Proceedings of Machine Learning Research vol 275:1217–1238, 2025 4th Conference on Causal Learning and Reasoning

Sample Complexity of Nonparametric Closeness Testing for Continuous Distributions and Its Application to Causal Discovery with Hidden Confounding

Fateme Jamshidi FATEME.JAMSHIDI@EPFL.CH
College of Management of Technology, EPFL, Lausanne, Switzerland

Sina Akbari SINA.AKBARI@EPFL.CH
Department of Computer and Communication Sciences, EPFL, Lausanne, Switzerland

Negar Kiyavash NEGAR.KIYAVASH@EPFL.CH
College of Management of Technology, EPFL, Lausanne, Switzerland

Editors: Biwei Huang and Mathias Drton

Abstract

We study the problem of closeness testing for continuous distributions and its implications for causal discovery. Specifically, we analyze the sample complexity of distinguishing whether two multidimensional continuous distributions are identical or differ by at least ϵ in terms of Kullback-Leibler (KL) divergence under non-parametric assumptions. To this end, we propose an estimator of KL divergence which is based on the von Mises expansion. Our closeness test attains optimal parametric rates under smoothness assumptions. Equipped with this test, which serves as a building block of our causal discovery algorithm to identify the causal structure between two multidimensional random variables, we establish sample complexity guarantees for our causal discovery method. To the best of our knowledge, this work is the first work that provides sample complexity guarantees for distinguishing cause and effect in multidimensional non-linear models with non-Gaussian continuous variables in the presence of unobserved confounding.

1. Introduction

The observation of a correlation between two variables A and B raises a fundamental question in causal inference: Does one variable cause the other, or is the correlation merely the result of a hidden confounding factor? As depicted in Figure 1, the explanation generally falls into one of three possibilities: A causes B, B causes A, or a hidden variable U causes both.

Distinguishing among these causal structures is challenging. It is well known that knowing the observational joint distribution $P(A, B)$ is not sufficient for this task, and it is necessary to perform *interventions*. In the framework of Pearl's do-calculus (Pearl, 1995), an intervention $do(A = a)$ forces A to a specific value a, allowing us to observe changes in the distribution of B. If the true structure is $A \to B$, intervening on A will affect B, but not vice versa. For $B \to A$, the reverse is true. In the case of $A \leftarrow U \to B$, interventions on A or B will not influence the other, as their correlation is only due to U.

A considerable body of work in the causal structure learning literature focuses on minimizing the number of required interventions for determining the causal structure (e.g., Eberhardt, 2007, 2012; Hauser and Bühlmann, 2012; Shanmugam et al., 2015; Kocaoglu et al., 2017; Lee and Bareinboim, 2018; Greenewald et al., 2019; Squires et al., 2020; Mokhtarian et al., 2022). However, most

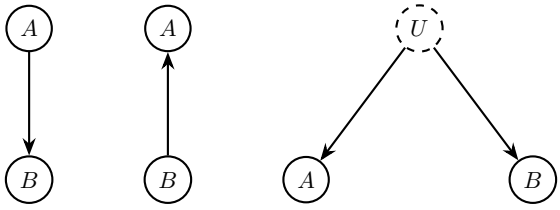

Figure 1: Three causal relationships between correlated variables A and B.

do so assuming access to an infinite number of samples, while in practice, the available data is often limited. Therefore, to ensure the reliability and applicability of causal discovery methods, it is necessary to establish sample complexity guarantees. A key question in this context is: How many samples are required to reliably infer the causal direction? In order to answer this question, we will begin by analyzing the sample complexity of estimating a measure of closeness between certain distributions, for instance, to distinguish between the observed distribution of B and the interventional distribution of B under $do(A = a)$. This will allow us to test whether there is indeed a causal influence from one variable to the other. Existing methods for closeness testing offer theoretical guarantees in discrete or structured domains (Diakonikolas et al., 2015, 2021, 2024). However, extending these guarantees to continuous variables remains challenging due to the infinite support and the complexity of estimating densities. Our work fills this gap by developing a closeness testing framework for *continuous* densities with theoretical guarantees. More specifically, we characterize the sample complexity of our test which allows us to apply it for causal discovery in the case of ***continuous***, ***multidimensional*** variables.

Contributions Our main contributions are as follows:

(i) We derive an exponential concentration inequality for an estimator of KL divergence based on Von Mises expansion when the joint densities are estimated through kernel density estimation (KDE).

(ii) Building on this result, we design a *closeness test* to decide whether two continuous probability densities are equal or differ by at least a given amount ϵ in terms of their KL divergence.

(iii) We harness this test to identify the causal structure between two multidimensional random variables and provide sample complexity guarantees for our approach, in the presence of hidden confounding.

Outline of the paper In Section 2, we briefly review the necessary background. In Section 3, we establish the exponential concentration properties of a Von Mises estimator and propose a closeness test using this estimator for KL divergence. In Section 4, we discuss how our closeness test can be harnessed to learn the causal relationship between d-dimensional continuous variables A and B and provide sample complexity guarantees for it. As a result of space constraints, some proofs are deferred to Appendix C.

1.1. Related Work

Learning the causal structure between variables of a system has been the focus of intense research for decades (Chickering et al., 1997; Spirtes et al., 2000; Glymour et al., 2019; Scanagatta et al.,

2019; Gong et al., 2023). Given the difficulty of the problem, these approaches, for the most part, assume ideal conditions, such as access to an infinite number of samples, which is rarely met in real-world scenarios. In recent years, attention has shifted toward the challenge of *finite sample complexity* in causal discovery. Several works have empirically studied the impact of sample size on the accuracy of causal inference (Eberhardt et al., 2010; Mooij et al., 2016; Yang et al., 2018). From the theoretical standpoint, Compton et al. (2022) provided formal finite-sample guarantees for two-variable systems under assumptions of causal sufficiency and an assumption on the entropy of the exogenous variable. Wadhwa and Dong (2021) studied the sample complexity of causal discovery for a network of *discrete* variables by integrating finite-sample conditional independence tests, proposed in (Canonne et al., 2018) into the causal framework of (Pearl and Verma, 1995). This result does not extend to the *continuous setting*. Recent work by Acharya et al. (2023) characterized the sample complexity required to distinguish cause from effect in bivariate one-dimensional *discrete* settings, with a success probability of at least $2/3$. They showed that the necessary number of interventional samples depends on the domain size and characterized the trade-off between observational and interventional data. Their work builds on the techniques developed by Diakonikolas et al. (2021), who focused on testing the closeness of *discrete* distributions with high probability, optimizing sample complexity as a function of parameters such as the error probability and domain size. While these studies offer valuable insights, they do not apply to the continuous setting of our interest. Indeed, causal discovery in *continuous settings* remains less explored, particularly in the finite-sample setting.

KL Divergence and Von Mises Estimators Over recent decades, considerable attention has focused on estimating KL divergence. Many of the existing methods are plug-in methods, i.e., they estimate the densities and evaluate the KL divergence functional based on these estimates. Singh and Póczos (2014) among others established convergence rates for plug-in estimators for KL divergence. In practice, evaluating the KL divergence numerically through its plug-in estimator becomes increasingly computationally expensive as the dimensionality of the variables grows. These estimators also suffer from slow convergence rates. Another simple yet effective method for the estimation of KL divergence is the k-nearest neighbors (kNN) based method (Wang et al., 2009; Póczos and Schneider, 2011, see e.g.,) although most of the work in this literature lacked convergence rate analyses. In this context, Noshad et al. (2017) provided convergence rate guarantees for a method of estimating Rényi and f-divergence measures via a graph-theoretic approach using kNN on joint data (A, B). However, both (Singh and Póczos, 2014) and (Noshad et al., 2017) have slower convergence rates and require stronger smoothness conditions than our approach in order to achieve similar convergence rates, a point we shall further discuss in the upcoming Section 3. More recently, Zhao and Lai (2020) among others studied the sample complexity of kNN-based estimators and showed that they are asymptotically optimal under different assumptions than ours, such as weak tail distribution conditions. Plug-in and kNN methods require undersmoothing the density estimate to achieve the best rate, and this smoothing parameter is in general unknown (Kandasamy et al., 2015).

