On the Sample Complexity of Conditional Independence Testing with Von Mises Estimator with Application to Causal Discovery

Fateme Jamshidi [1] **Luca Ganassali** [1] **Negar Kiyavash** [1]

Abstract

Motivated by conditional independence testing, an essential step in constraint-based causal discovery algorithms, we study the nonparametric Von Mises estimator for the entropy of multivariate distributions built on a kernel density estimator. We establish an exponential concentration inequality for this estimator. We design a test for conditional independence (CI) based on our estimator, called VM-CI, which achieves optimal parametric rates under smoothness assumptions. Leveraging the exponential concentration, we prove a tight upper bound for the overall error of VM-CI. This, in turn, allows us to characterize the sample complexity of any constraint-based causal discovery algorithm that uses VM-CI for CI tests. To the best of our knowledge, this is the first sample complexity guarantee for causal discovery for non-linear models and non-Gaussian continuous variables. Furthermore, we empirically show that VM-CI outperforms other popular CI tests in terms of either time, sample complexity, or both. This enhancement significantly improves the performance in structure learning as well.

1. Introduction

Causal discovery, the pursuit of uncovering the cause-and-effect relationships governing complex systems, has been the focus of research in machine learning, statistics, and various scientific domains over the past few decades. This is due to the extensive impact of causal inference, which enables us to make well-informed decisions and policies.

Current approaches for learning causal mechanisms with data can be divided into two categories: *score-based* (e.g., Chickering, 2002; Solus et al., 2021; Zheng et al., 2018; Zhu et al., 2019), and *constraint-based*, e.g., Peter-Clark (PC) algorithm (Spirtes et al., 2000) and grow-shrink (GS) algorithm (Margaritis & Thrun, 1999). Score-based approaches place restrictions on the functional causal model and/or the distribution of data. Consequently, these methods can struggle to identify an accurate causal graph when dealing with complex relationships between variables or in the presence of hidden variables. On the other hand, constraint-based methods often do not rely on the aforementioned assumptions and directly test for conditional independence (CI) relations between pairs of variables to determine causal connections.

Theoretical performance guarantees of constraint-based discovery algorithms in the literature nearly always hinge on the availability of a perfect CI oracle, which determines whether two random variables are conditionally independent. In practice, this oracle is substituted with a statistical conditional independence test, which assesses independence using a limited number of observed data points. Hence, to ensure the reliability and applicability of constraint-based causal discovery methods, it is imperative to establish robust sample complexity guarantees. Sample complexity of a causal discovery algorithm is the minimum number of data samples needed to infer the causal graph accurately at a given confidence level.

Unlike unconditional independence testing, conditional independence is not a testable hypothesis without further assumptions on the distribution. Shah & Peters (2020) proved this fundamental hardness result by showing that if (X, Y, Z) has an absolutely continuous distribution with respect to the Lebesgue measure and a given CI test (to identify if X and Y are independent given Z) has a level less than α, there is no alternative under which the test has a power greater than α. Neykov et al. (2021) showed that minimax optimal bounds could be obtained when defining an alternative by discarding distributions that are '$\varepsilon-$close' to the null hypothesis.

It is noteworthy that CI tests are well understood for discrete variables (see Canonne et al., 2018). Another solved case is that of linear models with Gaussian noise, where conditional

[1]College of Management of Technology, EPFL, Lausanne, Switzerland. Correspondence to: Fateme Jamshidi <fateme.jamshidi@epfl.ch>, Luca Ganassali <luca.ganassali@universite-paris-saclay.fr>, Negar Kiyavash <negar.kiyavash@epfl.ch>.

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

independence is equivalent to zero partial correlation, which is simple to assess.

In this paper, we derive sample complexity guarantees for CI tests for continuous distributions. Specifically, we design conditional independence tests built upon estimating conditional mutual information, a measure of conditional dependence between variables. The mutual information $I(X;Y \mid Z)$ between two random variables X, Y conditioned on Z is given by:

$$\iiint \log\left(\frac{p_Z(z)p_{X,Y,Z}(x,y,z)}{p_{X,Z}(x,z)p_{Y,Z}(y,z)}\right) p_{X,Y,Z}(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,.$$

Furthermore,

$$\begin{aligned} I(X;Y \mid Z) &= I(X;Y,Z) - I(X;Z) \\ &= H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z)\,. \end{aligned} \quad (1)$$

Hence, estimating $I(X;Y \mid Z)$ reduces to estimating entropy, the approach we shall take. Specifically, we use the *Von Mises estimator* $\hat{H}_{\mathrm{vm}}$, defined in Section 2.1, which has theoretical and practical advantages. First, this estimator is straightforward to compute by the usual trick of replacing the integration step with a Monte Carlo type summation. Second, when combined with a (nonparametric) kernel estimate of the density, under smoothness[1] assumptions on the joint, $\hat{H}_{\mathrm{vm}}$ converges at the parametric rate $O(n^{-1/2})$, and hence escapes the curse of dimensionality. Finally, it is computationally efficient: its time complexity is linear in the dimension and quadratic in the number of samples.

Contributions Our main contributions are as follows:

- (i) We establish an exponential concentration inequality for $\hat{H}_{\mathrm{vm}}$ when the joint density is computed via kernel density estimation (KDE). This allows for deriving a tighter sample complexity bound for $\hat{H}_{\mathrm{vm}}$ compared to those obtained by a standard appeal to Markov's inequality.
- (ii) We define a test for conditional independence, VM-CI, based on Von Mises estimators and establish its sample complexity when discriminating the null hypothesis, denoted by H_0, of conditional independence from an alternative of the form $H_1 := I(X;Y \mid \mathbf{Z}) > I_{\min}$, with a given level of confidence $1-\alpha$. These results are robust for *all* sufficiently smooth, compactly supported distributions with positive lower bounds.
- (iii) We show that the established sample complexity guarantees of VM-CI yield sample complexity bounds for any constraint-based causal discovery algorithm under mild smoothness assumptions. As an example, we present these bounds for two popular methods, PC and GS. To the best of our knowledge, these are the first sample complexity guarantees for causal discovery algorithms with non-linear models and non-Gaussian continuous variables.

[1]The appropriate notion of smoothness, $\beta-$Hölder smoothness, will be introduced in Definition 1.

Outline of the paper The kernel density estimator, as well as plug-in and Von Mises entropy estimators, are defined in Section 2. In Section 3, we establish the exponential concentration properties of $\hat{H}_{\mathrm{vm}}$, define a CI test based on the former, and derive its error rates. Section 4 is dedicated to causal discovery, where we derive sample complexity guarantees for PC and GS algorithms. Numerical experiments are presented in Section 5. Proofs of our theorems and corollaries are deferred to Appendix A, and further details on numerical experiments can be found in Appendix B.

Related Work

Conditional independence testing for continuous variables In the past decade, several methods for CI testing for continuous variables have been developed. One approach (see, e.g., Huang, 2010) is to discretize the conditioning set Z to a set of bins and perform simple independence tests in each bin. This strategy suffers from the curse of dimensionality, i.e., as the dimension of Z grows, the number of required samples increases drastically.

Another range of approaches is based on kernel methods. These procedures are comprised of two steps. In the first step, X and Y are separately regressed on Z via kernel ridge regression (Zhang et al., 2012). In the second step, the independence of the residuals is tested. This is often done using the Hilbert-Schmidt independence criterion (HSIC, Gretton et al., 2005) or variants of it (Zhang et al., 2012). Recently, a so-called generalized covariance measure was used in Shah & Peters (2020) to test the independence of the residuals. Theoretical guarantees for the second step, i.e., HSIC (as well as its most recent variants such as Nyström based independence criterion, see Kalinke & Szabó, 2023), are now well understood. The standard parametric rate $O(n^{-1/2})$ can be achieved as long as appropriate conditions on the decay rate of the eigenvalues of the corresponding covariance operator are satisfied. For the first step, namely kernel ridge regression, as stated in Shah & Peters (2020), achieving the parametric rate requires the function $f : z \mapsto \mathbb{E}[X \mid Z = z]$ to be $\beta-$smooth (say, $\beta-$Hölder) with $\beta > d/2$. The aforementioned approaches suffer from two main drawbacks. The first one is the time complexity of kernel ridge regression: it involves inverting a $n \times n$ matrix, which in general, takes $O(n^3)$ operations. This cubic time complexity pre-

vents the use of the method on large datasets, as we illustrate in Section 5. The time complexity of our proposed method is $O(dn^2)$ (see Remark 6), which significantly improves on the former. The second drawback is theoretical: to the best of our knowledge, only the convergence rate under the assumptions listed earlier is known, but for instance, no exponential concentration is established. As a consequence, with existing results, sample complexity guarantees for these methods are not as tight as ours.

CI testing can also rely on estimating conditional mutual information: the works by Liu et al. (2012) and Singh & Póczos (2016) are the most relevant to our study. In both articles, the authors consider kernel density estimate of the joint density, in dimension $d = 2$ in the latter and $d \geq 2$ in the former, and prove an exponential concentration for the *plug-in* estimator of entropy. In Poczos & Schneider (2011), the consistency of a plug-in estimator for Rényi divergences using a k-nearest neighbors (KNN) estimate of the density was studied. To the best of our knowledge, the convergence rate of this KNN-based estimator is not known. Instead, here we consider the Von Mises estimator combined with a Kernel density estimate (KDE), which is both easier and efficient to compute, and most importantly, converges with better rates than the plug-in estimator[2]. Empirically, as we shall see in Section 5, KNN-based estimators converge more slowly than the estimator using KDE. Another drawback of KNN is that it is not clear how to tune k in practice, while KDE hyperparameters can be tuned by cross-validation (see e.g., Wasserman, 2023). Convergence properties, asymptotic normality, and rates for Von Mises estimators were studied in Kandasamy et al. (2015). Our work complements these results by showing an exponential concentration inequality.

