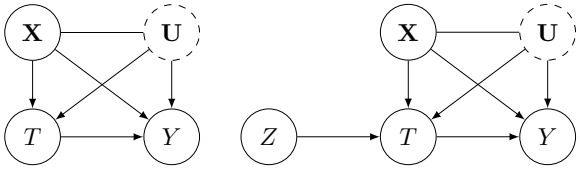
Two-Stage Shadow Inclusion Estimation: An IV Approach for Causal Inference under Latent Confounding and Collider Bias

Baohong Li [1] **Anpeng Wu** [1] **Ruoxuan Xiong** [2] **Kun Kuang** [1]

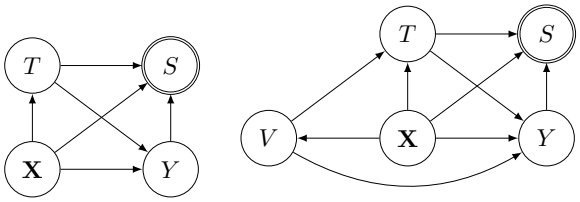
Abstract

Latent confounding bias and collider bias are two key challenges of causal inference in observational studies. Latent confounding bias occurs when failing to control the unmeasured covariates that are common causes of treatments and outcomes, which can be addressed by using the Instrumental Variable (IV) approach. Collider bias comes from non-random sample selection caused by both treatments and outcomes, which can be addressed by using a different type of instruments, i.e., shadow variables. However, in most scenarios, these two biases simultaneously exist in observational data, and the previous methods focusing on either one are inadequate. To the best of our knowledge, no approach has been developed for causal inference when both biases exist. In this paper, we propose a novel IV approach, Two-Stage Shadow Inclusion (2SSI), which can simultaneously address latent confounding bias and collider bias by utilizing the residual of the treatment as a shadow variable. Extensive experimental results on benchmark synthetic datasets and a real-world dataset show that 2SSI achieves noticeable performance improvement when both biases exist compared to existing methods.



(a) Latent confounding. (b) Instrumental variable approach.

Figure 1. Causal graphs for illustrating the latent confounding problem and the instrumental variable approach. The dashed node denotes that the variable is unmeasured.



(a) Collider bias. (b) Shadow variable approach.

Figure 2. Causal graphs for illustrating the collider bias problem and the shadow variable approach.

1. Introduction

Causal inference empowers machine learning methods to learn causality other than correlations from observational data. It has achieved remarkable success in trustworthy machine learning studies (Cui & Athey, 2022; Nilforoshan et al., 2022; Wang et al., 2022; Zhang et al., 2023; 2024). The primary task of causal inference in observational studies is to remove spurious correlations, of which the two most common sources are confounding bias and collider bias (Hernán & Robins, 2020).

Figure 1 shows the causal graph of (latent) confounding bias and a method to address it, where T denotes the treatment variable, $\mathbf{X}$ denotes the observed/measured covariates, $\mathbf{U}$ denotes the unobserved/unmeasured covariates, Y denotes the outcome variable, and Z denotes an Instrumental Variable (IV). As shown in Figure 1(a), confounding bias occurs when common causes of T and Y, namely confounders, are not measured and controlled. Confounding bias introduces spurious correlations between T and Y, e.g., the non-causal path $T \leftarrow \mathbf{U} \rightarrow Y$. In this paper, we focus on confounding bias caused by unmeasured confounders. The IV approach is commonly used to address latent confounding bias, that was initially proposed assuming the existence of well-defined

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China [2]Department of Quantitative Theory & Methods, Emory University, Atlanta, USA. Correspondence to: Kun Kuang <kunkuang@zju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

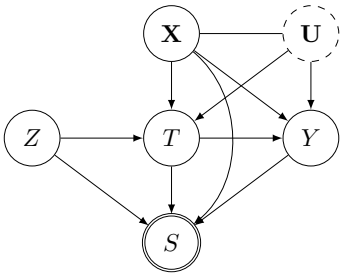


Figure 3. Causal graph for illustrating our problem setting. The undirected edge between $\mathbf{X}$ and $\mathbf{U}$ implies that $\mathbf{X} \rightarrow \mathbf{U}$ or $\mathbf{U} \rightarrow \mathbf{X}$, or they have more intricate interactions that are not straightforwardly directional.

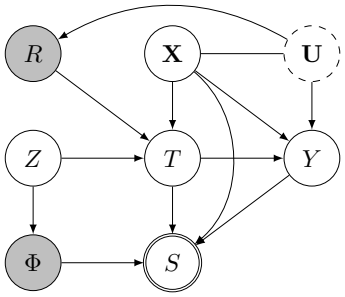


Figure 4. Causal graph for illustrating our motivation, where the gray node denotes the variable is learned from observed variables.

IVs under linearity settings (Angrist et al., 1996). As shown in Figure 1(b), a valid IV Z is a cause of T, not a direct cause of Y, and independent of $\mathbf{U}$. Recent studies have generalized the IV approach to nonlinear scenarios (Terza et al., 2008; Hartford et al., 2017; Xu et al., 2021). The basic idea of the IV approach is to regress Y on the estimated T by Z, such that $\mathbf{U}$ is conditional independent of the estimated T in the regression.

Figure 2 shows the causal graph for illustrating the collider bias and a method to address it, where S denotes the binary selection indicator and V denotes a shadow variable. S can be caused by both T and Y. If $S = 1$, then the unit is in the observational samples and all the variables are fully measured; otherwise, the unit is also in the observational samples, but her outcome value is missing (Heckman, 1979). Figure 2(a) shows the collider bias that arises from non-random sample selection conditional on S. Collider bias also introduces spurious correlations between T and Y, e.g., the non-causal path $T \rightarrow S \leftarrow Y$. Previous studies use shadow variables to address collider bias (d'Haultfoeuille, 2010; Miao & Tchetgen Tchetgen, 2016). As shown in Figure 2(b), a shadow variable V is associated with Y conditional on T in the $S = 1$ samples and not a direct cause of S. The idea of the shadow variable method is to generalize the results on the $S = 1$ samples to the $S = 0$ samples to remove the spurious correlation caused by S.

Although IV and shadow variable methods have successfully addressed either latent confounding bias or collider bias, they are inadequate when both biases exist. As shown in Figure 3, when both biases exist, IV approaches can only address the spurious correlation $T \leftarrow \mathbf{U} \rightarrow Y$ caused by latent confounding, and the estimate is still biased due to the other spurious correlations caused by conditioning on S. What is worse, collider bias also makes the estimation results on $S = 1$ samples inaccurate on the $S = 0$ samples because of the distribution shift between the $S = 1$ and $S = 0$ samples. Meanwhile, if there exists a shadow variable to address collider bias, this variable does not address the latent confounding bias.

In real-world scenarios, latent confounding bias and collider bias commonly both exist (Hernán & Robins, 2020; Griffith et al., 2020). For example, in the context of selecting pilots, we study the effect of a voluntary pilot training program (T) on the final testing results (Y), with covariates ($\mathbf{X}$) such as physical condition and family economic status. Officers/HR select whom to take the tests based on paperwork, i.e., whether taking training programs (T), height, weight, age ($\mathbf{X}$), etc., and release the admitted ones' score (the score Y is larger than a specific value). This selection mechanism (S) is not affected by candidates' latent characteristics ($\mathbf{U}$) (e.g., candidates' passion for the pilot profession). Only the test scores (Y) of selected pilots ($S = 1$) are released. We do not observe the test scores (Y) for unselected pilots ($S = 0$), leading to collider bias. Also, candidates who are passionate about the profession may take more time to prepare for the test and perform better in the test (so $\mathbf{U}$ affects Y) and also might be more willing to participate in training (so $\mathbf{U}$ affects T), leading to latent confounding bias. To the best of our knowledge, no approach has been developed for causal inference under both latent confounding and collider bias. Therefore, it is crucial to develop an approach for causal inference under both biases.

As previously mentioned, prior research on instrumental variables (IVs) has established identification theory under latent confounding when well-defined IVs are accessible. Similarly, studies on shadow variables have provided identification theory in the presence of collider bias when well-defined shadow variables are available. However, in practical scenarios, the simultaneous availability of both well-defined instrumental variables and shadow variables is rare. In this paper, we propose a novel approach: leveraging well-defined IVs to construct the shadow variable and address latent confounding and collider bias. Thus, we introduce Two-Stage Shadow Inclusion (2SSI), a method designed to achieve this objective. As shown in Figure 4, in the first

stage of 2SSI, we regress T on $\mathbf{X}$ and Z, and learn a decomposed representation Φ of Z that is only related to S and conditional independent of T. Consequently, we obtain the residual R of T that is a *shadow variable* because it is independent of S conditional on $\mathbf{X}, T, Y$, and Φ, and it is related to Y through $R \leftarrow \mathbf{U} \rightarrow Y$. In the second stage of 2SSI, we use R and Φ to address the spurious correlations caused by $T \leftarrow \mathbf{U} \rightarrow Y$ with $S = 1$ samples, and then generalize the results to the whole data space to address collider bias. We conducted extensive experiments on benchmark and real-world datasets. The experimental results show that the proposed method outperforms existing methods under latent confounding and collider bias.

