Causal Discovery via Conditional Independence Testing with Proxy Variables

Mingzhou Liu [1 2] **Xinwei Sun** ✉ [3] **Yu Qiao** [4] **Yizhou Wang** [2 1 5 6 7]

Abstract

Distinguishing causal connections from correlations is important in many scenarios. However, the presence of unobserved variables, such as the latent confounder, can introduce bias in conditional independence testing commonly employed in constraint-based causal discovery for identifying causal relations. To address this issue, existing methods introduced proxy variables to adjust for the bias caused by unobserveness. However, these methods were either limited to categorical variables or relied on strong parametric assumptions for identification. In this paper, we propose a novel hypothesis-testing procedure that can effectively examine the existence of the causal relationship over continuous variables, without any parametric constraint. Our procedure is based on discretization, which under completeness conditions, is able to asymptotically establish a linear equation whose coefficient vector is identifiable under the causal null hypothesis. Based on this, we introduce our test statistic and demonstrate its asymptotic level and power. We validate the effectiveness of our procedure using both synthetic and real-world data. Code is publicly available at https://github.com/lmz123321/proxy_causal_discovery.

1. Introduction

Distinguishing causation from correlation is a fundamental task in various fields, such as economics, healthcare, and policy formulation. By avoiding spurious correlations caused by confounding bias or mediation bias, it facilitates drawing reliable conclusions and thereby making informed decisions. Under the causal sufficiency condition where all confounders are observed, one can directly conduct conditional independence testing (Zhang et al., 2012) to examine the existence of causation, as employed by constraint-based methods for causal discovery (Glymour et al., 2019).

However, this procedure will suffer from nonignorable bias when *unobserved variables* are present, such as latent confounders or mediators, as illustrated in Fig. 1. Typically, these variables refer to attributes that are difficult to measure (*e.g.*, health status, living habits) but exhibit correlations to both variables within the specific pair of causal interests.

To address this challenge, recent studies in proximal causal learning introduced the proxy variable, which can be a noisy measurement (Kuroki & Pearl, 2014) or an observed descendant (Miao et al., 2018) of the unobserved variable, to adjust for the bias caused by unobserveness. Examples include (Tchetgen et al., 2020; Deaner, 2021; Sverdrup & Cui, 2023) that explored proxy variables for causal effect estimation; and (Ghassami et al., 2021; Dukes et al., 2023) that studied proximal mediation analysis. However, most of these works focused on causal inference. The works that are mostly related to us are (Miao et al., 2018) and (Miao et al., 2022), which introduced testing procedures to examine the null hypothesis of no causal relationship. However, their procedures were either limited to the discrete variables (Miao et al., 2018) or relied on strong parametric assumptions to ensure the testing validity (Miao et al., 2022).

In this paper, we propose a novel hypothesis-testing procedure that can effectively identify the causal relationship between two continuous variables, without any parametric assumption. Our strategy is to find a proper discretization procedure such that, under the causal null hypothesis, the probability matrices over discretized variables satisfy a linear equation whose coefficient vector is identifiable. For this purpose, we first introduce a discretization procedure that, under the completeness assumption, makes the probability matrices over discrete variables satisfy rank regularities that are crucial for the identifiability of the coefficient vector. Then, we proceed to analyze the errors in the linear equation after discretization and show that they can be controlled to be arbitrarily small, given that the bin numbers for discretiza-

---
[1]School of Computer Science, Peking University (liumingzhou@stu.pku.edu.cn) [2]Center on Frontiers of Computing Studies (CFCS), Peking University [3]Sch. of Data Science, Fudan University [4]Dep. of Automation, Shanghai Jiao Tong University [5]Inst. for Artificial Intelligence, Peking University [6]Nat'l Eng. Research Center of Visual Technology, Peking University [7]Nat'l Key Lab of General Artificial Intelligence, Peking University. Correspondence to: Xinwei Sun <sunxinwei@fudan.edu.cn>.

tion are sufficiently large. Guided by the above analyses, we introduce our testing statistic based on the residue of linear regression and establish the asymptotic level and power of our test. We demonstrate the validity and effectiveness of our procedure on both synthetic and real-world data.

We summarize our contributions as follows:

1. We propose a proximal hypothesis testing based on discretization for identifying causal relations between continuous variables.
2. We establish the asymptotic level and power of our procedure without any parametric assumptions.
3. We demonstrate the validity of our procedure on both synthetic and real-world datasets.

2. Related works

Proxy causal learning. The proximal causal learning was first introduced by (Kuroki & Pearl, 2014) for causal inference with unobserved confounders, and was later studied in (Miao et al., 2018; Tchetgen et al., 2020; Deaner, 2021; Sverdrup & Cui, 2023; Ghassami et al., 2023) for causal effect identification and in (Mastouri et al., 2021; Ghassami et al., 2022) for estimation. Recently, (Rojas-Carulla et al., 2017; Liu et al., 2023) leveraged the proxy variable for causal discovery; however, they respectively focused on causal orientation and time-series causal discovery, which are different from the problem considered in this paper. The works most relevant to ours are (Miao et al., 2018; 2022). In (Miao et al., 2018), the authors introduced a hypothesis testing procedure with the proxy variable to adjust for the bias introduced by latent confounders; however, their procedure is limited to discrete variables. For continuous variables, (Miao et al., 2022) constructed a testing procedure by examining the existence of the solution of the integral equation. However, they rely on parametric assumptions to establish the asymptotic behavior of the test statistic under the null hypothesis. **In this paper**, we propose a novel testing procedure to effectively identify causal relationships between continuous variables, applicable to both parametric and non-parametric models.

Conditional Independence (CI) testing. CI testing is a fundamental step for constraint-based causal discovery. To effectively test CI, many works have been proposed. Examples include (Zhang et al., 2012; Doran et al., 2014; Pfister et al., 2018; Strobl et al., 2019) that explored the kernel embedding and the Hilbert-Schimidt Independence Criterion (HSIC) for CI testing, (Runge, 2018; Mukherjee et al., 2020; Ai et al., 2022) that used the mutual information as a measure to quantify the CI, and (Margaritis, 2005; Huang, 2010; Neykov et al., 2021; Warren, 2021) that introduced discretization-based procedures for CI testing. However, these works rely on the causal sufficiency condition that all conditional variables are observable. **As a contrast**, our testing procedure can effectively test CI in the presence of unobserved variables.

3. Preliminary

In this section, we introduce the problem setup, notations, and background knowledge that is related to our work.

Problem setup. We consider the problem of identifying the causal relationship within a pair of variables X and Y. The system also includes an unobserved variable U, which can be either a sufficient confounder or mediator between X and Y. For this case, causal identification is generally not feasible, due to the confounding bias (Jager et al., 2008) or the mediation bias (Hicks & Tingley, 2011). To adjust for the bias, we assume the availability of a proxy variable W satisfying $X \perp\!\!\!\perp W | U$[1] (Kuroki & Pearl, 2014). In practice, the proxy variable can be a noisy measurement (Kuroki & Pearl, 2014) or an observed descendant (Liu et al., 2023) of the unobserved variable U. We illustrate the relationships between X, Y, U, and W in Fig. 1.

Under the Markovian and faithfulness assumptions (Pearl, 2009), the problem of identifying the causal relationship between X and Y is equivalent to that of testing the causal null hypothesis $\mathbb{H}_0 : X \perp\!\!\!\perp Y | U$ with the proxy variable W. $\mathbb{H}_0$ means there is no causal relationship, and a rejection of $\mathbb{H}_0$ is, therefore, evidence in favor of causation.

Remark 3.1. Our procedure can identify whether the causal relationship exists between X and Y. For orientation, one can exploit prior knowledge such as temporal order, acyclicity constraint, or other properties such as asymmetry and stability (Tu et al., 2022; Duong & Nguyen, 2022).

Notations. Suppose X, Y, U, W are random variables over the probability space $(\Omega, \mathcal{F}, P)$, with domains $\mathcal{X}, \mathcal{Y}, \mathcal{U}, \mathcal{W}$, respectively. We denote the probability law of X as $\mathcal{L}_X := P \circ X^{-1}$ and denote the cumulative density function (CDF) as F_X. For discrete variables, we denote their levels as l_X, l_Y, etc. We use the the column vector $P(X|y) := [p(x_1|y), ..., p(x_{l_X}|y)]^\top$, the row vector $P(x|Y) := [p(x|y_1), ..., p(x|y_{l_Y})]$, and the matrix $P(X|Y) := [P(X|y_1), ..., P(X|y_{l_Y})]$ to represent their transition probabilities. For a continuous variable X, we denote $\{\mathcal{X}_i\}_{l_X}$ a disjoint partition of $\mathcal{X}$ with l_X subspaces, such that $\cup_i^{l_X} \mathcal{X}_i = \mathcal{X}$ and $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ for $i \neq j$. We thereby define the discretized version of X as $\tilde{X}$, and denoting $\tilde{X} := \tilde{x}_i$ to represent the event $X \in \mathcal{X}_i$ for any i. We denote $F(x|y)$ as the conditional CDF of x given $Y = y$. Besides, we denote $F(x|\boldsymbol{y}_{[i]}) := [F(x|y_1), ..., F(x|y_i)]^\top$, $F(\boldsymbol{x}_{[i]}|y) := [F(x_1|y), ..., F(x_i|y)]^\top$, and $F(\boldsymbol{x}_{[i]}|\boldsymbol{y}_{[i]}) :=$

---

[1]Our proposed procedure can be naturally applied to the scenario where multiple proxies are available, while the single-proxy scenario considered in this paper is more challenging.
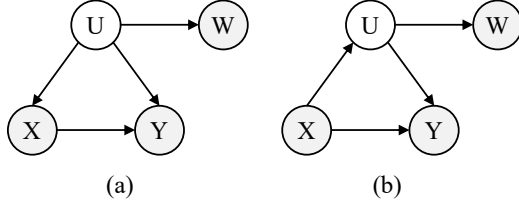
Figure 1. Causal diagrams illustrating causal discovery with proxy variables. (a) and (b) respectively represent the cases where U is a latent confounder and a latent mediator. Note that our procedure is not restricted to these diagrams, but can apply to any scenario satisfying $X \perp\!\!\!\perp W | U$ (Kuroki & Pearl, 2014).

$[F(x_1|\boldsymbol{y}_{[i]}), ..., F(x_i|\boldsymbol{y}_{[i]})]$, with $[i] := \{1, ..., i\}$. We let $A_1 \backslash A_2 := A_1 \cap A_2^c$ for two sets A_1, A_2. We denote n as the sample size.

The procedure described by (Miao et al., 2018). For discrete variables, (Miao et al., 2018) provided a hypothesis-testing procedure based on the proxy variable W[2]. In details, they connected $\mathbb{H}_0 : X \perp\!\!\!\perp Y | U$ with the following equation:

$$\underbrace{P(y|X)^\top}_{\text{response}} = \underbrace{P(W|X)^\top}_{\text{design matrix}} \underbrace{\{P(y|U)P(W|U)^{-1}\}^\top}_{\text{coefficient}}, \quad (1)$$

by noticing that:

$$P(y|X) = P(y|U)P(U|X) \quad \mathbb{H}_0 : X \perp\!\!\!\perp Y | U, \quad (2a)$$
$$P(W|X) = P(W|U)P(U|X) \quad X \perp\!\!\!\perp W | U. \quad (2b)$$

Under the assumptions that the matrix $P(W|X)$ has full row rank and $P(W|U)$ is invertible, the linear system described by Eq. (1) holds and the coefficient vector is identifiable under $\mathbb{H}_0$. In this regard, one can test $\mathbb{H}_0$ by investigating the residue via linear regression.

While it has proven effective for discrete variables, extending such a hypothesis-testing procedure to continuous variables can be challenging. This is because, the conditional **independence** property encoded in Eq. (2), which is key to establishing the linear relation in Eq. (1), may not hold after discretization. Additionally, to ensure the **identifiability** of the coefficient vector, one should design a valid discretization procedure to guarantee the full row rank of $P(W|X)$ after discretization. In the next section, we propose a novel approach to address these challenges.

4. Methodology

In this section, we introduce a hypothesis-testing procedure based on discretization. At the core of our procedure lies finding a discretization procedure that establishs the identifiability of Eq. (1).

---

[2]Their procedure was originally designed with two proxy variables W and Z, satisfying $l_X l_Z > l_W$. Indeed, we can drop Z and only use the single proxy W if $l_X > l_W$.

To achieve this goal, in Sect. 4.1, we first introduce a discretization procedure which, under the completeness assumption, can ensure the discretized $\tilde{X}$ and $\tilde{W}$ meet the rank regularity. Then, in Sect. 4.2, we analyze the errors induced during the discretization. Under smoothness conditions regarding conditional distributions, we show that they can be effectively controlled to be arbitrarily small if the bin number is sufficiently large. Finally, based on the above analyses, we introduce our testing statistic and discuss the asymptotic level and power of the test in Sect. 4.3.

4.1. Discretization under completeness

In this section, we introduce a discretization procedure to establish the identifiability of Eq. (1). As mentioned previously, this involves finding discretization for W and X such that the resulted $P(\tilde{W}|\tilde{X})$ has full row rank. To begin, we first introduce the following completeness conditions that have been similarly adopted for continuous variables in causal inference (Miao et al., 2018; 2022), which ensures the existence of such a discretization.

Assumption 4.1 (Completeness). We assume the conditional distributions $\mathcal{L}_{U|X}$ and $\mathcal{L}_{W|X}$ are complete. That is, for any bounded function g, we assume

$$\mathbb{E}[g(u)|x] = 0 \text{ almost surely iff } g(u) = 0 \text{ almost surely,}$$
$$\mathbb{E}[g(w)|x] = 0 \text{ almost surely iff } g(w) = 0 \text{ almost surely.}$$

To understand why we need Asm. 4.1 in our method, note that when W is discrete, the completeness of $\mathcal{L}_{W|X}$ means the conditional probability matrix $P(W|X)$ has full row rank, as required in the procedure for discrete variables (Miao et al., 2018). When W is continuous, this condition enables us to find a discretization such that $P(\tilde{W}|\tilde{X})$ has full row rank. Further, when combined with the completeness of $\mathcal{L}_{U|X}$, we can demonstrate the existence of discretization for U such that the matrix $P(\tilde{W}|\tilde{U})$ is invertible, which is also required for the identifiability of Eq. (1).

The following example shows that Asm. 4.1 can hold for a wide range of Addictive Noise Models (ANMs).

Example 4.2 (ANM with completeness). Suppose that X, U, W satisfy the ANM, *i.e.*, $X = g_X(U) + N_X, W = g_W(U) + N_W$, with N_X, N_W being the noises. If g_X is invertible with a non-zero derivative, and N_X, N_W have non-zero characterization functions, then the conditional distributions $\mathcal{L}_{U|X}$ and $\mathcal{L}_{W|X}$ are both complete.

Remark 4.3. We also provide an example where U is a mediator. Please refer to Appx. C.1 for details.

The completeness conditions have been similarly assumed in causal hypothesis testing and causal inference, with the difference that our condition is not restricted to square-integrable functions. For example, the completeness of

3

$\mathcal{L}_{U|X}$ was assumed by (Miao et al., 2018; 2022)[3]. Besides, a discrete version of the completeness of $\mathcal{L}_{W|X}$, i.e., $P(W|X)$ has full row rank, was required by the procedure described by (Miao et al., 2018). Indeed, in (Miao et al., 2018), they also assumed the completeness of $\mathcal{L}_{X|W}$[4]. However, this assumption is required primarily for causal effect estimation, which therefore is not required in this paper.