Belghazi et al. (2018) proposed the mutual information neural network estimator (MINE) for estimating mutual information between two continuous random variables. They rewrite the mutual information using the dual representation of KL divergence (Donsker & Varadhan, 1983), which allows us to formulate the estimation problem as a function optimization. They consider a family of functions parameterized by a deep neural network and solve the optimization using stochastic gradient descent. They derive a sample complexity bound for an estimator which approximates the true mutual information with ε error. The bound scales as $\Omega(d \log d/\varepsilon^2)$ and could be applied directly to derive sample complexity bounds for causal discovery algorithms. However, such bounds are overly dependent on dimension d. In practice, the estimate requires over 2×10^6 to begin to converge, which far exceeds the number of samples we require (see Section 5).

A recent work (Akbari et al., 2023) studies a different approach based on optimal transport (OT). The idea is first to learn a parametric lower triangular monotone map between the unknown joint distribution p and a reference distribution q, typically a standard isotropic Gaussian. Once this map is learned, they can estimate the joint distribution p and recover the conditional independence relationships. Although this method appears to be of practical interest, no theoretical guarantees are available, e.g. its consistency is not proved.

Sample complexity in causal discovery Sample complexity for causal discovery has been investigated in the past two decades. For the most part, these works rely on simplifying assumptions such as the linearity and/or Gaussianity of the model. For instance, Kalisch & Bühlman (2007); Ghoshal & Honorio (2017) study the case where variables are Gaussian, for which CI testing boils down to estimating partial correlations. Other works have considered linear models (Chen et al., 2019) or imposed additional assumptions on the variance of the noise (Park & Raskutti, 2018; Gao et al., 2020). Another well-studied case is that of discrete variables (Friedman & Yakhini, 2013; Zuk et al., 2012; Wadhwa & Dong, 2021). For instance, in this context, Wadhwa & Dong (2021) established the sample complexity of two causal discovery algorithms: inferred causation (IC) and PC, using the CI test introduced by Canonne et al. (2018). This CI test is designed for testing the conditional independence for discrete distributions p, namely testing $H_0 := X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$ vs $H_1 := \sup_q \mathrm{TV}(p, q) > \varepsilon$, where TV is the total variation distance, $\varepsilon > 0$. The supremum is over discrete probability mass functions such that $X \perp\!\!\!\perp_q Y \,|\, \mathbf{Z}$. They showed that the output of the CI test is correct with a probability of at least 2/3. In general, testing causal directions requires additional assumptions or information. In the bivariate discrete case, Acharya et al. (2023) recently established the sample complexity of distinguishing cause from effect when interventional data is available. They obtain a sample complexity which depends on the domain size and characterize the trade-off between the required number of observational and interventional samples.

2. Background on Kernel Density Estimation and Entropy Estimation

We begin by presenting some definitions and notations that appear throughout the paper. Notations $f = o(g)$, $f = O(g)$ and $f = \Theta(g)$ refer to standard Landau notations. The norm $\|\cdot\|_1$ denotes the L^1 norm of a vector in $\mathbb{R}^d$. We assume the $d-$dimensional vector $\mathbf{X}$ takes values in $\mathcal{X}$, a compact subset of $\mathbb{R}^d$. Given a tuple $\mathbf{s} = (s_1, \ldots, s_d)$ of non negative integers, we define $|\mathbf{s}| := \sum_{i=1}^d s_j$, $\mathbf{x}^{\mathbf{s}} := x_1^{s_1} \cdots x_d^{s_d}$, $\mathbf{s}! := s_1! \cdots s_d!$, and $D^{\mathbf{s}}$ denotes the operator

[2] As a consequence, our smoothness assumption required to obtain the parametric rate $O(n^{-1/2})$ – the best we can hope for – is weaker than those in Singh & Póczos (2016) ($\beta > d/2$ versus $\beta > d$).

$D^{\mathbf{s}} := \frac{\partial^{|\mathbf{s}|}}{\partial^{s_1} x_1 \ldots \partial^{s_d} x_d}$.

Definition 1. *(Hölder class, see e.g., Tsybakov, 2008, Definition 1.2) For $L > 0$ and a positive integer β, $f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ belongs to the* Hölder class $\Sigma(\beta, L)$ *on $\mathcal{X}$ if f is β times differentiable, and if for $\mathbf{s} = (s_1, \ldots, s_d)$ such that $|\mathbf{s}| = \beta$, $D^{\mathbf{s}} f$ is bounded by L, uniformly in $\mathbf{s}$ and $\mathbf{x}$, that is $\sup_{\mathbf{s}:|\mathbf{s}|=\beta} \sup_{\mathbf{x}\in\mathcal{X}} |D^{\mathbf{s}} f(\mathbf{x})| \leq L$. f is said to be β-*Hölder smooth *if $f \in \Sigma(\beta, L)$ for some $L > 0$.*

For any k times differentiable function g on $\mathcal{X} \subseteq \mathbb{R}^d$, and $\mathbf{a} \in \mathcal{X}$, we denote by $g_{k,\mathbf{a}}$ the truncated degree k Taylor expansion of g at $\mathbf{a}$, i.e.

$$g_{k,\mathbf{a}}(\mathbf{x}) := \sum_{s:|s|\leq k} \frac{D^{\mathbf{s}} g(\mathbf{a})}{\mathbf{s}!} (\mathbf{x} - \mathbf{a})^{\mathbf{s}} .$$

2.1. Plug-in versus Von Mises Estimator for Entropy

Assume we have access to n samples $(\mathbf{x}^{(i)})_{1\leq i\leq n} = ((x_1^{(i)}, \ldots, x_d^{(i)}))_{1\leq i\leq n}$ of a $d-$dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)$, with density p with a compact support $\mathcal{X}$ in $\mathbb{R}^d$. We seek to estimate the joint entropy

$$H(p) := H(X_1, \ldots, X_d) = -\int_{\mathcal{X}} p(\mathbf{x}) \log p(\mathbf{x}) \, \mathrm{d}\mathbf{x} . \quad (2)$$

The *plug-in* estimator of H is given by

$$\widehat{H}_{\mathrm{plug-in}} := -\int_{\mathcal{X}} \widehat{p}(\mathbf{x}) \log \widehat{p}(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \quad (3)$$

where $\widehat{p}$ is an estimate of the joint probability density. As discussed in the related work, this estimator was studied by Liu et al. (2012) and Singh & Póczos (2016). In practice, computing the numerical approximation of the integral in (3) is costly when dimension d increases. Herein, we study the *Von Mises* estimator defined as follows:

$$\widehat{H}_{\mathrm{vm}} := -\frac{2}{n} \sum_{i=n/2+1}^{n} \log \widehat{p}(\mathbf{x}^{(i)}) . \quad (4)$$

To estimate the entropy, the data is split into two parts. The first part is used to estimate the density $\widehat{p}_h$, and the second half is used to estimate $\widehat{H}_{\mathrm{vm}}$ using $\widehat{p}_h$ according to (4). Note that using Taylor expansion of $p \mapsto -p \log p$ around $\widehat{p}$ in (2) results in[3]:

$$\begin{aligned} H(p) &= H(\widehat{p}) - \textstyle\int_{\mathcal{X}} (\log \widehat{p}(\mathbf{x}) + 1)(p(\mathbf{x}) - \widehat{p}(\mathbf{x})) \, \mathrm{d}\mathbf{x} \\ &\quad + O\left(\textstyle\int_{\mathcal{X}} (p(\mathbf{x}) - \widehat{p}(\mathbf{x}))^2 \, \mathrm{d}\mathbf{x}\right) \\ &= -\textstyle\int_{\mathcal{X}} p(\mathbf{x}) \log \widehat{p}(\mathbf{x}) \, \mathrm{d}\mathbf{x} + O\left(\textstyle\int_{\mathcal{X}} (p(\mathbf{x}) - \widehat{p}(\mathbf{x}))^2 \, \mathrm{d}\mathbf{x}\right) , \end{aligned} \quad (5)$$

[3]This is up to justifying swapping the O and the integral, which will be done later in the proof of Theorem 2.

since $\int_{\mathcal{X}} p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \int_{\mathcal{X}} \widehat{p}(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 1$. This motivates the estimation of $H(p)$ with $-\int_{\mathcal{X}} p(\mathbf{x}) \log \widehat{p}(\mathbf{x}) \, \mathrm{d}\mathbf{x}$. $\widehat{H}_{\mathrm{vm}}$ in (4) is derived by replacing the integral with the Monte Carlo sum. Expansion (5) is often referred to as the *Von Mises expansion* (see Krishnamurthy et al., 2014).

As discussed earlier, to estimate the entropy H, we need to estimate the joint density p. We discuss an approach based on Kernel density estimation in the next section. Please refer to Tsybakov (2008) for more details.

2.2. Kernel Density Estimation

Multivariate kernel density estimation (KDE) provides an estimate of the density p of the following form. For all $\mathbf{x} = (x_1, \ldots, x_d)$ in $\mathcal{X}$,

$$\widehat{p}_h(\mathbf{x}) := \frac{2}{n} \sum_{i=1}^{n/2} \frac{1}{h^d} K_d\left(\frac{\mathbf{x}^{(i)} - \mathbf{x}}{h}\right), \quad (6)$$

where $h := h(n) > 0$ is the *bandwidth* and $K_d : \mathbb{R}^d \to \mathbb{R}$ is a *kernel*, satisfying $\int K_d(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 1$ to ensure that $\int_{\mathcal{X}} \widehat{p}_h(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 1$. Recall that we use the first half of the samples $(\mathbf{x}^{(i)})_{1\leq m\leq n/2}$ to compute $\widehat{p}_h$.

The choice of K_d is generally very open. However, when approximating smooth densities, kernels of order $\ell > 0$ are very useful. We define them below.