In summary, the contributions of this paper are as follows.

- We study a challenging and important problem of causal inference in observational studies, i.e., causal inference under latent confounding and collider bias.
- We propose a novel method that simultaneously addresses latent confounding and collider bias. To the best of our knowledge, this is the first approach developed for causal inference when both biases exist.
- Extensive experimental results on both synthetic and real-world datasets demonstrate the effectiveness of the proposed method.

2. Related Work

IV approaches for latent confounding. The series of IV approaches is a commonly employed way to address latent confounding. The most famous IV method is Two-Stage Least Regression (2SLS), which uses IVs to estimate the treatment and utilizes the estimated treatment to estimate the outcome under linear settings (Angrist et al., 1996; Angrist & Krueger, 2001; Brito & Pearl, 2002; Baiocchi et al., 2014). In nonlinear scenarios, Two-Stage Residual Inclusion (2SRI) is proposed to use the residual of the treatment from the first stage regression to estimate the outcome in the second stage (Terza et al., 2008). Recent studies utilize machine learning techniques to apply IV approaches to more complex real-world scenarios (Hartford et al., 2017; Bennett et al., 2019; Singh et al., 2019; Xu et al., 2021; Wu et al., 2022). All the above IV approaches can only address latent confounding. When there is also collider bias, since the Y values of $S = 0$ data are missing, they can only use the $S = 1$ samples for the second-stage regression, which brings two challenges for the above methods: (1) The estimate $\mathbb{E}[Y \mid \mathbf{X}, T, S = 1] \neq \mathbb{E}[Y \mid \mathbf{X}, T]$ because $Y \not\perp\!\!\!\perp S \mid \mathbf{X}, T$. (2) They can only eliminate the spurious correlations caused by $\mathbf{U} \rightarrow T$, but still suffer from the ones introduced by conditioning on S. Note that other methods for the latent confounding problem, like data fusion and negative control methods (Shi et al., 2020; Colnet et al., 2023), though being out of the scope of this paper because the key assumptions are different, also cannot address collider bias because of the challenges.

Methods for non-random sample selection. Collider bias can be regarded as a special case of the non-random sample selection problem, a.k.a. sample selection bias. Previous works on sample selection bias, including Heckit and its variants (Heckman, 1979; Ding, 2014; Ogundimu & Hutton, 2016; Wiemann et al., 2022), Inverse Probability of Sampling Weights (IPSW) (Cole & Stuart, 2010), and selection-backdoor adjustment (Bareinboim et al., 2022; Bareinboim & Tian, 2015), focus on the non-random sample selection caused by only the covariates and treatments and cannot deal with collider bias. d'Haultfoeuille (2010); Miao & Tchetgen Tchetgen (2016); Li et al. (2023) propose approaches that leverage a different type of IV, namely shadow variable, to address collider bias. These methods use a well-defined shadow variable V to generalize the estimate results obtained from the $S = 1$ samples to $S = 0$ data, i.e., to generalize $\mathbb{E}[Y \mid \mathbf{X}, T, V, S = 1]$ to $\mathbb{E}[Y \mid \mathbf{X}, T, V, S = 0]$, such that all the spurious correlations introduced by conditioning on S can be eliminated. However, their performance relies on the assumption that all the possible confounders are fully measured. If there is also latent confounding, $\mathbb{E}[Y \mid \mathbf{X}, T, V, S = 1]$ is not only biased by conditioning on S but also biased by the non-causal path $T \leftarrow \mathbf{U} \rightarrow Y$, making shadow variable approaches unavailable.

To the best of our knowledge, there is currently no method developed for causal inference under both latent confounding and collider bias. Therefore, we propose a novel IV approach to fill this gap in this paper.

3. Preliminaries

3.1. Problem Formulation

Suppose we have the observational data $\mathbb{D} = \{\mathbf{x}_i, t_i, y_i, z_i, s_i\}_{i=1}^n$ sampled from a super population $\mathcal{P}$, where n denotes the number of units and $s_i \in \{0, 1\}$ indicates whether a unit is selected into the sample, i.e., whether the value of its outcome can be observed. For a unit i with $s_i = 1$, we observe its treatment $t_i \in \mathcal{T}$ that can be continuous or binary, outcome $y_i \in \mathcal{Y}$, instrumental variable $z_i \in \mathcal{Z}$, and covariates $\mathbf{x}_i \in \mathcal{X}$. For a unit i with $s_i = 0$, we only observe $t_i, \mathbf{x}_i$, and z_i, while the value of y_i is missing. Our goal is to estimate $\mathbb{E}[Y \mid do(T = t), \mathbf{X}]$ (Pearl, 2009). However, because of latent confounding and collider bias, we can only estimate $\mathbb{E}[Y \mid \mathbf{X}, T, S = 1]$, which is generally different from $\mathbb{E}[Y \mid do(T = t), \mathbf{X}]$.

Following previous works, we make the following assumptions throughout this paper (Imbens & Rubin, 2015):

Assumption 3.1. Stable Unit Treatment Value Assumption. The distribution of the potential outcome of one unit is assumed to be independent of the treatment assignment of another unit.

Assumption 3.2. Overlap Assumption. A unit has a nonzero probability of being treated and selected, i.e., $0 < \mathbb{P}(T = 1 \mid \mathbf{X} = \mathbf{x}) < 1$ and $0 < \mathbb{P}(S = 1 \mid \mathbf{X} = \mathbf{x}) < 1$.

3.2. Preliminaries of the Instrumental Variable Approach

In this paper, we utilize the Instrumental Variable (IV) defined as follows (Angrist et al., 1996).

Definition 3.3. Instrumental Variable. An instrumental variable Z satisfies the following conditions: (1) Z is a cause of T, i.e., $\mathbb{P}(T \mid Z) \neq \mathbb{P}(T)$; (2) Z is not a direct cause of Y, i.e., $Z \perp\!\!\!\perp Y \mid \mathbf{X}, \mathbf{U}, T$; (3) Z is independent of $\mathbf{U}$ and $\mathbf{X}$, i.e., $Z \perp\!\!\!\perp \mathbf{U}, \mathbf{X}$.

The identification of IV approaches rely on the following additive model assumption (Newey & Powell, 2003; Heckman et al., 2006; Hernán & Robins, 2006; Terza et al., 2008; Imbens & Rubin, 2015; Hansen, 2022).

Assumption 3.4. Additive Model Assumption. The association between the IV and the treatment and between the treatment and the outcome follow additive noise models, i.e., $T = g(\mathbf{X}, Z) + e_t(\mathbf{U})$ and $Y = f(\mathbf{X}, T) + e_y(\mathbf{U})$, where $e_t(\mathbf{U})$ and $e_y(\mathbf{U})$ denote the noise term, a function of the latent variables $\mathbf{U}$.

Under Assumption 3.4, we can conduct a two-stage regression to address latent confounding. In the first stage, we regress T on $\mathbf{X}$ and Z and obtain the estimated T that is independent of $\mathbf{U}$. In the second stage, we regress Y on $\mathbf{X}$ and the estimated T that avoids the influence of $\mathbf{U} \to T$.

3.3. Preliminaries of the Shadow Variable Approach

Previous studies propose to use shadow variables to address collider bias under the assumption that there are no unmeasured confounders. The shadow variable is defined as follows (d'Haultfoeuille, 2010; Miao et al., 2024):

Definition 3.5. Shadow Variable. A shadow variable V needs to satisfy the following conditions: (1) V is related to Y conditional on $\mathbf{X}$ and T in the $S = 1$ samples, i.e., $V \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S = 1$; (2) V is not a direct cause of S, i.e., $V \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$.