The Von Mises estimators have become a valuable tool for estimating statistical functionals, such as entropy, mutual information, and divergence measures under nonparametric assumptions. These estimators, designed using the theory of influence functions and semi-parametric estimation (Fernholz, 2012; vd Vaart, 1998), are comprehensively studied for functionals of a single probability distribution (such as entropy). Kandasamy et al. (2015) proposed and analyzed estimators

for functionals of two densities (such as KL divergence) based on the Von Mises expansion. This approach, previously applied in semiparametric settings (Birgé and Massart, 1995; Robins et al., 2009), corrects for the first-order bias terms in estimation, resulting in faster convergence rates. Recently, building on this, Jamshidi et al. (2023) applied the nonparametric Von Mises estimator to estimate mutual information specifically to conditional independence (CI) testing, the core component of constraint-based causal discovery algorithms. Their work focused on recovering causal graphs up to the Markov equivalence class using only observational data. However, their framework is restricted to independence testing based on mutual information estimates with access to joint densities. This limitation prevents its application in our setting, where we need to test the *closeness* of two distributions using samples gathered under distinct conditions (e.g., either observational and interventional samples or two different interventions).

Sample complexity in closeness testing and causal discovery A line of work has focused on the sample complexity of closeness testing for discrete distributions to determine whether two distributions are identical or different. Diakonikolas et al. (2021) studied the sample complexity of distinguishing if two discrete distributions (with a constant support size) are identical or their total variation distance is greater than ϵ with probability at least $1-\delta$ for given parameters ϵ, δ and provided sample-optimal algorithms for this task.

In testing the closeness of two discrete distributions, sample complexity is largely dictated by the size of the domain. Testing for continuous distributions poses a significant challenge due to infinite domain size. In particular, if p and q are arbitrary continuous distributions, it is theoretically impossible to develop a finite-sample closeness tester with a constant probability of success (Batu et al., 2001). To address this challenge, two main approaches are commonly used in the literature. The first involves imposing structural assumptions on p and q. For instance, significant research has focused on closeness testing under Gaussianity and linearity assumptions (Diakonikolas et al., 2023; Ingster and Suslina, 2012; Verzelen and Villers, 2010). The second approach is to approximate the closeness measure through techniques such as discretizing continuous densities to achieve a finite domain size. For instance, the $\mathcal{A}_k$ distance measures the maximum l_1-distance between the reduced distributions derived from p and q over all partitions of the domain into at most k intervals. Diakonikolas et al. (2015) proposed a sample-optimal algorithm for closeness testing within univariate distribution families based on $\mathcal{A}_k$ distance. Following this, Diakonikolas et al. (2024) developed a closeness tester for multidimensional distributions, establishing upper and nearly-matching lower bounds for the sample complexity using tools from the Ramsey theory. In this work, we will follow the first approach by imposing smoothness assumptions on the densities.

2. Background

We will use kernel density estimators (KDE) throughout this work. For clarity, we include the formal definitions and properties of KDEs in Appendix A. We refer the interested reader to (Terrell and Scott, 1992; Chen, 2017) for further details. Here, we begin by reviewing a well-known exponential concentration result for these estimators. Building on this, we shall establish an exponential concentration inequality for a Von Mises estimator of KL divergence in the next section.

Before stating the concentration result, let us define a relevant notion, namely, the Hölder class, which is frequently used in the non-parametric estimation literature. For a tuple $\mathbf{s} = (s_1, \ldots, s_d)$ of non-negative integers, we define $|\mathbf{s}| = \sum_{i=1}^{d} s_i$ and let the operator $D^{\mathbf{s}}$ denote $D^{\mathbf{s}} := \frac{\partial^{|\mathbf{s}|}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}$.

Definition 1 (Hölder class) *For $L > 0$ and $\beta > 0$, the* Hölder class $\Sigma(\beta, L)$ *on $\mathcal{X} \subseteq \mathbb{R}^d$ is the set of functions $f : \mathcal{X} \to \mathbb{R}$ that are $\lfloor \beta \rfloor$ times differentiable and satisfy*

$$|D^{\mathbf{s}} f(x) - D^{\mathbf{s}} f(y)| \leq L\|x - y\|^{\beta - |\mathbf{s}|}$$

for all $\mathbf{s} = (s_1, \ldots, s_d)$ such that $|\mathbf{s}| \leq \lfloor \beta \rfloor$. A function f is called β-Hölder smooth *if $f \in \Sigma(\beta, L)$ for some $L > 0$.*

Let P be a probability measure on a compact space $\mathcal{X}$ that is absolutely continuous with respect to the Lebesgue measure, and let p denote its density function. Let $K_d : \mathbb{R}^d \to \mathbb{R}$ be a kernel and let $\hat{p}_h$ be the corresponding kernel density estimator with bandwidth h (see Eq. 13.) Under standard assumptions extensively studied in the literature (for instance, Assumption 3 in Appendix A), $\hat{p}_h$ is guaranteed to achieve minimax optimal rates of convergence to p. Assumption 3 results in optimal convergence rates, specifically outlined below. The proof of this result follows from standard bias analysis and results of (Rinaldo and Wasserman, 2010) – see (Jamshidi et al., 2023) for instance.

Proposition 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 3. 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 $n^{-\frac{\beta}{2\beta+d}} (\log n)^{1/2} \leq \varepsilon_n \leq \varepsilon_0$:*

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

2.1. Causal preliminaries

We use structural causal models (SCMs) as the semantic framework of our work (Pearl, 2009). In this framework, causal relationships between variables can be described through deterministic functions and independent noise terms. For instance, for the three structures in Figure 1, the following hold. If the causal structure is $A \to B$, the variables are generated by the SCM given as: $A := N_A, \quad B := f_B(A, N_B)$ where N_A and N_B are independent noise variables, and f_B is a deterministic function. The reverse structure, $B \to A$, can similarly be conceptualized as: $B := N_B, \quad A := f_A(B, N_A)$. In the presence of a hidden confounder U, causing both A and B, the model can be described as: $U := N_U, \quad A := f_A(U, N_A), \quad B := f_B(U, N_B)$ where N_U, N_A, and N_B are independent noise terms, with U as the latent variable inducing the correlation between A and B.

An *intervention* on the variable A refers to setting A to a specific value a, overriding the natural value it would have taken. This can be conceptualized as a modified SCM whereby $f_A(\cdot)$ is replaced by a constant function outputting the value a. We will use Pearl's *do* notation to represent interventions. In particular, $P(B|do(A = a))$ represents the probability distribution induced over the variable B in a modified SCM where the value of A is fixed at a. We will sometimes use the shorthand $P_a(B)$ when clear from context. Analogously, $P_b(A) = P(A|do(B = b))$ represents the interventional distribution of A under an intervention setting the value of B to b.

3. Exponential Concentration for KL Divergence Estimation

Here, we present our first main result: the exponential concentration bound for a KL divergence estimator based on Von Mises expansion. Let P and Q be two probability measures on a compact

set $\mathcal{X} \subseteq \mathbb{R}^d$ that are absolutely continuous with respect to the Lebesgue measure, with continuous densities p and q. The KL divergence between p and q is defined as

$$D_{\mathrm{KL}}(p\|q) := \mathbb{E}_P\big[\log \frac{p(x)}{q(x)}\big] = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}\, dx, \tag{1}$$

and $D_{\mathrm{KL}}(p\|q) = 0$ if and only if $p = q$ almost everywhere. One common approach to estimating $D_{\mathrm{KL}}(p\|q)$ is to first estimate the distributions and then plug them into the above integral, formally:

$$\hat{D}_{\mathrm{KL}}^{\text{plug-in}}(\hat{p}\|\hat{q}) = \int_{\mathcal{X}} \hat{p}(x) \log \frac{\hat{p}(x)}{\hat{q}(x)}\, dx,$$

where $\hat{p}$ and $\hat{q}$ denote the estimates of p and q, respectively. This method, although intuitive, becomes computationally intractable especially in higher dimensions d. Moreover, it results in slow convergence rates – see Remark 4. To address these issues, we will use an estimator based on the Von Mises expansion of KL divergence, which corrects for the first-order bias term, and, as we shall see, exhibits faster convergence rates. Suppose we have n and m i.i.d. samples, $\{x_1^i\}_{1\le i\le n}$ and $\{x_2^j\}_{1\le j\le m}$ from p and q, respectively. We will use an estimation procedure based on data split as follows. The dataset is divided into two subsets. Density estimates $\hat{p}$ and $\hat{q}$ are constructed using the first subset $\{x_1^i\}_{i=1}^{n/2}$ and $\{x_2^j\}_{j=1}^{m/2}$, respectively. These density estimates are plugged into the following Von Mises estimator using the second half of the data:

$$\hat{D}_{\mathrm{KL}}^{\mathrm{VM}}(\hat{p}\|\hat{q}) = \Big(\frac{2}{n}\sum_{i=n/2+1}^{n} \log \frac{\hat{p}(x_1^i)}{\hat{q}(x_1^i)}\Big) + \Big(1 - \frac{2}{m}\sum_{j=m/2+1}^{m} \frac{\hat{p}(x_2^j)}{\hat{q}(x_2^j)}\Big). \tag{2}$$