Definition 2 (Kernels of given order)**.** *Let ℓ be a positive integer. We say that a kernel $K_d : \mathbb{R}^d \to \mathbb{R}$ is a* kernel of order ℓ *if $\mathbf{x} \mapsto \mathbf{x}^{\mathbf{s}} K(\mathbf{x})$ is integrable for all $|\mathbf{s}| \leq \ell$ and*

$$\int K(\mathbf{x}) \mathrm{d}\mathbf{x} = 1 \text{ and } \int \mathbf{x}^{\mathbf{s}} K(\mathbf{x}) \mathrm{d}\mathbf{x} = 0 \text{ for } |\mathbf{s}| = 1, \ldots, \ell .$$

In particular, a kernel of order ℓ is orthogonal to any polynomial of degree $\leq \ell$ with no constant term.

Product kernels In our practical implementations, we will consider product kernels of the form

$$K_d(\mathbf{x}) = \otimes_d K(\mathbf{x}) := K(x_1) \cdot K(x_2) \cdots K(x_d),$$

where K is a one-dimensional kernel satisfying $\int K(u) \, \mathrm{d}u = 1$. We hence have

$$\widehat{p}_h(\mathbf{x}) := \frac{2}{n} \sum_{i=1}^{n/2} \frac{1}{h^d} K\left(\frac{x_1^{(i)} - x_1}{h}\right) \cdots K\left(\frac{x_d^{(i)} - x_d}{h}\right) .$$

Note that in view of Definition 2, if K is of order $\ell > 0$ then $\otimes_d K(\mathbf{x})$ is also of order ℓ.

Legendre kernels Tsybakov, 2008, in Section 1.2.2 provides a method to build a one-dimensional kernel supported on $[-1, 1]$ of any given order β as follows. Let

$\{\phi_m\}_{m\geq 0}$ be the orthonormal basis of Legendre polynomials $L^2([-1,1],\mathrm{d}x)$ defined by

$$\phi_m(x) := \sqrt{\frac{2m+1}{2}}\frac{1}{2^m m!}\frac{\mathrm{d}^m}{\mathrm{d}x^m}[(x^2-1)^m] \quad (7)$$

for all $m \geq 0$, with $\phi_0(x) = \frac{1}{\sqrt{2}}$ by convention. Then, for $\beta > 0$, the kernel K_β defined by

$$K_\beta(x) := \sum_{m=0}^{\beta} \phi_m(0)\phi_m(x)\mathbb{1}_{|x|\leq 1} \quad (8)$$

is of order[4] β. We will henceforth refer to kernels K_β as Legendre kernels.

3. Exponential Concentration for Entropy Estimation

In this section, we present one of our main results: the exponential concentration for our MI estimator. For pedagogical reasons, we begin by presenting the exponential concentration for the KDE estimator, a result known in the literature. We then proceed to establish an exponential concentration for the Von Mises estimator of entropy and, finally, the MI estimator. A tight upper bound on the error rate of our VM-CI test follows as a corollary. We recall that the kernel density estimator $\widehat{p}_h$ is defined in (6).

3.1. Exponential Concentration for Multivariate Kernel Density Estimation

To obtain the exponential concentration of $\widehat{p}_h$, we must first establish a few technical conditions on kernel K_d.

Assumption 1 (Assumptions on the kernel K_d)**.**

(1a) *K_d is uniformly upper bounded by some $\kappa > 0$,*

(1b) *K_d is of order β (see Definition 2),*

(1c) *The class of functions*

$$\mathcal{F} := \left\{ K_d\left(\frac{\mathbf{x}-\cdot}{h}\right), \mathbf{x}\in\mathbb{R}^d, h>0 \right\}$$

satisfies $\sup_Q N(\mathcal{F}, L^2(Q), \varepsilon\|F\|_{L^2(Q)}) \leq \left(\frac{A}{\varepsilon}\right)^v$,

where A and v are for two positive numbers, $N(T,d,\varepsilon)$ denotes the ε-covering number (see, e.g. John Lafferty, 2008-2010) of the metric space (T,d), F is the envelope function of $\mathcal{F}$ (i.e. $F(\mathbf{x}) := \sup_{f\in\mathcal{F}}|f(\mathbf{x})|$), and the supremum is taken over the set of all probability measures on $\mathbb{R}^d$. The quantity v is called the VC dimension of $\mathcal{F}$.

[4]Note that by symmetry of Legendre polynomials, $\phi_{2m+1}(0) = 0$ for all $m \geq 0$, hence $K_{2\ell} = K_{2\ell+1}$. We will hence often consider β to be odd so that K_β is exactly of order β and not of order $\beta+1$.

Assumption (1c) appears in Giné & Guillou (2002); Rinaldo & Wasserman (2010) and is at the heart of the exponential inequality obtained in Liu et al. (2012). This assumption is known to hold for a large class of kernels (van der Vaart & Wellner, 1996; Nolan & Pollard, 1987), such as compactly supported polynomial kernels and Gaussian kernels[5].

Remark 1. *Kernel K_β defined in (8) satisfies Assumption 1. Therefore, product kernel $K_d := \otimes_d K_\beta$ inherits the same property.*

Theorem 1 (Exponential concentration of $\|p-\widehat{p}_h\|_\infty$)**.** *Assume that p belongs to the* Hölder class $\Sigma(\beta, L)$ *on $\mathcal{X}$ for some $\beta, L > 0$ and that K_d satisfies Assumption 1. Let $h = h_n = \Theta(n^{-\frac{1}{2\beta+d}})$. Then, there exist $C_1, C_2, \varepsilon_0 > 0$ and $n_0 \geq 0$ such that for all $n^{-\frac{\beta}{2\beta+d}}(\log n)^{1/2} \leq \varepsilon_n \leq \varepsilon_0$:*

$$\forall n \geq n_0,\ \mathbb{P}\left(\|p-\widehat{p}_h\|_\infty > \varepsilon_n\right) \leq C_1\exp(-C_2 n^{\frac{2\beta}{2\beta+d}}\varepsilon_n^2)\,.$$

The proof of this result, which appears in Appendix A for the sake of completeness, follows from standard bias analysis and results in Rinaldo & Wasserman (2010).

3.2. Exponential Concentration for Entropy Estimation

Before stating our result on the exponential concentration of estimator $\widehat{H}_{\mathrm{vm}}$ – which we recall is defined in (4) – we describe the conditions which density p must satisfy.

Assumption 2 (Assumptions on the density p)**.**

(2a) *The support of p, $\mathcal{X}$, is a compact set in $\mathbb{R}^d$,*

(2b) *p is lower-bounded on $\mathcal{X}$ by some $p_{\min} > 0$,*

(2c) *p belongs to* Hölder class $\Sigma(\beta, L)$ *for some $L > 0$.*

Remark 2 (Positivity of $\widehat{p}_h$)**.** *The kernel K_d can take negative values[6], as does $\widehat{p}_h$. Hence, $\log\widehat{p}_h$ will not be defined in general, which poses issues in the definition of $\widehat{H}_{\mathrm{vm}}$. As proved by Giné and Guillou (see Theorem 2.3 in Giné & Guillou, 2002), Theorem 1 together with (14) and an application of Borel-Cantelli Lemma shows that almost surely, $n^{\frac{\beta}{2\beta+d}}(\log n)^{-1/2}\|p-\widehat{p}_h\|_\infty$ converges to some bounded random variable C. As a result, if p satisfies (2a), then almost surely there exists n_0 such that for $n \geq n_0$, $\widehat{p}_h$ is point-wise positive. In the sequel, indeed, we assume that n is large enough.*

[5]Assumption (1c) also holds in the following examples: if $K_d(\mathbf{x}) = \phi(T(\mathbf{x}))$, where T is a polynomial in $\mathbb{R}^d$ and ϕ is a bounded real function of bounded variation if the graph of K_d is a pyramid (truncated or not); or if $K_d = \mathbb{1}_{I_1\times\cdots\times I_d}$ where $I_1,\ldots,I_d$ are closed intervals of $\mathbb{R}$ (van der Vaart & Wellner, 1996; Nolan & Pollard, 1987).

[6]For instance, any kernel of order $\beta \geq 2$ needs to take negative values by definition.

Remark 3. *As seen in the next Theorem, the assumption that $p > p_{min}$ is necessary to establish that $\hat{H}_{vm}$ has an exponential convergence rate. In practice, for non-lower bounded densities, one can truncate the density on a compact interval which is large enough so that the entropies of the densities are close. Since the threshold used in VM-CI (9) is always a positive constant ($I_{min}/2$) in the end this truncation is always possible.*

Theorem 2 (Exponential concentration of $\widehat{H}_{\mathrm{vm}}$ in (4))**.** *Assume that K_d satisfies Assumption 1 and that p satisfies Assumption 2. Let $h = h_n = \Theta(n^{-\frac{1}{2\beta+d}})$. Then, there exist $C_1, C_2, C_1', C_2', \varepsilon_0 > 0$ and $n_0 \geq 0$, such that for all ε_n such that $\max(n^{-\frac{2\beta}{2\beta+d}} \log n, n^{-1/2}) \leq \varepsilon_n \leq \varepsilon_0$:*

$$\forall n \geq n_0,\ \mathbb{P}(|\widehat{H}_{\mathrm{vm}} - H(p)| > \varepsilon_n) \leq C_1 e^{-C_2 n^{\frac{2\beta}{2\beta+d}} \varepsilon_n} + C_1' e^{-C_2' n^{1/2} \varepsilon_n}\,.$$

Remark 4. *Note that when p is smooth enough ($\beta > d/2$), $\widehat{H}_{\mathrm{vm}}$ converges at parametric rate $O(n^{-1/2})$ which is the best rate we can hope for.*

Remark 5. *The well-known rate $O(n^{-\min(\frac{1}{2}, \frac{2\beta}{2\beta+d})})$ for Von Mises entropy estimation (see Wasserman, 2023) is immediate from Theorem 2. Note that when $d = 2$, we retrieve the concentration inequality of Liu et al. (2012). The minimax rates for entropy estimation are known to be slightly better, $O(n^{-\min(\frac{1}{2}, \frac{4\beta}{4\beta+d})})$ and can be achieved, but come at the cost of more complex estimators, requiring higher order corrections in the Von Mises expansion (5).*

3.3. Consequences for Error Rates of VM-CI

We start with an immediate corollary of Theorem 2, which states a dimension-free exponential concentration bound for conditional mutual information as long as the probability distributions are smooth enough. Given our application of interest, causal discovery, we assume X and Y are both one-dimensional, but $\mathbf{Z}$ is of dimension $d_{\mathbf{Z}}$. In view of (1), we can estimate $I(X;Y \mid \mathbf{Z})$ by

$$\widehat{I}_{\mathrm{vm}} := \widehat{H}_{\mathrm{vm}}(X, \mathbf{Z}) + \widehat{H}_{\mathrm{vm}}(Y, \mathbf{Z}) - \widehat{H}_{\mathrm{vm}}(\mathbf{Z}) - \widehat{H}_{\mathrm{vm}}(X, Y, \mathbf{Z})\,,$$

where $\widehat{H}_{\mathrm{vm}}$ is the Von Mises estimator in (4).

