

Sufficient Dimension Reduction Meets Two-Sample Regression Estimation

October 23, 2026

Abstract

When conducting regression analysis, researchers often face the situations in which important regressors are unavailable in a given dataset, whereas there is another dataset that contains the ‘missing’ regressors as well as other variables that overlap across the two datasets. In this environment, we can estimate regression coefficients consistently by combining the data. The matched-sample indirect inference (MSII) of Hirukawa and Prokhorov (2018) and plug-in least squares (PILS) of Hirukawa et al. (2023) do that; however, these estimators attain the parametric convergence rate only if the number of overlapping variables is three or less. We improve upon PILS and MSII by modelling the conditional expectation of each missing regressor as a single-index model with an unknown link function where the index coefficients are obtained using sufficient dimension reduction (SDR). We prove that building SDR into PILS permits to recover asymptotically normality with the parametric convergence rate for an arbitrarily fixed dimension of the conditioning set. Monte Carlo simulations confirm desirable finite-sample properties of PILS-SDR, and a real data example from wage modelling illustrates empirical relevance of the new estimator. The paper comes with extensive Supplementary Materials.

JEL Classification Codes: C13; C14; C21.

Keywords: Curse of dimensionality; higher-order kernel function; linearity condition; single-index model; sufficient dimension reduction; two-sample regression estimation.

1 Introduction

Consider the linear regression model

$$Y := X^\top \beta + u := \beta_0 + X_1^\top \beta_1 + X_2^\top \beta_2 + X_{3I}^\top \beta_3 + u, \quad (1)$$

where $X_1 \in \mathbb{R}^{d_1}$, $X_2 \in \mathbb{R}^{d_2}$, $X_{3I} \in \mathbb{R}^{d_{3I}}$, $X = (1, X_1^\top, X_2^\top, X_{3I}^\top)^\top \in \mathbb{R}^d$ and $d := d_1 + d_2 + d_{3I}$. If the dataset were complete in the sense that it contained all variables (Y, X_1, X_2, X_{3I}) , we could estimate $\beta = (\beta_0, \beta_1^\top, \beta_2^\top, \beta_3^\top)^\top$ consistently by OLS under standard assumptions. In this paper we deal with the situation in which X_2 is likely to be relevant but is missing from the dataset. A variable can be missing from a survey when it is regarded to be less important in light of the survey design or when collecting or hand-coding it is expensive. If not all relevant variables are available in a given dataset, it is common to resort to integrating information from diverse sources, also known as statistical data fusion (see, e.g., Ridder and Moffitt, 2007).

To be more explicit, suppose that the dataset at hand, called the *primary sample*, contains (Y, X_1, X_3) , where X_3 can be partitioned into $X_3 = (X_{3I}, X_{3E})$. That is, in addition to (Y, X_1, X_{3I}) , the primary sample contains a set of ‘extra’ variables $X_{3E} \in \mathbb{R}^{d_{3E}}$ that are excluded from (1), due to what is known as exclusion restrictions, but may be correlated with the missing regressor X_2 . While $d_3 := \dim(X_3) = d_{3I} + d_{3E} > 0$ must be the case, either d_{3I} or d_{3E} is allowed to be zero. Also suppose that there is another dataset, called the *auxiliary sample*, which consists of (X_2, X_3) and thus can be used to impute a proxy for X_2 in (1). X_3 overlaps across the primary and auxiliary samples and plays a key role in data fusion. In contrast, X_1 is available only in the primary sample. This setup

is encountered in many labor and health economics application (see, e.g., Murtazashvili et al., 2015; Bollinger and Hirsch, 2006).

Hirukawa and Prokhorov (2018), hereafter HP18, discuss how this setting arises in a wide range of labor market studies as a result of the Census practice known as “hot-deck imputation”, where missing records on earnings are imputed from the reported earnings of respondents with similar attributes. They propose the matched-sample indirect inference (MSII) estimator that corrects the resulting imputation bias analytically. Subsequently, Hirukawa et al. (2023), hereafter HMP23, proposed the plug-in least squares (PILS) estimator that imputes the missing variable nonparametrically using the overlapping variables in the two samples. HP18 and HMP23 establish point identification and $\sqrt{n}$ -consistency of MSII and PILS, where n is the number of observations in the primary sample. However, these estimators are subject to the curse of dimensionality. To explain this issue more precisely, partition $X_3 = (X_{3C}, X_{3D})$, where $X_{3C} \in \mathbb{R}^{d_{3C}}$ and $X_{3D} \in \mathbb{R}^{d_{3D}}$ are continuous and discrete components, respectively, and $d_3 = d_{3C} + d_{3D}$. HP18 and HMP23 demonstrate that d_{3C} must be three or less to attain $\sqrt{n}$ -consistency of MSII and PILS.

The purpose of this paper is to go beyond $d_{3C} = 3$. We focus on extending PILS rather than MSII, and we discuss the technical challenges associated with obtaining similar results for MSII in Supplement E of the Supplementary Materials. Because PILS essentially replaces X_2 with a kernel estimate of the conditional expectation $E(X_2|X_3)$, similar to the two-stage least squares (2SLS) estimation, one may think that the use of a higher-order kernel may mitigate the curse of dimensionality. However, this idea turns out to be less promising than one might expect. As the number of continuous components in X_3 increases, both the kernel order and degree of smoothness in $E(X_2|X_3)$ and in the joint density of X_{3C} must increase simultaneously. Yet, the mean squared error (MSE) of the estimator remains dominated by d_3 rather than by the kernel order.

Instead, we propose to reduce the *effective* dimension using sufficient dimension reduc-

tion (SDR) methods. SDR provides a way to project a high-dimensional set of covariates onto a lower-dimensional subspace spanned by a small number of linear indices while preserving the relevant information about the conditional expectation of the response given the covariates. This leads to a semiparametric single-index model (SIM) of the conditional expectation, with an unspecified link function. The SDR literature originates in statistics and has since been widely applied in machine learning, data mining, signal processing, and bioinformatics (see, e.g., Duan and Li, 1991; Naik and Tsai, 2000; Li et al., 2007). SIMs have been extensively studied in econometrics (see, e.g., Ichimura, 1993; Ahn, 1997) but their use in conjunction with SDR remains new to economic and business statistics.

We call the proposed estimator PILS-SDR. It involves using various SDR methods to estimate d_2 sets of index coefficients on X_3 , equation-by-equation, to obtain the proxy for $E(X_2|X_3)$ based on the auxiliary sample. How the SDR methods perform in comparison with conventional SIM methods, e.g., by Ichimura (1993), may be of independent interest. In the next step, we estimate d_2 unknown link functions using kernel regression smoothing. Although a higher-order kernel must still be employed, we adopt a fourth-order kernel regardless of d_{3C} . This leads to consistent estimation of (1) with X_2 replaced by $E(X_2|X_3)$.

There are very few applications of SDR in economics. Exceptions include Ghosh (2011) and Luo and Zhu (2020), who combine SDR with causal inference. One of this paper’s key contributions is to embed SDR into the two-sample regression framework, which, to the best of our knowledge, has not been done before. Moreover, we allow for heterogeneity in the population from which the two samples are drawn, particularly, relevant when datasets are collected in different ways and/or at different times (see, e.g., Freyberger et al., 2025).

Two other contributions of this paper concern missing discrete regressors and influence functions in SDR. The literature on missing data imputation typically assumes that the missing regressor is continuous, while applications in which proxies of missing binary variables are imputed from a different data source are abundant in economics, marketing and

other fields of business statistics. Our imputation scheme applies equally to both continuous and binary missing regressors. Additionally, asymptotic variance matrix for PILS-SDR involves outer products of influence functions and must be estimated using a combination of analytical and resampling methods. Most of the influence functions that arise in standard cases are analytical but for SDR they remain unspecified. We therefore develop a jackknife-based approach to estimating this unknown quantity in our derivations.

The remainder of the paper is organized as follows. Section 2 briefly reviews relevant literature, defines PILS-SDR and discusses its implementation. Section 3 discusses the regularity conditions and explores convergence properties of PILS-SDR. Section 4 presents Monte Carlo simulation results. Section 5 applies PILS-SDR to a return to schooling estimation. Section 6 concludes. The Appendix offers notational convention and a complete set of assumptions. A STATA code implementing PILS-SDR is available on GitHub.¹

Supplementary Materials available online contain six Supplements. Supplement A contains a detailed discussion of the regularity conditions. Supplement B provides technical proofs of the theoretical results. Supplement C offers a consistent estimator of the PILS-SDR covariance matrix. Supplement D contains a corrigendum for the variance formula given in HMP23. Supplement E covers various challenges involved in extending MSII to include SDR. Supplement F contains the full set of Monte Carlo simulations.

2 PILS-SDR: An Overview

2.1 Literature Review

Examples of missing regressors with proxies available from an alternative source are ubiquitous in economics and related fields. A textbook example is a measure of ability in a return to schooling estimation. An estimate of the return suffers from the so called “ability

¹https://github.com/flyingliudi/pils_sdr_public

bias” unless a proxy for ability is included in the regression (see, e.g., Card, 1995). While some micro-level datasets such as National Longitudinal Surveys (NLS) routinely report ability measures (e.g., standardized test scores), others such as Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID) do not contain this information.

Similarly, in studies of gender wage gap, work experience is an important regressor in wage regressions (e.g., Black et al., 1999). While NLS and PSID record actual work experience, the General Household Survey (GHS) and CPS contain no actual work experience. In studies of intergenerational income mobility, parental income is often missing (e.g., Murtaashvili et al., 2015). This occurs because the information about income linkage across generations is unavailable. In models of consumer expenditure, household wealth is a key regressor (e.g., Bostic et al., 2009). However, usually it is not recorded jointly with such variables as household income and expenditures in the same consumer survey data.

As an example from a cognate field, marketing researchers are often interested in determining which media advertisement produces the largest impact on sales for a target customer base (see, e.g., Kamakura and Wedel, 1997). A key variable in answering this question is the duration of exposure to a particular type of advertisement, not just purchase behavior. Panel data on consumption of nondurables often contain information on purchases but not on media exposure. In some cases, the data do not even indicate whether exposure occurred, which constitutes an example of a binary missing regressor.

There is a variety of estimation methods that combine information from two or more datasets. These methods fall into two broad categories. The first is characterized by set identification of the model parameters (see, e.g., D’Haultfoeuille et al., 2025; Hwang, 2026). The second category achieves point identification of the model parameters. This paper, along with HP18 and HMP23, belongs to this class.

A number of two-sample estimation methods have been proposed within the framework of instrumental variables (IV) and generalized method of moments (GMM), where the

required moments can be constructed separately from the two samples so that no matching is required (see, e.g., Inoue and Solon, 2010; Buchinsky et al., 2022). However, these approaches are not applicable to our setting because no sample analogue of $E(X_2Y)$ can be constructed within either of the two samples.

Another class of two-sample estimation methods relies on the primary sample for point identification of the econometric model, while an auxiliary sample is used to improve efficiency (e.g., Hellerstein and Imbens, 1999). Chen et al. (2008) and Graham et al. (2016) use an auxiliary sample to impute missing regressors, similarly to HP18 and HMP23. However, their focus is on parametric and semiparametric efficiency bounds, whereas this paper aims to improve nonparametric imputation methods.

