Sufficient Dimension Reduction With Missing Predictors
Lexin Li and Wenbin Lu

In high-dimensional data analysis, sufficient dimension reduction (SDR) methods are effective in reducing the predictor dimension, while retaining full regression information and imposing no parametric models. However, it is common in high-dimensional data that a subset of predictors may have missing observations. Existing SDR methods resort to the complete-case analysis by removing all the subjects with missingness in any of the predictors under inquiry. Such an approach does not make effective use of the data and is valid only when missingness is independent of both observed and unobserved quantities. In this article, we propose a new class of SDR estimators under a more general missingness mechanism that allows missingness to depend on the observed data. We focus on a widely used SDR method, sliced inverse regression, and propose an augmented inverse probability weighted sliced inverse regression estimator (AIPW-SIR). We show that AIPW-SIR is doubly robust and asymptotically consistent and demonstrate that AIPW-SIR is more effective than the complete-case analysis through both simulations and real data analysis. We also outline the extension of the AIPW strategy to other SDR methods, including sliced average variance estimation and principal Hessian directions.

KEY WORDS: Double robustness; Missing at random; Missing covariates; Sliced inverse regression; Sufficient dimension reduction.

1. INTRODUCTION

Sufficient dimension reduction (SDR; Cook 1998) methods have generated considerable interest in the analysis of high-dimensional data since SDR was first proposed by Li (1991). It can effectively reduce the dimension of the predictors, while retaining full regression information and imposing no parametric models. For a regression of a univariate response $Y$ given a $p \times 1$ vector of predictors $X$, SDR seeks a subspace $S$ of minimal dimension such that $Y \perp \! \! \! \perp X|PSX$, where $\perp \! \! \! \perp$ indicates independence and $PS$ is the orthogonal projection onto $S$ with respect to the usual inner product. Under minor conditions (Cook 1996), such a subspace exists and is unique. We call it the central subspace of regression of $Y$ on $X$, denote it by $S_{Y|X}$, and call its dimension, $d = \dim(S_{Y|X})$, the structural dimension of regression. The central subspace can be viewed as a parsimonious population parameter that captures all regression information of $Y$ given $X$. It allows reduction of the predictors from $X$ to $\Gamma^TX$, where $\Gamma$ is a $p \times d$ matrix whose columns form a basis of $S_{Y|X}$. The central subspace is the main object of interest in our dimension reduction inquiry.

A number of model-free methods have been proposed to estimate $S_{Y|X}$ and its variants, for instance, sliced inverse regression (Li 1991), sliced average variance estimation (Cook and Weisberg 1991), and principal Hessian directions (Li 1992). Among them, sliced inverse regression (SIR) is perhaps the most commonly used SDR method, and there have been many elaborations on the original methodology of SIR. Cook and Ni (2005) offered a review of the recent developments in SDR.

In practice, it is not uncommon that the information of some predictors is incomplete on some subjects, by design or by happenstance. Missingness is more often encountered in high-dimensional data, where a subset of predictors is observed for all study subjects, but the data of other predictors are incomplete. It is thus desirable to develop a statistical inferential method that comes as close as possible to the true analysis had data not been missing. The most commonly used method for handling missing data is the complete-case analysis, where all the subjects with missingness in any of the predictors under inquiry are removed. All existing SDR methods employ this strategy when there is missingness in the data. The complete-case analysis can be useful when the vast majority of the data are complete, but may be very inefficient when a fair proportion of the data are missing. Moreover, it is valid in an SDR setup only when the missingness mechanism is missing completely at random (MCAR); that is, missingness is independent of all the observed or unobserved quantities, and the method can yield biased estimates when MCAR is not true. For a comprehensive review of statistical methods handling missing data, see Little and Rubin (2002).

In this article, we consider sufficient dimension reduction estimation under a more general missing-data mechanism, by allowing missingness to depend on the observed data but not on the missing data. This mechanism is termed missing at random (MAR; Rubin 1976). Without loss of generality, we partition the predictor vector $X$ into $(X_1^T, X_2^T)^T$, where $X_1 \in \mathbb{R}^{p_1}$ has complete observations for all subjects, $X_2 \in \mathbb{R}^{p_2}$ contains predictors with missingness in a subset of subjects, and $p_1 + p_2 = p$. Furthermore, we introduce a missingness indicator $R$, which takes value 1 if there is no missingness for all the predictors and 0 otherwise. Then MAR is equivalent to the statement that $R \perp X_2|(Y, X_1)$. In general, the MAR assumption provides a better approximation to the reality and is also less restrictive than the MCAR assumption.

Under the MAR assumption, Robins, Rotnitzky, and Zhao (1994) introduced the augmented inverse probability weighted estimators for general regression problems with missing covariates. See also Scharfstein, Rotnitzky, and Robins (1999), Robins and Rotnitzky (2001), Van der Laan and Robins (2003), Lunceford and Davidian (2004), Bang and Robins (2005), Davidian, Tsiatis, and Leon (2005), and Tsiatis (2006). Motivated by their idea, we propose in this article a family of augmented inverse probability weighted (AIPW) sufficient dimension reduction estimators. Our contributions are twofold. First, the proposed methods enable estimation of the central
sliced inverse regression when a subset of predictors have missing observations, while missingness can depend on the observed data. The proposed estimators make use of data with both complete and incomplete observations and, as such, are more effective than the existing complete-case estimators. Second, whereas the majority of existing methods for handling missing predictors have primarily focused on the homoscedastic linear regression model (Little 1992), the proposed estimators can work with more flexible regression models with nonlinear mean or heteroscedastic variance. This is because the structure of the central subspace permits a variety of regression forms, for instance, the single-index model $Y = f(y_1'X) + \epsilon$, the heteroscedastic model $Y = f_1(y_1'X) + f_2(y_2'X) \times \epsilon$, and the generalized linear model $Y = \sum_{j=1}^{d} f_j(y_j'X) + \epsilon$, and the generalized linear model log $P(Y = 1|X)/P(Y = 0|X) = y_1'X$. In the preceding cases, the $f_j$’s are the smooth link functions, $\epsilon$ represents a random error independent of the predictors, and $y_j$’s form the basis of the central subspace $S_{Y|X}$.

The rest of the article is organized as follows. In Section 2, we first develop an augmented inverse probability weighted sliced inverse regression (AIPW–SIR) estimator. We show that AIPW–SIR is doubly robust, in a manner that we will describe later, and show it is $\sqrt{n}$-consistent under suitable regularity conditions. A consistent estimator of the structural dimension $d = \dim(S_{Y|X})$ is also proposed. The effectiveness of the proposed estimator is demonstrated by both simulations and real data analysis in Section 3. We conclude the article with a discussion in Section 4. An extension of the proposed AIPW strategy to other dimension reduction estimators, including sliced average variance estimation and principal Hessian directions, does not usually affect the SIR estimates. Consequently, SIR can be applied to both continuous and categorical responses. Following (1), Li (1991) showed that Span($\hat{y}_1, \ldots, \hat{y}_d$) is a $\sqrt{n}$-consistent estimator of $S_{Y|X}$. The structural dimension $d = \dim(S_{Y|X})$ can be estimated by the number of nonzero eigenvalues in (2), and there exist asymptotic tests (Li 1991; Bura and Cook 2001) and permutation test (Cook and Yin 2001) to determine $d$.