Note that $\hat{D}_{\mathrm{KL}}^{\mathrm{VM}}(p\|q)$ in Eq. (2) can be read off as the sample analogue of

$$\int_{\mathcal{X}} p(x) \log \frac{\hat{p}(x)}{\hat{q}(x)} dx + 1 - \int_{\mathcal{X}} q(x) \frac{\hat{p}(x)}{\hat{q}(x)} dx, \tag{3}$$

which is equal to $D_{\mathrm{KL}}(p\|q)$ when $\hat{p} = p$ and $\hat{q} = q$. In Appendix B, we explain in detail the derivation of the estimator in Eq. (2) based on the theory of influence functions (Newey, 1990; Ichimura and Newey, 2022), and show that this estimator corrects for the first-order bias terms in the estimation of KL divergence. In particular, using the Von Mises expansion of D_{KL} at (p, q),

$$\begin{aligned} D_{\mathrm{KL}}(p\|q) =& D_{\mathrm{KL}}(\hat{p}\|\hat{q}) + \int \Big(-D_{\mathrm{KL}}(\hat{p}\|\hat{q}) + log\frac{\hat{p}(x)}{\hat{q}(x)}\Big) p(x)\, dx + \int \Big(1 - \frac{\hat{p}(x)}{\hat{q}(x)}\Big) q(x)\, dx \\ &+ \mathcal{O}(\|p - \hat{p}\|_2^2) + \mathcal{O}(\|q - \hat{q}\|_2^2) \\ =& \int p(x) \log \frac{\hat{p}(x)}{\hat{q}(x)}\, dx + 1 - \int q(x) \frac{\hat{p}(x)}{\hat{q}(x)}\, dx + \mathcal{O}(\|p - \hat{p}\|_2^2) + \mathcal{O}(\|q - \hat{q}\|_2^2), \end{aligned} \tag{4}$$

which motivates the estimation of $D_{\mathrm{KL}}(p\|q)$ based on the form given in Eq. (2), the sample analogue of Eq. (3). In what follows, we will use kernel density estimators $\hat{p} = \hat{p}_h$, and $\hat{q} = \hat{q}_h$. To ensure that estimation is feasible, we require certain smoothness assumptions on the densities as follows.

Assumption 1 (Assumptions on the densities p and q.) *Densities p and q belong to the* Hölder classes $\Sigma(\beta_p, L)$ *and* $\Sigma(\beta_q, L)$*, respectively, for some $L > 0$. Moreover, p and q are lower-bounded on $\mathcal{X}$ by some $p_{\min} > 0$ and $q_{\min} > 0$.*

Remark 1 *Next theorem shows that assuming both p and q are bounded below by some positive constant $p_{\min}$ is essential to ensuring that the estimator $\hat{D}_{KL}^{VM}$ achieves an exponential convergence rate. When dealing with densities that do not meet this lower bound naturally, a practical solution is to truncate p and q on a sufficiently large compact interval, keeping the KL divergence close to its true value.*

Theorem 2 (Exponential concentration of $\hat{D}_{\mathbf{KL}}^{\mathbf{VM}}$ in Eq. 2) *Suppose that K_d satisfies Assumption 3 and that densities p and q both meet Assumption 1. Let the bandwidth of the kernel estimates for densities p and q equal $h_p(n) = \Theta(n^{-\frac{1}{2\beta_p+d}})$ and $h_q(m) = \Theta(m^{-\frac{1}{2\beta_q+d}})$, respectively. Then there exist positive constants n_0, m_0, $\{C_i, C_i'\}_{1\le i\le 4}$ and $\epsilon_0 > 0$ such that for any $n > n_0$, $m > m_0$, and $\max\{n^{-2\beta_p/(2\beta_p+d)}\log n, n^{-1/2}, m^{-2\beta_q/(2\beta_q+d)}\log m, m^{-1/2}\} \le \epsilon \le \epsilon_0$ such that*

$$Pr\Big(\Big|\hat{D}_{KL}^{VM}(\hat{p}_{h_p}\|\hat{q}_{h_q}) - D_{KL}(p\|q)\Big| > \epsilon\Big) \le C_1\exp\Big(-C_1' n^{1/2}\epsilon\Big) + C_2\exp\Big(-C_2' m^{1/2}\epsilon\Big) + C_3\exp\Big(-C_3' n^{\frac{2\beta_p}{2\beta_p+d}}\epsilon\Big) + C_4\exp\Big(-C_4' m^{\frac{2\beta_q}{2\beta_q+d}}\epsilon\Big). \tag{5}$$

The proof of Theorem 2 is given in Appendix C.

Remark 2 *When the smoothness parameters satisfy $\beta_p, \beta_q > \frac{d}{2}$, the estimator $\hat{D}_{KL}^{VM}$ achieves the optimal parametric convergence rate of $\mathcal{O}(n^{-1/2} + m^{-1/2})$ which is the best possible rate (Birgé and Massart, 1995; Laurent, 1996).*

Remark 3 *In the remainder of the paper, we assume $\beta_p = \beta_q = \beta$ to simplify the presentation.*

Remark 4 *From Theorem 2, we directly obtain the well-established convergence rate $\mathcal{O}(n^{-\lambda} + m^{-\lambda})$ where $\lambda = \min\{\frac{1}{2}, \frac{2\beta}{2\beta+d}\}$ for the Von Mises KL divergence estimator (see Kandasamy et al., 2015). In comparison, Singh and Póczos (2014) and Noshad et al. (2017) obtain the slower rate of $\mathcal{O}(n^{-\min\{\frac{\beta}{\beta+d}, 1/2\}})$ for n samples from p, q.*

3.1. Closeness Testing of Two Continuous Distributions

In this section, we use Theorem 2 to devise a test to distinguish between the following two hypotheses

$$H_0 := D_{\mathrm{KL}}(p\|q) = 0 \text{ vs. } H_1 := D_{\mathrm{KL}}(p\|q) > \epsilon\,.$$

We will use the Von Mises estimator of Eq. (2) to estimate $D_{\mathrm{KL}}(p\|q)$ given n and m samples from p and q, respectively, and use the following test criterion:

$$CT_{\mathrm{VM}}(n, m, \epsilon) := \begin{cases} H_1 & \text{if } \hat{D}_{\mathrm{KL}}^{\mathrm{VM}}(\hat{p}\|\hat{q}) > \epsilon/2, \\ H_0 & \text{otherwise.} \end{cases} \tag{6}$$

The following corollary is straightforward to verify using Theorem 2.

Corollary 3 (Sample Complexity of Closeness Testing) *Under the conditions in Theorem 2, given n and m i.i.d. samples drawn from p and q, respectively, $CT_{\mathrm{VM}}(n, m, \epsilon)$ distinguishes correctly between $p = q$ and $D_{KL}(p\|q) > \epsilon$ with probability at least $1 - \delta$ where ϵ and δ are positive constants as long as:*

$$n, m = \Omega\left((\frac{1}{\epsilon}\log\frac{1}{\delta})^{\tau}\right),$$

where $\tau = \max\{2, \frac{2\beta+d}{2\beta}\}$.

4. Determining Causal Relationships

We now turn our focus to the problem of identifying the causal structure between two correlated continuous d-dimensional random variables A and B using both observational and interventional data. In Subsection 4.3, we consider the case where only interventional data is accessible. We formalize the correlation between A and B as follows.

Assumption 2 *A and B are correlated. Formally, $D_{KL}\left(P(A, B)\|P(A)P(B)\right) > \epsilon$ for some $\epsilon > 0$.*

Our goal is to determine whether the relationship between A and B is that of direct causation, that is $A \to B$ or $B \to A$, or an unobserved confounder U influences both variables, resulting in the structure $A \leftarrow U \to B$ (See Figure 1). We will determine the causal structure by examining changes in the distributions under interventions. We explain this below.