With the help of a shadow variable, we can estimate $\mathbb{E}[Y \mid \mathbf{X}, T, V, S = 1]$ with the $S = 1$ samples and generalize it to the $S = 0$ samples by the following equation (Miao & Tchetgen Tchetgen, 2016; Miao et al., 2024):

$$\tau_0(\mathbf{X}, T, V) = \frac{\mathrm{OR}(\mathbf{X}, T, Y) \cdot \tau_1(\mathbf{X}, T, V)}{\mathbb{E}[\mathrm{OR}(\mathbf{X}, T, Y) \mid \mathbf{X}, T, V, S = 1]}, \quad (1)$$

where $\tau_0(\mathbf{X}, T, V)$ is equal to $\mathbb{E}[Y \mid \mathbf{X}, T, V, S = 0]$, $\tau_1(\mathbf{X}, T, V)$ is equal to $\mathbb{E}[Y \mid \mathbf{X}, T, V, S = 1]$, and $\mathrm{OR}(\mathbf{X}, T, Y)$ is the odds ratio function defined as

$$\mathrm{OR}(\mathbf{X}, T, Y) = \frac{\mathbb{P}(S = 0 \mid \mathbf{X}, T, Y) \cdot \mathbb{P}(S = 1 \mid \mathbf{X}, T, Y = 0)}{\mathbb{P}(S = 0 \mid \mathbf{X}, T, Y = 0) \cdot \mathbb{P}(S = 1 \mid \mathbf{X}, T, Y)}.$$

Note that $Y = 0$ is used as a reference value and can also be other values within the support of Y. Equation (1) shows that the key challenge of collider bias is that $\mathbb{E}[Y \mid \mathbf{X}, T, V, S = 0]$ is unidentifiable. This challenge can be addressed by using shadow variables through integrating the odds ratio function over the distribution of the $S = 1$ samples. Since $\mathbb{E}[Y \mid \mathbf{X}, T, V, S = 1]$ can be obtained from the fully observed $S = 1$ samples, the only problem is identifying the odds ratio function. The identification can be guaranteed by the following theorem.

Condition 3.6. *(Miao et al., 2024) For all square-integrable functions $h(\mathbf{X}, T, Y)$, $\mathbb{E}[h(\mathbf{X}, T, Y) \mid \mathbf{X}, \mathbf{Z}, T, S = 1] = 0$ almost surely if and only if $h(\mathbf{X}, T, Y) = 0$ almost surely.*

Theorem 3.7. *(Miao et al., 2024) If V satisfies the conditions in Definition 3.5, let $\widetilde{\mathrm{OR}}(\mathbf{X}, T, Y)$ denote $\mathrm{OR}(\mathbf{X}, T, Y)/\mathbb{E}[\mathrm{OR}(\mathbf{X}, T, Y) \mid \mathbf{X}, T, S = 1]$ and $\widetilde{\mathrm{OR}}_1(\mathbf{X}, T, Y)$ denote $\mathbb{E}[\widetilde{\mathrm{OR}}(\mathbf{X}, T, Y) \mid \mathbf{X}, T, V, S = 1]$, we have*

$$\mathrm{OR}(\mathbf{X}, T, Y) = \frac{\widetilde{\mathrm{OR}}(\mathbf{X}, T, Y)}{\widetilde{\mathrm{OR}}(\mathbf{X}, T, Y = 0)} \quad (2)$$

and

$$\widetilde{\mathrm{OR}}_1(\mathbf{X}, T, Y) = \frac{\mathbb{E}[V \mid \mathbf{X}, T, S = 0]}{\mathbb{E}[V \mid \mathbf{X}, T, S = 1]}. \quad (3)$$

Under Condition 3.6, Equation (3) has a unique solution. Consequently, $\tau_0(\mathbf{X}, T, V)$ and $\mathrm{OR}(\mathbf{X}, T, Y)$ can be identified, and thus $\mathbb{E}[Y \mid \mathbf{X}, T, V]$ can be identified.

Theorem 3.7 indicates that with $\mathbb{E}[V \mid \mathbf{X}, T, S = 0]$ and $\mathbb{E}[V \mid \mathbf{X}, T, S = 1]$ obtained from the observed data, we can obtain $\widetilde{\mathrm{OR}}(\mathbf{X}, T, Y)$ by Equation (3) and identify $\mathrm{OR}(\mathbf{X}, T, Y)$ by Equation (2). As $\tau_1(\mathbf{X}, T, V)$ and $\mathrm{OR}(\mathbf{X}, T, Y)$ can both be identified, $\tau_0(\mathbf{X}, T, V)$ can also be identified by Equation (1).

Note that in our problem setting, there is no well-defined shadow variable because any measured variable can be a direct cause of S. Meanwhile, because of the latent confounding problem, shadow variable approaches are not applicable since the assumption that there are no unmeasured confounders is violated.

4. Two-Stage Shadow Inclusion

4.1. Motivation

Under both latent confounding and collider bias, IV approaches are not applicable because of the two problems stated in Section 2:

- The basic idea of IV approaches is to leverage the estimated T by Z that is independent of $\mathbf{U}$ to eliminate the spurious correlations caused by $T \leftarrow \mathbf{U} \rightarrow Y$. However, when there is also collider bias, the other spurious correlations introduced by conditioning on S also bias the estimate.
- We can only estimate (biased) $\mathbb{E}[Y \mid \mathbf{X}, T, S = 1]$, which is different from $\mathbb{E}[Y \mid \mathbf{X}, T]$ because collider bias makes $Y \not\perp\!\!\!\perp S \mid \mathbf{X}, T$.

To address the above problems of IV methods and make causal inference under both biases possible, we propose to leverage an IV to accomplish two objectives simultaneously: (1) Addressing the spurious correlations introduced by $T \leftarrow \mathbf{U} \rightarrow Y$, and (2) automatically constructing a shadow variable and use it to generalize $\mathbb{E}[Y \mid \mathbf{X}, T, S = 1]$ to $\mathbb{E}[Y \mid \mathbf{X}, T]$, such that the spurious correlations introduced by conditioning on S is also avoided.

Throughout this paper, we make the following additional assumption.

Assumption 4.1. Unconfounded Sample Selection Assumption. The unmeasured confounders are not direct causes of the sample selection, i.e., $S \perp\!\!\!\perp \mathbf{U} \mid \mathbf{X}, T, Y, Z$.

Assumption 4.1 indicates that except for the unmeasured $\mathbf{U}$, the measured variables $\mathbf{X}, T, Y$, and Z can all be direct causes of S, as shown in Figure 3. Note that Assumption 4.1 only requires that $\mathbf{U}$ does not influence S, but *does not require* that all of $\mathbf{X}, T, Y$, and Z influence S. Instead, this assumption implies that our method can handle the complex scenarios where they all influence S. However, if some of them do not affect S, our method remains effective. The assumption is reasonable in many real-world scenarios because data collectors usually select data based on what they can observe. For example, in scenarios where selection is driven by data processors based on existing information (as exemplified earlier in this paper), $\mathbf{U}$ only indirectly affects S through the observed variables like $\mathbf{X}$.

Under Assumptions 3.1, 3.2, 3.4, and 4.1, we can use the IV Z in a two-stage regression way to address both latent confounding and collider bias.

- In the first stage, we regress T on $\mathbf{X}$ and Z and obtain the residual R in this regression. Simultaneously, we learn a decomposed representation Φ of Z that is conditional independent of T and is associated with S by regressing S on $\mathbf{X}, T$, and Φ.
- In the second stage, we incorporate R and Φ in the outcome regression process, using R as both a proxy for $\mathbf{U}$ and a shadow variable, i.e., we regress Y on $\mathbf{X}, T, \Phi$, and R.

The detailed explanation is as follows.

Proposition 4.2. *Under Assumption 3.4, it is possible to consistently estimate $f(\mathbf{X}, T)$ by including R and addressing the collider bias in the second stage.*

Proof. In the first stage, because $\mathbf{X}, T$, and Z are all fully measured, we can regress T on $\mathbf{X}$ and Z using both $S = 1$ and $S = 0$ data, which ensures that the first-stage regression is not conditional on S. Hence, the first-stage regression is not affected by collider bias. We can consistently estimate $g(\mathbf{X}, Z)$ in the first-stage regression and the residual R is a consistent estimator of $e_t(\mathbf{U})$. Then following the same argument in (Terza et al., 2008), by including R and properly addressing the collider bias in the second stage, it is possible to remove the bias from the term $e_y(\mathbf{U})$ and consistently estimate $f(\mathbf{X}, T)$.

□

Next we show how to address the collider bias. Our proposed approach is to learn a decomposed representation Φ of Z that is conditional independent of T and is associated with S during the first stage. Then the next proposition shows that by conditioning on Φ, $\mathbf{X}$, and T, the residual R satisfies the definition of shadow variable. Therefore, it is possible to address the collider bias by using R and Φ.

Proposition 4.3. *Conditional on Φ satisfying that $\Phi \not\perp\!\!\!\perp S$ and $\Phi \perp\!\!\!\perp T \mid Z$, the residual R is a shadow variable that satisfies all the conditions in Definition 3.5.*

Proof. Since $\Phi \not\perp\!\!\!\perp S$ and $\Phi \perp\!\!\!\perp T \mid Z$, we have $R \perp\!\!\!\perp S \mid \mathbf{X}, T, Y, \Phi$ and $R \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, \Phi, S = 1$ (because of $R \leftarrow \mathbf{U} \rightarrow Y$). Therefore, R is a shadow variable that satisfies the conditions in Definition 3.5. □

Based on Propositions 4.2 and 4.3, including R and Φ in the second stage can not only address the latent confounding bias, but also the collider bias. Thus, $f(\mathbf{X}, T)$ can be consistently estimated, as shown in the following proposition.

