Non-parametric Conditional Independence Testing for Mixed Continuous-Categorical Variables: A Novel Method and Numerical Evaluation

Oana-Iuliana Popescu OANA-IULIANA.POPESCU@TU-DRESDEN.DE
Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany
Technische Universität Dresden, Faculty of Computer Science, Dresden, Germany
Technische Universität Berlin, Institute of Computer Engineering and Microelectronics, Berlin, Germany

Andreas Gerhardus
German Aerospace Center (DLR), Institute of Data Science

Martin Rabel
Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany
Technische Universität Dresden, Faculty of Computer Science, Dresden, Germany

Jakob Runge
Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany
Technische Universität Dresden, Faculty of Computer Science, Dresden, Germany
Technische Universität Berlin, Institute of Computer Engineering and Microelectronics, Berlin, Germany
German Aerospace Center (DLR), Institute of Data Science

Editors: Biwei Huang and Mathias Drton

Abstract

Conditional independence testing (CIT) is a common task in machine learning, e.g., for variable selection, and a main component of constraint-based causal discovery. While most current CIT approaches assume that all variables in a dataset are of the same type, either numerical or categorical, many real-world applications involve mixed-type datasets that include both numerical and categorical variables. Non-parametric CIT can be conducted using conditional mutual information (CMI) estimators combined with a local permutation scheme. Recently, two novel CMI estimators for mixed-type datasets based on k-nearest-neighbors (k-NN) have been proposed. As with any k-NN method, these estimators rely on the definition of a distance metric. One approach computes distances by a one-hot encoding of the categorical variables, essentially treating categorical variables as discrete-numerical, while the other expresses CMI by entropy terms where the categorical variables appear as conditions only. In this work, we study these estimators and propose a variation of the former approach that does not treat categorical variables as numeric. Extensive numerical experiments show that our variant detects dependencies more robustly across different data distributions and preprocessing types.

Keywords: conditional independence testing, conditional mutual information, mixed-type data

1. Introduction

Conditional independence tests (CITs) are a central component of constraint-based causal discovery (CD) frameworks, e.g., in algorithms such as PC and FCI (Spirtes et al., 2000), and are used to infer causal relations from purely observational data. A good CIT is robust toward different types, distributions, and sample sizes of the data and achieves high statistical power to detect conditional dependence while simultaneously controlling false positives at the desired level. Many real-world

applications involve mixed-type datasets, for example, weather regimes and continuous temperatures in Earth science, or sensor data in telemetry. Both parametric and non-parametric CITs for mixed-type variables have been proposed. One type of non-parametric CIT for mixed variables uses conditional mutual information (CMI) as a non-parametric measure for conditional independence, since the conditional independence $X \perp\!\!\!\perp Y \mid Z$ holds if and only if $I(X;Y|Z) = 0$ (Gray, 2011). A CIT can be formulated by combining CMI estimators with a local permutation scheme to test the null hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$ (Runge, 2018; Kim et al., 2022; Berrett et al., 2020).

In this work, we first empirically study two recent k-NN estimators of CMI for mixed-type data that have been used for CIT. The estimator of Mesner and Shalizi (2021), recently used in Huegle et al. (2023) for a CIT, transforms categorical variables by one-hot encoding and then measures distances on the resulting product space of mixed continuous-discrete variables. Zan et al. (2022) propose a CIT with a CMI estimator that only requires distance notions on the quantitative subspaces by framing mixed-type datasets as consisting of quantitative and qualitative variables, and rewriting the CMI as a linear combination of entropies where qualitative variables appear as conditions only. We discuss the disadvantages of these estimators, such as increased bias or variance, and how they can affect the performance of the CITs. We investigate how the two estimators and their CITs perform under different choices of hyperparameters, data distributions, and combinations of variable types and dimensionalities. To address the challenges faced by the estimators in the context of CIT, we propose a variant of the Mesner and Shalizi (2021) estimator that does not rely on one-hot encoding of categorical variables. Through extensive numerical experiments, we show that our variant detects dependencies more robustly across different data distributions and preprocessing types.

In summary, our main contributions are **(1)** a new k-NN estimator for the CMI of mixed-type data that is a variant of the estimator of Mesner and Shalizi (2021), **(2)** an empirical evaluation of the three CMI estimators, and **(3)** an extensive and systematic numerical evaluation of CIT performance based on the three CMI estimators in combination with a local permutation scheme.