Suppose the true causal relationship is $A \to B$. In this case, the interventional distribution $P_a(B) := P(B|do(A = a))$ matches $P(B|A = a)$ for every a. To test for the latter, we can use interventional samples drawn from $P_a(B)$ and observational samples drawn from $P(A, B)$ to estimate the densities and use the hypothesis testing procedure outlined in Section 3.1. On the other hand, if there is no causal edge from A to B, i.e., $A \not\to B$, then $P_a(B)$ coincides with $P(B)$ for every a, which can be tested similarly. In case our analysis does not imply $A \to B$, we can analogously test for $B \to A$ by estimating the KL divergence between $P_b(A)$ and $P(A|B = b)$. If neither direction is implied by these tests, we conclude that the relationship is likely $A \leftarrow U \to B$, where U is an unobserved variable causing both A and B. To present our formal study, we will make use of two preliminary lemmas:

Lemma 1 *Under Assumption 2,*

$$\mathbb{E}_A\left[D_{KL}\left(P(B|A)\|P(B)\right)\right] > \epsilon \quad \textit{and} \quad \mathbb{E}_B\left[D_{KL}\left(P(A|B)\|P(A)\right)\right] > \epsilon.$$

See Appendix C for the proof.

Lemma 2 (Levin, 1985; see also Fact A.2 in Goldreich, 2014) *Let P be a probability measure, and let $h : supp(P) \to [0, 1]$ be a function with $\mathbb{E}[h(t)] > \epsilon$ for some $\epsilon \in (0, 1]$. Define $k = \lceil \log_2 \frac{2}{\epsilon} \rceil$, $\epsilon_j = 2^{-j}$, and $r_j = \frac{2^j \epsilon}{(k+5-j)^2}$. Then, there exists $j \in [k]$ such that $\Pr(h(t) > \epsilon_j) > r_j$.*

We begin by outlining the procedure for testing whether the edge $A \to B$ exists below.

4.1. Testing for $A \to B$ Using Observational and Interventional Data

If $A \not\to B$, then A has no causal effect on B, and $P_a(B) = P(B)$ for every a. Clearly,

$$\mathbb{E}_A \left[D_{\mathrm{KL}} \left(P(B|do(A)) \| P(B) \right) \right] = 0$$

in this case. Conversely if $A \to B$, i.e., A causally affects B, then $P_a(B) = P(B|A = a)$ for every a and hence from Lemma 1, $\mathbb{E}_A \left[D_{\mathrm{KL}} \left(P(B|do(A)) \| P(B) \right) \right] > \epsilon$. This brings us to the following criterion for testing the edge $A \to B$:

$$\begin{cases} A \to B & \Leftrightarrow \quad \mathbb{E}\left[h(A)\right] > \epsilon, \\ A \not\to B & \Leftrightarrow \quad \mathbb{E}\left[h(A)\right] = 0, \end{cases} \tag{7}$$

where

$$h(a) \coloneqq D_{\mathrm{KL}} \left(P(B|do(A = a) \| P(B) \right). \tag{8}$$

In the sequel, we show how the criterion of Eq. (7) can be verified with high probability, using the hypothesis testing procedure of Section 3.1 and Lemma 2. Before moving forward, we note that Lemma 2 requires $h(\cdot)$ to take values in $[0, 1]$. KL divergence is non-negative, but can be unbounded in general. However, under Assumption 1, KL divergence remains upper-bounded due to p and q being bounded away from zero. Hence, Lemma 2 remains valid with an appropriate scaling of h based on the constants p_{min} and q_{min}, which does not affect our asymptotic analysis.

Let k, ϵ_j, and r_j be defined as in Lemma 2. If the edge $A \to B$ exists, then Lemma 2 implies that there exists an index $j^* \in [k]$ for which:

$$\Pr \left(D_{\mathrm{KL}} \left(P(B|a) \| P(B) \right) > \epsilon_{j^*} \right) > r_{j^*},$$

since $\mathbb{E}[h(A)] > \epsilon$. This implies the following lemma.

Lemma 3 *Suppose the causal structure is $A \to B$, and let k, ϵ_j, and r_j be defined as in Lemma 2. There exists an index $j^* \in [k]$ such that for any $c > 0$ and given $l_{j^*} = \frac{2c+2}{r_{j^*}}$ i.i.d. samples $\{a_i\}_{i=1}^{l_{j^*}}$ from $P(A)$, at least one of these samples satisfies $h(a_i) > r_{j^*}$ with probability $1 - e^{-c}$.*

The proof of this lemma, which is included in Appendix C for the sake of completeness, goes through an application of the Chernoff inequality. Note that in the alternative case, i.e., $A \not\to B$, $h(a_i)$ is always zero. Therefore, our algorithm tests whether there is an edge $A \to B$ as follows: for each $j \in [k]$, we draw $l_j = \frac{c+2}{r_j}$ i.i.d. samples from A; for each sample a_i, we draw n_j and m_j samples from $p = P\left(B(A = a_i)\right)$ and $q = P(B)$, respectively; we run the hypothesis test of Section 3.1 to test whether $h(a_i) > \epsilon_j$ or $h(a_i) = 0$; finally, if any of these tests return the hypothesis H_1 (i.e., $h(a_i) > \epsilon_j$) then we conclude the edge $A \to B$ exists; otherwise we conclude $A \not\to B$. The pseudocode for this algorithm is presented as Algorithm 1.

The following result presents the sample complexity of our method.

Theorem 4 *Suppose Assumption 2 holds, kernel K_d satisfies Assumption 3, and that observational and interventional densities satisfy Assumption 1. For any $c > 0$, Algorithm 1 correctly distinguishes between $A \to B$ and $A \not\to B$ with probability $0.9(1 - e^{-c})$, using $\mathcal{O}\left(\frac{c}{\epsilon^\tau}\right)$ observational and $\mathcal{O}\left(\frac{c}{\epsilon^\tau}\right)$ interventional samples, where $\tau = \max\{2, (2\beta + d)/2\beta\}$.*

Algorithm 1: Algorithm for Testing the Edge $A \to B$ Using Observational Data

Input: parameter c, smoothness β, threshold ϵ, sample access to distributions $P(B)$ & $P_A(B)$
Output: Determine whether the causal structure is $A \to B$
Set params $\tau = \max\{2, \frac{2\beta+d}{2\beta}\}$, $k = \log\left(\frac{2}{\epsilon}\right)$, $\epsilon_j = 2^{-j}$, $r_j = \frac{2^j\epsilon}{(k+5-j)^2}$, $\delta_j = \frac{2^{j-k}}{(2c+2)(k+5-j)^4}$;
Define sample sizes: $n_j = m_j = (\frac{1}{\epsilon_j}\log\frac{1}{\delta_j})^\tau$;
for $j = 1$ **to** k **do**
  for $i = 1$ **to** $\left\lfloor \frac{2c+2}{r_j} \right\rfloor$ **do**
    Draw sample $a_i \sim P(A)$;
    if $CT_{VM}(n_j, m_j, \epsilon_j) = H_1$ *for* $p = P_{a_i}(B)$ *and* $q = P(B)$ **then**
      return $A \to B$
return $A \not\to B$;

The number of interventional samples used throughout the algorithm is

$$\begin{aligned}
\frac{\epsilon^\tau}{\epsilon^\tau}\sum_{j\in[k]}\frac{2c+2}{r_j}n_j &= \frac{2c+2}{\epsilon^\tau}\sum_{j=1}^{k}\left(\frac{\epsilon}{\epsilon_j}\right)^{\tau-1}(k+5-j)^2\log^\tau\frac{(2c+2)(k+5-j)^4}{2^{j-k}} \\
&= \frac{20}{\epsilon^\tau}\sum_{j=1}^{k}2^{-j(\tau-1)}(j+5)^2\log^\tau\left((2c+2)(j+5)^4 2^j\right) \\
&= \frac{20\times 2^{5(\tau-1)}}{\epsilon^\tau}\sum_{j=6}^{k}\frac{j^2}{(2^{\tau-1})^j}\log^\tau\left((2c+2)j^4 2^{j-5}\right) \\
&= \frac{1}{\epsilon^\tau}\times\mathcal{O}(c) = \mathcal{O}(\frac{c}{\epsilon^\tau}).
\end{aligned} \tag{9}$$

Similarly, the number of observational samples of $P(B)$ is the same. Finally, the number of observational samples from $P(A)$ is $\sum_{j\in[k]}\frac{2c+2}{r_j}$, which is fewer than that required for $P(B)$ and can be bounded in a similar fashion.

