Statistica Sinica **29** (2019), 1751-1777
doi:https://doi.org/10.5705/ss.202017.0288

SUFFICIENT DIMENSION REDUCTION UNDER DIMENSION-REDUCTION-BASED IMPUTATION WITH PREDICTORS MISSING AT RANDOM

Xiaojie Yang[1] and Qihua Wang[1,2]

[1] *Chinese Academy of Sciences and* [2] *Zhejiang Gongshang University*

Abstract: In some practical problems, a subset of predictors may be subject to missingness, especially when the dimension of the predictors is high. In this case, the standard sufficient dimension-reduction (SDR) methods cannot be applied directly to avoid the curse of dimensionality. Therefore, a dimension-reduction-based imputation method is developed such that any spectral-decomposition-based SDR method for full data can be applied to the case where predictors are missing at random. The sliced inverse regression (SIR) technique is used to illustrate this procedure. The proposed imputation estimator of the candidate matrix for the SIR, called the DRI-SIR estimator, is asymptotically normal under some mild conditions. Hence, the resulting estimator of the central subspace is root-n consistent. The finite-sample performance of the proposed method is evaluated through comprehensive simulations and real data are analyzed in an application of the method.

Key words and phrases: Kernel imputation, missing at random, missing predictors, sliced inverse regression, sufficient dimension reduction.

1. Introduction

Consider the regression of a univariate response variable Y on a $p\times 1$ covariate vector $\mathbf{X}$. Regression analyses typically focus on how the conditional distribution function $F(y|\mathbf{X}=\mathbf{x})$ changes as the value of $\mathbf{X}$ varies in its marginal sample space. When the dimension p is large, modeling a parametric structure for the regression is difficult, and nonparametric methods are not effective owing to *the curse of dimensionality*. As a result, sufficient dimension-reduction (SDR; Cook (1998a)) methods have been proposed in order to reduce the dimension of $\mathbf{X}$ while preserving full information for Y, without imposing specified regression parametric models. These methods replace $\mathbf{X}$ with $d \le p$ linear combinations, $\beta_1^T\mathbf{X},\ldots,\beta_d^T\mathbf{X}$, such that $Y \perp\!\!\!\perp \mathbf{X} \big| B^T\mathbf{X}$, where B is a $p\times d$ matrix with columns β_j, and $\perp\!\!\!\perp$ indicates statistical independence. The column space of B is called a dimension-reduction subspace (Li (1991)). Such subspaces always exist, but

they are not necessarily unique. Under mild, but fairly weak conditions, Cook (1996) showed that the intersection of all dimension-reduction subspaces is itself a dimension-reduction subspace and called the central subspace (CS) $\mathcal{S}_{Y|\mathbf{X}}$ for the regression of Y on $\mathbf{X}$, where its dimension $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ is called the structural dimension. It is clear that the CS provides the greatest reduction from $\mathbf{X}$ to $B^T\mathbf{X}$ and captures all regression information of Y on $\mathbf{X}$.

Since Li's pioneering work on the sliced inverse regression (SIR; Li (1991)), many SDR methods have been developed to estimate $\mathcal{S}_{Y|\mathbf{X}}$. These include spectral-decomposition-based methods, such as the SIR, sliced average variance estimation (SAVE; Cook and Weisberg (1991)), principal Hessian direction (PHD; Li (1992)), kernel inverse regression (Zhu and Fang (1996); Ferré and Yao (2005)), contour regression (Li, Zha and Chiaromonte (2005)), directional regression (DR; Li and Wang (2007)), and so on. Some other methods have been derived by numerically minimizing (or maximizing) nonparametric objective functions. These methods include the minimum average variance estimator (MAVE; Xia et al. (2002)), the information index method (Yin and Cook (2005)), and so forth. Cai and Chen (2010, Chap. 2) were the first to provide a selective review of SDR methods for regressions. Later, Ma and Zhu (2013) discussed recent developments in the SDR field.

Despite the growing number of the SDR literature with significant theoretical advances, only little attention has been paid to SDR with missing values, even though the problem of missing data is relatively common. For responses missing at random (MAR), Ding and Wang (2011) proposed a fusion-refinement (FR) procedure to handle dimension-reduction problems. In the context of predictors MAR, Li and Lu (2008) introduced the augmented inverse probability-weighted SIR estimator (AIPW-SIR), and Zhu, Wang and Zhu (2012) proposed a parametric imputation procedure for SIR (PI-SIR). Both methods require that parametric models are specified for the conditional expectations and the propensity function.

For ease of exposition, we write $\mathbf{X} = (X_1, \ldots, X_p)^T = (\mathbf{X}_{mis}^T, \mathbf{X}_{obs}^T)^T$, where $\mathbf{X}_{mis} = (X_1, \ldots, X_{p_1})^T \in R^{p_1}$ refers to predictors with missingness in a subset of subjects, and $\mathbf{X}_{obs} = (X_{p_1+1}, \ldots, X_p)^T \in R^{p-p_1}$ is always observed for all subjects. Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{p_1})^T$ denote a vector of missingness indicators for $\mathbf{X}_{mis}$, where δ_k takes the value one if there is no missingness for the k-th component X_k in $\mathbf{X}_{mis}$, and zero otherwise. Throughout this paper, we assume that $\mathbf{X}_{mis}$ is MAR; that is,

$$\boldsymbol{\delta} \perp\!\!\!\perp \mathbf{X}_{mis} \mid (\mathbf{X}_{obs}^T, Y)^T, \tag{1.1}$$

which essentially allows the missingness to depend only on the completely observed variables $(\mathbf{X}_{obs}^T, Y)^T$.

To make the SIR applicable to the case of missing predictors, the main difficulty is to estimate the candidate matrix $\{\mathrm{Cov}(\mathbf{X})\}^{-1}\mathrm{Cov}\{E(\mathbf{X}|Y)\}$ for the SIR. According to Zhu, Wang and Zhu (2012), we need to obtain consistent estimators of the quantities, $E(X_k), E(X_k\mathbf{X}_{obs}^T), E(X_k|Y), E(X_k^2)$, and $E(X_kX_l)$ $(k \neq l)$, where X_k (or X_l) denotes the k-th (or l-th) component in $\mathbf{X}_{mis}$, for $k, l = 1, \ldots, p_1$. Because there is no need to estimate the mixed moment $E(X_kX_l)$ $(k \neq l)$ for $p_1 = 1$, we focus on the general case $p_1 \geq 2$. Let $\mathbf{V} = (\mathbf{X}_{obs}^T, Y)^T \in R^{p-p_1+1}$. The double-expectation theorem yields that $E(X_k) = E\{E(X_k|\mathbf{V})\}$, $E(X_k\mathbf{X}_{obs}^T) = E\{E(X_k|\mathbf{V})\mathbf{X}_{obs}^T\}$, $E(X_k|Y) = E\{E(X_k|\mathbf{V})|Y\}$, $E(X_k^2) = E\{E(X_k^2|\mathbf{V})\}$, and $E(X_kX_l) = E\{E(X_kX_l|\mathbf{V})\}$ $(k \neq l)$. Thus, we need to handle $E(X_k|\mathbf{V})$, $E(X_k^2|\mathbf{V})$, and $E(X_kX_l|\mathbf{V})$, which poses two challenges:

- how to overcome the curse of dimensionality in the presence of missing predictors when estimating these conditional expectations; and
- how to obtain consistent estimators of these expectations or conditional expectations in the presence of missing predictors after the above problem is solved.

Existing methods fail to solve these two problems. To lessen the effect of high dimension, Li and Lu (2008) recommended using linear models or other proper parametric models for these conditional expectations. Then, Zhu, Wang and Zhu (2012) imposed linear models on $E(\delta_kX_k|\mathbf{V})$, $E(\delta_k|\mathbf{V})$, $E(\delta_kX_k^2|\mathbf{V})$, $E(\delta_k\delta_lX_k$ $X_l|\mathbf{V})$, and $E(\delta_k\delta_l|\mathbf{V})$, and constructed the estimators of these conditional expectations based on the equations $E(X_k|\mathbf{V}) = E(\delta_kX_k|\mathbf{V})/E(\delta_k|\mathbf{V})$, $E(X_k^2|\mathbf{V}) = E(\delta_kX_k^2|\mathbf{V})/E(\delta_k|\mathbf{V})$, and $E(X_kX_l|\mathbf{V}) = E(\delta_k\delta_lX_kX_l|\mathbf{V})/E(\delta_k\delta_l|\mathbf{V})$ indicated by the MAR assumption (1.1). Both methods might yield inconsistent estimators owing to the misspecification of the involved parametric models. Moreover, it is almost impossible to specify all of the parametric models correctly, in practice.

We are now in a position to develop nonparametric methods that avoid the parametric specification of the models, enabling us to resolve the two issues mentioned above. Our strategy is to seek a $q \times r$ matrix Γ, with $r < q = p - p_1 + 1$, for each conditional expectation such that Γ satisfies the following: (i) $\mathbf{V}$ in the conditional expectation can be replaced by its low-dimensional linear transformation $\Gamma^T\mathbf{V}$, without changing the conditional expectation, and (ii) the complete-case (CC) approach that simply removes all subjects with missing values can be used to obtain consistent estimators of Γ and the corresponding conditional expecta-

tion; we treat this as an intermediate step in the proposed method. It can be shown that any existing SDR methods based on a CC analysis can be used to obtain Γ. Furthermore, a dimension-reduction-based kernel imputation method is proposed to obtain consistent estimators of the expectations and conditional expectations in the candidate matrix for the SIR, and thus yield a consistent estimator of the candidate matrix.

The rest of the paper is organized as follows. In Section 2, we present the proposed method and our theoretical results. In Section 3, we check the finite-sample performance of the proposed method using simulated data. In Section 4, we analyze a real data set for illustration. We then conclude our paper with a discussion in Section 5. The proofs of the main results are given in the Appendix.

2. Dimension-reduction-based Kernel Imputation for SIR

In this section, we first briefly review the SIR under full data, and then develop a dimension-reduction-based kernel imputation method for the SIR with predictors MAR.

2.1. Review

The SIR is the most popular method for estimating $\mathcal{S}_{Y|\mathbf{X}}$. It relies on a typically reasonable *linearity condition*, in which for any basis matrix B of $\mathcal{S}_{Y|\mathbf{X}}$, $E(\mathbf{X}|B^T\mathbf{X})$ is linear in $B^T\mathbf{X}$. This condition holds approximately as the dimension of $\mathbf{X}$ increases, while d remains fixed (Hall and Li (1993)). Under this condition, $\text{Span}\{\Sigma_{\mathbf{X}}^{-1}\Sigma_{E(\mathbf{X}|Y)}\} \subseteq \mathcal{S}_{Y|\mathbf{X}}$, where $\Sigma_{\mathbf{X}} = \text{cov}(\mathbf{X}) \in R^{p\times p}$, $\Sigma_{E(\mathbf{X}|Y)} = \text{cov}\{E(\mathbf{X}|Y)\} \in R^{p\times p}$, and $\text{Span}\{A\}$ denotes the column space of a matrix A. The literature refers to the matrix $\Sigma_{\mathbf{X}}^{-1}\Sigma_{E(\mathbf{X}|Y)}$ as a candidate matrix for the SIR. Li (1991) divided the range of Y into H slices $I_1, \ldots, I_{_H}$, and provided an approximation of $\Sigma_{E(\mathbf{X}|Y)}$ as

$$\Lambda = \sum_{h=1}^{H} p_{_h}(m_{_h} - \mu)(m_{_h} - \mu)^T, \tag{2.1}$$

where $\mu = E(\mathbf{X}), p_{_h} = \Pr(Y \in I_h)$, and $m_{_h} = E(\mathbf{X}|Y \in I_{_h})$, for $h = 1, \ldots, H$. It also can be derived that $\text{Span}\{\Sigma_{\mathbf{X}}^{-1}\Lambda\} \subseteq \mathcal{S}_{Y|\mathbf{X}}$.

Given n independent and identically distributed (i.i.d) observations $\{(\mathbf{X}_{_1}, Y_{_1}), \ldots, (\mathbf{X}_{_n}, Y_{_n})\}$, by substituting the usual sample estimates of $p_{_h}, m_{_h}$, and μ into (2.1), the sample version $\widehat{\Lambda}$ of Λ can be obtained and used as an estimate of $\Sigma_{E(\mathbf{X}|Y)}$. Then, the sample estimate of the SIR follows from the spectral decomposition

$$\widehat{\Lambda}\hat{\beta}_j = \hat{\lambda}_j \widehat{\Sigma}_{\mathbf{X}} \hat{\beta}_j \quad \text{for } j = 1, \ldots, d,$$

where $\widehat{\Sigma}_{\mathbf{X}}$ is the usual sample estimate of $\Sigma_{\mathbf{X}}$, and $\hat{\beta}_1, \ldots, \hat{\beta}_d$ denote the eigenvectors corresponding to the d largest nonzero eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d > 0$ of the matrix $\widehat{\Sigma}_{\mathbf{X}}^{-1}\widehat{\Lambda}$. Mainly under the linearity condition and the coverage condition $\mathrm{Span}\{\Sigma_{\mathbf{X}}^{-1}\Sigma_{E(\mathbf{X}|Y)}\} = \mathcal{S}_{Y|\mathbf{X}}$, Li (1991) showed that $\mathrm{Span}\{\hat{\beta}_1, \ldots, \hat{\beta}_d\}$ is a $\sqrt{n}$-consistent estimator of $\mathcal{S}_{Y|\mathbf{X}}$.