Corollary 1 (Dimension-free exponential concentration of $\widehat{I}_{\mathrm{vm}}$ for smooth densities)**.** *Assume that*

- *joint distributions $p_{X,Y,\mathbf{Z}}$, $p_{X,\mathbf{Z}}$, $p_{Y,\mathbf{Z}}$, and $p_{\mathbf{Z}}$ satisfy Assumption 2 for some $\beta > 0$ such that $\beta > 1 + d_{\mathbf{Z}}/2$, and*
- *kernels involved in estimators $\widehat{H}_{\mathrm{vm}}(X, Y, \mathbf{Z})$ (resp. $\widehat{H}_{\mathrm{vm}}(X, \mathbf{Z})$, $\widehat{H}_{\mathrm{vm}}(Y, \mathbf{Z})$ and $\widehat{H}_{\mathrm{vm}}(\mathbf{Z})$) satisfy Assumption 1, with β given in the previous bullet.*

Choose bandwidth h_n as follows:

$$h_n = \begin{cases} \Theta(n^{-\frac{1}{2\beta+2+d_{\mathbf{Z}}}}) \text{ for } \widehat{H}_{\mathrm{vm}}(X, Y, \mathbf{Z}), \\ \Theta(n^{-\frac{1}{2\beta+1+d_{\mathbf{Z}}}}) \text{ for } \widehat{H}_{\mathrm{vm}}(X, \mathbf{Z}) \text{ and } \widehat{H}_{\mathrm{vm}}(Y, \mathbf{Z}), \\ \Theta(n^{-\frac{1}{2\beta+d_{\mathbf{Z}}}}) \text{ for } \widehat{H}_{\mathrm{vm}}(\mathbf{Z}), \end{cases}$$

then, there exist $C_1, C_2, \varepsilon_0 > 0$ and $n_0 \geq 0$ such that for all constant $0 < \varepsilon \leq \varepsilon_0$,

$$\forall n \geq n_0,\ \mathbb{P}(|\widehat{I}_{\mathrm{vm}} - I(X;Y \mid \mathbf{Z})| > \varepsilon) \leq C_1 \exp(-C_2 n^{1/2} \varepsilon)\,.$$

To provide performance guarantees for our CI test, we require the following mild assumption.

Assumption 3 (Minimum level of dependency)**.** *There exists $I_{min} > 0$ such that X, Y, and $\mathbf{Z}$ are either conditionally independent (i.e., $X \perp\!\!\!\perp Y \mid \mathbf{Z}$) or $I(X;Y \mid \mathbf{Z}) > I_{min}$.*

Under the Assumption 3, we can define the following hypothesis test.

$$H_0 := I(X;Y \mid \mathbf{Z}) = 0 \text{ vs. } H_1 := I(X;Y \mid \mathbf{Z}) > I_{\min}\,.$$

The test, VM-CI, is defined as follows.

$$T_{\mathrm{VM-CI}} := \begin{cases} 1 & \text{if } \widehat{I}_{\mathrm{vm}} > I_{\min}/2, \\ 0 & \text{elsehow.} \end{cases} \tag{9}$$

Corollary 2 (Error rates for VM-CI)**.** *Under Assumption 3 as well as the assumptions stated in Corollary 1, the sum of type one and type two errors for $T_{\mathrm{VM-CI}}$ is bounded by $O\left(\exp(-cn^{1/2} I_{\min})\right)$ for some $c > 0$. Hence, in order to achieve a confidence level $1 - \alpha \in [0, 1)$ it suffices that $n \geq \Omega\left(\frac{1}{I_{\min}^2} \log^2\left(\frac{1}{\alpha}\right)\right)$.*

Remark 6 (Time complexity of VM-CI)**.** *Assume that each evaluation of K_d is done in $O(d)$. Then, each appeal to $\widehat{p}_h$ takes $O(dn)$ operations. Hence, $\widehat{H}_{\mathrm{vm}}$, $\widehat{I}_{\mathrm{vm}}$, and VM-CI can be computed in $O(dn^2)$.*

4. Application: Sample Complexity Guarantees for Causal Discovery

In this section, we present a brief background on causal discovery and review two classic causal algorithms, PC and GS, before deriving their sample complexity when using the VM-CI test. Note that under appropriate assumptions, these sample complexities are optimal since they inherit the parametric convergence rate of VM-CI, which is the best we can hope for.

4.1. Background on Causal Discovery

A directed acyclic graph (DAG) is defined as $\mathcal{G} = (\mathbf{X}, E)$, where $\mathbf{X} = \{X_1, \ldots, X_m\}$ (resp. $E \subseteq \mathbf{X} \times \mathbf{X}$) denotes the set of vertices (resp. directed edges) of $\mathcal{G}$, such that $\mathcal{G}$

contains no directed cycle. Each vertex $X_k \in \mathbf{X}$ represents a random variable. Vertices $X, Y \in \mathbf{X}$ are called neighbors in $\mathcal{G}$ if (X, Y) or (Y, X) belongs to E. We denote the set of neighbors of X $\mathcal{G}$ by $N_{\mathcal{G}}(X)$. Causal discovery (a.k.a structure learning) is the task of learning the causal graph $\mathcal{G}$ from n i.i.d. samples drawn from the joint distribution p, commonly referred to as the observational distribution.

We assume that $\mathcal{G}$ and p satisfy *Markov* and *faithfulness* properties, which state that conditional independence relationships in p correspond to so-called d-separation (a graphical condition) in $\mathcal{G}$ [7]. Two DAGs satisfying *Markov* and *faithfulness* properties are *Markov equivalent* if they have the same set of d-separations (i.e., encode the same set of conditional independence). The equivalence class of a DAG $\mathcal{G}$ is called the Markov equivalence class (MEC) of $\mathcal{G}$. It is well-known that without further assumptions, we can only learn the underlying causal DAG up to its Markov equivalence from the observational data alone (Spirtes et al., 2000; Pearl, 2009).

4.2. PC Algorithm (Spirtes et al., 2000)

PC begins with a complete, undirected graph $\mathcal{C}$ on the vertex set $\mathbf{X}$. Starting from $\ell = 0$, the algorithm considers pairs of variables X and Y adjacent in $\mathcal{C}$ such that $|N_{\mathcal{C}}(X) \setminus Y| \geq \ell$. For all $\mathbf{Z} \subseteq N_{\mathcal{C}}(X) \setminus Y$ such that $|\mathbf{Z}| = \ell$, PC iteratively tests $X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$. If the conditional independence holds for a subset $\mathbf{Z}$, the edge $\{X, Y\}$ is removed in $\mathcal{C}$. After step $\ell = \Delta$, the maximum degree in $\mathcal{G}$, the process terminates. The last step of the algorithm consists of orienting the edges in $\mathcal{C}$, leveraging the information acquired in the previous phase as well as applying so-called Meek rules (see Meek, 1995). If all CI tests outputs are correct, the final graph $\mathcal{C}$ is the essential graph[8] of $\mathcal{G}$. Furthermore, recall that m is the number of nodes in $\mathcal{G}$.

Theorem 3 (Sample complexity of PC)**.** *Assume that all CI tests involving $(X, Y, \mathbf{Z})$ with $X, Y \in \mathbf{X}$ and $\mathbf{Z} \subseteq \mathbf{X}$, $|\mathbf{Z}| \leq \Delta$[9], Assumptions 1 (on the kernel), 2 (on the joint), and 3 (on the minimum level of dependency) are satisfied. Let $\alpha > 0$. Then, PC algorithm using VM-CI tests with threshold $I_{min}/2$ recovers the MEC of $\mathcal{G}$ with probability $\geq 1 - \alpha$, as long as $n \geq \Omega\left(\left(\frac{\Delta+1}{I_{min}} \log(m/\alpha)\right)^2\right)$.*

Remark 7. *The sample complexity result of Theorem 3 results from the exponential concentration derived in Theorem 2. The previously known rate (Wasserman, 2023) $\mathbb{E}[|\widehat{I}_{\text{vm}} - I|] \leq Cn^{-1/2}$ and applying Markov's inequality yields the much looser[10] bound $n \geq \Omega\left(\left(\frac{m^{\Delta+1}}{\alpha I_{min}}\right)^2\right)$.*

4.3. GS Algorithm (Margaritis & Thrun, 1999)

Definition 3 (Markov boundary)**.** *The Markov boundary of a random variable X in set $\mathbf{X}$, denoted by $\mathrm{MB}(X)$, is a minimal set $\mathbf{S} \subseteq \mathbf{X} \setminus \{X\}$ such that $X \perp\!\!\!\perp \mathbf{X} \setminus (\mathbf{S} \cup \{X\}) \,|\, \mathbf{S}$.*

The GS algorithm first recovers the Markov boundary of each variable $X \in \mathbf{X}$, as follows. Starting with $\mathrm{MB}(X) = \varnothing$,

1. (Growing phase) While $\exists\, Y \in \mathbf{X} \setminus \{X\}$ such that $Y \perp\!\!\!\perp X \,|\, \mathrm{MB}(X)$, add Y to $\mathrm{MB}(X)$,
2. (Shrinking phase) While $\exists\, Y \in \mathrm{MB}(X)$ such that $Y \not\perp\!\!\!\perp X \,|\, \mathrm{MB}(X) \setminus \{Y\}$, remove Y from $\mathrm{MB}(X)$.

