

Testability of Instrumental Variables in Additive Nonlinear, Non-Constant Effects Models

Xichen Guo*

GUOXICHEN0@GMAIL.COM

*Department of Applied Statistics, Beijing Technology and Business University
Beijing, 102488, China*

Zheng Li*

ZHENGLI0060@GMAIL.COM

*Department of Applied Statistics, Beijing Technology and Business University
Beijing, 102488, China*

Biwei Huang

BIH007@UCSD.EDU

*Halicioglu Data Science Institute (HDSI), University of California San Diego
La Jolla, San Diego, California, 92093, USA*

Yan Zeng

YANAZENG013@BTBU.EDU.CN

*Department of Applied Statistics, Beijing Technology and Business University
Beijing, 102488, China*

Zhi Geng

ZHIGENG@PKU.EDU.CN

*Department of Applied Statistics, Beijing Technology and Business University
Beijing, 102488, China
School of Mathematical Sciences, Peking University
Beijing, 100871, China*

Feng Xie[†]

FENGXIE@BTBU.EDU.CN

*Department of Applied Statistics, Beijing Technology and Business University
Beijing, 102488, China*

Editor: Ilya Shpitser

Abstract

We address the issue of the testability of instrumental variables derived from observational data. Most existing testable implications are centered on scenarios where the treatment is a discrete variable, e.g., instrumental inequality (Pearl, 1995), or where the effect is assumed to be constant, e.g., instrumental variables condition based on the principle of independent mechanisms (Buraue, 2023). However, treatments can often be continuous variables, such as drug dosages or nutritional content levels, and non-constant effects may occur in many real-world scenarios. In this paper, we consider an additive nonlinear, non-constant effects model with unmeasured confounders, in which treatments can be either discrete or continuous, and propose an Auxiliary-based Independence Test (AIT) condition to test whether a variable is a valid instrument. We first show that, under the completeness condition, if the candidate instrument is valid, then the AIT condition holds. Moreover, we illustrate the implications of the AIT condition and demonstrate that, under certain additional conditions, the AIT condition is necessary and sufficient to detect all invalid IVs. We also extend the AIT condition to include covariates and introduce a practical testing

*. Equal contribution

†. Corresponding author

algorithm. Experimental results on both synthetic and three different real-world datasets show the effectiveness of our proposed condition.

Keywords: instrumental variable, testability, unmeasured confounders, non-constant effects, causal graphical models

1. Introduction

Estimating causal effects from observational data is a fundamental task in understanding the underlying relationships between variables. The instrumental variables (IV) model is a well-established method for estimating the causal effect of a treatment (exposure) X on an outcome Y in the presence of unmeasured confounders and has been used in a range of fields, such as economics (Imbens, 2014; Imbens and Rubin, 2015), sociology (Pearl, 2009; Spirtes et al., 2000), epidemiology (Hernán and Robins, 2006; Baiocchi et al., 2014), and artificial intelligence (Chen et al., 2022; Wu et al., 2022). Generally speaking, given a causal relationship $X \rightarrow Y$, the valid IV Z is required to satisfy the following three conditions: $C1$. Z is related to the treatment (*relevance*), $C2$. Z is independent of the unmeasured confounders that affect the treatment and outcome (*exogeneity*), and $C3$. Z has no direct path to the outcome (*exclusion restriction*). Figure 1 illustrates the graphical criteria of the IV model, where Z is a valid IV relative to $X \rightarrow Y$ in the subgraph (a).

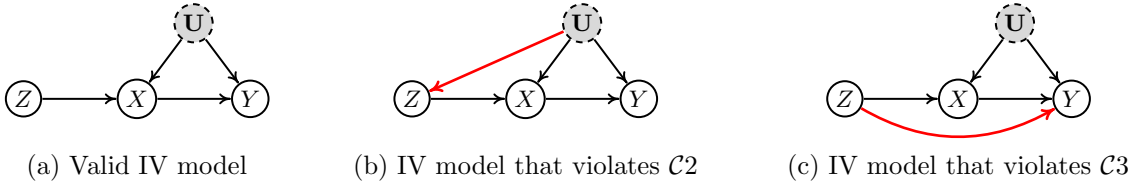


Figure 1: Graphical illustration of IV models, where $\mathbf{U}$ is the set of unmeasured confounders.

(a) Z is a valid IV. (b) Z is an invalid IV due to the edge $\mathbf{U} \rightarrow Z$ (Violate $C2$).

(c) Z is an invalid IV due to the edge $Z \rightarrow Y$ (Violate $C3$).

Due to the presence of unmeasured confounders $\mathbf{U}$, determining which variable serves as a valid IV is not always straightforward based solely on observational data, and often requires domain knowledge. A classic test for the IV model is the Durbin-Wu-Hausman test (Nakamura and Nakamura, 1981). Given a subset of valid IVs, it can identify whether other potential candidates are also valid IVs. However, it does not guide how to find the initial set of valid IVs. Meanwhile, given an invalid IV, it may not consistently identify the correct causal effect (Bound et al., 1995; Chu et al., 2001). Thus, it is vital to develop statistical methods for selecting IVs solely from observational data.

It is not feasible to ascertain the validity of IVs solely based on the joint distribution of observed variables, without incorporating additional assumptions (Pearl, 2009). Pearl (1995) introduced a seminal necessary criterion known as the *instrumental inequality*, which acts as a critical test for identifying potential IVs in models featuring discrete variables. Building on this groundwork, subsequent research by Manski (2003); Palmer et al. (2011); Kitagawa (2015); Wang et al. (2017) broadened the scope, exploring the applicability and limitations of IV validity tests across diverse scenarios. A notable advancement was made by Kédagni

and Mourifié (2020), who formulated a more encompassing set of criteria, the *generalized instrumental inequalities*. These criteria cater to scenarios with discrete treatment variables, removing constraints of the type of data on the IV and outcome variables and offering a robust framework to challenge the exogeneity condition. Intuitively, these methods mentioned above use the idea that if the IV is independent of the unmeasured confounders and the *exclusion restriction* (C3) holds, then changes in the IV should not have a significant impact on the outcome variable without altering the treatment variable, because the treatment variable mediates the influence of the instrumental variable on the outcome variable. However, these methods fail to work when treatment is a continuous variable. In reality, one may often be concerned about the causal effect of the continuous treatment on the outcome; see Skaaby et al. (2013); Martinussen et al. (2019).

Several contributions have been made to address continuous treatment settings under certain assumptions. In an additive linear, constant effects (ALICE) model, Kang et al. (2016); Bowden et al. (2016), and Windmeijer et al. (2019) have shown that if we assume more than half of the variables are valid IVs in the potential IVs (known as the majority rule), one may identify the valid IVs solely from observed data. Later, Hartwig et al. (2017); Guo et al. (2018); Windmeijer et al. (2021) relaxed the majority rule and assumed that the number of valid IVs is larger than any number of invalid IVs with the same ratio estimator limit (known as the plurality rule). They demonstrated that it is still possible to identify valid IVs under the plurality rule. Another interesting work by Silva and Shimizu (2017) proposed the IV-TETRAD algorithm, which uses the so-called Trek conditions (Sullivant et al., 2010; Spirtes, 2013) for selecting a valid IV set. This method requires at least two or more valid IVs in the system. However, although these methods have been used in a range of fields, they may fail to test whether a single IV is valid.

Recently, Xie et al. (2022) have demonstrated that a single IV imposes specific constraints within the linear non-Gaussian acyclic causal model. However, their method assumes that all noise terms are non-Gaussian and that the effects remain constant. More recently, Buraue (2023) have introduced a novel validity condition for Instrumental Variables based on the Principle of Independent Mechanisms, termed IV-PIM, within the linear IV framework. This condition is particularly notable as it applies to both continuous and discrete treatment variables. Nevertheless, its applicability is constrained by the presence of covariates, making it unsuitable when no covariates are available. Additionally, the condition is limited to scenarios with constant treatment effects.

Pearl (1995) conjectured that the validity of an instrument cannot be tested when dealing with continuous treatment variables without additional assumptions, a theory recently confirmed by Gunsilius (2021). Unlike the existing work that focuses on the parametric linear constant effects model, we consider IV validity in a more challenging additive non-parametric model, the Additive Nonlinear, Non-Constant Effects (ANINCE) Model. Rather surprisingly, although a single IV is in general not fully testable within the ANINCE model, we will show that a single variable Z , being a valid IV relative to $X \rightarrow Y$, imposes specific constraints in the ANINCE model under the completeness condition (i.e., for all functions $\psi(X)$ with finite expectation, $\mathbb{E}[\psi(X)|Z] = 0$ implies $\psi(X) = 0$).¹ Specifically, we make the following contributions:

1. See Section 2.2 for more detailed discussion of the completeness condition. For related context, we refer interested readers to Newey and Powell (2003); D'Haultfoeuille (2011); Hu and Shiu (2018).

1. We introduce a necessary condition, termed the Auxiliary-based Independence Test (AIT) condition, for detecting a single variable that cannot serve as an IV within the ANINCE model, under the completeness condition (Assumption 1). This condition is applicable to scenarios involving non-constant causal effects and both discrete and continuous treatment.
2. We provide the necessary and sufficient conditions for detecting all invalid IVs using the AIT condition under the ANINCE model. Specifically, we show that, under the partial non-Gaussianity assumption (Assumption 2), all observable violations of the IV exogeneity condition can be identified in the linear, constant effects model. Additionally, under the completeness condition (Assumption 1) and the distributional non-degeneracy condition assumption (Assumption 3), we can detect invalid IVs resulting from violations of either exogeneity or the exclusion restriction in the ANINCE model. We also present two notable types of non-identifiable invalid IVs (see Proposition 5), along with intuitive explanations for each.
3. We present a practical implementation of the AIT condition test that accounts for the presence of covariates with finite data. We establish its asymptotic validity, including control of Type I error and consistency (power tending to one) against alternatives.
4. We demonstrate the efficacy and applicability of the proposed approach on both synthetic and three real-world datasets with different scenarios.

The rest of this paper is organized as follows. In Section 2, we introduce notations, the additive non-parametric IV model, and the ANINCE model. In Section 3, we formulate the AIT condition for the single IV. We show the AIT condition is a necessary condition for IV validity in the ANINCE model under the completeness condition. We discuss the implications of AIT condition in the linear, constant effects model and the nonlinear, non-constant effects model, respectively. We show that, under additional assumptions, the AIT condition is a necessary and sufficient condition for IV validity. In Section 4, we address the practical scenario with covariates and provide the *AIT Condition* algorithm for implementing the test. We further provide a theoretical analysis establishing the asymptotic level and power of the AIT test. In Section 5, we present the efficacy and applicability of our method on both synthetic and three real-world datasets which contain continuous and discrete data. Conclusions are given in Section 6.

2. Preliminaries

2.1 Notations

This work is conducted within the framework of causal graphical models as elaborated by Pearl (2009) and Spirtes et al. (2000). Specifically, we represent causal relationships using the directed acyclic graph (DAG), denoted as $\mathcal{G}$, where nodes represent variables and directed edges (arrows) indicate causal links between those variables. Sets of variables are represented in bold, and individual variables and symbols for graphs are in italics. We use “instrumental variable (IV)” and “instrument” interchangeably. The main symbols used in this paper are summarized in Table 1.

Table 1: List of main symbols used in this paper

Symbol	Description
$\mathcal{G}$	A directed acyclic graph
X	Treatment (exposure)
Y	Outcome
Z	A candidate (potential) instrument
$\mathbf{U}$	The latent (unmeasured) confounders
$\mathbf{W}$	Covariates
$\mathcal{Z}$	The residual of Z after regressing on covariates $\mathbf{W}$
IV	Instrumental Variable
$A \perp\!\!\!\perp B C$	A is statistically independent of B given C
$A \not\perp\!\!\!\perp B C$	A is statistically dependent on B given C
$\{X, Y Z\}$	The candidate instrumental variable Z used to assess the independence of the auxiliary variable $\mathcal{A}_{X \rightarrow Y Z} = Y - h(X)$
$ \mathbf{W} $	The number of variables in set $\mathbf{W}$
$f(X, Z)$	The causal effect of X and Z on Y
$\tilde{f}_{bias}(X, Z)$	The bias between estimated causal effect of X on Y and ground-truth causal effect of (X, Z) on Y
ε_*	The noise term of a variable
$\varphi_*(\mathbf{U})$	The effect of the latent variables $\mathbf{U}$ on the observed variables
$g_*(Z)$	The effect of the instrument variable Z on other observed variables
$\mathbb{R}$	The field of real numbers
$\mathbb{R} \rightarrow \mathbb{R}$	A mapping from the real numbers to the real numbers
$\mathcal{I}(\cdot)$	The indicator function
$\mathbb{E}(X)$	The expected value of random variable X
$\frac{\partial^2 Y}{\partial X \partial Z}$	The second-order partial derivative of Y with respect to X and Z
$\mathcal{A}_{X \rightarrow Y Z}$	The auxiliary variable of causal relationship $X \rightarrow Y$ relative to Z . We often drop the subscript $X \rightarrow Y Z$ when there is no ambiguity (i.e., $\mathcal{A}$)
$\hat{h}(X, \mathbf{W})$	The empirical estimate of the function $h(X, \mathbf{W})$
$\hat{\mathcal{A}}_{X \rightarrow Y Z}$	The estimated auxiliary variable computed as $Y - \hat{h}(X, \mathbf{W})$. We often use $\hat{\mathcal{A}}$ as a shorthand when there is no ambiguity
k, l	The kernels $k : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}$ and $l : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$ used for variables in $\mathcal{U}$ and $\mathcal{V}$, respectively
K-test method	Kitagawa’s method from Kitagawa (2015)
IV-PIM method	Buraue’s method from Buraue (2023)

2.2 Additive Nonparametric Instrumental Variable Model

The instrumental variable approach offers a strategy for inferring the causal effect of interest in the presence of unmeasured confounders (Bowden and Turkington, 1990; Angrist et al., 1996; Pearl, 2009; Imbens and Rubin, 2015). Given a causal relationship $X \rightarrow Y$, a valid IV Z is required to satisfy the following three conditions:

- C1. (Relevance).** Z has a direct effect on the treatment X ;
- C2. (Exogeneity or Randomness).** Z is independent of the unmeasured confounders $\mathbf{U}$;
- C3. (Exclusion Restriction).** Z does not directly affect the outcome Y .

Definition 1 *A random variable Z is a valid IV for the causal relationship $X \rightarrow Y$ if the above three conditions $\mathcal{C}1 \sim \mathcal{C}3$ are satisfied.*

We here consider the additive non-parametric IV model presented in Newey and Powell (2003), which, for a valid IV Z , can be expressed as follows: ²

$$\begin{aligned} X &= g(Z) + \underbrace{\varphi_X(\mathbf{U})}_{\delta} + \varepsilon_X, \\ Y &= f(X) + \underbrace{\varphi_Y(\mathbf{U})}_{\epsilon} + \varepsilon_Y, \end{aligned} \tag{1}$$

where $\mathbb{E}[\varphi_Y(\mathbf{U}) + \varepsilon_Y | Z] = 0$, and the noise terms ε_X and ε_Y are statistically independent.

We here would like to mention that in nonparametric IV models, identifying the function $f(\cdot)$ amounts to solving the conditional moment restriction $\mathbb{E}[Y - f(X) | Z] = 0$, which can be formulated as a linear inverse problem of the form $Tf = q$, where $Tf := \mathbb{E}[f(X) | Z]$ and $q := \mathbb{E}[Y | Z]$. Under mild regularity conditions on the joint distribution of (X, Z) , the associated conditional expectation operator T is compact. As a consequence, solving $Tf = q$ constitutes an ill-posed inverse problem, and the existence of a solution is not guaranteed for arbitrary q . Classical results from inverse problem theory show that existence is characterized by a Picard criterion (Kress, 1989, 2013). ³ As noted by Darolles et al. (2011), restrictions that confine the function to a compact set are mathematically equivalent to enforcing the Picard criterion and thereby ensuring the existence of a solution to the associated inverse problem. Following common practice in the nonparametric IV literature (e.g., Newey and Powell 2003; Carrasco et al. 2007; Florens et al. 2011), we restrict the true function $f(\cdot)$ to belong to a compact set of functions, which ensures the existence of a solution to the conditional moment restriction $\mathbb{E}[Y - f(X) | Z] = 0$, that is, q lies in the range of the operator T . This condition is assumed to hold throughout the paper.

Assumption 1 (Completeness Condition) (Newey and Powell, 2003; D’Haultfoeulle, 2011; Hu and Shiu, 2018)) Given a valid IV Z , if for all measurable functions $\psi(X)$ such that $\mathbb{E}[|\psi(X)|] < +\infty$,

$$\mathbb{E}[\psi(X) | Z] = 0 \text{ almost surely} \Rightarrow \psi(X) = 0 \text{ almost surely.} \tag{2}$$

Intuitively, for a valid IV Z , completeness imposes a condition on the conditional probability density $k(X|Z)$ in the nonparametric IV model. Note that identifying the nonparametric IV model is generally nontrivial due to the challenges inherent in nonparametric settings. Consequently, the completeness condition has been extensively studied to establish identification in various nonparametric and semiparametric models, such as Newey and Powell (2003); Ai and Chen (2003); Hall and Horowitz (2005); Chen and Hu (2006); Blundell et al. (2007); Chernozhukov et al. (2007); Carrasco et al. (2007); Hu and Schennach (2008); Carroll

-
2. Here, we have slightly modified the model from Newey and Powell (2003) to explicitly represent the unmeasured confounders for subsequent analysis.
 3. The Picard criterion provides necessary and sufficient conditions for the existence of a solution to $\mathbb{E}[Y - f(X) | Z] = 0$. In particular, when the associated conditional expectation operator T is compact, existence can be characterized in terms of the operator T and the function q ; see Theorem 15.18 in Kress (1989, 2013) and the discussion in Horowitz (2012).

et al. (2010); D’Haultfoeuille (2011); Darolles et al. (2011); An and Hu (2012); Newey (2013); Canay et al. (2013); Shiu and Hu (2013); Fève and Florens (2014); Andrews (2017); Hu and Shiu (2018). For instance, Newey and Powell (2003) have shown that, given a valid IV Z , the causal effect $f(\cdot)$ of interest in the model specified by Equation (1) can be uniquely identified if the above completeness condition—i.e., Assumption 1—holds.

Remark 1 *Regarding the completeness condition, three key aspects are worth highlighting:*

1. *The testability of the completeness condition has attracted attention in recent literature. Canay et al. (2013) have shown that the completeness condition is generally untestable without additional restrictions. Therefore, for more specific models, testable conditions have been proposed. For example, Freyberger (2017) provided a test for restricted completeness by linking the outcome of the test to consistency of an estimator. Hu and Shiu (2022) provided a useful result for testing the completeness condition in a class of models based on convolution. Other tests include full-rank tests for completeness in discrete settings, as proposed by Robin and Smith (2000).*
2. *Newey and Powell (2003) have shown that the case of finite support and the exponential family setting constitute sufficient conditions for completeness. Many commonly used distributions—such as Gaussian, Poisson, Binomial, and certain multivariate forms of these—fall within the exponential family framework (Hu and Shiu, 2018). Building on this, other sufficient conditions for completeness have been further developed in the literature (D’Haultfoeuille, 2011; Chen et al., 2014; Andrews, 2017; Hu and Shiu, 2018). Notably, D’Haultfoeuille (2011) derived sufficient conditions for various forms of completeness of the endogenous variable X given the instrument Z , and applied these results to nonparametric IV regression. Furthermore, Hu and Shiu (2018) provided sufficient conditions for completeness of the distribution of treatment conditional on the instrument, without relying on a specific functional form.*
3. *Under the completeness condition, estimation methods for nonparametric IV models have also been extensively studied (Newey and Powell, 2003; Ai and Chen, 2003; Darolles et al., 2011; Chernozhukov et al., 2007; Newey, 2013; Singh et al., 2019; Bennett et al., 2019). Among them, Newey and Powell (2003) developed a nonparametric equivalent to the two-stage least squares estimator: they used linear-in-parameter series expansions of $\mathbb{E}[Y | Z]$ and $\mathbb{E}[g(X) | Z]$ in a generalized method of moments framework. Darolles et al. (2011) analyzed identification and overidentification of the nonparametric IV model, and proposed an estimator based on Tikhonov regularization to address the ill-posed inverse problem inherent in nonparametric IV estimation. A comprehensive review of several estimation techniques has been provided in Carrasco et al. (2007).*

2.3 Additive Nonlinear, Non-Constant Effects Model

Without loss of generality, we assume that all variables have a zero mean (otherwise can be centered) and that no covariates are present for simplicity. In Section 4, we address the practical scenario where covariates are included. In this paper, we focus our attention on the

Additive Nonlinear, Non-Constant Effects (ANINCE) Model. Specifically, the generation process satisfies the following structural causal model:

$$\begin{aligned} X &= g(Z) + \varphi_X(\mathbf{U}) + \varepsilon_X, \\ Y &= f(X, Z) + \varphi_Y(\mathbf{U}) + \varepsilon_Y, \end{aligned} \tag{3}$$

where $f(\cdot)$ denotes the true, unknown causal effect of interest, and $g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$, $f(\cdot) : \mathbb{R}^2 \rightarrow \mathbb{R}$, and $\varphi_*(\cdot) : \mathbb{R}^{|\mathbf{U}|} \rightarrow \mathbb{R}$ are smooth functions. The noise terms ε_X , ε_Y , ε_Z and $\varepsilon_{\mathbf{U}}$ are mutually independent. Note that Z and $\mathbf{U}$ may be dependent, which indicates that the *exogeneity* condition (C2) is violated, and the non-zero $f(\cdot, Z)$ function indicates that Z directly affects the outcome Y , implying that the *exclusion restriction* condition (C3) is violated.

A special case of the ANINCE model is the additive linear, constant effects model (ALICE), where functions $g(\cdot)$, $f(\cdot)$, and $\varphi_*(\cdot)$ are linear functions, which have been extensively studied in works such as those by Bowden et al. (2015); Kang et al. (2016); Silva and Shimizu (2017); Windmeijer et al. (2021). Compared to these works, we investigate the testability of IV in a more challenging scenario, where $g(\cdot)$, $f(\cdot)$, and $\varphi_*(\cdot)$ may be non-linear functions. Additionally, we focus on the testability of a single valid IV, whereas previous works have focused on the testability of a set of IVs (assuming that including at least two or more valid IVs among the candidate variables).

Remark 2 *Regarding the structural causal model (SCM)-based IV framework and its difference from the classical econometric IV framework, two key aspects are worth highlighting:*

1. *The ANINCE model is formulated within the SCM framework and therefore assumes mutually independent noise components. This SCM-based independence structure is essential for deriving additional distributional constraints that render IV validity testable; without it, valid instruments would not imply observable independence relations and no empirical test could be constructed. Moreover, the usefulness of independent-noise assumptions is well established in additive SCMs for causal discovery, where they enable identifiability of causal directions and have been validated across numerous applications (Hoyer et al., 2008; Peters et al., 2014; Peters and Bühlmann, 2014; Bühlmann et al., 2014; Peters et al., 2017; Glymour et al., 2019). Consequently, our adoption of independent-noise terms follows standard practice in SCM-based causal analysis rather than departing from conventional modeling assumptions.*
2. *In contrast to classical econometric IV frameworks, such as additive nonparametric IV models (Newey and Powell, 2003), which rely on the moment restriction $\mathbb{E}[\varepsilon | Z] = \mathbb{E}[\varphi_Y(\mathbf{U}) + \varepsilon_Y | Z] = 0$, the ANINCE model imposes mutual independence among the noise terms. This SCM-based formulation offers a key advantage: the independent-noise assumptions provide additional structure from which testable implications, such as the AIT condition (see Definition 2), can be derived for assessing IV validity. The trade-off is that these SCM assumptions encode structural restrictions beyond the moment restrictions commonly used in classical econometric IV models. Consequently, our results apply to IV settings that admit this SCM-based independent-noise representation, rather than to all settings satisfying only the classical moment conditions.*

Our Goal. The goal of this paper is to determine, from the observed dataset $\{X, Y, Z\}$ satisfying an ANINCE model, whether Z is related to X (i.e., *relevance* condition), Z is exogenous relative to (X, Y) (i.e., *exogeneity* condition) and Z does not directly affect outcome Y (i.e., *exclusion restriction* condition). Note that the first condition *relevance*, can be easily checked by the independence test because Z and X are observed variables. Therefore, we focus on the last two conditions of IV Z . In summary, we aim to provide a new necessary condition to detect whether a variable is a valid IV and investigate the necessary and sufficient conditions under which all invalid IVs can be detected.