2.2 PILS-SDR Estimation

Let the primary and auxiliary samples be denoted by $\mathcal{S}^A$ and $\mathcal{S}^B$, with sample sizes n and m , respectively. Then, $\mathcal{S}^A = \{(Y_i^A, X_{1i}^A, X_{3i}^A)\}_{i=1}^n$, and $\mathcal{S}^B = \{(X_{2j}^B, X_{3j}^B)\}_{j=1}^m$. These samples may or may not be drawn from the same population. We restate the regression (1) in terms of $\mathcal{S}^A$ as follows:

$$Y^A = X^{A\top} \beta + u^A = \beta_0 + X_1^{A\top} \beta_1 + X_2^{A\top} \beta_2 + X_{3I}^{A\top} \beta_3 + u^A, \quad (2)$$

where $X^A := (1, X_1^{A\top}, X_2^{A\top}, X_{3I}^{A\top})^\top$. The regression (2) would be complete if $X_2^A \in \mathcal{S}^A$.

For imputation purposes, we assume throughout that the conditional expectation $E(X_{2\ell}^s | X_3^s)$ for $\ell \in \{1, \dots, d_2\}$ and $s \in \{A, B\}$ can be expressed as a semiparametric SIM. Specifically,

$$E(X_{2\ell}^s | X_3^s) := g_{2\ell}^s(Z_\ell^s) := g_{2\ell}^s(\theta_\ell^{s\top} X_3^s), \quad (3)$$

where $g_{2\ell}^s(\cdot)$ is an unknown link function, and $Z_\ell^s = \theta_\ell^{s\top} X_3^s \in \mathbb{R}$ is a linear index with

coefficients $\theta_\ell^s \in \mathbb{R}^{d_3}$. Let Θ^s denote the $d_2 \times d_3$ index coefficient matrix, where $\Theta^{s^\top} := \begin{bmatrix} \theta_1^s & \dots & \theta_{d_2}^s \end{bmatrix}$, and define $Z^s := (Z_1^s, \dots, Z_{d_2}^s)^\top$ and $g_2^s(Z^s) := (g_{21}^s(Z_1^s), \dots, g_{2d_2}^s(Z_{d_2}^s))^\top$. Then, the PILS-SDR estimator we propose can be obtained in three steps:

Step 1 (index coefficient estimation). Assume that both the index coefficients and link functions are identical across the two samples, i.e., $\Theta^A = \Theta^B = \Theta$, where $\Theta^\top = \begin{bmatrix} \theta_1 & \dots & \theta_{d_2} \end{bmatrix}$, and $g_2^A(\cdot) = g_2^B(\cdot) =: g_2(\cdot)$. Under this *structural invariance* condition, index coefficients $\theta_1, \dots, \theta_{d_2}$ can be estimated using the auxiliary sample $\mathcal{S}^B$ on an equation-by-equation basis. The key here is to use the SDR direction estimates as described in the next section. This gives $\hat{\Theta}^\top = \begin{bmatrix} \hat{\theta}_1 & \dots & \hat{\theta}_{d_2} \end{bmatrix}$ and $\hat{Z}^s := \hat{\Theta} X_3^s = (\hat{\theta}_1^\top X_3^s, \dots, \hat{\theta}_{d_2}^\top X_3^s)^\top =: (\hat{Z}_1^s, \dots, \hat{Z}_{d_2}^s)^\top$ for $s \in \{A, B\}$.

Step 2 (link function estimation). The link functions $g_{21}(z), \dots, g_{2d_2}(z)$ are estimated equation-by-equation using the Nadaraya-Watson (NW) regression smoother

$$\hat{g}_{2\ell}(z) := \frac{\hat{r}_\ell(z)}{\hat{p}_\ell(z)} := \frac{\frac{1}{mh_\ell} \sum_{j=1}^m X_{2\ell,j}^B K\left(\frac{\hat{Z}_{\ell,j}^B - z}{h_\ell}\right)}{\frac{1}{mh_\ell} \sum_{j=1}^m K\left(\frac{\hat{Z}_{\ell,j}^B - z}{h_\ell}\right)}, \quad (4)$$

where $K(\cdot)$ is a univariate symmetric kernel function and $h_\ell (> 0)$ is the bandwidth for $\ell \in \{1, \dots, d_2\}$. This step yields estimates $\left\{ \hat{g}_2(\hat{Z}_i^A) \right\}_{i=1}^n$ of $\{g_2^A(Z_i^A)\}_{i=1}^n$. Note that even if the missing regressor is binary, the resulting proxy is continuous.

Step 3 (regression estimation). The regression coefficient β in (2) is estimated by OLS from the regression of Y^A on $\hat{X}_g^A := \left(1, X_1^{A^\top}, \hat{g}_2(\hat{Z}^A)^\top, X_{3I}^{A^\top}\right)^\top$. The PILS-SDR is

$$\hat{\beta}_{PS} := Q_{\hat{X}_g^A \hat{X}_g^A}^{-1} R_{\hat{X}_g^A Y^A} := \left(\frac{1}{n} \sum_{i=1}^n \hat{X}_{g,i}^A \hat{X}_{g,i}^{A^\top} \right)^{-1} \frac{1}{n} \sum_{i=1}^n \hat{X}_{g,i}^A Y_i^A. \quad (5)$$

2.3 Implementation

Index coefficient estimation. The step is implemented using as follows:

- (i) Reorder the components of X_3^B so that a continuous component appears first. This normalization facilitates identification of the index coefficients.
- (ii) Estimate the dimension reduction directions $\theta_1, \dots, \theta_{d_2}$ via an SDR method.
- (iii) Normalize each of the SDR estimates $\hat{\theta}_1, \dots, \hat{\theta}_{d_2}$ so that its first element equals one.

When the missing component $X_{2\ell}$ is continuous, the dimension reduction direction can be estimated using a variety of SDR methods. We focus on three methods based on their simplicity: the sliced inverse regression (SIR) of Li (1991) and Duan and Li (1991), the partial least squares (PLS) of Naik and Tsai (2000), and the partial inverse regression (PIR) of Li et al. (2007). For a comprehensive review of the SDR methods, see Dong (2021).

The SIR estimation is implemented in the following procedure:

1. Sort the observations $\{(X_{2\ell,j}^B, X_{3j}^B)\}_{j=1}^m$ by $X_{2\ell}^B$ and partition the range of $\{X_{2\ell,j}^B\}_{j=1}^m$ into H slices $I_1, \dots, I_H$ for some fixed $H \in \mathbb{N}$ so that each slice contains roughly the same number of observations.
2. Compute the sample mean of $\{X_{3j}^B\}_{j=1}^m$ within each slice I_k , $\bar{X}_{3(k)}^B := \sum_{j=1}^m \mathbb{I}\{X_{2\ell,j}^B \in I_k\} X_{3j}^B / m_k$, where $m_k := \sum_{j=1}^m \mathbb{I}\{X_{2\ell,j}^B \in I_k\}$ is the number of observations in slice I_k .
3. Compute the weighted covariance matrix $\hat{\Sigma}_H := \sum_{k=1}^H m_k \left(\bar{X}_{3(k)}^B - \bar{X}_3^B \right) \left(\bar{X}_{3(k)}^B - \bar{X}_3^B \right)^\top / m$, where $\bar{X}_3^B := \sum_{j=1}^m X_{3j}^B / m$. The SIR estimator $\hat{\theta}_\ell^{SIR}$ is then defined as the principal eigenvector of $\hat{\Sigma}_H$ relative to $\hat{\Sigma} := \sum_{j=1}^m (X_{3j}^B - \bar{X}_3^B) (X_{3j}^B - \bar{X}_3^B)^\top / m$.

Unlike SIR, PLS has a closed-form expression. It is defined as $\hat{\theta}_\ell^{PLS} := \hat{\mathcal{K}}_\ell \left(\hat{\mathcal{K}}_\ell^\top \hat{\Sigma} \hat{\mathcal{K}}_\ell \right)^{-1} \hat{\mathcal{K}}_\ell^\top \hat{\sigma}_\ell$, where $\hat{\mathcal{K}}_\ell = \hat{\mathcal{K}}_{\ell(q)} := \begin{bmatrix} \hat{\sigma}_\ell & \hat{\Sigma} \hat{\sigma}_\ell & \dots & \hat{\Sigma}^{q-1} \hat{\sigma}_\ell \end{bmatrix}$ is the $d_3 \times q$ Krylov (1931) matrix generated by $\hat{\Sigma}$ and $\hat{\sigma}_\ell$, and $\hat{\sigma}_\ell := \sum_{j=1}^m (X_{3j}^B - \bar{X}_3^B) (X_{2\ell,j}^B - \bar{X}_{2\ell}^B) / m$, where $\bar{X}_{2\ell}^B := \sum_{j=1}^m X_{2\ell,j}^B / m$. When the order of the Krylov subspace equals the dimension of the overlapping variables (i.e., $q = d_3$), the PLS estimator reduces to the OLS estimator.

PIR is a hybrid of SIR and PLS. Consider a $d_3 \times q$ Krylov matrix obtained by replacing $\hat{\sigma}_\ell$ in $\hat{\mathcal{K}}_\ell$ with $\hat{\Sigma}\hat{\theta}_\ell^{SIR}$. Define $\tilde{\mathcal{K}}_\ell = \tilde{\mathcal{K}}_{\ell(q)} := \begin{bmatrix} \hat{\Sigma}\hat{\theta}_\ell^{SIR} & \hat{\Sigma}^2\hat{\theta}_\ell^{SIR} & \dots & \hat{\Sigma}^q\hat{\theta}_\ell^{SIR} \end{bmatrix}$. The PIR estimator of θ_ℓ is then defined as $\hat{\theta}_\ell^{PIR} := \tilde{\mathcal{K}}_\ell \left(\tilde{\mathcal{K}}_\ell^\top \hat{\Sigma} \tilde{\mathcal{K}}_\ell \right)^{-1} \tilde{\mathcal{K}}_\ell^\top \hat{\theta}_\ell^{SIR}$.

Each of the SIR, PLS and PIR estimators involves user-specified tuning parameters, namely, the number of slices H for SIR and PIR and the order of the Krylov subspace q for PLS and PIR. The performance of SIR is often reported to be insensitive to the choice of H (see, e.g., Li, 1991; Naik and Tsai, 2000). Luo and Zhu (2020) suggest choosing H to be a small number (e.g., 5 or 10), and we adopt $H = 10$. For PLS and PIR, Li et al. (2007) suggest a threshold approach for selecting q . Consider the $d_3 \times d_3$ Krylov matrix $\check{\mathcal{K}}_\ell = \check{\mathcal{K}}_{\ell(d_3)} := \begin{bmatrix} \varsigma_\ell & \hat{\Sigma}\varsigma_\ell & \dots & \hat{\Sigma}^{d_3-1}\varsigma_\ell \end{bmatrix}$, where ς_ℓ is either $\hat{\sigma}_\ell$ for PLS or $\hat{\Sigma}\hat{\theta}_\ell^{SIR}$ for PIR. Let $\lambda_1, \dots, \lambda_{d_3}$ be the eigenvalues of $\check{\mathcal{K}}_\ell \check{\mathcal{K}}_\ell^\top$ in the descending order. Then, Li et al. (2007) propose to select the value

$$q = \sum_{p=1}^{d_3-1} \mathbb{I} \left\{ \frac{\lambda_p}{\lambda_{p+1}} > a \right\} \quad (6)$$

for some prespecified threshold value a . It is also reported by Li et al. (2007) that $a = 1.5$ performs well in practice, and we adopt this value.

