PDF Download
3735969.pdf
25 December 2025
Total Citations: 3
Total Downloads: 776

Published: 13 June 2025
Online AM: 15 May 2025
Accepted: 29 April 2025
Revised: 11 December 2024
Received: 04 September 2023

SURVEY

Instrumental Variables in Causal Inference and Machine Learning: A Survey

ANPENG WU, Zhejiang University, Hangzhou, Zhejiang, China

KUN KUANG, Zhejiang University, Hangzhou, Zhejiang, China

RUOXUAN XIONG, Emory University, Atlanta, GA, United States

FEI WU, Zhejiang University, Hangzhou, Zhejiang, China

Open Access Support provided by:

Emory University

Zhejiang University

ACM Computing Surveys, Volume 57, Issue 11 (November 2025)
https://doi.org/10.1145/3735969
EISSN: 1557-7341

Instrumental Variables in Causal Inference and Machine Learning: A Survey

ANPENG WU, Zhejiang University, Hangzhou, China
KUN KUANG, Zhejiang University, Hangzhou, China
RUOXUAN XIONG, Emory University, Atlanta, United States
FEI WU, Zhejiang University, Hangzhou, China

Causal inference is the process of drawing conclusions about causal relationships between variables using a combination of assumptions, study designs, and estimation strategies. In machine learning, causal inference is crucial for uncovering the mechanisms behind complex systems and making informed decisions. This article provides a comprehensive overview of using Instrumental Variables (IVs) in causal inference and machine learning, with a focus on addressing unobserved confounding that affects both treatment and outcome variables. We review identification conditions under standard assumptions in the IV literature. In this article, we explore three key research areas of IV methods: Two-Stage Least Squares (2SLS) regression, control function (CFN) approaches, and recent advances in IV learning methods. These methods cover both classical causal inference approaches and recent advancements in machine learning research. Additionally, we provide a summary of available datasets and algorithms for implementing these methods. Furthermore, we introduce a variety of applications of IV methods in real-world scenarios. Lastly, we identify open problems and suggest future research directions to further advance the field. A toolkit of reviewed IV methods with machine learning (MLIV) is available at https://github.com/causal-machine-learning-lab/mliv.

CCS Concepts: • **Computing methodologies → Machine learning**; **Causal reasoning and diagnostics;**

Additional Key Words and Phrases: Causal machine learning, instrumental variable, control function, unmeasured confounders

ACM Reference Format:
Anpeng Wu, Kun Kuang, Ruoxuan Xiong, and Fei Wu. 2025. Instrumental Variables in Causal Inference and Machine Learning: A Survey. *ACM Comput. Surv.* 57, 11, Article 292 (June 2025), 36 pages. https://doi.org/10.1145/3735969

1 Introduction

Causal inference and machine learning have emerged as two important and complementary research areas, driving innovation across a wide range of disciplines, including economics [85, 181],

This work was supported in part by the National Key Research and Development Program of China (2024YFE0203700), National Natural Science Foundation of China (62376243, 62441605), and "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02037).

Authors' Contact Information: Anpeng Wu, Zhejiang University, Hangzhou, China; e-mail: anpwu@zju.edu.cn; Kun Kuang (Corresponding author), Zhejiang University, Hangzhou, Zhejiang, China; e-mail: kunkuang@zju.edu.cn; Ruoxuan Xiong (Corresponding author), Emory University, Atlanta, Georgia, United States; e-mail: ruoxuan.xiong@emory.edu; Fei Wu, Zhejiang University, Hangzhou, Zhejiang, China; e-mail: wufei@zju.edu.cn.

ACM 0360-0300/2025/06-ART292
https://doi.org/10.1145/3735969

social sciences [111, 118], epidemiology [126, 128], healthcare [118, 133], computer vision [73, 94], and natural language processing [45, 160]. Causal inference provides a principled framework for identifying cause-and-effect relationships in complex systems, enabling reliable predictions and informed decision-making. Meanwhile, machine learning provides a rich toolkit—ranging from flexible function approximators to representation learning techniques—for modeling high-dimensional and nonlinear relationships. The integration of causal inference and machine learning has garnered increasing attention across various domains [136]. However, the presence of observed and unmeasured confounders, which are common causes of both the treatment and response variables, poses challenges in identifying causal relationships. Even when all observed confounders are controlled, systematic differences in unobserved key variables across treatment groups may still introduce bias, commonly referred to as unmeasured confounding bias [136].

In causal inference and machine learning, most existing methods rely on the strong assumption of unconfoundedness, where all relevant confounders are observed. Under this assumption, recent surveys by Guo et al. [58] and Yao et al. [177] provide comprehensive reviews of approaches for estimating causal effects from observational data, including (1) re-weighting, (2) stratification, (3) matching, (4) tree-based methods, (5) representation learning, (6) multi-task learning, and (7) meta-learning. However, when unmeasured confounding is present, these methods may fail to provide valid causal estimates. In such cases, **instrumental variable** (**IV**) methods offer a promising alternative and are increasingly being incorporated into machine learning methods [58, 177].

IVs are exogenous variables that are correlated with the treatment (explanatory variables) but affect the outcome (response variable) only through their influence on the treatment [3, 5, 66, 114]. By leveraging such instruments, researchers can isolate exogenous variation in treatment and identify causal effects even in the presence of unmeasured confounding. IV methods are foundational in econometrics and are frequently introduced in textbooks as a primary strategy for addressing endogeneity issues [31, 62, 72, 78, 167, 169]. However, these presentations often focus on linear models and traditional estimation techniques such as the **Generalized Method of Moments** (**GMM**), with limited attention to their integration with modern machine learning frameworks. In applied domains, IV methods have gained traction through **Mendelian Randomization** (**MR**), which uses genetic variants as instruments to infer causal effects of risk factors on health outcomes. Burgess et al. [28], Mokry et al. [110], and Sanderson et al. [133] provide comprehensive reviews on the widespread adoption of MR in epidemiology and genomics.

In this article, we present a comprehensive survey of IV techniques in causal inference and machine learning, emphasizing their theoretical foundations, methodological developments, and practical applications. We review the identification conditions required for IV methods and discuss how machine learning can improve traditional treatment effect estimators, such as **two-stage least squares** (**2SLS**) and the **control function** (**CFN**) approach [59, 66, 83, 84, 113]. Additionally, we examine the use of machine learning algorithms to synthesize IVs [28, 29, 67, 96, 179], and how IVs can be applied in machine learning [136, 154, 178, 184]. Several existing surveys discuss causal effect estimation methods under the unconfoundedness assumption, such as those introduced by [58, 177]. To summarize, our contributions are as follows:

- *Comprehensive review.* We survey IV methods within the potential outcomes framework, including identification conditions, two-stage regression methods, and CFN algorithms.
- *General settings.* In cases where valid instruments are not directly observable, we review methods for IV testing and data-driven IV discovery.
- *Resource aggregation.* We compile state-of-the-art methods, benchmark datasets, open-source implementations, and illustrative applications.

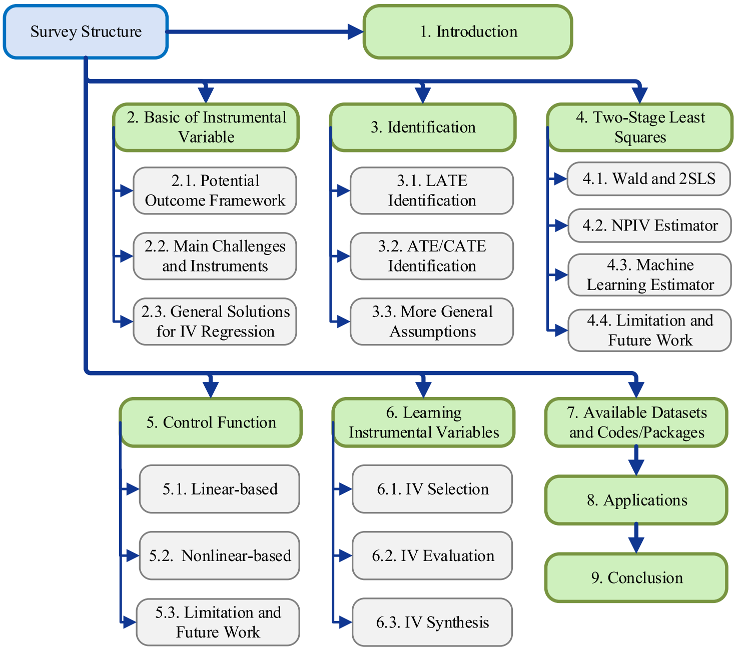

Survey Structure
1. Introduction
2. Basic of Instrumental Variable
2.1. Potential Outcome Framework
2.2. Main Challenges and Instruments
2.3. General Solutions for IV Regression
3. Identification
3.1. LATE Identification
3.2. ATE/CATE Identification
3.3. More General Assumptions
4. Two-Stage Least Squares
4.1. Wald and 2SLS
4.2. NPIV Estimator
4.3. Machine Learning Estimator
4.4. Limitation and Future Work
5. Control Function
5.1. Linear-based
5.2. Nonlinear-based
5.3. Limitation and Future Work
6. Learning Instrumental Variables
6.1. IV Selection
6.2. IV Evaluation
6.3. IV Synthesis
7. Available Datasets and Codes/Packages
8. Applications
9. Conclusion

Fig. 1. Outline of the survey.

— *Reproducibility.* We develop the MLIV platform, a curated toolkit that integrates existing implementations of IV methods in machine learning, available at https://github.com/causal-machine-learning-lab/mliv.

The structure of the article is outlined in Figure 1. In Section 2, we introduce the potential outcomes framework and IV methodology, including key definitions, assumptions, and challenges. In Section 3, we present the structural assumptions for the identification of causal effects using IV methods. We review the 2SLS-based and CFN-based methods in Sections 4 and 5, respectively. In Section 6, we review the literature on IV selection and IV synthesis. Section 7 provides practical guidance for empirical applications, followed by a review of IV-based applications in Section 8. Section 9 concludes with open challenges and future research directions.

2 Basic of IVs

The concept of IV dates back to Philip Wright's work, *The Tariff on Animal and Vegetable Oils* [170], published in 1928. In this study, Wright introduced the IV regression as a way to address the endogeneity issue in economic models, particularly when the explanatory variables are correlated with the error term. By employing IVs that are correlated with the endogenous explanatory variables but uncorrelated with the error term, Wright provided a framework for obtaining consistent parameter estimates, laying the foundation for modern econometrics [82, 150]. Later, Haavelmo [60] and Reiersøl [124] applied a similar approach in the context of errors-in-variables models and further advanced the IV framework. The classical IV estimator is commonly implemented via the 2SLS method. In the first stage, the endogenous treatment variable is regressed on the instruments to obtain predicted values. In the second stage, the outcome variable is regressed on these predicted values, yielding an estimate of the treatment effect that is robust to endogeneity.

In this article, we adopt the following notation conventions: capital letters (e.g., X) denote random variables or vectors, lowercase letters (e.g., x) denote their realizations, and calligraphic letters (e.g., $\mathcal{X}$) denote their support. The subscript i indicates the i-th unit, and bold symbols (e.g., $\mathbf{X} = [X_1, X_2, \ldots, X_n]$) denote matrices or collections of variables across n units. Next, we introduce the potential outcomes framework for IV estimation [129, 130], outline key challenges in causal effect estimation, and discuss methods for treatment effect estimation using instruments.

2.1 Potential Outcome Framework

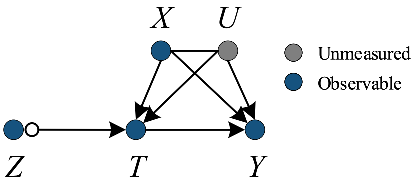


Fig. 2. The causal framework, where $\circ\!\!\rightarrow$ encodes a direct effect or a common cause of Z and T.

The Rubin causal model [129–131], also known as the potential outcome framework, provides a foundational structure for analyzing causal effects and is widely used in IV methods. In this article, we adopt this framework to study the causal effects of treatments on outcomes. For simplicity, we first consider the case of a single treatment and a single instrument. Extensions to multi-dimensional treatments and instruments will be discussed in Sections 3.3, 4.2, and 5.2. Let the observed dataset be denoted as $\mathcal{D} = \{Z, X, T, Y\}$, , as illustrated in Figure 2. Here, the treatment variable is $T \in \mathcal{T} \subset \mathbb{R}$, representing an intervention or exposure, and $Y \in \mathcal{Y} \subset \mathbb{R}$ denotes the outcome. $Z \in \mathcal{Z} \subset \mathbb{R}$ is the IV, which affects the treatment T but has no direct effect on the outcome Y other than through T. The variable $X \in \mathcal{X} \subset \mathbb{R}$ represents observed covariates that may confound the relationship between T and Y.

In practice, there may also exist unmeasured confounders $U \in \mathcal{U} \subset \mathbb{R}$ that affect both T and Y. We consider Z, X, and U to be pre-treatment variables, collectively denoted as $V \in \mathcal{V} \subset \mathbb{R}$, which may influence the assignment of treatment and the resulting outcome.

To formalize these relationships, we assume the following linear structural equations:

$$T(Z, X, U) = \alpha Z + \theta_T X + \epsilon_T, \qquad Y(T, X, U) = \beta T + \theta_Y X + \epsilon_Y \, . \tag{1}$$

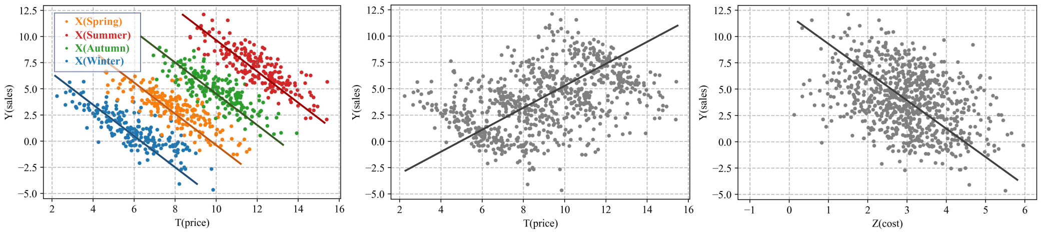
where $\alpha, \beta, \theta_T, \theta_Y$ are unknown coefficients, and β is the causal parameter of interest, representing the effect of treatment T on outcome Y. The terms ϵ_T and ϵ_Y represent random noise or unobserved disturbances, which may be arbitrarily correlated with the unmeasured confounders U. We denote $T(Z, X, U)$ and $Y(T, X, U)$ are potential treatments and potential outcomes, respectively, under fixed values of Z (or T), X, and U. For notational simplicity, we use $Y(T)$ to denote the potential outcome under treatment level T, implicitly acknowledging its dependence on X and U. When a unit receives treatment level $T = t$, the observed outcome is $Y = Y(T = t)$, referred to as the factual outcome. Conversely, $Y(T = t')$ for $t' \neq t$ represents a counterfactual outcome, corresponding to the outcome that would have been observed under an alternative treatment level.

In observational data, the goal of causal inference is to estimate treatment effects, defined as differences in potential outcomes of the form $Y(T = t) - Y(T = 0)$, which quantify the causal impact of receiving treatment level t relative to no treatment. These effects can be studied at various levels of granularity—population-wide, conditional on covariates, or at the individual level.

*Definition 2.1 (**Average Treatment Effect** (ATE)).*

$$\mathbf{ATE}(t) = \mathbb{E}[Y_i(T = t) - Y_i(T = 0)] \, , \tag{2}$$

which represents the expected effect of treatment t across the entire population.



(a) Relationship of Prices and Sales (Observed Confounder) (b) Relationship of Prices and Sales (Unmeasured Confounder) (c) Relationship of Cost and Sales (Instrumental Variable)

Fig. 3. Toy example: Exploring the causal relationship between ice cream prices and sales considering (a) Observed confounders, (b) Unmeasured confounders, and (c) Instrumental variables.

Definition 2.2 (***Conditional Average Treatment Effect*** (**CATE**)).

$$\mathbf{CATE}(t, x) = \mathbb{E}[Y_i(T = t) - Y_i(T = 0) \mid X_i = x], \tag{3}$$

which captures the expected treatment effect conditional on observed covariates $X = x$.

Definition 2.3 (***Individual Treatment Effect*** (**ITE**)).

$$\tau_i(t) = Y_i(T = t) - Y_i(T = 0), \tag{4}$$

which measures the causal effect of treatment t for a specific individual i.

These notions of treatment effect provide a comprehensive framework for understanding causal relationships, from average effects across the entire population to more granular effects at the subgroup and individual levels.

2.2 Main Challenges and Instruments

In observational data, estimating treatment effects can be challenging due to two primary issues:

(1) **Unobserved Counterfactuals:** For each individual, only the outcome corresponding to the treatment received is observed, while outcomes for alternative treatments remain unknown. This missing information prevents direct comparison of causal effects.
(2) **Confounding Bias:** Non-random treatment assignment often results in imbalances in confounders, which influence both the treatment and the outcome, leading to biased estimates.

Taking ice cream sales as a toy example, consider the treatment T as the price of ice cream and the outcome Y as the sales. Confounders such as seasons (X) may influence both T (e.g., price adjustments based on season) and Y (e.g., higher sales during hot summer days). As illustrated by the colorful lines in Figure 3(a), if the confounders are observable, we can apply causal techniques to control and balance them, concluding that sales (Y) and prices (T) are negatively correlated. Although numerous works in causality and machine learning have addressed biases from observed confounders [48, 58, 140, 177], the unmeasured confounders pose additional challenges. As shown in Figure 3(b), if we ignore unmeasured season confounders and treat all nodes as a whole, the relationship between T and Y may be biased. In such cases, if we directly apply **ordinary least squares** (**OLS**) regression, we may find that higher prices are associated with higher sales, leading to the wrong conclusion that increasing prices boost sales.