Next, we show that, under Asm. 4.1, we can find a discretization procedure such that the $P(\tilde{W}|\tilde{X})$ has full row rank and the matrix $P(\tilde{W}|\tilde{U})$ is invertible.

Proposition 4.4. *Suppose Asm. 4.1 holds. Then, for any discretization $\tilde{W}$ of W, there exists a discretization $\tilde{X}$ of X such that the matrix $P(\tilde{W}|\tilde{X})$ has full row rank. Similarly, there also exists a discretization $\tilde{U}$ of U such that the matrix $P(\tilde{W}|\tilde{U})$ is invertible.*

Remark 4.5. We leave the proof to Appx. C.2. Roughly speaking, we first show that given $\tilde{W}$ after discretizing W, the functions $F(x|\tilde{w}_1), ..., F(x|\tilde{w}_{l_W})$ are linear independent under the completeness condition of $\mathcal{L}_{W|X}$. Then we prove with induction that there exists $\tilde{x}_1, ..., \tilde{x}_{l_X}$ with $l_X > l_W$, such that $\left[F(\tilde{\boldsymbol{x}}_{[l_X]}|\tilde{w}_1), ..., F(\tilde{\boldsymbol{x}}_{[l_X]}|\tilde{w}_{l_W}) \right]$ are linear independent, implying the full column rank of $P(\tilde{X}|\tilde{W})$ and therefore the full row rank of $P(\tilde{W}|\tilde{X})$.

Remark 4.6. While we cannot discretize U due to its unobserved nature, this is not necessary as we do not need to directly compute $P(\tilde{W}|\tilde{U})$ in the linear regression. Instead, we only require it to be invertible with a proper discretization of U, as ensured by Prop. 4.4, to achieve identifiability of the coefficient vector in Eq. (1).

Inspired by Prop. 4.4, we introduce our discretization procedure in Alg. 1. The algorithm mainly contains three steps. First, in line 1, we uniformly discretize W into l_W bins to obtain $\tilde{W}$. Indeed, the choice of discretization for W is arbitrary, as indicated by Prop. 4.4. Next, in lines 2-8, we discretize X to construct $\tilde{X}'$ that also contains l_W bins, such that $P(\tilde{X}'|\tilde{W})$ and therefore $P(\tilde{W}|\tilde{X}')$ is invertible. Finally, we add another $(l_X - l_W)$ bins to $\tilde{X}'$ to obtain $\tilde{X}$ in line 9. In this regard, $P(\tilde{W}|\tilde{X})$ has full row rank.

In the second step (lines 2-8), we iteratively search for x_i such that the matrix $M_i := F(\boldsymbol{x}_{[i]}|\tilde{\boldsymbol{w}}_{[i]})$ is invertible given $x_1, ..., x_{i-1}$. This can be achieved by noting that the functions $F(x|\tilde{w}_1), ..., F(x|\tilde{w}_{l_W})$ are linear independent under the completeness condition. Once we have the invertibility of M_i, we have $P(\tilde{X}|\tilde{W}) := \{P(\tilde{x}_i|\tilde{w}_j)\}_{i,j}$ is invertible, with $P(\tilde{x}_i|\tilde{w}_j) := F(x_i|W \in \mathcal{W}_j) - F(x_{i-1}|W \in \mathcal{W}_j)$ and $\mathcal{X}_i := [x_{i-1}, x_i)$ $((-\infty, x_1)$ for $i = 1)$. To find such a x_i, note that $\det(M_i)$ can be expanded as:

---

[3]See Condition (ii) in (Miao et al., 2018) (by dropping Z) and Asm. 3 in (Miao et al., 2022).

[4]See Condition (iii), where z should be replaced with x to induce variation for identification.

Algorithm 1 Discretizing W and X.

Input: Continuous data $\{x_i, w_i\}_{i=1}^n$; $l_X > l_W$.
Output: $\{\mathcal{X}_i\}_{i=1}^{l_X}, \{\mathcal{W}_j\}_{j=1}^{l_W}$ s.t. $P(\tilde{W}|\tilde{X})$ has full row rank.

1: Discretize $\mathcal{W}$ into l_W uniform bins to obtain $\{\mathcal{W}_j\}_{j=1}^{l_W}$.
2: **for** $i = 1 : l_W$ **do**
3:     **for** $j = 1 : i$ **do**
4:         Estimate $d_j = \det(F(\boldsymbol{x}_{[i-1]}|\tilde{\boldsymbol{w}}_{[i]\setminus j}))$.
5:     **end for**
6:     Search for x_i s.t. $\sum_{j=1}^i (-1)^{1+j} d_j F(x_i|\tilde{w}_j) \neq 0$.
7:     Set $\mathcal{X}_i = [x_{i-1}, x_i)$.
8: **end for**
9: Discretize $\mathcal{X}\setminus[x_1, x_{l_W})$ into $(l_X - l_W)$ uniform bins to obtain $\{\mathcal{X}_i\}_{i=l_W+1}^{l_X}$.

$$\det(M_i)(x_i) = \sum_{j=1}^i (-1)^{1+j} d_j F(x_i|\tilde{w}_j),$$

where $d_j := \det(F(\boldsymbol{x}_{[i-1]}|\tilde{\boldsymbol{w}}_{[i]\setminus j}))$. By induction, there exists a non-zero $(i - 1)$ dimensional minor among $\{d_j\}$. Since $F(x|\tilde{\boldsymbol{w}}_{[i]})$ are linear independent, there exists x_i s.t. $\det(M_i)(x_i) \neq 0$ and therefore $F(\boldsymbol{x}_{[i]}|\tilde{\boldsymbol{w}}_{[i]})$ is invertible.

Choosing the bin number. The bin number is a trade-off between the discretization error and the estimation of discrete probabilities. While a larger bin number effectively controls discretization error (as explained in the next section), it may lead to inaccurate probability estimation due to insufficient samples in each bin. To balance this trade-off, we follow the heuristics proposed by (Dougherty et al., 1995) to set the bin number to $\lfloor \log(n) \rfloor$.

4.2. Discretization error analysis

In this section, we discuss the discretization error control. Given $\mathbb{H}_0 : X \perp\!\!\!\perp Y|U$ and $X \perp\!\!\!\perp W|U$, let $\tilde{y}$ represent the event $Y \in \tilde{\mathcal{Y}}$ for some $\tilde{\mathcal{Y}} \subset \mathcal{Y}$, our goal is to have:

$$P(\tilde{y}|\tilde{X}) = P(\tilde{y}|\tilde{U})P(\tilde{U}|\tilde{X}) + E_y, \tag{3a}$$

$$P(\tilde{W}|\tilde{X}) = P(\tilde{W}|\tilde{U})P(\tilde{U}|\tilde{X}) + E_w, \tag{3b}$$

with the error terms

$$E_y(i) = \sum_j \{P(\tilde{y}|\tilde{x}_i, \mathcal{U}_j) - P(\tilde{y}|\mathcal{U}_j)\}P(\mathcal{U}_j|\tilde{x}_i),$$

$$E_w(i) = \sum_j \{P(\tilde{W}|\tilde{x}_i, \mathcal{U}_j) - P(\tilde{W}|\mathcal{U}_j)\}P(\mathcal{U}_j|\tilde{x}_i),$$

being controlled to zeros asymptotically for all i. In this regard, we can establish Eq. (1) asymptotically.

We first consider the case where $\mathcal{U}$ is a compact interval, then extend the results to $\mathcal{U} = \mathbb{R}^d$ $(d \in \mathbb{Z}_+)$. For a compact $\mathcal{U}$, we can partition $\mathcal{U}$ into sufficiently many bins $\{\mathcal{U}_j\}_{l_U}$, such that in each small bin $\mathcal{U}_j$, the errors $|P(\tilde{y}|\tilde{x}_i, U \in \mathcal{U}_j) - P(\tilde{y}|U \in \mathcal{U}_j)|$ and $|P(\tilde{W}|\tilde{x}_i, U \in \mathcal{U}_j) - P(\tilde{W}|U \in$

$\mathcal{U}_j)|$ can be effectively controlled. The key to this property is the following smoothness condition, which ensures the conditional distributions do not vary too much in each small bin $\mathcal{U}_j$.

Assumption 4.7 (TV smoothness)**.** We assume that the maps $u \mapsto \mathcal{L}_{Y|U=u}$ and $u \mapsto \mathcal{L}_{W|U=u}$ are Lipschitz continuous wrt the Total Variation (TV) distance. That is, we assume there exist L_Y and L_W such that for any $u, u' \in \mathcal{U}$:

$$\text{TV}(\mathcal{L}_{Y|U=u}, \mathcal{L}_{Y|U=u'}) \leq L_Y|u - u'|,$$
$$\text{TV}(\mathcal{L}_{W|U=u}, \mathcal{L}_{W|U=u'}) \leq L_W|u - u'|,$$

with $\text{TV}(\mathcal{L}_1, \mathcal{L}_2) := \sup_{F \in \mathcal{F}} |\mathcal{L}_1(F) - \mathcal{L}_2(F)|$.

To understand the role of Asm. 4.7 in controlling the error $|P(\tilde{y}|\tilde{x}_i, U \in \mathcal{U}_j) - P(\tilde{y}|U \in \mathcal{U}_j)|$, note that for the small bin $\mathcal{U}_j$ and any $u \in \mathcal{U}_j$, Asm. 4.7 means that the distance

$$\text{TV}(\mathcal{L}_{Y|U \in \mathcal{U}_j}, \mathcal{L}_{Y|U=u})$$

is small. Similarly, it also means that the distance

$$\text{TV}(\mathcal{L}_{Y|X \in \mathcal{X}_i, U \in \mathcal{U}_j}, \mathcal{L}_{Y|X \in \mathcal{X}_i, U=u})$$

is small. Since we have $\mathcal{L}_{Y|X \in \mathcal{X}_i, U=u} = \mathcal{L}_{Y|U=u}$ under $\mathbb{H}_0 : X \perp\!\!\!\perp Y|U$, by applying the triangle inequality, we can control the distance $\text{TV}(\mathcal{L}_{Y|X \in \mathcal{X}_i, U \in \mathcal{U}_j}, \mathcal{L}_{Y|U \in \mathcal{U}_j})$ and hence control the term $|P(\tilde{y}|\tilde{x}_i, U \in \mathcal{U}_j) - P(\tilde{y}|U \in \mathcal{U}_j)|$.

The following example shows that Asm. 4.7 can hold for a wide range of ANMs.

Example 4.8 (ANM with TV smoothness)**.** Suppose that Y, U, W satisfy the ANM, *i.e.*, $Y = g_Y(U) + N_Y, W = g_W(U) + N_W$, with N_Y, N_W being the noises. If g_Y, g_W are differentiable, and the probability densities of N_Y, N_W have absolutely integrable derivatives, then the maps $u \mapsto \mathcal{L}_{Y|U=u}$ and $u \mapsto \mathcal{L}_{W|U=u}$ are Lipschitz continuous with respect to TV distance.

Under Asm. 4.7, we can control the discretization error as long as the length of each bin is sufficiently small.

Proposition 4.9. *Suppose that $\mathcal{U}$ is a compact interval, and that $\{\mathcal{U}_j\}_{l_U}$ is a partition of $\mathcal{U}$ satisfying*

$$len(\mathcal{U}_j) \leq \frac{1}{2}min\{\epsilon/L_Y, \epsilon/L_W\}, \qquad (4)$$

for every bin $\mathcal{U}_j$. Then, under Asm. 4.7 and $\mathbb{H}_0$, for any i, we have that:

$$|P(\tilde{y}|\tilde{x}_i, U \in \mathcal{U}_j) - P(\tilde{y}|U \in \mathcal{U}_j)| \leq \epsilon,$$
$$|P(\tilde{W}|\tilde{x}_i, U \in \mathcal{U}_j) - P(\tilde{W}|U \in \mathcal{U}_j)| \leq \epsilon,$$

and therefore $E_y(i) \leq \epsilon, E_w(i) \leq \epsilon$ for any i.

According to Prop. 4.9, when $\mathcal{U}$ is a compact interval, we can (implicitly) partition $\mathcal{U}$ into sufficiently refined bins, by setting the bin number large enough. However, when $\mathcal{U}$ is unbounded, it is intractable to partition it into finite bins with finite length. Therefore, to extend the above result to $\mathcal{U} = \mathbb{R}^d$, we additionally assume $\mathcal{L}_{U|X}$ to be tight, meaning that for large values of u's, $p(u|x)$ is close to zero. The formal definition of a tight distribution is given below:

Definition 4.10 (Tight distribution)**.** The marginal distribution $\mathcal{L}_U$ is said to be tight if $\forall \epsilon > 0, \exists t_0, \forall t \geq t_0, \mathcal{L}(|U| \geq t) \leq \epsilon$. The conditional distribution $\mathcal{L}_{U|X}$ is said to be uniformly tight in $[a, b]$ if $\forall \epsilon > 0, \exists t_0, \forall t \geq t_0, \forall x \in [a, b], \mathcal{L}(|U| \geq t|x) \leq \epsilon$.

Now we introduce the tightness condition below:

Assumption 4.11. We assume the conditional distribution $\mathcal{L}_{U|X}$ is uniformly tight for any compact interval $[a, b]$.

To understand how Asm. 4.11 can be applied in the case when $\mathcal{U} = \mathbb{R}^d$, note that we can write $p(\tilde{y}|\tilde{x}_i)$ as:

$$p(\tilde{y}|\tilde{x}_i) = \underbrace{p\left(\tilde{y}||U| \leq t_0, \tilde{x}_i\right) p\left(|U| \leq t_0|\tilde{x}_i\right)}_{\text{compact}}$$
$$+ \underbrace{p\left(\tilde{y}||U| > t_0, x\right) p\left(|U| > t_0|\tilde{x}_i\right)}_{\text{non-compact}}.$$

As the t_0 goes to infinity, the "non-compact" term approaches 0 and hence the error terms $\{E_y(i)\}_i, \{E_w(i)\}_i$ are only accounted for by the "compact" term.

The following example shows that Asm. 4.11 can easily hold when the distribution of the noise is tight.

Example 4.12 (ANM with tightness)**.** Suppose that X and U satisfy the ANM, *i.e.*, $X = g_X(U) + N_X$, with N_X being the noise. If g_X is invertible and the distribution of N_X is tight (e.g., distributions from the exponential family), then $\mathcal{L}_{U|X}$ is uniformly tight in any compact interval $[a, b]$ with $-\infty < a < b < \infty$.

Under Asm. 4.11, we have the following result for controlling the error induced by the non-compact part of $\mathcal{U}$.

Lemma 4.13. *Suppose $\mathcal{L}_{U|X}$ is uniformly tight in $[a, b]$, and the map $x \mapsto \mathcal{L}_{U|X=x}$ is L_U-Lipschitz continuous wrt TV distance. Suppose $\{\mathcal{X}_i\}_{l_X}$ is a partition of $[a, b]$ satisfying*

$$len(\mathcal{X}_i) \leq \epsilon/(2L_U). \qquad (5)$$

Then, $\forall \epsilon > 0, \exists t_0$ such that $P(\|U\|_2 > t_0|\tilde{x}_i) < \epsilon$ for any i.