Proposition 4.4. *Under Assumptions 3.1, 3.2, 3.4, and 4.1, we can consistently estimate $f(\mathbf{X}, T)$ under latent confounding and collider bias by including R and Φ in the second stage.*

Proof. As stated in the proof of Proposition 4.2, the first-stage regression is not conditional on S and thus is not collider-biased. Meanwhile, latent confounding does not affect the estimation of the parameters of Z and $\mathbf{X}$ in the first-stage regression. Therefore, the first-stage regression is unbiased even when these two biases exist.

Therefore, including both R and Φ in the second stage can not only address the spurious correlations caused by $T \leftarrow \mathbf{U} \rightarrow Y$ through blocking $\mathbf{U} \rightarrow T$ by R, but also address the spurious correlations introduced by conditioning on S through using R as a shadow variable. As a result, with

the help of the IV, we can construct the shadow variable R and consistently estimate $E[Y \mid \mathbf{X}, T, \Phi, R]$ under latent confounding and collider bias. □

4.2. Implementation

Based on the above motivation, we propose a novel IV approach that simultaneously addresses latent confounding and collider bias, namely Two-Stage Shadow Inclusion (2SSI). Following previous works (Hartford et al., 2017; Xu et al., 2021), we utilize deep neural networks to learn deep features of the instruments, treatments, and covariates, which allows us to fit highly nonlinear basis functions.

The first-stage regression. In the first stage of 2SSI, we regress T on the fully observed $\mathbf{X}$ and Z with both $S = 1$ and $S = 0$ data by a treatment prediction function $f_t(\mathbf{X}, Z)$. Meanwhile, we also learn a representation Φ of Z satisfying that $\Phi \not\perp\!\!\!\perp S$ and $\Phi \perp\!\!\!\perp T \mid Z$ by a representation function $f_\Phi(Z)$ and a selection indicator prediction function $f_s(\mathbf{X}, T, f_\Phi(Z))$. The loss function is

$$\begin{aligned}\mathcal{L}_t = &\frac{1}{n}\sum_{i=1}^{n}(t_i - f_t(\mathbf{x}_i, z_i))^2 \\ &- \frac{1}{n}\sum_{i=1}^{n}\big[s_i \cdot \log(f_s(\mathbf{x}_i, t_i, f_\Phi(z_i))) \\ &+ (1 - s_i) \cdot \log(1 - f_s(\mathbf{x}_i, t_i, f_\Phi(z_i)))\big] \\ &+ \lambda \cdot \mathrm{disc}(T, f_\Phi(Z)),\end{aligned}$$

where λ is a hyperparameter and $\mathrm{disc}(T, f_\Phi(Z))$ denotes the distributional discrepancy measurement used for making Φ independent of T conditional Z. Following previous works (Shalit et al., 2017; Hassanpour & Greiner, 2020; Wu et al., 2022), for binary treatments, we use the Integral Probability Metric (IPM) to minimize the distance between $\mathbb{P}(f_\Phi(Z) \mid T = 1)$ and $\mathbb{P}(f_\Phi(Z) \mid T = 0)$, i.e., $\mathrm{disc}(T, f_\Phi(Z)) = \mathrm{IPM}(\{f_\Phi(z_i)\}_{i:t_i=1}, \{f_\Phi(z_i)\}_{i:t_i=0})$. For continuous treatments, we use Contrastive Log-ratio Upper Bound (CLUB) (Cheng et al., 2020) to minimize the Mutual Information (MI) of Φ and T, i.e., $\mathrm{disc}(T, f_\Phi(Z)) = \mathrm{MI}(T, f_\Phi(Z))$. The second and third lines in $\mathcal{L}_t$ are the likelihood to predict S using $\mathbf{X}$, T, and $f_\Phi(Z)$, ensuring that the learned Φ is associated with S, while being conditionally independent of T given Z (from $\mathrm{disc}(T, f_\Phi(Z))$).

Subsequently, we obtain the residuals R of T by $R = T - \hat{T}$, where $\hat{T}$ is the predicted value of T. Note that for binary treatments, a generalized version of residual (Gourieroux et al., 1987) is a viable option.

The second-stage regression. In the second stage of 2SSI, we incorporate R and Φ into the regression of Y. Because the value of Y for the $S = 0$ samples are missing, we first learn a selected outcome prediction function $f_{y_1}(\mathbf{X}, T, \Phi, R)$ to estimate $\mathbb{E}[Y \mid \mathbf{X}, T, \Phi, R, S = 1]$ with the following loss function

$$\mathcal{L}_{y_1} = \frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - f_{y_1}(\mathbf{x}_i, t_i, f_\Phi(z_i), r_i))^2,$$

where n_1 denotes the number of $S = 1$ units in $\mathbb{D}$. This estimation avoids the spurious correlations caused by $T \leftarrow \mathbf{U} \rightarrow Y$ because R blocks this path. However, it is still biased due to the other spurious correlations introduced by the collider bias. Therefore, we propose to use R as a shadow variable to address this problem.

Based on Theorem 3.7, we separately use functions $f_{r_1}(\mathbf{X}, T, \Phi)$ and $f_{r_0}(\mathbf{X}, T, \Phi)$ to estimate $\mathbb{E}[R \mid \mathbf{X}, T, \Phi, S = 1]$ and $\mathbb{E}[R \mid \mathbf{X}, T, \Phi, S = 0]$ with $S = 1$

Algorithm 1 Two-Stage Shadow Inclusion

Input: $\mathbb{D} = \{\mathbf{x}_i, t_i, y_i, z_i, s_i\}_{i=1}^n$, λ, mini-batch sizes $m_1, m_{2(\mathrm{A})}, m_{2(\mathrm{B})}$, number of updates e_1, e_2.
Output: The estimated conditional expectation of Y on the target population, i.e., $\mathbb{E}[Y \mid \mathbf{X}, T, \Phi, R]$.
Initialize parameters in f_Φ, f_s, f_t, f_{y_1}, f_{r_1}, f_{r_0}, $f_{\widetilde{\mathrm{or}}}$, f_{y_0}, f_p.
repeat
  Sample m_1 units from $\mathbb{D}$ as Batch 1.
  for $j = 1$ to e_1 **do**
    Optimize f_Φ, f_s, and f_t by $\mathcal{L}_t$ using Batch 1.
  end for
  $R \leftarrow T - f_t(\mathbf{X}, Z)$.
  Sample $m_{2(\mathrm{A})}$ units with $S = 1$ as Batch 2(A).
  Sample $m_{2(\mathrm{B})}$ units with $S = 0$ as Batch 2(B).
  for $j = 1$ to e_2 **do**
    Optimize f_p by $\mathcal{L}_p$ using Batch 1.
    Optimize f_{y_1} by $\mathcal{L}_{y_1}$ using Batch 2(A).
    Optimize f_{r_1} by $\mathcal{L}_{r_1}$ using Batch 2(A).
    Optimize f_{r_0} by $\mathcal{L}_{r_0}$ using Batch 2(B).
  end for
  for $j = 1$ to e_2 **do**
    Optimize $f_{\widetilde{\mathrm{or}}}$ by $\mathcal{L}_{\widetilde{\mathrm{or}}}$ using Batch 2(A).
  end for
  Calculate $\mathrm{OR}(\mathbf{X}, T, \Phi, Y)$ and $\tilde{Y}$ in Batch 2(A) by Equations (1) and (3).
  for $j = 1$ to e_2 **do**
    Optimize f_{y_0} by $\mathcal{L}_{y_0}$ using Batch 2(A).
  end for
until convergence
$\mathbb{E}[Y \mid \mathbf{X}, T, \Phi, R] \leftarrow f_{y_0}(\mathbf{X}, T, \Phi, R) \cdot (1 - f_p(\mathbf{X}, T, \Phi, R)) + f_{y_1}(\mathbf{X}, T, \Phi, R) \cdot f_p(\mathbf{X}, T, \Phi, R)$.
return $\mathbb{E}[Y \mid \mathbf{X}, T, \Phi, R]$

and $S = 0$ data. The loss functions are

$$\mathcal{L}_{r_1} = \frac{1}{n_1}\sum_{i=1}^{n_1}(r_i - f_{r_1}(\mathbf{x}_i, t_i, f_\Phi(z_i)))^2,$$
$$\mathcal{L}_{r_0} = \frac{1}{n_0}\sum_{i=1}^{n_0}(r_i - f_{r_0}(\mathbf{x}_i, t_i, f_\Phi(z_i)))^2,$$

where n_0 denotes the number of $S = 0$ units in $\mathbb{D}$. Following this, we can obtain $\widetilde{\mathrm{OR}}_1(\mathbf{X}, T, \Phi, Y)$ by Equation (3) and further estimate $\widetilde{\mathrm{OR}}(\mathbf{X}, T, \Phi, Y)$ using a function $f_{\widetilde{\mathrm{or}}}(\mathbf{X}, T, \Phi, Y)$. The loss function is

$$\mathcal{L}_{\widetilde{\mathrm{or}}} = \frac{1}{n_1}\sum\nolimits_{i:s_i=1}(f_{\widetilde{\mathrm{or}}}(\mathbf{x}_i, t_i, f_\Phi(z_i), y_i) - f_{r_0}(\mathbf{x}_i, t_i, f_\Phi(z_i))/f_{r_1}(\mathbf{x}_i, t_i, f_\Phi(z_i)))^2,$$