Error analysis To evaluate the probability of error, note that the total number of closeness tests performed is bounded by $\sum_{j\in[k]}\frac{2c+2}{r_j}$, where each has an error probability of at most $\delta_j = \frac{2^{j-k}}{(2c+2)(k+5-j)^4}$. Using union bound, the probability of event E that at least one test results in an error is at most

$$\begin{aligned}
Pr(E) &\le \sum_{j\in[k]}\frac{2c+2}{r_j}\cdot\delta_j \\
&= \sum_{j\in[k]}\frac{(2c+2)(k+5-j)^2}{2^j\epsilon}\cdot\frac{2^{j-k}}{(2c+2)(k+5-j)^4} \\
&= \frac{1}{2}\sum_{j\in[k]}\frac{1}{(k+5-j)^2} = \frac{1}{2}\sum_{j=5}^{k+6}\frac{1}{j^2} \\
&< \frac{1}{2}\int_5^\infty\frac{1}{x^2}dx = 0.1,
\end{aligned} \tag{10}$$

which implies that *all* tests return correct answers with probability at least 0.9. Finally, from Lemma 3, if the true structure is $A \to B$, then at least one test will return H_1 with probability $1 - e^{-c}$, and therefore the output of the algorithm is correct with probability at least $0.9(1 - e^{-c})$. Conversely, if $A \not\to B$, all tests return H_0 and the algorithm returns the correct output with probability 0.9.

4.2. Complete Causal Discovery Algorithm

We presented the procedure for deciding between $A \to B$ and $A \not\to B$. The same procedure can be used to decide between $A \leftarrow B$ and $A \not\leftarrow B$. Finally, if we conclude that none of the edges $A \to B$ and $A \leftarrow B$ exist, the correlation between A and B is explained through the hidden confounder U: $A \leftarrow U \to B$. Based on Theorem 4 and a symmetrical result for $B \to A$ versus $B \not\to A$, along with an application of union bound we obtain the following result.

Corollary 5 *Under Assumption 1 for observational and interventional densities, Assumption 2, and Assumption 3 for the kernel, for any $c > 0$, the causal structure among $\{A \to B, A \leftarrow B, A \leftarrow U \to B\}$ can be correctly identified with probability at least $0.8(1 - 2.25e^{-c})$ given $\mathcal{O}(\epsilon^{-\tau})$ observational and interventional samples, where $\tau = \max\{2, (2\beta + d)/2\beta\}$.*

Remark 5 *Note that although this causal discovery method is initially presented for a constant probability of success, it can be boosted to achieve a success rate of $1 - \delta$ for any $\delta > 0$ by employing the median trick (Jerrum et al., 1986). This approach involves enumerating the possible causal structures and running our causal discovery method $\log(1/\delta)$ times and returning the median result, achieving an arbitrarily high probability of success with only a logarithmic factor increase in sample complexity.*

4.3. Testing for $A \to B$ Without Observational Samples

In certain applications such as online learning, observational samples might not be available. Herein, we analyze how the causal structure can be identified using only interventional data. Analogous to the previous section, we begin by presenting a method to decide between $A \to B$ and $A \not\to B$, which can be combined with its counterpart for the reverse direction ($B \to A$ versus $B \not\to A$) to form a complete causal discovery algorithm. We will use the following lemma, which is the analogue of Lemma 1. The proof is deferred to Appendix C.

Lemma 4 *Under Assumption 2,*

$$\mathbb{E}_{(A,\tilde{A})\sim P(A)\times P(A)}\left[D_{KL}\left(P(B|A)\|P(B|\tilde{A})\right)\right] > \epsilon, \quad \textit{and}$$
$$\mathbb{E}_{(B,\tilde{B})\sim P(B)\times P(B)}\left[D_{KL}\left(P(A|B)\|P(A|\tilde{B})\right)\right] > \epsilon.$$

As discussed before, if the true structure is $A \to B$, then $P(B|do(A = a)) = P(B|A = a)$ and therefore $\mathbb{E}_{(A,\tilde{A})\sim P(A)\times P(A)}\left[D_{\mathrm{KL}}\left(P(B|do(A))\|P(B|do(\tilde{A}))\right)\right] > \epsilon$. Otherwise, i.e., if $A \not\to B$, this expectation is 0 since simply $P(B|do(A)) = P(B)$. In this setting, we have the following criterion for testing the existence of the edge $A \to B$ (analogue of Eq. 7):

$$\begin{cases} A \to B & \Leftrightarrow \quad \mathbb{E}\big[h_2(A, \tilde{A})\big] > \epsilon, \\ A \not\to B & \Leftrightarrow \quad \mathbb{E}\big[h_2(A, \tilde{A})\big] = 0, \end{cases} \tag{11}$$

Algorithm 2: Algorithm for Testing the Edge $A \to B$ Without Observational Data

Input: parameter c, smoothness β, threshold ϵ, sample access to distributions $P_B(A)$ & $P_A(B)$

Output: Determine whether the causal structure is $A \to B$

Set params $\tau = \max\{2, \frac{2\beta+d}{2\beta}\}$, $k = \log\left(\frac{2}{\epsilon}\right)$, $\epsilon_j = 2^{-j}$, $r_j = \frac{2^j \epsilon}{(k+5-j)^2}$, $\delta_j = \frac{2^{j-k}}{(2c+2)(k+5-j)^4}$;

Define sample sizes: $n_j = m_j = (\frac{1}{\epsilon_j} \log \frac{1}{\delta_j})^\tau$;

for $j = 1$ **to** k **do**

  for $i = 1$ **to** $\left\lfloor \frac{2c+2}{r_j} \right\rfloor$ **do**

    Draw i.i.d. samples $a_i, \tilde{a}_i \sim P_b(A)$;

    if $CT_{VM}(n_j, m_j, \epsilon_j) = H_1$ *for* $p = P_{a_i}(B)$ *and* $q = P_{\tilde{a}_i}(B)$ **then**

      return $A \to B$

return $A \not\to B$;

where

$$h_2(a, \tilde{a}) \coloneqq D_{\mathrm{KL}}\big(P(B|do(A=a)) \| P(B|do(A=\tilde{a}))\big). \tag{12}$$

Based on this criterion, we adopt a similar testing method, outlined as Algorithm 2. The workings of this algorithm is similar to Algorithm 1. The following result presents the sample complexity of this algorithm.

Theorem 6 *Suppose Assumption 2 holds, the kernel K_d satisfies Assumption 3, and that interventional densities satisfy Assumption 1. For any $c > 0$, Algorithm 2 correctly distinguishes between $A \to B$ and $A \not\to B$ with probability $0.9(1 - e^{-c})$, using $\mathcal{O}\big(\frac{c}{\epsilon^\tau}\big)$ interventional samples, where $\tau = \max\{2, (2\beta + d)/2\beta\}$.*

The proof is identical to Theorem 4. Note that as before, the median trick can be applied to achieve arbitrarily low error probabilities. Furthermore, the edge $B \to A$ can be tested in a similar fashion, and if neither $A \to B$ nor $B \to A$ are confirmed, we conclude that $A \leftarrow U \to B$ is the true causal structure.

5. Conclusion

We presented an exponential concentration bound for a first-order Von Mises estimator of KL divergence. We developed a hypothesis testing procedure based on this estimator and analyzed its sample complexity. We then studied the problem of causal discovery involving multidimensional continuous variables in the presence of hidden confounding and developed an algorithm that correctly identifies the true causal structure with a constant probability and analyzed its sample complexity. Boosting approaches can be applied to achieve higher success probabilities at the cost of logarithmic factor increase in sample complexity.

Acknowledgments

This research was in part supported by the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40_180545.

References

Jayadev Acharya, Sourbh Bhadane, Arnab Bhattacharyya, Saravanan Kandasamy, and Ziteng Sun. Sample complexity of distinguishing cause from effect. In *International Conference on Artificial Intelligence and Statistics*, pages 10487–10504. PMLR, 2023.

Tugkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451. IEEE, 2001.

Lucien Birgé and Pascal Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, pages 11–29, 1995.

Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing conditional independence of discrete distributions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 735–748, 2018.

Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.

David Maxwell Chickering, David Heckerman, and Christopher Meek. A bayesian approach to learning bayesian networks with local structure. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 80–89, 1997.

Spencer Compton, Kristjan Greenewald, Dmitriy A Katz, and Murat Kocaoglu. Entropic causal inference: Graph identifiability. In *International Conference on Machine Learning*, pages 4311–4343. PMLR, 2022.

Ilias Diakonikolas, Daniel M Kane, and Vladimir Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1183–1202. IEEE, 2015.

Ilias Diakonikolas, Themis Gouleakis, Daniel M Kane, John Peebles, and Eric Price. Optimal testing of discrete distributions with high probability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 542–555, 2021.

Ilias Diakonikolas, Daniel M Kane, and Ankit Pensia. Gaussian mean testing made simple. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 348–352. SIAM, 2023.

Ilias Diakonikolas, Daniel M Kane, and Sihan Liu. Testing closeness of multivariate distributions via ramsey theory. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 340–347, 2024.