Combining Prop. 4.9, we then have the following result for discretization error control when $\mathcal{U} = \mathbb{R}^d$:

Proposition 4.14. *Suppose $\mathcal{U} = \mathbb{R}^d$ and assumptions in Prop. 4.9 and Lem. 4.13 hold. Denote $\{\mathcal{U}_j\}_{l_U}$ and $\{\mathcal{X}_i\}_{l_X}$*

as partitions of $[-t_0, t_0]$ and $[a, b]$, respectively, s.t. (4) and (5) hold. Then, for the corresponded discretized variables $\tilde{U}, \tilde{X}$, it holds that $E_y(i) \leq \epsilon$ and $E_w(i) \leq \epsilon$ for any i.

4.3. Hypothesis test

In this section, we present our hypothesis-testing procedure. Specifically, we adopt the test statistic proposed by (Miao et al., 2018) and analyze its asymptotic distribution under the null hypothesis. Additionally, we investigate the behavior of the statistic under the alternative hypothesis and establish the asymptotic power of our test.

We first introduce our test statistic. The idea is that we can assess whether $\mathbb{H}_0$ is correct by looking at the residue of the least-square solution of Eq. (1). If $\mathbb{H}_0$ is correct, then the linear relation in Eq. (1) holds, hence the least-square residue should be close to zero. Specifically, we let

$$q := P(\tilde{y}|\tilde{X})^\top, \quad Q := P(\tilde{W}|\tilde{X}).$$

Using the Maximum Likelihood Estimation (MLE)[5], we can obtain estimators $(\hat{q}, \hat{Q})$ satisfying that:

$$n^{1/2}(\hat{q} - q) \to N(0, \Sigma) \text{ in distribution,}$$
$$\hat{Q} \to Q, \hat{\Sigma} \to \Sigma \text{ in probability,}$$

with $\hat{\Sigma}, \Sigma$ both positive definite. Then, we have the least-square residue of regressing $\hat{\Sigma}^{-1/2}\hat{q}$ on $\hat{\Sigma}^{-1/2}\hat{Q}$ to be:

$$\xi := \{I - \hat{\Sigma}^{-1/2}\hat{Q}^\top(\hat{Q}\hat{\Sigma}^{-1}\hat{Q}^\top)^{-1}\hat{Q}\hat{\Sigma}^{-1/2}\}\hat{\Sigma}^{-1/2}\hat{q}.$$

Following (Miao et al., 2018), we adopt the squared residue $T := n\xi^\top\xi$ as our test statistic.

Next, we discuss the asymptotic distribution of T. Suppose that W and X are respectively discretized into $l_W = l$ and $l_X = l + r$ bins, with r a positive number. Then, as n, l both go to infinity, the statistic $T_{n,l}$ follows an asymptotic chi-squared distribution χ_r^2, where the degree of freedom r is determined by the additional levels that $\tilde{X}$ has beyond $\tilde{W}$. Formally, denote $z_{1-\alpha}$ as the $(1-\alpha)$-th quantile of χ_r^2, we have the following result:

Theorem 4.15. *Suppose that Asm. 4.1, 4.7, 4.11 hold, then under $\mathbb{H}_0$, it holds that:*

$$T_{n,l} \to \chi_r^2 \text{ in distribution}$$

as $n, l \to \infty$. As a result, for any $\alpha \in (0, 1)$, we have:

$$\lim_{n,l\to\infty} P(T_{n,l} > z_{1-\alpha}) = \alpha.$$

That means, our procedure has uniform asymptotic level α.

[5]Please refer to Appx. E.1 for details.

In general, we cannot expect that the test has power against alternatives to $\mathbb{H}_0$ without further restrictions on the alternative distribution. This is analogous to other conditional independence tests (Zhang et al., 2012; Runge, 2018; Christgau et al., 2023) that are used for causal discovery. Therefore, to establish the power of our test, we put the following restriction on the alternative distribution:

Assumption 4.16. We consider the alternative $\mathbb{H}_1$, such that the following integral equation has no solution:

$$p(y|x) = \int h(y, w)p(w|x)dw. \tag{6}$$

Remark 4.17. Empirically, we find that Asm. 4.16 can hold in many cases under $X \not\perp\!\!\!\perp Y|U$, as shown in Appx. E.3. Besides, we can prove that $\mathbb{H}_1$ indeed does not cover all cases where $\mathbb{H}_0$ fails, that is, there exists distributions satisfying $X \not\perp\!\!\!\perp Y|U$ but do not satisfy Asm. 4.16. See Appx. E.3 for details.

To understand the role of Asm. 4.16 in controlling the type II error, we first note that the residues in T should deviate from zero under the alternative model, which requires that Eq. (1) has no solution. Intuitively, this means the variability of $P(\tilde{y}|\tilde{X})$ can not be fully explained $P(\tilde{W}|\tilde{X})$, which then corresponds to Asm. 4.16 for continuous variables. On the other hand, we can show in Appx. E.4 that the solution in Eq. (6) exists under $\mathbb{H}_0$. Therefore, we can employ it to effectively distinguish $\mathbb{H}_0$ from the alternative.

Under this assumption, we can establish the asymptotic power for our testing procedure.

Theorem 4.18. *Suppose that Asm. 4.1, 4.7, 4.11, 4.16 hold, then for any $\alpha \in (0, 1)$, it holds that:*

$$\lim_{n,l\to\infty} P(T_{n,l} > z_{1-\alpha}) = 1.$$

In other words, the limiting power of our test is one.

5. Experiment

In this section, we apply our method to synthetic data and causal discovery on a real-world dataset.

Compared baselines. We compare our method to the following baselines: **i)** the Kernel-based Conditional Independence test (KCI) (Zhang et al., 2012) that tested the null hypothesis of $X \perp\!\!\!\perp Y|W$ using kernel matrices; **ii)** the Integrated Conditional Moment test (ICM) that tested whether the conditional moment equation $\mathbb{E}[Y - h(W)|X]$ has a solution; and **iii)** the procedure described by (Miao et al., 2018) (Miao for simplicity) that was designed for causal discovery over discrete variables using the proxy variable.

Evaluation metrics. We report the rate of type I and type II errors. The type I error is the mistaken rejection of a null
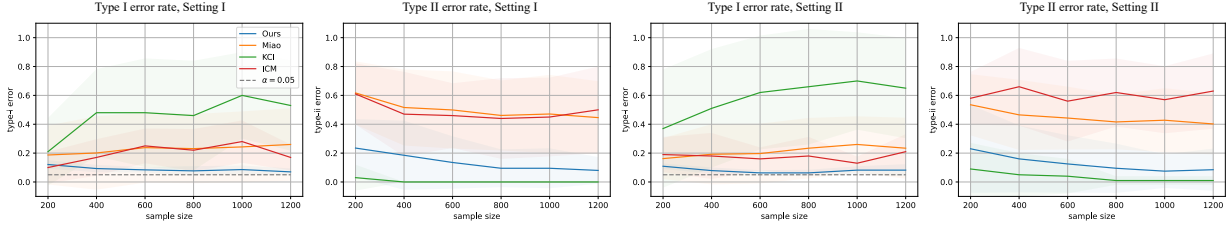
Figure 2. Type I and type II error rates of our testing procedure and baseline methods. Note that for a valid testing procedure, the type I error should be close to the significant level α (the dashed line), and the type II error should be close to zero.

hypothesis when it is true, and the type II error means the failure to reject a null hypothesis when it is false. Ideally, for a valid testing procedure, the type I error rate should be close to the significance level α, and the type II error rate should be close to zero.

Implementation details. We set the significance level α to 0.05. We set the bin numbers of our method to $l_X = 14$, $l_W = 12$, respectively. For KCI and ICM, we adopt the implementations provided in the causallearn[6] and mliv[7] packages, respectively. For the procedure described by Miao, we implement the R code released in the paper and set $l_X = 3, l_W = 2$ by default.

5.1. Synthetic data

Data generation. We consider two settings for data generation, using the confounding graph in Fig. 1 (a) and the mediation graph in Fig. 1 (b), respectively. For both settings, we follow the ANM to generate V, *i.e.*, $V = f_V(\mathbf{PA}_V) + N_V$, for each $V \in \{X, Y, U, W\}$, with $\mathbf{PA}_V$ and N_V denoting the associated parent nodes and the noise, respectively. For each V, the function f_V is randomly chosen from $\{\text{linear}, \tanh, \sin, \text{sigmoid}\}$, and the noise distribution is random selected from $\{\text{Gaussian}, \text{uniform}, \text{exponential}, \text{gamma}\}$. To remove the effect of randomness, we repeat the process 20 times. For each time, we generate 100 replications under $\mathbb{H}_0$ and $\mathbb{H}_1$, respectively, to record the type I and type II error rates.

Type I and type II error rates. In Fig. 2, we report the performance of our testing procedure and baseline methods. As shown, the type I error rate of our method approximates the desired level $\alpha = 0.05$, while other methods exhibit considerably higher rates. Besides, our type II error rate is getting close to 0 as n increases. To explain, note that conditioning on the proxy variable W of KCI cannot eliminate the confounding/mediation bias. The higher type I error rate of ICM may come from its limitation in hypothesis testing for only parametric models, as well as the estimation error.
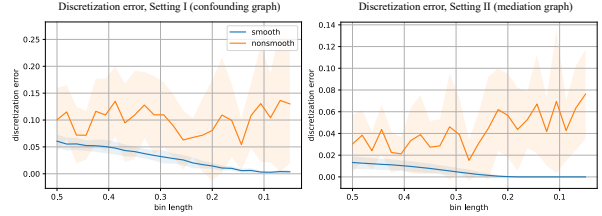
Figure 3. Discretization error with respect to the bin length. Left: setting I with the confounding graph Fig. 1 (a). Right: setting II with the mediation graph Fig. 1 (b). For both settings, the blue line corresponds to the case where the smoothness condition, *i.e.*, Asm. 4.7, holds, whereas the orange line corresponds to the case where the data is generated from a nonsmooth model.

Besides, this procedure suffers from a large type II error rate. Finally, the additional errors of Miao's procedure may come from the violation of the linear equation in Eq. (1) under $\mathbb{H}_0$ arising from coarse discretization.

Controlling the discretization error. To demonstrate the effectiveness of our method in controlling the discretization error, we choose a pair of (i, j) and compute the $e_{\text{dis}} := |P(\tilde{y}|\tilde{x}_i, \mathcal{U}_j) - P(\tilde{y}|\mathcal{U}_j)|$ with different bin lengths and present the results with blue lines in Fig. 3. As we can see, when the smoothness condition in Asm. 4.7 is satisfied, the e_{dis} quickly decreases to zero as the bin length becomes small. This result aligns with our analysis in Prop. 4.9.

As a contrast, for the nonsmooth model (orange lines) in which Asm. 4.7 is violated, it is interesting to observe that the e_{dis} is consistently large for all bin lengths. This may be due to the large variation of the conditional distribution $\mathcal{L}_{Y|U=u}$ in the same discretization bin.

Influence of the bin number and the sample size. In Fig. 4, we report the type I and type II error rates for different bin numbers $l_W = \{4, 6, 8, 10, 12, 14\}$ (with $l_X = 14$ constantly) and sample sizes $n = \{200, 400, 600, 800, 1200\}$. As shown, as the bin number and the sample size grow, the type I error rate approaches the significance level α, and the type II error rate becomes close to zero. These results align with the level and power of our testing procedure, as claimed
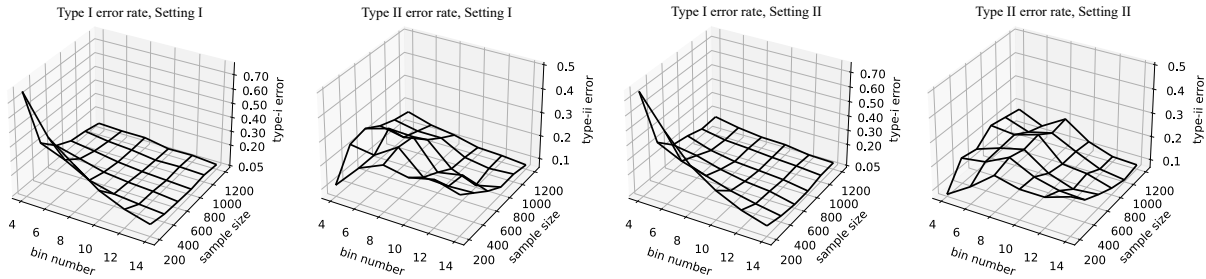
Figure 4. Type I and type II error rates with respect to the bin number and sample size. We consider two settings for data generating, with setting I (left) using the confounding graph in Fig. 1 (a), and setting II (right) using the mediation graph in Fig. 1 (b).

in Thm. 4.15 and Thm. 4.18, respectively. Moreover, for small sample sizes (e.g., $n = 200$), we can observe that the type II error rate slightly increases as the bin number grows, which may be due to insufficient samples in each bin to estimate the discrete probability.

5.2. Application to sepsis disease

In this section, we apply the proposed procedure to sepsis disease, which is a systemic inflammatory response to infection (Matot & Sprung, 2001) and the leading cause of death in the intensive care unit (ICU) (O'Brien et al., 2007).

Specifically, we examine two pairs of causal relationships: Vancomycin → White Blood Cell count (WBC) and Morphine → WBC. Vancomycin and morphine are two common medications for sepsis, with vancomycin being used to control the bacterial infection (Moellering, 2006) and morphine being used for pain management (McQuay, 1999), and the WBC is a common biomarker used to evaluate patient's response to medications (Castelli et al., 2006).

In both pairs, it is believed that the confounder U is accounted for by the patient's health status, which can influence the medication received and the WBC at the same time. To adjust for the latent confounding, we use blood pressure as the proxy variable (W).

Golden standards from RCTs. We take the results from two random controlled trials (RCTs) as the golden standards for causal discovery. Specifically, (Rosini et al., 2015) found that vancomycin can causally influence the WBC; whereas (Anand et al., 2004) found that morphine has no causal effect on WBC. In other words, the correlation between morphine usage and an increase in WBC (as found in (Zhang et al., 2018)) is likely due to the confounding bias.

Data extraction from MIMIC database. We consider the Medical Information Mart for Intensive Care (MIMIC III) database (Johnson et al., 2016), which contains the electronic health records for patients in the ICU. From MIMIC III, we extracted data for 3,251 patients diagnosed with sepsis during their ICU stays. Among these patients, 1,888
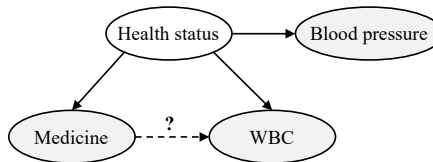

Figure 5. Illustration of causal discovery in sepsis disease. Observable variables are marked in gray. WBC denotes the count of White Blood Cells, which is a common biomarker used to assess patient's response to medicines. By using the blood pressure as the proxy variable (W) for the health status (U), our goal is to determine whether the edge Medicine $\longrightarrow$ WBC exists or not.

Table 1. Identified causal relationships in sepsis disease. We use the results from two random controlled trials (RCTs) (Anand et al., 2004; Rosini et al., 2015) as the golden standards. Results that contradict the golden standards are highlighted in red.

Method	Vancomycin→WBC	Morphine→ WBC
RCT (gold. stand.)	√	×
KCI	√	√
ICM	√	√
Miao	×	√
Ours	√	×

individuals received vancomycin, and 559 individuals received morphine.

Results analysis. In Tab. 1, we report the identified causal relationships using our procedure and baseline methods. To highlight, we marked the results that are inconsistent with the result from RCT in red. As shown, in both pairs, our procedure identifies causal relations that well align with the RCT results, which demonstrate the effectiveness and utility of our method.