2.2. Extension to the case of missing predictors

When the predictors are MAR, the major difficulty is to develop a consistent estimating approach for the candidate matrix $\Sigma_{\mathbf{X}}^{-1}\Sigma_{E(\mathbf{X}|Y)}$ of the SIR. According to Zhu, Wang and Zhu (2012), $\Sigma_{E(\mathbf{X}|Y)} = \Phi_1 - \Phi_0$ and $\Sigma_{\mathbf{X}} = \Phi_2 - \Phi_0$, where

$$\begin{aligned}
\Phi_0 &= E(\mathbf{X})E(\mathbf{X}^T) = \begin{pmatrix} E(\mathbf{X}_{mis})E(\mathbf{X}_{mis}^T) & E(\mathbf{X}_{mis})E(\mathbf{X}_{obs}^T) \\ E(\mathbf{X}_{obs})E(\mathbf{X}_{mis}^T) & E(\mathbf{X}_{obs})E(\mathbf{X}_{obs}^T) \end{pmatrix}, \\
\Phi_1 &= E\{E(\mathbf{X}|Y)E(\mathbf{X}^T|Y)\} \\
&= \begin{pmatrix} E\{E(\mathbf{X}_{mis}|Y)E(\mathbf{X}_{mis}^T|Y)\} & E\{E(\mathbf{X}_{mis}|Y)E(\mathbf{X}_{obs}^T|Y)\} \\ E\{E(\mathbf{X}_{obs}|Y)E(\mathbf{X}_{mis}^T|Y)\} & E\{E(\mathbf{X}_{obs}|Y)E(\mathbf{X}_{obs}^T|Y)\} \end{pmatrix}, \\
\Phi_2 &= E(\mathbf{X}\mathbf{X}^T) = \begin{pmatrix} E(\mathbf{X}_{mis}\mathbf{X}_{mis}^T) & E(\mathbf{X}_{mis}\mathbf{X}_{obs}^T) \\ E(\mathbf{X}_{obs}\mathbf{X}_{mis}^T) & E(\mathbf{X}_{obs}\mathbf{X}_{obs}^T) \end{pmatrix},
\end{aligned}$$

by the partition $\mathbf{X} = (\mathbf{X}_{mis}^T, \mathbf{X}_{obs}^T)^T$. To implement the SIR, we need to estimate the following expectations $E(\mathbf{X}_{obs})$, $E\{E(\mathbf{X}_{obs}|Y)E(\mathbf{X}_{obs}^T|Y)\}$, $E(\mathbf{X}_{obs}\mathbf{X}_{obs}^T)$, $E\{E(\mathbf{X}_{mis}|Y)E(\mathbf{X}_{obs}^T|Y)\}$, $E\{E(\mathbf{X}_{mis}|Y)E(\mathbf{X}_{mis}^T|Y)\}$, $E(\mathbf{X}_{mis})$, $E(\mathbf{X}_{mis}\mathbf{X}_{obs}^T)$, and $E(\mathbf{X}_{mis}\mathbf{X}_{mis}^T)$. The first three quantities can be estimated using standard methods because they involve only the completely observed variables $(\mathbf{X}_{obs}^T, Y)^T$. However, the last five quantities involve the missing covariate vector $\mathbf{X}_{mis}$. Thus, we need to develop new methods to obtain their consistent estimators. In an element-wise manner, this problem reduces to estimating $E(X_k|Y), E(X_k), E(X_k\mathbf{X}_{obs}^T), E(X_k^2)$, and $E(X_kX_l)$ $(k \neq l)$, with X_k (or X_l) denoting the k-th (or l-th) component in $\mathbf{X}_{mis}$, for $k, l = 1, 2, \ldots, p_1$. As discussed in the introduction, estimating these expectations essentially reduces to estimating several conditional expectations, given $\mathbf{V} = (\mathbf{X}_{obs}^T, Y)^T$.

In particular, we describe how to estimate $E(X_k)$, for $k = 1, \ldots, p_1$, and note that the principle of estimating other quantities is similar. Here, we focus on estimating $E(X_k|\mathbf{V})$, owing to $E(X_k) = E\{E(X_k|\mathbf{V})\}$. A parametric regression model is most efficient when the dimension of $\mathbf{V}$ is small and the relationship

between X_k and $\mathbf{V}$ is correctly specified (Yates (1933); Matloff (1981)). In this case, we can estimate $E(X_k)$ by $n^{-1}\sum_{i=1}^{n}\widehat{m}(\mathbf{V}_i)$, with $\widehat{m}(\cdot)$ being an estimator of the parametric model $m(\cdot)$ imposed on $E(X_k|\mathbf{V})$. Such an estimator is inconsistent if $m(\cdot)$ is misspecified. A nonparametric method also can be employed to estimate $E(X_k|\mathbf{V})$ without requiring a parametric specification, but it will most likely suffer from the curse of dimensionality. A natural idea is to replace $\mathbf{V}$ in $E(X_k|\mathbf{V})$ with its low-dimensional transformation $S_k(\mathbf{V}) : R^q \mapsto R^{q^*}$ $(q^* < q)$, such that $E(X_k|\mathbf{V}) = E\{X_k|S_k(\mathbf{V})\}$. The simplest form of $S_k(\mathbf{V})$ is the linear transformation $\Gamma_k^T\mathbf{V}$, where Γ_k denotes a $q \times r_{_k}$ matrix with $r_{_k} < q$. It then follows that

$$E(X_k) = E\{E(X_k|\Gamma_k^T\mathbf{V})\}, \tag{2.2}$$

$$E(X_k\mathbf{X}_{obs}^T) = E\{E(X_k|\mathbf{V})\mathbf{X}_{obs}^T\} = E\{E(X_k|\Gamma_k^T\mathbf{V})\mathbf{X}_{obs}^T\}, \tag{2.3}$$

$$E(X_k|Y) = E\{E(X_k|\mathbf{V})|Y\} = E\{E(X_k|\Gamma_k^T\mathbf{V})|Y\}. \tag{2.4}$$

Following this idea, let Υ_k be a $q \times \tilde{r}_{_k}$ matrix and let Γ_{kl} $(k \neq l)$ be a $q \times r_{_{kl}}$ matrix with $\tilde{r}_{_k}, r_{_{kl}} < q$ and $k, l = 1, \ldots, p_{_1}$. Then, we have

$$E(X_k^2) = E\{E(X_k^2|\Upsilon_k^T\mathbf{V})\}, \tag{2.5}$$

$$E(X_kX_l) = E\{E(X_kX_l|\Gamma_{kl}^T\mathbf{V})\} \quad (k \neq l). \tag{2.6}$$

To estimate the above-mentioned five quantities based on(2.2)–(2.6), we need to first find suitable Γ_k, Υ_k, and Γ_{kl}, as well as their sample counterparts, and then employ kernel smoothing to estimate $E(X_k|\Gamma_k^T\mathbf{V})$, $E(X_k^2|\Upsilon_k^T\mathbf{V})$, and $E(X_kX_l|\Gamma_{kl}^T\mathbf{V})$ $(k \neq l)$.

As stated in the introduction, obtaining Γ_k, Υ_k, and Γ_{kl} as an intermediate step should avoid intensive computations. It is well known that the CC method is a simple, but useful approach in some cases, although it often yields biased and inefficient estimators in many other cases. To make the CC method applicable to constructing consistent estimators of $E(X_k|\Gamma_k^T\mathbf{V})$, $E(X_k^2|\Upsilon_k^T\mathbf{V})$, and $E(X_kX_l|\Gamma_{kl}^T\mathbf{V})$ $(k \neq l)$ with $k, l = 1, 2, \ldots, p_1$, the matrices Γ_k, Υ_k, and Γ_{kl} should respectively satisfy

$$E(X_k|\mathbf{V}) = E(X_k|\Gamma_k^T\mathbf{V}) = E(X_k|\Gamma_k^T\mathbf{V}, \delta_k = 1), \tag{2.7}$$

$$E(X_k^2|\Upsilon_k^T\mathbf{V}) = E(X_k^2|\Upsilon_k^T\mathbf{V}, \delta_k = 1), \tag{2.8}$$

$$E(X_kX_l|\Gamma_{kl}\mathbf{V}) = E(X_kX_l|\Gamma_{kl}^T\mathbf{V}, \delta_k = 1, \delta_l = 1) \quad (k \neq l). \tag{2.9}$$

We next describe how to obtain Γ_k, Υ_k, and Γ_{kl}, for $k, l = 1, \ldots, p_{_1}$.

(i) Derivation of Γ_k, for $k = 1, \ldots, p_{_1}$.

We gain insights from the condition $E(X_k|\mathbf{V}) = E(X_k|\Gamma_k^T\mathbf{V})$. This makes the basis matrix of the central mean subspace (CMS; Cook and Li (2002))

$\mathcal{S}_{E(X_k|\mathbf{V})}$ a natural choice for Γ_k, where $\mathcal{S}_{E(X_k|\mathbf{V})}$ is the minimal mean subspace $\mathcal{S}$ satisfying $E(X_k|\mathbf{V}) = E(X_k|P_{\mathcal{S}}\mathbf{V})$, with $P_{\mathcal{S}}$ being a projection operator onto $\mathcal{S}$ in a standard inner product. With such a Γ_k and the MAR assumption, we further obtain that

$$\begin{aligned} E(X_k|\Gamma_k^T\mathbf{V}, \delta_k = 1) &= E\big[E(X_k|\mathbf{V}, \delta_k = 1)\big|\Gamma_k^T\mathbf{V}, \delta_k = 1\big] \\ &= E\big[E(X_k|\mathbf{V})\big|\Gamma_k^T\mathbf{V}, \delta_k = 1\big] \\ &= E\big[E(X_k|\Gamma_k^T\mathbf{V})\big|\Gamma_k^T\mathbf{V}, \delta_k = 1\big] \\ &= E(X_k|\Gamma_k^T\mathbf{V}). \end{aligned} \tag{2.10}$$

This implies that taking Γ_k as a basis matrix of $\mathcal{S}_{E(X_k|\mathbf{V})}$ satisfies condition (2.7). However, this raises the question of how to obtain a consistent estimator of a method-specific basis Γ_k of $\mathcal{S}_{E(X_k|\mathbf{V})}$ in the presence of missing predictors. In this paper, we use the phrase "a method-specific basis" to avoid ambiguity caused by the non-uniqueness of the basis; at times, we may omit it for simplicity. SDR methods based on the CC analysis can yield an estimator of the partial central mean subspace (Li, Cook and Chiaromonte (2003)) $\mathcal{S}^{\{\delta_k=1\}}_{E(X_k|\mathbf{V})}$, which is the minimal partial mean subspace $\mathcal{S}$ satisfying $E(X_k|\mathbf{V}, \delta_k = 1) = E(X_k|P_{\mathcal{S}}\mathbf{V}, \delta_k = 1)$. Proposition 1 states that a consistent estimator of Γ_k can be obtained using $\mathcal{S}^{\{\delta_k=1\}}_{E(X_k|\mathbf{V})}$.

Proposition 1. *Suppose that the MAR assumption* (1.1) *holds,* $\mathbf{V}$ *has support* R^q*, and* $P(\delta_k = 1|\mathbf{V}) > 0$*, for* $k = 1, \ldots, p_1$*. Then, we have* $\mathcal{S}_{E(X_k|\mathbf{V})} = \mathcal{S}^{\{\delta_k=1\}}_{E(X_k|\mathbf{V})}$*, for* $k = 1, \ldots, p_1$*.*

Cook and Li (2002) proposed the iterative Hessian transformation (IHT) method for estimating the basis of a CMS. By Proposition 1, the IHT method can be applied to the completely observed dataset $\{(X_{ki}, \mathbf{V}_i) : \delta_{ki} = 1, i = 1, \ldots, n\}$ to obtain a consistent estimator $\widehat{\Gamma}_k$ of Γ_k.

(ii) Derivation of Υ_k, for $k = 1, \ldots, p_1$.

Let $\widetilde{X}_k = X_k^2$. Similar arguments to (2.10) can be used to verify that condition (2.8) holds, provided that Υ_k is taken as a basis matrix of the CMS $\mathcal{S}_{E(\widetilde{X}_k|\mathbf{V})}$. Proposition 2 ensures that we can obtain a consistent estimator of a method-specific basis Υ_k using the partial central mean subspace $\mathcal{S}^{\{\delta_k=1\}}_{E(\widetilde{X}_k|\mathbf{V})}$, where $\mathcal{S}^{\{\delta_k=1\}}_{E(\widetilde{X}_k|\mathbf{V})}$ is the minimal partial mean subspace $\mathcal{S}$ satisfying $E(\widetilde{X}_k|\mathbf{V}, \delta_k = 1) = E(\widetilde{X}_k|P_{\mathcal{S}}\mathbf{V}, \delta_k = 1)$.

