<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-2017-0288</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
<tr>
<td><strong>Complete List of Authors</strong></td>
</tr>
<tr>
<td><strong>Corresponding Author</strong></td>
</tr>
<tr>
<td><strong>E-mail</strong></td>
</tr>
</tbody>
</table>

Notice: Accepted version subject to English editing.
Sufficient Dimension Reduction under
Dimension-reduction-based Imputation with
Predictors Missing at Random

Xiaojie Yang and Qihua Wang

*Academy of Mathematics and Systems Science,
Chinese Academy of Sciences.*

*Abstract:* In some practical problems, a subset of predictors are frequently subject to missingness, especially when the dimension of the predictors is high. For this case, the standard sufficient dimension reduction (SDR) methods cannot be applied directly to avoid the curse of dimensionality. A dimension-reduction-based imputation method is developed in this article such that any of spectral-decomposition-based SDR methods for full data is applicable to the case of predictors missing at random. The sliced inverse regression (SIR) is used to illustrate this procedure. The proposed dimension-reduction-based imputation estimator of the candidate matrix for SIR, termed as DRI-SIR estimator, is asymptotically normal under some mild conditions and 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 a real data set is analyzed for illustration.
Key words and phrases: Kernel imputation, Missing at random, Missing predictors, Sufficient dimension reduction, Sliced inverse regression.

1. Introduction

Consider the regression of an univariate response variable \( Y \) on a \( p \times 1 \) covariate vector \( X \). Regression analysis typically focuses on how the conditional distribution function \( F(y|X = x) \) changes as the value of \( X \) varies in its marginal sample space. When the dimension \( p \) is large, modeling a parametric structure for the regression is difficult, and nonparametric methods also cannot work well due to the curse of dimensionality. Sufficient dimension reduction (SDR; Cook, 1998a) has been proposed to reduce the dimension of \( X \) while preserving full information for \( Y \) without imposing specified regression parametric models. It is aimed at replacing \( X \) by \( d \leq p \) linear combinations, \( \beta_1^T X, \cdots, \beta_d^T X \), such that \( Y \perp \!\!\!\perp X | B^T X \), where \( B \) is a \( p \times d \) matrix with columns \( \beta_j \), 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) \( S_{Y|X} \) for the regression of \( Y \) on \( X \); its dimension \( d = \dim(S_{Y|X}) \), is called the structural dimension. It is clear
that the central subspace provides the greatest reduction from $\mathbf{X}$ to $\mathbf{B}^T \mathbf{X}$ and captures all regression information of $Y$ on $\mathbf{X}$.

Since Li’s pioneering work, say sliced inverse regression (SIR; Li, 1991), many SDR methods have been developed to estimate $\mathcal{S}_{Y|X}$. Some are described as the spectral-decomposition-based methods, including 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 are derived from numerical minimization (or maximization) of some nonparametric objective functions, including 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) first offered a selective review of SDR methods in regression and later Ma and Zhu (2013) discussed the recent developments in the SDR field.

Despite a growing number of SDR literature with significant theoretical advances, only little attention has been paid to SDR with missing values. In many applications, however, missing data are often encountered. With missing response at random, Ding and Wang (2011) proposed a fusion-refinement (FR) procedure to handle the dimension-reduction problems. In
the context of missing predictors at random, \cite{Li2008} introduced
the augmented inverse probability weighted SIR estimator (AIPW-SIR),
while \cite{Zhu2012} proposed a parametric imputation proce-
dure for SIR (PI-SIR). Both methods require to specify parametric models
for the involved conditional expectations and the propensity function.

For ease of exposition, write \(X = (X_1, \cdots, X_p)^T = (X_{mis}^T, X_{obs}^T)^T\), where
\(X_{mis} = (X_1, \cdots, X_{p_1})^T \in \mathbb{R}^{p_1}\) refers to predictors with missingness in a
subset of subjects, and \(X_{obs} = (X_{p_1+1}, \cdots, X_p)^T \in \mathbb{R}^{p-p_1}\) is always observed
for all subjects. Let \(\delta = (\delta_1, \cdots, \delta_{p_1})^T\) denote a vector of missingness
indicators for \(X_{mis}\), where \(\delta_k\) takes value 1 if there is no missingness for the
\(k\)-th component \(X_k\) in \(X_{mis}\) and 0 otherwise. Throughout this paper, we
assume that \(X_{mis}\) is missing at random (MAR), i.e.,
\[
\delta \perp X_{mis} \mid (X_{obs}^T, Y)^T, \tag{1.1}
\]
which essentially allows the missingness to depend only on the completely
observed variables \((X_{obs}^T, Y)^T\).

To make SIR applicable to the case of missing predictors, a main
challenge is to estimate the candidate matrix \(\{\text{Cov}(X)\}^{-1}\text{Cov}\{E(X|Y)\}\)
for SIR. According to \cite{Zhu2012}, we need obtain the
consistent estimators of these quantities, i.e. \(E(X_k), E(X_kX_{obs}^T), E(X^2_k),
E(X_kX_l)\ (k \neq l)\) and \(E(X_k|Y)\), where \(X_k\) (or \(X_l\)) denotes the \(k\)-th (or \(l\)-th)
component in $X_{\text{mis}}$, for $k, l = 1, \cdots, p_1$. As there is no need to estimate the mixed moment $E(X_k X_l) \ (k \neq l)$ for $p_1 = 1$, we focus on the general case $p_1 \geq 2$. Let $V = (X_{\text{obs}}^T, Y)^T \in \mathbb{R}^{p-p_1+1}$. The double-expectation theorem leads to that $E(X_k) = E\{E(X_k|V)\}$, $E(X_k X_{\text{obs}}^T) = E\{E(X_k|V)X_{\text{obs}}^T\}$, $E(X_k X_l) = E\{E(X_k X_l|V)\} \ (k \neq l)$, $E(X_k^2) = E\{E(X_k^2|V)\}$ and $E(X_k|Y) = E\{E(X_k|V)|Y\}$. Then the key point is to handle $E(X_k|V)$, $E(X_k^2|V)$ and $E(X_k X_l|V)$, which poses two challenges,

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

However, the existing methods fail to solve these two issues. To lessen the effect of high dimension, Li and Lu (2008) recommended the linear models or other proper parametric models for these conditional expectations, while Zhu, Wang and Zhu (2012) imposed linear models on $E(\delta_k X_k|V)$, $E(\delta_k|V)$, $E(\delta_k X_k^2|V)$, $E(\delta_k \delta_l X_k X_l|V)$ and $E(\delta_k \delta_l|V)$, and then constructed the estimators of these conditional expectations based on the equations $E(X_k|V) = E(\delta_k X_k|V)/E(\delta_k|V)$, $E(X_k^2|V) = E(\delta_k X_k^2|V)/E(\delta_k|V)$ and $E(X_k X_l|V) = E(\delta_k \delta_l X_k X_l|V)/E(\delta_k \delta_l|V)$ that are indicated by the MAR
assumption (1.1). Both methods might yield inconsistent estimators due to the misspecification of the involved parametric models. Moreover, it is almost impossible to specify all the parametric models correctly in practice.

We are now in a position to develop some nonparametric methods to avoid the parametric specification of those involved models and work out the two issues mentioned above. Our strategy is to seek a $q \times r$ matrix $\Gamma$ with $q = p - p_1 + 1 < r$ for each conditional expectation such that $\Gamma$ satisfies that, (i) $V$ in this conditional expectation can be replaced by its low-dimensional linear transformation $\Gamma^T V$ without changing this conditional expectation, and (ii) the complete-case (CC) approach that simply removes all the subjects with missing values can be used to obtain consistent estimators of $\Gamma$ and the corresponding conditional expectation, which is actually treated as an intermediate step for the proposed method. It can be shown that any existing SDR methods based on CC analysis can be used to obtain such $\Gamma$. Furthermore, a dimension-reduction-based kernel imputation method is proposed to obtain consistent estimators of those expectations and conditional expectations in the candidate matrix for SIR and finally 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 theoretical results. In Section 3, we check the
finite sample performance of the proposed method through simulated data.
In Section 4, we analyze a real data set for illustration. We then conclude
our paper with a discussion in Section 5. Proofs of the main results are
given in Appendix.

2. Dimension-reduction-based Kernel Imputation for SIR

In this section, we first briefly review SIR under full data, and then de-
velop a dimension-reduction-based kernel imputation method for SIR with
predictors missing at random.

2.1 Review

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

\[ \Lambda = \sum_{h=1}^{H} p_h (m_h - \mu)(m_h - \mu)^T, \]  

(2.1)

where \( \mu = E(X), p_h = \Pr(Y \in I_h) \) and \( m_h = E(X|Y \in I_h) \), for \( h = 1, \cdots, H \). It also can be derived that \( \text{Span}\{\Sigma^{-1}_X \Lambda\} \subseteq S_{Y|X} \).