In contrast, there is a large discrepancy between the results of the baselines and the RCTs. For instance, KCI and ICM detect that morphine has a causal influence on WBC, which can be a false positive result due to the failure to adjust for the confounding bias; and Miao's procedure fails to reject

the null hypothesis that vancomycin has no causal effect on WBC, which may be attributed to the loss of ability in the proxy variable without careful discretization.

6. Conclusions and Discussions

In this paper, we propose a proxy-based hypothesis testing procedure for bivariate causal discovery. Our procedure involves the discretization of continuous variables, for which we control the discretization errors and establish the identifiability of a linear system. We then construct testing statistics based on the residue of linear regression, which enjoys asymptotically level and power.

Limitations and future work. Our method relies on discretization and hence may lose its power when the dimension of the conditional variable is high. Indeed, such a curse of dimensionality issue is common in all conditional independence (CI) tests (Ramdas et al., 2015). For alleviation, we will investigate theories on high dimensional discretization (Van Handel 2014, Lem. 5.12). Besides, we are also interested in plugging our test into multivariate causal discovery algorithms such as the FCI (Spirtes et al., 2000), to enhance their ability for structure identification.

Acknowledgements

Impact Statement

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

References

Ai, C., Sun, L.-H., Zhang, Z., and Zhu, L. Testing unconditional and conditional independence via mutual information. *Journal of Econometrics*, 2022.

Anand, K., Hall, R. W., Desai, N., Shephard, B., Bergqvist, L. L., Young, T. E., Boyle, E. M., Carbajal, R., Bhutani, V. K., Moore, M. B., et al. Effects of morphine analgesia in ventilated preterm neonates: primary outcomes from the neopain randomised trial. *The Lancet*, 363(9422): 1673–1682, 2004.

Banerjee, S. and Roy, A. *Linear algebra and matrix analysis for statistics*. Crc Press, 2014.

Blöbaum, P., Janzing, D., Washio, T., Shimizu, S., and Schölkopf, B. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pp. 900–909. PMLR, 2018.

Casella, G. and Berger, R. L. *Statistical inference*. Cengage Learning, 2021.

Castelli, G., Pognani, C., Cita, M., Stuani, A., Sgarbi, L., Paladini, R., et al. Procalcitonin, c-reactive protein, white blood cells and sofa score in icu: diagnosis and monitoring of sepsis. *Minerva anestesiologica*, 72(1/2):69, 2006.

Chang, J. T. and Pollard, D. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.

Christgau, A. M., Petersen, L., and Hansen, N. R. Nonparametric conditional local independence testing. *The Annals of Statistics*, 51(5):2116–2144, 2023.

Deaner, B. Many proxy controls. *arXiv preprint arXiv:2110.03973*, 2021.

Doran, G., Muandet, K., Zhang, K., and Schölkopf, B. A permutation-based kernel conditional independence test. In *UAI*, pp. 132–141, 2014.

Dougherty, J., Kohavi, R., and Sahami, M. Supervised and unsupervised discretization of continuous features. In *Machine learning proceedings 1995*, pp. 194–202. Elsevier, 1995.

Dukes, O., Shpitser, I., and Tchetgen Tchetgen, E. J. Proximal mediation analysis. *Biometrika*, 110(4):973–987, 2023.

Duong, B. and Nguyen, T. Bivariate causal discovery via conditional divergence. In *Conference on Causal Learning and Reasoning*, pp. 236–252. PMLR, 2022.

Durrett, R. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Ghassami, A., Shpitser, I., and Tchetgen, E. T. Proximal causal inference with hidden mediators: Frontdoor and related mediation problems. *arXiv preprint arXiv:2111.02927*, 2021.

Ghassami, A., Ying, A., Shpitser, I., and Tchetgen, E. T. Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 7210–7239. PMLR, 2022.

Ghassami, A., Shpitser, I., and Tchetgen, E. T. Partial identification of causal effects using proxy variables. *arXiv preprint arXiv:2304.04374*, 2023.

Glymour, C., Zhang, K., and Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Gyenis, Z., Hofer-Szabó, G., and Rédei, M. Conditioning using conditional expectations: the borel–kolmogorov paradox. *Synthese*, 194(7):2595–2630, 2017.

Hicks, R. and Tingley, D. Causal mediation analysis. *The Stata Journal*, 11(4):605–619, 2011.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Hu, S., Chen, Z., Partovi Nia, V., Chan, L., and Geng, Y. Causal inference and mechanism clustering of a mixture of additive noise models. *Advances in neural information processing systems*, 31, 2018.

Huang, T.-M. Testing conditional independence using maximal nonlinear conditional correlation. 2010.

Jager, K., Zoccali, C., Macleod, A., and Dekker, F. Confounding: what it is and how to deal with it. *Kidney international*, 73(3):256–260, 2008.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Liu, M., Sun, X., Hu, L., and Wang, Y. Causal discovery from subsampled time series with proxy variables. *Advances in neural information processing systems*, 2023.

Margaritis, D. Distribution-free learning of bayesian network structure in continuous domains. In *AAAI*, volume 5, pp. 825–830, 2005.

Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., and Muandet, K. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International Conference on Machine Learning*, pp. 7512–7523. PMLR, 2021.

Matot, I. and Sprung, C. L. Definition of sepsis. *Intensive care medicine*, 27(14):S3–S9, 2001.

McQuay, H. Opioids in pain management. *The Lancet*, 353 (9171):2229–2232, 1999.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Miao, W., Hu, W., Ogburn, E. L., and Zhou, X.-H. Identifying effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association*, pp. 1–15, 2022.

Mitrovic, J., Sejdinovic, D., and Teh, Y. W. Causal inference via kernel deviance measures. *Advances in neural information processing systems*, 31, 2018.

Moellering, R. C. Vancomycin: a 50-year reassessment, 2006.

Mukherjee, S., Asnani, H., and Kannan, S. Ccmi: Classifier based conditional mutual information estimation. In *Uncertainty in artificial intelligence*, pp. 1083–1093. PMLR, 2020.

Neykov, M., Balakrishnan, S., and Wasserman, L. Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177, 2021.

Nuttall, A. Some integrals involving the qm function. *Naval Underwater Systems Center (NUSC) Technical Report*, 4297, 1972.

O'Brien, J. M., Ali, N. A., Aberegg, S. K., and Abraham, E. Sepsis. *The American journal of medicine*, 120(12): 1012–1022, 2007.

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

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):5–31, 2018.

Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. 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.

Reick, N., Wiewel, F., Bartler, A., and Yang, B. Estimation of bivariate structural causal models by variational gaussian process regression under likelihoods parametrised by normalising flows. *arXiv preprint arXiv:2109.02521*, 2021.

Rojas-Carulla, M., Baroni, M., and Lopez-Paz, D. Causal discovery using proxy variables. *arXiv preprint arXiv:1702.07306*, 2017.

Rosini, J. M., Laughner, J., Levine, B. J., Papas, M. A., Reinhardt, J. F., and Jasani, N. B. A randomized trial of loading vancomycin in the emergency department. *Annals of Pharmacotherapy*, 49(1):6–13, 2015.

Runge, J. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 938–947. PMLR, 2018.

Sgouritsa, E., Janzing, D., Hennig, P., and Schölkopf, B. Inference of cause and effect with unsupervised inverse regression. In *Artificial intelligence and statistics*, pp. 847–855. PMLR, 2015.

Shao, J. *Mathematical statistics*. Springer Science & Business Media, 2003.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

Strobl, E. V., Zhang, K., and Visweswaran, S. Approximate kernel-based conditional independence tests for fast nonparametric causal discovery. *Journal of Causal Inference*, 7(1):20180017, 2019.

Sverdrup, E. and Cui, Y. Proximal causal learning of heterogeneous treatment effects. *arXiv preprint arXiv:2301.10913*, 2023.

Tagasovska, N., Chavez-Demoulin, V., and Vatter, T. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *International Conference on Machine Learning*, pp. 9311–9323. PMLR, 2020.

Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X., and Miao, W. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Tu, R., Zhang, K., Kjellström, H., and Zhang, C. Optimal transport for causal discovery. *arXiv preprint arXiv:2201.09366*, 2022.

Van Handel, R. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Warren, A. Wasserstein conditional independence testing. *arXiv preprint arXiv:2107.14184*, 2021.

Williams, D. *Probability with martingales*. Cambridge university press, 1991.

Zhang, K. and Hyvarinen, A. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Zhang, R., Meng, J., Lian, Q., Chen, X., Bauman, B., Chu, H., Segura, B., and Roy, S. Prescription opioids are associated with higher mortality in patients diagnosed with sepsis: a retrospective cohort study using electronic health records. *PLoS One*, 13(1):e0190362, 2018.

Appendix

A. Notations & definitions

We introduce some extra notations and definitions that will be used in the proof. f_X is the probability density function (pdf) of X, and similarly for the joint density $f_{X,Y}$ and conditional density $f_{X|Y}$. We say a pdf $f : \mathcal{X} \mapsto [0,1]$ is absolutely integrable if $\int_{\mathcal{X}} |f(x)| dx < \infty$. For clarity, we sometimes use the notations $\{\mathcal{X}_i\}_{i=1}^I, \{\mathcal{Y}_j\}_{j=1}^J, \{\mathcal{U}_k\}_{k=1}^K, \{\mathcal{W}_l\}_{l=1}^L$ to denote partitions of $\mathcal{X}, \mathcal{Y}, \mathcal{U}$, and $\mathcal{W}$, respectively. For the matrix A, we use $\det(A)$ to denote its determinant. Besides, we use $\ker(A) := \{x : Ax = 0\}$ and $\mathrm{im}(A) := \{y : y = Ax \text{ for some } x\}$ to denote its kernel space and the image space, respectively.

B. Discussion with works on causal direction identification

There have been extensive studies in bivariate causal discovery that aim to identify the direction of causal influence ($X \to Y$ v.s. $X \leftarrow Y$). Within this area, one branch of these works distinguished cause from effect by exploiting asymmetry properties in functional causal models, which include but are not limited to the linear non-Gaussian model (Shimizu et al., 2006), additive noise model (Hoyer et al., 2008; Hu et al., 2018; Tu et al., 2022), and post non-linear model (Zhang & Hyvarinen, 2012; Reick et al., 2021). Other methods are based on the independence of cause and mechanism (Sgouritsa et al., 2015; Blöbaum et al., 2018; Tagasovska et al., 2020), and the stability of causation (Janzing & Schölkopf, 2010; Mitrovic et al., 2018; Duong & Nguyen, 2022). However, these works focused on learning causal direction, rather than establishing causal relations that can be more fundamental in real applications. Besides, the identifiability in these works relied on the strong ignorability condition, which may not hold in many scenarios. **In this paper**, we consider the causal relation identification problem in the presence of latent variables.

C. Discretization under completeness

C.1. Details of Exam. 4.2: ANM with completeness

Example C.1. (ANM when U is a mediator) Suppose that X, U, W satisfy the ANM, *i.e.*, $U = g_U(X) + N_U, W = g_W(U) + N_W$, with N_X, N_W being the noises. If N_X, N_W both have non-zero characterization functions, then the conditional distributions $\mathcal{L}_{U|X}$ and $\mathcal{L}_{W|X}$ are complete.

Proof. We first show $\mathcal{L}_{U|X}$ is complete. Denote $h(x) := f_{N_U}(-x)$, for any bounded function m, we have

$$\mathbb{E}[m(u)|x] = \int m(u) f_{N_U}(u - g_U(x)) du = \int m(u) h(g_U(x) - u) du = \{m * h\}(g_U(x)),$$

where $*$ denotes the convolution operator. Therefore, $\mathbb{E}[m(u)|x] = 0$ a.s. means $m * h = 0$ a.s. Applying the convolution theorem, we have $\hat{m} \cdot \hat{h} = 0$ a.s., where $\hat{m}, \hat{h}$ are the Fourier transforms of m and h, respectively. Since the characterization function of N_U is non-zero, we have $\hat{h} > 0$ and hence $\hat{m} = 0$ a.s. Applying the inverse Fourier transform, we have $m = 0$ a.s., which means $\mathcal{L}_{U|X}$ is complete.

Similarly, we can prove $\mathcal{L}_{W|U}$ is complete. Next, given that $\mathcal{L}_{U|X}$ and $\mathcal{L}_{W|U}$ are both complete, we show $\mathcal{L}_{W|X}$ is also complete. Specifically, for any bounded function m, we have

$$\mathbb{E}[m(w)|x] = \int m(w) f_{W|X}(w|x) dw = \iint m(w) f_{W|U}(w|u) dw f_{U|X}(u|x) du = 0 \text{ a.s.}$$

if and only if $\int m(w) f_{W|U}(w|u) dw = 0$ a.s., because $\mathcal{L}_{U|X}$ is complete. As $\mathcal{L}_{W|U}$ is also complete, we have $\int m(w) f_{W|U}(w|u) dw = 0$ a.s. if and only if $m(w) = 0$ a.s. Therefore, we have $\mathbb{E}[m(w)|x] = 0$ a.s. if and only if $m(w) = 0$ a.s., which concludes the proof. $\square$

Example 4.2. (ANM when U is a confounder) Suppose that X, U, W satisfy the ANM, *i.e.*, $X = g_X(U) + N_X, W = g_W(U) + N_W$, with N_X, N_W being the noises. If g_X is invertible with a non-zero derivative, and N_X, N_W have non-zero characterization functions, then the conditional distributions $\mathcal{L}_{U|X}$ and $\mathcal{L}_{W|X}$ are both complete.

Proof. Similar to the proof in the Exam. C.1, we first have the completeness of $\mathcal{L}(W|U)$. Therefore, if we can prove the completeness for $\mathcal{L}(U|X)$, we can have the completeness of $E\{g(w)|x\}$, following the procedure in the mediation model.

To show $\mathcal{L}(W|U)$ is complete, note that we have $f_{U|X}(u|x) = f_{N_X}(x - g_X(u))|g'_X(u)|$. Therefore, the completeness of $\mathcal{L}(U|X)$ means, for any bounded function m,

$$\int m(u)|g'_X(u)|f_{N_X}(x - g_X(u))du = 0 \text{ a.s. if and only if } m(u) = 0 \text{ a.s.}$$

Denote $m_1(u) := m(u)|g'_X(u)|$. We first show that $m_1(u) = 0$ a.s. For this purpose, let $t = g_X(u)$ and hence $u = g_X^{-1}(t)$. Further, let $h(t) = g_X^{-1}(t)$ with $h'(t) = \frac{1}{g'_X(g_X^{-1}(t))} \neq 0$. Then, we have

$$\int m_1(u)f_{N_X}(x - g_X(u))du = \int m_1(h(t))h'(t)f_{N_X}(x - t)dt.$$

Since N_X has a non-zero characteristic function, we first have $\int m_1(h(t))h'(t)f_{N_X}(x - t)dt = 0$ a.s. implies $m_1(h(t))h'(t) = 0$ a.s. Since $h'(t) \neq 0$, we then have $m_1(h(t)) = m_1(g_X^{-1}(t)) = 0$, a.s. Since u takes values in $\mathbb{R}$, we have the range of g_X^{-1} is $\mathbb{R}$, and therefore $m_1(u) = 0$, a.s. Since $g'_X(u) \neq 0$, we have $m(u) = 0$ a.s., which concludes the proof. $\square$

C.2. Proof of Prop. 4.4: Rank-preserving discretization

Proposition 4.4. *Suppose Asm. 4.1 holds. Then, for any discretization $\tilde{W}$ of W, there exists a discretization $\tilde{X}$ of X such that the matrix $P(\tilde{W}|\tilde{X})$ has full row rank. Similarly, there also exists a discretization $\tilde{U}$ of U such that the matrix $P(\tilde{W}|\tilde{U})$ is invertible.*

Proof. We first show that for any partition $\{\mathcal{W}_l\}_{l=1}^L$, there exists a partition $\{\mathcal{X}_i\}_{i=1}^I$ (with $I \geq L$) such that the matrix $P(\tilde{W}|\tilde{X})$ has rank L. The proof contains two parts: in **part I**, we first show that for any $\{\mathcal{W}_l\}_{l=1}^L$, the functions $\{f_{W|X}(W \in \mathcal{W}_l|x)\}_{l=1}^L$ are linear independent. Then, in **part II**, we use this conclusion to construct I points $\{x_i\}_{i=1}^I$ and therefore the partition $\{\mathcal{X}_i\}_{i=1}^I$ such that the matrix $P(\tilde{W}|\tilde{X})$ has rank L.