To overcome the issues from unmeasured confounders in observational data, researchers introduced IVs to estimate causal effects [66, 83, 84, 113]. In the ice cream example (Figure 3), the cost (Z) of producing ice cream serves as a valid IV, provided that consumer demand remains unaffected by the underlying production expenses. Then, we can estimate causal effect by comparing the correlation coefficients between Z and T and between Z and Y, i.e., $\beta = \frac{\mathrm{Corr}(Z,Y)}{\mathrm{Corr}(Z,T)}$. Conceptually, an IV is related to the treatment but not directly related to the outcome, except through its influence on the treatment. Formally, we follow Hernan and Robins [78] and refer to $Z \in \mathcal{Z}$ as an IV if it meets the following three instrumental conditions.

Definition 2.4 (IV). Variable Z is an IV if it satisfies the following three conditions:

(a) *Relevance*, Z is associated with T, i.e., $Z \perp T$ does not hold;
(b) *Exclusion*, Z does not directly affect Y, except through its potential effect on T, i.e., $Y(t, z) = Y(t, z') = Y(t)$ for all z, z' and all t;[1]
(c) *Exogeneity*, Z and Y do not share causes, i.e., $Y(t, z) \perp Z$ for all t and z.

When the three conditions are satisfied, the variable Z is referred to as a valid IV, and it can be used to identify treatment effects. However, if either the exclusion or the exogeneity condition is violated, Z becomes an invalid IV, leading to unreliable results from IV regression. The exclusion condition is frequently implicitly imposed when the structural equation of Y does not include Z.[2] In the ice cream example, certain production costs (Z) that are unaffected by seasonal fluctuations, such as labor costs or employee wages, influence only the pricing but do not directly affect consumer demand, making Z a valid IV. Therefore, we can use Z to isolate and estimate the causal effect of pricing on ice cream sales by leveraging the relationships between variables Z, T, and Y.

In practical applications, the exclusion and exogeneity conditions are inherently untestable. Therefore, researchers often rely on expert knowledge, structural assumptions, or randomized intent-to-treatment designs to identify and validate IVs. Furthermore, even when the three IV conditions are satisfied, additional assumptions are often required to identify causal effects. To this end, a growing literature has focused on relaxing these identification assumptions for more general scenarios [2, 3, 68, 74, 79, 113, 114], which will be discussed in Section 3.

Remark 2.5. In the special case of binary instruments and binary treatments, Angrist, Imbens, and Rubin [3, 5] suggest that the IV approach allows us to identify local treatment effects, but typically not the global effect. Specifically, in binary cases, where instrument variables (Z) are different intervention assignments and treatment variables are individuals' responses to assignments ($T(Z)$), we can define four different compliance types by the pair of values $(T(Z = 0), T(Z = 1))$ [83]:

$$i \in \begin{cases} n \text{ (never taker)} & \text{if } T_i(0) = T_i(1) = 0 \\ c \text{ (complier)} & \text{if } T_i(0) = 0, T_i(1) = 1 \\ d \text{(defier)} & \text{if } T_i(0) = 1, T_i(1) = 0 \\ a \text{(always taker)} & \text{if } T_i(0) = T_i(1) = 1 \end{cases} \tag{5}$$

In such cases, the average treatment effects for compliers, referred to as the **local average treatment effects (LATE)**, can be identified and is formally defined as follows:

Definition 2.6 (LATE).

$$\mathbf{LATE} = \mathbb{E}[Y_i(T = 1) - Y_i(T = 0) \mid i \in \mathit{complier}]. \tag{6}$$

2.3 Two General Solutions of IV Regression

IV regression is a robust framework for addressing endogeneity in causal inference. Below, we discuss two general solutions commonly employed in IV regression: 2SLS [46, 66, 112, 146] and CFN Method [19, 120, 123, 168]. Both 2SLS and CFN share similar asymptotic properties in large samples, ensuring consistency and unbiased causal effect estimation. The 2SLS approach

[1] $Y(t, z)$ denotes the potential outcomes indexed by $T = t$ and $Z = z$. Since the potential outcomes do not depend on $Z = z$ given $T = t$, for the remaining parts of this article, we only index potential outcomes by $T = t$.

[2] See for example Equation (2.1) in [114], Equation (12.4) in [62] and Equation (15.2) in [169].

is often preferred for its intuitive nature, as it directly provides regression coefficient estimates after addressing endogeneity. However, in small sample settings, CFN can outperform 2SLS in terms of efficiency, as it incorporates additional information about the endogenous structure through the residuals, allowing for a more precise adjustment of endogeneity-related biases. To illustrate the differences between these methods, consider the simplified linear models derived from Equation (1), i.e., $T = \alpha Z + \epsilon_T$ and $Y = \beta T + \epsilon_Y$, omitting covariates X for clarity. We adopt 2SLS and CFN to predict the causal parameter β.

2.3.1 2SLS. 2SLS is a two-step statistical method for addressing endogeneity and estimating causal effects in IV regression. It establishes the relationship between instruments and treatments in the first stage and then uses this relationship to estimate the effect of treatments on outcomes in the second stage. The procedure is as follows:
Stage 1: Regress treatments $\mathbf{T}$ on instruments $\mathbf{Z}$ with $\mathbf{T} = t_{1,\ldots,n}$ and $\mathbf{Z} = z_{1,\ldots,n}$:

$$\hat{\alpha} = \arg\min_{\alpha} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (t_i - \alpha z_i)^2, \tag{7}$$

Let $\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, and the predicted treatment is $\widehat{\mathbf{T}} = \mathbf{Z}\hat{\alpha} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{T} = \mathbf{P}_Z\mathbf{T}$.

Stage 2: Regress outcomes $\mathbf{Y}$ on the predicted values $\widehat{\mathbf{T}}$ with $\mathbf{Y} = y_{1,\ldots,n}$ and $\widehat{\mathbf{T}} = \hat{t}_{1,\ldots,n}$:

$$\hat{\beta}_{2SLS} = \arg\min_{\beta} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (y_i - \beta \hat{t}_i)^2, \tag{8}$$

that means $\hat{\beta}_{\text{2SLS}} = (\mathbf{T}'\mathbf{P}_Z\mathbf{T})^{-1}\mathbf{T}'\mathbf{P}_Z\mathbf{Y}$. This requires a strong linear relationship between covariates, instruments, treatments, and outcomes. Nonlinear variants of 2SLS are discussed in Section 4.

2.3.2 CFN Method. CFN estimates the effect of a treatment on the outcome by first calculating residual variables from the treatment regression stage, then using these residuals and the true treatment to estimate the outcome in the outcome regression stage:
Stage 1: Regress treatments $\mathbf{T}$ on instruments $\mathbf{Z}$ with $\mathbf{T} = t_{1,\ldots,n}$ and $\mathbf{Z} = z_{1,\ldots,n}$:

$$\hat{\alpha} = \arg\min_{\alpha} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (t_i - \alpha z_i)^2, \tag{9}$$

and the predicted residuals is: $\hat{\boldsymbol{\epsilon}} = \mathbf{T} - \mathbf{Z}\hat{\alpha} = \mathbf{T} - \mathbf{P}_Z\mathbf{T}$.
Stage 2: Regress outcomes $\mathbf{Y}$ on the predicted residuals $\hat{\boldsymbol{\epsilon}}$ with $\mathbf{Y} = y_{1,\ldots,n}$ and $\hat{\boldsymbol{\epsilon}} = \hat{\epsilon}_{1,\ldots,n}$:

$$\hat{\beta}_{CFN}, \hat{\theta}_{\epsilon} = \arg\min_{\beta,\theta_\epsilon} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (y_i - \beta t_i - \theta_\epsilon \hat{\epsilon}_i)^2, \tag{10}$$

which gives $(\hat{\beta}_{\text{CFN}}, \hat{\theta}_{\epsilon}) = ((\mathbf{T}, \hat{\boldsymbol{\epsilon}})'(\mathbf{T}, \hat{\boldsymbol{\epsilon}}))^{-1}(\mathbf{T}, \hat{\boldsymbol{\epsilon}})'\mathbf{Y}$, where $(\mathbf{A}, \mathbf{B})$ means the concatenation of vectors/matrices $\mathbf{A}$ and $\mathbf{B}$. Nonlinear variants of CFN are discussed in Section 5.

2.3.3 Development of IVs. In this review, we strive to present a comprehensive overview of IV methods in causal inference and machine learning. The development of IV methodologies is depicted in Figure 4, highlighting key milestones and innovations. Our primary focus is on machine learning-based approaches, particularly those building upon 2SLS and CFN frameworks. Furthermore, we discuss recent advancements in learning IVs from candidate sets.

3 Identification

IV methods have been widely used in causal inference to address endogeneity issues in causal effect estimation. However, even when we have a well-defined and valid instrument, identifying causal effects remains challenging without additional structural assumptions. A fundamental assumption in IV methods is the classical linearity framework, where the treatment and outcome are modeled as $T = \alpha Z + \theta_T X + \epsilon_T$, $Y = \beta T + \theta_Y X + \epsilon_Y$, as shown in Equation (1). Then, we can identify the causal parameter by $\beta = \frac{\text{Corr}(Z,Y)}{\text{Corr}(Z,T)}$. However, these structural assumptions, particularly linearity, often fail to capture complex nonlinear relationships or heterogeneous effects in

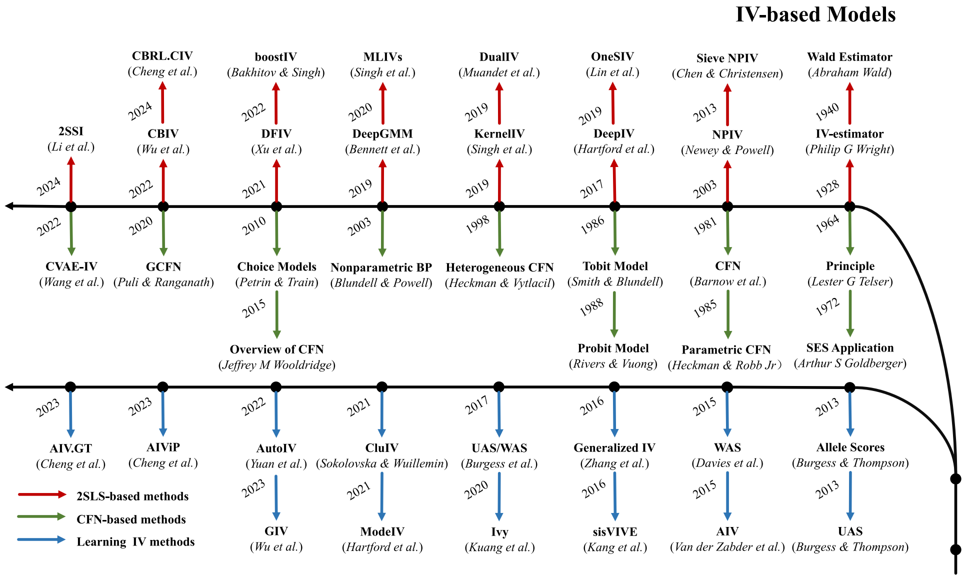


Fig. 4. The development of IV (mainly focus on machine learning methods).

real applications. To address these challenges, researchers have proposed more flexible approaches, including frameworks for identifying LATE for specific sub-populations [3, 5] and nonparametric methods for estimating CATE and ATE [66, 113]. In this section, we will explore the identification assumptions for LATE, CATE, ATE, and other common approaches[68, 69, 79, 163].

3.1 LATE Identification

As illustrated in Remark 2.5, Angrist, Imbens, and Rubin [3, 5] sought to estimate sub-populations whose average effects can be identified, specifically for the compliers (Equation (5)), instead of the overall average effect. Below we list two assumptions. If either holds true, it is sufficient to identify LATE.

Assumption 3.1 (Zero Probability [2]). *$\mathbb{E}[Y_i(0) \mid Z_i = z]$ and $\mathbb{E}[T_i \mid Z_i = z]$ are nontrivial function for all z, and there is a set $\mathcal{Z}_0$ such that $0 < \mathbb{P}(Z \in \mathcal{Z}_0) < 1$ and $\mathbb{P}(T = 1 \mid Z = z) = 0$ for all $z \in \mathcal{Z}_0$.*

The Zero Probability Assumption ensures the existence of a value or a set of values, $\mathcal{Z}_0$, which is realized with non-zero probability and for which the probability of participation is zero. This assumption is sufficient for the identification of treatment effects. To see why, let A be an indicator for the event $Z \notin \mathcal{Z}_0$, i.e., $A_i = \mathbb{1}\{Z_i \notin \mathcal{Z}_0\}$. Then, we have $\mathbb{E}[Y_i \mid A_i = 0] = \mathbb{E}[Y_i(0) \mid Z_i \in \mathcal{Z}_0] = \mathbb{E}[Y_i(0)]$ following that when $Z \in \mathcal{Z}_0$, $\mathbb{P}(T = 0 \mid Z = z) = 1$ and $\mathbb{E}[Y_i(0) \mid Z_i = z] = \mathbb{E}[Y_i(0)]$. Moreover, we have $\mathbb{E}[Y_i \mid A_i = 1] = \mathbb{E}[Y_i(0) \mid A_i = 1] + \mathbb{P}(T = 1 \mid A = 1) \cdot \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1, A_i = 1] = \mathbb{E}[Y_i(0)] + \mathbb{P}(T = 1 \mid A = 1) \cdot \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1]$. Since we can consistently estimate $\mathbb{P}(T = 1 \mid A = 1)$, $\mathbb{E}[Y_i \mid A_i = 0]$ and $\mathbb{E}[Y_i \mid A_i = 1]$, we can identify **LATE** $= \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1] = (\mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0]) / \mathbb{P}(T = 1 \mid A = 1)$.

Assumption 3.2 (Monotonicity [3]). *For all z, the triple $(Y(0), Y(1), T(z))$ is jointly independent of Z and $\mathbb{E}[T_i \mid Z_i = z]$ is a nontrivial function of z. For all z and w, either $T(z) \geq T(w)$ or $T(z) \leq T(w)$ always holds for all subjects.*

The Monotonicity Assumption ensures that the instrument affects the participation or selection decision in a monotone way. If, on average, subjects are more likely to participate given $Z = z$ than $Z = w$, then anyone who would participate given $Z = w$ must also participate given $Z = z$. Under this assumption, we have $\mathbb{E}[Y_i \mid Z_i = z] - \mathbb{E}[Y_i \mid Z_i = w] = (\mathbb{P}(T = 1 \mid Z = z) - \mathbb{P}(T = 1 \mid Z = w)) \cdot \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i(z) - T_i(w) = 1]$. As we can consistently estimate $\mathbb{E}[Y_i \mid Z_i = z]$ and $\mathbb{P}(T = 1 \mid Z = z)$ for any z, we can then identify **LATE** $= \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i(z) - T_i(w) = 1]$.

3.2 ATE/CATE Identification

Recently, a growing stream of literature has focused on identifying ATE and CATE under weaker structural assumptions on the outcomes. In this line of research, Newey and Powell (2003) [114], among others, introduced the *Additive Noise Assumption* and proposed the flexible **nonparametric instrumental variables** (**NPIVs**) for the identification of treatment effects.

Assumption 3.3 (Additive Noise Assumption [114]). *Consider the outcome model $Y = g(X, T) + \epsilon_Y$, where the measurement errors ϵ_Y can be arbitrarily correlated with U and satisfy $\mathbb{E}[\epsilon_Y \mid X_i, Z_i] = 0$.*

Under the *Additive Noise Assumption*, the relationship between the outcome process and the reduced form can be represented as a first-order Fredholm integral equation [93]. To obtain a consistent estimation, [66, 113, 114] identified the causal effect as the solution to an integral:

$$\mathbb{E}[y \mid x, z] = \int g(x, t) dF(t \mid x, z), \tag{11}$$

where F denotes the conditional cumulative distribution of t given $\{x, z\}$. Since $\mathbb{E}[y \mid x, z]$ and $F(t \mid x, z)$ are functionals of the distribution function for the observable random vector (y, x, z), they are identified. The identification of g and ATE/CATE thus depends on the existence of a unique solution to the integral Equation (11). By substituting $g(x, t)$ by another function $\tilde{g}(x, t)$ in Equation (11), we can see that the identification is equivalent to the nonexistence of any function $\delta(x, t) = g(x, t) - \tilde{g}(x, t) \neq 0$ such that the conditional expectation $\mathbb{E}[\delta(x, t) \mid z] = 0$.

When g can be identified, we can identify CATE and ATE:

$$\mathbf{CATE}(t, x) = g(x, t) - g(x, 0), \qquad \mathbf{ATE}(t) = \mathbb{E}[\mathbf{CATE}(t, x)]\,. \tag{12}$$

[66, 114] characterized identification of structural functions as completeness of certain conditional distributions $\mathbb{E}[\epsilon_T \mid z] = \mathbb{E}[\epsilon_Y \mid z] = 0$, where ϵ_T and ϵ_Y might be arbitrarily correlated with U.

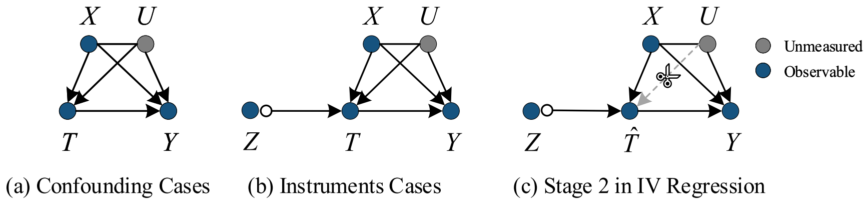
3.3 More General Assumptions

In the econometrics literature [68, 69], *Homogeneity Assumption* is a more general version than *Monotonicity Assumption* and *Additive Noise Assumption* for multi-dimensional instruments and treatments. Next, we describe two general *Homogeneity Assumptions* and *No Effect Modification Assumption*. The previous assumptions can be viewed as a special case of Homogeneity Assumptions.