Proposition 2. *Assuming the same conditions as those in Proposition* 1*, we have* $\mathcal{S}_{E(\widetilde{X}_k|\mathbf{V})} = \mathcal{S}^{\{\delta_k=1\}}_{E(\widetilde{X}_k|\mathbf{V})}$*, for* $k = 1, \ldots, p_1$*.*

According to Proposition 2, the IHT method can be applied to the completely observed dataset $\{(\widetilde{X}_{ki}, \mathbf{V}_i) : \widetilde{X}_{ki} = X_{ki}^2, \delta_{ki} = 1, i = 1, \ldots, n\}$ to obtain a consistent estimator $\widehat{\Upsilon}_k$ of Υ_k.

(iii) Derivation of Γ_{kl} $(k \neq l)$, for $k, l = 1, \ldots, p_{\text{1}}$.

Observing that $\Gamma_{kl} = \Gamma_{lk}$, with $k \neq l$, we consider the case of $k < l$ only. Let $Z^{(kl)} = X_k X_l$, with $k < l$. When Γ_{kl} is taken as a basis matrix of the CMS $\mathcal{S}_{E(Z^{(kl)}|\mathbf{V})}$, together with the MAR assumption, we can show that

$$\begin{aligned}
&E(X_k X_l | \Gamma_{kl}^T \mathbf{V}, \delta_k = 1, \delta_l = 1) \\
&= E\big[E(X_k X_l | \mathbf{V}, \delta_k = 1, \delta_l = 1) \big| \Gamma_{kl}^T \mathbf{V}, \delta_k = 1, \delta_l = 1\big] \\
&= E\big[E(X_k X_l | \mathbf{V}) \big| \Gamma_{kl}^T \mathbf{V}, \delta_k = 1, \delta_l = 1\big] \\
&= E\big[E(X_k X_l | \Gamma_{kl}^T \mathbf{V}) \big| \Gamma_{kl}^T \mathbf{V}, \delta_k = 1, \delta_l = 1\big] \\
&= E(X_k X_l | \Gamma_{kl}^T \mathbf{V}) \quad (k < l).
\end{aligned}$$

In other words, Γ_{kl} satisfies condition (2.9). Proposition 3 states that a consistent estimator of a method-specific basis Γ_{kl} of $\mathcal{S}_{E(Z^{(kl)}|\mathbf{V})}$ can be obtained using the partial central mean subspace $\mathcal{S}^{\{\delta_k=1,\delta_l=1\}}_{E(Z^{(kl)}|\mathbf{V})}$, which is the minimal partial mean subspace $\mathcal{S}$ satisfying $E\{Z^{(kl)}|\mathbf{V}, \delta_k = 1, \delta_l = 1\} = E\{Z^{(kl)}|P_{\mathcal{S}}\mathbf{V}, \delta_k = 1, \delta_l = 1\}$.

Proposition 3. *Assuming the same conditions as those in Proposition* 1*, we have* $\mathcal{S}_{E(Z^{(kl)}|\mathbf{V})} = \mathcal{S}^{\{\delta_k=1,\delta_l=1\}}_{E(Z^{(kl)}|\mathbf{V})}$*, for* $k < l$ *and* $k, l = 1, \ldots, p_{\text{1}}$.

Proposition 3 ensures that the IHT method can be applied to the completely observed dataset $\{(Z_i^{(kl)}, \mathbf{V}_i) : Z_i^{(kl)} = X_{ki}X_{li}, \delta_{ki}\delta_{li} = 1, k < l, i = 1, \ldots, n\}$ to obtain a consistent estimator $\widehat{\Gamma}_{kl}$ of Γ_{kl}.

2.3. Dimension-reduction-based imputation for SIR (DRI-SIR)

We now employ the kernel method to derive a dimension-reduction-based imputation estimator of the candidate matrix for the SIR.

Given n i.i.d observations $\{(\mathbf{X}_{mis,i}, \mathbf{X}_{obs,i}, Y_i, \delta_{1i}, \ldots, \delta_{p_{\text{1}}i})\}_{i=1}^n$, where $\mathbf{X}_{mis,i} = (X_{1i}, \ldots, X_{p_{\text{1}}i})^T$ is subject to missingness, $\mathbf{V}_i = (\mathbf{X}_{obs,i}^T, Y_i)^T = (X_{p_{\text{1}}+1,i}, \ldots, X_{pi}, Y_i)^T$ is always observed, and $\delta_{ki} = 1$ if X_{ki} is observed and $\delta_{ki} = 0$ otherwise, for $k = 1, \ldots, p_{\text{1}}$. For ease of exposition, let $M_k(\Gamma_k^T\mathbf{V}) = E(X_k|\Gamma_k^T\mathbf{V}) = E(X_k|\Gamma_k^T\mathbf{V}, \delta_k = 1), Q_k(\Upsilon_k^T\mathbf{V}) = E(X_k^2|\Upsilon_k^T\mathbf{V}) = E(X_k^2|\Upsilon_k^T\mathbf{V}, \delta_k = 1)$, and $R_{kl}(\Gamma_{kl}^T\mathbf{V}) = E(X_kX_l|\Gamma_{kl}^T\mathbf{V}) = E(X_kX_l|\Gamma_{kl}^T\mathbf{V}, \delta_k = 1, \delta_l = 1)$ $(k \neq l)$, for $1 \leq k, l \leq p_{\text{1}}$. Then, after obtaining the consistent estimators $\widehat{\Gamma}_k, \widehat{\Upsilon}_k$, and $\widehat{\Gamma}_{kl}$ of Γ_k, Υ_k, and Γ_{kl} using the method presented in Subsection 2.2, $M_k(\Gamma_k^T\mathbf{V}), Q_k(\Upsilon_k^T$

$\mathbf{V}$), and $R_{kl}(\Gamma_{kl}^T\mathbf{V})$ can be estimated nonparametrically as

$$\widehat{M}_k(\widehat{\Gamma}_k^T\mathbf{V}) = \frac{\sum_{j=1}^n K_h(\widehat{\Gamma}_k^T\mathbf{V}_j - \widehat{\Gamma}_k^T\mathbf{V})\delta_{kj}X_{kj}}{\sum_{j=1}^n K_h(\widehat{\Gamma}_k^T\mathbf{V}_j - \widehat{\Gamma}_k^T\mathbf{V})\delta_{kj}}, \tag{2.11}$$

$$\widehat{Q}_k(\widehat{\Upsilon}_k^T\mathbf{V}) = \frac{\sum_{j=1}^n K_h(\widehat{\Upsilon}_k^T\mathbf{V}_j - \widehat{\Upsilon}_k^T\mathbf{V})\delta_{kj}X_{kj}^2}{\sum_{j=1}^n K_h(\widehat{\Upsilon}_k^T\mathbf{V}_j - \widehat{\Upsilon}_k^T\mathbf{V})\delta_{kj}}, \tag{2.12}$$

$$\widehat{R}_{kl}(\widehat{\Gamma}_{kl}^T\mathbf{V}) = \frac{\sum_{j=1}^n K_h(\widehat{\Gamma}_{kl}^T\mathbf{V}_j - \widehat{\Gamma}_{kl}^T\mathbf{V})\delta_{kj}\delta_{lj}X_{kj}X_{lj}}{\sum_{j=1}^n K_h(\widehat{\Gamma}_{kl}^T\mathbf{V}_j - \widehat{\Gamma}_{kl}^T\mathbf{V})\delta_{kj}\delta_{lj}}, \tag{2.13}$$

where $K_h(\mathbf{u}) = h^{-r}\prod_{i=1}^r K(u_i/h)$ is a multivariate product kernel function, with r denoting the dimension of $\widehat{\Gamma}_k^T\mathbf{V}, \widehat{\Upsilon}_k^T\mathbf{V}$, or $\widehat{\Gamma}_{kl}^T\mathbf{V}$, and the bandwidth h might take different values when it appears in different palaces.

From (2.11) to (2.13), the dimension-reduction-based imputation estimators $\widehat{E}(X_k)$, $\widehat{E}(X_k^2)$, and $\widehat{E}(X_kX_l)$ $(k \neq l)$ of $E(X_k)$, $E(X_k^2)$, and $E(X_kX_l)$ with $1 \leq k, l \leq p_1$, can be respectively expressed as

$$\widehat{E}(X_k) = n^{-1}\sum_{i=1}^n\{\delta_{ki}X_{ki} + (1-\delta_{ki})\widehat{M}_k(\widehat{\Gamma}_k^T\mathbf{V}_i)\}, \tag{2.14}$$

$$\widehat{E}(X_k^2) = n^{-1}\sum_{i=1}^n\{\delta_{ki}X_{ki}^2 + (1-\delta_{ki})\widehat{Q}_k(\widehat{\Upsilon}_k^T\mathbf{V}_i)\}, \tag{2.15}$$

$$\widehat{E}(X_kX_l) = n^{-1}\sum_{i=1}^n\{\delta_{ki}\delta_{li}X_{ki}X_{li} + (1-\delta_{ki}\delta_{li})\widehat{R}_{kl}(\widehat{\Gamma}_{kl}^T\mathbf{V}_i)\}. \tag{2.16}$$

Based on (2.3), $E(X_k\mathbf{X}_{obs}^T)$ can be estimated using

$$\widehat{E}(X_k\mathbf{X}_{obs}^T) = n^{-1}\sum_{i=1}^n\{\delta_{ki}X_{ki} + (1-\delta_{ki})\widehat{M}_k(\widehat{\Gamma}_k^T\mathbf{V}_i)\}\mathbf{X}_{obs,i}^T. \tag{2.17}$$

Let $T_k(Y) = E(X_k|Y)$. Based on (2.4), we can estimate $T_k(Y)$ using

$$\widehat{T}_k(Y) = n^{-1}\sum_{j=1}^n \frac{K_h(Y_j - Y)\{\delta_{kj}X_{kj} + (1-\delta_{kj})\widehat{M}_k(\widehat{\Gamma}_k^T\mathbf{V}_j)\}}{\widehat{f}(Y)}, \tag{2.18}$$

where $\widehat{f}(Y) = n^{-1}\sum_{j=1}^n K_h(Y_j - Y)$ is the kernel estimator of the density function of Y. Consequently, the kl-th element of $E\{E(\mathbf{X}_{mis}|Y)E(\mathbf{X}_{mis}^T|Y)\}$ can be estimated using $n^{-1}\sum_{i=1}^n \widehat{T}_k(Y_i)\widehat{T}_l(Y_i)$. In addition, let $H(Y) = E(\mathbf{X}_{obs}^T|Y)$, with its kernel estimator given by

$$\widehat{H}(Y) = n^{-1}\sum_{i=1}^n \frac{K_h(Y_i - Y)\mathbf{X}_{obs,i}^T}{\widehat{f}(Y)}. \tag{2.19}$$

Then, we use $n^{-1}\sum_{i=1}^{n}\widehat{T}_k(Y_i)\widehat{H}(Y_i)$ to estimate $E\{E(X_k|Y)E(\mathbf{X}_{obs}^T|Y)\}$, which is the k-th raw of $E\{E(\mathbf{X}_{mis}|Y)E(\mathbf{X}_{obs}^T|Y)\}$.

Finally, by replacing the expectations and conditional expectations in Φ_0, Φ_1, and Φ_2 with their corresponding estimators, we obtain an estimator of the candidate matrix for the SIR, say $\widehat{\Sigma}_{\mathbf{X}}^{-1}\widehat{\Sigma}_{E(\mathbf{X}|Y)}$, where $\widehat{\Sigma}_{\mathbf{X}}=\widehat{\Phi}_1-\widehat{\Phi}_0$ and $\widehat{\Sigma}_{E(\mathbf{X}|Y)}=\widehat{\Phi}_2-\widehat{\Phi}_0$. For ease of exposition, $\widehat{\Sigma}_{\mathbf{X}}^{-1}\widehat{\Sigma}_{E(\mathbf{X}|Y)}$, constructed using our proposed method, is called the DRI-SIR estimator of $\Sigma_{\mathbf{X}}^{-1}\Sigma_{E(\mathbf{X}|Y)}$. Then, the eigenvectors corresponding to the first d largest nonzero eigenvalues of $\widehat{\Sigma}_{\mathbf{X}}^{-1}\widehat{\Sigma}_{E(\mathbf{X}|Y)}$ form an estimator of the CS $\mathcal{S}_{Y|\mathbf{X}}$.

Remark 1. If the dimension of $\mathcal{S}_{E(X_k|\mathbf{V})}, \mathcal{S}_{E(\widetilde{X}_k|\mathbf{V})}$, or $\mathcal{S}_{E(Z^{(kl)}|\mathbf{V})}$ (i.e., $\Gamma_k^T\mathbf{V}$, $\Upsilon_k^T\mathbf{V}$, or $\Gamma_{kl}^T\mathbf{V}$) is greater than three, our method might not perform well. In fact, existing dimension-reduction techniques only partly solve the high-dimension problems. In some cases, the dimension might be as small as one, two or three. However, in other cases, the dimension may be larger than three, but smaller than the dimension of the predictors, in which case the subsequent statistical inference can be improved, but might not work well. In fact, it is a common problem that existing dimension-reduction techniques are limited when the structural dimension is not small. On the other hand, as illustrated in several studies, the low structural dimension might be sufficient for many practical problems. For example, Cook (1998b) analyzed motor octane data, selecting a structural dimension of one using his proposed chi-square test. Xia et al. (2002) chose a dimension of two for Hitter's salary data using cross-validation. Ma and Zhu (2012) used bootstrap procedure to determine a dimension of one for employee's salary data from the Fifth National Bank of Springfield. Zhu et al. (2011) also noted that, for the purpose of dimension reduction, the structural dimension is, in general, assumed to be small, taking values one, two or three.