Given \( n \) iid observations \( \{(X_1,Y_1), \cdots, (X_n,Y_n)\} \), by substituting the usual sample estimates of \( p_h, m_h \) and \( \mu \) into (2.1), the sample version \( \hat{\Lambda} \) of \( \Lambda \) can be obtained and used as an estimate of \( \Sigma_{E(X|Y)} \). Then, the sample estimate of SIR follows from the spectral decomposition,

\[ \hat{\lambda}_j \hat{\beta}_j = \hat{\lambda}_j \hat{\Sigma}_X \hat{\beta}_j, \quad \text{for} \ j = 1, \cdots, d, \]

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

2.2 Extension to the Case of Missing Predictors

When the predictors are missing at random, the major challenge is to develop a consistent estimating approach of the candidate matrix \( \Sigma_{E(X|Y)}^{-1} \Sigma_X \) for SIR. According to Zhu, Wang and Zhu (2012), \( \Sigma_{E(X|Y)} = \Phi_1 - \Phi_0 \) and \( \Sigma_X = \Phi_2 - \Phi_0 \), where
2.2 Extension to the Case of Missing Predictors

\[ \Phi_0 = E(X)E(X^T) = \begin{pmatrix}
E(X_{mis})E(X^T_{mis}) & E(X_{mis})E(X^T_{obs}) \\
E(X_{obs})E(X^T_{mis}) & E(X_{obs})E(X^T_{obs})
\end{pmatrix}, \]

\[ \Phi_1 = E\{E(X|Y)E(X^T|Y)\} = \begin{pmatrix}
E\{E(X_{mis}|Y)E(X^T_{mis}|Y)\} & E\{E(X_{mis}|Y)E(X^T_{obs}|Y)\} \\
E\{E(X_{obs}|Y)E(X^T_{mis}|Y)\} & E\{E(X_{obs}|Y)E(X^T_{obs}|Y)\}
\end{pmatrix}, \]

\[ \Phi_2 = E(XX^T) = \begin{pmatrix}
E(X_{mis}X^T_{mis}) & E(X_{mis}X^T_{obs}) \\
E(X_{obs}X^T_{mis}) & E(X_{obs}X^T_{obs})
\end{pmatrix}, \]

by the partition \( X = (X^T_{mis}, X^T_{obs})^T \). To implement SIR, one must estimate these expectations: \( E(X_{obs}), E(X_{obs}X^T_{obs}), E\{E(X_{obs}|Y)E(X^T_{obs}|Y)\}, \)
\( E\{E(X_{mis}|Y)E(X^T_{obs}|Y)\}, E\{E(X_{mis}|Y)E(X^T_{mis}|Y)\}, E(X_{mis}), E(X_{mis}X^T_{obs}) \)
and \( E(X_{mis}X^T_{mis}) \). The first three quantities can be estimated by the standard methods, since they involve only the completely observed variables \( (X^T_{obs}, Y)^T \). For the last five quantities involving the missing covariate vector \( X_{mis} \), however, it is desired to develop new methods to obtain their consistent estimators. In an element-wise manner, this problem reduces to estimating \( E(X_k|Y), E(X_k), E(X_kX^T_{obs}), E(X_k^2) \) and \( E(X_k X_l) (k \neq l) \)
with \( X_k \) (or \( X_l \)) denoting the \( k \)-th (or \( l \)-th) component in \( X_{mis} \), for \( k, l = 1, 2, \ldots, p \). As discussed in the introduction, estimating these expectations substantially resorts to the estimation of some conditional expectat-
2.2 Extension to the Case of Missing Predictors

tions given \( V = (X^T_{\text{obs}}, Y)^T \).

We particularly describe how to estimate \( E(X_k) \), for \( k = 1, \ldots, p_1 \), and state that the principle of estimating other quantities is roughly the same. The focus is actually on the estimation of \( E(X_k|V) \) by noting that \( E(X_k) = E\{E(X_k|V)\} \). Regression is most efficient when the dimension of \( V \) is small and the relationship between \( X_k \) and \( V \) is correctly specified (Yates, 1933; Matloff, 1981). In this case, one can estimate \( E(X_k|V) \) by \( n^{-1} \sum_{i=1}^n \hat{m}(V_i) \) with \( \hat{m}(\cdot) \) being an estimator of the parametric model \( m(\cdot) \) imposed on \( E(X_k|V) \). Such an estimator is inconsistent if \( m(\cdot) \) is misspecified. The nonparametric regression certainly can be employed to estimate \( E(X_k|V) \) without requiring parametric specification but probably suffers from the curse of dimensionality. A natural idea is to replace \( V \) in \( E(X_k|V) \) with its low-dimensional transformation \( S_k(V) : \mathbb{R}^q \mapsto \mathbb{R}^{q^*} \quad (q^* < q) \), such that \( E(X_k|V) = E\{X_k|S_k(V)\} \). The simplest form of \( S_k(V) \) is the linear transformation \( \Gamma_k^TV \), where \( \Gamma_k \) denotes a \( q \times r_k \) matrix with \( r_k < q \). It then follows that

\[
E(X_k) = E\{E(X_k|\Gamma_k^TV)\} \, ,
\]

\[
E(X_kX_{\text{obs}}^T) = E\{E(X_k|V)X_{\text{obs}}^T\} = E\{E(X_k|\Gamma_k^TV)X_{\text{obs}}^T\} \, ,
\]

\[
E(X_k|Y) = E\{E(X_k|V)|Y\} = E\{E(X_k|\Gamma_k^TV)|Y\} \, .
\]

Along this idea, let \( \Upsilon_k \) be a \( q \times \tilde{r}_k \) matrix and \( \Gamma_{kl} \quad (k \neq l) \) be a \( q \times r_{kl} \) matrix.
2.2 Extension to the Case of Missing Predictors

with $\tilde{r}_k, r_{kl} < q$ and $k, l = 1, \cdots, p_1$, we have

$$E(X_k^2) = E\{E(X_k^2|Y_k^T V)\}, \quad (2.5)$$

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

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

As stated in the introduction, obtaining $\Gamma_k$, $\Upsilon_k$ and $\Gamma_{kl}$ 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 $E(X_k|\Gamma_k^T V)$, $E(X_k^2|Y_k^T V)$ and $E(X_k X_l|\Gamma_{kl}^T V)$ ($k \neq l$) with $k, l = 1, 2, \cdots, p_1$, the matrices $\Gamma_k$, $\Upsilon_k$ and $\Gamma_{kl}$ should respectively satisfy that

$$E(X_k|V) = E(X_k|\Gamma_k^T V) = E(X_k|\Gamma_k^T V, \delta_k = 1), \quad (2.7)$$

$$E(X_k^2|Y_k^T V) = E(X_k^2|\Upsilon_k^T V, \delta_k = 1), \quad (2.8)$$

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

We next describe how to obtain $\Gamma_k$, $\Upsilon_k$ and $\Gamma_{kl}$, for $k, l = 1, \cdots, p_1$.

(i) Derivation of $\Gamma_k$, for $k = 1, \cdots, p_1$. 

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)
2.2 Extension to the Case of Missing Predictors

We gain insights from the condition $E(X_k|\mathbf{V}) = E(X_k|\Gamma_k^T \mathbf{V})$. It makes the basis matrix of the central mean subspace (CMS; Cook and Li 2002) $S_{E(X_k|\mathbf{V})}$ be a natural choice for $\Gamma_k$, where $S_{E(X_k|\mathbf{V})}$ is the minimal mean subspace $S$ satisfying $E(X_k|\mathbf{V}) = E(X_k|P_S \mathbf{V})$ with $P_S$ being a projection operator onto $S$ in a standard inner product. With such a $\Gamma_k$ and the MAR assumption, we further obtain that

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

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

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

$$= E(X_k|\Gamma_k^T \mathbf{V}). \quad (2.10)$$

This implies that $\Gamma_k$ taken as a basis matrix of $S_{E(X_k|\mathbf{V})}$ really satisfy the condition (2.7). However, another question arises, how to obtain a consistent estimator of a method-specific basis $\Gamma_k$ of $S_{E(X_k|\mathbf{V})}$ in the presence of missing predictors? We here and after use the phrase “a method-specific basis” to avoid the ambiguity caused by the non-uniqueness of the basis and sometimes we may omit it for simplicity. The SDR methods based on the CC analysis can yield an estimator of the partial central mean subspace (Li, Cook and Chiaromonte 2003) $S_{E(X_k|\mathbf{V})}^{\{\delta_k = 1\}}$, which is the minimal partial mean subspace $S$ satisfying $E(X_k|\mathbf{V}, \delta_k = 1) = E(X_k|P_S \mathbf{V}, \delta_k = 1)$. Proposition
2.2 Extension to the Case of Missing Predictors

1 states that a consistent estimator of $\Gamma_k$ can be obtained via $S_{E(X_k|V)}^{(\delta_k=1)}$.

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

Cook and Li (2002) proposed the Iterative Hessian Transformation (IHT) method to estimate a basis of the central mean subspace. By Proposition 1, IHT method can be applied to the completely observed data set $\{(X_{ki}, V_i) : \delta_{ki} = 1, i = 1, \cdots, n\}$ to obtain a consistent estimator $\hat{\Gamma}_k$ of $\Gamma_k$.

(ii) Derivation of $\Upsilon_k$, for $k = 1, \cdots, p_1$.

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

**Proposition 2.** Assuming the same conditions as Proposition 1, we have $S_{E(\tilde{X}_k|V)} = S_{E(\tilde{X}_k|V)}^{(\delta_k=1)}$, for $k = 1, \cdots, p_1$.

According to Proposition 2, the IHT method can be applied to the completely observed data set $\{((\tilde{X}_{ki}, V_i) : \tilde{X}_{ki} = X_{ki}^2, \delta_{ki} = 1, i = 1, \cdots, n\}$.
2.2 Extension to the Case of Missing Predictors

to gain a consistent estimator \( \hat{\Upsilon}_k \) of \( \Upsilon_k \).

(iii) Derivation of \( \Gamma_{kl} (\neq l) \), for \( k, l = 1, \ldots, p \).

Observing that \( \Gamma_{kl} = \Gamma_{lk} \) with \( k \neq l \), we only consider the case of \( k < l \).

