\mathrm{SIR}}$ (SE)	$\hat{\boldsymbol{\theta}}_{0,\mathrm{CUME}}$ (SE)	AC
$\mathbf{Z}_1$	higher education enrollment rate in 1960	N/A	N/A	
$\mathbf{Z}_2$	absolute latitude	0.4185 (0.0081)	0.294 (0.0082)	$\surd$
$\mathbf{Z}_3$	urbanization rate (fraction in cities)	N/A	N/A	
$\mathbf{Z}_4$	lforce index in 1960	0.9082 (0.0294)	0.9558 (0.0292)	$\surd$

Appendix

In this Appendix, we state the assumptions and give the proofs of the main results.

A.1. Assumptions

The following are the regularity conditions for our asymptotic results.

(A1) $\sup_{1\le i\le p_n} EX_i^4 < C_0$, $\sup_{1\le j\le q_n} EZ_j^4 < C_1$ for some constants $C_0 > 0$, $C_1 > 0$.

(A2) $\Sigma = \mathrm{Cov}(\mathbf{T})$ is positive definite, and all of its eigenvalues are bounded between two positive $\underline{c}$ and $\bar{C}$ for all p_n and q_n.

(A3) The function $E(\mathbf{X}_0|\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_0 = \boldsymbol{\theta}_{10}^\tau z_0)$ and the density function $f_{\theta_{10}^\tau \mathbf{Z}_0}(\boldsymbol{\theta}_{10}^\tau z_0)$ of the random variable $\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_0$ are both three times continuously differentiable with respect to z_0. The third derivatives are uniformly Lipschitz continuous on $\mathcal{T}_{\theta_{10}} = \{\boldsymbol{\theta}^\tau z_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}, z_0 \in \mathcal{Z}_0 \subset \mathbb{R}^{p_0}\}$, where $\mathcal{Z}_0$ is a compact support set. Furthermore, the density function $f_{\theta_{10}^\tau \mathbf{Z}_0}(\boldsymbol{\theta}_{10}^\tau z)$ is bounded away from 0 on $\mathcal{T}_{\theta_{10}}$.

(A4) The kernel function $K(\cdot)$ is a bounded, continuous and symmetric probability density function satisfying $\int_{-\infty}^{\infty} |u|^j K(u)du < \infty$ for $j = 1, 2$, and $\int_{-\infty}^{\infty} u^2 K(u)du \neq 0$. Moreover, $K(\cdot)$ satisfies a Lipschitz condition on $\mathbb{R}^1$.

(A5) The bandwidth h satisfies $h \to 0$, and $\log^2 n/nh^2 \to 0$, and $nh^3 \to \infty$.

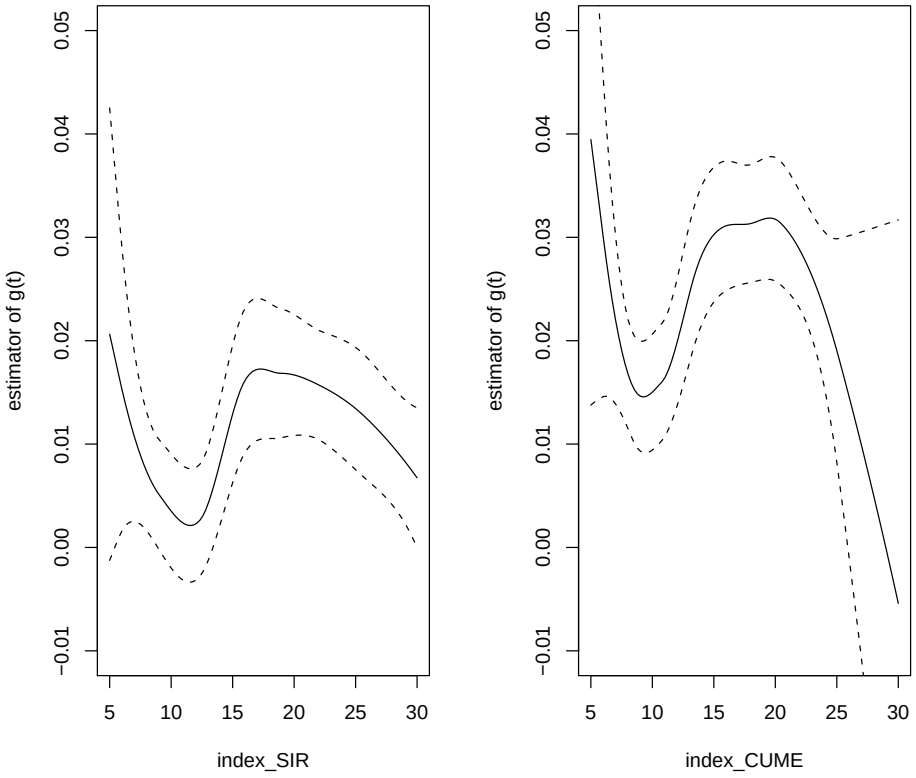

FIG 1. *The estimated curves (solid lines) of the single-index function $g(\cdot)$ and the associated 95% pointwise confidence intervals (dotted lines). The left panel: the SIR dimension reduction method; the right panel: CUME dimension reduction method.*

Condition (A1) is a technical condition imposed on the moments of $\mathbf{X}$ and $\mathbf{Z}$ in the context of diverging parameters. See more detailed discussions in Zhu and Zhu (2009a); Zhu et al. (2006). Condition (A2) is imposed on the covariance matrix of $(\mathbf{X}^\tau, \mathbf{Z}^\tau)^\tau$ to avoid the ill-conditioned problem of estimator $\Sigma_n^{-1/2}$. Because $\Sigma_n^{-1/2}$ is needed in the CISE estimation procedure, the full rank condition of Σ guarantees that even when the dimensions of $\boldsymbol{\beta}_0$, $\boldsymbol{\theta}_0$ diverge. Once n is large enough, the estimator $\Sigma_n^{-1/2}$ is of full rank, and Σ_n is invertible. See more details in Remark 2 of Zhu et al. (2006). Condition (A3) entails the density function $f_{\theta_{10}^\tau \mathbf{Z}_0}(\cdot)$ is positive, which implies that the denominators involved in the nonparametric estimators bounded away from 0. The three times continuously derivatives of $E(\mathbf{X}_0|\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_0 = \boldsymbol{\theta}_{10}^\tau z_0)$ and $f_{\theta_{10}^\tau \mathbf{Z}_0}(\boldsymbol{\theta}_{10}^\tau z_0)$ are standard smoothness conditions in nonparametric estimation. Condition (A4) is a standard assumption in the nonparametric regression literature. The Gaussian

and quadratic kernels satisfy this condition. Condition (A5) indicates that the "optimal" bandwidth can be routinely used.

A.2. Notations and Definitions

As Chen et al. (2010) mentioned, the CISE procedure hinges operationally on Grassmann manifold optimization. In order to prove the results of Theorems 3.1 and 3.2, we introduce some notations and definitions for ease of illustration.

Define the Stiefel manifold $St(p,d)$ as

$$St(p,d) = \{\eta \in \mathbb{R}^{p\times d} : \eta^\tau \eta = \mathbf{I}_d\}.$$

Let $\lfloor v \rfloor$ be the subspace spanned by the columns of η, then $\eta \in G_r(p,d)$, where $G_r(p,d)$ stands for the Grassmann manifold. The projection operator $L : \mathbb{R}^{p\times d} \to St(p,d)$ onto the Stiefel manifold $St(p,d)$ is defined to be

$$L(\eta) = \arg \min_{\mu \in St(p,d)} \|\eta - \mu\|_2^2.$$

The tangent space $T_\eta(p,d)$ of $\eta \in St(p,d)$ is defined by

$$T_\eta(p,d) = \left\{\gamma \in \mathbb{R}^{p\times d} : \gamma = \eta J + \eta^c K, J \in \mathbb{R}^{d\times d}, J + J^\tau = 0_{d\times d}, K \in \mathbb{R}^{(p-d)\times d}\right\},$$

where $\eta^c \in \mathbb{R}^{p\times(p-d)}$ is the complement of η satisfying $\eta^\tau \eta^c = 0$ and $(\eta^c)^\tau \eta^c = \mathbf{I}_{(p-d)}$.

Next, we define the neighborhood of $\lfloor \eta \rfloor$. For any matrix $\omega \in \mathbb{R}^{p\times d}$ and $\delta \in \mathbb{R}$, the perturbed point around η in the Stiefel manifold can be expressed by $L(\eta + \delta\omega)$, and the perturbed point around $\lfloor \eta \rfloor$ in the Grassmann manifold can be expressed by $\lfloor L(\eta + \delta\omega) \rfloor$. According to Lemma 8 of Manton (2002), w can be uniquely decomposed as $\omega = \eta J + \eta^c K + \eta D$, where $J \in \mathbb{R}^{d\times d}$ is a skew-symmetric matrix, $K \in \mathbb{R}^{(p-d)\times d}$ is an arbitrary matrix, and $D \in \mathbb{R}^{d\times d}$ is a symmetric matrix. As Chen et al. (2010) showed that, for a sufficiently small δ, $\lfloor L(\eta + \delta\omega) \rfloor = \lfloor L(\eta + \delta\eta^c K) \rfloor$, which indicates that the movement from $\lfloor \eta \rfloor$ in the near neighborhood only depends on the $\eta^c K$. In other words, it suffices to consider perturbed points like $L(\eta + \delta\vartheta)$ in the following proofs, where $\vartheta = \eta J + \eta^c K$ and $\|K\|_t = \sqrt{\operatorname{tr}(K^\tau K)} = C$ for some given constant C, $\operatorname{tr}(\cdot)$ is the trace operator.

A.3. Identifiability

In this section we provide an identifiability condition to guarantee that we have nonzero φ_1 and φ_2. To this end, let Y^0 be an independent copy of Y. Write $\mathcal{M}_X = E_{Y^0}\left[E_{X,Z,Y}\left\{(\mathbf{X} - E\mathbf{X})I(Y \le Y^0)\right\}\right]^{\otimes 2}$, $\mathcal{M}_Z = E_{Y^0}\left[E_{X,Z,Y}\left\{(\mathbf{Z} - E\mathbf{Z})I(Y \le Y^0)\right\}\right]^{\otimes 2}$ and $C_{X,Z} = E_{Y^0}\left[E_{X,Z,Y}\left\{(\mathbf{X}-E\mathbf{X})I(Y \le Y^0)\right\}E_{X,Z,Y}\left\{(\mathbf{Z}-E\mathbf{Z})^\tau I(Y \le Y^0)\right\}\right]$, where $E_{X,Z,Y}$ and E_{Y^0} stand for the expectation over $(\mathbf{X},\mathbf{Z},Y)$ and Y^0, respectively. Let $\Sigma_{X,Z} = \operatorname{Cov}(\mathbf{X},\mathbf{Z})$, and Σ_X and Σ_Z be the covariance matrices of $\mathbf{X}$ and $\mathbf{Z}$, respectively. Then we have the following proposition.

Proposition A.1. *Suppose*

$$\boldsymbol{\beta}_0^\tau C_{X,Z}\boldsymbol{\theta}_0 \neq \frac{\boldsymbol{\beta}_0^\tau \Sigma_{X,Z}\boldsymbol{\theta}_0}{\boldsymbol{\beta}_0^\tau \Sigma_X \boldsymbol{\beta}_0}\boldsymbol{\beta}_0^\tau \mathcal{M}_X \boldsymbol{\beta}_0 \quad \textit{and} \quad \boldsymbol{\beta}_0^\tau C_{X,Z}\boldsymbol{\theta}_0 \neq \frac{\boldsymbol{\beta}_0^\tau \Sigma_{X,Z}\boldsymbol{\theta}_0}{\boldsymbol{\theta}_0^\tau \Sigma_Z \boldsymbol{\theta}_0}\boldsymbol{\theta}_0^\tau \mathcal{M}_Z \boldsymbol{\theta}_0. \tag{A.1}$$

Then $\varphi_1 \neq 0$ *and* $\varphi_2 \neq 0$.