2. Background and related work

We first introduce foundational concepts and discuss current advances in CIT for continuous and mixed variables. Then, we introduce the two CMI estimation approaches most relevant to our work.

Preliminaries Let $X : \Omega \to \mathcal{X}$, $Y : \Omega \to \mathcal{Y}$ and $Z : \Omega \to \mathcal{Z}$ be (vectors of) random variables with $\dim(\mathcal{X}) = m_X$, $\dim(\mathcal{Y}) = m_Y$, and $\dim(\mathcal{Z}) = m_Z$. We demand that $\mathcal{X} = \mathcal{X}_1 \times \ldots \mathcal{X}_{m_X}$ where $\mathcal{X}_a$ with $1 \leq a \leq m_X$ is $\mathbb{R}$ or a discrete set; similarly for $\mathcal{Y}$ and $\mathcal{Z}$. Let P_{XYZ} be the probability measure on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ induced by the joint vector (X, Y, Z). We assume that the conditional probability measure $P_{XY|Z}$ exists and is absolutely continuous with respect to the product measure $P_{X|Z} \times P_{Y|Z}$. These assumptions are fulfilled if every component of (X, Y, Z) is either discrete (i.e., absolutely continuous with respect to the counting measure) or non-singular continuous (i.e., absolutely continuous with respect to the Lebesgue measure) or a mixture of these two cases (for simplicity, we refer to "non-singular continuous" as "continuous"). We categorize discrete random variables V as: "discrete numeric", where the values of V can be a discrete subset of $\mathbb{R}$ such that distance notions on $\mathbb{R}$ other than the discrete metric are semantically meaningful, and "non-numeric" if the values are on an ordinal or nominal/categorical scale. For mixed continuous-categorical variables, we distinguish the following cases: **(1)** all of X, Y, Z are either fully discrete or fully continuous, **(2)** contain both discrete and continuous components (but no dimension of (X, Y, Z) is a mixture variable) and **(3)** at least one component of (X, Y, Z) is a mixture variable,

with both discrete and continuous values. For case 2, we denote the discrete component of variable W as W^d and the continuous components as W^c, such that $W = (W^c, W^d)$ (see concrete examples of cases 2 and 3 in App. A.1). We focus on cases 1 and 2, but our approach works for case 3 as well.

2.1. Conditional independence testing

A one-sided CIT assesses whether two random variables X, Y are independent given the values of an additional variable Z, using a test statistic $T : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ to reject the null hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$ if the test statistic value $T \geq c$, with threshold c chosen to control the type I error (false positives). The p-value indicates the probability of obtaining a test statistic at least as extreme as the observed test value. Under H_0, the p-value follows a uniform distribution over $[0, 1]$. The significance level α defines the threshold for rejecting H_0, bounding the probability of making a Type I error (false positives) by α. Type II errors (false negatives) occur when the false null hypothesis is not rejected. The power of a test is defined as $1 - \beta$, where β is the probability of a type II error. Parametric tests need further assumptions about the underlying data distributions and have higher power when these assumptions hold. Non-parametric tests do not assume specific distributions and are thus more robust and flexible. Shah and Peters (2020) show that conditional independence testing is fundamentally hard: Without assumptions to restrict the null hypothesis, no CIT controls type I error while maintaining non-trivial power. In practice, for high-dimensional settings, the null and alternative become indistinguishable, are hard to estimate, or dependencies are difficult to detect.

CIT for the continuous case A widely used parametric test for fully continuous X, Y, Z is the partial correlation test, which assumes that the variables are linearly dependent with additive Gaussian errors. Non-parametric kernel-based tests (Zhang et al., 2011) relax these assumptions but can be computationally expensive. Other non-parametric CITs, such as the generalized covariance measure (GCM) CIT of Shah and Peters (2020), rely on regression. Non-parametric CIT using CMI has been introduced in Runge (2018) and uses the CMI, estimated using the k-nearest neighbor (k-NN) estimator of Frenzel and Pompe (2007) (FP, see App. A), as a statistic in a local permutation-based test. Li and Fan (2020) provides an overview of further non-parametric tests for continuous variables. Further approaches use deep learning, e.g., to approximate the null distribution (Bellot and van der Schaar, 2019; Shi et al., 2021), or pose CIT as a classification problem (Sen et al., 2017).

