Instrumental Variables in Causal Inference and Machine Learning: A Survey

Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Fei Wu, *Senior Member, IEEE*

Abstract—Causal inference is the process of using assumptions, study designs, and estimation strategies to draw conclusions about the causal relationships between variables based on data. This allows researchers to better understand the underlying mechanisms at work in complex systems and make more informed decisions. In many settings, we may not fully observe all the confounders that affect both the treatment and outcome variables, complicating the estimation of causal effects. To address this problem, a growing literature in both causal inference and machine learning proposes to use Instrumental Variables (IV). This paper serves as the first effort to systematically and comprehensively introduce and discuss the IV methods and their applications in both causal inference and machine learning. First, we provide the formal definition of IVs and discuss the identification problem of IV regression methods under different assumptions. Second, we categorize the existing work on IV methods into three streams according to the focus on the proposed methods, including two-stage least squares with IVs, control function with IVs, and evaluation of IVs. For each stream, we present both the classical causal inference methods, and recent developments in the machine learning literature. Then, we introduce a variety of applications of IV methods in real-world scenarios and provide a summary of the available datasets and algorithms. Finally, we summarize the literature, discuss the open problems and suggest promising future research directions for IV methods and their applications. We also develop a toolkit of IVs methods reviewed in this survey at https://github.com/causal-machine-learning-lab/mliv.

Index Terms—Causal Inference, Instrument Variable, Identification.

1 Introduction

Nowadays, traditional machine learning and statistical modeling explore correlation patterns among observational variables for data mining and explanatory analysis, and have made amazing achievements in many domains over the past year [1]–[3], especially in speech recognition, image recognition, natural language processing and recommender systems. As correlation-based algorithms, machine learning techniques gain striking performance from the overfitting in training distributions under the IID hypothesis that training and testing data are independently sampled from the identical distribution. However, these models will degrade performance when the test distribution undergoes uncontrolled and unknown distribution shifts [4], [5], i.e., Out of Distribution (OOD) setting. Essentially, the accuracy drop of current models is mainly caused by the spurious correlation between the features and labels [6], [7], refered as confounding bias[1]. For example, if we do not consider the peak season, we may mistakenly conclude that higher airline ticket prices will lead to higher sales, as the peak season will lead to changes in both prices and demand for airline tickets. Lack of interpretability, actionability and stability from causality, correlation-based models has poor generalization performance on OOD data [5], [9].

To address these issues, machine learning community has tried to develop causality-inspired models by incorporating causal inference paradigms. The substantive content of these paradigms is to exploit the invariant causal relationships in the data to build models and establish stable and interpretable predictions. Scholkopf and Bengio (2022) [5] collectively refer to these approaches as structural causal models to answer counterfactual questions and make the model imaginative, that is, models can give a correct prediction in unseen scenarios. To identify the stable causal effects rather than unstable correlation patterns, the gold standard approach is to perform Randomized Controlled Trials (RCTs), where different treatments are randomly assigned to units. Nevertheless, RCTs are unrealistic in some settings due to ethical and cost issues. Hence, various methods are developed to draw inference of causal effects from observational datasets, commonly under the unconfoundedness assumption, e.g., propensity score [10], [11], covariate balance [12]–[14], back-door criteria [8], [15] and representation learning [16]. However, in practice, regardless of the approach that one adopts to control confounding in observational studies, there always exists the possibility of bias, when unmeasured confounders exist.

To control for unmeasured confounding, we introduce a third variable, named instrumental variable (IV), which is a cause of input features, has no direct effect on the outcome and does not share common causes with outcome. Using an instrumental variable to identify the hidden (unmeasured) correlation allows one to see the true correlation between the explanatory variable and response variable. For instance, the cost of fuel was used as an instrument in [17] to estimate the impact of ticket prices on sales. Thus, changes in the cost of fuel create movement in ticket prices that is independent of unmeasured confounders, and this movement is equivalent to randomization for the purposes of causal inference

- *A. Wu, K. Kuang and F. Wu are with the College of Computer Science and Technology, Zhejiang University, China. (E-mail: anpwu@zju.edu.cn; kunkuang@zju.edu.cn; wufei@cs.zju.edu.cn).*
- *R. Xiong is with the Department of Quantitative Theory & Methods, Emory University, USA. (E-mail: ruoxuan.xiong@emory.edu).*
- *K. Kuang is the corresponding author.*

1. As introduced in Chapter 3.3 in Causality [8], the **confounding bias** between the feature input and target output can be defined as the bias of causal effect estimation when imbalanced confounders exist. Confounders are common causes of feature and ouput of interest.

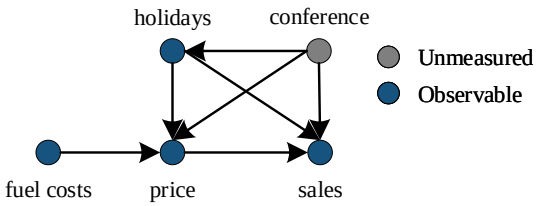


Fig. 1: The airline demand example.

[17]. See Fig. 1 for a graphical illustration of this example and of the general class of causal graphs that we consider.

Two commonly used estimators for using an instrumental variable to estimate treatment effects are the two stage least squares estimator (2SLS) and the control function estimator (CFN) [18], [19]: (1) 2SLS identifies the probability distribution over the treatment conditioned on the IVs in the treatment regression stage, and regresses the outcome based on the conditional distribution of the treatment (obtained from the treatment stage) in the outcome regression stage. Based on the adopted model, we divide 2SLS and its variants into three categories as vanilla 2SLS estimator for linear models, sieve estimator and machine learning estimator for non-linear models. (2) CFN constructs the residual variables (called control functions) in the estimation of treatment from the treatment regression stage, and estimates the outcome from the observed treatment and the residual in the outcome regression stage. Based on the structural assumption, we divide CFN and its variants into two categories as linear estimator and non-linear estimator.

In linearity setting, [19] show that CFN estimator is a 2SLS estimator with an augmented set (i.e., control function for unmeasured confounders) from instrumental variables. If these augmented variables are valid, then the control function estimator, while less robust than two stage least squares, might be much more precise because it keeps the treatment variables in the second stage [18]. However, if the augmented variables are not valid, then CFN estimator may be inconsistent, which is common in non-linear models. Fortunately, with more flexible kernel methods and neural network functions, machine learning methods have developed conditional density estimators, mutual information estimators, representational equilibrium models, etc. that can learn automatically from the data, for IV regression. This relaxes the linearity assumption and allows us to explore more complex causal systems and data.

One limitation is that these standard methods and variants of instrumental variable (IV) analysis require a pre-defined strong valid IV. These methods are reliable only when the pre-defined IV only affects the outcome through its strong association with the cause variable of interest, in practice, which is hardly satisfied due to the untestable exclusion association with outcome. Therefore, in addition to lagged values and prior knowledge [20]–[22], researchers usually implement Randomized Controlled Trials (RCTs) to sample a random variable as IV to intervene the received treatments, called intention-to-treat variable, such as Oregon health insurance experiment [23] and effects of military service on lifetime earnings [24], which are too expensive to be universally available. To save the human effort selecting pre-defined IVs, a growing number of machine learning methods have been proposed to summary existing IV candidates to generate a valid IV representations [25]–[29].

In this paper, we provide a comprehensive review of the instrumental variable methods under the potential outcome framework. We first introduce the background of the potential outcome framework and instrumental variable, including the basic definitions, corresponding assumptions, and the fundamental problems with their general solutions. To identify the causal effect from instrumental variable, then, we further summarize most of the identification conditions of instrumental variables. Combined with machine learning, subsequently, we introduce the two-stage least-squares method (2SLS) and the traditional control function method (CFN) to estimate the average treatment effects. To void human effort selecting pre-defined IVs, we also discuss how to use machine learning algorithms to synthesize a summary-IV to plug into IV-based methods. Then, we provide the related experimental information, including the available datasets that are commonly adopted in the experiments, and the open-source codes of the above methods. We also develop a toolkit of IVs methods reviewed in this survey at https://github.com/causal-machine-learning-lab/mliv.

Machine learning methods provide more flexible network models and conditional moment constraint models, which promote the development of causal inference. Meanwhile, causal inference also contributes to the development of machine learning methods. Recently, advent works [5], [7], [30] have revealed the existence and pervasiveness of variant and invariant(stable) features in data-driven algorithms and pointed out that the unstable features can provoke unexpected estimation bias for predictions. Lacking a causal perspective, machine learning algorithms are prone to exploit subtle statistical correlations present in the training distribution for predictions, which is effective when testing data and training data are independently sampled from identical distribution, i.e., IID hypothesis. In practice, however, unbalanced samples and attribute-wise imbalance are common across different scenarios [30], unlike high-quality experimental data. That means estimators tend to regard high-frequency features from the training as predictive features and view low-frequency stable features as noise, which is unstable in other distributions and even bring additional bias. Due to low-quality observational data and some key unmeasured factors, there still exists a lack of common consensus on underlying invariant features in data-driven algorithms, albeit comprehensive endeavors [5]. Therefore, researchers proposed instrumental variable regression to develop causality-inspired models, and the real-world applications that the discussed methods have great potential to benefit are discussed, including the social networks, recommendation system, computer vision, genome project, and domain adaptation as the representative examples.

To the best of our knowledge, this is the first paper that provides a comprehensive survey for instrumental variable methods under the potential outcome framework. There also exist several surveys that discuss the causal effect estimation methods under the unconfoundedness assumption, [31], [32] introduce. To summarize, our contributions of this survey are as follows:

- *Comprehensive review.* We provide a comprehensive

survey for instrumental variable methods under the potential outcome framework, including identification conditions, two-stage regression methods and control function algorithms.
- *General setting*. When we cannot access a valid instrumental variable directly, we survey a line of IV testing methods and IV synthesis methods.
- *Abundant resources*. In this survey, we list the state-of-art methods, the benchmark data sets, open-source codes, and representative applications.
- *Reproducible*. We integrate the existing resources and codes, and provide a unified interface and parameters to facilitate reproduction.

The rest of the paper is organized as follows. In section 2, we introduce the background of the instrumental variable, including the basic definitions, the assumptions, and the fundamental problems with their general solutions. In section 3, we elaborate the structural assumption for identification of causal effect in IV regression. For estimation, the 2SLS-based methods and CFN-based methods are presented in Section 4 & 5. In Section 6, we list a series of literature about IV selection and IV synthesis. Afterward, we provide experimental guidelines in section 7, and the typical applications of causal in Section 8. Final, in Section 9, we conclude several IV-based open problems and future directions.

2 Basic of Instrumental Variable

Although machine learning techniques have provided breakthroughs in statistics, econometrics, epidemiology and related disciplines, they usually suffer from low generalizability, instability, and inexplicability, due to the spurious relationship in which two or more events or variables are associated but not causally related [31], [32]. For example, in airplane sales (Example 2.1), holidays and conferences may confound the causal relationship between prices and sales and introduce additional bias; in hospital (Example 2.2), comorbidities and physical fitness would distort the causal relationships between the treatments and outcomes. Spurious relationship, deriving from confounders that are common causes of treatments and outcomes, is a common phenomenon in real-world scenarios. Hence, it is incredibly imperative and highly demanding to eliminate bias from conofunders and develop stable approaches to infer causal effect in observational studies, known as causal inference.

Example 2.1. *In the relationship (Fig. 1) between airline ticket prices and sales, prices and demand rise and fall through the seasons, being affected by other events, such as holidays and conferences [17], [33]. These events are called confounders that are the common causes of prices (cause variable, often referred to as the 'treatment') and sales (target outcome).*

Example 2.2. *In a hospital for infectious diseases, we study the effect of injection different from taking medicine (treatments) on patients' cure time (outcomes) from historical data. The patients' severity level of comorbidities and physical fitness are common causes of the treatments and outcomes, which we define as confounders. We may observe that patients with severe comorbidity have an injection, but the cure time is longer than those with mild comorbidity taking medicine, distorting the causal relationships between the treatments and outcomes.*

In causal inference, the causal effects of treatment variable on target outcome, often referred to as the treatment effect, can be estimated using control experiments, regression models, matching estimators, re-weighting techniques, and instrumental variable (IV) [31], [32]. Among these approaches, the gold standard for treatment effect estimation is to perform Randomized Controlled Trials (RCTs), in which one of two or more treatments (cause variables) are randomly assigned to samples. With enough participants, RCTS would achieve sufficient control over confounding factors and deliver a useful comparison of the treatments studied. Considering the cost and ethical issues [34], [35], fully RCTs are not always feasible in practical. Thus, in observational studies, there are a substantial number of regression models [36], [37], matching estimators [38], [39], and re-weighting techniques are developed to control or adjust the confounders to reduce the confounding bias under unconfoundedness assumption, i.e., all common causes of treatments and outcomes have been observed in data.

Nevertheless, in real-world scenarios, it is common that unmeasured confounders exist, violating the unconfounderness assumption and posing a big challenge in estimating treatment effects from observational data. Regardless of the approach that one adopts to control confounding in observational studies, there always exists the possibility of bias due to unmeasured confounders [40], e.g., it is hard to obtain all conferences information in airline demand example (Fig. 1). To overcome unmeasured confounder problems, researchers introduced an instrumental variable (fuel costs), an exogenous variable that induces changes in the treatments (prices) but has no independent effect on the outcomes (sales) [17], [41], allowing researchers to uncover the causal effect of the treatment on the outcome under a series of identification assumptions developed by [42], [43].

In 1928, the economist Philip Wright (Sewall's father) introduced IV, IV-estimator, and the equivalent two step least squares estimator, possibly in co-authorship with Sewall Wright, in the context of simultaneous equations in his book *The Tariff on Animal and Vegetable Oils*[2] [41]. Later, Haavelmo [46] and Reiersl [47] also applied the similar approach in the context of errors-in-variables models and contributed to the development of IVs unaware of the contributions of the Wrights. In linearity cases, IV estimators implement a two-stage least squares (2SLS) regression analysis for treatment effect estimation: stage 1 performs linear regression from the IVs to the treatments; and stage 2 performs linear regression from the conditional expectation of the treatments (obtained from stage 1) to the outcomes and the corresponding coefficient is used as a measure of treatment effect. To relax linearity assumption, [42], [48], [49] customized a series of identification assumptions for various scenarios, which would be elaborated in Section 3.

The framework used by IV is essentially similar to potential outcome framework outlined by Rubin [50], [51]. Next, we introduce the notations used in the IV estimator [8], and present the main challenges for causal effect estimation as well as general solutions for treatment effect.

2. Based on [44], [45].

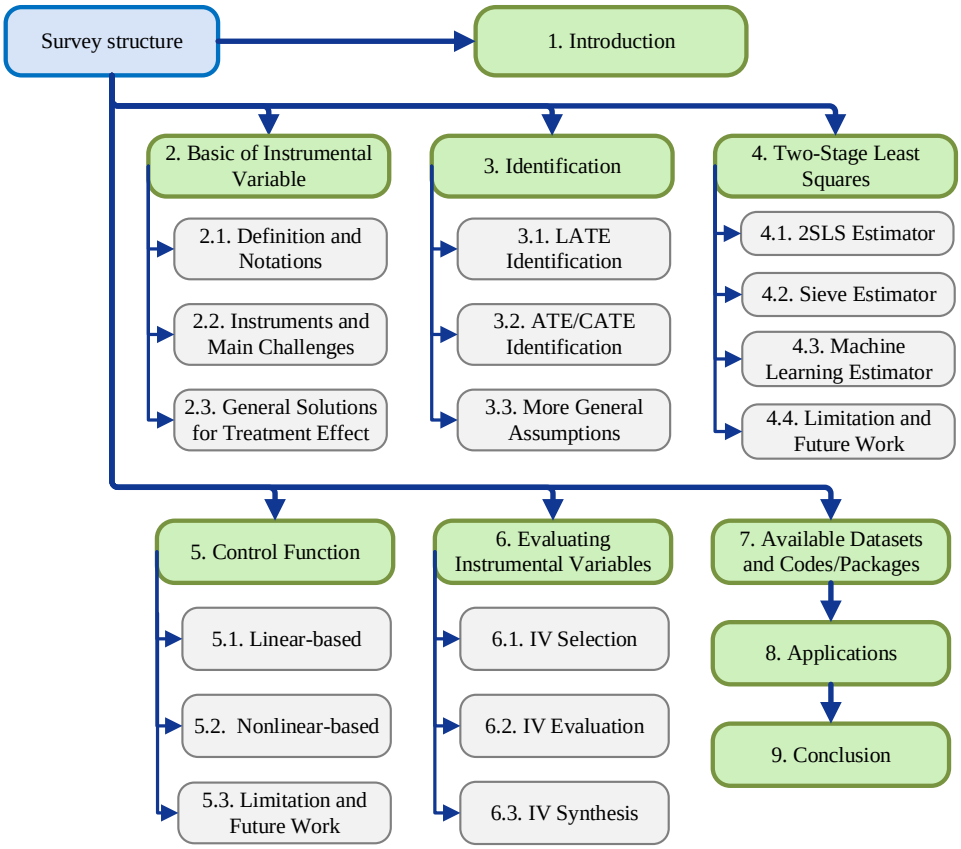


Fig. 2: Outline of the Survey.

2.1 Definition and Notations

The Rubin causal model [50]–[52], also known as the potential outcome framework, is a standard approach for IV analysis and treatment effect estimation, named after Donald Rubin. Similar to [32], we define the notations under the potential outcome framework (Fig 3).

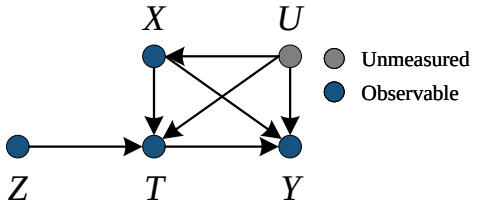


Fig. 3: The causal framework.

Note, in this paper, we use capital letters for random variables (X), small letters for their values (x), bold letters for vectors/sets of variables $(\mathbf{X})$ and their values $(\mathbf{x})$, and calligraphic letters for the spaces where they are defined $(\mathcal{X})$ if not explicitly stated. In addition, we use the subscript i to represent the a variable X_i belongs to the i-th unit, and view x_i as specific value of X_i. To simplify notation, we consistently use the shorthand $p(x)$ to represent probabilities or densities $p(X = x)$. For three random variables X, Y, Z, the conditional independence statement "X is conditionally independent of Y given $Z = z$" is written as $X \perp Y \mid Z$.

Definition 2.3. Unit/Sample i. A unit/sample denotes a single item or a collection of items from a larger whole or group. In the observational data, we can get a subset of n samples from whole population, and we use the lowercase letter i to mark each unit, $i = 1, 2, \cdots, n$.

Definition 2.4. Treatment T. Treatment refers to an intervention that applies (exposes, or subjects) to a unit. Based on the properties of treatments, we flesh out two cases: (1) in binary treatment cases, different treatment arms $T \in \{0, 1\}$ denote different intervention (receive treatment or not) and researchers spilt all samples as the treated group $(T = 1)$ and the control group $(T = 0)$; (2) in multi-valued or continuous treatment cases, practitioners generalize the binary treatment effects framework and discrete or continuous interventions are used, called dose or dosage, i.e., $T \in \mathcal{T}, \mathcal{T} \subset \mathbb{R}$.

Definition 2.5. Potential outcome $Y(T)$. Potential outcome is a core element of potential outcome framework, which defines causal effect as a comparison between two states of the world, i.e., factual state and counterfactual state of the world. In the factual state, **the factual outcome** $Y(T = t)$ is **the observed outcome** of the treatment $T = t$ that is actually applied; in the counterfactual state, one would question "what would have happened if another treatment is applied" and imagine that same man takes another treatment $T = t'$ and get the **the counterfactual outcome** $\{Y(T = t')\}_{t' \neq t, t' \in \mathcal{T}}$. The above outcomes are called potential outcome $Y(T)$, which means a proposition stating what would have happened had a potential treatment T

been applied.

In the observational data, besides the treatment of interest and observed outcome, practitioners would collect other information for each units, which can be separated as pre-treatment variables and the post-treatment variables.

Definition 2.6. Pre-treatment variables $\mathbf{V} = \{Z, X, U, \cdots\}$ are background variables that occur before the treatment T is applied and will not be affected by the treatment T. Instead, a portion of the pre-treatment variables may be the causes of the treatment, and then researchers will assign treatment based on these variables for obtaining the desired outcome. **Post-treatment variables** $\mathbf{W} = \{Y, \cdots\}$ are variables that are affected by the treatment T, and these events will occur after the treatment is accepted. In practice, based on the sequence of events and treatments occurring, Pre-treatment variables and Post-treatment variables are easily distinguished. In the following sections, we focus on the the pre-treatment variable $\mathbf{V}$ for causal inference, and we refer the terminology variable to the pre-treatment variable unless otherwise specified.

Both in decision-making applications and in the scientific literature, one tends to choose the level of the treatment to most efficiently pursue their objectives given the constraints they face [53]. That means that the above pre-treatment variables may affect practitioners' treatment assignment, leading to unbalanced data distributions across different levels of the treatment. Recently, several works [8], [54] shows such a unbalanced data would produce a spurious association, called confounding, because it tends to confound our judgment and to bias our estimate of the causal effect studied. For example, high-frequency but unrelated daily products are likely to be considered to exhibit correlation. Thus, [8] claims that if a third variable X that influences both T and Y, the real and stable causal relationship may be confounded and spurious association would introduce additional bias for stable prediction. Such a variable is then called a confounder.

Definition 2.7. Confounders X & U. In the causal relationship graph (Fig 3), confounders are some special pre-treatment variables, which simultaneously affect the treatment assignment and the outcome being studied (T, Y) so that the effect estimation may not reflect the actual relationship $(T \rightarrow Y)$ between the variables under study. In this paper, we denote $\mathbf{X}$ by **observable confounders** in observational data. For missing key variables in the record that may confound the relationship between the variables being studied (T, Y), we refer to them as **unmeasured confounders** $\mathbf{U}$. Confounders $\mathbf{X}$ & $\mathbf{U}$ are both pre-treatment variables $\mathbf{V}$.

Causal Inference. After introducing the key terminologies and definition for causal inference, the causal effect can be quantitatively defined using the above definitions. In observational dataset, the treatment can be either binary, multi-valued or continuous. For notational simplicity, we uniformly use $Y(T = t)$ to represent the potential outcome with treatment $T = t$. Then, the definition of the treatment effect is the difference $Y(T = t) - Y(T = 0)$, which can be measured at the population, subgroup, and individual levels.

Definition 2.8. Average Treatment Effect (ATE).

$$\mathbf{ATE}(t) = \mathbb{E}[Y(T = t) - Y(T = 0)], \tag{1}$$

Definition 2.9. Conditional Average Treatment Effect (CATE).

$$\mathbf{CATE}(t, \mathbf{x}) = \mathbb{E}[Y(T = t) - Y(T = 0) \mid \mathbf{X} = \mathbf{x}], \tag{2}$$

which has an another name:
Individual Treatment Effect (ITE).

$$\mathbf{ITE}_i(t) = Y_i(T = t) - Y_i(T = 0). \tag{3}$$

2.2 Instruments and Main Challenges

In many circumstances, running Randomized Controlled Trials (RCTs) are not possible due to ethical or cost concerns. In the presence of unmeasured confounders $\mathbf{U}$, estimating treatment effect from observational data is challenging due to following reasons:

- **Counterfactual**. We only realize the outcome $y_i(T = t_i)$ with a specific treatment value t_i applied to individual i, but cannot obtain the counterfactual outcomes $y_i(T \neq t_i)$ that would potentially happened if a different treatment option was assigned.
- **Imbalanced observed Covariates**. The treatments are typically not assigned at random and the covariate distributions can be quite different between different treatment arms. Some high-frequency but unrelated variables $\mathbf{X}$ would confound the causal effect of treatment on outcome of interest.
- **Imbalanced Unmeasured Covariates**. Even if we control all observed variables and adjust confounding differences from observational covariates, unmeasured key variables and differences $\mathbf{U}$ may distort the causal relationships in the observational data.