2. Sliced Inverse Regression with Missing Predictors

2.1 Sliced Inverse Regression

Sliced inverse regression (Li 1991) is perhaps the most widely used method for estimating $S_{Y|X}$. It requires a typically mild linearity condition that $E(X\Gamma'X = x)$ be a linear function of $x$, with $\Gamma$ denoting a basis of $S_{Y|X}$. This condition involves only the marginal distribution of $X$ and holds to a reasonable approximation as $p$ increases (Hall and Li 1993). Given this condition,

$$\text{Span}((\Sigma_X^{-1}\Sigma_{E(X|Y)}) \subseteq S_{Y|X}, \quad (1)$$

where $\Sigma_{E(X|Y)} = \text{Cov}[E(X|Y)] \in \mathbb{R}^{p \times p}$ denotes the covariance matrix of the inverse mean, $\Sigma_X = \text{Cov}(X) \in \mathbb{R}^{p \times p}$ is the predictor covariance. Equation (1) connects the central subspace with the inverse regression of $X$ on $Y$ and is the foundation of SIR.

Given $n$ iid observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, the sample estimate of SIR can be obtained by the spectral decomposition:

$$\hat{\Sigma}_{E(X|Y)} \hat{\gamma}_j = \hat{\lambda}_j \hat{\xi}_j \hat{\gamma}_j \quad \text{for } j = 1, \ldots, d, \quad (2)$$

with $\hat{\gamma}_1, \ldots, \hat{\gamma}_d$ denoting the eigenvectors corresponding to the nonzero eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d > 0$. Here $\hat{\Sigma}_{E(X|Y)}$ and $\hat{\Sigma}_X$ denote the usual sample estimates of $\Sigma_{E(X|Y)}$ and $\Sigma_X$, respectively, where $\hat{\Sigma}_{E(X|Y)} = \text{Cov}[E(X|Y)]$, $\hat{\Sigma}_X = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' / n$, and $\bar{x} = \sum_{i=1}^{n} x_i / n$. In $\hat{\Sigma}_{E(X|Y)}$, the term $E(X|Y)$ can be obtained using the conditional sample mean when $Y$ is discrete, that is, by averaging all the $x$’s with the same value of $y$. When $Y$ is continuous, Li (1991) proposed replacing $Y$ with a discrete version by partitioning the range of $Y$ into $H$ fixed slices. The number of slices $H$ is a tuning parameter, but the choice of $H$ does not usually affect the SIR estimates. Consequently, SIR can be applied to both continuous and categorical responses. Following (1), Li (1991) showed that Span($\hat{y}_1, \ldots, \hat{y}_d$) is a $\sqrt{n}$-consistent estimator of $S_{Y|X}$. The structural dimension $d = \dim(S_{Y|X})$ can be estimated by the number of nonzero eigenvalues in (2), and there exist asymptotic tests (Li 1991; Bura and Cook 2001) and permutation test (Cook and Yin 2001) to determine $d$.

2.2 Augmented Inverse Probability Weighted Estimator

When some covariate values are missing, the usual SDR methods such as SIR cannot be applied directly. The complete-case analysis is currently the widely adopted strategy in SDR estimation with missing predictors. However, it does not effectively use information in the data and may yield biased estimates when the missing completely at random assumption is violated. It is also noteworthy that, in the usual linear regression setup, the complete-case analysis can still be valid when missingness only depends on those missing predictors (i.e., $X_2$) alone (see Little 1992). However, this is no longer true in the SDR setup. Use SIR as an example and, for simplicity, assume $p_2 = 1$ for the moment. As we will see later, SIR requires estimation of $E(X_2)$. When missingness only depends on $X_2$, for instance, most observations with a large value of $X_2$ are missing, estimation of $E(X_2)$ based on the observed data only would be biased, which in turn would cause the complete-case SIR to be biased. To improve efficiency and robustness of the complete-case analysis, we propose a class of new estimators, under a more general missing at random assumption, of the central subspace $S_{Y|X}$ via sliced inverse regression.

For partition $X = (X_1^T, X_2^T)^T$ as defined in Section 1, consider the corresponding partitions:

$$\Phi_0 = E(X)E(X)^T = \left(\begin{array}{cc} E(X_1)E(X_1)^T & E(X_1)E(X_2)^T \\ E(X_2)E(X_1)^T & E(X_2)E(X_2)^T \end{array} \right),$$

$$\Phi_1 = E(X|Y)E(X|Y)^T = \left(\begin{array}{cc} E(X_1|Y)E(X_1|Y)^T & E(X_1|Y)E(X_2|Y)^T \\ E(X_2|Y)E(X_1|Y)^T & E(X_2|Y)E(X_2|Y)^T \end{array} \right),$$

$$\Phi_2 = E(X|X)^T = \left(\begin{array}{cc} E(X_1|X)^T & E(X_1|X)^T \\ E(X_2|X)^T & E(X_2|X)^T \end{array} \right),$$

where all $\Phi_j$’s are $p \times p$ matrices. Then $\Sigma_{E(X|Y)} = \Phi_1 - \Phi_0$ and $\Sigma_X = \Phi_2 - \Phi_0$. To implement SIR, one needs to estimate the following terms: $E(X_1)$, $E(X_1|X_1)$, $E(X_1|Y)$, $E(X_2)$, $E(X_2|X_1)$, $E(X_2|X_2)$, and $E(X_2|Y)$. Note that the first three terms can be estimated as usual, because they involve no missing observations. However, new consistent estimators need to be derived for the last four terms involving $X_2$.

Define the population quantity $\pi = \pi(Y, X_1) = P(R = 1|Y, X_1) \in \mathbb{R}^1$, the probability of no missingness given the completely observed variables. For the case where there is more
than one missing predictor in $\mathbf{X}_2$, that is, $p_2 > 1$, we start with a common missingness indicator $R$, which equals 0 as long as there is one variable in $\mathbf{X}_2$ with missingness. We will later introduce individual missingness indicator $R_i$ for each variable $X_{2j}$, $j = 1, \ldots, p_2$. Throughout the article, we assume that $\pi(\mathbf{Y}, \mathbf{X}_1) > 0$ for all the values of $\mathbf{Y}$ and $\mathbf{X}_1$. In practice, $\pi$ is unlikely to be known beforehand; thus, we posit a parametric model, $\pi(\mathbf{Y}, \mathbf{X}_1; \beta)$, indexed by the parameter $\beta$, on the missingness indicator $R$ given $\mathbf{Y}$ and $\mathbf{X}_1$. We first consider the following inverse probability weighted (IPW) estimators:

$$
\hat{\mathbf{x}}_{\text{IPW}} = \frac{R}{\pi(\beta)} \mathbf{X}_2 \in \mathbb{R}^{p_2} \quad \text{and}
$$

$$
(\mathbf{x}_2^{\text{IPW}})^\top = \frac{R}{\pi(\beta)} \mathbf{x}_2^{\text{IPW}} \in \mathbb{R}^{p_2 \times p_2}.
$$

The following property of the IPW estimators is observed.

**Lemma 1.** Given that $\pi(\mathbf{Y}, \mathbf{X}_1)$ is correctly specified and that the MAR assumption $\mathcal{R} \perp \mathbf{X}_2(\mathbf{Y}, \mathbf{X}_1)$ holds, we have $\mathbb{E}(\hat{\mathbf{X}}_{\text{IPW}}) = \mathbb{E}(\mathbf{X}_2)$, $\mathbb{E}(\hat{\mathbf{X}}_2^{\text{IPW}}) = \mathbb{E}(\mathbf{X}_2^{\text{IPW}})$, $\mathbb{E}(\hat{\mathbf{X}}_2^{\text{IPW}}) = \mathbb{E}(\mathbf{X}_2^{\text{IPW}})$, and $\mathbb{E}(\hat{\mathbf{X}}_2^{\text{IPW}} | \mathbf{Y}) = \mathbb{E}(\mathbf{X}_2 | \mathbf{Y})$.

