Sufficient Dimension Reduction Meets Two-Sample Regression Estimation

Masayuki Hirukawa*
Ryukoku University

Di Liu†
Stata Corp

Artem Prokhorov‡
University of Sydney,
CEBDA & CIREQ

June 8, 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 extend the applicability of PILS and MSII by modelling the conditional expectation of each missing regressor given the overlapping variables as a single-index with an unknown link function. We show how to obtain the index coefficients using sufficient dimension reduction (SDR), and we prove that the estimator that combines PILS with SDR is asymptotically normal with the parametric convergence rate, regardless of the dimensionality 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 PILS-SDR. 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.

*e-mail: hirukawa@econ.ryukoku.ac.jp.
†e-mail: flyingliudi@gmail.com.
‡e-mail: artem.prokhorov@sydney.edu.au.

1 Introduction

Suppose that using an anonymous, individual-level survey dataset, we wish to estimate the following linear regression:

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

where $X_1 \in \mathbb{R}^{d_1}$, $X_2 \in \mathbb{R}^{d_2}$, $X_{3I} \in \mathbb{R}^{d_{3I}}$, $X = \left(1, X_1^\top, X_2^\top, X_{3I}^\top\right)^\top \in \mathbb{R}^d$ 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 $\beta = \left(\beta_0, \beta_1^\top, \beta_2^\top, \beta_3^\top\right)^\top$ consistently by the ordinary least squares (OLS) under standard assumptions. In this paper, however, we deal with the situation in which X_2 is relevant ($\beta_2 \neq 0$) but missing in the dataset. The absence of a variable in a survey can occur when the variable is regarded as less important in light of the survey design or when collecting it or the hand-coding required to measure 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 make the problem 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 $X_3 = (X_{3I}, X_{3E})$. In other words, 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 we can find another dataset, called the *auxiliary sample*, which consists of (X_2, X_3) and can thus 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 and not used for data fusion.

The above setup follows Hirukawa and Prokhorov (2018) and Hirukawa et al. (2023), abbreviated hereinafter as HP18 and HMP23, respectively. Such data and model structures are encountered in many labor and health economics application (see, e.g., Murtazashvili et al., 2015; Bollinger and Hirsch, 2006). HP18 discuss how this setting arises in a wide range of studies of labor market dynamics and outcomes as a result of the Census practice know as "hot deck imputation", where missing records on earnings are imputed using the reported earnings of respondents with similar recorded attributes.

The objective of HP18 and HMP23 was to establish point identification and $\sqrt{n}$-consistent estimation of β in such settings, where n is the number of observations in the primary sample. As $\sqrt{n}$-consistent estimators of β, HP18 and HMP23 proposed the matched-sample indirect inference (MSII) and plug-in least squares (PILS) estimators, respectively. These estimators offer ways of imputing a proxy for the missing regressor through utilizing the information contained in the overlapping variables in the two samples.

However, MSII and PILS are subject to the curse of dimensionality as both construct the proxies nonparametrically. To describe the issue more precisely, let the set of overlapping variables X_3 admit yet another 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 $d_3 = d_{3C} + d_{3D}$ holds. To attain $\sqrt{n}$-consistency for MSII or PILS using X_3, HP18 and HMP23 require d_{3C} to be no greater than three. If $d_{3C} > 3$ and we still wish to use MSII or PILS, we either need to select at most three most relevant continuous components from all d_{3C} candidates or give up the claim of $\sqrt{n}$-consistency.

The purpose of this paper is to expand on HP18 and HMP23 when $d_{3C} > 3$. We focus on extending PILS rather than MSII, and we discuss the technical challenges associated with obtaining similar results for MSII in the Supplement C of the Supplementary Materials. In a nutshell, PILS 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 order of the kernel and the required 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 by d_3 rather than by the kernel order.

Instead of using the overlapping variables as they are, we propose to reduce the *effective* dimension using a dimension reduction technique. More specifically, we assume that the conditional expectation can be expressed as a semiparametric single-index model (SIM) with an unspecified link function. SIMs have been extensively studied in the econometric literature (see, e.g., Ichimura, 1993; Ahn, 1997). In fact, many commonly used generalized linear models arise as special cases of SIMs, including linear regression, Probit and Logit models.

The specific way we achieve dimension reduction is through sufficient dimension reduction (SDR). SDR provides a way to projects a high-dimensional set of covariates onto a lower-dimensional subspace spanned by a small number of linear indices

while preserving the the relevant information about the conditional expectation of the response given the covariates. The SDR literature originated in statistics and has since been widely applied in machine learning, data mining, signal processing, and bioinformatics. Different SDR methods vary in how they estimate the so-called dimension reduction directions. SDR methods are particularly natural in the SIM setting because index coefficients in SIMs correspond to these directions and can be uniquely identified through an appropriate normalization. Examples of studies in this area include Duan and Li (1991), Naik and Tsai (2000) and Li et al. (2007).

We call the proposed estimator PILS-SDR. The estimation proceeds in three steps. A proxy of $E(X_2|X_3)$ is constructed in the first two steps, and this is done equation-by-equation. In the first step, d_2 sets of index coefficients on X_3 are estimated using the auxiliary sample via well-established SDR methods, and estimated indices are obtained. An investigation of how the SDR algorithms perform in comparision with the method by Ichimura (1993) may be of independent interest. In the second step, d_2 unknown link functions are estimated using the indices from the first step and kernel regression smoothing. Although a higher-order kernel must still be employed, in practice we may adopt a fourth-order kernel regardless of d_{3C}. After obtaining a proxy of $E(X_2|X_3)$, the third and final step estimates (1) by OLS using the imputed values in place of X_2.

The paper's contribution to the literature is fourfold. First, while SDR has been an active area of research in statistics for decades, there are very few applications of SDR in economics. Exceptions include Ghosh (2011) and Luo and Zhu (2020), who combine SDR with causal inference. However, little is known about the use of SDR to other areas in econometrics. We incorporate SDR into two-sample regression framework, which, to the best of our knowledge, has not been done before.

Second, we allow for heterogeneity in the population from which the two samples are drawn. Such heterogeneity often arises in practice when datasets are collected in different ways and/or at different times. Accounting for heterogeneous populations reflects recent developments in the literature on regression estimation via data combination or in the presence of missing data. Examples include Zhao et al. (2019), Dai and Shao (2024) and Freyberger et al. (2025), among others. Accordingly, the regularity conditions in Section 3.1 are developed under the possibility of population heterogeneity.

Third, the literature on missing data imputation typically assumes that the missing regressor is continuous. However, applications in which proxies of a missing binary variable are imputed from a different data source are commonplace in marketing and

related fields; see Section 2.1 for further discussion. Our imputation scheme applies equally to both continuous and binary missing regressors.

Fourth, estimating the asymptotic covariance matrix of PILS-SDR is of independent interest. The core component of the covariance matrix involves outer products of two independent influence functions and must be estimated using a combination of analytical and resampling methods. Most of these influence functions can be derived analytically. However, the influence function of the SDR estimator, which is a key component, remains unspecified. We therefore adopt a resampling-based approach and replace this unknown quantity with the jackknife influence function à la Efron (1992).

The remainder of the paper is organized as follows. Section 2 sets the stage through a brief literature review and defines the PILS-SDR estimator. The implementation details, including the estimation of index coefficients and link functions, are discussed there. Section 3 provides a set of regularity conditions and explores convergence properties of PILS-SDR. Consistency and asymptotic normality is shown there. Section 4 conducts Monte Carlo simulations to explore finite-sample properties of PILS-SDR. As a real data example, Section 5 applies PILS-SDR in a return to schooling application. Section 6 concludes. A STATA code implementing PILS-SDR is available on GitHub[1].

In addition, Supplementary Materials available online contain five Supplements. Supplement A provides all technical proofs of the main theoretical results. Supplement B offers a consistent estimator of the asymptotic covariance matrix of PILS-SDR. Supplement C covers various challenges involved in extending MSII to include SDR. Supplement D presents the complete set of Monte Carlo simulation results. Supplement E contains a corrigendum for the variance formula given in HMP23.

The following notational convention is adopted throughout: we say '$a_n \asymp b_n$' if there exist constants $0 < c_1 < c_2 < \infty$ so that $c_1 a_n \leq b_n \leq c_2 a_n$; '$\overset{d}{=}$' denotes equality in distribution; $\|A\|$ is the Frobenius norm of matrix A, i.e., $\|A\| = \left\{\text{tr}\left(A^\top A\right)\right\}^{1/2}$; $\mathbb{I}\{\cdot\}$ denotes an indicator function; $h^{(p)}(x) = d^p h(x)/dx^p$ is the pth-order derivative of a function $h(x)$; $0_{p\times q}$ signifies the $p \times q$ zero matrix; and the symbol '$>$' applied to matrices means positive definiteness.