Next, we get $\mathrm{OR}(\mathbf{X}, T, \Phi, Y)$ in Equation (2) and the counterfactual $S = 0$ outcomes of the $S = 1$ samples in Equation (1), denoted as $\tilde{Y}$, i.e., $\tilde{Y} = f_{\widetilde{\mathrm{or}}}(\mathbf{X}, T, \Phi, Y)/f_{\widetilde{\mathrm{or}}}(\mathbf{X}, T, \Phi, f_{y_1}(\mathbf{X}, T, \Phi, R)) \cdot f_{y_1}(\mathbf{X}, T, \Phi, R)$. Now, we can learn another unselected outcome prediction function $f_{y_0}(\mathbf{X}, T, \Phi, R)$ to estimate $\mathbb{E}[Y \mid \mathbf{X}, T, \Phi, R, S = 0]$, with the loss function being

$$\mathcal{L}_{y_0} = \frac{1}{n_1}\sum_{i=1}^{n_1}(\tilde{y}_i - f_{y_0}(\mathbf{x}_i, t_i, f_\Phi(z_i), r_i))^2.$$

All the above learning processes in the second stage share the same features of $\mathbf{X}$, T, Φ, and R and can thus be jointly optimized. To finally obtain $\mathbb{E}[Y \mid \mathbf{X}, T, \Phi, R]$, we also learn a sample selection function $f_p(\mathbf{X}, T, \Phi, R)$ that estimates $\mathbb{P}(S = 1 \mid \mathbf{X}, T, \Phi, R)$ with the loss function being

$$\mathcal{L}_p = -\frac{1}{n}\sum_{i=1}^{n}(s_i \cdot \log(f_p(\mathbf{x}_i, t_i, f_\Phi(z_i), r_i)) + (1 - s_i) \cdot \log(1 - f_p(\mathbf{x}_i, t_i, f_\Phi(z_i), r_i))).$$

Following this, we simultaneously address latent confounding and collider bias and estimate $\mathbb{E}[Y \mid \mathbf{X}, T, \Phi, R]$ by

$$\mathbb{E}[Y \mid \mathbf{X}, T, \Phi, R] = f_{y_0}(\mathbf{X}, T, \Phi, R) \cdot (1 - f_p(\mathbf{X}, T, \Phi, R)) + f_{y_1}(\mathbf{X}, T, \Phi, R) \cdot f_p(\mathbf{X}, T, \Phi, R).$$

The pseudo-codes are in Algorithm 1, and the source code is available at https://github.com/ZJUBaohongLi/2SSI.

5. Experiments

5.1. Baselines

We compare the proposed 2SSI with two groups of methods. One group is **IV approaches for addressing latent confounding**, including (1) 2SRI (Terza et al., 2008), (2) DeepIV (Hartford et al., 2017), (3) Kernel IV (Singh et al., 2019), (4) DeepGMM (Bennett et al., 2019), (5) DFIV (Xu et al., 2021), (6) CB-IV (Wu et al., 2022). The other group is **approaches for addressing non-random sample selection**, including (1) Heckit (Heckman, 1979), (2) IPSW (Cole & Stuart, 2010), and (3) Shadow variable estimation (SHADOW) (Miao & Tchetgen Tchetgen, 2016).

Because there is no shadow variable available in our problem setting, we used Z as a shadow variable to implement SHADOW. Specifically, with the help of the decomposed representation Φ learned from the first-stage regression of 2SSI, Z satisfies that $Z \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, \Phi, S = 1$ and $Z \perp\!\!\!\perp S \mid \mathbf{X}, T, Y, \Phi$. Therefore, SHADOW can be regarded as an ablation version of 2SSI without incorporating the residual into the second-stage regression. Note that 2SRI can be regarded as another ablation version of 2SSI without learning Φ in the first-stage regression and incorporating it to make the residual a shadow variable in the second-stage regression.

5.2. Experiments on Synthetic Data

5.2.1. Datasets

Following previous works (Hartford et al., 2017; Xu et al., 2021; Wu et al., 2022), we use the benchmark datasets of IV studies, i.e., **Demand datasets**, to evaluate the performance of 2SSI and the baselines. We adopted the same data generation process (DGP) following previous works (Hartford et al., 2017; Xu et al., 2021) with minor changes because the original datasets are not collider-biased. That is, we additionally generated a selection indicator S that satisfies Assumption 4.1 to introduce collider bias into Demand datasets. We also introduced two additional parameters, i.e., α and β, into the DGP as measurements of the collider bias and latent confounding strengths (the bias is stronger when the value is larger), respectively. We conducted experiments on the low-dimensional and high-dimensional settings of Demand datasets. The detailed description and DGP of the collider-biased Demand datasets is in Appendix C.

5.2.2. Results

Following previous works (Hartford et al., 2017; Xu et al., 2021; Wu et al., 2022), we use the Mean Square Error (MSE) as the evaluation metric. To clearly evaluate the performance of these methods under collider bias, we separately report the results on $S = 1$ and $S = 0$ samples. Under the low-dimensional setting of Demand datasets, we changed the collider bias strength α and latent confounding strength β in the experiments to test the robustness of 2SSI and the baselines as the biases strengthen. That is, we first fixed $\beta = 10$ and conducted experiments with $\alpha = \{5, 10, 15\}$ and then fixed $\alpha = 10$ and conducted experiments with

Table 1. Out-of-sample MSE (mean ± std) on Demand datasets under different collider bias strengths α with a fixed latent confounding strength $\beta = 10$. The results in the table are scaled by a factor of 10^3 for clarity. The best results are in bold.

	$\alpha = 5$		$\alpha = 10$		$\alpha = 15$	
ESTIMATOR	$S = 1$ DATA	$S = 0$ DATA	$S = 1$ DATA	$S = 0$ DATA	$S = 1$ DATA	$S = 0$ DATA
HECKIT	0.648±0.038	5.685±0.324	0.724±0.060	6.741±0.359	0.714±0.057	6.763±0.507
2SRI	0.998±0.070	6.076±0.262	1.052±0.097	6.673±0.219	1.016±0.063	6.799±0.363
IPSW	1.075±0.082	12.72±1.070	1.063±0.093	12.39±0.656	1.056±0.085	12.45±0.442
SHADOW	0.852±0.083	4.619±0.269	0.956±0.116	5.233±0.289	0.957±0.088	5.258±0.449
DEEPIV	0.630±0.048	12.06±0.832	0.633±0.051	12.46±0.424	0.641±0.053	12.57±1.007
KERNEL IV	0.297±0.097	6.450±0.367	0.400±0.175	6.995±0.526	0.428±0.109	7.220±0.903
DEEPGMM	0.655±0.105	8.138±1.268	0.688±0.108	8.399±1.555	0.699±0.104	9.163±1.419
DFIV	0.572±0.090	12.72±0.686	0.580±0.106	12.97±1.667	0.620±0.052	13.60±1.181
CB-IV	1.202±0.228	7.910±0.482	1.270±0.225	8.442±0.574	1.343±0.153	8.771±0.712
2SSI	**0.147±0.020**	**1.320±0.335**	**0.154±0.016**	**1.278±0.155**	**0.155±0.024**	**1.275±0.227**

Table 2. Out-of-sample MSE (mean ± std) on Demand datasets under different latent confounding strengths β with a fixed collider bias strength $\alpha = 10$. The results in the table are scaled by a factor of 10^3 for clarity. The best results are in bold.