The proof of Lemma 2 is given in the appendix. We next describe the augmented inverse probability weighted sliced inverse regression estimator.

### 2.3 AIPW–SIR Estimator

Given the data $\{(\mathbf{x}_1, y_1, r_1), \ldots, (\mathbf{x}_n, y_n, r_n)\}$, where $r_i$ is the missingness indicator for the $i$th observation, $i = 1, \ldots, n$, the AIPW–SIR estimator requires estimation of $\pi(\beta)$, $\psi_1(\theta_1)$, and $\psi_2(\theta_2)$. Because $R$ is binary, a natural choice for $\pi(\beta)$ is the logistic regression, but other parametric models can be easily accommodated. Because the missingness indicator $R$ only depends on the observed quantities $\mathbf{Y}$ and $\mathbf{X}_1$, an estimate $\hat{\beta}$ of $\beta$ can be obtained by maximizing the likelihood function $\prod_{i=1}^n \pi(\mathbf{y}_i, \mathbf{x}_1; \hat{\beta})^r(1 - \pi(\mathbf{y}_i, \mathbf{x}_1; \hat{\beta}))^{1-r_i}$. Similarly for $\psi_1(\theta_1)$ and $\psi_2(\theta_2)$, a convenient choice is the linear regression model, but other parametric models can be accommodated as well. Given the MAR assumption, both $\theta_1$ and $\theta_2$ can be estimated by their least squares estimates, $\hat{\theta}_1$ and $\hat{\theta}_2$, provided that the parametric models $\psi_1(\theta_1)$ and $\psi_2(\theta_2)$ are both correctly specified. Write $\hat{\pi}_i = \hat{\pi}(\mathbf{y}_i, \mathbf{x}_1; \hat{\beta})$, $\hat{\psi}_1(\theta_1)$, and $\hat{\psi}_2(\theta_2)$, respectively. The augmented inverse probability weighted (AIPW) variables are then

$$
\hat{\mathbf{x}}_{\text{AIPW}} = \left( \begin{array}{c} r_1 \hat{\mathbf{x}}_{2i} - r_1 - \hat{\pi}_i \hat{\psi}_1 \theta_1 \\ \hat{\mathbf{x}}_{2i} \end{array} \right)
$$

$$
(\hat{\mathbf{x}}_{2i}^{\text{AIPW}})^\top = \left( \begin{array}{c} r_1 \hat{\mathbf{x}}_{2i} - r_1 - \hat{\pi}_i \hat{\psi}_1 \theta_1 \\ \hat{\mathbf{x}}_{2i} \end{array} \right)^\top.
$$

In the preceding estimators, the parametric model $\pi(\beta)$ is fitted based on all the observations $(r_i, y_i, x_{1i})$, $i = 1, \ldots, n$; thus, we can obtain $\hat{\pi}_i$ for every subject $i$. On the other hand, the models $\psi_1(\theta_1)$ and $\psi_2(\theta_2)$ are fitted with only the complete cases $(\mathbf{x}_{2i}, y_i, x_{1i})$ with corresponding $r_i = 1$. Based on the fitted models, we then obtain the estimated values of $\hat{\psi}_1$ and $\hat{\psi}_2$ for all subjects $i = 1, \ldots, n$. For those subjects with missing $\mathbf{x}_{2i}$’s, their inverse probability weighted components are all 0’s, because the corresponding $r_i$’s are 0’s. Subsequently, $\hat{\mathbf{x}}_{2i}^{\text{AIPW}}$ and $\hat{\mathbf{x}}_{2i}^{\text{AIPW}}$ reduce to $\hat{\psi}_1$ and $\hat{\psi}_2$, respectively.

Following Hsing and Carroll (1992) and Zhu and Ng (1995), we assume that each slice $h$ contains $c$ data points, where $c \geq 2$ is a fixed constant. To compute $\sum \hat{\mathbf{X}}(\mathbf{Y}|\mathbf{X})$ and $\sum \hat{\mathbf{X}}$ in SIR, we first obtain

$$
\hat{\mathbf{F}}^{\text{AIPW}}_0 = \left( \begin{array}{c} \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{x}}_{1i}^{\text{AIPW}} \\ \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{x}}_{2i}^{\text{AIPW}} \end{array} \right)^\top,
$$

$$
\hat{\mathbf{F}}^{\text{AIPW}}_2 = \frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{x}}_{2i}^{\text{AIPW}})^\top.
$$

where, for $i = 1, \ldots, n$,

$$
\hat{\mathbf{x}}_{1i}^{\text{AIPW}} = \left( \begin{array}{c} \hat{\mathbf{x}}_{1i}^{\text{AIPW}} \\ \hat{\mathbf{x}}_{2i}^{\text{AIPW}} \end{array} \right),
$$

$$
(\hat{\mathbf{x}}_{2i}^{\text{AIPW}})^\top = \left( \begin{array}{c} \hat{\mathbf{x}}_{2i}^{\text{AIPW}} \end{array} \right)^\top.
$$

$\sum \hat{\mathbf{X}}_{\text{AIPW}}$ and $\sum \hat{\mathbf{X}}_{\text{AIPW}}$ have the same properties as $\sum \hat{\mathbf{X}}$ and $\sum \hat{\mathbf{X}}$.
Next, noting that $\Phi_1 = \Sigma_x(E(Y|X)) + \Phi_0 = \text{Cov}(E(Y|X)) = E(\text{Cov}(Y|X)) + \Phi_0 = E(\text{Cov}(Y|X)) - \Sigma_x$, we thus have

$$\hat{\Phi}_{1\text{APW}} = \frac{1}{n} \sum_{i=1}^{n} \hat{x}_{i\text{APW}} (\hat{x}_{i\text{APW}})^\top - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{c} \sum_{j=1}^{c} \left( \hat{x}_{i\text{APW}}^{(j)(h)} - \frac{1}{c} \sum_{j=1}^{c} \hat{x}_{i\text{APW}}^{(j)(h)} \right) \times \left( \hat{x}_{i\text{APW}}^{(j)(h)} - \frac{1}{c} \sum_{j=1}^{c} \hat{x}_{i\text{APW}}^{(j)(h)} \right)^\top,$$

where the index $j(h)$ denotes the $j$th element in the $h$th slice.

Finally, we obtain the AIPW estimators $\hat{Y}_{1\text{APW}}, \ldots, \hat{Y}_{d\text{APW}}$ and $\hat{\Sigma}_{1\text{APW}}, \ldots, \hat{\Sigma}_{d\text{APW}}$ respectively. We then perform the spectral decomposition following (2) and let $\hat{x}_{i\text{APW}}^{(j)}$ denote the eigenvectors corresponding to the nonzero eigenvalues $\hat{\lambda}_{i\text{APW}} \geq \cdots \geq \hat{\lambda}_{d\text{APW}} > 0$. We call $\text{Span}(\hat{x}_{1\text{APW}}^{(1)}, \ldots, \hat{x}_{d\text{APW}}^{(d)})$ the augmented inverse probability weighted sliced inverse regression estimator of the central subspace $S_{Y|X}$.