Remark 3 *Existing approaches have attempted to detect violations of the exogeneity for a single IV in settings with discrete variables, such as through instrumental inequality and its extensions. In contrast to prior work, we focus on continuous-variable settings. Notably, instrument validity is generally untestable in continuous variable settings without additional assumptions. Hence, in this paper, we introduce additive function constraints, allowing us to detect the invalid IVs when the treatment variable is continuous.*

3. AIT Condition and Its Implications in ANINCE Models

In this section, we first formulate the Auxiliary-based Independence Test condition (AIT condition) and show that it is a necessary condition for evaluating IV validity under the completeness condition (Assumption 1). We further present theoretical results regarding the implications of the AIT condition in the linear, constant effects model and nonlinear, non-constant effects model.

3.1 AIT Condition

Below, we give the AIT condition, which defines the independent relationship between the “Auxiliary variable” and candidate IV. Note that the concept of “auxiliary variable” has been developed to address different tasks (Drton and Richardson, 2004; Chen et al., 2017; Cai et al., 2019), but our formalization is different from theirs (see Equation (4)). To the best of our knowledge, it has not been realized that the independence property involving such an auxiliary variable reflects the validity of the IV in the ANINCE model.

Definition 2 (AIT Condition) *Suppose the treatment X , the outcome Y , and a candidate IV Z are nodes in a causal graph $\mathcal{G}$. Define the auxiliary variable of the causal relationship $X \rightarrow Y$ relative to Z as*

$$\mathcal{A}_{X \rightarrow Y||Z} := Y - h(X), \quad (4)$$

where $h(\cdot)$ satisfies $\mathbb{E}[\mathcal{A}_{X \rightarrow Y||Z} | Z] = 0$ and $h(\cdot) \neq 0$. We say that $\{X, Y | Z\}$ follows the AIT condition if and only if $\mathcal{A}_{X \rightarrow Y||Z}$ is independent from Z .

The AIT condition states that if there exists a function $h(\cdot)$ such that $\mathbb{E}[Y - h(X) | Z] = 0$, then $\{X, Y | Z\}$ satisfies the AIT condition if and only if the auxiliary variable $\mathcal{A}_{X \rightarrow Y||Z}$ is independent of Z . If no non-zero function $h(\cdot)$ exists that satisfies the conditional moment restriction—i.e., $\mathbb{E}[Y - h(X) | Z] \neq 0$ for all $h(\cdot)$ —then $Y - h(X)$ is always correlated with Z (recall that all variables are assumed to be mean-centered). This implies that $Y - h(X)$ is

dependent on Z , and hence the AIT condition does not hold. It is noteworthy that under the completeness condition, there exists a unique function $h(\cdot)$ satisfying $\mathbb{E}[Y - h(X) | Z] = 0$ in the additive nonparametric IV model (Equation (1)); this function coincides with the true causal function $f(\cdot)$ of X on Y . For the sake of conciseness, we often drop the subscript $X \rightarrow Y || Z$ from $\mathcal{A}_{X \rightarrow Y || Z}$ when there is no ambiguity. The following theorem shows the testability of an IV in light of the AIT condition in an ANINCE model.

Theorem 1 (Necessary Condition for IV) *Let X , Y , and Z be the treatment, outcome, and candidate IV in an ANINCE model, respectively. Suppose that X , Y , and Z are correlated and that Assumption 1 holds. If Z is a valid IV relative to $X \rightarrow Y$, then $\{X, Y | Z\}$ always satisfies the AIT condition.*

Proof If Z is a valid IV relative to $X \rightarrow Y$, then the ANINCE model can be rephrased as additive nonparametric IV models:

$$X = g(Z) + \varphi_X(\mathbf{U}) + \varepsilon_X, \quad Y = f(X) + \varphi_Y(\mathbf{U}) + \varepsilon_Y, \quad (5)$$

where $\mathbf{U} = \varepsilon_U$, $Z = \varepsilon_Z$, and $\mathbb{E}[\varphi_Y(\mathbf{U}) + \varepsilon_Y | Z] = 0$ due to IV validity.

Taking conditional expectations of Y with respect to Z yields

$$\mathbb{E}[Y | Z] = \mathbb{E}[f(X) + \varphi_Y(\mathbf{U}) + \varepsilon_Y | Z] = \mathbb{E}[f(X) | Z] = \int f(X) k(X | Z) dX, \quad (6)$$

where $k(X | Z)$ is the conditional probability distribution function of X given Z . Thus, $f(\cdot)$ is a solution to the integral equation (6).

Now let $h(\cdot)$ be any function that satisfies the conditional moment restriction

$$E[Y - h(X) | Z] = 0 \iff E[h(X) | Z] = E[Y | Z]. \quad (7)$$

Since $f(\cdot)$ also satisfies $E[Y - f(X) | Z] = 0$, both $f(\cdot)$ and $h(\cdot)$ solve the same integral equation. Subtracting the two displays yields

$$\mathbb{E}[h(X) - f(X) | Z] = \mathbb{E}[h(X) | Z] - \mathbb{E}[f(X) | Z] = \mathbb{E}[Y | Z] - \mathbb{E}[Y | Z] = 0. \quad (8)$$

Define $\psi(X) := h(X) - f(X)$. By the completeness condition (Assumption 1), the condition $\mathbb{E}[\psi(X) | Z] = 0$ implies $\psi(X) = 0$ almost surely, and thus $h(X) = f(X)$ almost surely. Therefore, the function $h(\cdot)$ that solves $\mathbb{E}[Y - h(X) | Z] = 0$ is uniquely identified and coincides with the true causal effect function $f(\cdot)$ of X on Y . This is the standard completeness-based identification argument in the additive non-parametric IV literature (see, e.g., Newey and Powell 2003; Newey 2013, Singh et al. 2019, and Bennett et al. 2019).

Thus, we have the auxiliary variable

$$\mathcal{A}_{X \rightarrow Y || Z} = Y - h(X) = Y - f(X) = \varphi_Y(\mathbf{U}) + \varepsilon_Y. \quad (9)$$

By Theorem 2.2.5 and its extension both in Meester (2008), if random variables are mutually independent, then any measurable functions applied to disjoint subsets of them yield independent random variables (see Theorem 5 and Corollary 3 in Appendix A for further details). Based on this result, we next show that the auxiliary variable $\mathcal{A}_{X \rightarrow Y || Z}$ and Z are

statistically independent. Specifically, since the noise terms ε_Z , ε_Y , and ε_U are mutually independent, we can obtain that ε_Z is also independent of $\varphi_Y(\varepsilon_U) + \varepsilon_Y$. Furthermore, combining the equations $U = \varepsilon_U$ and $Z = \varepsilon_Z$, we conclude that Z is independent of $\varphi_Y(U) + \varepsilon_Y$. Therefore, the auxiliary variable $\mathcal{A}_{X \rightarrow Y|Z}$ and Z are statistically independent, i.e., $\mathcal{A}_{X \rightarrow Y|Z} \perp\!\!\!\perp Z$. This implies that $\{X, Y|Z\}$ satisfies the AIT condition. ■

Theorem 1 means that if $\{X, Y|Z\}$ violates the AIT condition, then Z is an invalid IV relative to $X \rightarrow Y$. Otherwise, Z may or may not be valid.

3.2 Implications of AIT Condition in Additive Linear, Constant Effects Models

In this section, we focus our attention on a special type of ANINCE model, the linear, constant effects model, which has been widely studied (Bowden et al., 2015; Kang et al., 2016; Silva and Shimizu, 2017; Windmeijer et al., 2021). Specifically, we assume that the underlying causal structure of the system can be represented by a DAG $\mathcal{G}$, and that the data-generating process follows a linear structural equation model associated with $\mathcal{G}$. In particular, each variable V_i satisfies $V_i = \sum_{V_j \in pa(V_i)} \alpha_{ij} V_j + \varepsilon_{V_i}$, $i = 1, 2, \dots, r$, where the noise terms $\varepsilon_{V_1}, \dots, \varepsilon_{V_r}$ are independent of each other, and α_{ij} is the direct effect of $V_j \rightarrow V_i$. Hence, the ANINCE model in Equation (3) can be expressed as follows:

$$X = \tau Z + \boldsymbol{\rho}^T \mathbf{U} + \varepsilon_X, \quad Y = \beta X + \nu Z + \boldsymbol{\kappa}^T \mathbf{U} + \varepsilon_Y, \quad (10)$$

where $Z = \boldsymbol{\gamma}^T \mathbf{U} + \varepsilon_Z$. When $\boldsymbol{\gamma}^T = \mathbf{0}$ (satisfying the *exogeneity* condition) and $\nu = 0$ (satisfying the *exclusion restriction* condition), Z qualifies as a valid IV relative to $X \rightarrow Y$. Below, we show the implications of the AIT condition in this model.

Motivating Examples: Firstly, we illustrate with two simple examples that while a valid IV does not impose any restrictions on the joint marginal distribution of the observed variables within the linear Gaussian model, it does impose certain constraints in the linear partial non-Gaussian model, which can be identified using AIT condition. Consider the causal graph in Figure 1(b), where Z is an invalid IV for $X \rightarrow Y$, as it violates the exogeneity condition. Let $\mathcal{N}(0, 1)$ denote the standard normal distribution, and $\exp(0.5)$ denote the exponential distribution with a parameter of 0.5. Suppose the generating mechanisms of these models are as follows:

- *Linear Gaussian model.* $U = \varepsilon_U$, $Z = 2U + \varepsilon_Z$, $X = 1.5Z + 0.8U + \varepsilon_X$, $Y = X + 3.5U + \varepsilon_Y$, and $\varepsilon_U, \varepsilon_Z, \varepsilon_X, \varepsilon_Y \sim \mathcal{N}(0, 1)$.
- *Linear partial non-Gaussian model.* $U = \varepsilon_U$, $Z = 2U + \varepsilon_Z$, $X = 1.5Z + 0.8U + \varepsilon_X$, $Y = X + 3.5U + \varepsilon_Y$, $\varepsilon_U \sim \exp(0.5)$, and $\varepsilon_Z, \varepsilon_X, \varepsilon_Y \sim \mathcal{N}(0, 1)$.

The difference between the above two models lies in the noise term ε_U ; the first follows a Gaussian distribution, while the second follows an exponential distribution (non-Gaussian distribution). Figure 2 shows the scatter plots of $\mathcal{A}_{X \rightarrow Y|Z}$ versus the invalid IV Z for two models. Interestingly, we find that, in the linear Gaussian model, $\mathcal{A}_{X \rightarrow Y|Z}$ and Z are statistically independent (satisfying AIT condition), while in the linear partial non-Gaussian

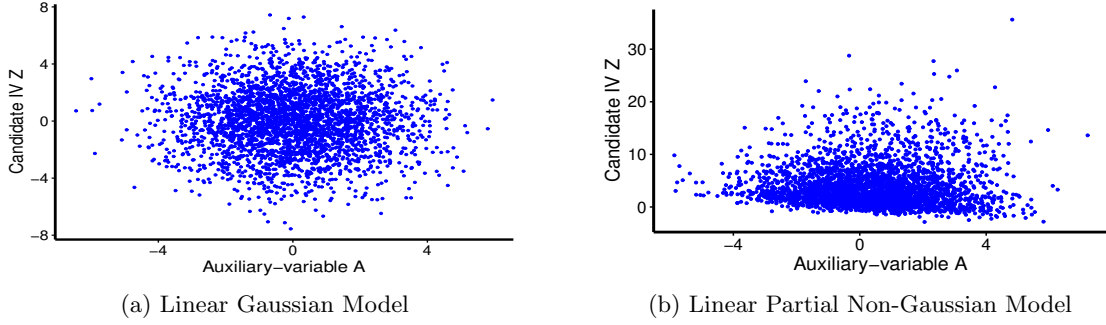


Figure 2: Scatter plots of Candidate IV Z and Auxiliary-variable $\mathcal{A}$ under the linear models. (a) All noise terms follow Gaussian distributions. (b) Some noise terms follow non-Gaussian distributions.

model, $\mathcal{A}_{X \rightarrow Y|Z}$ and Z are statistically dependent (violating AIT condition). These results suggest that non-Gaussianity is beneficial for identifying the invalid IV.

The following Propositions 1 and 2 formalize the phenomena discussed above.

Proposition 1 (Non-Testability in Linear Gaussian Models) *Let X , Y , and Z be the treatment, outcome, and candidate IV in a linear model (Equation (10)), respectively. Suppose that X , Y , and Z are correlated. If all noise terms of variables follow Gaussian distributions, then regardless of whether Z is a valid IV relative to $X \rightarrow Y$ or not, $\{X, Y|Z\}$ always satisfies the AIT condition.*

Proof The proof of Proposition 1 is straightforward. Let Z represent any candidate IV, which may or may not be valid. In the linear Gaussian model, by the definition of the AIT condition, there exists a function $h(X) = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} \cdot X$ such that $\mathbb{E}[Y - h(X) | Z] = 0$. Hence, $\mathbb{E}[\mathcal{A}_{X \rightarrow Y|Z} | Z] = \mathbb{E}[Y - \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} X | Z] = 0$, which implies that $\text{Cov}(\mathcal{A}_{X \rightarrow Y|Z}, Z) = 0$. Since zero correlation implies independence in the linear Gaussian model (Bain and Engelhardt, 1992), we conclude that $\mathcal{A}_{X \rightarrow Y|Z}$ is independent of the candidate IV Z . Consequently, $\{X, Y|Z\}$ always satisfies the AIT condition. $\blacksquare$

Proposition 1 states that checking the AIT condition in a linear Gaussian causal model (second-order statistics) does not provide any useful information for identifying invalid IVs. Below, we show that one can leverage higher-order statistics of noise terms to identify certain types of invalid IVs that violate *exogeneity* condition.⁴ Before presenting the result, we give the key assumption.

Assumption 2 (Partial Non-Gaussianity) *At least one of the following conditions holds: (i) there exists at least one variable $U_i \in \mathbf{U}$ whose noise term follows a non-Gaussian distribution and cause Z ; (ii) the noise term of Z follows a non-Gaussian distribution.*

4. Higher-order statistics mean beyond the second-order moments in statistics of the data, such as skewness and kurtosis.

Assumption 2 states the non-Gaussianity of data, which is expected to be widespread, as suggested by Cramér Decomposition Theorem (Cramér, 1962). Considerable works have already been built on this assumption (Shimizu et al., 2006; Salehkaleybar et al., 2020). For additional references, see Spirtes and Zhang (2016); Shimizu (2022).

We now show that the AIT condition can access the validity of *exogeneity* condition in linear models under Assumption 2.

Proposition 2 (Testability of Exogeneity in Linear Models) *Let X , Y , and Z be the treatment, outcome, and candidate IV in a linear model (Equation (10)), respectively. Suppose that X , Y , and Z are correlated, and that Assumption 2 holds. If Z violates the exogeneity condition, i.e., at least one variable $U_i \in \mathbf{U}$ causes Z , then $\{X, Y||Z\}$ violates the AIT condition.*

Proof Roughly speaking, by the Darmois–Skitovich theorem (Darmois, 1953; Skitovitch, 1953), if $\mathcal{A}_{X \rightarrow Y||Z}$ shares any common non-Gaussian noise terms ε_{U_i} or ε_Z with Z , $\mathcal{A}_{X \rightarrow Y||Z}$ is statistically dependent on Z . This implies that $\{X, Y||Z\}$ violates the AIT condition. See Appendix A.1 for its complete proof. ■

Remark 4 *When all noise terms follow the non-Gaussian distributions, the linear partial non-Gaussian model becomes the well-known Linear Non-Gaussian Acyclic Model (LiNGAM), which has been extensively studied (Shimizu et al., 2006; Salehkaleybar et al., 2020). Consequently, according to Proposition 2, an invalid IV that violates the exogeneity condition in LiNGAM can be detected in light of the AIT condition.*

Proposition 2 states that we can detect an invalid IV that violates *exogeneity* condition using the AIT condition based on the observational data in the linear model under Assumption 2. A natural question that arises is whether we can detect IVs that violate the *exclusion restriction* condition within the same framework. Unfortunately, in practice, we cannot detect an invalid IV that solely violates the *exclusion restriction* condition in the linear model, as shown in the following example.

Example 1 *Let’s consider the causal structure illustrated in Figure 1 (a), where Z is a valid IV. We assume the following relationship: $Z = \varepsilon_Z$, $X = \tau Z + \boldsymbol{\rho}^T \mathbf{U} + \varepsilon_X$, $Y = \beta X + \boldsymbol{\kappa}^T \mathbf{U} + \varepsilon_Y$. Next, we demonstrate how to construct another causal structure as shown in Figure 1 (c), where Z becomes an invalid IV (violating solely the exclusion restriction condition). Specifically, let $\beta' = \beta - \frac{\nu}{\tau}$, $Z' = Z$, $X' = X$, and $Y' = \beta' X' + \nu Z' + \boldsymbol{\kappa}^T \mathbf{U} + \varepsilon_Y + \frac{\nu}{\tau}(\boldsymbol{\rho}^T \mathbf{U} + \varepsilon_X) = Y$. Thus, (Z, X, Y) has the same distribution as (Z', X', Y') . This implies that a variable being an instrument imposes no constraints on the joint marginal distribution of the observed variables. The same result is also discussed in Section 3 of Chu et al. (2001).*

The following proposition states the above phenomenon in the linear model.

Proposition 3 (Non-Testability of Exclusion Restriction in Linear Models) *Let X , Y , and Z be the treatment, outcome, and candidate IV in a linear model (Equation (10)), respectively. Suppose that X , Y , and Z are correlated. If Z satisfies the exogeneity condition, regardless of whether Z violates the exclusion restriction condition or not, then $\{X, Y||Z\}$ always satisfies the AIT condition.*

Proof This proof is straightforward. The auxiliary variable $\mathcal{A}_{X \rightarrow Y||Z}$ shares no common noise terms with candidate IV Z , whether the noise terms are Gaussian or non-Gaussian. By the Darmois-Skitovich theorem (Darmois, 1953; Skitovitch, 1953), $\mathcal{A}_{X \rightarrow Y||Z}$ is statistically independent from Z . This implies that $\{X, Y||Z\}$ always satisfies the AIT condition. See Appendix A.2 for its complete proof. ■

Based on Theorem 1 and Propositions 1 $\sim$ 3, we derive the following theorem, which provides a necessary and sufficient condition for detecting invalid IVs that violate the *exogeneity* condition within a linear causal model.

Theorem 2 (Necessary and Sufficient Conditions in Linear Models) *Let X, Y , and Z be the treatment, outcome, and candidate IV in a linear model (Equation (10)), respectively. Suppose that X, Y , and Z are correlated and that Assumption 2 holds. $\{X, Y||Z\}$ violates the AIT condition if and only if the candidate IV Z is invalid due to a violation of the *exogeneity* condition.*

Proof See Appendix A.3 for its proof. ■

This theorem states that the AIT condition is necessary and sufficient to detect the candidate IV violations of the *exogeneity* condition when Assumption 2 holds.

3.3 Implications of AIT Condition in Additive Nonlinear, Non-Constant Effects Models

In this section, we investigate the implications of the AIT condition on the ANINCE model. Before giving our main results, we first show a simple example to show that nonlinearity is beneficial in identifying the invalid IV.

A Motivating Example: Continue to consider the causal graph in Figure 1 (b), where Z serves as an invalid IV for the causal relationship $X \rightarrow Y$, violating the *exogeneity* condition. Here, we modify the generation mechanism of the linear Gaussian model by introducing a nonlinear function between U and Z , specifically as follows:

- *Partial Non-linear Gaussian model.* $U = \varepsilon_U$, $Z = \exp(U) + \varepsilon_Z$, $X = 1.5Z + 0.8U + \varepsilon_X$, $Y = X + 3.5U + \varepsilon_Y$, and $\varepsilon_U, \varepsilon_Z, \varepsilon_X, \varepsilon_Y \sim \mathcal{N}(0, 1)$.

Figure 3 presents the scatter plots of $\mathcal{A}_{X \rightarrow Y||Z}$ versus the candidate IV Z in the partial non-linear Gaussian model. Compared to the linear Gaussian model, we transformed the functional relationship between U and Z from a linear function ($2U$) to an exponential function ($\exp(U)$). Interestingly, in the partial non-linear Gaussian model, $\mathcal{A}_{X \rightarrow Y||Z}$ and Z are statistically dependent (violating AIT condition). Note that Proposition 1 shows that the AIT condition is always satisfied in the linear Gaussian model. These findings suggest that nonlinearity is beneficial in assessing the validity of *exogeneity* condition.

We now investigate the conditions under which the invalid IV can be detected in terms of AIT condition. It is noteworthy that the ANINCE model (Equation (3)) is flexible as functions g , f , φ_X , and φ_Y might be any unknown functions. Consequently, without imposing further parametric assumptions, it is impossible to determine the explicit forms of

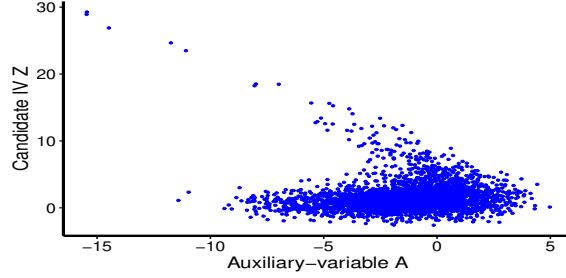


Figure 3: Scatter plot of Candidate IV Z and Auxiliary-variable $\mathcal{A}$ when all noise terms follow Gaussian distribution in the partially non-linear invalid IV model.

the estimated f and $\mathcal{A}_{X \rightarrow Y \| Z}$. Hence, let $h(\cdot)$ be a function satisfying $\mathbb{E}[Y - h(X)|Z] = 0$ and $h(\cdot) \neq 0$. According to the definition of the AIT condition, the auxiliary variable $\mathcal{A}_{X \rightarrow Y \| Z}$ is given by:

$$\mathcal{A}_{X \rightarrow Y \| Z} = Y - h(X) = \underbrace{f(X, Z) - h(X)}_{\tilde{f}_{bias}(X, Z)} + \varphi_Y(\mathbf{U}) + \varepsilon_Y, \quad (11)$$

where $\tilde{f}_{bias}(X, Z) = f(X, Z) - h(X)$. It is important to note that for a valid instrumental variable Z , $\tilde{f}_{bias}(X, Z) = 0$.

Below, we give the key assumption regarding the second-order partial derivative in the nonlinear model.

Assumption 3 (Distributional Non-degeneracy Condition) *Assume that the joint density $p(\mathcal{A}_{X \rightarrow Y \| Z}, Z)$ is twice continuously differentiable and satisfies*

$$\frac{\partial^2}{\partial \mathcal{A}_{X \rightarrow Y \| Z} \partial Z} \log p(\mathcal{A}_{X \rightarrow Y \| Z}, Z) \neq 0 \quad (12)$$

on a set with non-zero Lebesgue measure.