Assumption 3.4 (Homogeneous Instrument-Treatment Association [26, 68, 163]). *The association between the IV and the treatment is homogeneous in the different level of unmeasured confounders, i.e., $\mathbb{E}[T \mid Z = z, U] - \mathbb{E}[T \mid Z = w, U] = \mathbb{E}[T \mid Z = z] - \mathbb{E}[T \mid Z = w]$.*

Assumption 3.5 (Homogeneous Treatment-Outcome Association [68, 77, 79]). *The association between the treatment and the outcome is homogeneous in the different level of unmeasured confounders, i.e., $\mathbb{E}[Y \mid T = z, U] - \mathbb{E}[Y \mid T = w, U] = \mathbb{E}[Y \mid T = z] - \mathbb{E}[Y \mid T = w]$.*

As long as either Assumption 3.4 or 3.5 holds, then CATE is identifiable. The **No Effect Modification** (**NEM**) assumption can be used to identify **individual treatment effects** (**ITEs**), but it may not be plausible in many instances [68, 79].



(a) Confounding Cases (b) Instruments Cases (c) Stage 2 in IV Regression

Fig. 5. The causal diagram in different cases.

Assumption 3.6 (No Effect Modification [68, 79, 163]). *The unmeasured confounders U would not modify the causal effect of T on Y.*

Multi-Instruments and Treatments: We systematically review the conditions for identifying the treatment effect of a uni-dimensional treatment. Additionally, these identification results can be easily extended to multi-dimensional treatments. When studying multi-dimensional treatments ($T \in \mathbb{R}^{d_T}$) and multi-dimensional instruments ($Z \in \mathbb{R}^{d_Z}$), as long as each treatment has at least one pre-specified valid instrument, its treatment effect can be identified using the conclusions presented in this section. Then, the model is just-identified if $d_Z = d_T$ and over-identified if $d_Z > d_T$ [62, 151].

4 2SLS

In the presence of unmeasured confounders, as shown in Figure 5(a), even though we can control the confounding effect from observed confounders X, the causality obtained by direct regression (Ordinary Least Squares, OLS) will be biased by unmeasured confounders U. To address this issue, the classical statistical method of IVs has been proposed, specifically the 2SLS technique. In this section, we discuss 2SLS and its machine learning variants. As illustrated in Figure 6, we categorize them into three groups: (1) Wald and 2SLS and Estimator; (2) NPIV estimator [33, 114]; (3) Machine learning estimator for further estimation. There are four main research lines from machine learning estimator, including Kernel-based estimator [112, 146], Deep-based estimator [12, 66, 106, 175], Moment conditions estimator [18, 46, 145] and confounder balance estimator. Finally, we conclude by summarizing the limitations of these approaches and discussing future work.

4.1 Wald and 2SLS Estimator

For ease of understanding, let us consider a basic scenario of uni-dimensional treatment and instrument, and no observed confounding variables. The simplified linear models derived from Equation (1), i.e., $T = \alpha Z + \epsilon_T$ and $Y = \beta T + \epsilon_Y$, omitting covariates X for clarity. The corresponding causal diagram is shown in Figure 5(b). Under linear specifications, we would like to adopt most commonly used methods, i.e., Wald and 2SLS estimator, to predict the causal parameter β.

4.1.1 Wald Estimator. In the 1940s, Wald introduced the Wald estimator, a method for identifying the **average treatment effect** (**ATE**) when the instrument is binary [161]. The ATE is calculated by $\hat{\beta}_{\text{Wald}} = \frac{\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]}{\mathbb{E}[T|Z=1]-\mathbb{E}[T|Z=0]}$. The Wald estimator $\hat{\beta}_{\text{Wald}}$ estimate β by plugging the sample average $\hat{\mathbb{E}}[Y \mid Z = 1]$, $\hat{\mathbb{E}}[Y \mid Z = 0]$, $\hat{\mathbb{E}}[T \mid Z = 1]$ and $\hat{\mathbb{E}}[T \mid Z = 0]$ into the above equation.

4.1.2 2SLS. As elaborated in Section 2.3.1, the 2SLS estimator is a widely used method for estimating causal effects in linear models, leveraging IVs to address endogeneity. It proceeds in two stages: In stage 1, the estimator directly regresses treatment $\mathbf{T} = t_{1,\ldots,n}$ on the IVs $\mathbf{Z} = z_{1,\ldots,n}$: $\hat{\alpha} = \arg\min_{\alpha} \frac{1}{n}\sum_{i=1}^{n}(t_i - \alpha z_i)^2$, and obtain predicted treatment values $\widehat{\mathbf{T}} = \mathbf{Z}\hat{\alpha} = \mathbf{P}_Z\mathbf{T}$. In stage 2, the outcome variable $\mathbf{Y} = y_{1,\ldots,n}$ is regressed on $\widehat{\mathbf{T}} = \hat{t}_{1,\ldots,n}$, minimizing $\hat{\beta}_{2SLS} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta \hat{t}_i)^2$. Then, the ATE is computed as $\hat{\beta}_{\text{2SLS}} = (\mathbf{T}'\mathbf{P}_Z\mathbf{T})^{-1} \mathbf{T}'\mathbf{P}_Z\mathbf{Y}$.

The 2SLS approach is limited by its linear assumptions, although alternative structural assumptions have been proposed by Angrist et al. [3, 5] to identify the LATE. Nevertheless, deterministic monotonicity may not always be realistic, and LATE can be limited in certain cases. To address these issues, motivated by the works [35, 53] on sieve estimation, [113, 114] propose an NPIV method, with uniform convergence rates [33].

4.2 NPIV Estimator

To relax the linearity assumption (Equation (1)), we consider a more general model:

$$T = f(Z, X) + \epsilon_T, \quad Y = g(T, X) + \epsilon_Y, \tag{13}$$

where f and g are the structural function we would like to estimate, both T and Z can be uni-dimensional or multi-dimensional. To achieve the NPIV estimator of f and g [34, 81], let $\{\psi_i\}$ be an orthonormal basis for the function space $L_2[0,1]$ that satisfies $\|g - \sum_{j=1}^{J} g_j \psi_j\| \leq CJ^{-s}$ for a smoothness parameter s, coefficient g_j and J.[3] Intuitively, the more orthonormal basis functions J we use, the better g can be approximated. Common basis choices include trigonometric functions, orthogonal polynomials, and splines [32]. Then we can approximate $f(z,x)$ by $\hat{f}(z,x) = \sum_{i=1}^{J_n} \sum_{j=1}^{J_n} \alpha_{ij} \psi_j(z) \psi_k(x)$ and approximate $g(t,x)$ by $\hat{g}(z,x) = \sum_{k=1}^{J_n} \sum_{j=1}^{J_n} \beta_{kj} \psi_k(z) \psi_j(x)$, where a_{ij} and β_{kj} are unknown coefficients, and the integer J_n is a truncation point of the number of basis functions that increases at a suitable rate as the number of samples $n \longrightarrow \infty$.

Then the 2SLS approach can be extended to the NPIV estimator through a two-stage process for estimating f and g. In the treatment regression stage, we estimate α_{ij} by minimizing the quadratic loss between f and $\hat{f}$, and possibly with a regularization term. In the outcome regression stage, we estimate β_{kj} by minimizing the quadratic loss between g and $\hat{g}$ using $\hat{f}$ from the first stage, and possibly with a regularization term. The regularization term bypasses estimating higher-order coefficients and reduces the estimation variance. To ensure consistency, the regularization parameter decreases with the sample size n, which balances model flexibility and estimation stability.

In the two-stage regression of the NPIV estimator, the challenge is how to define appropriate basis functions [33]. Thus, recent works [112, 146] introduce machine learning algorithms to obtain the basis functions and estimate causal effects.

4.3 Machine Learning Estimator

To implement further estimation, as shown in Figure 6, there are four main research lines from machine learning estimator (Figure 6), including Kernel-based Estimator [33, 112, 146], Deep-based Estimator [12, 66, 106, 175], Moment-based Estimator [18, 46, 145] and Confounder Balanced Estimator [171].

4.3.1 Kernel-based Estimator.

Motivated by Sieve NPIV [33] and **predictive state representation (PSR)** models [24, 75], [146] proposes **kernel instrumental variable (KernelIV)** regressionto model relations among Z, X, T, and Y as nonlinear functions in **reproducing kernel Hilbert spaces (RKHSs)** [149], and prove the consistency of KernelIV.

KernelIV. As shown in Figure 7, KernelIV defines two measurable positive definite kernels $k_{\mathcal{T}} : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R}$ and $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ corresponding to scalar-valued RKHSs $\mathcal{H}_{\mathcal{T}}$ and $\mathcal{H}_{\mathcal{Z}}$, where the mappings are given by $\psi : \mathcal{T} \rightarrow \mathcal{H}_{\mathcal{T}}, t \mapsto k_{\mathcal{T}}(t, \cdot)$ and $\phi : \mathcal{Z} \rightarrow \mathcal{H}_{\mathcal{Z}}, z \mapsto k_{\mathcal{Z}}(z, \cdot)$, respectively,

[3] The function space $L_2[0,1]$ is the set of functions square integrable on $[0,1]$, i.e., $L_2[0,1] = \{h : \int_0^1 h(x)^2 dx < \infty\}$. $\{\psi_i\}$ is an orthonormal basis if $\int_0^1 \psi_i(x)\psi_j(x)dx = 1$ for $i = j$ and 0 otherwise. The norm $\|h\| = [\int_0^1 h(x)^2 dx]^{1/2}$.

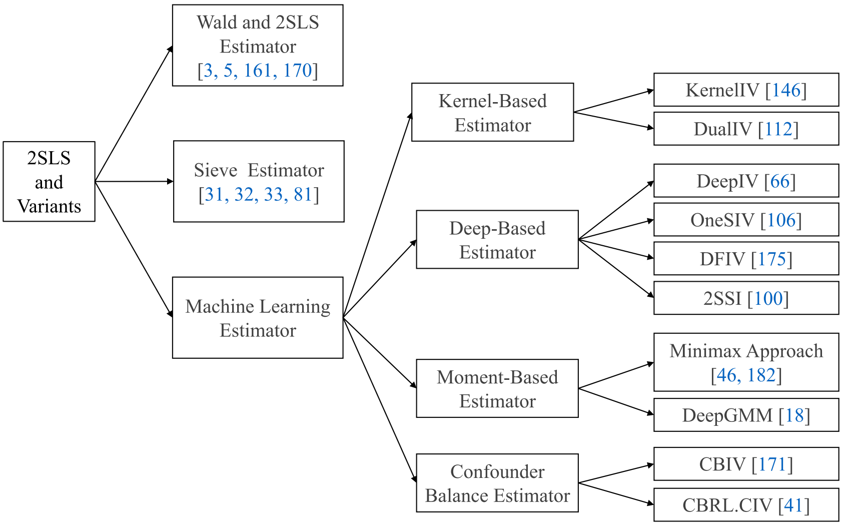


Fig. 6. Categorization of 2SLS and variants.

serving as the basis functions of $\mathcal{T}$ and $\mathcal{Z}$. In this subsection, we default $\mathcal{Z}$ as the horizontal concatenation of IVs Z and confounders X, while $\mathcal{T}$ as the horizontal concatenation of treatments T and confounders X. Following this, KernelIV reformulates the problem as $e \in E$, where $E : \mathcal{H}_\mathcal{Z} \to \mathcal{H}_\mathcal{T}$, and $h \in H$, where $H : \mathcal{H}_\mathcal{T} \to \mathcal{H}_\mathcal{Y}$. Then, KernelIV implements a two-stage regression.

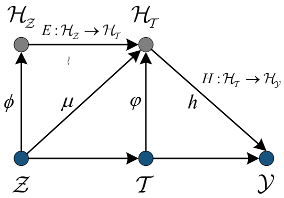

$\mathcal{H}_\mathcal{Z}$
$E:\mathcal{H}_\mathcal{Z} \to \mathcal{H}_\mathcal{T}$
$\mathcal{H}_\mathcal{T}$
ℓ
$H:\mathcal{H}_\mathcal{T} \to \mathcal{H}_\mathcal{Y}$
ϕ
μ
φ
h
$\mathcal{Z}$
$\mathcal{T}$
$\mathcal{Y}$

Fig. 7. The KernelIV framework.

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 with $\hat{\phi}(t_i) = \mu(z_i) = e(\psi(z_i))$, where $\psi(z_i) \in \mathcal{H}_\mathcal{Z}$ and $\hat{\phi}(t_i) \in \mathcal{H}_\mathcal{T}$. KernelIV constructs an objective for optimizing $e \in E$ by kernel ridge regression, i.e., $e^*_\lambda = \text{argmin}_{e \in E} \mathbb{E}||\mu(z_i) - \phi(t_i)||^2 + \lambda||e||^2$, where λ is a hyper-parameter and $||e||^2$ is a penalty term for function e. Then, we have $\hat{t}_i = \mu(z_i) = [e^*\psi](z_i)$.

In stage 2, to estimate the structural function $g(\cdot)$ in Equation (13), KernelIV predicts the potential outcome function with the conditional mean embedding $\hat{\phi}(t_i)$, $\hat{y}_i = g(t_i) = h(\phi(t_i)) = \mathbb{E}[y_i \mid \hat{\phi}(t_i)]$. Then, KernelIV constructs an objective for optimizing $h \in H$ by kernel ridge regression:

$$h^*_\lambda = \text{argmin}_{h \in H} \mathbb{E}||h(\phi(t_i)) - y_i||^2 + \lambda||h||^2, \text{ or } h^*_\lambda = \text{argmin}_{h \in H} \mathbb{E}||h(\mu(z_i)) - y_i||^2 + \lambda||h||^2, \quad (14)$$

where $||h||^2$ is a penalty term for function h. Indeed, $\widehat{\mathbf{Y}} = g(\mathbf{T}) = [h\phi](\mathbf{T}) = [h\mu](\mathbf{Z})$.

DualIV. Inspired by stochastic programming [43, 141], DualIV [112] shows that two-stage IV-based regression can be reformulated as a convex-concave saddle-point problem. It develops a simple kernel-based algorithm and simplifies traditional two-stage methods via a dual formulation. First, Muandet et al. [112] reformulate the problem as an empirical risk minimization problem:

$$\min_{g \in \mathcal{G}} R(g) = \mathbb{E}_{YZ}\left[\ell\left(Y, \mathbb{E}_{T|Z,X}[g(T,X)]\right)\right], \quad \mathbb{E}_{T|Z,X}[g(T,X)] = \int g(T,X)dF(T \mid Z,X). \quad (15)$$

where $\ell(\cdot)$ denotes the loss function, and $dF(T \mid Z,X)$ is the conditional treatment distribution.

Applying the interchangeability and Fenchel duality [43, 141] to Equation (15), we can obtain:

$$R(g) = \max_{u \in \mathcal{U}} \mathbb{E}_{ZXTY}[g(T,X)u(Y,Z,X)] - \mathbb{E}_{ZXY}\left[\ell^\star(Y, u(Y,Z,X))\right], \quad (16)$$

where $\mathcal{U}(\Omega) = \{u(\cdot) : \Omega \to \mathbb{R}\}$ denotes the function family defined on the support $\Omega = \{\mathcal{Y}, \mathcal{Z}, \mathcal{X}\}$. ℓ is a convex and lower semi-continuous loss function 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 denote by $W := Y \oplus Z \oplus X$ and $T := T \oplus X$ as concatenation. Then, the saddle-point problem is $\min_{g \in \mathcal{G}} \max_{u \in \mathcal{U}} \mathbb{E}_{TW}[g(T)u(W)] - \mathbb{E}_W\left[\ell^\star(Y, u(W))\right]$. With $\ell^\star(y, u) = uy + \frac{1}{2}u^2$, DualIV reduce the traditional two-stage methods as

$$\min_{g \in \mathcal{G}} \max_{u \in \mathcal{U}} \Psi(g, u), \quad \Psi(g, u) = \mathbb{E}_{TW}\{[g(T) - Y]u(W)\} - \tfrac{1}{2}\mathbb{E}_W[u(W)^2]. \tag{17}$$

Motivated by the RKHSs [149], the objective is rewritten as

$$\Psi(f, u) = \mathbb{E}_{TW}[f(T)u(W)] - \mathbb{E}_{YZ}[Yu(Y, Z)] - \frac{1}{2}\mathbb{E}_W\left[u(W)^2\right] = \langle C_{WT}f - b, u\rangle_{\mathcal{U}} - \frac{1}{2}\langle u, C_W u\rangle_{\mathcal{U}},$$

where $b := \mathbb{E}_{YZ}[Y\varphi(Y, Z)] \in \mathcal{U}, C_W := \mathbb{E}_W[\varphi(W) \otimes \varphi(W)] \in \mathcal{U} \otimes \mathcal{U}$ is a covariance operator, $\{\phi, \varphi\}$ are canonical feature maps [137], and $C_{WT} := \mathbb{E}_{WT}[\varphi(W) \otimes \phi(T)] \in \mathcal{U} \otimes \mathcal{F}$ is a cross-covariance operator. The generalized least squares solution in RKHS is

$$f^* = \arg\min_{f \in \mathcal{F}} \tfrac{1}{2}\left\langle C_{WT}f - b, C_W^{-1}\left(C_{WT}f - b\right)\right\rangle_{\mathcal{U}} = \left(C_{TW}C_W^{-1}C_{WT}\right)^{-1} C_{TW}C_W^{-1}b. \tag{18}$$

Equation (18) gives a solution for IV-based regression in closed form.

4.3.2 Deep-based Estimator. Recent machine learning methods, such as KernelIV [146] and DualIV [112], have extended the 2SLS estimator to non-linear settings using infinite dictionaries of basis functions from RKHS. However, these methods have limited flexibility as basis functions require pre-specification or feature engineering [66, 175]. On the other hand, deep learning methods, while lacking a formal theoretical foundation, make fewer assumptions about the data-generating process. They automatically learn flexible feature mappings for high-dimensional and non-linear data, reducing the need for manual selection of basis functions and enhancing the accuracy of causal effect estimation.