Part I. We first show that for any $\{\mathcal{W}_l\}_{l=1}^L$, the functions $\{f_{W|X}(W \in \mathcal{W}_l|x)\}_{l=1}^L$ are linear independent. Prove by contradiction. Suppose that there exists a partition $\{\mathcal{W}_l\}_{l=1}^L$ such that $\{f_{W|X}(W \in \mathcal{W}_l|x)\}_{l=1}^L$ are linear dependent. That is, there are L numbers $\{a_l\}_{l=1}^L$ that are not all zero such that $\sum_{l=1}^L a_l f_{W|X}(W \in \mathcal{W}_l|x) = 0$. Let $g(w)$ be the piece-wise constant function defined by $g(w) := a_l, w \in \mathcal{W}_l$, we then have $\int_{\mathcal{W}} g(w)f_{W|X}(w|x) = 0$ for all x. Since $g(w) \neq 0$, this contradicts with the bounded completeness of $\mathcal{L}_{W|X}$ as assumed in Asm. 4.1.

We then show for any $\{\mathcal{W}_l\}_{l=1}^L$, the functions $\{F_l(x) := F_{X|W}(x|W \in \mathcal{W}_l)\}_{l=1}^L$ are linear independent. To see this, note that $f_{X|W}(x|W \in \mathcal{W}_l) = \frac{f_{W|X}(W \in \mathcal{W}_l|x)f_X(x)}{P(W \in \mathcal{W}_l)}$, therefore it is easy to obtain the linear independence of $\{f_{X|W}(x|W \in \mathcal{W}_l)\}_{l=1}^L$ through the independence of $\{f_{W|X}(W \in \mathcal{W}_l|x)\}_{l=1}^L$. Besides, since $f_{X|W}(x|W \in \mathcal{W}_l) = F'_{X|W}(x|W \in \mathcal{W}_l)$, we can also obtain the linear independence of $\{F_{X|W}(x|W \in \mathcal{W}_l)\}_{l=1}^L$.

Part II. Given that $\{F_l(x)\}_{l=1}^L$ are linear independent, we now show there exists I points $\{x_i\}_{i=1}^I$ such that the matrix $P(\tilde{W}|\tilde{X})$ has rank L. For this purpose, we only have to show that there exists I points such that the L vectors $F_1(\boldsymbol{x}), ..., F_L(\boldsymbol{x})$ are linear independent, with $F_l(\boldsymbol{x}) := [F_l(x_1), ..., F_l(x_I)]^T$ for each $l = 1, .., L$. This is because, if this holds, we can complete the proof by showing that

$$\text{rank}(P(\tilde{W}|\tilde{X})) = \text{rank}(P(\tilde{X}|\tilde{W})) = \text{rank}([F_1(\boldsymbol{x}), ..., F_L(\boldsymbol{x})]) = L.$$

We first show the case where $I = L$. Denote the matrix $M_L = [F_1(\boldsymbol{x}), ..., F_L(\boldsymbol{x})]$. We show $F_1(\boldsymbol{x}), ..., F_L(\boldsymbol{x})$ are linear independent by showing $\det(M_L) \neq 0$. Prove by induction over L:

(1) Base: when $L = 1$, we have $\exists x_1, F_1(x_1) \neq 0$. Therefore, we have $\det(M_1) = F_1(x_1) \neq 0$.

(2) Induction hypothesis: assume that for any $L - 1$ linear independent functions $F_1, ..., F_{L-1}$, there exists $L - 1$ points $x_1, ..., x_{L-1}$ such that $\det(M_{L-1}) \neq 0$.

(3) Induction: show that for any L linear independent functions $F_1, ..., F_L$, there exists L points $x_1, ..., x_L$ such that $\det(M_L) \neq 0$. To this end, we first have the Laplace expansion of the matrix determinant

$$\det(M_L)(x_L) = \sum_{l=1}^L (-1)^{1+l}\det(A_l^L)F_l(x_L), \tag{7}$$

where $A_{L,l}$ denotes the matrix obtained by removing the L-th row and l-th column from M_L. That is, $A_{L,l} = [F_1(\boldsymbol{x}'), ..., F_{l-1}(\boldsymbol{x}'), F_{l+1}(\boldsymbol{x}'), ..., F_L(\boldsymbol{x}')]$, where $F_l(\boldsymbol{x}') = [F_l(x_1), ..., F_l(x_{L-1})]^T$. Then, since $\{F_1, ..., F_{l-1}, F_{l+1}, ..., F_L\}$ are $L-1$ linear independent functions, according to the induction hypothesis, there exists $L-1$ points $x_1, ..., x_{L-1}$ such that $\det(A_{L,l}) \neq 0$. That is, at least one of the cofactors $(-1)^{1+l}\det(A_l^L)$ in the expansion is not zero. Finally, since $F_1, ..., F_L$ are linear independent, there exists x_L such that $\det(M_L)(x_L) \neq 0$ for x_L, which concludes the induction.

For cases where $I > L$, We have $F_1(\boldsymbol{x}), ..., F_L(\boldsymbol{x})$ (where $F_l(\boldsymbol{x}) := [F_l(x_1), ..., F_l(x_I)]^T$) are linear independent because $F_1(\boldsymbol{x}'), ..., F_L(\boldsymbol{x}')$ (where $F_l(\boldsymbol{x}') := [F_l(x_1), ..., F_l(x_L)]^T$) are linear independent.

Below, we show for any partition $\{\mathcal{W}_l\}_{l=1}^L$, there exists a partition $\{\mathcal{U}_k\}_{k=1}^K$ (with $K = L$) such that the matrix $P(\tilde{W}|\tilde{U})$ is invertible. We prove this by showing that under Assumption 4.1, $\mathcal{L}_{W|U}$ is complete. Then, repeating the proof above, we have $P(\tilde{W}|\tilde{U})$ is invertible. To show this, according to the completeness of $\mathcal{L}_{W|X}$, for any bounded function g, we have

$$\int g(w)f(w|x)dw = \int \left(\int g(w)f(w|u)dw \right) f(u|x)du = 0 \text{ a.s.} \tag{8}$$

if and only if $g(w) = 0$ a.s. Moreover, this is also equivalent to $\int g(w)f(w|u)dw = 0$ a.s., since $\mathcal{L}_{U|X}$ is complete. Therefore, we have $\int g(w)f(w|u)dw = 0$ a.s. if and only if $g(w) = 0$ a.s., which concludes the proof. $\square$

D. Discretization error analysis

D.1. Definition of conditioning on a zero-probability event

The TV smoothness condition in Asm. 4.7 involves conditioning on the zero-probability event $U = u$, which may cause ambiguity without careful definition (Gyenis et al., 2017). In the following, we use the regular conditional probability to provide a rigorous definition. We start from the standard Kolmogorov definition of conditional expectation.

Definition D.1 (Condition expectation, (Williams 1991, Def. 9.2)). Let $(\Omega, \mathcal{F}, P)$ be the probability space, X a random variable with $E(|X|) < \infty$, and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$. Then there exists a random variable Y such that:

1. Y is $\mathcal{G}$-measurable,
2. $E[|Y|] < \infty$,
3. for every set $G \in \mathcal{G}$, we have:

$$\int_G Y dP = \int_G X dP.$$

Moreover, if Y' is another random variable with these properties then $Y' = Y$, a.e., that is $P(\{w : Y(w) = Y'(w)\}) = 1$. A random variable Y with the above properties is called a version of the conditional expectation $E[X|\mathcal{G}]$ of X given $\mathcal{G}$, and we write $Y = E[X|\mathcal{G}]$, a.s.

Remark D.2. The existence of Y satisfying the above properties is guaranteed by the Radon-Nikodym theorem (Durrett 2019, Thm. A.4.8). $Y = dX/dP$ is also called the Radon-Nikodym derivative.

We often write $E[X|Z]$ in short for $E[X|\sigma(Z)]$, which is the conditional expectation of X given $\sigma(Z)$. For $F \in \mathcal{F}$, define the conditional probability $P(F|Z)$ as a version of $E[\mathbb{1}_F|Z]$. In this regard, the map $P(\cdot, \cdot) : \Omega \times \mathcal{F} \mapsto [0, 1)$ defined by $P(\omega, F) := P(F|Z)(\omega)$ is a measurable function that maps a tuple (ω, F) to a real value in $[0, 1)$.

Then, according to the disintegration theorem (Chang & Pollard 1997, Def. 1 and Thm. 1), there exists a **regular conditional probability** $P(\cdot, \cdot) : \Omega \times \mathcal{F} \mapsto [0, 1)$ satisfying that:

1. for $F \in \mathcal{F}$, the function $\omega \mapsto P(\omega, F)$ is a version of $P(F|\mathcal{G})$,
2. for almost every ω, the map $F \mapsto P(\omega, F)$ is a probability measure on $\mathcal{F}$.

Remark D.3. Note that for those $\{\omega\}$ where $P(\omega, F)$ is not well-defined (which is a set of measure zero), one can redefine $P(\omega, F) := \lim_{r \to 0} \frac{1}{|B_r(\omega)|} \int_{B_r(\omega)} P(t, F)dt$ (where $B_r(\omega)$ is a ball with radius r), so that $P(\omega, F)$ is now well-defined strictly everywhere.

Equipped with the above definition of regular conditional probability (rcp), we then define the the conditional probability $P(X \in F|Z(\omega) = z)$ given the zero-probability event $Z = z$ as the rcp $P(\omega, F)$ (and similarly for the conditional probabilities $P(Y|U = u)$ and $P(W|U = u)$). Thanks to the above properties 1-2, our definition is well-defined.

D.2. Details of Exam. 4.8: ANM with TV smoothness

Example 4.8. Suppose that Y, U, W satisfy the ANM, *i.e.*, $Y = g_Y(U) + N_Y$, $W = g_W(U) + N_W$, with N_Y, N_W being the noises. If g_Y and g_W are differentiable, and the probability densities of N_Y and N_W have absolutely integrable derivatives, then the maps $u \mapsto \mathcal{L}_{Y|U=u}$ and $u \mapsto \mathcal{L}_{W|U=u}$ are Lipschitz continuous with respect to TV distance.

Proof. We show $u \mapsto \mathcal{L}_{Y|U=u}$ is Lipschitz continuous and the Lipschitzness of $u \mapsto \mathcal{L}_{W|U=u}$ can be similarly derived. Specifically, according to the definition of TV distance, we have:

$$
\begin{aligned}
\mathrm{TV}(\mathcal{L}_{Y|U=u}, \mathcal{L}_{Y|U=u'}) &= \frac{1}{2}\int |f_{Y|U}(y|u) - f_{Y|U}(y|u')|dy, \\
&\stackrel{(1)}{=} \frac{1}{2}\int |f_{N_Y}(y - g_Y(u)) - f_{N_Y}(y - g_Y(u'))|dy, \\
&\stackrel{(2)}{=} \int \frac{1}{2}|f'_{N_Y}(y - g_Y(u_c)) \cdot g'_Y(u_c)| \cdot |u - u'|dy.
\end{aligned}
$$

for some $u_c \in (u, u')$, where "(1)" is due to ANM that can write $f_{Y|U}(y|u)$ as $f_{N_Y}(y - g_Y(u))$ and "(2)" is due to the mean value theorem. Let $L_Y := \frac{1}{2}\int |f'_{N_Y}(y - g_Y(u_c))g'_Y(u_c)|dy$, we have $\mathrm{TV}(\mathcal{L}_{Y|U=u}, \mathcal{L}_{Y|U=u'}) \leq L_Y|u - u'|$, which concludes the proof. □

D.3. Proof of Prop. 4.9: Discretization error control under smoothness

Proposition 4.9. *Suppose that $\{\mathcal{U}_k\}_{k=1}^K$ is a partition of $\mathcal{U}$, and that for every bin $\mathcal{U}_k$,*

$$
len(\mathcal{U}_k) \leq \frac{1}{2}min\{\epsilon/L_Y, \epsilon/L_W\}.
$$

Then, under Asm. 4.7 and $\mathbb{H}_0 : X \perp\!\!\!\perp Y|U$, we have:

$$
|p(\tilde{y}|\tilde{u}_k) - p(\tilde{y}|\tilde{x}, \tilde{u}_k)| \leq \epsilon, \quad |p(\tilde{w}|\tilde{u}_k) - p(\tilde{w}|\tilde{x}, \tilde{u}_k)| \leq \epsilon.
$$

Proof. First recall that we use the notations $\{\mathcal{X}_i\}_{i=1}^I, \{\mathcal{Y}_j\}_{j=1}^J, \{\mathcal{U}_k\}_{k=1}^K, \{\mathcal{W}_l\}_{l=1}^L$ to denote the partitions of $\mathcal{X}, \mathcal{Y}, \mathcal{U}$, and $\mathcal{W}$, respectively. Below, we show $|P(Y \in \mathcal{Y}_j|X \in \mathcal{X}_i, U \in \mathcal{U}_k) - P(Y \in \mathcal{Y}_j|U \in \mathcal{U}_k)| \leq \epsilon$ and the bounding of $|P(W \in \mathcal{W}_l|X \in \mathcal{X}_i, U \in \mathcal{U}_k) - P(W \in \mathcal{W}_l|U \in \mathcal{U}_k)|$ can be similarly derived.

The proof consists of two parts. In **part I**, we show the smoothness of $u \mapsto \mathcal{L}_{Y|U=u}$, which together with $len(\mathcal{U}_k) \leq \epsilon/(2L_Y)$, implies $\mathrm{TV}(\mathcal{L}_{Y|U=u}, \mathcal{L}_{Y|U\in\mathcal{U}_k}) \leq \epsilon/2$. Then, in **part II**, we show the control $|P(Y \in \mathcal{Y}_j|X \in \mathcal{X}_i, U \in \mathcal{U}_k) - P(Y \in \mathcal{Y}_j|U \in \mathcal{U}_k)|$ via that of $\mathrm{TV}(\mathcal{L}_{Y|U=u}, \mathcal{L}_{Y|U\in\mathcal{U}_k})$.