Assumption 3 is a natural condition that one expects to hold for detecting all invalid IVs in the ANINCE model. Intuitively, if the auxiliary variable $\mathcal{A}_{X \rightarrow Y \| Z}$ is independent of the instrument Z (i.e., AIT holds), then the joint probability density of $(\mathcal{A}_{X \rightarrow Y \| Z}, Z)$ factorizes into the product of the marginal densities: $p(\mathcal{A}_{X \rightarrow Y \| Z}, Z) = p(\mathcal{A}_{X \rightarrow Y \| Z}) \cdot p(Z)$. According to Lin (1997), for a set of independent random variables whose joint density is twice differentiable, the Hessian matrix of the logarithm of their joint density is diagonal everywhere (see Theorem 7 in Appendix A for further details). Taking the logarithm of $p(\mathcal{A}_{X \rightarrow Y \| Z}, Z)$ and computing its second-order mixed partial derivative yields: $\frac{\partial^2 \log p(\mathcal{A}_{X \rightarrow Y \| Z}, Z)}{\partial \mathcal{A}_{X \rightarrow Y \| Z} \partial Z} = 0$. In contrast, if AIT does not hold—i.e., $\mathcal{A}_{X \rightarrow Y \| Z}$ and Z are dependent—then $p(\mathcal{A}_{X \rightarrow Y \| Z}, Z) \neq p(\mathcal{A}_{X \rightarrow Y \| Z}) \cdot p(Z)$, which implies that $\frac{\partial^2 \log p(\mathcal{A}_{X \rightarrow Y \| Z}, Z)}{\partial \mathcal{A}_{X \rightarrow Y \| Z} \partial Z} \neq 0$ (Equation (12)).

Below, we provide two nontrivial analytic examples that satisfy the distributional non-degeneracy condition.

Example 2 (Violation of the Exogeneity Condition) Consider the causal graph in Figure 1 (b), where Z violates exogeneity. Let the structural model be $U = \varepsilon_U$, $Z = \gamma U + \varepsilon_Z$, $X = \exp(Z) + \rho U + \varepsilon_X$, and $Y = \beta X + \kappa U + \varepsilon_Y$, where $\gamma \neq 0$ and $(\varepsilon_U, \varepsilon_Z, \varepsilon_X, \varepsilon_Y) \stackrel{\text{ind}}{\sim} \mathcal{N}(0, 1)$. According to the definition of auxiliary variables, we obtain $\mathcal{A}_{X \rightarrow Y|Z} = (\beta - \hat{\beta}) \exp(Z) + ((\beta - \hat{\beta})\rho + \kappa)U + (\beta - \hat{\beta})\varepsilon_X + \varepsilon_Y$, where $\hat{\beta} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} = \beta + \frac{\kappa\gamma}{\sigma_Z^2 \exp(\sigma_Z^2/2) + \rho\gamma}$ and $\sigma_Z^2 = \gamma^2 + 1$. Conditioned on $Z = z$, the auxiliary variable $\mathcal{A}_{X \rightarrow Y|Z}$ is a linear combination of Gaussian variables $\varepsilon_U, \varepsilon_X$, and ε_Y , implying $\mathcal{A} | Z = z \sim \mathcal{N}(m(z), v)$, where $m(z) = (\beta - \hat{\beta}) \exp(z) + ((\beta - \hat{\beta})\rho + \kappa) \frac{\gamma}{\sigma_Z^2} z$ and $v = ((\beta - \hat{\beta})\rho + \kappa)^2 \frac{1}{\sigma_Z^2} + (\beta - \hat{\beta})^2 + 1$. Using the factorization $p(\mathcal{A}, Z) = p(\mathcal{A} | Z)p(Z)$, we take the derivative of the joint density of $(\mathcal{A}, Z)$ and obtain:

$$\frac{\partial^2 \log p(\mathcal{A}, Z)}{\partial \mathcal{A} \partial Z} = \frac{\partial^2 \log p(\mathcal{A} | Z)}{\partial \mathcal{A} \partial Z} = \frac{m'(Z)}{v} = \frac{\kappa\gamma - \beta_{\text{bias}}[\rho\gamma + \exp(Z)(\gamma^2 + 1)]}{(\kappa - \beta_{\text{bias}}\rho)^2 + (\gamma^2 + 1)(\beta_{\text{bias}}^2 + 1)},$$

where $\beta_{\text{bias}} = \frac{\kappa\gamma}{\sigma_Z^2 \exp(\sigma_Z^2/2) + \rho\gamma}$. Since the numerator $\kappa\gamma - \beta_{\text{bias}}[\rho\gamma + \exp(Z)(\gamma^2 + 1)]$ is not identically zero, the mixed derivative is nonzero on a set with non-zero Lebesgue measure, and therefore Assumption 3 holds. A detailed derivation is provided in Appendix B.1.

Example 3 (Violation of the Exclusion Restriction Condition) Consider the causal graph in Figure 1 (c), where Z violates the exclusion restriction. Let the structural model be $U = \varepsilon_U$, $Z = \varepsilon_Z$, $X = \exp(Z) + \rho U + \varepsilon_X$, and $Y = \beta X + \nu Z + \kappa U + \varepsilon_Y$, where $\nu \neq 0$ and $(\varepsilon_U, \varepsilon_Z, \varepsilon_X, \varepsilon_Y) \stackrel{\text{ind}}{\sim} \mathcal{N}(0, 1)$. According to the definition of auxiliary variables, we obtain $\mathcal{A}_{X \rightarrow Y|Z} = (\beta - \hat{\beta}) \exp(Z) + \nu Z + ((\beta - \hat{\beta})\rho + \kappa)U + (\beta - \hat{\beta})\varepsilon_X + \varepsilon_Y$, where $\hat{\beta} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} = \beta + \nu \exp(-1/2)$. Conditioned on $Z = z$, the auxiliary variable $\mathcal{A}_{X \rightarrow Y|Z}$ is a linear combination of Gaussian variables U, ε_X , and ε_Y , implying $\mathcal{A} | Z = z \sim \mathcal{N}(m(z), v)$, where $m(z) = (\beta - \hat{\beta}) \exp(z) + \nu z$, and $v = ((\beta - \hat{\beta})\rho + \kappa)^2 + (\beta - \hat{\beta})^2 + 1$. Using the factorization $p(\mathcal{A}, Z) = p(\mathcal{A} | Z)p(Z)$, we take the derivative of the joint density of $(\mathcal{A}, Z)$ and obtain:

$$\frac{\partial^2 \log p(\mathcal{A}, Z)}{\partial \mathcal{A} \partial Z} = \frac{\partial^2 \log p(\mathcal{A} | Z)}{\partial \mathcal{A} \partial Z} = \frac{m'(Z)}{v} = \frac{\nu[1 - \exp(Z - 1/2)]}{[\kappa - \nu \exp(-1/2)]^2 + \nu^2 \exp(-1) + 1}.$$

Since the numerator $\nu[1 - \exp(Z - 1/2)]$ is not identically zero, the mixed derivative is nonzero on a set with non-zero Lebesgue measure, and therefore Assumption 3 holds. A detailed derivation is provided in Appendix B.2.

Proposition 4 (Testability of IV in ANINCE Models) Let X, Y , and Z be the treatment, outcome, and candidate IV in an ANINCE model, respectively. Suppose that X, Y , and Z are correlated and that Assumptions 1 and 3 hold. If the candidate IV Z is invalid, then $\{X, Y | Z\}$ violates the AIT condition.

Proof See Appendix A.4 for its proof. ■

The proposition above shows that, under Assumptions 1 and 3, the AIT condition can be used to detect invalid IVs. Although it is not obvious whether Assumption 3 (the distributional non-degeneracy condition) holds in general, some solutions under which the Assumption 3 does not hold are worth reporting, as shown in the following proposition.

Proposition 5 *Let X , Y , and Z be the treatment, outcome, and candidate IV in an ANINCE model, respectively. Assumption 3 does not hold if any of the following conditions is satisfied:*

- (i) *the candidate IV Z solely violates the exclusion restriction condition and the direct causal effect of $Z \rightarrow Y$ is a linear function of the direct causal effect from $Z \rightarrow X$ in the ANINCE model, i.e., $g_Y(Z) = a \cdot g_X(Z) + b$, where a is the non-zero constant;*
- (ii) *all noise terms associated with the variables follow Gaussian distributions within the linear causal model.*

Proof See Appendix A.5 for its proof. ■

Condition (i) implies that the AIT condition is always satisfied when the candidate IV Z satisfies the *exogeneity* condition and the relationship of effect $g_Y(Z) = a \cdot g_X(Z) + b$ holds in the ANINCE model. In other words, if the direct causal effect of $Z \rightarrow Y$ is not a linear function of the direct causal effect of $Z \rightarrow X$, we can identify invalid IVs that solely violate the exclusion restriction condition using the AIT condition. Intuitively, the distribution of the invalid IV model under condition (i), in which candidate IV Z solely violates the exclusion restriction condition and satisfies the specific linear relationship, can be transformed into the distribution of the valid IV model, as illustrated in Example 4. Note that the constraint $g_Y(Z) = a \cdot g_X(Z) + b$ naturally arises in linear causal models. Thus, condition (i) further implies that when the candidate IV Z solely violates the exclusion restriction condition in the linear causal model, Assumption 3 does not hold, which is consistent with the result of Proposition 3. In addition, condition (ii) shows that the AIT condition is always satisfied in the linear Gaussian model, which aligns with the conclusion of Proposition 1.

Example 4 *Continue to consider the causal graph in Figure 1 (c), where Z is an invalid IV relative to $X \rightarrow Y$. The generating mechanism is as follows:*

$$Z = \varepsilon_Z, \quad X = g_X(Z) + \varphi_X(\mathbf{U}) + \varepsilon_X, \quad Y = f(X) + g_Y(Z) + \varphi_Y(\mathbf{U}) + \varepsilon_Y, \quad (13)$$

where $g_Y(Z) = a \cdot g_X(Z) + b$. We now construct another model based on the causal graph shown in Figure 1 (a), where Z is a valid IV relative to $X \rightarrow Y$. Let $Z' = Z$, $X' = X$, and $f'(X') = f(X') + a \cdot g_X(Z') + b$. Furthermore, Y' is expressed as $Y' = f'(X') + \varphi_Y(\mathbf{U}) + \varepsilon_Y = f(X) + g_Y(Z) + \varphi_Y(\mathbf{U}) + \varepsilon_Y = Y$. Hence, we conclude that the distribution of (Z, X, Y) and (Z', X', Y') are identical. Because we can solely observe the variables (Z, X, Y) , it is impossible to determine from the distribution of (Z, X, Y) whether the data come from Figure 1 (a) or Figure 1 (c). In other words, we cannot ascertain whether Z is a valid IV or not.

According to Proposition 5, not all invalid IVs can be identified solely from the joint distribution of observational data. Below, based on Theorem 1 and Proposition 4, we introduce the necessary and sufficient conditions for invalid IV in the additive nonlinear, non-constant effects model.

Theorem 3 (Necessary and Sufficient Conditions for IV in ANINCE Models) *Let X , Y , and Z be the treatment, outcome, and candidate IV in an ANINCE model (Equation (3)), respectively. Suppose that X , Y , and Z are correlated, and that Assumption 1 holds.*

- *If Z is a valid IV relative to $X \rightarrow Y$, then $\{X, Y|Z\}$ always satisfies the AIT condition.*
- *If Z is an invalid IV relative to $X \rightarrow Y$ and Assumption 3 holds, then $\{X, Y|Z\}$ always violates the AIT condition.*

Proof See Appendix A.6 for its proof. ■

Theorem 3 outlines two scenarios where a candidate IV would be considered invalid: either the IV doesn't meet the *exogeneity* condition or it violates the *exclusion restriction* condition within a nonlinear model. Figure 4 presents how the AIT condition relates to the validity of IVs under different assumptions. The completeness condition (Assumption 1) alone guarantees necessity, while adding a distributional non-degeneracy condition (Assumption 3) ensures sufficiency.

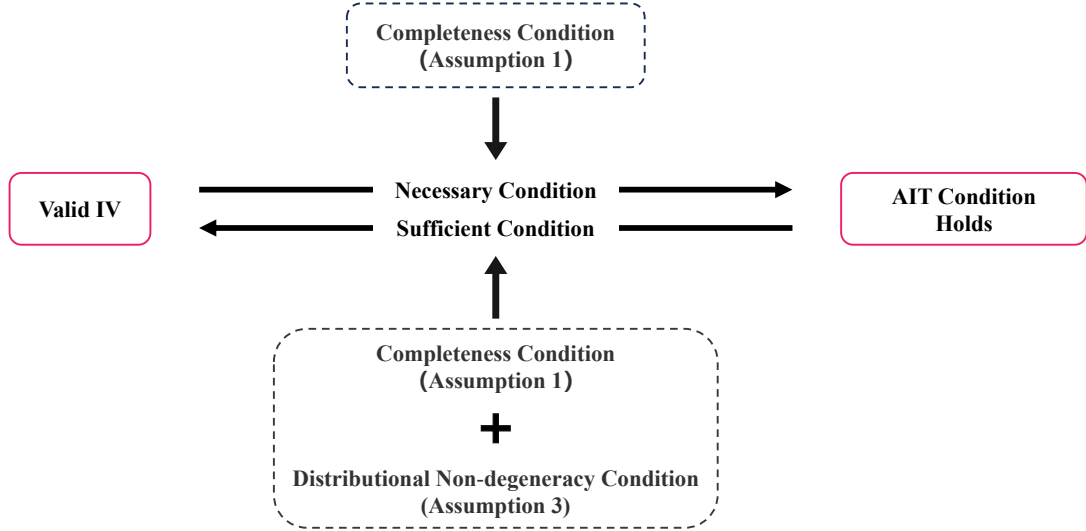


Figure 4: Flowchart showing how assumptions lead to the AIT condition being necessary or sufficient for IV validity.

4. Practical Implementation of AIT Condition

In this section, we discuss the practical implementation of the AIT condition. In Section 4.1, we first address how to implement the AIT condition when covariates are present. In Section 4.2, we then propose a sample splitting procedure for implementing the AIT condition with finite samples. Finally, in Section 4.3, we establish the asymptotic validity of the AIT test.

4.1 AIT Condition with Covariates

In practice, there are scenarios where covariates $\mathbf{W}$ are present. For instance, age may influence how the treatment method affects patient recovery speed. Specifically, we focus on the generation of the ANINCE model with covariates as follows:

$$\begin{aligned} X &= g(\mathbf{W}, Z) + \varphi_X(\mathbf{U}) + \varepsilon_X, \\ Y &= f(X, \mathbf{W}, Z) + \varphi_Y(\mathbf{U}) + \varepsilon_Y, \end{aligned} \quad (14)$$

where $g(\cdot) : \mathbb{R}^{|\mathbf{W}|+1} \rightarrow \mathbb{R}$, $f(\cdot) : \mathbb{R}^{|\mathbf{W}|+2} \rightarrow \mathbb{R}$, and $\varphi_*(\cdot) : \mathbb{R}^{|\mathbf{U}|} \rightarrow \mathbb{R}$ are smooth functions. The noise terms ε_X , ε_Y , ε_Z , and $\varepsilon_{\mathbf{U}}$ are statistically independent. Below, we first extend the AIT condition from Definition 2 to account for the presence of covariates, as stated in the following definition.

Definition 3 (AIT Condition with Covariates) *Suppose the treatment X , the outcome Y , the covariates $\mathbf{W}$, and a candidate IV Z are nodes in a causal graph $\mathcal{G}$. Define the auxiliary variable of the causal relationship $X \rightarrow Y$ relative to $(Z, \mathbf{W})$ as*

$$\mathcal{A}_{X \rightarrow Y|(Z, \mathbf{W})} := Y - h(X, \mathbf{W}), \quad (15)$$

where $h(\cdot)$ satisfies $\mathbb{E}[\mathcal{A}_{X \rightarrow Y|(Z, \mathbf{W})} | Z, \mathbf{W}] = 0$ and $h(\cdot) \neq 0$. Define the residual of Z after regressing on $\mathbf{W}$ as:

$$\mathcal{Z} := Z - \mathbb{E}[Z | \mathbf{W}]. \quad (16)$$

We say that $\{X, Y | (Z, \mathbf{W})\}$ follows the AIT condition if and only if $\mathcal{A}_{X \rightarrow Y|(Z, \mathbf{W})}$ is independent from $\mathcal{Z}$.

Remark 5 *It is worth noting that covariates $\mathbf{W}$ can provide additional constraints that can be exploited in the IV validity test. Specifically, under the ANINCE model, incorporating covariates generalizes the AIT condition to the independence $\mathcal{A}_{X \rightarrow Y|(Z, \mathbf{W})} \perp\!\!\!\perp (Z, \mathbf{W})$, meaning that the auxiliary variable must be independent of the joint vector $(Z, \mathbf{W})$. However, directly testing independence against a high-dimensional vector is statistically challenging: nonparametric tests typically suffer from low power in high dimensions due to the curse of dimensionality and concentration of measure (Zhang et al., 2011; Ramdas et al., 2015). To mitigate this issue, we use a statistically more tractable surrogate condition: $\mathcal{A}_{X \rightarrow Y|(Z, \mathbf{W})} \perp\!\!\!\perp \mathcal{Z}$, where $\mathcal{Z} := Z - \mathbb{E}[Z | \mathbf{W}]$, which tests independence between the auxiliary variable and the residual part of Z that is conditionally orthogonal to $\mathbf{W}$. This preserves the intended inferential target under the ANINCE model while improving statistical power.*

Based on Definition 3 and Theorem 1, we obtain the necessary condition for IV in the presence of covariates $\mathbf{W}$, as described in the following corollary.

Corollary 1 (Necessary Condition for IV with Covariates) *Let X , Y , $\mathbf{W}$, and Z be the treatment, outcome, covariates, and candidate IV in an ANINCE model, respectively. Suppose that X , Y , $\mathbf{W}$, and Z are correlated and that Assumption 1 holds. If Z is a valid IV relative to $X \rightarrow Y$ given $\mathbf{W}$, then $\{X, Y | (Z, \mathbf{W})\}$ always satisfies the AIT condition.*

Proof See Appendix A.7 for its proof. ■

Corollary 1 means that if $\{X, Y | (Z, \mathbf{W})\}$ violates the AIT condition, then Z is an invalid IV relative to $X \rightarrow Y$ given $\mathbf{W}$. Otherwise, Z may or may not be valid.

4.2 AIT Condition with Finite Data

Below, we provide the practical implementation of the AIT condition with finite data. For a candidate IV Z , under the ANINCE model, we need to test the following hypothesis:

$$H_0 : Z \text{ is a valid IV, } H_1 : Z \text{ is an invalid IV.} \quad (17)$$

Since the AIT test is implemented as a plug-in procedure involving estimated functions, including the function $\hat{h}(X, \mathbf{W})$ and, when covariates are present, the regression function $\hat{\pi}(\mathbf{W})$ of $\mathbb{E}[Z | \mathbf{W}]$, we employ sample splitting to separate estimation and testing (Doran et al., 2014; Chernozhukov et al., 2018; Maia Polo et al., 2023; Ren et al., 2025; Scheidegger et al., 2025). Specifically, we randomly split the observed dataset $\mathcal{D}$ with $|\mathcal{D}| = N$ into two disjoint subsets, $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$, where $|\mathcal{D}_1| = n = \lfloor \rho N \rfloor$, $\rho \in (0, 1)$ is the splitting ratio, and $|\mathcal{D}_2| = m = N - n$. The functions $\hat{h}(\cdot)$ and $\hat{\pi}(\cdot)$ (if covariates are present) are estimated using dataset $\mathcal{D}_1$, and the statistical independence test between $\hat{\mathcal{A}}_{X \rightarrow Y \parallel (Z, \mathbf{W})} = Y - \hat{h}(X, \mathbf{W})$ and $\hat{\mathcal{Z}} = Z - \hat{\pi}(\mathbf{W})$ is then implemented using dataset $\mathcal{D}_2$.

Next, we describe how to address the two main issues above, including estimating the functions $\hat{h}(\cdot)$ and $\hat{\pi}(\cdot)$ on $\mathcal{D}_1$, followed by conducting the independence test on $\mathcal{D}_2$.

Issue 1: Estimating the Functions on $\mathcal{D}_1$. Regarding the estimation of the function $\hat{h}(\cdot)$, the intuitive idea is to apply a standard IV estimator to obtain $\hat{h}(X, \mathbf{W})$. Note that solving a nonparametric IV problem is, in general, highly nontrivial and may require regularization to address issues such as slow convergence rates, as thoroughly discussed in Darolles et al. (2011). Newey and Powell (2003) show that, given a valid IV and under the completeness condition, the estimated function $\hat{h}(X, \mathbf{W})$ converges to the true function $f(X, \mathbf{W})$ in the population (infinite-sample) limit for exponential families or for discrete data with finite support. Conversely, if the IV is invalid, the population-limit estimator $\hat{h}(X, \mathbf{W})$ is biased relative to the true model. There are now many IV estimators for Additive Non-Parametric IV Models (Guo and Small, 2016; Singh et al., 2019; Bennett et al., 2019). Rather than proposing a new estimation strategy for nonparametric IV models, our paper focuses on developing a testable condition for identifying valid instruments using existing IV estimation methods, under an oracle scenario where the estimator is assumed to have access to the true structural function. Hence, we here adopt the control function IV estimator proposed by Guo and Small (2016), which is a two-stage approach. In the first stage, it regresses the treatment variable X on covariates $\mathbf{W}$ and instruments $(Z, \tau_2(Z), \dots, \tau_i(Z))$ (a known vector of linearly independent functions of Z), obtaining the predicted value $\hat{X}$ and the residual $e_1 = X - \hat{X}$. In the second stage, the outcome Y is regressed on the treatment X , covariates $\mathbf{W}$, and the residual e_1 from the first stage regression. The coefficients of the second stage regression are taken as the control function estimates. Note that nonlinear functions can be estimated using basis functions. In our experiments, we apply polynomial basis functions as a predefined set of linearly independent functions, leveraging this control function method. Regarding the regression function $\hat{\pi}(\cdot)$, when the covariates $\mathbf{W}$ are present, we use the random forest regression method (Pedregosa et al., 2011; Breiman, 2001) to fit the function $\hat{\pi}(\mathbf{W})$.

Issue 2: Testing Independence on $\mathcal{D}_2$. As a first step, we compute the estimated auxiliary variable $\hat{\mathcal{A}}_{X \rightarrow Y \parallel (Z, \mathbf{W})} = Y - \hat{h}(X, \mathbf{W})$ and the residualized instrument $\hat{\mathcal{Z}} = Z - \hat{\pi}(\mathbf{W})$ using the dataset $\mathcal{D}_2$, where the functions $\hat{h}(\cdot)$ and $\hat{\pi}(\cdot)$ were estimated on $\mathcal{D}_1$ in

the previous Issue 1. To check whether $\hat{\mathcal{A}}_{X \rightarrow Y \parallel (Z, \mathbf{W})}$ and $\hat{\mathcal{Z}}$ are statistically independent, we employ the Large-Scale HSIC Test, a Hilbert-Schmidt Independence Criterion (HSIC)-based test proposed by Zhang et al. (2018) for independence testing. If the output p_{value} is less than the preset significance level α , we reject the null hypothesis H_0 , indicating that the candidate IV Z is invalid. Conversely, if we fail to reject the null hypothesis, it suggests that Z is a valid IV.

Based on the above discussions, the complete *AIT Condition* test procedure is given in Algorithm 1. To allow for greater flexibility in the algorithm and considering that prior knowledge, such as a constant causal effect, may sometimes be available, we include a two-stage least squares estimator suitable for linear effects (Basmann, 1957; Henckel et al., 2024) to estimate the causal effect.