DeepIV. Under *Additive Noise Assumption*, DeepIV [66] provides a unique solution for the inverse problem using deep learning techniques. Specifically, in the treatment regression stage, DeepIV [66] uses a deep neural network $\pi_\phi(Z, X)$ with parameters ϕ to model the conditional density function of treatment $F(T \mid Z, X)$, i.e., $\min \mathcal{L}_1 = \frac{1}{n}\sum_{i=1}^n l(t_i, \pi_\phi(z_i, x_i))$, where $l(t_i, \pi_\phi(z_i, x_i))$ would be an l_2-loss for continuous outcomes or a log-loss for binary outcomes. To obtain the conditional probability estimation $\pi_\phi(Z, X)$ of treatments, DeepIV models a mixture of Gaussian distributions with component $\pi_{\phi,k}(Z, X)$ and sub-networks $[\mu_{\phi,k}(Z, X)], \sigma_{\phi,k}(Z, X)]$ for Gaussian distribution parameters $G(\mu, \sigma)$. With enough mixture components, the network $\pi_\phi(Z, X)$ can approximate arbitrary smooth densities. Then, in the outcome regression stage, DeepIV models a counterfactual prediction network h_θ with parameters θ, to approximate the potential outcome, i.e., $\min \mathcal{L}_2 = \frac{1}{n}\sum_{i=1}^n (y_i - \int_t h_\theta(t, x_i)d\hat{F}_\phi(t \mid z_i, x_i))^2$, where $\hat{F}_\phi(T \mid Z, X)$ is obtained from the stage 1. Then, we can identify the counterfactual function $h_\theta(t, x_i)$ for treatment effect estimation. 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 [106]. To address this issue, Lin et al. [106] propose a OneSIV algorithm to merge the two stages by integrating the outcome regression $h_\theta(T, X)$ with the treatment distribution estimation $\hat{F}_\phi(T \mid Z, X)$.

DFIV. Combining the theoretical advantages of kernel-based methods and the empirical advantages of deep learning methods, DFIV [175] uses **deep neural networks** (**DNNs**) to adaptively learn deep features as kernel basis, which fits structural functions with highly nonlinear flexibility. [175] develops three DNNs $\{f_\phi, g_\xi, u_\psi\}$ to learn the corresponding feature mappings for $\{Z, X, T\}$, respectively. DFIV reformulates the IV-based regression as

$$u_{\psi,k}(T) = \sum_{i=1}^{d^Z}\sum_{j=1}^{d^X} \alpha_{i,j}^k f_{\phi,i}(Z)g_{\xi,j}(X) + \epsilon_T, Y = \sum_{k=1}^{d^T}\sum_{j=1}^{d^X} \beta_{k,j}u_{\psi,k}(T)g_{\xi,j}(X) + \epsilon_Y, \tag{19}$$

where $f_{\phi,i}(Z)$ denotes the ith element in the instrument representation $f_\phi(Z)$, $g_{\xi,j}(X)$ is the jth element in the covariate representation $g_\xi(X)$, and $u_{\psi,k}(T)$ is the kth element in the treatment representation $u_\psi(T)$. $\{d^Z, d^X, d^T\}$ denotes the dimensions of the representation $f_\phi(Z)$, $g_\xi(X)$, and $u_\psi(T)$. $A = [\alpha_{i,j}^k]_{i,j,k}$ and $B = [\beta_{i,j}]_{i,j}$ denote the corresponding coefficients in the linear associations between features $\{f_\phi(Z), g_\xi(X), u_\psi(T), Y\}$. Additionally, the boostIV algorithm takes a different approach by replacing sieve functions with additional weak learners [12]. It focuses on optimizing the conditional expectation $\mathbb{E}[y \mid z] = \alpha_0 + \sum_{m=1}^{M} \alpha_m \mathbb{E}\left[\varphi\left(x;\theta_m\right) \mid z\right]$, where $\varphi\left(x;\theta_m\right)$ are generated by simple algorithms called weak learners and α_m are the corresponding weights.

Two-Stage Shadow Inclusion (2SSI). The 2SSI method [100] is an extension of the 2SLS framework, incorporating innovations to address both unmeasured confounding bias and collider biases. While it shares similarities with 2SLS, such as the two-step structure of regressing the treatment on IVs in the first stage and outcome regression in the second stage, 2SSI introduces residuals as shadow variables to handle collider biases under the Shadow Variable Theorem [108]. This theorem establishes conditions for identifying causal effects in the presence of missing data with shadow variables, allowing for generalization from observed ($S = 1$) to unobserved ($S = 0$) data. The key generalization formula is $\tau_0(X,T,V) = \text{OR}(X,T,Y) \cdot \tau_1(X,T,V)/E[\text{OR}(X,T,Y)|X,T,V,S=1]$, where τ_0 and τ_1 represent the outcomes for unobserved and observed data, respectively, and $\text{OR}(X,T,Y)$ is the odds ratio function modeling the selection mechanism. The odds ratio is defined as $\text{OR}(X,T,Y) = \frac{P(S=0|X,T,Y)\cdot P(S=1|X,T,Y=0)}{P(S=0|X,T,Y=0)\cdot P(S=1|X,T,Y)}$, ensuring that the residuals satisfy the shadow variable conditions of being unrelated to selection ($V \perp S|X,T,Y$) while remaining informative about the outcome ($V \not\perp Y|X,T,S=1$). This method enables 2SSI to effectively handle collider bias while ensuring unbiased causal inference.

4.3.3 Moment-based Estimator. The GMM is a prevalent estimation technique in econometrics that leverages moment conditions to estimate model parameters [16, 64]. By minimizing the discrepancy between empirical moments and their theoretical counterparts implied by the model, GMM yields robust parameter estimates. 2SLS is a special case of GMM designed to address unmeasured confounding issues through a two-stage regression process [63, 72, 167]. Under valid IV conditions and the additive noise assumption, the IVs Z are orthogonal to the unmeasured factors, i.e., $Z \perp U, \epsilon_T, \epsilon_Y$, and the moment conditions for $Z \in \mathbb{R}^{n\times d^Z}$ can be formulated as $\mathbb{E}[z_i \cdot \epsilon_i] = 0$, yielding a set of d^Z moments: $M = (m_1, m_2, \cdots, m_{d^Z})', m_i = z_i'\epsilon_{Y,i} = z_i'(y_i - g(t_i,x_i))$ for $i = 1, 2, \ldots, d^Z$. The GMM minimizes the discrepancies between sample and theoretical moments by solving

$$g^* = \arg\min_{g\in\mathcal{G}} \|M\|^2, \quad \|M\|^2 = M'WM = \sum_{j=1}^{d^Z} m_j^2 = \sum_{j=1}^{d^Z} \left[z_j'(y_j - g(t_j,x_j))\right]^2, \qquad (20)$$

where $W = I$ is an identified matrix, meaning the average effect. Although GMM is an incredibly flexible estimator, in practice, 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 [17, 18, 42, 104].

Minimax Approaches. With advancements in machine learning, 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 [18, 46, 182]. Following this, Lewis et al. [46] formulate the expectation minimization problem as the maximum moment deviation over the set of potential functions, referred to as **Adversarial GMM** (**AGMM**):

$$h^* = \text{arginf}_{h\in\mathcal{H}} \text{sup}_{f\in\mathcal{F}} \mathbb{E}[(y_i - h(t_i,x_i))f(z_i,x_i)]. \qquad (21)$$

Similar to Wasserstein GANs [9, 101], 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. Lewis et al. [46] offer main theorems for several hypothesis spaces of practical interest including RKHS, functions defined via shape restrictions, random forests, and neural networks.

Given observational data $\{z_i, x_i, t_i, y_i\}_{i=1,\ldots,n}$, to obtain optimal h_ϕ and f_ψ, AGMM [46] minimizes the empirical analogue of the minimax objective:

$$\phi^* = \operatorname{arginf}_{\phi\in\Phi}\sup_{\psi\in\Psi}\mathbb{E}[(y_i - h_\phi(t_i, x_i))f_\psi(z_i, x_i)] - \lambda_1\|\psi\|^2 - \mathbb{E}[f_\psi(z_i, x_i)^2] + \lambda_2\|\phi\|^2. \tag{22}$$

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 an unweighted vector norm can lead to significant inefficiencies in the minimization of objective Equation (20) [64, 65]. [64, 65] 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 GMM [17, 18, 64], DeepGMM [18] construct an optimal combination of moment conditions via adversarial training, with the objective:

$$\phi^* = \operatorname{arginf}_{\phi\in\Phi}\sup_{\psi\in\Psi}\mathbb{E}[(y_i - h_\phi(t_i, x_i))f_\psi(z_i, x_i)] - \tfrac{1}{4}\mathbb{E}[(y_i - h_\phi(t_i, x_i))^2 f_\psi^2(z_i, x_i)]. \tag{23}$$

Notably, DeepGMM [18] 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 the minimax problem are developed by [42, 104].

MLIVs. To address the issue of limited variation derived from weak IVs, Singh et al. [145] propose a machine learning algorithm to learn a class of IV functions $H(Z;\eta)$ parameterized by η to obtain optimal MLIVs in the GMM architecture: $\min_\eta[\mathcal{L}(\eta) = \operatorname{tr}(\hat{V}_N(\theta;\eta))]$, s.t. $\theta \in \arg\min_{\theta'} \hat{Q}_N(\theta';\eta)$, where $\hat{Q}_N$ refers to the sample GMM criterion, and $\hat{V}$ is variance-covariance matrix of model parameters θ. This sets up a bi-level optimization problem and then they iteratively update θ and η using Gauss Newton Regression with cross-fitting, which preserves asymptotic theory.

4.3.4 *Confounder Balance Estimator.*

With advancements in machine learning, recent IV models have been developed to address complex scenarios involving interactions among multiple variables, but often overlook the joint influence of IVs Z and covariates X on treatment assignment T. They incorporate observed Z and X to jointly predict the conditional distribution of treatments $P(T \mid Z, X)$, aiming to obtain a resampled $\hat{T}$ unaffected by unmeasured confounders U. However, as depicted in Figure 5(c), while these methods block the connection between the predicted treatment $\hat{T}$ and unmeasured confounders U, the observed confounders X persist in the outcome stage, introducing additional confounding bias in potential outcome regression.

Confounder Balanced IV (CBIV). To jointly address the confounding bias from observed and unmeasured confounders, under *Homogeneity Assumptions*, Wu et al. [171] propose a CBIV Regression algorithm and model a more general causal relationship by relaxing the additive assumption to multiplicative assumption on response-outcome function as $Y = g_1(T, X) + g_2(T)g_3(U) + g_4(X, U)$, where $g_j(\cdot)$ are unknown and potentially non-linear continuous functions. The completeness of $\mathbb{P}(T \mid Z, X)$ and $\mathbb{P}(Y \mid T, X)$ guarantees uniqueness of the solution [114].

Like previous works, the CBIV algorithm takes the same treatment regression in stage 1 with loss $\mathcal{L}_T = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(t_i - \hat{t}_i^j)^2$, resampling m treatments $\hat{t}i^j j = 1, \ldots, m \sim \hat{P}(t_i|z_i, x_i)$ to approximate the true treatment t_i, but differs in stage 2 by learning a balanced representation $C = f_\theta(X)$ through **mutual information (MI)** minimization: first, CBIV uses variational distribution $Q_\psi(\hat{T} \mid C) = \mathcal{N}(\mu_\psi(C), \sigma_\psi(C))$ parameterized by $\{\mu_\psi, \sigma_\psi\}$ to approximate the true conditional distribution $P(\hat{T} \mid C)$; then, they minimize the log-likelihood loss function $Q_\psi(\hat{T} \mid C)$ with

n samples to estimate MI, i.e., $\mathrm{disc}(\hat{T}, C) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[\log Q_\psi(\hat{t}_i \mid c_i) - \log Q_\psi(\hat{t}_j \mid c_i)]$, where $C = f_\theta(X)$. With the learned balanced representation, CBIV regresses the outcome on the estimated treatment $\hat{T}$ and the representation C. The objective function is $\mathcal{L}_Y = \frac{1}{n}\sum_{i=1}^{n}(y_i - h_\xi(\hat{t}_i, f_\theta(x_i)))^2$. Theoretically and empirically, CBIV tackles and solves the inverse issues of the response-outcome function. By simultaneously removing biases from the observed and unmeasured confounders, the method improves the accuracy of treatment effect estimations [171].

CBRL.CIV. In linear settings, Cheng et al. [39, 40] propose a CIV.VAE method to learn CIVs and their conditioning sets based on **variational autoencoder** (**VAE**), and estimate the causal parameter by $\beta = \sigma_{z*y*\mathbf{c}}/\sigma_{z*t*\mathbf{c}}$, where $\sigma_{z*y*\mathbf{c}}$ and $\sigma_{z*t*\mathbf{c}}$ are the estimated causal effect of Z on Y conditioning on $\mathbf{C}$ and the causal effect of Z on T conditioning on $\mathbf{C}$, respectively. Inspired by CBIV [171], Cheng et al. [41] further develop the CBRL.CIV algorithm, a three-stage approach for causal effect estimation in non-linear settings. In the CIV regression stage, CBRL.CIV predicts $P(Z|X)$ using a network $\phi_\mu(X)$ with the loss function $L_Z = \frac{1}{n}\sum_{i=1}^{n}[z_i \log \phi_\mu(x_i) + (1 - z_i)\log(1 - \phi_\mu(x_i))]$, and learns a representation $C = \psi_\theta(X)$ by minimizing the discrepancy $D(Z, C) = \mathrm{IPM}(\psi_\theta(X)|Z = 0, \psi_\theta(X)|Z = 1)$. In the treatment regression stage, it estimates $P(T|Z, X)$ using a network $\phi v(Z, X)$ with the loss $L_T = \frac{1}{n}\sum_{i=1}^{n}[t_i \log \phi_v(z_i, x_i) + (1 - t_i)\log(1 - \phi_v(z_i, x_i))]$ and balances the learned representation W by minimizing $D(T, C) = \mathrm{IPM}(\psi_\theta(X)|T = 0, \psi_\theta(X)|T = 1)$. Finally, in the outcome regression stage, the method predicts the potential outcomes $\{Y(T = 0), Y(T = 1)\}$ via two separate networks, minimizing the loss $L_Y = \frac{1}{n}\sum i = 1^n(y_i - \sum_{t\in 0,1} f_{\gamma_t}(\psi_\theta(X))P(t|z_i, x_i))^2$. The final objective integrates the confounding balance terms as regularization: $\min_{\theta,\gamma_0,\gamma_1} L_Y + \alpha D(T, C) + \beta D(Z, C)$, where α and β control the tradeoff. The **average causal effect** (**ACE**) is computed as $\hat{\beta} = \mathbb{E}[f_{\gamma_1}(\psi_\theta(X)) - f_{\gamma_0}(\psi_\theta(X))]$. This method ensures confounding balance and robust causal effect estimation with **conditional IVs** (**CIVs**).

4.4 Limitation and Future Work

4.4.1 Limitation. The effectiveness of IV methods relies on the validity and strength of the predefined IVs. However, identifying valid and strong IVs is a challenging task due to the untestable exclusion condition [171, 179]. In this section, we discuss the limitations of IVs and their implications for causal inference.

Validity of IVs. The validity of IVs is crucial for obtaining reliable causal estimates. However, ensuring the validity of IVs is often difficult, as it requires expert knowledge and is not guaranteed even with careful selection. To address this issue, researchers in the IV literature often implement **Randomized Controlled Trials** (**RCTs**) to obtain exogenous IVs, such as the Oregon health insurance experiment [49] and the effects of military service on lifetime earnings [4]. However, RCTs can be expensive and are not universally available.

Weak IVs and Model Mis-specification. IVs may have little causal effect on the treatment variables, which we call weak IV. Weak IVs can lead to biased and imprecise causal estimates, as the IV method relies on the strength of the relationship between the IV and the treatment variable. This limitation poses a significant challenge for researchers in obtaining reliable and accurate causal estimates. Machine learning algorithms tend to combine observed confounders and IVs to predict the conditional distribution of the treatments to eliminate unmeasured confounding bias. However, 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 [171].

4.4.2 Future Work. The IV has been a significant tool for causal inference, and several promising directions for future work have emerged. In this section, we outline four key areas for future work: causal discovery and **ancestral instrumental variables** (**AIVs**), the GMM, and confounder balance.

Causal Discovery and AIVs. While sufficient conditions for discovering valid IVs are elusive, causal discovery methods can identify potential sets of IVs by leveraging structural assumptions and testable constraints. For example, Silva and Shimizu [144] discuss how causal discovery methods can identify potential IVs under structural and non-Gaussian assumptions. Cheng et al. [37, 38] use the RFCI (Really Fast Causal Inference) algorithm to learn **partial ancestral graphs** (**PAGs**) from observational data. Using these recovered PAGs, they propose algorithms like AIV.GT to identify AIVs, which are a special class of CIVs, along with their corresponding conditioning sets directly from data. These advancements demonstrate how causal discovery can identify potential sets of IVs, expanding the toolkit for causal inference in practical application.

GMM. The GMM is a highly flexible IV estimator that relies on a large number of moment conditions and IV independence conditions. Recent advancements in machine learning algorithms have led to the development of nonlinear independence detection algorithms, which have demonstrated superior performance compared to first-order moment independence methods. One potential avenue for future research is to incorporate independent testing algorithms, such as HSIC-X [132], into the GMM framework as a means of constraining the IV regression. This approach may yield more accurate and reliable causal estimates in the presence of nonlinear relationships.

Confounder Balance. Conventional methods for complex non-linear IV regression would suffer from the bias from the observed confounders, which are imbalanced in stage 2. To address this problem, researchers have proposed CBIV regression algorithms, such as CBIV [171]. By combining confounder balancing techniques with IV regression, these methods aim to remove bias from both observed and unobserved variables. Theoretical analyses and numerical experiments provided by [171] demonstrate the effectiveness of this approach, suggesting that confounder balance should be a key consideration in future IV research.

5 CFN

The CFN method, also known as two-stage residual inclusion, is another statistical approach to address unmeasured confounding bias [168]. This technique can be traced back to early works by Telser [156] and Goldberger [55]. A CFN is a variable that renders known cause variables (i.e., treatments) appropriately exogenous in the outcome regression [14, 31, 55]. In observational data, conditional on the CFN or confounders,[4] the CFN estimator makes the treatment appropriately exogenous in the regression equation. While both 2SLS and CFN share similar asymptotic properties for treatment effect estimation, CFN can be more efficient in complex relationships by leveraging additional information about unmeasured confounders. Figure 8 demonstrates how CFN is applied in causal inference and machine learning, encompassing both linear and non-linear scenarios.