Hence, to overcome unmeasured confounder problems in observational data where the treatments are non-random assigned, researchers introduced an instrumental variable, an exogenous variable that induces changes in the treatments but has no direct effect on the outcomes, to estimate treatment effect. The instrumental variable is defined as follows:

Definition 2.10. Instrument Variable Z is an exogenous variable that affects the treatment T, but does not directly affect the outcome Y, as shown in Fig 3. Besides, an valid instrument variable satisfies the following three restrictions:
Relevance: Z is a cause of T, i.e., $\mathbb{P}(T \mid Z) \neq \mathbb{P}(T)$.
Exclusion: Z does not directly affect the outcome Y, i.e., $Z \perp Y \mid T, \mathbf{X}, \mathbf{U}$.
Independent: Z is independent of all confounders, including $\mathbf{X}$ and $\mathbf{U}$, i.e., $Z \perp \mathbf{X}, \mathbf{U}$

Nevertheless, IV methods are reliable when the pre-defined IV is a valid IV that only affects the outcome through its strong association with treatment options, called exclusion assumption. Besides, they also need some strong structural assumptions, e.g., linear models. To sum up, IV regression has the following main challenges:

- **Strict Structural Assumption**. Even if the instrument Z satisfies three restrictions in the definition, at least

one structural assumption is required to identify the treatment effect of T on Y [42], [43], [49]. The most common structural assumption is the linearity assumption, which requires that the causal relationships between all variables are linear.

- **Untestable Exclusion and Independent**. We do not have access to the unmeasured confounders in observational data, and therefore we cannot test for independence between instrumental and unmeasured variables. In addition, we cannot test whether instrumental variables have additional causality on the outcome variable.
- **Invalid and Weak IV**. In instrumental variables regression, the instruments are called *weak IV* if their correlation with the endogenous regressors is close to zero, or *invalid IV* if there is a direct effect or a hidden common cause between the instrument and the outcome. Due to untestable exclusion and independent restrictions, the predefined hand-made IVs could be weak or erroneous by violating the conditions of valid IVs.

Although IV has been used in tons of empirical papers, these thorny facts hinder the further application of the IV-based methods for treatment effect estimation. Recently, several works devote to relax or resolving these restrictions.

For structural assumptions, a substantial number of IV works have been developed to relax the unconfoundedness assumption and the identification assumption for various scenarios [43], [49], [55]–[59], which would be elaborated in Section 3. Although Exclusion and Independent are not testable, thanks to machine learning algorithms, researchers have developed Summary IV methods to automatically synthesize valid strong instrumental variables from a candidate set of instrumental variables [26], [29], [60]. Unless otherwise stated, in the following, we assume that the instrumental variables obtained from the observational data are valid strong IVs.

Remark 2.11. Angrist, Imbens and Rubin [42], [49] abandoned the effort to draw inference for the overall average effect, and focused on sub-populations for which the average effect could be identified, the so-called compliers. In binary cases, where instrument variables (Z) are different intervention assignments and treatment variables are individuals' respond to assignments ($T(Z)$), four different compliance types defined by the pair of values $(T(Z=0), T(Z=1))$ [53]:

$$i \in \begin{cases} n \text{ (never } - \text{ taker)} & \text{if } T_i(0) = T_i(1) = 0 \\ c \text{ (complier)} & \text{if } T_i(0) = 0, T_i(1) = 1 \\ d(\text{ defier)} & \text{if } T_i(0) = 1, T_i(1) = 0 \\ a \text{ (always } - \text{ taker)} & \text{if } T_i(0) = T_i(1) = 1 \end{cases} \tag{4}$$

The local average treatment effect or complier average causal effect is identified:

Definition 2.12. **Local Average Treatment Effect (LATE)**.

$$\textbf{LATE} = \mathbb{E}[Y_i(T=1) - Y_i(T=0) | i \in complier] \tag{5}$$

Under the monotonicity assumption 3.5[3], the proportion of compliers can be obtained from the remainder:

$$P(i \in c) = 1 - P(T=1|Z=0) - P(T=0|Z=1), \tag{6}$$

3. Monotonicity Assumption would be elaborated in Section 3.

Thus, monotonicity assumption is a sufficient identification assumption for LATE estimation. In this paper, we focus on reviewing more general identification assumption for ATE/CATE estimation (and thus also for LATE).

2.3 General Solutions for Treatment Effect

In IV Regression, there are two main frameworks for causal inference, i.e., Two-stage Least Squares (2SLS) [17], [61]–[63] and Control Function Method (CFN) [64]–[67]. The former uses the conditional expectation of the treatment variable to estimate the causal effect, while the latter recovers the unmeasured confounders to estimate the causal effect. In linearity assumption, we let $Z = [z_1, z_2, \cdots, z_n]'$ and $T = [t_1, t_2, \cdots, t_n]'$ and assume that the observational data is generated by:

$$T = Z\alpha + \epsilon, Y = T\beta + \epsilon, \tag{7}$$

where $\epsilon \sim \mathcal{N}(0,1)$, and $\{\alpha, \beta\}$ are the coefficients in the linear equation. Besides, our target is to predict the causal parameter β as treatment effect estimation. Details of the implementation of 2SLS and CFN are as follows.

2.3.1 Two-stage Least Squares (2SLS)

2SLS identifies the probability distribution over the treatment conditioned on the IVs in the treatment regression stage, and regresses the outcome based on the conditional distribution of the treatment (obtained from the treatment stage) in the outcome regression stage. In linearity models, the predicted values from 2SLS are obtained:
Stage 1: Regress treatments T on instruments Z:

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{i=1}^{n} (t_i - \alpha z_i)^2 = (Z'Z)^{-1} Z'T \tag{8}$$

Let $P_Z = Z(Z'Z)^{-1}Z'$, and the predicted treatment is:

$$\widehat{T} = Z\hat{\alpha} = Z(Z'Z)^{-1}Z'T = P_Z T \tag{9}$$

Stage 2: Regress Y on the predicted values $\widehat{T}$ from stage 1:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta \hat{t}_i)^2 \tag{10}$$

which gives:

$$\beta_{\text{2SLS}} = \left(X' P_Z^T P_Z T\right)^{-1} T' P_Z Y \tag{11}$$

This method requires a strong linear relationship between the instrumental variables and the treatment variables, which will be not applicable if the unmeasured noise ϵ is large. More nonlinear variants of 2SLS are detailed in Section 4.

2.3.2 Control Function Method (CFN)

CFN constructs the residual variables (called control functions) in the estimation of treatment from the treatment regression stage, and estimates the outcome from the true treatment and the residual in the outcome regression stage. In linearity models, the predicted values from CFN are

obtained:
Stage 1: Regress treatments T on instruments Z:

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{i=1}^{n} (t_i - \alpha z_i)^2 = (Z'Z)^{-1} Z'T \tag{12}$$

and the predicted residuals is:

$$\hat{\epsilon} = T - Z\hat{\alpha} = T - P_Z T \tag{13}$$

Stage 2: Regress Y on the predicted residuals from stage 1:

$$\hat{\beta}, \hat{\beta}_\epsilon = \arg\min_{\beta, \beta_\epsilon} \sum_{i=1}^{n} (y_i - \beta \hat{t}_i - \beta_\epsilon \hat{\epsilon}_i)^2 \tag{14}$$

which gives:

$$(\beta_{\text{CFN}}, \beta_\epsilon) = ((T, \epsilon)^T (T, \epsilon))^{-1} (T, \epsilon)^T Y \tag{15}$$

where (A, B) means a concatenate of vectors/matrices A and B. This method is valid for large unmeasured confounding bias. More nonlinear variants of CFN are detailed in the Section 5.

3 IDENTIFICATION

In this section, we will discuss the structural assumptions or restrictions on data and model for precise inference to be possible, i.e., identification. After obtaining an infinite number of observations from population, if it is theoretically possible to learn the true values of a model's underlying parameters, then the model is identifiable. Otherwise, it is not non-identifiable. Even if the instrument satisfies IV's three constraints, we might not be able to identify causal effects unless there are additional structural assumptions [49], [55], [56], [68].

Example 3.1. ***Non-identifiability*** *Without loss of generality, we take continuous treatment cases as an example and assume an unmeasured confound U is a random variable from a standard normal distribution $\mathcal{N}(0, 1)$:*

$$T = ZU, Y = TU, U \sim \mathcal{N}(0, 1). \tag{16}$$

Due to the unmeasured confounders, the relationships between Z & T, Z & Y and T & Y can no longer be accurately regressed by any parametric or non-parametric models.

In econometric program evaluation, to address the non-identifiability problem, a standard method is to build structural equation model with linearity assumptions for treatment effect identification [69], [70].

Assumption 3.2. ***Linearity Assumptions [42], [70].*** *For experimental or observational data, let Y be the observed outcome of interest, let T be the observed treatment, and let Z be the observed instrument variable. For continuous treatment cases, a standard structural assumption for the identification of treatment effect would have the form:*

$$T = \alpha_0 + \alpha_1 Z + \epsilon_T, \tag{17}$$

$$Y = \beta_0 + \beta_1 T + \epsilon_Y, \tag{18}$$

where $\{\alpha_0, \alpha_1, \beta_0, \beta_1\}$ are corresponding scalar coefficients, as well as ϵ_T and ϵ_Y are additive confounding effect from unmeasured confounders, which influence both the treatment and the outcome ($\epsilon_T \not\perp \epsilon_Y$). In the model β_1 represents the causal effect of T on Y.
For binary treatment, the structural Eq. (17) could be reformulated as:

$$T = 1\{\alpha_0 + \alpha_1 Z + \epsilon_T \geq 0\}, \tag{19}$$

where $1\{\cdot\}$ is a indicator function.

Following IV's independent restriction, we have $Z \perp \epsilon_T, \epsilon_T$. Then, the absence of Z in Eq. 18 denotes that any effect of Z on Y must be through an effect of Z on T in Eq. 17/ 19. Thus, Z can be considered as a strong and valid IV for treatment effect estimation (i.e., β_1). The IV estimator is defined as the ratio of sample covariance [42], [71]. For binary instrument and treatment cases,

$$\hat{\beta}_1 = \text{cov}(Y, Z) / \text{cov}(T, Z) \tag{20}$$

$$= \frac{\mathbb{E}(YZ)/\mathbb{E}(Z) - \mathbb{E}(Y(1-Z))/\mathbb{E}(1-Z)}{\mathbb{E}(TZ)/\mathbb{E}(Z) - \mathbb{E}(T(1-Z))/\mathbb{E}(1-Z)} \tag{21}$$

For continuous instrument and treatment cases,

$$\hat{\beta}_1 = \text{cov}(Y, Z) / \text{cov}(T, Z) \tag{22}$$

$$= \frac{\mathbb{E}(Y - (\mathbb{E}(Y))(Z - \mathbb{E}(Z)))}{\mathbb{E}(T - (\mathbb{E}(T))(Z - \mathbb{E}(Z)))} \tag{23}$$

Without the linearity assumption in real-world scenarios, even if the instrument satisfies IV's three constraints, we might not be able to identify causal effects. However, the linearity assumption (Eq. 17, 18 & 19) have not found widespread use in real-world scenarios, and exists only in theoretical studies. Besides, the apparently unreproducible experimental results also prevent the application of instrumental variable parameter models under the linear assumption [72]. To relax the linear assumption and avoid parametric evaluation models, researchers had devoted to establishing conditions that guarantee nonparametric identification of treatment effects in observational studies, i.e. identification without relying on functional form restrictions or distributional assumptions [42], [43], [49], [59], [73].
Identifiability. Briefly speaking, even if the IVs are valid, further assumptions are required for the identification of treatment effect. Following criticism of parametric evaluation models [72], instead of sticking to the average treatment effects in a population of interest, researchers use some weaker assumptions to identify the average effect for the compliers sub-population, i.e., Local Average Treatment Effect (LATE). Sufficient assumptions for this include: Constant/Additive Treatment Effect [68], Zero Probability on Some IV Value [55], [56] and Monotonicity [42], [49], [73]. However, under these assumptions, we can identify the average treatment effect for the group of compliers but not for the specific members.

It was not until 2003, when Newey and Powell (2003) gave the identification and estimation results for nonparametric conditional moment restrictions, that practitioners started to focus on the identifiability of the structure function for outcome, i.e., Conditional Average Treatment Effect (CATE) or ATE [17], [43], [57]. Subsequently, some more general homogeneity assumptions are developed one after another [58], [59], [74]–[76] for ATE(CATE). In the econometrics literature, Homogeneity Assumption is a more general version than Monotonicity Assumption and Additive Noise Assumption [59], [77].

Based on assumption for LATE or CATE, we are going to elaborate on these assumptions as LATE Identification Assumptions, CATE Identification Assumptions, and More General Assumptions.

3.1 LATE Identification

As illustrated in Remark 2.11, Angrist, Imbens and Rubin [42], [49] abandoned the effort to draw inference for the overall average effect, and focused on sub-populations for which the average effect could be identified, the so-called compliers. Four different compliance types are defined in Eq. 4. The compliers means that the groups of people who can be induced to change treatments by assigning different instruments. In this section, we will list some sufficient assumptions or conditions for identifying treatment effect of the compliers, as follows.

Assumption 3.3. ***Constant Treatment Effect [68].*** *To prevent above problem, one condition is the treatment effect is constant: $\alpha = Y(1) - Y(0)$ for any unit i. Then $\mathbb{E}[Y \mid Z = z] - \mathbb{E}[Y \mid Z = w]$ is equal to:*

$$\begin{aligned} & \mathbb{E}[Y \mid Z = z] - \mathbb{E}[Y \mid Z = w] \\ = \; & \mathbb{E}[T(z)Y(1) + (1 - T(z)) \cdot Y(0) \mid Z = z] \\ - \; & \mathbb{E}[T(w)Y(1) + (1 - T(w)) \cdot Y(0) \mid Z = w] \\ = \; & Y(1)P[T = 1 \mid Z = z] + Y(0)P[T = 0 \mid Z = z] \\ - \; & Y(1)P[T = 1 \mid Z = w] + Y(0)P[T = 0 \mid Z = w] \\ = \; & [Y(1)P(z) + Y(0)(1 - P(z))] \\ - \; & [Y(1)P(w) + Y(0)(1 - P(w))] \\ = \; & (Y(1) - Y(0))(P(z) - P(w)) \\ = \; & \alpha(P(z) - P(w)) \end{aligned} \tag{24}$$

Assumption 3.4. ***Zero Probability [55], [56].*** *A second approach is to assume the existence of some value of the instrument, $w \in \mathcal{W}$, such that the probability of participation conditional on that value is equal to zero, i.e., $P(w) = 0$. Then $P(T(z) - T(w) = -1) = 0$:*

$$\begin{aligned} & \mathbb{E}[Y \mid Z = z] - \mathbb{E}[Y \mid Z = w] \\ = \; & [Y(1)P(z) + Y(0)(1 - P(z))] - [Y(1)P(w) + Y(0)(1 - P(w))] \\ = \; & Y(1)P(z) + Y(0)(1 - P(z)) - Y(0) \\ = \; & Y(1)P(z) - Y(0)P(z) \\ = \; & P(z)\mathbb{E}[Y(1) - Y(0) \mid T(z) = 1] \end{aligned} \tag{25}$$

To identify the causal effect, we need to know at least one value w of $\mathcal{W}$.

Let A be an indicator for the event $Z \notin \mathcal{W}$, i.e., $A = \mathbb{1}\{Z \notin \mathcal{W}\}$. Then:

$$\begin{aligned} & \mathbb{E}[Y \mid A = 0] \\ = \; & \mathbb{E}[Y \mid T(w) = 1] \cdot P(w) \\ + \; & \mathbb{E}[Y \mid T(w) = 0] \cdot (1 - P(w)) \\ = \; & \mathbb{E}[Y(0)], w \in \mathcal{W}. \end{aligned} \tag{26}$$

and

$$\begin{aligned} & \mathbb{E}[Y \mid A = 1] \\ = \; & \mathbb{E}[Y_0 \mid A = 1] + P(z)\mathbb{E}[Y_1 - Y_0 \mid T(z) = 1] \\ = \; & \mathbb{E}[Y_0] + P(z)\mathbb{E}[Y_1 - Y_0 \mid T(z) = 1], z \notin \mathcal{W}. \end{aligned} \tag{27}$$

Since we can estimate $P(z)$, $\mathbb{E}[Y \mid A = 0]$ and $\mathbb{E}[Y \mid A = 1]$ and we know $z \notin \mathcal{W}$, then we can identify ATE:

$$\mathbb{E}[Y_1 - Y_0 \mid T(z) = 1] = \frac{\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]}{P(z)}. \tag{28}$$

Assumption 3.5. ***Monotonicity [42], [49], [73].*** *For all possible value of instrument, z and w, either $T(z) \geq T(w)$ for any unit i, or $T(z) \leq T(w)$ for any unit i. Without loss of generality, the assumption is satisfied with $T(z) \geq T(w)$:*

$$\begin{aligned} & \mathbb{E}[Y \mid Z = z] - \mathbb{E}[Y \mid Z = w] \\ = \; & (P(z) - P(w)) \cdot \mathbb{E}[Y(1) - Y(0) \mid T(z) - T(w) = 1] \end{aligned} \tag{29}$$

3.2 ATE/CATE Identification

For LATE models, assumption 3.3, 3.4 or 3.5 are sufficient for identification [49], [55], [56], [68]. Besides, in a linear outcome process, where the outcome process is a sum of the causal effect and zero-mean noise, zero covariance between the instruments Z and unmeasured disturbances (confounders) $\mathbf{U}$, suffices for identify the causal effect [43], [57]. In a nonparametric IV (NPIV) model for CATE, the moment restrictions that unmeasured disturbances has conditional mean zero given instruments is a necessary restriction for identification, i.e., $\mathbb{E}[\mathbf{U} \mid Z] = 0$.

Assumption 3.6. ***Additive Noise Assumption / Separability Assumption [17], [43].*** *In the parametric/nonparametric model (Eq. (30)), the identification/uniqueness of $\hat{g}(\mathbf{X}, T)$ is equivalent to the nonexistence of any function $\delta(\mathbf{X}, T) := g(\mathbf{X}, T) - \hat{g}(\mathbf{X}, T) \neq 0$ such that $\mathbb{E}[\delta(\mathbf{X}, T) \mid Z] = 0$.*

In the nonparametric setting, the relationship between the outcome process and reduced form belongs a 1st Fredholm integral equation [78] and leads an ill-posed inverse problem [43]. Considering the identification of a general nonparametric model:

$$Y = g(\mathbf{X}, T) + \mathbf{U}, \mathbb{E}[\mathbf{U} \mid Z] = \mathbb{E}[\mathbf{U}] = 0. \tag{30}$$

where $g(\cdot)$ denotes a true, unknown structural function of interest. For a consistency estimation, [17], [43], [57] identified the causal effect as the solution of an integral equation:

$$\begin{aligned} \mathbb{E}[Y \mid Z, \mathbf{X}] &= \mathbb{E}[g(\mathbf{X}, T) \mid Z, \mathbf{X}] + \mathbb{E}[\mathbf{U} \mid \mathbf{X}] \\ &= \int [g(\mathbf{X}, T) + \mathbb{E}[\mathbf{U} \mid \mathbf{X}]] \, dF(T \mid Z, \mathbf{X}) \\ &= \int \hat{g}(\mathbf{X}, T) dF(T \mid Z, \mathbf{X}) \end{aligned} \tag{31}$$

where F denotes the conditional cumulative distribution function of T given $\{Z, \mathbf{X}\}$. Given two observable functions $\mathbb{E}[Y \mid Z, \mathbf{X}]$ and $F(T \mid Z, \mathbf{X})$, $\hat{g}(\mathbf{X}, T)$ is the solution of the inverse problem. Then, we can identify ATE:

$$\text{ATE} = \hat{g}(\mathbf{X}, T) - \hat{g}(\mathbf{X}, 0) = g(\mathbf{X}, T) - g(\mathbf{X}, 0). \tag{32}$$

Therefore, [17], [43] characterized identification of structural functions as completeness of certain conditional distributions $\mathbb{E}[\mathbf{U} \mid Z] = 0$.

3.3 More General Assumptions

In the econometrics literature [59], [77], Homogeneity Assumption is a more general version than Monotonicity Assumption and Additive Noise Assumption. Next, we describe two general Homogeneity Assumptions and the No Effect Modification Assumption. Note that the previous assumptions (except the Monotonicity Assumption) can be viewed as a special case of the Homogeneity Assumptions.

Assumption 3.7. ***Homogeneous Instrument-Treatment Association [59], [75], [76]****: The association between the IV and the treatment is homogeneous in the different level of unmeasured confounders, i.e.,* $\mathbb{E}[T|Z=a,\mathbf{U}]-\mathbb{E}[T|Z=b,\mathbf{U}]=\mathbb{E}[T|Z=a]-\mathbb{E}[T|Z=b]$.

Assumption 3.8. ***Homogeneous Treatment-Outcome Association [58], [59], [74]****: The association between the treatment and the outcome is homogeneous in the different level of unmeasured confounders, i.e.,* $\mathbb{E}[Y|T=a,\mathbf{U}]-\mathbb{E}[Y|T=b,\mathbf{U}]=\mathbb{E}[Y|T=a]-\mathbb{E}[Y|T=b]$.

Meanwhile, No effect modification of the treatment effect (NEM) is weaker than Homogeneity Assumptions, but may not be plausible in many instances [58], [59].

Assumption 3.9. ***No Effect Modification [58], [59], [76]****: The unmeasured confounders* $\mathbf{U}$ *would not modify the causal effect of* T *on* Y.

4 Two-Stage Least Squares

As discussed above, we know that when there are unmeasured confounder in the data, the causality obtained by direct regression (Ordinary Least Squares, OLS) will be distorted. In this section, we will explain the inconsistency of ordinary least squares for causal effect and introduce some typical IV-based methods for consistency estimation, i.e., two-stage least squares and its variants with machine learning. As a classical statistical method for causal effect estimation, the two-stage least squares performs linear regression from the instruments Z to the treatments T in stage 1, and fit the counterfactual outcome function to predict the outcomes Y from the conditional expectation of the treatments $\mathbb{E}[T \mid Z]$ (obtained from stage 1) in stage 2.

Under the nonparametric identification of ATE/CATE in observational studies (See Section 3.2), based on traditional linear methods and advanced non-linear variants, as shown in Fig. 5, we divide two-stage least squares and its variants into three categories: (1) Vanilla 2SLS and Wald Estimator for linear models (OLS is not applicable to causal effects); (2) Sieve estimator for non-linear models [43], [79]; (3) Machine Learning for further estimation. There are four main research lines from machine learning estimator, incliuding: Kernel-based estimator [61], [62], Deep-based estimator [17], [80], [81], Moment conditions estimator [63], [82] and confounder balance estimator.Finally we will summarize the limitations of these approaches and future works.