When the missing component $X_{2\ell}$ is binary, the implementation of SDR differs from the continuous case. The following methods perform well in practice: (i) SIR with two slices as proposed by Cook and Lee (1999); and (ii) maximum likelihood estimation (MLE) using a Probit or Logit model. For option (i), the two slices are naturally determined by whether the binary component equals 0 or 1. Categorical $X_{2\ell}$ can be handled by converting them into a set of binary indicators. It is worth noting that the Probit and Logit coefficient estimators are consistent up to scale under the linearity condition (see Section 3.1), even if the error distribution is misspecified (see, e.g., Ruud, 1983).

There is a rich literature on estimation of index coefficients in semiparametric SIMs, outside the SDR framework. It can be roughly divided into iterative (e.g., Ichimura, 1993;

Klein and Spady, 1993) and non-iterative (e.g., Powell et al., 1989; Härdle and Stoker, 1989). The former is known to carry heavy computational costs because it often involves solving nonlinear optimization problems numerically, with objective functions that are not necessarily convex. The latter is often based on average derivative estimation and is subject to the curse of dimensionality (see, e.g., Hristache et al., 2001).

Link Function Estimation. The choice of kernel and bandwidth is the most important practical issue in implementing this step. We use a higher-order kernel. As discussed in Section 3.1, the kernel order can be set equal to four regardless of d_{3C} . An example of a fourth-order kernel is the twiced Gaussian kernel (see, e.g., Newey et al., 2004)

$$K_{TG}(u) = \frac{2 \exp(-u^2/2)}{\sqrt{2\pi}} - \frac{\exp(-u^2/4)}{\sqrt{4\pi}}. \quad (7)$$

The bandwidths $h_1, \dots, h_{d_2}$ should satisfy $h_1, \dots, h_{d_2} \asymp (\log m/m)^\alpha$ for some $\alpha \in (1/8, 1/6)$, again for theoretical reasons. In practice, a simple rule of thumb bandwidth

$$\hat{h}_\ell := \frac{1}{2} \hat{\sigma}_{\hat{Z}_\ell^B} \left(\frac{\log m}{m} \right)^{\bar{\alpha}} \quad (8)$$

works, where $\hat{\sigma}_{\hat{Z}_\ell^B}$ is the sample standard deviation of $\left\{ \hat{Z}_{\ell,j}^B \right\}_{j=1}^m$ for $\ell \in \{1, \dots, d_2\}$ and $\bar{\alpha} = 3/20 \in (1/8, 1/6)$. We use $K_{TG}(u)$ and $\hat{h}_\ell$ in our Monte Carlo and empirical sections.

3 Convergence Properties of PILS-SDR

In this section, we explore the asymptotic properties of the PILS-SDR estimator. We defer a detailed discussion of regularity conditions to Supplement A of the Supplementary Materials and estimation of the asymptotic covariance matrix to Supplement C as both require considerable space. All convergence results are established under the asymptotic regime $n, m \rightarrow \infty$ with $n/m \rightarrow \kappa \in (0, \infty)$. Throughout we assume that $d_{3C} > 3$ and

fixed.² To our knowledge, these results are not available in the literature. Therefore, we derive them fully, starting from the basic regularity conditions.

3.1 Regularity Conditions

We group the regularity conditions into general (Assumptions G1-G3), i.e., assumed to hold for all the dimension reduction methods discussed in Section 2.2, and specific, i.e., assumed for individual components of our approach (Assumptions S1-S7 for SDR and Assumptions P1-P4 for PILS). The assumptions are summarized in the Appendix, and their detailed discussions are provided in Supplement A of the Supplementary Materials.

We allow for possibly heterogeneous populations, in line with recent literature on two-sample estimation (Assumption G1), and we do not explicitly impose relevance of X_2 , or, equivalently, that $\beta_2 \neq 0_{d_2 \times 1}$, because even when the regressor turns out to be irrelevant, the estimation procedure remains valid (Assumption G2). We assume that the error η_2 in the reduced form of X_2 satisfies $E(\eta_2 | X_1^A, X_3^A) = E(\eta_2 | X_3^A) = 0_{d_2 \times 1}$ (Assumption G3). Thus, we can write (2) as a regression of Y^A on $\left(1, X_1^{A\top}, E(X_2^A | X_3^A)^\top, X_{3I}^{A\top}\right)^\top$:

$$Y^A = \beta_0 + X_1^{A\top} \beta_1 + E(X_2^A | X_3^A)^\top \beta_2 + X_{3I}^{A\top} \beta_3 + \epsilon^A, \quad (9)$$

where all regressors are orthogonal to the composite error term $\epsilon^A := u^A + \eta_2^{A\top} \beta_2$. Therefore, if $E(X_2^A | X_3^A)$ were observable, β could be consistently estimated by OLS.

To enable modelling $E(X_{2\ell}^s | X_3^s)$ as a SIM, we postulate the existence of an index coefficient θ_ℓ^s such that the linear index Z_ℓ^s is sufficient for $E(X_{2\ell}^s | X_3^s)$ (Assumption S2). This is known as the *central mean subspace* condition, which allows us to specify $E(X_{2\ell}^s | X_3^s)$ as a (possibly nonlinear) function of Z_ℓ^s (see, e.g., Cook and Li, 2002). It may appear that

²As indicated in Section 5, SIMs actually work for $d_{3C} \geq 2$. Moreover, Corollary 2.1 of He and Shao (2000) suggests that letting d_{3C} diverge to infinity at a suitable rate may still retain $\sqrt{n}$ -consistency of M-estimators. However, asymptotic analysis for this case is beyond the scope of this paper.

we achieve dimensionality reduction by imposing overly stringent structural restrictions in Assumption S2. For example, consistent estimation of θ_ℓ^s under Assumption S2 requires that $g_{3\ell}^s(z) := E(X_3^s | Z_\ell^s = z)$ is linear in z , *the linearity condition*. This would be the case if the marginal distribution of X_3^s were *elliptically symmetric*, e.g., multivariate normal, but is immediately violated once discrete or categorical components are included in X_3^s . However, Lemma 1 of Soale (2023) broadens the applicability of SDR methods to a mix of continuous and categorical covariates. Based on Theorems 5 and 6 in Bickel et al. (2018), this lemma implies that for any vector t , $E(t^\top X_3^s | Z_\ell^s = z)$ is linear in z for a sufficiently large m , and thus $E(X_3^s | Z_\ell^s = z)$ can be approximated by a linear function of z regardless of the distribution of X_3^s .

We require the conditions for identification and consistent estimation of Θ^B using the auxiliary sample (Assumptions S1-S5). Additionally, we impose *structural invariance* that allows us to recover Θ^A of the primary sample (Assumption S6). A similar assumption appears in Freyberger et al. (2025). Assumption S6 allows us to write $\Theta^A = \Theta^B = \Theta$ and $g_2^A(\cdot) = g_2^B(\cdot) = g_2(\cdot)$ and to estimate $g_2(Z^A)$ with the aid of the auxiliary sample.

Influence functions and asymptotic linear representations for SDR estimators are generally complicated, except in a few cases (see, e.g., the influence functions for Probit and Logit in Supplement C of the Supplementary Materials). Theorem 3 of Saracco (1997) and Proposition 10 of Cook et al. (2013) provide the influence functions of SIR and PLS, respectively. The corresponding influence function for PIR is considerably more complex (see Li et al., 2007, Proposition 2). Therefore, we leave the influence function $\varphi_j(\Theta)$ unspecified (Assumption S7). This assumption is a high-level one and ensures that for each SDR estimator, we can write

$$\sqrt{m}(\hat{\theta}_\ell - \theta_\ell) = \frac{1}{\sqrt{m}} \sum_{j=1}^m \varphi_{\ell j}^B(\theta_\ell) + o_p(1) \xrightarrow{d} N(0_{d_3 \times 1}, V_{\varphi_\ell^B}), \quad (10)$$

where $\varphi_{\ell j}^B(\theta_\ell) := \varphi_\ell(X_{2j}^B, X_{3j}^B; \theta_\ell)$ is the influence function and $V_{\varphi_\ell^B} := E \left\{ \varphi_{\ell j}^B(\theta_\ell) \varphi_{\ell j}^B(\theta_\ell)^\top \right\}$ is the asymptotic covariance matrix.

We assume point identification of β (Assumption P1) as well as conditions for consistent kernel-based estimation of link functions $g_2(\cdot)$ (Assumptions P2-P4). Assumption P3 also determines the order of the kernel ν . Suppose that we put $h \asymp (\log m/m)^\alpha$ for some $\alpha > 0$. A simple calculation shows that the range of α satisfying this assumption is $(1/(2\nu), 1/6)$, provided that $2\nu > 6$. Therefore, the kernel order must be at least four, and $\alpha \in (1/8, 1/6)$ is the case for $\nu = 4$, as illustrated in Section 2. As emphasized by Hansen (2008), Assumption P2 is not at all restrictive. It covers almost all commonly used higher-order kernels such as higher-order polynomial kernels and higher-order Gaussian kernels including (7). Assumption P4 is an additional requirement for weak uniform consistency of the “oracle” NW regression estimator which uses the true index Z_ℓ^B , rather than $\hat{Z}_\ell^B$, in (4). This is equivalent to Assumption 4(iii) of HMP23, and a similar condition can also be found in Hansen (2008).

3.2 Consistency and Asymptotic Normality

We now state the convergence properties of the PILS-SDR estimator. Proofs of the theorems below are presented in Supplement B of the Supplementary Materials.

Theorem 1. *If Assumptions G1-G3, S1-S7, and P1-P4 hold, then $\hat{\beta}_{PS} \xrightarrow{p} \beta$ as $n, m \rightarrow \infty$ so that $n/m \rightarrow \kappa \in (0, \infty)$.*

Theorem 2. *If Assumptions G1-G3, S1-S7, and P1-P4 hold, then*

$$\sqrt{n} \left(\hat{\beta}_{PS} - \beta \right) \xrightarrow{d} N \left(0_{(d+1) \times 1}, V_{PS} \right) := N \left(0_{(d+1) \times 1}, \Phi_{PS}^{-1} \Omega_{PS} \Phi_{PS}^{-1} \right)$$

as $n, m \rightarrow \infty$ so that $n/m \rightarrow \kappa \in (0, \infty)$, where $\Phi_{PS} := E \left(X_g^A X_g^{A\top} \right)$, $X_g^A := \left(1, X_1^{A\top}, g_2(Z^A)^\top, X_{3I}^{A\top} \right)^\top$, $\Omega_{PS} := \Omega_{PS,1} + \kappa \Omega_{PS,2} := E \left(\psi_{1i} \psi_{1i}^\top \right) + \kappa E \left(\psi_{2j} \psi_{2j}^\top \right)$, $\psi_{1i} := X_{gi}^A \epsilon_i^A = X_{gi}^A \left(u_i^A + \eta_{2i}^{A\top} \beta_2 \right)$,

$\psi_{2j} := \psi_{21j} + \psi_{22j} := \sum_{\ell=1}^{d_2} G_\ell^A(Z_{\ell,j}^B) \eta_{2\ell,j}^B \beta_{2\ell} + \sum_{\ell=1}^{d_2} E \left[X_g^A \{X_3^A - g_{3\ell}^B(Z_\ell^A)\}^\top g_{2\ell}^{(1)}(Z_\ell^A) \right] \varphi_{\ell,j}^B(\theta_\ell) \beta_{2\ell},$
 $G_\ell^A(z) := E(X_g^A | Z_\ell^A = z) = \left(1, E(X_1^A | Z_\ell^A = z)^\top, E\{g_2(Z^A) | Z_\ell^A = z\}^\top, E(X_{3I}^A | Z_\ell^A = z)^\top\right)^\top,$
 and $\varphi_{\ell,j}^B(\theta_\ell)$ is given in (10).