5.1 Linear-based CFN

Consider a basic scenario with a uni-dimensional treatment, uni-dimensional instrument, and no covariates X for simplicity. Following the linearity assumption (Equation (1)), the CFN estimator maintains the spirit of the earlier definitions and estimations [168]: $T = \alpha Z + \epsilon_T$, $Y = \beta T + \epsilon_Y$. To predict the causal parameter β, we first apply the CFN estimator under these linear specifications.

5.1.1 CFN Estimator. As elaborated in Section 2.3.2, the CFN estimator serves as a general solution for estimating causal effects in linear models [71], employing a two-stage regression. In stage 1, the estimator directly regresses treatment $\mathbf{T} = t_{1,\ldots,n}$ from IV $\mathbf{Z} = z_{1,\ldots,n}$: $\hat{\alpha} = \arg\min_{\alpha} \frac{1}{n}\sum_{i=1}^{n}(t_i -$

[4] The role of the CFN in IV regression is to complement unmeasured or missing data.

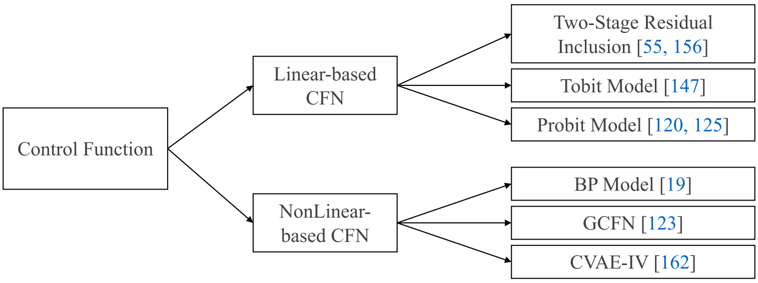


Fig. 8. Categorization of CFN estimators.

$\alpha z_i)^2$, and the residuals is $\hat{\mathbf{r}}_\epsilon = \mathbf{T} - \mathbf{Z}\hat{\alpha}$. In stage 2, the estimator regresses the outcome $\mathbf{Y} = y_{1,\dots,n}$ on the predicted residuals $\hat{\mathbf{r}}_\epsilon$, i.e., $\hat{\beta}_{CFN}, \hat{\theta}_\epsilon = \arg\min_{\beta,\theta_\epsilon} \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta\hat{t}_i - \theta_\epsilon\hat{r}_i)^2$, where $(\mathbf{A},\mathbf{B})$ means the concatenation of vectors/matrices $\mathbf{A}$ and $\mathbf{B}$. The causal effect is then estimated as $(\hat{\beta}_{\text{CFN}}, \hat{\theta}_\epsilon) = ((\mathbf{T}, \hat{\mathbf{r}}_\epsilon)'(\mathbf{T}, \hat{\mathbf{r}}_\epsilon))^{-1}(\mathbf{T}, \hat{\mathbf{r}}_\epsilon)'\mathbf{Y}$.

5.1.2 Binary Treatment Effects. The binary treatment $T = \{0, 1\}$ is a special case for CFN. We can utilize the binary nature of treatment T and replace the linear regression with a logistic regression model. The structural equation would be $T = \mathbb{1}\{\alpha Z + \epsilon_T > 0\}, \quad Y = \beta T + \epsilon_Y$, where $\mathbb{1}\{\cdot\}$ is the indicator function. Then, the treatment assignment can be regarded as a probit model: $P(T = 1 \mid Z) = \Phi(\alpha Z)$, where $\Phi(\cdot)$ is the normal cumulative distribution function. Then we can derive the CFN [120, 167]. In stage 1, we build the probit model and estimate the *generalized residual* by $\hat{\mathbf{r}}_\epsilon = \mathbf{T}\lambda(\alpha\mathbf{Z}) - (1 - \mathbf{T})\lambda(-\alpha\mathbf{Z})$, where $\lambda(\cdot) = \frac{\phi}{\Phi}(\cdot)$ is the well-known inverse Mills ratio [57]. In stage 2, we control the generalized residual $\hat{\mathbf{r}}_\epsilon$ to estimate the conditional average causal effect of treatments T on outcomes Y, i.e., $\mathbf{CATE}(t, r_\epsilon) = \mathbb{E}[Y(T = t) - Y(T = 0) \mid r_\epsilon]$.

One limitation for CFN in binary/discrete treatment cases is that the results are 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 unbiased causal effect.

5.2 NonLinear-based CFN

In the previous section, we have introduced CFN estimators employed for linear models. Recently, some works have broadened the scope of the CFN applications to a more general scenario [125, 147, 168]. Here, we detail the flexibility of the CFN estimator in a more general model, where both T and Z can be uni-dimensional or multi-dimensional. For example, in the multiplicative cases:

$$T = \alpha Z + \theta_T X + \epsilon_T, \quad Y = \beta TX + \theta_Y X + \epsilon_Y, \tag{24}$$

where, TX denotes the multiplicative interactions of T and X, and the unmeasured confounders U influence T and Y through the noise terms ϵ_T and ϵ_Y. According to the IV's conditions, we have that $Z \perp \{X, \epsilon_T, \epsilon_Y\}$, then we can obtain the residual in stage 1 $\hat{\mathbf{r}}_\epsilon = \mathbf{T} - \hat{\alpha}\mathbf{Z} + \hat{\theta}_T\mathbf{X}$, where $\{\hat{\alpha}, \hat{\theta}_T\}$ is the corresponding estimated coefficients. Sequentially, we can perform the outcome regression on $\{\hat{r}_\epsilon, X\}$ and interaction TX: $\mathbf{CATE}(t, x, r_\epsilon) = \mathbb{E}[Y(T = t) - Y(T = 0) \mid r_\epsilon, X = x, T = t, TX = tx]$. A similar estimator can be built for a discrete treatment case, in the discrete model [120, 157].

5.2.1 Non-Parametric BP Estimator. The CFN is an adaptable approach for addressing confounding bias in both linear and non-linear scenarios. In general models, the causal structural function may exhibit complex non-linear relationships. Blundell and Powell [19] proposes a non-parametric extension of the Rivers-Vuong approach [125], which is applicable in a general setting:

$$T = f(Z, X) + \epsilon_T, \qquad Y = g(X, T, U), \tag{25}$$

where $f(\cdot)$ and $g(\cdot)$ are the structural functions.

BP model. Blundell and Powell [19] propose a BP model to estimate the **Average Structural Function** (**ASF**) of the outcome, defined as $\text{ASF}(X, T) = \mathbb{E}[g(X, T, U) \mid X, T]$. This implies that the unmeasured confounders U are averaged out in the population, conditional on the fixed X and T, i.e., $\mathbb{E}_U[g(X, T, U)] = \mathbb{E}[g(X, T, U) \mid X, T]$. In stage 1 of IV regression, the BP model [19] computes the residual $\hat{r}\epsilon$ using $\hat{r}\epsilon = T - f(Z, X)$, where $f(Z, X)$ is identified as $f(Z, X) = \mathbb{E}[T \mid Z, X]$. Thus, $\hat{r}_\epsilon = T - \mathbb{E}[T \mid Z, X] = T - \hat{f}(Z, X)$. Machine learning methods, such as kernel-based regression and neural networks, can be used to estimate the expectation $\mathbb{E}[T \mid Z, X]$ effectively. In stage 2, the conditional distribution of the unmeasured confounders U is related to Z, X, T only through the residual $\hat{r}\epsilon$ [166, 168], i.e., $P(U \mid Z, X, T) = P(U \mid Z, X, \hat{r}\epsilon) = P(U \mid \hat{r}_\epsilon)$. Then, we have

$$\hat{\text{ASF}}(X, T) = \mathbb{E}[\hat{g}'(X, T, \hat{r}_\epsilon) \mid X, T] = \mathbb{E}_{\hat{r}_\epsilon}[\hat{g}'(X, T, \hat{r}_\epsilon)], \hat{g}'(X, T, \hat{r}_\epsilon) = \mathbb{E}[Y \mid X, T, \hat{r}_\epsilon], \tag{26}$$

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

5.2.2 *General CFN Estimator.* Although CFN estimators have been widely used for addressing unmeasured confounders, residual regression in CFN usually breaks down under complex nonlinear models, because it assumes residuals are a perfect proxy of unmeasured confounders, which may not hold in real applications. Based on VAE [91], some studies propose using proxy variables to reconstruct unmeasured confounders [107, 174, 183]. Building on this, [123] developed the **General Control Function** (**GCFN**) Method to construct CFNs satisfying ignorability and positivity assumptions. Assuming the data is generated by

$$T = f(Z, X, \hat{r}_\epsilon), \qquad Y = g(X, T, U). \tag{27}$$

The causal effect is identified by the joint distribution $q(Z, X, \hat{r}_\epsilon, T)$ over the CFN $\hat{r}_\epsilon$ and the observables $\{Z, X, T\}$: $\mathbb{E}_{\hat{r}_\epsilon}[Y \mid T, \hat{r}_\epsilon] = \mathbb{E}_{\hat{r}_\epsilon}[Y \mid \text{do}(T), \hat{r}_\epsilon] = \mathbb{E}[Y \mid \text{do}(T)]$.

GCFN. Following [59, 80, 107], GCFN's first stage, called **variational decoupling** (**VDE**), constructs GCFNs by using VAE and recovering the residual variation in the treatment given IVs [123]. This yields **evidence lower bound** (**ELBO**) of VAE to reconstruct the latent variables:

$$L(\theta, \phi, \xi \mid Z, X, T) = (1 + \lambda)\mathbb{E}_{q_\theta(\hat{r}_\epsilon | Z, X, T)} \log p_\phi(T \mid Z, X, \hat{r}_\epsilon) - \lambda D_{KL}(q_\theta(\hat{r}_\epsilon \mid Z, X, T) \| p_\xi(\hat{r}_\epsilon)), \tag{28}$$

where λ is the hyper-parameter that is used to balance the reconstruction term and the KL term in the beta-VAE. $p_\xi(\hat{r}_\epsilon)$ and $p_\phi(T \mid Z, X, \hat{r}_\epsilon)$ are real (posterior) probability distributions, $q_\theta(\hat{r}_\epsilon \mid Z, X, T)$ 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, VDE stage provides a GCFN $\hat{r}_\epsilon$ from the observables $\{Z, X, T\}$. Using learned $\hat{r}_\epsilon$, in stage 2, GCFN estimates causal effects via existing confounder adjusting/control methods like matching/balancing methods [10, 95, 103], doubly robust methods [13] and representation learning methods [70, 87, 140, 173]. Furthermore, Puli et al. [123] develop semi-supervised GCFN to construct GCFNs using subsets of data that have both IV and confounders.

5.2.3 *Conditional VAE Estimator.* Due to untestable exclusion and exogeneity conditions, finding a valid IV is always a tricky problem. To relax the restriction, Wang et al. [162] focus on estimating treatment effects with more accessible conditional instruments that violate the exogeneity assumption, i.e., $\{Z_1, Z_2, \cdots, Z_m\} \not\perp U$. Inspired by deep conditional VAE, CVAE-IV [162] model a substitute $\hat{r}_\epsilon$ based on the statistical principle $Y \perp \{Z_i\}_{i=1}^m \mid T, X, \hat{r}_\epsilon$, which states that the outcome and IV candidates are conditionally independent given the treatment, observed covariates and the generated residual $\hat{r}_\epsilon$.

CVAE-IV [162] construct a conditional VAE to generate the confounder substitute $\hat{r}_\epsilon$ using multiple CIVs $Z = \{Z_i\}_{i=1}^{m}$. They apply the variational inference to model the conditional distribution $P(Y, Z \mid T, X)$. In stage 1 of IV regression, CVAE-IV [162] uses networks f_Y and f_Z to regress z_i and y_i as well as minimize the ELBO of CVAE as objective to reconstruct the latent variables $\hat{r}_\epsilon$:

$$\mathcal{L} = \textstyle\sum_i^n [(y_i - f_Y(t_i, x_i, \hat{r}_{\epsilon,i}))^2]/Var(Y) + \sum_i^n [(z_i - f_Z(t_i, x_i, \hat{r}_{\epsilon,i}))^2] + \lambda \mathcal{L}_{KL}, \tag{29}$$

where λ controls the variance of the reconstructed output. In stage 2, CVAE-IV [162] fits 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: $\mathcal{L}_{Reg} = \sum_i^n [(y_i - g_{\psi_1}(t_i, x_i) - g_{\psi_2}(\hat{r}_{\epsilon,i}))^2]$. Then, they predict the counterfactual outcome $Y(t)$ and CATE with the regression model $\{\psi_1, \psi_2\}$:

$$Y(t, x, \hat{r}_\epsilon) = g_{\psi_1}(t, x) + g_{\psi_2}(\hat{r}_\epsilon), \qquad \mathbf{CATE}(t, x, r_\epsilon) = Y(t, x, \hat{r}_\epsilon) - Y(0, x, \hat{r}_\epsilon). \tag{30}$$

By constructing the CVAE-IV model to generate an ignorable confounder substitute, one isolates the influence of the unmeasured confounder from the estimation of conditional treatment effect.

5.3 Limitation and Future Work

5.3.1 *Limitation.*

In this section, we discuss the limitations of the CFN method, focusing on the inverse problem of residuals, and the challenges of invalid or weak IVs.

Inverse Relationship. A key assumption in the CFN method is the structural assumption, which implicitly requires a one-to-one mapping (or inverse relationship) between the residuals $\hat{r}_\epsilon$ from treatment regression and the unmeasured confounders U. Otherwise, even if we recover the residuals perfectly, we cannot control the unmeasured confounders. For instance, if $\hat{r}_\epsilon = sin(U)$, controlling for $\hat{r}_\epsilon = 1$ still leaves U still has infinitely many possibilities, making it impossible to discuss and analyze the impact of these confounders.

Invalid IV and Weak IV. The effectiveness of the CFN method heavily relies on the well-defined IVs that satisfy three instrumental conditions: (1) the IV does not have a direct effect on the outcome variable, (2) the IV only affects the outcome indirectly through the treatment variable, and (3) the IV is not correlated with any unmeasured confounders. These conditions are untestable, which makes finding a valid IV more of an art than a science. Consequently, the problem of using invalid or weak IVs to implement the CFN method remains an open challenge.

5.3.2 *Future Work.*

The CFN method has demonstrated its potential in addressing causal inference challenges, and several promising directions for future work have emerged. In this section, we outline two key areas for future exploration: VAEs for unmeasured confounders recovery and conditional IVs.

VAE. Based on the concept of VAE [91], some works study the proxy variable for unmeasured confounders and attempt to use these proxies to reconstruct the unmeasured confounders [107, 174, 183]. Motivated by this, [123] developed the GCFN method to learn the distribution of unmeasured confounders and estimate the effects.

CIV. Recently, researchers [162] have proposed using CIVs, which have a direct effect on the outcome variable but indirectly through the treatment and confounders, as a substitute for valid IVs to recover unmeasured confounders. By considering the conditional independence between CIVs and the outcomes, the CVAE-IV method [162] generates a substitute for the unmeasured confounder using a conditional VAE. As a result, the exploration of CIVs represents a promising research direction for the future.

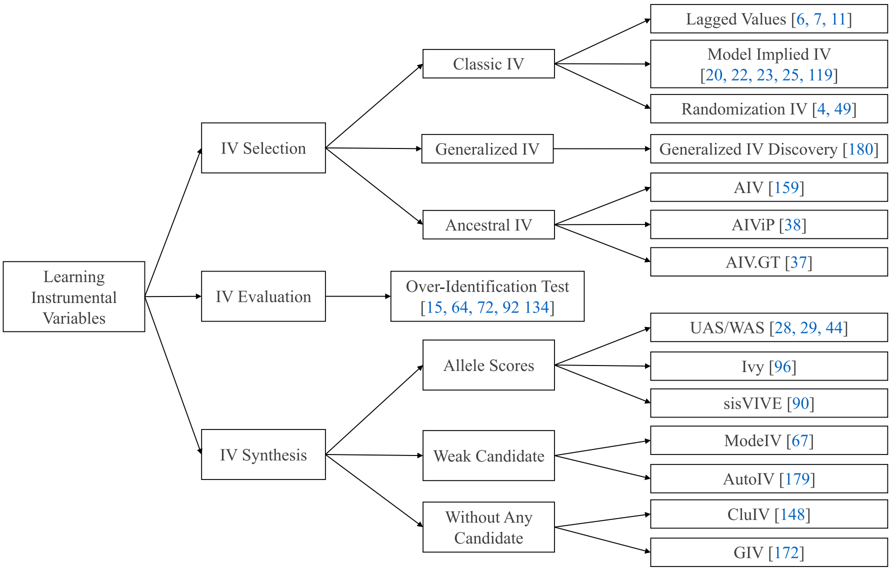


Fig. 9. Categorization of learning instrumental variables (IVs) methods.

6 Learning IVs

In Sections 4 and 5, we introduce the implementation of two-stage regression with IVs for treatment effect estimation. However, a major limitation of these methods is the requirement of a strong and valid IV [115] for treatment regression, which is often difficult to find in real-world scenarios. In this section, we provide a comprehensive overview of three research lines for finding classical IVs, generalized IVs, and Ancestral IVs. We also discuss the over-identification test for evaluating the exclusion condition of IVs. Moreover, we investigate several machine learning algorithms for learning strong IVs, known as Summary IVs. This provides a solid foundation for understanding the challenges and potential solutions in causal inference and machine learning. The overall structure of this section is depicted in Figure 9.

6.1 IV Selection

For IV-based methods to be reliable, the selected IVs must affect outcomes exclusively through their association with treatments. However, identifying suitable IVs remains a persistent challenge [21]. To this end, researchers have developed various methods to identify or test IVs.