2.4. Asymptotic properties

In this section, we study the asymptotic behavior of the proposed DRI-SIR estimator. Let $f_{_0}(\cdot), f_k(\cdot), \tilde{f}_k(\cdot)$, and $f_{kl}(\cdot)$ respectively denote the density functions of $Y, \Gamma_k^T\mathbf{V}, \Upsilon_k^T\mathbf{V}$, and $\Gamma_{kl}^T\mathbf{V}$ $(k\neq l)$, for $k,l=1,\ldots,p_{_1}$. For ease of interpretation, we also introduce the following notations:

$$\pi_k(\Gamma_k^T\mathbf{V})=P(\delta_k=1|\Gamma_k^T\mathbf{V}), \qquad \widetilde{\pi}_k(\Upsilon_k^T\mathbf{V})=P(\delta_k=1|\Upsilon_k^T\mathbf{V}),$$
$$\pi_{kl}(\Gamma_{kl}^T\mathbf{V})=P(\delta_k\delta_l=1|\Gamma_{kl}^T\mathbf{V}), \qquad m_k(\Gamma_k^T\mathbf{V})=E(\delta_k X_k|\Gamma_k^T\mathbf{V}),$$
$$q_k(\Upsilon_k^T\mathbf{V})=E(\delta_k X_k^2|\Upsilon_k^T\mathbf{V}), \qquad w_{kl}(\Gamma_{kl}^T\mathbf{V})=E(\delta_k\delta_l X_k X_l|\Gamma_{kl}^T\mathbf{V}),$$

$$g_k(\Gamma_k^T\mathbf{V}) = \pi_k(\Gamma_k^T\mathbf{V})f_k(\Gamma_k^T\mathbf{V}), \qquad G_k(\Gamma_k^T\mathbf{V}) = m_k(\Gamma_k^T\mathbf{V})f_k(\Gamma_k^T\mathbf{V}),$$
$$a_k(\Upsilon_k^T\mathbf{V}) = \widetilde{\pi}_k(\Upsilon_k^T\mathbf{V})\tilde{f}_k(\Upsilon_k^T\mathbf{V}), \qquad A_k(\Upsilon_k^T\mathbf{V}) = q_k(\Upsilon_k^T\mathbf{V})\tilde{f}_k(\Upsilon_k^T\mathbf{V}),$$
$$b_{kl}(\Gamma_{kl}^T\mathbf{V}) = \pi_{kl}(\Gamma_k^T\mathbf{V})f_{kl}(\Gamma_{kl}^T\mathbf{V}), \qquad B_{kl}(\Gamma_{kl}^T\mathbf{V}) = w_{kl}(\Gamma_k^T\mathbf{V})f_{kl}(\Gamma_{kl}^T\mathbf{V}),$$
$$S_k(Y) = T_k(Y)f_0(Y), \qquad W(Y) = H(Y)f_0(Y).$$

Next,we list a set of regularity conditions to facilitate our technical derivations of the main results.

Condition 1. The symmetric and continuous kernel function $K(\cdot)$ has support in the interval $[-1,1]$. Moreover, for some positive integer m, the function satisfies $\int_{-1}^{1} K(u)du = 1, \int_{-1}^{1} u^i K(u)du = 0$, with $1 \le i \le m-1, 0 \ne \int_{-1}^{1} |u|^m K(u)du < \infty$, and $\int_{-1}^{1} K^2(u)du < \infty$.

Condition 2. The $(m-1)$-th-order derivatives of the functions $f_0(\cdot), f_k(\cdot), \tilde{f}_k(\cdot), f_{kl}(\cdot), g_k(\cdot), G_k(\cdot), a_k(\cdot), A_k(\cdot), b_{kl}(\cdot), B_{kl}(\cdot), T_k(\cdot), S_k(\cdot), H(\cdot)$, and $W(\cdot)$ are locally Lipschitz continuous.

Condition 3. The bandwidth h satisfies $nh^{2m} \to 0, nh^{2(r_k+1)}/(\log n)^2 \to \infty$, $nh^{2(\tilde{r}_k+1)}/(\log n)^2 \to \infty$, and $nh^{2(r_{kl}+1)}/(\log n)^2 \to \infty$ $(k \ne l)$ as $n \to \infty$ and $h \to 0$, for $k, l = 1, \ldots, p_1$, where $r_k = \dim\{\mathcal{S}_{E(X_k|\mathbf{V})}\}, \tilde{r}_k = \dim\{\mathcal{S}_{E(\widetilde{X}_k|\mathbf{V})}\}$, and $r_{kl} = \dim\{\mathcal{S}_{E(Z^{(kl)}|\mathbf{V})}\}$.

Condition 4. $f_0(\cdot), g_k(\cdot), a_k(\cdot)$, and $b_{kl}(\cdot)$ have compact supports, and there exist positive constants c_1, c_2, c_3, and c_4 such that $\inf_y f_0(y) \ge c_1$, $\inf_{\Gamma_k^T\mathbf{V}} g_k(\Gamma_k^T\mathbf{V}) \ge c_2$, $\inf_{\Upsilon_k^T\mathbf{V}} a_k(\Upsilon_k^T\mathbf{V}) \ge c_3$, and $\inf_{\Gamma_{kl}^T\mathbf{V}} b_{kl}(\Gamma_{kl}^T\mathbf{V}) \ge c_4$.

Condition 5. Each entry in $\mathbf{X}\mathbf{X}^T$ has a finite fourth-order moment.

Here, we briefly discuss these conditions. Condition 1 is commonly used in the literature. Condition 2 presents the smooth properties of density functions and regression curves. Condition 3 is needed technically for Lemmas B.1–B.3 in the Appendix to ensure the desired convergence rate. In particular, condition 3 indicates that $m \ge 4$ is required in our method to reduce the order of the bias of the kernel estimators such that the $\sqrt{n}$ rate of consistence can be achieved. Condition 4 is widely used in the literature to avoid the boundary effect of the related kernel estimators. Condition 5 assumes several finite moments, and is necessary for asymptotic normality.

Theorem 1. *Suppose that the MAR assumption* (1.1) *and the regularity conditions 1–5 hold, and that the dimensions* $r_k, \tilde{r}_k$, *and* r_{kl} *of the subspaces* $\mathcal{S}_{E(X_k|\mathbf{V})}$, $\mathcal{S}_{E(\widetilde{X}_k|\mathbf{V})}$, *and* $\mathcal{S}_{E(Z^{(kl)}|\mathbf{V})}$ $(k \ne l)$, *respectively are known. Then, we have the following:*

(i) $\sqrt{n}\{vec(\widehat{\Sigma}_{\mathbf{X}}^{-1}\widehat{\Sigma}_{E(\mathbf{X}|Y)}) - vec(\Sigma_{\mathbf{X}}^{-1}\Sigma_{E(\mathbf{X}|Y)})\}$ *converges in distribution to a multivariate normal distribution with mean* $\mathbf{0}$ *as* $n \to \infty$, *where "vec" denotes*

an operator that stacks all columns of a matrix to a vector.

(ii) We further assume that the linearity condition and the coverage condition hold, and that the dimension d of the CS $\mathcal{S}_{Y|\mathbf{X}}$ is known. Let $\hat{\beta}_1, \ldots, \hat{\beta}_d$ denote the eigenvectors corresponding to the first d nonzero eigenvalues of $\widehat{\Sigma}_{\mathbf{X}}^{-1}\widehat{\Sigma}_{E(\mathbf{X}|Y)}$. Then, Span$\{\hat{\beta}_1, \ldots, \hat{\beta}_d\}$ is a $\sqrt{n}$-consistent estimator of $\mathcal{S}_{Y|\mathbf{X}}$.

When $r_k, \widetilde{r}_k, r_{kl}$, and d are unknown, but their respective consistent estimators $\widehat{r}_k, \widehat{\widetilde{r}}_k, \widehat{r}_{kl}$, and $\widehat{d}$, are available, that is, $\widehat{r}_k \to r_k, \widehat{\widetilde{r}}_k \to \widetilde{r}_k, \widehat{r}_{kl} \to r_{kl}$, and $\widehat{d} \to d$ in probability, the proposed estimator of $\mathcal{S}_{Y|\mathbf{X}}$ is still $\sqrt{n}$-consistent.

2.5. Estimation of the structural dimension

The structural dimension d of the CS $\mathcal{S}_{Y|\mathbf{X}}$ is, in general, unknown, and thus needs to be estimated. Here,we employ the modified Bayes information criterion (BIC), initially developed by (Zhu, Miao and Peng (2006)) and later modified by Zhu et al. (2010), to estimate the true dimension d of $\mathcal{S}_{Y|\mathbf{X}}$:

$$\widehat{d} = \arg\max_{s=1,\ldots,p} \left\{ \frac{n}{2} \times \frac{\sum_{i=1}^{s}\{\log(\hat{\lambda}_i + 1) - \hat{\lambda}_i\}}{\sum_{i=1}^{p}\{\log(\hat{\lambda}_i + 1) - \hat{\lambda}_i\}} - C_n \times \frac{s(s+1)}{p} \right\}, \tag{2.20}$$

where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$ are the eigenvalues of $\widehat{\Sigma}_{\mathbf{X}}^{-1}\widehat{\Sigma}_{E(\mathbf{X}|Y)}$, and C_n is a penalty constant. Theorem 2 states the consistency of the estimated dimension $\widehat{d}$ in the presence of missing predictors.

Theorem 2. *Suppose that $\lim_{n\to\infty} C_n/n = 0$ and $\lim_{n\to\infty} C_n = \infty$. If the conditions in Theorem 1 hold, then $\widehat{d}$ converges to d in probability.*

The proof of Theorem 2 is similar to that presented by Zhu et al. (2010) and, hence, is omitted here.

The choice of C_n remains an open problem. Zhu, Miao and Peng (2006) recommend a practical form $C_n = c^{-1}W_n$, where c denotes the number of observations per slice and $W_n = a\log(n) + bn^{1/3}$, for some scalar constants a and b. In fact, c^{-1} can be absorbed into a and b. In our simulation studies, we choose $C_n = 6\log(n) + 3n^{1/3}$ in all of the model settings.

Remark 2. Our proposed method also needs to estimate the generally unknown dimensions $r_k = \dim\{\mathcal{S}_{E(X_k|\mathbf{V})}\}, \widetilde{r}_k = \dim\{\mathcal{S}_{E(\widetilde{X}_k|\mathbf{V})}\}$, and $r_{kl} = \dim\{\mathcal{S}_{E(Z^{(kl)}|\mathbf{V})}\}$ $(k \neq l)$, for $k, l = 1, \ldots, p_1$. Their respective consistent estimators, $\widehat{r}_k, \widehat{\widetilde{r}}_k$, and $\widehat{r}_{kl}$ can be obtained by solving similar minimization problems to that given in (2.20). The major change is that we substitute $\hat{\lambda}_i$ in (2.20) with the eigenvalues of the estimated candidate matrices for the corresponding subspaces $\mathcal{S}_{E(X_k|\mathbf{V})}, \mathcal{S}_{E(\widetilde{X}_k|\mathbf{V})}$, and $\mathcal{S}_{E(Z^{(kl)}|\mathbf{V})}$.

3. Simulation Studies

In this section, we check the finite-sample performance of the proposed DRI-SIR estimator. In our simulations, we also compare the results with those of five other estimations:

- Full-SIR Without missingness, the SIR is applied to all n observations.
- CC-SIR The subjects with missing values are removed, and the SIR is applied to the remaining completely observed data.
- AIPW-SIR in Li and Lu (2008) Only one missingness indicator δ is introduced; $\delta = 1$ if there is no missingness for all of the predictors, and 0 otherwise.
- MAIPW-SIR in Li and Lu (2008) Here, $p_{_1}$ missingness indicators $(\delta_1, \ldots, \delta_{p_{_1}})$ are introduced; $\delta_k = 1$ if there is no missingness for the k-th component X_k of $\mathbf{X}_{mis}$, and 0 otherwise, for $k = 1, \ldots, p_{_1}$.
- PI-SIR in Zhu, Wang and Zhu (2012).

(i) Evaluation Criteria.