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

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 in economics is a measure of ability in a return to schooling estimation of Mincer's (1974) wage regression. An estimate of the return suffers from the so called "ability bias" unless a proxy for ability is included in the regression; see Card (1995) for more details. 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 generally contain this piece of information.

A similar situation arises in many other settings in economics. In studies of gender wage gap, work experience is an important regressor in wage regressions (e.g., Zabalza and Arrufat, 1985; 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 a missing regressor (e.g., Björklund and Jäntti, 1997; Murtazashvili et al., 2015). The missingness occurs because the information about income linkage across generations is unavailable. In models of consumer expenditure, household wealth is an important 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 to economics, 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.

In response to the problem of missing regressors, econometricians have developed a variety of estimation methods that combine information from two or more datasets. In this paper, we focus on methods in which two samples jointly identify the regression model of interest. These methods fall into two broad categories. The first is characterized by set identification of the model parameters. Examples include Pacini

(2019), D'Haultfœuille et al. (2024, 2025), and Hwang (2026). The second category achieves point identification of the model parameters. This paper, along with HP18 and HMP23, belongs to this class.

Similarly to HP18 and HMP23, Chen et al. (2008) and Graham et al. (2016) use an auxiliary sample to impute missing regressors. Their primary focus, however, is on parametric and semiparametric efficiency bounds in moment-based two-sample estimation, whereas this paper (like HP18 and HMP23) aims to improve nonparametric imputation methods for missing regressors.

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 (e.g., Klevmarken, 1982; Angrist and Krueger, 1992, 1995; Arellano and Meghir, 1992; Inoue and Solon, 2010; Murtazashvili et al., 2015; Pacini and Windmeijer, 2016; Zhao et al., 2019; Buchinsky et al., 2022). However, these approaches are not applicable to linear regression models with missing regressors because no sample analogue of $E(X_2 Y)$ can be constructed within either the primary or the auxiliary sample.

Finally, 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. Examples include Imbens and Lancaster (1994), Hellerstein and Imbens (1999), and Dai and Shao (2024).

2.2 Model Setup

To introduce the model, let the primary and auxiliary samples be denoted by $\mathcal{S}^A$ and $\mathcal{S}^B$, respectively. Let n and m be the numbers of observations in $\mathcal{S}^A$ and $\mathcal{S}^B$, respectively. Superscripts A and B indicate variables from the primary or auxiliary samples. Accordingly,

$$\mathcal{S}^A = \mathcal{S}^A_n = \left\{\left(Y^A_i, X^A_{1i}, X^A_{3i}\right)\right\}_{i=1}^n, \qquad \mathcal{S}^B = \mathcal{S}^B_m = \left\{\left(X^B_{2j}, X^B_{3j}\right)\right\}_{j=1}^m.$$

As mentioned in the previous section, these samples may or may not be drawn from the same population.

We restate the regression in (1) in terms of variables in $\mathcal{S}^A$ as the regression of Y^A on $X^A := \left(1, X_1^{A\top}, X_2^{A\top}, X_{3I}^{A\top}\right)^\top$:

$$Y^A = X^{A\top}\beta + u^A = \beta_0 + X_1^{A\top}\beta_1 + X_2^{A\top}\beta_2 + X_{3I}^{A\top}\beta_3 + u^A. \tag{2}$$

The regression in (2) would be complete if the missing regressor X_2^A were observable in $\mathcal{S}^A$.

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

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

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 index coefficient $\theta_\ell^s \in \mathbb{R}^{d_3}$. Let Θ^s denote the $d_2 \times d_3$ index coefficient matrix, where $\Theta^{s\top} := \begin{bmatrix} \theta_1^s & \cdots & \theta_{d_2}^s \end{bmatrix}$, and define $Z^s := \left(Z_1^s, \ldots, Z_{d_2}^s\right)^\top$ and $g_2^s\left(Z^s\right) := \left(g_{21}^s\left(Z_1^s\right), \ldots, g_{2d_2}^s\left(Z_{d_2}^s\right)\right)^\top$.

2.3 PILS-SDR Estimation

The PILS-SDR estimation procedure consists of three steps. In Steps 1 and 2 below, we estimate the conditional expectation $E\left(X_2^A \middle| X_3^A\right) = g_2^A\left(Z^A\right)$ consistently by combining the primary and auxiliary samples. To accomplish this task, we assume that both the index coefficient matrix and the link functions are identical across the two samples, i.e., $\Theta^A = \Theta^B = \Theta$, where $\Theta^\top = \begin{bmatrix} \theta_1 & \cdots & \theta_{d_2} \end{bmatrix}$, and $g_2^A(\cdot) = g_2^B(\cdot) =: g_2(\cdot)$. This "structural invariance" condition allows these quantities to be estimated using the auxiliary sample $\mathcal{S}^B$. Similar assumptions appear in Zhao et al. (2019) and Freyberger et al. (2025); see Section 3.1.2 for further discussion.

Step 1 (index coefficient estimation). Index coefficients $\theta_1, \ldots, \theta_{d_2}$ are estimated using the auxiliary sample $\mathcal{S}^B$ on an equation-by-equation basis. We estimate these coefficients using SDR directions as described in the next section. This step produces an SDR estimate of the index coefficient matrix $\hat{\Theta}^\top = \begin{bmatrix} \hat{\theta}_1 & \cdots & \hat{\theta}_{d_2} \end{bmatrix}$ and the estimated indices $\hat{Z}^s := \hat{\Theta} X_3^s = \left(\hat{\theta}_1^\top X_3^s, \ldots, \hat{\theta}_{d_2}^\top X_3^s\right)^\top =: \left(\hat{Z}_1^s, \ldots, \hat{Z}_{d_2}^s\right)^\top$ for $s \in \{A, B\}$.

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

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

where $K(\cdot)$ is a univariate symmetric kernel function and $h_\ell\,(>0)$ is the bandwidth for $\ell \in \{1,\ldots,d_2\}$. This step yields estimates $\left\{\hat{g}_2\left(\hat{Z}_i^A\right)\right\}_{i=1}^n$ of $\left\{g_2^A\left(Z_i^A\right)\right\}_{i=1}^n$. 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 $\hat{X}_g^A := \left(1, X_1^{A\top}, \hat{g}_2\left(\hat{Z}^A\right)^\top, X_{3I}^{A\top}\right)^\top$. The PILS-SDR estimator of β is

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

We show in Section 3.2 that this estimator is $\sqrt{n}$-consistent for β.

2.4 Implementation Details

2.4.1 Index Coefficient Estimation

The procedure in this step is 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 $\theta_1,\ldots,\theta_{d_2}$ via an SDR algorithm.

(iii) Normalize each of the SDR estimates $\hat{\theta}_1,\ldots,\hat{\theta}_{d_2}$ so that its first element equals one.

The estimation procedure in (ii) depends on whether the missing component $X_{2\ell}$ is continuous or binary.

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

To estimate the dimension reduction direction θ_ℓ by SIR, we follow three steps:

1. Sort the observations $\left\{\left(X_{2\ell,j}^B, X_{3j}^B\right)\right\}_{j=1}^m$ by $X_{2\ell}^B$ and partition the range of $\left\{X_{2\ell,j}^B\right\}_{j=1}^m$ into H slices $I_1,\ldots,I_H$ for some fixed $H \in \mathbb{N}$ so that each slice contains roughly the same number of observations.

2. Compute the sample mean of $\left\{X_{3j}^B\right\}_{j=1}^m$ within each slice I_k,

$$\bar{X}_{3(k)}^B := \sum_{j=1}^m \mathbb{I}\left\{X_{2\ell,j}^B \in I_k\right\} X_{3j}^B / m_k,$$

where $m_k := \sum_{j=1}^m \mathbb{I}\left\{X_{2\ell,j}^B \in I_k\right\}$ is the number of observations in slice I_k.

3. Compute the weighted covariance matrix

$$\hat{\Sigma}_H := \frac{1}{m} \sum_{k=1}^H m_k \left(\bar{X}_{3(k)}^B - \bar{X}_3^B\right)\left(\bar{X}_{3(k)}^B - \bar{X}_3^B\right)^\top,$$

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 $\hat{\Sigma} := \sum_{j=1}^m \left(X_{3j}^B - \bar{X}_3^B\right)\left(X_{3j}^B - \bar{X}_3^B\right)^\top/m$.