Let \( Z^{(kl)} = X_kX_l \) with \( k < l \). When \( \Gamma_{kl} \) is taken as a basis matrix of the central mean subspace \( \mathcal{S}_{E(Z^{(kl)}|V)} \), together with the MAR assumption we can show that

\[
E(X_kX_l|\Gamma^T_{kl}V, \delta_k = 1, \delta_l = 1) = E[E(X_kX_l|V, \delta_k = 1, \delta_l = 1)|\Gamma^T_{kl}V, \delta_k = 1, \delta_l = 1]
\]

\[
= E[E(X_kX_l|V)|\Gamma^T_{kl}V, \delta_k = 1, \delta_l = 1]
\]

\[
= E[E(X_kX_l|\Gamma^T_{kl}V)|\Gamma^T_{kl}V, \delta_k = 1, \delta_l = 1]
\]

\[
= E(X_kX_l|\Gamma^T_{kl}V), \quad (k < l).
\]

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

\[
E\{Z^{(kl)}|V, \delta_k = 1, \delta_l = 1\} = E\{Z^{(kl)}|P_SV, \delta_k = 1, \delta_l = 1\}.
\]

**Proposition 3.** Assuming the same conditions as Proposition 1, we have

\[
\mathcal{S}_{E(Z^{(kl)}|V)} = \mathcal{S}_{E\{Z^{(kl)}|V, \delta_k = 1, \delta_l = 1\}} \text{, for } k < l \text{ and } k, l = 1, \ldots, p.
\]

Proposition 3 ensures that the IHT method can be applied to the com-
2.3 Dimension-reduction-based Imputation for SIR (DRI-SIR)

Partially observed data set \(\{(Z_i^{(kl)}, V_i) : Z_i^{(kl)} = X_{ki}X_{li}, \delta_{ki}\delta_{li} = 1, k < l, i = 1, \cdots, n\}\) to gain a consistent estimator \(\hat{\Gamma}_{kl}\) of \(\Gamma_{kl}\).

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 SIR.

Given \(n\) iid observations \(\{(X_{\text{mis},i}, X_{\text{obs},i}, Y_i, \delta_{1i}, \cdots, \delta_{pi}) : i = 1, \cdots, n\}\), where \(X_{\text{mis},i} = (X_{ki}, \cdots, X_{pi})^T\) is subject to missingness, \(V_i = (X_{\text{obs},i}^T, Y_i)^T = (X_{p+1,i}, \cdots, X_{pi}, Y_i)^T\) is always observed, and \(\delta_{ki} = 1\) if \(X_{ki}\) is observed and \(\delta_{ki} = 0\) otherwise, for \(k = 1, \cdots, p_1\). For ease of exposition, let \(M_k(\Gamma_k^TV) = E(X_k|\Gamma_k^TV) = E(X_k|\Gamma_k^TV, \delta_k = 1)\), \(Q_k(\Upsilon_k^TV) = E(X_k^2|\Upsilon_k^TV) = E(X_k^2|\Upsilon_k^TV, \delta_k = 1)\) and \(R_{kl}(\Gamma_{kl}^TV) = E(X_kX_l|\Gamma_{kl}^TV) = E(X_kX_l|\Gamma_{kl}^TV, \delta_k = 1, \delta_l = 1)\) \((k \neq l)\), for \(1 \leq k, l \leq p_1\). Then, after obtaining the consistent estimators \(\hat{\Gamma}_k, \hat{\Upsilon}_k\) and \(\hat{\Gamma}_{kl}\) of \(\Gamma_k, \Upsilon_k\) and \(\Gamma_{kl}\) by the method mentioned in Subsection 2.2, \(M_k(\Gamma_k^TV), Q_k(\Upsilon_k^TV)\) and \(R_{kl}(\Gamma_{kl}^TV)\) can be nonparametrically estimated by

\[
\hat{M}_k(\Gamma_k^TV) = \frac{\sum_{j=1}^n K_h(\hat{\Gamma}_k^TV_j - \hat{\Gamma}_k^TV)\delta_{kj}X_{kj}}{\sum_{j=1}^n K_h(\hat{\Gamma}_k^TV_j - \hat{\Gamma}_k^TV)\delta_{kj}},
\]

\[
\hat{Q}_k(\Upsilon_k^TV) = \frac{\sum_{j=1}^n K_h(\hat{\Upsilon}_k^TV_j - \hat{\Upsilon}_k^TV)\delta_{kj}X_{kj}^2}{\sum_{j=1}^n K_h(\hat{\Upsilon}_k^TV_j - \hat{\Upsilon}_k^TV)\delta_{kj}},
\]

\(\hat{\Gamma}_{kl}\) by the method mentioned in Subsection 2.2, \(M_k(\Gamma_k^TV), Q_k(\Upsilon_k^TV)\) and \(R_{kl}(\Gamma_{kl}^TV)\) can be nonparametrically estimated by

\[
\hat{M}_k(\Gamma_k^TV) = \frac{\sum_{j=1}^n K_h(\hat{\Gamma}_k^TV_j - \hat{\Gamma}_k^TV)\delta_{kj}X_{kj}}{\sum_{j=1}^n K_h(\hat{\Gamma}_k^TV_j - \hat{\Gamma}_k^TV)\delta_{kj}},
\]

\[
\hat{Q}_k(\Upsilon_k^TV) = \frac{\sum_{j=1}^n K_h(\hat{\Upsilon}_k^TV_j - \hat{\Upsilon}_k^TV)\delta_{kj}X_{kj}^2}{\sum_{j=1}^n K_h(\hat{\Upsilon}_k^TV_j - \hat{\Upsilon}_k^TV)\delta_{kj}},
\]
2.3 Dimension-reduction-based Imputation for SIR (DRI-SIR)

\[
\hat{R}_{kl}(\hat{\Gamma}_T^{T}V) = \frac{\sum_{j=1}^{n} K_{h}(\hat{\Gamma}_T^{T}V_{j} - \hat{\Gamma}_T^{T}V)\delta_{kj}\delta_{lj}X_{kj}X_{lj}}{\sum_{j=1}^{n} K_{h}(\hat{\Gamma}_T^{T}V_{j} - \hat{\Gamma}_T^{T}V)\delta_{kj}\delta_{lj}},
\]

(2.13)

where \( K_{h}(u) = h^{-r} \prod_{i=1}^{r} K(u_i/h) \) is a multivariate product kernel function with \( r \) denoting the dimension of \( \hat{\Gamma}_T^{T}V_{k} \), \( \hat{\Upsilon}_T^{T}V_{k} \) or \( \hat{\Gamma}_T^{T}V_{kl} \), and the bandwidth \( h \) might take different values when it appears in different places.

With (2.11)-(2.13), the dimension-reduction-based imputation estimators \( \hat{E}(X_{k}), \hat{E}(X_{k}^{2}) \) and \( \hat{E}(X_{k}X_{l}) \) \((k \neq l)\) of \( E(X_{k}), E(X_{k}^{2}) \) and \( E(X_{k}X_{l}) \) with \( 1 \leq k, l \leq p_{1} \), can be respectively expressed as

\[
\hat{E}(X_{k}) = n^{-1} \sum_{i=1}^{n} \{ \delta_{ki}X_{ki} + (1 - \delta_{ki})\hat{M}_{k}(\hat{\Gamma}_T^{T}V_{i}) \},
\]

(2.14)

\[
\hat{E}(X_{k}^{2}) = n^{-1} \sum_{i=1}^{n} \{ \delta_{ki}X_{ki}^{2} + (1 - \delta_{ki})\hat{Q}_{k}(\hat{\Upsilon}_T^{T}V_{i}) \},
\]

(2.15)

\[
\hat{E}(X_{k}X_{l}) = n^{-1} \sum_{i=1}^{n} \{ \delta_{kl}\delta_{li}X_{ki}X_{li} + (1 - \delta_{kl}\delta_{li})\hat{R}_{kl}(\hat{\Gamma}_T^{T}V_{i}) \}.
\]

(2.16)

Based on (2.3), \( E(X_{k}X_{l}^{T}) \) can be estimated by

\[
\hat{E}(X_{k}X_{l}^{T})_{obs} = n^{-1} \sum_{i=1}^{n} \{ \delta_{ki}X_{ki} + (1 - \delta_{ki})\hat{M}_{k}(\hat{\Gamma}_T^{T}V_{i}) \}X_{obs,l}^{T}.
\]

(2.17)

Let \( T_{k}(Y) = E(X_{k}|Y) \). With (2.4), we can estimate \( T_{k}(Y) \) by

\[
\hat{T}_{k}(Y) = n^{-1} \sum_{j=1}^{n} \frac{K_{h}(Y_{j} - Y)\{ \delta_{kj}X_{kj} + (1 - \delta_{kj})\hat{M}_{k}(\hat{\Gamma}_T^{T}V_{j}) \}}{\hat{f}(Y)},
\]

(2.18)

where \( \hat{f}(Y) = n^{-1} \sum_{j=1}^{n} K_{h}(Y_{j} - Y) \) is the kernel estimator of the density function of \( Y \). Consequently, the \( kl \)-th element of \( E\{ E(X_{mis}|Y)E(X_{mis}^{T}|Y) \} \)
2.3 Dimension-reduction-based Imputation for SIR (DRI-SIR)

can be estimated by $n^{-1} \sum_{i=1}^{n} \hat{T}_k(Y_i)\hat{T}_i(Y_i)$. Besides, let $H(Y) = E(X_{obs}^T|Y)$, and its kernel estimator is given by

$$\hat{H}(Y) = n^{-1} \sum_{i=1}^{n} \frac{K_h(Y_i - Y)X_{obs,i}^T}{\hat{f}(Y)}.$$  \hspace{1cm} (2.19)

We then use $n^{-1} \sum_{i=1}^{n} \hat{T}_k(Y_i)\hat{H}(Y_i)$ to estimate $E\{E(X_k|Y)E(X_{obs}^T|Y)\}$, which exactly is the $k$-th raw of $E\{E(X_{mis}|Y)E(X_{obs}^T|Y)\}$.