We assess the performance of the above estimators from two aspects. First, assuming that the structural dimension d of $\mathcal{S}_{Y|\mathbf{X}}$ is known, we use the *trace correlation coefficient* (TCC; Hooper (1959)) to measure the closeness between the estimated subspace and the true subspace. Let $B_{p\times d}$ be the true basis matrix of $\mathcal{S}_{Y|\mathbf{X}}$. For an estimator $\widehat{B}$ of B, the TCC is defined as the positive square root of $r^2 = d^{-1}\sum_{i=1}^{d}\phi_i^2$, where $1 \geq \phi_1^2 \geq \phi_2^2 \geq \cdots \phi_d^2 \geq 0$ are the eigenvalues of the matrix $\widehat{B}_0^T(B_0B_0^T)\widehat{B}_0$, with $\widehat{B}_0$ and B_0 denoting the orthonormalized versions of $\widehat{B}$ and B, respectively. A TCC closer to one indicates a better estimate of the CS. Second, assuming d is unknown, we report the empirical distribution (in percentages) of the estimated dimension $\widehat{d}$ to evaluate the efficacy of the various methods in determining the structural dimension.

(ii) Simulation Settings.

The simulations for each model are repeated 500 times, where each sample is of size $n = 400$. We set the slice number to $H = 10$, as required for the Full-SIR, CC-SIR, AIPW-SIR, and MAIPW-SIR. For the PI-SIR and the proposed DRI-SIR, which involve kernel smoothing, we use a multivariate product kernel $K_h(\mathbf{u}) = h^{-r}\prod_{i=1}^{r}K(u_i/h)$, where $K(u) = (2\pi)^{-1/2}\exp(-u^2/2)$ and r is the dimension of the kernel. Because the kernel method for a global estimator is

insensitive to the choice of bandwidth (Wang and Rao (2002)), we simply choose the classical bandwidth $h \propto n^{-1/(4+r)}$.

For comparison purposes, we consider the following three models:

$$Y = (\beta_1^T \mathbf{X})(\beta_1^T \mathbf{X} + \beta_2^T \mathbf{X} + 3) + 0.5\varepsilon, \tag{3.1}$$

$$Y = \frac{\beta_1^T \mathbf{X}}{(\beta_2^T \mathbf{X} + 1.5)^2 + 0.5} + 0.5\varepsilon, \tag{3.2}$$

$$Y = 0.5\beta_1^T \mathbf{X} + (\beta_2^T \mathbf{X} + 2)\varepsilon, \tag{3.3}$$

where $\mathbf{X} = (X_1, \ldots, X_p)^T$ follows a multivariate normal distribution, with mean $\mathbf{0}$ and covariance $0.3^{|k-l|}$ between X_k and X_l, with $1 \le k, l \le p$, and ε follows a standard normal distribution and is independent of $\mathbf{X}$. The predictor effects exist only in the conditional mean of $Y|\mathbf{X}$ for models (3.1)–(3.2), but also appear in the conditional variance of model (3.3). We set $p = 15$, $p_1 = 3, 5$, and 10 (the dimension of the missing predictors), $\beta_1 = (0.5 \times \mathbf{1}_{p_1-1}, \mathbf{0}_{p-p_1-2}, 0.5, -1, -1)^T$, and $\beta_2 = (\mathbf{0}_{p_1-1}, 0.5, -0.5, -0.5, 0.5, 0.5, 0.5, \mathbf{0}_{p-p_1-5})^T$, where $\mathbf{1}_s$ and $\mathbf{0}_s$ are $1 \times s$ vectors with all elements being one and zero respectively. For the three models, the CS $\mathcal{S}_{Y|\mathbf{X}} = \text{Span}\{\beta_1, \beta_2\}$ and, thus, the true structural dimension is $d = 2$. In addition to the logistic linear missingness mechanism,

$$P(\delta_k = 1|\mathbf{V}) = \frac{\exp(c_0 + \gamma_0^T \mathbf{V})}{1 + \exp(c_0 + \gamma_0^T \mathbf{V})} \quad k = 1, \ldots, p_1, \tag{3.4}$$

we also consider the logistic quadratic missingness mechanism,

$$P(\delta_k = 1|\mathbf{V}) = \frac{\exp(c_0 + (\gamma_1^T \mathbf{V})^2 + \gamma_2^T \mathbf{V})}{1 + \exp(c_0 + (\gamma_1^T \mathbf{V})^2 + \gamma_2^T \mathbf{V})} \quad k = 1, \ldots, p_1, \tag{3.5}$$

where $\mathbf{V} = (X_{p_1+1}, \ldots, X_p, Y)^T$ is always observed, and c_0 is a scalar constant to control the missing proportion. Here, the same form of model (3.4) or (3.5) is used for $P(\delta_k = 1|\mathbf{V})$ with different k, which does not affect the performance evaluation of the proposed method. To investigate the effect of different missing proportions on the efficacy of the various methods, we take three values of c_0 for each case to control the corresponding missing proportions around 20%, 35%, and 50%. We set $\gamma_0 = (-1, -1, -1, 0, \ldots, 0, 0.5, 0.5, 0.25)^T$ with $q - 6$ zeros, $\gamma_1 = (0.5, 0, \ldots, 0, -1, 0.25)^T$ with $q - 3$ zeros, and $\gamma_2 = (0, 1, \ldots, 1, 0, 0)^T$ with $q - 3$ ones.

As discussed in the introduction, it is necessary to assume parametric models when implementing the AIPW-SIR, MAIPW-SIR, or PI-SIR. From a theoretical point of view , we should evaluate the performance of these three methods for two cases, where all or some of the required parametric models are specified correctly.

Table 1. Comparison of the median TCC of the SDR estimations for model (3.1) under the missingness mechanism in (3.4), with different p_1 and missing proportions (mp).

p_1	C_0	mp	Full-SIR	**DRI-SIR**	CC-SIR	AIPW-SIR	MAIPW-SIR	PI-SIR
3	1.7	20.54%	0.9518	**0.9470**	0.8935	0.9311	0.9381	0.7473
	0.4	35.49%	0.9518	**0.9375**	0.8102	0.9063	0.9145	0.6943
	−0.7	50.03%	0.9530	**0.9242**	0.6731	0.8372	0.7979	0.7354
5	1.6	20.84%	0.9477	**0.9378**	0.8301	0.9007	0.9186	0.6446
	0.3	35.33%	0.9480	**0.9266**	0.6919	0.8337	0.8428	0.6192
	−0.9	50.62%	0.9480	**0.9058**	0.5592	0.6416	0.6570	0.6119
10	1.5	20.63%	0.9476	**0.9316**	0.7258	0.8201	0.8906	0.5314
	0.2	34.96%	0.9484	**0.9034**	0.5770	0.5915	0.6833	0.6133
	−1	50.27%	0.9476	**0.8380**	0.5077	0.4978	0.4808	0.6381

However, it is nontrivial to specify parametric models correctly for all of the quantities, including $E(X_k|\mathbf{V})$, $E(X_k^2|\mathbf{V})$, $E(X_kX_l|\mathbf{V})$, and $P(\delta_k = 1|\mathbf{V})$ for the AIPW-SIR and MAIPW-SIR, and $P(\delta_k = 1|\mathbf{V})$, $E(\delta_kX_k|\mathbf{V})$, $E(\delta_kX_k^2|\mathbf{V})$, $E(\delta_k\delta_lX_kX_l|\mathbf{V})$, and $E(\delta_k\delta_l|\mathbf{V})$ for the PI-SIR. Existing studies consider only the case in which the missingness mechanism is specified correctly, without regard for the correctness of other involved parametric models. Thus, we conduct our comparisons under two special cases, where $P(\delta_k = 1|\mathbf{V})$, for $k = 1, \ldots, p_1$, are specified correctly and incorrectly, respectively. Specifically, regardless of whether the missingness mechanism in (3.4) or (3.5) holds, we always use a logistic linear form of the missingness mechanism when implementing the AIPW-SIR, MAIPW-SIR and PI-SIR. Then, it corresponds to a correct specification of the missingness mechanism if the missingness mechanism (3.4) is true, and a misspecification of the missingness mechanism otherwise.

(iii) Simulation Results.

Tables 1–2 give the simulation results under model (3.1) with the missingness mechanism (3.4).

Table 1 reports the median TCCs between the true subspace and the estimated subspace for each method, with different p_1 and three missing proportions, in 500 replications. Here, the missingness mechanism is specified correctly. First, in most situations, the proposed DRI-SIR performs uniformly better than the CC-SIR, AIPW-SIR, MAIPW-SIR, and PI-SIR do, and even shows comparable performance to that of the Full-SIR under small missing proportions. Second, as the missing proportion increases, the CC-SIR, AIPW-SIR, MAIPW-SIR, and PI-SIR perform increasingly poorly, but our method is relatively robust. Third, even though the missing proportion exceeds 50%, the DRI-SIR still performs

Table 2. Distribution (in percentages) of the estimated structural dimension for model (3.1) under the missingness mechanism in (3.4), with different p_1 and missing proportions (mp).

p_1	Method	$\widehat{d}$	1	2	>2	1	2	>2	1	2	>2
			mp=20.54%			mp=35.49%			mp=50.03%		
3	Full-SIR		0.0020	**0.9980**	0.0000	0.0020	**0.9980**	0.0000	0.0040	**0.9960**	0.0000
	DRI-SIR		0.0000	**1.0000**	0.0000	0.0000	**1.0000**	0.0000	0.0000	**0.9980**	0.0020
	CC-SIR		0.2180	**0.7820**	0.0000	0.6720	**0.3280**	0.0000	0.9780	**0.0220**	0.0000
	AIPW-SIR		0.0080	**0.9760**	0.0160	0.0140	**0.9360**	0.0500	0.0280	**0.8120**	0.1600
	MAIPW-SIR		0.0060	**0.9900**	0.0040	0.0200	**0.9280**	0.0520	0.0580	**0.7540**	0.1880
	PI-SIR		0.1700	**0.6620**	0.1680	0.1200	**0.5920**	0.2880	0.0880	**0.5880**	0.3240
			mp=20.84%			mp=35.33%			mp=50.62%		
5	Full-SIR		0.0000	**1.0000**	0.0000	0.0020	**0.9980**	0.0000	0.0020	**0.9980**	0.0000
	DRI-SIR		0.0000	**1.0000**	0.0000	0.0000	**1.0000**	0.0000	0.0000	**0.9940**	0.0060
	CC-SIR		0.4860	**0.5140**	0.0000	0.9240	**0.0760**	0.0000	0.9960	**0.0040**	0.0000
	AIPW-SIR		0.0100	**0.9500**	0.0400	0.0320	**0.8580**	0.1100	0.1060	**0.6360**	0.2580
	MAIPW-SIR		0.0100	**0.9620**	0.0280	0.0240	**0.8220**	0.1540	0.0760	**0.5240**	0.4000
	PI-SIR		0.2780	**0.4660**	0.2560	0.1980	**0.4780**	0.3240	0.1740	**0.4680**	0.3580
			mp=20.63%			mp=34.96%			mp=50.27%		
10	Full-SIR		0.0000	**1.0000**	0.0000	0.0040	**0.9960**	0.0000	0.0000	**1.0000**	0.0000
	DRI-SIR		0.0000	**1.0000**	0.0000	0.0000	**1.0000**	0.0000	0.0060	**0.9060**	0.0880
	CC-SIR		0.8980	**0.1020**	0.0000	0.9920	**0.0080**	0.0000	1.0000	**0.0000**	0.0000
	AIPW-SIR		0.0260	**0.9100**	0.0640	0.3060	**0.5020**	0.1920	0.2200	**0.5240**	0.2560
	MAIPW-SIR		0.0120	**0.9520**	0.0360	0.0320	**0.7120**	0.2560	0.1060	**0.4180**	0.4760
	PI-SIR		0.2420	**0.5280**	0.2300	0.2020	**0.5680**	0.2300	0.1780	**0.6040**	0.2180

well, especially with the low-dimensional missing predictors, whereas the CC-SIR, AIPW-SIR, MAIPW-SIR and PI-SIR, perform quite poorly.

Table 2 reports the empirical distribution of the estimated dimension $\widehat{d}$ for each method, with different p_1 and three missing proportions, over 500 replications. In nearly all cases, the proposed DRI-SIR selects the true structural dimension with a probability much close to one, clearly outperforming the other methods. In fact, $\widehat{d}$ obtained using the modified BIC in (2.20) is determined by both the penalty constant C_n and the eigenvalues of the estimated candidate matrix. Our finite simulation studies also reveal that a well-chosen C_n may result in the performance of $\widehat{d}$ coinciding well with that of the estimated candidate matrix for $\mathcal{S}_{Y|\mathbf{X}}$. However, as noted in Section 2.5, the choice of C_n requires further research.