4.1 2SLS Estimator

Followed the linear Gaussian assumption in traditional 2SLS, without intercept for notational convenience, we assume that the observational data is generated by:

$$T = Z\alpha + f(\mathbf{U}) = Z\alpha + \epsilon_T, \tag{33}$$

$$Y = T\beta + g(\mathbf{U}) = T\beta + \epsilon_U, \tag{34}$$

where $\{\alpha, \beta\}$ are the coefficients in the linear equation. Without interactions between unmeasured confounders and treatment, we can represent the effect of infinitely many unmeasured causes $\{f(\mathbf{U}), g(\mathbf{U})\}$ as an additive noise $\{\epsilon_T, \epsilon_Y\}$ regardless of how they interact among themselves, where $f(\cdot)$ and $g(\cdot)$ can be any continuous functions. Besides, the instrumental variable Z is correlated with the independent variable T and uncorrelated with the unmeasured confounder U. We call Eqs. (33) and (34) the structural equations or primary equations, especially, Eq. (33) is treatment-assignment function and Eq. (34) is counterfactual function in continuous setting. The corresponding causal diagram is shown in Fig. 6(c).

4.1.1 Inconsistency of Ordinary Least Squares

In the presence of unmeasured confounder in the observational data, the causality obtained by direct regression (Ordinary Least Squares, OLS) will be distorted. In causal inference, the goal of regression analysis is to estimate the conditional expectation function $\mathbb{E}[Y \mid do(T)]$, i.e., to recover coefficient β from the scalar regression model[4]. Recall that ordinary least squares (OLS) solves for $\hat{\beta}$ by minimize the sum of squared errors:

$$\min_{\beta} \quad (Y - T\beta)'(Y - T\beta). \tag{35}$$

The first-order condition is $T'(Y - T\hat{\beta}) = T'\hat{\epsilon}_Y = 0$. The regression results are reliable only when T and $\mathbf{U}$ are independent $\mathbb{P}(T \mid \mathbf{U}) = \mathbb{P}(T)$, i.e., $\epsilon_T = f(\mathbf{U}) \equiv 0$ in dose-assignment function Eq. (33), as shown in Fig. 6(a). Then the treatment variable T affects the outcome variable Y only through $T\beta$, and there is no association between T and $\mathbf{U}$.

But in real-world scenarios, there may exist some unmeasured confounders $\mathbf{U}$ that are the common causes of the treatment T and the outcome Y, i.e., $\epsilon_T = f(\mathbf{U}) \not\equiv 0$ in dose-assignment function Eq. (33), as shown in Fig. 6(b). Now there is an association between T and $\mathbf{U}$, i.e., ϵ_T. Then, the true model is believed to have $T'\mathbf{U} \neq 0$ in the presence of unmeasured confounders $\mathbf{U}$.

Based on non-zero $\mathbf{U}$-T association ($f(\mathbf{U}) \not\equiv 0$), from Eq. (34) there is a direct effect ($T\beta$) and an indirect effect via $\mathbf{U}$ affecting T which in turn generates an additional false correlation term between T and Y. If we directly perform OLS regression (Eq. (35)) to estimate the causal effect, OLS will combine these two effects to give a bias result, i.e., $\hat{\beta}_{\text{OLS}} \neq \beta$. In this case, the coefficient on the treatment T is given:

$$\begin{aligned}\hat{\beta}_{\text{OLS}} &= (T'T)^{-1}T'Y = (T'T)^{-1}T'(T\beta + \epsilon_Y) \\ &= \beta + (T'T)^{-1}T'\epsilon_Y \qquad (36)\\ \hat{\beta}_{\text{OLS}} &= \frac{dY}{dT} = \beta + \frac{d\epsilon_Y}{dT} \qquad (37)\end{aligned}$$

Therefore, the OLS estimates the bias effect $\beta + d\epsilon_Y/dT$ rather than the true effect β. In a conclusion, the OLS

4. $do(\cdot)$ denotes do-operation which manipulates the value of treatments as T in experimental study.

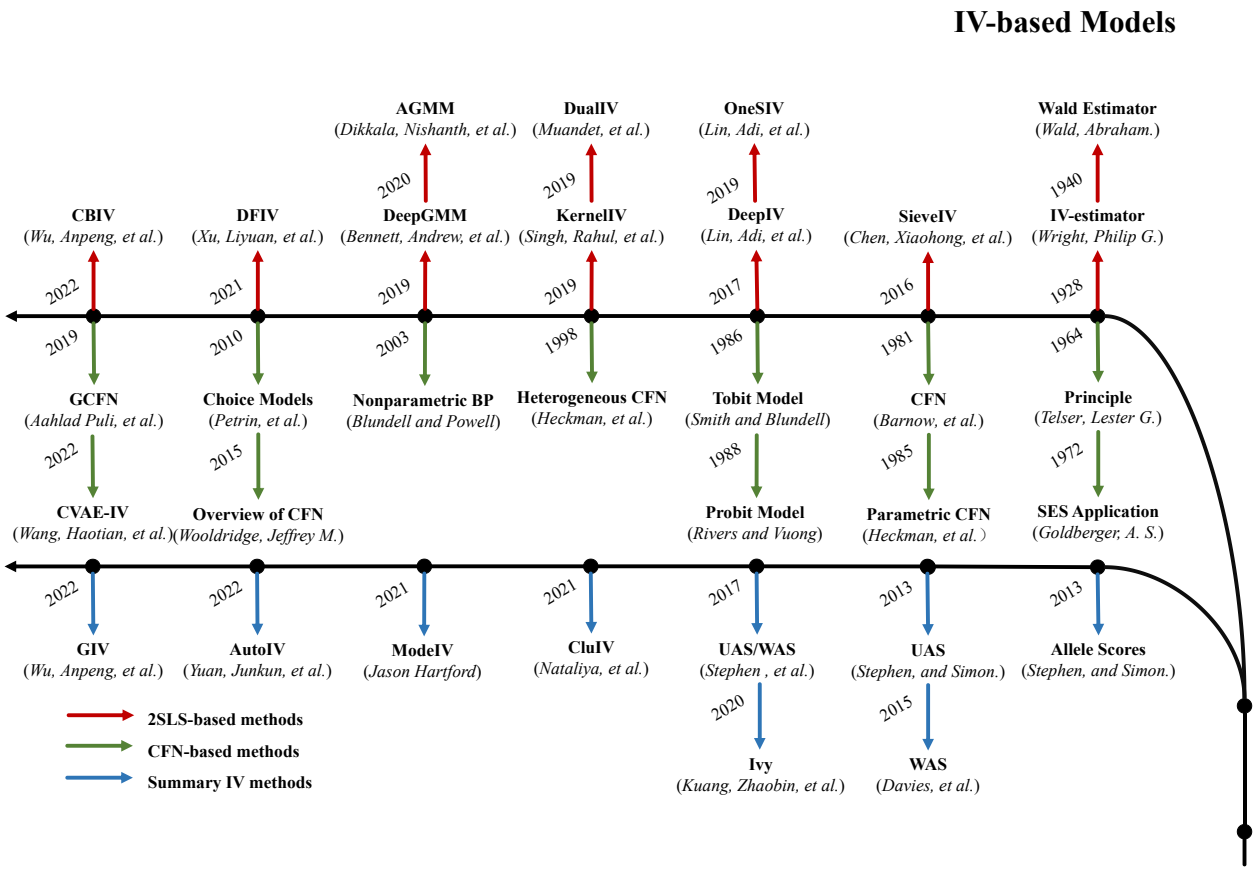


Fig. 4: Key milestones in the development of instrumental variable.

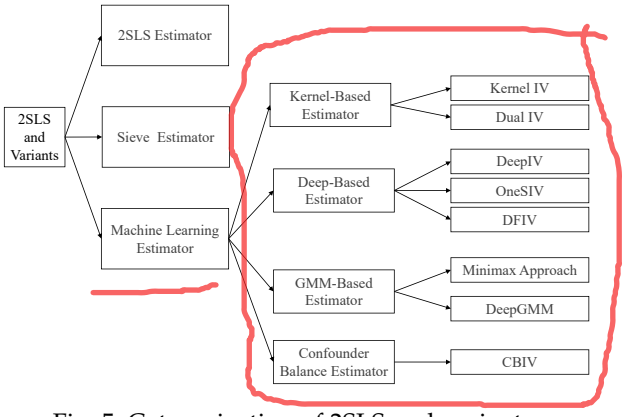


Fig. 5: Categorization of 2SLS and variants.

estimator is biased and inconsistent for causal inference in the presence of unmeasured confoudners. Therefore, the researchers proposed a two-stage regression method to eliminate confounding bias $d\epsilon_Y/dT$ [41].

4.1.2 Two-Stage Least Squares

In the linear Gaussian model (Eqs. (33) and (34)) discussed above, assumption 3.5 is automatically satisfied. Thus, we can identify the causal effect via 2SLS using IVs Z, which is not related to $\mathbf{U}$, i.e., $Z'\epsilon_T = Z'\epsilon_Y = 0$.

The Treatment Regression Stage: in stage 1 of 2SLS, estimator regresses treatment T from IVs Z:

$$\hat{\alpha} = (Z'Z)^{-1} Z'T = (Z'Z)^{-1} Z'(Z\alpha + \epsilon_T) = \alpha \quad (38)$$

$$\hat{T} = \mathbb{E}[T \mid Z] = Z\alpha \quad (39)$$

According to the IVs' unconfounded assumption, there is no association between $\hat{T}$ and $\mathbf{U}$. Hence, as shown in Fig. 6(d) the first-order condition $\hat{T}'(Y - T\beta) = \hat{T}'\epsilon_Y = 0$ is satisfied.

The Outcome Regression Stage: in stage 2 of 2SLS, estimator regresses the outcome Y based on the conditional expectation of the treatment $\hat{T}$ (obtained from stage 1):

$$\hat{\beta}_{\text{2SLS}} = \left(\hat{T}'\hat{T}\right)^{-1} \hat{T}'Y = \left(\hat{T}'\hat{T}\right)^{-1} \hat{T}'(\hat{T}\beta + \epsilon_Y) = \beta \quad (40)$$

$$\hat{Y} = \mathbb{E}[Y \mid \hat{T}] = \hat{T}\beta \quad (41)$$

Then, we can get the counterfactual function by replacing $\hat{T}$ with T:

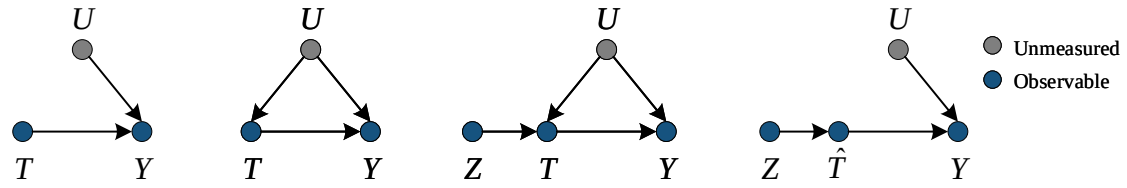
$$\hat{Y} = \mathbb{E}[Y \mid do(T)] = T\beta. \quad (42)$$

4.1.3 Wald Estimator

In 1940s, the economist Wald proposed the wald estimator for a non-continuous IV case where the instruments Z is a binary instrument [83]. Denote the sub-sample averages of Y and T by $\bar{Y}_1$ and $\bar{T}_1$ when $Z = 1$ and by $\bar{Y}_0$ and $\bar{T}_0$ when $Z = 0$. Then, we can get the derivatives:

$$\frac{dY}{dZ} = \bar{Y}_1 - \bar{Y}_0 \quad (43)$$

$$\frac{dT}{dZ} = \bar{T}_1 - \bar{T}_0 \quad (44)$$



(a) Ideal Cases (b) Confounding Cases (c) Instruments Cases (d) Target of IV Regression

Fig. 6: The causal diagram in different cases.

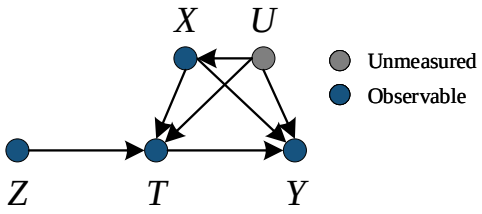


Fig. 7: The causal diagram in more general cases.

therefore, the causal effect is:

$$\hat{\beta}_{\text{Wald}} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{T}_1 - \bar{T}_0}. \tag{45}$$

Wald Estimator is a binary IV version of 2SLS, under linearity assumption.

4.2 Sieve Estimator

To satisfy the identification conditions, 2SLS simplifies the IV estimation problem by assuming linear models. To generalize 2SLS to the nonlinear setting, motivated by the works [84], [85] on sieve estimation, [43], [57] propose a non-parametric two-stage basis expansion approach, called Sieve NPIV, with uniform convergence rates [79]. Under a more general case ((Fig. 7)), Sieve IV defines an appropriate finite dictionary of basis functions (Hermite polynomial or a set of indicator functions) for the treatments regression T and the outcomes regression Y with instruments Z and observed covariates $\mathbf{X}$, and specifies the number of basis expansion functions [86], [87]. Specifically, under homogeneity assumption 3.8, Sieve IV focus on identification of the models:

$$T = f(Z, \mathbf{X}) + \epsilon_T, \tag{46}$$
$$Y = g(T, \mathbf{X}) + \epsilon_Y, \tag{47}$$

where $g(\cdot)$ denotes the true, unknown structural function of interest, ϵ_T and ϵ_T are joint errors from unobserved variables $\mathbf{U}$, and the unmeasured confounders ϵ_T and ϵ_T are additive noise, that is independent with the instruments Z, i.e., $\mathbb{E}[\epsilon_T \mid Z] = \mathbb{E}[\epsilon_Y \mid Z] = 0$.

Based on these sieve bases, Sieve IV implement a two-stage regression to estimate causal effect.

Sieve IV.
Formally, Sieve IV estimates the structure function using an appropriate finite dictionary of basis functions and we can reformulate the structure function as:

$$T = \sum_{i=1}^{d^Z} \sum_{j=1}^{d^X} \alpha_{i,j} \phi_i(Z) \xi_j(\mathbf{X}) + \epsilon_T, \tag{48}$$
$$Y = \sum_{k=1}^{d^T} \sum_{j=1}^{d^X} \beta_{k,j} \psi_k(T) \xi_j(\mathbf{X}) + \epsilon_Y, \tag{49}$$

where $\{\phi_i\}_{i=1}^{d^Z}$ is the sieve basis for IVs Z with degree d^Z, $\{\xi_i\}_{i=1}^{d^X}$ is the sieve basis for confounders $\mathbf{X}$ with degree d^X, $\{\phi_k\}_{k=1}^{d^T}$ is the sieve basis for treatments T with degree d^T, and $\{\alpha_{i,j}, \beta_{i,j}\}$ are the corresponding coefficients. Each of the ϕ_i is a function from $\mathcal{Z}$ into $\mathbb{R}$, each of the ξ_j is a function from $\mathcal{X}$ into $\mathbb{R}$, and each of the ψ_k is a function from $\mathcal{T}$ into $\mathbb{R}$.

Then the goal of Sieve IV is to estimate:

$$\text{CATE}(x, t) = \sum_{k=1}^{d^T} \sum_{j=1}^{d^X} \beta_{k,j} \xi_j(x) [\psi_k(T = t) - \psi_k(T = 0)]. \tag{50}$$

In **the treatment regression stage**, different than 2SLS, Sieve IV regresses each of the treatment basis functions ($\mathbb{E}[\psi_k(T) \mid \phi_i(Z)\xi_j(\mathbf{X})]$) on the basis features $\{\phi_i(Z)\xi_j(\mathbf{X})\}$ rather than the conditional expectation treatment distribution ($\mathbb{E}[T \mid Z, \mathbf{X}]$).

$$\hat{\alpha} = \text{argmin}_\alpha \text{MSE}(\psi(\sum_{i=1}^{d^Z} \sum_{j=1}^{d^X} [\alpha_{i,j} \phi_i(Z) \xi_j(\mathbf{X})]), \psi(T)). \tag{51}$$

In **the outcome regression stage**, Sieve IV estimates the expectation outcome onto these estimated functions $\mathbb{E}[\psi_k(T) \mid \phi_i(Z)\xi_j(\mathbf{X})]$ (obtained by the stage 1) and bases $\xi_j(\mathbf{X})$ to identify the coefficients $\beta_{k,j}$.

$$\hat{\beta} = \text{argmin}_\beta \text{MSE}(\sum_{k=1}^{d^T} \sum_{j=1}^{d^X} [\beta_{k,j} \psi_k(T) \xi_j(\mathbf{X})], Y). \tag{52}$$

In the two-stage regression of Sieve IV, the challenge is how to define an appropriate series basis functions [79]. Thus, recent works [61], [62] introduce machine learning algorithm to obtain the basis functions and estimate causal effect.

Z和X估计T，T和X估计Y

4.3 Machine Learning Estimator

To implement further estimation, as shown in Fig. 5, there are four main research lines from machine learning estimator (Fig. 5), incliuding: Kernel-based Estimator [61], [62], [79], Deep-based methods [17], [80], [81], Moment conditions methods [63], [82] and Confounder Balanced Estimator [54].

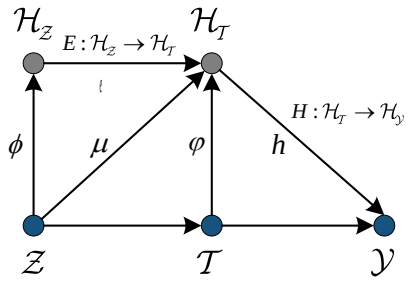


Fig. 8: The Structural Function of KernelIV.

4.3.1 Kernel-based Estimator

Motivated by Sieve NPIV [79] and predictive state representation models (PSRs) [88] and [89], [61] proposes kernel instrumental variable regression (KernelIV) to model relations among Z, $\mathbf{X}$, T, and Y as nonlinear functions in reproducing kernel Hilbert spaces (RKHSs) [90], and prove the consistency of KernelIV.

Kernel IV.
As shown in Fig. 8, KernelIV defines two measurable positive definite kernels $k_\mathcal{T}:\mathcal{T}\times\mathcal{T}\to\mathbb{R}$ and $k_\mathcal{Z}:\mathcal{Z}\times\mathcal{Z}\to\mathbb{R}$ corresponding to scalar-valued RKHSs $\mathcal{H}_\mathcal{T}$ and $\mathcal{H}_\mathcal{Z}$:

$$\psi:\mathcal{T}\to\mathcal{H}_\mathcal{T}, t\mapsto k_\mathcal{T}(t,\cdot),\quad \phi:\mathcal{Z}\to\mathcal{H}_\mathcal{Z}, z\mapsto k_\mathcal{Z}(z,\cdot) \tag{53}$$

where ψ and ϕ are the basis functions of $\mathcal{Z}$ and $\mathcal{T}$. In this section, $\mathcal{Z}$ means the the horizontal concatenation of IVs $\mathcal{Z}$ and confounders $\mathcal{X}$, and $\mathcal{T}$ means the the horizontal concatenation of treatments $\mathcal{T}$ and confounders $\mathcal{X}$, i.e., $\mathcal{Z}=\mathcal{Z}\oplus\mathcal{X}$ and $\mathcal{T}=\mathcal{T}\oplus\mathcal{X}$. Then, KernelIV reformulates the problem as:

$$e\in E, E:\mathcal{H}_\mathcal{Z}\to\mathcal{H}_\mathcal{T}, \tag{54}$$
$$h\in H, H:\mathcal{H}_\mathcal{T}\to\mathcal{H}_\mathcal{Y}. \tag{55}$$

In stage 1, KernelIV learns a conditional mean embedding to model the relations between $\mathcal{Z}$ and $\mathcal{T}$ by two kernel functions ψ and ϕ and a conditional expectation operator E:

$$\hat{\phi}(T)=\mu(Z)=e(\psi(Z))=\mathbb{E}[\phi(T)\mid Z], \tag{56}$$
$$\psi(Z)\in\mathcal{H}_\mathcal{Z},\quad \hat{\phi}(T),\phi(T)\in\mathcal{H}_\mathcal{T},$$

KernelIV constructs a objective for optimizing $e\in E$ by kernel ridge regression:

$$e^*_\lambda=\text{argmin}_{e\in E}\mathbb{E}\|[e\psi](Z)-\phi(T)\|^2+\lambda\|[e\|^2, \tag{57}$$
$$e^*_\lambda=\text{argmin}_{e\in E}\mathbb{E}\|\mu(Z)-\phi(T)\|^2+\lambda\|[e\|^2, \tag{58}$$

where λ is a hyper-parameter and $\|[e\|^2$ is a penalty term for function e. Indeed, $\hat{T}=\mu(Z)=[e^*\psi](Z)$. Analogously, in 2SLS $\hat{T}=\mathbb{E}[T\mid Z]=\hat{\alpha}Z$ for stage 1 linear regression parameter $\hat{\alpha}$.

In stage 2, to estimate the structural function $g(\cdot)$ (Eq. (47)), KernelIV predicts the potential outcome function onto the conditional mean embedding $\hat{\phi}(T)\in\mathcal{H}_\mathcal{T}$:

$$\hat{Y}=g(T)=h(\phi(T))=[h\mu](Z)=\mathbb{E}[Y\mid\hat{\phi}(T)], \tag{59}$$

KernelIV constructs a objective for optimizing $h\in H$ by kernel ridge regression:

$$h^*_\lambda=\text{argmin}_{h\in H}\mathbb{E}\|h(\phi(T))-Y\|^2+\lambda\|[h\|^2, \tag{60}$$
$$h^*_\lambda=\text{argmin}_{h\in H}\mathbb{E}\|h(\mu(Z))-Y\|^2+\lambda\|[h\|^2, \tag{61}$$

where $\|[h\|^2$ is a penalty term for function h. Indeed, $\hat{Y}=g(T)=[h\phi](T)=[h\mu](Z)$. Analogously, in 2SLS $\hat{Y}=\mathbb{E}[Y\mid T]=\hat{\beta}T$ for stage 2 linear regression parameter $\hat{\beta}$.

Dual IV.
Inspired by stochastic programming [91], [92], DualIV [62] shows that two-stage IV-based regression can be reformulated as a convex-concave saddle-point problem. Then, [62] develops a simple kernel-based algorithm and simplifies traditional two-stage methods via a dual formulation.

Based on the outcome structural function, the expectation of Eq. (47) w.r.t. Y conditioned on $\{Z,\mathbf{X}\}$ yields [43]:

$$\begin{aligned}\mathbb{E}[Y\mid Z,\mathbf{X}] &= \mathbb{E}[g(T,\mathbf{X})\mid Z,\mathbf{X}]+\mathbb{E}[\epsilon_Y\mid\mathbf{X}]\\ &= \int g(T,\mathbf{X})dF(T\mid Z,\mathbf{X}),\end{aligned} \tag{62}$$

where, $dF(T\mid Z,\mathbf{X})$ is the conditional treatment distribution obtained from the treatment regression. [62] reformulate the equation as an empirical risk minimization problem:

$$\min_{g\in\mathcal{G}}R(g)=\mathbb{E}_{YZ}\left[\ell\left(Y,\mathbb{E}_{T|Z,\mathbf{X}}[g(T,\mathbf{X})]\right)\right] \tag{63}$$

where $\ell(y,y')=(y-y')^2$ denotes the mean squared error.