Algorithm 1 AIT Condition

Input: Observed dataset $\mathcal{D} = \{X, Y, \mathbf{W}, Z\}$, where X is the treatment, Y is the outcome, $\mathbf{W}$ denotes covariates, and Z is a candidate IV; block size B ; data split ratio ρ ; significance level α

- 1: Initialize: Result $\leftarrow$ Do not reject H_0
- 2: **Sample Splitting.** Randomly split the dataset $\mathcal{D}$ with $|\mathcal{D}| = N$ into two disjoint subsets, $\mathcal{D}_1$ and $\mathcal{D}_2$, where $|\mathcal{D}_1| = n = \lfloor \rho N \rfloor$, and $|\mathcal{D}_2| = m = N - n$.
 - Step 1: Estimate functions $h(\cdot)$ and $\pi(\cdot)$ on $\mathcal{D}_1$**
 - 3: **if** the constant-effect assumption is adopted **then**
 - 4: $\hat{h}(\cdot) \leftarrow$ Two-Stage Least Squares Estimator $(X, Y, \mathbf{W}, Z)$ on $\mathcal{D}_1$
 - 5: **else**
 - 6: $\hat{h}(\cdot) \leftarrow$ Control Function IV Estimator $(X, Y, \mathbf{W}, Z)$ on $\mathcal{D}_1$
 - 7: **end if**
 - 8: **if** $\mathbf{W} \neq \emptyset$ **then**
 - 9: Fit the regression function $\hat{\pi}(\cdot)$ using the random forest regression method on $\mathcal{D}_1$
 - 10: **end if**
 - Step 2: Test the AIT Condition on $\mathcal{D}_2$**
 - 11: Compute the auxiliary variable $\hat{\mathcal{A}}_{X \rightarrow Y \parallel (Z, \mathbf{W})} \leftarrow Y - \hat{h}(X, \mathbf{W})$ on $\mathcal{D}_2$
 - 12: **if** $\mathbf{W} \neq \emptyset$ **then**
 - 13: Compute the residualized instrument $\hat{\mathcal{Z}} \leftarrow Z - \hat{\pi}(\mathbf{W})$ on $\mathcal{D}_2$
 - 14: **else**
 - 15: Set $\hat{\mathcal{Z}} \leftarrow Z$ on $\mathcal{D}_2$
 - 16: **end if**
 - 17: $p_{value} \leftarrow$ Large-Scale HSIC Test($\hat{\mathcal{A}}_{X \rightarrow Y \parallel (Z, \mathbf{W})}, \hat{\mathcal{Z}}, B$)
 - 18: **if** $p_{value} < \alpha$ **then**
 - 19: Result $\leftarrow$ Reject H_0
 - 20: **else**
 - 21: **return** Result
 - 22: **end if**

Output: Result

4.3 Theoretical Analysis of Type I and Type II Errors in the AIT Test

In this section, we investigate the asymptotic properties of the AIT test, focusing on both Type I and Type II errors. Despite being a plug-in procedure, the AIT test achieves asymptotic control of the Type I error at the nominal level under the null hypothesis and appropriate regularity conditions. At the same time, the AIT test exhibits desirable power properties against fixed alternatives.

In the implementation of the algorithm, we adopt the block-based HSIC statistic introduced in Zhang et al. (2018). We first briefly review the relevant theoretical results on the Hilbert–Schmidt independence criterion for testing independence between two random variables; see Zhang et al. (2018); Gretton et al. (2005, 2008). Let U and V be random variables taking values in domains $\mathcal{U}$ and $\mathcal{V}$, respectively. Given kernels $k : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}$, $l : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$, and the independently and identically distributed samples $S_{i=1}^m = \{u_i, v_i\}_{i=1}^m$, the samples are split into $\frac{m}{B}$ disjoint blocks, each of size B , denoted by $\{\{u_i^{(b)}, v_i^{(b)}\}_{i=1}^B\}_{b=1}^{m/B}$. For each block $b \in \{1, \dots, \frac{m}{B}\}$, the unbiased HSIC statistic is

$$\hat{\eta}_b := \frac{1}{B(B-3)} \left[\text{tr} \left(\tilde{K}^{(b)} \tilde{L}^{(b)} \right) + \frac{\mathbb{1}^T \tilde{K}^{(b)} \mathbb{1} \mathbb{1}^T \tilde{L}^{(b)} \mathbb{1}}{(B-1)(B-2)} - \frac{2}{B-2} \mathbb{1}^T \tilde{K}^{(b)} \tilde{L}^{(b)} \mathbb{1} \right], \quad (18)$$

where $\tilde{K} = K - \text{diag}(K)$, with $K_{i,j} := k(u_i, u_j)$, i.e., the kernel matrix with diagonal elements set to zero, and similarly for $\tilde{L}$. Here, $\mathbb{1}$ denotes a vector of ones of relevant dimension. Throughout, we assume that the block size satisfies $B \geq 4$; when the sample size m is not divisible by B , the number of blocks is taken to be $\lfloor m/B \rfloor$, so that the estimator is well defined.

The block-based HSIC estimator is then given by $\widehat{\text{HSIC}} = \frac{B}{m} \sum_{b=1}^{m/B} \hat{\eta}_b$. According to Zhang et al. (2018), in the asymptotic regime where $m \rightarrow \infty$, $B \rightarrow \infty$, and $m/B \rightarrow \infty$, we have

$$\sqrt{mB} \widehat{\text{HSIC}} \xrightarrow{D} \mathcal{N}(0, \sigma^2),$$

where σ^2 is the variance of the null distribution.

For notational convenience, let k and l denote the kernels used to construct $\mathcal{A}$ and $\mathcal{Z}$ in the oracle setting, as well as $\hat{\mathcal{A}}$ and $\hat{\mathcal{Z}}$ in the estimated setting of the AIT test. For simplicity, we show the asymptotic validity of the AIT test without covariates. Before that, we first introduce regularity assumptions on the kernels required for our theoretical results.

Assumption 4 (Boundedness and Lipschitz Kernels) *Let $\mathcal{U} \subset \mathbb{R}^{d_u}$ and $\mathcal{V} \subset \mathbb{R}^{d_v}$, and let $\|\cdot\|$ denote the Euclidean norm.*

- (1) *The kernel $l : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$ is bounded, i.e., there exists a constant $M > 0$ such that $|l(v, v')| \leq M$ for all $v, v' \in \mathcal{V}$.*
- (2) *The kernel $k : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}$ is Lipschitz continuous, i.e., there exists a constant $Q > 0$ such that $|k(u, u') - k(\tilde{u}, \tilde{u}')| \leq Q(\|u - \tilde{u}\| + \|u' - \tilde{u}'\|)$ for all $u, u', \tilde{u}, \tilde{u}' \in \mathcal{U}$.*

These two conditions are standard in the kernel literature. In particular, the Gaussian radial basis function (RBF) kernel satisfies both boundedness and Lipschitz continuity on compact domains. Moreover, similar regularity assumptions have been employed in kernel-based independence testing, including the regression-based conditional independence tests studied by Maia Polo et al. (2023); Ren et al. (2025).

Assumption 5 ($\sqrt{n}$ -Consistency of $\hat{h}$) *The estimator $\hat{h}$ satisfies the $\sqrt{n}$ -convergence rate:*

$$\|h - \hat{h}\|_{L_2(P_X)} = O_p(n^{-1/2}).$$

where h denotes the oracle function, and $\|\cdot\|_{L_2(P_X)}$ denotes the L_2 norm with respect to the distribution of X .

Note that fully nonparametric IV estimators generally fail to achieve the $\sqrt{n}$ -rate due to the ill-posedness of the associated inverse problem (Horowitz, 2012; Chen and Reiss, 2011; Hall and Horowitz, 2005). In the nonlinear case, we employ the control function IV estimator proposed by Guo and Small (2016), which is semiparametric and attains the $\sqrt{n}$ -rate under standard regularity conditions (Guo and Small, 2016; Murphy and Topel, 2002). In the linear case, the two-stage least squares (2SLS) estimator also satisfies Assumption 5 (Basmann, 1957; Henckel et al., 2024; Wooldridge, 2010).

Theorem 4 (Asymptotic Level and Power of the AIT Test) *Suppose that Assumptions 4 and 5 hold, and that the sample sizes satisfy $mB = o(n)$ as $n \rightarrow \infty$.*

- **Type I Error:** *Under the null hypothesis H_0 ,*

$$P_{H_0}(\text{Type I error}) = P_{H_0}\left(\sqrt{mB}\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) > c_{m,\alpha}\right) \leq \alpha + o(1),$$

- **Type II Error:** *Under any fixed alternative hypothesis H_1 such that $\text{HSIC}(\mathcal{A}, \mathcal{Z}) = \eta > 0$,*

$$P_{H_1}(\text{Type II error}) = P_{H_1}\left(\sqrt{mB}\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) \leq c_{m,\alpha}\right) = O\left(\frac{1}{\sqrt{mB}}e^{-cmB}\right), \quad mB \rightarrow \infty,$$

where $c_{m,\alpha}$ denotes the critical value, and the constant $c = \frac{\eta^2}{2\sigma_{H_1}^2} > 0$, and $\sigma_{H_1}^2$ is the asymptotic variance.

Proof See Appendix A.8 for its proof. ■

Theorem 4 establishes bounds on both Type I and Type II errors of the AIT test while simultaneously accounting for errors in both the estimation and independence testing stages. Specifically, the test controls the Type I error at the nominal level α asymptotically and achieves vanishing Type II error under fixed alternatives with $\widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) > 0$.

We next consider the theoretical asymptotic validity of the AIT test with covariates $\mathbf{W}$. To this end, we introduce three assumptions that will be used to establish the result.

Assumption 6 (Boundedness and Lipschitz Kernels) *Let $\mathcal{U} \subset \mathbb{R}^{d_u}$ and $\mathcal{V} \subset \mathbb{R}^{d_v}$, and let $\|\cdot\|$ denote the Euclidean norm. The kernels $k : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}$ and $l : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$ satisfy the following conditions:*

- (1) *There exists a constant $M > 0$ such that $|k(u, u')| \leq M$ and $|l(v, v')| \leq M$ for all $u, u' \in \mathcal{U}$ and all $v, v' \in \mathcal{V}$.*

- (2) *There exists a constant $Q > 0$ such that $|k(u, u') - k(\tilde{u}, \tilde{u}')| \leq Q(\|u - \tilde{u}\| + \|u' - \tilde{u}'\|)$ and $|l(v, v') - l(\tilde{v}, \tilde{v}')| \leq Q(\|v - \tilde{v}\| + \|v' - \tilde{v}'\|)$ for all $u, u', \tilde{u}, \tilde{u}' \in \mathcal{U}$ and all $v, v', \tilde{v}, \tilde{v}' \in \mathcal{V}$.*

With covariates, $\mathcal{Z}$ is estimated. Unlike Assumption 4, which requires boundedness and Lipschitz continuity for only one kernel (k or l), Assumption 6 imposes these conditions on both.

Assumption 7 ($\sqrt{n}$ -Consistency of $\hat{h}$ with Covariates) *The estimator $\hat{h}$ satisfies the $\sqrt{n}$ -convergence rate:*

$$\|h - \hat{h}\|_{L_2(P_{(X, \mathbf{W})})} = O_p(n^{-1/2}).$$

where h denotes the oracle function, and $\|\cdot\|_{L_2(P_{(X, \mathbf{W})})}$ denotes the L_2 norm with respect to the distribution of $(X, \mathbf{W})$.

Assumption 7 is an extension of Assumption 5 that takes covariates into account.

Assumption 8 (n^q -Consistency of $\hat{\pi}$) *The estimator $\hat{\pi}$ satisfies the convergence rate*

$$\|\pi - \hat{\pi}\|_{L_2(P_{\mathbf{W}})} = O_p(n^{-q}),$$

for some $q \in (0, 1/2)$, where $\|\cdot\|_{L_2(P_{\mathbf{W}})}$ denotes the L_2 norm with respect to the distribution of $\mathbf{W}$.

Assumption 8 is formulated as a high-level rate condition on the regression estimator. In our implementation, we employ random forests to estimate $\hat{\pi}$. Random forests are known to be consistent under suitable structural assumptions, such as additive or low-dimensional models (Scornet et al., 2015). More generally, under standard nonparametric smoothness assumptions (e.g., Hölder smoothness of order s in dimension d), regression estimators typically achieve convergence rates of the form $n^{-s/(2s+d)}$, which corresponds to some $q \in (0, 1/2)$ (Stone, 1982; Yang and Tokdar, 2015). Therefore, Assumption 8 is compatible with commonly used nonparametric estimators, including random forests.

Corollary 2 (Asymptotic Level and Power of the AIT Test with Covariates)

Suppose that Assumptions 6 ~ 8 hold, and that the sample sizes satisfy $mB = o(n^{2q})$ as $n \rightarrow \infty$.

- **Type I Error:** *Under the null hypothesis,*

$$P(\text{Type I error}) = P_{H_0}(\sqrt{mB} \widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) > c_{m,\alpha}) \leq \alpha + o(1),$$

- **Type II Error:** *Under any fixed alternative hypothesis H_1 such that $\text{HSIC}(\mathcal{A}, \mathcal{Z}) = \eta > 0$,*

$$P_{H_1}(\text{Type II error}) = P_{H_1}(\sqrt{mB} \widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) \leq c_{m,\alpha}) = O\left(\frac{1}{\sqrt{mB}} e^{-cmB}\right), \quad mB \rightarrow \infty,$$

where $c_{m,\alpha}$ denote the critical value, and the constant $c = \frac{\eta^2}{2\sigma_{H_1}^2} > 0$, and $\sigma_{H_1}^2$ is the asymptotic variance.

Proof See Appendix A.9 for its proof. ■

Corollary 2 establishes that the proposed test controls the Type I error asymptotically and achieves consistency under fixed alternatives, with power converging to one as $mB \rightarrow \infty$.

Remark 6 *We would like to mention two main tuning parameters in practical implementations: the data split ratio ρ and the block size B . Regarding the data split ratio ρ : Theoretically, our method requires the rate condition $mB = o(n)$ (or $mB = o(n^{2q})$ in the presence of covariates). In practice, we heuristically set the data split ratio to $\rho = 0.7$, following Maia Polo et al. (2023); Wang et al. (2025). Moreover, the experimental results in Section 5 show that this choice leads to favorable performance. Regarding the block size B : The sample size of each block B can be guided by Zaremba et al. (2013); Zhang et al. (2018), who showed that the null distribution is well approximated by the central limit theorem when B is small, which thus helps the Type I error approach the nominal level. However, a smaller B may lead to reduced statistical power for a given sample size. In our experiments (Section 5), we heuristically set $B = m/15$ in the continuous setting and $B = m/6$ in the discrete setting, which demonstrates superior performance.*

5. Experiments

In this section, we evaluated the performance of the proposed AIT condition for instrument validity across both synthetic data and three real-world datasets. Our source code is available at https://github.com/zhengli0060/AIT_Condition.

5.1 Synthetic Data

We conducted simulation experiments from four perspectives. First, in Section 5.1.1, we verified the correctness of our theoretical results. Then, in Section 5.1.2, we compared the IV-PIM method proposed by Bureau (2023), which is designed for continuous treatments with covariates. Moreover, in Section 5.1.3, we compared the proposed method with Kitagawa’s method, abbreviated as the K-test method (Kitagawa, 2015), which is designed for discrete treatments. Finally, in Section 5.1.4, we validate our theoretical results under an oracle setting, where the auxiliary variable is correctly specified (i.e., no estimation error in Step 1 of Algorithm 1). In all experiments, we evaluated the performance of our method using the following metrics:

- **Valid IVs Misidentification Ratio (abbreviated as Valid MR)**: the ratio of the number of valid IVs incorrectly identified in the output.
- **Invalid IVs Misidentification Ratio (abbreviated as Invalid MR)**: the ratio of the number of invalid IVs incorrectly identified in the output.

5.1.1 THEORETICAL VALIDATION OF PROPOSED METHOD

In this section, we conducted simulations to verify our theoretical results for both the linear model and the ANINCE model. Table 2 summarizes the propositions validated by the

experimental results reported in various tables, along with the corresponding model settings. Specifically, we tested our method’s ability to identify invalid IVs that violate the *exogeneity* condition under three scenarios: linear constant-effect setting with varied distributions, partially nonlinear constant conditions (with Gaussian noise terms), and partially nonlinear non-constant conditions with varied functional forms but a consistent Uniform distribution. The performance results are presented in Table 3, Table 4, and Table 5, respectively. Further, we examined the proposed method’s effectiveness in identifying invalid IVs that violate the *exclusion restriction* condition under nonlinear constant and nonlinear non-constant conditions, considering various functional forms as well as Beta and Uniform distributions, respectively. Results for these tests are shown in Table 6 and Table 7.

Table 2: Summary of propositions validated by the experimental results, along with the corresponding model settings.

Table	Proposition Demonstrated	Model	IV Violation
Table 3	Propositions 1 and 2	Linear	Exogeneity
Table 4	Proposition 4	Partial Non-Linear Constant Effect	Exogeneity
Table 5	Proposition 4	Partial Non-Linear Non-Constant Effect	Exogeneity
Table 6	Proposition 4	Partial Non-Linear Constant Effect	Exclusion Restriction
Table 7	Proposition 4	Non-Linear Non-Constant Effect	Exclusion Restriction

Note: “Proposition Demonstrated” refers to the propositions that are supported by the experimental results presented in each table.

Experimental Design: We generated data according to the model in Equation (3), where each candidate IV set contains both a valid IV and an invalid IV. For the nonlinear non-constant effects model, we applied five functions—*logarithmic*, *quadratic polynomial*, *cubic polynomial*, *logarithmic quadratic polynomial*, *exponent quadratic polynomial*—for $g(\cdot)$, $f(\cdot)$, and $\varphi_*(\cdot)$. In contrast, these functions were linear under linear, constant settings. For the error term ε_* , we selected six distributions: *Gaussian*, *Uniform*, *T*, *Beta*, *Gamma*, *Log-normal*.⁵ In all experiments, the significance level was set to 0.05 for sample sizes smaller than 1000, and to 0.01 otherwise, with the sample splitting ratio fixed at $\rho = 0.7$. Each experiment was repeated 100 times with randomly generated data, and the reported results were averaged. The sample sizes are chosen from 3000 (3k), 5000 (5k), and 7000 (7k). Additional experimental details can be found in Appendix C.

Table 3: Performance in Testing Violations of Exogeneity within Linear Models.

Distribution	Size=3K		Size=5K		Size=7K	
	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓
Uniform	0.00	0.04	0.00	0.00	0.00	0.00
Beta	0.01	0.12	0.00	0.03	0.00	0.01
T-distribution	0.02	0.08	0.00	0.04	0.00	0.01
Gamma	0.01	0.08	0.00	0.01	0.00	0.01
Log-normal	0.01	0.08	0.00	0.02	0.00	0.00
Gaussian	0.00	1.00	0.00	1.00	0.00	1.00
Mixed	0.01	0.20	0.00	0.15	0.00	0.05

Note: ↓ means a lower value is better, and vice versa.

5. The “Mixed distribution” refers to that obtained by randomly selecting from the mentioned distributions.

Table 4: Performance in Testing Violations of Exogeneity within Partial Non-Linear Constant Effect Models.

Function	Size=3K		Size=5K		Size=7K	
	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓
Log	0.00	0.01	0.00	0.00	0.00	0.00
Quadratic polynomial	0.00	0.00	0.00	0.00	0.00	0.00
Cubic polynomial	0.00	0.00	0.00	0.00	0.00	0.00
Log(quadratic)	0.00	0.00	0.00	0.00	0.00	0.00
Exp(quadratic)	0.00	0.00	0.00	0.00	0.00	0.00

Note: ↓ means a lower value is better, and vice versa.

Table 5: Performance in Testing Violations of Exogeneity within Partial Non-Linear Non-Constant Effect Models.

Function	Size=3K		Size=5K		Size=7K	
	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓
Log	0.00	0.00	0.00	0.00	0.00	0.00
Quadratic polynomial	0.00	0.00	0.00	0.00	0.00	0.00
Cubic polynomial	0.00	0.00	0.00	0.00	0.00	0.00
Log(quadratic)	0.00	0.00	0.00	0.00	0.00	0.00
Exp(quadratic)	0.00	0.01	0.00	0.00	0.00	0.00

Note: ↓ means a lower value is better, and vice versa.

Table 6: Performance in Testing Violations of Exclusion Restriction within Partial Non-Linear Constant Effect Models.

Function	Size=3K		Size=5K		Size=7K	
	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓
Log	0.02	0.11	0.01	0.08	0.00	0.04
Quadratic polynomial	0.00	0.09	0.00	0.05	0.00	0.02
Cubic polynomial	0.01	0.12	0.00	0.07	0.00	0.01
Log(quadratic)	0.00	0.16	0.00	0.09	0.00	0.04
Exp(quadratic)	0.01	0.09	0.00	0.05	0.00	0.03

Note: ↓ means a lower value is better, and vice versa.

Table 7: Performance in Testing Violations of Exclusion Restriction within Non-Linear Non-Constant Effect Models.

Function	Size=3K		Size=5K		Size=7K	
	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓
Log	0.00	0.16	0.00	0.12	0.00	0.04
Quadratic polynomial	0.00	0.02	0.00	0.00	0.00	0.00
Cubic polynomial	0.00	0.02	0.00	0.00	0.00	0.00
Log(quadratic)	0.01	0.06	0.00	0.02	0.00	0.00
Exp(quadratic)	0.00	0.00	0.00	0.00	0.00	0.00

Note: ↓ means a lower value is better, and vice versa.

Results: As shown in Tables 3 ~ 7, both metrics generally improve significantly with increasing sample sizes across various distributions and functions. Those facts suggest that our method can correctly identify invalid IVs that violate the *exogeneity* condition or violate the *exclusion restriction* condition. More specifically, Table 3 highlights that non-Gaussianity is beneficial for identifying IV in linear models, as demonstrated in Figure 2 and supported by Proposition 2. Notably, the Invalid MR value in the Gaussian distribution of the linear model in Table 3 is 1, indicating that our method cannot detect invalid IVs in a linear Gaussian model. This finding is consistent with the conclusions presented in Proposition 1. Tables 4 ~ 5 further show that even a slight degree of nonlinearity facilitates the assessment of IV validity under the exogeneity condition in both nonlinear constant and non-constant effect models, as illustrated in Figure 3 and stated in Proposition 4. Lastly, Tables 6 ~ 7 reveal that in nonlinear models, when the direct causal effect of $Z \rightarrow Y$ does not follow a linear function of the effect of $Z \rightarrow X$, it becomes possible to assess the validity of IVs solely concerning the exclusion restriction condition, as highlighted in the negative of Proposition 5 (i) and elaborated in Proposition 4.

5.1.2 COMPARISON WITH IV-PIM IN CONTINUOUS TREATMENT SETTING

In this section, we compared the proposed *AIT Condition* with IV-PIM, as proposed by Buraue (2023), in continuous treatment settings. Note that since IV-PIM requires covariates, we introduce covariates here for a fair comparison.

Experimental Design: The specific generation mechanism with covariates $\mathbf{W}$ in the linear model is defined as follows: $U = \varepsilon_U$, $\mathbf{W} = \varepsilon_{\mathbf{W}}$, $Z_1 = \mathcal{I}(U + \mathbf{W} + \varepsilon_{Z_1})$, $Z_2 = \mathcal{I}(\mathbf{W} + \varepsilon_{Z_2})$, $X = 0.5Z_1 + 0.5Z_2 + \lambda\mathbf{W} + \delta$, and $Y = X + \mathbf{W} + \epsilon$, where $\varepsilon_U \sim T(5)$, $\varepsilon_{Z_1} \sim \text{Beta}(0.5, 0.1)$, $\varepsilon_{Z_2} \sim \mathcal{N}(0, 1)$, and $\delta, \epsilon \sim T(5)$. Here, $\mathcal{I}(\cdot)$ is the indicator function such that $\mathcal{I}(\cdot) > \text{mean}(\cdot)$ equals 1; otherwise, it is 0. The coefficient λ is randomly drawn from a normalized standard normal distribution. The noise terms $\varepsilon_{\mathbf{W}}$ follow a multidimensional normal distribution and are consistent with IV-PIM, with the dimensionality of covariates $\mathbf{W}$ varying across $|\mathbf{W}| = \{2, 3, 5\}$. The remaining settings are the same as in Section 5.1.1.

Results: As shown in Table 8, our method outperforms experimental results of IV-PIM with covariates under both Valid MR and Invalid MR. Interestingly, IV-PIM’s performance improves as the dimensionality of covariates increases, consistent with findings in Buraue (2023). Additionally, Table 8 highlights the practicality of the AIT condition with covariates, as presented in Corollary 1.

Table 8: Performance in Testing Instrumental Variables with Covariates.

W	Condition	Size=3K		Size=5K		Size=7K	
		Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓
2	IV-PIM method (Buraue, 2023)	0.22	0.25	0.22	0.18	0.32	0.22
	AIT condition	0.00	0.11	0.00	0.01	0.00	0.00
3	IV-PIM method (Buraue, 2023)	0.15	0.27	0.20	0.21	0.15	0.16
	AIT condition	0.01	0.05	0.00	0.00	0.00	0.00
5	IV-PIM method (Buraue, 2023)	0.11	0.16	0.06	0.25	0.08	0.23
	AIT condition	0.00	0.02	0.00	0.00	0.00	0.00

Note: ↓ means a lower value is better, and vice versa.