Part I. We show $\mathrm{TV}(\mathcal{L}_{Y|U=u}, \mathcal{L}_{Y|U\in\mathcal{U}_k}) \leq \epsilon/2$. Note that we first have

$$
\mathcal{L}_{Y|U\in\mathcal{U}_k} = \fint_{\mathcal{U}_k} \mathcal{L}_{Y|U=u}d\mathcal{L}_U(u) := \frac{\int_{\mathcal{U}_k} \mathcal{L}_{Y|U=u}d\mathcal{L}_U(u)}{\int_{\mathcal{U}_k} d\mathcal{L}_U(u)},
$$

where $\fint$ denotes the average integral (Warren, 2021). To see this, according to the definition of $\mathcal{L}$, for any $\mathcal{Y}_j \in \mathcal{Y}$, the LHS equals $\mathcal{L}_{Y|U\in\mathcal{U}_k}(\mathcal{Y}_j) = P(Y \in \mathcal{Y}_j|U \in \mathcal{U}_k)$, and the RHS equals $\int_{\mathcal{U}_k} \mathcal{L}_{Y|U=u}(\mathcal{Y}_j)d\mathcal{L}_U(u)/P(U \in \mathcal{U}_k) = P(Y \in \mathcal{Y}_j, U \in \mathcal{U}_k)/P(U \in \mathcal{U}_k) = \mathrm{LHS}$.

Then, due to the assumed TV smoothness, for any $u, u' \in \mathcal{U}_K$, we have $\mathrm{TV}(\mathcal{L}_{Y|U=u}, \mathcal{L}_{Y|U=u'}) := \sup_{\mathcal{Y}_j \in \mathcal{Y}}|\mathcal{L}_{Y|U=u}(\mathcal{Y}_j) - \mathcal{L}_{Y|U=u'}(\mathcal{Y}_j)| \leq L_Y|u - u'| \leq \epsilon/2$. Now, for **any** $\mathcal{Y}_j \subseteq \mathcal{Y}$, we have (note that u is a fixed point and u' varies in the integration domain):

$$
|\mathcal{L}_{Y|U=u}(\mathcal{Y}_j) - \mathcal{L}_{Y|U\in\mathcal{U}_k}(\mathcal{Y}_j)| = \left|\mathcal{L}_{Y|U=u}(\mathcal{Y}_j) - \fint_{\mathcal{U}_k} \mathcal{L}_{Y|U=u'}(\mathcal{Y}_j)d\mathcal{L}_U(u')\right|
$$

16

$$= \left| \fint_{\mathcal{U}_k} \mathcal{L}_{Y|U=u}(\mathcal{Y}_j) d\mathcal{L}_U(u') - \fint_{\mathcal{U}_k} \mathcal{L}_{Y|U=u'}(\mathcal{Y}_j) d\mathcal{L}_U(u') \right| \qquad \# \fint_{\mathcal{U}_k} d\mathcal{L}_U(u') = 1$$

$$\leq \fint_{\mathcal{U}_k} \left| \mathcal{L}_{Y|U=u}(\mathcal{Y}_j) - \mathcal{L}_{Y|U=u'}(\mathcal{Y}_j) \right| d\mathcal{L}_U(u') \qquad \# \text{ Triangle inequity}$$

$$\leq \fint_{\mathcal{U}_k} \epsilon/2 \, d\mathcal{L}_U(u') = \epsilon/2,$$

which means $\sup_{\mathcal{Y}_j \in \mathcal{Y}} |\mathcal{L}_{Y|U=u}(\mathcal{Y}_j) - \mathcal{L}_{Y|U\in\mathcal{U}_k}(\mathcal{Y}_j)| =: \mathrm{TV}(\mathcal{L}_{Y|U=u}, \mathcal{L}_{Y|U\in\mathcal{U}_k}) \leq \epsilon/2$. In a similar way, we can prove $\mathrm{TV}(\mathcal{L}_{Y|X\in\mathcal{X}_i,U=u}, \mathcal{L}_{Y|X\in\mathcal{X}_i,U\in\mathcal{U}_k}) \leq \epsilon/2$. To be specific, under $\mathbb{H}_0 : X \perp\!\!\!\perp Y|U$, we have $\mathcal{L}_{Y|X\in\mathcal{X}_i,U=u} = \mathcal{L}_{Y|U=u}$ for any $\mathcal{X}_i \subseteq \mathcal{X}$. As a result, the map $u \mapsto \mathcal{L}_{Y|X\in\mathcal{X}_i,U=u}$ is also L_Y-Lipschitz continuous wrt TV distance. The rest of the proof is the same as above.

Part II. We show $|P(Y \in \mathcal{Y}_j|X \in \mathcal{X}_i, U \in \mathcal{U}_k) - P(Y \in \mathcal{Y}_j|U \in \mathcal{U}_k)| \leq \epsilon$. Specifically, since $\mathrm{TV}(\mathcal{L}_{Y|U=u}, \mathcal{L}_{Y|U\in\mathcal{U}_k}) \leq \epsilon/2$, we have $|P(Y \in \mathcal{Y}_j|U = u) - P(Y \in \mathcal{Y}_j|U \in \mathcal{U}_k)| \leq \epsilon/2$. Since $\mathrm{TV}(\mathcal{L}_{Y|X\in\mathcal{X}_i,U=u}, \mathcal{L}_{Y|X\in\mathcal{X}_i,U\in\mathcal{U}_k}) \leq \epsilon/2$, we also have $|P(Y \in \mathcal{Y}_j|X \in \mathcal{X}_i, U = u) - P(Y \in \mathcal{Y}_j|X \in \mathcal{X}_i, U \in \mathcal{U}_k)| \leq \epsilon/2$. According to the triangle inequality, we have

$$|P(Y \in \mathcal{Y}_j|U \in \mathcal{U}_k) - P(Y \in \mathcal{Y}_j|X \in \mathcal{X}_i, U \in \mathcal{U}_k)|$$
$$\leq |P(Y \in \mathcal{Y}_j|U \in \mathcal{U}_k) - P(Y \in \mathcal{Y}_j|U = u)| + |P(Y \in \mathcal{Y}_j|U = u) - P(Y \in \mathcal{Y}_j|X \in \mathcal{X}_i, U \in \mathcal{U}_k)|$$
$$= |P(Y \in \mathcal{Y}_j|U \in \mathcal{U}_k) - P(Y \in \mathcal{Y}_j|U = u)| + |P(Y \in \mathcal{Y}_j|X \in \mathcal{X}_i, U = u) - P(Y \in \mathcal{Y}_j|X \in \mathcal{X}_i, U \in \mathcal{U}_k)| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

where the 2nd equality is due to $P(Y \in \mathcal{Y}_j|U = u) = P(Y \in \mathcal{Y}_j|X \in \mathcal{X}_i, U = u)$ under $\mathbb{H}_0 : X \perp\!\!\!\perp Y|U$. $\qquad \square$

D.4. Details of Exam. 4.12: ANM with tightness

Example 4.12. (ANM when U is a confounder) Suppose that X and U satisfy the ANM, *i.e.*, $X = g_X(U) + N_X$, with N_X being the noise. If g_X is invertible and the distribution of N_X is tight, then $\mathcal{L}_{U|X}$ is uniformly tight in an arbitrary compact interval $[a, b]$.

Proof. Since $\mathcal{L}_{N_U}$ is tight, we first have $\forall \epsilon > 0, \exists t_0, \forall t \geq t_0, \int_t^\infty f_{N_U}(u)du \leq \epsilon$. Then we can obtain the uniform tightness of $\mathcal{L}_{U|X}$ by noting that $\forall \epsilon > 0$, we have $s_0 := g_X^{-1}(a - t_0)$ such that $\forall s \geq s_0$,

$$\int_{-\infty}^t f_{U|X}(u|x)du = -\int_{-\infty}^t f_{N_U}(x - g_X(u))d(x - g_X(u))$$

$$= \int_{x-g_X(t)}^{x-g_X(-\infty)} f_{N_U}(n)dn \leq \int_{x-g_X(t)}^\infty f_{N_U}(n)dn \leq \epsilon,$$

where we use the substitution $n = x - g_X(u)$, and the fact that $x - g_X(t) \geq x - a + t_0 \geq t_0$. $\qquad \square$

Example D.4 (ANM when U is a mediator)**.** Suppose that X and U satisfy the ANM, *i.e.*, $U = g_U(X) + N_U$, with N_U being the noise. If g_U is bounded in $[a, b]$ and the distribution of N_U is tight, then $\mathcal{L}_{U|X}$ is uniformly tight in $[a, b]$.

Proof. Since $\mathcal{L}_{N_U}$ is tight, we first have $\forall \epsilon > 0, \exists t_0, \forall t \geq t_0, \int_t^\infty f_{N_U}(u)du \leq \epsilon$. Besides, we denote m_0 as the upper bound of $g_U(x)$, such that $|g_U(x)| \leq m_0$ for all $x \in [a, b]$. Then we can obtain the uniform tightness of $\mathcal{L}_{U|X}$ by noting that $\forall \epsilon > 0$, we have $s_0 := t_0 + m_0$ such that $\forall s \geq s_0$,

$$\int_s^\infty f_{U|X}(u|x)du = \int_s^\infty f_{N_U}(u - g_X(x))du = \int_{s-g_U(x)}^\infty f_{N_U}(u)du \leq \epsilon,$$

where we use the fact that $s \geq t_0 + h_0 \geq t_0 + |g_U(x)| \geq t_0 + g_U(x)$ therefore $s - g_U(x) \geq t_0$. $\qquad \square$

D.5. Proof of Prop. 4.14: Discretization error control under tightness

Proposition 4.14. *Suppose that $\mathcal{L}_{U|X}$ is uniformly tight in $[a, b]$, and that the map $x \mapsto \mathcal{L}_{U|X=x}$ is L_U-Lipschitz continuous wrt TV distance. Suppose that $\{\mathcal{X}_i\}_{l_X}$ is a partition of $[a, b]$ satisfying $len(\mathcal{X}_i) \leq \epsilon/(2L_U)$. Then, $\forall \epsilon > 0, \exists t_0$, let $\mathcal{U}_{nc} := \{u : \|U\|_2 > t_0\}$, we have $p(\tilde{u}_{nc}|\tilde{x}_i) \leq \epsilon$.*

Proof. We only show the proof of $P(Y \in \mathcal{Y}_j, \|U\|_2 > t_0 | X \in \mathcal{X}_i) \le \epsilon$, as the control of $P(W \in \mathcal{W}_l, \|U\|_2 > t_0 | X \in \mathcal{X}_i)$ can be similarly derived. We obtain this conclusion by showing $P(\|U\|_2 > t_0 | X \in \mathcal{X}_i) \le \epsilon$ and using the fact that

$$P(Y \in \mathcal{Y}_j, \|U\|_2 > t_0 | X \in \mathcal{X}_i) = P(Y \in \mathcal{Y}_j | \|U\|_2 > t_0, X \in \mathcal{X}_i) P(\|U\|_2 > t_0 | X \in \mathcal{X}_i) \le P(\|U\|_2 > t_0 | X \in \mathcal{X}_i) \le \epsilon.$$

Specifically, given that $\mathcal{L}_{U|X}$ is uniformly tight, we first have $\forall \epsilon/2 > 0, \exists t_0, \forall t \ge t_0, \forall x \in \mathcal{X}_i, P(\|U\|_2 > t | x) \le \epsilon/2$. Then, since $x \mapsto \mathcal{L}_{U|X=x}$ is L_U-Lipschitz, we have $\forall \epsilon/2 > 0, |P(|U| > t | x) - P(|U| > t | \mathcal{X}_i)| \le \epsilon/2$ for $x \in \mathcal{X}_i$. Applying the triangle inequality, we have $P(|U| > t | \mathcal{X}_i) \le \epsilon$, which concludes the proof. $\qquad\square$

E. Hypothesis-testing procedure

E.1. Maximum likelihood estimation

Consider discrete random variables X, Y that respectively take values $\{x_1, ..., x_I\}$ and $\{y_1, ..., y_J\}$. Suppose that we have n iid samples. Denote Z_{ij} the number of samples that take the value (x_i, y_j), and denote $Z_{i\cdot} := \sum_j Z_{ij}$. Then, the MLE estimator of conditional probability vector $P(y|X)$ is $\hat{P}(y|X) = [\hat{p}(y|x_1), ..., \hat{p}(y|x_I)]$, where $\hat{p}(y_j|x_i) = Z_{ij}/Z_{i\cdot}$.

The MLE estimator satisfies that $\hat{P}(y|X) \to P(y|X)$ in probability, and that $n^{1/2}(\hat{P}(y|X) - P(y|X)) \to N(0, \Sigma)$ in distribution. The covariance matrix is $\Sigma = \text{diag}[\frac{p(y_j|x_1)(1-p(y_j|x_1))}{p(x_1)}, ..., \frac{p(y_j|x_I)(1-p(y_j|x_I))}{p(x_I)}]$ and is positive definite. The corresponding estimator $\hat{\Sigma}$ that satisfies $\hat{\Sigma} \to \Sigma$ in probability can be obtained by replacing $p(y_j|x_i)$ with $\hat{p}(y_j|x_i)$, and replacing $p(x_i)$ with $\hat{p}(x_i) = Z_{i\cdot}/n$. These results are obtained by noting that Z_{ij} follows the multinomial distribution, and by applying the Fisher information matrix (in the sandwich form) and the Cramer-Rao lower bound (Casella & Berger 2021, Thm. 7.3.9).

E.2. Proof of Thm. 4.15: Asymptotic level of the test

Theorem 4.15. *Suppose that Asm. 4.1, 4.7, 4.11 hold, then under H_0, it holds that:*

$$T_{n,l} \to \chi_r^2 \text{ in distribution}$$

as $n, l \to \infty$. As a result, for any $\alpha \in (0, 1)$, we have:

$$\lim_{n,l\to\infty} P(T_{n,l} > z_{1-\alpha}) = \alpha.$$

In other words, the testing procedure we define has uniform asymptotic level α.

Proof. First recall that we use the notations $\{\mathcal{X}_i\}_{i=1}^I, \{\mathcal{Y}_j\}_{j=1}^J, \{\mathcal{U}_k\}_{k=1}^K, \{\mathcal{W}_l\}_{l=1}^L$ to denote the partitions of $\mathcal{X}, \mathcal{Y}, \mathcal{U}$, and $\mathcal{W}$, respectively. We denote $r := I - L$. We denote $q := P(\tilde{y}|\tilde{X})^T$ and $Q := P(\tilde{W}|\tilde{X})$.

The proof consists of two parts. In **part I**, we first show the asymptotic linearity between q^T and Q. Then, in **part II**, we analyze the asymptotic distribution of the test statistics.

Part I. In the following, we show the asymptotic linearity $q^T = P(\tilde{y}|\tilde{U})P(\tilde{W}|\tilde{U})^{-1}Q + e$ with $\lim_{L\to\infty} e = \mathbf{0}$.

We first consider cases where $\mathcal{U}$ is a compact interval. To show the asymptotic linearity, we first show:

$$q^T = P(\tilde{y}|\tilde{U})P(\tilde{U}|\tilde{X}) + e_1, \quad \lim_{L\to\infty} e_1 = \mathbf{0}, \tag{9}$$

$$Q = P(\tilde{W}|\tilde{U})P(\tilde{U}|\tilde{X}) + e_2, \quad \lim_{L\to\infty} e_2 = \mathbf{0}. \tag{10}$$

In this regard, under Asm. 4.1, we have:

$$q^T = P(\tilde{y}|\tilde{U})P(\tilde{W}|\tilde{U})^{-1}Q + e, \quad \lim_{L\to\infty} e = \lim_{L\to\infty}\{e_1 - P(\tilde{y}|\tilde{U})P(\tilde{W}|\tilde{U})^{-1}e_2\} = \mathbf{0},$$

which concludes **part I**. To show Eq. (9), note that we first have

$$q^T := P(\tilde{y}|\tilde{X}) = P(\tilde{y}_l|\tilde{U})P(\tilde{U}|\tilde{X}) + e_1, \quad e_1 = [\{P(\tilde{y}_l|\tilde{U}, \tilde{x}_i) - P(\tilde{y}_l|\tilde{U})\}P(\tilde{U}|\tilde{x}_i)]_{i=1}^I.$$

Then, according to Prop. 4.9 and the mean inequality, we have $\lim_{L \to \infty} e_1 = \mathbf{0}$. Similarly, we can show

$$Q = P(\tilde{W}|\tilde{U})P(\tilde{U}|\tilde{X}) + e_2, \text{ where } e_2 = [\{P(\tilde{W}|\tilde{U}, \tilde{x}_i) - P(\tilde{W}|\tilde{U})\}P(\tilde{U}|\tilde{x}_i)]_{i=1}^I,$$

and $\lim_{L \to \infty} e_2 = \mathbf{0}$.