Applying the interchangeability and Fenchel duality [91], [92] to Eq. (63):

$$\begin{aligned}R(g)&=\mathbb{E}_{YZ\mathbf{X}}\left[\max_{u\in\mathbb{R}}\left\{\mathbb{E}_{T|Z,\mathbf{X}}[g(T,\mathbf{X})]u-\ell^\star(Y,u)\right\}\right]\\ &=\max_{u\in\mathcal{U}}\mathbb{E}_{Z\mathbf{X}Y}\left[\mathbb{E}_{T|Z,\mathbf{X}}[g(T,\mathbf{X})]u(Y,Z,\mathbf{X})-\ell^\star(Y,u(Y,Z,\mathbf{X}))\right]\\ &=\max_{u\in\mathcal{U}}\mathbb{E}_{Z\mathbf{X}TY}[g(T,\mathbf{X})u(Y,Z,\mathbf{X})]-\mathbb{E}_{Z\mathbf{X}Y}\left[\ell^\star(Y,u(Y,Z,\mathbf{X}))\right]\end{aligned}$$

where $\mathcal{U}(\Omega)=\{u(\cdot):\Omega\to\mathbb{R}\}$ is the entire space of functions defined on the support Ω, and Ω is the corresponding space of random variables $\mathcal{Y}\oplus\mathcal{Z}\oplus\mathcal{X}$. $\ell:\mathbb{R}\times\mathbb{R}\to\mathbb{R}_+$ is a proper, convex, and lower semi-continuous loss function for any value in its first argument and $\ell^\star_y=\ell^\star(y,\cdot)$ is a convex conjugate of $\ell_y=\ell(y,\cdot)$.

To simplify notation, in this section, we denotes by $W=Y\oplus Z\oplus\mathbf{X}$ and $T=T\oplus\mathbf{X}$. Then, the saddle-point problem is:

$$\min_{g\in\mathcal{G}}\max_{u\in\mathcal{U}}\mathbb{E}_{TW}[g(T)u(W)]-\mathbb{E}_W\left[\ell^\star(Y,u(W))\right] \tag{64}$$

With $\ell^\star(y,u)=uy+\frac{1}{2}u^2$, DualIV reduce the traditional two-stage methods as:

$$\min_{g\in\mathcal{G}}\max_{u\in\mathcal{U}}\Psi(g,u), \tag{65}$$
$$\Psi(g,u)=\mathbb{E}_{TW}\{[g(T)-Y]u(W)\}-\tfrac{1}{2}\mathbb{E}_W[u(W)^2]. \tag{66}$$

Motivated by the reproducing kernel Hilbert spaces (RKHSs) [90], DualIV introduces positive definite kernels $k:\mathcal{T}\times\mathcal{T}\to\mathbb{R}$ and $l:\mathcal{W}\times\mathcal{W}\to\mathbb{R}$ for $\mathcal{G}$ and $\mathcal{U}$, respectively. [93] introduces the canonical feature maps:

$$\phi:t\mapsto k(t,\cdot),\varphi:w\mapsto l(w,\cdot). \tag{67}$$

The objective can be rewritten as:

$$\begin{aligned}\Psi(f,u) &= \mathbb{E}_{TW}[f(T)u(W)]\\ &- \mathbb{E}_{YZ}[Yu(Y,Z)]-\frac{1}{2}\mathbb{E}_W\left[u(W)^2\right]\\ &= \langle\mathcal{C}_{WT}f-\mathbf{b},u\rangle_\mathcal{U}-\frac{1}{2}\langle u,\mathcal{C}_W u\rangle_\mathcal{U}.\end{aligned} \tag{68}$$

where $\mathbf{b} := \mathbb{E}_{YZ}[Y\varphi(Y,Z)] \in \mathcal{U}, \mathcal{C}_W := \mathbb{E}_W[\varphi(W) \otimes \varphi(W)] \in \mathcal{U} \otimes \mathcal{U}$ is a covariance operator, and $\mathcal{C}_{WT} := \mathbb{E}_{WT}[\varphi(W) \otimes \phi(T)] \in \mathcal{U} \otimes \mathcal{F}$ is a cross-covariance operator. The generalized least squares solution in RKHS is:

$$\begin{aligned} f^* &= \arg\min_{f \in \mathcal{F}} \frac{1}{2} \left\langle \mathcal{C}_{WT} f - \mathbf{b}, \mathcal{C}_W^{-1} \left(\mathcal{C}_{WT} f - \mathbf{b} \right) \right\rangle_{\mathcal{U}} \\ &= \left(\mathcal{C}_{TW} \mathcal{C}_W^{-1} \mathcal{C}_{WT} \right)^{-1} \mathcal{C}_{TW} \mathcal{C}_W^{-1} \mathbf{b} \end{aligned} \tag{69}$$

Eq. (69) gives a solution for IV-based regression in closed form.

4.3.2 Deep-based Estimator

Originally, 2SLS performs linear regressions in both stages under linearity assumption. Recent machine learning methods extend it to non-linear settings with infinite dictionaries of basis functions from reproducing kernel Hibert spaces (RKHS), such as KernelIV [61] and DualIV [62]. Although these methods enjoy desirable theoretical properties, the flexibility of the model is limited, since the basis functions are pre-specified by human-hand or feature engineering [17], [81].

DeepIV.
DeepIV builds upon deep-based methods, i.e., deep neural networks [17]. Although there is little theory to justify when learning with neural networks can identify a true model, deep methods make substantially weaker assumptions about the data generating process and automatically learn flexible feature mappings for high-dimension and non-linear data, which saves the human effort selecting pre-defined basis functions and improves the accuracy of causal effect estimation. Under additive noise assumption or linearity assumption, [17] provide an unique solution for the inverse problem with the learned representation, as follows.

Taking the expectation of both sides of Eq. (47) conditioned on $\{Z, \mathbf{X}\}$ and applying assumptions formulates the relationship [43]:

$$\begin{aligned} \mathbb{E}[Y \mid Z, \mathbf{X}] &= \mathbb{E}[g(T, \mathbf{X}) \mid Z, \mathbf{X}] + \mathbb{E}[\epsilon_Y \mid \mathbf{X}] \\ &= \int g(T, \mathbf{X}) dF(T \mid Z, \mathbf{X}), \end{aligned} \tag{70}$$

where, again, $dF(T \mid Z, \mathbf{X})$ is the conditional treatment distribution obtained from the treatment regression. The relationship defines an inverse problem in structural function identification. Given observational data $\{z_i, \mathbf{x}_i, t_i, y_i\}$, the counterfactual functions are recovered by minimizing the objective:

$$\hat{g} = \text{argmin}_{g \in \mathcal{G}} \sum_{i=1}^{n} \left(y_i - \int_t g(t, x_i) dF(t \mid z_i, \mathbf{x}_i) \right)^2. \tag{71}$$

Furthermore, in estimation, DeepIV develops a two-stages procedure. To obtain the the conditional probability estimation $dF(t \mid z_i, \mathbf{x}_i)$ of treatments, deep methods use conditional density estimation model as treatment regression module in stage 1 [17], [94]. Then, they perform a joint mapping from re-sampled treatments $\hat{T}$ and confounders $\mathbf{X}$ to the counterfactual outcomes Y in stage 2.

Treatment Regression Stage. Specifically, we use a deep neural network $\pi_\phi(Z, \mathbf{X})$ with parameters ϕ to model the conditional density function of treatment $F(T \mid Z, \mathbf{X})$. The objective can be written as:

$$\min \mathcal{L}_1 = l(T, \pi_\phi(Z, \mathbf{X})), \tag{72}$$

where $l(T, \pi_\phi(Z, \mathbf{X}))$ would be an l_2-loss for continuous outcomes or a log-loss for binary outcomes. For discrete treatments T, we model $\pi_\phi(Z, \mathbf{X})$ with $P(T = k) = \pi_{\phi,k}(Z, \mathbf{X})$ for each treatment arm $T = k$ and where $\pi_{\phi,k}(Z, \mathbf{X})$ is given by the k-th element of softmax output in a DNN. For continuous treatments T, we model a mixture of Gaussian distributions with component $\pi_{\phi,k}(Z, \mathbf{X})$ and sub-networks $[\mu_{\phi,k}(Z, \mathbf{X})], \sigma_{\phi,k}(Z, \mathbf{X})]$ for Gaussian distribution parameters $G(\mu, \sigma)$. With enough mixture components, the network $\pi_\phi(Z, \mathbf{X})$ can approximate arbitrary smooth densities.

Outcome Regression Stage. We model a counterfactual prediction network h_θ with parameters θ, to approximate the potential outcome. The objective can be written as:

$$\min \mathcal{L}_2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \int_t h_\theta(t, x_i) d\hat{F}_\phi(t \mid z_i, x_i) \right)^2, \tag{73}$$

where $\hat{F}_\phi(T \mid Z, \mathbf{X})$ is from the stage 1. Then, we can can optimize the $\hat{F}_\phi(T \mid Z, \mathbf{X})$ and $h_\theta(T, \mathbf{X})$ by minimizing the loss $\mathcal{L}_1(\phi)$ and $\mathcal{L}_2(\theta)$ using gradient descent, respectively.

OneSIV.
However, existing deep-based methods require two stages to separately estimate the conditional treatment distribution and the potential outcome function, which is not sufficiently effective [80]. Lin et al. [80] claims that the information from the outcome regression is one significant component for joint distribution of observations, and we should utilizing this information to improve the conditional treatment distribution estimation.

One Stage Regression. Further, they merge the two stages to leverage the outcome regression $h_\theta(T, \mathbf{X})$ to the treatment distribution estimation $\hat{F}_\phi(T \mid Z, \mathbf{X})$ through a cleverly designed deep neural network structure. Then, they present a joint trade-off objective, as follows:

$$\min_{\phi,\theta} w_1 \mathcal{L}_1 + w_2 \mathcal{L}_2, \tag{74}$$

where w_1 and w_2 are the hyper-parameters to control the relative importance of treatment regression $\mathcal{L}_1$ (Eq. 72) and outcome regression $\mathcal{L}_2$ (Eq. 73) obtained from DeepIV. Minimizing this objective, the treatment regression network and the outcome regression network can promote each other's evolution, i.e., Co-evolution.

DFIV.
Combining the theoretical advantages of kernel-based methods and the empirical advantages of deep learning methods, DFIV [81] uses deep neural networks (DNNs) to adaptively learn deep features as kernel basis in the 2SLS approach, which fits structural functions with highly nonlinear flexibility. [81] develops three DNNs $\{f_\phi, g_\xi, u_\psi\}$ to learn the corresponding feature mappings for $\{Z, \mathbf{X}, T\}$, respectively.

Similar to Eqs. (48)(49), we can reformulate the IV-based regression as:

$$u_{\psi,k}(T) = \sum_{i=1}^{d^Z}\sum_{j=1}^{d^X} \alpha_{i,j}^k f_{\phi,i}(Z) g_{\xi,j}(\mathbf{X}) + \epsilon_T, \tag{75}$$

$$Y = \sum_{k=1}^{d^T}\sum_{j=1}^{d^X} \beta_{k,j} u_{\psi,k}(T) g_{\xi,j}(\mathbf{X}) + \epsilon_Y, \tag{76}$$

where $f_{\phi,i}(Z)$ denotes the i-th element in the outcome vector of instrument representation network $f_\phi(Z)$, $g_{\xi,j}(\mathbf{X})$ is the j-th element in the outcome vector of covariate representation network $g_\xi(\mathbf{X})$, and $u_{\psi,k}(T)$ is the k-th element in the outcome vector of treatment representation network $u_\psi(T)$. $\{d^Z, d^X, d^T\}$ denotes the dimension of the outcome vector $f_\phi(Z)$, $g_\xi(\mathbf{X})$, and $u_\psi(T)$. $\mathbf{A} = [\alpha_{i,j}^k]_{i,j,k}$ and $\mathbf{B} = [\beta_{i,j}]_{i,j}$ denote the corresponding coefficients in the linear associations between features $\{f_\phi(Z), g_\xi(\mathbf{X}), u_\psi(T), Y\}$.

Treatment Regression Stage. Fixing the parameter ψ of the treatment representation network $u_\psi(\cdot)$ and the parameter ξ of the covariate representation network $g_\xi(\cdot)$ during stage 1, DFIV aims to regress the conditional expectation $\mathbb{E}[u_\psi(T) \mid f_\phi(Z) \otimes g_\xi(\mathbf{X})]$ by learning the network parameter ϕ and the coefficient matrix $\mathbf{A} \in \mathbb{R}^{d^T \times (d^Z \cdot d^X)}$, where $f_\phi(Z) \otimes g_\xi(\mathbf{X})$ denotes the multiplication combination set $[f_{\phi,i}(Z) g_{\xi,j}(\mathbf{X})]_{i,j}$.

$$\phi^* = \text{argmin}_\phi \mathcal{L}_1(\phi), \tag{77}$$

$$\mathcal{L}_1(\phi) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \left[\|u_\psi(t_i) - \mathbf{A} f_\phi(z_i) \otimes g_\xi(x_i)\|^2 + \lambda_1 \|\mathbf{A}\|^2\right] \tag{78}$$

$$\mathbf{A}(\phi) = u_\psi(T)'\mathbf{C}(\mathbf{C}'\mathbf{C} + n\lambda_1 I)^{-1} \tag{79}$$

To simplify notation, in this section, we denotes by $\mathbf{C} = f_\phi(Z) \otimes g_\xi(\mathbf{X}) \in \mathbb{R}^{n \times (d^Z \cdot d^X)}$. We can then learn the parameters ϕ of the instrument representation network $f_\phi(\cdot)$ by minimizing the loss $\mathcal{L}_1(\phi)$ using gradient descent.

Outcome Regression Stage. Fixing the parameters ϕ of the instrument representation network $f_\phi(\cdot)$ and the parameter ξ of the covariate representation network $g_\xi(\cdot)$ during stage 2, DFIV predicts the structural function $\mathbb{E}[Y \mid u_\psi(T) \otimes g_\xi(\mathbf{X})]$ by learning the network parameter ψ and the coefficient matrix $\mathbf{B} \in \mathbb{R}^{1 \times (d^T \cdot d^X)}$. To simplify notation, in this section, we use $\mathbf{D} = f_\psi(T) \otimes g_\xi(\mathbf{X})$ denotes the multiplication combination set $[f_{\psi,k}(T) g_{\xi,j}(\mathbf{X})]_{k,j}$.

$$\psi^* = \text{argmin}_\psi \mathcal{L}_2(\psi), \tag{80}$$

$$\mathcal{L}_2(\psi) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \left[\|y_i - \mathbf{B} f_\psi(t_i) \otimes g_\xi(x_i)\|^2 + \lambda_2 \|\mathbf{B}\|^2\right] \tag{81}$$

$$\mathbf{B}(\psi) = Y'\mathbf{D}(\mathbf{D}'\mathbf{D} + n\lambda_2 I)^{-1} \tag{82}$$

We can then learn the parameters ψ of the treatment representation network $f_\psi(\cdot)$ by minimizing the loss $\mathcal{L}_2(\psi)$ using gradient descent.

Note that the covariate representation network $g_\xi(\cdot)$ is fixed during stage 1 and stage 2. To update the covariate network $g_\xi(\cdot)$, fixing the parameters ψ and ϕ, we minimize the loss $\mathcal{L}_1(\xi) + \mathcal{L}_2(\xi)$ using gradient descent. Then, we adopt an alternating training strategy to iteratively optimize the representations for $g_\psi(\cdot)$, $g_\phi(\cdot)$ and $g_\xi(\cdot)$.

4.3.3 GMM-based Estimator

In the presence of heteroskedasticity, although the counterfactual function estimation of the standard IV estimators and some variants is consistent with the true potential outcomes, the standard errors are inconsistent, preventing valid inference [95]. Assuming observational data can be formalized in moment conditions, when facing heteroskedasticity of unknown form, we can make use of the conditional moment restrictions to allow for efficient estimation. That is, instrumental variable regression and 2SLS can be seen as special cases of generalized method of moments (GMM), introduced by [96], which is a prototypical (non-)parametric estimator [97]–[99].

The standard IV estimator is a special case of GMM. Satisfying the IV assumptions, the instruments Z is correlated with the endogenous treatments T and orthogonal to the unmeasured confounders ϵ_T/ϵ_Y at the same time, i.e., $Z \perp U$. Then, we can design a IV-based GMM estimator to satisfy the orthogonality conditions with the overidentified context. Under the additive noise assumption (Eq. (46)(47)), the moment conditions for instruments $Z \in \mathbb{R}^{n \times d^Z}$ can be formulated as $\mathbb{E}[Z\epsilon_T] = \mathbb{E}[Z\epsilon_Y] = 0$. The d^Z instruments give a set of d^Z moments:

$$l_i(g) = z_i' u_i = z_i'(y_i - g(t_i, x_i)), i = 1, \cdots, n \tag{83}$$

$$\mathbb{E}[l(g)] = \frac{1}{n}\sum_{i=1}^n l_i(g) = \frac{1}{n}\sum_{i=1}^n z_i'(y_i - g(t_i, x_i)) = \mathbf{0}. \tag{84}$$

where $\mathbb{E}[l(g)]$ is a d^Z vector, and we set $L_j = \mathbb{E}[l(g)]_j$ to denote the j-th element in the expectation error vector $\mathbb{E}[l(g)]$. The intuition of GMM is to choose an estimator for function g, and set these d^Z moments as close to zero as possible.

In the estimation of potential outcome function, if the number of unknown parameter is exactly d^Z, the estimated equation is exactly identified the d^Z moment conditions and the d^Z parameters in regression function. If we have less unknown parameters than conditional moment restrictions, then the estimated equation is overidentified, and we cannot find a prediction function g to set all d^Z sample moment conditions $[L_j = \mathbb{E}[l(g)]_j]_{j=1,\cdots,d^z}$ to exactly zero. Thus, GMM estimator replace the theoretical expected value $\mathbb{E}[\cdot]$ with its empirical analogsample average:

$$\mathcal{J}(\theta) = \sum_{j=1}^{d^Z} L_j^2(\theta) = \|L(\theta)\|^2 = L(\theta)' W L(\theta) = \sum_{j=1}^{d^Z} [l(g_\theta)]_j]^2. \tag{85}$$

where $W = I$ is an identify matrix, meaning the average effect. Then we minimize the norm of this expression with respect to function g_θ. The minimizing function of g_θ is our estimate for g.

Although GMM is an incredibly flexible estimator, in practical, there are an infinite number of moment conditions with IV independence assumptions. Imposing all of them is infeasible with finite data. Therefore, recent literature proposes a series of minimax approaches to reformulate the minimax optimization problem.

Minimax Approachs.

There has also been a recent surge in interest with minimax approaches that reformulate conditional moment conditions as a minimax optimization problem. For example, Lewis & Syrgkanis (2018); Zhang et al. (2020) use the reformulation $\sup_{h \in L_2(Z_i)} \left(\mathbb{E}\left[h\left(Z_i\right)\left(Y_i - f^*\left(T_i, \mathbf{X}_i\right)\right)\right]\right)^2$. Bennett et

al. (2019), Bennett & Kallus (2020), Muandet et al. (2020), while Dikkala et al. (2020), Chernozhukov et al. (2020), and Liao et al. (2020), employ other reformulations, i.e., $\sup_{h\in L_2(Z_i,\mathbf{X}_i)}\left(\mathbb{E}\left[h\left(Z_i,\mathbf{X}_i\right)\left(Y_i-f^*\left(T_i,\mathbf{X}_i\right)\right)\right]\right)^2$.

With the rapid development of machine learning algorithms, researchers apply adaptive non-parametric learners such as reproducing kernel Hilbert spaces, random forests, and neural networks to reformulate GMM estimation to the minimax optimization problem [63], [82], [100]. In machine learning and statistics, researchers formulate the target estimand as an objective minimization problem. Then, Lewis et al. [63] formulate the expectation minimization problem as the maximum moment deviation over the set of potential functions, refered as Adversarial GMM (AGMM):

$$h^* = \text{arginf}_{h\in\mathcal{H}}\sup_{f\in\mathcal{F}}\mathbb{E}[(Y-h(T,\mathbf{X}))f(Z,\mathbf{X})]. \quad (86)$$

Similar to Wasserstein and MMD GANs [101], [102], the formulation proposes a learner network h to set moments as close to zero as possible, and an adversary network f to identify moments that are violated for the chosen h. [63] offers main theorems and applications for several hypothesis spaces of practical interest including reproducing kernel Hilbert spaces (RKHS), functions defined via shape restrictions, random forests, and neural networks.

Given observational data $\{z_i, x_i, t_i, y_i\}_{i=1,\cdots,n}$, to obtain optimal h_ϕ and f_ψ, AGMM [63] minimizes the empirical analogue of the minimax objective:

$$\begin{aligned}\phi^* &= \text{arginf}_{\phi\in\Phi}\sup_{\psi\in\Psi}\mathbb{E}[(Y-h_\phi(T,\mathbf{X}))f_\psi(Z,\mathbf{X})] \\ &- \lambda_1\|\psi\|^2-\mathbb{E}[f_\psi(Z,\mathbf{X})^2]+\lambda_2\|\phi\|^2. \quad (87)\end{aligned}$$

where $\{\lambda_1,\lambda_2\}$ are the hyper-parameters for penalty items $\|\phi\|^2$ and $\|\psi\|^2$.

DeepGMM.
With infinite moment conditions, using identify matrix I as unweighted vector norm can lead to significant inefficiencies in the minimization of objective Eq. (85) [96], [103]. [96], [103] claim that weighting moment conditions by their inverse covariance would yield minimal variance estimates, and it is sufficient to consistently estimate this covariance. Based on the optimally weighted Generalized Method of Moments (GMM) [82], [96], [104], DeepGMM [82] construct an optimal combination of moment conditions via adversarial training, with the objective:

$$\begin{aligned}\phi^* &= \text{arginf}_{\phi\in\Phi}\sup_{\psi\in\Psi}\mathbb{E}[(Y-h_\phi(T,\mathbf{X}))f_\psi(Z,\mathbf{X})] \\ &- \frac{1}{4}\mathbb{E}[(Y-h_\phi(T,\mathbf{X}))^2 f_\psi^2(Z,\mathbf{X})]. \quad (88)\end{aligned}$$

Notably, DeepGMM [82] has a few tuning parameters: the models $\mathcal{F}$ and $\mathcal{H}$ (i.e., the neural network architectures) and whatever parameters the optimization method uses. Besides, other reformulations of minimax problem are developed by [105], [106]

4.3.4 Confounder Balance Estimator

With the development of machine learning, instrumental variables are no longer limited to simple linear models. The recent IV models described above have focused on various complex setting, where interactions between various variables may exist, such as $T = ZX + X + U$. At this point, if we do not consider the joint effect of modeling covariates and IVs, the effect of IVs on the treatment variables will be very limited, i.e., weak IV. Therefore, these algorithms combine observed confounders and IVs to predict the conditional distribution of the treatments to eliminate unmeasured confounding bias in stage 1. However, this introduces additional bias due to imbalanced covariates X on different treatment arms in stage 2 (Fig. 9).

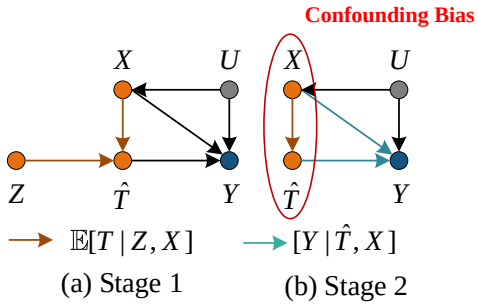


Fig. 9: Confounding bias from observed confounders.