6.1.1 Classical IV. There are three approaches for finding classical IVs: Lagged Values, Prior Knowledge of Causal Graphs, and RCTs.

Lagged Values. In panel data, a common strategy for finding IVs is using lagged values as IVs for current treatments [7]. For instance, [6] estimated the causal effect of compulsory schooling on earnings using the quarter of birth as an IV for education. Similarly, [11] employed characteristics of the respondent's childhood, husband's childhood, and parents and husband's parent as IVs to predict the respondent's probability of sending their children to school in the future. The predicted value from this model was then used as an independent variable in predicting contraceptive use.

Model Implied Instrumental Variables (**MIIVs**). Another strategy for deriving IVs from observed variables is MIIVs [20, 22, 23]. MIIVs utilize prior knowledge of a causal graph to construct the model structure, which indicates which observed variables can serve as IVs. Related to the

MIIV method is the **directed acyclic graph** (**DAG**) [25, 119], which provides rules for selecting variables that can serve as IVs: a variable's correlation with the residual term of the outcome prediction equation should be zero [22].

Randomization IVs. In the IV literature, researchers often implement RCTs to sample a random variable as an IV for intervening in the received treatments, known as the intention-to-treat variable, such as Oregon health insurance experiment [49] and the effects of military service on lifetime earnings [4]. However, RCTs can be expensive and not universally available. Sometimes, randomization is introduced by "nature" [127], resulting in natural experiments such as twin births, gender, and weather events.

6.1.2 Generalized IV. Zander et al. [180] introduce generalized IVs to extend the traditional IV framework for identifying causal effects in linear structural equation models. Generalized IVs allow for the simultaneous identification of multiple causal parameters. A set of variables $\{Z_1, \ldots, Z_k\}$ qualifies as a generalized IV relative to a set of causal variables $\{T_1, \ldots, T_k\}$ and an outcome Y if certain graphical conditions are met: (i) each Z_i has an unblocked directed path to its corresponding T_i that includes the edge $T_i \rightarrow Y$, (ii) Z_i is d-separated from Y by a set of variables C_i, and (iii) the paths $\pi_1, \ldots, \pi_k$ between $Z_1, \ldots, Z_k$ and $T_1, \ldots, T_k$ are mutually incompatible. Zander et al. propose efficient algorithms for identifying Generalized IVs in restricted cases, with a polynomial runtime for small sets.

Generalized IV Discovery. For classical IVs, they provide an $O(nm)$ algorithm that checks the basic graphical criteria, ensuring d-separation and path compatibility. For CIVs, their approach involves computing "nearest separators" C_i that block undesired paths while preserving the required connections between Z_i and T_i. This method leverages flow graph transformations and max-flow algorithms with vertex capacities to find incompatible paths efficiently, achieving a runtime of $O(nd + 3)$ for small sets of size $k = d$. For the more complex case of Generalized IVs, where multiple parameters need to be identified simultaneously, they demonstrate that testing is NP-complete but provides a polynomial-time algorithm when the size k of the set is bounded. This involves enumerating all permutations of Z and T, constructing nearest separators, and solving a **generalized vertex-disjoint paths problem** (**GVDPP**) using a pebbling game algorithm, with a runtime of $O(k(k!)^2 n^{3k+1})$. Thus, their framework systematically addresses IVs, CIVs, and Generalized IVs, offering scalable solutions for restricted cases and expanding the applicability of IV methods in causal inference.

6.1.3 AIV. While traditional IVs are widely used for causal inference, their applicability is constrained by strict exogeneity assumptions. CIVs extend the flexibility of IVs by allowing exogeneity to hold conditionally on a set of covariates $\mathbf{C}$; however, identifying CIVs is computationally expensive, requiring exponential time and rendering the problem NP-hard. To address this limitation, Van der Zander et al. [159] proposed the concept of AIVs, a specialized type of CIVs.

Definition 6.1 (AIV [159]). Variable Z is an AIV if there exists a conditional set $\mathbf{C} \subset \mathbf{X}$ such that

(1) Z correlates with T conditional on $\mathbf{C}$.
(2) $\mathbf{C}$ d-separates Z and Y.
(3) $\mathbf{C}$ consists of ancestors of Z or Y or both, and are non-descendants of Y.

AIV. This definition ensures AIVs maintain the causal identification properties of CIVs while being computationally easier to find. Van der Zander et al. [159] demonstrate that identifying AIVs can be done in polynomial time and propose the nearest separator approach to achieve this efficiently. Nearest separator $\mathbf{C}$ is a minimal set of covariates that d-separates Z and Y while ensuring that $\mathbf{C}$ only includes ancestors of Y or Z and excludes descendants of Y. This is achieved using a greedy selection process based on moralization, where undirected edges are added to represent collider relationships and simplify the identification of separating sets. After constructing **C**, the algorithm verifies whether Z remains conditionally correlated with T given **C**. If these conditions are satisfied, Z is identified as an AIV. This method operates in $O(n^2m)$, where n is the number of nodes and m is the number of edges, offering a computationally tractable solution for AIV identification.

AIV.GT. Under the definition of AIVs, Cheng et al. [38] first propose AIViP, a data-driven algorithm leveraging PAGs to identify conditioning sets for instrumentalizing AIVs without requiring a complete causal DAG. The algorithm uses the D-SEP(Z, Y) criterion to extract potential ancestors of the AIV and the outcome, ensuring Z and T are m-connected and Z and Y are m-separated. However, in practical applications, pre-specifying an AIV is often challenging. To address this, Cheng et al. [37] further propose AIV.GT, which directly identifies AIVs (Z) and their corresponding conditioning sets (**C**) from observational data using a generalized tetrad condition and consistency scoring mechanism. Specifically, AIV.GT first extracts candidate AIVs from the adjacency set of the treatment (T) and outcome (Y) variables in the learned PAG, denoted as $Z \in \text{Adj}(T \cup Y) \setminus T, Y$. For each candidate Z_i, the conditioning set $\mathbf{C}_i$ is determined as $\text{PossAn}(Z_i \cup Y) \setminus \{Z_i, T, Y\}$, where PossAn represents the set of possible ancestors. Second, AIV.GT evaluates pairs of candidate AIVs (Z_i, Z_j) using the generalized tetrad condition:

$$\sigma_{z_i * y * \mathbf{c}_i} \sigma_{z_j * t * \mathbf{c}_j} - \sigma_{z_i * t * \mathbf{c}_i} \sigma_{z_j * y * \mathbf{c}_j} = 0, \tag{31}$$

where σ denotes partial covariance. Then, the authors propose a consistency score to assess which paired variables are the most likely AIVs based on the generalized tetrad condition. Define $\epsilon_{ij} = |\sigma_{z_i * y * \mathbf{c}_i} \sigma_{z_j * t * \mathbf{c}_j} - \sigma_{z_i * t * \mathbf{c}_i} \sigma_{z_j * y * \mathbf{c}_j}|$ as the tetrad consistency, and $\delta_{ij} = |\hat{\beta}_i - \hat{\beta}_j|$ as the difference between the causal effect estimates $\hat{\beta}_i$ and $\hat{\beta}_j$ obtained using Z_i and Z_j as instruments, respectively. The consistency score is defined as $\lambda_{ij} = |\epsilon_{ij} - \delta_{ij}|$. The pair of AIVs (Z_i, Z_j) with the smallest λ_{ij}, along with their conditioning sets $(\mathbf{C}_i, \mathbf{C}_j)$, is then used for unbiased causal effect estimation through 2SLS regression.

6.2 IV Evaluation

Regardless of the method used to select IVs, it is crucial to evaluate the IV's quality. A valid IV should only affect the outcome through its strong association with treatment options, known as the exclusion assumption. If the structural assumptions for IVs are not satisfied and the correlation is weak, the instrument may provide misleading inferences about parameter estimates [62, 115].

Over-Identification Test. When there are more IVs than the number of treatments, which is known as over-identification, one can test the exclusion of IVs. The over-identification tests construct a null hypothesis that all IVs are exogenous variables, while the alternative hypothesis states that at least one IV violates exclusion (correlates with the residuals from the two-stage IV regression). In a linear setting, [134] provided a well-known over-identification test for IVs: In linear setting, [134] gave a known over-identification tests for IVs: $p = \frac{\epsilon' \bar{Z}(\bar{Z}'\bar{Z})^{-1}\bar{Z}'\epsilon}{\epsilon'\epsilon/n} \sim \mathcal{X}^2$, where ϵ are the residuals from the two-stage IV regression, and $\bar{Z}$ represents another IV (Over Identification) not involved in the causal effects estimation. Asymptotically, the test statistic p asymptotically follows a chi-square distribution, with degrees of freedom equal to the number of IVs beyond just-identification requirements [167]. Additionally, [15] introduced a similar over-identification test using the F-distribution. [92] proposed several variants for homoscedastic disturbances. In heteroscedastic-consistent cases, [64, 72] develops a statistic test for GMM.

6.3 IV Synthesis

In practice, it is challenging to find strong and valid IVs. 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 can combine IV candidates, which may not necessarily be strong or valid, into a single summary variable that is used in place of an IV for causal effect estimation [96].

6.3.1 *Allele Scores.*

In MR [30], a growing number of works have been proposed to synthesize a summary IV by combining widely available IV candidates. [27] demonstrates that summary IVs can be reproduced using summarized data on genetic associations with the treatment and the outcome. The representative approaches for combining IV candidates into a summary variable are the **unweighted/weighted allele scores** (**UAS/WAS**) [28, 29, 44]. UAS/WAS synthesize a summary variable of genetic contribution towards elevating the risk factor, which can serve as reliable IVs for inferring causal effects among clinical variables, provided that all genetic variants associated with a risk factor are independent and valid IVs [28, 139].

UAS and WAS. In MR, we can use genetic variants as IV candidates for IV synthesis. We assume K genetic variants $G = \{G_1, G_2, \ldots, G_K\}$ are actually independent weak IVs, and use them as IV candidates. Then we can obtain UAS: $UAS_{IV} = \frac{1}{K}\sum_{j=1}^{K} G_j$, where K denotes the number of IV candidates, and G_j denotes the jth IV candidate. Factually, UAS takes the average of IV candidates. 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: $WAS_{IV} = \frac{1}{K}\sum_{j=1}^{K} W_j G_j$, where W_j denotes the weights that are the same as the coefficients from the treatment regression stage.

Ivy. Allele scores rely on strong assumptions and require all IV candidates being weak IVs for estimation. To relax it, Ivy [96] only requires more than half of IV candidates should be valid. They then introduce a generalized allele score to combine valid and invalid IV candidates in a robust manner, using 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.

sisVIVE. Another innovative approach for addressing the challenge of potentially invalid IVs in MR is sisVIVE (Some Invalid Some Valid Instrumental Variables Estimator) [90]. Kang et al. [90] introduced this method to estimate the causal effects of **Body Mass Index** (**BMI**) on **Health-Related Quality of Life** (**HRQL**), employing **Single Nucleotide Polymorphisms** (**SNPs**) as IVs, even in the presence of invalid instruments. The intuition is based on the additive linear potential outcomes framework: $Y = Z^\top \alpha^* + T\beta^* + \epsilon$, where Z represents instruments, T is the treatment, Y is the outcome, and α^* and β^* are the parameters of interest. α^* captures both direct effects of instruments on outcomes and violations of the IV assumptions, while β^* is the causal effect of interest. The sisVIVE estimator solves the penalized optimization problem: $(\alpha^*, \beta^*) = \arg\min_{\alpha,\beta} \frac{1}{2}\|P_Z(Y - Z\alpha - T\beta)\|_2^2 + \lambda\|\alpha\|_1$, where P_Z is the projection matrix onto the space spanned by Z, and $\|\alpha\|_1$ is the ℓ_1-norm encouraging sparsity in α^*. This formulation assumes fewer than 50% of instruments are invalid and is computationally efficient due to its connection with Lasso. Cross-validation or theoretical guidance is used to tune λ. The method ensures robust estimation by allowing identification even without knowing the exact validity of individual instruments, and it provides theoretical guarantees under certain sparsity and regularization conditions.

6.3.2 *Generation Methods.*

Most 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 weaker assumptions for IV Synthesis.

ModeIV. [67] no longer requires more than half the number of valid IVs 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 [67] will iterate over all the elements in the set of IV candidates $G = \{G_1, G_2, \cdot, G_K\}$ and plug G_j into the IV regression method to estimate the causal effects τ_{G_j}. Then, the outcomes $\{\tau_{G_j}\}_{j=1}^{K}$ from the valid IVs must all converge to the same value, and IV candidates in the tightest cluster of estimation points just are valid IVs.

AutoIV. Furthermore, AutoIV [179] generate IV representations based on independence conditions and MI, with the assumption that all variables in the IV candidates G are independent of the unmeasured confounders U, i.e., $G \perp U$. Given the observational data $D = \{X, G, T, Y\}$, AutoIV [179] learn a disentangled representation $Z = \phi(G)$ based on independence conditions, $\hat{\phi} = \arg\min_{\phi} \mathrm{MSE}(f(\phi(\mathbf{G}), \mathbf{X}), \mathbf{T})^2, \quad \text{s.t. } \phi(\mathbf{G}) \perp \mathbf{X}, \phi(\mathbf{G}) \perp \mathbf{Y} \mid \mathbf{T}, \mathbf{X}$, where $f(\cdot)$ denotes a regression network of $\phi(G), X$ to predict treatment variables. According to the independence conditions, AutoIV [179] obtain valid IVs that do not have a direct effect on the outcome variable, only indirectly through the treatment variable. To learn relevance and exclusion, AutoIV [179] constructs $\mathcal{L}_{XY}^{LLD}$ and $\mathcal{L}_{XY|V}^{MI}$ to denote the MI and conditional MI. Then, AutoIV [179] (1) maximize $\mathcal{L}_{GT}^{MI}$ to optimize the IV representations $\phi(G)$ for relevance condition; (2) minimize $\mathcal{L}_{GY|T}^{MI}$ to optimize the IV representations $\phi(G)$ for exclusion condition; and (3) minimize $\mathcal{L}_{GX}^{MI}$ to optimize the IV representations $\phi(G)$ for exogeneity condition.

6.3.3 Without Any Candidates. Although existing IV generation methods have moved away from manually selecting pre-defined IVs, they still rely on high-quality candidate sets with at least half valid IVs or exogeneity assumptions, which is sometimes impractical due to cost constraints and limited expert knowledge. Therefore, it is highly demanded to automatically obtain valid IVs directly from observed variables $\{X, T, Y\}$. In 2021, clustering-based methods, such as CluIV [148] and **Group IV** (**GIV**) [172], emerged to address these issues. GIV [172], in particular, introduces a novel algorithm (Meta-EM) that models latent IVs and implements a data-driven approach to automatically reconstruct valid GIVs from observed variables, beyond hand-made IV candidates.

GIV. With the advent of the big data era, various observation databases sourced from different sources have emerged, which may contain the same treatment effect mechanism but different treatment assignment mechanisms. 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, Wu et al. [172] 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 superiority of the Meta-EM approach over existing methods.

7 Available Datasets and Codes/Packages

7.1 Datasets

In real-world applications, finding a strictly valid IV from observational data is challenging due to untestable exclusion and exogeneity conditions. Pre-defined IVs and IV candidates, selected through human effort, may be invalid if they do not strictly satisfy valid IV conditions, especially when prior knowledge is insufficient. Furthermore, ground truth dose-response functions (ATE, ATT, CATE, or ITE) are unavailable in observational datasets due to the absence of counterfactual outcomes. Consequently, IV-based works often employ (semi-)synthetic datasets, such as Demand [66] and Toy Datasets [18, 46], or combine prior specific knowledge with observational control datasets.

Low-dimensional Toy [18, 46]. In low-dimensional cases, [18] generated data via the following process: $Y = g(T) + U + \delta, T = Z + U + \gamma$, where $Z \sim \text{Uniform}(-3, 3), U \sim \mathcal{N}(0, 1), \delta, \gamma \sim \mathcal{N}(0, 0.1)$.

Similarity, [46] consider the following data-generating processes: $Y = g(T) + U + \delta, T = \gamma Z + (1 - \gamma)U + \gamma$, where $Z \sim \mathcal{N}(0, 2), U \sim \mathcal{N}(0, 2), \delta, \gamma \sim \mathcal{N}(0, 0.1)$. Keeping the data generating process fixed, [18, 46] design various true response function $g(T) = sin(x)/|x|/x$.

MNIST (http://yann.lecun.com/exdb/mnist/) [18]. Similar to [66], in high-dimensional cases, [18] use same data generating process introduced in Low-dimensional Toy, based on the MNIST dataset [98], 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 (https://github.com/jhartford/DeepIV) [66]. The demand simulation design is from [66], 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). The study incorporates fuel price (Z) as the IV, customer types (X_1), time of year (X_2) as an observable confounder, and conferences (U) as an unmeasured confounder.

Infant Health and Development Program (IHDP) (http://www.fredjo.com) [87, 140]. The IHDP, from an RCT, assesses whether the future cognition of premature infants is affected by specialist home visits. 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, [172] generate 2-dimension random variables for each unit. Then, [172] select a subset of pre-treatment variables as the confounders unobserved confounders U. With known treated and control potential outcome (accessible in IHDP), [172] designs the treatment assignment policy as: $P(T \mid Z, X) = \frac{1}{1+\exp\left(-(\sum_{i=1}^{2} Z_i + \sum_{i=1}^{m_X} X_i) + \sum_{i=1}^{m_U} U_i)\right)}, T \sim Bernoulli(P(T \mid Z, X))$, where $Z_1, Z_2 \sim \mathcal{N}(0, 1)$, m_X and m_U are the dimensions of X and U selected from the IHDP.

PISA (https://www.oecd.org/pisa/data/) [121]. The PISA survey aims to evaluate the students' ability to apply their knowledge and skills to real-life situations [121], covering three main domains: reading (131 items), mathematics (35 items), and science (53 items). [47, 138] selected 4,951 participants in March 2009, 4,041 participants in October 2009, and 3,989 participants in April 2010 and there were 3,472 students who participated in all three rounds. The distance to school was expressed in the number of minutes as an instrument. Gender and type of school (General comprehensive, Vocational with comprehensive program, and Basic vocational school) are used as covariates.