Then, GS recovers the non-oriented graph structure. For every $X \in \mathbf{X}$ and $Y \in \mathrm{MB}(X)$, a non-oriented edge $\{X, Y\}$ is added if

3. for all $\mathbf{S} \subseteq \mathbf{T}$, where $\mathbf{T}$ is the set with the smaller cardinality between $\mathrm{MB}(X) \setminus \{Y\}$ and $\mathrm{MB}(Y) \setminus \{X\}$, it holds that $X \not\perp\!\!\!\perp Y \,|\, \mathbf{S}$.

Finally, every edge $\{X, Y\}$ is oriented $Y \to X$ if

4. $\exists\, Z \in N(X) \setminus (N(Y) \cup \{Y\})$ such that for all $\mathbf{S} \subseteq \mathbf{W}$, where $\mathbf{W}$ is the set with the smaller cardinality between $\mathrm{MB}(Y) \setminus \{X, Z\}$ and $\mathrm{MB}(Z) \setminus \{X, Y\}$, it holds that $Y \not\perp\!\!\!\perp Z \,|\, \mathbf{S} \cup \{X\}$.

The last step of GS is the same as in PC, namely, it applies the Meek rules.

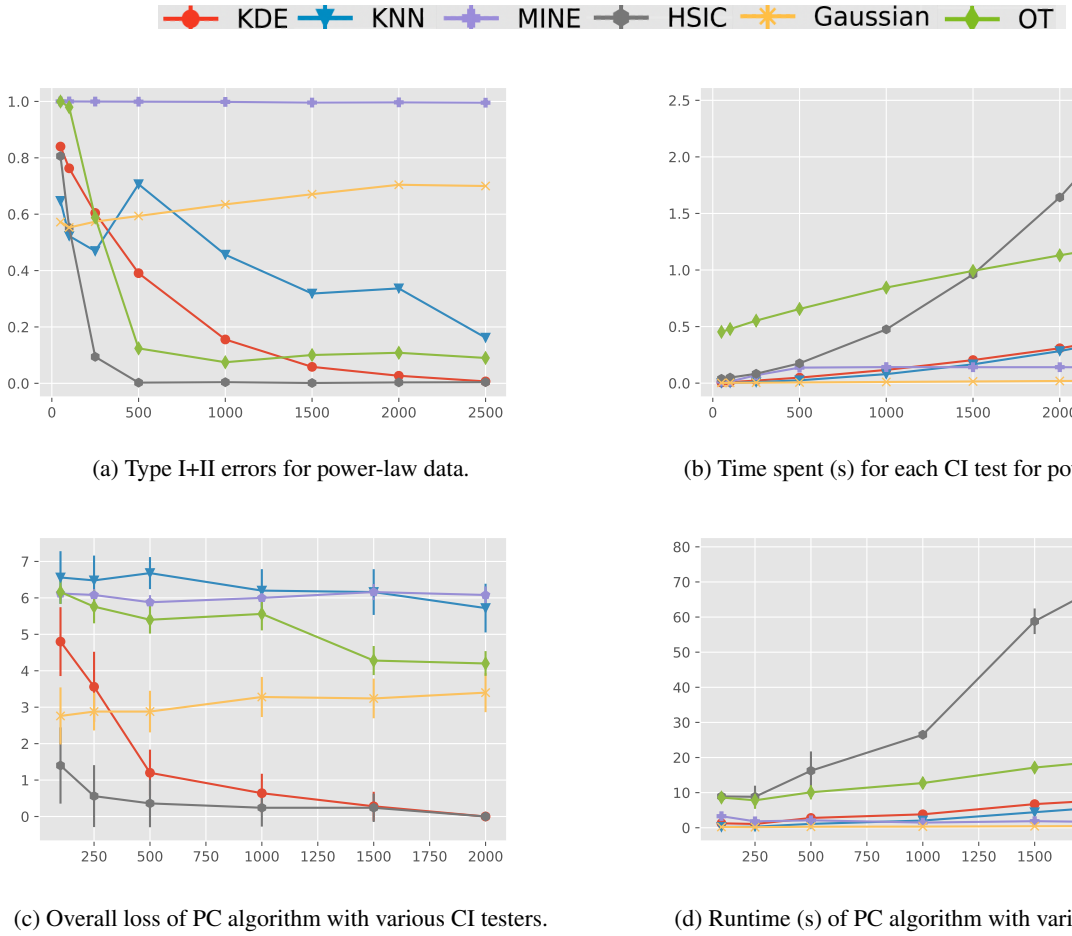
Theorem 4 (Sample complexity of GS)**.** *Assume that $\max_{X \in \mathbf{X}} |\mathrm{MB}(X)| \leq \Gamma$ and that for all CI test involving $(X, Y, \mathbf{Z})$ with $X, Y \in \mathbf{X}, \mathbf{Z} \subseteq \mathbf{X}, |\mathbf{Z}| \leq \Gamma$, Assumptions 1 (on the kernel), 2 (on the joint), and 3 (on the minimum level of dependency) are satisfied. Then, GS algorithm using VM-CI tests with threshold $I_{min}/2$ recovers the MEC of $\mathcal{G}$ with probability $\geq 1 - \alpha$, as long as $n \geq \Omega\left(\frac{1}{I_{min}} \log\left(\frac{m^2 + m\Gamma^2 2^\Gamma}{\alpha}\right)^2\right)$.*

[7] We refer the reader to Pearl (2009) for definitions and further discussion on this topic.

[8] The essential graph of $\mathcal{G}$ represents the Markov equivalence class of $\mathcal{G}$. Namely, it has the same skeleton and v-structures (see Pearl, 2009).

[9] Δ is the maximum degree or an upper bound on the maximum degree in $\mathcal{G}$.

[10] Note that the sample complexity of PC obtained when MINE (Belghazi et al., 2018) is used to estimate the mutual information scales as $\Omega\left(\left(\frac{(m/2+\Delta+1)}{I_{min}} \log(m/\alpha)\right)^2\right)$, which already improves over Markov inequality but is still much looser than ours.



(a) Type I+II errors for power-law data.

(b) Time spent (s) for each CI test for power-law data.

(c) Overall loss of PC algorithm with various CI testers.

(d) Runtime (s) of PC algorithm with various CI testers.

Figure 1: Results of the numerical experiments, with the x-axis representing the number of samples (n). The red curves correspond to our method.

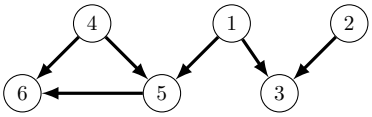


Figure 2: Underlying causal graph in the experiments.

5. Numerical Experiments

5.1. Experiments for Single Conditional Independence Test

We compared VM-CI to other CI tests discussed in the related work, including the KNN-based estimator (Poczos & Schneider, 2011), MINE (Belghazi et al., 2018), the HSIC-based CI test [11](Zhang et al., 2012), the OT-based method (Akbari et al., 2023), and the standard Gaussian partial correlation test.

We conducted the experiments using power-law distributed synthetic data. For each value of n, we ran $n_{\text{exp}} = 5000$ experiments, the first half with $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, the second half with $I(X;Y \mid \mathbf{Z}) > I_{min}$. The resulting estimated errors (sum of type I and type II errors) are depicted as a function of n in Figure 1a. In the figures, our method is denoted as 'KDE'. More details regarding the generative models and parameters can be found in Appendix B.

These results illustrate that when dealing with non-Gaussian data, VM-CI outperforms most other methods in terms of type I and type II errors, except for the HSIC-based method (Zhang et al., 2012). However, VM-CI competes favorably with HSIC when the sample size n exceeds 1500, and it is significantly faster, as demonstrated in Figure 1b. Among the methods we explored, the OT-based approach of Akbari et al. (2023) comes close to that of VM-CI, although its performance and time complexity fall slightly short of VM-CI.

[11]Provided in Kalainathan et al. (2020).

5.2. Experiments for Causal Discovery Algorithms

In addition to CI tests, we ran experiments to assess PC performance using VM-CI compared to PC performance using other CI tests. These experiments were performed on non-Gaussian synthetic data generated using a Structural Equation Model (SEM) with the causal graph depicted in Figure 2.

For each value of n, we conducted 25 experiments and depicted the overall loss in Figures 1c. The overall loss is defined as the total number of missing, extra, and misoriented edges in the resulting graph. Additional details regarding the SEM and its parameters, along with similar experiments for GS, can be found in Appendix B.

As shown, VM-CI outperforms all methods except HSIC. Once again, VM-CI is competitive with HSIC when $n \geq 1500$, and it significantly outpaces HSIC in terms of computational efficiency. It is noteworthy that, despite being an efficient CI test, the OT-based method performs poorly in our example. This may be attributed to several factors: (i) the absence of theoretical guarantees such as consistency for the OT method; (ii) the lack of robustness of PC/GS to errors in CI tests[12]; and (iii) the strong dependence of the performance of this method on the dimension. The remaining results are consistent with the performance of the CI tests in Figure 1a.

To conclude, we emphasize that among the reviewed methods, those that compete favorably with VM-CI either suffer from the lack of theoretical guarantees or have a prohibitive time complexity, or a combination of both.