CBIV.
Wu et al. [54] focus on treatment effect estimation with IV regression under homogeneity assumptions, and they propose a Confounder Balanced IV Regression (CB-IV) algorithm to further remove the confounding bias from observed confounders by balancing in nonlinear scenarios.

Based on the Homogeneous Instrument-Treatment Assumption, Wu et al. [54] model a more general causal relationship by relaxing the additive assumption to multiplicative assumption on response-outcome function as:

$$T = f_1(Z,X)+f_2(X,U) \quad (89)$$
$$Y = g_1(T,X)+g_2(T)g_3(U)+g_4(X,U), Z\perp U, X \quad (90)$$

where $f_i(\cdot), g_j(\cdot)$ are unknown and potentially non-linear continuous functions. $g_2(T)g_3(U)$ denotes the multiplicative terms of U with T (e.g., $U^2T-UT+U$). The completeness of $\mathbb{P}(T\mid Z,X)$ and $\mathbb{P}(Y\mid T,X)$ guarantees uniqueness of the solution [43].

The CB-IV algorithm contains the following three main components:

Treatment Regression in Stage 1: For continuous treatment T, CBIV regresses treatment T with IVs Z and observed confounders X.

$$\mathcal{L}_T = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(t_i-\hat{t}_i^j\right)^2, \hat{t}_i^j\sim\hat{P}(t_i|z_i,x_i), \quad (91)$$

we sample m (the larger the better) treatment $\{\hat{t}_i^j\}_{j=1,\ldots,m}$ for each unit $\{z_i,x_i\}$ to approximate the true treatment t_i. Empirically, the above objective (Eq. (91)) is sufficient to accurately estimate causal effects in continuous CB-IV framework.

Confounder Balance in Stage 2: For continuous treatment T, we learn a "balanced" representation (i.e., C) of the observed confounders X as $C = f_\theta(X)$ via mutual information (MI) minimization constraints: firstly, we use variational distribution $Q_\psi(\hat{T}\mid C) = \mathcal{N}(\mu_\psi(C),\sigma_\psi(C))$

parameterized by neural networks $\{\mu_\psi, \sigma_\psi\}$ to approximate the true conditional distribution $P(\hat{T} \mid C)$; then, we minimize the log-likelihood loss function of variational approximation $Q_\psi(\hat{T} \mid C)$ with n samples to estimate MI:

$$\text{disc}(\hat{T}, C) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\log Q_\psi \left(\hat{t}_i \mid c_i \right) - \log Q_\psi \left(\hat{t}_j \mid c_i \right) \right]. \quad (92)$$

where, $C = f_\theta(X)$. We adopt an alternating training strategy to iteratively optimize $Q_\psi(\hat{T} \mid C)$ and the network $C = f_{\theta(X)}$ to implement balanced representation in the Confounder Balancing.

Outcome Regression: Finally, we propose to regress the outcome with the estimated treatment $\hat{T} \sim P(T|Z, X)$ obtained in treatment regression module and the representation of confounders $C = f_\theta(X)$ obtained in confounder balancing module:

$$\mathcal{L}_Y = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - h_\xi(\hat{t}_i, f_\theta(x_i)) \right)^2 \quad (93)$$

where $\hat{t}_i \sim \hat{P}(T|Z, X)$ and $f_\theta(x_i)$ are derived from treatment regression module and confounder balancing module, respectively.

Theoretically and empirically, CBIV confirms that eliminating confounding bias in the outcome regression stage will contribute to more accurate treatment effect estimation.

4.4 Limitation and Future Work

4.4.1 Limitation

Invalid IV. The above methods are reliable only if the predefined IVs are valid and strongly correlated with the treatment variable. However, such valid IVs are hardly satisfied due to the untestable exclusion association with outcome [54]. Therefore, we have to rely on expert knowledge to select the instrumental variables, but this often does not guarantee the validity of the instrumental variables: IV does not have a direct effect on the outcome variable, only indirectly through the treatment variable. As an alternative, in instrumental variable literature, researchers usually implement Randomized Controlled Trials (RCTs) to obtain exogenous IVs, such as Oregon health insurance experiment [23] and effects of military service on lifetime earnings [24], which are too expensive to be universally available.

Weak IV and Mis-specified Model. In the real world, ones always consider a large number of variables (i.e., pretreatment variables) that are relevant to the outcome and then choose treatments in the hope of obtaining the optimal results. The instrumental variables are usually only a few, or even non-existent. Besides, the potential mechanisms of data generation are complex, and there may be interactions between various variables, such as $T = ZX + X + U$. In other words, IVs may have little causal effect on the treatment variables, which we call weak IV. Therefore, machine learning algorithms tend to combine observed confounders and IVs to predict the conditional distribution of the treatments to eliminate unmeasured confounding bias. Wu et al. [54] points out that these methods would make the predicted treatments $\hat{T}$ correlate with the observed variables X and imbalanced variables X will bring additional confounding bias for outcome regression, if the outcome model is misspecified (Fig. 9).

Limited Sample. Machine learning algorithms are data-driven algorithms, and their performance is highly dependent on the number of samples. When the sample size is infinite, we can obtain unbiased estimates by the above algorithm. However, in finite samples, machine learning algorithms are prone to overfitting, leading to errors in the regression of the intervening variables, which will further lead to failure in the regression of the resulting coefficients. In addition, imbalanced covariates can also induce overfitting and introduce sample selection bias.

4.4.2 Future Work

Causal Discovery. When we have access to a large number of variables, we can try to mine the instrumental variables from the data by using causal discovery algorithms with latent variable, including constraint-based methods, score-based methods and model-based methods, such as SCORE [107].

Generalized Method of Moments. GMM is an incredibly flexible IV estimator that relies on a large number of moment conditions with IV independence conditions. With the advancement of machine learning algorithms, nonlinear independence detection algorithms have also been developed, which has outperformed first-order moment independent etc. Therefore, a natural idea is to use independent testing algorithms instead of moment conditions to constrain the instrumental variable regression, such as HSIC-X [108]

Confounder Balance. In the presence of unmeasured confounders and the above IV methods raises a very interesting bias problem in non-linear IV methods. These methods would suffer from the bias from the observed confounders, which are imbalanced in the second stage of IV regression. To address this problem, CBIV [54] proposes a confounder balanced IV regression algorithm by a novel combination of the confounder balancing and IV regression, where the confounder balancing is designed for removing the bias from the observed confounders and the IV regression is for removing the bias from the unobserved variables. In the provided theoretical analyses and numerical experiments, [54] demonstrates the effectiveness of the proposed algorithm. In the future, confounder balance is an issue that has to be considered in instrumental variable regression.

5 Control Function

Another statistical method to correct for unmeasured confounding bias is control function (CFN), also know as two-stage residual inclusion. The principle of control function can be traced back to some early works[5] [109], [110], a control function is a variable that renders known cause variables (i.e., Treatments) appropriately exogenous in the outcome regression [110]–[112]. In observational data, conditional on control function or confounders[6], CFN estimator makes the treatment appropriately exogenous in the regression queation. CFN is a two-stage residual inclusion method, which deponds on the parameters estimated by

5. Based on [66].

6. Under the unconfoundedness assumption, the role of control function in regression is consistent with that of confounding variables

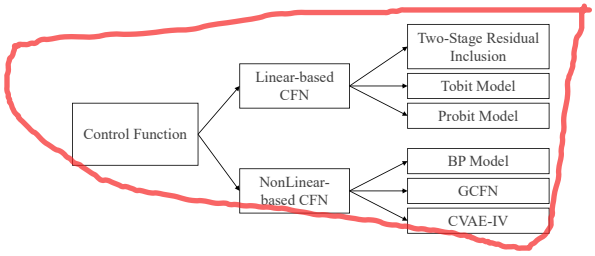


Fig. 10: Categorization of Control Function Estimators.

treatments T and valid IVs Z in stage 1 [69]. And, it is not only useful in linear cases, but also in the non-linear scenarios to elimate bias for endogeneity.

In stage 1, based on the variation induced by exogenous IVs in the treatment regression from IVs Z to treatments T, we can obtain a generalized residual that serves as control function. As for stage 2, conditional on control function estimated in stage 1, the treatment become appropriately exogenous in the outcome regression. Next, we show how CFN regression works in cansal inference and machine learning, including linear and non-linear scenarios, as shown in Fig. 10.

5.1 Linear-based CFN

5.1.1 Control Function Estimations

For the most part, the usage of CFN maintains the spirit of the earlier definitions and estimations [66]. In the presence of unmeasured confounders $\mathbf{U}$, we assume $\mathbf{V} = f(\mathbf{U})$ as unmeasured noise for treatments and model structural linearity function in constant coefficients:

$$T = Z\alpha + f(\mathbf{U}) = Z\alpha + \mathbf{V}, \tag{94}$$
$$Y = T\beta + \mathbf{U} = T\beta + f^{-1}(\mathbf{V}), \tag{95}$$

where instrumental variables are independent of unmearsured confoudners, i.e., $\mathbb{E}(Z\mathbf{U}) = 0$ and $\mathbb{E}(\mathbf{U} \mid Z) = \mathbb{E}(\mathbf{U})$. Similarity, the IVs are uncorrelated with $f(\mathbf{U})$. In linearity, we model the $\mathbf{U}$-T association (i.e., the residuals) as $\mathbf{V} = f(\mathbf{U}) = \mathbf{U}/\rho$, and f^{-1} is the inverse function of association f. Then we can obtain:

$$f^{-1}(\mathbf{V}) = \rho\mathbf{V}, \tag{96}$$

where ρ is the population regression coefficient. We plug it into the Eq. (95):

$$Y = T\beta + \rho\mathbf{V}. \tag{97}$$

In the observational data $\mathcal{D} = \{Z, \mathbf{U}, T, Y\}$, we do not observe $\mathbf{U}$ or the residuals $\mathbf{V} = f(\mathbf{U})$. Nevertheless, based on Eq. (94), we can get $\mathbf{V} = T - Z\alpha$. Because Z is uncorrelated with $\mathbf{V}$ in the linear model, we can consistently estimate the coefficient α by OLS. The two-step control function procedure is as follows:

The Residual Learning Stage: in stage 1 of CFN, we perfrom the regression of the treatments T on exogenous IVs Z:

$$\hat{\alpha} = (Z'Z)^{-1}Z'T = (Z'Z)^{-1}Z'(Z\alpha + \mathbf{V}) = \alpha, \tag{98}$$
$$\hat{\mathbf{V}} = T - Z\hat{\alpha} = \mathbf{V}. \tag{99}$$

The Outcome Regression Stage: in stage 2 of CFN, based on the association between residuals $\mathbf{V}$ and unmeasured confounders $\mathbf{U}$, we can regard residuals $\mathbf{V}$ as a control function for unmeasured confounders. Then we can control the residuals $\mathbf{V}$ to estimate the conditional average causal effect of treatments T on outcomes Y:

$$\text{CATE} = \mathbb{E}[Y(T = t) - Y(T = 0) \mid \mathbf{V}]. \tag{100}$$

or dose-response function (ITE):

$$\text{ITE} = Y(T = t, \mathbf{V}) - Y(T = 0, \mathbf{V}). \tag{101}$$

The coefficents on Z and T from CFN estimator arenumerically identical to that of 2SLS estimator [113]. In above linear setting, CFN estimator does not lead to a novel estimator different from 2SLS. In fact, if we perform OLS in the outcome regression, we find it is hard to obtain unbiased causal effect and we need to control the CFN/confounders.

5.1.2 Binary/Discrete treatment effects

Binary/Discrete Treatment $T = \{0, 1\}$ is a special case for CFN. When the treatment is a binary random variable, that is also applicable to discrete variables, a choice is to utilize the binary nature of treatment T and replace the linear regression with a binary response model. The structural equation is supplemented with the continuous models in Eqs. (94) and (95):

$$T = \mathbb{1}\{Z\alpha + \mathbf{V} > 0\}, \tag{102}$$
$$Y = T\beta + \mathbf{U}, \tag{103}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, and $\{\mathbf{U}, \mathbf{V}\}$ are independent of Z. There is a linear causal relationship between $\mathbf{U}$ and $\mathbf{V}$. Without loss of generality, we assume that the residual satisfies $\mathbf{V} \sim \mathcal{N}(0, 1)$. Thus, the treatment assignment can be regarded as a probit model:

$$P(T = 1 \mid Z) = \Phi(Z\alpha), \tag{104}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Then we can derive a CFN for binary treatment cases [65], [98].

In stage 1, we estimate the probit model in Eq. (104) and obtain the *generalized residual*:

$$r_{\hat{\mathbf{V}}} = T\lambda(Z\alpha) - (1 - T)\lambda(-Z\alpha) \tag{105}$$

where $\lambda(\cdot) = \frac{\phi}{\Phi}(\cdot)$ is the well-known inverse Mills ratio [114].

In stage 2, we control the generalized residual $r_{\mathbf{V}}$ to estimate the conditional average causal effect of treatments T on outcomes Y:

$$\text{CATE} = \mathbb{E}[Y(T = t) - Y(T = 0) \mid r_{\mathbf{V}}]. \tag{106}$$

One limitation for CFN in binary/discrete treatment cases is that the results is reliable only when the designed probit model for T is correct. If the probit model is correctly specified, then the CFN estimator would give an unbias causal effect.

5.1.3 Heterogeneous treatment effects

When the coefficients in the structural function is correlated the treatment variable, there are heterogeneous treatment effects in observational data. The random coeffient setting is called a "correlated random coefficient" (CRC) model [115], [116]. Consider the outcome structural function as:

$$T = Z\alpha + \mathbf{V}, \tag{107}$$
$$Y = TU_1 + U_2, \tag{108}$$

where all unobservables are independent of IVs, i.e., $Z \perp \{U_1, U_2, \mathbf{V}\}$, and the unobservables U_1 and U_2 are linearly correlated with the residual $\mathbf{V}$:

$$\mathbb{E}[U_1 \mid \mathbf{V}] = \eta\mathbf{V} + \beta, \mathbb{E}[U_2 \mid \mathbf{V}] = \psi\mathbf{V} + c, \tag{109}$$

where $\{\beta, c\}$ are constant terms, and $\{\eta, \psi\}$ are the corresponding regression coefficients.

In the heterogeneous treatment effects dataset, there are two sources of unmeasured confounding bias from U_1 and U_2. In this cases, we focus on the average treatment effect, i.e., $\beta = \mathbb{E}(U_1)$. Then, we set $U_1 = \mathbb{E}(U_1) + R, \mathbb{E}(R) = 0$, and reformulate the outcome structural function as:

$$Y = T\beta + TR + U_2, \tag{110}$$

where $\mathbb{E}[R] = \eta\mathbf{V}$ and the correlation between T and R satisfies the assumption: $\text{Cov}(T, R \mid Z) = \text{Cov}(T, R)$ [116]. Then we formulate the CFN estimator as:

$$\begin{aligned}\mathbb{E}[Y(T) \mid U_1, U_2] &= \mathbb{E}[Y(T) \mid \mathbf{V}, T\mathbf{V}] \\ &= T\beta + \eta T\mathbf{V} + \psi\mathbf{V} + c. \end{aligned} \tag{111}$$

In stage 1, we regress the treatment T on the exogenous IVs Z:

$$\hat{\alpha} = (Z'Z)^{-1} Z'T = (Z'Z)^{-1} Z'(Z\alpha + \mathbf{V}) = \alpha \tag{112}$$

Thus, the residual is.

$$\hat{\mathbf{V}} = T - Z\hat{\alpha} = \mathbf{V}. \tag{113}$$

In stage 2, we control the residual $\mathbf{V}$ and the multiplicative interaction $T\mathbf{V}$ to estimate the conditional average causal effect of treatments T on outcomes Y:

$$\text{CATE} = \mathbb{E}[Y(T = t) - Y(T = 0) \mid \mathbf{V}, T\mathbf{V}]. \tag{114}$$

Similar CFNs are also applicable to discrete treatment cases.

5.2 NonLinear-based CFN

In the previous section, we have introduced contron function methods employed for linear models, including Probit and Tobit. [66], [117], [118] broaden the scope of the CFN applications. Here, we detail the flexibility of the CFN estimator in the complex non-linear models using machine learning methods.

Consider a simple nonlinear model (observed confounders $\mathbf{X}$ includes a multiplicative interaction $T\mathbf{X}$):

$$T = Z\alpha_1 + \mathbf{X}\alpha_2 + \mathbf{V}, \tag{115}$$
$$Y = \mathbf{X}\beta_1 + T\mathbf{X}\beta_2 + \mathbf{U}, \mathbf{U} = \mathbf{V}\rho. \tag{116}$$

According to the IV's three conditions, we have that $Z \perp \{\mathbf{X}, \mathbf{U}, \mathbf{V}\}$. In this model, the treatment is continuous, then we obtain the residual in the stage 1.

$$\hat{\mathbf{V}} = T - \mathbb{E}[T \mid Z, \mathbf{X}] = T - (Z, \mathbf{X})(\hat{\alpha_1}, \hat{\alpha_2})' = \mathbf{V} \tag{117}$$

where $(Z, \mathbf{X})$ denotes the joint vector of Z and $\mathbf{X}$, and $(\hat{\alpha_1}, \hat{\alpha_2})$ is the corresponding coefficients. Sequentially, we can perform the outcome regression on $\mathbf{V}$, $\mathbf{X}$, and the interaction $T\mathbf{X}$:

$$\text{CATE} = \mathbb{E}[Y(T = t) - Y(T = 0) \mid \mathbf{V}, \mathbf{X}, T\mathbf{X}]. \tag{118}$$

A similar estimator can be built for a discrete treatment case, in the discrete model Eq. (106) [65], [119]. The limitation is that the results are reliable only when we have modeled the correct model for non-linear relationship with the prior knowledge of interaction $T\mathbf{X}$. In the next section, we will give a general solution through probit models.

5.2.1 Non-Parametric BP Estimator

For more general models, there may be some more complex non-linear relationship in the causal structural function. Based on the probit model [118], Blundell and Powell (BP) [64] proposes a non-parametric extension of the Rivers-Vuong approach [118], which is applicable in most general setting:

$$T = f(Z, \mathbf{X}) + \mathbf{V}, \tag{119}$$
$$Y = g(\mathbf{X}, T, \mathbf{U}). \tag{120}$$

where $f(\cdot)$ and $g(\cdot)$ are the structural functions. The target of BP approach is to estimate the Average Structural Function (ASF) of outcome, defined as follows:

$$\text{ASF}(\mathbf{X}, T) = \mathbb{E}[g(\mathbf{X}, T, \mathbf{U}) \mid \mathbf{X}, T]. \tag{121}$$

The notation means that the unmeasured confounders $\mathbf{U}$ are averaged out in the population conditional on the fixed $\mathbf{X}$ and T, i.e., $\mathbb{E}_{\mathbf{U}}[g(\mathbf{X}, T, \mathbf{U})] = \mathbb{E}[g(\mathbf{X}, T, \mathbf{U}) \mid \mathbf{X}, T]$.

BP Model.

In the first stage, we can obtain the residual $\mathbf{V}$ from $\mathbf{V} = T - f(Z, \mathbf{X})$, and $f(Z, \mathbf{X})$ can be identified by $f(Z, \mathbf{X}) = \mathbb{E}[T \mid Z, \mathbf{X}]$:

$$\hat{\mathbf{V}} = T - \mathbb{E}[T \mid Z, \mathbf{X}] = T - \hat{f}(Z, \mathbf{X}). \tag{122}$$

where we can use machine learning methods to estimate the expectation $\mathbb{E}[T \mid Z, \mathbf{X}]$, such as kernel-based regression and neural networks regression.

In the second stage, the conditional distribution of the unmeasured confounders $\mathbf{U}$ is related to $\{Z, \mathbf{X}, T\}$ only through the residual $\mathbf{V}$ [66], [120]:

$$P(\mathbf{U} \mid Z, \mathbf{X}, T) = P(\mathbf{U} \mid Z, \mathbf{X}, \mathbf{V}) = P(\mathbf{U} \mid \mathbf{V}). \tag{123}$$

Then, the consistent estimator of the ASF is:

$$\hat{g}'(\mathbf{X}, T, \mathbf{V}) = \mathbb{E}[Y \mid \mathbf{X}, T, \mathbf{V}], \tag{124}$$
$$\text{ASF}(\mathbf{X}, T) = \mathbb{E}[\hat{g}'(\mathbf{X}, T, \mathbf{V}) \mid \mathbf{X}, T] = \mathbb{E}_{\mathbf{V}}[\hat{g}'(\mathbf{X}, T, \mathbf{V})], \tag{125}$$
$$\hat{\text{ASF}}(\mathbf{X}, T) = \frac{1}{n}\sum_{i=1}^{n} \hat{g}'(\mathbf{X}, T, \mathbf{V}). \tag{126}$$

where we can use machine learning methods to estimate the expectation $\hat{g}'(\mathbf{X}, T, \mathbf{V})$, such as kernel-based regression and neural networks regression.

5.2.2 General CFN Estimator

Althrough CFN estimators have been widely used for solving the unmeasured confounders in causal inference, one critical limitation is that CFN usually breakdown under complex non-linear models. Besided, CFN requires that the residual obtained from the treatment outcome regression is linearly related to the unmeasured confounder, i.e., the structural treatment process assumptions, and the results is reliable only when the models are specified correctly.

Based on the concept of variational autoencoder (VAE) [121], some works study the proxy variable for unmeasured confounders and try to use the proxy to reconstruct the unmeasured confounders [122]–[124]. Motivated by this, [67] develop the general control function method (GCFN) to construct general control functions and estimate effects.

With the control function that satisfies the ignorability and positivity assumptions, GCFN does not need the additive separation assumption and simplify the causal effect estimation as outcome regression on the treatment and the control function. The observation data can be sampled from:

$$T = f(Z, \mathbf{X}, \mathbf{V}), \tag{127}$$
$$Y = g(\mathbf{X}, T, \mathbf{U}). \tag{128}$$

Then, the control functions can be characterized:

Theorem 5.1. ***Meta-identification.*** *The causal effect is identified by the joint distribution $q(Z, \mathbf{X}, \mathbf{V}, T)$ over the control function $\hat{\mathbf{V}}$ and the observables $\{Z, \mathbf{X}, T\}$:*

$$\mathbb{E}_{\hat{\mathbf{V}}}[Y \mid T, \hat{\mathbf{V}}] = \mathbb{E}_{\hat{\mathbf{V}}}[Y \mid do(T), \hat{\mathbf{V}}] = \mathbb{E}[Y \mid do(T)]. \tag{129}$$

With the following assumptions:

- *(A1) $\hat{\mathbf{V}}$ satisfies the reconstruction property: the treatment T can be represented by $\{Z, \mathbf{X}, \hat{\mathbf{V}}\}$;*
- *(A2) The IVs Z are independent of control functions, confounders and residuals, i.e., $Z \perp \{\mathbf{X}, \mathbf{U}, \mathbf{V}, \hat{\mathbf{V}}\}$;*
- *(A3) Fixing the general control function $\hat{\mathbf{V}}$, the strong IVs can set treatment to any value.*

Then, the control function $\hat{\mathbf{V}}$ satisfies ignorability and positivity:

$$q(Y \mid T, \hat{\mathbf{V}}) = q(Y \mid \mathrm{do}(T), \hat{\mathbf{V}}), \tag{130}$$
$$q(\hat{\mathbf{V}}) > 0 \Rightarrow q(T \mid \hat{\mathbf{V}}) > 0. \tag{131}$$

GCFN.