Unlike SIR, PLS has a closed-form expression. The PLS estimator of θ_ℓ 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

$$\hat{\mathcal{K}}_\ell = \hat{\mathcal{K}}_{\ell(q)} := \left[\begin{array}{cccc} \hat{\sigma}_\ell & \hat{\Sigma}\hat{\sigma}_\ell & \cdots & \hat{\Sigma}^{q-1}\hat{\sigma}_\ell \end{array}\right]$$

is the $d_3 \times q$ Krylov matrix generated by $\hat{\Sigma}$ and $\hat{\sigma}_\ell$ (Krylov, 1931), and

$$\hat{\sigma}_\ell := \frac{1}{m} \sum_{j=1}^m \left(X_{3j}^B - \bar{X}_3^B\right)\left(X_{2\ell j}^B - \bar{X}_{2\ell}^B\right),$$

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 another $d_3 \times q$ Krylov matrix obtained by replacing $\hat{\sigma}_\ell$ in $\hat{\mathcal{K}}_\ell$ with $\hat{\Sigma}\hat{\theta}_\ell^{SIR}$. Define

$$\tilde{\mathcal{K}}_\ell = \tilde{\mathcal{K}}_{\ell(q)} := \left[\begin{array}{cccc} \hat{\Sigma}\hat{\theta}_\ell^{SIR} & \hat{\Sigma}^2\hat{\theta}_\ell^{SIR} & \cdots & \hat{\Sigma}^q\hat{\theta}_\ell^{SIR} \end{array}\right].$$

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. Such parameters include the number of slices H for SIR and PIR and the order of the Krylov subspace q for PLS and PIR. For the former, 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 such as 5 or 10; we adopt $H = 10$. For the latter, Li et al. (2007) suggests a threshold approach for selecting q. Consider the $d_3 \times d_3$ Krylov matrix

$$\check{\mathcal{K}}_\ell = \check{\mathcal{K}}_{\ell(d_3)} := \left[\begin{array}{cccc} \varsigma_\ell & \hat{\Sigma}\varsigma_\ell & \cdots & \hat{\Sigma}^{d_3-1}\varsigma_\ell \end{array} \right],$$

where ς_ℓ is either $\hat{\sigma}_\ell$ for PLS or $\hat{\Sigma}\hat{\theta}_\ell^{SIR}$ for PIR. Let $\lambda_1, \ldots, \lambda_{d_3}$ 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

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

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

Binary Case. When the missing component $X_{2\ell}$ is binary, the estimation strategy for the dimension reduction direction differs from that used in the continuous cases. 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. More general discrete cases of $X_{2\ell}$ can be handled by converting them into a set of binary indicators.

One may ask why MLE works in this context. As demonstrated by Ruud (1983, 1986), the Probit and Logit coefficient estimators are consistent up to scale under the linearity condition (see Section 3.1.2), even if the error distribution is misspecified. This provides an example in which correct specification is not required for consistency of MLE, as discussed by White (1982, p.4).

2.4.2 Link Function Estimation

Step 2 is based on kernel regression estimation. The choice of kernel and bandwidth in this setting is the most important practical issue. Several practical guidelines are available for these choices.

For the reasons discussed in Section 3.1.3, we should use a higher-order kernel. Specifically, the order of the kernel should be greater than two. For instance, the kernel order can be set to four regardless of the dimensionality d_{3C}. An example of a fourth-order kernel is the twiced Gaussian kernel (e.g., Stuetzle and Mittal, 1979;

Newey et al., 2004)

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

It is constructed using the Gaussian kernel $K_G(u) = \exp(-u^2/2)/\sqrt{2\pi}$ and can also be recognized as the $G_{4,c}$ kernel of Wand and Schucany (1990), with $c = 1/\sqrt{2}$.

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

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

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

3 Convergence Properties of PILS-SDR

In this section, we explore the asymptotic properties of the PILS-SDR estimator. We defer the estimation of its asymptotic covariance matrix to the Appendix B as it requires a considerable amount of space. All convergence results are established under the asymptotic regime $n, m \to \infty$ with $n/m \to \kappa \in (0, \infty)$. 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

The regularity conditions account for specific methods of obtaining the dimension reduction directions discussed in Section 2.4.1. We categorize the regularity conditions into three groups: general conditions, conditions applicable to SDR/SIM, and conditions applicable to PLS. These three groups are labelled by letters G, S and P, respectively.

[2] As 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.

3.1.1 General Conditions

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

Assumption G2. $E\left(u^A \middle| X_1^A, X_3^A\right) = 0$, $\sigma_{u^A}^2\left(X_1^A, X_3^A\right) := E\left\{\left(u^A\right)^2 \middle| X_1^A, X_3^A\right\} \in (0, \infty)$, and $\beta_2 \neq 0_{d_2 \times 1}$ hold.

Assumption G3. Let $\eta_2^s = \left(\eta_{21}^s, \ldots, \eta_{2d_2}^s\right)^\top := X_2^s - E\left(X_2^s \middle| X_3^s\right)$ 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.

Assumption G1 allows for possibly heterogeneous populations, as in the recent literature on two-sample regression estimation including Zhao et al. (2019) and Dai and Shao (2024). Assumption G2 implies that the missing regressor X_2^A in (2) is relevant. Assumption G3 is a key condition for consistency of PILS-SDR and weaker than the conditional independence of η_2^A from X_1^A given X_3^A (e.g., Assumption 3(ii) of HMP23). It follows from this condition and the definition of η_2^A that $E\left(\eta_2^A \middle| X_1^A, X_3^A\right) = E\left(\eta_2^A \middle| X_3^A\right) = 0_{d_2 \times 1}$. A condition such as Assumption G3 is standard in the proxy variable literature. As discussed by Wooldridge (2010, p.68), a conditional mean independence assumption is routinely made or implied in all proxy-based estimators, and our approach is no different.

Now (2) can be rewritten as a regression of Y^A on $\left(1, X_1^{A\top}, E\left(X_2^A \middle| X_3^A\right)^\top, X_{3I}^{A\top}\right)^\top$ as follows:

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

where all regressors are orthogonal to the composite error term $\epsilon^A := u^A + \eta_2^{A\top}\beta_2$ under Assumptions G2 and G3. Therefore, if $E\left(X_2^A \middle| X_3^A\right)$ were observable (and linearly independent of X_{3I}^A), β could be consistently estimated by OLS.

3.1.2 Conditions for SDR and SIM

Assumption 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.

Assumption S2. For each $\ell \in \{1, \ldots, d_2\}$ 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:

(i) $X_{2\ell}^s \perp\!\!\!\perp X_3^s | \, Z_\ell^s$.

(ii) $g_{3\ell}^s(z) := E\,(X_3^s | \, Z_\ell^s = z)$ is a linear function of z.

Assumption 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 $\ell \in \{1, \ldots, d_2\}$ and $s \in \{A, B\}$.

Assumptions S1-S3 are fundamental to modelling of $E\,(X_{2\ell}^s | \, X_3^s)$ as a SIM. Although the first two conditions of Assumption S1 are appropriate when a proxy of $E\left(X_2^A \middle| \, X_3^A\right)$ is constructed using the proximity between the overlapping variables X_3^A and X_3^B, these conditions are less natural when the relevant notion of proximity is between the indices Z_ℓ^A and Z_ℓ^B. In fact, even if Assumption S1 holds, this does not automatically warrant Assumption S3, as argued in Abadie and Imbens (2016). Condition (iii) of Assumption S1 is standard in the SDR literature.

The first condition of Assumption S2 is a key condition for SDR and justifies the SIM specification of $E\,(X_{2\ell}^s | \, X_3^s)$. It postulates the existence of an index coefficient θ_ℓ^s such that the linear index Z_ℓ^s is sufficient for modelling of $X_{2\ell}^s$. By Proposition 1 of Cook and Li (2002), this condition implies that $E\,(X_{2\ell}^s | \, X_3^s)$ can be expressed as a function of Z_ℓ^s.