Theorems 1 and 2 jointly establish $\sqrt{n}$ -consistency of the PILS-SDR estimator. Assumptions P1-P4 ensure that $\tilde{g}_{2\ell}(z) \xrightarrow{P} g_{2\ell}(z)$ uniformly in $z \in \mathbb{Z}_\ell$. Assumption P1 also guarantees that Φ_{PS} has full column rank, and thus $\Phi_{PS} > 0$ is guaranteed.

The asymptotic linear representation of $\sqrt{n}(\hat{\beta}_{PS} - \beta)$ in Theorem 2 consists of two independent influence functions, ψ_{1i} and ψ_{2j} . The first influence function ψ_{1i} reflects the sampling error. The proxy $\hat{g}_2(\hat{Z}^A)$ in place of the missing regressor X_2^A is an example of *the generated regressor*, and it inflates the asymptotic variance of $\sqrt{n}(\hat{\beta}_{PS} - \beta)$ by introducing the second influence function ψ_{2j} . This influence function corresponds to the approximation error in $\hat{g}_2(\hat{Z}^A)$, and the two components, ψ_{21j} and ψ_{22j} , account for the estimation errors in the link functions and index coefficients, respectively. As shown in the proof of Lemma B4 in Supplement B of the Supplementary Materials, the derivation of ψ_{21j} largely relies on asymptotic analysis for *two-sample U-statistics*.

Finally, Ω_{PS} , the core part of the asymptotic covariance matrix V_{PS} , is the sum of the outer products of ψ_{1i} and ψ_{2j} , and the latter takes a highly complicated form. For brevity, we defer the discussion of consistent estimation of V_{PS} to Supplement B of the Supplementary Materials. In particular, when the index coefficient θ_ℓ is estimated by SIR, PLS or SIR, its influence function $\varphi_{\ell,j}^B(\theta_\ell)$ is replaced by the jackknife influence function.

4 Finite-Sample Performance

The simulations in this section examine how PILS-SDR performs under different SDR methods used to estimate the index coefficients. The data generating process is given in (2) with $X_1^A \in \mathbb{R}$, $X_2^A = (X_{21}^A, X_{22}^A)^\top \in \mathbb{R}^2$, $X_{3I}^A = (X_{3I1}^A, X_{3I2}^A)^\top \in \mathbb{R}^2$, $\beta_2 = (\beta_{21}, \beta_{22})^\top$,

and $\beta_3 = (\beta_{31}, \beta_{32})^\top$. All true coefficients are unity and the coefficients β_{21} and β_{31} are of interest. The regressors X_{21}^A (continuous) and X_{22}^A (binary) are missing in the primary sample and obtained from an auxiliary sample. The overlapping variables across the two samples, X_3^s for $s \in \{A, B\}$, can be decomposed into $X_3^s = (X_{3I}^{s\top}, X_{3E}^{s\top})^\top \in \mathbb{R}^2 \times \mathbb{R}^8$. Additionally, X_3^s can be partitioned as $X_3^s = (X_{3C}^{s\top}, X_{3D}^{s\top})^\top \in \mathbb{R}^6 \times \mathbb{R}^4$.

To estimate $\beta := (\beta_0, \beta_1, \beta_{21}, \beta_{22}, \beta_{31}, \beta_{32})^\top$, we consider four scenarios. In the first (infeasible) scenario, the complete sample $\mathcal{S}^* = \{(Y_i^A, X_{1i}^A, X_{2i}^A, X_{3Ii}^A)\}_{i=1}^n$ is observed and regression (2) can be estimated by OLS. We denote this benchmark estimator by OLS*. In the second scenario, we simply run OLS on the short regression with X_2^A omitted from (2). This estimator, labelled OLS-S, is generally inconsistent.

The two remaining scenarios cover the PILS-SDR estimator. In the third scenario, we assume that the true index coefficients are known. In this case, the “oracle” primary and auxiliary samples are available and denoted $\mathcal{S}^{A*} = \{(Y_i^A, X_{1i}^A, X_{3i}^A, Z_i^A)\}_{i=1}^n$ and $\mathcal{S}^{B*} = \{(X_{2j}^B, X_{3j}^B, Z_j^B)\}_{j=1}^m$, respectively. Variables Z_i^A and Z_j^B represent the true index values. The corresponding “oracle” PILS-SDR estimator is labelled PILS-SDR*. The fourth (most realistic) scenario involves estimating the index coefficients. In this case, we obtain $\mathcal{S}^A = \{(Y_i^A, X_{1i}^A, X_{3i}^A, \hat{Z}_i^A)\}_{i=1}^n$ and $\mathcal{S}^B = \{(X_{2j}^B, X_{3j}^B, \hat{Z}_j^B)\}_{j=1}^m$ from the estimated indices. The corresponding PILS-SDR estimator is labelled PILS-SDR.

The continuous component $X_{3C}^s \in \mathbb{R}^6$ for $s \in \{A, B\}$ is generated according to one of the two distributions: [Distribution 1] a truncated multivariate normal distribution, i.e., $X_{3C}^s \stackrel{iid}{\sim} TMVN\left(0, V_C^s; \sqrt{X_{3C}^{s\top} (V_C^s)^{-1} X_{3C}^s} \leq 3\right)$, where V_C^s is the covariance matrix with the typical element $(\rho_N^s)^{|p-q|}$, $p, q \in \{1, \dots, 6\}$; and [Distribution 2] an equal-weight mixture of truncated normal and uniform distributions, i.e., each component of X_{3C}^s is distributed as $X_{3C,p}^s \sim (TN_p + U_p)/2$, for $p \in \{1, \dots, 6\}$, where TN_p denotes a standard normal variable truncated so that $|X_{3C,p}^s| \leq 3$, and U_p ’s are uniform random variables on $[-2, 2]$, correlated over p through an underlying multivariate normal with the typical

covariance $(\rho_U^s)^{|p-q|}$, $p, q \in \{1, \dots, 6\}$. Distribution 1 is elliptically symmetric and provides a favorable scenario for SDR in terms of the linearity condition discussed in Section 3.1, while Distribution 2 is a non-elliptical distribution, less favorable for SDR.

The binary component $X_{3D}^s \in \mathbb{R}^4$ for $s \in \{A, B\}$ consists of correlated Rademacher random variables. Each element $X_{3D,p}^s$ is generated as $X_{3D,p}^s := 2\mathbb{I}\{Z_p^\dagger > 0\} - 1$, $p \in \{1, \dots, 4\}$, where the four-dimensional vector $Z^\dagger$ is multivariate normal, $Z^\dagger \stackrel{iid}{\sim} MVN(0, V_D^s)$ with the variance matrix V_D^s having the typical element $(\rho_D^s)^{|p-q|}$, $p, q \in \{1, \dots, 4\}$.

The correlation parameters ρ_N^s, ρ_U^s and ρ_D^s for $s \in \{A, B\}$ are set to $\rho_N^A = 0.40, \rho_N^B = \pm 0.40$; $\rho_U^A = 0.35, \rho_U^B = \pm 0.35$; and $\rho_D^A = 0.40, \rho_D^B = \pm 0.40$. The two populations are homogeneous if $(\rho_N^A, \rho_D^A) = (\rho_N^B, \rho_D^B)$ or $(\rho_U^A, \rho_D^A) = (\rho_U^B, \rho_D^B)$ and heterogeneous otherwise. Finally, the true indices $(Z_1^s, Z_2^s) = (\theta_1^\top X_3^s, \theta_2^\top X_3^s)$ are computed using $X_3^s = (X_{3C}^{s\top}, X_{3D}^{s\top})^\top$ and the true values of the index coefficients are as follows: $\theta_1 = (1, 0, -1, 0, 1, 0, 1/2, 0, -1/2, 0)^\top$ and $\theta_2 = (1, 0, -1, 0, 1, 0, 0, 1/2, 0, -1/2)^\top$.

The remaining variables are generated as follows. For $(u^A, \eta_1^A, v, \eta_{21}) \stackrel{iid}{\sim} MVN(0, I_4)$, $X_1^A := X_{3I1}^A + X_{3I2}^A + \eta_1^A (= X_{3C1}^A + X_{3D1}^A + \eta_1^A)$ and $X_{21}^s := g_{21}(Z_1^s) + \eta_{21}$, $s \in \{A, B\}$. The link function $g_{21}(z)$ is specified using one of the following formulas: [Model A] $g_{21}(z) = z$; [Model B] $g_{21}(z) = z + 9\phi(z)$, where $\phi(\cdot)$ is the standard normal density function; and [Model C] $g_{21}(z) = [2\mathbb{I}\{z > 0\} + 1] \sqrt{|z|}$. Model A is an identity link, and X_{21}^s reduces to a linear regression in this case. Model B is smooth but non-monotone. Model C is non-monotone and non-differentiable at the origin, and the left and right derivatives diverge as $z \rightarrow 0$. The binary missing regressor X_{22}^s for $s \in \{A, B\}$ is given by $X_{22}^s = \mathbb{I}\{Z_2^s + v > 0\}$. The correct specification of X_{22}^s is a Probit model, and thus its link function reduces to $g_{22}(z) = E(X_{22}^s | Z_2^s = z) = \Phi(z)$. Therefore, the error term η_{22}^s can be defined as $\eta_{22}^s := X_{22}^s - \Phi(Z_2^s)$. Finally, Y^A is generated using (2).

We estimate the first index coefficient θ_1 using SIR, PLS and PIR, and the second index coefficient θ_2 using SIR with two slices, Probit, and Logit. This yields nine combina-

tions of estimators: (SIR, SIR) , $(SIR, Probit)$, $(SIR, Logit)$, (PLS, SIR) , $(PLS, Probit)$, $(PLS, Logit)$, (PIR, SIR) , $(PIR, Probit)$, and $(PIR, Logit)$. For the estimates of β_{21} and β_{31} , we report the following performance measures: (i) *Mean* (average of parameter estimates over simulations); (ii) *SD* (standard deviation of the parameter estimates over simulations); (iii) *RMSE* (root mean-squared error of parameter estimates over simulations); (iv) *MedSE* (median of the standard error over simulations); and (v) *CR* (coverage rate of the nominal 95% confidence interval). Standard errors for OLS* and OLS-S are heteroskedasticity-robust. Covariance estimation procedures for PILS-SDR and PILS are described in Supplements C and D of the Supplementary Materials. Standard errors for PILS-SDR* are from the estimate of its asymptotic covariance matrix $\Phi_{PS}^{-1} \Omega_{PS^*} \Phi_{PS}^{-1}$, where $\Omega_{PS^*} := E(\psi_{1i} \psi_{1i}^\top) + \kappa E(\psi_{21j} \psi_{21j}^\top)$ with Φ_{PS} , ψ_{1i} and ψ_{21j} given in Theorem 2. The sample sizes are $(n, m) = (2000, 1000)$, and the number of Monte Carlo replications is 1000.