Finally, by replacing those expectations and conditional expectations in $\Phi_0$, $\Phi_1$ and $\Phi_2$ with their corresponding estimators mentioned above, we can obtain the estimator of the candidate matrix for SIR, say $\hat{\Sigma}^{-1}X\hat{\Sigma}E(X|Y)$, where $\hat{\Sigma}_x = \hat{\Phi}_1 - \hat{\Phi}_0$ and $\hat{\Sigma}_{E(X|Y)} = \hat{\Phi}_2 - \hat{\Phi}_0$. For ease of exposition, $\hat{\Sigma}_x^{-1}\hat{\Sigma}_{E(X|Y)}$ constructed by our proposed method is called the DRI-SIR estimator of $\Sigma^{-1}_{x}E(X|Y)$. Then the eigenvectors corresponding to the first $d$ largest nonzero eigenvalues of $\hat{\Sigma}_x^{-1}\hat{\Sigma}_{E(X|Y)}$ form an estimator of the central subspace $S_{Y|X}$.

**Remark 1.** If the dimension of $S_{E(X_k|V)}, S_{E(\tilde{X}_k|V)}$ or $S_{E(Z_{kl}|V)}$ (i.e. $\Gamma_k^T V$, $\Upsilon_k^T V$ or $\Gamma_{kl}^T V$) is greater than 3, our method might not perform well. In fact, the existing dimension reduction techniques only solve the high-dimension problems partly. In some cases, the dimension might be as small as 1, 2 or 3, while in other cases the dimension may be larger than 3 but smaller than the dimension of the predictors and hence the subsequent statistical inference could be improved, but might not work well. Actually, it is a common
problem that the application of the existing dimension reduction techniques is limited when the structural dimension is not small. On the other hand, as illustrated in some existing literature, the low structural dimension might be sufficient for many practical problems. For example, Cook (1998b) analyzed the Motor Octane data and selected the structural dimension as 1 by his proposed chi-square test; Xia et al. (2002) chose the dimension as 2 for the Hitter’s Salary data using the cross-validation; Ma and Zhu (2012) used the bootstrap procedure to determine the dimension as 1 for the Employee’s Salary data from the fifth National Bank of Springfield. Zhu et al. (2011) also pointed out that, by the purpose of dimension reduction, the structural dimension is generally assumed to be small and takes values 1, 2 or 3.

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 \( Y, \Gamma_k^T V, \Upsilon_k^T V \) and \( \Gamma_{kl}^T V \) \((k \neq l)\), for \( k, l = 1, \ldots, p_1 \). For ease of interpretation, we also introduce the following notations,

\[
\pi_k(\Gamma_k^T V) = P(\delta_k = 1|\Gamma_k^T V) , \quad \tilde{\pi}_k(\Upsilon_k^T V) = P(\delta_k = 1|\Upsilon_k^T V), \\
\pi_{kl}(\Gamma_{kl}^T V) = P(\delta_k \delta_l = 1|\Gamma_{kl}^T V) , \quad m_k(\Gamma_k^T V) = E(\delta_k X_k|\Gamma_k^T V), \\
q_k(\Upsilon_k^T V) = E(\delta_k X_k^2|\Upsilon_k^T V) , \quad w_{kl}(\Gamma_{kl}^T V) = E(\delta_k \delta_l X_k X_l|\Gamma_{kl}^T V),
\]
We now 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 the support in the interval $[-1, 1]$. Moreover, for some positive integer $m$, it satisfies that, $\int_{-1}^{1} K(u)du = 1$, $\int_{-1}^{1} u^i K(u)du = 0$ with $1 \leq i \leq m - 1$, $0 \neq \int_{-1}^{1} u^m K(u)du < \infty$ and $\int_{-1}^{1} K^2(u)du < \infty$.

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

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

**Condition 4.** $f_o(\cdot), g_k(\cdot), a_k(\cdot)$ and $b_{kl}(\cdot)$ have compact supports, and there exist positive constants $c_1, c_2, c_3$ and $c_4$ such that $\inf_y f_o(y) \geq$
2.4 Asymptotic Properties

\[ c_1, \inf_{\Gamma_k^T V} g_k(\Gamma_k^T V) \geq c_2, \inf_{\Gamma_k^T V} a_k(\Gamma_k^T V) \geq c_3 \text{ and } \inf_{\Gamma_{kl}^T V} b_{kl}(\Gamma_{kl}^T V) \geq c_4. \]

**Condition 5.** Each entry in \( XX^T \) has finite fourth-order moment.

We here 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 technically needed for Lemmas A.1-A.3 in Appendix to ensure the desired convergence rate. Specially, condition 3 also indicates that \( m \geq 4 \) is required for our method to reduce the order of the bias of the involved 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 some finite moments, which is necessary for asymptotic normality.

**Theorem 1.** Suppose that the MAR assumption \([1,1]\) and the regularity conditions 1-5 hold, and the dimensions \( r_k, \tilde{r}_k \) and \( r_{kl} \) of the subspaces \( S_{E(X_k|V)} \), \( S_{E(\tilde{X}_k|V)} \) and \( S_{E(Z_{(kl)}|V)} \) \((k \neq l)\) are known, then we have

(i) \( \sqrt{n}\{\text{vec}(\Sigma^{-1}\Sigma_{E(X|Y)}) - \text{vec}(\Sigma^{-1}\Sigma_{E(X|Y)})\} \) converges in distribution to a multivariate normal distribution with mean 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 the dimension \( d \) of the central subspace \( S_{Y|X} \) is known. Let \( \hat{\beta}_1, \ldots, \hat{\beta}_d \) denote the eigenvectors corresponding to the first \( d \) nonzero
2.5 Estimation of Structural Dimension

The structural dimension \( d \) of the central subspace \( S_{Y|X} \) is generally unknown and needed to be estimated. The modified Bayes information criterion (BIC), which is initially developed by Zhu, Miao and Peng (2006) and later modified by Zhu et al. (2010), is employed to estimate the true dimension \( d \) of \( S_{Y|X} \), i.e.

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

where \( \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0 \) are the eigenvalues of \( \hat{\Sigma}_X^{-1} \hat{\Sigma}_{E(X|Y)} \), and \( C_n \) is a penalty constant. Theorem 2 states the consistency of the estimated dimension \( \hat{d} \) 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 \( \hat{d} \) 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) recommended 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 $C_n = 6\log(n) + 3n^{1/3}$ for all model settings.

Remark 2. Our proposed method also need to estimate the generally unknown dimensions $r_k = \dim\{\mathcal{S}_{E(X_k|\mathbf{V})}\}$, $\tilde{r}_k = \dim\{\mathcal{S}_{E(\tilde{X}_k|\mathbf{V})}\}$ and $r_{kl} = \dim\{\mathcal{S}_{E(Z^{(kl)}|\mathbf{V})}\}$ ($k \neq l$), for $k, l = 1, \cdots, p_1$. Their consistent estimators $\hat{r}_k, \tilde{r}_k$ and $\hat{r}_{kl}$ can be obtained by solving similar minimization problems to (2.20), with the major changes being the substitutes of $\hat{\lambda}_i$ in (2.20) by those eigenvalues of the estimated candidate matrices for the corresponding subspaces $\mathcal{S}_{E(X_k|\mathbf{V})}, \mathcal{S}_{E(\tilde{X}_k|\mathbf{V})}$ 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. Five other estimators are also compared in our simulations.

- Full-SIR : Without missingness, SIR is applied to all $n$ observations.
- CC-SIR : The subjects with missing values are removed, and SIR is
applied to the remaining completely observed data.

- AIPW-SIR in [Li and Lu (2008)]: Only one missingness indicator $\delta$ is introduced, and $\delta = 1$ if there is no missingness for all the predictors and 0 otherwise.

- MAIPW-SIR in [Li and Lu (2008)]: There are $p_1$ missingness indicators $(\delta_1, \cdots, \delta_{p_1})$ introduced, and $\delta_k = 1$ if there is no missingness for the $k$-th component $X_k$ of $X_{mis}$ and 0 otherwise, for $k = 1, \cdots, p_1$.


(i) **Evaluation Criteria.**

We assess the performances of the above estimators from two aspects. On the one hand, when we assume that the structural dimension $d$ of $S_{Y|X}$ is known, the trace correlation coefficient (TCC; Hooper 1959) is employed to measure the closeness between the estimated subspace and the true subspace. Let $B_{p \times d}$ be a true basis matrix of $S_{Y|X}$. For an estimator $\hat{B}$ of $B$, the trace correlation coefficient is defined as the positive square root of

$$r^2 = d^{-1} \sum_{i=1}^{d} \phi_i^2,$$

where $1 \geq \phi_1^2 \geq \phi_2^2 \geq \cdots \geq \phi_d^2 \geq 0$ are the eigenvalues of the matrix $\hat{B}_0^T (B_0 B_0^T) \hat{B}_0$, with $\hat{B}_0$ and $B_0$ denoting the orthonormalized versions of $\hat{B}$ and $B$ respectively. A trace correlation coefficient which is closer to 1 means better estimation of the central subspace. On the other hand, when we assume that $d$ is unknown, the empirical distribution (in
percentage) of the estimated dimension $\hat{d}$ is reported to evaluate the efficacy of various methods for determining the structural dimension.

**(ii) Simulation Settings.**

The simulations for each model are repeated 500 times with each of sample size $n = 400$. We take the slice number $H = 10$ that is required by Full-SIR, CC-SIR, AIPW-SIR and MAIPW-SIR. For PI-SIR and the proposed DRI-SIR that involve kernel smoothing, we use a multivariate product kernel $K_h(u) = h^{-r} \prod_{i=1}^{r} K(u_i/h)$, where $K(u) = (2\pi)^{-1/2} \exp(-u^2/2)$ and $r$ is the dimension of the kernel. As the kernel method for a global estimator is insensitive to the choice of bandwidth ([Wang and Rao] 2002), we simply choose a classical bandwidth $h \propto n^{-1/(4+r)}$.