To facilitate implementation, a summary of the numerical calculations involved in construction of the AIPW–SIR estimator follows:

1. Choose the models $\pi(\beta), \psi_1(\theta_1)$, and $\psi_2(\theta_2)$ and obtain the sample estimates $\hat{\pi}_i = \pi(Y_i; X_i; \hat{\beta}), \hat{\psi}_1 = \psi_1(Y_i; X_i; \hat{\theta}_1)$, and $\hat{\psi}_2 = \psi_2(Y_i; X_i; \hat{\theta}_2)$.
2. Compute the sample AIPW estimates $\hat{x}_{i\text{APW}}$ and $(\hat{x}_{i\text{APW}}, \hat{x}_{i\text{APW}})$ as in (5).
3. Construct $\hat{\Phi}_{1\text{APW}}, \hat{\Phi}_{2\text{APW}}$, and $\hat{\Phi}_{2\text{APW}}$ as in (6) and (7).
4. Construct $\hat{\Sigma}_{1\text{APW}}(X) = \hat{\Phi}_{1\text{APW}} - \hat{\Phi}_{0\text{APW}}$ and $\hat{\Sigma}_{2\text{APW}} = \hat{\Phi}_{2\text{APW}} - \hat{\Phi}_{0\text{APW}}$. Then perform the spectral decomposition following (2) to obtain an estimate of the basis of $S_{Y|X}$.

2.4 Asymptotic Properties of AIPW–SIR

We study the asymptotic behavior of the proposed AIPW–SIR estimator in this section. Hsing and Carroll (1992) and Zhu and Ng (1995) derived the asymptotic properties of the usual SIR estimator under a set of regularity conditions. We adopt similar conditions.

More specifically, let $\Pi_n(B)$ denote the collection of all the $n$-point partitions, $-B \leq y_{1}^* \leq \cdots \leq y_{n}^* \leq B$, of the close interval $[-B, B]$, where $B > 0$ and $n \geq 1$. Any vector-valued or real-valued function $m(y)$ is said to have a total variation of order $r$ if, for any fixed $B > 0$,

$$\lim_{n \to \infty} \frac{1}{n^r} \sup_{B \in \Pi_n(B)} \sum_{i=1}^{n} \|m(y_{i+1}) - m(y_{i})\| = 0.$$

Furthermore, $m(y)$ is said to be nonexpansive in the metric of $M$ on both sides of $B_0$, if there exist a nondecreasing real-valued function $M$ and a real positive number $B_0$ such that, for any two points, say, $y_1$ and $y_2$, both in $(-\infty, -B_0]$ or both in $[B_0, \infty)$,

$$\|m(y_1) - m(y_2)\| \leq |M(y_1) - M(y_2)|.$$

For the AIPW estimator, let $\beta^*, \theta_1^*, \theta_2^*$ denote the limits of $\hat{\beta}, \hat{\theta}_1$, and $\hat{\theta}_2$, respectively, which are assumed to be interior points of some compact parameter space. Let $\Sigma_{1\text{APW}}, \Sigma_{2\text{APW}}, \Sigma_{3\text{APW}}, \Sigma_{4\text{APW}}, \Sigma_{5\text{APW}}, \Sigma_{6\text{APW}}$, $i = 1, \ldots, n$, denote the corresponding terms obtained from $\Sigma_{1\text{APW}}$ and $(\Sigma_{2\text{APW}}, \Sigma_{3\text{APW}})$ by replacing $\hat{\beta}, \hat{\theta}_1$, and $\hat{\theta}_2$ with $\beta^*, \theta_1^*$, and $\theta_2^*$, respectively. Define the inverse regression function $m(y) = \Sigma_{1\text{APW}}, \Sigma_{2\text{APW}}, \Sigma_{3\text{APW}}, \Sigma_{4\text{APW}}, \Sigma_{5\text{APW}}, \Sigma_{6\text{APW}}$, $i = 1, \ldots, n$, where the second equality is the result of Lemma 2, assuming either $\pi(\beta)$ or $\psi_1(\theta_1)$ and $\psi_2(\theta_2)$ are correctly specified and the MAR assumption holds. Define $\epsilon_i = x_i - m(y_i)$ and $\hat{\epsilon}_{1\text{APW}} = x_{1\text{APW}} - m(y_i)$. Similarly as in Zhu and Ng (1995), we assume that the following conditions are satisfied:

(i) $E(\|x_i\|^4) < \infty$ and $E(\|x_{1\text{APW}}\|^4) < \infty$.
(ii) The inverse regression function $m(y)$ has a total variation of order $r = 1/4$.
(iii) $m(y)$ is not expansive in the metric of $M(y)$ on both sides of a positive number $B_0$ such that $M^4(t)P(y > t) \to 0$ as $t \to \infty$ and $n \to \infty$.
(iv) The elements of Cov($\epsilon_i|y_i$) and Cov($\epsilon_{1\text{APW}}|y_i$) all have total variation of order $r = 1$.

In addition, we assume the conditions of Lemma 2:

(v) Either $\pi(\beta)$ is correctly specified or $\psi_1(\theta_1)$ and $\psi_2(\theta_2)$ are correctly specified.
(vi) The MAR assumption $R \perp X_2(Y, X_1)$ holds.

We then have the following theorem. The proof is given in the appendix.

**Theorem 1.** Assume conditions (i)–(vi) hold

1. $\|\hat{\Phi}_{k\text{APW}} - \Phi_k\| = O_p(n^{-1/2})$ for $k = 0, 1, 2$ as $n \to \infty$.
2. Furthermore, if the linearity condition holds, then $\text{Span}(\hat{\Phi}_{1\text{APW}}, \ldots, \hat{\Phi}_{d\text{APW}})$ is a $\sqrt{n}$-consistent estimator of the central subspace $S_{Y|X}$.

**Remark 1.** Theorem 1 states the convergence rate of $\hat{\Phi}_k, k = 0, 1, 2$. In the proof of the theorem, we actually show that $\sqrt{n}(\text{vec}(<\hat{\Phi}_k\text{APW}> - \text{vec}(\Phi_k)))$, $k = 0, 1, 2$, converges in distribution to a multivariate normal distribution with mean 0 as $n \to \infty$. Here vec$(\cdot)$ denotes the matrix operator that stacks all columns of a matrix to a vector.

**Remark 2.** We have shown the asymptotic behavior of the AIPW–SIR estimator for the case where each slice contains an arbitrary but fixed number $c$ of observations. Similarly, we can obtain the $\sqrt{n}$-convergence of AIPW–SIR for the case where the number of observations $c$ in each slice goes to $\infty$, with a rate following condition (iv) of Theorem 2 of Zhu and Ng (1995), as the sample size goes to $\infty$. The proof is parallel to that of Theorem 1 as given in the appendix and the proof of Theorem 2 of Zhu and Ng (1995), which showed the $\sqrt{n}$-convergence of the usual SIR estimator. Details of the proof are skipped here.

**Remark 3.** More recently, Zhu, Miao, and Peng (2006) studied the asymptotic properties of the usual SIR estimator for the case where the number of predictors $p$ goes to $\infty$ along with the
sample size $n$. We speculate that similar properties can be obtained for our AIPW–SIR estimator under suitable conditions. However, the rate of $p$ and the detailed proof are to be further investigated.