Table 1 reports the results for the estimators of β_{21} and β_{31} , the coefficients on missing and overlapping regressors, respectively. To save space we report the results only for Model A under Distribution 1 (Panel A) and Model C under Distribution 2 (Panel F). The complete set of results is available in Tables F1-F2 in Supplement F of the Supplementary Materials. It is immediately clear from the table that OLS* outperforms all other estimators as would be expected. The results for OLS-S are extremely poor. The error term in the short regression includes the relevant regressor X_2^A , causing a substantial omitted variable bias in the estimates of β_{31} and additional dispersion of these estimates across all specifications. We do not report coverage rates for β_{31} as they are consistently 0%.

Our inspection of alternative estimators starts from Distribution 1, Model A, and homogeneous populations. Compared with OLS*, PILS-SDR* generates a visible finite-sample bias and loses considerable efficiency. These are the price to pay for switching from a one-sample to two-sample estimation. In contrast, the cost of estimating the index coefficient is marginal at best, as RMSEs of PILS-SDR are often smaller than those of PILS-SDR*.

Table 1: Monte Carlo Results

Panel A: Distribution 1 (elliptical) - Model A: $g_{21}(z) = z$

Estimator	Mean	SD	RMSE	MedSE	CR
Homogeneous Populations: $(\rho_N^A, \rho_D^A) = (\rho_N^B, \rho_D^B) = (0.40, 0.40)$					
For β_{21} :					
OLS*	1.0004	0.0152	0.0152	0.0151	95%
OLS-S	—	—	—	—	—
PILS-SDR*	0.9415	0.1549	0.1656	0.0476	90%
PILS-SDR:					
(SIR, SIR)	0.9320	0.1708	0.1838	0.0640	90%
(SIR, Probit)	0.9337	0.1703	0.1827	0.0646	91%
(SIR, Logit)	0.9340	0.1702	0.1825	0.0703	93%
(PLS, SIR)	0.9405	0.1530	0.1641	0.0578	91%
(PLS, Probit)	0.9422	0.1524	0.1630	0.0586	92%
(PLS, Logit)	0.9426	0.1524	0.1628	0.0651	94%
(PIR, SIR)	0.9316	0.1716	0.1847	0.0642	90%
(PIR, Probit)	0.9333	0.1711	0.1836	0.0648	91%
(PIR, Logit)	0.9336	0.1710	0.1834	0.0705	93%
For β_{31} :					
OLS*	0.9979	0.0356	0.0357	0.0359	94%
OLS-S	2.0559	0.0709	1.0583	0.0687	—
PILS-SDR*	1.0163	0.0697	0.0716	0.0548	92%
PILS-SDR:					
(SIR, SIR)	1.0207	0.0889	0.0912	0.0820	94%
(SIR, Probit)	1.0213	0.0889	0.0914	0.0830	95%
(SIR, Logit)	1.0213	0.0891	0.0916	0.0901	97%
(PLS, SIR)	1.0182	0.0840	0.0859	0.0719	94%
(PLS, Probit)	1.0189	0.0845	0.0866	0.0730	95%
(PLS, Logit)	1.0189	0.0847	0.0868	0.0803	96%
(PIR, SIR)	1.0209	0.0889	0.0913	0.0823	94%
(PIR, Probit)	1.0215	0.0889	0.0915	0.0835	95%
(PIR, Logit)	1.0215	0.0891	0.0917	0.0905	97%
Heterogeneous Populations: $(\rho_N^A, \rho_D^A) = (0.40, 0.40), (\rho_N^B, \rho_D^B) = (-0.40, -0.40)$					
For β_{21} :					
OLS*	1.0004	0.0152	0.0152	0.0151	95%
OLS-S	—	—	—	—	—
PILS-SDR*	0.9440	0.1505	0.1605	0.0472	90%
PILS-SDR:					
(SIR, SIR)	0.9398	0.1500	0.1617	0.0712	92%
(SIR, Probit)	0.9413	0.1496	0.1607	0.0719	92%
(SIR, Logit)	0.9416	0.1496	0.1605	0.0801	94%
(PLS, SIR)	0.9398	0.1517	0.1632	0.0620	89%
(PLS, Probit)	0.9414	0.1514	0.1623	0.0631	90%
(PLS, Logit)	0.9417	0.1514	0.1622	0.0716	93%
(PIR, SIR)	0.9394	0.1509	0.1626	0.0714	92%
(PIR, Probit)	0.9409	0.1504	0.1616	0.0720	92%
(PIR, Logit)	0.9412	0.1504	0.1615	0.0804	95%
For β_{31} :					
OLS*	0.9979	0.0356	0.0357	0.0359	94%
OLS-S	2.0559	0.0709	1.0583	0.0687	—
PILS-SDR*	1.0147	0.0664	0.0680	0.0549	92%
PILS-SDR:					
(SIR, SIR)	1.0180	0.0986	0.1002	0.1032	96%
(SIR, Probit)	1.0183	0.0977	0.0994	0.1056	96%
(SIR, Logit)	1.0185	0.0977	0.0994	0.1194	99%
(PLS, SIR)	1.0187	0.0973	0.0991	0.0871	93%
(PLS, Probit)	1.0190	0.0964	0.0983	0.0897	94%
(PLS, Logit)	1.0192	0.0965	0.0984	0.1026	97%
(PIR, SIR)	1.0182	0.0988	0.1005	0.1034	96%
(PIR, Probit)	1.0185	0.0978	0.0996	0.1058	95%
(PIR, Logit)	1.0187	0.0978	0.0996	0.1197	99%

Table 1: (continued)

Panel F: Distribution 2 (non-elliptical) - Model C: $g_{21}(z) = [2\mathbb{I}\{z > 0\} + 1] \sqrt{|z|}$

Estimator	Mean	SD	RMSE	MedSE	CR
Homogeneous Populations: (ρ_U^A, ρ_D^A) = (ρ_U^B, ρ_D^B) = (0.35, 0.40)					
For β_{21} :					
OLS*	0.9994	0.0137	0.0138	0.0138	95%
OLS-S	—	—	—	—	—
PILS-SDR*	0.9742	0.1356	0.1381	0.0418	92%
PILS-SDR:					
(SIR, SIR)	0.9742	0.1280	0.1306	0.0487	94%
(SIR, Probit)	0.9738	0.1277	0.1304	0.0490	93%
(SIR, Logit)	0.9739	0.1279	0.1305	0.0503	94%
(PLS, SIR)	0.9710	0.1353	0.1384	0.0490	93%
(PLS, Probit)	0.9704	0.1351	0.1383	0.0493	93%
(PLS, Logit)	0.9704	0.1351	0.1383	0.0508	94%
(PIR, SIR)	0.9742	0.1272	0.1298	0.0504	94%
(PIR, Probit)	0.9738	0.1269	0.1296	0.0508	94%
(PIR, Logit)	0.9738	0.1270	0.1297	0.0518	95%
For β_{31} :					
OLS*	1.0004	0.0318	0.0318	0.0321	96%
OLS-S	1.7696	0.0646	0.7723	0.0659	—
PILS-SDR*	1.0056	0.0526	0.0529	0.0494	94%
PILS-SDR:					
(SIR, SIR)	1.0032	0.0666	0.0666	0.0744	97%
(SIR, Probit)	1.0044	0.0659	0.0660	0.0758	98%
(SIR, Logit)	1.0045	0.0659	0.0661	0.0810	99%
(PLS, SIR)	1.0014	0.0709	0.0709	0.0741	97%
(PLS, Probit)	1.0024	0.0701	0.0701	0.0757	97%
(PLS, Logit)	1.0026	0.0702	0.0702	0.0814	98%
(PIR, SIR)	1.0034	0.0667	0.0668	0.0790	98%
(PIR, Probit)	1.0045	0.0658	0.0659	0.0806	98%
(PIR, Logit)	1.0047	0.0659	0.0661	0.0861	99%
Heterogeneous Populations: (ρ_U^A, ρ_D^A) = (0.35, 0.40), (ρ_U^B, ρ_D^B) = (-0.35, -0.40)					
For β_{21} :					
OLS*	0.9994	0.0137	0.0138	0.0138	95%
OLS-S	—	—	—	—	—
PILS-SDR*	0.9875	0.1047	0.1055	0.0417	94%
PILS-SDR:					
(SIR, SIR)	0.9697	0.1252	0.1288	0.0500	95%
(SIR, Probit)	0.9693	0.1249	0.1287	0.0504	95%
(SIR, Logit)	0.9694	0.1249	0.1286	0.0521	96%
(PLS, SIR)	0.9706	0.1230	0.1265	0.0501	93%
(PLS, Probit)	0.9701	0.1227	0.1263	0.0504	94%
(PLS, Logit)	0.9702	0.1227	0.1263	0.0520	95%
(PIR, SIR)	0.9698	0.1223	0.1260	0.0516	95%
(PIR, Probit)	0.9695	0.1220	0.1258	0.0518	95%
(PIR, Logit)	0.9696	0.1220	0.1258	0.0535	95%
For β_{31} :					
OLS*	1.0004	0.0318	0.0318	0.0321	96%
OLS-S	1.7696	0.0646	0.7723	0.0659	—
PILS-SDR*	1.0038	0.0502	0.0503	0.0494	95%
PILS-SDR:					
(SIR, SIR)	1.0056	0.0696	0.0699	0.0789	98%
(SIR, Probit)	1.0065	0.0690	0.0693	0.0807	98%
(SIR, Logit)	1.0067	0.0691	0.0695	0.0872	99%
(PLS, SIR)	1.0036	0.0734	0.0735	0.0776	97%
(PLS, Probit)	1.0044	0.0727	0.0728	0.0794	97%
(PLS, Logit)	1.0046	0.0727	0.0728	0.0861	98%
(PIR, SIR)	1.0058	0.0694	0.0697	0.0835	98%
(PIR, Probit)	1.0067	0.0690	0.0693	0.0848	98%
(PIR, Logit)	1.0069	0.0690	0.0694	0.0918	99%

Notes: *Mean* = simulation average of the parameter estimate; *SD* = simulation standard deviation of the parameter estimate; *RMSE* = root mean-squared error of the parameter estimate; *MedSE* = simulation median of the standard error; and *CR* = coverage rate against the nominal 95% confidence interval.

A closer look at the PILS-SDR results reveals a tendency to underestimate β_{21} – the coefficient on the missing regressor – and to overestimate β_{31} – the coefficient on the overlapping regressor. While Probit is the correct model specification for the missing binary regressor, misspecifying it as Logit does not distort the quality of the PILS-SDR estimate for β_{21} or β_{31} . All three estimators (SIR, Probit, or Logit) perform well, as expected given the results of Ruud (1983).