Frederick Eberhardt. Causation and intervention. *Unpublished doctoral dissertation, Carnegie Mellon University*, 93, 2007.

Frederick Eberhardt. Almost optimal intervention sets for causal discovery. *arXiv preprint arXiv:1206.3250*, 2012.

Frederick Eberhardt, Patrik Hoyer, and Richard Scheines. Combining experiments to discover linear cyclic models with latent variables. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 185–192. JMLR Workshop and Conference Proceedings, 2010.

Luisa Turrin Fernholz. *Von Mises calculus for statistical functionals*, volume 19. Springer Science & Business Media, 2012.

Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Oded Goldreich. On multiple input problems in property testing. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2014.

Chang Gong, Di Yao, Chuzhe Zhang, Wenbin Li, and Jingping Bi. Causal discovery from temporal data: An overview and new perspectives. *arXiv preprint arXiv:2303.10112*, 2023.

Kristjan Greenewald, Dmitriy Katz, Karthikeyan Shanmugam, Sara Magliacane, Murat Kocaoglu, Enric Boix Adsera, and Guy Bresler. Sample efficient active learning of causal trees. *Advances in Neural Information Processing Systems*, 32, 2019.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13 (1):2409–2464, 2012.

Hidehiko Ichimura and Whitney K Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61, 2022.

Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2012.

Fateme Jamshidi, Luca Ganassali, and Negar Kiyavash. On sample complexity of conditional independence testing with von mises estimator with application to causal discovery. *arXiv preprint arXiv:2310.13553*, 2023.

Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical computer science*, 43:169–188, 1986.

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

Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, et al. Nonparametric von mises estimators for entropies, divergences and mutual informations. *Advances in Neural Information Processing Systems*, 28, 2015.

Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. Cost-optimal learning of causal graphs. In *International Conference on Machine Learning*, pages 1875–1884. PMLR, 2017.

Béatrice Laurent. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? *Advances in neural information processing systems*, 31, 2018.

Leonid A Levin. One-way functions and pseudorandom generators. In *Proceedings of the seventeenth annual ACM symposium on Theory of computing*, pages 363–365, 1985.

Han Liu, Larry Wasserman, and John Lafferty. Exponential concentration for mutual information estimation with application to forests. *Advances in Neural Information Processing Systems*, 25, 2012.

Ehsan Mokhtarian, Sina Akbari, Fateme Jamshidi, Jalal Etesami, and Negar Kiyavash. Learning bayesian networks in the presence of structural side information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7814–7822, 2022.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

Whitney K Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2): 99–135, 1990.

Deborah Nolan and David Pollard. 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.

Morteza Noshad, Kevin R Moon, Salimeh Yasaei Sekeh, and Alfred O Hero. Direct estimation of information divergence using nearest neighbor ratios. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 903–907. IEEE, 2017.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

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

Judea Pearl and Thomas S Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995.

Barnabás Póczos and Jeff Schneider. On the estimation of alpha-divergences. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 609–617. JMLR Workshop and Conference Proceedings, 2011.

Alessandro Rinaldo and Larry Wasserman. Generalized density clustering, 2010.

James Robins, Lingling Li, Eric Tchetgen, and Aad W van der Vaart. Quadratic semiparametric von mises calculus. *Metrika*, 69:227–247, 2009.

Mauro Scanagatta, Antonio Salmerón, and Fabio Stella. A survey on bayesian network structure learning from data. *Progress in Artificial Intelligence*, 8(4):425–439, 2019.

Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. *Advances in Neural Information Processing Systems*, 28, 2015.

Shashank Singh and Barnabás Póczos. Exponential concentration of a density functional estimator. *Advances in Neural Information Processing Systems*, 27, 2014.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

Chandler Squires, Sara Magliacane, Kristjan Greenewald, Dmitriy Katz, Murat Kocaoglu, and Karthikeyan Shanmugam. Active structure learning of causal dags via directed clique trees. *Advances in Neural Information Processing Systems*, 33:21500–21511, 2020.

George R Terrell and David W Scott. Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265, 1992.

A. van der Vaart and J.A. Wellner. *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.

AW vd Vaart. Asymptotic statistics. cambridge series in statistical and probabilistic mathematics, 1998.

Nicolas Verzelen and Fanny Villers. Goodness-of-fit tests for high-dimensional gaussian linear models, 2010.

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

Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5): 2392–2405, 2009.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pages 5541–5550. PMLR, 2018.

Puning Zhao and Lifeng Lai. Minimax optimal estimation of kl divergence for continuous distributions. *IEEE Transactions on Information Theory*, 66(12):7787–7811, 2020.

Appendix A. Multivariate Kernel Density Estimation

Multivariate kernel density estimation (KDE) provides an approximation of the density p given by of the following form. For all

$$\widehat{p}_h(x) := \frac{2}{n}\sum_{i=1}^{n/2}\frac{1}{h^d}K_d\left(\frac{x^i - x}{h}\right), \tag{13}$$

where $x = (x_1, \ldots, x_d)$ in $\mathcal{X}$. Here, $h := h(n) > 0$ is the *bandwidth* and $K_d : \mathbb{R}^d \to \mathbb{R}$ is a *kernel* function with $\int K_d(x)\,dx = 1$, ensuring $\int_{\mathcal{X}} \widehat{p}_h(x)\,dx = 1$. Recall that for this estimation we only use the first half of the samples, i.e., $(x^i)_{1\le m\le n/2}$.

While the selection of K_d remains flexible, higher-order kernels (order $\ell > 0$) are particularly effective for approximating smooth densities. We provide their formal definition 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 $x \mapsto x^{\mathbf{s}}K(x)$ is integrable for all $|\mathbf{s}| \le \ell$ and*

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

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

Assumption 3 (Assumptions on the kernel K_d) *The kernel K_d satisfies the following:*

$(1a)$ *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*

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

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

where A and v are 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. $F(x) := \sup_{f\in\mathcal{F}}|f(x)|$), 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 $\mathcal{F}$.

Assumption $(1c)$ is a widely used condition, appearing in works such as Giné and Guillou (2002); Rinaldo and Wasserman (2010) and is fundamental to deriving the exponential inequality in Liu et al. (2012). This assumption holds for a broad class of kernels, such as polynomial kernels with compact support and Gaussian kernels (van der Vaart and Wellner, 1996; Nolan and Pollard, 1987).

Appendix B. Von Mises Expansion for KL Divergence

We review a few formal definitions from the theory of influence functions here and derive our estimator for $D_{\mathrm{KL}}(\cdot\|\cdot)$ based on the Von Mises expansion. The definitions are often given for functionals of a single distribution. Since D_{KL} is a functional of two distributions, we require the extended definitions that apply to functionals of multiple distributions. We follow (Kandasamy et al., 2015) for this purpose.

Let $\mathcal{X}$ be a compact metric space. Let $\mathcal{M}$ denote the set of all probability measures that are absolutely continuous with respect to Lebesgue[1], and with Radon-Nikodym derivities lying in $L_2(\mathcal{X})$. For $P, Q, H \in \mathcal{M}$ and a functional $T : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$, the maps $T'_P, T'_Q : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ where

$$T'_P(H; P, Q) = \frac{\partial T(P + tH, Q)}{\partial t}\Big|_{t=0}, \quad T'_Q(H; P, Q) = \frac{\partial T(P, Q + tH)}{\partial t}\Big|_{t=0},$$

are called the Gâteaux derivatives at (P, Q) if the derivatives exist and are linear in H. We say $T(\cdot, \cdot)$ is Gâteaux differentiable at (P, Q) if the Gâteaux derivatives exist at (P, Q). For a functional T that is Gâteaux differentiable at (P, Q), functions $\psi_P, \psi_Q : \mathcal{X} \to \mathbb{R}$ which satisfy the following equations are said to be the influence functions of T with respect to P and Q:

$$T'_P(H_1 - P; P, Q) = \int_{\mathcal{X}} \psi_P(x; P, Q) dH_1(x), \quad T'_Q(H_2 - Q; P, Q) = \int_{\mathcal{X}} \psi_Q(x; P, Q) dH_2(x).$$

It can be shown that the influence functions calculated below satisfy the equation above (Fernholz, 2012):

$$\psi_P(x; P, Q) = T'_P(\delta_x - P; P, Q) = \frac{\partial T\left((1-t)P + t\delta_x, Q\right)}{\partial t}\Big|_{t=0},$$
$$\psi_Q(x; P, Q) = T'_Q(\delta_x - Q; P, Q) = \frac{\partial T\left(P, (1-t)Q + t\delta_x\right)}{\partial t}\Big|_{t=0}.$$