The simulation results under model (3.1) with the missingness mechanism in (3.5) are given in Tables 3–4. In this case, the missingness mechanism is misspecified for the AIPW-SIR, MAIPW-SIR, and PI-SIR. The proposed DRI-SIR performs uniformly better than the CC-SIR, AIPW-SIR, MAIPW-SIR, and

Table 3. Comparison of the median TCC of the SDR estimations for model (3.1) under the missingness mechanism in (3.5), with different p_1 and missing proportions (mp).

p_1	C_0	mp	Full-SIR	**DRI-SIR**	CC-SIR	AIPW-SIR	MAIPW-SIR	PI-SIR
3	1.5	20.06%	0.9532	**0.9524**	0.9180	0.9008	0.9270	0.6730
	−0.9	35.18%	0.9535	**0.9454**	0.8802	0.8172	0.8784	0.6274
	−3.3	50.74%	0.9542	**0.9355**	0.8304	0.7160	0.7832	0.6334
5	1.3	19.86%	0.9483	**0.9434**	0.8966	0.8573	0.9091	0.6319
	−1.1	35.41%	0.9480	**0.9387**	0.8430	0.7158	0.8420	0.5485
	−3.3	50.42%	0.9480	**0.9255**	0.7858	0.5996	0.6890	0.5057
10	0.1	20.56%	0.9490	**0.9406**	0.8621	0.8195	0.9220	0.9014
	−1.6	35.03%	0.9465	**0.9272**	0.8251	0.7307	0.8789	0.8140
	−3.4	49.84%	0.9465	**0.9068**	0.7778	0.4410	0.8000	0.6159

Table 4. Distribution (in percentages) of the estimated structural dimension for model (3.1) under the missingness mechanism in (3.5), with different p_1 and missing proportions (mp).

p_1	Method	$\widehat{d}$	1	2	>2	1	2	>2	1	2	>2
			mp=20.06%			mp=35.18%			mp=50.74%		
3	Full-SIR		0.0000	**1.0000**	0.0000	0.0000	**1.0000**	0.0000	0.0000	**1.0000**	0.0000
	DRI-SIR		0.0000	**1.0000**	0.0000	0.0000	**1.0000**	0.0000	0.0000	**1.0000**	0.0000
	CC-SIR		0.1180	**0.8820**	0.0000	0.3200	**0.6800**	0.0000	0.7140	**0.2860**	0.0000
	AIPW-SIR		0.0400	**0.8020**	0.1580	0.0700	**0.7180**	0.2120	0.0620	**0.6480**	0.2900
	MAIPW-SIR		0.0200	**0.8960**	0.0840	0.0460	**0.7840**	0.1700	0.0500	**0.6800**	0.2700
	PI-SIR		0.1780	**0.5840**	0.2380	0.2300	**0.4760**	0.2940	0.1340	**0.4620**	0.4040
			mp=19.86%			mp=35.41%			mp=50.42%		
5	Full-SIR		0.0000	**1.0000**	0.0000	0.0020	**0.9980**	0.0000	0.0020	**0.9980**	0.0000
	DRI-SIR		0.0000	**1.0000**	0.0000	0.0000	**1.0000**	0.0000	0.0000	**1.0000**	0.0000
	CC-SIR		0.1360	**0.8640**	0.0000	0.4440	**0.5560**	0.0000	0.8000	**0.2000**	0.0000
	AIPW-SIR		0.0580	**0.7280**	0.2140	0.0740	**0.6600**	0.2660	0.1000	**0.5280**	0.3720
	MAIPW-SIR		0.0180	**0.8960**	0.0860	0.0320	**0.7720**	0.1960	0.0620	**0.6000**	0.3380
	PI-SIR		0.2820	**0.5260**	0.1920	0.2780	**0.5100**	0.2120	0.2740	**0.4900**	0.2360
			mp=20.56%			mp=35.03%			mp=49.84%		
10	Full-SIR		0.0000	**1.0000**	0.0000	0.0000	**1.0000**	0.0000	0.0000	**1.0000**	0.0000
	DRI-SIR		0.0000	**1.0000**	0.0000	0.0020	**0.9980**	0.0000	0.0000	**0.9880**	0.0120
	CC-SIR		0.0460	**0.9540**	0.0000	0.1760	**0.8240**	0.0000	0.4920	**0.5080**	0.0000
	AIPW-SIR		0.0140	**0.8700**	0.1160	0.0360	**0.7860**	0.1780	0.1500	**0.5220**	0.3280
	MAIPW-SIR		0.0000	**0.9800**	0.0200	0.0000	**0.9560**	0.0440	0.0140	**0.8660**	0.1200
	PI-SIR		0.0700	**0.9180**	0.0120	0.2460	**0.7340**	0.0200	0.4920	**0.4400**	0.0680

PI-SIR in all simulation settings. In particular, the AIPW-SIR, MAIPW-SIR, and PI-SIR even perform worse than the CC-SIR in most cases.

The simulation results under models (3.2)–(3.3) with the missingness mechanisms in (3.4)–(3.5), given in Tables 5–12 of the supplementary material, indicate similar features to those of Tables 1-4. The results reinforce the general quanti-

tative patterns observed in Tables 1-4, and show the superiority of our proposed DRI-SIR over other methods.

4. Real-Data Analysis

Here, We apply the proposed method to an automobile data set, which is available from the Machine Learning Repository at the University of California-Irvine (`http://mlr.cs.umass.edu/ml/datasets/Automobile`). Our primary goal is to describe the relationship between the car price and a set of car attributes. We choose 14 features with continuous values as predictors, including normalized losses, wheelbase, length, width, height, curb-weight, engine size, bore, stroke, compression ratio, horsepower, peak-rpm, city-mpg, and highway-mpg. The response Y is the logarithm of the car price. The original data set consists of 205 sample points. For simplicity, we first remove four sample points with missing responses. For bore and stroke, four sample points contain missing values, and for horsepower and peak-rpm, two sample points have missing values. Because the number of sample points with missing values in these four predictors is very small relative to the sample size, we simply delete these six points. Among the remaining 195 observations, 35 observations have missing values in the predictor "normalized losses." To eliminate the influence of the scale, we standardize each predictor. In particular, for the missing predictor "normalized losses," standardization is only implemented for the 160 completely observed data points.

The structural dimension of $\mathcal{S}_{Y|\mathbf{X}}$ is chosen as one, from (2.20). Using our proposed method, the first dimension-reduction direction is estimated as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{DRI-SIR} = (&0.0598, -0.1448, 0.2465, -0.2284, -0.0745, 0.5577, -0.0483,\\ &0.0229, 0.1325, -0.1851, -0.5381, -0.0739, 0.3152, -0.3112)^T.\end{aligned}$$

Figure 1 shows a scatterplot of the log price versus the estimated linear combination of the standardized predictors using only the 160 completely observed data points. The figure shows a significant linear trend. This indicates that our method not only reduces the dimension of the predictors effectively, but also provides a reference for modeling the parametric structure of a regression of Y on the linear combination of the standardized predictors.

5. Concluding Remarks

It is a common practice to develop imputation or inverse probability-weighted

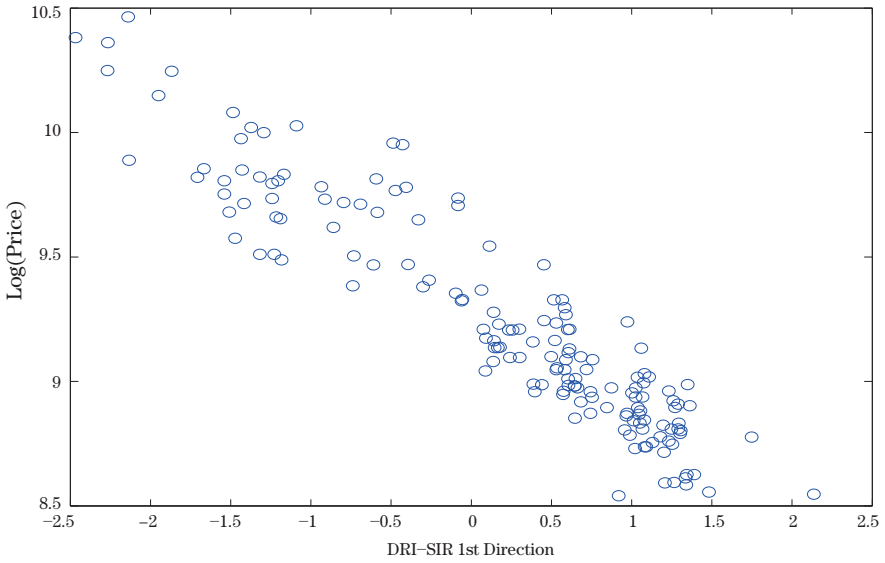

Figure 1. Sufficient summary plot for the log price versus the first linear combination of the standardized predictors based on the DRI-SIR.

methods such that the standard statistical methods for full data can be applied to the case of missing data. Our proposed method belongs to the former group. Future research interests in the SDR field might only need to focus on SDR methods for full data, because when they satisfy certain conditions, these methods can always be applied to the case of missing predictors with the aid of our proposed imputation procedure.

Our proposed method possesses typical nonparametric properties. It is quite different to existing semiparametric methods, such as the AIPW-SIR (Li and Lu (2008)) and PI-SIR (Zhu, Wang and Zhu (2012)), which assume parametric models and, hence, are difficult to apply in some practical problems. In particular, we describe the differences between the proposed DRI-SIR and the methods of Zhu, Wang and Zhu (2012), who consider the problem of the SDR with missing predictors under two types of missingness mechanisms. First, they assume all predictors are MAR, or equivalently, $\boldsymbol{\delta} \perp\!\!\!\perp \mathbf{X} \mid Y$. They consider this special case to avoid the curse of dimensionality. We take the estimation of $E(X_k)$ as an example to illustrate this point. They construct the estimator of $E(X_k)$ based on $E(X_k) = E\{E(X_k|Y)\} = E\{E(X_k|Y, \delta_k = 1)\}$ which is derived from this type of missingness mechanism. They also consider the same MAR assumption as that in

(1.1) in our paper. However, their proposed PI-SIR imposes parametric models on $E(\delta_k X_k|\mathbf{V})$, $E(\delta_k|\mathbf{V})$, $E(\delta_k X_k^2|\mathbf{V})$, $E(\delta_k\delta_l X_k X_l|\mathbf{V})$, and $E(\delta_k\delta_l|\mathbf{V})$ to avoid the curse of dimensionality. It is very difficult to specify all parametric models correctly; hence the PI-SIR runs a significant risk of misspecified parametric models. In contrast, our proposed method not only avoids assuming parametric models, but also overcomes the curse of dimensionality. This is the main reason why the numerical performance of our method is uniformly better than that of the other methods.

The proposed method can be applied in broader contexts. As pointed out by an anonymous referee, a direct extension of our work would be to estimate a general class of conditional expectations, where the variables treated as responses are MAR and the given variables are high-dimensional. The estimation of mean functionals with missing responses Cheng (1994) is the most typical example. Another important extension would be to apply all spectral-decomposition-based SDR methods to the case of predictors MAR using the proposed dimension-reduction imputation procedure. Part S1 in the supplementary material demonstrates how to extend our method to the SAVE and PHD. In addition, the proposed method works when Y is discrete or categorical. We describe the details and conduct a simulation study with a discrete response in Part S2 of the supplementary material. These extensions would greatly expand the scope of the applicability of our method.

Supplementary Materials

The supplementary material is available online. It contains two extensions of our proposed method, the simulation results under models (3.2)–(3.3) and technical proofs for Lemmas B.1–B.3 in the Appendix.

Acknowledgment

Qihua Wang is the corresponding author of this paper. We are grateful to the Editor, Associate Editor, and two referees for their constructive comments, which have helped to improve our paper. We also thank Professor Liping Zhu and Dr. Xiaobo Ding for their useful discussion and suggestions, which lead to a much improved presentation. Wang's research was supported by the National Natural Science Foundation of China (General Program 11871460 and Key Program 11331011) and the program for Creative Research Group of National Natural Science Foundation of China (Grant No. 61621003).

Appendix A: Proof of Proposition 1.

First, we show that $\mathcal{S}^{\{\delta_k=1\}}_{E(X_k|\mathbf{V})} \subseteq \mathcal{S}_{E(X_k|\mathbf{V})}$, for $k = 1, \ldots, p_1$. Suppose that A is a basis matrix of $\mathcal{S}_{E(X_k|\mathbf{V})}$, that is, $E(X_k|\mathbf{V}) = E(X_k|A^T\mathbf{V})$. Under the MAR assumption, $E(X_k|\mathbf{V}, \delta_k = 1) = E(X_k|\mathbf{V})$. Then, we have $E(X_k|\mathbf{V}, \delta_k = 1) = E(X_k|A^T\mathbf{V})$, which implies that

$$\begin{aligned} E(X_k|A^T\mathbf{V}, \delta_k = 1) &= E\{E(X_k|\mathbf{V}, \delta_k = 1)|A^T\mathbf{V}, \delta_k = 1\} \\ &= E\{E(X_k|A^T\mathbf{V})|A^T\mathbf{V}, \delta_k = 1\} \\ &= E(X_k|A^T\mathbf{V}) \\ &= E(X_k|\mathbf{V}, \delta_k = 1). \end{aligned}$$

It follows that $\mathcal{S}_{E(X_k|\mathbf{V})}$ is also a partial mean dimension-reduction subspace for the regression of X_k on $\mathbf{V}$ under the condition $\delta_k = 1$. Then, $\mathcal{S}^{\{\delta_k=1\}}_{E(X_k|\mathbf{V})} \subseteq \mathcal{S}_{E(X_k|\mathbf{V})}$, because $\mathcal{S}^{\{\delta_k=1\}}_{E(X_k|\mathbf{V})}$ is the minimal partial mean dimension-reduction subspace of X_k on $\mathbf{V}$, given $\delta_k = 1$.