The second condition of Assumption S2 is the so-called *linearity condition*. This condition is essential for consistency of all estimation methods discussed in Section 2.4.1. A sufficient condition for Assumption S2(ii) is that the marginal distribution of X_3^s be *elliptically symmetric*.[3] If X_3^s indeed obeys an elliptically symmetric distribution, then it holds that

$$g_{3\ell}^s(z) = \mu^s + \frac{z - \theta_\ell^{s\top}\mu^s}{\theta_\ell^{s\top}\Sigma^s\theta_\ell^s}\Sigma^s\theta_\ell^s, \tag{10}$$

[3]Goodness-of-fit testing for elliptical symmetry has been a long-standing topic in statistics at least since Beran (1979). Recent literature on this testing problem includes Albisetti et al. (2020), Babić et al. (2021), Tang and Li (2024), and Wang and Lopes (2025), to name a few.

where μ^s and Σ^s are the mean and covariance matrix of X_3^s, respectively. An example of elliptically symmetric distributions is the multivariate normal distribution. The linearity condition also holds approximately for general large-dimensional covariates (e.g., Diaconis and Freedman, 1984; Hall and Li, 1993), and it is not considered restrictive in applications.

Assumption S4. For each $\ell \in \{1, \ldots, d_2\}$ 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, 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.

Assumption S5. Suppose that, for some $\ell \in \{1, \ldots, d_2\}$, the binary variable $X_{2\ell}^B$ admits the latent-index representation $X_{2\ell}^B := \mathbb{I}\left\{X_{2\ell}^{B*} > 0\right\}$, where the latent variable $X_{2\ell}^{B*}$ is generated by $X_{2\ell}^{B*} = \theta_\ell^{B\top} X_3^B + v_\ell^B$ for some regression error v_ℓ^B. Also suppose that the conditional density of $v_\ell^B \big| X_3^B$ is (mis)specified as $f_\ell\left(v_\ell^B \big| X_3^B; \theta_\ell^B\right)$. Moreover, let $\log f_{\ell,j}(\vartheta)$ denote $\log f_\ell\left(v_{\ell,j}^B \big| X_{3j}^B; \vartheta\right)$, the log-likelihood contribution of the j-th observation. Assume θ_ℓ^B is estimated by MLE and the parameter space $\bar{\Theta}_\ell^B$ is a convex and compact subset of $\mathbb{R}^{d_3}$. Assume that the following conditions hold:

(i) $\sup_{\vartheta \in \bar{\Theta}_\ell^B} \left| \sum_{j=1}^m \log f_{\ell,j}(\vartheta)/m - E\{\log f_{\ell,j}(\vartheta)\} \right| \xrightarrow{p} 0$.

(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 $\log f_{\ell,j}(\vartheta)$ with respect to ϑ are continuous and bounded uniformly on $\vartheta \in \bar{\Theta}_\ell^B$.

(iv) $\sup_{\vartheta \in \bar{\Theta}_\ell^B} E\left\|\partial \log f_{\ell,j}(\vartheta)/\partial\vartheta\right\|^2, \sup_{\vartheta \in \bar{\Theta}_\ell^B} E\left\|\partial^2 \log f_{\ell,j}(\vartheta)/\partial\vartheta\partial\vartheta^\top\right\| < \infty$.

Assumption S6. $\Theta^A = \Theta^B$, $Z_\ell^A \overset{d}{=} Z_\ell^B$ for all $\ell \in \{1, \ldots, d_2\}$, and $g_2^A(\cdot) = g_2^B(\cdot)$.

Assumptions S4-S6 are additional requirements for index coefficient estimation. Under Assumption S2, the index coefficient θ_ℓ^s can be estimated consistently up to a scaling factor. The first condition of Assumption S4 refers to a normalization of θ_ℓ^s required for its point identification. The second condition of Assumption S4 is

relevant when there are discrete overlapping variables, and a similar condition can be found in Ichimura (1993, Assumption 4.2).

Assumption S5 is relevant only when the index coefficient for a missing binary regressor is estimated by MLE. This assumption, jointly with the linearity condition in Assumption S2(ii), establishes consistency with the parametric convergence rate of the MLE. Ruud (1986) presents similar conditions to Assumption S5. In contrast to MLE, the SDR algorithms are free of distributional assumptions on the error terms, and thus the linearity condition is the primary requirement for consistency with the parametric convergence rate.

Assumptions S1-S5 enable us to estimate Θ^B consistently using the auxiliary sample. However, these conditions alone do not suffice for our objective of recovering Θ^A. Combining the primary and auxiliary samples must allow us to establish that $\left(E\left(X_{2\ell}^A \middle| X_3^A\right) =\right) E\left(X_{2\ell}^A \middle| Z_\ell^A\right) = E\left(X_{2\ell}^B \middle| Z_\ell^B\right) \left(= E\left(X_{2\ell}^B \middle| X_3^B\right)\right)$, and Assumption S6 is imposed precisely for this purpose. This assumption is an example of "structural invariance" conditions of Zhao et al. (2019) and Freyberger et al. (2025). Under Assumption S6, we may write $\Theta^A = \Theta^B = \Theta$ and $g_2^A(\cdot) = g_2^B(\cdot) = g_2(\cdot)$ so that their estimates are denoted by $\hat{\Theta}$ and $\hat{g}_2(\cdot)$, respectively, as they have already been defined in Section 2.2.

It is also worth emphasizing that $Z_\ell^A \overset{d}{=} Z_\ell^B$ can hold even when $X_3^A \overset{d}{=} X_3^B$ does not. Furthermore, let $r_\ell^s(\cdot) := g_{2\ell}^s(\cdot)\, p_\ell^s(\cdot)$ for $\ell \in \{1, \ldots, d_2\}$ and $s \in \{A, B\}$, where $p_\ell^s(\cdot)$ is defined in Assumption S3. Then, by Assumptions S1 and S6, the following conditions hold: (i) $\mathbb{Z}_\ell^A = \mathbb{Z}_\ell^B := \mathbb{Z}_\ell$; (ii) $p_\ell^A(\cdot) = p_\ell^B(\cdot) := p_\ell(\cdot)$ on $\mathbb{Z}_\ell$; and (iii) $r_\ell^A(\cdot) = r_\ell^B(\cdot) := r_\ell(\cdot)$ on $\mathbb{Z}_\ell$. In conclusion, these results allow us to estimate $\Theta = \Theta^A$ and $g_2(\cdot) = g_2^A(\cdot)$ consistently with the aid of the auxiliary sample. For later use, we define $\mathbb{Z} := \mathbb{Z}_1 \times \cdots \times \mathbb{Z}_{d_2}$, the Cartesian product of $\mathbb{Z}_1, \ldots, \mathbb{Z}_{d_2}$.

Assumption S7. The SDR estimate of the index coefficient matrix $\hat{\Theta}$ admits an asymptotic linear representation

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

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

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

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 B 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 expected to be considerably more complex (see Proposition 2 of Li et al., 2007). With this in mind, the influence function $\varphi_j(\Theta)$ in Assumption S7 is left unspecified. Assumption S7 ensures that for each SDR estimator $\hat{\theta}_\ell$,

$$\sqrt{m}\left(\hat{\theta}_\ell - \theta_\ell\right) = \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\varphi_{\ell j}^{B}(\theta_\ell) + o_p(1) \xrightarrow{d} N\left(0_{d_3\times 1}, V_{\varphi_\ell^B}\right), \tag{11}$$

where $\varphi_{\ell j}^{B}(\theta_\ell) := \varphi_\ell\left(X_{2j}^{B}, X_{3j}^{B}; \theta_\ell\right)$ 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. This result plays an important role in the analysis of the convergence properties of PILS-SDR.

3.1.3 Conditions for PILS

While dimension reduction only requires estimation of the index coefficients, estimation of β in (9) requires a consistent estimator of the unknown link functions. We now state the conditions required for consistent estimation of the link functions.

Assumption P1. Let $\theta_{\ell I}$ and $\theta_{\ell E}$ denote the coefficients on X_{3I}^{s} and X_{3E}^{s}, respectively. For each $\ell \in \{1,\ldots,d_2\}$ and $s \in \{A,B\}$, $\theta_{\ell I}$ and $\theta_{\ell E}$ are included in θ_ℓ. Assume that either (a) or (b) holds:

(a) If X_{3E}^{s} is non-empty and $\theta_{\ell E} \neq 0_{d_3\times 1}$, 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 $\theta_{\ell E} = 0_{d_3\times 1}$, then $g_{2\ell}(\cdot)$ is strictly nonlinear on $\mathbb{Z}_\ell$.

Assumption P1 is an identification condition for β and corresponds to Assumption 3(iii) of HMP23. In case (a), the excluded overlapping variable X_{3E}^{A} introduces additional randomness in $g_{2\ell}(\cdot)$, which can circumvent linear dependence between X_{3I}^{A} and $g_{2\ell}\left(Z^A\right)$.