Below, we derive these influence functions for $T \equiv D_{\mathrm{KL}}$. We further restrict our attention to measures with continuous densities. In what follows, p, q denote the densities corresponding to $P, Q \in \mathcal{M}$.

$$\begin{aligned}
\psi_p(x; p, q) &= \frac{\partial D_{\mathrm{KL}}\left((1-t)p + t\delta_x \| q\right)}{\partial t}\Big|_{t=0} = \frac{\partial}{\partial t}\int_{\mathcal{X}} \left((1-t)p + t\delta_x\right) \log\left(\frac{(1-t)p + t\delta_x}{q}\right) d\tilde{x}\Big|_{t=0} \\
&= \int_{\mathcal{X}} (-p + \delta_x) \log(\frac{p}{q}) + \int_{\mathcal{X}} (-p + \delta_x)\, d\tilde{x} \\
&= -D_{\mathrm{KL}}(p\|q) + \log\frac{p(x)}{q(x)}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\psi_q(x; p, q) &= \frac{\partial D_{\mathrm{KL}}\left(p \| (1-t)q + t\delta_x\right)}{\partial t}\Big|_{t=0} = \frac{\partial}{\partial t}\int_{\mathcal{X}} p \log\left(\frac{p}{(1-t)q + t\delta_x}\right) d\tilde{x}\Big|_{t=0} \\
&= \int_{\mathcal{X}} p - \frac{p}{q}\delta_x\, d\tilde{x} = 1 - \frac{p(x)}{q(x)}.
\end{aligned}$$

1. In general, the definitions can be adapted to an arbitrary measure μ. We work with the Lebesgue measure for simplicity.

Given these influence functions and approximations $\hat{p}$ and $\hat{q}$ of p and q, the first-order Von Mises expansion can be written as

$$\begin{aligned} D_{\mathrm{KL}}(p\|q) =& D_{\mathrm{KL}}(\hat{p}\|\hat{q}) + \int \psi_{\hat{p}}(x;\hat{p},\hat{q})p(x)\,dx + \int \psi_{\hat{q}}(x;\hat{p},\hat{q})q(x)\,dx \\ &+ \mathcal{O}(\|p-\hat{p}\|_2^2) + \mathcal{O}(\|q-\hat{q}\|_2^2) \\ =& D_{\mathrm{KL}}(\hat{p}\|\hat{q}) + \int \left(-D_{\mathrm{KL}}(\hat{p}\|\hat{q}) + log(\frac{\hat{p}}{\hat{q}})\right) p(x)\,dx + \int \left(1-\frac{\hat{p}}{\hat{q}}\right) q(x)\,dx \\ &+ \mathcal{O}(\|p-\hat{p}\|_2^2) + \mathcal{O}(\|q-\hat{q}\|_2^2) \\ =& \int p(x)\log\frac{\hat{p}(x)}{\hat{q}(x)}\,dx - \int q(x)\frac{\hat{p}(x)}{\hat{q}(x)}\,dx + 1 \\ &+ \mathcal{O}(\|p-\hat{p}\|_2^2) + \mathcal{O}(\|q-\hat{q}\|_2^2). \end{aligned} \tag{14}$$

It is clear from this expansion that the difference between $D_{\mathrm{KL}}(p\|q)$ and

$$\int p(x)\log\frac{\hat{p}(x)}{\hat{q}(x)}\,dx - \int q(x)\frac{\hat{p}(x)}{\hat{q}(x)}\,dx + 1 \tag{15}$$

is bounded by second-order remainder terms. To construct an estimator based on Von Mises expansion, we use data split. In particular, we estimate $\hat{p},\hat{q}$ using one half of the data, while using the other half to compute the sample analogue of Eq. (15). Specifically, given iid samples $\{x_1^i\}_{i=1}^m$ and $\{x_2^j\}_{j=1}^n$ drawn from p and q respectively, our data-split estimator based on Von Mises expansion is

$$\hat{D}_{\mathrm{KL}}^{\mathrm{VM}} = \frac{2}{n}\sum_{i=n/2+1}^{n}\log\frac{\hat{p}(x_1^i)}{\hat{q}(x_1^i)} - \frac{2}{m}\sum_{j=m/2+1}^{m}\frac{\hat{p}(x_2^j)}{\hat{q}(x_2^j)} + 1,$$

where samples $\{x_1^i\}_{i=1}^{m/2}$ and $\{x_2^j\}_{j=1}^{n/2}$ are used to estimate $\hat{p}$ and $\hat{q}$, respectively.

Appendix C. Omitted Proofs

Theorem 2 (Exponential concentration of $\hat{D}_{\mathrm{KL}}^{\mathrm{VM}}$ in Eq. 2) *Suppose that K_d satisfies Assumption 3 and that densities p and q both meet Assumption 1. Let the bandwidth of the kernel estimates for densities p and q equal $h_p(n) = \Theta(n^{-\frac{1}{2\beta_p+d}})$ and $h_q(m) = \Theta(m^{-\frac{1}{2\beta_q+d}})$, respectively. Then there exist positive constants n_0, m_0, $\{C_i, C_i'\}_{1\le i\le 4}$ and $\epsilon_0 > 0$ such that for any $n > n_0$, $m > m_0$, and $\max\{n^{-2\beta_p/(2\beta_p+d)}\log n, n^{-1/2}, m^{-2\beta_q/(2\beta_q+d)}\log m, m^{-1/2}\} \le \epsilon \le \epsilon_0$ such that*

$$\begin{aligned} Pr\Big(\Big|\hat{D}_{KL}^{VM}(\hat{p}_{h_p}\|\hat{q}_{h_q}) - D_{KL}(p\|q)\Big| > \epsilon\Big) \le& C_1\exp\Big(-C_1' n^{1/2}\epsilon\Big) + C_2\exp\Big(-C_2' m^{1/2}\epsilon\Big) \\ &+ C_3\exp\Big(-C_3' n^{\frac{2\beta_p}{2\beta_p+d}}\epsilon\Big) + C_4\exp\Big(-C_4' m^{\frac{2\beta_q}{2\beta_q+d}}\epsilon\Big). \end{aligned} \tag{5}$$

Proof. In order to avoid the notational burden, we drop the subscripts and denote the kernel density estimators simply by $\hat{p}$ and $\hat{q}$ throughout this proof. From Eq. (14) and by definition of $\hat{D}_{\text{KL}}^{\text{VM}}$,

$$\begin{aligned}\hat{D}_{\text{KL}}^{\text{VM}} - D_{\text{KL}}(p\|q) =& \Big(\frac{2}{n}\sum_{i=n/2+1}^{n}\log\frac{\hat{p}(x^i)}{\hat{q}(x^i)} - \int p(x)\log\frac{\hat{p}(x)}{\hat{q}(x)}\,dx\Big)\\ &- \Big(\frac{2}{m}\sum_{j=m/2+1}^{m}\frac{\hat{p}(y^j)}{\hat{q}(y^j)} - \int q(x)\frac{\hat{p}(x)}{\hat{q}(x)}\,dx\Big)\\ &+ \mathcal{O}(\|p-\hat{p}\|_2^2) + \mathcal{O}(\|q-\hat{q}\|_2^2).\end{aligned} \tag{16}$$

We will bound each term on the right-hand side separately. First note that from Proposition 1, $\hat{p}$ and $\hat{q}$ uniformly converge to p and q, respectively. Since $p(x) > p_{\min}$ and $q(x) > q_{\min}$ on the (compact) set $\mathcal{X}$, for large enough n and m (represented by constant thresholds n_0 and m_0, respectively), we have that

$$\hat{p}(x) > p_{\min}/2, \quad \hat{q}(x) > q_{\min}/2,$$

almost surely. Thus, every term in the sum $\sum_{i=n/2+1}^{n}\frac{2}{n}\log\frac{\hat{p}(x^i)}{\hat{q}(x^i)}$ is almost surely bounded by c/n where c is a positive constant. Azuma-Hoeffding inequality implies

$$\begin{aligned}Pr\left(\left|\frac{2}{n}\sum_{i=n/2+1}^{n}\log\frac{\hat{p}(x^i)}{\hat{q}(x^i)} - \int p(x)\log\frac{\hat{p}(x)}{\hat{q}(x)}\,dx\right| > \epsilon/4\right) &\le 2\exp\left(\frac{-(\frac{\epsilon}{4})^2}{2\sum_{i=n/2+1}^{n}(\frac{c}{n})^2}\right)\\ &\le C_1\exp\left(-C_1' n^{1/2}\epsilon\right),\end{aligned} \tag{17}$$

where the last inequality holds since $n^{1/2}\epsilon > 1$ by assumption.