5.1.3 COMPARISON WITH K-TEST IN DISCRETE TREATMENT SETTING

In this section, we compared the proposed instrument validity test with the K-test, which was proposed by Kitagawa (2015) for discrete treatment settings without covariates. The source code for the K-test is available at <https://rdrr.io/github/CarrThomas/TestforInstrumentValidity/>.

Experimental Design: The discrete-treatment data were generated to simulate violations of the *exogeneity* and *exclusion restriction* conditions as follows: $U = \varepsilon_U$, $Z = \mathcal{I}(\varphi_Z(U) + \varepsilon_Z)$, $X = \mathcal{I}(g_X(Z) + \varphi_X(U) + \varepsilon_X)$, $Y = \beta X + g_Y(Z) + \varphi_Y(U) + \varepsilon_Y$, and $\varepsilon_* \sim \mathcal{N}(0, 1)$, where the causal effect $\beta = 1$, and $\mathcal{I}(\cdot)$ is the indicator function such that $\mathcal{I}(\cdot) > \text{mean}(\cdot)$ equals 1; otherwise, it is 0. The functions $\varphi_*(U)$ and $g_*(Z)$ are nonlinear and randomly selected from the following set: *cos*, *sin*, *square*, *cubic* (third-degree polynomials), *logarithmic*, *exponential*. The significance level is set to be the same for both methods at each sample size. The remaining settings are the same as in Section 5.1.1.

Results: As shown in Table 9, the proposed AIT condition exhibits competitive performance relative to the K-test in terms of Invalid MR. Specifically, AIT achieves lower Invalid MR at sample sizes of 3K and 7K, while yielding comparable performance at 5K. For the *Valid MR* metric, both methods achieve a value of 0, indicating that neither method mistakenly identifies valid IVs as invalid.

Table 9: Performance in Testing Instrumental Variables with Discrete Treatment.

Condition	Size=3K		Size=5K		Size=7K	
	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓	Valid MR↓	Invalid MR↓
K-test method (Kitagawa, 2015)	0.00	0.20	0.00	0.10	0.00	0.08
AIT condition	0.00	0.15	0.00	0.13	0.00	0.07

Note: ↓ means a lower value is better, and vice versa.

5.1.4 THEORETICAL VALIDATION WITH THE AUXILIARY VARIABLE CORRECTLY SPECIFIED

In this section, we present additional simulation experiments to verify the theoretical results with the auxiliary variable correctly specified. Specifically, we report results for both valid and invalid IV testing cases.

- For **valid IV testing**, we evaluate the Valid MR under three different model settings: the Linear Model, the Partial Non-Linear Model with Constant Causal Effect, and the ANINCE Model.
- For **invalid IV testing**, we evaluate the Invalid MR under various scenarios where IV assumptions are violated. These include: the Linear Model with exogeneity violated; the Linear Model with both exogeneity and exclusion restriction violated; the Partial Non-Linear Model with Constant Causal Effect where exogeneity is violated; the Partial Non-Linear Model with Constant Causal Effect where exclusion restriction is violated; and the Partial Non-Linear Model with Constant Causal Effect where both exogeneity and exclusion restriction are violated.

It is worth noting that for the nonparametric invalid IV model, we are unable to derive the explicit closed-form solution for the auxiliary variable. As a result, we cannot report the

Invalid MR for that setting. Besides, in each model, we include only a single candidate IV to focus on the core objective of the test—assessing the validity of that IV. The remaining settings are the same as in Section 5.1.1.

Results. As shown in Tables 10 and 11, both Valid MR and Invalid MR are low across different sample sizes and model specifications. Specifically,

- The Valid MR is consistently low in both the constant and non-constant effect models, indicating that the test appropriately controls the false positive rate when the auxiliary variable is correctly specified.
- The Invalid MR is also very low in the constant effect model with an invalid IV, suggesting that the test has strong power to detect violations of IV validity.

In summary, these results provide empirical support for the validity of our test when the auxiliary variable is assumed to be correctly specified.

Table 10: Performance in Testing Valid IV with the Auxiliary Variable Correctly Specified in Different Cases.

Cases	Valid MR		
	Size=3K	Size=5K	Size=7K
Linear Model	0.00	0.00	0.00
Partial Non-Linear Model with Constant Causal Effect	0.00	0.00	0.00
ANINCE Model	0.00	0.00	0.00

Table 11: Performance in Testing Invalid IV with the Auxiliary Variable Correctly Specified in Different Cases.

Cases	Invalid MR		
	Size=3K	Size=5K	Size=7K
Linear Model: Exogeneity Violated	0.01	0.00	0.00
Linear Model: Exogeneity & Exclusion Restriction Violated	0.00	0.00	0.00
Partial Non-Linear Model with Constant Causal Effect: Exogeneity Violated	0.00	0.00	0.00
Partial Non-Linear Model with Constant Causal Effect: Exclusion Restriction Violated	0.02	0.00	0.00
Partial Non-Linear Model with Constant Causal Effect: Exogeneity & Exclusion Restriction Violated	0.01	0.00	0.00

5.2 Real-World Datasets

In this section, we evaluated the effectiveness of the proposed method by applying it to three real-world datasets from different domains.

5.2.1 SCHOOLING-RETURNS DATA

We consider the application of our method to the study by Card (1993). This study investigates the impact of education levels on earnings using data from the Young Men Cohort of the National Longitudinal Survey.

Data Description: The dataset is a sample of 3010 men taken from the US National Longitudinal Survey of Young Men (NLSY). It includes variables such as *LivedNearCollege*,

Schooling, *Returns*, and a set of **Covariates** including $\{Experience, Experience\ square, Black, Smsa, Smsa66, Region\ information\ (reg662-reg669), South\}$, among others. The hypothesized model of Card (1993) is presented in Figure 5. The hypothesized data generation mechanism is described as follows:

$$\begin{aligned} Schooling &= \alpha_0 + \alpha_1 \cdot LivedNearCollege + \alpha^\top \cdot \mathbf{Covariates} + \delta, \\ Returns &= \beta_0 + \beta_1 \cdot Schooling + \beta^\top \cdot \mathbf{Covariates} + \epsilon, \end{aligned} \quad (19)$$

where δ and ϵ are dependent.

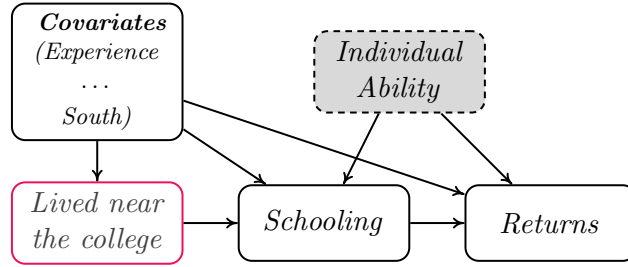


Figure 5: Graphical illustration of an IV model for estimating the causal effect of *Schooling* on *Returns* (Card, 1993).

Results: Card (1993) demonstrated that *LivedNearCollege* can serve as a valid IV for the causal relationship $Schooling \rightarrow Returns$, while controlling for the **Covariates**. For consistency, we adopt the causal effect of *Schooling* on *Returns*, i.e., $\beta_1 = 0.1315$, as well as the coefficients β of the covariates from Card (1993), as the estimated parameter $\hat{\beta}$ under the AIT condition. We then obtain the residual $\widetilde{LivedNearCollege}$ by regressing *LivedNearCollege* on the **Covariates**. The *P*-value of the independence test between the auxiliary variable and the residual $\widetilde{LivedNearCollege}$ is 0.73, indicating that we cannot reject *LivedNearCollege* as a valid IV. This result further supports the validity of using *LivedNearCollege* as an IV, consistent with the findings in Card (1993).

5.2.2 COLONIAL ORIGINS DATA

We apply our method to the study by Acemoglu et al. (2001), which estimates the impact of colonial history on the economic development of different regions using the Colonial Origins of Comparative Development dataset.

Data Description: The dataset includes five key variables across 64 countries, after excluding samples with missing data. These variables are: *Mortality*, *Euro1990*, *Latitude*, *Institutions*, and *Economic Development*. The hypothesized model proposed by Acemoglu et al. (2001) is illustrated in Figure 6, and the hypothesized data generation mechanism is

described as follows:

$$\begin{aligned} \text{Institutions} &= \gamma + (\gamma_1 \quad \gamma_2 \quad \gamma_3) \begin{pmatrix} \text{Mortality} \\ \text{Latitude} \\ \text{Euro1990} \end{pmatrix} + \delta, \\ \text{Economic Development} &= \beta + (\beta_1 \quad \beta_2) \begin{pmatrix} \text{Institutions} \\ \text{Latitude} \end{pmatrix} + \epsilon, \end{aligned} \tag{20}$$

where δ and ϵ are dependent. Here, we adjust the parameter to $B = m/10$ to avoid overly small subsamples and improve the stability of estimation, due to the small sample size ($m = 64$) in this dataset.

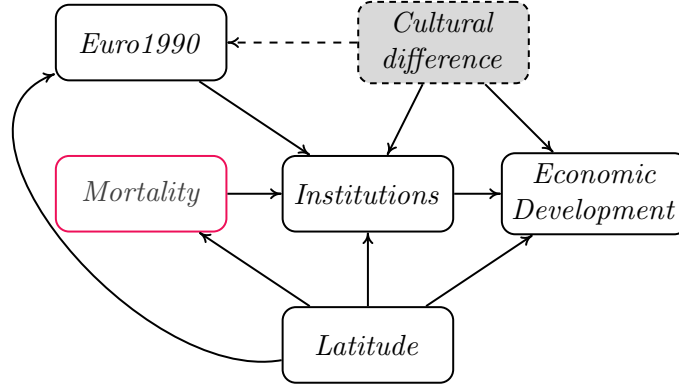


Figure 6: Graphical illustration of an IV model for estimating the causal effect of *Institutions* on *Economic Development* (Acemoglu et al., 2001).

Results: Acemoglu et al. (2001) demonstrated that both *Mortality* and *Euro1990* can serve as valid IVs, conditional on *Latitude*, with respect to *Institutions* and *Economic Development*. To verify this, we test their validity using the AIT condition. For consistency, we adopt the causal effects of *Institutions* on *Economic Development* as reported by Acemoglu et al. (2001), specifically $\beta_1 = 0.9458$ and $\beta_2 = -0.5971$, as the estimated parameters $\hat{\beta}_i (i = 1, 2)$ in the AIT condition. We next obtain the residuals *Mortality* and *Euro1990* by regressing *Mortality* and *Euro1990* on covariate *Latitude*, respectively. The validity test for *Mortality* yields a *P*-value of 0.61, whereas the test for *Euro1990* yields a *P*-value of 0.25. These results indicate that *Euro1990* is more likely to be an invalid IV compared to *Mortality*, suggesting that the exogeneity of *Euro1990* is weaker than that of *Mortality*. These findings are consistent with those of Acemoglu et al. (2001), and we cannot reject the validity of *Mortality* and *Euro1990* as IVs, aligning with their conclusions.

5.2.3 CONFLICT AND TIME PREFERENCE DATA

We consider the application of our method to the study by Voors et al. (2012). This study uses the Conflict and Time Preference Data to investigate the impact of violence on a person's patience.

Data Description: The dataset consists of 302 observations and fifteen variables, as described in Voors et al. (2012). Treatment variable, *Violence*, is measured by the percentage

dead in attacks in the area the person lived, while the person's *Patience* (outcome variable) is assessed by a person's discount rate for willingness to receive larger amounts of money in the future compared to smaller amounts of money now. Other variables include *Distance* (distance to Bujumbura) and *Altitude*, and **Covariates** such as *{whether the respondent is literate, the respondent's age, the respondent's sex, the total land holding per capita, land Gini coefficient, distance to market, conflict over land, ethnic homogeneity, socioeconomic homogeneity, population density, per capita total expenditure}*. The **Covariates** used in the study represent exogenous personal and geographical information variables. As discussed in Voors et al. (2012), violence may be targeted in a non-random way, potentially related to community patience, which makes violence endogenous. The hypothesized model from Guo and Small (2016) is illustrated in Figure 7, and the hypothesized generation mechanism is as follows:

$$\begin{aligned}
 \text{Violence} &= \alpha_0 + (\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \alpha_4 \quad \alpha_5 \quad \alpha_6^\top) \begin{pmatrix} \text{Distance} \\ \text{Altitude} \\ \text{Distance}^2 \\ \text{Altitude}^2 \\ \text{Distance} \cdot \text{Altitude} \\ \text{Covariates} \end{pmatrix} + \delta, \\
 \text{Patience} &= \beta_0 + (\beta_1^\top \quad \beta_2 \quad \beta_3) \begin{pmatrix} \text{Covariates} \\ \text{Violence} \\ \text{Violence}^2 \end{pmatrix} + \epsilon,
 \end{aligned} \tag{21}$$

where δ and ϵ are dependent.

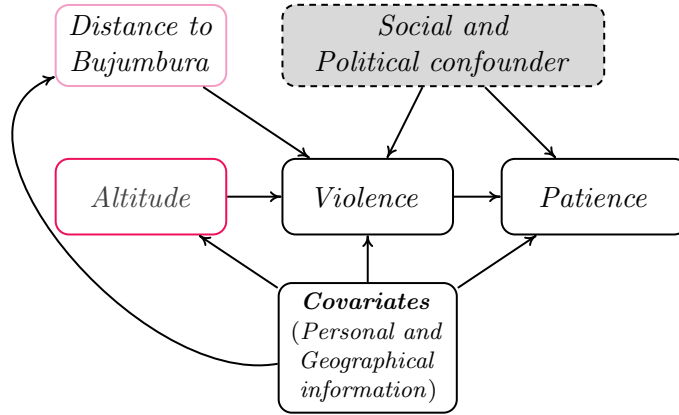


Figure 7: Graphical illustration of an IV model for estimating the causal effect of *Violence* on a person's *Patience* (Voors et al., 2012).

Results: Voors et al. (2012) showed that both *Distance* and *Altitude* can serve as valid IVs, conditional on **Covariates**, with respect to *Violence* on *Patience*. To verify this, we test their validity using the AIT condition. For consistency, we adopt the causal effects of *Violence* on *Patience* as reported by Guo and Small (2016), specifically $\beta_2 = 2.054$ and $\beta_3 = 0.049$, as the estimated parameters $\hat{\beta}_i (i = 2, 3)$ in the AIT condition. We obtain

the residual $\widetilde{Distance}$ and $\widetilde{Altitude}$ by regressing $Distance$ and $Altitude$ on **Covariates**, respectively. We first test the validity of $Distance$, which yields a P -value of 0.33, suggesting that $Distance$ cannot be rejected as a valid instrumental variable. This is consistent with the findings of Voors et al. (2012). Similarly, testing $Altitude$ as a potential IV results in a P -value of 0.76, indicating that $Altitude$ also cannot be rejected as a valid IV. These results align with the conclusions of Voors et al. (2012).

6. Conclusions

In this paper, we explored the testability of single IVs in the additive nonlinear, non-constant effects (ANINCE) model, where the treatment variable can be either discrete or continuous. To this end, we introduced a necessary condition, termed the AIT condition. We showed that, under the completeness condition, the AIT condition can detect whether a variable is a valid IV without requiring knowledge of whether any other variable serves as an instrument. Furthermore, we provided additional precision conditions for identifying all invalid IVs in linear and ANINCE models. We then proposed the practical AIT condition test algorithm with covariates and finite data, and established its asymptotic level and power. Experimental results using both simulation data and three real datasets have further validated the usefulness of our algorithm.

A limitation of the current asymptotic analysis of the AIT test is its reliance on the $\sqrt{n}$ -consistency of the IV estimator $\hat{h}$ (Assumption 5). Although this rate is plausible for parametric and some semiparametric IV estimators, it is generally difficult to attain for fully nonparametric IV estimators because of the ill-posed inverse problem. Future work could therefore develop more efficient testing procedures that maintain asymptotic validity under weaker rate requirements. More broadly, this paper studies testability in the single-IV setting, whereas some applications involve multiple candidate instruments. It would be useful to examine whether the AIT condition can be extended to provide testable implications for IV sets, including cases where some instruments may be invalid.

Acknowledgments

We thank the AE and the anonymous reviewers for their insightful comments and constructive suggestions, which have substantially improved the quality of this paper. We are also grateful to Kun Zhang for valuable discussions and helpful suggestions. We thank Patrick Buraue for kindly sharing his R implementation of IV-PIM. FX acknowledges support from the National Natural Science Foundation of China (Grant No. 62306019). YZ would like to acknowledge the support of the Beijing Natural Science Foundation (4264116). FX and YZ were supported by the Beijing Key Laboratory of Applied Statistics and Digital Regulation, and the BTBU Digital Business Platform Project by BMEC. XCG acknowledges the support of the Graduate Research Ability Enhancement Program Project Funding at Beijing Technology and Business University in 2024.

Appendix A. Proofs

Before presenting the proofs, we introduce three important theorems since these are used to prove our results. We begin by recalling the theorem on the properties of independent random variables from Meester (2008). These results serve as the foundation for the proofs of Theorem 1 and Corollary 1.

Theorem 5 (Theorem 2.2.5 in Meester (2008)) *Let $X_1, X_2, \dots, X_m$ be mutually independent random variables, and for $i = 1, \dots, m$, g_i be a function $g_i : \mathbb{R} \rightarrow \mathbb{R}$. Then the random variables $g_1(X_1), g_2(X_2), \dots, g_m(X_m)$ are also mutually independent.*

Corollary 3 *Let X_1, X_2 and X_3 be mutually independent random variables; and let function $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ and function $h : \mathbb{R} \rightarrow \mathbb{R}$. Then, $g(X_1, X_2)$ and $h(X_3)$ are independent random variables.*

Proof Let $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ and $h : \mathbb{R} \rightarrow \mathbb{R}$. We aim to show that the random variables $g(X_1, X_2)$ and $h(X_3)$ are independent. We here provide the proof for the case where the random variables are discrete. The argument for continuous random variables is similar and can be obtained by replacing the summation with integration.

We compute:

$$\begin{aligned}
 & P(g(X_1, X_2) = a, h(X_3) = b) \\
 &= \sum_{\substack{x_1, x_2, x_3: \\ g(x_1, x_2) = a, h(x_3) = b}} P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \\
 &= \sum_{\substack{x_1, x_2, x_3: \\ g(x_1, x_2) = a, h(x_3) = b}} P(X_1 = x_1) P(X_2 = x_2) P(X_3 = x_3) \\
 &= \left(\sum_{\substack{x_1, x_2: \\ g(x_1, x_2) = a}} P(X_1 = x_1) P(X_2 = x_2) \right) \left(\sum_{\substack{x_3: \\ h(x_3) = b}} P(X_3 = x_3) \right) \\
 &= P(g(X_1, X_2) = a) \cdot P(h(X_3) = b)
 \end{aligned}$$

Therefore, $g(X_1, X_2)$ and $h(X_3)$ are independent. ■

We then quote the Darmois–Skitovitch theorem that characterizes the independence of two linear statistics (Darmois, 1953; Skitovitch, 1953). This theorem provides the foundation for proving Propositions 2 ~ 3 and Theorem 2.

Theorem 6 (Darmois–Skitovitch Theorem) *Define two random variables V_1 and V_2 as linear combinations of independent random variables $\varepsilon_1, \dots, \varepsilon_p$:*

$$V_1 = \sum_{i=1}^p \alpha_i \varepsilon_i, \quad V_2 = \sum_{i=1}^p \beta_i \varepsilon_i, \quad (22)$$

where the α_i, β_i are constant coefficients. If V_1 and V_2 are independent, then the random variables ε_j for which $\alpha_j \beta_j \neq 0$ are Gaussian.

The above theorem states that if there exists a non-Gaussian ε_j for which $\alpha_j\beta_j \neq 0$, V_1 and V_2 are dependent.

Next, we introduce a local geometric information theorem that characterizes the independence of two nonlinear statistics (Lin, 1997). This result provides the foundation for proving Propositions 4 ~ 5, and Theorem 3.

Theorem 7 (Proposition in Lin (1997)) *The Hessian H_f of function f is block diagonal everywhere, $\partial_i\partial_j f|_{\vec{s}_0} = 0$ for all points $\vec{s}_0$ and all $i \leq k, j > k$, if and only if f is separable into a sum $f(s_1, \dots, S_m) = g(s_1, \dots, s_k) + h(s_{k+1}, \dots, S_m)$ for some functions g and h .*

Theorem 7 states that a function f is separable if and only if its mixed second-order partial derivative is zero.

A.1 Proof of Proposition 2: Testability of Exogeneity in Linear Models

Proof Under the assumption of a linear model, Equation (3) can be expressed as follows:

$$\begin{aligned} \mathbf{U} &= \varepsilon_U, & Z &= \boldsymbol{\gamma}^T \mathbf{U} + \varepsilon_Z, \\ X &= \tau Z + \boldsymbol{\rho}^T \mathbf{U} + \varepsilon_X, & Y &= \beta X + \nu Z + \boldsymbol{\kappa}^T \mathbf{U} + \varepsilon_Y. \end{aligned} \quad (23)$$

Let $h(X) = \hat{\beta}X = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)}X$. According to the AIT condition, in the linear model, the function $h(X)$ satisfies the conditional moment restriction, i.e., $\mathbb{E}[Y - h(X) | Z] = 0$. By Equation (23), $\hat{\beta}$ can be expressed as follows:

$$\hat{\beta} = \beta + \underbrace{\frac{(\nu\boldsymbol{\gamma}^T + \boldsymbol{\kappa}^T)\text{Cov}(\mathbf{U})\boldsymbol{\gamma} + \nu \text{Var}(\varepsilon_Z)}{(\tau\boldsymbol{\gamma}^T + \boldsymbol{\rho}^T)\text{Cov}(\mathbf{U})\boldsymbol{\gamma} + \tau \text{Var}(\varepsilon_Z)}}_{\beta_{\text{bias}}}. \quad (24)$$

Hence, we have

$$\begin{aligned} \mathcal{A}_{X \rightarrow Y|Z} &= Y - \hat{\beta}X \\ &= \nu Z + \boldsymbol{\kappa}^T \mathbf{U} + \varepsilon_Y - \beta_{\text{bias}}X \\ &= [\nu\boldsymbol{\gamma}^T + \boldsymbol{\kappa}^T - (\nu\tau + \boldsymbol{\rho}^T)\beta_{\text{bias}}]\boldsymbol{\varepsilon}_U + (\nu - \tau\beta_{\text{bias}})\varepsilon_Z - \beta_{\text{bias}}\varepsilon_X + \varepsilon_Y. \end{aligned} \quad (25)$$

Since Z is an invalid IV violating exogeneity, without loss of generality, we assume that U_i causes the candidate IV Z , i.e., $\gamma_i \neq 0$, then $\beta_{\text{bias}} \neq 0$. This will imply that $[\nu\gamma_i + \kappa_i - (\nu\tau + \rho_i)\beta_{\text{bias}}] \neq 0$ for ε_{U_i} , and $(\nu - \tau\beta_{\text{bias}}) \neq 0$ for ε_Z . Furthermore, because of Assumption 2, i.e., (i) there exists at least one variable $U_i \in \mathbf{U}$ whose noise term follows a non-Gaussian distribution and cause Z or, (ii) the noise term of Z follows a non-Gaussian distribution, then at least one of the non-Gaussian noise terms ε_{U_i} or ε_Z , is common between $\mathcal{A}_{X \rightarrow Y|Z}$ and Z . Due to the Darmois-Skitovitch Theorem, we have $\mathcal{A}_{X \rightarrow Y|Z}$ is dependent on Z . That is to say, $\{X, Y|Z\}$ violates the AIT condition. ■

A.2 Proof of Proposition 3: Non-Testability of Exclusion Restriction in Linear Models

Proof Because candidate IV Z satisfies exogeneity condition, the model of Equation (3) can be written as:

$$\begin{aligned} \mathbf{U} &= \varepsilon_U, & Z &= \varepsilon_Z, \\ X &= \tau Z + \boldsymbol{\rho}^T \mathbf{U} + \varepsilon_X, & Y &= \beta X + \nu Z + \boldsymbol{\kappa}^T \mathbf{U} + \varepsilon_Y. \end{aligned} \quad (26)$$

Let $h(X) = \hat{\beta}X = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)}X$. According to the AIT condition, in the linear model, the function $h(X)$ satisfies the conditional moment restriction, i.e., $\mathbb{E}[Y - h(X) | Z] = 0$. By Equation (26), $\hat{\beta}$ can be expressed as follows:

$$\hat{\beta} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} = \beta + \frac{\nu}{\tau} = \beta + \beta_{\text{bias}}.$$

Hence, we obtain

$$\begin{aligned} \mathcal{A}_{X \rightarrow Y || Z} &= Y - \hat{\beta}X \\ &= \nu Z + \boldsymbol{\kappa}^T \mathbf{U} + \varepsilon_Y - \beta_{\text{bias}}X \\ &= \left(\boldsymbol{\kappa}^T - \frac{\boldsymbol{\rho}^T \nu}{\tau} \right) \varepsilon_U - \frac{\nu}{\tau} \varepsilon_X + \varepsilon_Y. \end{aligned} \quad (27)$$

When $\nu = 0$, Equation (26) implies that Z is a valid IV, we have $\mathcal{A}_{X \rightarrow Y || Z} = \boldsymbol{\kappa}^T \varepsilon_U + \varepsilon_Y$. Since Z only contains the noise term ε_Z , in both cases—whether Z is an invalid IV that violates solely the exclusion restriction condition or a valid IV—there are no shared noise terms between $\mathcal{A}_{X \rightarrow Y || Z}$ and Z . By the Darmois–Skitovitch Theorem, we have $\mathcal{A}_{X \rightarrow Y || Z}$ is independent of Z . Thus, $\{X, Y || Z\}$ always *satisfies* the AIT condition. ■

A.3 Proof of Theorem 2: Necessary and Sufficient Conditions in Linear Models

Proof Below, we prove the necessary and sufficient condition for identifying invalid IV in the linear model.