	$\beta = 5$		$\beta = 10$		$\beta = 15$	
ESTIMATOR	$S = 1$ DATA	$S = 0$ DATA	$S = 1$ DATA	$S = 0$ DATA	$S = 1$ DATA	$S = 0$ DATA
HECKIT	0.089±0.007	6.008±0.230	0.724±0.060	6.741±0.359	1.819±0.122	9.490±0.778
2SRI	0.114±0.011	3.317±0.155	1.052±0.097	6.673±0.219	2.918±0.251	10.88±0.452
IPSW	0.136±0.035	5.588±0.777	1.063±0.093	12.39±0.656	3.150±0.396	24.65±2.918
SHADOW	0.099±0.009	3.220±0.149	0.956±0.116	5.233±0.289	2.512±0.260	7.350±0.520
DEEPIV	0.065±0.007	4.697±0.278	0.633±0.051	12.46±0.424	1.945±0.117	24.97±1.596
KERNEL IV	0.052±0.015	3.734±0.182	0.400±0.175	6.995±0.526	1.759±0.747	12.07±0.611
DEEPGMM	0.040±0.009	3.893±0.443	0.688±0.108	8.399±1.555	2.445±0.396	13.59±2.546
DFIV	0.072±0.008	5.056±0.435	0.580±0.106	12.97±1.667	1.809±0.302	28.27±1.978
CB-IV	0.099±0.021	3.737±0.276	1.270±0.225	8.442±0.574	4.031±0.412	14.46±0.615
2SSI	**0.038±0.020**	**1.995±0.088**	**0.154±0.016**	**1.278±0.155**	**0.522±0.411**	**2.655±0.630**

$\beta = \{5, 10, 15\}$. For each setting, we randomly sampled $10,000$ units and performed 20 replications to report the mean and the standard deviation (std) of the MSE. The results are reported in Tables 1 and 2.

From the results, we observe that: (1) For all methods, the performance on $S = 0$ data is much worse than that on $S = 1$ data, which proves the harm of the collider bias problem. (2) The methods for addressing non-random sample selection caused by T and $\mathbf{X}$, i.e., Heckit and IPSW, perform poorly in all settings, and the performance gets worse as the latent confounding strength β gets larger. The reason is that these methods cannot address either collider bias or latent confounding. (3) SHADOW, as an ablation version of 2SSI, performs better than Heckit and IPSW because it is designed to address collider bias. However, its performance is still much worse than 2SSI since it cannot address the latent confounding problem. This observation also demonstrates the necessity of using the residual as a proxy for $\mathbf{U}$ in the second-stage regression. (4) The performance of all the IV baselines is bad and worsens as the collider bias strength α increases because they all suffer from collider bias. (5) Moreover, 2SRI, as an ablation version of 2SSI, also performs worse than 2SSI. It demonstrates the necessity of decomposed representation learning in the first-stage regression and using the residual as a shadow variable in the second-stage regression. (6) 2SSI achieves the best performance under all settings of α and β, and the performance remains robust in the face of varying strengths of both biases. An interesting observation is that the performance under $\beta = 5$ is worse than that under $\beta = 10$. The reason is that if the latent confounding is too weak, the conditional dependence between R and Y also gets weak. Therefore, just as most IV approaches need, before conducting an IV analysis, we can first analyze how much latent confounding there is and whether using IVs is needed by scientific consideration and statistical tests (Baiocchi et al., 2014). These observations prove the effectiveness of 2SSI in addressing both biases.

We also conducted experiments on Demand datasets under high-dimensional settings, as well as ablation experiments studying the first-stage decomposed representation learning module. The results and observations can be found in Appendix A.

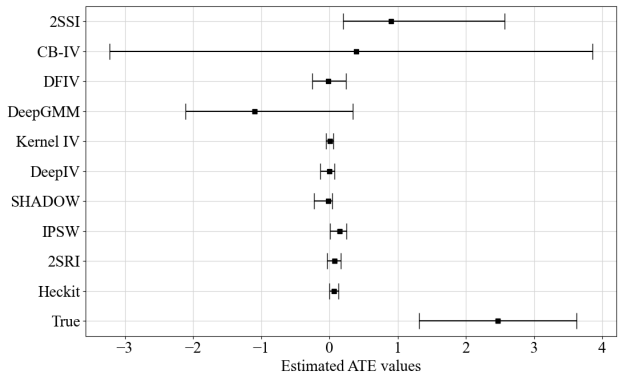


Figure 5. The ATE estimation results on the Fertil2 dataset, where True denotes the reference for the ground truth ATE reported in Wooldridge (2010).

5.3. Experiments on Real-World Data

5.3.1. Dataset

To evaluate the performance of different methods in real-world scenarios where both latent confounding and collider bias exist, we conducted experiments on a real-world dataset with the latent confounding problem, i.e., **Fertil2 dataset** (Wooldridge, 2010). This dataset has a well-defined IV. To introduce collider bias into the Fertil2 dataset, we non-randomly selected a subset of samples from the original dataset. Since the treatment is binary in the Fertil2 dataset, the goal is to estimate the Average Treatment Effect (ATE) on the population of the original Fertil2 dataset using only $S = 1$ samples. ATE is defined as $\mathbb{E}[Y \mid do(T = 1), \mathbf{X}] = \mathbb{E}[Y \mid do(T = 0), \mathbf{X}]$. The detailed description of the collider-biased Fertil2 dataset is in Appendix C.

5.3.2. Results

We aim to compare the estimators' performance of using the collider-biased samples with latent confounding to estimate the ATE on the target population, i.e., the population of the original Fertil2 dataset without collider bias. Following previous studies (Ding et al., 2017), we regarded the ATE estimation results in Wooldridge (2010) as a reference for the ground truth ATE. Therefore, we can tell an estimator is better when facing both latent confounding and collider bias if the result is more similar to this reference. To further evaluate the robustness of the estimators, we performed ten replications for each estimator. We report the mean values and error bars of the estimated in Figure 5.

From the results, we observe that: (1) The ground truth ATE on the target population is supposed to be positive, but the estimated ATE range of most baselines is close to zero, and the mean values of DeepGMM, SHADOW, and DFIV are even negative. It demonstrates that these methods cannot simultaneously address latent confounding and collider bias in treatment effect estimation. (2) Although CB-IV sometimes obtained results similar to the ground truth, it also achieved the worst estimation results in stark contrast to the reference range. Therefore, CB-IV is not applicable to address both biases. (3) 2SSI achieves the best overall performance among all estimators and can consistently achieve positive ATE estimates, demonstrating the effectiveness and robustness of 2SSI for treatment effect estimation under both biases. (4) The standard deviation of estimated ATEs by 2SSI in Fertil2 is larger than those of some baselines, which is not observed in Demand. This observation is reasonable because of the differing methodologies employed for repeated experimentation across the two datasets. Since Demand is a semi-synthetic dataset, we regenerated data based on the DGPs in each experiment. This leads to varying values for T, Y, and S in each experiment, resulting in larger stds across all methods on this dataset. Conversely, Fertil2 is a real-world dataset where T, Y, and S values remain constant across experiments, with variations only in the training and testing set splits. Hence, most methods exhibit smaller stds on this dataset. However, those methods showing lower stds have not escaped the curse of latent confounding and collider bias in each experiment; hence, they consistently show a stable high bias in each experiment. On the contrary, our method shows a relatively higher std (though still very small) because it manages to address the biases, albeit to varying degrees across different experiments.

6. Conclusion

In this paper, we studied the problem of causal inference under both latent confounding and collider bias and proposed a novel IV approach, i.e., 2SSI, to address it. In the first stage of 2SSI, we simultaneously regress the treatment to obtain the residual and regress the selection indicator to learn a decomposed representation. In the second stage, we incorporate the residual and decomposed representation into the outcome regression process and use the residual as both a proxy for the unmeasured confounders and a shadow variable. To the best of our knowledge, it is the first approach developed for simultaneously addressing the two biases. One limitation of our work is that the proposed method requires well-defined IVs, necessitating expert knowledge. Additionally, as most IV approaches need, we should first analyze how much latent confounding there is and whether using IV approaches is necessary. It will be interesting to test whether the large amounts of existing approaches for finding valid IVs and testing latent confounding are still available when there is also collider bias, which is left to future work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62376243, 62441605, U20A20387), and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Angrist, J. D. and Krueger, A. B. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4):69–85, 2001.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455, 1996.

Baiocchi, M., Cheng, J., and Small, D. S. Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340, 2014.

Bareinboim, E. and Tian, J. Recovering causal effects from selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Bareinboim, E., Tian, J., and Pearl, J. Recovering from selection bias in causal and statistical inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 433–450. ACM, 2022.

Bennett, A., Kallus, N., and Schnabel, T. Deep generalized method of moments for instrumental variable analysis. *Advances in Neural Information Processing Systems*, 2019.

Brito, C. and Pearl, J. Generalized instrumental variables. In *Proceedings of the Conference in Uncertainty in Artificial Intelligence*, 2002.

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pp. 1779–1788. PMLR, 2020.