Remark 3. (A.1) actually depicts the correlation relationship between $\boldsymbol{\beta}_0^\tau \mathbf{X} I(Y \leq Y^0)$ and $\boldsymbol{\theta}_0^\tau \mathbf{Z} I(Y \leq Y^0)$. The proposition means that if we hope the eigenvector corresponding to the largest eigenvalue $\lambda_{\max}$ to recover two directions $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$, then $\boldsymbol{\beta}_0^\tau \mathbf{X} I(Y \leq Y^0)$ and $\boldsymbol{\theta}_0^\tau \mathbf{Z} I(Y \leq Y^0)$ cannot be linearly related to each other. This requirement is reasonable.

Proof of Proposition A.1. Note that the CUME based kernel matrix (Zhu, Zhu and Feng, 2010) can be expressed as

$$\mathcal{M} = E_{Y^0}\left[E_{X,Z,Y}\left\{\mathbf{T} I(Y \leq Y^0)\right\}\right]^{\otimes 2} = \begin{pmatrix} \mathcal{M}_{\mathbf{X}} & C_{X,Z} \\ C_{X,Z}^\tau & \mathcal{M}_Z \end{pmatrix}.$$

Recall $\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{X,Z} \\ \Sigma_{X,Z}^\tau & \Sigma_Z \end{pmatrix}$, $v_0 = (\boldsymbol{\beta}_0^\tau \varphi_1, \boldsymbol{\theta}_0^\tau \varphi_2)^\tau$ and $\mathcal{M} v_0 = \lambda_{\max} \Sigma v_0$. We have the following two equations:

$$\mathcal{M}_X \boldsymbol{\beta}_0 \varphi_1 + C_{X,Z}\boldsymbol{\theta}_0 \varphi_2 = \lambda_{\max} \Sigma_X \boldsymbol{\beta}_0 \varphi_1 + \lambda_{\max} \Sigma_{X,Z}\boldsymbol{\theta}_0 \varphi_2, \tag{A.2}$$

$$C_{X,Z}^\tau \boldsymbol{\beta}_0 \varphi_1 + \mathcal{M}_Z \boldsymbol{\theta}_0 \varphi_2 = \lambda_{\max} \Sigma_{X,Z}^\tau \boldsymbol{\beta}_0 \varphi_1 + \lambda_{\max} \Sigma_Z \boldsymbol{\theta}_0 \varphi_2. \tag{A.3}$$

Note that $v_0^\tau \Sigma v_0 = 1$, then at least one of φ_1, φ_2 is non-zero. Without loss of generality, we assume $\varphi_1 \neq 0$, but $\varphi_2 = 0$. A direct simplification from (A.2) and (A.3) yields that

$$\mathcal{M}_X \boldsymbol{\beta}_0 = \lambda_{\max} \Sigma_X \boldsymbol{\beta}_0 \quad \text{and} \quad C_{X,Z}^\tau \boldsymbol{\beta}_0 = \lambda_{\max} \Sigma_{X,Z}^\tau \boldsymbol{\beta}_0. \tag{A.4}$$

Multiplying $\boldsymbol{\beta}_0^\tau$ and $\boldsymbol{\theta}_0^\tau$ from the left of (A.4), respectively, we have $\lambda_{\max} = \frac{\boldsymbol{\beta}_0^\tau \mathcal{M}_X \boldsymbol{\beta}_0}{\boldsymbol{\beta}_0^\tau \Sigma_X \boldsymbol{\beta}_0}$ and further

$$\boldsymbol{\theta}_0^\tau C_{X,Z}^\tau \boldsymbol{\beta}_0 = \frac{\boldsymbol{\beta}_0^\tau \mathcal{M}_X \boldsymbol{\beta}_0}{\boldsymbol{\beta}_0^\tau \Sigma_X \boldsymbol{\beta}_0} \boldsymbol{\theta}_0^\tau \Sigma_{X,Z}^\tau \boldsymbol{\beta}_0, \tag{A.5}$$

which contradicts condition (A.1). Then it is not possible that φ_2 is zero. In the same way, we can prove that when $\varphi_2 \neq 0$, $\varphi_1 \neq 0$ too. We complete the proof. □

A.4. Proof of Theorem 3.1

Recall Σ_n is the estimator of Σ defined in Section 2. Write $\tilde{\xi}_n = \Sigma_n^{1/2} \tilde{v}_n$, $\xi_0^* = \Sigma_n^{1/2} v_0$, $\xi_0 = \Sigma^{1/2} v_0$. We finish the proof in two steps. In the first step we study

the relationship between $\|\tilde{v}_n - v_0\|_2^2$ and $\|\tilde{\xi}_n - \xi_0^*\|_2^2$. In the second step we derive the order of $\|\tilde{\xi}_n - \xi_0^*\|_2^2$.

Step A.1. From the definition of $\tilde{v}_n$ in Section 3.1, we can derive that

$$\begin{aligned}\|\tilde{v}_n - v_0\|_2^2 &= (\tilde{v}_n - v_0)^\tau(\tilde{v}_n - v_0) = (\Sigma_n^{-1/2}\tilde{\xi}_n - \Sigma^{-1/2}\xi_0)^\tau(\Sigma_n^{-1/2}\tilde{\xi}_n - \Sigma^{-1/2}\xi_0)\\ &= \tilde{\xi}_n^\tau(\Sigma_n^{-1/2} - \Sigma^{-1/2})^2\tilde{\xi}_n + \tilde{\xi}_n^\tau(\Sigma_n^{-1/2} - \Sigma^{-1/2})\Sigma^{-1/2}(\tilde{\xi}_n - \xi_0)\\ &\quad + (\tilde{\xi}_n - \xi_0)^\tau\Sigma^{-1/2}(\Sigma_n^{-1/2} - \Sigma^{-1/2})\tilde{\xi}_n + (\tilde{\xi}_n - \xi_0)^\tau\Sigma^{-1}(\tilde{\xi}_n - \xi_0)^\tau.\end{aligned}$$

For any $s \times s$ symmetric matrix A and any $s \times 1$ vector x, $x^\tau Ax \le \lambda_{\max}(A)x^\tau x$. Note that Condition (A2) indicates $\lambda_{\min}(\Sigma) > 0$. Then by Cauchy-Schwarz inequality and the equality $\tilde{\xi}_n^\tau\tilde{\xi}_n = 1$, we have

$$\begin{aligned}\|\tilde{v}_n - v_0\|_2^2 &\le \tilde{\xi}_n^\tau(\Sigma_n^{-1/2} - \Sigma^{-1/2})^{\otimes 2}\tilde{\xi}_n + 2\lambda_{\min}^{-1}(\Sigma)\big\{\tilde{\xi}_n^\tau(\Sigma_n^{-1/2} - \Sigma^{-1/2})^{\otimes 2}\\ &\qquad \tilde{\xi}_n\big\}^{1/2}\|\tilde{\xi}_n - \xi_0\|_2 + \lambda_{\min}^{-1}(\Sigma)\|\tilde{\xi}_n - \xi_0\|_2^2\\ &\le \lambda_{\max}\big\{(\Sigma_n^{-1/2} - \Sigma^{-1/2})^{\otimes 2}\big\} + 2\lambda_{\min}^{-1}(\Sigma)\lambda_{\max}^{1/2}\big\{(\Sigma_n^{-1/2} - \Sigma^{-1/2})^{\otimes 2}\big\}\\ &\qquad \|\tilde{\xi}_n - \xi_0\|_2 + \lambda_{\min}^{-1}(\Sigma)\|\tilde{\xi}_n - \xi_0\|_2^2.\end{aligned}$$

In the following, we show that $\lambda_{\max}\big\{(\Sigma_n^{-1/2} - \Sigma^{-1/2})^{\otimes 2}\big\} = O_P(d_n^2/n)$. For any symmetric matrix A and any positive semi-definite matrix B, we have the following inequality:

$$\lambda_{\min}(B)\lambda_{\max}(AA^\tau) \le \lambda_{\max}(ABA^\tau) \le \lambda_{\max}(B)\lambda_{\max}(AA^\tau).$$

Taking $A = \Sigma_n^{1/2} - \Sigma^{1/2}$, $B = (\Sigma_n^{1/2} + \Sigma^{1/2})^{\otimes 2}$, we then have

$$\lambda_{\max}\big\{(\Sigma_n^{1/2} - \Sigma^{1/2})^{\otimes 2}\big\} \le \frac{\lambda_{\max}\big\{(\Sigma_n - \Sigma)^{\otimes 2}\big\}}{\lambda_{\min}\big\{(\Sigma_n^{1/2} + \Sigma^{1/2})^{\otimes 2}\big\}} \le \lambda_{\min}^{-1}(\Sigma)\lambda_{\max}\big\{(\Sigma_n - \Sigma)^{\otimes 2}\big\}.$$

Note that $\lambda_{\max}\big\{(\Sigma_n - \Sigma)^{\otimes 2}\big\} \le \|\Sigma_n - \Sigma\|_F^2$. We know that all the elements of $(\Sigma_n - \Sigma)$ are of order $O_P(n^{-1/2})$. It follows that $\|\Sigma_n - \Sigma\|_F = E\|\Sigma_n - \Sigma\|_F + O_P(\sqrt{\mathrm{Var}(\|\Sigma_n - \Sigma\|_F)}) = O_P(\sqrt{E\|\Sigma_n - \Sigma\|_F^2}) = O_P\left(d_n^2/n\right)$. Thus, we have $\lambda_{\max}\big\{(\Sigma_n^{1/2} - \Sigma^{1/2})^{\otimes 2}\big\} = \left(d_n^2/n\right)$. Furthermore, $\lambda_{\max}\big\{(\Sigma_n^{-1/2} - \Sigma^{-1/2})^{\otimes 2}\big\} = \lambda_{\max}\big(\Sigma_n^{-1}(\Sigma_n^{1/2} - \Sigma^{1/2})^{\otimes 2}\Sigma^{-1}\big) \le \lambda_{\min}^{-1}(\Sigma_n)\lambda_{\min}^{-1}(\Sigma)\lambda_{\max}\big\{(\Sigma_n^{1/2} - \Sigma^{1/2})^{\otimes 2}\big\} = \left(d_n^2/n\right)$. These statements yield

$$\|\tilde{v}_n - v_0\|_2^2 = O_P\left(\frac{d_n^2}{n}\right) + O_P\left(\frac{d_n}{\sqrt{n}}\right)\|\tilde{\xi}_n - \xi_0\|_2 + \lambda_{\min}^{-1}(\Sigma)\|\tilde{\xi}_n - \xi_0\|_2^2.$$

Observe that

$$\begin{aligned}\|\tilde{\xi}_n - \xi_0\|_2^2 &\le 2\|\tilde{\xi}_n - \xi_0^*\|_2^2 + 2\|\xi_0^* - \xi_0\|_2^2\\ &\le 2\|\tilde{\xi}_n - \xi_0^*\|_2^2 + \lambda_{\max}\big\{(\Sigma_n^{1/2} - \Sigma^{1/2})^{\otimes 2}\big\}\\ &= 2\|\tilde{\xi}_n - \xi_0^*\|_2^2 + O_P(d_n^2/n).\end{aligned}$$

Then we have

$$\|\tilde{v}_n - v_0\|_2^2 = O_P\left(\frac{d_n^2}{n}\right) + O_P\left(\frac{d_n}{\sqrt{n}}\right)\|\tilde{\xi}_n - \xi_0^*\|_2 + 2\lambda_{\min}^{-1}(\Sigma)\|\tilde{\xi}_n - \xi_0^*\|_2^2. \quad \text{(A.6)}$$

Step A.2. In this step, we show that $\|\tilde{\xi}_n - \xi_0^*\|_2^2 = O_P(d_n^2/n)$ and then we finally obtain the result $\|\tilde{v}_n - v_0\|_2^2 = O_P(d_n^2/n)$ together with the result of (A.6). It suffices to show that, for any given small $\epsilon > 0$, there is a large constant C such that, for large enough n,

$$P\left\{\inf_{\vartheta \in T_{\xi_0^*}(p_n+q_n,1):\|K\|_2=C} Q_n\left(L\left(\xi_0^* + \frac{d_n}{\sqrt{n}}\vartheta\right); \mathbf{G}_n, \Sigma_n\right) > Q_n(\xi_0^*; \mathbf{G}_n, \Sigma_n)\right\} > 1 - \epsilon. \quad \text{(A.7)}$$

Then we conclude that there exists a local minimizer $\tilde{\xi}_n$ of $Q_n(\vartheta; \mathbf{G}_n, \Sigma_n)$, with probability approaching one, such that $\|\tilde{\xi}_n - \xi_0^*\|_2 = O_P(d_n/\sqrt{n})$.