($\Rightarrow$): According to Theorem 1, we know that if Z is a valid IV relative to $X \rightarrow Y$, then $\{X, Y || Z\}$ always satisfies the AIT condition. This indicates that if $\{X, Y || Z\}$ violates the AIT condition, then Z is an invalid IV. Below, we processed with a proof by contradiction to prove that candidate IV Z violates the *exogeneity* condition. If Z solely violates the *exclusion restriction* condition, then by Proposition 3, $\{X, Y || Z\}$ always satisfies the AIT condition. This contradicts the assumption that $\{X, Y || Z\}$ violates the AIT condition. As a result, candidate IV Z violates the *exogeneity* condition.

($\Leftarrow$): To establish that Assumption 2 holds and that the candidate IV Z is invalid due to a violation of the exogeneity condition, we need to show that $\{X, Y || Z\}$ consequently violates the AIT condition. This conclusion is equivalent to Proposition 2, thereby proving the theorem. ■

A.4 Proof of Proposition 4: Testability of IV in ANINCE Models

Proof If Z violates the IV conditions, the generating mechanism can be described as follows:

$$\begin{aligned} \mathbf{U} &= \varepsilon_{\mathbf{U}}, & Z &= \varphi_Z(\mathbf{U}) + \varepsilon_Z, \\ X &= g(Z) + \varphi_X(\mathbf{U}) + \varepsilon_X, & Y &= f(X, Z) + \varphi_Y(\mathbf{U}) + \varepsilon_Y. \end{aligned} \quad (28)$$

Without loss of generality, let $h(\cdot)$ be a function satisfying $\mathbb{E}[Y - h(X)|Z] = 0$ and $h(\cdot) \neq 0$. According to the definition of the AIT condition, we have

$$\begin{aligned} \mathcal{A}_{X \rightarrow Y|Z} &= Y - h(X) \\ &= \underbrace{f(X, Z) - h(X)}_{\tilde{f}_{bias}(X, Z)} + \varphi_Y(\mathbf{U}) + \varepsilon_Y, \end{aligned} \quad (29)$$

where $\tilde{f}_{bias}(X, Z) = f(X, Z) - h(X)$.

Combining Equations (28) and (29), we have

$$\begin{aligned} \mathcal{A}_{X \rightarrow Y|Z} &= \tilde{f}_{bias}(X, Z) + \varphi_Y(\mathbf{U}) + \varepsilon_Y, \\ Z &= \varphi_Z(\mathbf{U}) + \varepsilon_Z. \end{aligned} \quad (30)$$

Below, we prove this proposition using the linear separability of the logarithm of the joint density of independent variables, which states the fact that for a set of independent random variables whose joint density is twice differentiable, the Hessian of the logarithm of their density is diagonal everywhere from Lin (1997) (See Theorem 7 in Appendix A for further details). According to Assumption 3, we know that the second-order partial derivative $\frac{\partial^2 \log p(\mathcal{A}_{X \rightarrow Y|Z}, Z)}{\partial \mathcal{A}_{X \rightarrow Y|Z} \partial Z} \neq 0$. Furthermore, according to the local geometric information theorem (Theorem 7), we have $\mathcal{A}_{X \rightarrow Y|Z} \not\perp Z$. This implies that $\{X, Y|Z\}$ violates the AIT condition. $\blacksquare$

A.5 Proof of Proposition 5

Below, we provide separate proofs for each of the two cases in Proposition 5.

A.5.1 PROOF OF PROPOSITION 5 (I)

Proof Because candidate IV Z satisfies exogeneity condition, the model of Equation (3) can be written as:

$$\begin{aligned} \mathbf{U} &= \varepsilon_{\mathbf{U}}, & Z &= \varepsilon_Z, \\ X &= g_X(Z) + \varphi_X(\mathbf{U}) + \varepsilon_X, & Y &= f(X) + g_Y(Z) + \varphi_Y(\mathbf{U}) + \varepsilon_Y. \end{aligned} \quad (31)$$

If the direct causal effect of $Z \rightarrow Y$ is a linear function of the direct causal effect of $Z \rightarrow X$, i.e., $g_Y(Z) = a \cdot g_X(Z) + b$, then it is possible to construct a valid IV model that shares the same distribution as the above invalid IV model (31). The model for this valid IV, which has an identical distribution, is as follows:

$$\begin{aligned} \mathbf{U}' &= \mathbf{U}, & Z' &= Z, \\ X' &= X, & Y' &= f'(X') + \varphi_Y(\mathbf{U}) + \varepsilon_Y = Y, \end{aligned} \quad (32)$$

where $f'(X') = f(X) + g_Y(Z)$. According to Equation (32), we know that Z' is a valid IV relative to $X' \rightarrow Y'$.

Based on Theorem 1, the auxiliary variable $\mathcal{A}_{X \rightarrow Y||Z}$ is independent of IV Z . As a result, the joint probability density of $(\mathcal{A}_{X \rightarrow Y||Z}, Z)$ factorizes as the product of the marginal densities:

$$p(\mathcal{A}_{X \rightarrow Y||Z}, Z) = p(\mathcal{A}_{X \rightarrow Y||Z}) \cdot p(Z).$$

According to the local geometric information theorem (Theorem 7), taking the logarithm of both sides and computing the second-order partial derivative yields:

$$\frac{\partial^2 \log p(\mathcal{A}_{X \rightarrow Y||Z}, Z)}{\partial \mathcal{A}_{X \rightarrow Y||Z} \partial Z} = \frac{\partial^2 [\log p(\mathcal{A}_{X \rightarrow Y||Z}) + \log p(Z)]}{\partial \mathcal{A}_{X \rightarrow Y||Z} \partial Z} = 0.$$

This result indicates that Assumption 3 does not hold. ■

A.5.2 PROOF OF PROPOSITION 5 (II)

Proof Due to the assumption of linearity in the model, the model in Equation (3) can be written as:

$$\begin{aligned} \mathbf{U} &= \varepsilon_U, & Z &= \gamma^T \mathbf{U} + \varepsilon_Z, \\ X &= \tau Z + \rho^T \mathbf{U} + \varepsilon_X, & Y &= \beta X + \nu Z + \kappa^T \mathbf{U} + \varepsilon_Y, \end{aligned} \quad (33)$$

where all noise terms are Gaussian.

Let $h(X) = \hat{\beta}X = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)}X$. According to the AIT condition, in the linear model, the function $h(X)$ satisfies the conditional moment restriction, i.e., $\mathbb{E}[Y - h(X) \mid Z] = 0$. Since all noise terms are Gaussian and the model is linear, the vector (X, Y, Z) is jointly Gaussian. Because $\mathcal{A}_{X \rightarrow Y||Z} = Y - \hat{\beta}X$ is a linear combination of (X, Y) , the pair $(\mathcal{A}_{X \rightarrow Y||Z}, Z)$ is also jointly Gaussian.

Hence, we have

$$\text{Cov}(\mathcal{A}_{X \rightarrow Y||Z}, Z) = \text{Cov}(Y - \hat{\beta}X, Z) = \text{Cov}(Y, Z) - \hat{\beta} \text{Cov}(X, Z) = 0. \quad (34)$$

Therefore, $\mathcal{A}_{X \rightarrow Y||Z}$ and Z are uncorrelated. For jointly Gaussian variables, uncorrelatedness implies independence, and thus

$$p(\mathcal{A}_{X \rightarrow Y||Z}, Z) = p(\mathcal{A}_{X \rightarrow Y||Z}) p(Z) \implies \log p(\mathcal{A}_{X \rightarrow Y||Z}, Z) = \log p(\mathcal{A}_{X \rightarrow Y||Z}) + \log p(Z). \quad (35)$$

It follows immediately that

$$\frac{\partial^2}{\partial \mathcal{A}_{X \rightarrow Y||Z} \partial Z} \log p(\mathcal{A}_{X \rightarrow Y||Z}, Z) \equiv 0, \quad (36)$$

which implies that Assumption 3 does not hold. ■

A.6 Proof of Theorem 3: Necessary and Sufficient Conditions for IV in ANINCE Models

Proof Below, we prove the necessary and sufficient conditions for identifying invalid IV in the ANINCE model.

- ($\Rightarrow$): According to Theorem 1, we know that if Z is a valid IV relative to $X \rightarrow Y$, then $\{X, Y \parallel Z\}$ always satisfies the AIT condition.
- ($\Leftarrow$): To establish that Assumptions 1 and 3 hold and that the candidate IV Z is invalid, we need to show that $\{X, Y \parallel Z\}$ consequently violates the AIT condition. This conclusion is equivalent to Proposition 4, thereby proving the theorem. ■

A.7 Proof of Corollary 1: Necessary Condition for IV with Covariates

Proof Below, we apply the same proof technique used in Theorem 1 to demonstrate Corollary 1. Suppose Z is a valid IV relative to $X \rightarrow Y$ given $\mathbf{W}$, the generating mechanism can be expressed as follows:

$$\begin{aligned} \mathbf{U} &= \varepsilon_{\mathbf{U}}, \quad \mathbf{W} = t_{\mathbf{W}}(\mathbf{PA}_{\mathbf{W}}) + \varepsilon_{\mathbf{W}}, \quad Z = t_Z(\mathbf{W}) + \varepsilon_Z, \\ X &= g(\mathbf{W}, Z) + \varphi_X(\mathbf{U}) + \varepsilon_X, \quad Y = f(X, \mathbf{W}) + \varphi_Y(\mathbf{U}) + \varepsilon_Y, \end{aligned} \quad (37)$$

where functions $g(\cdot)$, $f(\cdot)$ and $\varphi_*(\cdot)$ are smooth functions, $\mathbf{PA}_{\mathbf{W}}$ denotes the set of parent variables for each variable in $\mathbf{W}$, and $\mathbf{PA}_{\mathbf{W}} \subseteq \mathbf{W}$. According to Newey and Powell (2003), Singh et al. (2019), Bennett et al. (2019), under the completeness condition and given that Z is a valid IV under the covariates $\mathbf{W}$, the function $h(\cdot)$ that satisfies the conditional moment restriction $\mathbb{E}[Y - h(X) \mid Z, \mathbf{W}] = 0$ is uniquely identified, and coincides with the true causal effect function $f(\cdot)$ of $(X, \mathbf{W})$ on Y , that is, $h(\cdot) = f(\cdot)$.

Thus, we have

$$\mathcal{A}_{X \rightarrow Y \parallel (Z, \mathbf{W})} = Y - h(X, \mathbf{W}) = Y - f(X, \mathbf{W}) = \varphi_Y(\mathbf{U}) + \varepsilon_Y. \quad (38)$$

Let $\mathcal{Z}$ denote the residual from the regression Z on $\mathbf{W}$. Thus, $\mathcal{Z} = \varepsilon_Z$.

By Theorem 2.2.5 and its extension in Meester (2008), if random variables are mutually independent, then any measurable functions applied to disjoint subsets of them yield independent random variables (see Theorem 5 and Corollary 3 in Appendix A for further details). Based on this result, we next show that the auxiliary variable $\mathcal{A}_{X \rightarrow Y \parallel (Z, \mathbf{W})}$ and $\mathcal{Z}$ are statistically independent. Specifically, since the noise terms ε_Z , ε_Y , and $\varepsilon_{\mathbf{U}}$ are mutually independent, we can obtain that ε_Z is also independent of $\varphi_Y(\varepsilon_{\mathbf{U}}) + \varepsilon_Y$. Furthermore, combining the equations $\mathbf{U} = \varepsilon_{\mathbf{U}}$ and $\mathcal{Z} = \varepsilon_Z$, we conclude that $\mathcal{Z}$ is independent of $\varphi_Y(\mathbf{U}) + \varepsilon_Y$. Hence, the auxiliary variable $\mathcal{A}_{X \rightarrow Y \parallel (Z, \mathbf{W})}$ and $\mathcal{Z}$ are statistically independent, i.e., $\mathcal{A}_{X \rightarrow Y \parallel (Z, \mathbf{W})} \perp\!\!\!\perp \mathcal{Z}$. Consequently, $\{X, Y \parallel (Z, \mathbf{W})\}$ satisfies the AIT condition. ■

A.8 Proof of Theorem 4: Asymptotic Level and Power of the AIT Test

Proof Let $h(\cdot)$ denote the oracle function and $\hat{h}(\cdot)$ its estimator trained on the dataset $\mathcal{D}_1 \subset \mathcal{D}$. We establish our theoretical results using an independent dataset $\mathcal{D}_2 = \{(X_i, Y_i, Z_i)\}_{i=1}^m \subset \mathcal{D}$. For notational simplicity, we below use (X, Y, Z) to denote a generic observation drawn from $\mathcal{D}_2$. Recall that $\mathcal{A} = Y - h(X)$ and $\hat{\mathcal{A}} = Y - \hat{h}(X)$. In the absence of covariates, we simply have $\mathcal{Z} = Z$. To establish the asymptotic level and power of the AIT test, we analyze the difference between $\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z})$ and its oracle counterpart $\widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})$, where $\widehat{\text{HSIC}}(\cdot, \cdot)$ denotes the block-based estimator of Zhang et al. (2018). Specifically,

$$\begin{aligned}
 & \left| \sqrt{mB} \widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) - \sqrt{mB} \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) \right| \\
 &= \left| \sqrt{mB} \cdot \frac{B}{m} \sum_{b=1}^{m/B} \hat{\eta}_b(\hat{\mathcal{A}}) - \sqrt{mB} \cdot \frac{B}{m} \sum_{b=1}^{m/B} \hat{\eta}_b(\mathcal{A}) \right| \\
 &= \sqrt{mB} \cdot \frac{B}{m} \left| \sum_{b=1}^{m/B} [\hat{\eta}_b(\hat{\mathcal{A}}) - \hat{\eta}_b(\mathcal{A})] \right| \quad (\text{by the triangle inequality}) \\
 &\leq \sqrt{mB} \cdot \frac{B}{m} \sum_{b=1}^{m/B} \left| [\hat{\eta}_b(\hat{\mathcal{A}}) - \hat{\eta}_b(\mathcal{A})] \right| \\
 &= \sqrt{mB} \cdot \frac{B}{m} \sum_{b=1}^{m/B} \cdot \left| \underbrace{\left[\frac{1}{B(B-3)} \left[\text{tr} \left(\tilde{K}^{(b)} \tilde{L}^{(b)} \right) - \text{tr} \left(\tilde{K}^{(b)} \tilde{L}^{(b)} \right) \right] \right]}_{\text{Part I}} \right. \\
 &\quad + \underbrace{\frac{1}{B(B-3)} \left[\frac{\mathbb{1}^T \tilde{K}^{(b)} \mathbb{1} \mathbb{1}^T \tilde{L}^{(b)} \mathbb{1}}{(B-1)(B-2)} - \frac{\mathbb{1}^T \tilde{K}^{(b)} \mathbb{1} \mathbb{1}^T \tilde{L}^{(b)} \mathbb{1}}{(B-1)(B-2)} \right]}_{\text{Part II}} \\
 &\quad \left. - \underbrace{\frac{1}{B(B-3)} \left[\frac{2}{B-2} \mathbb{1}^T \tilde{K}^{(b)} \tilde{L}^{(b)} \mathbb{1} - \frac{2}{B-2} \mathbb{1}^T \tilde{K}^{(b)} \tilde{L}^{(b)} \mathbb{1} \right]}_{\text{Part III}} \right|, \tag{39}
 \end{aligned}$$

where $\tilde{K}^{(b)}$ denotes the kernel matrix for the estimated auxiliary variables $\hat{\mathcal{A}}$ in the b -th block.

Below, we prove the asymptotic level and power of the AIT condition test, respectively.

1. Proof of the asymptotic level of the AIT condition.

To establish the asymptotic level of the AIT test under the null hypothesis, we proceed in three steps. First, we derive a bound on the estimation error within a single block. Second, we extend this bound to the full HSIC estimator. Finally, we study the asymptotic behavior of the resulting test statistic.

- **Step 1: Single-block Error Bound for $|\hat{\eta}_b(\hat{\mathcal{A}}) - \hat{\eta}_b(\mathcal{A})|$.** For a fixed block b (omitting the superscript (b) for brevity), let $l_{ij} = l(\mathcal{Z}_i, \mathcal{Z}_j)$. We decompose the difference into the three components appearing in (39) (Parts I–III) and bound them separately.

(1) **Part I in Equation (39):**

$$\begin{aligned} & \frac{1}{B(B-3)} \operatorname{tr}(\tilde{K}\tilde{L}) - \frac{1}{B(B-3)} \operatorname{tr}(\tilde{K}\tilde{L}) \\ &= \frac{1}{B(B-3)} \left[\sum_{i,j:i \neq j} \left(k(\hat{\mathcal{A}}_i, \hat{\mathcal{A}}_j) - k(\mathcal{A}_i, \mathcal{A}_j) \right) l_{ij} \right] \end{aligned}$$

Taking absolute value of the above expression, and by the boundedness and Lipschitz continuity (Assumption 4), we have

$$\begin{aligned} & \left| \frac{1}{B(B-3)} \left[\sum_{i,j:i \neq j} \left(k(\hat{\mathcal{A}}_i, \hat{\mathcal{A}}_j) - k(\mathcal{A}_i, \mathcal{A}_j) \right) l_{ij} \right] \right| \\ & \leq \frac{M}{B(B-3)} \sum_{i,j:i \neq j} \left| \left(k(\hat{\mathcal{A}}_i, \hat{\mathcal{A}}_j) - k(\mathcal{A}_i, \mathcal{A}_j) \right) \right| \quad (\text{by Assumption 4 (1)}) \\ & \leq \frac{MQ}{B(B-3)} \sum_{i,j:i \neq j} \left(\|\hat{\mathcal{A}}_i - \mathcal{A}_i\| + \|\hat{\mathcal{A}}_j - \mathcal{A}_j\| \right) \quad (\text{by Assumption 4 (2)}) \quad (40) \\ & \leq \frac{MQ}{B(B-3)} \left[(B-1) \cdot \sum_i \|\hat{\mathcal{A}}_i - \mathcal{A}_i\| + (B-1) \cdot \sum_j \|\hat{\mathcal{A}}_j - \mathcal{A}_j\| \right] \\ & \leq \frac{2MQ(B-1)}{B(B-3)} \sum_i \|\hat{\mathcal{A}}_i - \mathcal{A}_i\|. \end{aligned}$$

(2) **Part II in Equation (39):**

$$\begin{aligned} & \frac{1}{B(B-3)} \frac{\mathbb{1}^T \tilde{K} \mathbb{1} \mathbb{1}^T \tilde{L} \mathbb{1}}{(B-1)(B-2)} - \frac{1}{B(B-3)} \frac{\mathbb{1}^T \tilde{K} \mathbb{1} \mathbb{1}^T \tilde{L} \mathbb{1}}{(B-1)(B-2)} \\ &= \frac{1}{B(B-3)(B-1)(B-2)} \sum_{i,j:i \neq j} \left[k(\hat{\mathcal{A}}_i, \hat{\mathcal{A}}_j) - k(\mathcal{A}_i, \mathcal{A}_j) \right] \cdot \sum_{q,r:q \neq r} l_{qr}. \end{aligned}$$

Taking absolute value of the above expression, and by the boundedness and Lipschitz continuity (Assumption 4), we have

$$\begin{aligned} & \left| \frac{1}{B(B-3)(B-1)(B-2)} \sum_{i,j:i \neq j} \left[k(\hat{\mathcal{A}}_i, \hat{\mathcal{A}}_j) - k(\mathcal{A}_i, \mathcal{A}_j) \right] \sum_{q,r:q \neq r} l_{qr} \right| \\ & \leq \frac{MB(B-1)}{B(B-3)(B-1)(B-2)} \sum_{i,j:i \neq j} \left| k(\hat{\mathcal{A}}_i, \hat{\mathcal{A}}_j) - k(\mathcal{A}_i, \mathcal{A}_j) \right| \quad (\text{by Assumption 4 (1)}) \\ & \leq \frac{MQ(B-1)}{(B-3)(B-2)} \left(\sum_i \|\hat{\mathcal{A}}_i - \mathcal{A}_i\| + \sum_j \|\hat{\mathcal{A}}_j - \mathcal{A}_j\| \right) \quad (\text{by Assumption 4 (2)}) \\ & = \frac{2MQ(B-1)}{(B-3)(B-2)} \sum_i \|\hat{\mathcal{A}}_i - \mathcal{A}_i\|. \end{aligned} \quad (41)$$

(3) **Part III in Equation (39):**

$$\begin{aligned} & \frac{-1}{B(B-3)} \frac{2}{B-2} \mathbb{1}^T \tilde{K} \tilde{L} \mathbb{1} + \frac{1}{B(B-3)} \frac{2}{B-2} \mathbb{1}^T \tilde{K} \tilde{L} \mathbb{1} \\ &= \frac{-2}{B(B-3)(B-2)} \sum_{i,j,r:r \neq i, j \neq i} \left[k(\hat{\mathcal{A}}_i, \hat{\mathcal{A}}_j) - k(\mathcal{A}_i, \mathcal{A}_j) \right] l_{ir}. \end{aligned}$$

Taking absolute value of the above expression, and by the boundedness and Lipschitz continuity (Assumption 4), we have

$$\begin{aligned} & \left| \frac{-2}{B(B-3)(B-2)} \sum_{i,j,r:r \neq i, j \neq i} \left[k(\hat{\mathcal{A}}_i, \hat{\mathcal{A}}_j) - k(\mathcal{A}_i, \mathcal{A}_j) \right] l_{ir} \right| \\ & \leq \frac{2}{B(B-3)(B-2)} \sum_i \sum_{r:r \neq i} \left| k(\hat{\mathcal{A}}_i, \hat{\mathcal{A}}_r) - k(\mathcal{A}_i, \mathcal{A}_r) \right| \cdot \sum_{j:j \neq i} |l_{ji}| \\ & \leq \frac{2M(B-1)}{B(B-3)(B-2)} \sum_i \sum_{r:r \neq i} \left| k(\hat{\mathcal{A}}_i, \hat{\mathcal{A}}_r) - k(\mathcal{A}_i, \mathcal{A}_r) \right| \quad (\text{by Assumption 4 (1)}) \\ & \leq \frac{2MQ(B-1)}{B(B-3)(B-2)} \sum_i \left(\sum_{r:r \neq i} \left\| \hat{\mathcal{A}}_r - \mathcal{A}_r \right\| + (B-1) \left\| \hat{\mathcal{A}}_i - \mathcal{A}_i \right\| \right) \quad (\text{by Assumption 4 (2)}) \\ & \leq \frac{4MQ(B-1)^2}{B(B-3)(B-2)} \sum_i \left\| \hat{\mathcal{A}}_i - \mathcal{A}_i \right\|. \end{aligned} \tag{42}$$

Combining the bounds derived in (40)–(42), we obtain

$$|\hat{\eta}_b(\hat{\mathcal{A}}) - \hat{\eta}_b(\mathcal{A})| = \frac{C(B)}{B} \sum_{i \in \mathcal{I}_b} \left\| \hat{\mathcal{A}}_i - \mathcal{A}_i \right\|,$$

where the constant $C(B) := \frac{2MQ(B-1)}{(B-3)} \left[1 + \frac{B}{B-2} + \frac{2(B-1)}{B-2} \right]$, and $\mathcal{I}_b$ denotes the index set of the b -th block with $|\mathcal{I}_b| = B$.