### 2.5 A Marginal Version of AIPW–SIR

The method developed in the previous sections is based on a common missingness indicator $R$. When the number of missing predictors in $X_2$ is larger than 1, that is, $p_2 > 1$, we may achieve greater estimation accuracy by introducing an individual missingness indicator $R_j$ for each variable $X_{2j}$, $j = 1, \ldots, p_2$.

More specifically, define $R_j$ as the binary indicator that takes value 1 if there is no missingness for $X_{2j}$ and 0 otherwise. We still work under the MAR assumption, that is, $R_j \perp \independent X_2| (Y, X_1)$. Parallel to developments in Section 2.2, we posit a parametric model $\pi_j(\beta_j) = \pi_j(Y, X_1; \beta_j) = P(R_j = 1|Y, X_1) \in \mathbb{R}^1$ on the missingness indicator $R_j$ given the fully observed data $Y$ and $X_1$, $1 \leq j \leq p_2$, and a parametric model $\pi_{jk}(\beta_{jk}) = \pi_{jk}(Y, X_1; \beta_{jk}) = P(R_jR_k = 1|Y, X_1) \in \mathbb{R}^1$ on the product of missingness indicators $R_jR_k$, $1 \leq j, k \leq p_2$.

We then consider a marginal version of the inverse probability weighted estimator:

$$X_{2j}^{IPW_m} = \frac{R_j}{\pi_j(\beta_j)} X_{2j} \in \mathbb{R}^1, \quad 1 \leq j \leq p_2,$$

(8)

$$(X_2X_2^T)^{IPW_m}_{jk} = \frac{R_jR_k}{\pi_{jk}(\beta_{jk})} X_{2j}X_{2k} \in \mathbb{R}^1, \quad 1 \leq j, k \leq p_2.
$$

Similarly, we posit a parametric model $\psi_j(\theta_{1j}) = \psi_j(Y, X_1; \theta_{1j}) \in \mathbb{R}^1$ for the conditional mean $E(X_{2j}| Y, X_1)$, $1 \leq j \leq p_2$, a parametric model $\psi_{jk}(\theta_{jk}) = \psi_{jk}(Y, X_1; \theta_{jk}) \in \mathbb{R}^1$ for the conditional mean $E(X_{2j}X_{2k}| Y, X_1)$, $1 \leq j, k \leq p_2$, and introduce a marginal version of the augmented inverse probability weighted estimator:

$$X_{2j}^{AIPW_m} = \frac{R_j}{\pi_j(\beta_j)} X_{2j} - \frac{R_j - \pi_j(\beta_j)}{\pi_j(\beta_j)} \psi_{1j}(\theta_{1j}) \in \mathbb{R}^1,$$

$$1 \leq j \leq p_2,$$

(9)

$$(X_2X_2^T)^{AIPW_m}_{jk} = \frac{R_jR_k}{\pi_{jk}(\beta_{jk})} X_{2j}X_{2k} - \frac{R_jR_k - \pi_{jk}(\beta_{jk})}{\pi_{jk}(\beta_{jk})} \psi_{2jk}(\theta_{2jk}) \in \mathbb{R}^1,$$

$$1 \leq j, k \leq p_2.$$

For brevity, we will refer to the new estimators in (8) and (9) as IPW$_m$ and AIPW$_m$ estimators, respectively, where the subscript $m$ indicates that the marginal models $\pi_j(\beta_j)$, $\psi_{1j}(\theta_{1j})$, and $\psi_{2jk}(\theta_{2jk})$ are employed for individual components of $X_2$.

Because the relative behavior of IPW versus IPW$_m$ is similar to that of AIPW versus AIPW$_m$, we focus our discussion on comparing AIPW with AIPW$_m$. We first note that AIPW$_m$ makes use of the observed data more effectively than AIPW when $p_2 > 1$. For AIPW$_m$, only those observations with missing values in the component $X_{2j}$ are replaced with the corresponding sample estimates $\hat{\psi}_{1j}$’s. This is in contrast to the AIPW estimator in (4), where the entire vector of $X_2$ is replaced even if there is only one variable in $X_2$ missing. On the other hand, AIPW$_m$ requires specification of more marginal models and, as such, estimation of more parameters than AIPW. Consequently, if missingness tends to occur among a common subset of observations, AIPW may be preferred, while if missingness scatters among different observations, AIPW$_m$ might be a better choice. The empirical comparison of the two estimators is given in Section 3.

Under regularity conditions similar to conditions (i)–(vi) in Section 2.4, the AIPW$_m$ estimator is doubly robust and $\sqrt{n}$-consistent, as summarized in the next theorem. A detailed description of the conditions and the proof are given in the appendix.

**Theorem 2.** Assume the conditions (i’)-(vi’) in the appendix, along with the linearity condition. Then Span($\hat{\mathbf{Y}}_1^{AIPW_m}, \ldots, \hat{\mathbf{Y}}_{d_B}^{AIPW_m}$) is a $\sqrt{n}$-consistent estimator of the central subspace $S_{Y|X}$, where $\{\hat{\mathbf{Y}}_1^{AIPW_m}, \ldots, \hat{\mathbf{Y}}_{d_B}^{AIPW_m}\}$ are the eigenvectors constructed from the AIPW$_m$ estimator following (9).

### 2.6 Estimation of Structural Dimension

The structural dimension $d$ of the central subspace $S_{Y|X}$ can be estimated via the number of nonzero eigenvalues in the spectral analysis (2). Usually, $d$ is determined by a sequence of tests of hypotheses $d = m$ versus $d > m$ for $m = 0, \ldots, p-1$. The estimate of $d$ is taken as the minimum $m$ such that the null hypothesis $d = m$ is not rejected at a given nominal level (Li 1991). However, such a sequential test procedure is not consistent in general. Alternatively, Zhu et al. (2006) proposed a consistent Bayesian information criterion type of formula to estimate $d$. We adopt the same criterion formula for our AIPW–SIR estimator.

Let $\Omega = \Sigma_{E(X|Y)} + I_p$, where $I_p$ is the $p \times p$ identity matrix, and $\hat{\Omega}_{AIPW}$ denotes a sample estimate of $\Omega$, say $\hat{\Omega}_{AIPW} = \hat{\Sigma}_{E(X|Y)} + I_p$. Let $\hat{\delta}_1, \ldots, \hat{\delta}_p$ denote the eigenvalues of $\hat{\Omega}$ and let $\kappa$ be the number of $\hat{\delta}_i$’s greater than 1. Then an estimate of $d$ can be obtained as

$$\hat{d} = \arg \max_m \left\{ \frac{1}{2} \sum_{k=1+\min(1,m)}^{p} \log(\hat{\delta}_k) + 1 - \hat{\delta}_k \right\} - C_n(m(2p - m + 1)),$$

(10)

where $C_n$ is a penalty constant. We will discuss the practical choice of $C_n$ in Section 3.3. We next state the strong and weak consistency of $\hat{d}$.

**Theorem 3.** Assume that $\lim_{n \to \infty} C_n/n = 0$ and $\lim_{n \to \infty} C_n = \infty$. Then $\hat{d} - d = o_p(1)$ or $= o(1)$ almost surely.

Given the first conclusion of Theorem 1, we have $||\hat{\Omega}_{AIPW} - \Omega|| = O_p(n^{-1/2})$. Then following theorem 2 of Zhu et al. (2006), the conclusion of Theorem 3 follows.