Assumption P2. The kernel function $K:\mathbb{R}\to\mathbb{R}$ 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 $p \in \{1,\ldots,\nu-1\}$, 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 $\left|K^{(1)}(u) - K^{(1)}(u^*)\right| \leq L|u-u^*|$ for all $|u|, |u^*| \leq M$.
- **(b)** $\left|K^{(1)}(u)\right| \leq L$ for all $u \in \mathbb{R}$, $\left|K^{(1)}(u)\right| \leq L|u|^{-\rho}$ for all $|u| > M$, and $\left|K^{(1)}(u) - K^{(1)}(u^*)\right| \leq L|u-u^*|$ for all $u, u^* \in \mathbb{R}$.

Assumption P3. The bandwidths $h_1,\ldots,h_{d_2}$ satisfy $h_1,\ldots,h_{d_2} \asymp h > 0$ and

$$h + nh^{2\nu} + \frac{\log m}{mh^6} \to 0$$

as $n, m \to \infty$ so that $n/m \to \kappa \in (0,\infty)$.

Assumption P4. For each $\ell \in \{1,\ldots,d_2\}$, the νth-order derivatives of $p_\ell^{(1)}(z)$, $r_\ell^{(1)}(z)$, $E\left(X_1^A \middle| Z_\ell^A = z\right)$, and $E\left\{g_2\left(Z^A\right) \middle| Z_\ell^A = z\right\}$ are Lipschitz continuous and bounded uniformly on $z \in \mathbb{Z}_\ell$. In addition, there exist constants $\gamma \in (0,\infty)$ and $C \in [1,\infty)$ such that

$$\sup_{z\in\mathbb{Z}_\ell} E\left(\left|X_{2\ell}^B\right|^{2+\gamma} \middle| Z_\ell^B = z\right) \leq C.$$

Assumptions P2-P4 are conditions for kernel estimation of link functions $g_2(\cdot)$. Assumption P2 combines kernel conditions from Hansen (2008) and Ahn (1997). The former delivers conditions for weak uniform consistency of $\hat{g}_2(\cdot)$, whereas the latter provides additional requirements for SIMs. While the condition $nh^{2\nu} \to 0$ in Assumption P3 makes the dominant bias term in $\hat{g}_2(\cdot)$ asymptotically negligible, the condition $1/(mh^6) \to 0$, which follows from $\log m/(mh^6) \to 0$, controls the rates of the remainder terms.

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 2.4.2. 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 by Müller (1984) and higher-order Gaussian kernels by Wand and Schucany (1990) including (7). Similar conditions to Assumptions P2

and P3 can be commonly found in the literature on semiparametric SIMs. Examples include Powell et al. (1989), Härdle and Stoker (1989), Klein and Spady (1993), and Ahn (1997).

Assumption P4 is an additional requirement for weak uniform consistency of the hypothetical "oracle" NW regression estimator

$$\tilde{g}_{2\ell}(z) := \frac{\tilde{r}_\ell(z)}{\tilde{p}_\ell(z)} := \frac{\frac{1}{mh_\ell}\sum_{j=1}^m X_{2\ell,j}^B K\left(\frac{Z_{\ell,j}^B - z}{h_\ell}\right)}{\frac{1}{mh_\ell}\sum_{j=1}^m K\left(\frac{Z_{\ell j}^B - z}{h_\ell}\right)}, \tag{12}$$

which uses the true index Z_ℓ^B as if it were observed. 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. We provide proofs of the theorems below in Supplement A of the Supplementary Materials.

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

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 $n, m \to \infty$ *so that* $n/m \to \kappa \in (0, \infty)$, *where* $\Phi_{PS} := E\left(X_g^A X_g^{A\top}\right)$,

$$\begin{aligned}
X_g^A &:= \left(1, X_1^{A\top}, g_2\left(Z^A\right)^\top, X_{3I}^{A\top}\right)^\top, \\
\Omega_{PS} &:= \Omega_{PS,1} + \kappa\Omega_{PS,2} := E\left(\psi_{1i}\psi_{1i}^\top\right) + \kappa E\left(\psi_{2j}\psi_{2j}^\top\right), \\
\psi_{1i} &:= X_{gi}^A \epsilon_i^A = X_{gi}^A\left(u_i^A + \eta_{2i}^{A\top}\beta_2\right), \\
\psi_{2j} &:= \psi_{21j} + \psi_{22j} \\
&:= \sum_{\ell=1}^{d_2} G_\ell^A\left(Z_{\ell,j}^B\right)\eta_{2\ell,j}^B\beta_{2\ell} + \sum_{\ell=1}^{d_2} E\left[X_g^A\left\{X_3^A - g_{3\ell}^B\left(Z_\ell^A\right)\right\}^\top g_{2\ell}^{(1)}\left(Z_\ell^A\right)\right]\varphi_{\ell,j}^B(\theta_\ell)\,\beta_{2\ell}, \\
G_\ell^A(z) &:= E\left(X_g^A \middle| Z_\ell^A = z\right) \\
&= \left(1, E\left(X_1^A \middle| Z_\ell^A = z\right)^\top, E\left\{g_2\left(Z^A\right) \middle| Z_\ell^A = z\right\}^\top, E\left(X_{3I}^A \middle| Z_\ell^A = z\right)^\top\right)^\top,
\end{aligned}$$

and $\varphi_{\ell j}^B(\theta_\ell)$ *is given in (11).*

Theorems 1 and 2 jointly establish the $\sqrt{n}$-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. These results play a key role in the proofs of the two theorems.

The asymptotic linear representation of $\sqrt{n}\left(\hat{\beta}_{PS} - \beta\right)$ 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 $\hat{g}_2\left(\hat{Z}^A\right)$ in place of the missing regressor X_2^A is an example of *a generated regressor* in the sense of Pagan (1984, 1986), and it inflates the asymptotic variance of $\sqrt{n}\left(\hat{\beta}_{PS} - \beta\right)$ by introducing the second influence function ψ_{2j}. This influence function corresponds to the approximation error in $\hat{g}_2\left(\hat{Z}^A\right)$, and the two components, ψ_{21j} and ψ_{22j}, account for the estimation errors in the link functions and index coefficients, respectively. Furthermore, as shown in the proof of Lemma A4 in the Supplement, the derivation of ψ_{21j} largely relies on asymptotic analysis for *two-sample U-statistics*; see Section 6.1 of Lehmann (1999) for an excellent reference.

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 discussion on consistent estimation of V_{PS} to Supplement B of the Supplementary Materials.

4 Finite-Sample Performance

4.1 Monte Carlo Design

4.1.1 The Model

The simulations in this section examine how PILS-SDR performs under different SDR methods used to estimate the index coefficients. Consider the following linear regression model with five regressors:

$$Y^A = \beta_0 + X_1^A\beta_1 + X_2^{A\top}\beta_2 + X_{3I}^{A\top}\beta_3 + u^A, \tag{13}$$

where $X_1^A \in \mathbb{R}$, $X_2^A = \left(X_{21}^A, X_{22}^A\right)^\top \in \mathbb{R}^2$, $X_{3I}^A = \left(X_{3I1}^A, X_{3I2}^A\right)^\top \in \mathbb{R}^2$, $\beta_2 = (\beta_{21}, \beta_{22})^\top$, and $\beta_3 = (\beta_{31}, \beta_{32})^\top$. All coefficients are set equal to one. Suppose that the coefficients β_{21} and β_{31} are of particular interest. The corresponding regressors X_{21}^A (continuous) and X_{22}^A (binary) are missing in the primary sample and instead obtained from an auxiliary sample. The overlapping variables across the two samples, X_3^s for $s \in \{A, B\}$, consist of two included components (X_{3I}^s) and eight excluded components (X_{3E}^s). Alternatively, X_3^s can be partitioned into six continuous components (X_{3C}^s) and four binary components (X_{3D}^s). The first elements of X_{3C}^A and X_{3D}^A enter regression (13) as X_{3I1}^A (continuous) and X_{3I2}^A (binary), respectively.

4.1.2 Samples and Estimators

To estimate $\beta := (\beta_0, \beta_1, \beta_{21}, \beta_{22}, \beta_{31}, \beta_{32})^\top$, we consider four scenarios. In the first scenario, the complete sample $\mathcal{S}^* = \left\{\left(Y_i^A, X_{1i}^A, X_{2i}^A, X_{3Ii}^A\right)\right\}_{i=1}^n$ is observed. This is usually infeasible in practice. In this case, OLS can be applied to regression (13). This benchmark estimator is labelled OLS*.