Cole, S. R. and Stuart, E. A. Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. *American journal of epidemiology*, 172(1): 107–115, 2010.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2023.

Cui, P. and Athey, S. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, 4(2):110–115, 2022.

Ding, P. Bayesian robust inference of sample selection using selection-t models. *Journal of Multivariate Analysis*, 124: 451–464, 2014.

Ding, P., VanderWeele, T., and Robins, J. M. Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*, 104(2):291–302, 2017.

d'Haultfoeuille, X. A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154 (1):1–15, 2010.

Gourieroux, C., Monfort, A., Renault, E., and Trognon, A. Generalised residuals. *Journal of Econometrics*, 34(1): 5–32, 1987.

Greiner, N. H. R. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020.

Griffith, G. J., Morris, T. T., Tudball, M. J., Herbert, A., Mancano, G., Pike, L., Sharp, G. C., Sterne, J., Palmer, T. M., Davey Smith, G., et al. Collider bias undermines our understanding of covid-19 disease risk and severity. *Nature communications*, 11(1):5749, 2020.

Hansen, B. *Econometrics*. Princeton University Press, 2022.

Hartford, J. S., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, volume 70, pp. 1414–1423. PMLR, 2017.

Hassanpour, N. and Greiner, R. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020.

Heckman, J. J. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 47(1): 153–161, 1979.

Heckman, J. J., Urzua, S., and Vytlacil, E. Understanding instrumental variables in models with essential heterogeneity. *The review of economics and statistics*, 88(3): 389–432, 2006.

Hernán, M. A. and Robins, J. M. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, pp. 360–372, 2006.

Hernán, M. A. and Robins, J. M. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, volume 37, pp. 448–456, 2015.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

LeCun, Y. and Cortes, C. Mnist handwritten digit database, 2010. URL http://yann.lecun.com/exdb/mnist/.

Li, W., Miao, W., and Tchetgen Tchetgen, E. Nonparametric inference about mean functionals of nonignorable non-response data without identifying the joint distribution. *J R Stat Soc Series B Stat Methodol*, 85(3): 913–935, 2023.

Miao, W. and Tchetgen Tchetgen, E. J. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016.

Miao, W., Liu, L., Li, Y., Tchetgen Tchetgen, E. J., and Geng, Z. Identification and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *ACM/JMS Journal of Data Science*, 1(2):1–23, 2024.

Newey, W. K. and Powell, J. L. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5): 1565–1578, 2003.

Nilforoshan, H., Gaebler, J. D., Shroff, R., and Goel, S. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pp. 16848–16887. PMLR, 2022.

Ogundimu, E. O. and Hutton, J. L. A sample selection model with skew-normal distribution. *Scandinavian Journal of Statistics*, 43(1):172–190, 2016.

P., H. L. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, 50(4):1029–1054, 1982.

Pearl, J. *Causality*. Cambridge university press, 2009.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, volume 70, pp. 3076–3085. PMLR, 2017.

Shi, X., Miao, W., and Tchetgen, E. T. A selective review of negative control methods in epidemiology. *Current epidemiology reports*, 7:190–202, 2020.

Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Terza, J. V., Basu, A., and Rathouz, P. J. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of health economics*, 27 (3):531–543, 2008.

Wang, H., Fan, J., Chen, Z., Li, H., Liu, W., Liu, T., Dai, Q., Wang, Y., Dong, Z., and Tang, R. Optimal transport for treatment effect estimation. *Advances in Neural Information Processing Systems*, 2023.

Wang, X., Wu, Y., Zhang, A., Feng, F., He, X., and Chua, T.-S. Reinforced causal explainer for graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2297–2309, 2022.

Wiemann, P. F. V., Klein, N., and Kneib, T. Correcting for sample selection bias in bayesian distributional regression models. *Computational Statistics & Data Analysis*, 168: 107382, 2022.

Wooldridge, J. M. *Econometric analysis of cross section and panel data*. MIT press, 2010.

Wu, A., Kuang, K., Li, B., and Wu, F. Instrumental variable regression with confounder balancing. In *International Conference on Machine Learning*, pp. 24056–24075. PMLR, 2022.

Xu, L., Chen, Y., Srinivasan, S., de Freitas, N., Doucet, A., and Gretton, A. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2021.

Zhang, M., Yuan, J., He, Y., Li, W., Chen, Z., and Kuang, K. MAP: Towards balanced generalization of iid and ood through model-agnostic adapters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11921–11931, 2023.

Zhang, M., Li, H., Wu, F., and Kuang, K. Metacoco: A new few-shot classification benchmark with spurious correlation. In *International Conference on Learning Representations*, 2024.

Zhang, W., Liu, L., and Li, J. Treatment effect estimation with disentangled latent factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2021.

A. Supplementary Experiments

A.1. Experiments on Demand Datasets under the High-Dimensional Setting

Under the high-dimensional setting of Demand datasets, we used the same data generation process as in experiments under the low-dimensional setting with $\alpha = 10$ and $\beta = 10$. The results are shown in Table 3. From the results, we observe that: (1) The methods for addressing non-random sample selection caused by T and $\mathbf{X}$, i.e., Heckit and IPSW, perform the worst among all estimators because they cannot address either collider bias or latent confounding and suffer from model misspecification problems. (2) SHADOW performs better than Heckit and IPSW because it is designed to address collider bias. However, its performance is still much worse than 2SSI since it cannot address the latent confounding problem. Meanwhile, SHADOW also performs worse than IV approaches because the latent confounding problem is more significant in high-dimensional settings. Specifically, since $\mathbf{X}$ is high-dimensional and hard to control, the non-causal path $\mathbf{U} \to \mathbf{X} \to T$ also biases the estimation. (3) The performance of all the IV baselines is better than the baselines designed for addressing non-random sample selection. As mentioned earlier, the reason is that the latent confounding problem is more significant in high-dimensional settings. However, these methods still perform worse than 2SSI because they cannot address collider bias. (4) 2SSI achieves the best performance, which proves the effectiveness of 2SSI in addressing both biases under high-dimensional settings.

A.2. Ablation Study of the Decomposed Representation Learning Module of 2SSI

In the first stage of 2SSI, we propose to learn a decomposed representation Φ satisfying that $\Phi \not\perp\!\!\!\perp S$ and $\Phi \perp\!\!\!\perp T \mid Z$, such that the residual R of T is a shadow variable that satisfies all the conditions in Definition 3.5 conditional on Φ. To do so, we propose the loss function of the first-stage regression as

$$\begin{aligned}\mathcal{L}_t = & \frac{1}{n}\sum_{i=1}^{n}(t_i - f_t(\mathbf{x}_i, z_i))^2 \\ & - \frac{1}{n}\sum_{i=1}^{n}(s_i \cdot \log(f_p(\mathbf{x}_i, t_i, f_\Phi(z_i))) \\ & + (1 - s_i) \cdot \log(1 - f_s(\mathbf{x}_i, t_i, f_\Phi(z_i)))) \\ & + \lambda \cdot \mathrm{disc}(T, f_\Phi(Z)),\end{aligned}$$

In this section, we aim to investigate the performance of 2SSI under two scenarios: (1) when Φ is not learned well, i.e., $\lambda = 0$, and (2) when Φ cannot be learned, i.e., Z does not influence S. Therefore, we conducted ablation experiments on Demand datasets (with $\alpha = 10$ and $\beta = 10$) by setting $\lambda = 0$ and by setting the coefficient of Z on S in the data generation process to 0. As shown in Table 4, the first scenario hurts the performance of 2SSI since if Φ is not learned well, the residual R will not satisfy the conditions of a shadow variable. On the contrary, the second scenario does not affect the performance of 2SSI because if Z itself does not directly cause S, there is no need to learn Φ since R already satisfies the conditions of a shadow variable ($R \not\perp\!\!\!\perp Y \mid X, T, S = 1$, and $R \perp\!\!\!\perp S \mid X, T, Y$).

B. Implementation Details

We provide an overview diagram of 2SSI, as shown in Figure 6, to help readers better understand the proposed 2SSI.

We implemented the proposed method and the baselines using Python 3.9 with PyTorch 1.13.0. The hardware used was a Windows 11 operating system with the 13th Gen Intel(R) Core(TM) i7-13700K CPU and NVIDIA GeForce RTX 3080 GPU (with CUDA version 12.1). Following Xu et al. (2021), we used multi-layer perceptrons with ReLU activation function to implement each module of 2SSI. We used the Adam optimizer (Kingma & Ba, 2015) with batch normalization (Ioffe & Szegedy, 2015) in the training process. Following Wu et al. (2022), we implemented the IPM with the Maximum Mean Discrepancy (MMD) metric and the MI with Contrastive Log-ratio Upper Bound (CLUB) (Cheng et al., 2020). The hyperparameters of 2SSI on different datasets are in Table 5. Note that we used the same learning rates and weight decays for all the modules in 2SSI except for the network learning covariate features, of which the weight decay is 0.1.