### 3. SIMULATIONS AND REAL DATA ANALYSIS

We examine in this section the finite-sample performance of the proposed IPW–SIR and AIPW–SIR estimators and their corresponding marginal versions, IPW$_m$–SIR and AIPW$_m$–SIR, respectively. Moreover, we compare the proposed estimators with the complete-case analysis, as well as the imputation method.
3.1 A Single Missing Predictor

For illustration, we start with a mean regression model with only one missing predictor:

\[ Y = y_1^\top X (y_1^\top X + y_2^\top X + 3) + \epsilon, \]

(11)

where \( X = (V_1, \ldots, V_p)^\top \) follows a multivariate normal distribution with mean 0, and the correlation between \( V_i \) and \( V_j \), for \( 1 \leq i < j \leq p \), is \( .3^{i-j} \). The error \( \epsilon \) is standard normally distributed and is independent of \( X \). The central subspace \( \mathcal{S}_Y|X = \text{Span}(y_1, y_2) \), where \( y_1 = (1, 0, \ldots, 0)^\top \) and \( y_2 = (0, 1, 0, \ldots, 0)^\top \), and the structural dimension \( d = 2 \). We chose \( n = 200 \) and \( p = 5 \). The predictor \( V_1 \) has missing observations, with the missingness probability \( \pi \) given the response \( Y \) and all other predictors equal to

\[ \pi = P(R = 1|Y, V_2, \ldots, V_p) = \frac{\exp(c_0 + .25Y + .5V_2 - V_p)}{1 + \exp(c_0 + .25Y + .5V_2 - V_p)}, \]

where \( c_0 \) is a scalar constant that controls the proportion of missingness. The vector correlation coefficient \( \varphi \) (Hotelling 1936) is employed to evaluate the estimation accuracy of various estimators. This measure describes the “closeness” of the estimated and the true subspaces, and it ranges between 0 and 1, with a larger value indicating a better estimate.

Figure 1 summarizes the vector correlation coefficient, out of 100 data replications, versus the missingness proportion. Four estimators are compared: the complete-case SIR (the dotted line), IPW–SIR (the dashed line), AIPW–SIR (the solid line), and a typical SIR estimator given all the simulated data (the horizontal line). The last one serves as a benchmark, indicating the best possible estimate that all methods could achieve, because there is completely no missingness. For both IPW–SIR and AIPW–SIR, the missingness probability \( \pi \) was estimated by a logistic regression of the missingness indicator \( R \) given the response \( Y \) and all the predictors with no missingness, \( (V_2, \ldots, V_p) \). For AIPW–SIR, the augmentation components \( \psi_1 \) and \( \psi_2 \) were estimated by ordinary least squares of \( V_1 \) and \( V_1^2 \) given \( Y \) and \( (V_2, \ldots, V_p) \), respectively. Note that in this example the probability model for \( \pi \) is correctly specified while the models for \( \psi_1 \) and \( \psi_2 \) are misspecified. Figure 1(a) shows the median of \( \varphi \), and Figure 1(b) shows the median absolute deviation of \( \varphi \). It is seen from the plot that both IPW–SIR and AIPW–SIR yield more accurate as well as less variable estimates than the complete-case SIR in our simulated scenario, particularly when the missingness proportion is moderate to large. In addition, both IPW and AIPW estimators achieve high estimation accuracy that is comparable to the benchmark estimator in this example.

We also examined a heteroscedastic model:

\[ Y = (y_1^\top X + 2)^2 + .75y_1^\top X \times \epsilon, \]

(12)

where the predictors \( X \), the error \( \epsilon \), and the missingness indicator \( R \) were generated the same way as before. The central subspace \( \mathcal{S}_Y|X = \text{Span}(y_1) \), where \( y_1 = (1, 0, \ldots, 0)^\top \), and the structural dimension \( d = 1 \). We chose \( n = 100 \) and \( p = 10 \). In this example, the predictor effects are present in both the conditional mean and the conditional variance of \( Y|X \), while in (11), the predictor effects are in the conditional mean only. The simulation results (not shown here) reinforce the general qualitative

Figure 1. Median (a) and median absolute deviation (b) of vector correlation versus the overall missingness proportion. Four estimators are compared, the complete-case SIR (dotted line), IPW–SIR (dashed line), AIPW–SIR (solid line), and the benchmark estimator with full data observations (horizontal line).
patterns we have observed that both IPW and AIPW estimators outperform the complete-case analysis. In addition, based on various simulations we have conducted, the AIPW estimator works better than the IPW estimator, in particular when the missingness proportion is large. One possible reason is that the IPW estimator is prone to unstable inverse probability estimation at the tails, whereas the AIPW estimator makes more effective use of the data and as such can offer further improvement to IPW. There also exist alternative options to deal with weakness of IPW under tail problems; for instance, one could bound the estimated missingness probabilities from below by a positive constant (Hothorn, Bühlmann, Dudoit, Molinaro, and Van der Laan 2006).

3.2 Structural Dimension Estimation

For the structural dimension estimation criterion (10), Zhu et al. (2006) recommended a practical form for the penalty term $C_n = c^{-1}W_n$, where $c$ denotes the number of observations per slice, as defined before, $W_n = a \log(n) + bn^{1/3}$, for some scalar constants $a$ and $b$. In our simulations we have experimented with a number of combinations of $(a, b)$ and have obtained very similar results. Here we only present the outcome for $a = 1$ and $b = 0$.

The response models (11) and (12) are examined, with three missingness proportions, 25%, 50%, and 75%. Table 1 reports the distribution (in percentage) of the estimated $\hat{d}$ out of 100 data replications for the four estimators: the benchmark estimator using all the data, the complete-case SIR, IPW–SIR, and AIPW–SIR. The complete-case estimator is seen to clearly overestimate $d$, in particular when the missingness proportion is large. Both IPW–SIR and AIPW–SIR improve the accuracy of estimating $d$ by a great amount. Moreover, AIPW–SIR achieves an accuracy that is close to the benchmark. We also note that we focused on $d = 1$ and $d = 2$ in our simulations, whereas the proposed methodology applies to $d > 2$ cases as well.

3.3 Multiple Missing Predictors

We next examine the response model (11) but with multiple missing predictors. Here $n = 400$ and $p = 20$. Among the 20 predictors, the first $q$ ones $V_1, \ldots, V_q$ have missingness, and the remaining $p - q$ variables have full observations. The number of missing predictors $q$ takes three values, 5, 10, and 15. The central subspace is spanned by $Y_1 = (1, 0, \ldots, 0)^\top$ and $Y_2 = (0, 0, 0, \ldots, 0)^\top$, where in $Y_2$ there are $q$ 0’s preceding 1. Two missingness mechanisms are investigated. First, for each missing predictor, missingness is completely at random with a common missingness proportion for each missing predictor, and the overall missingness proportion (mp) is recorded.

The remaining $p - q$ variables have full observations. The number of missing predictors $q$ takes three values, 5, 10, and 15. The central subspace is spanned by $Y_1 = (1, 0, \ldots, 0)^\top$ and $Y_2 = (0, 0, 0, \ldots, 0)^\top$, where in $Y_2$ there are $q$ 0’s preceding 1. Two missingness mechanisms are investigated. First, for each missing predictor, missingness is completely at random with a common missingness proportion. As such, missingness is to occur among different subsets of observations for different missing predictor $X_{2j}$. In the second case, missingness depends on the response variable $Y$ such that, if $Y$ is greater than a threshold value, then the observation is missing with probability .99; otherwise, it is missing with probability .01. Consequently, the observation with a larger value of $Y$ is more likely to miss, and missingness tends to occur among a common subset of observations for all missing predictors $X_{2j}$.