For the purpose of comparison, we consider the following three models,

\begin{align*}
    Y &= (\beta_1^T X)(\beta_1^T X + \beta_2^T X + 3) + 0.5\varepsilon, \quad (3.1) \\
    Y &= \frac{\beta_1^T X}{(\beta_1^T X + 1.5)^2 + 0.5} + 0.5\varepsilon, \quad (3.2) \\
    Y &= 0.5\beta_1^T X + (\beta_2^T X + 2)\varepsilon, \quad (3.3)
\end{align*}

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

\[
P(\delta_k = 1|V) = \frac{\exp(c_0 + \gamma^T_0 V)}{1 + \exp(c_0 + \gamma^T_0 V)}, \quad k = 1, \cdots, p_1, \tag{3.4}
\]

we also consider the logistic quadratic missingness mechanism

\[
P(\delta_k = 1|V) = \frac{\exp(c_0 + (\gamma^T_1 V)^2 + \gamma^T_2 V)}{1 + \exp(c_0 + (\gamma^T_1 V)^2 + \gamma^T_2 V)}, \quad k = 1, \cdots, p_1, \tag{3.5}
\]

where \( V = (X_{p_1+1}, \cdots, X_p, Y)^T \) is always observed and \( c_0 \) is a scalar constant to control the missing proportion (mp). Here the same form of model (3.4) or (3.5) is used for \( P(\delta_k = 1|V) \) 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 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 \( \gamma_0 = (-1, -1, -1, 0, \cdots, 0, 0.5, 0.5, 0.25)^T \) with \( q-6 \) zeros, \( \gamma_1 = (0.5, 0, \cdots, 0, -1, 0.25)^T \) with \( q-3 \) zeros and \( \gamma_2 = (0, 1, \cdots, 1, 0, 0)^T \) with \( q-3 \) ones.

As discussed in the introduction, it is necessary to assume parametric
models for implementing AIPW-SIR, MAIPW-SIR and PI-SIR. Theoretically speaking, we should evaluate the performance of these three methods in two cases, where all or a part of the required parametric models are specified correctly. However, it is nontrivial to specify parametric models correctly for all the involved quantities including $E(X_k|V)$, $E(X_k^2|V)$, $E(X_kX_l|V)$ and $P(\delta_k = 1|V)$ for both AIPW-SIR and MAIPW-SIR, and $P(\delta_k = 1|V)$, $E(\delta_kX_k|V)$, $E(\delta_kX_k^2|V)$, $E(\delta_k\delta_lX_kX_l|V)$ and $E(\delta_k\delta_l|V)$ for PI-SIR. In the literature, only the case in which the missingness mechanism is specified correctly is considered, without regard to the correctness of other involved parametric models. In view of these facts, we here only consider the comparisons under two special cases, where $P(\delta_k = 1|V)$ with $k = 1, \ldots, p_1$, are specified correctly and incorrectly, respectively. To be specific, whether the missingness mechanism (3.4) or (3.5) holds true, one always imposes the logistic linear forms on the missingness mechanism when implementing AIPW-SIR, MAIPW-SIR and PI-SIR. The former corresponds to the situation where the missingness mechanism is specified correctly, while the latter indicates the misspecification of the missingness mechanism.

(iii) Simulation Results.

Tables 1-2 give the simulation results under model (3.1) with the missingness mechanism (3.4).
Table 1: Comparison of SDR estimation in median TCC for model (3.1) under the missingness mechanism (3.4) with different $p_1$ and different missing proportion (mp)

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

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 when the missingness mechanism is specified correctly. First, the proposed DRI-SIR in most situations performs uniformly better than CC-SIR, AIPW-SIR, MAIPW-SIR and PI-SIR, and even has comparable performances to Full-SIR under small missing proportions. Second, with the increase of the missing proportion, CC-SIR, AIPW-SIR, MAIPW-SIR and PI-SIR perform more and more poorly, but our method is relatively robust. Third, even though the missing proportion exceeds 50%, DRI-SIR still performs well especially with the low-dimensional missing predictors, while CC-SIR, AIPW-SIR, MAIPW-SIR and PI-SIR perform quite poorly.
Table 2: Distribution (in percentage) of the estimated structural dimension for model (3.1) under the missingness mechanism (3.4) with different $p_1$ and different missing proportion (mp)

<table>
<thead>
<tr>
<th>$p_1$</th>
<th>Method</th>
<th>$\hat{d}$</th>
<th>1</th>
<th>2</th>
<th>&gt; 2</th>
<th>1</th>
<th>2</th>
<th>&gt; 2</th>
<th>1</th>
<th>2</th>
<th>&gt; 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Full-SIR</td>
<td>0.0020</td>
<td>0.9980</td>
<td>0.0000</td>
<td></td>
<td>0.0040</td>
<td>0.9960</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DRI-SIR</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td></td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CC-SIR</td>
<td>0.2180</td>
<td>0.7820</td>
<td>0.0000</td>
<td></td>
<td>0.6720</td>
<td>0.3280</td>
<td>0.0000</td>
<td></td>
<td>0.9780</td>
<td>0.0220</td>
</tr>
<tr>
<td></td>
<td>AIPW-SIR</td>
<td>0.0080</td>
<td>0.9760</td>
<td>0.0160</td>
<td></td>
<td>0.0140</td>
<td>0.9360</td>
<td>0.0500</td>
<td></td>
<td>0.0280</td>
<td>0.8120</td>
</tr>
<tr>
<td></td>
<td>MAIPW-SIR</td>
<td>0.0060</td>
<td>0.9900</td>
<td>0.0040</td>
<td></td>
<td>0.0200</td>
<td>0.9280</td>
<td>0.0520</td>
<td></td>
<td>0.0580</td>
<td>0.7540</td>
</tr>
<tr>
<td></td>
<td>PI-SIR</td>
<td>0.1700</td>
<td>0.6620</td>
<td>0.1680</td>
<td></td>
<td>0.1200</td>
<td>0.5920</td>
<td>0.2880</td>
<td></td>
<td>0.0880</td>
<td>0.5880</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Full-SIR</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td></td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DRI-SIR</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td></td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CC-SIR</td>
<td>0.4860</td>
<td>0.5140</td>
<td>0.0000</td>
<td></td>
<td>0.9240</td>
<td>0.0760</td>
<td>0.0000</td>
<td></td>
<td>0.9960</td>
<td>0.0040</td>
</tr>
<tr>
<td></td>
<td>AIPW-SIR</td>
<td>0.0100</td>
<td>0.9500</td>
<td>0.0400</td>
<td></td>
<td>0.0320</td>
<td>0.8580</td>
<td>0.1100</td>
<td></td>
<td>0.1060</td>
<td>0.6360</td>
</tr>
<tr>
<td></td>
<td>MAIPW-SIR</td>
<td>0.0100</td>
<td>0.9620</td>
<td>0.0280</td>
<td></td>
<td>0.0240</td>
<td>0.8220</td>
<td>0.1540</td>
<td></td>
<td>0.0760</td>
<td>0.5240</td>
</tr>
<tr>
<td></td>
<td>PI-SIR</td>
<td>0.2780</td>
<td>0.4660</td>
<td>0.2560</td>
<td></td>
<td>0.1980</td>
<td>0.4780</td>
<td>0.3240</td>
<td></td>
<td>0.1740</td>
<td>0.4680</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Full-SIR</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td></td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DRI-SIR</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td></td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CC-SIR</td>
<td>0.8980</td>
<td>0.1020</td>
<td>0.0000</td>
<td></td>
<td>0.9920</td>
<td>0.0080</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>AIPW-SIR</td>
<td>0.0260</td>
<td>0.9100</td>
<td>0.0640</td>
<td></td>
<td>0.3060</td>
<td>0.5020</td>
<td>0.1920</td>
<td></td>
<td>0.2200</td>
<td>0.5240</td>
</tr>
<tr>
<td></td>
<td>MAIPW-SIR</td>
<td>0.0120</td>
<td>0.9520</td>
<td>0.0360</td>
<td></td>
<td>0.0320</td>
<td>0.7120</td>
<td>0.2560</td>
<td></td>
<td>0.1060</td>
<td>0.4180</td>
</tr>
<tr>
<td></td>
<td>PI-SIR</td>
<td>0.2420</td>
<td>0.5280</td>
<td>0.2300</td>
<td></td>
<td>0.2020</td>
<td>0.5680</td>
<td>0.2300</td>
<td></td>
<td>0.1780</td>
<td>0.6040</td>
</tr>
</tbody>
</table>

Table 2 reports the empirical distribution of the estimated dimension $\hat{d}$ for each method with different $p_1$ and three missing proportions over 500 replications. In nearly all cases, the proposed DRI-SIR can select the true structural dimension with a probability much close to 1, and obviously outperforms other methods. Actually, $\hat{d}$ obtained via the modified BIC (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 make the performance of $\hat{d}$ coincide well with that of the estimated candidate matrix for $S_{Y|X}$. As noted in Section 2.5, however, the choice of $C_n$ requires further research.

The simulation results under model (3.1) with the missingness mechanism (3.5) are presented in Tables 3-4. In this case, the missingness mechanism is misspecified for AIPW-SIR, MAIPW-SIR and PI-SIR. The proposed DRI-SIR in all simulation settings performs uniformly better than CC-SIR, AIPW-SIR, MAIPW-SIR and PI-SIR. In particular, AIPW-SIR, MAIPW-SIR and PI-SIR even have worse performance than CC-SIR in most cases.