- **Step 2: Overall HSIC Error Bound** $|\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) - \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})|$. Recall that the block-based HSIC estimator is defined as $\widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) = \frac{B}{m} \sum_{b=1}^{m/B} \hat{\eta}_b(\mathcal{A})$. Hence, summing over all m/B blocks:

$$\begin{aligned} |\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) - \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})| &= \frac{B}{m} \sum_{b=1}^{m/B} |\hat{\eta}_b(\hat{\mathcal{A}}) - \hat{\eta}_b(\mathcal{A})| \\ &\leq \frac{B}{m} \sum_{b=1}^{m/B} \frac{C(B)}{B} \sum_{i \in \mathcal{I}_b} \left\| \hat{\mathcal{A}}_i - \mathcal{A}_i \right\| \leq \frac{C(B)}{m} \sum_{i=1}^m \left\| \hat{\mathcal{A}}_i - \mathcal{A}_i \right\| \end{aligned}$$

Substituting the auxiliary variable $\mathcal{A} = Y - h(X)$, we obtain $\hat{\mathcal{A}}_i - \mathcal{A}_i = h(X_i) - \hat{h}(X_i)$. Thus,

$$|\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) - \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})| \leq \frac{C(B)}{m} \sum_{i=1}^m \|h(X_i) - \hat{h}(X_i)\|.$$

By the law of large numbers, $\frac{1}{m} \sum_{i=1}^m \|h(X_i) - \hat{h}(X_i)\| \xrightarrow{p} \mathbb{E}[\|h(X_i) - \hat{h}(X_i)\|]$. Moreover, since $\hat{h}$ is estimated using another independent dataset $\mathcal{D}_1$ of sample size n , and is $\sqrt{n}$ -consistent by Assumption 5, we have $\|h - \hat{h}\|_{L_2(P_X)} = O_p(n^{-1/2})$. Combining the above results yields

$$|\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) - \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})| = O_p(n^{-1/2}). \quad (43)$$

- **Step 3: Asymptotic Validity $\sqrt{mB} \widehat{\text{HSIC}}$.** By the assumption that the sample sizes satisfy $mB = o(n)$ as $n \rightarrow \infty$, together with Equation (43), Equation (39) can be expressed as follows:

$$|\sqrt{mB} \widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) - \sqrt{mB} \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})| = \sqrt{mB} \cdot O_p(n^{-1/2}) = o_p(1).$$

Hence,

$$\sqrt{mB} \widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) = \sqrt{mB} \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) + o_p(1).$$

Under H_0 , let $c_{m,\alpha}$ be a critical value such that the oracle statistic satisfies

$$P_{H_0}(\sqrt{mB} \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) > c_{m,\alpha}) \leq \alpha + o(1).$$

By Slutsky's theorem,

$$P_{H_0}(\sqrt{mB} \widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) > c_{m,\alpha}) = P_{H_0}(\sqrt{mB} \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) + o_p(1) > c_{m,\alpha}) \leq \alpha + o(1).$$

Consequently,

$$P_{H_0}(\text{Type I error}) = P_{H_0}(\sqrt{mB} \widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) > c_{m,\alpha}) \leq \alpha + o(1).$$

2. Proof of the asymptotic power of the AIT condition.

Under the fixed alternative H_1 , assume that $\text{HSIC}(\mathcal{A}, \mathcal{Z}) = \eta > 0$. Recall that the block-based HSIC estimator can be written as

$$\widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) = \frac{1}{M} \sum_{b=1}^M \hat{\eta}_b, \quad M = m/B.$$

Under H_1 , the block statistics $\{\hat{\eta}_b\}_{b=1}^M$ are i.i.d. with $\mathbb{E}[\hat{\eta}_b] = \eta > 0$ and finite variance $\text{Var}(\hat{\eta}_b) = \tau^2 < \infty$. By the classical central limit theorem,

$$\sqrt{M} (\widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) - \eta) \xrightarrow{D} \mathcal{N}(0, \tau^2).$$

Since $M = m/B$, multiplying both sides by B , this is equivalent to

$$\sqrt{mB} (\widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) - \eta) \xrightarrow{D} \mathcal{N}(0, \sigma_{H_1}^2),$$

where $\sigma_{H_1}^2 = \tau^2 B^2$.

By the plug-in error bound established in Equation (43), $\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) = \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) + O_p(n^{-1/2})$, and under $mB = o(n)$,

$$\sqrt{mB} \left(\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) - \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) \right) = o_p(1).$$

By Slutsky's theorem,

$$\sqrt{mB} \left(\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) - \eta \right) \xrightarrow{D} \mathcal{N}(0, \sigma_{H_1}^2).$$

The Type II error probability is

$$P_{H_1}(\text{Type II error}) = P_{H_1} \left(\sqrt{mB} \widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) \leq c_{m,\alpha} \right),$$

where $c_{m,\alpha}$ is defined as the $(1 - \alpha)$ -quantile of the null limiting distribution. Subtracting $\sqrt{mB}\eta$ and standardizing yields

$$P_{H_1} \left(\frac{\sqrt{mB} \left(\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \mathcal{Z}) - \eta \right)}{\sigma_{H_1}} \leq \frac{c_{m,\alpha} - \sqrt{mB}\eta}{\sigma_{H_1}} \right) := P_{H_1}(T_{m,B} \leq \Delta_{m,B}),$$

where $T_{m,B} \xrightarrow{D} \mathcal{N}(0, 1)$. Since the null distribution is non-degenerate, its quantiles are finite, and hence $c_{m,\alpha} = O_p(1)$. Moreover, because $\eta > 0$,

$$\Delta_{m,B} = -\frac{\eta}{\sigma_{H_1}} \sqrt{mB} + O(1) \rightarrow -\infty.$$

Hence,

$$P_{H_1}(\text{Type II error}) = \Phi(\Delta_{m,B}) + o(1), \text{ as } mB \rightarrow \infty.$$

Applying Mills' ratio for $\Delta_{m,B} < 0$ gives

$$\Phi(\Delta_{m,B}) \leq \frac{1}{\sqrt{2\pi}(-\Delta_{m,B})} \exp\left(-\frac{\Delta_{m,B}^2}{2}\right) = O\left(\frac{1}{\sqrt{mB}} e^{-cmB}\right),$$

where $c = \eta^2/(2\sigma_{H_1}^2) > 0$. Therefore,

$$P_{H_1}(\text{Type II error}) = O\left(\frac{1}{\sqrt{mB}} e^{-cmB}\right).$$

■

A.9 Proof of Corollary 2: Asymptotic Level and Power of the AIT Test with Covariates

Proof Let $h(\cdot)$ and $\pi(\cdot)$ denote the oracle functions, and let $\hat{h}(\cdot)$ and $\hat{\pi}(\cdot)$ denote their estimators trained on the dataset $\mathcal{D}_1 \subset \mathcal{D}$. We establish our theoretical results using an independent dataset $\mathcal{D}_2 = \{(X_i, Y_i, \mathbf{W}_i, Z_i)\}_{i=1}^m \subset \mathcal{D}$. For notational simplicity, below we use

$(X, Y, \mathbf{W}, Z)$ to denote a generic observation drawn from $\mathcal{D}_2$. Recall that $\mathcal{A} = Y - h(X)$ and $\hat{\mathcal{A}} = Y - \hat{h}(X)$, and that $\mathcal{Z} = Z - \pi(\mathbf{W})$ and $\hat{\mathcal{Z}} = Z - \hat{\pi}(\mathbf{W})$. To establish the asymptotic level and power of the AIT test, we analyze the difference between $\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}})$ and its oracle counterpart $\widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})$, where $\widehat{\text{HSIC}}(\cdot, \cdot)$ denotes the block-based estimator of Zhang et al. (2018), i.e., $|\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) - \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})|$.

Applying the triangle inequality, we have

$$\begin{aligned} & |\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) - \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})| \\ & \leq \underbrace{|\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) - \widehat{\text{HSIC}}(\mathcal{A}, \hat{\mathcal{Z}})|}_{\text{Part I: error induced by } \hat{\mathcal{A}}} + \underbrace{|\widehat{\text{HSIC}}(\mathcal{A}, \hat{\mathcal{Z}}) - \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})|}_{\text{Part II: error induced by } \hat{\mathcal{Z}}}. \end{aligned} \quad (44)$$

We next show that, by arguments analogous to those in the proof of Theorem 4, the two terms can be bounded separately. Specifically, by Assumptions 6 and 7, we have

$$|\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) - \widehat{\text{HSIC}}(\mathcal{A}, \hat{\mathcal{Z}})| = O_p(n^{-1/2}).$$

Moreover, by Assumptions 6 and 8, it holds that

$$|\widehat{\text{HSIC}}(\mathcal{A}, \hat{\mathcal{Z}}) - \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})| = O_p(n^{-q}), \quad q \in (0, 1/2).$$

Combining the above two results, we obtain

$$|\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) - \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})| = O_p(n^{-1/2}) + O_p(n^{-q}), \quad q \in (0, 1/2).$$

Since n^{-q} dominates $n^{-1/2}$ for $q < 1/2$, the overall convergence rate is $O_p(n^{-q})$, i.e., $|\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) - \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})| = O_p(n^{-q})$.

By the assumption that the sample sizes satisfy $mB = o(n^{2q})$, which ensures that:

$$|\sqrt{mB} \widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) - \sqrt{mB} \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z})| = o_p(1). \quad (45)$$

Below, we prove the asymptotic level and power of the AIT condition test, respectively.

1. Proof of the asymptotic level of the AIT condition with covariates.

Under H_0 , let $c_{m,\alpha}$ be a critical value such that the oracle statistic satisfies

$$P_{H_0}(\sqrt{mB} \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) > c_{m,\alpha}) \leq \alpha + o(1).$$

By Slutsky's theorem, we further have

$$P_{H_0}(\sqrt{mB} \widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) > c_{m,\alpha}) = P_{H_0}(\sqrt{mB} \widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) + o_p(1) > c_{m,\alpha}) \leq \alpha + o(1).$$

Consequently,

$$P_{H_0}(\text{Type I error}) = P_{H_0}(\sqrt{mB} \widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) > c_{m,\alpha}) \leq \alpha + o(1).$$

2. Proof of the asymptotic power of the AIT condition with covariates.

We proceed analogously to the proof of Theorem 4. In particular, applying the same block-averaging and a central limit theorem argument to the oracle variables $(\mathcal{A}, \mathcal{Z})$ yields

$$\sqrt{mB} (\widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) - \eta) \xrightarrow{D} \mathcal{N}(0, \sigma_{H_1}^2), \quad \sigma_{H_1}^2 = \tau^2 B^2,$$

where $\eta = \text{HSIC}(\mathcal{A}, \mathcal{Z}) > 0$ and $\tau^2 = \text{Var}(\hat{\eta}_b) < \infty$.

According to Equation (45), we have

$$\sqrt{mB} \left(\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) - \eta \right) = \sqrt{mB} \left(\widehat{\text{HSIC}}(\mathcal{A}, \mathcal{Z}) - \eta \right) + o_p(1).$$

By Slutsky's theorem,

$$\sqrt{mB} \left(\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) - \eta \right) \xrightarrow{D} \mathcal{N}(0, \sigma_{H_1}^2).$$

The Type II error probability is

$$P_{H_1}(\text{Type II error}) = P_{H_1} \left(\sqrt{mB} \widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) \leq c_{m,\alpha} \right),$$

where $c_{m,\alpha}$ is defined as the $(1 - \alpha)$ -quantile of the null limiting distribution. Subtracting $\sqrt{mB}\eta$ and standardizing yields

$$P_{H_1} \left(\frac{\sqrt{mB} \left(\widehat{\text{HSIC}}(\hat{\mathcal{A}}, \hat{\mathcal{Z}}) - \eta \right)}{\sigma_{H_1}} \leq \frac{c_{m,\alpha} - \sqrt{mB}\eta}{\sigma_{H_1}} \right) := P_{H_1}(T_{m,B} \leq \Delta_{m,B}),$$

where $T_{m,B} \xrightarrow{D} \mathcal{N}(0, 1)$. Since the null distribution is non-degenerate, its quantiles are finite, and hence $c_{m,\alpha} = O_p(1)$. Moreover, because $\eta > 0$, we have

$$\Delta_{m,B} = -\frac{\eta}{\sigma_{H_1}} \sqrt{mB} + O_p(1) \rightarrow -\infty.$$

Hence,

$$P_{H_1}(\text{Type II error}) = \Phi(\Delta_{m,B}) + o(1), \text{ as } mB \rightarrow \infty.$$

Applying Mills' ratio for $\Delta_{m,B} < 0$ gives

$$\Phi(\Delta_{m,B}) \leq \frac{1}{\sqrt{2\pi}(-\Delta_{m,B})} \exp\left(-\frac{\Delta_{m,B}^2}{2}\right) = O\left(\frac{1}{\sqrt{mB}} e^{-cmB}\right),$$

where $c = \eta^2/(2\sigma_{H_1}^2) > 0$. Therefore,

$$P_{H_1}(\text{Type II error}) = O\left(\frac{1}{\sqrt{mB}} e^{-cmB}\right).$$

■

Appendix B. Two Analytic Examples Satisfying Assumption 3

In this section, we provide two fully analytic data-generating processes and verify Assumption 3 in a straightforward manner. Our verification is based on the following observation: since $p(\mathcal{A}, Z) = p(\mathcal{A} \mid Z)p(Z)$, we have

$$\log p(\mathcal{A}, Z) = \log p(\mathcal{A} \mid Z) + \log p(Z), \quad \Rightarrow \quad \frac{\partial^2}{\partial \mathcal{A} \partial Z} \log p(\mathcal{A}, Z) = \frac{\partial^2}{\partial \mathcal{A} \partial Z} \log p(\mathcal{A} \mid Z),$$

because $\log p(Z)$ does not depend on $\mathcal{A}$. Thus, it suffices to show that the mixed derivative of $\log p(\mathcal{A} \mid Z)$ is nonzero on a set with non-zero Lebesgue measure.

Throughout, we define the auxiliary variable for the constant effect model as

$$\mathcal{A}_{X \rightarrow Y \parallel Z} := Y - \hat{\beta}X, \quad \hat{\beta} := \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)}.$$

B.1 Example 2: Violation of the Exogeneity Condition ($Z \not\perp U$)

Consider the nonlinear Gaussian noise model

$$\begin{aligned} U &= \varepsilon_U, & Z &= \gamma U + \varepsilon_Z, \\ X &= \exp(Z) + \rho U + \varepsilon_X, & Y &= \beta X + \kappa U + \varepsilon_Y, \end{aligned} \tag{46}$$

where $(\varepsilon_U, \varepsilon_Z, \varepsilon_X, \varepsilon_Y) \stackrel{\text{ind}}{\sim} \mathcal{N}(0, 1)$. Since $Z = \gamma U + \varepsilon_Z$, we have $Z \not\perp U$ whenever $\gamma \neq 0$, i.e., exogeneity is violated.

Step 1: Compute $\hat{\beta} = \text{Cov}(Y, Z) / \text{Cov}(X, Z)$.

Let $\sigma_Z^2 := \text{Var}(Z) = \gamma^2 + 1$. For $Z \sim \mathcal{N}(0, \sigma_Z^2)$, the moment identity gives

$$\text{Cov}(\exp(Z), Z) = \mathbb{E}[Z e^Z] = \left. \frac{d}{dt} \mathbb{E}[e^{tZ}] \right|_{t=1} = \left. \frac{d}{dt} \exp\left(\frac{1}{2} \sigma_Z^2 t^2\right) \right|_{t=1} = \sigma_Z^2 \exp\left(\frac{1}{2} \sigma_Z^2\right). \tag{47}$$

Moreover, $\text{Cov}(U, Z) = \text{Cov}(U, \gamma U + \varepsilon_Z) = \gamma$. Hence

$$\text{Cov}(X, Z) = \text{Cov}(\exp(Z), Z) + \rho \text{Cov}(U, Z) = \sigma_Z^2 \exp(\sigma_Z^2/2) + \rho\gamma.$$

Since $Y = \beta X + \kappa U + \varepsilon_Y$ and $\varepsilon_Y \perp Z$,

$$\text{Cov}(Y, Z) = \beta \text{Cov}(X, Z) + \kappa \text{Cov}(U, Z) = \beta(\sigma_Z^2 \exp(\sigma_Z^2/2) + \rho\gamma) + \kappa\gamma.$$

Therefore,

$$\hat{\beta} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} = \beta + \underbrace{\frac{\kappa\gamma}{\sigma_Z^2 \exp(\sigma_Z^2/2) + \rho\gamma}}_{\beta_{\text{bias}}}. \tag{48}$$

Step 2: Derive $\mathcal{A}$ and $p(\mathcal{A} \mid Z)$.

Using $Y = \beta X + \kappa U + \varepsilon_Y$,

$$\begin{aligned} \mathcal{A} &= Y - \hat{\beta}X \\ &= (\beta - \hat{\beta})X + \kappa U + \varepsilon_Y \\ &= (\beta - \hat{\beta}) \exp(Z) + ((\beta - \hat{\beta})\rho + \kappa)U + (\beta - \hat{\beta})\varepsilon_X + \varepsilon_Y. \end{aligned}$$

Conditional on $Z = z$, the random variable $U \mid Z = z$ is Gaussian since (U, Z) is jointly Gaussian. Specifically,

$$\mathbb{E}[U \mid Z = z] = \frac{\text{Cov}(U, Z)}{\text{Var}(Z)} z = \frac{\gamma}{\sigma_Z^2} z, \quad \text{Var}(U \mid Z = z) = 1 - \frac{\text{Cov}(U, Z)^2}{\text{Var}(Z)} = \frac{1}{\sigma_Z^2}.$$

Since $(\varepsilon_X, \varepsilon_Y) \perp\!\!\!\perp Z$ and are Gaussian, it follows that $\mathcal{A} \mid Z = z$ is Gaussian with mean $m(z)$ and variance v given by

$$\begin{aligned} m(z) &= (\beta - \hat{\beta}) \exp(z) + ((\beta - \hat{\beta})\rho + \kappa) \frac{\gamma}{\sigma_Z^2} z, \\ v &= ((\beta - \hat{\beta})\rho + \kappa)^2 \frac{1}{\sigma_Z^2} + (\beta - \hat{\beta})^2 + 1. \end{aligned} \tag{49}$$

In particular, $v > 0$ is constant (independent of z).

Step 3: Verify Assumption 3.

Since $\mathcal{A} \mid Z = z \sim \mathcal{N}(m(z), v)$,

$$\log p(\mathcal{A} \mid Z = z) = -\frac{(\mathcal{A} - m(z))^2}{2v} - \frac{1}{2} \log(2\pi v).$$

Hence

$$\frac{\partial}{\partial \mathcal{A}} \log p(\mathcal{A} \mid Z = z) = -\frac{\mathcal{A} - m(z)}{v}, \quad \frac{\partial^2}{\partial \mathcal{A} \partial z} \log p(\mathcal{A} \mid Z = z) = \frac{m'(z)}{v}.$$

Using Equation (49), we have the derivative

$$m'(z) = (\beta - \hat{\beta}) \exp(z) + ((\beta - \hat{\beta})\rho + \kappa) \frac{\gamma}{\sigma_Z^2}. \tag{50}$$

Therefore, combining Equations (49) and (50) yields

$$\frac{\partial^2}{\partial \mathcal{A} \partial Z} \log p(\mathcal{A}, Z) = \frac{\partial^2}{\partial \mathcal{A} \partial Z} \log p(\mathcal{A} \mid Z) = \frac{m'(Z)}{v} = \frac{\kappa\gamma - \beta_{bias}[\rho\gamma + \exp(Z)(\gamma^2 + 1)]}{(\kappa - \beta_{bias}\rho)^2 + (\gamma^2 + 1)(\beta_{bias}^2 + 1)}. \tag{51}$$

Since $m'(z)$ is not identically zero (which holds generically; e.g., if $\kappa\gamma - \beta_{bias}[\rho\gamma + \exp(z)(\gamma^2 + 1)] \neq 0$), the set $\{z : m'(z) \neq 0\}$ has non-zero Lebesgue measure. Consequently, Equation (51) is nonzero on a set with non-zero Lebesgue measure, and Assumption 3 holds.

B.2 Example 3: Violation of the Exclusion Restriction Condition ($Z \rightarrow Y$ Direct Effect)

Consider the model

$$\begin{aligned} U &= \varepsilon_U, & Z &= \varepsilon_Z, \\ X &= \exp(Z) + \rho U + \varepsilon_X, & Y &= \beta X + \nu Z + \kappa U + \varepsilon_Y, \end{aligned} \tag{52}$$

where $(\varepsilon_U, \varepsilon_Z, \varepsilon_X, \varepsilon_Y) \stackrel{\text{ind}}{\sim} \mathcal{N}(0, 1)$. Here $Z \perp\!\!\!\perp U$ (exogeneity holds), but exclusion is violated when $\nu \neq 0$ due to the direct effect νZ .

Step 1: Compute $\hat{\beta} = \text{Cov}(Y, Z) / \text{Cov}(X, Z)$.

Since $Z \sim \mathcal{N}(0, 1)$, by Equation (47) with $\sigma_Z^2 = 1$,

$$\text{Cov}(\exp(Z), Z) = \exp(1/2).$$

Moreover, $U \perp\!\!\!\perp Z$ and $\varepsilon_X \perp\!\!\!\perp Z$, hence

$$\text{Cov}(X, Z) = \text{Cov}(\exp(Z), Z) = \exp(1/2).$$

Since $Y = \beta X + \nu Z + \kappa U + \varepsilon_Y$ and $(U, \varepsilon_Y) \perp\!\!\!\perp Z$,

$$\text{Cov}(Y, Z) = \beta \text{Cov}(X, Z) + \nu \text{Var}(Z) = \beta \exp(1/2) + \nu.$$

Therefore,

$$\hat{\beta} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} = \beta + \underbrace{\nu \exp(-1/2)}_{\beta_{\text{bias}}}. \quad (53)$$

Step 2: Derive $\mathcal{A}$ and $p(\mathcal{A} | Z)$.

We have

$$\begin{aligned} \mathcal{A} &= Y - \hat{\beta}X \\ &= (\beta - \hat{\beta})X + \nu Z + \kappa U + \varepsilon_Y \\ &= (\beta - \hat{\beta}) \exp(Z) + \nu Z + ((\beta - \hat{\beta})\rho + \kappa)U + (\beta - \hat{\beta})\varepsilon_X + \varepsilon_Y. \end{aligned}$$

Conditional on $Z = z$, U remains $\mathcal{N}(0, 1)$ and is independent of Z . Thus $\mathcal{A} | Z = z$ is Gaussian with mean and variance

$$\begin{aligned} m(z) &= (\beta - \hat{\beta}) \exp(z) + \nu z, \\ v &= ((\beta - \hat{\beta})\rho + \kappa)^2 + (\beta - \hat{\beta})^2 + 1, \end{aligned} \quad (54)$$

where $v > 0$ is constant.