Following [19], [122] and [125], GCFN's first stage called variational decoupling (VDE) constructs general control functions by using VAE and recovering the residual variation in the treatment given the IV. This yields an evidence lower bound (ELBO) of VAE to reconstruct the latent variables:

$$\begin{aligned} & L(\theta, \phi, \xi \mid Z, \mathbf{X}, T) \\ = \; & (1+\lambda)\mathbb{E}_{q_\theta(\mathbf{V} \mid Z, \mathbf{X}, T)} \log p_\phi(T \mid Z, \mathbf{X}, \mathbf{V}) \\ - \; & \lambda D_{KL}(q_\theta(\mathbf{V} \mid Z, \mathbf{X}, T) \| p_\xi(\mathbf{V})) \end{aligned} \tag{132}$$

where λ is the hype-parameter that is used to balance the reconstruction term and the KL term in the beta-VAE. $p_\xi(\mathbf{V})$ and $p_\phi(T \mid Z, \mathbf{X}, \mathbf{V})$ are real (posterior) probability distributions, $q_\theta(\mathbf{V} \mid Z, \mathbf{X}, T)$ is the estimated probability distributions by neural networks with parameter θ. $D_{KL}(\cdot)$ denotes the Kullback-Leibler (KL) divergence. By maximizing the above objective function, we can sample the control function $\hat{\mathbf{V}}$ from the observables $\{Z, \mathbf{X}, T\}$.

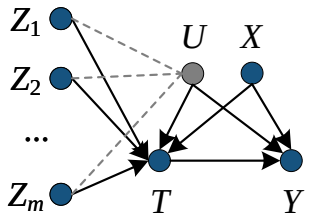


Fig. 11: Causal graph of Confounded IV. Dashed lines represent unknown causality.

VDE provides a general control function $\hat{\mathbf{V}}$ and its marginal distribution $q_\theta(\mathbf{V})$. Using VDE's control function, GCFN's second stage estimates effects via regression. Other confounder adjusting/control methods like matching/balancing methods [12], [39], [126], doubly robust methods [10] and representation learning methods [16], [36], [37], [127] can be used for outcome regression:

$$\mathrm{CATE} = \mathbb{E}[Y(T=t) - Y(T=0) \mid \mathbf{V}, \mathbf{X}]. \tag{133}$$

Further, [67] develop semi-supervised GCFN to construct general control functions using subsets of data that have both IV and confounders observed as supervision; this needs no structural treatment process assumptions.

5.2.3 Conditional Variational Autoencoder Estimator

Due to untestable Exclusion and Independent restrictions, finding a valid IV is always a tricky problem. To relax the restriction, Wang et al. [128] focus on estimating treatment effects with more accessible confounded instruments that violate the unconfounded instruments assumption, i.e., $\{Z_1, Z_2, \cdots, Z_m\} \not\perp \mathbf{U}$. Inspired by deep conditional variational autoencoder, they aim to generate a substitute of unmeasured confounder that obeys strong ignorability, such that $Y \perp T \mid \mathbf{U}, \mathbf{X}$. To achieve the ignorability, CVAE-IV [128] model a substitute $\hat{\mathbf{V}}$ based on the statistical principle $Y \perp \{Z_i\}_{i=1}^m \mid T, \mathbf{X}, \hat{\mathbf{V}}$, which states that the outcome and IV candidates are conditionally independent given the treatment, observed covariates and the generated $\hat{\mathbf{V}}$.

CVAE-IV.

In the first stage, with multiple confounded IVs $\mathbf{Z} = \{Z_i\}_{i=1}^m$, as shown in Fig. 11, CVAE-IV [128] constructs a conditional variational autoencoder to generate the confounder substitute $\hat{\mathbf{V}}$. Specifically, they apply the variational inference to model the conditional distribution $P(Y, \mathbf{Z} \mid T, \mathbf{X})$ as follow:

$$\begin{aligned} \log P(Y, \mathbf{Z} \mid T, \mathbf{X}) \geq \; & \mathbb{E}\left[\log P_\theta\left(Y, \mathbf{Z} \mid T, \mathbf{X}, \hat{\mathbf{V}}\right)\right] \\ & - D_{KL}\left(Q_\phi\left(\hat{\mathbf{V}} \mid T, Y, \mathbf{Z}, \mathbf{X}\right) \| P\left(\hat{\mathbf{V}} \mid T, \mathbf{X}\right)\right) \end{aligned} \tag{134}$$

where D_{KL} refers to the KL-divergence between variational posterior and the underlying one, P_θ is the decoder model and Q_ϕ is the encoder model. By forcing the underlying posterior $P\left(\hat{\mathbf{V}} \mid T, \mathbf{X}\right)$ to follow the normal distribution.

We use networks f_Y and f_Z to regress the outcome and instruments as well as minimize the evidence lower bound (ELBO) of CVAE as objective to reconstruct the latent variables $\hat{\mathbf{V}}$:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{Rec} + \mathcal{L}_{Chol} + \lambda\mathcal{L}_{KL} \quad (135)\\ \mathcal{L}_{Rec} &= \sum_i^n [(y_i - f_Y(t_i, \mathbf{x}_i, \mathbf{v}_i))^2]/Var(Y),\\ \mathcal{L}_{Chol} &= \sum_i^n [(\mathbf{z}_i - f_Z(t_i, \mathbf{x}_i, \mathbf{v}_i))^2],\\ \mathcal{L}_{KL} &= D_{KL}\left(Q_\phi\left(\hat{\mathbf{V}} \mid T, Y, \mathbf{Z}, \mathbf{X}\right) \| P\left(\hat{\mathbf{V}} \mid T, \mathbf{X}\right)\right),\end{aligned}$$

where the λ controls the variance of the reconstructed output.

In the second stage, we fit the observational outcome using two regression functions g_{ψ_1} and g_{ψ_2} , which are parametrized by deep networks with ψ_1 and ψ_2:

$$\mathcal{L}_{Reg} = \sum_i^n [(y_i - g_{\psi_1}(t_i, \mathbf{x}_i) - g_{\psi_2}(\mathbf{v}_i))^2]. \quad (136)$$

Then, we predict the counterfactual outcome $Y(t)$ and CATE with the trained regression model $\{\psi_1, \psi_2\}$:

$$Y(t, \mathbf{x}, \mathbf{v}) = g_{\psi_1}(t, \mathbf{x}) + g_{\psi_2}(\mathbf{v}), \quad (137)$$
$$CATE = Y(t, \mathbf{x}, \mathbf{v}) - Y(0, \mathbf{x}, \mathbf{v}). \quad (138)$$

By constructing the CVAE-IV model to generate a ignorable confounder substitute, we isolate the influence of the unmeasured confounder from the estimation on conditional treatment effect.

5.3 Limitation and Future Work

5.3.1 Limitation

Inverse Relationship. In the structural assumption, CFN implicitly require a one-to-one mapping (or Inverse Relationship) between the residuals $\mathbf{V}$ from treatment regression and the unmeasured confounders $\mathbf{U}$. Otherwise, even if we recover the residuals perfectly, we cannot control the unmeasured confounders. For example, if $\mathbf{V} = sin(\mathbf{U})$, then we control for $\mathbf{V} = 1$, but $\mathbf{U}$ still has infinitely many possibilities, which we cannot discuss and analyze.

Invalid IV and Weak IV. The performance of these methods relies on the well-predefined IVs that satisfy three instruments restrictions (i.e., IV does not have a direct effect on the outcome variable, only indirectly through the treatment variable), which is untestable and leads to finding a valid IV becomes an art rather than science. Therefore, how to use invalid IV or wark IV to implement CFN is still an open problem.

5.3.2 Future Work

Variational Autoencoder. Inverse relationship between the residuals $\mathbf{V}$ and the unmeasured confounders $\mathbf{U}$ means that we can achieve indirect control of $\mathbf{U}$ by controlling the residuals $\mathbf{V}$. So, naturally, why don't we just recover $\mathbf{U}$? Based on the concept of variational autoencoder (VAE) [121], some works study the proxy variable for unmeasured confounders and try to use the proxy to reconstruct the unmeasured confounders [122]–[124]. Motivated by this, [67] develop the general control function method (GCFN) to learn the distribution of unmeasured confoudners and estimate effects.

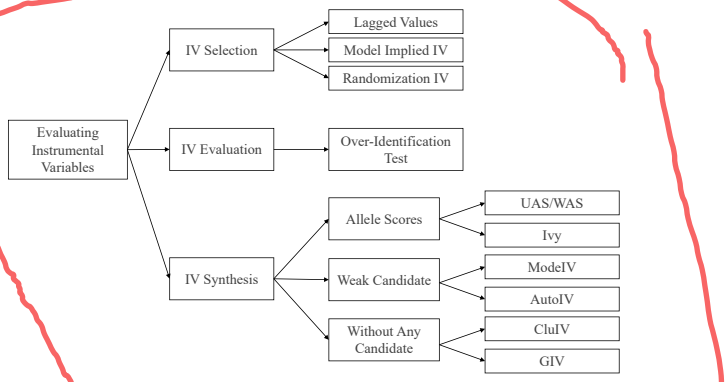


Fig. 12: Evaluating of Instrumental Variables.

Confounded IV. In reality, the acquisition of valid IV is a tricky project, so Wang et al. [128] proposes to use confounded IV, having a direct effect on the outcome variable but indirectly through the treatment and confounders, instead of valid IV to recover unmeasured confounders. By considering the conditional independence between confounded instruments and the outcomes, CVAE-IV [128] generates a substitute of the unmeasured confounder with a conditional variational autoencoder. Therefore, the exploration of invalid IV is a promising research line for the future.

6 Evaluating Instrumental Variables

In Section 4 & 5, we have introduce how to implement two-stage regression with IV for treatment effect estimation. One limitation is that, these methods require a strong and valid IV[7] for treatment regression, which is rare in reality. In this Section, we summarize three methods for selecting IV, i.e., Lagged Values, Prior Knowledge of Causal Graph and Randomized Controlled Trials, and provide over-identification test for IV's exclusion restriction. Subsequently, we also introduce several machine learning algorithms for strong IV generation, i.e., Summary IVs. The overall skeleton is shown in the Fig. 12.

6.1 IV Selection

The above IV methods are reliable if and only if the IVs we found only affect the outcomes through its strong association with treatments. Finding suitable IVs still is a challenge for the IV methods [130]. Next, we will introduce several methods to find or test IVs.

Lagged Values. With panel data, a common strategy of finding IV is to use the lagged values as IVs for the current treatments [131]. For example, [132] estimated the causal effect of compulsory schooling on earnings by using quarter of birth as an IV for education. [20] used characteristics of

7. The instrument must be correlated with the endogenous treatment variables. If this correlation is strong, then the instrument is said to have a strong first stage. A weak correlation may provide misleading inferences about parameter estimates and standard errors [129].

the respondent's childhood, husband's childhood, and parents and husband's parent as IVs to predict the respondent's probability to send their children to school in the future, and then used the predicted value from this model as an independent variable in the prediction of contraceptive use.

Model Implied Instrumental Variables. A second strategy draws IVs from among the observed variables is Model Implied Instrumental Variables (MIIVs), taken from [21], [22], [133]. In MIIVs, a prior knowledge of causal graph is used to build the model structure, which tells the researcher which observed variables can serve as IVs and which cannot. Closely related to the MIIV method is the directed acyclical graph (DAG), [134], [135] gave rules to select the variables that can serve as IVs: the correlation of a variable with the residual term of the outcome predict equation is zero [22].

Randomization Instrumental Variables. In instrumental variable literature, researchers usually implement Randomized Controlled Trials (RCTs) to sample a random variable as IV to intervene the received treatments, called intention-to-treat variable, such as Oregon health insurance experiment [23] and effects of military service on lifetime earnings [24], which are too expensive to be universally available. Sometimes, there might be randomization introduced by nature [136], called natural experiments, such as twin births, gender, and weather events.

6.2 IV Evaluation

Regardless of IVs selected by which prior, we must evaluate the IVs' quality: a valid IV that only affects the outcome through its strong association with treatment options, called exclusion assumption. If the structure assumptions for IV are dissatisfied and the correlation is weak, then the instrument may provide misleading inferences about parameter estimates and standard errors [129], [137].

Over-Identification Test. When the number of IVs is more than the need for just-identification, i.e., there are more IVs than the number of treatments, one can test the exogeneity of IVs. The over-identification tests construct a null hypothesis that all IVs are exogenous variables versus the alternative hypothesis that at least one IV violates exogeneity (correlates with the residuals from the two-stage IV regression). In linear setting, [138] gave a known over-identification tests for IVs:

$$p = \frac{\epsilon' \bar{Z}(\bar{Z}'\bar{Z})^{-1}\bar{Z}'\epsilon}{\epsilon'\epsilon/n} \sim \mathcal{X}^2, \tag{139}$$

where ϵ are the residuals from the two-stage IV regression, and $\bar{Z}$ is another instrumental variable (Over Identification) not involved in the regression of causal effects. Asymptotically, the test statistic p follows a chi square distribution and the degrees of freedom equal to the number of IVs beyond the need for just-identification [98]. Besides, [139] proposed a similar over-identification tests with F-distribution. [140] developed several variants for homoscedastic disturbances. Considering heteroscedastic-consistent, [96], [99] designed a test statistic for GMM-IV models.

6.3 IV Synthesis

Strong and valid IVs are hardly satisfied in practice. Fortunately, with the advent of machine learning, researchers have found some data-driven algorithms to automatically synthesize strong IV from additional data information under some assumptions. Practitioners combine more commonly available IV candidateswhich are not necessarily strong, or even valid, IVsinto a single summary that is plugged into causal effect estimators in place of an IV [27].

6.3.1 Allele Scores

In Mendelian randomization (MR) [141], a growing number of works have been proposed to synthesize a summary IV by combining widely availabel IV candidates. [142] shows that summary IV can be reproduced using summarized data on genetic assocaitions with the treatment and the outcome, and a representative approach that combines the IV candidates into a summary variable is unweighted/weighted allele scores [25], [26], [143] (UAS/WAS). UAS/WAS synthesize a summary variable of genetic contribution towards elevating the risk factor, which serve as reliable IVs to infer causal effect among clinical variables, only if genetic variants associated with a risk factor are actually all independent valid IVs [25], [144].

UAS.
In Mendelian randomization (MR), we can use genetic variants to as IV candidates for IV synthesis. We assume K genetic variants $\mathbf{G} = \{G_1, G_2, \cdots, G_K\}$ are actually independent weak IVs, and use them as IV candidates. Then we can obtain UAS:

$$UAS_{IV} = \frac{1}{K}\sum_{j=1}^{K} G_j, \tag{140}$$

where K denotes the number of IV candidates, and G_j denotes the j-th IV candidate. Factually, UAS takes the average of IV candidates.

WAS.
In addition to an unweighted standard allele score where each risk-increasing allele contributed the same value to the allele score, WAS weights each candidate based on the associations with the treatment:

$$WAS_{IV} = \frac{1}{K}\sum_{j=1}^{K} W_j G_j, \tag{141}$$

where W_j denotes the weights that are the same as the coefficients from the treatment regression stage in the 2SLS analysis. In addition, some other weight estimation methods for calculating relevance and importance can be used as an alternative.

Ivy.
Allele scores require strong assumptions, i.e., all IV candidates are weak IVs for estimation. To relax these assumptions, [27] require more than half of the variables in the IV candidates are valid, and then propose a generalized allele scores to combine valid IV candidates and invalid candidates in a robust manner, with the following steps: (1) Identify Valid IV Candidates and their Dependencies; (2) Estimate Parameters of the Candidate Model; and (3) Synthesize IV and Estimate Causal Effect.

6.3.2 Weak Candidates

Most of Allele Scores follow the assumption that IV candidates are actually all independent weak IVs, which is actually difficult to meet. In this subsection, we review some more weaker assumptions for IV Synthesis.

ModeIV.
[28] no longer requires more than half the number of valid instrumental variables in the candidate set, but proposes that each estimate in the tightest cluster of estimation points from each IV candidate is approximately causal effects and these IV candidates are valid. ModeIV [28] will iterate over all the elements in the set of instrumental variable candidates $\mathbf{G} = \{G_1, G_2, \cdot, G_K\}$ and plug G_j into the instrumental variable regression method to estimate the causal effects τ_{G_j}. Then, the outcomes $\{\tau_{G_j}\}_{j=1}^{K}$ from the valid instrumental variables must all converge to the same value, and IV candidates in the tightest cluster of estimation points just are valid IVs.

AutoIV.
Furthermore, AutoIV [29] generate IV representations based on independence conditions and mutual information, with the assumption that all variables in the IV candidates $\mathbf{G}$ are independent of the unmeasured confoudners $\mathbf{U}$, i.e., $\mathbf{G} \perp \mathbf{U}$. Given the obaservational data $D = \{\mathbf{X}, \mathbf{G}, T, Y\}$, AutoIV [29] learn a disentangled representation $\mathbf{Z} = \phi(\mathbf{G})$ based on independence conditions:

$$\begin{aligned} & \hat{\phi} = \arg\min_{\phi} (T - f(\phi(\mathbf{G}), \mathbf{X}))^2, \\ \text{s.t.} \quad & \phi(\mathbf{G}) \perp \mathbf{X}, \\ & \phi(\mathbf{G}) \perp Y \mid T, \mathbf{X}, \end{aligned} \tag{142}$$

where $f(\cdot)$ denotes a regression network of $\phi(\mathbf{G}), \mathbf{X}$ to predict treatment variables. According to the independence conditions, AutoIV [29] obtain valid IVs that does not have a direct effect on the outcome variable, only indirectly through the treatment variable.

To learn relevance and exclusion, AutoIV [29] construct a mutual information estimation network to optimize the network. Take two any random variables X and Y as an example, the log-likelihood loss function of variational approximation $Q_{\theta_{XY}}(Y|\phi_X(X))$ with n samples is given as:

$$\mathcal{L}_{XY}^{LLD} = -\frac{1}{n}\sum_{i=1}^{n} \log Q_{\theta_{XY}}(y_i|\phi_X(x_i)). \tag{143}$$

They minimize Eq. (143) to get optimal variational approximation $Q_{\hat{\theta}_{XY}}(Y|\phi_X(X))$ with parameters $\hat{\theta}_{XY}$. To increase the relevance between the IV representations and the treatment, they maximize the mutual information between them:

$$\begin{aligned} \mathcal{L}_{XY}^{MI} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (&\log Q_{\theta_{XY}}(y_i|\phi_X(x_i)) - \\ &\log Q_{\theta_{XY}}(y_j|\phi_X(x_i))), \end{aligned} \tag{144}$$

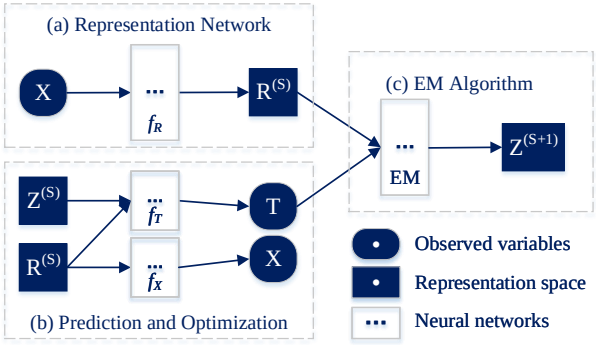


Fig. 13: Overview of Meta-EM Architecture.

Besides, they also model the conditional mutual information $\mathcal{L}_{XY|V}^{MI}$ conditional on random variable Z as:

$$\begin{aligned} \mathcal{L}_{XY|V}^{MI} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (&\omega_{ij}(\log Q_{\theta_{XY}}(y_i|\phi_X(x_i)) - \\ &\log Q_{\theta_{XY}}(y_j|\phi_X(x_i)))), \end{aligned} \tag{145}$$

where, $\omega_{ij} = \text{softmax}(e^{-\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{2\sigma^2}})$ is the conditional weight of each pair of positive and negative samples.

Based on the $\mathcal{L}_{XY}^{LLD}$ and $\mathcal{L}_{XY|V}^{MI}$ operators, AutoIV [29] (1) maximize $\mathcal{L}_{\mathbf{G}T}^{MI}$ to optimize the IV representations $\phi(\mathbf{G})$ for relevance condition; (2) minimize $\mathcal{L}_{\mathbf{G}Y|T}^{MI}$ to optimize the IV representations $\phi(\mathbf{G})$ for exclusion condition; and (3) minimize $\mathcal{L}_{\mathbf{G}\mathbf{X}}^{MI}$ to optimize the IV representations $\phi(\mathbf{G})$ for observed confounders independence condition.

6.3.3 Without Any Candidates

Limitation. Although the above IV generation methods no longer require manually selected pre-defined IVs selected, they all require a high-quality IV candidates' set with at least half valid IVs or unconfounded IV assumption, which is unrealistic in practice due to cost issues and lack of expert knowledge. These methods still cannot get rid of the dependence on predefined candidate sets. Therefore, it is highly demanded to model IVs and implement a data-driven approach to automatically obtain valid IVs directly from the observed variables $\{\mathbf{X}, T, Y\}$.

In 2021, the idea of using clustering methods to generate instrumental variables started to present, such as CluIV [145] and GIV [146]. Under a more practical setting without any candidates, GIV [146] proposes a novel algorithm (Meta-EM) to model latent GIV and implement a data-driven approach to automatically reconstruct valid Group IVs directly from the observed variables, beyond hand-made IV candidates.

GIV.
With the advent of the big data era, a variety of observation databases collected from different sources have been established, which may contain the same treatment effect mechanism (from treatment to outcome) but different treatment assignment mechanisms (from covariates to treatment). Here, the omitted source label can serve as a latent multi-valued IV, which only affects the outcome through its strong association with offer decisions.

Therefore, as shown in Fig. 13, Wu et al. [146] propose a non-linear Meta-EM to (1) map the raw data into a representation space to construct Linear Mixed Models for the assigned treatment variable; (2) estimate the distribution differences and model the GIV for the different treatment assignment mechanisms; and (3) adopt an alternating training strategy to iteratively optimize the representations and the joint distribution to model GIV for IV regression. Empirical results demonstrate the advantages of our Meta-EM compared with state-of-the-art methods.

7 Available Datasets and Codes/Packages

7.1 Datasets

In real-world applications, it's thorny to find a strictly valid instrumental variable from observational data, due to the untestable exclusion and unconfounded conditions. In short, the predefined IVs and IV candidates selected by human effort might be invalid IVs that do not strictly satisfy the conditions of the valid IVs, without enough prior knowledge for valid IVs. Besides, in observational dataset, the ground truth dose-response function (ATE, ATT, CATE or ITE) is not available, due to the lack of the counterfactual outcome. Hence, the datasets used in the IV-based works are often (semi-)synthetic datasets, such as Demand [17] and Toy Datasets [63], [82]. Some datasets combine the prior specific knowledge and the observational control dataset together to create the datasets. We detail the available benchmark datasets, as follows:

Low-dimensional Toy [63], [82]. In low-dimensional cases, [82] generated data via the following process:

$$Y = g(T) + U + \delta, T = Z + U + \gamma. \quad (146)$$
$$Z \sim \text{Uniform}(-3, 3), U \sim \mathcal{N}(0, 1), \delta, \gamma \sim \mathcal{N}(0, 0.1).$$

Similarity, [63] consider the following data generating processes:

$$Y = g(T) + U + \delta, T = \gamma Z + (1 - \gamma)U + \gamma. \quad (147)$$
$$Z \sim \mathcal{N}(0, 2), U \sim \mathcal{N}(0, 2), \delta, \gamma \sim \mathcal{N}(0, 0.1).$$

Keeping the data generating process fixed, [63], [82] design various true response function g between the following cases:

$$\textbf{sin:}\quad g(T) = sin(x), \quad \textbf{step:}\ g(T) = 0,$$
$$\textbf{abs:}\quad g(T) = |x|, \quad \textbf{linear:}\ g(T) = x.$$

MNIST [82]. Similar to [17], in high-dimensional cases, [82] use same data generating process introduced in Low-dimensional Toy, based on the MNIST dataset [147], but replace T and Z with MNIST images:

$$T := \text{RandomImage}(\pi(T)), Z := \text{RandomImage}(\pi(Z)).$$

where $\pi(t) = \text{round}(\min(\max(1.5t + 5, 0), 9))$ is a transformation function that maps input t to an integer range from 0 to 9, and the RandomImage(d) is a function that samples a image from the digit label d. The images are $28 \times 28 = 784$-dimensional digit matrices.