Table 3: Comparison of SDR estimation in median TCC for model (3.1) under the missingness mechanism (3.5) with different $p_1$ and different missing proportion (mp)

<table>
<thead>
<tr>
<th>$p_1$</th>
<th>$C_0$</th>
<th>mp</th>
<th>Full-SIR</th>
<th>DRI-SIR</th>
<th>CC-SIR</th>
<th>AIPW-SIR</th>
<th>MAIPW-SIR</th>
<th>PI-SIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>20.06%</td>
<td>0.9532</td>
<td>0.9524</td>
<td>0.9180</td>
<td>0.9008</td>
<td>0.9270</td>
<td>0.6730</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.9</td>
<td>35.18%</td>
<td>0.9535</td>
<td>0.9454</td>
<td>0.9454</td>
<td>0.8802</td>
<td>0.8172</td>
<td>0.8784</td>
</tr>
<tr>
<td></td>
<td>-3.3</td>
<td>50.74%</td>
<td>0.9542</td>
<td>0.9355</td>
<td>0.9355</td>
<td>0.8304</td>
<td>0.7160</td>
<td>0.7832</td>
</tr>
<tr>
<td>5</td>
<td>-1.1</td>
<td>35.41%</td>
<td>0.9483</td>
<td>0.9434</td>
<td>0.9434</td>
<td>0.8966</td>
<td>0.8573</td>
<td>0.9091</td>
</tr>
<tr>
<td></td>
<td>-3.3</td>
<td>50.42%</td>
<td>0.9480</td>
<td>0.9387</td>
<td>0.9387</td>
<td>0.8430</td>
<td>0.7158</td>
<td>0.8420</td>
</tr>
<tr>
<td>10</td>
<td>-1.6</td>
<td>35.03%</td>
<td>0.9465</td>
<td>0.9272</td>
<td>0.9272</td>
<td>0.8251</td>
<td>0.7307</td>
<td>0.8789</td>
</tr>
<tr>
<td></td>
<td>-3.4</td>
<td>49.84%</td>
<td>0.9465</td>
<td>0.9068</td>
<td>0.9068</td>
<td>0.7778</td>
<td>0.4410</td>
<td>0.8000</td>
</tr>
</tbody>
</table>
Table 4: Distribution (in percentage) of the estimated structural dimension for model (3.1) under the missingness mechanism (3.5) with different $p_1$ and different missing proportion (mp)

<table>
<thead>
<tr>
<th>$p_1$</th>
<th>Method</th>
<th>$d$</th>
<th>1</th>
<th>2</th>
<th>$&gt;2$</th>
<th>1</th>
<th>2</th>
<th>$&gt;2$</th>
<th>1</th>
<th>2</th>
<th>$&gt;2$</th>
<th>1</th>
<th>2</th>
<th>$&gt;2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Full-SIR</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DRI-SIR</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CC-SIR</td>
<td>0.1180</td>
<td>0.8820</td>
<td>0.0000</td>
<td>0.3200</td>
<td>0.6800</td>
<td>0.0000</td>
<td>0.7140</td>
<td>0.2860</td>
<td>0.0000</td>
<td>0.6480</td>
<td>0.2900</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AIPW-SIR</td>
<td>0.0400</td>
<td>0.8020</td>
<td>0.0000</td>
<td>0.0700</td>
<td>0.9300</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.6480</td>
<td>0.2900</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAIPW-SIR</td>
<td>0.0200</td>
<td>0.8960</td>
<td>0.0840</td>
<td>0.0460</td>
<td>0.7840</td>
<td>0.1700</td>
<td>0.0500</td>
<td>0.6800</td>
<td>0.2700</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PI-SIR</td>
<td>0.1780</td>
<td>0.5840</td>
<td>0.2380</td>
<td>0.2300</td>
<td>0.4760</td>
<td>0.2940</td>
<td>0.1340</td>
<td>0.4620</td>
<td>0.4040</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Full-SIR</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.9980</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.9980</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DRI-SIR</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CC-SIR</td>
<td>0.1360</td>
<td>0.8640</td>
<td>0.0000</td>
<td>0.4440</td>
<td>0.5560</td>
<td>0.0000</td>
<td>0.8000</td>
<td>0.2000</td>
<td>0.0000</td>
<td>0.6000</td>
<td>0.3380</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AIPW-SIR</td>
<td>0.0580</td>
<td>0.7280</td>
<td>0.2140</td>
<td>0.0740</td>
<td>0.6600</td>
<td>0.2660</td>
<td>0.1000</td>
<td>0.5280</td>
<td>0.3720</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAIPW-SIR</td>
<td>0.0180</td>
<td>0.8960</td>
<td>0.0860</td>
<td>0.0320</td>
<td>0.7720</td>
<td>0.1960</td>
<td>0.0620</td>
<td>0.6000</td>
<td>0.3380</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PI-SIR</td>
<td>0.2820</td>
<td>0.5260</td>
<td>0.1920</td>
<td>0.2780</td>
<td>0.5100</td>
<td>0.2120</td>
<td>0.2740</td>
<td>0.4900</td>
<td>0.2360</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Full-SIR</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DRI-SIR</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CC-SIR</td>
<td>0.0460</td>
<td>0.9540</td>
<td>0.0000</td>
<td>0.1760</td>
<td>0.8240</td>
<td>0.0000</td>
<td>0.4920</td>
<td>0.5080</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AIPW-SIR</td>
<td>0.0140</td>
<td>0.8700</td>
<td>0.1160</td>
<td>0.0360</td>
<td>0.7860</td>
<td>0.1780</td>
<td>0.1500</td>
<td>0.5220</td>
<td>0.3280</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAIPW-SIR</td>
<td>0.0000</td>
<td>0.9800</td>
<td>0.0200</td>
<td>0.0000</td>
<td>0.9560</td>
<td>0.0440</td>
<td>0.0140</td>
<td>0.8660</td>
<td>0.1200</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PI-SIR</td>
<td>0.0700</td>
<td>0.9180</td>
<td>0.0120</td>
<td>0.2460</td>
<td>0.7340</td>
<td>0.0200</td>
<td>0.4920</td>
<td>0.4400</td>
<td>0.0680</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>

The simulation results under models (3.2)-(3.3) with the missingness mechanisms (3.4)-(3.5), given in Tables 5-12 of the supplementary file, indicate the similar features to Tables 1-4. These results indeed reinforce the general quantitative patterns that we have observed in Tables 1-4, and also show the superiority of our proposed DRI-SIR to other methods.
4. Real Data Analysis

We now apply the proposed method to the 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). The primary interest 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, wheel-base, 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 4 sample points with missing responses. For bore and stroke, four sample points contain missing values, while for horsepower and peak-rpm, there are two sample points with missing values. Because the number of sample points with missing values in these four predictors is much small relative to the sample size, we simply delete these 6 points. Among the remaining 195 observations, there are 35 observations with missing values appearing in the predictor “normalized-losses”. To eliminate the influence of the scale, we standardize each predictor. Specially, for the missing predictor “normalized-losses”, standardization is only implemented among the 160 completely observed data points.
The structural dimension of $S_{Y\mid X}$ is chosen as 1 via (2.20). By our proposed method, the first dimension reduction direction is estimated by

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

Figure 1 gives a scatterplot of the log price versus the estimated linear combination of the standardized predictors only using 160 completely observed data points. It 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 a parametric structure for the regression of $Y$ on the linear combination of the standardized predictors.

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

It is a common practice to develop imputation or inverse probability weighted methods such that the standard statistical methods for full data can be applied to the case of missing data. Our proposed method obviously belongs to the former. The further research interests in the SDR field might only need to focus on SDR methods for full data, in the sense that these SDR methods satisfying certain conditions always can be applied to the case of missing predictors with the aid of our proposed imputation procedure.

Our proposed method possesses a typical nonparametric flavour. It is much different from the existing semiparametric methods, i.e. AIPW-SIR (Li and Lu, 2008) and PI-SIR (Zhu, Wang and Zhu, 2012), which require to assume parametric models and hence are hard to be applied to some practical problems. We especially describe the differences between the proposed DRI-SIR and the methods of (Zhu, Wang and Zhu, 2012). They consider the problem of sufficient dimension reduction with missing predictors under two types of missingness mechanism. One is to assume all predictors missing at random, or equivalently \( \delta \perp \!\!\!\perp X \mid Y \). 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 \( E(X_k) = E\{E(X_k \mid Y)\} = E\{E(X_k \mid Y, \delta_k = 1)\} \), which is derived from...
such type of missingness mechanism. They also consider the same MAR assumption as (1.1) in our paper. However, their proposed PI-SIR imposes parametric models on $E(\delta_k X_k|V)$, $E(\delta_k|V)$, $E(\delta_k X_k^2|V)$, $E(\delta_k \delta_l X_k X_l|V)$ and $E(\delta_k \delta_l|V)$ to avoid the curse of dimensionality. It is terribly hard to specify all parametric models correctly and hence PI-SIR runs the great risk of the misspecified parametric models. Instead, our proposed method not only avoids assuming parametric models, but also overcomes the curse of dimensionality. This is exactly the main reason why the numerical performance of our method is uniformly better than others.

The proposed method can be applicable to broader contexts. As pointed out by an anonymous referee, a direct extension is to estimate a general class of conditional expectations where the variables treated as the responses are missing at random and meanwhile the given variables are high-dimensional. The estimation of mean functionals with missing response [Cheng (1994)] is just the most typical example. Another important extension is that all of spectral-decomposition-based SDR methods are applicable to the case of predictors missing at random by the proposed dimension-reduction imputation procedure. Part S1 in the supplementary file demonstrates how to extend to the sliced average variance estimation (SAVE) and principal Hessian direction (PHD). Besides, the proposed method also can work 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 file. These extensions greatly expand the scope of the applicability of our method.