In the second scenario, we simply run OLS on the short regression with X_2^A omitted from (13). 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 $\mathcal{S}^{A*} = \left\{\left(Y_i^A, X_{1i}^A, X_{3i}^A, Z_i^A\right)\right\}_{i=1}^n$ and $\mathcal{S}^{B*} = \left\{\left(X_{2j}^B, X_{3j}^B, Z_j^B\right)\right\}_{j=1}^m$, 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 and most realistic scenario involves estimating the index coefficients. In this case, we obtain primary and auxiliary samples using the estimated indices $\mathcal{S}^A = \left\{\left(Y_i^A, X_{1i}^A, X_{3i}^A, \hat{Z}_i^A\right)\right\}_{i=1}^n$ and $\mathcal{S}^B = \left\{\left(X_{2j}^B, X_{3j}^B, \hat{Z}_j^B\right)\right\}_{j=1}^m$. The corresponding PILS-SDR estimator is labelled PILS-SDR.

The sample sizes are $(n, m) = (2000, 1000)$, and the number of Monte Carlo replications is 1000.

4.1.3 Data Generation

The continuous component $X_{3C}^s \in \mathbb{R}^6$ for $s \in \{A, B\}$ is generated according to either Distribution 1 or Distribution 2. Distribution 1 is a truncated multivariate normal, defined as $X_{3C}^s \overset{iid}{\sim} TMVN\left(0, V_C^s; \sqrt{X_{3C}^{s\top}\left(V_C^s\right)^{-1} X_{3C}^s} \leq 3\right)$, where V_C^s is the covariance matrix with the typical element $\left(\rho_N^s\right)^{|p-q|}, p, q \in \{1, \ldots, 6\}$. This distribution is elliptically symmetric and therefore satisfies the linearity condition.

Distribution 2 is an equal-weight mixture of truncated normal and uniform distributions. More specifically, each component of X_{3C}^s is distributed as $X_{3C,p}^s \sim \left(TN_p + U_p\right)/2$, for $p \in \{1, \ldots, 6\}$, where TN_p denotes a standard normal variable truncated so that $\left|X_{3C,p}^s\right| \leq 3$, and U_p is a uniform random variable on $[-2, 2]$. The uniform variables are correlated over p and obtained by first generating $Z^* \overset{iid}{\sim} MVN\left(0, V_U^s\right)$, where V_U^s is the variance matrix with the typical element $\left(\rho_U^s\right)^{|p-q|}, p, q \in \{1, \ldots, 6\}$, and then applying the probability integral transform to each dimension p as $U_p = 4\Phi\left(Z_p^*\right) - 2$, where $\Phi(\cdot)$ is the standard normal distribution function. Clearly, Distribution 2 is an example of non-elliptical distributions.

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

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 $\left(\rho_N^A, \rho_D^A\right) = \left(\rho_N^B, \rho_D^B\right)$ or $\left(\rho_U^A, \rho_D^A\right) = \left(\rho_U^B, \rho_D^B\right)$ and heterogeneous otherwise. Finally, the true indices $\left(Z_1^s, Z_2^s\right) = \left(\theta_1^\top X_3^s, \theta_2^\top X_3^s\right)$ are computed using $X_3^s = \left(X_{3C}^{s\top}, X_{3D}^{s\top}\right)^\top$ 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 $\theta_2 = (1, 0, -1, 0, 1, 0, 0, 1/2, 0, -1/2)^\top$. Even when populations of X_3^A and X_3^B differ, $\left(Z_1^A, Z_2^A\right) \overset{d}{=} \left(Z_1^B, Z_2^B\right)$.

The remaining variables are generated as follows. For $\left(u^A, \eta_1^A, v, \eta_{21}\right) \overset{iid}{\sim} MVN\left(0, I_4\right)$, the regressor X_1^A is generated by $X_1^A := X_{3I1}^A + X_{3I2}^A + \eta_1^A \left(= X_{3C1}^A + X_{3D1}^A + \eta_1^A\right)$. The continuous missing regressors X_{21}^s for $s \in \{A, B\}$ are generated as $X_{21}^s := g_{21}\left(Z_1^s\right) + \eta_{21}$. The link function $g_{21}(z)$ is specified using one of the following formulas:

$$g_{21}(z) = \begin{cases} z & \text{[Model A]} \\ z + 9\phi(z) & \text{[Model B]} \\ \left[2\mathbb{I}\{z > 0\} + 1\right]\sqrt{|z|} & \text{[Model C]} \end{cases},$$

where $\phi(\cdot)$ is the standard normal density function. 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 $z \to 0$. 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 $g_{22}(z) = E\left(X_{22}^s \middle| Z_2^s = z\right) = \Phi(z)$. Therefore, the error term η_{22}^s can be defined as $\eta_{22}^s := X_{22}^s - \Phi\left(Z_2^s\right)$. Finally, Y^A is generated using Eq. (13).

4.1.4 Implementation Details and Performance Measures

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 combinations of estimators: (SIR, SIR), $(SIR, Probit)$, $(SIR, Logit)$, (PLS, SIR), $(PLS, Probit)$, $(PLS, Logit)$, (PIR, SIR), $(PIR, Probit)$, and $(PIR, Logit)$.

We first describe several implementation details. First, the number of slices for SIR and PIR of θ_1 is fixed at $H = 10$, and the order of Krylov subspaces q for PLS and PIR is chosen via the threshold approach using Eq. (6) with $a = 1.5$. Second,

all index coefficient estimates are normalized so that the first element equals one. Finally, for the link function estimation in PILS-SDR* and PILS-SDR, we use the twiced Gaussian kernel in (7) and the rule-of-thumb bandwidth in (8).

For the estimators 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 are computed as follows. Heteroskedasticity-robust standard errors by Eicker (1963) and White (1980) are calculated for OLS* and OLS-S. It can be shown that $\sqrt{n}\left(\hat{\beta}_{PS^*} - \beta\right) \xrightarrow{d} N\left(0_{(d+1)\times 1}, V_{PS^*}\right) := N\left(0_{(d+1)\times 1}, \Phi_{PS}^{-1}\Omega_{PS^*}\Phi_{PS}^{-1}\right)$, where the subscript "PS*" stands for "PILS-SDR*", $\Omega_{PS^*} := E\left(\psi_{1i}\psi_{1i}^\top\right) + \kappa E\left(\psi_{21j}\psi_{21j}^\top\right)$, and Φ_{PS}, ψ_{1i} and ψ_{21j} are defined in Theorem 2. Estimating the asymptotic covariance matrix V_{PS^*} as described in Supplement B yields standard errors for PILS-SDR*. Standard errors for all PILS-SDR estimators are obtained from the covariance estimation procedure described in Supplement B.

4.2 Results

Table 1 reports the performance measures 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 with Distribution 1 (Panel A) and Model C with Distribution 2 (Panel F). The complete set of results is available in Table D1 in Supplement D of the Supplementary Materials.

It is immediately clear from the table that OLS* outperforms all other estimators as would be expected. It is nearly unbiased and most efficient in all cases. Median standard errors are quite close to standard deviations of the corresponding parameter estimates. Coverage rates coincide with the nominal 95% level of confidence in most cases. However, OLS* is infeasible, and its performance should be interpreted just as a benchmark.

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: $(\rho_U^A, \rho_D^A) = (\rho_U^B, \rho_D^B) = (0.35, 0.40)$					
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: $(\rho_U^A, \rho_D^A) = (0.35, 0.40)$, $(\rho_U^B, \rho_D^B) = (-0.35, -0.40)$					
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.

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 the 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. This is the price to pay for switching from a one-sample to two-sample estimation. In contrast, it appears that the cost of estimating the index coefficient is marginal at best, as RMSEs of PILS-SDR are often smaller than those of corresponding PILS-SDR*.

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, 1986).

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 D. Even when the linearity condition is violated under a non-elliptical distribution (Distribution 2), the qualitative conclusions obtained under elliptical symmetry continue to hold.

Finally, it is difficult to determine which combination of SDR algorithms 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

5.1 Data Construction

In this section, we explore the empirical relevance of PILS-SDR within the classical framework of the human capital earnings equation of Mincer (1974), where education is endogenous and correlated with potentially unobservable skills and ability. We follow HMP23 and consider the following wage regression:

$$\begin{aligned}\log(income) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 abil \\ + \beta_5 married + \beta_6 black + \beta_7 south + \beta_8 urban + u,\end{aligned} \tag{14}$$

where $\log(income)$ is the natural logarithm of the total annual labor income; *educ* is years of schooling; *exper* is actual work experience; *abil* is an ability measure; *married* is an indicator for being married; *black* is an indicator for being black; *south* is an indicator for currently living in the Southern region of the United States; *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 in (14), we use the same two US micro datasets as HMP23, except that 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 (IQ) measured by the sentence completion test. Consequently, we may estimate (14) by OLS using 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 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 have the following two estimation strategies available to them. The first strategy is to estimate the short regression with IQ dropped from (14). 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 coefficient between *educ*

and IQ is 0.4496. To address the inconsistency, researchers 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.