Since $\vartheta \in T_{\xi_0^*}(p_n + q_n, 1)$, we have $\vartheta = \xi_0^{*c} K$, $K \in \mathbb{R}^{(p_n+q_n-1)}$. Thus, as $d_n^2/n \to 0$, applying Lemma 1 of Chen et al. (2010), we have

$$\begin{aligned}
&Q_n\left(L\left(\xi_0^* + \frac{d_n}{\sqrt{n}}\vartheta\right); \mathbf{G}_n, \Sigma_n\right) - Q_n(\xi_0^*; \mathbf{G}_n, \Sigma_n) \\
&\quad = \left\{\xi_0^{*\tau}\mathbf{G}_n\xi_0^* - \left(\xi_0^* + \frac{d_n}{\sqrt{n}}\vartheta - \frac{d_n^2}{2n}\xi_0^*\vartheta^\tau\vartheta + O_P\left(\frac{d_n^3}{n^3}\right)\right)^\tau \mathbf{G}_n \right. \\
&\qquad\qquad \left. \times \left(\xi_0^* + \frac{d_n}{\sqrt{n}}\vartheta - \frac{d_n^2}{2n}\xi_0^*\vartheta^\tau\vartheta + O_P\left(\frac{d_n^3}{n^{3/2}}\right)\right)\right\} \\
&\qquad + \left\{\rho_n\left(\Sigma_n^{-1/2}\left(\xi_0^* + \frac{d_n}{\sqrt{n}}\vartheta - \frac{d_n^2}{2n}\xi_0^*\vartheta^\tau\vartheta + O_P\left(\frac{d_n^3}{n^{3/2}}\right)\right)\right)\right. \\
&\qquad\qquad \left. - \rho_n\left(\Sigma_n^{-1/2}\xi_0^*\right)\right\} \\
&\quad \stackrel{\text{def}}{=} \Upsilon_{1n} + \Upsilon_{2n}.
\end{aligned}$$

We first deal with Υ_{1n}. Denote by 1_n the $(p_n + q_n)$ vector of ones.

$$\begin{aligned}
\Upsilon_{1n} = &\frac{d_n^2}{n}(\xi_0^{*\tau}\mathbf{G}_n\xi_0^*\vartheta^\tau\vartheta - \vartheta^\tau\mathbf{G}_n\vartheta) - \frac{d_n}{\sqrt{n}}(2\vartheta^\tau\mathbf{G}_n\xi_0^*) \\
&+ O_P\left(\frac{d_n^3}{n^{3/2}}(\xi_0^{*\tau}\mathbf{G}_n 1_n + \vartheta^\tau\mathbf{G}_n\xi_0^*\vartheta^\tau\vartheta)\right) \\
&+ O_P\left(\frac{d_n^4}{n^2}(\vartheta^\tau\mathbf{G}_n 1_n + \xi_0^{*\tau}\mathbf{G}_n\xi_0^*(\vartheta^\tau\vartheta)^2)\right) \\
&+ O_P\left(\frac{d_n^5}{n^{5/2}}\xi_0^{*\tau}\mathbf{G}_n 1_n\vartheta^\tau\vartheta\right).
\end{aligned}$$

Note that $\vartheta = \xi_0^{*c}K$, and $\xi_0^{*c\tau}\xi_0^{*c} = \mathbf{I}_{p_n+q_n-1}$, $K^\tau K = C^2$. Similar to $\|\Sigma_n - \Sigma\|_F^2 = O_P(d_n^2/n)$, we have $\|\mathcal{M}_n - \mathcal{M}\|_F^2 = O_P(d_n^2/n)$, and then $\lambda_{\max}\{(\mathcal{M}_n -$

$\mathcal{M})^{\otimes 2}\}=O_P(d_n^2/n)$. It follows that

$$\begin{aligned}\xi_0^{*\tau}\mathbf{G}_n\xi_0^*\vartheta^\tau\vartheta &= v_0^\tau(\mathcal{M}_n-\mathcal{M})v_0\vartheta^\tau\vartheta+v_0^\tau\mathcal{M}v_0\vartheta^\tau\vartheta\\ &= v_0^\tau(\mathcal{M}_n-\mathcal{M})v_0K^\tau K+\lambda_1K^\tau K\\ &\geq C^2\left\{\lambda_{\max}^{-1/2}(\Sigma)O_P\left(\frac{d_n}{\sqrt{n}}\right)+\lambda_1\right\}.\end{aligned}$$

Next we consider the term $\vartheta^\tau\mathbf{G}_n\vartheta$. Let $v_0,v_1,\ldots,v_{p_n+q_n-1}$ be the eigenvectors defined by: $\mathcal{M}v=\lambda\Sigma v$. v_0 and v_1 are the eigenvectors corresponding to the two largest eigenvalues λ_1 and λ_2, and $v_2,\ldots,v_{p_n+q_n-1}$ are the eigenvectors corresponding to the zero eigenvalues. Since the columns of $v^*=(\Sigma^{1/2}v_1,\ldots,\Sigma^{1/2}v_{p_n+q_n-1})$ consist of an orthogonal basis of $\mathbb{R}^{p_n+q_n-1}$, there exists a B^* such that $\xi_0^{*c}=v^*B^*$. Noting that $v^{*\tau}v^*=\xi_0^{*c\tau}\xi_0^{*c}=\mathbf{I}_{p_n+q_n-1}$, we also have $B^{*\tau}B^*=\mathbf{I}_{p_n+q_n-1}$. Thus,

$$\begin{aligned}\vartheta^\tau\mathbf{G}_n\vartheta &= K^\tau\xi_0^{*c\tau}\mathbf{G}_n\xi_0^{*c}K=K^\tau B^{*\tau}v^{*\tau}\mathbf{G}_nv^*B^*K\\ &= K^\tau B^{*\tau}v^{*\tau}(\Sigma_n^{-1/2}\mathcal{M}_n\Sigma_n^{-1/2}-\Sigma^{-1/2}\mathcal{M}\Sigma^{-1/2})v^*B^*K\\ &\qquad+K^\tau B^{*\tau}v^{*\tau}\Sigma^{-1/2}\mathcal{M}\Sigma^{-1/2}v^*B^*K\\ &\leq\{K^\tau B^{*\tau}B^*K\}^{1/2}\Big\{K^\tau B^{*\tau}v^{*\tau}\\ &\qquad\times\left(\Sigma_n^{-1/2}\mathcal{M}_n\Sigma_n^{-1/2}-\Sigma^{-1/2}\mathcal{M}\Sigma^{-1/2}\right)^{\otimes 2}v^*B^*K\Big\}^{1/2}\\ &\quad+K^\tau B^{*\tau}B^*K\lambda_{\max}\left(v^{*\tau}\Sigma^{-1/2}\mathcal{M}\Sigma^{-1/2}v^*\right)=CO_P\left(\frac{d_n}{\sqrt{n}}\right)+C^2\lambda_2.\end{aligned}$$

Note that $v^{*\tau}\Sigma^{1/2}v_0=0$, then

$$\begin{aligned}\vartheta^\tau\mathbf{G}_n\xi_0^* &= K^\tau\xi_0^{*c\tau}\Sigma_n^{-1/2}(\mathcal{M}_n-\mathcal{M})v_0+K^\tau\xi_0^{*c\tau}\Sigma_n^{-1/2}\mathcal{M}v_0\\ &\leq \{K^\tau\xi_0^{*c\tau}\xi_0^{*c}K\}^{1/2}\left\{v_0^\tau\Sigma_n^{-1/2}(\mathcal{M}_n-\mathcal{M})^{\otimes 2}\Sigma_n^{-1/2}v_0\right\}^{1/2}\\ &\qquad+\lambda_1K^\tau B^{*\tau}v^{*\tau}(\Sigma_n^{-1/2}-\Sigma^{-1/2})\Sigma v_0\\ &\leq C\Big[\lambda_{\min}^{-1/2}(\Sigma_n)\lambda_{\min}^{-1/2}(\Sigma)\lambda_{\max}^{1/2}\{(\mathcal{M}_n-\mathcal{M})^{\otimes 2}\}\Big]\\ &\qquad+C\lambda_1\lambda_{\max}^{1/2}\left((\Sigma_n^{-1/2}-\Sigma^{-1/2})^{\otimes 2}\right)\\ &= CO_P\left(\frac{d_n}{\sqrt{n}}\right).\end{aligned}$$

Furthermore,

$$\begin{aligned}\xi_0^{*\tau}\mathbf{G}_n1_n &= 1_n^\tau\Sigma_n^{-1/2}(\mathcal{M}_n-\mathcal{M})v_0+1_n^\tau\Sigma_n^{-1/2}\mathcal{M}v_0\\ &\leq \sqrt{1_n^\tau1_n}\left\{v_0^\tau(\mathcal{M}_n-\mathcal{M})\Sigma_n^{-1}(\mathcal{M}_n-\mathcal{M})v_0\right\}^{1/2}\\ &\quad+\sqrt{1_n^\tau1_n}\lambda_{\max}^{1/2}(v_0^\tau\Sigma\Sigma_n^{-1}\Sigma v_0)\\ &= O_P(d_n^{1/2}),\\ \vartheta^\tau\mathbf{G}_n1_n &= K^c\xi_0^{*c\tau}\mathbf{G}_n1_n\leq\sqrt{1_n^\tau1_n}\{\vartheta^\tau\mathbf{G}_n^{\otimes 2}\vartheta\}^{1/2}=O_P(d_n^{1/2}).\end{aligned}$$

Thus, we have

$$\begin{aligned}\Upsilon_{1n} \quad &\geq \quad \frac{d_n^2}{n}\left\{C^2(\lambda_1-\lambda_2)+CO_P(1)+(C^2-C)O_P\left(\frac{d_n}{\sqrt{n}}\right)\right\}\\ &\quad + \frac{d_n^2}{n}\left\{O_P\left(\sqrt{\frac{d_n^3}{n}}\right)+\lambda_1 O_P\left(\sqrt{\frac{d_n^5}{n^2}}\right)\right\}.\end{aligned}$$

So if $d_n^3/n \to 0$, as long as the constant C is sufficiently large, Υ_{1n} is positive because $\lambda_1 > \lambda_2$. In the following, we consider the second term Υ_{2n}. Let $e_r = (0,0,\ldots,0,1,0,\ldots,0)^\tau$ be a (p_n+q_n) vector with the rth element being 1 and 0 otherwise. Note that $e_r^\tau \Sigma_n^{-1/2}\xi_0^* = e_r^\tau v_0 = 0$ for any $r > p_0+q_0$ by the definition of v_0. Then we have

$$\begin{aligned}\Upsilon_{2n} &= \sum_{r=1}^{p_n+q_n}\left\{\alpha_r\left(\left\|e_r^\tau\Sigma_n^{-1/2}\left(\xi_0^*+\frac{d_n}{\sqrt{n}}\vartheta-\frac{d_n^2}{2n}\xi_0^*\vartheta^\tau\vartheta+O_P\left(\frac{d_n^3}{n^{3/2}}\right)\right)\right\|_2\right.\right.\\ &\qquad \left.\left.-\left\|e_r\Sigma_n^{-1/2}\xi_0^*\right\|_2\right)\right\}\\ &\geq \sum_{r=1}^{p_0+q_0}\left\{\alpha_r\left(\left\|e_r^\tau\Sigma_n^{-1/2}\left(\xi_0^*+\frac{d_n}{\sqrt{n}}\vartheta-\frac{d_n^2}{2n}\xi_0^*\vartheta^\tau\vartheta+O_P\left(\frac{d_n^3}{n^{3/2}}\right)\right)\right\|_2\right.\right.\\ &\qquad \left.\left.-\left\|e_r\Sigma_n^{-1/2}\xi_0^*\right\|_2\right)\right\}\\ &= \sum_{r=1}^{p_0+q_0}\left\{\alpha_r\mathrm{sign}(e_r^\tau v_0)\right.\\ &\qquad \left.\times\left(\frac{d_n}{\sqrt{n}}e_r^\tau\Sigma_n^{-1/2}\vartheta-\frac{d_n^2}{2n}e_r^\tau v_0\vartheta^\tau\vartheta+e_r^\tau\Sigma_n^{-1/2}1_nO_P\left(\frac{d_n^3}{n^{3/2}}\right)\right)\right\}\\ &\geq -\sum_{r=1}^{p_0+q_0}\frac{d_n}{\sqrt{n}}\left\{\alpha_r\mathrm{sign}(e_r^\tau v_0)\right.\\ &\qquad \left.\times\left(C\lambda_{\min}^{-1}(\Sigma_n)+\frac{d_n}{2\sqrt{n}}C^2+\lambda_{\min}^{-1}(\Sigma_n)O_P\left(\frac{d_n^{5/2}}{n}\right)\right)\right\}\\ &\geq -\frac{\sqrt{n}}{d_n}\max_{r\leq p_0+q_0}\{\alpha_r\}(p_0+q_0)\frac{d_n^2}{n}\\ &\qquad \times\left(C\lambda_{\min}^{-1}(\Sigma_n)+\frac{d_n}{2\sqrt{n}}C^2+\lambda_{\min}^{-1}(\Sigma_n)O_P\left(\frac{d_n^{5/2}}{n}\right)\right).\end{aligned}$$