Table 2 reports the median of the vector correlation $q$ based on 100 replications for the first missingness mechanism. We examined a series of missingness proportions, [.01, .05, .075, .1], for individual missing predictors and recorded the overall missingness proportion in the table. Depending on the number of missing predictors $q$, the overall missingness proportion can be

<table>
<thead>
<tr>
<th>$q$</th>
<th>mp = 25%</th>
<th>AIPW</th>
<th>AIPWm</th>
<th>Impute</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>.049</td>
<td>.869</td>
<td>.878</td>
<td>.876</td>
<td>.875</td>
</tr>
<tr>
<td></td>
<td>.225</td>
<td>.835</td>
<td>.865</td>
<td>.872</td>
<td>.874</td>
</tr>
<tr>
<td></td>
<td>.324</td>
<td>.814</td>
<td>.857</td>
<td>.872</td>
<td>.874</td>
</tr>
<tr>
<td></td>
<td>.411</td>
<td>.793</td>
<td>.849</td>
<td>.873</td>
<td>.875</td>
</tr>
<tr>
<td>10</td>
<td>.097</td>
<td>.870</td>
<td>.880</td>
<td>.888</td>
<td>.889</td>
</tr>
<tr>
<td></td>
<td>.401</td>
<td>.800</td>
<td>.827</td>
<td>.883</td>
<td>.887</td>
</tr>
<tr>
<td></td>
<td>.540</td>
<td>.753</td>
<td>.786</td>
<td>.877</td>
<td>.881</td>
</tr>
<tr>
<td></td>
<td>.649</td>
<td>.636</td>
<td>.695</td>
<td>.875</td>
<td>.878</td>
</tr>
<tr>
<td>15</td>
<td>.136</td>
<td>.856</td>
<td>.867</td>
<td>.886</td>
<td>.888</td>
</tr>
<tr>
<td></td>
<td>.538</td>
<td>.712</td>
<td>.742</td>
<td>.877</td>
<td>.882</td>
</tr>
<tr>
<td></td>
<td>.691</td>
<td>.645</td>
<td>.612</td>
<td>.872</td>
<td>.879</td>
</tr>
<tr>
<td></td>
<td>.794</td>
<td>.444</td>
<td>.328</td>
<td>.865</td>
<td>.875</td>
</tr>
</tbody>
</table>

NOTE: Five estimators are compared, the complete-case analysis (CC), AIPW, AIPWm, a single imputation, and a hybrid method. Missingness is completely at random with a common missingness proportion for each missing predictor, and the overall missingness proportion (mp) is recorded.

<table>
<thead>
<tr>
<th>$d$</th>
<th>mp = 25%</th>
<th>0</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>
</tr>
<tr>
<td></td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>----</td>
</tr>
</tbody>
</table>

Table 1. Distribution (in percentage) of the estimated structural dimension $d$

<table>
<thead>
<tr>
<th>$d$</th>
<th>mp = 25%</th>
<th>0</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>
</tr>
<tr>
<td></td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>----</td>
</tr>
</tbody>
</table>

Table 2. Comparison of SDR estimators in median vector correlation coefficient

NOTE: Three missingness proportions (mp) and four estimators are examined, the benchmark estimator with full observations (Full), the complete-case estimator (CC), IPW, and AIPW. The column corresponding to the true value of $d$ for each model is shown in boldface.
as large as about 80%. We make two observations. First, the complete-case analysis is generally inferior to AIPW or its marginal version, although the missing completely at random assumption is satisfied in this setup. This is mainly attributed to a large overall missingness proportion when there are multiple missing predictors. Second, the AIPW estimator is comparable to AIPWm when the number of missing predictors $q$ is small, but is outperformed by AIPWm when $q$ is large. This is because missingness occurs among different subsets of observations for different missing predictors, and AIPWm can make use of the observed data more effectively than AIPW, as discussed in Section 2.5.

Table 3 reports the results for the second missingness mechanism where missingness depends on the response variable and the missing at random assumption is satisfied. A series of threshold values for missingness were examined, and the corresponding overall missingness proportions were recorded. We again make two comparisons. First, the complete-case estimator performs worst among all methods, as one would expect. Second, the AIPW estimator is comparable and sometimes slightly better than the AIPWm estimator in this example. This is due to the fact that missingness tends to occur among a common subset of observations for all missing predictors, and AIPW requires less parameter estimation compared with AIPWm.

3.4 Imputation

An alternative approach to analyzing incomplete data is imputation, that is, filling in missing data with plausible values. However, to our knowledge, imputation in the context of sufficient dimension reduction has not yet been studied. In this section, we briefly compare the AIPW estimator with a single imputation method. We also suggest a hybrid approach combining imputation and imputation method is used often in practice, it does not follow some established principles for imputation (see, e.g., Little 1992; Little and Rubin 2002, secs. 4.2 and 4.3). For example, it does not condition on all of the observed data such as $\sum$ and the observed values of $X_2$, and thus it could potentially result in biases under missingness at random (as seen in our second simulation setup later) as well as inefficiencies. In addition, it uses predictions rather than draws from the predictive distribution, and thus it could result in biases for estimating quantities that are not linear in the missing values, such as $X_2X_2^\top$.

It is also noted that the augmentation introduced in the AIPW estimator in effect fills in all missing observations with the augmented components $\psi$. We thus consider a hybrid of augmented inverse probability weighting and imputation. Specifically, we replace $X_2$ with $X_2^{\text{AIPWm}} = (X_2^{\text{AIPW1}}, \ldots, X_2^{\text{AIPWm}})^\top$ as defined in (9) and proceed with a typical SIR given the imputed predictor vector $(X_1^\top, X_2^{\text{AIPWm}})^\top$. In effect, the hybrid method introduces weight to the observed components in $X_2$ and at the same time imputes the missing components of $X_2$ with their regression prediction values. Comparing the hybrid estimator with AIPWm, the difference arises in the estimation of the $E(X_2X_2^\top)$.

The simulation setup in Section 3.3 was employed, and the results are reported in Tables 2 and 3. We first note that the imputation method outperforms the AIPWm estimator slightly in the first setup, because missingness is completely at random, whereas for the second missingness mechanism, the single imputation is outperformed by AIPWm. Second, the hybrid approach is seen to perform best among all estimators for both missingness mechanisms. One possible reason, we believe, is that under the given simulation setup, the average of $E(X_2^{\text{AIPWm}}(X_2^{\text{AIPWm}})^\top)$’s estimates $E(X_2X_2^\top)$ reasonably well, and it requires less parameter estimation than AIPWm. However, it is generally not true that $E(X_2^{\text{AIPWm}}(X_2^{\text{AIPWm}})^\top) = E(X_2X_2^\top)$. Consequently, the hybrid estimator may not be consistent. The theoretical justification for the hybrid method as well as the regression imputation method in the context of SDR estimation is not clear and requires further investigation.

3.5 Automobile Data

We analyzed the automobile data (available from the Machine Learning Repository at the University of California–Irvine) to illustrate the proposed method. The objective of this analysis is to depict the relation between the car price and a set of car features. We focused on 14 attributes with continuous measurements as predictors: Wheelbase, Length, Width, Height, Curb weight, Engine size, Bore, Stroke, Compression ratio, Horsepower, Peak rpm, City mpg, Highway mpg, and Losses. The response is the logarithm of the car price. There were originally 205 samples. Among them, four instances had...
missing response values and were removed from the analysis. For Bore and Stroke, four instances had missing values, and for Horsepower and Peak rpm, two were missing. Because the number of missing values was quite small compared to the total sample size, we simply removed those six instances, and the effect was negligible. For the remaining 195 instances, the predictor Losses measures the average loss per car per year, and it had 35 missing observations. In our analysis, Losses plays the role of $X_2$, and $X_1$ consists of the remaining of 13 predictors. With the aid of a predictor scatterplot matrix (Cook and Weisberg 1994), we transformed Horsepower to its inverse to alleviate potential nonlinear confounding among the predictors.