The second strategy is to find an auxiliary sample and estimate (14) by PILS or PILS-SDR. Following HMP23, we adopt the `CARD` dataset as the auxiliary sample. The dataset is available with a popular textbook by Wooldridge (2013) and was used in Card (1995). 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. As argued by HMP23, the KWW score is a better measure of unobserved ability. All other regressors except *exper* are recorded in both the 1966 and 1976 interview.[4]

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 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 of the data construction, we obtain the primary dataset $\mathcal{S}^A$ with $(n =) 1303$ observations and the auxiliary dataset $\mathcal{S}^B$ with $(m =) 1516$ observations. The summary statistics of all variables in the two datasets are presented in Table 2. The data characteristics are close to those used by HMP23, except for *exper* and *age* in $\mathcal{S}^A$.

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

Table 2: Sample Characteristics

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

5.2 Additional Setup for PILS and PILS-SDR

We begin by categorizing the variables in $\mathcal{S}^A$ and $\mathcal{S}^B$. The variables in $\mathcal{S}^A$ are as follows:

$$
\begin{aligned}
Y^A &= \log(income)\,; X_1^A = \left(exper, exper^2\right); X_2^A = IQ; \\
X_{3I}^A &= (educ, married, black, south, urban)\,; \text{ and} \\
X_{3E}^A &= (fatheduc, motheduc, age, south\_in\_youth)\,.
\end{aligned}
$$

Similarly, the variables in $\mathcal{S}^B$ can be categorized as follows:

$$X_2^B = KWW; X_{3I}^B = (educ, married, black, south, urban); \text{ and}$$
$$X_{3E}^B = (fatheduc, motheduc, age, south\_in\_youth).$$

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 scenario uses the variables

$$\{educ, age, married, black, south, urban, south\_in\_youth\}$$

in this order. Because two of these variables are continuous (i.e., $d_{3C} = 2$), both PILS and PLIS-SDR are $\sqrt{n}$-consistent. The second scenario uses the variables

$$\{educ, fatheduc, motheduc, age, married, black, south, urban, south\_in\_youth\}$$

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.

We implement PILS and PILS-SDR as follows. For PILS, each continuous overlapping variable is smoothed by the Epanechnikov kernel and each binary overlapping variable is smoothed by the discrete kernel of Li and Racine (2003). We use the rule-of-thumb bandwidths $\hat{h} = \hat{\sigma}_B (\log m/m)^c$ for the Epanechnikov and $\hat{\lambda} = (\log m/m)^{2c}$ for the discrete kernel, where $\hat{\sigma}_B$ is the sample standard deviation of the continuous variable from $\mathcal{S}^B$ and c is set at 0.3 for the case $d_{3C} = 2$ and 0.225 for the case $d_{3C} = 4$. PILS loses $\sqrt{n}$-consistency in the latter case, and the choice of c simply ensures uniform consistency of the nonparametric regression estimator as a proxy for the missing regressor.

For PILS-SDR, we normalize the index coefficient on $educ$ to be one and we fix the number of slices for SIR and PIR at $H = 10$. For the case $d_{3C} = 2$, the threshold approach using (6) with $a = 1.5$ yields estimated orders of the Krylov subspaces $q = 5$ for PLS and $q = 4$ for PIR. For the case $d_{3C} = 4$, the same approach yields $q = 4$ for both PLS and PIR. For link function estimation, we again employ the twiced Gaussian kernel in (7) together with the rule-of-thumb bandwidth in (8).

5.3 Estimation Results

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 B and E of the Supplementary Materials, respectively, 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) of Table 3. 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 $\mathcal{S}^A$ and $\mathcal{S}^B$, while *abil* is missing in $\mathcal{S}^A$. 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 (14) 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-

Table 3: Estimation Results

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

Notes: The dependent variable is log (*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*.

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 regression estimation methods and instrument-based methods.

6 Conclusion

In regression analysis using a primary sample, we often face the situation in which some important regressors are unavailable. If an auxiliary sample exists that contains the missing regressors along with variables overlapping with the primary sample, we can estimate the regression model consistently. The PILS estimator by HMP23 is an example of such two-sample estimation methods. However, similar to other nonparametric estimators, this method suffers from the curse of dimensionality. PILS attains the parametric rate of convergence if the number of continuous overlapping variables no greater than three.

We develop a natural extension to PILS, which incorporates sufficient dimension reduction (SDR) studied extensively over the past years in other settings. The dimension reduction step relies on the assumption that the conditional mean of each missing regressor given the overlapping variables can be specified as a semiparametric single-index model (SIM) with an unknown link function.

We derive asymptotic properties of PILS-SDR and show how to implement it in three steps. In the first step, the index coefficients are estimated through SDR algorithms. Given the estimated indices, the second step estimates the link functions by kernel regression smoothing. In the third and final step, we replace the missing regressors by their proxies and proceed to OLS estimation. PILS-SDR can accommodate both continuous and binary missing regressors and remains useful when the two samples are drawn from potentially different populations.

We establish asymptotic normality of PILS-SDR and demonstrate that it attains the parametric rate of convergence even when there are more than three continuous overlapping variables. Desirable finite-sample properties of PILS-SDR are confirmed through a Monte Carlo study. An empirical application to Mincer's (1974) wage regression illustrates the practical relevance of the estimator.

References

[1] Abadie, A., and G. W. Imbens (2016): "Matching on the Estimated Propensity Score," *Econometrica*, 84, 781 - 807.

[2] Ahn, H. (1997): "Semiparametric Estimation of a Single-Index Model with Nonparametrically Generated Regressors," *Econometric Theory*, 13, 3 - 31.

[3] Albisetti, I., F. Balabdaoui, and H. Holzmann (2020): "Testing for Spherical and Elliptical Symmetry," *Journal of Multivariate Analysis*, 180, Article No.104667.

[4] Angrist, J., and A. Krueger (1992): "The effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, 87(418), 328 - 336.

[5] Angrist, J., and A. Krueger (1995): "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business and Economic Statistics*, 13(2), 225 - 235.

[6] Arellano, M., and C. Meghir (1992): "Female Labour Supply and on the Job Search: An Empirical Model Estimated Using Complementary Data Sets," *Review of Economic Studies*, 59, 537 - 559.

[7] Babić, S., L. Gelbgras, M. Hallin, and C. Ley (2019): "Optimal Tests for Elliptical Symmetry: Specified and Unspecified Location," *Bernoulli*, 27, 2189 - 2216.

[8] Beran, R. (1979): "Testing for Ellipsoidal Symmetry of a Multivariate Density," *Annals of Statistics*, 7, 150 - 162.

[9] Björklund, J., and M. Jäntti (1997): "Intergenerational Income Mobility in Sweden Compared to the United States," *American Economic Review*, 87, 1009 - 1018.

[10] Black, B., M. Trainor, and J. E. Spencer (1999): "Wage Protection Systems, Segregation and Gender Pay Inequalities: West Germany, the Netherlands and Great Britain," *Cambridge Journal of Economics*, 23, 449 - 464.

[11] Bollinger, C.R., and B. T. Hirsch (2006): "Match bias from earnings imputation in the current population survey: The case of imperfect matching, " *Journal of Labor Economics*, 24, 483 - 519.

[12] Bostic, R., S. Gabriel, and G. Painter (2009): "Housing Wealth, Financial Wealth, and Consumption: New Evidence from Micro Data," *Regional Science and Urban Economics*, 39, 79 - 89.

[13] Buchinsky, M., F. Li, and Z. Liao (2022): "Estimation and Inference of Semiparametric Models Using Data from Several Sources," *Journal of Econometrics*, 226, 80 - 103.

[14] Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in 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.

[15] Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric Efficiency in GMM Models with Auxiliary Data," *Annals of Statistics*, 36, 808 - 843.

[16] Cook, R. D., I. S. Helland, and Z. Su (2013): "Envelopes and Partial Least Squares Regression," *Journal of the Royal Statistical Society, Series B*, 75, 851 - 877.

[17] Cook, R. D., and H. Lee (1999): "Dimension Reduction in Binary Response Regression," *Journal of the American Statistical Association*, 94, 1187 - 1200.