6. Conclusion

We established an exponential concentration inequality for the nonparametric Von Mises estimator. Using this estimator, we designed VM-CI to test conditional independence. This test achieves optimal parametric rates under smoothness assumptions and provides a tight upper bound for the error. This further allowed us to compute the sample complexity of causal discovery algorithms using VM-CI, the first such guarantee for non-linear models and non-Gaussian continuous variables. Our empirical findings show that VM-CI overall outperforms other popular conditional independence tests in terms of time, sample complexity, or both. Methods competing with VM-CI either require excessive time complexity or suffer from a lack of theoretical guarantees.

[12] A single error in one of the many CI tests in PC/GS can lead to a drastically different graph.

Impact Statement

This work introduces VM-CI, a novel conditional independence test utilizing a nonparametric Von Mises estimator for the entropy of multivariate distributions. By establishing an exponential concentration inequality for this estimator, we design a test that achieves optimal parametric rates under smoothness conditions. This provides the first sample complexity guarantee for causal discovery in non-linear models with non-Gaussian continuous variables. This method has broad implications across multiple domains, including policy analysis, public health, economics, and environmental research, by offering more robust and efficient tools for causal inference. We anticipate no specific ethical issues or negative societal impacts resulting from this work.

References

Acharya, J., Bhadane, S., Bhattacharyya, A., Kandasamy, S., and Sun, Z. Sample complexity of distinguishing cause from effect. In *International Conference on Artificial Intelligence and Statistics*, pp. 10487–10504. PMLR, 2023.

Akbari, S., Ganassali, L., and Kiyavash, N. Learning causal graphs via monotone triangular transport maps, 2023.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Canonne, C. L., Diakonikolas, I., Kane, D. M., and Stewart, A. Testing conditional independence of discrete distributions, 2018.

Chen, W., Drton, M., and Wang, Y. S. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4): 973–980, 2019.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of machine learning research*, 3 (Nov):507–554, 2002.

Donsker, M. and Varadhan, S. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36 (2):183–212, March 1983. ISSN 0010-3640. doi: 10.1002/cpa.3160360204.

Friedman, N. and Yakhini, Z. On the sample complexity of learning bayesian networks. *arXiv preprint arXiv:1302.3579*, 2013.

Gao, M., Ding, Y., and Aragam, B. A polynomial-time algorithm for learning nonparametric causal graphs. *Advances in Neural Information Processing Systems*, 33: 11599–11611, 2020.

Ghoshal, A. and Honorio, J. Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. *Advances in Neural Information Processing Systems*, 30, 2017.

Giné, E. and Guillou, A. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907–921, 2002. ISSN 0246-0203. doi: https://doi.org/10.1016/S0246-0203(02)01128-7. URL https://www.sciencedirect.com/science/article/pii/S0246020302011287.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. *Algorithmic Learning Theory: 16th International Conference, ALT 2005, 63-78 (2005)*, 3734, 10 2005. doi: 10.1007/11564089_7.

Huang, T.-M. Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, 38(4):2047 – 2091, 2010. doi: 10.1214/09-AOS770. URL https://doi.org/10.1214/09-AOS770.

John Lafferty, Han Liu, L. W. Lecture notes: Statistical methods for machine learning, 2008-2010. URL https://www.stat.cmu.edu/~larry/=sml/Concentration.pdf. (Chapter 7, concentration of measure).

Kalainathan, D., Goudet, O., and Dutta, R. Causal discovery toolbox: Uncovering causal relationships in python. *The Journal of Machine Learning Research*, 21(1):1406–1410, 2020.

Kalinke, F. and Szabó, Z. Nyström m-Hilbert-Schmidt independence criterion. In Evans, R. J. and Shpitser, I. (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1005–1015. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/kalinke23a.html.

Kalisch, M. and Bühlman, P. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Kandasamy, K., Krishnamurthy, A., Poczos, B., Wasserman, L., and robins, j. m. Nonparametric von mises estimators for entropies, divergences and mutual informations. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/06138bc5af6023646ede0e1f7c1eac75-Paper.pdf.

Krishnamurthy, A., Kandasamy, K., Poczos, B., and Wasserman, L. Nonparametric estimation of renyi divergence and friends. In *International Conference on Machine Learning*, pp. 919–927. PMLR, 2014.

Liu, H., Lafferty, J., and Wasserman, L. Exponential concentration for mutual information estimation with application to forests. *Advances in Neural Information Processing Systems*, 4:2537 – 2545, 2012. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84877734502&partnerID=40&md5=7b970d4dcec9bf08b3137384060fa955.

Margaritis, D. and Thrun, S. Bayesian network induction via local neighborhoods. *Advances in neural information processing systems*, 12, 1999.

Meek, C. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 403–410, 1995.

Neykov, M., Balakrishnan, S., and Wasserman, L. Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151 – 2177, 2021. doi: 10.1214/20-AOS2030. URL https://doi.org/10.1214/20-AOS2030.

Nolan, D. and Pollard, D. U-Processes: Rates of Convergence. *The Annals of Statistics*, 15(2):780 – 799, 1987. doi: 10.1214/aos/1176350374. URL https://doi.org/10.1214/aos/1176350374.

Park, G. and Raskutti, G. Learning quadratic variance function (qvf) dag models via overdispersion scoring (ods). *Journal of Machine Learning Research*, 18(224):1–44, 2018.

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

Poczos, B. and Schneider, J. On the estimation of α-divergences. In Gordon, G., Dunson, D., and Dudík, M. (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 609–617, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/poczos11a.html.

Rinaldo, A. and Wasserman, L. Generalized density clustering. *The Annals of Statistics*, 38(5):2678 – 2722, 2010. doi: 10.1214/10-AOS797. URL https://doi.org/10.1214/10-AOS797.

Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), jun 2020. doi: 10.1214/19-aos1857. URL https://doi.org/10.1214%2F19-aos1857.

Singh, S. and Póczos, B. Exponential concentration of a density functional estimator, 2016.

Solus, L., Wang, Y., and Uhler, C. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108(4):795–814, 2021.

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

Tsybakov, A. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008. ISBN 9780387790527. URL https://books.google.ch/books?id=mwB8rUBsbqoC.

van der Vaart, A. and Wellner, J. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. ISBN 9780387946405. URL https://books.google.ch/books?id=OCenCW9qmp4C.

Wadhwa, S. and Dong, R. On the sample complexity of causal discovery and the value of domain expertise. *arXiv preprint arXiv:2102.03274*, 2021.

Wasserman, L. Lecture notes: Statistical methods for machine learning (36-708), 2023. (Density estimation).

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.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Zhu, S., Ng, I., and Chen, Z. Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477*, 2019.

Zuk, O., Margel, S., and Domany, E. On the number of samples needed to learn the correct structure of a bayesian network. *arXiv preprint arXiv:1206.6862*, 2012.

A. Proofs

A.1. Proof of Theorem 1

Proof of Theorem 1. To prove our result, we use the standard bias-variance decomposition:

$$|p(\mathbf{x}) - \widehat{p}_h(\mathbf{x})| \leq \underbrace{|p(\mathbf{x}) - p_h(\mathbf{x})|}_{\text{bias}} + \underbrace{|p_h(\mathbf{x}) - \widehat{p}_h(\mathbf{x})|}_{\text{variance}}, \tag{10}$$

where we recall that $p_h := \mathbb{E}[\widehat{p}_h]$. We bound the bias and variance terms separately.

Bounding the bias.

$$\begin{aligned} p(\mathbf{x}) - p_h(\mathbf{x}) &= p(\mathbf{x}) - \frac{2}{n}\sum_{i=1}^{n/2} \int \frac{1}{h^d} K_d\left(\frac{\mathbf{x}^{(i)} - \mathbf{x}}{h}\right) p(\mathbf{x}^{(i)})\, \mathrm{d}\mathbf{x}^{(i)} \\ &= p(\mathbf{x}) - \int \frac{1}{h^d} K_d\left(\frac{\mathbf{x}' - \mathbf{x}}{h}\right) p(\mathbf{x}')\, \mathrm{d}\mathbf{x}' \\ &\overset{(a)}{=} \int K_d(\mathbf{y})(p(\mathbf{x}) - p(\mathbf{x} + h\mathbf{y}))\, \mathrm{d}\mathbf{y}, \end{aligned} \tag{11}$$

where (a) results from change of variable $\mathbf{y} = (\mathbf{x}' - \mathbf{x})/h$. We now take advantage of the fact that functions in $\Sigma(\beta, L)$ are well approximated by their Taylor expansions. Namely, we have the following classical result:

Lemma 1. *If $g \in \Sigma(\beta, L)$ on $\mathcal{X} \subseteq \mathbb{R}^d$, then for all $\mathbf{a}, \mathbf{x} \in \mathcal{X}$,*

$$|g(\mathbf{x}) - g_{\beta-1,\mathbf{a}}(\mathbf{x})| \leq L\frac{\|\mathbf{x} - \mathbf{a}\|_1^\beta}{\beta!}\,. \tag{12}$$

Proof of Lemma 1. We apply Taylor's theorem at the order $\beta - 1$. There exists $c \in [0, 1]$ such that

$$g(\mathbf{x}) = \sum_{|s| \leq \beta - 1} \frac{D^{\mathbf{s}} g(\mathbf{a})}{\mathbf{s}!}(\mathbf{x} - \mathbf{a})^{\mathbf{s}} + \sum_{|s| = \beta} \frac{D^{\mathbf{s}} g(\mathbf{a} + c(\mathbf{x} - \mathbf{a}))}{\mathbf{s}!}(\mathbf{x} - \mathbf{a})^{\mathbf{s}}$$