Supplementary Materials

A supplementary file 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 A.1-A.3 in Appendix.

Acknowledgements

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 11171331 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.

Firstly, we show that $S_{E(X_k|V)}^{\{\delta_k=1\}} \subseteq S_{E(X_k|V)}$, for $k = 1, \cdots, p_1$. Suppose
that \( A \) is a basis matrix of \( S_{E(X_k|V)} \), that is, \( E(X_k|V) = E(X_k|A^T V) \).

Under the MAR assumption, \( E(X_k|V, \delta_k = 1) = E(X_k|V) \). Then we have
\[
E(X_k|V, \delta_k = 1) = E(X_k|A^T V),
\]
which implies that
\[
E(X_k|A^T V, \delta_k = 1) = E\{E(X_k|V, \delta_k = 1)|A^T V, \delta_k = 1\}
= E\{E(X_k|A^T V)|A^T V, \delta_k = 1\}
= E(X_k|A^T V)
= E(X_k|V, \delta_k = 1).
\]

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

Secondly, we prove \( S_{E(X_k|V)}^{\{\delta_k=1\}} \supseteq S_{E(X_k|V)} \) in a similar way. Assume that \( A \) is a basis matrix of \( S_{E(X_k|V)}^{\{\delta_k=1\}} \), then we have \( E(X_k|V, \delta_k = 1) = E(X_k|A^T V, \delta_k = 1) \). Together with the MAR assumption, we have
\[
E(X_k|V) = E(X_k|A^T V, \delta_k = 1),
\]
which indicates that
\[
E(X_k|A^T V)
= E\{E(X_k|V)|A^T V\}
= E\{E(X_k|A^T V, \delta_k = 1)|A^T V\}
= E(X_k|A^T V, \delta_k = 1)
= E(X_k|V).
\]
It follows that $S_{E(X_k|V)}^{\delta_k=1}$ is also a mean dimension-reduction subspace for the regression of $X_k$ on $V$. Then $S_{E(X_k|V)}^{\delta_k=1} \supseteq S_{E(X_k|V)}$, because $S_{E(X_k|V)}$ is the minimal mean dimension-reduction subspace of $X_k$ on $V$. This completes the proof of Proposition 1.

Since the technical proofs of Proposition 2-3 are almost similar to that of Proposition 1, we omit the details.

**Appendix B: Proof of Theorem 1.**

We begin with three lemmas to facilitate the proof of Theorem 1. All the technical proofs for these lemmas are provided in the supplementary file. We remark that the notations used here have been defined in Section 2.

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

(i) $\hat{E}(X_k) - E(X_k) = \frac{1}{n} \sum_{i=1}^{n} J_1(X_{ki}, \delta_{ki}, \Gamma_k^T V_i) + o_p(n^{-1/2})$ ,

(ii) $\hat{E}(X_k^2) - E(X_k^2) = \frac{1}{n} \sum_{i=1}^{n} J_2(X_{ki}, \delta_{ki}, \Upsilon_k^T V_i) + o_p(n^{-1/2})$ ,

(iii) $\hat{E}(X_kX_l) - E(X_kX_l) = \frac{1}{n} \sum_{i=1}^{n} J_3(X_{ki}, X_{li}, \delta_{ki}, \delta_{li}, \Gamma_{kl}^T V_i) + o_p(n^{-1/2}), (k \neq l)$,

(iv) $\hat{E}(X_kX_{obs}^T) - E(X_kX_{obs}^T) = \frac{1}{n} \sum_{i=1}^{n} J_4(X_{ki}, X_{obs,i}, \delta_{ki}, \Gamma_k^T V_i) + o_p(n^{-1/2})$ ,

where $\hat{E}(X_k), \hat{E}(X_k^2), \hat{E}(X_kX_l)$ and $\hat{E}(X_kX_{obs}^T)$ are respectively defined in
\[(2.14)-(2.17), \text{ and}
\]
\[
J_1(X_{ki}, \delta_{ki}, \Gamma^T_k V_i) = M_k(\Gamma^T_k V_i) + \frac{\delta_{ki}}{\pi_k(\Gamma^T_k V_i)} \{X_{ki} - M_k(\Gamma^T_k V_i)\} - E(X_k),
\]
\[
J_2(X_{ki}, \delta_{ki}, \Upsilon^T_k V_i) = Q_k(\Upsilon^T_k V_i) + \frac{\delta_{ki}}{\pi_k(\Upsilon^T_k V_i)} \{X^2_{ki} - Q_k(\Upsilon^T_k V_i)\} - E(X^2_k),
\]
\[
J_3(X_{ki}, X_{li}, \delta_{ki}, \delta_{li}, \Gamma^T_k V_i) = R_{kl}(\Gamma^T_{kl} V_i) + \frac{\delta_{ki} \delta_{li}}{\pi_{kl}(\Gamma^T_{kl} V_i)} \{X_{ki} X_{li} - R_{kl}(\Gamma^T_{kl} V_i)\} - E(X_k X_l),
\]
\[
J_4(X_{ki}, X_{obs,i}, \delta_{ki}, \Gamma^T_k V_i) = \left\{\delta_{ki} X_{ki} + (1 - \delta_{ki}) M_k(\Gamma^T_k V_i)\right\} X_{obs,i}^T - E(X_k X_{obs}^T)
\]
\[+ \frac{\delta_{ki}}{\pi_k(\Gamma^T_k V_i)} \{X_{ki} - M_k(\Gamma^T_k V_i)\} E[(1 - \delta_k) X_{obs}^T | \Gamma^T_k V_i]
\]

with \(M_k(\cdot), Q_k(\cdot), R_{kl}(\cdot), \pi_k(\cdot), \tilde{\pi}_k(\cdot)\) and \(\pi_{kl}(\cdot)\) defined in Subsection 2.3-2.4.

**Lemma A.2** Suppose that conditions 1-4 hold. Then, for the case \(1 \leq k, l \leq p_1\), we have the following results,

(i) \[
\frac{1}{n} \sum_{i=1}^n \tilde{T}_k(Y_i) \tilde{T}_i(Y_i) - E\{E(X_k|Y)E(X_i|Y)\} = \frac{1}{n} \sum_{i=1}^n J_5(X_{ki}, X_{li}, \delta_{ki}, \delta_{li}, Y_i, \Gamma^T_k V_i, \Gamma^T_i V_i) + o_p(n^{-1/2}) ,
\]

(ii) \[
\frac{1}{n} \sum_{i=1}^n \hat{T}_k(Y_i) \hat{H}(Y_i) - E\{E(X_k|Y)E(X_{obs}^T|Y)\} = \frac{1}{n} \sum_{i=1}^n J_6(X_{ki}, \delta_{ki}, X_{obs,i}, Y_i, \Gamma^T_k V_i) + o_p(n^{-1/2}) ,
\]

where \(\tilde{T}_k(\cdot)\) and \(\hat{H}(\cdot)\) are defined in (2.18) and (2.19) respectively, and
\[ J_5(X_{ki}, X_{li}, \delta_{ki}, \delta_{li}, Y_i, \Gamma_k^T \mathbf{V}_i, \Gamma_l^T \mathbf{V}_i) \]
\[ = \{\delta_{ki}X_{ki} + (1 - \delta_{ki})M_k(T_k^T \mathbf{V}_i)\}T_l(Y_i) + \{\delta_{li}X_{li} + (1 - \delta_{li})M_l(T_l^T \mathbf{V}_i)\}T_k(Y_i) \]
\[ + E \left\{ \left[ \delta_k X_k + (1 - \delta_k)M_k(T_k^T \mathbf{V}_i) \right] | Y = Y_i \right\} T_l(Y_i) \]
\[ + E \left\{ \left[ \delta_l X_l + (1 - \delta_l)M_l(T_l^T \mathbf{V}_i) \right] | Y = Y_i \right\} T_k(Y_i) \]
\[ + \frac{\delta_{ki}[X_{ki} - M_k(T_k^T \mathbf{V}_i)]}{\pi_k(\Gamma_k^T \mathbf{V}_i)}E[(1 - \delta_k)T_l(Y)|\Gamma_k^T \mathbf{V} = \Gamma_k^T \mathbf{V}_i] \]
\[ + \frac{\delta_{li}[X_{li} - M_l(T_l^T \mathbf{V}_i)]}{\pi_l(\Gamma_l^T \mathbf{V}_i)}E[(1 - \delta_l)T_k(Y)|\Gamma_l^T \mathbf{V} = \Gamma_l^T \mathbf{V}_i] \]
\[ - E \left\{ [\delta_k X_k + (1 - \delta_k)M_k(T_k^T \mathbf{V}_i)] T_l(Y) \right\} - E \left\{ [\delta_l X_l + (1 - \delta_l)M_l(T_l^T \mathbf{V}_i)] T_k(Y) \right\} \]
\[ + E \left\{ T_k(Y)T_l(Y) \right\} - 3T_k(Y_i)T_l(Y_i) + o_p(n^{-1/2}) , \]
\[ J_6(X_{ki}, \delta_{ki}, X_{obs,i}, Y_i, \Gamma_k^T \mathbf{V}_i) \]
\[ = \{\delta_{ki}X_{ki} + (1 - \delta_{ki})M_k(T_k^T \mathbf{V}_i)\} H(Y_i) + T_k(Y_i)X_{obs,i}^T - 2E[T_k(Y)H(Y)] \]

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

**Lemma A.3** Suppose that conditions 1-4 hold. Then we have
\[
\frac{1}{n} \sum_{i=1}^{n} \hat{H}(Y_i) - E\{E(X_{obs}|Y)E(X_{obs}^T|Y)\} = \frac{1}{n} \sum_{i=1}^{n} J_7(X_{obs,i}, Y_i) + o_p(n^{-1/2}) ,
\]
where \( J_7(X_{obs,i}, Y_i) = X_{obs,i}H(Y_i) + H(Y_i)^T X_{obs,i}^T - 2E[H(Y)^T H(Y)] \).