While there are no substantial differences between standard deviations of PILS-SDR* and PILS-SDR, median standard errors of the latter are uniformly greater than those of the former. This reflects the additional variability introduced in PILS-SDR through the index coefficient estimation. We also observe that PILS-SDR median standard errors for β_{21} are considerably smaller than standard deviations, whereas the two measures are roughly equal for β_{31} . The former observation explains the tendency toward undercoverage of the nominal 95% confidence intervals, while the latter results in coverage closer to the nominal level.

These findings remain largely unchanged under elliptical symmetry, even when the two samples are drawn from heterogeneous populations and/or when the link function loses monotonicity (Model B) or smoothness (Model C) - see Supplement F of the Supplementary Materials. Even when the distribution of the continuous component X_{3C}^s is replaced with a non-elliptical one (Distribution 2), the qualitative conclusions obtained under Distribution 1 continue to hold.

Finally, it is difficult to determine which combination of SDR methods performs best. The results are mixed in terms of $RMSE$, and no uniformly superior method emerges. Overall, all nine combinations perform similarly well.

5 Empirical Application

In this section, we explore the empirical relevance of PILS-SDR within the classical framework of the human capital earnings equation, where education is endogenous and correlated with potentially unobservable skills and ability. We follow HMP23 and consider the following wage regression:

$$\begin{aligned} \log(\textit{income}) = & \beta_0 + \beta_1\textit{educ} + \beta_2\textit{exper} + \beta_3\textit{exper}^2 + \beta_4\textit{abil} \\ & + \beta_5\textit{married} + \beta_6\textit{black} + \beta_7\textit{south} + \beta_8\textit{urban} + u, \end{aligned} \quad (11)$$

where $\log(\textit{income})$ is the natural logarithm of the total annual labor income; $\textit{educ}$ is years of schooling; $\textit{exper}$ is actual work experience; $\textit{abil}$ is an ability measure; $\textit{married}$ is an indicator for being married; $\textit{black}$ is an indicator for being black; $\textit{south}$ is an indicator for currently living in the Southern region of the United States; $\textit{urban}$ is an indicator for living in urban areas while the respondent was growing up; and u is the error term.

Our goal is to compare PILS-SDR with PILS and other estimators reported in HMP23. To estimate the regression (11), we use the same two US micro datasets as HMP23, except the respondent's age is limited between 18 and 40 years old. The point is to compare one-sample and two-sample approaches for similar age groups. An ability measure is typically unavailable in the primary sample, and we need an auxiliary sample that contains an ability measure and/or is potentially more reliable within (roughly) the same age group.

As the primary sample, we employ the male sub-sample of the 1972 wave of PSID. This wave of PSID is particularly useful because, unlike most other waves, it records IQ scores ($\textit{IQ}$) measured by the sentence completion test. Consequently, we may estimate (11) by OLS using $\textit{IQ}$ as a proxy for unobserved skills and ability. We are interested in this scenario as a benchmark. At the same time, we wish to examine how the estimation results change if $\textit{IQ}$ is missing. Accordingly, we view the OLS results as infeasible and label them as

OLS*, similar to the previous section.

When the primary sample does not contain IQ , researchers may estimate the short regression with IQ dropped from (11). This is likely to result in an upward inconsistent estimation as the respondent’s years of education are positively correlated with their skills and ability. In our example, the sample correlation between $educ$ and IQ is 0.4496.

Alternatively, the researcher may seek valid instruments for $educ$ and proceed to IV or GMM estimation. Typical instruments for $educ$ are $fatheduc$ (years of schooling for the respondent’s father) and $motheduc$ (years of schooling for the respondent’s mother). The former is available in the same 1972 wave, whereas the latter can be taken from the 1974 wave of PSID. These instruments are in fact strong, with sample correlations between $educ$ and $fatheduc$ and between $educ$ and $motheduc$ being 0.4313 and 0.4162, respectively. In IV estimation, we use $fatheduc$ as the instrument for $educ$. In two-step GMM, both $fatheduc$ and $motheduc$ are used as instruments for $educ$, and 2SLS serves as the first-step estimate. The labels we use for OLS, IV and GMM estimators of the short regression (without IQ) based on only the primary sample are OLS-S, IV-S and GMM-S, respectively.

Our primary interest is in the use of an auxiliary sample for PILS or PILS-SDR. Following HMP23, we adopt the dataset used in Card (1995) as the auxiliary sample. It contains observations from the 1976 interview of the NLS Young Men (NLSYM) Cohort. NLSYM maintains scores of the “Knowledge of the World of Work” (KWW) test that respondents took in the 1966 interview, and we employ this variable as an alternative ability measure to the 1972 IQ test. The KWW score is commonly argued to be a better measure of unobserved ability. All other regressors except $exper$ are recorded in both the 1966 and 1976 interview.³

To show the benefits of using PILS-SDR as compared to PILS, we extract from PSID and NLSYM two additional variables, age (respondent’s age) and $south_in_youth$ (an indicator

³The actual work experience is unavailable in NLSYM. Because of this, Card (1995) suggests computing $age - educ - 6$ as a proxy for $exper$.

for living in the Southern geographical region while growing up). Together with *fatheduc* and *motheduc*, these variables form the four excluded overlapping variables between the two samples. The use of these additional variables makes our samples somewhat different from those of HMP23.

As a result, we obtain the primary dataset $\mathcal{S}^A$ with ($n =$) 1303 observations and the auxiliary dataset $\mathcal{S}^B$ with ($m =$) 1516 observations. The variables in $\mathcal{S}^A$ are $Y^A = \log(\text{income})$, $X_1^A = (\text{exper}, \text{exper}^2)$, $X_2^A = IQ$, $X_{3I}^A = (\text{educ}, \text{married}, \text{black}, \text{south}, \text{urban})$, and $X_{3E}^A = (\text{fatheduc}, \text{motheduc}, \text{age}, \text{south_in_youth})$. The variables in $\mathcal{S}^B$ are $X_2^B = KWW$, $X_{3I}^B = (\text{educ}, \text{married}, \text{black}, \text{south}, \text{urban})$ and $X_{3E}^B = (\text{fatheduc}, \text{motheduc}, \text{age}, \text{south_in_youth})$. The summary statistics of all variables are presented in Table 2. The data characteristics are close to those used by HMP23, except for *exper* and *age* in $\mathcal{S}^A$.

We treat *IQ* as the missing continuous regressor. Among the nine overlapping variables, *educ*, *fatheduc*, *motheduc*, and *age* are treated as continuous. All overlapping variables are demeaned within the primary and auxiliary samples. The two ability measures, *IQ* and *KWW*, are also demeaned before the analysis.

To compare PILS and PILS-SDR, we consider two scenarios that differ in the overlapping variables they use. The first uses the variables *educ*, *age*, *married*, *black*, *south*, *urban*, *south_in_youth*, in this order. Because two of these variables are continuous ($d_{3C} = 2$), both PILS and PILS-SDR are $\sqrt{n}$ -consistent. The second scenario uses *educ*, *fatheduc*, *motheduc*, *age*, *married*, *black*, *south*, *urban*, *south_in_youth*, in this order. Here, $d_{3C} = 4$ and it follows that PILS is no longer $\sqrt{n}$ -consistent due to the curse of dimensionality, while PILS-SDR retains $\sqrt{n}$ -consistency.

To implement PILS, we employ the Epanechnikov kernel for continuous overlapping variables – with bandwidth $\hat{h} = \hat{\sigma}_B (\log m/m)^c$ and sample standard deviation $\hat{\sigma}_B$ – and the discrete kernel from HMP23 for binary overlapping variables – with bandwidth $\hat{\lambda} = (\log m/m)^{2c}$. The constant c is set at 0.3 for $d_{3C} = 2$ and 0.225 for $d_{3C} = 4$.

Table 2: Sample Characteristics

Variable	<i>Mean</i>	<i>SD</i>	<i>Min</i>	<i>Med</i>	<i>Max</i>
Primary Sample $\mathcal{S}^A$ (PSID; $n = 1303$)					
$\log(\text{income})$	8.81	0.70	3.91	8.92	10.82
<i>educ</i>	12.85	2.66	5	12	17
<i>exper</i>	9.77	5.98	0	9	36
<i>IQ</i>	9.63	2.15	0	10	13
<i>married</i>	0.88	0.33	0	1	1
<i>black</i>	0.27	0.45	0	0	1
<i>south</i>	0.42	0.49	0	0	1
<i>urban</i>	0.31	0.46	0	0	1
<i>fatheduc</i>	9.92	3.23	0	8	17
<i>motheduc</i>	10.62	2.86	0	11	17
<i>age</i>	28.67	6.00	18	28	40
<i>south_in_youth</i>	0.46	0.50	0	0	1
Auxiliary Sample $\mathcal{S}^B$ (NLSYM; $m = 1516$)					
<i>KWW</i>	36.02	7.37	13	36.5	56
<i>educ</i>	14.15	2.26	8	14	18
<i>married</i>	0.74	0.44	0	1	1
<i>black</i>	0.10	0.30	0	0	1
<i>south</i>	0.33	0.47	0	0	1
<i>urban</i>	0.69	0.46	0	1	1
<i>fatheduc</i>	10.61	3.37	0	12	18
<i>motheduc</i>	11.06	2.83	0	12	18
<i>age</i>	28.25	2.95	24	28	34
<i>south_in_youth</i>	0.32	0.47	0	0	1

Table 3 presents the estimation results for OLS*, OLS-S, IV-S, GMM-S, PILS, and the three versions of PILS-SDR, with standard errors reported in parentheses. The PILS-SDR and PILS standard errors come from the covariance matrix estimation described in Supplements C and D of the Supplementary Materials and heteroskedasticity-robust standard errors are used for all other estimators. No standard errors are reported for the case $d_{3C} = 4$ because PILS is not $\sqrt{n}$ -consistent in this case.

The results of OLS*, OLS-S, IV-S, GMM-S, and PILS for $d_{3C} = 2$ are similar to those reported in HMP23. For the benchmark OLS*, the signs of the coefficient estimates on

$educ$, $exper$, $exper^2$ and $abil (= IQ)$ are as expected, and their values are all significant at the 1% level. Columns two to four report the results for OLS-S, IV-S and GMM-S. While the signs of the coefficient estimates on $educ$, $exper$ and $exper^2$ are again as expected, the OLS-S results should be interpreted with caution, as they are likely to be inconsistent. Moreover, the value of the J -test statistic for over-identifying restrictions in GMM-S is 4.7 (with one degree of freedom), suggesting evidence of the model misspecification.

The PILS-SDR results for $d_{3C} = 2$ using both S^A and S^B are reported in columns (6)-(8). The signs of the coefficient estimates on $educ$, $exper$, $exper^2$, and $abil (= KWW)$ stay the same as for OLS*. Standard errors of PILS-SDR are inflated compared to OLS*. Nonetheless, quite a few PILS-SDR coefficient estimates are significantly positive at the 5% level although their magnitudes vary across specifications. It is also evident that the PILS-SDR estimates of β_1 , the return to schooling, are usually higher than the benchmark, whereas the estimates of β_4 , the coefficient on $abil$, are lower. Recall that $educ$ is observed in both S^A and S^B , while $abil$ is missing in S^A . Thus, the empirical results confirm the pattern observed in the Monte Carlo study: PILS-SDR tends to underestimate the coefficient on the missing regressor and overestimate the coefficient on the overlapping variable.