Hence

$$|g(\mathbf{x}) - g_{\beta-1,\mathbf{a}}(\mathbf{x})| \leq \sum_{|s|=\beta} L\frac{|\mathbf{x} - \mathbf{a}|^{\mathbf{s}}}{\mathbf{s}!} = L\frac{\|\mathbf{x} - \mathbf{a}\|_1^\beta}{\beta!},$$

by the multinomial theorem. □

With Lemma 1 in mind, (11) becomes

$$\begin{aligned} |p(\mathbf{x}) - p_h(\mathbf{x})| \leq &\left| \int K_d(\mathbf{y})(p(\mathbf{x}) - p_{\beta-1,\mathbf{x}}(\mathbf{x} + h\mathbf{y}))\, \mathrm{d}\mathbf{y} \right| \\ &+ \int |K_d(\mathbf{y})(p(\mathbf{x} + h\mathbf{y}) - p_{\beta-1,\mathbf{x}}(\mathbf{x} + h\mathbf{y}))|\, \mathrm{d}\mathbf{y} \end{aligned} \tag{13}$$

Note that $p(\mathbf{x}) - p_{\beta-1,\mathbf{x}}(\mathbf{x} + h\mathbf{y})$ is a polynomial in $\mathbf{y}$, of degree $\leq \beta$, and with no constant term. Since K_d is of order β, the first term of the RHS of (13) evaluates to 0. This gives in turn, applying Lemma 1,

$$|p(\mathbf{x}) - p_h(\mathbf{x})| \leq Lh^\beta \int |K_d(\mathbf{y})| \|\mathbf{y}\|_1^\beta\, \mathrm{d}\mathbf{y} \leq Ch^\beta\,, \tag{14}$$

for some constant $C > 0$, since $\mathbf{y} \mapsto |K_d(\mathbf{y})|\|\mathbf{y}\|_1^\beta$ is integrable by assumption. Note that the bound (14) is uniform in $\mathbf{x} \in \mathcal{X}$.

Bounding the variance. The variance satisfies an exponential concentration property, thanks to Assumption $(1c)$. We leverage on a result from (Rinaldo & Wasserman, 2010), obtained by applying some previously established results from (Giné & Guillou, 2002).

Proposition 1 (Proposition 9 in (Rinaldo & Wasserman, 2010)). *Assume that K_d satisfies $(1a)$ and $(1c)$. Then, for any $D_1 > 0$ there exists constants $D_2, D_3, \varepsilon_0 > 0$ and $n_0 > 0$ such that, if $h_n \to 0$, $\frac{h_n^d}{|\log h_n|} \to 0$, and $D_1\sqrt{\frac{|\log h_n|}{nh_n^d}} \leq \varepsilon_n \leq \varepsilon_0$, then*

$$\mathbb{P}\left(\sup_{\mathbf{x}\in\mathcal{X}} |p_h(\mathbf{x}) - \widehat{p}_h(\mathbf{x})| > \varepsilon_n\right) \leq D_2 \exp\left(-D_3 n h_n^d \varepsilon_n^2\right) \tag{15}$$

Completing the proof. Taking $h_n = \Theta(n^{-\frac{1}{2\beta+d}})$ (minimizing the MISE $\Theta(h^{2\beta}) + \Theta(\frac{1}{nh^d})$), $D_1 = 1$ and $n^{-\frac{\beta}{2\beta+d}}(\log n)^{1/2} \leq \varepsilon_n/2 \leq \varepsilon_0$ ensures that Proposition 1 applies for $\varepsilon_n/2$. Applying (14) to $h = h_n$ gives that almost surely $\|p - p_h\|_\infty = O(n^{-\frac{\beta}{2\beta+d}}) < \varepsilon_n/2$ for n large enough. Now, for n large enough,

$$\begin{aligned}\mathbb{P}(\|p - \widehat{p}_h\|_\infty > \varepsilon_n) &\leq \mathbb{P}(\|p - p_h\|_\infty > \varepsilon_n/2) + \mathbb{P}(\|p_h - \widehat{p}_h\|_\infty > \varepsilon_n/2) \\ &\leq 0 + D_2 \exp\left(-C_2 n^{-\frac{2\beta}{2\beta+d}} \varepsilon_n^2\right),\end{aligned}$$

which ends the proof of Theorem 1. □

A.2. Proof of Theorem 2

Proof of Theorem 2. As stated in (5), the first step is to rigorously justify the Von Mises expansion. Note that since $(-y\log y)' = -\log y - 1$ and $(-y\log y)'' = -1/y$, then for a given $\mathbf{x} \in \mathcal{X}$,

$$|-p(\mathbf{x})\log p(\mathbf{x}) + \widehat{p}_h(\mathbf{x})\log\widehat{p}_h(\mathbf{x}) + (\log\widehat{p}_h(\mathbf{x}) + 1)(p(\mathbf{x}) - \widehat{p}_h(\mathbf{x}))| \leq \left(\sup_{x\in\mathcal{X}} \frac{1}{|\widehat{p}_h(\mathbf{x})|}\right)(p(\mathbf{x}) - \widehat{p}_h(\mathbf{x}))^2, \tag{16}$$

and since p is lower bounded by $p_{\min} > 0$, then by Remark 2, for n large enough, $\sup_{x\in\mathcal{X}} \frac{1}{|\widehat{p}_h(\mathbf{x})|} \leq 2/p_{\min}$ and we can integrate of (16) over $\mathcal{X}$ to indeed get

$$\begin{aligned}H(p) &= H(\widehat{p}_h) - \int_{\mathcal{X}} (\log\widehat{p}_h(\mathbf{x}) + 1)(p(\mathbf{x}) - \widehat{p}_h(\mathbf{x}))\,\mathrm{d}\mathbf{x} + O\left(\int_{\mathcal{X}} (p(\mathbf{x}) - \widehat{p}_h(\mathbf{x}))^2\,\mathrm{d}\mathbf{x}\right) \\ &= -\int_{\mathcal{X}} p(\mathbf{x})\log\widehat{p}_h(\mathbf{x})\,\mathrm{d}\mathbf{x} + O\left(\int_{\mathcal{X}} (p(\mathbf{x}) - \widehat{p}_h(\mathbf{x}))^2\,\mathrm{d}\mathbf{x}\right).\end{aligned} \tag{17}$$

This in turn implies that

$$\widehat{H}_{\mathrm{vm}} - H(p) = -\frac{2}{n}\sum_{i=n/2+1}^{n} \log\widehat{p}_h(\mathbf{x}^{(i)}) + \int_{\mathcal{X}} p(\mathbf{x})\log\widehat{p}_h(\mathbf{x})\,\mathrm{d}\mathbf{x} + O\left(\int_{\mathcal{X}} (p(\mathbf{x}) - \widehat{p}_h(\mathbf{x}))^2\,\mathrm{d}\mathbf{x}\right). \tag{18}$$

The first two terms are the difference between an empirical mean and its expectation w.r.t. p. Recall that n is large enough so that $\|p - \widehat{p}_h\|_\infty < p_{\min}/2$ (Remark 2). Hence, since p is bounded on the compact set $\mathcal{X}$, so is $\widehat{p}_h$. Every term in the sum $\sum_{i=n/2+1}^{n} \frac{2}{n}\log\widehat{p}_h(\mathbf{x}^{(i)})$ is almost surely bounded by c/n where $c > 0$ is a constant. Azuma-Hoeffding inequality yields

$$\begin{aligned}\mathbb{P}\left(\left|-\frac{2}{n}\sum_{i=n/2+1}^{n} \log\widehat{p}_h(\mathbf{x}^{(i)}) + \int_{\mathcal{X}} p(\mathbf{x})\log\widehat{p}_h(\mathbf{x})\,\mathrm{d}\mathbf{x}\right| > \varepsilon_n/2\right) &\leq 2\exp\left(-\frac{\varepsilon_n^2}{8\sum_{i=n/2+1}^{n}(c/n)^2}\right) \\ &= 2\exp\left(-\frac{\varepsilon_n^2 n}{4c}\right) \\ &\leq C_1'\exp\left(-C_2' n^{1/2}\varepsilon_n\right),\end{aligned} \tag{19}$$

since $n^{1/2}\varepsilon_n > 1$ by assumption. The second part of the result comes from the inequality

$$\mathbb{P}\left(\int_{\mathcal{X}} (p(\mathbf{x}) - \widehat{p}_h(\mathbf{x}))^2\,\mathrm{d}\mathbf{x} > t\right) \leq \mathbb{P}\left(\sup_{\mathbf{x}\in\mathcal{X}} |p(\mathbf{x}) - \widehat{p}_h(\mathbf{x})| > \sqrt{t}/\mathrm{Vol}(\mathcal{X})\right), \tag{20}$$

and appealing to Theorem 1 with a deviation $\sqrt{\varepsilon_n}/C_4$ where $C_4 > 0$ is some constant (depending on $p_{\min}$ and $\mathrm{Vol}(\mathcal{X})$). □

A.3. Proof of Corollary 1

Proof of Corollary 1. The proof of Corollary 1 is a straightforward application of Theorem 2 to $\widehat{H}_{\mathrm{vm}}(X, Y, \mathbf{Z})$, $\widehat{H}_{\mathrm{vm}}(X, \mathbf{Z})$, $\widehat{H}_{\mathrm{vm}}(Y, \mathbf{Z})$ and $\widehat{H}_{\mathrm{vm}}(\mathbf{Z})$. The dimension-free rate comes from the assumption $\beta > 1 + d_{\mathbf{Z}}/2$, which implies that $\frac{2\beta}{2\beta+(2+d_{\mathbf{Z}})}$, $\frac{2\beta}{2\beta+(1+d_{\mathbf{Z}})}$ and $\frac{2\beta}{2\beta+d_{\mathbf{Z}}}$ are always larger than $1/2$. □

A.4. Proof of Corollary 2

Proof of Corollary 2. Let $I := I(X; Y \mid \mathbf{Z})$. The sum of type one and type two errors of T is easily bounded for n large enough by applying Corollary 1 as follows.