We applied both the complete-case analysis and AIPW–SIR to the data. The estimated structural dimension based on (10) was 1 by both methods. Figure 2 shows the log price versus the estimated linear combination of the predictors based on AIPW–SIR. Such a plot is called a sufficient summary plot (Cook 1998), because it offers a comprehensive and informative view of the data. A strong linear relation between the two quantities can be clearly observed. Moreover, the complete-case SIR and AIPW–SIR yielded similar estimates for the data; the resulting vector correlation coefficient $\varphi$ equals .91. It is possible that for this particular dataset the missingness mechanism follows missing completely at random.

To further illustrate the effect of the AIPW estimator, we treated the complete data of 160 observations (with 35 instances of missing Losses variable removed) as the full data. We then artificially deleted 80 observations for Losses, following a missing at random mechanism, where the nonmissingness probability $\pi$ followed $1/1 + \exp(-.5 - .1 \log(Price) - \text{Curb weight} - \text{Horsepower})$. We then applied the regular SIR to the full data of 160 observations, the complete-case SIR to the remaining 80 complete observations, and the AIPW–SIR to the 160 observations with 50% missingness in Losses. The resulting coefficient estimates are summarized in Table 4. The estimated standard error for the SIR estimate, based on an approximate formula given in Chen and Li (1998), is also reported (in parentheses). It is clearly seen that the AIPW estimator is pretty close to the one obtained from the full data, with $\varphi = .99$. By contrast, the complete-case SIR is far off the target, with $\varphi = .64$ only.

### 4. DISCUSSION

In this article, we have proposed a class of augmented inverse probability weighted estimators to tackle the general sufficient dimension reduction problem when a subset of predictors has missing observations. The new estimators are $\sqrt{n}$-consistent under suitable conditions and enjoy the desirable property of double robustness, which leads to unbiased estimation as long as one of the regression models involved coincides with the true relationship. As such, it affords the analyst a broader protection against misspecification.

Because the complete-case analysis is currently the widely adopted strategy in SDR estimation when predictors have missing observations, we have primarily focused our comparison of the newly proposed estimators with the complete-case analysis. Although it retains the nonparametric flavor of SDR and is not subject to any potential model misspecification, the complete-case analysis is seen to perform unsatisfactorily in our empirical analysis. We also briefly compare the proposed method with a regression imputation method and suggest a hybrid method of combining augmented inverse probability weighting and imputation. The study, however, is by no means comprehensive and the theoretical justification of both imputation and hybrid methods in an SDR setup is not clear. A thorough investigation of various imputation strategies in conjunction with SDR estimation is warranted for future research.

Throughout this article, we assume the missing at random mechanism. We did not consider the case where the probability of missingness may also depend on the unobservable part of the data ($X_2$). In this latter case, one may have nonidentifiability issues and may not be able to verify the model using the observable data (see also a discussion in Tsiatis 2006, p. 151). Restricting attention to missing at random alone seems to have allowed us a great deal of flexibility and, as such, is the primary focus of this article. We also note that the most general form of an MAR mechanism allows missingness to depend on the observed components in $X_2$ as well, that is, $R \perp X_{2[\text{obs}]} | (Y, X_1)$. Here we have adopted a different version of MAR by requiring that $R \perp X_2 | (Y, X_1)$ and note that this form of MAR has been commonly used in the literature.

### Table 4. Coefficient estimation for the automobile data

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Full</th>
<th>CC</th>
<th>AIPW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wheelbase</td>
<td>.03(08)</td>
<td>.09(08)</td>
<td>.03(09)</td>
</tr>
<tr>
<td>Length</td>
<td>.20(09)</td>
<td>.47(08)</td>
<td>.20(10)</td>
</tr>
<tr>
<td>Width</td>
<td>.09(07)</td>
<td>-.18(07)</td>
<td>.10(08)</td>
</tr>
<tr>
<td>Height</td>
<td>.13(05)</td>
<td>.04(05)</td>
<td>.08(05)</td>
</tr>
<tr>
<td>Curb weight</td>
<td>.58(12)</td>
<td>.55(11)</td>
<td>.63(13)</td>
</tr>
<tr>
<td>Engine size</td>
<td>-.10(11)</td>
<td>.21(13)</td>
<td>-.14(13)</td>
</tr>
<tr>
<td>Bore</td>
<td>-.01(04)</td>
<td>-.18(05)</td>
<td>.00(05)</td>
</tr>
<tr>
<td>Stroke</td>
<td>-.03(04)</td>
<td>-.02(04)</td>
<td>-.03(04)</td>
</tr>
<tr>
<td>Compression ratio</td>
<td>.18(05)</td>
<td>-.02(04)</td>
<td>.19(05)</td>
</tr>
<tr>
<td>Horsepower</td>
<td>-.69(09)</td>
<td>-.48(10)</td>
<td>-.66(17)</td>
</tr>
<tr>
<td>Peak rpm</td>
<td>-.03(04)</td>
<td>-.06(04)</td>
<td>-.01(04)</td>
</tr>
<tr>
<td>City mpg</td>
<td>-.15(14)</td>
<td>.27(14)</td>
<td>-.19(17)</td>
</tr>
<tr>
<td>Highway mpg</td>
<td>.14(14)</td>
<td>-.23(14)</td>
<td>.14(16)</td>
</tr>
<tr>
<td>Losses</td>
<td>.16(03)</td>
<td>.02(04)</td>
<td>.11(05)</td>
</tr>
</tbody>
</table>

NOTE: Three estimators are compared, the full data (160 observations) estimate (Full), the complete-case analysis (CC), and the AIPW estimator (160 observations, with 80 missing in the last predictor).
of augmented inverse probability weighting estimation (see the references in the Introduction).

In an SDR framework, no traditional parametric model assumptions are imposed, whereas in our proposals, models are introduced to estimate the AIPW components $\pi$, $\psi_1$, and $\psi_2$. As such, some of the nonparametric flavor is lost in our SDR estimation with missing predictors. In our limited simulation studies, however, the proposed AIPW estimators work well empirically, even though simple models such as logistic regression or ordinary least squares were used. Moreover, the applicability of the proposed AIPW estimators inherit that of the usual SDR estimators had data not been missing. Because SIR is limited to all continuous predictors, the proposed AIPW–SIR does not apply to categorical or mixture types of predictors. On the other hand, the augmented inverse probability weighting strategy can be extended to SDR methods that accommodate mixture predictors, for instance, partial sliced inverse regression (Chiaromonte, Cook, and Li 2002). We also point out that the empirical performance of the proposed estimators hinges on the number of variables with full data observations ($i.e., p_1$). Not surprisingly, they work best when there exist a relatively large number of variables with full observations that are available to explain missingness. In practice, given a variety of missingness scenarios, a combination of various strategies in dealing with missing observations, including the complete-case analysis, AIPW estimation, and imputation, may offer the best practical solution.

[Received October 2006. Revised January 2008.]

REFERENCES