Recall that $\sqrt{n}\max_{r\leq p_0+q_0}\{\alpha_r\}\to 0$. Then we have that $\Upsilon_{2n} = o_P(1)d_n^2/nC$. Consequently, Υ_{2n} is dominated by Υ_{1n}. This implies that the probability inequality (A.7) holds, and the proof is completed. □

A.5. Proof of Theorem 3.2

Step B.1. To prove the first conclusion, we first show that there exists an $r > p_0+q_0$ such that $\tilde{v}_{n,r} = 0$. Suppose that all $\tilde{v}_{n,r}$ are non-zero for $r > p_0+q_0$. Then, according to the proof of Theorem 2 in Chen et al. (2010), $\tilde{v}_n$ satisfies the following equation:

$$2\tilde{H}_n\mathcal{M}_n\tilde{v}_n = \tilde{H}_n\iota_n, \tag{A.8}$$

where $\tilde{H}_n = \mathbf{I}_{p_n+q_n} - \frac{\Sigma_n\tilde{v}_n\tilde{v}_n^\tau\Sigma_n}{\tilde{v}_n^\tau\Sigma_n^2\tilde{v}_n}$ and $\iota_n = (\alpha_1\text{sign}(\tilde{v}_{n,1}), \alpha_2\text{sign}(\tilde{v}_{n,2}), \ldots, \alpha_{p_n+q_n}\text{sign}(\tilde{v}_{n,p_n+q_n}))^\tau$. It follows from the expression of $\tilde{H}_n$ that $\tilde{H}_n$ is an idempotent matrix with rank p_n+q_n-1, and $\Sigma_n\tilde{v}_n$ is the eigenvector of $\tilde{H}_n$ corresponding to its eigenvalue 0. Let $(l_1, \ldots, l_{p_n+q_n-1})$ be the eigenvectors of $\tilde{H}_n$ corresponding to its eigenvalue 1. Thus, $(\Sigma_n\tilde{v}_n, l_1, \ldots, l_{p_n+q_n-1})$ is an independent basis of the space $\mathbb{R}^{p_n+q_n}$.By using this independent basis, there exist two sequences of constants $\{a_r\}_{r=1}^{p_n+q_n}$, $\{a'_r\}_{r=1}^{p_n+q_n}$ such that $M_n\tilde{v}_n = a_0\Sigma_n\tilde{v}_n + \sum_{r=1}^{p_n+q_n-1} a_r l_r$, and $\iota_n = a'_0\Sigma_n\tilde{v}_n + \sum_{r=1}^{p_n+q_n-1} a'_r l_r$. Plugging these two expressions into (A.8), we obtain that

$$2M_n\tilde{v}_n - 2a_0\Sigma_n\tilde{v}_n = \iota_n - a'_0\Sigma_n\tilde{v}_n. \tag{A.9}$$

From (A.9), we obtain that for $r > p_0 + q_0$,

$$\alpha_r\text{sign}(\tilde{v}_{n,r}) = 2e_r^\tau M_n\tilde{v}_n - (2a_0 - a'_0)e_r^\tau\Sigma_n\tilde{v}_n, \tag{A.10}$$

where e_r is a $(p_n + q_n)$ vector with 1 in the rth position and 0 elsewhere for $r > p_0 + q_0$.

We have supposed that $\tilde{v}_{n,r}$ are non-zero for all $r > p_0+q_0$, then $[\text{sign}(\tilde{v}_{n,r})]^2 = 1$. Note that $\tilde{v}_n^\tau\Sigma_n\tilde{v}_n = 1$, $\|\Sigma_n - \Sigma\|_F = O_p(d_n/\sqrt{n})$, $\|\tilde{v}_n - v_0\|_2 = O_p(d_n/\sqrt{n})$ and $\|\mathcal{M}_n - \mathcal{M}\|_F = O_p(d_n/\sqrt{n})$, from (A.10) we have

$$\begin{aligned}\sum_{r=p_0+q_0+1}^{p_n+q_n}\alpha_r^2 &= \sum_{r=p_0+q_0+1}^{p_n+q_n}\left\{2e_r\mathcal{M}_n\tilde{v}_n - (2a_0 - a'_0)e^r\Sigma_n\tilde{v}_n\right\}^2 \\ &\le \sum_{r=p_0+q_0+1}^{p_n+q_n}\left\{\lambda_1^2\lambda_{\max}(\Sigma) + O_p\left(\frac{d_n}{\sqrt{n}}\right)\right\} = O(d_n).\end{aligned} \tag{A.11}$$

Furthermore,

$$\tilde{v}_n^\tau\iota_n = \sum_{r=1}^{p_0+q_0}\alpha_r|\tilde{v}_{n,r}| + \sum_{r=p_0+q_0+1}^{p_n+q_n}\alpha_r|\tilde{v}_{n,r}| = 2\tilde{v}_n^\tau\mathcal{M}_n\tilde{v}_n - (2a_0 - a'_0)\tilde{v}_n^\tau\Sigma_n\tilde{v}_n,$$

$$\tilde{v}_n^\tau M_n\tilde{v}_n = \lambda_1 v_0^\tau\Sigma v_0 - (2a_0 - a'_0) + O_p\left(\frac{d_n}{\sqrt{n}}\right) = \lambda_1 - (2a_0 - a'_0) + O_p\left(\frac{d_n}{\sqrt{n}}\right).$$

As $\frac{\sqrt{n}}{d_n}\max_{r\le p_0+q_0}\{\alpha_r\} \to 0$, we have $\alpha_r = o\left(\frac{d_n}{\sqrt{n}}\right)$ for $r \le p_0 + q_0$. Thus, we obtain

$$\sum_{r=p_0+q_0+1}^{p_n+q_n}\alpha_r|\tilde{v}_{n,r}| = \lambda_1 - (2a_0 - a'_0) + O_p\left(\frac{d_n}{\sqrt{n}}\right). \tag{A.12}$$

By the result of Theorem 3.1 and the definition of $v_0 = (\beta_0^\tau \phi_1, \theta_0^\tau \phi_2)^\tau$, we know that $\|\tilde{v}_n - v_0\|_2^2 = O_P\left(d_n^2/n\right)$ and $\sum_{r=p_0+q_0+1}^{p_n+q_n} |\tilde{v}_{n,r}|^2 = O_P\left(d_n^2/n\right)$. It follows from (A.11) and (A.12) that

$$\left\{\lambda_1 - (2a_0 - a_0') + O_p\left(\frac{d_n}{\sqrt{n}}\right)\right\}^2 \le \left(\sum_{r=p_0+q_0+1}^{p_n+q_n} \alpha_r |\tilde{v}_{n,r}|\right)^2 \le \left(\sum_{r=p_0+q_0+1}^{p_n+q_n} \alpha_r^2\right)\left(\sum_{r=p_0+q_0+1}^{p_n+q_n} |\tilde{v}_{n,r}|^2\right) = O_P\left(\frac{d_n^3}{n}\right).$$

Together with (A.12) and that $\sqrt{n} \min_{r>p_0+q_0} \{\alpha_r\}/d_n \to \infty$, we have

$$\min_{r>p_0+q_0} \{|\tilde{v}_{n,r}|\} \le \frac{\sum_{r=p_0+q_0+1}^{p_n+q_n} \alpha_r |\tilde{v}_{n,r}|}{d_n \min_{r>p_0+q_0} \{\alpha_r\}} = o_P\left(\sqrt{\frac{1}{d_n}}\right). \tag{A.13}$$

So we have $1 = \text{sign}\big(\min_{r>p_0+q_0}\{|\tilde{v}_{n,r}|\}\big) = \text{sign}\big(o_P(\sqrt{1/d_n})\big) \xrightarrow{P} 0$, a contradiction, which means that, as $n \to \infty$, there is at least $r_0 > p_0 + q_0$ such that $\tilde{v}_{n,r_0} = 0$ with probability going to 1. We now further confirm that, for all $r > p_0 + q_0$, $\tilde{v}_{n,r} = 0$ with probability going to 1. This can be proved in a way similar to the proof of Theorem 2 in Chen et al. (2010). The details are omitted.

Step B.2. In this step, we show that $\tilde{v}_{n(\tilde{0})} = \hat{v}_n^I(1 + o_P(1/\sqrt{n}))$, where $\tilde{v}_{n(\tilde{0})}$ is the first $(\widetilde{p}_0 + \widetilde{q}_0)$ elements of $\tilde{v}_n$. $\hat{v}_n^I = \Sigma_{nI}^{-1/2} \hat{u}_n^I$, where $\hat{u}_n^I$ is the minimizer as

$$\hat{u}_n^I = \arg \min_{u \in \mathbb{R}^{(\widetilde{p}_0+\widetilde{q}_0)\times 1}} Q_{nI}(u; \mathbf{G}_{nI}, \Sigma_{nI}) \quad \text{subject to} \quad u^\tau u = 1, \tag{A.14}$$

where, $Q_{nI}(u; \mathbf{G}_{nI}, \Sigma_{nI}) = -u^\tau \mathbf{G}_{nI} u + \rho_{nI}(\Sigma_{nI}^{-1/2} u)$. Write $\tilde{\varrho} = \max_{\mathrm{r} \le \widetilde{p}_0+\widetilde{q}_0} \{\alpha_r\}$. Similar to the proof of Theorem 3.1, we first prove that, for large n and arbitrarily small $\epsilon > 0$, there exists a sufficiently large constant C such that

$$P\Big\{\inf_{\vartheta^I \in T_{\hat{u}_n^I}(\widetilde{p}_0+\widetilde{q}_0,1):\|K^I\|_2=C} Q_{nI}\left(L\left(\hat{u}_n^I + \tilde{\varrho}\vartheta^I\right); \mathbf{G}_{nI}, \Sigma_{nI}\right) > Q_{nI}\left(\hat{u}_n^I; \mathbf{G}_{nI}, \Sigma_{nI}\right)\Big\} > 1 - \epsilon. \tag{A.15}$$