Avon Longitudinal Study of Parents and Children (ALSPAC) (http://www.alspac.bris.ac.uk) [117, 158]. The ALSPAC is a longitudinal birth cohort study of 14541 pregnant women resident in Avon, UK, with expected dates of delivery ranging from April 1991 to December 1992 [56]. Similar to [158], through selection, Palmer et al. [117] uses 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 (https://www.mrbase.org/) [76]. Hemani et al. [76] develops an MR-Base platform that integrates a curated database of complete GWAS results, which uses genetic variants as IVs. The database comprises 11 billion SNP-trait associations from 1,673 GWAS and is under updated.

7.2 Codes/Packages

In this part, we provide a summary of the available code resources for IVs and causal inference, as shown in Table 1. Furthermore, we have reproduced and integrated these codes, resulting in the creation of an open-source toolbox called **MLIV**.[5]

[5] Project Page: https://github.com/causal-machine-learning-lab/mliv

Table 1. Available Codes of Methods for IVs and Causal Inference

IV-based Methods (Mainly focus on Machine Learning)		
Method	**Language**	**Link**
DeepIV [66]	python	https://github.com/jhartford/DeepIV
KernelIV [146]	Matlab	https://github.com/r4hu1-5in9h/KIV
DualIV [112]	Matlab	https://github.com/krikamol/DualIV-NeurIPS2020
DFIV [175]	python	https://github.com/liyuan9988/DeepFeatureIV
DeepGMM [18]	python	https://github.com/CausalML/DeepGMM
AGMM [46]	python	https://github.com/microsoft/AdversarialGMM
CBIV [171]	python	https://github.com/anpwu/CB-IV
CBRL.CIV [41]	python	https://openreview.net/forum?id=qDhq1icpO8
AutoIV [179]	python	https://github.com/junkunyuan/AutoIV
AIV.GT [37]	R	https://github.com/chengdb2016/AIV.GT
econML	python	https://github.com/microsoft/EconML
CausalDCD	python	https://github.com/anpwu/Awesome-Instrumental-Variable

8 Applications

In practice, unmeasured confounders are frequently encountered, making the use of IVs regression algorithms essential in various machine learning applications, such as graph network learning, **reinforcement learning** (**RL**), recommendation system, and computer vision.

8.1 Sociology and Social Sciences

In sociology and social sciences, causal inference aims to study the association between social networks and interference, also known as peer effects [51, 86, 116]. Peer effects refer to the influence of an individual's peers (those they are connected to or alter) on their behavior [116]. However, eliminating peer effects in observational data is challenging due to contextual confounding, peer selection, simultaneity bias, and measurement error [1].

Taking the city-level characteristics serve as instruments, [52] 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 one's chance of dropping out, [51] used characteristics of the local labor market (or city) as instruments, and the results suggested that neighborhood conditions do influence an individual's likelihood of finishing high school. Besides, researchers and data scientists have an increasing interest in social network services in Facebook, X (formerly known as Twitter), WeChat, and so on [61], which are collectively called "social media". Adopting an IVs approach, [61] explored the effect of social network services on social capital. [61] suggested that high-intensity users are higher in network social capital than non-users of social network services.

8.2 RL

In RL, an agent aims to maximize cumulative rewards by taking actions in an environment [109, 165]. Many RL concepts are analogous to causal inference, such as treatment as the agent's action, the environment as the confounder, and cumulative reward as the outcome [50, 54]. Due to the Markov property, reinforcement learners typically have access to environment information, satisfying the unconfoundedness assumption [88, 89, 99]. Common methods in RL for unbiased reward estimation include importance sampling weighting and doubly robust policy evaluation [48, 122]. Under unconfoundedness, various approaches can estimate the state-action value (Q-function) [97, 152, 153, 185]. To relax the unconfoundedness assumption, IVs have been introduced for policy optimization [36, 102, 105, 176]. In the context of **offline policy evaluation** (**OPE**), improved Q-function estimators have been proposed using IV techniques, leading to competitive new methods [36]. By leveraging IVs, a **conditional moment restriction** (**CMR**) has been derived, and an **IV-aided Value Iteration** (**IVVI**) algorithm has been developed based on a primal-dual reformulation of CMR [105]. Furthermore, recent work applies IV Regression to correct bias in RL algorithms under time-dependent noise [102].

8.3 Recommendation System

Another application, highly correlated with the treatment effect estimation, is the recommendation system [135, 153, 164, 177]. Exposing the user to an item can be viewed as a specific treatment and the user's behavior (click or activity) is the corresponding outcome. To eliminate the bias from the unmeasured confounders and the self-selection of the users, [142] proposed an IV 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, [143] embedded users' search behaviors as IVs and implemented a two-stage regression for an unbiased estimate of causal effect.

8.4 Computer Vision

Computer Vision is a typical field of **artificial intelligence** (**AI**), suffering from unstable learning and lacking generalization ability [136]. To achieve a proactive defense against adversarial examples, [155] proposed to use the IV that achieves causal intervention. Using "retinotopic sampling" as IV [8], **Causal intervention by instrumental Variable** (**CiiV**) [155] algorithm implements a spatial data augmentation using different retinotopic sampling masks and learns features linearly responding to spatial interpolations. In Domain Adaptation, [178] claimed that the input features of one domain are valid IVs for other domains. Inspired by this finding, we design a simple yet effective framework to learn the **Domain-invariant Relationship with Instrumental Variables** (**DRIVE**) via a two-stage IV method.

9 Conclusion

9.1 Future Direction

Relaxing IV Assumptions An instrument must meet three assumptions: relevance assumption, exclusion assumption, and exogeneity assumption. To relax the exclusion assumption, we can use mediators to block out the direct effect of the IV on the outcomes. For CIVs, we can attempt to recover the unmeasured confounders affecting the IV based on conditional independence constraints and adjust accordingly.

Combining IV Regression with Confounder Control Traditional IV regression methods often overlook the bias caused by observed confounding variables. Even with the use of CFN, investigators do not control for confounding of the recovered residuals. Developing a more robust IV regression method that considers confounder balance is a promising direction.

Reducing Unmeasured Confounding without IVs In reality, IVs may not always be available. Historically, observational datasets and randomized controlled experiments have been considered separately. However, even if randomized controlled experiments are expensive, small-scale randomized controlled experiments can still be conducted. By considering small intervention data and a large amount of observational data, i.e., data fusion, it is possible to establish causality without confounding bias. This approach presents an opportunity to further advance the field and improve our understanding of causal relationships in complex real-world problems.

9.2 Conclusion

In this survey, we provide a comprehensive overview of IVs in causal inference and machine learning, emphasizing their importance in addressing unmeasured confounders affecting treatment and outcome variables. We delved into the identification conditions of IV regression methods under common assumptions and explored three research areas: 2SLS regression, CFN approaches, and IV generation techniques, covering both classical causal inference and recent advancements in machine learning. We summarized available datasets, algorithms, and real-world applications in fields like graph network learning, reinforcement learning, recommendation systems, and computer vision. Additionally, we identify open problems and suggest future research directions to advance causal inference and machine learning. By bridging the gap between causal inference and machine learning, this survey aims to inspire further research and development, leading to a better understanding of complex systems and more informed decision-making processes.

References

[1] Weihua An. 2015. Instrumental variables estimates of peer effects in social networks. *Social Science Research* 50 (2015), 382–394.

[2] Joshua D. Angrist and Guido W. Imbens, 1991. Sources of identifying information in evaluation models. *NBER Technical Working Papers 0117*. National Bureau of Economic Research, Inc. (1991).

[3] Joshua D. Angrist and Guido W. Imbens. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62, 2 (1994), 467–475.

[4] Joshua D. Angrist. 1990. Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review* 80, 3 (1990), 313–336.

[5] Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91, 434 (1996), 444–455.

[6] Joshua D. Angrist and Alan B. Keueger. 1991. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics* 106, 4 (1991), 979–1014.

[7] Luc Anselin. 1988. *Spatial Econometrics: Methods and Models*. Springer Science and Business Media.

[8] Michael J. Arcaro, Stephanie A. McMains, Benjamin D. Singer, and Sabine Kastner. 2009. Retinotopic organization of human ventral visual cortex. *Journal of Neuroscience* 29, 34 (2009), 10638–10652.

[9] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*. PMLR, 214–223.

[10] Susan Athey, Guido W. Imbens, and Stefan Wager. 2018. Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B* 80, 4 (2018), 597–623.

[11] William G. Axinn and Jennifer S. Barber. 2001. Mass education and fertility transition. *American Sociological Review* 66, 4 (2001), 481–505.

[12] Edvard Bakhitov and Amandeep Singh. 2022. Causal gradient boosting: Boosted instrumental variable regression. In *Proceedings of the 23rd ACM Conference on Economics and Computation*. 604–605.

[13] Heejung Bang and James M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61, 4 (2005), 962–973.

[14] Burt Barnow, Glen Cain, and Arthur Goldberg. 1981. Selection on observables. *Evaluation Studies Review Annual* 5 (1981), 43–59.

[15] Robert L. Basmann. 1960. On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association* 55, 292 (1960), 650–659.

[16] Christopher F. Baum, Mark E. Schaffer, and Steven Stillman. 2003. Instrumental variables and GMM: Estimation and testing. *The Stata Journal* 3, 1 (2003), 1–31.

[17] Andrew Bennett and Nathan Kallus. 2023. The variational method of moments. 85, 3 (2023), 810–841.

[18] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. 2019. Deep generalized method of moments for instrumental variable analysis. *Advances in Neural Information Processing Systems* 32 (2019), 3564–3574.

[19] Richard Blundell and James L. Powell. 2003. Endogeneity in nonparametric and semiparametric regression models. *Econometric Society Monographs* 36 (2003), 312–357.

[20] Kenneth A. Bollen. 1996. An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika* 61, 1 (1996), 109–121.

[21] Kenneth A. Bollen. 2012. Instrumental variables in sociology and the social sciences. *Annual Review of Sociology* 38 (2012), 37–72.

[22] Kenneth A. Bollen. 2019. Model implied instrumental variables (MIIVs): An alternative orientation to structural equation modeling. *Multivariate Behavioral Research* 54, 1 (2019), 31–46.

[23] Kenneth A. Bollen and Daniel J. Bauer. 2004. Automating the selection of model-implied instrumental variables. *Sociological Methods and Research* 32, 4 (2004), 425–452.

[24] Byron Boots, Geoffrey Gordon, and Arthur Gretton. 2013. Hilbert space embeddings of predictive state representations. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence* 13 (2013), 92–101.

[25] Carlos Brito and Judea Pearl. 2002. A graphical criterion for the identification of causal effects in linear models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2002), 533–539.

[26] M. Alan Brookhart and Sebastian Schneeweiss. 2007. Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *The International Journal of Biostatistics* 3, 1 (2007), 1–23.

[27] Stephen Burgess, Frank Dudbridge, and Simon G. Thompson. 2016. Combining information on multiple instrumental variables in Mendelian randomization: Comparison of allele score and summarized data methods. *Statistics in Medicine* 35, 11 (2016), 1880–1906.

[28] Stephen Burgess, Dylan S. Small, and Simon G. Thompson. 2017. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research* 26, 5 (2017), 2333–2355.

[29] Stephen Burgess and Simon G. Thompson. 2013. Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology* 42, 4 (2013), 1134–1144.

[30] Stephen Burgess and Simon G. Thompson. 2015. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation.* CRC Press.

[31] A. Colin Cameron and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications.* Cambridge University Press.

[32] Xiaohong Chen. 2007. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics* 6, B (2007), 5549–5632.

[33] Xiaohong Chen and Timothy M. Christensen. 2018. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics* 9, 1 (2018), 39–84.

[34] Xiaohong Chen and Demian Pouzo. 2012. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica* 80, 1 (2012), 277–321.

[35] Xiaohong Chen and Xiaotong Shen. 1998. Sieve extremum estimates for weakly dependent data. *Econometrica* 66, 2 (1998), 289–314.

[36] Yutian Chen, Liyuan Xu, Caglar Gulcehre, Tom Le Paine, Arthur Gretton, Nando de Freitas, and Arnaud Doucet. 2021. On instrumental variable regression for deep offline policy evaluation. arXiv:2105.10148. Retrieved from https://arxiv.org/abs/2105.10148

[37] Debo Cheng, Jiuyong Li, Lin Liu, Kui Yu, Thuc Duy Le, and Jixue Liu. 2023. Discovering ancestral instrumental variables for causal inference from observational data. *IEEE Transactions on Neural Networks and Learning Systems* 35, 8 (2023), 11542–11552.

[38] Debo Cheng, Jiuyong Li, Lin Liu, Jiji Zhang, Thuc Duy Le, and Jixue Liu. 2022. Ancestral instrument method for causal inference without complete knowledge. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI-22.* Lud De Raedt (Ed.), International Joint Conferences on Artificial Intelligence Organization, 4843–4849. DOI:https://doi.org/10.24963/ijcai.2022/671 Main Track.

[39] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Thuc Duy Le, and Jixue Liu. 2023. Learning conditional instrumental variable representation for causal effect estimation. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 525–540.

[40] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, and Thuc Duy Le. 2023. Causal inference with conditional instruments using deep generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 7122–7130.

[41] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, and Thuc Duy Le. 2024. Conditional instrumental variable regression with representation learning for causal inference. In *Proceedings of the 12th International Conference on Learning Representations.* Retrieved from https://openreview.net/forum?id=qDhq1icpO8

[42] Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. 2020. Adversarial estimation of riesz representers. arXiv:2101.00009. Retrieved from https://arxiv.org/abs/2101.00009

[43] Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. 2017. Learning from conditional distributions via dual embeddings. In *Proceedings of the Artificial Intelligence and Statistics.* PMLR, 1458–1467.

[44] Neil M. Davies, Stephanie von Hinke Kessler Scholder, Helmut Farbmacher, Stephen Burgess, Frank Windmeijer, and George Davey Smith. 2015. The many weak instruments problem and Mendelian randomization. *Statistics in Medicine* 34, 3 (2015), 454–468.

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805

[46] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. 2020. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems* (2020), 12248–12262.

[47] Henryk Domański, Michał Federowicz, Artur Pokropek, Dariusz Przybysz, Michał Sitek, Marek Smulczyk, and Tomasz Żółtak. 2012. From school to work: Individual and institutional determinants of educational and occupational career trajectories of young Poles. *ASK: Research and Methods* 21, 1 (2012), 123–141.

[48] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. *Proceedings of the 28th International Conference on International Conference on Machine Learning* (2011). 1097–1104.

[49] Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. 2012. The Oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics* 127, 3 (2012), 1057–1106.

[50] Andrew Forney, Judea Pearl, and Elias Bareinboim. 2017. Counterfactual data-fusion for online reinforcement learners. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1156–1164.

[51] E. Michael Foster. 1997. Instrumental variables for logistic regression: An illustration. *Social Science Research* 26, 4 (1997), 487–504.

[52] E. Michael Foster and Sara McLanahan. 1996. An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school? *Psychological Methods* 1, 3 (1996), 249.

[53] A. Ronald Gallant. 1987. Identification and consistency in seminonparametric regression. *Econometrics* 1 (1987), 145–170.

[54] Samuel J. Gershman. 2017. Reinforcement learning and causal models. *The Oxford Handbook of Causal Reasoning* 1, 17 (2017), 295–306.

[55] Arthur S. Goldberger. 1972, reprinted 2008. Selection bias in evaluating treatment effects: Some formal illustrations. In *Proceedings of the Modelling and Evaluating Treatment Effects in Econometrics*. Emerald Group Publishing Limited.

[56] Jean Golding, Marcus Pembrey, Richard Jones, and ALSPAC Study Team. 2001. ALSPAC -the avon longitudinal study of parents and children. I. study methodology. *Paediatric and Perinatal Epidemiology* 15, 1 (2001), 74–87.

[57] William H. Greene. 2003. *Econometric Analysis*. Pearson Education India.

[58] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys* 53, 4 (2020), 1–37.

[59] Zijian Guo and Dylan S. Small. 2016. Control function instrumental variable estimation of nonlinear causal effect models. *The Journal of Machine Learning Research* 17, 1 (2016), 3448–3482.

[60] Trygve Haavelmo. 1943. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society* 11, 1 (1943), 1–12.

[61] Sehee Han and Kyung-Gook Park. 2020. Social network services and their effects on network social capital: An instrumental variables approach. *International Journal of Mobile Communications* 18, 4 (2020), 386–404.

[62] Bruce Hansen. 2022. *Econometrics*. Princeton University Press.

[63] Bruce E. Hansen. 2000. Testing for structural change in conditional models. *Econometrics* 97, 1 (2000), 93–115.

[64] Lars Peter Hansen. 1982. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 50, 4 (1982), 1029–1054.

[65] Lars Peter Hansen, John Heaton, and Amir Yaron. 1996. Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics* 14, 3 (1996), 262–280.

[66] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1414–1423.

[67] Jason S. Hartford, Victor Veitch, Dhanya Sridhar, and Kevin Leyton-Brown. 2021. Valid causal inference with (some) invalid instruments. In *Proceedings of the International Conference on Machine Learning*. PMLR, 4096–4106.

[68] Fernando Pires Hartwig, Linbo Wang, George Davey Smith, and Neil Martin Davies. 2023. Average causal effect estimation via instrumental variables: The no simultaneous heterogeneity assumption. *Epidemiology* 34, 3 (2023), 325–332.

[69] Fernando Pires Hartwig, Linbo Wang, George Davey Smith, and Neil Martin Davies. 2021. 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:2107.01070. Retrieved from https://arxiv.org/abs/2107.01070

[70] Negar Hassanpour and Russell Greiner. 2020. Learning disentangled representations for counterfactual regression. In *Proceedings of the International Conference on Learning Representations*.