With the same reasoning, we can conclude that each term in the sum $\sum_{j=m/2+1}^{m}\frac{2}{m}\frac{\hat{p}(y^j)}{\hat{q}(y^j)}$ is almost surely bounded by c'/m where c' is a positive constant. Again, by Azuma-Hoeffding inequality we get

$$\begin{aligned}Pr\left(\left|\frac{2}{m}\sum_{j=m/2+1}^{m}\frac{\hat{p}(y^j)}{\hat{q}(y^j)} - \int q(x)\frac{\hat{p}(x)}{\hat{q}(x)}\,dx\right| > \epsilon/4\right) &\le 2\exp\left(\frac{-(\frac{\epsilon}{4})^2}{2\sum_{j=m/2+1}^{m}(\frac{c'}{m})^2}\right)\\ &\le C_2\exp\left(-C_2' m^{1/2}\epsilon\right)\end{aligned} \tag{18}$$

where the last inequality holds since $m^{1/2}\epsilon > 1$ by assumption.

For the third term on the right-hand side of Eq. (16),

$$\begin{aligned}Pr(\|p-\hat{p}\|_2^2 > \epsilon/4) &= Pr\left(\int_{\mathcal{X}}(p(x)-\hat{p}(x))^2\,dx > \epsilon/4\right)\\ &\le Pr\left(\sup_{x\in\mathcal{X}}|p(x)-\hat{p}(x)| > \frac{\sqrt{\epsilon/4}}{\text{Vol}(\mathcal{X})}\right) \le C_3\exp\left(-C_3' n^{\frac{2\beta_p}{2\beta_p+d}}\epsilon\right),\end{aligned} \tag{19}$$

where the last inequality follows from Proposition 1. In a similar way, we get:

$$Pr(\|q-\hat{q}\|_2^2 > \epsilon/4) = Pr\left(\int_{\mathcal{X}}(q(x)-\hat{q}(x))^2\,dx > \epsilon/4\right) \le C_4\exp\left(-C_4' m^{\frac{2\beta_q}{2\beta_q+d}}\epsilon\right). \tag{20}$$

Finally, combining Equations (17), (18), (19), and (20) with a union bound, we arrive at the desired inequality:

$$Pr\Big(\Big|\hat{D}_{\text{KL}}^{\text{VM}}(\hat{p}_{h_p}\|\hat{q}_{h_q}) - D_{\text{KL}}(p\|q)\Big| > \epsilon\Big) \leq C_1 \exp\Big(-C_1' n^{1/2}\epsilon\Big) + C_2 \exp\Big(-C_2' m^{1/2}\epsilon\Big) + C_3 \exp\Big(-C_3' n^{\frac{2\beta_p}{2\beta_p+d}}\epsilon\Big) + C_4 \exp\Big(-C_4' m^{\frac{2\beta_q}{2\beta_q+d}}\epsilon\Big).$$

■

Lemma 1 *Under Assumption 2,*

$$\mathbb{E}_A\left[D_{KL}\left(P(B|A)\|P(B)\right)\right] > \epsilon \quad \textit{and} \quad \mathbb{E}_B\left[D_{KL}\left(P(A|B)\|P(A)\right)\right] > \epsilon.$$

Proof.

By Assumption 2 we have:

$$D_{\text{KL}}\left(P(A,B)\|P(A)P(B)\right) > \epsilon = \int\int P(a,b)\log\frac{P(a,b)}{P(a)P(b)}\,dadb > \epsilon.$$

Therefore,

$$\int\left(\int P(b|a)\log\frac{P(b|a)}{P(b)}\,db\right)P(a)da > \epsilon$$
$$\implies \quad \mathbb{E}_B\left[D_{\text{KL}}\left(P(b|a)\|P(b)\right)\right] > \epsilon.$$

Similarly, we can conclude that:

$$\mathbb{E}_A\left[D_{\text{KL}}\left(P(a|b)\|P(a)\right)\right] > \epsilon.$$

■

Lemma 3 *Suppose the causal structure is $A \to B$, and let k, ϵ_j, and r_j be defined as in Lemma 2. There exists an index $j^* \in [k]$ such that for any $c > 0$ and given $l_{j^*} = \frac{2c+2}{r_{j^*}}$ i.i.d. samples $\{a_i\}_{i=1}^{l_{j^*}}$ from $P(A)$, at least one of these samples satisfies $h(a_i) > r_{j^*}$ with probability $1 - e^{-c}$.*

Proof. Define event $\mathcal{E}_i = \mathbb{1}\{h(a_i) > 2^{-j^*}\}$, $\mathcal{E} = \sum_{i=1}^{l_j^*}\mathcal{E}_i$, and $\mu := \mathbb{E}[\mathcal{E}]$. From Lemma 2, $\mathbb{E}[\mathcal{E}_i] > r_j^*$ for every $i \in [l_{j^*}]$, and therefore,

$$\mu > l_{j^*} r_{j^*}. \tag{21}$$

Additionally, let $\gamma := 2c + 1 - \sqrt{2c(2c+2)}$. Note that since $(2c+1)^2 > 2c(2c+2)$ therefore $\gamma > 0$.

It can also be verified that:

$$(1 - \frac{1+\gamma}{2c+2})^2 = \frac{c}{c+1}. \tag{22}$$

Now, set $\alpha = 1 - \frac{1+\gamma}{\mu}$. Observe that $0 < \alpha < 1$, since $\mu > l_{j^*} r_{j^*} = 2c + 2 > \gamma + 1$.

Next, we bound $Pr(\mathcal{E} < 1)$ as follows:

$$\begin{aligned} Pr(\mathcal{E} < 1) \leq Pr\left(\mathcal{E} \leq 1+\gamma\right) = Pr\left(\mathcal{E} \leq (1-\alpha)\mu\right) &\overset{(a)}{\leq} \exp(-\frac{\mu\alpha^2}{2}) \\ = \exp\left(-\frac{\mu\left(1-\frac{1+\gamma}{\mu}\right)^2}{2}\right) &\overset{(b)}{\leq} \exp(-\frac{l_{j^*}r_{j^*}(1-\frac{1+\gamma}{l_{j^*}r_{j^*}})^2}{2}) \\ = \exp\left(-\frac{(2c+2)\left(1-\frac{1+\gamma}{2c+2}\right)^2}{2}\right) & \\ \overset{(c)}{=} \exp\left(-\frac{(2c+2)\frac{c}{c+1}}{2}\right) = e^{-c} & \end{aligned}$$

Here, (a) follows from the Chernoff bound, (b) from (21), and (c) from (22).

Finally, we conclude that:

$$Pr\left(\exists \quad \text{sample } a_i \text{ that satisfies} \quad h(a_i) > r_{j^*}\right) = 1 - Pr(\mathcal{E} < 1) \geq 1 - e^{-c}.$$

■

Lemma 4 *Under Assumption 2,*

$$\begin{aligned} &\mathbb{E}_{(A,\tilde{A})\sim P(A)\times P(A)}\left[D_{KL}\left(P(B|A)\|P(B|\tilde{A})\right)\right] > \epsilon, \quad \textit{and} \\ &\mathbb{E}_{(B,\tilde{B})\sim P(B)\times P(B)}\left[D_{KL}\left(P(A|B)\|P(A|\tilde{B})\right)\right] > \epsilon. \end{aligned}$$

Proof.

$$\begin{aligned} &\mathbb{E}_{(A,\tilde{A})\sim P(A)\times P(A)}\left[D_{\mathrm{KL}}\left(P(B|A)\|P(B|\tilde{A})\right)\right] \\ &\qquad \overset{(a)}{\geq} \mathbb{E}_{A\sim P(A)}\left[D_{\mathrm{KL}}\left(P(B|A)\|\mathbb{E}_{\tilde{A}\sim P(A)}\left[P(B|\tilde{A})\right]\right)\right] \\ &\qquad = \mathbb{E}_{A\sim P(A)}\left[D_{\mathrm{KL}}\left(P(B|A)\|\sum_{\tilde{A}} P(\tilde{A})P(B|\tilde{A})\right)\right] \\ &\qquad = \mathbb{E}_{A\sim P(A)}\left[D_{\mathrm{KL}}\left(P(B|A)\|P(B)\right)\right] \overset{(b)}{>} \epsilon. \end{aligned}$$

Here, (a) follows from Jensen's inequality and (b) holds by Lemma 1.

The proof of the second inequality is identical.

■