Applying Lemma 1 (ii) of Chen et al. (2010), we have

$$\begin{aligned}
& Q_{nI}\left(L\left(\hat{u}_n^I + \tilde{\varrho}\vartheta^I\right); \mathbf{G}_{nI}, \Sigma_{nI}\right) - Q_{nI}(\hat{u}_n^o; \mathbf{G}_{nI}, \Sigma_{nI}) \\
= & \left\{\hat{u}_n^{I\tau} \mathbf{G}_{nI} \hat{u}_n^I - \left(\hat{u}_n^I + \tilde{\varrho}\vartheta^I - \frac{1}{2}\tilde{\varrho}^2 \hat{u}_n^I \vartheta^{I\tau}\vartheta^I + O_P(\tilde{\varrho}^3)\right)^\tau \mathbf{G}_{nI} \right. \\
& \qquad \left. \times \left(\hat{u}_n^I + \tilde{\varrho}\vartheta^I - \frac{1}{2}\tilde{\varrho}^2 \hat{u}_n^I \vartheta^{I\tau}\vartheta^I + O_P(\tilde{\varrho}^3)\right)\right\} \\
& + \left\{\rho_{nI}\left(\Sigma_{nI}^{-1/2}\left(\hat{u}_n^I + \varrho\vartheta^I - \frac{1}{2}\tilde{\varrho}^2 \hat{u}_n^I \vartheta^{I\tau}\vartheta^I + O_P(\tilde{\varrho}^3)\right)\right) - \rho_{nI}(\Sigma_{nI}^{-1/2}\hat{u}_n^I)\right\} \\
\stackrel{\text{def}}{=} & \ \Xi_{n1} + \Xi_{n2}.
\end{aligned}$$

Using the arguments similar to the proofs of Υ_{n1} and Υ_{n2} in Theorem 3.1, we have $\Xi_{n1} = C^2\tilde{\varrho}^2(\tilde{\lambda}_1 - \tilde{\lambda}_2) + o_P(\tilde{\varrho}^2)$ and $\Xi_{n2} = Co_P(\tilde{\varrho}^2)$, where $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ are the two largest eigenvalues of $\mathbf{G}_{nI}$. Note that $P(\mathcal{A}_n = \mathcal{A}_0) \to 1$ in **Step B.1**, thus, on the set $\{\mathcal{A}_n = \mathcal{A}_0\}$, we have $\tilde{\lambda}_1 \xrightarrow{P} \lambda_1$ and $\tilde{\lambda}_2 \xrightarrow{P} \lambda_2$. So $\tilde{\lambda}_1 > \tilde{\lambda}_1$ holds in probability, which implies that the probability inequality (A.15) holds for large C. On the set $\{\mathcal{A}_n = \mathcal{A}_0\}$, following the similar arguments in Theorem 2 of Chen et al. (2010), $\Sigma_{nI}^{1/2}\tilde{v}_{n(\tilde{0})}$ is also a local minimizer of $Q_{nI}(u; \mathbf{G}_{nI}, \Sigma_{nI})$. The inequality (A.15) implies that the local minimizer $\Sigma_{nI}^{1/2}\tilde{v}_{n(\tilde{0})}$ satisfies $\|\Sigma_{nI}^{1/2}\tilde{v}_{n(\tilde{0})} - \hat{u}_n^I\|_2^2 = O_P(\tilde{\varrho}^2)$. Note that on the set $\{\mathcal{A}_n = \mathcal{A}_0\}$, $\widetilde{p}_0 = p_0$, $\widetilde{q}_0 = q_0$, so $\sqrt{n}\tilde{\varrho} = \sqrt{n}\varrho = o_P(1)$ and $\|\Sigma_{nI}^{1/2}\tilde{v}_{n(\tilde{0})} - \hat{u}_n^I\|_2^2 = o_P(1/n)$. Thus, $\|\tilde{v}_{n(\tilde{0})} - \hat{v}_n^I\|_2^2 = (\Sigma_{nI}^{1/2}\tilde{v}_{n(\tilde{0})} - \Sigma_{nI}^{1/2}\hat{v}_n^I)^\tau\Sigma_{nI}^{-1}(\Sigma_{nI}^{1/2}\tilde{v}_{n(\tilde{0})} - \Sigma_{nI}^{1/2}\hat{v}_n^I) \le \lambda_{\min}^{-1}(\Sigma_{nI})\|\Sigma_{nI}^{1/2}\tilde{v}_{n(\tilde{0})} - \hat{u}_n^I\|_2^2 = o_P(1/n)$. We complete the proof. □

A.6. Proof of Theorem 3.3

Define $\mathbf{T}_{i0} = (\mathbf{X}_{i0}^\tau, \mathbf{Z}_{i0}^\tau)^\tau$ with $\mathbf{X}_{i0}^\tau = (X_{i1}, \ldots, X_{ip_0})$, and $\mathbf{Z}_{i0}^\tau = (Z_{i1}, \ldots, Z_{iq_0})$ for $i = 1, \ldots, n$. By replacing $\mathbf{T}_i$ by $\mathbf{T}_{i0}$, we define $\Sigma_{n(0)}$, $\bar{\mathbf{T}}_0$, $\mathbf{G}_{n(0)}$, $\rho_{n(0)}$ and $\mathcal{M}_{n(0)}$, $\mathbf{m}_{n(0)}(Y_i)$ in the same way as the corresponding quantities for (2.3), and denote $\hat{u}_n^{(1)}$ as the first eigenvector of $\mathbf{G}_{n(0)}$ corresponding to its largest eigenvalue. Write $\mathbf{G}_{(0)} = (g_{ij})$ for $1 \le i, j \le (p_0+q_0)$. The standard perturbation theory gives the following chain-rule formulas (Zhu and Fang, 1996):

$$\frac{\partial u_{(0)}}{\partial g_{ij}} = \sum_{m=2}^{p_0+q_0} \frac{u_{(0)}^{(m)} u_{(0)}^{(m)\tau} \left(\partial \mathbf{G}_{(0)}/\partial g_{ij}\right) u_{(0)}^{(m)}}{\lambda_1 - \lambda_m}.$$

Recall that the eigenvalues of $\mathbf{G}_{(0)}$ satisfy $\lambda_1 \ge \lambda_2 \ge \lambda_3 = \cdots = \lambda_{p_0+q_0} = 0$. Thus, by the argument similar to Zhu and Fang (1996), we have the following expression:

$$\begin{aligned} &\sqrt{n}\left(\hat{u}_n^{(1)} - u_{(0)}\right) \qquad (A.16)\\ = \;& \sqrt{n}\sum_{m=2}^{p_0+q_0} \frac{u_{(0)}^{(m)} u_{(0)}^{(m)\tau}\left(\mathbf{G}_{n(0)} - \mathbf{G}_{(0)}\right) u_0^{(m)}}{\lambda_1 - \lambda_m} + o_P\left(\sqrt{n}\|\mathbf{G}_{n(0)} - \mathbf{G}_{(0)}\|_{L_1}\right), \end{aligned}$$

where $\|A\|_{L_1} = \sum_{1\le s,t,\le k} |a_{st}|$ for any $k \times k$ matrix $A = (a_{st})_{1\le s,t,\le k}$.

We now establish an asymptotic expansion of $\mathbf{G}_{n(0)} - \mathbf{G}_{(0)}$. Write $\mathbf{C}_{n1} = (\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2})\mathcal{M}_{(0)}\Sigma_{(0)}^{-1/2}$, $\mathbf{C}_{n2} = \Sigma_{(0)}^{-1/2}(\mathcal{M}_{n(0)} - \mathcal{M}_{(0)})\Sigma_{(0)}^{-1/2}$, and $\mathbf{C}_{n3} = \Sigma_{(0)}^{-1/2}(\mathcal{M}_{n(0)} - \mathcal{M}_{(0)})(\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2}) + (\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2})(\mathcal{M}_{n(0)} - \mathcal{M}_{(0)})\Sigma_{(0)}^{-1/2} + (\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2})(\mathcal{M}_{n(0)} - \mathcal{M}_{(0)})(\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2}) + (\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2})\mathcal{M}_{(0)}(\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2})$. Then we have

$$\begin{aligned} \sqrt{n}(\mathbf{G}_{n(0)} - \mathbf{G}_{(0)}) &= \sqrt{n}(\Sigma_{n(0)}^{-1/2}\mathcal{M}_{n(0)}\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2}\mathcal{M}_{(0)}\Sigma_{(0)}^{-1/2}) \\ &= \sqrt{n}\{\mathbf{C}_{n1} + \mathbf{C}_{n1}^\tau + \mathbf{C}_{n2}\} + \sqrt{n}\mathbf{C}_{n3}. \end{aligned}$$

Applying the asymptotic expansion of $\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2}$ in Lemma 2.4 of Yu et al. (2011), we have

$$\sqrt{n}\left(\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \aleph_{\Sigma_{(0)}^{-1/2}}(\mathbf{T}_{i0}) + O_P\left(\frac{1}{\sqrt{n}}\right).$$

Thus,

$$\begin{aligned}\sqrt{n}(\mathbf{C}_{n1} + \mathbf{C}_{n1}^{\tau}) &= \sqrt{n}\Big\{(\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2})\mathcal{M}_{(0)}\Sigma_{(0)}^{-1/2} \\ &\qquad + \Sigma_{(0)}^{-1/2}\mathcal{M}_{(0)}(\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2})\Big\} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\aleph_{\Sigma_{(0)}^{-1/2}}(\mathbf{T}_{i0})\mathcal{M}_{(0)}\Sigma_{(0)}^{-1/2} \\ &\qquad + \Sigma_{(0)}^{-1/2}\mathcal{M}_{(0)}\aleph_{\Sigma_{(0)}^{-1/2}}(\mathbf{T}_{i0})\Big\} + o_P(1).\end{aligned}$$

It follows from (A.15) in Zhu, Zhu and Feng (2010) that

$$\sqrt{n}\left(\mathcal{M}_{n(0)} - \mathcal{M}_{(0)}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\aleph_{\mathcal{M}}(\mathbf{T}_{i0}, Y_i) + O_P\left(\frac{1}{\sqrt{n}}\right).$$

So

$$\sqrt{n}\mathbf{C}_{n2} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Sigma_{(0)}^{-1/2}\aleph_{\mathcal{M}}(\mathbf{T}_{i0}, Y_i)\,\Sigma_{(0)}^{-1/2} + o_P(1).$$

From the asymptotic expansions of $\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2}$ and $\mathcal{M}_{n(0)} - \mathcal{M}_{(0)}$, we conclude that $\sqrt{n}\mathbf{C}_{n3} = o_P(1)$. As a consequence, we have

$$\begin{aligned}\sqrt{n}\left(\mathbf{G}_{n(0)} - \mathbf{G}_{(0)}\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Sigma_{(0)}^{-1/2}\Big\{\Sigma_{(0)}^{1/2}\aleph_{\Sigma_{(0)}^{-1/2}}(\mathbf{T}_{i0})\,\mathcal{M}_{(0)} \\ &\qquad + \mathcal{M}_{(0)}\aleph_{\Sigma_{(0)}^{-1/2}}(\mathbf{T}_{i0})\,\Sigma_{(0)}^{1/2} + \aleph_{\mathcal{M}}(\mathbf{T}_{i0}, Y_i)\Big\}\Sigma_{(0)}^{-1/2} + o_P(1) \\ &\stackrel{\text{def}}{=} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{\Phi}_{i0} + o_P(1).\end{aligned}$$

Note that on the set $\{\mathcal{A}_n = \mathcal{A}_0\}$, $\mathbf{G}_{n(0)} = \mathbf{G}_{nI}$, $\Sigma_{n(0)} = \Sigma_{nI}$, $\mathcal{M}_{n(0)} = \mathcal{M}_{nI}$, $\widetilde{p}_0 = p_0$ and $\widetilde{q}_0 = q_0$. As eigenvalue decomposition is a linear operator, the condition $\sqrt{n}\min_{r\le p_0+q_0}\{\alpha_r\} \to 0$ and the perturbation theory entail that $\|\hat{u}_n^I - \hat{u}_n^{(1)}\|_2 = o_P(1/\sqrt{n})$ on the set $\{\mathcal{A}_n = \mathcal{A}_0\}$. Then, by (A.16), we have

$$\sqrt{n}\left(\hat{u}_n^I - u_{(0)}\right) = \sqrt{n}\sum_{m=2}^{p_0+q_0} \frac{u_0^{(m)} u_0^{(m)\tau}\left(\mathbf{G}_{n(0)} - \mathbf{G}_{(0)}\right) u_0^{(m)}}{\lambda_1 - \lambda_m} + o_P\left(\sqrt{n}\|\mathbf{G}_{n(0)} - \mathbf{G}_{(0)}\|_{L_1}\right)$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{m=2}^{p_0+q_0} \frac{u_0^{(m)} u_0^{(m)\tau}\mathbf{\Phi}_{i0} u_0^{(m)}}{\lambda_1 - \lambda_m} + o_P(1).$$