Finally, it is informative to compare the return to schooling estimates across all the estimation methods. OLS* yields a small estimated magnitude, while estimating (11) without $abil$ yields a substantially larger estimated return to schooling. Consistent estimation using IV and GMM further inflates the estimated return to schooling relative to the inconsistent OLS-S, which is already expected to be upward biased. Columns (9)-(12) permit comparison between the two scenarios. Expanding the set of overlapping variables increases the PILS estimate of the return to schooling (although it is no longer $\sqrt{n}$ -consistent in this scenario). In contrast, the three versions of PILS-SDR look fairly stable and closer to OLS*. While we cannot tell which estimate is closer to the true value, the results indicate that PILS-SDR is a viable alternative to existing two-sample and IV-based methods.

Table 3: Estimation Results

Regressor	(1) OLS*	(2) OLS-S	(3) IV-S	(4) GMM-S	(5)		(6)		(7)		(8)		(9)		(10)		(11)		(12)	
					PILS	SIR	PILS	SIR	PILS	PILS	PILS	PILS	PILS	PILS	SIR	SIR	PILS	PILS	PILS	PILS
<i>educ</i>	0.0412 (0.0070)	0.0461 (0.0065)	0.0722 (0.0219)	0.0983 (0.0186)	0.0400 (0.0073)	0.0426 (0.0086)	0.0299 (0.0098)	0.0458 (0.0065)	0.0449 (-)	0.0413 (0.0089)	0.0458 (0.0065)	0.0458 (0.0065)	0.0449 (-)	0.0413 (0.0089)	0.0413 (0.0089)	0.0413 (0.0089)	0.0389 (0.0078)	0.0389 (0.0078)	0.0436 (0.0068)	0.0436 (0.0068)
<i>exper</i>	0.1067 (0.0115)	0.1072 (0.0114)	0.1047 (0.0112)	0.1014 (0.0110)	0.1049 (0.0117)	0.1033 (0.0145)	0.0902 (0.0164)	0.1064 (0.0114)	0.1063 (-)	0.1015 (0.0162)	0.1064 (0.0114)	0.1064 (0.0114)	0.1063 (-)	0.1015 (0.0162)	0.1015 (0.0162)	0.1015 (0.0162)	0.1000 (0.0132)	0.1000 (0.0132)	0.1043 (0.0121)	0.1043 (0.0121)
<i>exper</i> ²	-0.0028 (0.0005)	-0.0029 (0.0005)	-0.0027 (0.0005)	-0.0025 (0.0005)	-0.0028 (0.0005)	-0.0028 (0.0005)	-0.0027 (0.0006)	-0.0028 (0.0005)	-0.0028 (-)	-0.0028 (0.0006)	-0.0028 (0.0005)	-0.0028 (0.0005)	-0.0028 (-)	-0.0028 (0.0006)	-0.0028 (0.0006)	-0.0028 (0.0006)	-0.0028 (0.0005)	-0.0028 (0.0005)	-0.0028 (0.0005)	-0.0028 (0.0005)
<i>abil</i>	0.0180 (0.0088)	- (-)	- (-)	- (-)	0.0097 (0.0053)	0.0035 (0.0064)	0.0162 (0.0083)	0.0007 (0.0001)	0.0045 (-)	0.0049 (0.0072)	0.0007 (0.0001)	0.0007 (0.0001)	0.0045 (-)	0.0049 (0.0072)	0.0049 (0.0072)	0.0049 (0.0072)	0.0062 (0.0046)	0.0062 (0.0046)	0.0022 (0.0018)	0.0022 (0.0018)
<i>married</i>	0.3776 (0.0660)	0.3816 (0.0657)	0.3970 (0.0679)	0.4162 (0.0680)	0.3834 (0.0660)	0.3819 (0.0660)	0.3795 (0.0664)	0.3815 (0.0658)	0.3821 (-)	0.3782 (0.0679)	0.3815 (0.0658)	0.3815 (0.0658)	0.3821 (-)	0.3782 (0.0679)	0.3782 (0.0679)	0.3782 (0.0679)	0.3770 (0.0667)	0.3770 (0.0667)	0.3814 (0.0660)	0.3814 (0.0660)
<i>black</i>	-0.1473 (0.0390)	-0.1702 (0.0380)	-0.1254 (0.0513)	-0.0829 (0.0480)	-0.1720 (0.0383)	-0.1538 (0.0515)	-0.0902 (0.0581)	-0.1681 (0.0381)	-0.1690 (-)	-0.1477 (0.0545)	-0.1681 (0.0381)	-0.1681 (0.0381)	-0.1690 (-)	-0.1477 (0.0545)	-0.1477 (0.0545)	-0.1477 (0.0545)	-0.1529 (0.0407)	-0.1529 (0.0407)	-0.1654 (0.0394)	-0.1654 (0.0394)
<i>south</i>	-0.0988 (0.0342)	-0.1052 (0.0341)	-0.0882 (0.0370)	-0.0707 (0.0366)	-0.1055 (0.0343)	-0.1052 (0.0344)	-0.1037 (0.0342)	-0.1034 (0.0341)	-0.1055 (-)	-0.1040 (0.0366)	-0.1037 (0.0342)	-0.1037 (0.0342)	-0.1055 (-)	-0.1040 (0.0366)	-0.1040 (0.0366)	-0.1040 (0.0366)	-0.0957 (0.0339)	-0.0957 (0.0339)	-0.1012 (0.0345)	-0.1012 (0.0345)
<i>urban</i>	0.1409 (0.0365)	0.1472 (0.0368)	0.1275 (0.0417)	0.1072 (0.0413)	0.1487 (0.0368)	0.1436 (0.0376)	0.1318 (0.0376)	0.1457 (0.0368)	0.1467 (-)	0.1434 (0.0401)	0.1457 (0.0368)	0.1457 (0.0368)	0.1467 (-)	0.1434 (0.0401)	0.1434 (0.0401)	0.1434 (0.0401)	0.1414 (0.0371)	0.1414 (0.0371)	0.1452 (0.0370)	0.1452 (0.0370)
Data combination?	NO	NO	NO	NO	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
<i>abil</i>	<i>IQ</i>	-	-	-	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>	<i>KWW</i>
Sample size:	<i>n</i>	1303	1303	1303	1303	1303	1303	1303	1303	1303	1303	1303	1303	1303	1303	1303	1303	1303	1303	1303
	<i>m</i>	-	-	-	1516	1516	1516	1516	1516	1516	1516	1516	1516	1516	1516	1516	1516	1516	1516	1516
<i>d_{3C}</i>	-	-	-	-	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Notes: The dependent variable is $\log(\text{income})$. Numbers in parentheses are standard errors. For IV-S, *fatheduc* is an instrument for *educ*. For GMM-S, *fatheduc* and *motheduc* are instruments for *educ*, and 2SLS is used as the first step estimate. The included overlapping variables for (5)-(12) are *educ*, *married*, *black*, *south*, and *urban*. The excluded overlapping variables for (5)-(8) are *age* and *south.in.youth*, whereas those for (9)-(12) are *fatheduc*, *motheduc*, *age*, and *south.in.youth*.

6 Conclusion

To mitigate the curse of dimensionality that arises when proxies of missing regressors are generated nonparametrically, we improve on the PILS estimator of HMP23 through incorporating sufficient dimension reduction. We derive asymptotic properties of the resulting PILS-SDR estimator and illustrate how to implement it. PILS-SDR can accommodate both continuous and binary missing regressors and remains useful when the two samples are drawn from potentially different populations. Most importantly, it attains the parametric rate of convergence even when there are more than three continuous overlapping variables used in imputing the missing regressors. Desirable finite-sample properties and empirical relevance are confirmed through a Monte Carlo study and a wage regression application.

Conflict of Interest Statement

The authors report there are no competing interests to declare.

References

- Ahn, H. (1997). Semiparametric estimation of a single-index model with nonparametrically generated regressors. *Econometric Theory*, 13:3–31.
- Bickel, P. J., Kur, G., and Nadler, B. (2018). Projection pursuit in high dimensions. *Proceedings of the National Academy of Sciences*, 115:9151–9156.
- Black, B., Trainor, M., and Spencer, J. E. (1999). Wage protection systems, segregation and gender pay inequalities: West Germany, the Netherlands and Great Britain. *Cambridge Journal of Economics*, 23:449–464.

- Bollinger, C. and Hirsch, B. T. (2006). Match bias from earnings imputation in the current population survey: The case of imperfect matching. *Journal of Labor Economics*, 24:483–519.
- Bostic, R., Gabriel, S., and Painter, G. (2009). Housing wealth, financial wealth, and consumption: New evidence from micro data. *Regional Science and Urban Economics*, 39:79–89.
- Buchinsky, M., Li, F., and Liao, Z. (2022). Estimation and inference of semiparametric models using data from several sources. *Journal of Econometrics*, 226:80–103.
- Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. L. N. Christophides, E. K. Grant, and R. Swidinsky (eds.), *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*. Toronto: University of Toronto Press, 201 - 222.
- Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric efficiency in gmm models with auxiliary data. *Annals of Statistics*, 36:808–843.
- Cook, R. D., Helland, I. S., and Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society, Series B*, 75:851–877.
- Cook, R. D. and Lee, H. (1999). Dimension reduction in binary response regression. *Journal of the American Statistical Association*, 94:1187–1200.
- Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics*, 30:455–474.
- D’Haultfoeuille, X., Gaillac, C., and Maurel, A. (2025). Partially linear models under data combinations. *Review of Economic Studies*, 92:238–267.

- Dong, Y. (2021). A brief review of linear sufficient dimension reduction through optimization. *Journal of Statistical Planning and Inference*, 211:154–161.
- Duan, N. and Li, K.-C. (1991). Slicing regression: A link-free regression method. *Annals of Statistics*, 19:505–530.
- Freyberger, J., Hoepfner, B., Neuhierl, A., and Weber, M. (2025). Missing data in asset pricing panels. *Review of Financial Studies*, 38:760–802.
- Ghosh, D. (2011). Propensity score modelling in observational studies using dimension reduction methods. *Statistics & Probability Letters*, 81:813–820.
- Graham, B. S., de Xavier Pinto, C. C., and Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business & Economic Statistics*, 34:288–301.
- Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24:726–748.
- Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84:986–995.
- He, X. and Shao, Q.-M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73:120–135.
- Hellerstein, J. K. and Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *Review of Economics and Statistics*, 81:1–14.
- Hirukawa, M., Murtazashvili, I., and Prokhorov, A. (2023). Yet another look at the omitted variable bias. *Econometric Reviews*, 42:1–27.

- Hirukawa, M. and Prokhorov, A. (2018). Consistent estimation of linear regression models using matched data. *Journal of Econometrics*, 203:344–358.
- Hristache, M., Juditsky, A., and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, 29:595–623.
- Hwang, Y. (2026). Bounding omitted variable bias using auxiliary data: With an application to estimate neighborhood effects. *Journal of Business & Economic Statistics*, 44:1167–1180.
- Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58:71–120.
- Inoue, A. and Solon, G. (2010). Two-sample instrumental variables estimators. *Review of Economics and Statistics*, 92:557–561.
- Kamakura, W. A. and Wedel, M. (1997). Statistical data fusion for cross-tabulation. *Journal of Marketing Research*, 34:485–498.
- Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, 61:387–421.
- Krylov, A. N. (1931). On the numerical solution of the equation by which in technical questions frequencies of small oscillations of material systems are determined. *Izvestiya Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk*, 7:491–539.
- Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:316–327.
- Li, L., Cook, R. D., and Tsai, C.-L. (2007). Partial inverse regression. *Biometrika*, 94:615–625.