Second, we prove $\mathcal{S}^{\{\delta_k=1\}}_{E(X_k|\mathbf{V})} \supseteq \mathcal{S}_{E(X_k|\mathbf{V})}$ in a similar way. Assume that A is a basis matrix of $\mathcal{S}^{\{\delta_k=1\}}_{E(X_k|\mathbf{V})}$. Then, we have $E(X_k|\mathbf{V}, \delta_k = 1) = E(X_k|A^T\mathbf{V}, \delta_k = 1)$. Together with the MAR assumption, we have $E(X_k|\mathbf{V}) = E(X_k|A^T\mathbf{V}, \delta_k = 1)$, which indicates that

$$\begin{aligned} E(X_k|A^T\mathbf{V}) &= E\{E(X_k|\mathbf{V})|A^T\mathbf{V}\} \\ &= E\{E(X_k|A^T\mathbf{V}, \delta_k = 1)|A^T\mathbf{V}\} \\ &= E(X_k|A^T\mathbf{V}, \delta_k = 1) \\ &= E(X_k|\mathbf{V}). \end{aligned}$$

It follows that $\mathcal{S}^{\{\delta_k=1\}}_{E(X_k|\mathbf{V})}$ is also a mean dimension-reduction subspace for the regression of X_k on $\mathbf{V}$. Then, $\mathcal{S}^{\{\delta_k=1\}}_{E(X_k|\mathbf{V})} \supseteq \mathcal{S}_{E(X_k|\mathbf{V})}$, because $\mathcal{S}_{E(X_k|\mathbf{V})}$ is the minimal mean dimension-reduction subspace of X_k on $\mathbf{V}$. This completes the proof of Proposition 1.

Because the technical proofs of Proposition 2–3 are almost similar to that of Proposition 1, we omit the details here.

Appendix B: Proof of Theorem 1

We begin with three lemmas to facilitate the proof of Theorem 1. All technical proofs for these lemmas are provided in the supplementary material. Note that the notation used here is defined in Section 2.

Lemma B.1. *Suppose that conditions* 1–4 *hold. Then, for the case* $1 \leq k, l \leq p_1$,

we have the following results:

(i) $\widehat{E}(X_k) - E(X_k) = \dfrac{1}{n}\sum_{i=1}^{n} J_1(X_{ki}, \delta_{ki}, \Gamma_k^T \mathbf{V}_i) + o_p(n^{-1/2})$.

(ii) $\widehat{E}(X_k^2) - E(X_k^2) = \dfrac{1}{n}\sum_{i=1}^{n} J_2(X_{ki}, \delta_{ki}, \Upsilon_k^T \mathbf{V}_i) + o_p(n^{-1/2})$.

(iii) $\widehat{E}(X_k X_l) - E(X_k X_l) = \dfrac{1}{n}\sum_{i=1}^{n} J_3(X_{ki}, X_{li}, \delta_{ki}, \delta_{li}, \Gamma_{kl}^T \mathbf{V}_i) + o_p(n^{-1/2}), (k \neq l)$.

(iv) $\widehat{E}(X_k \mathbf{X}_{obs}^T) - E(X_k \mathbf{X}_{obs}^T) = \dfrac{1}{n}\sum_{i=1}^{n} J_4(X_{ki}, \mathbf{X}_{obs,i}, \delta_{ki}, \Gamma_k^T \mathbf{V}_i) + o_p(n^{-1/2})$.

where $\widehat{E}(X_k)$, $\widehat{E}(X_k^2)$, $\widehat{E}(X_k X_l)$, and $\widehat{E}(X_k \mathbf{X}_{obs}^T)$ are defined in (2.14)–(2.17), respectively, and

$$J_1(X_{ki}, \delta_{ki}, \Gamma_k^T \mathbf{V}_i) = M_k(\Gamma_k^T \mathbf{V}_i) + \frac{\delta_{ki}}{\pi_k(\Gamma_k^T \mathbf{V}_i)}\{X_{ki} - M_k(\Gamma_k^T \mathbf{V}_i)\} - E(X_k),$$

$$J_2(X_{ki}, \delta_{ki}, \Upsilon_k^T \mathbf{V}_i) = Q_k(\Upsilon_k^T \mathbf{V}_i) + \frac{\delta_{ki}}{\widetilde{\pi}_k(\Upsilon_k^T \mathbf{V}_i)}\{X_{ki}^2 - Q_k(\Upsilon_k^T \mathbf{V}_i)\} - E(X_k^2),$$

$$\begin{aligned} J_3(X_{ki}, X_{li}, \delta_{ki}, \delta_{li}, \Gamma_{kl}^T \mathbf{V}_i) &= R_{kl}(\Gamma_{kl}^T \mathbf{V}_i) + \frac{\delta_{ki}\delta_{li}}{\pi_{kl}(\Gamma_{kl}^T \mathbf{V}_i)}\{X_{ki}X_{li} - R_{kl}(\Gamma_{kl}^T \mathbf{V}_i)\} \\ &\quad - E(X_k X_l), \\ J_4(X_{ki}, \mathbf{X}_{obs,i}, \delta_{ki}, \Gamma_k^T \mathbf{V}_i) &= \big\{\delta_{ki} X_{ki} + (1-\delta_{ki}) M_k(\Gamma_k^T \mathbf{V}_i)\big\}\mathbf{X}_{obs,i}^T - E(X_k \mathbf{X}_{obs}^T) \\ &\quad + \frac{\delta_{ki}}{\pi_k(\Gamma_k^T \mathbf{V}_i)}\{X_{ki} - M_k(\Gamma_k^T \mathbf{V}_i)\} E[(1-\delta_k)\mathbf{X}_{obs}^T | \Gamma_k^T \mathbf{V}_i], \end{aligned}$$

with $M_k(\cdot)$, $Q_k(\cdot)$, $R_{kl}(\cdot)$, $\pi_k(\cdot)$, $\widetilde{\pi}_k(\cdot)$, and $\pi_{kl}(\cdot)$ defined in Subsections 2.3–2.4.

Lemma B.2. *Suppose that conditions* 1–4 *hold. Then, for the case* $1 \leq k, l \leq p_1$, *we have the following results:*

(i) $$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \widehat{T}_k(Y_i)\widehat{T}_l(Y_i) - E\{E(X_k|Y)E(X_l|Y)\} \\ &= \frac{1}{n}\sum_{i=1}^{n} J_5(X_{ki}, X_{li}, \delta_{ki}, \delta_{li}, Y_i, \Gamma_k^T \mathbf{V}_i, \Gamma_l^T \mathbf{V}_i) + o_p(n^{-1/2}), \end{aligned}$$

(ii) $$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \widehat{T}_k(Y_i)\widehat{H}(Y_i) - E\{E(X_k|Y)E(\mathbf{X}_{obs}^T|Y)\} \\ &= \frac{1}{n}\sum_{i=1}^{n} J_6(X_{ki}, \delta_{ki}, \mathbf{X}_{obs,i}, Y_i, \Gamma_k^T \mathbf{V}_i) + o_p(n^{-1/2}), \end{aligned}$$

where $\widehat{T}_k(\cdot)$ and $\widehat{H}(\cdot)$ are defined in (2.18) and (2.19), respectively, and

$$
\begin{aligned}
&J_5(X_{ki}, X_{li}, \delta_{ki}, \delta_{li}, Y_i, \Gamma_k^T\mathbf{V}_i, \Gamma_l^T\mathbf{V}_i)\\
&= \big\{\delta_{ki}X_{ki} + (1-\delta_{ki})M_k(\Gamma_k^T\mathbf{V}_i)\big\}T_l(Y_i) + \big\{\delta_{li}X_{li} + (1-\delta_{li})M_l(\Gamma_l^T\mathbf{V}_i)\big\}T_k(Y_i)\\
&\quad + E\Big\{\big[\delta_k X_k + (1-\delta_k)M_k(\Gamma_k^T\mathbf{V})\big]\Big|Y = Y_i\Big\}T_l(Y_i)\\
&\quad + E\Big\{\big[\delta_l X_l + (1-\delta_l)M_l(\Gamma_l^T\mathbf{V})\big]\Big|Y = Y_i\Big\}T_k(Y_i)\\
&\quad + \frac{\delta_{ki}\big[X_{ki} - M_k(\Gamma_k^T\mathbf{V}_i)\big]}{\pi_k(\Gamma_k^T\mathbf{V}_i)}E\big[(1-\delta_k)T_l(Y)\big|\Gamma_k^T\mathbf{V} = \Gamma_k^T\mathbf{V}_i\big]\\
&\quad + \frac{\delta_{li}\big[X_{li} - M_l(\Gamma_l^T\mathbf{V}_i)\big]}{\pi_l(\Gamma_l^T\mathbf{V}_i)}E\big[(1-\delta_l)T_k(Y)\big|\Gamma_l^T\mathbf{V} = \Gamma_l^T\mathbf{V}_i\big]\\
&\quad - E\Big\{\big[\delta_k X_k + (1-\delta_k)M_k(\Gamma_k^T\mathbf{V})\big]T_l(Y)\Big\} - E\Big\{\big[\delta_l X_l + (1-\delta_l)M_l(\Gamma_l^T\mathbf{V})\big]T_k(Y)\Big\}\\
&\quad + E\big\{T_k(Y)T_l(Y)\big\} - 3T_k(Y_i)T_l(Y_i) + o_p(n^{-1/2}),\\
&J_6(X_{ki}, \delta_{ki}, \mathbf{X}_{obs,i}, Y_i, \Gamma_k^T\mathbf{V}_i)\\
&= \big\{\delta_{ki}X_{ki} + (1-\delta_{ki})M_k(\Gamma_k^T\mathbf{V}_i)\big\}H(Y_i) + T_k(Y_i)\mathbf{X}_{obs,i}^T - 2E\big[T_k(Y)H(Y)\big],
\end{aligned}
$$

with $T_k(\cdot)$, $M_k(\cdot)$, and $H(\cdot)$ defined in Subsection 2.2.

Lemma B.3. *Suppose that conditions* 1–4 *hold. Then, we have*

$$
\frac{1}{n}\sum_{i=1}^n \widehat{H}(Y_i)^T\widehat{H}(Y_i) - E\{E(\mathbf{X}_{obs}|Y E(\mathbf{X}_{obs}^T|Y)\} = \frac{1}{n}\sum_{i=1}^n J_7(\mathbf{X}_{obs,i}, Y_i) + o_p(n^{-1/2}),
$$

where $J_7(\mathbf{X}_{obs,i}, Y_i) = \mathbf{X}_{obs,i}H(Y_i) + H(Y_i)^T\mathbf{X}_{obs,i}^T - 2E[H(Y)^T H(Y)]$.

Remark B.1. Lemmas B.1–B.2 only consider the nontrivial cases. For the case of $p_1 + 1 \le k, l \le p$, both $\widehat{E}(X_k)$ and $\widehat{E}(X_k X_l)$ are the usual sample estimates such that each is naturally a sum of i.i.d random variables.

Proof of Theorem 1. Observe that

$$
\begin{aligned}
&\widehat{\Sigma}_{\mathbf{X}}^{-1}\widehat{\Sigma}_{E(\mathbf{X}|Y)} - \Sigma_{\mathbf{X}}^{-1}\Sigma_{E(\mathbf{X}|Y)}\\
&= \Sigma_{\mathbf{X}}^{-1}(\Sigma_{\mathbf{X}} - \widehat{\Sigma}_{\mathbf{X}})\widehat{\Sigma}_{\mathbf{X}}^{-1}(\widehat{\Sigma}_{E(\mathbf{X}|Y)} - \Sigma_{E(\mathbf{X}|Y)}) + \Sigma_{\mathbf{X}}^{-1}(\Sigma_{\mathbf{X}} - \widehat{\Sigma}_{\mathbf{X}})\widehat{\Sigma}_{\mathbf{X}}^{-1}\Sigma_{E(\mathbf{X}|Y)}\\
&\quad + \Sigma_{\mathbf{X}}^{-1}(\widehat{\Sigma}_{E(\mathbf{X}|Y)} - \Sigma_{E(\mathbf{X}|Y)}).
\end{aligned}
\tag{B.1}
$$