From Theorem 3.2 and $\lambda_m = 0$ for $m \geq 3$, we know that $\sqrt{n}(\tilde{v}_{n(\tilde{0})} - v_{(0)})$ equals $\sqrt{n}\Sigma_{n(0)}^{-1/2}(\hat{u}_n^I - u_{(0)}) + \sqrt{n}(\Sigma_{n(0)}^{-1/2} - \Sigma_{(0)}^{-1/2})\Sigma_{(0)}^{1/2} v_{(0)}$, which can be expressed as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\Sigma_{(0)}^{-1/2}\sum_{m=2}^{p_0+q_0}\frac{u_0^{(m)} u_0^{(m)\tau}\mathbf{\Phi}_{i0} u_0^{(m)}}{\lambda_1 - \lambda_m I(m \leq 2)} + \aleph_{\Sigma_{(0)}^{-1/2}}\left(\mathbf{T}_{i(0)}\right)\Sigma_{(0)}^{1/2} v_{(0)}\right\} + o_P(1). \tag{A.17}$$

Together with $P(\mathcal{A}_n = \mathcal{A}_0) \to 1$ and (A.17), $\sqrt{n}(\tilde{v}_{n(\tilde{0})} - \hat{v}_0)$ converges to $N(0_{(p_0+q_0)\times 1}, \mathbf{\Omega}_0)$ in distribution. We complete the proof. □

A.7. Proof of Theorem 3.4

Recall that $\mathbf{J}_{\theta_{10}} = 1/\|\boldsymbol{\pi}_1 v_{(0)}\|_2 \mathbf{I}_{q_0} - 1/\|\boldsymbol{\pi}_1 v_{(0)}\|_2^3 [\boldsymbol{\pi}_1 v_{(0)}]^{\otimes 2}$. On the set $\{\mathcal{A}_n = \mathcal{A}_0\}$, $\tilde{\boldsymbol{\pi}}_1 = \boldsymbol{\pi}$, $\widetilde{p}_0 = p_0$ and $\widetilde{q}_0 = q_0$ and $P(\mathcal{A}_n = \mathcal{A}_0) \to 1$. Using Delta method and expression (A.17), we have

$$\sqrt{n}\left(\frac{\tilde{\boldsymbol{\pi}}_1\tilde{v}_{n(\tilde{0})}}{\|\tilde{\boldsymbol{\pi}}_1\tilde{v}_{n(\tilde{0})}\|_2} - \frac{\boldsymbol{\pi}_1 v_{(0)}}{\|\boldsymbol{\pi}_1 v_{(0)}\|_2}\right) \xrightarrow{L} N(0_{q_0\times 1}, \mathbf{J}_{\theta_{10}}\boldsymbol{\pi}_1\mathbf{\Omega}_0\boldsymbol{\pi}_1^\tau\mathbf{J}_{\theta_{10}}).$$

Note that $\boldsymbol{\pi}_1 v_{(0)} = \left(v_{(0),p_0+1}, v_{(0),p_0+2}, \ldots, v_{(0),p_0+q_0}\right)^\tau$, and this is proportional to $\boldsymbol{\theta}_{10}$. From the asymptotic expression given in (A.17), we have

$$\sqrt{n}\left(\tilde{v}_{n(\tilde{0}),\widetilde{p}_0+1} - v_{(0),p_0+1}\right) \xrightarrow{L} N(0, \sigma_{p_0+1}^2)$$

for some $\sigma_{p_0+1}^2$. Since $v_{(0),p_0+1}$ is non-zero, and when $w \neq 0$ the sign function $\text{sign}(w)$ is a continuous function and the first derivation of $\text{sign}(w)$ is zero, it follows that

$$\sqrt{n}\left(\text{sign}(\tilde{v}_{n(\tilde{0}),\widetilde{p}_0+1}) - \text{sign}(v_{(0),p_0+1})\right) \xrightarrow{L} 0,$$

i,e, $\text{sign}(\tilde{v}_{n(\tilde{0}),\widetilde{p}_0+1}) = \text{sign}(v_{(0),p_0+1}) + o_P(1/\sqrt{n})$. Note that $\text{sign}(v_{(0),p_0+1}) \times \frac{\boldsymbol{\pi}_1 v_{(0)}}{\|\boldsymbol{\pi}_1 v_{(0)}\|_2} = \boldsymbol{\theta}_{10}$. As a consequence, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{10} - \boldsymbol{\theta}_{10}\right) = \sqrt{n}\left(\text{sign}(\tilde{v}_{n(\tilde{0}),\widetilde{p}_0+1})\tilde{\boldsymbol{\pi}}_1\tilde{v}_{n(\tilde{0})}/\|\tilde{\boldsymbol{\pi}}_1\tilde{v}_{n(\tilde{0})}\|_2 - \boldsymbol{\theta}_{10}\right),$$

which converges to $N(0_{q_0\times 1}, \mathbf{J}_{\theta_{10}}\boldsymbol{\pi}_1\mathbf{\Omega}_0\boldsymbol{\pi}_1^\tau\mathbf{J}_{\theta_{10}})$ in distribution. We complete the proof. □

A.8. Proof of Theorem 3.5

Step C.1. The numerator of $\hat{\boldsymbol{\kappa}}$ can be decomposed as:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-\hat{r}_Y(\hat{\boldsymbol{\theta}}_{10}^{\tau}\mathbf{Z}_{iI};\hat{\boldsymbol{\theta}}_{10})\right\}\left\{\mathbf{X}_{iI}-\hat{r}_{\mathbf{X}_I}(\hat{\boldsymbol{\theta}}_{10}^{\tau}\mathbf{Z}_{iI};\hat{\boldsymbol{\theta}}_{10})\right\}^{\tau}(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})})\\
&\quad=\left[\frac{1}{n}\sum_{i=1}^{n}\{Y_i-r_Y(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{iI};\boldsymbol{\theta}_{10})\}\{\mathbf{X}_{iI}-r_{\mathbf{X}_I}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{iI};\boldsymbol{\theta}_{10})\}^{\tau}(\boldsymbol{\pi}_2 v_{(0)})\right.\\
&\qquad+\frac{1}{n}\sum_{i=1}^{n}\{Y_i-r_Y(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{iI};\boldsymbol{\theta}_{10})\}\{\mathbf{X}_{iI}-r_{\mathbf{X}_I}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{iI};\boldsymbol{\theta}_{10})\}^{\tau}\\
&\qquad\left.(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})}-\boldsymbol{\pi}_2 v_{(0)})\right]+R_{n1}\stackrel{\text{def}}{=}I_{n1}+R_{n1},
\end{aligned}$$

where

$$\begin{aligned}
R_{n1}=&\frac{1}{n}\sum_{i=1}^{n}\{Y_i-r_Y(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{iI};\boldsymbol{\theta}_{10})\}\\
&\qquad\left\{r_{\mathbf{X}_I}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{iI};\boldsymbol{\theta}_{10})-\hat{r}_{\mathbf{X}_I}(\hat{\boldsymbol{\theta}}_{10}^{\tau}\mathbf{Z}_{iI};\hat{\boldsymbol{\theta}}_{10})\right\}^{\tau}(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})})\\
&+\frac{1}{n}\sum_{i=1}^{n}\left\{r_Y(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{iI};\boldsymbol{\theta}_{10})-\hat{r}_Y(\hat{\boldsymbol{\theta}}_{10}^{\tau}\mathbf{Z}_{iI};\hat{\boldsymbol{\theta}}_{10})\right\}\\
&\qquad\{\mathbf{X}_{iI}-r_{\mathbf{X}_I}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{iI};\boldsymbol{\theta}_{10})\}^{\tau}(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})})\\
&+\frac{1}{n}\sum_{i=1}^{n}\left\{r_Y(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{iI};\theta_{10})-\hat{r}_Y(\hat{\boldsymbol{\theta}}_{10}^{\tau}\mathbf{Z}_{iI};\hat{\boldsymbol{\theta}}_{10})\right\}\\
&\qquad\left\{r_{\mathbf{X}_I}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{iI};\boldsymbol{\theta}_{10})-\hat{r}_{\mathbf{X}_I}(\hat{\boldsymbol{\theta}}_{10}^{\tau}\mathbf{Z}_{iI};\hat{\boldsymbol{\theta}}_{10})\right\}^{\tau}\\
\times(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})})&\stackrel{\text{def}}{=}R_{n1}^{(1)}+R_{n1}^{(2)}+R_{n1}^{(3)}.
\end{aligned}$$

Recall that on the set $\{\mathcal{A}_n=\mathcal{A}_0\}$, $\widetilde{p}_0=p_0$, $\widetilde{q}_0=q_0$, $\tilde{\boldsymbol{\pi}}_2=\boldsymbol{\pi}_2$, $\mathbf{Z}_{iI}=\mathbf{Z}_{i0}$ and $\mathbf{X}_{iI}=\mathbf{X}_{i0}$ for $i=1,\ldots,n$, and $r_{\mathbf{X}_I}(t;\varsigma)=r_{\mathbf{X}_0}(t;\varsigma)$, $\hat{r}_{\mathbf{X}_I}(t;\varsigma)=\hat{r}_{\mathbf{X}_0}(t;\varsigma)$. Using $\boldsymbol{\beta}_{10}=\boldsymbol{\kappa}\times(\boldsymbol{\pi}_2 v_{(0)})$ and the result in Theorem 3.3 that $\tilde{v}_{n(\tilde{0})}-v_{(0)}=O_P(1/\sqrt{n})$, we can obtain the following asymptotic expression:

$$\begin{aligned}
I_{n1}\quad=\quad&\boldsymbol{\kappa}(\boldsymbol{\pi}_2 v_{(0)})^{\tau}\frac{1}{n}\sum_{i=1}^{n}\breve{\mathbf{X}}_{i0}\breve{\mathbf{X}}_{i0}^{\tau}(\boldsymbol{\pi}_2 v_{(0)})+\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\breve{\mathbf{X}}_{i0}^{\tau}(\boldsymbol{\pi}_2 v_{(0)})\\
&+\boldsymbol{\kappa}(\boldsymbol{\pi}_2 v_{(0)})^{\tau}\frac{1}{n}\sum_{i=1}^{n}\breve{\mathbf{X}}_{i0}\breve{\mathbf{X}}_{i0}^{\tau}\{\boldsymbol{\pi}_2(\tilde{v}_{n(\tilde{0})}-v_{(0)})\}+o_P\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Next, we show that R_{n1} is of order $o_P(1/\sqrt{n})$. First, we deal with $R_{n1}^{(1)}$ on the set $\{\mathcal{A}_n=\mathcal{A}_0\}$.

$$R_{n1}^{(1)} = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i + \boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0})\left\{r_{\mathbf{X}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0};\boldsymbol{\theta}_{10}) - \hat{r}_{\mathbf{X}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0};\boldsymbol{\theta}_{10})\right\}^{\tau}(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})})$$
$$+ \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i + \boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0})\left\{\hat{r}_{\mathbf{X}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0};\boldsymbol{\theta}_{10}) - \hat{r}_{\mathbf{X}_0}(\hat{\boldsymbol{\theta}}_{10}^{\tau}\mathbf{Z}_{i0};\hat{\boldsymbol{\theta}}_{10})\right\}^{\tau}(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})})$$
$$\stackrel{\text{def}}{=} R_{n1}^{(1)A} + R_{n1}^{(1)B}.$$

By applying Proposition 1*(iii)* in Cui et al. (2011), we can obtain that

$$\hat{r}_{\mathbf{X}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0};\boldsymbol{\theta}_{10}) - \hat{r}_{\mathbf{X}_0}(\hat{\boldsymbol{\theta}}_{10}^{\tau}\mathbf{Z}_{i0};\hat{\boldsymbol{\theta}}_{10}) \tag{A.18}$$
$$= \left\{r'_{\mathbf{X}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0};\boldsymbol{\theta}_{10})\,\breve{\mathbf{Z}}_{i0} + O_P\left(\sqrt{h^4 + n^{-1}h^{-3}}\right)\right\}^{\tau}(\hat{\boldsymbol{\theta}}_{10} - \boldsymbol{\theta}_{10}).$$