- Luo, W. and Zhu, Y. (2020). Matching using sufficient dimension reduction for causal inference. *Journal of Business & Economic Statistics*, 38:888–900.
- Murtazashvili, I., Liu, D., and Prokhorov, A. (2015). Two-sample nonparametric estimation of intergenerational income mobility in the United States and Sweden. *Canadian Journal of Economics*, 48:1733–1761.
- Naik, P. and Tsai, C.-L. (2000). Partial least squares estimator for single-index models. *Journal of the Royal Statistical Society, Series B*, 62:763–771.
- Newey, W. K., Hsieh, F., and Robins, J. M. (2004). Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, 72:947–962.
- Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57:1403–1430.
- Ridder, G. and Moffitt, R. (2007). The econometrics of data combination. J. J. Heckman and E. E. Leamer (eds.), *Handbook of Econometrics, Vol. 6, Part B*. Amsterdam: Elsevier, Chapter 75, 5469 - 5547.
- Ruud, P. A. (1983). Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica*, 51:225–228.
- Saracco, J. (1997). An asymptotic theory for sliced inverse regression. *Communications in Statistics - Theory and Methods*, 26:2141–2171.
- Soale, A.-N. (2023). Projection expectile regression for sufficient dimension reduction. *Computational Statistics & Data Analysis*, 180:Article No.107666.

A Appendix

A.1 Notational Convention

We say ‘ $a_n \asymp b_n$ ’ if there exist constants $0 < c_1 < c_2 < \infty$ so that $c_1 a_n \leq b_n \leq c_2 a_n$; ‘ $\perp\!\!\!\perp$ ’ means stochastic independence; $\|A\|$ is the Frobenius norm of matrix A , i.e., $\|A\| = \{\text{tr}(A^\top A)\}^{1/2}$; $\mathbb{I}\{\cdot\}$ signifies an indicator function; $h^{(p)}(x) = d^p h(x)/dx^p$ is the p th-order derivative of a function $h(x)$; $0_{p \times q}$ denotes the $p \times q$ zero matrix; and the symbol ‘ $>$ ’ applied to matrices means positive definiteness.

A.2 Assumptions

G1. Two random samples $\mathcal{S}^A = \mathcal{S}_n^A = \{(Y_i^A, X_{1i}^A, X_{3i}^A)\}_{i=1}^n$ and $\mathcal{S}^B = \mathcal{S}_m^B = \{(X_{2j}^B, X_{3j}^B)\}_{j=1}^m$ are drawn independently from possibly non-identical populations of the random vector (Y, X_1, X_2, X_3) with finite fourth-order moments.

G2. $E(u^A | X_1^A, X_3^A) = 0$ and $\sigma_{u^A}^2(X_1^A, X_3^A) := E\{(u^A)^2 | X_1^A, X_3^A\} \in (0, \infty)$ hold.

G3. Let $\eta_2^s = (\eta_{21}^s, \dots, \eta_{2d_2}^s)^\top := X_2^s - E(X_2^s | X_3^s)$ be the error term in the reduced form of X_2^s for $s \in \{A, B\}$. Then, η_2^A is conditionally mean independent of X_1^A given X_3^A .

S1. Overlapping variables X_3^s , for $s \in \{A, B\}$, satisfy the following conditions: (i) The support of X_{3C}^A and X_{3C}^B is identical, and so is the support of X_{3D}^A and X_{3D}^B ; (ii) The common support $\mathbb{X}_{3C} := \text{supp}(X_{3C}^s)$ is a convex and compact subset of $\mathbb{R}^{d_{3C}}$, and the joint densities of X_{3C}^s are bounded and bounded away from zero on $\mathbb{X}_{3C}$ for $s \in \{A, B\}$; (iii) Both X_3^A and X_3^B are centered.

S2. For each $\ell \in \{1, \dots, d_2\}$ and $s \in \{A, B\}$, there exists a $d_3 \times 1$ vector θ_ℓ^s that defines the linear index $Z_\ell^s = \theta_\ell^{s\top} X_3^s$ and satisfies $X_{2\ell}^s \perp\!\!\!\perp E(X_{2\ell}^s | X_3^s) | Z_\ell^s$.

- S3.** The density of Z_ℓ^s , denoted by $p_\ell^s(\cdot)$, is bounded and bounded away from zero on $\mathbb{Z}_\ell^s := \text{supp}(Z_\ell^s) \subseteq \mathbb{R}$ for all $\ell \in \{1, \dots, d_2\}$ and $s \in \{A, B\}$.
- S4.** For each $\ell \in \{1, \dots, d_2\}$ and $s \in \{A, B\}$, the following conditions hold: (i) The first element of X_3^s is continuous, and the index coefficient on this element in θ_ℓ^s is unity; (ii) If X_{3D}^s is non-empty, then varying the values of X_{3D}^s does not divide the support of $\mathbb{Z}_\ell^s$ into disjoint subsets, and the link function $g_{2\ell}^s(\cdot)$ is not periodic.
- S5.** For some $\ell \in \{1, \dots, d_2\}$, the binary variable $X_{2\ell}^B$ admits the latent-index representation $X_{2\ell}^B := \mathbb{I}\{X_{2\ell}^{B*} > 0\}$, where $X_{2\ell}^{B*} = \theta_\ell^{B\top} X_3^B + v_\ell^B$ for some regression error v_ℓ^B . The conditional density of $v_\ell^B | X_3^B$ is (mis)specified as $f_\ell(v_\ell^B | X_3^B; \theta_\ell^B)$. If θ_ℓ^B is estimated by MLE over $\bar{\Theta}_\ell^B$, a convex and compact subset of $\mathbb{R}^{d_3}$, then the following conditions hold: (i) $\sup_{\vartheta \in \bar{\Theta}_\ell^B} \left| \sum_{j=1}^m \log f_{\ell,j}(\vartheta) / m - E\{\log f_{\ell,j}(\vartheta)\} \right| \xrightarrow{p} 0$, where $\log f_{\ell,j}(\vartheta) := \log f_\ell(v_{\ell,j}^B | X_{3j}^B; \vartheta)$, the log-likelihood contribution of the j -th observation; (ii) $E\{\log f_{\ell,j}(\vartheta)\}$ is uniquely maximized at $\vartheta = \theta_\ell^B$, an interior point of $\bar{\Theta}_\ell^B$; (iii) Second-order derivatives of $\log f_{\ell,j}(\vartheta)$ w.r.t. ϑ are continuous and bounded uniformly on $\bar{\Theta}_\ell^B$; (iv) $\sup_{\vartheta \in \bar{\Theta}_\ell^B} E\|\partial \log f_{\ell,j}(\vartheta) / \partial \vartheta\|^2, \sup_{\vartheta \in \bar{\Theta}_\ell^B} E\|\partial^2 \log f_{\ell,j}(\vartheta) / \partial \vartheta \partial \vartheta^\top\| < \infty$.
- S6.** $\Theta^A = \Theta^B$ and $g_2^A(\cdot) = g_2^B(\cdot)$.
- S7.** The SDR estimate of index coefficients $\hat{\Theta}$ admits an asymptotic linear representation

$$\sqrt{m} \left\{ \text{vec}(\hat{\Theta}^\top) - \text{vec}(\Theta^\top) \right\} = \frac{1}{\sqrt{m}} \sum_{j=1}^m \varphi_j^B(\Theta) + o_p(1)$$

as $m \rightarrow \infty$, where $\varphi_j^B(\Theta) := \varphi(X_{2j}^B, X_{3j}^B; \Theta)$ is the influence function that satisfies $E\{\varphi_j^B(\Theta)\} = 0_{d_2 d_3 \times 1}$, $E\|\varphi_j^B(\Theta)\|^2 < \infty$, and

$$\frac{1}{\sqrt{m}} \sum_{j=1}^m \varphi_j^B(\Theta) \xrightarrow{d} N(0_{d_2 d_3 \times 1}, V_{\varphi^B}) := N\left(0_{d_2 d_3 \times 1}, E\left\{\varphi_j^B(\Theta) \varphi_j^B(\Theta)^\top\right\}\right).$$

P1. Let $\theta_{\ell I}$ and $\theta_{\ell E}$ denote the coefficients on X_{3I}^s and X_{3E}^s , respectively. For each $\ell \in \{1, \dots, d_2\}$ and $s \in \{A, B\}$, $\theta_{\ell I}$ and $\theta_{\ell E}$ are included in θ_ℓ . Either condition holds:
(a) If X_{3E}^s is non-empty and $\theta_{\ell E} \neq 0$, then $g_{2\ell}(\cdot)$ is non-constant on $\mathbb{Z}_\ell$; (b) If X_{3E}^s is empty or if X_{3E}^s is non-empty but $\theta_{\ell E} = 0$, then $g_{2\ell}(\cdot)$ is strictly nonlinear on $\mathbb{Z}_\ell$.

P2. The kernel function $K : \mathbb{R} \rightarrow \mathbb{R}$ satisfies the following conditions: (i) $K(u)$ is symmetric, uniformly bounded, and differentiable; (ii) $K(u)$ is of order ν , i.e., $\int K(u) du = 1$, $\int u^p K(u) du = 0$ for $p \in \{1, \dots, \nu - 1\}$, and $\int u^\nu K(u) du \neq 0$; (iii) There exist constants $L, M \in (0, \infty)$ and $\rho \in (1, \infty)$ such that either (a) or (b) holds: (a) $K(u) = K^{(1)}(u) = 0$ for all $|u| > M$, $|K(u) - K(u^*)| \leq L|u - u^*|$ for all $u, u^* \in \mathbb{R}$, and $|K^{(1)}(u) - K^{(1)}(u^*)| \leq L|u - u^*|$ for all $|u|, |u^*| \leq M$; or (b) $|K^{(1)}(u)| \leq L$ for all $u \in \mathbb{R}$, $|K^{(1)}(u)| \leq L|u|^{-\rho}$ for all $|u| > M$, and $|K^{(1)}(u) - K^{(1)}(u^*)| \leq L|u - u^*|$ for all $u, u^* \in \mathbb{R}$.

P3. The bandwidths $h_1, \dots, h_{d_2}$ satisfy $h_1, \dots, h_{d_2} \asymp h > 0$ and $h + nh^{2\nu} + \log m / (mh^6) \rightarrow 0$ as $n, m \rightarrow \infty$ so that $n/m \rightarrow \kappa \in (0, \infty)$.

P4. For each $\ell \in \{1, \dots, d_2\}$, the ν th-order derivatives of $p_\ell^{(1)}(z)$, $r_\ell^{(1)}(z)$, $E(X_1^A | Z_\ell^A = z)$, and $E\{g_2(Z^A) | Z_\ell^A = z\}$ are Lipschitz continuous and bounded uniformly on $z \in \mathbb{Z}_\ell$. There exist constants $\gamma \in (0, \infty)$ and $C \in [1, \infty)$ such that $\sup_{z \in \mathbb{Z}_\ell} E\left(|X_{2\ell}^B|^{2+\gamma} | Z_\ell^B = z\right) \leq C$.