[18] Cook, R. D., and B. Li (2002): "Dimension Reduction for Conditional Mean in Regression," *Annals of Statistics*, 30, 455 - 474.

[19] Dai, C.-S., and J. Shao (2024): "Kernel Regression Utilizing Heterogeneous Datasets," *Statistical Theory and Related Fields*, 8, 51 - 68.

[20] D'Haultfœuille, X., C. Gaillac, and A. Maurel (2024): "Linear Regressions with Combined Data," Working Paper No.1602, Toulouse School of Economics.

[21] D'Haultfœuille, X., C. Gaillac, and A. Maurel (2025): "Partially Linear Models Under Data Combinations," *Review of Economic Studies*, 92, 238 - 267.

[22] Diaconis, P., and D. Freedman (1984): "Asymptotics of Graphical Projection Pursuit," *Annals of Statistics*, 12, 793 - 815.

[23] Duan, N., and K.-C. Li (1991): "Slicing Regression: A Link-Free Regression Method," *Annals of Statistics*, 19, 505 - 530.

[24] Efron, B. (1992): "Jackknife-After-Bootstrap Standard Errors and Influence Functions," *Journal of the Royal Statistical Society, Series B*, 54, 83 - 127.

[25] Eicker, F. (1963): "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," *Annals of Mathematical Statistics*, 34, 447 - 456.

[26] Freyberger, J., B. Hoeppner, A. Neuhierl, and M. Weber (2025): "Missing Data in Asset Pricing Panels," *Review of Financial Studies*, 38, 760 - 802.

[27] Ghosh, D. (2011): "Propensity Score Modelling in Observational Studies Using Dimension Reduction Methods," *Statistics & Probability Letters*, 81, 813 - 820.

[28] Graham, B. S., C. C. de Xavier Pinto, and D. Egel (2016): "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)," *Journal of Business & Economic Statistics*, 34, 288 - 301.

[29] Hall, P., and K.-C. Li (1993): "On Almost Linearity of Low Dimensional Projections from High Dimensional Data," *Annals of Statistics*, 21, 867 - 889.

[30] Hansen, B. E. (2008): "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory*, 24, 726 - 748.

[31] Härdle, W., and T. M. Stoker (1989): "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986 - 995.

[32] He, X., and Q.-M. Shao (2000): "On Parameters of Increasing Dimensions," *Journal of Multivariate Analysis*, 73, 120 - 135.

[33] Hellerstein, J. K., and G. W. Imbens (1999): "Imposing Moment Restrictions from Auxiliary Data by Weighting," *Review of Economics and Statistics*, 81, 1 - 14.

[34] Hirukawa, M., I. Murtazashvili, and A. Prokhorov (2023): "Yet Another Look at the Omitted Variable Bias," *Econometric Reviews*, 42, 1 - 27.

[35] Hirukawa, M., and A. Prokhorov (2018): "Consistent Estimation of Linear Regression Models Using Matched Data," *Journal of Econometrics*, 203, 344 - 358.

[36] Hwang, Y. (2026): "Bounding Omitted Variable Bias Using Auxiliary Data: With an Application to Estimate Neighborhood Effects," *Journal of Business & Economic Statistics*, forthcoming.

[37] Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58, 71 - 120.

[38] Imbens, G. W., and T. Lancaster (1994): "Combining Micro and Macro Data in Microeconometric Models," *Review of Economic Studies*, 61, 655 - 680.

[39] Inoue, A., and G. Solon (2010): "Two-Sample Instrumental Variables Estimators," *Review of Economics and Statistics*, 92, 557 - 561.

[40] Kamakura, W. A., and M. Wedel (1997): "Statistical Data Fusion for Cross-Tabulation," *Journal of Marketing Research*, 34, 485 - 498.

[41] Klein, R. W., and R. H. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, 387 - 421.

[42] Klevmarken, A. (1982): "Missing Variables and Two-Stage Least Squares Estimation from More than One Data Set," IFN Working Paper, No. 62, Stockholm: Research Institute of Industrial Economics (IFN).

[43] 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.

[44] Lehmann, E. L. (1999): *Elements of Large-Sample Theory*. New York: Springer-Verlag.

[45] Li, K.-C. (1991): "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316 - 327.

[46] Li, L., R. D. Cook, and C.-L. Tsai (2007): "Partial Inverse Regression," *Biometrika*, 94, 615 - 625.

[47] Li, Q., and J. Racine (2003): "Nonparametric Estimation of Distributions with Categorical and Continuous Data," *Journal of Multivariate Analysis*, 86, 266 - 292.

[48] Luo, W., and Y. Zhu (2020): "Matching Using Sufficient Dimension Reduction for Causal Inference," *Journal of Business & Economic Statistics*, 38, 888 - 900.

[49] Mincer, J. A. (1974): *Schooling, Experience and Earnings*. New York: National Bureau of Economic Research.

[50] Müller, H.-G. (1984): "Smooth Optimum Kernel Estimators of Densities, Regression Curves and Modes," *Annals of Statistics*, 12, 766 - 774.

[51] Murtazashvili, I., D. Liu, and A. Prokhorov (2015): "Two-Sample Nonparametric Estimation of Intergenerational Income Mobility in the United States and Sweden," *Canadian Journal of Economics*, 48, 1733 - 1761.

[52] Naik, P., and C.-L. Tsai (2000): "Partial Least Squares Estimator for Single-Index Models," *Journal of the Royal Statistical Society, Series B*, 62, 763 - 771.

[53] Newey, W. K., F. Hsieh, and J. M. Robins (2004): "Twicing Kernels and a Small Bias Property of Semiparametric Estimators," *Econometrica*, 72, 947 - 962.

[54] Pagan, A. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25, 221 - 247.

[55] Pagan, A. (1986): "Two Stage and Related Estimators and Their Applications," *Review of Economic Studies*, 53, 517 - 538.

[56] Pacini, D. (2019): "Two-Sample Least Squares Projection," *Econometric Reviews*, 38, 95 - 123.

[57] Pacini, D. and F. Windmeijer (2016): "Robust Inference for the Two-Sample 2SLS estimator," *Economics Letters*, 146, 50 - 54.

[58] Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403 - 1430.

[59] Ridder, G., and R. Moffitt (2007): "The Econometrics of Data Combination," in J. J. Heckman and E. E. Leamer (eds.), *Handbook of Econometrics*, Vol. 6, Part B. Amsterdam: Elsevier, Chapter 75, 5469 - 5547.

[60] 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.

[61] Ruud, P. A. (1986): "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution," *Journal of Econometrics*, 32, 157 - 187.

[62] Saracco, J. (1997): "An Asymptotic Theory for Sliced Inverse Regression," *Communications in Statistics - Theory and Methods*, 26, 2141 - 2171.

[63] Stuetzle, W., and Y. Mittal (1979): "Some Comments on the Asymptotic Behavior of Robust Smoothers," in T. Gasser and M. Rosenblatt (eds.), *Smoothing Techniques for Curve Estimation: Proceedings of a Workshop Held in Heidelberg, April 2 - 4, 1979*. Berlin: Springer-Verlag, 191 - 195.

[64] Tang, Y., and B. Li (2024): "A Nonparametric Test for Elliptical Distribution Based on Kernel Embedding of Probabilities," *Annals of Statistics*, 52, 2349 - 2374.

[65] Wand, M. P., and W. R. Schucany (2001): "Gaussian-Based Kernels," *Canadian Journal of Statistics*, 18, 197 - 204.

[66] Wang, S., and M. E. Lopes (2025): "Testing Elliptical Models in High Dimensions," *Journal of the American Statistical Association*, forthcoming.

[67] White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817 - 838.

[68] White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1 - 25.

[69] Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, 2nd Edition. Cambridge, MA: MIT Press.

[70] Wooldridge, J. M. (2013): *Introductory Econometrics: A Modern Approach*, 5th Edition. Mason, OH: South-Western Cengage Learning.

[71] Zabalza, A., and J. L. Arrufat (1985): "The Extent of Sex Discrimination in Great Britain," in A. Zabalza and Z. Tzannatos (eds.), *Women and Equal Pay: The Effects of Legislation on Female Employment and Wages in Britain*, Cambridge, U.K.: Cambridge University Press, 70 - 96.

[72] Zhao, Q., J. Wang, W. Spiller, J. Bowden, and D. S. Small (2019): "Two-Sample Instrumental Variable Analyses Using Heterogeneous Samples," *Statistical Science*, 34, 317 - 333.