Demand [17]. The demand simulation design is from [17], which describes an airline scenario. In this simulation, the airline wants to estimate the effect of prices T (i.e., treatment) on passenger ticket sales Y (i.e., outcome). We assume that the fuel price Z, the customer types X_1, the time of year X_2, and the conferences U are the pre-treatment variables V, where the instrumental varialble is Z, the observable confounders are $X = \{X_1, X_2\}$ and the unmeasured confounder is U. The simulation data is generated by:

$$\begin{aligned} T &= 25 + (Z + 3)\psi(X_2) + U, && (148)\\ Y &= 100 + (10 + T)X_1\psi(X_2) - 2T + \epsilon, && (149)\\ \psi(X_2) &= 2\left(\tfrac{1}{600}(X_2 - 5)^4 + \exp[-4(X_2 - 5)^2] + \tfrac{X_2}{10} - 2\right), \\ X_1 &\in \{1, \cdots, 7\}, \quad X_2 \sim \text{unif}(0, 10), \\ Z, U &\sim \mathcal{N}(0, 1), \quad \epsilon \sim \mathcal{N}(\rho U, 1 - \rho^2). \end{aligned}$$

where, the simulation generates the latent errors ϵ with a parameter ρ that is used to smoothly vary the unmeasured confounding bias in causal model.

The target dose-response function, i.e., counterfactual function is $g(T, X) = (10 + T)X_1\psi(X_2) - 2T$.

IHDP[8] **[16], [36]** . The Infant Health and Development Program (IHDP), from a Randomized Controlled Trial (RCT), assesses whether the future cognitive of of premature infants is affected by specialist home visits. To reduce the randomness and create a observational data, [148] removed a non-random subset of the treated group to induce selection bias. The dataset comprises 747 units (139 treated, 608 control) with 25 pre-treatment variables related to the children and their mothers. The treatment is the specialist home visits and the outcome is the cognitive test scores in the future. To develop instrument variables, [146] generate 2-dimension random variables for each unit. Then, [146] select a subset of pre-treatment variables as the confounders unobserved confounders U. With known treated and control potential outcome (accessible in IHDP), [146] designs the treatment assignment policy as:

$$P(T \mid Z, X) = \frac{1}{1+\exp\left(-(\sum_{i=1}^{2} Z_i + \sum_{i=1}^{m_X} X_i) + \sum_{i=1}^{m_U} U_i)\right)}, \quad (150)$$
$$T \sim Bernoulli(P(T \mid Z, X)), Z_1, Z_2 \sim \mathcal{N}(0, 1) \quad (151)$$

where m_X and m_U are the dimensions of X and U selected from the IHDP.

PISA [149]. The PISA survey aims to evaluate the students ability to apply their knowledge and skills to real-life situations [149], covering three main domains: reading (131 items), mathematics (35 items), and science (53 items). [150], [151] selected 4951 participants in March 2009, 4041 participants in October 2009 and 3989 participants in April 2010 and there are 3472 students participated in all three rounds. The distance to school was expressed in the number of minutes is an instrument. Gender and type of school (General comprehensive, Vocational with comprehensive program, and Basic vocational school) are used as covariates.

ALSPAC[9] **[152], [153]**. The Avon Longitudinal Study of Parents and Children (ALSPAC) is a longitudinal, population-based birth cohort study from 14541 pregnant women resident in Avon, UK, with expected dates of delivery range from April 1991 to December 1992 [154]. Sim-

8. http://www.fredjo.com
9. http://www.alspac.bris.ac.uk

ilar to [152], through selection, [153] used four adiposity-associated genetic variants as IVs for estimating the effect of fat mass on kid's bone density, based on 5509 birth cohorts.

MR-base[10] **[155]**. [155] developed a MR-Base platform that integrates a curated database of complete GWAS results, which used genetic variants as instrumental variables. The database comprises 11 billion single nucleotide polymorphism-trait associations from 1673 GWAS and is under updated.

7.2 Codes/Packages

In this part, we summarize the available codes for instrumental variables and causal inference, see Table 1. Besides, we merge these codes into a tool-box **CausalDCD**.

8 Applications

In practical, unmeasured confounder is a common setting. Therefore, in the presence of unmeasured confounders, IV regression algorithms have a variety of applications in real-world scenarios.

8.1 Mendelian Randomization

According to the Mendels First and Second Laws of Inheritance, when applied to independent heritable units, genotype is independent of unmeasured confounders. Therefore, Mendelian randomization (MR) analysis (first used by [156]), using genetic variants as instrumental variables to estimate causal effects in the presence of unmeasured confounders [157]–[159], is receiving increasing attention from economists, statisticians, epidemiologists and social scientists are focus [153], [160], [161]. The growing availability in genome-wide association studies (GWAS) facilitated discovery of genetic variants, that only affects the outcomes through its strong association with treatment factors of interest [153], [162].

By comparing outcomes in patients with and without human leukocyte antigen (HLA)-compatible siblings, [156] first proposed 'Mendelian randomization' method to explore the effect of allogenic sibling bone marrow transplantation on the treatment of acute myeloid leukaemia (AML). Mendelian randomization provides one method for assessing the causal nature of some treatment exposures [158]. [152] used two independent genetic markers (FTO and MC4R genes) of obesity as IVs and found a positive effect of fat mass on bone mineral density (BMD), i.e., higher fat mass caused increased accrual of bone mass in childhood. [153] used multiple genetic variants as instrumental variables for increasing statistical precision of IV estimates and for testing underlying IV assumptions.

Use of Mendelian randomisation is growing rapidly [153], [163]. Recently, MR has been used successfully across a wide range of domains, i.e., drug target validation, drug target repurposing, side effect identification, and interpretation of high-dimensional omics studies [164], [165]. [164] reviewed recent developments in Mendelian randomization Studies and detailed the extensions to the basic MR design: including two-sample Mendelian randomization [141], [166], [167], bidirectional Mendelian randomization [168], two-step Mendelian randomization [169], multivariable Mendelian randomization [170], [171] and factorial Mendelian randomization [172]. In all, MR is a flexible and robust statistical method, which uses genetic variants as IVs to identify the causal relationships from observational studies.

10. https://www.mrbase.org/

8.2 Sociology and Social Sciences

In sociology and social sciences, the purpose of causal inference is to examine the association between social network and behaviors, also known as peer effects, social contagion or induction [173]–[175]. The peer effect means that the behavior, traits, or characteristics of an individual's peers (those he is connected to or alters) would affect his behavior [175]. Due to contextual confounding, peer selection, simultaneity bias and measurement error, [176] points that it is very difficult to estimate the peer effects from observational data but instrumental variables (IVs) can help to address these problems.

Taking the city-level characteristics serve as instruments, [177] study the effect of the neighborhood dropout rate on the individual's chance of finishing high school. To explore whether moving to a lower dropout rate would lower ones' chance of dropping out, [174] used characteristics of the local labor market (or city) as instruments and the results suggested that neighborhood conditions do influence an individuals likelihood of finishing high school.

Besides, researchers and data scientists have an increasing interest on the social network services in the Facebook, Twitter, Wechat and etc [178], which are collectively called 'social media'. Adopting an instrumental variables approach, [178] explored the effect of social network services on social capital. [178] suggested that high intensity users are higher in network social capital than non-users of social network services.

8.3 Reinforcement Learning

In reinforcement learning (RL), an agent would take actions in an environment in order to maximize the cumulative reward [179], [180]. Many concepts of reinforcement learning can be found in causal inference: the treatment is the action taken by agents, the environment can be viewed as the confounder and the cumulative reward is the outcome in causal inference [181], [182].

For reinforcement learners, the environment information is usually accessible, due to the Markov property [183], [184], which satisfies the unconfoundedness assumption [185]. To obtain an unbiased reward estimation, importance sampling weighting and doubly robust policy evaluation [186], [187] are common methods adopted in RL. Under unconfoundedness assumption, there are a substantial number of variants can estimate the state-action value (Q-function) [188]–[191].

To relax the unconfoundedness assumption, [192]–[195] introduced instrumental variables to optimize the policy for maximizing the reward. In the context of offline policy evaluation (OPE), [193] proposed improved Q-function estimators with different IV techniques and obtain competitive new techniques in recovering previously proposed OPE

TABLE 1: Available Codes of Methods for Instrumental Variables and Causal Inference.

IV-based Methods		
Method	**Language**	**Link**
DeepIV	python	https://github.com/jhartford/DeepIV
KernelIV	matlab	https://github.com/r4hu1-5in9h/KIV
DualIV	matlab	https://github.com/krikamol/DualIV-NeurIPS2020
DFIV	python	https://github.com/liyuan9988/DeepFeatureIV
DeepGMM	python	https://github.com/CausalML/DeepGMM
AGMM	python	https://github.com/microsoft/AdversarialGMM
CBIV	python	https://github.com/anpwu/CB-IV
AutoIV	python	https://github.com/junkunyuan/AutoIV
econML	python	https://github.com/microsoft/EconML
CausalDCD	python	https://github.com/anpwu/Awesome-Instrumental-Variable

methods. Using IVs, [194] derived a conditional moment restriction (CMR) and propose a IV-aided Value Iteration (IVVI) algorithm based on a primal-dual reformulation of CMR. In addition, [195] developed a new techniques to apply IV Regression to correct for the bias in RL algorithm in the presence of time-dependency noise.

8.4 Recommendation System

Another application, highly correlated with the treatment effect estimation, is recommendation system [11], [32], [189], [196], [197]. Exposing the user to an item can be viewed as a specific treatment and the user's behaviour (click or activity) is the corresponding outcome. To elimate the bias form the unmeasured confounders and the self-selection of the users, [198] proposed an instrumental variable estimate of the click-through rate, where the shock is the instrument, the treatment is exposure to the focal product, and the outcome is click-through to the recommended product. Jointly considering users behaviors in search scenarios and recommendation scenarios, [199] embedded users search behaviors as instrumental variables (IVs) and implemented a two-stage regression for an unbiased estimate of causal effect.

8.5 Computer Vision

Computer Vision is a typical field of artificial intelligence (AI), suffering from unstable learning and lacking of generalization ability [5]. To achieve a proactive defense against adversarial examples, [200] proposed to use the instrumental variable that achieves causal intervention. Using retinotopic sampling as IV [201], Causal intervention by instrumental Variable (CiiV) [200] algorithm implements a spatial data augmentation using different retinotopic sampling masks and learns features linearly responding to spatial interpolations. In Domian Adaptation, [202] claimed that the input features of one domain are valid instrumental variables for other domains. Inspired by this finding, we design a simple yet effective framework to learn the Domain-invariant Relationship with Instrumental VariablE (DRIVE) via a two-stage IV method.

9 Conclusion

9.1 Future Direction

Instrumental Variable has been an attractive research topic for a long time as it provides an effective way to uncover causal relationships in real-world problems. In this section, we point several lines for further research.

9.1.1 How to find a valid IV?

The exclusion restriction is the most critical and typically most controversial assumption underlying instrumental variables methods and we don't have any means to test it. In traditional literature, researchers implement randomized controlled trials (RCTs) to sample a random variable as IV to intervene the received treatments, which are too costly to be universally available. Therefore, its highly demanding to develop a data-driven approach to automatically obtain valid IVs. Fortunately, machine learning and Bayesian learning provide tools for modeling latent variables. Based on conditional independence test, it is likely to disentangle instrumental variables from these hidden variables. In addition, causal discovery algorithms are also a promising direction to help us automatically find instrumental variables from observed covariates.

9.1.2 How to relax the IV assumptions?

An instrument meets the following three assumptions: relevance assumption, exclusion assumption and unconfoundedness assumption. For exclusion assumption, we can use some mediators to block out the direct effect of IV on the outcomes to relax it. For confounded IV, we can also try to recover the unmeasured confounders affecting IV based on conditional independence constraints, and adjust it.

9.1.3 How to combine IV Regression with Confounder Control?

In traditional instrumental variable regression methods, researchers always ignore the bias caused by the observed confounding variables. Even by CFN, the investigators did not control for confounding of the recovered residuals. Considering confounder balance, a more robust instrumental variable regression method is a promising direction.

9.1.4 How to reduce unmeasured confounding without IV?

However, in real life, instrumental variables may not always exist, which is the norm. In the past, we have always considered observational datasets or randomized controlled experiments separately. But in fact, even if randomized controlled experiments are expensive, we can still conduct small-scale randomized controlled experiments. Considering small intervention data and a large amount of observational data,

i.e., data fusion, it is possible to establish causality without confounding bias.

9.2 Conclusion

In this survey, we provide a comprehensive review of the connection between the instrumental variable methods and machine learning models. Combined with machine learning, we mainly introduce two typical types of methods to estimate the average treatment effect: two-stage least squares (vanilla 2SLS estimator for linear models and machine learning estimator for non-linear models) and the traditional control function method (linear estimator and non-linear estimator). As IV-based framework relies on one structural assumption and three restrictions for identification of causal effects, we also review the traditional identifiability assumptions that apply in various scenarios and how to find or generate a valid IV towards these restrictions. The available benchmark datasets and open-source codes of those methods are also listed. Finally, some representative real-world applications of causal inference are introduced, such as advertising, recommendation, medicine, and reinforcement learning.

REFERENCES

[1] C. Wu, P. Jiang, C. Ding, F. Feng, and T. Chen, "Intelligent fault diagnosis of rotating machinery based on one-dimensional convolutional neural network," *Computers in Industry*, vol. 108, pp. 53–61, 2019.

[2] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph neural networks in recommender systems: a survey," *ACM Computing Surveys (CSUR)*, 2020.

[3] S. Lee and D. Kim, "Deep learning based recommender system using cross convolutional filters," *Information Sciences*, vol. 592, pp. 112–122, 2022.

[4] Y. He, Z. Shen, and P. Cui, "Towards non-iid image classification: A dataset and baselines," *Pattern Recognition*, vol. 110, p. 107383, 2021.

[5] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Towards causal representation learning," *arXiv preprint arXiv:2102.11107*, 2021.

[6] Z. Shen, P. Cui, T. Zhang, and K. Kunag, "Stable learning via sample reweighting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5692–5699.

[7] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5372–5382.

[8] J. Pearl, *Causality*. Cambridge university press, 2009.

[9] P. Cui and S. Athey, "Stable learning establishes some common ground between causal inference and machine learning," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 110–115, Feb. 2022. [Online]. Available: https://doi.org/10.1038/s42256-022-00445-z

[10] H. Bang and J. M. Robins, "Doubly robust estimation in missing data and causal inference models," *Biometrics*, vol. 61, no. 4, pp. 962–973, 2005.

[11] X. Wang, R. Zhang, Y. Sun, and J. Qi, "Doubly robust joint learning for recommendation on data missing not at random," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6638–6647.

[12] S. Athey, G. W. Imbens, and S. Wager, "Approximate residual balancing: debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 4, pp. 597–623, 2018.

[13] J. R. Zubizarreta, "Stable weights that balance covariates for estimation with incomplete outcome data," *Journal of the American Statistical Association*, vol. 110, no. 511, pp. 910–922, 2015.

[14] J. Hainmueller, "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political analysis*, vol. 20, no. 1, pp. 25–46, 2012.

[15] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.

[16] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3076–3085.

[17] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy, "Deep iv: A flexible approach for counterfactual prediction," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1414–1423.

[18] G. Imbens and J. Wooldridge, "Control function and related methods," *Whats new in Econometrics*, 2007.

[19] Z. Guo and D. S. Small, "Control function instrumental variable estimation of nonlinear causal effect models," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3448–3482, 2016.

[20] W. G. Axinn and J. S. Barber, "Mass education and fertility transition," *American Sociological Review*, pp. 481–505, 2001.

[21] K. A. Bollen and D. J. Bauer, "Automating the selection of model-implied instrumental variables," *Sociological Methods & Research*, vol. 32, no. 4, pp. 425–452, 2004.

[22] K. A. Bollen, "Model implied instrumental variables (miivs): An alternative orientation to structural equation modeling," *Multivariate behavioral research*, vol. 54, no. 1, pp. 31–46, 2019.

[23] A. Finkelstein, S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and O. H. S. Group, "The oregon health insurance experiment: evidence from the first year," *The Quarterly journal of economics*, vol. 127, no. 3, pp. 1057–1106, 2012.

[24] J. D. Angrist, "Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records," *The american economic review*, pp. 313–336, 1990.

[25] S. Burgess, D. S. Small, and S. G. Thompson, "A review of instrumental variable estimators for mendelian randomization," *Statistical methods in medical research*, vol. 26, no. 5, pp. 2333–2355, 2017.

[26] S. Burgess and S. G. Thompson, "Use of allele scores as instrumental variables for mendelian randomization," *International journal of epidemiology*, vol. 42, no. 4, pp. 1134–1144, 2013.

[27] Z. Kuang, F. Sala, N. Sohoni, S. Wu, A. Córdova-Palomera, J. Dunnmon, J. Priest, and C. Ré, "Ivy: Instrumental variable synthesis for causal inference," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 398–410.

[28] J. S. Hartford, V. Veitch, D. Sridhar, and K. Leyton-Brown, "Valid causal inference with (some) invalid instruments," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4096–4106.

[29] J. Yuan, A. Wu, K. Kuang, B. Li, R. Wu, F. Wu, and L. Lin, "Auto iv: Counterfactual prediction via automatic instrumental variable decomposition," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 16, no. 4, pp. 1–20, 2022.

[30] K. Tang, M. Tao, J. Qi, Z. Liu, and H. Zhang, "Invariant feature learning for generalized long-tailed classification," *arXiv preprint arXiv:2207.09504*, 2022.

[31] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–37, 2020.

[32] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 5, pp. 1–46, 2021.

[33] M. Kato, H. Kakehi, K. McAlinn, and S. Yasui, "Learning causal relationships from conditional moment conditions by importance weighting," *arXiv preprint arXiv:2108.01312*, 2021.

[34] R. Kohavi and R. Longbotham, "Unexpected results in online controlled experiments," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 31–35, 2011.

[35] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual reasoning and learning systems: The example of computational advertising." *Journal of Machine Learning Research*, vol. 14, no. 11, 2013.

[36] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*. PMLR, 2016, pp. 3020–3029.

[37] N. Hassanpour and R. Greiner, "Learning disentangled representations for counterfactual regression," in *International Conference on Learning Representations*, 2020.

[38] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[39] S. Li, N. Vlassis, J. Kawale, and Y. Fu, "Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns." in *IJCAI*, 2016, pp. 3768–3774.

[40] M. A. Brookhart, T. St"urmer, R. J. Glynn, J. Rassen, and S. Schneeweiss, "Confounding control in healthcare database research: challenges and potential approaches," *Medical care*, vol. 48, no. 6 0, p. S114, 2010.

[41] P. G. Wright, *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

[42] J. D. Angrist, G. W. Imbens, and D. B. Rubin, "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, vol. 91, no. 434, pp. 444–455, 1996.

[43] W. K. Newey and J. L. Powell, "Instrumental variable estimation of nonparametric models," *Econometrica*, vol. 71, no. 5, pp. 1565–1578, 2003.

[44] J. H. Stock and F. Trebbi, "Retrospectives: Who invented instrumental variable regression?" *Journal of Economic Perspectives*, vol. 17, no. 3, pp. 177–194, 2003.

[45] Ø. Hoveid, "Constructing valid instrumental variables in generalized linear causal models from directed acyclic graphs," *arXiv preprint arXiv:2102.08056*, 2021.

[46] T. Haavelmo, "The statistical implications of a system of simultaneous equations," *Econometrica, Journal of the Econometric Society*, pp. 1–12, 1943.

[47] O. Reiersøl, "Identifiability of a linear relation between variables which are subject to error," *Econometrica: Journal of the Econometric Society*, pp. 375–389, 1950.

[48] J. Pearl *et al.*, "Models, reasoning and inference," *Cambridge, UK: CambridgeUniversityPress*, vol. 19, 2000.

[49] J. Angrist and G. Imbens, "Identification and estimation of local average treatment effects," 1995.

[50] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[51] ——, "Bayesian inference for causal effects: The role of randomization," *The Annals of statistics*, pp. 34–58, 1978.

[52] ——, "Comment: Neyman (1923) and causal inference in experiments and observational studies," *Statistical Science*, vol. 5, no. 4, pp. 472–480, 1990.

[53] G. Imbens, "Instrumental variables: an econometrician's perspective," National Bureau of Economic Research, Tech. Rep., 2014.

[54] A. Wu, K. Kuang, B. Li, and F. Wu, "Instrumental variable regression with confounder balancing," in *International Conference on Machine Learning*. PMLR, 2022, pp. 24 056–24 075.

[55] J. Heckman, "Varieties of selection bias," *The American Economic Review*, vol. 80, no. 2, pp. 313–318, 1990.

[56] J. Angrist and G. Imbens, "Sources of identifying information in evaluation models," 1991.

[57] W. K. Newey, "Nonparametric instrumental variables estimation," *American Economic Review*, vol. 103, no. 3, pp. 550–56, 2013.

[58] M. A. Hernán and J. M. Robins, "Instrumental variable estimation," *Causal Inference: What If*, pp. 193–206, 2020.

[59] F. P. Hartwig, L. Wang, G. D. Smith, and N. M. Davies, "Average causal effect estimation via instrumental variables: the no simultaneous heterogeneity assumption," *arXiv preprint arXiv:2010.10017*, 2020.

[60] L. E. Mokry, O. Ahmad, V. Forgetta, G. Thanassoulis, and J. B. Richards, "Mendelian randomisation applied to drug development in cardiovascular disease: a review," *Journal of medical genetics*, vol. 52, no. 2, pp. 71–79, 2015.

[61] R. Singh, M. Sahani, and A. Gretton, "Kernel instrumental variable regression," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 4593–4605.

[62] K. Muandet, A. Mehrjou, S. Le Kai, and A. Raj, "Dual instrumental variable regression," in *NeurIPS 2020*, 2020.

[63] N. Dikkala, G. Lewis, L. Mackey, and V. Syrgkanis, "Minimax estimation of conditional moment models," in *NeurIPS 2020*, 2020.

[64] R. Blundell and J. L. Powell, "Endogeneity in nonparametric and semiparametric regression models," *Econometric society monographs*, vol. 36, pp. 312–357, 2003.

[65] A. Petrin and K. Train, "A control function approach to endogeneity in consumer choice models," *Journal of marketing research*, vol. 47, no. 1, pp. 3–13, 2010.