To prove the asymptotic normality of $\sqrt{n}\big\{\mathrm{vec}(\widehat{\Sigma}_{\mathbf{X}}^{-1}\widehat{\Sigma}_{E(\mathbf{X}|Y)}) - \mathrm{vec}(\Sigma_{\mathbf{X}}^{-1}\Sigma_{E(\mathbf{X}|Y)})\big\}$, it suffices to prove that both $\widehat{\Sigma}_{\mathbf{X}} - \Sigma_{\mathbf{X}}$ and $\widehat{\Sigma}_{E(\mathbf{X}|Y)} - \Sigma_{E(\mathbf{X}|Y)}$ can be asymptotically represented as sums of i.i.d random variables. Clearly, we only need to deal with the kl-th element of these two matrices. It is easy to show that the kl-th elements of the matrices $\widehat{\Sigma}_{\mathbf{X}} - \Sigma_{\mathbf{X}}$ and $\widehat{\Sigma}_{E(\mathbf{X}|Y)} - \Sigma_{E(\mathbf{X}|Y)}$, denoted by $\big(\widehat{\Sigma}_{\mathbf{X}} - \Sigma_{\mathbf{X}}\big)_{kl}$ and $\big(\widehat{\Sigma}_{E(\mathbf{X}|Y)} - \Sigma_{E(\mathbf{X}|Y)}\big)_{kl}$, respectively, with $1 \le k, l \le p$ can be

written as

$$\begin{aligned}
&\big(\widehat{\Sigma}_{\mathbf{X}} - \Sigma_{\mathbf{X}}\big)_{kl} \\
&= \widehat{E}(X_k X_l) - E(X_k X_l) - \{\widehat{E}(X_k)\widehat{E}(X_l) - E(X_k)E(X_l)\} \\
&= \widehat{E}(X_k X_l) - E(X_k X_l) - \{\widehat{E}(X_k) - E(X_k)\}E(X_l) - \{\widehat{E}(X_l) - E(X_l)\}E(X_k) \\
&\quad - \{\widehat{E}(X_k) - E(X_k)\}\{\widehat{E}(X_l) - E(X_l)\}, \qquad \text{(B.2)}
\end{aligned}$$

and

$$\begin{aligned}
&\big(\widehat{\Sigma}_{E(\mathbf{X}|Y)} - \Sigma_{E(\mathbf{X}|Y)}\big)_{kl} \\
&= \widehat{E}\big\{\widehat{E}(X_k|Y)\widehat{E}(X_l|Y)\big\} - E\big\{E(X_k|Y)E(X_l|Y)\big\} - \{\widehat{E}(X_k)\widehat{E}(X_l) - E(X_k)E(X_l)\} \\
&= \widehat{E}\big\{\widehat{E}(X_k|Y)\widehat{E}(X_l|Y)\big\} - E\big\{E(X_k|Y)E(X_l|Y)\big\} - \{\widehat{E}(X_k) - E(X_k)\}E(X_l) \\
&\quad - \{\widehat{E}(X_l) - E(X_l)\}E(X_k) - \{\widehat{E}(X_k) - E(X_k)\}\{\widehat{E}(X_l) - E(X_l)\}, \qquad \text{(B.3)}
\end{aligned}$$

where $\widehat{E}(X_k)$, $\widehat{E}(X_k X_l)$ and $\widehat{E}\big\{\widehat{E}(X_k|Y)\widehat{E}(X_l|Y)\big\}$ denote the estimates of $E(X_k)$, $E(X_k X_l)$ and $E\{E(X_k|Y)E(X_l|Y)\}$, respectively. It then suffices to prove that $\widehat{E}(X_k) - E(X_k)$, $\widehat{E}(X_k X_l) - E(X_k X_l)$ and $\widehat{E}\big\{\widehat{E}(X_k|Y)\widehat{E}(X_l|Y)\big\} - E\big\{E(X_k|Y)$ $E(X_l|Y)\big\}$ can be asymptotically represented as sums of i.i.d random variables. These are presented in Lemmas B.1 – B.2 for the case of $1 \le k, l \le p_1$, and in Lemma B.3 and Remark B.1 for the case of $p_1 + 1 \le k, l \le p$.

Next, we show details for the asymptotic representations of $\widehat{\Sigma}_{\mathbf{X}} - \Sigma_{\mathbf{X}}$ and $\widehat{\Sigma}_{E(\mathbf{X}|Y)} - \Sigma_{E(\mathbf{X}|Y)}$. Let

$$\widehat{\Sigma}_{\mathbf{X}} - \Sigma_{\mathbf{X}} := n^{-1}\sum_{i=1}^{n} A^{(i)} + o_p(n^{-1/2}), \qquad \text{(B.4)}$$

where the block matrix

$$A^{(i)} = \begin{pmatrix} A_1^{(i)} & A_2^{(i)} \\ A_2^{(i)T} & A_3^{(i)} \end{pmatrix}_{p\times p}$$

corresponds to the partition $\mathbf{X} = (\mathbf{X}_{mis}^T, \mathbf{X}_{obs}^T)^T$, with $A_1^{(i)}$, $A_2^{(i)}$, and $A_3^{(i)}$ denoting $p_1 \times p_1, p_1 \times (p - p_1)$, and $(p - p_1) \times (p - p_1)$ matrices, respectively. By (B.2) and Lemma B.1 (i)–(iii), the kl-th element $a_{1kl}^{(i)}$ of the sub-matrix $A_1^{(i)}$ can be expressed as

$$\begin{aligned}
a_{1kl}^{(i)} &= J_3(X_{ki}, X_{li}, \delta_{ki}, \delta_{li}, \Gamma_{kl}^T \mathbf{V}_i) - J_1(X_{ki}, \delta_{ki}, \Gamma_k^T \mathbf{V}_i)E(X_l) \\
&\quad - J_1(X_{li}, \delta_{li}, \Gamma_l^T \mathbf{V}_i)E(X_k) \quad \text{for } 1 \le k \ne l \le p_1
\end{aligned}$$

and

$$a_{1kk}^{(i)} = J_2(X_{ki}, \delta_{ki}, \Upsilon_k^T \mathbf{V}_i) - 2J_1(X_{ki}, \delta_{ki}, \Gamma_k^T \mathbf{V}_i)E(X_k) \quad \text{for } 1 \le k \le p_1.$$

By (B.2), Lemma B.1 (iv), and Remark B.1, the k-th raw $\mathbf{a}_{2k}^{(i)}$ of the sub-matrix $A_2^{(i)}$ can be expressed as

$$\begin{aligned}\mathbf{a}_{2k}^{(i)} &= J_4(X_{ki}, \mathbf{X}_{obs,i}, \delta_{ki}, \Gamma_k^T \mathbf{V}_i) \\ &\quad - J_1(X_{ki}, \delta_{ki}, \Gamma_k^T \mathbf{V}_i) E(\mathbf{X}_{obs}^T) - [\mathbf{X}_{obs,i} - E(\mathbf{X}_{obs})]^T E(X_k).\end{aligned}$$

In addition, the sub-matrix $A_3^{(i)}$ only involves the completely observed data and, hence, can be expressed as

$$\begin{aligned}A_3^{(i)} &= \mathbf{X}_{obs,i}\mathbf{X}_{obs,i}^T - \mathbf{X}_{obs,i}E(\mathbf{X}_{obs}^T) \\ &\quad - E(\mathbf{X}_{obs})\mathbf{X}_{obs,i}^T - E(\mathbf{X}_{obs}\mathbf{X}_{obs}^T) + 2E(\mathbf{X}_{obs})E(\mathbf{X}_{obs}^T).\end{aligned}$$

Similarly, we write

$$\widehat{\Sigma}_{E(\mathbf{X}|Y)} - \Sigma_{E(\mathbf{X}|Y)} := n^{-1}\sum_{i=1}^{n} B^{(i)} + o_p(n^{-1/2}), \tag{B.5}$$

where the block matrix

$$B^{(i)} = \begin{pmatrix} B_1^{(i)} & B_2^{(i)} \\ B_2^{(i)T} & B_3^{(i)} \end{pmatrix}_{p\times p}$$

corresponds to the partition $\mathbf{X} = (\mathbf{X}_{mis}^T, \mathbf{X}_{obs}^T)^T$, with $B_1^{(i)}, B_2^{(i)}$, and $B_3^{(i)}$ denoting $p_1 \times p_1, p_1 \times (p-p_1)$, and $(p-p_1)\times(p-p_1)$ matrices, respectively. By (B.3), Lemma B.1 (i), and Lemma B.2 (i), the kl-th element $b_{1kl}^{(i)}$ of the sub-matrix $B_1^{(i)}$ can be expressed as

$$\begin{aligned}b_{1kl}^{(i)} &= J_5(X_{ki}, X_{li}, \delta_{ki}, \delta_{li}, Y_i, \Gamma_k^T\mathbf{V}_i, \Gamma_l^T\mathbf{V}_i) - J_1(X_{ki}, \delta_{ki}, \Gamma_k^T\mathbf{V}_i)E(X_l) \\ &\quad - J_1(X_{li}, \delta_{li}, \Gamma_l^T\mathbf{V}_i)E(X_k).\end{aligned}$$

By (B.3), Lemma B.2 (ii), and Remark A, the k-th raw $\mathbf{b}_{2k}^{(i)}$ of the sub-matrix $B_2^{(i)}$ can be expressed as

$$\begin{aligned}\mathbf{b}_{2k}^{(i)} &= J_6(X_{ki}, \delta_{ki}, \mathbf{X}_{obs,i}, Y_i, \Gamma_k^T\mathbf{V}_i) - J_1(X_{ki}, \delta_{ki}, \Gamma_k^T\mathbf{V}_i)E(\mathbf{X}_{obs}^T) \\ &\quad - [\mathbf{X}_{obs,i} - E(\mathbf{X}_{obs})]^T E(X_k).\end{aligned}$$

In addition, (B.3) and Lemma B.3 jointly yield that

$$B_3^{(i)} = J_7(\mathbf{X}_{obs,i}, Y_i) - \mathbf{X}_{obs,i}E(\mathbf{X}_{obs}^T) - E(\mathbf{X}_{obs})\mathbf{X}_{obs,i}^T + 2E(\mathbf{X}_{obs})E(\mathbf{X}_{obs}^T).$$

Finally, from (B.1), (B.4), and (B.5), simple algebraic calculations give the result

$$\widehat{\Sigma}_{\mathbf{X}}^{-1}\widehat{\Sigma}_{E(\mathbf{X}|Y)} - \Sigma_{\mathbf{X}}^{-1}\Sigma_{E(\mathbf{X}|Y)} = \frac{1}{n}\sum_{i=1}^{n}\left\{\Sigma_{\mathbf{X}}^{-1}B^{(i)} - \Sigma_{\mathbf{X}}^{-1}A^{(i)}\Sigma_{\mathbf{X}}^{-1}\Sigma_{E(\mathbf{X}|Y)}\right\} + o_p(n^{-1/2}) \tag{B.6}$$

which implies that each element of the matrix $\widehat{\Sigma}_{\mathbf{X}}^{-1}\widehat{\Sigma}_{E(\mathbf{X}|Y)} - \Sigma_{\mathbf{X}}^{-1}\Sigma_{E(\mathbf{X}|Y)}$ can be asymptotically expanded as a sum of i.i.d random variables. Then, the central limit theorem leads to conclusion (i) of Theorem 1, as a result of which, conclusion (ii) of Theorem 1 also holds.

References

Cai, T. and Chen, X. (2010). *Highdimensional Data Analysis.*, Volume II. Higher Education Press: Beijing.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association* **89**, 81–87.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced Inverse Regression for Dimension Reduction" by K. C. Li. *Journal of the American Statistical Association* **86**, 328–332.

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

Cook, R. D. (1998a). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, Wiley: New York.

Cook, R. D. (1998b). Principal Hessian directions revisited. *Journal of the American Statistical Association* **93**, 84–94.

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics* **30**, 455–474.

Ding, X. and Wang, Q. (2011). Fusion-refinement procedure for dimension reduction with missing response at random. *Journal of the American Statistical Association* **106**, 1193–1207.

Ferré, L. and Yao, A. (2005). Smoothed functional inverse regression. *Statistica Sinica* **15**, 665–683.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics* **21**, 867–889.

Hooper, J. W. (1959). Simultaneous equations and canonical correlation theory. *Econometrica: Journal of the Econometric Society* **27**, 245–256.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–327.

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.

Li, B., Cook, R. D. and Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *The Annals of Statistics* **31**, 1636–1668.

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics* **33**, 1580–1616.

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

Li, L. and Lu, W. (2008). Sufficient dimension reduction with missing predictors. *Journal of the American Statistical Association* **103**, 822–831.

Li, L., Zhu, L. P. and Zhu, L. X. (2011). Inference on the primary parameter of interest with

the aid of dimension reduction estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 59–80.

Matloff, N. S. (1981). Use of regression functions for improved estimation of means. *Biometrika* **68**, 685–689.

Ma, Y. and Zhu, L. P. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association* **107**, 168–179.

Ma, Y. and Zhu, L. P. (2013). A review on dimension reduction. *International Statistical Review* **81**, 134–150.

Wang, Q. and Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics* **89**, 896–924.

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 (Statistical Methodology)* **64**, 363–410.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture* **1**, 129–142.

Yin, X. and Cook, R. D. (2005). Direction estimation in single-index regressions. *Biometrika* **92**, 371–384.

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

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.

Zhu, L. P. and Zhu, L. X. (2007). On kernel method for sliced average variance estimation. *Journal of Multivariate Analysis* **98**, 970–991.

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

Zhu, L. P., Li, L., Li, R. and Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106**, 1464–1475.

Zhu, L. P., Wang, T. and Zhu, L. X. (2012). Sufficient dimension reduction in regressions with missing predictors. *Statistica Sinica* **22**, 1611–1637.

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, NO. 55, Zhongguancun East Road, Haidian District, Beijing, 100190, China.

E-mail: yangxiaojie@amss.ac.cn

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, NO. 55, Zhongguancun East Road, Haidian District, Beijing, 100190, China.
School of Statistics and Mathematics, Zhejiang Gongshang University Hangzhou, Zhejiang 310018, China

E-mail: qhwang@amss.ac.cn

(Received November 2016; accepted January 2018)