CI testing for the mixed case Tsagris et al. (2018) propose a likelihood-ratio test by fitting regression models for X using Z and $Z \cup Y$, but these models make implicit distributional assumptions. Cui et al. (2016) develop a CIT under the assumption that data is drawn from a Gaussian copula, but are limited to binary and ordinal discrete variables. Handhayani and Cussens (2020) transform continuous and discrete components using kernel methods and then generate an alignment matrix used to compute partial correlation, which can be computationally expensive and is not suitable for mixture data. Another non-parametric approach uses normalizing flows (Duong and Nguyen, 2023), but mapping categorical variables to a continuous space may lead to information loss and overfitting, and can become prohibitively expensive in the context of repeated CIT, e.g. for causal discovery. Other works from the CD community deal with mixed data using score-based CD algorithms, e.g. Huang et al. (2018), but these methods do not directly compare to our approach. The approach of combining k-NN-based CMI estimation methods with a permutation-based statistical test has been adapted to the mixed-type case in Huegle et al. (2023) and Zan et al. (2022). In this work, we focus on this approach because of its flexibility, simplicity, few hyperparameters, and lower computational requirements.

2.2. *k*-NN CMI estimation in the mixed variables case

We now introduce the CMI and the two CMI estimation approaches most related to our work. These build upon previous entropy, MI and CMI estimators for continuous and mixed variables, which we describe in App. A.2. The CMI $I(X;Y|Z)$ of X and Y given Z can be defined as below, where the argument of the $\log$ is the Radon-Nikodym derivative of $P_{XY|Z}$ with respect to $P_{X|Z} \times P_{Y|Z}$ (see Gray (2011)):

$$I(X;Y|Z) = \int \log \left(\frac{dP_{XY|Z}}{d(P_{X|Z} \times P_{Y|Z})} \right) dP_{XY|Z}. \tag{1}$$

If all components of (X, Y, Z) are discrete, the rhs of eq. (1) reduces to the form in eq. (2) in terms of the probability mass functions (pmfs) $p_{\cdot}$. If all components are continuous, integrals replace sums and probability density functions (pdfs) $f_{\cdot}$ replace pmfs in eq. (2).

$$\sum_{x,y,z} p_{XYZ}(x,y,z) \log \frac{p_{XY|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)} \tag{2}$$

MS estimator The CMI estimator of Mesner and Shalizi (2021) is motivated by two observations: **(1)** the observation of Gao et al. (2017) that, in the mixed case, it is possible that the distance ρ_i of point $w_i = (x_i, y_i, z_i)$ to its k-NN in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ is $\rho_i = 0$ if w_i is fully discrete and **(2)** the observation that for fully discrete points, there is also a non-zero probability that different pairs of points have the same distance. Thus, the k-NN of w_i is non-unique with non-zero probability. A non-unique k-NN is equivalent to $k < \tilde{k}_{XYZ,i}$ with $\tilde{k}_{XYZ,i}$ as defined by eq. (3)

$$\tilde{k}_{W,i} = |\{w_j \mid \|w_j - w_i\| \leq \rho_i,\ j \neq i\}| \tag{3}$$

for $W = XYZ$ and where ρ_i is the distance of w_i to its k-th nearest neighbor. Thus, instead of $\rho_i = 0$, MS consider the event $k < \tilde{k}_{XYZ,i}$ to indicate that w_i is fully discrete. Specifically, their estimator takes the form

$$\hat{I}^{MS}(X;Y|Z) = \frac{1}{n} \cdot \sum_{i=1}^{n} \underbrace{\left[g(\tilde{k}_{XYZ,i}) + g(\tilde{k}_{Z,i}) - g(\tilde{k}_{XZ,i}) - g(\tilde{k}_{YZ,i}) \right]}_{\equiv \hat{\xi}_i^{MS}(X;Y|Z)} \tag{4}$$

where $g(\cdot) = \psi(\cdot)$, with ψ the Digamma function if $\tilde{k}_{XYZ,i} = k$ and $g(\cdot) = \log(\cdot)$ if $\tilde{k}_{XYZ,i} > k$.[1] The authors prove consistency of their estimator, and also show that it suffers from the curse of dimensionality: For fixed m_X and m_Y, if the dimension m_Z of $\mathcal{Z}$ increases to infinity and $H(Z)/m_Z$ is non-zero in this limit, then $\hat{I}^{MS}(X;Y|Z)$ converges to 0 in probability as $m_Z \to \infty$. The MS estimator equips the discrete components of $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ with the discrete metric, which is equivalent to a one-hot encoding of the components and raises the conceptual problem that the corresponding distance notions might not be semantically meaningful. The MS estimator outperforms the FP estimator (see App. A.2) and the estimator of Frenzel and Pompe (2007) for continuous variables (FP estimator, see App. A.2) which MS apply by treating discrete variables as continuous. We thus do not include RAVK and FP in our experiments. In their experiments, the authors heuristically set $k = n \cdot 0.1$ with n the sample size.