Thus, (A.18) entails that $R_{n1}^{(1)B} = o_P(1/\sqrt{n})$. Now, we consider the first term $R_{n1}^{(1)A}$ on the set $\{\mathcal{A}_n = \mathcal{A}_0\}$, which can be expressed as follows.

$$R_{n1}^{(1)A} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{j=1}^{n}(\varepsilon_i + \boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0})$$
$$\left\{\frac{\psi_j(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0},\boldsymbol{\theta}_{10})\big(r_{\mathbf{X}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0},\boldsymbol{\theta}_{10}) - \mathbf{X}_{j0}\big)}{\frac{1}{n}\sum_{j=1}^{n}\psi_j(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0},\boldsymbol{\theta}_{10})}\right\}^{\tau}(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})}).$$

We first derive asymptotic expansions of $V_{n,0}(\theta_0^{\tau}\mathbf{Z}_{i0},\boldsymbol{\theta}_{10})$, $V_{n,1}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0},\boldsymbol{\theta}_{10})$ and $V_{n,2}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0},\boldsymbol{\theta}_{10})$, which were defined in the first paragraph of Section 3.4, and then that of $\psi_j(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0},\boldsymbol{\theta}_{10})$ for $j = 1,\ldots,n$. It follows from the arguments similar to the proof of Theorem 3.1 in Fan and Gijbels (1996) that, as $h \to 0$ and $nh \to \infty$,

$$E\left\{\left|\frac{1}{n}V_{n,2}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0},\boldsymbol{\theta}_{10}) - f_{\theta_{10}^{\tau}\mathbf{Z}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0})\right|\right\}^2 = O\left(h^4 + \frac{1}{nh}\right),$$
$$E\left\{\left|\frac{1}{nh}V_{n,1}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0},\boldsymbol{\theta}_{10}) - hf'_{\theta_{10}^{\tau}\mathbf{Z}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0})\int K(w)w^2dw\right|\right\}^2 = O\left(h^6 + \frac{1}{nh}\right),$$
$$E\left\{\left|\frac{1}{nh^2}V_{n,2}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0},\theta_{10}) - f_{\theta_{10}^{\tau}\mathbf{Z}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0})\int K(w)w^2dw\right|\right\}^2 = O\left(h^4 + \frac{1}{nh}\right).$$

Thus, $R_{n1}^{(1)A}$ can be asymptotically expressed as

$$R_{n1}^{(1)A} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_h(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{j0} - \boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0})\left(\varepsilon_i + \boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0}\right)$$
$$\left\{\frac{r_{\mathbf{X}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0},\boldsymbol{\theta}_{10}) - \mathbf{X}_{j0}}{f_{\theta_{10}^{\tau}\mathbf{Z}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0})}\right\}^{\tau} \times (\boldsymbol{\pi}_2 v_{(0)})(1 + o_P(1)).$$

We know that, if $nh^2 \to \infty$,

$$\frac{1}{nh^2}\sum_{i=1}^{n}K(0)(\varepsilon_i + \boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0})\frac{\{r_{\mathbf{X}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0},\boldsymbol{\theta}_{10}) - \mathbf{X}_{i0}\}^{\tau}(\boldsymbol{\pi}_1 v_{(0)})}{f_{\theta_{10}^{\tau}\mathbf{Z}_0}(\boldsymbol{\theta}_{10}^{\tau}\mathbf{Z}_{i0})} = o_P(1/\sqrt{n})$$

by the law of large numbers. Using the arguments similar to those used by Zhu and Fang (1996), we can prove that the summation for $i \neq j$ within $R_{n1}^{(1)A}$ is a standard U-statistic with a varying kernel with the bandwidth h; that is,

$$\frac{2c_n}{n(n-1)} \sum_{1 \le i < j \le n} \mathbf{H}\big\{(\mathbf{X}_{i0}, \boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}, \varepsilon_i), (\mathbf{X}_{j0}, \boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{j0}, \varepsilon_j)\big\}(1 + o_P(1)),$$

where $c_n = (n-1)/n$. Note that $K(\cdot)$ is a symmetric function and the symmetric U-statistic kernel is $\mathbf{H}\big\{(\mathbf{X}_{i0}, \boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}, \varepsilon_i), (\mathbf{X}_{j0}, \boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{j0}, \varepsilon_j)\big\}$ given as

$$\begin{aligned} \frac{1}{2} \Bigg\{ & \frac{\big(\varepsilon_i + \boldsymbol{\beta}_{10}^\tau \breve{\mathbf{X}}_{i0}\big)\big(r_{\mathbf{X}_0}(\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}, \boldsymbol{\theta}_{10}) - \mathbf{X}_{j0}\big)}{f_{\theta_{10}^\tau \mathbf{Z}_0}(\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0})} \\ & + \frac{\big(\varepsilon_j + \boldsymbol{\beta}_{10}^\tau \breve{\mathbf{X}}_{j0}\big)\big(r_{\mathbf{X}_0}(\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{j0}, \boldsymbol{\theta}_{10}) - \mathbf{X}_{i0}\big)}{f_{\theta_{10}^\tau \mathbf{Z}_0}(\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{j0})} \Bigg\}^\tau \\ & \times (\boldsymbol{\pi}_2 v_{(0)}) K_h(\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{j0} - \boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}). \end{aligned}$$

Using the projection of U-statistics (Serfling, 1980, Section 5.3.1), we obtain that

$$R_{n1}^{(1)A} = \frac{1}{n} \sum_{i=1}^n \mathbf{H}^*(\mathbf{X}_{i0}, \boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}, \varepsilon_i)(1 + o_P(1)) + o_P\left(\frac{1}{\sqrt{n}}\right)$$

with $\mathbf{H}^*(\mathbf{X}_{i0}, \boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}, \varepsilon_i)$ being $E[\mathbf{H}\{(\mathbf{X}_{i0}, \boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}, \varepsilon_i), (\mathbf{X}_{j0}, \boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{j0}, \varepsilon_j)\} | (\mathbf{X}_{i0}, \boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}, \varepsilon_i)]$, which can be expressed as

$$\begin{aligned} \frac{1}{2} \frac{\varepsilon_i + \boldsymbol{\beta}_{10}^\tau \breve{\mathbf{X}}_{i0}}{f_{\theta_{10}^\tau \mathbf{Z}_0}(\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0})} \Big\{ & (r'_{\mathbf{X}_0}(\theta_{10}^\tau \mathbf{Z}_{i0}, \boldsymbol{\theta}_{10}) f'_{\theta_{10}^\tau \mathbf{Z}_0}(\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}) \\ & + \frac{1}{2} r''_{\mathbf{X}_0}(\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}, \boldsymbol{\theta}_{10}) f_{\theta_{10}^\tau \mathbf{Z}_0}(\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}) \Big\} h^2. \end{aligned}$$

Note that $E\mathbf{H}^*(\mathbf{X}_{i0}, \boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}, \varepsilon_i) = 0$, then $1/\sqrt{n} \sum_{i=1}^n \mathbf{H}^*(\mathbf{X}_{i0}, \boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}, \varepsilon_i) = O_P(1)h^2$. It follows from $h \to 0$ that $\sqrt{n} R_{n1}^{(1)A} = o_P(1)$. Using a similar analysis to the proof for $R_{n1}^{(1)}$, we obtain $\sqrt{n} R_{n1}^{(2)} = o_P(1)$. Applying Lemma A.4 in Wang et al. (2010) and Cauchy-Schwarz inequality, we have, as $(\log n)^2/(nh^2) \to 0$,

$$\begin{aligned} R_{n1}^{(3)} \quad \le \quad & \frac{1}{n} \left[\sum_{i=1}^n \left\{ r_Y\left(\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}; \boldsymbol{\theta}_{10}\right) - \hat{r}_Y(\hat{\boldsymbol{\theta}}_{10}^\tau \mathbf{Z}_{i0}; \hat{\boldsymbol{\theta}}_{10}) \right\}^2 \right]^{1/2} \\ & \times \left(\sum_{i=1}^n \left[\left\{ r_{\mathbf{X}_0}(\boldsymbol{\theta}_{10}^\tau \mathbf{Z}_{i0}; \boldsymbol{\theta}_{10}) - \hat{r}_{\mathbf{X}_0}(\hat{\boldsymbol{\theta}}_{10}^\tau \mathbf{Z}_{i0}; \hat{\boldsymbol{\theta}}_{10}) \right\}^\tau \tilde{\boldsymbol{\pi}}_2 \tilde{v}_{n(\tilde{0})} \right]^2 \right)^{1/2} \\ = \quad & O_P\left(\frac{\log n}{nh}\right) = o_P\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

So on the set $\{\mathcal{A}_n = \mathcal{A}_0\}$, we have that $R_{n1}^{(1)}$, $R_{n1}^{(2)}$ and $R_{n1}^{(3)}$ are all $o_P(1/\sqrt{n})$. Again as $P(\mathcal{A}_n = \mathcal{A}_0) \to 1$, we obtain that $R_{n1} = o_P(1/\sqrt{n})$.

Step C.2. We now decompose the denominator of $\hat{\boldsymbol{\kappa}}$ on the set $\{\mathcal{A}_n = \mathcal{A}_0\}$ as follows.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left[\left\{\mathbf{X}_{i0}-\hat{r}_{\mathbf{X}_0}(\hat{\boldsymbol{\theta}}_{10}^{\tau}\mathbf{Z}_{i0};\hat{\boldsymbol{\theta}}_{10})\right\}^{\tau}(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})})\right]^2 &= \left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\pi}_2 v_{(0)})^{\tau}\breve{\mathbf{X}}_{i0}\right.\\ &\left.\breve{\mathbf{X}}_{i0}^{\tau}(\boldsymbol{\pi}_2 v_{(0)})+\frac{2}{n}\sum_{i=1}^{n}(\boldsymbol{\pi}_1 v_{(0)})^{\tau}\breve{\mathbf{X}}_{i0}\breve{\mathbf{X}}_{i0}^{\tau}\{(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})}-\boldsymbol{\pi}_2 v_{(0)})\}\right]+R_{n2}\\ &\stackrel{\text{def}}{=} I_{n2}+R_{n2}.\end{aligned}$$

Similar to the analysis of R_{n1}, we obtain that $R_{n2}=o_P\left(1/\sqrt{n}\right)$. Recalling that the asymptotic expression of $\tilde{v}_{n(\tilde{0})}$ in (A.17) and $P(\mathcal{A}_n=\mathcal{A}_0)\to 1$, we have

$$\begin{pmatrix} F_{n1}\\ F_{n2}\end{pmatrix}\stackrel{\text{def}}{=}\begin{pmatrix}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0}\\ \boldsymbol{\beta}_{10}^{\tau}\Sigma_{\breve{\mathbf{X}}_0}\left[\sqrt{n}(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})}-\boldsymbol{\pi}_2 v_{(0)})\right]\end{pmatrix}\xrightarrow{L}N\left(0_{2\times 1},\mathbf{W}\right),$$

where $\mathbf{W}$ is defined in (3.2). As a consequence, we have

$$\begin{aligned}\sqrt{n}\begin{pmatrix} I_{n1}-\frac{1}{n\boldsymbol{\kappa}}\sum_{i=1}^{n}\boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0}\breve{\mathbf{X}}_{i0}^{\tau}\boldsymbol{\beta}_{10}\\ I_{n2}-\frac{1}{n\boldsymbol{\kappa}^2}\sum_{i=1}^{n}\boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0}\breve{\mathbf{X}}_{i0}^{\tau}\boldsymbol{\beta}_{10}\end{pmatrix}&=\begin{pmatrix}1/\boldsymbol{\kappa} & 1\\ 0 & 2/\boldsymbol{\kappa}\end{pmatrix}\begin{pmatrix}F_{n1}\\ F_{n2}\end{pmatrix}\\ &\xrightarrow{L}N\left(0,\mathbf{W}^*\right),\end{aligned}$$

where $\mathbf{W}^*=\begin{pmatrix}1/\boldsymbol{\kappa} & 1\\ 0 & 2/\boldsymbol{\kappa}\end{pmatrix}\mathbf{W}\begin{pmatrix}1/\boldsymbol{\kappa} & 0\\ 1 & 2/\boldsymbol{\kappa}\end{pmatrix}$.