We then consider the case where $\mathcal{U} = \mathbb{R}$. Still, we want to establish Eq. (9)- (10). We first derive the form of e_1. For a given t, divide $\mathcal{U}$ into $\mathcal{U}' := \{u : |u| > t\}$ and $\mathcal{U}'' := \mathcal{U} \backslash \mathcal{U}'$, and further partition $\mathcal{U}''$ into K bins $\{\mathcal{U}_k\}_{k=1}^K$. We then have

$$p(\tilde{y}|\tilde{x}_i) = P(\tilde{y}|\tilde{U}, \tilde{x}_i)P(\tilde{U}|\tilde{x}_i) + p(\tilde{y}|U \in \mathcal{U}', \tilde{x}_i)p(U \in \mathcal{U}'|\tilde{x}_i).$$

Therefore, we have:

$$P(\tilde{y}|\tilde{X}) = P(\tilde{y}|\tilde{U})P(\tilde{U}|\tilde{X}) + e_c + e_{nc},$$
$$e_c = \{[P(\tilde{y}|\tilde{U}, \tilde{x}_i) - P(\tilde{y}|\tilde{U})]P(\tilde{U}|\tilde{x}_i)\}_{i=1}^I,$$
$$e_{nc} = \{p(\tilde{y}, U \in \mathcal{U}'|\tilde{x}_i)p(U \in \mathcal{U}'|\tilde{x}_i)\}_{i=1}^I.$$

where e_c is the error in the compact interval $\mathcal{U}'$ and e_{nc} is the error in the non-compact $\mathcal{U}''$. We then have $\lim_{L \to \infty} e_1 = \lim_{L \to \infty} e_c + e_{nc} = \mathbf{0}$, since $\lim_{L \to \infty} e_c = \mathbf{0}$ by repeating the previous proof and $\lim_{L \to \infty} e_{nc} = \mathbf{0}$ can be shown in Prop. 4.14. Similarly, we have:

$$Q = P(\tilde{W}|\tilde{U})P(\tilde{U}|\tilde{X}) + e_2, \ e_2 = \{[P(\tilde{W}|\tilde{U}, \tilde{x}_i) - P(\tilde{W}|\tilde{U})]P(\tilde{U}|\tilde{x}_i)\}_{i=1}^I + \{P(\tilde{W}|U \in \mathcal{U}', \tilde{x}_i)P(U \in \mathcal{U}'|\tilde{x}_i)\}_{i=1}^I,$$

and $\lim_{t, L \to \infty} x_2 = \mathbf{0}$.

Part II. We analyze the asymptotic distribution of the statistics $T := n^{1/2}\xi^T\xi$. We first show $n^{1/2}\xi \to N(0, \Omega)$. Specifically, given that $\hat{Q} \to Q, \hat{\Sigma} \to \Sigma$ in probability, and that $n^{1/2}(\hat{q} - q) \to N(0, \Sigma)$ in distribution, apply Slutsky's theorem, we have:

$$n^{1/2}(\xi - \Omega\Sigma^{1/2}q_l) \to_d N(0, \Omega_l), \ \Omega := I_I - \Sigma^{-1/2}Q^T(Q\Sigma^{-1}Q^T)^{-1}Q\Sigma^{-1/2}. \tag{11}$$

Then, under $\mathbb{H}_0 : X \perp\!\!\!\perp Y|U$, according to the asymptotic linearity in **part I**, we have $\Omega\Sigma^{-1/2}q \to 0$ as $L \to \infty$. Applying Slutsky's theorem again, we have $n^{1/2}\xi \to N(0, \Omega)$ in distribution.

Then, we show $\text{rank}(\Omega) = r$. Specifically, we have $\text{rank}(Q) = L$ according to Prop. 4.4. Since $\Sigma^{-1/2}$ is invertible, we have $\text{rank}(Q\Sigma^{-1/2}) = L$. Thus, the idempotent matrix $M := \Sigma^{-1/2}Q^T(Q\Sigma^{-1}Q^T)^{-1}Q\Sigma_l^{-1/2}$ has rank L. According to Coro. 11.5 of (Banerjee & Roy, 2014), M has K eigenvalues equal to one and $I - L$ eigenvalues equal to zero. Thus, we have $\Omega = I - M$ (I denotes the identity matrix) being an idempotent matrix with rank $r = I - L$.

Finally, we show $T \to \chi_r^2$ in distribution. To be specific, given that $n^{1/2}\xi \to N(0, \Omega)$, we have $T := n\xi^T\xi \to N(0, \Omega)^T N(0, \Omega)$ according to Thm. 1.12 of (Shao, 2003). What is left is to show $N(0, \Omega)^T N(0, \Omega) = \chi_r^2$. To see this, note that we have the orthogonal decomposition $V\Omega V^T = \text{diag}(1, ..., 1, 0, ..., 0)$, where V is an orthogonal matrix and $\text{diag}(1, ..., 1, 0, ..., 0)$ is a diagonal matrix with r 1s and $(I - r)$ 0s in the diagonal. Therefore, we have $V N(0, \Omega) = N(0, \text{diag}(1, ..., 1, 0, ..., 0))$ and $N(0, \Omega)^T N(0, \Omega) = [V N(0, \Omega)]^T [V N(0, \Omega)] = \chi_r^2$. $\qquad \square$

E.3. Details of Rem. 4.17: Characterization of Asm. 4.16

The asymptotic power of our test (Thm. 4.18) is built on a local alternative $\mathbb{H}_1$ that the integral equation $f(y|x) = \int h(y, w)f(w|x)dw$ does not have a solution. Intuitively, $\mathbb{H}_1$ means that the correlation between X and Y can not be fully explained by that between X and W. Empirically, we find this condition can hold in many cases under $X \not\perp\!\!\!\perp Y|U$. Specifically, we conduct numerical experiments with the following models:

- Linear Gaussian model: $U = X + N_1, Y = aX + U + N_2, W = U + N_3, \ X, N_1, N_2, N_3 \sim N(0, 0.1)$, where a ranges from 1 to 10.

- Linear non-Gaussian model: $U = X + N_1, Y = aX + U + N_2, W = U + N_3, X \sim \text{Uniform}(-0.5, 0.5), N_1, N_2, N_3 \sim \text{Exp}(0.1)$, where a ranges from 1 to 10.
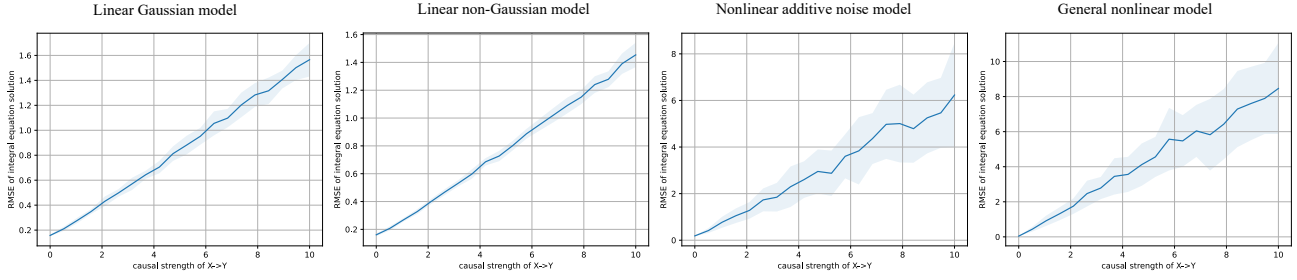
Figure 6. Insolubility of the integral equation when $X \not\perp\!\!\!\perp Y|U$. We consider four examples, namely the linear Gaussian model, the linear nonGaussian model, the nonlinear additive noise model, and the general nonlinear models. For each example, we report the averaged residual of the best fit of the integral equation $f(y|x) = \int h(y,w)f(w|x)dw$ wrt to the correlation strength of X and Y. As we can see, when X and Y become strongly correlated ($X \not\perp\!\!\!\perp Y|U$), the residual becomes non-zero, indicating the insolubility of the integral equation.

- ANM: $U = X^2 + N_1, Y = ae^X + U + N_2, W = \tan(U) + N_3, X \sim \text{Uniform}(-0.5, 0.5), N_1, N_2, N_3 \sim \text{Exp}(0.1)$, where a ranges from 1 to 10.

- General nonlinear model: $U = |X|N_1, Y = ae^{X+N_2} + U, W = \tan(U + N_3), X \sim \text{Uniform}(-0.5, 0.5), N_1, N_2, N_3 \sim \text{Exp}(0.1)$, where a ranges from 1 to 10.

For each model, we implement the Maximum Moment Restriction (MMR) with Gaussian kernels in the mliv package[8] to find the solution to the conditional moment equation $E[Y - h(W)|X] = 0$. The results are shown in Fig. 6. As we can see, when X and Y becomes strongly correlated (*i.e.*, $X \not\perp\!\!\!\perp Y|U$ holds), there is no valid solution to the conditional moment function, which indicates the insolubility of the integral equation.

Nonetheless, this condition may not constantly hold given $X \not\perp\!\!\!\perp Y|U$, which means $\mathbb{H}_1$ is indeed a local alternative. Specifically, denote $(\lambda_n, \phi_n, \psi_n)_{n=1}^{\infty}$ a singular value decomposition of K, we then have the following theorem:

Theorem E.1. *Suppose that the following conditions hold:*

1. *Completeness of $\mathcal{L}_{X|U}$.*
2. $\iint f(u|x)f(x|u)dxdu < \infty$.
3. $\int f^2(y|x)f(x)dx < \infty$.
4. $\sum_{n=1}^{\infty} \lambda_n^{-2} |\langle f(y|x), \psi_n\rangle| < +\infty$, *where* $\langle f(y|x), \psi_n\rangle := \int f^2(y|x)\psi_n(x)f(x)dx$.

Then there exists $h(u,y)$ such that $\int h(u,y)f(u|\mathrm{x}) = f(y|x)$ holds for all (x,y).

Remark E.2. These conditions should be compatible with our completeness condition in Asm. 4.1. Particularly, the completeness of $\mathcal{L}(U|X)$ and that of $\mathcal{L}(X|U)$ roughly means the U and X have similar levels of variation. To explain, note that if U and X are discrete, then we have $P(U|X)$ and $P(X|U)$ are both invertible.

Remark E.3. This result, which means Asm. 4.16 does not always hold, means it is reasonable to take Asm. 4.16 as a condition, instead of a theorem or a claim/hypothesis.

Proof. Simply follow the proof of Prop. 1 of (Miao et al., 2018) with the following replacements:

- replace $f(w|z, x)$ with $f(u|x)$,

- replace $f(z|w, x)$ with $f(x|u)$,

- replace $f(y|z, x)$ with $f(y|x)$,

- replace $\left(K_x, L^2(f(w|x)), L^2(f(z|x)), \lambda_{x,n}, \phi_{x,n}, \psi_{x,n}\right)$ with $\left(K, L^2(f(u)), L^2(f(x)), \lambda_n, \phi_n, \psi_n\right)$.

$\square$

---

[8]https://github.com/causal-machine-learning-lab/mliv

E.4. Proof of Thm. 4.18: Asymptotic power of the test

Theorem 4.18. *Suppose that Asm. 4.1, 4.7, 4.11, 4.16 hold, then for any $\alpha \in (0, 1)$, it holds that:*

$$\lim_{n,l \to \infty} P(T_{n,l} > z_{1-\alpha}) = 1.$$

In other words, the limiting power of our test is one.

Remark E.4. According to the following proof, we have that under the same condition, there exists a solution h such that $f(y|x) = \int h(y, w)f(w|x)dw$ under $\mathbb{H}_0$.

Proof. First recall that we use the notations $\{\mathcal{X}_i\}_{i=1}^I, \{\mathcal{Y}_j\}_{j=1}^J, \{\mathcal{U}_k\}_{k=1}^K, \{\mathcal{W}_l\}_{l=1}^L$ to denote the partitions of $\mathcal{X}, \mathcal{Y}, \mathcal{U}$, and $\mathcal{W}$, respectively. Moreover, we use l and L interchangeably. We denote $q := P(\tilde{y}|\tilde{X})^T$ and $Q := P(\tilde{W}|\tilde{X})$. The proof consists of two parts. In **part I**, we show that under $\mathbb{H}_1$, $q \notin \operatorname{im}(Q)$ as $L \to \infty$. Then, in **part II**, given that $q \notin \operatorname{im}(Q)$, we show $P(T > z_{1-\alpha})$ can be controlled to zero.

Part I. We show that under $\mathbb{H}_1$, $q \notin \operatorname{im}(Q)$ as $L \to \infty$. To see this, we first show that there exists L_0 such that as long as $I > K = L > L_0$, $q \notin \operatorname{im}(Q)$. Specifically, let

$$\operatorname{res}(h) := \iint |f(y|x) - \int h(y, w)f(w|x)dw|dyf(x)dx,$$

and let $h_0 := \arg\min_h \operatorname{res}(h)$. Since $f(y|x) \geq 0$ and $f(w|x) \geq 0$, since h_0 minimizes $\operatorname{res}(h)$, we have $h_0 \geq 0$ a.s. Besides, we have the following by taking M sufficiently large,

$$\int_{|x| \leq M} \int_{|y| \leq M} \left| f(y|x) - \int h(y, w)f(w|x)dw \right| dyf(x)dx > \frac{1}{2}\epsilon_0, \text{ for any } h \geq 0.$$

To see this, note that for any M, we have $\operatorname{res}(h) = \operatorname{res}_{\leq M}h + \operatorname{res}_{>M}(h)$, where

$$\operatorname{res}_{\leq M}(h) := \int_{|x| \leq M} \int_{|y| \leq M} |f(y|x) - \int h(y, w)f(w|x)dw|dyf(x)dx$$

$$\operatorname{res}_{>M}(h) := \left(\int_{|x|>M} \int_{|y| \leq M} + \int_{|x| \leq M} \int_{|y|>M} + \int_{|x|>M} \int_{|y|>M} \right) \left| f(y|x) - \int h(y, w)f(w|x)dw \right| dyf(x)dx.$$

For $\operatorname{res}_{>M}(h)$, we have

$$\operatorname{res}_{>M}(h) \geq \left(\int_{|x|>M} \int_{|y| \leq M} + \int_{|x| \leq M} \int_{|y|>M} + \int_{|x|>M} \int_{|y|>M} \right) f(y|x)f(x)dydx$$

$$= P(|X|> M) + P(|X| \leq M, |Y|> M) < \frac{1}{2}\epsilon_0$$

if M is sufficiently large. Then, without loss of generality, we have $\operatorname{res}(h) > \epsilon_0$ for some ϵ_0 and $M > 0$ such that $|X| \leq M, |Y| \leq M$. Since $f(y|x)$ is continuous, there exists M_0 such that $f(y|x) \leq M_0$ on $\{(x, y) : |x| \leq M, |y| \leq M\}$. Therefore, we have $0 \leq h_0 \leq M_0$, which means we only need to consider h with $0 \leq h \leq M_0$.