$$\begin{aligned}\mathbb{P}(\text{reject } H_0 \mid H_0) + \mathbb{P}(\text{accept } H_0 \mid H_1) &\leq \mathbb{P}(|\widehat{I}_{\mathrm{vm}} - I| > I_{\min}/2) + \mathbb{P}(|\widehat{I}_{\mathrm{vm}} - I| > I_{\min}/2) \\ &\leq 2C_1 \exp\left(-C_2 n^{1/2} I_{\min}/2\right).\end{aligned}$$

Finding n such that the RHS of the above is less than α concludes the proof. □

A.5. Proof of Theorem 3

Proof of Theorem 3. By definition, the number of CI tests required by this algorithm to recover the MEC is upper bounded by $2\binom{m}{2}\sum_{i=0}^{\Delta-1}\binom{m-1}{i} = O(m^{\Delta+1})$. Using Corollary 2 and the union bound, the probability that at least one of the outputs of these CI tests is incorrect is less than:

$$C_1 m^{\Delta+1} \exp\left(-C_2 n^{1/2} I_{\min}/2\right).$$

Finding n such that the RHS of the above is less than α gives $n \geq \Omega\left(\left(\frac{\Delta+1}{I_{min}}\log(m/\alpha)\right)^2\right)$ and concludes the proof. □

A.6. Proof of Theorem 4

Proof of Theorem 4. Steps 1-2 conduct $O(m)$ CI tests in the worst case, hence $O(m^2)$ CI tests are needed to recover all Markov boundaries. Recall $\max_{X\in\mathbf{X}} |\mathrm{MB}(X)| \leq \Gamma$. Then Step 3 needs $O(m\Gamma 2^{\Gamma})$ CI tests. Finally, Step 4 performs $O(m\Gamma^2 2^{\Gamma})$ tests at the worst case. The rest of the steps of the algorithm do not require CI tests. Therefore, GS requires $O(m^2 + m\Gamma^2 2^{\Gamma})$ number of CI tests.

Using Corollary 2 and the union bound, the probability that at least one of the outputs of these CI tests is incorrect is less than:

$$(m^2 + m\Gamma^2 2^{\Gamma}) m^{\Delta+1} \exp\left(-C_2 n^{1/2} I_{\min}/2\right).$$

Finding n such that the RHS of the above is less than α gives $n \geq \Omega\left(\frac{1}{I_{min}}\log\left(\frac{m^2+m\Gamma^2 2^{\Gamma}}{\alpha}\right)^2\right)$ and concludes the proof. □

B. Further on Numerical Experiments

B.1. Single Conditional Independence Test

Model In our tests, X and Y are one dimensional and $\mathbf{Z} = (Z_1, Z_2)$ is two dimensional. X, Y, Z_1, Z_2 are distributed on $[0, 1]$ with same marginal distributions $p_\beta(x) = (\beta + 1.15)x^{\beta+0.15}\mathbf{1}_{[0,1]}(x)$ for some positive integer β. Note that this distribution – often referred to as *power law distribution* – is $\beta-$Hölder smooth (see Definition 1). Next, we denote by $\mathcal{U}([0, 1])$ the uniform law on $[0, 1]$. We generate the data via inverse transform sampling as follows:

$$\begin{aligned}U_{Z,1} &\sim \mathcal{U}([0,1]) \\ U_{Z,2} &\sim \mathcal{U}([0,1]) \\ U_X \mid (U_{Z,1}, U_{Z,2}) &\sim t_1\delta_{U_{Z,1}} + t_2\delta_{U_{Z,2}} + (1 - t_1 - t_2)\,\mathcal{U}([0,1]) \\ U_Y \mid (U_{Z,1}, U_{Z,2}, U_X) &\sim t_1\delta_{U_{Z,1}} + t_2\delta_{U_{Z,2}} + t_{xy}\delta_{U_X} + (1 - t_1 - t_2 - t_{xy})\,\mathcal{U}([0,1]),\end{aligned}$$

where t_1, t_2, t_{xy} are non-negative real numbers such that $t_1 + t_2 + t_{xy} < 1$. Then X, Y, Z_1 and Z_2 are obtained as follows: $X = (U_X)^{\frac{1}{\beta+1.15}}, Y = (U_Y)^{\frac{1}{\beta+1.15}}, Z_1 = (U_{Z,1})^{\frac{1}{\beta+1.15}}$, and $Z_2 = (U_{Z,2})^{\frac{1}{\beta+1.15}}$. Note that it suffices to take $t_{xy} = 0$ to get conditional independence of X and Y given $\mathbf{Z}$. In the case where $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ we took $\beta = 3$ and $(t_1, t_2, t_{xy}) = (0.2, 0.2, 0)$. For $X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$ we took $\beta = 3$ and $(t_1, t_2, t_{xy}) = (0.2, 0.1, 0.3)$ and $I_{\min} = 0.11$.

Parameters We present in Table 1 the parameters used for numerical experiments on CI tests.

Method	Reference	Parameters	Values
KDE + Von Mises	This paper	β	3
		I_{min}	0.11
		γ s.t. $h_n = \gamma n^{-\frac{1}{2\beta+2+2}}$	0.35
KNN + Von Mises	(Poczos & Schneider, 2011)	I_{min}	0.05
		number of bins k	$\lfloor\sqrt{n}\rfloor$
MINE	(Belghazi et al., 2018)	I_{min}	0.11
		number of epochs	10 if $n \leq 100$, 50 if $n = 250$, 100 otherwise
HSIC	(Zhang et al., 2012)	statistical significance α	0.001
Gaussian	–	statistical significance α	0.05
OT-based	(Akbari et al., 2023)	threshold δ	1.7

Table 1: Parameters for CI tests in numerical experiments

Further comments on performance of MINE As shown in Figure 1a, MINE ((Belghazi et al., 2018)) performs very poorly in our experiments; the total error is close to 1. This is because the number of samples at which we work is way smaller than the number of samples required for the method to work, namely $\sim 2 \times 10^6$.

B.2. PC and GS Algorithms

The model For our experiments in Section 5.2, we used the following Structural Equation Model (SEM) to generate the data:

$$
\begin{aligned}
X_1 &:= U_1 \\
X_2 &:= U_2 \\
X_3 &:= X_1^2 + X_2 + U_3 \\
X_4 &:= U_4 \\
X_5 &:= 0.5 \times X_1^2 - 0.5 \times X_4^2 + U_5 \\
X_6 &:= X_4^3 - X_5 + U_6,
\end{aligned}
$$

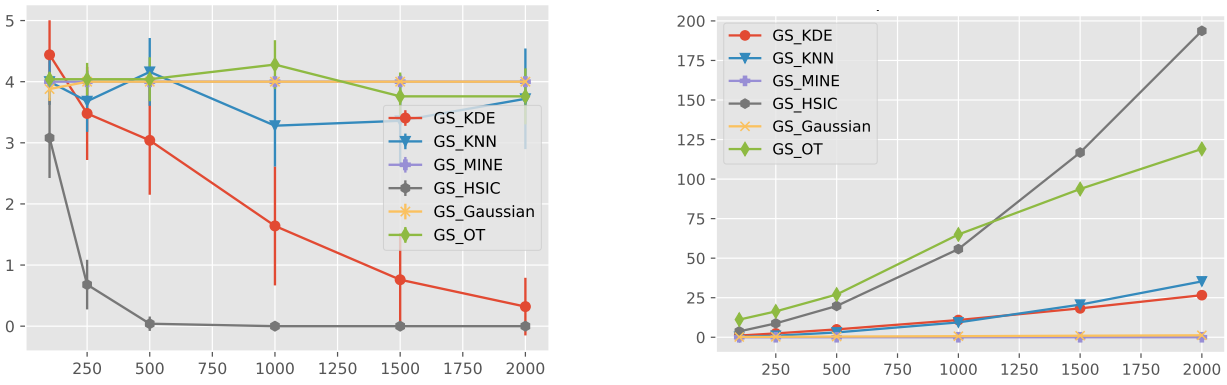
where U_i variables are i.i.d. power-law distributed with density $p_\beta(x) = (\beta + 1.15)x^{\beta+0.15}\mathbf{1}_{[0,1]}(x)$. It is clear from this SEM that the corresponding causal graph is the one displayed in Figure 2.

Parameters Table 2 provides the parameters employed in numerical experiments for the PC and GS algorithms with various CI testers.

Experiments for GS Similar to Section 5.2, we conducted experiments to evaluate the performance of the GS algorithm when using VM-CI as a CI tester vs. other CI testers. To do so, we used the aforementioned SEM. Results for GS are shown in Figure 3. Similar to the results observed for the PC algorithm, this figure illustrates that VM-CI surpasses the majority of approaches, with HSIC being the only exception. Nevertheless, analogous to PC, VM-CI competes with HSIC when the number of samples increases and offers significantly better computational efficiency than HSIC.

Method	Reference	Parameters	Values
KDE + Von Mises	This paper	β I_{min} γ s.t. $h_n = \gamma n^{-\frac{1}{2\beta+2+2}}$	3 0.01 0.35
KNN + Von Mises	(Poczos & Schneider, 2011)	I_{min} number of bins k	0.05 $\lfloor\sqrt{n}\rfloor$
MINE	(Belghazi et al., 2018)	I_{min} number of epochs	0.01 10 if $n \leq 100$, 50 if $n = 250$, 100 otherwise
HSIC	(Zhang et al., 2012)	statistical significance α	0.001
Gaussian	–	statistical significance α	0.05
OT-based	(Akbari et al., 2023)	thresholds $\delta(d_{\mathbf{Z}} = 2, \ldots, 6)$	[1.9, 1.8, 1.2, 0.4, 0.4]

Table 2: Parameters for PC and GS tests in numerical experiments



(a) Overall loss of GS algorithm with various CI testers.

(b) Runtime (s) of GS algorithm with various CI testers.

Figure 3: Results of the numerical experiments for GS (on the x-axis: number of samples n). The red curves correspond to our method.