Step 3: Verify Assumption 3.

As in Example B.1, using Equations (53) and (54), we obtain $m'(z) = (\beta - \hat{\beta}) \exp(z) + \nu$, which implies

$$\frac{\partial^2}{\partial \mathcal{A} \partial Z} \log p(\mathcal{A}, Z) = \frac{\partial^2}{\partial \mathcal{A} \partial Z} \log p(\mathcal{A} | Z) = \frac{m'(Z)}{v} = \frac{\nu[1 - \exp(Z - 1/2)]}{[\kappa - \nu \exp(\frac{-1}{2})]^2 + \nu^2 \exp(-1) + 1}. \quad (55)$$

Since $\nu \neq 0$ (i.e., exclusion is violated), and $m'(z) \neq 0$ for all $z \neq 1/2$, $\{z : m'(z) \neq 0\}$ has non-zero Lebesgue measure. Therefore, the mixed second derivative (Equation (55)) is nonzero on a set with non-zero Lebesgue measure, and Assumption 3 holds.

Appendix C. More Details on Simulation Experiments in Section 5

In this section, we provide details of the simulation experiments corresponding to Tables 3 ~ 11. Specifically, the generation mechanism for each table is as follows:

- Table 3. The model setup is as follows: $U = \varepsilon_U$, $Z_1 = U + \varepsilon_{Z1}$, $Z_2 = \varepsilon_{Z2}$, $X = \tau_1 Z_1 + \tau_2 Z_2 + \rho U + \varepsilon_X$, $Y = X + \kappa U + \varepsilon_Y$, where all constant coefficients are randomly selected from a uniform distribution with parameters $\min = 0.5$ and $\max = 1.5$. The noise terms $\varepsilon_U, \varepsilon_{Z1}, \varepsilon_{Z2}, \varepsilon_X$, and ε_Y follow the specific distributions listed in each row. The final row indicates that all noise terms are randomly drawn from one of six distributions.⁶

6. These six distributions include Uniform, Beta, T, Gamma, Lognormal, and Gaussian.

- Table 4. The model setup is as follows: $U = \varepsilon_U$, $Z_1 = \varphi_{Z1}(U) + \varepsilon_{Z1}$, $Z_2 = \varepsilon_{Z2}$, $X = \tau_1 Z_1 + \tau_2 Z_2 + \rho U + \varepsilon_X$, $Y = \beta X + \kappa U + \varepsilon_Y$, where all constant coefficients are set to 1, and all noise terms follow the Gaussian distribution with mean 0 and standard deviation 1. The nonlinear function $\varphi_{Z1}(U)$ matches the corresponding function provided in each row. The details of the nonlinear function are as follows:

$$\begin{aligned}
 \text{Log: } Y &= \log_e(0.2|X - 1|); \\
 \text{Quadratic polynomial: } Y &= X^2 - 2 \cdot X + 1; \\
 \text{Cubic polynomial: } Y &= X^3 - 0.5 \cdot X^2 + 0.2 \cdot X; \\
 \text{Log (quadratic): } Y &= \log_e(|0.5 \cdot X^2 + X|); \\
 \text{Exp (quadratic): } Y &= e^{0.3 \cdot X^2 + 1};
 \end{aligned} \tag{56}$$

- Table 5. The model setup is as follows: $U = \varepsilon_U$, $Z_1 = \gamma U + \varepsilon_{Z1}$, $Z_2 = \varepsilon_{Z2}$, $X = \tau_1 Z_1 + \tau_2 Z_2 + \rho U + \varepsilon_X$, $Y = f(X) + \kappa U + \varepsilon_Y$, where all constant coefficients are set to 1, and all noise terms follow the Uniform distribution with parametric min = -2 and max = 2. The nonlinear function $f(X)$ corresponds to the specific function listed in each row. The details of the nonlinear function are as follows:

$$\begin{aligned}
 \text{Log: } Y &= 8 \cdot \log_e(|X|); \\
 \text{Quadratic polynomial: } Y &= (5 \cdot X + 2)^2 - 10; \\
 \text{Cubic polynomial: } Y &= 5 \cdot X^3 + 2 \cdot X^2 + 2 \cdot X - 3; \\
 \text{Log (quadratic): } Y &= 8 \cdot \log_e(2 \cdot X^2 + X + 1); \\
 \text{Exp (quadratic): } Y &= 3 \cdot e^{1.5 \cdot X^2 + 0.5 \cdot X + 2};
 \end{aligned} \tag{57}$$

- Table 6. The model setup is as follows: $U = \varepsilon_U$, $Z_1 = \varepsilon_{Z1}$, $Z_2 = \varepsilon_{Z2}$, $X = \text{sign}(Z_1) + g_X(Z_2) + \rho U + \varepsilon_X$, $Y = X + g_Y(Z_1) + \kappa U + \varepsilon_Y$, where all constant coefficients are set to 1, and all noise terms follow a Beta distribution with parameters alpha = 0.5 and beta = 0.1. The $\text{sign}(\cdot)$ denotes sign function, where $\cdot > 0$ equals 1, $\cdot = 0$ equals 0, and otherwise, it equals -1. The nonlinear functions $g_X(Z_2)$, $g_Y(Z_1)$ are defined by the specific functions provided in each row. The specific nonlinear functions are as follows:

$$\begin{aligned}
 \text{Log: } Y &= 4.5 \cdot \log_e(2 \cdot |X| + 0.3); \\
 \text{Quadratic polynomial: } Y &= (1.5 \cdot X + 2)^2 - 8; \\
 \text{Cubic polynomial: } Y &= 3 \cdot X^3 + 2 \cdot X^2 + 5.5 \cdot X + 5; \\
 \text{Log (quadratic): } Y &= \log_e(|4.5 \cdot X^2 - 0.1 \cdot X - 0.1|); \\
 \text{Exp (quadratic): } Y &= e^{2.5 \cdot X^2 + 1.9 \cdot X};
 \end{aligned} \tag{58}$$

The model setup is as follows: $U = \varepsilon_U$, $Z_1 = \varepsilon_{Z1}$, $Z_2 = \varepsilon_{Z2}$, $X = \text{sign}(Z_1) + g_X(Z_2) + \varphi_X(U) + \varepsilon_X$, $Y = f(X) + g_Y(Z_1) + \varphi_Y(U) + \varepsilon_Y$, where $\text{sign}(\cdot)$ denotes sign function, and all noise terms follow the Uniform distribution with parametric min = -2 and max = 2. The nonlinear functions correspond to the specific functions provided in each row. The detailed forms of these nonlinear functions are as follows:

$$\begin{aligned}
 \text{Log: } Y &= \log_e(|X - 0.01|) - 0.01; \\
 \text{Quadratic polynomial: } Y &= 0.1 \cdot (0.5 \cdot X - 0.1)^2 - 10; \\
 \text{Cubic polynomial: } Y &= X^3 + 0.5 \cdot X^2 + 0.2 \cdot X - 0.05; \\
 \text{Log (quadratic): } Y &= \log_e(|0.5 \cdot X^2 + X| + 1); \\
 \text{Exp (quadratic): } Y &= e^{0.2 \cdot X^2};
 \end{aligned} \tag{59}$$

- Table 8. The specific generation mechanism for the linear model with covariates $\mathbf{W}$ is defined as follows: $U = \varepsilon_U$, $\mathbf{W} = \varepsilon_{\mathbf{W}}$, $Z_1 = \mathcal{I}(U + \mathbf{W} + \varepsilon_{Z1})$, $Z_2 = \mathcal{I}(\mathbf{W} + \varepsilon_{Z2})$, $X = 0.5Z_1 + 0.5Z_2 + \lambda\mathbf{W} + \delta$, and $Y = X + \mathbf{W} + \epsilon$, where $\varepsilon_U \sim T(5)$, $\varepsilon_{Z1} \sim \text{Beta}(0.5, 0.1)$, $\varepsilon_{Z2} \sim \mathcal{N}(0, 1)$, and $\delta, \epsilon \sim T(5)$. Here, $\mathcal{I}(\cdot)$ is the indicator function such that $\mathcal{I}(\cdot) > \text{mean}(\cdot)$ equals 1; otherwise, it is 0. The coefficient λ is randomly drawn from a normalized standard normal distribution. The noise terms $\varepsilon_{\mathbf{W}}$ follow a multidimensional normal distribution, consistent with IV-PIM, with the dimensionality of covariates $\mathbf{W}$ varying across $|\mathbf{W}| = \{2, 3, 5\}$.

In the IV-PIM method, the parameters are set as follows: the number of bootstrap samples $B_{bootstrap} = 5$, the kappa method is specified as *spectral*, and the synthetic treatment variable method is set to *knockoff*.

- Table 9. The discrete treatment data that simulates violations of the *exogeneity* and *exclusion restriction* conditions as follows: $U = \varepsilon_U$, $Z = \mathcal{I}(\varphi_Z(U) + \varepsilon_Z)$, $X = \mathcal{I}(g_X(Z) + \varphi_X(U) + \varepsilon_X)$, $Y = \beta X + g_Y(Z) + \varphi_Y(U) + \varepsilon_Y$, and $\varepsilon_* \sim \mathcal{N}(0, 1)$, where $\beta = 1$, and $\mathcal{I}(\cdot)$ is the indicator function such that $\mathcal{I}(\cdot) > \text{mean}(\cdot)$ equals 1; otherwise, it is 0. The functions $\varphi_*(U)$ and $g_*(Z)$ are nonlinear and randomly selected from the following: *cos*, *sin*, *square*, *cubic(third-degree polynomials)*, *logarithmic*, *exponential function*.
- Table 10. The model settings are as follows:
 - Linear Model: $U = \varepsilon_U$, $Z = \varepsilon_Z$, $X = Z + U + \varepsilon_X$, $Y = X + U + \varepsilon_Y$, where $\varepsilon_U, \varepsilon_Z \sim \mathcal{U}[-1, 0) \cup (0, 1]$, and $\varepsilon_X, \varepsilon_Y \sim \mathcal{N}(0, 1)$.
 - Partial Non-Linear Model with Constant Causal Effect: $U = \varepsilon_U$, $Z = \varepsilon_Z$, $X = \sin(Z) + U + \varepsilon_X$, $Y = X + U^2 + \varepsilon_Y$, where all independent noise follow the standard Gaussian distribution $\varepsilon_* \sim \mathcal{N}(0, 1)$.
 - ANINCE Model: $U = \varepsilon_U$, $Z = \varepsilon_Z$, $X = Z^2 + U + \varepsilon_X$, $Y = e^X + U + \varepsilon_Y$, where all independent noise follow the standard Gaussian distribution $\varepsilon_* \sim \mathcal{N}(0, 1)$.
- Table 11. The model settings are as follows:
 - Linear Model and Exogeneity Violated: $U = \varepsilon_U$, $Z = 2U + \varepsilon_Z$, $X = Z + U + \varepsilon_X$, $Y = X + U + \varepsilon_Y$, where $\varepsilon_U, \varepsilon_Z \sim \text{Gamma}(1, 2)$, and $\varepsilon_X, \varepsilon_Y \sim \text{Beta}(1, 2)$.
 - Linear Model and Exogeneity & Exclusion Restriction Violated: $U = \varepsilon_U$, $Z = 3U + \varepsilon_Z$, $X = Z + U + \varepsilon_X$, $Y = X + 2Z + U + \varepsilon_Y$, where $\varepsilon_U, \varepsilon_Z \sim \text{Gamma}(1, 2)$, and $\varepsilon_X, \varepsilon_Y \sim \mathcal{N}(0, 1)$.

- Partial Non-Linear Model with Constant Causal Effect and Exogeneity Violated: $U = \varepsilon_U$, $Z = U^3 + \varepsilon_Z$, $X = Z + U + \varepsilon_X$, $Y = X + U + \varepsilon_Y$, where $\varepsilon_U, \varepsilon_Z \sim \mathcal{N}(0, 1)$, and $\varepsilon_X, \varepsilon_Y \sim \mathcal{U}[-1, 0) \cup (0, 1]$.
- Partial Non-Linear Model with Constant Causal Effect and Exclusion Restriction Violated: $U = \varepsilon_U$, $Z = \varepsilon_Z$, $X = Z + U + \varepsilon_X$, $Y = X + \sin(Z) + U^2 + \varepsilon_Y$, where $\varepsilon_U, \varepsilon_Z \sim \mathcal{N}(0, 1)$, and $\varepsilon_X, \varepsilon_Y \sim \mathcal{U}[-1, 0) \cup (0, 1]$.
- Partial Non-Linear Model with Constant Causal Effect and Exogeneity & Exclusion Restriction Violated: $U = \varepsilon_U$, $Z = U + \varepsilon_Z$, $X = Z + U + \varepsilon_X$, $Y = X + e^Z + 2U + \varepsilon_Y$, where $\varepsilon_U, \varepsilon_Z \sim \mathcal{U}[-1, 0) \cup (0, 1]$, and $\varepsilon_X, \varepsilon_Y \sim \mathcal{N}(0, 1)$.

References

- Daron Acemoglu, Simon Johnson, and James A Robinson. The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5):1369–1401, 2001.
- Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.
- Yonghong An and Yingyao Hu. Well-posedness of measurement error models for self-reported data. *Journal of Econometrics*, 168(2):259–269, 2012.
- Donald WK Andrews. Examples of l2-complete and boundedly-complete distributions. *Journal of Econometrics*, 199(2):213–220, 2017.
- Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.
- Lee J Bain and Max Engelhardt. *Introduction to probability and mathematical statistics*, volume 4. Duxbury Press Belmont, CA, 1992.
- Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.
- Robert L Basmann. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica: Journal of the Econometric Society*, pages 77–83, 1957.
- Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in Neural Information Processing Systems*, 32, 2019.
- Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.
- John Bound, David A Jaeger, and Regina M Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450, 1995.

- Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International Journal of Epidemiology*, 44(2):512–525, 2015.
- Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314, 2016.
- Roger J Bowden and Darrell A Turkington. *Instrumental variables*. Cambridge University Press, 1990.
- Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.
- Patrick F Burael. Evaluating instrument validity using the principle of independent mechanisms. *Journal of Machine Learning Research*, 24(176):1–56, 2023.
- Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. *Advances in Neural Information Processing Systems*, 32, 2019.
- Ivan A Canay, Andres Santos, and Azeem M Shaikh. On the testability of identification in some nonparametric models with endogeneity. *Econometrica*, 81(6):2535–2559, 2013.
- David Card. Using geographic variation in college proximity to estimate the return to schooling, 1993.
- Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics*, 6:5633–5751, 2007.
- Raymond J Carroll, Xiaohong Chen, and Yingyao Hu. Identification and estimation of nonlinear models using two samples with nonclassical measurement errors. *Journal of Nonparametric Statistics*, 22(4):379–399, 2010.
- Bryant Chen, Daniel Kumor, and Elias Bareinboim. Identification and model testing in linear structural equation models using auxiliary variables. In *International Conference on Machine Learning*, pages 757–766. PMLR, 2017.
- Xiaohong Chen and Yingyao Hu. Identification and inference of nonlinear models using two samples with arbitrary measurement errors. *Cowles Foundation Discussion Papers*, 2006.
- Xiaohong Chen and Markus Reiss. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27(3):497–521, 2011.
- Xiaohong Chen, Victor Chernozhukov, Sokbae Lee, and Whitney K Newey. Local identification of nonparametric and semiparametric models. *Econometrica*, 82(2):785–809, 2014.

- Yutian Chen, Liyuan Xu, Caglar Gulcehre, Tom Le Paine, Arthur Gretton, Nando De Freitas, and Arnaud Doucet. On instrumental variable regression for deep offline policy evaluation. *The Journal of Machine Learning Research*, 23(1):13635–13674, 2022.
- Victor Chernozhukov, Guido W Imbens, and Whitney K Newey. Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, 139(1):4–14, 2007.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, pages C1–C68, 2018.
- Tianjiao Chu, Richard Scheines, and Peter Spirtes. Semi-instrumental variables: a test for instrument admissibility. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 83–90, 2001.
- H. Cramér. *Random variables and probability distributions*. Cambridge University Press, Cambridge, 2nd edition, 1962.
- George Darmois. Analyse générale des liaisons stochastiques: etude particulière de l’analyse factorielle linéaire. *Revue de l’Institut international de statistique*, pages 2–8, 1953.
- Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.
- Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 132–141, 2014.
- Mathias Drton and Thomas S Richardson. Iterative conditional fitting for gaussian ancestral graph models. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, pages 130–137, 2004.
- Xavier D’Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460–471, 2011.
- Frédérique Fève and Jean-Pierre Florens. Non parametric analysis of panel data models with endogenous variables. *Journal of Econometrics*, 181(2):151–164, 2014.
- Jean-Pierre Florens, Jan Johannes, and Sébastien Van Bellegem. Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory*, 27(3):472–496, 2011.
- Joachim Freyberger. On completeness and consistency in nonparametric instrumental variable models. *Econometrica*, 85(5):1629–1644, 2017.
- Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019.
- Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005.

- Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, pages 585–592, 2008.
- Florian F Gunsilius. Nontestability of instrument validity under continuous treatments. *Biometrika*, 108(4):989–995, 2021.
- Zijian Guo and Dylan S Small. Control function instrumental variable estimation of nonlinear causal effect models. *Journal of Machine Learning Research*, 17(100):1–35, 2016.
- Zijian Guo, Hyunseung Kang, T Tony Cai, and Dylan S Small. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):793–815, 2018.
- Peter Hall and Joel L Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929, 2005.
- Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 46(6):1985–1998, 2017.
- Leonard Henckel, Martin Buttenschoen, and Marloes H Maathuis. Graphical tools for selecting conditional instrumental sets. *Biometrika*, 111(3):771–788, 2024.
- Miguel A Hernán and James M Robins. Instruments for causal inference: an epidemiologist’s dream? *Epidemiology*, pages 360–372, 2006.
- Joel L Horowitz. Specification testing in nonparametric instrumental variable estimation. *Journal of Econometrics*, 167(2):383–396, 2012.
- Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21, 2008.
- Yingyao Hu and Susanne M Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.
- Yingyao Hu and Ji-Liang Shiu. Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Econometric Theory*, 34(3):659–693, 2018.
- Yingyao Hu and Ji-Liang Shiu. A simple test of completeness in a class of nonparametric specification. *Econometric Reviews*, 41(4):373–399, 2022.
- Guido W. Imbens. Instrumental variables: An econometrician’s perspective. *Statistical Science*, 29(3):323–358, 2014.
- Guido W Imbens and Donald B Rubin. Causal inference for statistics, social, and biomedical sciences: An introduction. *Cambridge University Press*, 2015.

- Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.
- Désiré Kédagni and Ismael Mourifié. Generalized instrumental inequalities: testing the instrumental variable independence assumption. *Biometrika*, 107(3):661–675, 2020.
- Toru Kitagawa. A test for instrument validity. *Econometrica*, 83(5):2043–2063, 2015.
- Rainer Kress. *Linear integral equations*, volume 82. Springer, 1989.
- Rainer Kress. *Linear Integral Equations*, volume 82. Springer Science & Business Media, 2013.
- Juan Lin. Factorizing multivariate function classes. *Advances in Neural Information Processing Systems*, 10, 1997.
- Felipe Maia Polo, Yuekai Sun, and Moulinath Banerjee. Conditional independence testing under misspecified inductive biases. *Advances in Neural Information Processing Systems*, 36:58577–58612, 2023.
- Charles F Manski. *Partial identification of probability distributions*. Springer Science & Business Media, 2003.
- Torben Martinussen, Ditte Nørbo Sørensen, and Stijn Vansteelandt. Instrumental variables estimation under a structural cox model. *Biostatistics*, 20(1):65–79, 2019.
- Ronald Meester. *A natural introduction to probability theory*. Springer, 2008.
- Kevin M Murphy and Robert H Topel. Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 20(1):88–97, 2002.
- Alice Nakamura and Masao Nakamura. On the relationships among several specification error tests presented by durbin, wu, and hausman. *Econometrica: Journal of the Econometric Society*, pages 1583–1588, 1981.
- Whitney K Newey. Nonparametric instrumental variables estimation. *American Economic Review*, 103(3):550–556, 2013.
- Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.
- Tom M Palmer, Roland R Ramsahai, Vanessa Didelez, and Nuala A Sheehan. Nonparametric bounds for the causal effect in a binary instrumental-variable model. *The Stata Journal*, 11(3):345–367, 2011.
- Judea Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 435–443, 1995.

- Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edition, 2009.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.
- Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.
- Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.
- Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.
- Yixin Ren, Juncai Zhang, Yewei Xia, Ruxin Wang, Feng Xie, Jihong Guan, Hao Zhang, and Shuigeng Zhou. Regression-based conditional independence test with adaptive kernels. *Artificial Intelligence*, page 104391, 2025.
- Jean-Marc Robin and Richard J Smith. Tests of rank. *Econometric Theory*, 16(2):151–175, 2000.
- Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang. Learning linear non-gaussian causal models in the presence of latent variables. *Journal of Machine Learning Research*, 21(39):1–24, 2020.
- Cyrrill Scheidegger, Malte Londschieen, and Peter Bühlmann. A residual prediction test for the well-specification of linear instrumental variable models. *arXiv preprint arXiv:2506.12771*, 2025.
- Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.
- Shohei Shimizu. *Statistical Causal Discovery: LiNGAM Approach*. Springer, 2022.
- Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7 (Oct):2003–2030, 2006.
- Ji-Liang Shiu and Yingyao Hu. Identification and estimation of nonlinear dynamic panel data models with unobserved covariates. *Journal of Econometrics*, 175(2):116–131, 2013.

- Ricardo Silva and Shohei Shimizu. Learning instrumental variables with structural and non-gaussianity assumptions. *Journal of Machine Learning Research*, 18(120):1–49, 2017.
- Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.
- Tea Skaaby, Lise Lotte Nystrup Husemoen, Torben Martinussen, Jacob P Thyssen, Michael Melgaard, Betina Heinsbæk Thuesen, Charlotta Pisinger, Torben Jørgensen, Jeanne D Johansen, Torkil Menné, et al. Vitamin d status, filaggrin genotype, and cardiovascular risk factors: a mendelian randomization approach. *PloS one*, 8(2):e57647, 2013.
- VP Skitovitch. On a property of the normal distribution. *DAN SSSR*, 89:217–219, 1953.
- Peter Spirtes. Calculation of entailed rank constraints in partially non-linear and cyclic models. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 606–615. AUAI Press, 2013.
- Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied Informatics*, volume 3, pages 1–28. SpringerOpen, 2016.
- Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.
- Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.
- Seth Sullivant, Kelli Talaska, Jan Draisma, et al. Trek separation for gaussian graphical models. *The Annals of Statistics*, 38(3):1665–1685, 2010.
- Maarten J Voors, Eleonora E M Nillesen, Philip Verwimp, Erwin H Bulte, Robert Lensink, and Daan P Van Soest. Violent conflict and behavior: a field experiment in burundi. *American Economic Review*, 102(2):941–964, 2012.
- Linbo Wang, James M Robins, and Thomas S Richardson. On falsification of the binary instrumental variable model. *Biometrika*, 104(1):229–236, 2017.
- Wenjie Wang, Mingming Gong, Biwei Huang, James Bailey, Bo Han, Kun Zhang, and Feng Liu. Practical kernel selection for kernel-based conditional independence test. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- Frank Windmeijer, Helmut Farbmacher, Neil Davies, and George Davey Smith. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1339–1350, 2019.
- Frank Windmeijer, Xiaoran Liang, Fernando P Hartwig, and Jack Bowden. The confidence interval method for selecting valid instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):752–776, 2021.
- Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, 2010.

- Anpeng Wu, Kun Kuang, Bo Li, and Fei Wu. Instrumental variable regression with confounder balancing. In *International Conference on Machine Learning*, pages 24056–24075. PMLR, 2022.
- Feng Xie, Yangbo He, Zhi Geng, Zhengming Chen, Ru Hou, and Kun Zhang. Testability of instrumental variables in linear non-gaussian acyclic causal models. *Entropy*, 24(4):512, 2022.
- Yun Yang and Surya T Tokdar. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, pages 652–674, 2015.
- Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. *Advances in Neural Information Processing Systems*, 26, 2013.
- Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813. AUAI Press, 2011.
- Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.