[66] J. M. Wooldridge, "Control function methods in applied econometrics," *Journal of Human Resources*, vol. 50, no. 2, pp. 420–445, 2015.

[67] A. Puli and R. Ranganath, "General control functions for causal effect estimation from ivs," *Advances in neural information processing systems*, vol. 33, pp. 8440–8451, 2020.

[68] G. Chamberlain, "Asymptotic efficiency in semi-parametric models with censoring," *journal of Econometrics*, vol. 32, no. 2, pp. 189–218, 1986.

[69] J. J. Heckman and R. Robb Jr, "Alternative methods for evaluating the impact of interventions: An overview," *Journal of econometrics*, vol. 30, no. 1-2, pp. 239–267, 1985.

[70] J. J. Heckman and V. J. Hotz, "Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training," *Journal of the American statistical Association*, vol. 84, no. 408, pp. 862–874, 1989.

[71] J. Durbin, "Errors in variables," *Revue de l'institut International de Statistique*, pp. 23–32, 1954.

[72] R. J. LaLonde, "Evaluating the econometric evaluations of training programs with experimental data," *The American economic review*, pp. 604–620, 1986.

[73] D. Chetverikov and D. Wilhelm, "Nonparametric instrumental variable estimation under monotonicity," *Econometrica*, vol. 85, no. 4, pp. 1303–1320, 2017.

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

[75] M. A. Brookhart and S. Schneeweiss, "Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results," *The international journal of biostatistics*, vol. 3, no. 1, 2007.

[76] L. Wang and E. Tchetgen Tchetgen, "Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 3, pp. 531–550, 2018.

[77] F. P. Hartwig, L. Wang, G. D. Smith, and N. M. Davies, "Homogeneity in the instrument-treatment association is not sufficient for the wald estimand to equal the average causal effect for a binary instrument and a continuous exposure," *arXiv preprint arXiv:2107.01070*, 2021.

[78] R. Kress, V. Maz'ya, and V. Kozlov, *Linear integral equations*. Springer, 1989, vol. 82.

[79] X. Chen and T. M. Christensen, "Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression," *Quantitative Economics*, vol. 9, no. 1, pp. 39–84, 2018.

[80] A. Lin, J. Lu, J. Xuan, F. Zhu, and G. Zhang, "One-stage deep instrumental variable method for causal inference from observational data," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 419–428.

[81] L. Xu, Y. Chen, S. Srinivasan, N. de Freitas, A. Doucet, and A. Gretton, "Learning deep features in instrumental variable regression," 2021.

[82] A. Bennett, N. Kallus, and T. Schnabel, "Deep generalized method of moments for instrumental variable analysis," *Advances in neural information processing systems*, vol. 32, 2019.

[83] A. Wald, "The fitting of straight lines if both variables are subject to error," *The annals of mathematical statistics*, vol. 11, no. 3, pp. 284–300, 1940.

[84] A. R. Gallant, "Identification and consistency in seminonparametric regression," *Advances in Econometrics*, vol. 1, pp. 145–170, 1987.

[85] X. Chen and X. Shen, "Sieve extremum estimates for weakly dependent data," *Econometrica*, pp. 289–314, 1998.

[86] J. L. Horowitz, "Applied nonparametric instrumental variables estimation," *Econometrica*, vol. 79, no. 2, pp. 347–394, 2011.

[87] X. Chen and D. Pouzo, "Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals," *Econometrica*, vol. 80, no. 1, pp. 277–321, 2012.

[88] B. Boots, G. Gordon, and A. Gretton, "Hilbert space embeddings of predictive state representations," *arXiv preprint arXiv:1309.6819*, 2013.

[89] A. Hefny, C. Downey, and G. J. Gordon, "Supervised learning for dynamical system learning," *Advances in neural information processing systems*, vol. 28, 2015.

[90] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 961–968.

[91] B. Dai, N. He, Y. Pan, B. Boots, and L. Song, "Learning from conditional distributions via dual embeddings," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1458–1467.

[92] A. Shapiro, D. Dentcheva, and A. Ruszczynski, "Lectures on stochastic programming: Modeling and theory," 2014.

[93] B. Schölkopf, A. J. Smola, F. Bach *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[94] S. Darolles, Y. Fan, J.-P. Florens, and E. Renault, "Nonparametric instrumental regression," *Econometrica*, vol. 79, no. 5, pp. 1541–1565, 2011.

[95] C. F. Baum, M. E. Schaffer, and S. Stillman, "Instrumental variables and gmm: Estimation and testing," *The Stata Journal*, vol. 3, no. 1, pp. 1–31, 2003.

[96] L. P. Hansen, "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the econometric society*, pp. 1029–1054, 1982.

[97] B. E. Hansen, "Testing for structural change in conditional models," *Journal of Econometrics*, vol. 97, no. 1, pp. 93–115, 2000.

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

[99] F. Hayashi, *Econometrics*. Princeton University Press, 2011.

[100] R. Zhang, M. Imaizumi, B. Schölkopf, and K. Muandet, "Maximum moment restriction for instrumental variable regression," *arXiv preprint arXiv:2010.07684*, 2020.

[101] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[102] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos, "Mmd gan: Towards deeper understanding of moment matching network," *Advances in neural information processing systems*, vol. 30, 2017.

[103] L. P. Hansen, J. Heaton, and A. Yaron, "Finite-sample properties of some alternative gmm estimators," *Journal of Business & Economic Statistics*, vol. 14, no. 3, pp. 262–280, 1996.

[104] A. Bennett and N. Kallus, "The variational method of moments," *arXiv preprint arXiv:2012.09422*, 2020.

[105] L. Liao, Y.-L. Chen, Z. Yang, B. Dai, M. Kolar, and Z. Wang, "Provably efficient neural estimation of structural equation models: An adversarial approach," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8947–8958, 2020.

[106] V. Chernozhukov, W. Newey, R. Singh, and V. Syrgkanis, "Adversarial estimation of riesz representers," *arXiv preprint arXiv:2101.00009*, 2020.

[107] P. Rolland, V. Cevher, M. Kleindessner, C. Russell, D. Janzing, B. Schölkopf, and F. Locatello, "Score matching enables causal discovery of nonlinear additive noise models," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 18 741–18 753. [Online]. Available: https://proceedings.mlr.press/v162/rolland22a.html

[108] S. Saengkyongam, L. Henckel, N. Pfister, and J. Peters, "Exploiting independent instruments: Identification and distribution generalization," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 18 935–18 958. [Online]. Available: https://proceedings.mlr.press/v162/saengkyongam22a.html

[109] L. G. Telser, "Iterative estimation of a set of linear regression equations," *Journal of the American Statistical Association*, vol. 59, no. 307, pp. 845–862, 1964.

[110] A. S. Goldberger, "Selection bias in evaluating treatment effects: Some formal illustrations," in *Modelling and Evaluating Treatment Effects in Econometrics*. Emerald Group Publishing Limited, 1972, reprinted 2008.

[111] B. Barnow, G. Cain, and A. Goldberg, "Selection on observables," *Evaluation Studies*, 1981.

[112] A. C. Cameron and P. K. Trivedi, *Microeconometrics: methods and applications*. Cambridge university press, 2005.

[113] J. A. Hausman, "Specification tests in econometrics," *Econometrica: Journal of the econometric society*, pp. 1251–1271, 1978.

[114] W. H. Greene, *Econometric analysis*. Pearson Education India, 2003.

[115] J. Heckman and E. Vytlacil, "Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling," *Journal of Human Resources*, pp. 974–987, 1998.

[116] D. Card, "Estimating the return to schooling: Progress on some persistent econometric problems," *Econometrica*, vol. 69, no. 5, pp. 1127–1160, 2001.

[117] R. J. Smith and R. W. Blundell, "An exogeneity test for a simultaneous equation tobit model with an application to labor supply," *Econometrica: journal of the Econometric Society*, pp. 679–685, 1986.

[118] D. Rivers and Q. H. Vuong, "Limited information estimators and exogeneity tests for simultaneous probit models," *Journal of econometrics*, vol. 39, no. 3, pp. 347–366, 1988.

[119] J. V. Terza, A. Basu, and P. J. Rathouz, "Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling," *Journal of health economics*, vol. 27, no. 3, pp. 531–543, 2008.

[120] J. M. Wooldridge, "Unobserved heterogeneity and estimation of average partial effects," *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, pp. 27–55, 2005.

[121] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[122] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," *Advances in neural information processing systems*, vol. 30, 2017.

[123] W. Zhang, L. Liu, and J. Li, "Treatment effect estimation with disentangled latent factors," *arXiv preprint arXiv:2001.10652*, 2020.

[124] P. A. Wu and K. Fukumizu, "β-intact-VAE: Identifying and estimating causal effects under limited overlap," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=q7n2RngwOM

[125] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[126] K. Kuang, L. Li, Z. Geng, L. Xu, K. Zhang, B. Liao, H. Huang, P. Ding, W. Miao, and Z. Jiang, "Causal inference," *Engineering*, vol. 6, no. 3, pp. 253–263, 2020.

[127] A. Wu, J. Yuan, K. Kuang, B. Li, R. Wu, Q. Zhu, Y. T. Zhuang, and F. Wu, "Learning decomposed representations for treatment effect estimation," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[128] H. Wang, W. Yang, L. Yang, A. Wu, L. Xu, J. Ren, F. Wu, and K. Kuang, "Estimating individualized causal effect with confounded instruments," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 1857–1867.

[129] A. Nichols *et al.*, "Weak instruments: An overview and new techniques," in *Stata 5th North American Meeting Presentation*, 2006.

[130] K. A. Bollen, "Instrumental variables in sociology and the social sciences," *Annual Review of Sociology*, vol. 38, pp. 37–72, 2012.

[131] L. Anselin, *Spatial econometrics: methods and models*. Springer Science & Business Media, 1988, vol. 4.

[132] J. D. Angrist and A. B. Keueger, "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics*, vol. 106, no. 4, pp. 979–1014, 1991.

[133] K. A. Bollen, "An alternative two stage least squares (2sls) estimator for latent variable equations," *Psychometrika*, vol. 61, no. 1, pp. 109–121, 1996.

[134] C. Brito and J. Pearl, "A graphical criterion for the identification of causal effects in linear models," *AAAI/IAAI*, vol. 2002, pp. 533–539, 2002.

[135] J. Pearl, "The foundations of causal inference," *Sociological Methodology*, vol. 40, no. 1, pp. 75–149, 2010.

[136] M. R. Rosenzweig and K. I. Wolpin, "Natural" natural experiments" in economics," *Journal of Economic Literature*, vol. 38, no. 4, pp. 827–874, 2000.

[137] B. Hansen, *Econometrics*, 2022.

[138] J. D. Sargan, "The estimation of economic relationships using instrumental variables," *Econometrica: Journal of the Econometric Society*, pp. 393–415, 1958.

[139] R. L. Basmann, "On finite sample distributions of generalized classical linear identifiability test statistics," *Journal of the American Statistical Association*, vol. 55, no. 292, pp. 650–659, 1960.

[140] J. B. Kirby and K. A. Bollen, "10. using instrumental variable tests to evaluate model specification in latent variable structural equation models," *Sociological Methodology*, vol. 39, no. 1, pp. 327–355, 2009.

[141] S. Burgess and S. G. Thompson, *Mendelian randomization: methods for using genetic variants in causal estimation*. CRC Press, 2015.

[142] S. Burgess, F. Dudbridge, and S. G. Thompson, "Combining information on multiple instrumental variables in mendelian randomization: comparison of allele score and summarized data methods," *Statistics in medicine*, vol. 35, no. 11, pp. 1880–1906, 2016.

[143] N. M. Davies, S. von Hinke Kessler Scholder, H. Farbmacher, S. Burgess, F. Windmeijer, and G. D. Smith, "The many weak instruments problem and mendelian randomization," *Statistics in medicine*, vol. 34, no. 3, pp. 454–468, 2015.
[144] P. Sebastiani, N. Solovieff, and J. Sun, "Naïve bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: not so different after all!" *Frontiers in genetics*, vol. 3, p. 26, 2012.
[145] N. Sokolovska and P.-H. Wuillemin, "The role of instrumental variables in causal inference based on independence of cause and mechanism," *Entropy*, vol. 23, no. 8, p. 928, 2021.
[146] A. Wu, K. Kuang, R. Xiong, M. Zhu, Y. Liu, B. Li, F. Liu, Z. Wang, and F. Wu, "Treatment effect estimation with unmeasured confounders in data fusion," *arXiv preprint arXiv:2208.10912*, 2022.
[147] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[148] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.
[149] A. Pokropek, "Introduction to instrumental variables and their application to large-scale assessment data," *Large-scale Assessments in Education*, vol. 4, no. 1, pp. 1–20, 2016.
[150] W. Schulz, J. Ainley, and J. Fraillon, "Iccs 2009 technical report," 2011.
[151] H. Domański, M. Federowicz, A. Pokropek, D. Przybysz, M. Sitek, M. Smulczyk, and T. Żółtak, "From school to work: Individual and institutional determinants of educational and occupational career trajectories of young poles," *ASK: Research & Methods*, vol. 21, no. 1, pp. 123–141, 2012.
[152] N. J. Timpson, A. Sayers, G. Davey-Smith, and J. H. Tobias, "How does body fat influence bone mass in childhood? a mendelian randomization approach," *Journal of Bone and Mineral Research*, vol. 24, no. 3, pp. 522–533, 2009.
[153] T. M. Palmer, D. A. Lawlor, R. M. Harbord, N. A. Sheehan, J. H. Tobias, N. J. Timpson, G. D. Smith, and J. A. Sterne, "Using multiple genetic variants as instrumental variables for modifiable risk factors," *Statistical methods in medical research*, vol. 21, no. 3, pp. 223–242, 2012.
[154] J. Golding, M. Pembrey, R. Jones *et al.*, "Alspac–the avon longitudinal study of parents and children. i. study methodology." *Paediatric and perinatal epidemiology*, vol. 15, no. 1, pp. 74–87, 2001.
[155] G. Hemani, J. Zheng, B. Elsworth, K. H. Wade, V. Haberland, D. Baird, C. Laurin, S. Burgess, J. Bowden, R. Langdon *et al.*, "The mr-base platform supports systematic causal inference across the human phenome," *elife*, vol. 7, p. e34408, 2018.
[156] R. Gray and K. Wheatley, "How to avoid bias when comparing bone marrow transplantation with chemotherapy." *Bone marrow transplantation*, vol. 7, pp. 9–12, 1991.
[157] L. Youngman, B. Keavney, A. Palmer, S. Parish, S. Clark, J. Danesh, M. Delepine, M. Lathrop, R. Peto, and R. Collins, "Plasma fibrinogen and fibrinogen genotypes in 4685 cases of myocardial infarction and in 6002 controls: Test of causality by" mendelian randomisation"," *Circulation*, vol. 102, no. 18, 2000.
[158] G. Davey Smith and S. Ebrahim, "mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease?" *International journal of epidemiology*, vol. 32, no. 1, pp. 1–22, 2003.
[159] D. C. Thomas and D. V. Conti, "Commentary: the concept of mendelian randomization," *International journal of epidemiology*, vol. 33, no. 1, pp. 21–25, 2004.
[160] G. D. Smith, "Capitalizing on mendelian randomization to assess the effects of treatments," *Journal of the Royal Society of Medicine*, vol. 100, no. 9, pp. 432–435, 2007.
[161] G. Thanassoulis and C. J. ODonnell, "Mendelian randomization: nature's randomized trial in the post–genome era," *Jama*, vol. 301, no. 22, pp. 2386–2388, 2009.
[162] S. Von Hinke, G. D. Smith, D. A. Lawlor, C. Propper, and F. Windmeijer, "Genetic markers as instrumental variables," *Journal of Health Economics*, vol. 45, pp. 131–148, 2016.
[163] N. J. Timpson, D. A. Lawlor, R. M. Harbord, T. R. Gaunt, I. N. Day, L. J. Palmer, A. T. Hattersley, S. Ebrahim, G. D. Lowe, A. Rumley *et al.*, "C-reactive protein and its role in metabolic syndrome: mendelian randomisation study," *The Lancet*, vol. 366, no. 9501, pp. 1954–1959, 2005.
[164] J. Zheng, D. Baird, M.-C. Borges, J. Bowden, G. Hemani, P. Haycock, D. M. Evans, and G. D. Smith, "Recent developments in mendelian randomization studies," *Current epidemiology reports*, vol. 4, no. 4, pp. 330–345, 2017.
[165] N. M. Davies, M. V. Holmes, and G. D. Smith, "Reading mendelian randomisation studies: a guide, glossary, and checklist for clinicians," *Bmj*, vol. 362, 2018.
[166] S. Burgess, R. A. Scott, N. J. Timpson, G. Davey Smith, and S. G. Thompson, "Using published data in mendelian randomization: a blueprint for efficient identification of causal risk factors," *European journal of epidemiology*, vol. 30, no. 7, pp. 543–552, 2015.
[167] F. P. Hartwig, N. M. Davies, G. Hemani, and G. Davey Smith, "Two-sample mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique," pp. 1717–1726, 2016.
[168] N. J. Timpson, B. G. Nordestgaard, R. M. Harbord, J. Zacho, T. M. Frayling, A. Tybjærg-Hansen, and G. Davey Smith, "C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal mendelian randomization," *International journal of obesity*, vol. 35, no. 2, pp. 300–308, 2011.
[169] S. Burgess, R. M. Daniel, A. S. Butterworth, S. G. Thompson, and E.-I. Consortium, "Network mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways," *International journal of epidemiology*, vol. 44, no. 2, pp. 484–495, 2015.
[170] S. Burgess and S. G. Thompson, "Multivariable mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects," *American journal of epidemiology*, vol. 181, no. 4, pp. 251–260, 2015.
[171] J. P. Kemp, A. Sayers, G. D. Smith, J. H. Tobias, and D. M. Evans, "Using mendelian randomization to investigate a possible causal relationship between adiposity and increased bone mineral density at different skeletal sites in children," *International journal of epidemiology*, vol. 45, no. 5, pp. 1560–1572, 2016.
[172] B. A. Ference, J. J. Kastelein, H. N. Ginsberg, M. J. Chapman, S. J. Nicholls, K. K. Ray, C. J. Packard, U. Laufs, R. D. Brook, C. Oliver-Williams *et al.*, "Association of genetic variants related to cetp inhibitors and statins with lipoprotein levels and cardiovascular risk," *Jama*, vol. 318, no. 10, pp. 947–956, 2017.
[173] C. Jencks and S. E. Mayer, "The social consequences of growing up in a poor neighborhood," *Inner-city poverty in the United States*, vol. 111, p. 186, 1990.
[174] E. M. Foster, "Instrumental variables for logistic regression: an illustration," *Social Science Research*, vol. 26, no. 4, pp. 487–504, 1997.
[175] A. J. O'Malley, F. Elwert, J. N. Rosenquist, A. M. Zaslavsky, and N. A. Christakis, "Estimating peer effects in longitudinal dyadic data using instrumental variables," *Biometrics*, vol. 70, no. 3, pp. 506–515, 2014.
[176] W. An, "Instrumental variables estimates of peer effects in social networks," *Social Science Research*, vol. 50, pp. 382–394, 2015.
[177] E. M. Foster and S. McLanahan, "An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school?" *Psychological Methods*, vol. 1, no. 3, p. 249, 1996.
[178] S. Han and K.-G. Park, "Social network services and their effects on network social capital: an instrumental variables approach," *International Journal of Mobile Communications*, vol. 18, no. 4, pp. 386–404, 2020.
[179] M. L. Minsky, *Theory of neural-analog reinforcement systems and its application to the brain-model problem*. Princeton University, 1954.
[180] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.
[181] A. Forney, J. Pearl, and E. Bareinboim, "Counterfactual data-fusion for online reinforcement learners," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1156–1164.
[182] S. J. Gershman, "Reinforcement learning and causal models," *The Oxford handbook of causal reasoning*, vol. 1, p. 295, 2017.
[183] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.
[184] C. Lei, "Deep reinforcement learning," in *Deep Learning and Practice with MindSpore*. Springer, 2021, pp. 217–243.
[185] N. Kallus and A. Zhou, "Confounding-robust policy evaluation in infinite-horizon reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 293–22 304, 2020.
[186] D. Precup, "Eligibility traces for off-policy policy evaluation," *Computer Science Department Faculty Publication Series*, p. 80, 2000.
[187] M. Dudík, J. Langford, and L. Li, "Doubly robust policy evaluation and learning," *arXiv preprint arXiv:1103.4601*, 2011.

[188] A. Swaminathan and T. Joachims, "Counterfactual risk minimization: Learning from logged bandit feedback," in *International Conference on Machine Learning*. PMLR, 2015, pp. 814–823.
[189] A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudik, J. Langford, D. Jose, and I. Zitouni, "Off-policy evaluation for slate recommendation," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[190] H. Zou, K. Kuang, B. Chen, P. Chen, and P. Cui, "Focused context balancing for robust offline policy evaluation," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 696–704.
[191] D. Kumor, J. Zhang, and E. Bareinboim, "Sequential causal imitation learning with unobserved confounders," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
[192] X. Xu, H.-g. He, and D. Hu, "Efficient reinforcement learning using recursive least-squares methods," *Journal of Artificial Intelligence Research*, vol. 16, pp. 259–292, 2002.
[193] Y. Chen, L. Xu, C. Gulcehre, T. L. Paine, A. Gretton, N. de Freitas, and A. Doucet, "On instrumental variable regression for deep offline policy evaluation," *arXiv preprint arXiv:2105.10148*, 2021.
[194] L. Liao, Z. Fu, Z. Yang, Y. Wang, M. Kolar, and Z. Wang, "Instrumental variable value iteration for causal offline reinforcement learning," *arXiv preprint arXiv:2102.09907*, 2021.
[195] J. Li, Y. Luo, and X. Zhang, "Causal reinforcement learning: An instrumental variable approach," *Available at SSRN 3792824*, 2021.
[196] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, "Recommendations as treatments: Debiasing learning and evaluation," in *international conference on machine learning*. PMLR, 2016, pp. 1670–1679.
[197] A. Lada, A. Peysakhovich, D. Aparicio, and M. Bailey, "Observational data for heterogeneous treatment effects with application to recommender systems," in *Proceedings of the 2019 ACM Conference on Economics and Computation*, 2019, pp. 199–213.
[198] A. Sharma, J. M. Hofman, and D. J. Watts, "Estimating the causal impact of recommendation systems from observational data," in *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 2015, pp. 453–470.
[199] Z. Si, X. Han, X. Zhang, J. Xu, Y. Yin, Y. Song, and J.-R. Wen, "A model-agnostic causal learning framework for recommendation using search data," *arXiv preprint arXiv:2202.04514*, 2022.
[200] K. Tang, M. Tao, and H. Zhang, "Adversarial visual robustness by causal intervention," *arXiv preprint arXiv:2106.09534*, 2021.
[201] M. J. Arcaro, S. A. McMains, B. D. Singer, and S. Kastner, "Retinotopic organization of human ventral visual cortex," *Journal of neuroscience*, vol. 29, no. 34, pp. 10 638–10 652, 2009.
[202] J. Yuan, X. Ma, K. Kuang, R. Xiong, M. Gong, and L. Lin, "Learning domain-invariant relationship with instrumental variable for domain generalization," *arXiv preprint arXiv:2110.01438*, 2021.