1. Mesner and Shalizi (2021) define their estimator by additionally computing a maximum of the estimate with 0, since $I(X;Y|Z) \geq 0$. However, their implementation does not seem to apply this maximum, and our preliminary experiments show that it can be detrimental to CIT. We do not apply the maximum with 0 in any experiments.

ZMADG estimator Zan et al. (2022) exclude mixture variables and propose a CMI estimator that avoids defining a distance between qualitative components. They split X, Y, and Z in their respective quantitative (discrete) components X^d, Y^d, Z^d and qualitative (continuous) components X^c, Y^c, Z^c and express the CMI as

$$I(X;Y|Z) = H(X^c, Z^c|X^d, Z^d) + H(Y^c, Z^c|Y^d, Z^d) - H(X^c, Y^c, Z^c|X^d, Y^d, Z^d) \\ - H(Z^c|Z^d) + H(X^c, Z^d) + H(Y^d, Z^d) - H(X^d, Y^d, Z^d) - H(Z^d)\,, \quad (5)$$

where the first four terms on the rhs are (conditional) differential entropies and the last four are (conditional) entropies. The entropies are estimated with the standard plug-in estimator using empirical frequencies. The differential entropies are computed using the estimator of Kozachenko and Leonenko (1987) (App. A.2) on the sample subsets defined by fixed qualitative component values. The entropies are then averaged based on the empirical frequencies of the qualitative values. The parameter k of the KL estimates is set to $k = \max\{\lfloor n_{cluster}/10 \rfloor, 1\}$ with $n_{cluster}$ the number of samples in the respective subsets determined by the values of the qualitative components (i.e., k is separately chosen for each subset of samples). As a sum of consistent estimators, the estimator is consistent. The ZMADG estimator does not seem to suffer from the curse of dimensionality as strongly as the MS estimator, but, as discussed below, we believe it incurs higher variance.

2.3. Non-parametric CIT using CMI and a local permutation scheme

To statistically test the null hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$ of conditional independence, the null distribution of the estimate $\hat{I}(X;Y|Z)$ under H_0 or an approximation thereof is needed. If $X \perp\!\!\!\perp Y \mid Z$, then the component values x_i and y_i within the subset of samples determined by the value z_i can be permuted arbitrarily without changing the distribution of the estimated CMI. Thus, setting $\tilde{x}_i = x_{\pi(i)}$ with a permutation π such that for all i both x_i and $x_{\pi(i)}$ are in the subset of samples determined by z_i, the estimators $\hat{I}(X;Y|Z)$ and $\hat{I}(\tilde{X};Y|Z)$ have the same distribution. A null distribution can be obtained since this equality holds for any such permutation. For fully discrete Z, the subset of samples determined by z_i are all samples (x_j, y_j, z_j) with $z_j = z_i$. For fully continuous Z, Runge (2018) uses a k-NN approach to determine the subsets of samples for which $z_j \approx z_i$ according to the L^∞-distance by taking the k_{perm} nearest neighbors of the i-th sample point in the Z subspace. Zan et al. (2022) and Huegle et al. (2023) adapt this method to the mixed data case: The sample (x_j, y_j, z_j) with $z_j = (z_j^c, z_j^c)$, where z_j^c is the continuous and z_j^d the discrete component, is part of the subset of samples determined by z_i if and only if $z_j^d = z_i^d$ and $z_j^c \approx z_i^c$. App. A.3 describes how p-values are obtained. As discussed in Kim et al. (2022), for discrete or mixed-type Z, CIT remains a difficult problem for high-dimensional Z, as the probability of repeatedly observing the same value of Z decreases.