We have shown that $\sqrt{n}R_{n1}=o_P(1)$, $\sqrt{n}R_{n2}=o_P(1)$, and $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0}\times\breve{\mathbf{X}}_{i0}^{\tau}\boldsymbol{\beta}_{10}=\boldsymbol{\beta}_{10}^{\tau}\Sigma_{\breve{\mathbf{X}}_0}\boldsymbol{\beta}_{10}$, a.s. The asymptotic distribution of $\hat{\boldsymbol{\kappa}}$ can be obtained by the following expression and a direct calculation:

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\kappa}}-\boldsymbol{\kappa})&=\sqrt{n}\left(\frac{I_{n1}+R_{n1}}{I_{n2}+R_{n2}}-\boldsymbol{\kappa}\right)=\sqrt{n}\left(\frac{I_{n1}}{I_{n2}}-\boldsymbol{\kappa}\right)+o_P(1)\\ &=\sqrt{n}\left(\frac{I_{n1}}{I_{n2}}-\frac{\frac{1}{n\boldsymbol{\kappa}}\sum_{i=1}^{n}\boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0}\breve{\mathbf{X}}_{i0}^{\tau}\boldsymbol{\beta}_{10}}{\frac{1}{n\boldsymbol{\kappa}^2}\sum_{i=1}^{n}\boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0}\breve{\mathbf{X}}_{i0}^{\tau}\boldsymbol{\beta}_{10}}\right)+o_P(1)\xrightarrow{L}N\left(0,\sigma_{\boldsymbol{\kappa}}^2\right).\end{aligned}$$

□

A.9. Proof of Theorem 3.6

Note that on the set $\{\mathcal{A}_n=\mathcal{A}_0\}$, $\widetilde{p}_0=p_0$, $\widetilde{q}_0=q_0$, then

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{10}-\boldsymbol{\beta}_{10}\right)&=\sqrt{n}\left\{\hat{\boldsymbol{\kappa}}\times(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})})-\boldsymbol{\kappa}\times(\boldsymbol{\pi}_2 v_{(0)})\right\}\\ &=\sqrt{n}(\hat{\boldsymbol{\kappa}}-\boldsymbol{\kappa})\times\left(\boldsymbol{\pi}_2 v_{(0)}\right)+\boldsymbol{\kappa}\left\{\sqrt{n}\left(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})}-\boldsymbol{\pi}_2 v_{(0)}\right)\right\}+o_P(1).\end{aligned}\tag{A.19}$$

Taylor expansion implies

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\kappa}}-\boldsymbol{\kappa}) &= \sqrt{n}\Big(I_{n1}-\frac{1}{n\boldsymbol{\kappa}}\sum_{i=1}^{n}\boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0}\breve{\mathbf{X}}_{i0}^{\tau}\boldsymbol{\beta}_{10}\Big)\frac{\boldsymbol{\kappa}^2}{\boldsymbol{\beta}_{10}^{\tau}\Sigma_{\breve{\mathbf{X}}_0}\boldsymbol{\beta}_{10}} \\ &\quad -\sqrt{n}\Big(I_{n2}-\frac{1}{n\boldsymbol{\kappa}^2}\sum_{i=1}^{n}\boldsymbol{\beta}_{10}^{\tau}\breve{\mathbf{X}}_{i0}\breve{\mathbf{X}}_{i0}^{\tau}\boldsymbol{\beta}_{10}\Big)\frac{\boldsymbol{\kappa}^3}{\boldsymbol{\beta}_{10}^{\tau}\Sigma_{\breve{\mathbf{X}}_0}\boldsymbol{\beta}_{10}}+o_P(1) \\ &= \frac{\boldsymbol{\kappa}^2}{\boldsymbol{\beta}_{10}^{\tau}\Sigma_{\breve{\mathbf{X}}_0}\boldsymbol{\beta}_{10}}\left(\frac{1}{\boldsymbol{\kappa}}F_{n1}+F_{n2}\right)-\frac{2}{\boldsymbol{\kappa}}\frac{\boldsymbol{\kappa}^3}{\boldsymbol{\beta}_{10}^{\tau}\Sigma_{\breve{\mathbf{X}}_0}\boldsymbol{\beta}_{10}}F_{n2}+o_P(1).\end{aligned}$$

It follows that

$$\sqrt{n}(\hat{\boldsymbol{\kappa}}-\boldsymbol{\kappa})=\frac{\boldsymbol{\kappa}}{\boldsymbol{\beta}_{10}^{\tau}\Sigma_{\breve{\mathbf{X}}_0}\boldsymbol{\beta}_{10}}F_{n1}-\frac{\boldsymbol{\kappa}^2}{\boldsymbol{\beta}_{10}^{\tau}\Sigma_{\breve{\mathbf{X}}_0}\boldsymbol{\beta}_{10}}F_{n2}+o_P(1). \tag{A.20}$$

Note that $\boldsymbol{\beta}_{10}=\boldsymbol{\kappa}\times(\boldsymbol{\pi}_2 v_{(0)})$. A combination of (A.17), (A.19) and (A.20) and $P(\mathcal{A}_n=\mathcal{A}_0)\to 1$ yields

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{10}-\boldsymbol{\beta}_{10}\right) &= \frac{\boldsymbol{\beta}_{10}}{\boldsymbol{\beta}_{10}^{\tau}\Sigma_{\breve{\mathbf{X}}_0}\boldsymbol{\beta}_{10}}F_{n1}-\frac{\boldsymbol{\kappa}\boldsymbol{\beta}_{10}}{\boldsymbol{\beta}_{10}^{\tau}\Sigma_{\breve{\mathbf{X}}_0}\boldsymbol{\beta}_{10}}F_{n2} \\ &\quad +\boldsymbol{\kappa}\left\{\sqrt{n}\left(\tilde{\boldsymbol{\pi}}_2\tilde{v}_{n(\tilde{0})}-\boldsymbol{\pi}_2 v_{(0)}\right)\right\}+o_P(1)\xrightarrow{L}N\left(0,\Sigma_{\beta_{10}}\right).\end{aligned}$$

We complete the proof. □

Acknowledgements

The authors greatly thank the Editor and referees for their constructive comments that substantially improved an earlier version of this paper.

References

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics* **36**(6): 2577–2604. MR2485008

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**(1): 199–227. MR2387969

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106**(494): 672–684. MR2847949

Carroll, R., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association* **92**: 477–489. MR1467842

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *The Annals of Statistics* **16**: 136–146. MR0924861

Chen, X., Zou, C., and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics* **38**: 3696–3723. MR2766865

Cook, R. D. (1996a). Added-variable plots and curvature in linear regression. *Technometrics* **38**: 275–278. MR1411884

Cook, R. D. (1996b). Graphics for regressions with a binary response. *Journal of the American Statistical Association* **91**: 983–992. MR1424601

Cook, R. D. (1998). *Regression Graphics*. John Wiley & Sons Inc., New York. Ideas for studying regressions through graphics, A Wiley-Interscience Publication. MR1645673

Cook, R. D. and Weisberg, S. (1991). Comment on "sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**: 328–332. MR1137117

Cui, X., Härdle, W., and Zhu, L.-X. (2011). The EFM approach for single-index models. *The Annals of Statistics* **39**: 1658–1688. MR2850216

Cuzick, J. (1992). Efficient estimates in semiparametric additive regression models with unknown error distribution. *The Annals of Statistics* **20**: 1129–1136. MR1165611

Engle, R., Granger, C., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* **81**: 310–320.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London. MR1383587

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. MR1946581

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *International Congress of Mathematicians. Vol. III*, pages 595–622. Eur. Math. Soc., Zürich. MR2275698

Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *The Annals of Statistics* **32**: 928–961. MR2065194

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics* **21**: 157–178. MR1212171

Härdle, W., Liang, H., and Gao, J. T. (2000). *Partially Linear Models*. Springer Physica, Heidelberg. MR1787637

Heckman, N. E. (1986). Spline smoothing in partly linear models. *Journal of the Royal Statistical Society, Series B* **48**: 244–248. MR0868002

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer, New York. MR2535631

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics* **38**: 2281–2313. MR2676890

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58**: 71–120. MR1230981

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics* **37**(6B): 4254–4278. MR2572459

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**: 997–1008. MR2354409

Li, G. R., Peng, H., Zhang, J., and Zhu, L. X. (2012). Robust rank correlation based screening. *The Annals of Statistics* **40**: 1846–1877.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86**: 316–342. MR1137117

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association* **87**: 1025–1039. MR1209564

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94**: 603–613. MR2410011

Liang, H., Härdle, W., and Carroll, R. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *The Annals of Statistics* **27**: 1519–1535. MR1742498

Liang, H., Liu, X., Li, R., and Tsai, C. L. (2010). Estimation and testing for partially linear single-index models. *The Annals of Statistics* **38**: 3811–3836. MR2766869

Liang, H., Wang, S., Robins, J., and Carroll, R. (2004). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association* **99**: 357–367. MR2062822

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in smoothing spline analysis of variance models, **34**(5): 2272–2297. MR2291500

Manton, J. H. (2002). Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing* **50**: 635–650. MR1895067

Meier, L., van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics* **37**: 3779–3821. MR2572443

Naik, P. A. and Tsai, C.-L. (2001). Single-index model selections. *Biometrika* **88**: 821–832. MR1859412

Ni, X., Zhang, H. H., and Zhang, D. (2009). Automatic model selection for partially linear models. *Journal of Multivariate Analysis* **100**: 2100–2111. MR2543089

Ravikumar, P., Lafferty, H., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society, Series B* **71**: 1009–1030. MR2750255

Sala-I-Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review* **87**: 178–183.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons Inc., New York. MR0595165

Speckman, P. E. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, 50: 413–436. MR0970977

Wahba, G. (1984). Partial spline models for the semiparametric estimation of functions of several variables. In *Statistical Analyses for Time Series*, pages 319–329, Tokyo. Institute of Statistical Mathematics. Japan-US Joint Seminar.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association* **103**: 811–821. MR2524332

Wang, J. L., Xue, L. G., Zhu, L. X., and Chong, Y. S. (2010). Estimation for a partial-linear single-index model. *The Annals of Statistics* **38**: 246–274. MR2589322

Wang, L. and Yang, L. (2009). Spline estimation of single-index models. *Statistica Sinica* **19**: 765–783. MR2514187

Wang, T. and Zhu, L. X. (2011). Consistent model selection and estimation in a general single-index model with "large p and small n". Technical report, Department of Mathematics, Hong Kong Baptist University, Hong Kong.

Wu, Y. and Li, L. (2011). Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statistica Sinica* **21**: 707–730. MR2829852

Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis* **97**: 1162–1184. MR2276153

Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society, Series B* **64**: 363–410. MR1924297

Xie, H. and Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics* **37**: 673–696. MR2502647

Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional kth moment in regression. *Journal of the Royal Statistical Society, Series B* **64**(2): 159–175. MR1904698

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* **97**: 1042–1054. MR1951258

Yu, Z., Li, B., and Zhu, L. X. (2011). Asymptotic expansion for dimension reduction methods and its application to bias correction. Technical report, The Pennsylvania State University.

Zhang, H. H., Cheng, G., and Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *Journal of the American Statistical Association* **106**: 1099–1112. MR2894767

Zhu, L. P., Wang, T., Zhu, L. X., and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika* **97**: 295–304. MR2650739

Zhu, L. P. and Zhu, L. X. (2009a). Dimension reduction for conditional variance in regressions. *Statistica Sinica* **19**: 869–883. MR2514192

Zhu, L. P. and Zhu, L. X. (2009b). On distribution-weighted partial least squares with diverging number of highly correlated predictors. *Journal of the Royal Statistical Society, Series B* **71**: 525–548. MR2649607

Zhu, L. P., Zhu, L. X., and Feng, Z. H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association* **105**: 1455–1466. MR2796563

Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics* **24**: 1053–1068. MR1401836

Zhu, L. X., Miao, B., and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association* **101**: 630–643. MR2281245