Further, there exists $M_w > 0$ such that

$$\int_{|x| \leq M} \int_{|y| \leq M} \left| f(y|x) - \int_{|w| \leq M_w} h(w, y)f(w|x)dw \right| dyf(x)dx > \frac{3}{4}\epsilon_0$$

for any $0 \leq h \leq M_0$. To see this, note that

$$\operatorname{res}(h) = \int_{|x| \leq M} \int_{|y| \leq M} \left| f(y|x) - \int_{|w| \leq M_w} h(y, w)f(w|x)dw \right| dyf(x)dx$$

$$+ \int_{|x| \leq M} \int_{|y| \leq M} \int_{|w|>M_w} h(w, y)f(w|x)dwdyf(x)dx$$

21

$$\leq \int_{|x|\leq M} \int_{|y|\leq M} |f(y|x) - \int_{|w|\leq M_w} h(y,w)f(w|x)dw|dyf(x)dx + M \cdot M_0 P(|W| > M_w, |X| \leq M)$$

$$\leq \int_{|x|\leq M} \int_{|y|\leq M} |f(y|x) - \int_{|w|\leq M_w} h(y,w)f(w|x)dw|dyf(x)dx + M \cdot M_0 P(|W| > M_w).$$

Therefore, for any $0 \leq h \leq M_0$, we have $\text{res}^{M_w}(h) > \frac{3}{4}\epsilon_0$. Note that

$$\text{res}^{M_w}(h) = \sum_{i=1}^{I} \sum_{j=1}^{J} \int_{\mathcal{X}_i} \int_{\mathcal{Y}_j} \left| f(y|x) - \int_{|w|\leq M_w} h(w,y)f(w|x)dw \right| dyf(x)dx.$$

Since $f(y|x)$ and $f(w|x)$ are respectively uniformly continuous in $\{(x,y) : |x| \leq M, |y| \leq M\}$ and $\{(x,w) : |x| \leq M, |w| \leq M_w\}$, there exists $I_{0,1} J_{0,1}$ such that for any $I > I_{0,1}, J > J_{0,1}$, we have

$$\left| \text{res}^{M_w}(h) - \sum_{i=1}^{I} \sum_{j=1}^{J} \left| f(\bar{y}_j|\bar{x}_i) - \int_{|w|\leq M_w} h(w,y)f(w|\bar{x}_i)dw \right| P(X \in \mathcal{X}_i)\text{len}(\mathcal{Y}_j) \right| < \frac{1}{4}\epsilon_0 \tag{12}$$

for any $\bar{x}_i \in \mathcal{X}_i, \bar{y}_j \in \mathcal{Y}_j$, and h that is constant for any w and $y \in \mathcal{Y}_j$. To see this, note that $f(y|x)$ and $f(w|x)$ are uniformly continuous. Then $\forall \epsilon > 0$, there exists $\delta > 0$ such that

$$|f(w_1|x_1) - f(w_2|x_2)| \leq \epsilon, \ |f(y_1|x_1) - f(y_2|x_2)| \leq \epsilon$$

for any $\sqrt{(w_1 - w_2)^2 + (x_1 + x_2)^2} < \delta$ and $\sqrt{(y_1 - y_2)^2 + (x_1 + x_2)^2} < \delta$. Therefore, as long as the partition satisfies that for each $i, j, \text{len}(\mathcal{X}_i) \leq \delta, \text{len}(\mathcal{Y}_j) < \delta$, then for any h that is constant for any w and $y \in \mathcal{Y}_j$, we have

$$\iint_{\mathcal{X}_i, \mathcal{Y}_j} |P(\bar{y}_j|\bar{x}_i) - f(y|x)|f(x)dxdy \leq \epsilon\delta,$$

$$\iint_{\mathcal{X}_i, \mathcal{Y}_j} \left| \int_{|w|\leq M_w} h(w,\bar{y}_j)f(w|\bar{x}_i)dw - \int_{|w|\leq M_w} h(w,y)f(w|x)dw \right| f(x)dxdy$$

$$\leq \iint_{\mathcal{X}_i, \mathcal{Y}_j} \int_{|w|\leq M_w} h(w,\bar{y}_j) |f(w|\bar{x}_i) - f(w|x)| dwf(x)dxdy \leq M_0 M_w \epsilon\delta,$$

for some $\bar{x}_i \in \mathcal{X}_i$ and $\bar{y}_j \in \mathcal{Y}_i$. By taking the ϵ small enough, we have the desired conclusion.

Similarly, there exists $I_{0,2}, J_{0,2}$ such that as long as $I > I_{0,2}, J > J_{0,2}$, we have

$$\left| \left| \sum_{i=1}^{I} \sum_{j=1}^{J} \left| \int_{\mathcal{X}_i} \int_{\mathcal{Y}_j} f(y|x) - \int_{|w|\leq M_w} h(w,y)f(w|x)dwdyf(x)dx \right| - \right.\right.$$

$$\left.\left. \sum_{i=1}^{I} \sum_{j=1}^{J} \left| f(\bar{y}_j|\bar{x}_i) - \int_{|w|\leq M_w} h(w,y)f(w|\bar{x}_i)dw \right| P(X \in \mathcal{X}_i)\text{len}(\mathcal{Y}_j) \right| < \frac{1}{4}\epsilon_0 \right. \tag{13}$$

for any h that is constant for any w and $y \in \mathcal{Y}_j$. In addition, by noting that

$$\int_{\mathcal{X}_i} \int_{\mathcal{Y}_j} \left(f(y|x) - \int h(w,y)f(w|x)dw \right) dyf(x)dx = P(\tilde{y}_j|\tilde{x}_i) - \int g_l(w)f(w|X \in \mathcal{X}_i)dwP(X \in \mathcal{X}_i),$$

where $g_l(w) := \int_{Y \in \mathcal{Y}_j} h(w,y)dy$, we have:

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \left| \int_{\mathcal{X}_i} \int_{\mathcal{Y}_j} \left(f(y|x) - \int h(w,y)f(w|x)dw \right) dyf(x)dx \right|$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} \left| P(\tilde{y}_j|\tilde{x}_i) - \int g_l(w)f(w|X \in \mathcal{X}_i)dw \right| P(X \in \mathcal{X}_i). \tag{14}$$

22

By combining Eq. (12), (13), (14), for any $I > \max(I_{0,1}, I_{0,2})$, $J > \max(J_{0,1}, J_{0,2})$, we have:

$$\left| \text{res}^{M_w}(h) - \sum_{i=1}^{I} \sum_{j=1}^{J} \left| P(\tilde{y}_j | \tilde{x}_i) - \int g_l(w) f(w | X \in \mathcal{X}_i) dw \right| P(X \in \mathcal{X}_i) \right| \leq \frac{1}{2} \epsilon_0.$$

Then for any I, J, there exist L_0 such that for any $L > L_0$, we have:

$$\left| \text{res}^{M_w}(h) - \sum_{i=1}^{I} \sum_{j=1}^{J} \left| P(\tilde{y}_j | \tilde{x}_i) - \sum_k P(\tilde{w}_k | \tilde{x}_i) g_l(w) f(w | X \in \mathcal{X}_i) dw \right| P(X \in \mathcal{X}_i) \right| \leq \frac{3}{4} \epsilon_0.$$

If $q = Q^T c$ for some $c = [[c_{1,1}, ..., c_{L,1}], ..., [c_{1,J}, ..., C_{L,J}]]^T$, with $c_{l,i,j} = c_{l,i',j}$, then we let $h(w, y) = c_{k,l}$ for $w \in \mathcal{W}_l, y \in \mathcal{Y}_j$ and have $p(\tilde{y}_j | \tilde{x}_i) = \sum_l p(\tilde{w}_k | \tilde{x}_i) c_{l,j}$. Therefore, we have $\sum_{i=1}^{I} \sum_{j=1}^{J} |p(\tilde{y}_j | \tilde{x}_i) - \sum_l p(\tilde{w}_l | \tilde{x}_i) g_l(\tilde{u}_l)| P(X \in \mathcal{X}_i) = 0$ for some h that is constant in each $\mathcal{Y}_j$. Therefore, we have $\text{res}^{M_w}(h) < \frac{3}{4}\epsilon_0$, which obtains the contradiction.

Part II. Given that $q \notin \text{im}(Q)$, we show $\lim_{n,L \to \infty} P_{\mathbb{H}_1}(T \leq z_{1-\alpha}) = 0$. To show this, we first discuss the distribution of T given $q \notin \text{im}(Q)$. First, we have $||\Omega \Sigma_y^{-1/2} q_y||_2^2 > 0$. To see this, we first have that $\ker(\Omega) = \text{im}(\Sigma^{-1/2} Q^T)$.

To see this, we have $\forall \vec{a} \in \text{im}(\Sigma^{-1/2} Q^T), \exists \vec{b}, s.t. \vec{a} = \Sigma^{-1/2} Q^T \vec{b}$. According to the definition of Ω in Eq. (11), we have $\Omega \vec{a} = \Omega \Sigma^{-1/2} Q^T \vec{b} = \vec{0}$, which means $\text{im}(\Sigma^{-1/2} Q^T) \subseteq \ker(\Omega)$. What is left is to show $\dim(\text{im}(\Sigma^{-1/2} Q^T)) = \dim(\ker(\Omega))$. First, we have $\dim(\text{im}(\Sigma^{-1/2} Q^T)) = \text{rank}(\Sigma^{-1/2} Q^T) = L(J-1)$. Since $\text{rank}(\Omega) = (I-L)(J-1)$ and $\Omega \in \mathbb{R}^{I(J-1) \times I(J-1)}$, we have $\text{nul}(\Omega) = I * (J-1) - (I-L)(J-1) = L(J-1)$ ($\text{nul}(\Omega)$ is the null space of Ω) and therefore $\dim(\ker(\Omega)) = L(J-1) = \dim(\text{im}(\Sigma^{-1/2} Q^T))$. Therefore, we have $\ker(\Omega) = \text{im}(\Sigma^{-1/2} Q^T)$.

Therefore, $q \notin \text{im}(Q)$ means $\Sigma^{-1/2} q \notin \text{img}(\Sigma^{-1/2} Q^T) = \ker(\Omega)$, and we have $\Omega \Sigma^{-1/2} q \neq \vec{0}$ and $||\Omega \Sigma^{-1/2} q||_2^2 > 0$. Second, let $\vec{c} := \Omega \Sigma^{-1/2} q$, we show T follows the noncentral chi square distribution $\chi^2(r, \lambda)$ with $\lambda = n ||\vec{c}||_2^2$ and $r = (I-L)(J-1)$. To be specific, we have $\sqrt{n}(\xi - \vec{c}) \to N(0, \Omega)$ and therefore $\sqrt{n} \xi \to N(\sqrt{n} \vec{c}, \Omega)$. Repeating the orthogonal decomposition trick used in the proof of Theorem 4.15, we have:

$$T := n\xi^T \xi \to N(\sqrt{n}\vec{c}, \Omega)^T N(\sqrt{n}\vec{c}, \Omega) = \{VN(\sqrt{n}\vec{c}, \Omega)\}^T \{VN(\sqrt{n}\vec{c}, \Omega)\}$$
$$= N(\sqrt{n}V\vec{c}, \text{diag}(1, ..., 1, 0, ..., 0))^T N(\sqrt{n}V\vec{c}, \text{diag}(1, ..., 1, 0, ..., 0)) = \chi^2(r, \lambda),$$

Given that T follows the noncentral $\chi^2(r, \lambda)$ distribution as $L \to \infty$, we then show $\lim_{n,L \to \infty} P_{\mathbb{H}_1}(T \leq z_{1-\alpha}) = 0$, which is equivalent to show that $\forall \epsilon > 0, \exists N, \forall n \geq N, P_{\mathbb{H}_1}(T_y \leq z_{1-\alpha}) \leq \epsilon$. In details, denote $F(x; r, \lambda)$ the cdf of $\chi^2(r, \lambda)$ and let $f(\lambda) := F(x; r, \lambda)$. We first show for fixed (x, r) (the x is $z_{1-\alpha}$ in our problem), $\lim_{\lambda \to \infty} f(\lambda) = 0$. According to (Nuttall, 1972), we can represent the cdf with Marcum Q function, i.e., $F(\lambda; x, r) = 1 - Q_{r/2}(\sqrt{\lambda}, \sqrt{x})$, where

$$Q_v(a, b) = 1 + H_v(a, b) + C_v(a, b)$$
$$H_v(a, b) = \frac{(a/b)^{1-v}}{2\pi} \exp\left(-\frac{1}{2}(a^2 + b^2)\right) \int_0^{2\pi} \frac{\cos(v-1)\theta - (a/b)\cos(v\theta)}{1 - 2(a/b)\cos\theta + (a/b)^2} \exp(ab\cos\theta) d\theta,$$
$$C_v(a, b) = \frac{\sin(v\pi)}{\pi} \exp\left(-\frac{1}{2}(a^2 + b^2)\right) \int_0^1 \frac{x^{v-1}(b/a)^{v-1}}{(a/b) + x} \exp(-\frac{ab}{2}(x + \frac{1}{x})) dx.$$

Therefore, to show $\lim_{\lambda \to \infty} f(\lambda) = 0$, we show $\lim_{\lambda \to \infty} Q_{k/2}(\sqrt{\lambda}, \sqrt{x}) = 0$, which needs $\lim_{a \to \infty} H_v(a, b) = 0$ and $\lim_{a \to \infty} C_v(a, b) = 0$. To see $\lim_{a \to \infty} H_v(a, b) = 0$, note that:

$$\lim_{a \to \infty} \frac{(a/b)^{1-v}}{2\pi} = 0, \quad \# v \text{ is a large number, so } 1 - v < 0,$$
$$\lim_{a \to \infty} \exp\left(-\frac{1}{2}(a^2 + b^2)\right) \exp(ab\cos\theta) = 0, \quad \# \text{because } \lim_{a \to \infty} a^2 - 2ab\cos\theta + b^2 = \infty,$$
$$\lim_{a \to \infty} \frac{\cos(v-1)\theta - (a/b)\cos(v\theta)}{1 - 2(a/b)\cos\theta + (a/b)^2} = 0 \quad \# \text{because } \lim_{a \to \infty} \frac{a}{1 - a + a^2} = 0.$$

23

Therefore, use Lebesgue's dominated convergence theorem (our integrand can be dominated by 1 when a is sufficiently large, and 1 is an integrable function) to exchange the limitation and integration, we have $\lim_{a \to \infty} H_v(a, b) = 0$. A similar process can show $\lim_{a \to \infty} C_v(a, b) = 0$. Now, with $\lim_{\lambda \to \infty} f(\lambda) = 0$, we have $\forall \epsilon > 0, \exists \Lambda, \forall \lambda > \Lambda, f(\lambda) := P_{\chi^2(r,\lambda)}(T_y \leq z_{1-\alpha}) \leq \epsilon$. Therefore, we have $\forall \epsilon > 0, \exists N := \frac{\Lambda}{||\vec{c}||_2^2}, \forall n \geq N, \lambda := n||\vec{c}||_2^2 \geq \Lambda$ and hence $P_{\mathbb{H}_1}(T_y \leq z_{1-\alpha}) \leq \epsilon$, which concludes the proof. $\qquad \square$