**Remark A.** Lemmas A.1-A.2 only consider the non-trivial cases. For the case of \( p_1 + 1 \leq k, l \leq p \), both \( \hat{E}(X_k) \) and \( \hat{E}(X_k X_l) \) are the usual sample estimates so that each of them is naturally a sum of iid random variables.
Proof of Theorem 1. Observe that,

\[
\hat{\Sigma}_X^{-1}\hat{\Sigma}_{E(X|Y)} - \Sigma_X^{-1}\Sigma_{E(X|Y)} = \\
\Sigma_X^{-1}(\Sigma_X - \hat{\Sigma}_X)\hat{\Sigma}_X^{-1}(\hat{\Sigma}_{E(X|Y)} - \Sigma_{E(X|Y)}) + \Sigma_X^{-1}(\Sigma_X - \hat{\Sigma}_X)\hat{\Sigma}_X^{-1}\Sigma_{E(X|Y)} \\
+ \Sigma_X^{-1}(\hat{\Sigma}_{E(X|Y)} - \Sigma_{E(X|Y)}).
\]  

(A.1)

To prove the asymptotic normality of \(\sqrt{n}\{\text{vec}(\hat{\Sigma}_X^{-1}\hat{\Sigma}_{E(X|Y)}) - \text{vec}(\Sigma_X^{-1}\Sigma_{E(X|Y)})\}\), it suffices to prove that both \(\hat{\Sigma}_X - \Sigma_X\) and \(\hat{\Sigma}_{E(X|Y)} - \Sigma_{E(X|Y)}\) can be asymptotically represented as sums of iid 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 \(\hat{\Sigma}_X - \Sigma_X\) and \(\hat{\Sigma}_{E(X|Y)} - \Sigma_{E(X|Y)}\), denoted by \((\hat{\Sigma}_X - \Sigma_X)_{kl}\) and \((\hat{\Sigma}_{E(X|Y)} - \Sigma_{E(X|Y)})_{kl}\) with \(1 \leq k, l \leq p\), respectively, can be written as

\[
(\hat{\Sigma}_X - \Sigma_X)_{kl} = \hat{E}(X_kX_l) - E(X_kX_l) - \{\hat{E}(X_k)\hat{E}(X_l) - E(X_k)E(X_l)\} \\
= \hat{E}(X_kX_l) - E(X_kX_l) - \{\hat{E}(X_k) - E(X_k)\}E(X_l) - \{\hat{E}(X_l) - E(X_l)\}E(X_k) \\
- \{\hat{E}(X_k) - E(X_k)\}\{\hat{E}(X_l) - E(X_l)\} \tag{A.2}
\]

and...
$$\hat{\Sigma}_{E(X|Y)} - \Sigma_{E(X|Y)} = \hat{\Sigma}_{E(X|Y)} - \Sigma_{E(X|Y)} = \sum_{k=l}^{\hat{E}(X_k|Y)\hat{E}(X_l|Y)} - \sum_{k=l}^{\hat{E}(X_k)\hat{E}(X_l)}$$

\[= \hat{E}\left\{\hat{E}(X_k|Y)\hat{E}(X_l|Y)\right\} - \sum_{k=l}^{\hat{E}(X_k)\hat{E}(X_l)} \hat{E}(X_k) - \hat{E}(X_l) = \hat{E}(X_k) - \hat{E}(X_l) = \hat{E}(X_k) - \hat{E}(X_l),\]

(A.3)

where \(\hat{E}(X_k), \hat{E}(X_k|X_l)\) and \(\hat{E}\left\{\hat{E}(X_k|Y)\hat{E}(X_l|Y)\right\}\) denotes the estimates of \(E(X_k), E(X_k|X_l)\) and \(E\left\{E(X_k|Y)E(X_l|Y)\right\}\), respectively. It then suffices to prove that \(\hat{E}(X_k) - E(X_k) , \hat{E}(X_k|X_l) - E(X_k|X_l)\) and \(\hat{E}\left\{\hat{E}(X_k|Y)\hat{E}(X_l|Y)\right\} - E\left\{E(X_k|Y)E(X_l|Y)\right\}\) can be asymptotically represented as sums of iid random variables, which have been presented in Lemmas A.1-A.2 for the case of \(1 \leq k, l \leq p\), and Lemma A.3 and Remark A for the case of \(p+1 \leq k, l \leq p\).

Next we show details for the asymptotically representations of \(\hat{\Sigma}_X - \Sigma_X\) and \(\hat{\Sigma}_{E(X|Y)} - \Sigma_{E(X|Y)}\). Let

$$\hat{\Sigma}_X - \Sigma_X := n^{-1} \sum_{i=1}^{n} A^{(i)} + o_p(n^{-1/2}),$$

(A.4)

where the block matrix

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

is corresponding to the partition \(X = (X^{T}_{mis}, X^{T}_{obs})^{T}\), 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 (A.2) and Lemma A.1 (i)-(iii), the \( kl \)-th element \( a_{1kl}^{(i)} \) of the sub-matrix \( A_1^{(i)} \) can be expressed as
\[
a_{1kl}^{(i)} = J_3(X_{ki}, X_{li}, \delta_{ki}, \delta_{li}, \Gamma_{kl}^T V_i) - J_1(X_{ki}, \delta_{ki}, \Gamma_k^T V_i) E(X_i) \\
- J_1(X_{li}, \delta_{li}, \Gamma_l^T V_i) E(X_k) \text{ , for } 1 \leq k \neq l \leq p_1 ,
\]
and
\[
a_{1kk}^{(i)} = J_2(X_{ki}, \delta_{ki}, \Gamma_k^T V_i) - 2J_1(X_{ki}, \delta_{ki}, \Gamma_k^T V_i) E(X_k) \text{ , for } 1 \leq k \leq p_1 .
\]

By (A.2), Lemma A.1 (iv) and Remark A, the \( k \)-th raw \( a_{2k}^{(i)} \) of the sub-matrix \( A_2^{(i)} \) can be given by
\[
a_{2k}^{(i)} = J_4(X_{ki}, X_{obs,i}, \delta_{ki}, \Gamma_k^T V_i) - J_1(X_{ki}, \delta_{ki}, \Gamma_k^T V_i) E(X_{obs}) - [X_{obs,i} - E(X_{obs})] ^T E(X_k) .
\]

Besides, the sub-matrix \( A_3^{(i)} \) only involves the completely observed data and hence can be expressed as
\[
A_3^{(i)} = X_{obs} X_{obs}^T - X_{obs} E(X_{obs}) - E(X_{obs}) X_{obs}^T - E(X_{obs} X_{obs}^T) + 2E(X_{obs}) E(X_{obs}^T) .
\]

Similarly, we write
\[
\widehat{\Sigma}_{E(X|Y)} - \Sigma_{E(X|Y)} := n^{-1} \sum_{i=1}^{n} B^{(i)} + o_p(n^{-1/2}) , \quad (A.5)
\]
where the block matrix
\[
B^{(i)} = \begin{pmatrix}
B_1^{(i)} & B_2^{(i)} \\
B_2^{(i)T} & B_3^{(i)}
\end{pmatrix}_{p \times p}
\]
is corresponding to the partition \( X = (X_{\text{mis}}^T, X_{\text{obs}}^T)^T \), 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 (A.3), Lemma A.1 (i) and Lemma A.2 (i), the \( kl \)-th element \( b_{1kl}^{(i)} \) of the sub-matrix \( B_1^{(i)} \) can be expressed as

\[
b_{1kl}^{(i)} = J_5(X_{ki}, \delta_{ki}, Y_i, \Gamma_k^T V_i, \Gamma_l^T V_i) - J_1(X_{ki}, \delta_{ki}, \Gamma_k^T V_i) E(X_i) - J_1(X_{li}, \delta_{li}, \Gamma_l^T V_i) E(X_k).
\]

By (A.3), Lemma A.2 (ii) and Remark A, the \( k \)-th raw \( b_{2k}^{(i)} \) of the sub-matrix \( B_2^{(i)} \) can be given by

\[
b_{2k}^{(i)} = J_6(X_{ki}, \delta_{ki}, X_{\text{obs}, i}, Y_i, \Gamma_k^T V_i) - J_1(X_{ki}, \delta_{ki}, \Gamma_k^T V_i) E(X_i^T) - [X_{\text{obs}, i} - E(X_{\text{obs}})]^T E(X_k).
\]

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

\[
B_3^{(i)} = J_7(X_{\text{obs}, i}, Y_i) - X_{\text{obs}, i} E(X_{\text{obs}}^T) - E(X_{\text{obs}}) X_{\text{obs}, i}^T + 2 E(X_{\text{obs}}) E(X_{\text{obs}}^T).
\]

Finally, according to (A.1), (A.4) and (A.5), some simple algebraic calculations give the result

\[
\hat{\Sigma}_X^{-1} \hat{\Sigma}_{E(X|Y)} - \Sigma_X^{-1} \Sigma_{E(X|Y)} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \Sigma_X^{-1} B^{(i)} - \Sigma_X^{-1} A^{(i)} \Sigma_X^{-1} \Sigma_{E(X|Y)} \right\} + o_p(n^{-1/2}),
\]

(A.6)

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

References


Statistical Association* 91, pp. 983-992.

York: Wiley.

Association* 93, pp. 84-94.


Education Press.


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