3. Proposed novel estimator

3.1. Motivation: Problems of the MS and ZMADG estimators

The MS and ZMADG estimators suffer from a few problems that motivate us to introduce a novel estimator. We first highlight *three issues* of the **MS estimator**: **(1)** MS suffers from the conceptual problem that—because the k-NNs can come from different *clusters* (defined as the subsets of samples points with equal values of the discrete variable)—it implicitly assumes local constancy (see

App. A.2) across different clusters. However, different clusters might be entirely unrelated to each other. For example, it could be that dependence exists in only one of the clusters, e.g., dependencies that change with seasonal regimes. Yet, the MS estimator might estimate the local contribution of a point by combining neighbours from both clusters. Besides the conceptual complications, this can also negatively affect statistical power. **(2)** Due to the one-hot encoding of discrete non-numeric variables, the MS estimator is not invariant under scaling all variables with a common factor, as opposed to CMI. **(3)** As discussed in Mesner and Shalizi (2021), the MS estimator is biased towards 0 in high-dimensional settings. To exemplify, say the continuous and discrete numeric variables are scaled to $[0, 1]$ in preprocessing. Then, due to one-hot encoding and the L^∞-metric, the maximum distance between any two sample points is 1. If the cluster of the i-th sample point contains at most k points, then $\rho_i = 1$ (as there are not enough points in the cluster), which in turn implies $k_{i,XZ} = k_{i,YZ} = k_{i,XYZ} = k_{i,Z} = n$, with n the number of samples, and hence $\xi_i^{MS}(X;Y|Z) = 0$. Zan et al. (2022) discuss further cases in which the MS estimators suffer from local zero estimates. A bias towards zero can affect CI test performance as it can lead to false conclusions of independence.

The **ZMADG estimator** reduces these problems by considering each discrete cluster individually and adaptively reducing k (in the estimation of entropies). However, this approach can lead to *another issue* that has not yet been discussed in detail. The ZMADG estimator is a sum of up to eight individual estimated terms, where each term is derived from a subset of the data that is smaller than the total sample size. The smaller sample size of the individual terms can lead to increased variance for these terms. Consequently, aggregating these terms can increase the overall variance, compared to the variance obtained when a single estimator would be used.

3.2. Definition and intuition of the $\text{MS}_{0-\infty}$ estimator

To address these problems, we introduce a **novel CMI estimator $\text{MS}_{0-\infty}$** that combines ideas from the MS and ZMADG estimators. $\text{MS}_{0-\infty}$ can be understood as a variant of MS with the following two modifications. **(1)** Instead of one-hot encoding non-numeric variables, we keep the original space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ and equip it with the following $0-\infty$ "metric"[2]

$$||w_i - w_j||_{0-\infty} = \begin{cases} ||w_i^c - w_j^c||_{L^\infty} & \text{if } w_i^d = w_j^d \\ \infty & \text{otherwise} \end{cases}, \tag{6}$$

where we split the point $w_k = (x_k, y_k, z_k)$ into its continuous components $w_k^c = (x_k^c, y_k^c, z_k^c)$ and its discrete components $w_k^d = (x_k^d, y_k^d, z_k^d)$; similarly for the subspaces $\mathcal{X} \times \mathcal{Z}$, $\mathcal{Y} \times \mathcal{Z}$, and $\mathcal{Z}$. Thus, if w_i and w_j are in the same cluster (i.e., $w_i^d = w_j^d$), then their distance is finite and measured by the L^∞-distance, else their distance is ∞. **(2)** We adopt a heuristic (named the "local" heuristic) to adaptively set $k = \lfloor k_c \cdot n_{cl,\min} \rfloor$, where $0 < k_c < 1$ is a hyperparameter and $n_{cl,\min} = \min_{i \in [[1,n]]} |\{w_j : ||w_i - w_j||_{L^\infty} \neq \infty\}|$ is the number of points in the "smallest" cluster. The necessity of such a heuristic stems from the fact that, unlike in the infinite sample case, in practice some clusters might contain less than $k+1$ points. We choose the "local" heuristic after a comparison of performance on the CMI estimation models from Sec. 4, as described in App. E.