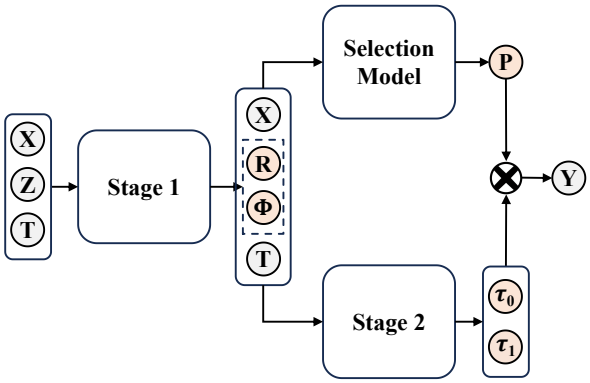


Figure 6. An overview of the proposed method. In the first stage, we construct a shadow variable by learning a decomposed representation Φ from Z and the residual R of the treatment T. In the second stage, we incorporate R and Φ into the regression of Y and use R as both a proxy for $\mathbf{U}$ and a shadow variable to estimate τ_1 and τ_0. Finally, with the help of a selection model that estimates $\mathbb{P}(S = 1 \mid \mathbf{X}, T, \Phi, R)$, we can achieve an unbiased estimate of $E[Y \mid \mathbf{X}, T, \Phi, R]$.

C. Detailed Description of the Datasets

C.1. Demand Datasets

In Demand datasets (Hartford et al., 2017), we aim to estimate the effect of the ticket price (T) on the customer's decision about whether to buy a ticket (Y). There are two measured confounders, i.e., the time of year $X_1 \in [0, 10]$ and customer type $X_2 \in \{1, \cdots, 7\}$ categorized by the levels of price sensitivity. The latent confounding problem is introduced by making the noise term U in Y associated with T. In order to address latent confounding, the cost of fuel (Z) is used as an instrumental variable. Demand datasets also have a high-dimensional setting, in which the customer type X_2 is replaced with the pixels of the corresponding handwritten digit from the MNIST dataset (LeCun & Cortes, 2010). We adopted the same data generation process following previous works (Hartford et al., 2017; Xu et al., 2021) with minor changes because the original datasets are not collider-biased. That is, we additionally generated a selection indicator S that satisfies Assumption 4.1 to introduce collider bias into Demand datasets. We also introduced two additional parameters, i.e., α and β, as measurements of the collider bias and latent confounding strengths, respectively. The detailed data generation process of the collider-biased Demand datasets is as

$$Y = 100 + (10 + T)X_2\psi_t - 2T + U,$$

$$T = 25 + (Z + 3)\psi_t + \beta U + E,$$

and

$$S = \text{Bernoulli}(1 + e^{-T+0.1(X_1+X_2+Z)+\alpha Y}),$$

where

$$\psi_t = 2\left(\frac{(T-5)^4}{600} + e^{-4(T-5)^2} + \frac{T}{10} - 2\right),$$

$Z, U, E \sim \mathcal{N}(0, 1)$, and $\text{Bernoulli}(\cdot)$ denotes the Bernoulli distribution.

C.2. Fertil2 Dataset

The Fertil2 dataset aims to estimate the effect of at least seven years of education for a woman (T) on the number of living children in the family (Y). Several observed confounders are included in the dataset, such as age, religion, the ideal number of children, and whether the woman lived in urban areas. The instrumental variable Z is a binary indicator of whether the woman was born in the first half of the year. To introduce collider bias into the Fertil2 dataset, we non-randomly selected a subset of samples from the original dataset. Specifically, we set $S = 0$ for those whose ideal number of children is larger

Table 3. Out-of-sample MSE (mean ± std) on Demand datasets under the high-dimensional setting with $\alpha = 10$ and $\beta = 10$. The results in the table are scaled by a factor of 10^3 for clarity. The best results are in bold.

ESTIMATOR	$S = 1$ DATA	$S = 0$ DATA
HECKIT	> 10000	> 10000
2SRI	0.869±0.057	8.828±0.506
IPSW	> 10000	> 10000
SHADOW	4.046±8.505	16.54±24.86
DEEPIV	0.496±0.100	8.428±0.391
KERNEL IV	0.730±0.126	6.635±0.871
DEEPGMM	1.562±0.156	8.268±0.393
DFIV	0.498±0.095	9.261±1.038
CB-IV	1.409±0.212	8.223±0.505
2SSI	**0.489±0.033**	**5.680±0.168**

Table 4. Out-of-sample MSE (mean ± std) of Ablation Experiments on Demand datasets. The results in the table are scaled by a factor of 10^3 for clarity.

SCENARIO	$S = 1$ DATA	$S = 0$ DATA
Z DOES NOT DIRECTLY CAUSE S	0.166 ± 0.032	1.251 ± 0.174
$\lambda = 0$	0.168 ± 0.039	1.908 ± 0.943
THE ORIGINAL SETTING	0.154 ± 0.016	1.278 ± 0.155

than the number of living children and those who did not live in urban areas and had less than seven years of education. Intuitively, the former group might subjectively refuse to report their outcomes, while the outcomes of the latter group might be objectively hard to collect.

D. Discussions of the Difference Between 2SSI and Other Related Works

Most previous causal inference studies have developed various techniques to control the measured confounders and shown great success in addressing the confounding bias caused by fully measured covariates (Shalit et al., 2017; Greiner, 2020; Zhang et al., 2021; Wang et al., 2023). However, when facing unmeasured confounders, these methods are not applicable. Therefore, we believe these methods are out of the scope of this paper.

The IV approaches are widely used to address latent confounding. The basis of IV approaches is Two-Stage Least Regression (2SLS), which uses IVs to regress the treatment and utilizes the estimated treatment to regress the outcome under linear settings (Angrist et al., 1996; Angrist & Krueger, 2001; Brito & Pearl, 2002; Baiocchi et al., 2014). In nonlinear scenarios, recent studies utilize machine learning techniques to apply IV approaches to more complex real-world scenarios. DeepIV (Hartford et al., 2017) utilizes deep models to estimate the conditional probability distribution of the treatment in the first stage. Kernel IV (Singh et al., 2019) learns relations among variables in 2SLS as nonlinear functions in reproducing kernel Hilbert spaces (RKHSs). DeepGMM (Bennett et al., 2019) and DFIV (Xu et al., 2021) respectively leverage Generalized Method of Moments (GMM) (P., 1982) and deep networks to learn the nonlinear basis functions as deep features. CB-IV (Wu et al., 2022) is proposed to further balance the deep features of the measure covariates in the second stage. The above IV approaches can only address latent confounding and are not applicable when collider bias exists. This is because collider bias results in missing Y values of $S = 0$ data, and thus the outcome regression process can only be conducted conditional on S, which introduces more spurious correlations than only $T \leftarrow \mathbf{U} \rightarrow Y$.

The shadow variable approaches are developed for addressing collider bias (d'Haultfoeuille, 2010; Miao & Tchetgen Tchetgen, 2016; Li et al., 2023). However, in our problem setting, no shadow variable is available because all measured variables can directly cause Y, violating the conditional independence condition in Definition 3.5. The baseline SHADOW implemented in our experiments is indeed an ablation version of 2SSI because only with the decomposed representation learned by 2SSI can Z be used as a shadow variable.

The most relevant work is the Two-Stage Residual Inclusion (2SRI) approach (Terza et al., 2008), which also introduces the residuals of T from the first-stage regression into the second stage. However, they differ in the following aspects.

Table 5. Hyperparameters of 2SSI on different datasets, where $\text{Demand}_{\text{Low}}$ denotes Demand datasets under the low-dimensional setting and $\text{Demand}_{\text{High}}$ denotes Demand datasets under the high-dimensional setting.

SETTING	$\text{DEMAND}_{\text{Low}}$	$\text{DEMAND}_{\text{High}}$	FERTIL2
EPOCH	100	100	100
LEARNING RATE	0.001	0.001	0.001
WEIGHT DECAY	0.0001	0.01	0.0001
λ	0.1	0.1	0.1

- **The studied problem is different.** In this paper, we propose a novel IV method for causal inference under both latent confounding and collider bias, while 2SRI is only applicable for addressing the latent confounding problem solely and cannot address collider bias.
- **The regression process of the first stage is different.** 2SRI only conducts the regression on T to obtain R in the first stage, while the proposed method additionally conducts the regression on S to learn a decomposed representation Φ of Z.
- **The regression process of the second stage is different.** 2SRI simply incorporates R into the outcome regression process, while the proposed method incorporates R and Φ into the regression and uses R as a shadow variable to address collider bias further.