[71] Jerry A. Hausman. 1978. Speciffcation tests in econometrics. *Econometrica: Journal of the Econometric Society* 46, 6 (1978), 1251–1271.

[72] Fumio Hayashi. 2011. *Econometrics*. Princeton University Press.

[73] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[74] James Heckman. 1990. Varieties of selection bias. *The American Economic Review* 80, 2 (1990), 313–318.

[75] Ahmed Hefny, Carlton Downey, and Geoffrey J. Gordon. 2015. Supervised learning for dynamical system learning. *Advances in Neural Information Processing Systems* 28 (2015).

[76] Gibran Hemani, Jie Zheng, Benjamin Elsworth, Kaitlin H. Wade, Valeriia Haberland, Denis Baird, Charles Laurin, Stephen Burgess, Jack Bowden, Ryan Langdon, et al. 2018. The MR-Base platform supports systematic causal inference across the human phenome. *Elife* 7 (2018), e34408.

[77] Miguel A. Hernán and James M. Robins. 2006. Instruments for causal inference: An epidemiologist's dream? *Epidemiology* 17, 4 (2006), 360–372.

[78] Miguel A. Hernan and James M. Robins. 2010. Causal Inference.

[79] Miguel A. Hernán and James M. Robins. 2024. Causal Inference: What If. CRC Press. (2024).

[80] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations* (2017).

[81] Joel L. Horowitz. 2011. Applied nonparametric instrumental variables estimation. *Econometrica* 79, 2 (2011), 347–394.

[82] Øyvind Hoveid. 2021. Constructing valid instrumental variables in generalized linear causal models from directed acyclic graphs. arXiv:2102.08056. Retrieved from https://arxiv.org/abs/2102.08056

[83] Guido Imbens. 2014. *Instrumental Variables: An Econometrician's Perspective.* Technical Report. National Bureau of Economic Research.

[84] Guido Imbens and Jeffrey Wooldridge. 2007. Control function and related methods. *What's New in Econometrics* 6 (2007), 420–445.

[85] Guido W. Imbens and Jeffrey M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47, 1 (2009), 5–86.

[86] Christopher Jencks and Susan E. Mayer. 1990. The social consequences of growing up in a poor neighborhood. *Innercity Poverty in the United States* 111 (1990), 111–186.

[87] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3020–3029.

[88] Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. 1996. Reinforcement learning: A survey. *Journalof Artiffcial Intelligence Research* 4, 1 (1996), 237–285.

[89] Nathan Kallus and Angela Zhou. 2020. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 22293–22304.

[90] Hyunseung Kang, Anru Zhang, T. Tony Cai, and Dylan S. Small. 2016. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association* 111, 513 (2016), 132–144.

[91] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. *International Conference on Learning Representations* (2014).

[92] James B. Kirby and Kenneth A. Bollen. 2009. 10. Using instrumental variable tests to evaluate model specification in latent variable structural equation models. *Sociological Methodology* 39, 1 (2009), 327–355.

[93] Rainer Kress, V. Maz'ya, and V. Kozlov. 1989. *Linear Integral Equations.* Vol. 82. Springer.

[94] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM* 60, 6 (2017), 84–90.

[95] Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, and Zhichao Jiang. 2020. Causal inference. *Engineering* 6, 3 (2020), 253–263.

[96] Zhaobin Kuang, Frederic Sala, Nimit Sohoni, Sen Wu, Aldo Córdova-Palomera, Jared Dunnmon, James Priest, and Christopher Ré. 2020. Ivy: Instrumental variable synthesis for causal inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics.* PMLR, 398–410.

[97] Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. 2021. Sequential causal imitation learning with unobserved confounders. *Advances in Neural Information Processing Systems* 34 (2021), 14669–14680.

[98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86, 11 (1998), 2278–2324.

[99] Chen Lei. 2021. Deep reinforcement learning. In *Proceedings of the Deep Learning and Practice with MindSpore.* Springer, 217–243.

[100] Baohong Li, Anpeng Wu, Ruoxuan Xiong, and Kun Kuang. 2024. Two-stage shadow inclusion estimation: An IV approach for causal inference under latent confounding and collider bias. In *Proceedings of the 41st International Conference on Machine Learning*. Retrieved from https://openreview.net/forum?id=YRWdiaupCr

[101] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. 2017. Mmd gan: Towards deeper understanding of moment matching network. *Advances in Neural Information Processing Systems* 30 (2017).

[102] Jin Li, Ye Luo, Zigan Wang, Xiaowei Zhang. 2024. Asymptotic theory for IV-based reinforcement learning with potential endogeneity. arXiv:2103.04021. Retrieved from https://arxiv.org/abs/2103.04021

[103] Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. 2016. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns.. In *Proceedings of the IJCAI*. 3768–3774.

[104] Luofeng Liao, You-Lin Chen, Zhuoran Yang, Bo Dai, Mladen Kolar, and Zhaoran Wang. 2020. Provably efficient neural estimation of structural equation models: An adversarial approach. *Advances in Neural Information Processing Systems* 33 (2020), 8947–8958.

[105] Luofeng Liao, Zuyue Fu, Zhuoran Yang, Yixin Wang, Dingli Ma, Mladen Kolar, and Zhaoran Wang. 2024. Instrumental variable value iteration for causal offline reinforcement learning. *Journal of Machine Learning Research* 25, 303 (2024), 1–56.

[106] Adi Lin, Jie Lu, Junyu Xuan, Fujin Zhu, and Guangquan Zhang. 2019. One-stage deep instrumental variable method for causal inference from observational data. In *Proceedings of the ICDM 2019*. IEEE, 419–428.

[107] Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems* 30 (2017).

[108] Wang Miao, Lan Liu, Yilin Li, Eric J. Tchetgen Tchetgen, and Zhi Geng. 2024. Identification and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *ACM/JMS Journal of Data Science* 1, 2 (2024), 1–23.

[109] Marvin Lee Minsky. 1954. *Theory of Neural-analog Reinforcement Systems and its Application to the Brain-model Problem*. Princeton University.

[110] Lauren E. Mokry, Omar Ahmad, Vincenzo Forgetta, George Thanassoulis, and J. Brent Richards. 2015. Mendelian randomisation applied to drug development in cardiovascular disease: A review. *Journal of Medical Genetics* 52, 2 (2015), 71–79.

[111] Stephen L. Morgan and Christopher Winship. 2015. *Counterfactuals and Causal Inference*. Cambridge University Press.

[112] Krikamol Muandet, Arash Mehrjou, Si Le Kai, and Anant Raj. 2020. Dual instrumental variable regression. In *Proceedings of the NeurIPS 2020*.

[113] Whitney K. Newey. 2013. Nonparametric instrumental variables estimation. *American Economic Review* 103, 3 (2013), 550–56.

[114] Whitney K. Newey and James L. Powell. 2003. Instrumental variable estimation of nonparametric models. *Econometrica* 71, 5 (2003), 1565–1578.

[115] Austin Nichols. 2006. Weak instruments: An overview and new techniques. In *Proceedings of the Stata 5th North American Meeting Presentation*.

[116] A. James O'Malley, Felix Elwert, J. Niels Rosenquist, Alan M. Zaslavsky, and Nicholas A. Christakis. 2014. Estimating peer effects in longitudinal dyadic data using instrumental variables. *Biometrics* 70, 3 (2014), 506–515.

[117] Tom M. Palmer, Debbie A. Lawlor, Roger M. Harbord, Nuala A. Sheehan, Jon H. Tobias, Nicholas J. Timpson, George Davey Smith, and Jonathan AC Sterne. 2012. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research* 21, 3 (2012), 223–242.

[118] Judea Pearl. 2009. *Causality*. Cambridge University Press.

[119] Judea Pearl. 2010. The foundations of causal inference. *Sociological Methodology* 40, 1 (2010), 75–149.

[120] Amil Petrin and Kenneth Train. 2010. A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research* 47, 1 (2010), 3–13.

[121] Artur Pokropek. 2016. Introduction to instrumental variables and their application to large-scale assessment data. *Large-scale Assessments in Education* 4, 1 (2016), 1–20.

[122] Doina Precup. 2000. Eligibility traces for off-policy policy evaluation. *International Conference on Machine Learning* (2000), 759–766.

[123] Aahlad Puli and Rajesh Ranganath. 2020. General control functions for causal effect estimation from IVs. *Advances in Neural Information Processing Systems* 33 (2020), 8440–8451.

[124] Olav Reiersøl. 1950. Identifiability of a linear relation between variables which are subject to error. *Econometrica: Journal of the Econometric Society* 18, 4 (1950), 375–389.

[125] Douglas Rivers and Quang H. Vuong. 1988. Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics* 39, 3 (1988), 347–366.

[126] James M. Robins, Miguel Angel Hernan, and Babette Brumback. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11, 5 (2000), 550–560.

[127] Mark R. Rosenzweig and Kenneth I. Wolpin. 2000. Natural "natural experiments" in economics. *Journal of Economic Literature* 38, 4 (2000), 827–874.

[128] Kenneth J. Rothman and Sander Greenland. 2005. Causation and causal inference in epidemiology. *American Journal of Public Health* 95, S1 (2005), S144–S150.

[129] Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 5 (1974), 688.

[130] Donald B. Rubin. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* 6, 1 (1978), 34–58.

[131] Donald B. Rubin. 1990. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* 5, 4 (1990), 472–480.

[132] Sorawit Saengkyongam, Leonard Henckel, Niklas Pfister, and Jonas Peters. 2022. Exploiting independent instruments: Identification and distribution generalization. In *Proceedings of the International Conference on Machine Learning*. PMLR, 18935–18958.

[133] Eleanor Sanderson, M. Maria Glymour, Michael V. Holmes, Hyunseung Kang, Jean Morrison, Marcus R. Munafò, Tom Palmer, C. Mary Schooling, Chris Wallace, Qingyuan Zhao, et al. 2022. Mendelian randomization. *Nature Reviews Methods Primers* 2, 1 (2022), 6.

[134] John D. Sargan. 1958. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society* 26, 3 (1958), 393–415.

[135] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the ICML*. PMLR, 1670–1679.

[136] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Towards causal representation learning. arXiv:2102.11107. Retrieved from https://arxiv.org/abs/2102.11107

[137] Bernhard Schölkopf and Alexander J. Smola. 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press. (2002).

[138] Wolfram Schulz, John Ainley, and Julian Fraillon. 2011. ICCS 2009 technical report. (2011).

[139] Paola Sebastiani, Nadia Solovieff, and Jenny Sun. 2012. Naïve Bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: Not so different after all! *Frontiers in Genetics* 3 (2012), 26. https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2012.00026/full

[140] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3076–3085.

[141] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. 2014. Lectures on stochastic programming: Modeling and theory. *SIAM* 16 (2014).

[142] Amit Sharma, Jake M. Hofman, and Duncan J. Watts. 2015. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the 16th ACM Conference on Economics and Computation*. 453–470.

[143] Zihua Si, Xueran Han, Xiao Zhang, Jun Xu, Yue Yin, Yang Song, and Ji-Rong Wen. 2022. A model-agnostic causal learning framework for recommendation using search data. arXiv:2202.04514. Retrieved from https://arxiv.org/abs/2202.04514

[144] Ricardo Silva and Shohei Shimizu. 2017. Learning instrumental variables with structural and non-gaussianity assumptions. *Journal of Machine Learning Research* 18, 120 (2017), 1–49.

[145] Amandeep Singh, Kartik Hosanagar, and Amit Gandhi. 2020. Machine learning instrument variables for causal inference. In *Proceedings of the 21st ACM Conference on Economics and Computation*. 835–836.

[146] Rahul Singh, Maneesh Sahani, and Arthur Gretton. 2019. Kernel instrumental variable regression. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 4593–4605.

[147] Richard J. Smith and Richard W. Blundell. 1986. An exogeneity test for a simultaneous equation Tobit model with an application to labor supply. *Econometrica: Journal of the Econometric Society* 54, 3 (1986), 679–685.

[148] Nataliya Sokolovska and Pierre-Henri Wuillemin. 2021. The role of instrumental variables in causal inference based on independence of cause and mechanism. *Entropy* 23, 8 (2021), 928.

[149] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. 2009. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the Annual International Conference on Machine Learning*. 961–968.

[150] James H. Stock and Francesco Trebbi. 2003. Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives* 17, 3 (2003), 177–194.

[151] James H. Stock, Mark W. Watson, et al. 2003. *Introduction to Econometrics*. Addison Wesley Boston.

[152] Adith Swaminathan and Thorsten Joachims. 2015. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the International Conference on Machine Learning*. PMLR, 814–823.

[153] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. 2017. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems* 30 (2017).

[154] Kaihua Tang, Mingyuan Tao, Jiaxin Qi, Zhenguang Liu, and Hanwang Zhang. 2022. Invariant feature learning for generalized long-tailed classiffcation. *European Conference on Computer Vision. Cham: Springer Nature Switzerland* (2022), 709–726.

[155] Kaihua Tang, Mingyuan Tao, and Hanwang Zhang. 2021. Adversarial visual robustness by causal intervention. arXiv:2106.09534. Retrieved from https://arxiv.org/abs/2106.09534

[156] Lester G. Telser. 1964. Iterative estimation of a set of linear regression equations. *Journal of the American Statistical Association* 59, 307 (1964), 845–862.

[157] Joseph V. Terza, Anirban Basu, and Paul J. Rathouz. 2008. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics* 27, 3 (2008), 531–543.

[158] Nicholas J. Timpson, Adrian Sayers, George Davey-Smith, and Jonathan H. Tobias. 2009. How does body fat influence bone mass in childhood? A Mendelian randomization approach. *Journal of Bone and Mineral Research* 24, 3 (2009), 522–533.

[159] Benito Van der Zander, Johannes Textor, and Maciej Liskiewicz. 2015. Efficiently finding conditional instruments for causal inference. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence.*

[160] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).

[161] Abraham Wald. 1940. The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics* 11, 3 (1940), 284–300.

[162] Haotian Wang, Wenjing Yang, Longqi Yang, Anpeng Wu, Liyang Xu, Jing Ren, Fei Wu, and Kun Kuang. 2022. Estimating individualized causal effect with confounded instruments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1857–1867.

[163] Linbo Wang and Eric Tchetgen Tchetgen. 2018. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B* 80, 3 (2018), 531–550.

[164] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly robust joint learning for recommendation on data missing not at random. In *Proceedings of the International Conference on Machine Learning*. PMLR, 6638–6647.

[165] Christopher John Cornish Hellaby Watkins. 1989. Learning from delayed rewards. (1989).

[166] Jeffrey M. Wooldridge. 2005. Unobserved heterogeneity and estimation of average partial effects. *Identiffcation and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* (2005), 27–55.

[167] Jeffrey M. Wooldridge. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

[168] Jeffrey M. Wooldridge. 2015. Control function methods in applied econometrics. *Journal of Human Resources* 50, 2 (2015), 420–445.

[169] Jeffrey M. Wooldridge. 2015. *Introductory Econometrics: A Modern Approach*. Cengage Learning.

[170] Philip Green Wright. 1928. *Tariff on Animal and Vegetable Oils*. Macmillan Company, New York.

[171] Anpeng Wu, Kun Kuang, Bo Li, and Fei Wu. 2022. Instrumental variable regression with confounder balancing. In *Proceedings of the International Conference on Machine Learning*. PMLR, 24056–24075.

[172] Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Minqin Zhu, Yuxuan Liu, Bo Li, Furui Liu, Zhihua Wang, and Fei Wu. 2023. Learning instrumental variable from data fusion for treatment effect estimation. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 9 (2023), 10324–10332.

[173] Anpeng Wu, Junkun Yuan, Kun Kuang, Bo Li, Runze Wu, Qiang Zhu, Yueting Zhuang, and Fei Wu. 2023. Learning decomposed representations for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering* 35, 5 (2023), 4989–5001. DOI: https://doi.org/10.1109/TKDE.2022.3150807

[174] Pengzhou Abel Wu and Kenji Fukumizu. 2022. beta-Intact-VAE: Identifying and estimating causal effects under limited overlap. In *Proceedings of the International Conference on Learning Representations.*

[175] Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. 2021. Learning deep features in instrumental variable regression. In *Proceedings of the International Conference on Learning Representations.*

[176] Xin Xu, Han-gen He, and Dewen Hu. 2002. Efffcient reinforcement learning using recursive least-squares methods. *Journal of Artiffcial Intelligence Research* 16, 1 (2002), 259–292.

[177] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data* 15, 5 (2021), 1–46.

[178] Junkun Yuan, Xu Ma, Kun Kuang, Ruoxuan Xiong, MingmingGong, and Lanfen Lin. 2021. Learning domain-invariant relationship with instrumental variable for domain generalization. *ACM Transactions on Knowledge Discovery from Data* 17, 8 (2023), 1–21.

[179] Junkun Yuan, Anpeng Wu, Kun Kuang, Bo Li, Runze Wu, Fei Wu, and Lanfen Lin. 2022. Auto IV: Counterfactual prediction via automatic instrumental variable decomposition. *ACM Transactions on Knowledge Discovery from Data* 16, 4 (2022), 1–20.

[180] Benito Zander and Maciej Liśkiewicz. 2016. On searching for generalized instrumental variables. In *Proceedings of the Artificial Intelligence and Statistics*. PMLR, 1214–1222.

[181] Jingwen Zhang, Yifang Chen, and Amandeep Singh. 2022. Causal bandits: Online decision-making in endogenous settings. *NeurIPS 2022 Workshop CDS*. (2022).

[182] Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. 2023. Instrumental variable regression via kernel maximum moment loss. *Journal of Causal Inference* 11, 1 (2023), 1–42.

[183] Weijia Zhang, Lin Liu, and Jiuyong Li. 2021. Treatment effect estimation with disentangled latent factors. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 12 (2021), 10923–10930.

[184] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. 2021. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5372–5382.

[185] Hao Zou, Kun Kuang, Boqi Chen, Peixuan Chen, and Peng Cui. 2019. Focused context balancing for robust offline policy evaluation. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 696–704.

Received 4 September 2023; revised 11 December 2024; accepted 29 April 2025