To formally specify our estimator, we first define the counts

$$\tilde{k}_{W,i}^{0-\infty} = |\{w_j \mid \|w_j - w_i\|_{0-\infty} \leq \rho_i,\, j \neq i\}|\,, \tag{7}$$

2. Formally, the $0-\infty$ "metric" is not a metric due to the value $+\infty$. We use this terminology to highlight the similarity with MS.

where $\rho_i < \infty$ is the $0-\infty$ distance of w_i to its k-NN (here, the distance equals the L^∞-distance as ρ_i is finite) and W stands for the vector of variables XYZ, XZ, YZ or Z. Our estimator then reads

$$\hat{I}^{0-\infty}(X;Y|Z) = \frac{1}{n} \cdot \sum_{i=1}^{n} \underbrace{\left[g(\tilde{k}^{0-\infty}_{XYZ,i}) + g(\tilde{k}^{0-\infty}_{Z,i}) - g(\tilde{k}^{0-\infty}_{XZ,i}) - g(\tilde{k}^{0-\infty}_{YZ,i})\right]}_{\equiv \hat{\xi}_i^{0-\infty}(X;Y|Z)}, \tag{8}$$

where $g(\cdot) = \psi(\cdot)$ if $\tilde{k}^{0-\infty}_{XYZ,i} = k$ and $g(\cdot) = \log(\cdot)$ if $\tilde{k}^{0-\infty}_{XYZ,i} > k$.

Despite possibly appearing minor, the modifications introduced by our estimator specifically address the above explained issues of the MS and ZMADG estimators: First, our estimator restricts all nearest neighbours of a point to the cluster of that point by construction, and thus does not assume local constancy across different clusters. Second, our estimator is invariant under a common scaling of all variables and chooses k adaptively, which reduces bias towards zero compared to the MS estimator. For example, in the case presented in Sec. 3.1, our estimator would not generically have $\xi_i^{0-\infty}(X;Y|Z) = 0$. A discussion of all cases in which MS and $\text{MS}_{0-\infty}$ have local zero estimates is, however, out of scope. Thus, an empirical evaluation of the bias towards zero is called for. Fourth, unlike the ZMADG estimator, our estimator is not the sum of up to 8 entropy terms but retains the same general form as the MS estimator. Thus, our estimator is not expected to incur increased variance, which is another hypothesis subject to empirical evaluation.

3.3. Theoretical guarantees

We provide theoretical guarantees for our estimator under the assumptions (asm.) presented in App. B.1. Besides Asm. 1 to 3, also assumed in Mesner and Shalizi (2021), we make two additional assumptions. Asm. 4 assumes that there are at most finitely many clusters as defined by the non-numeric components of XYZ. We do not see a solution for discrete variables with infinitely many values in a cluster-based approach. Asm. 5 states that all numeric components of XYZ have a finite range. We are confident that the theoretical guarantees also hold without Asm. 5, and that, to prove them, only mild adaptations of the corresponding proofs in Mesner and Shalizi (2021) are needed. Especially since this assumption is met with many types of data preprocessing (e.g., normalization), we leave an adaption to future work. [3] Due to space limitations, we provide all proofs in App. B.

Lemma 1. *Let q be a positive integer, and let $X'Y'Z'$ be obtained by applying a common non-constant affine function $h : \mathbb{R} \to \mathbb{R}$ to all numeric components of XYZ such that the ranges of all numeric components of $X'Y'Z'$ are contained within the open interval $(0,1)$.[4] Then, the difference $\hat{I}^{0-\infty}(X;Y|Z) - \hat{I}^{MS}(X';Y'|Z')$ converges to the constant 0 in L^q-norm, that is,*

$$\lim_{n\to\infty} \mathbb{E}\left[\left|\hat{I}^{0-\infty}(X;Y|Z) - \hat{I}^{MS}(X';Y'|Z')\right|^q\right] = 0\,. \tag{9}$$

Writing the difference $\hat{I}^{0-\infty}(X;Y|Z) - I(X;Y|Z)$ as $[\hat{I}^{0-\infty}(X;Y|Z) - \hat{I}^{MS}(X';Y'|Z')] + [\hat{I}^{MS}(X';Y'|Z') - I(X;Y|Z)]$ and using $I(X;Y|Z) = I(X';Y'|Z')$, Lemma 1 transfers the convergence results of the MS estimator to our estimator. Specifically, we get the following.

3. Note that our above heuristic choice of k, which is only intended for finite sample size n, does not lead to $k/n \to 0$ as $n \to \infty$. For the infinite sample case $n \to \infty$ (and thus for the theoretical results), we require that $k/n \to 0$.
4. Such a function h exists due to the Asm. 5.

Theorem 2. *Our CMI estimator $\hat{I}^{0-\infty}(X;Y|Z)$ is L^1-consistent in the k-NN limit, that is*

$$\lim_{n\to\infty} \mathbb{E}\left[\left|\hat{I}^{0-\infty}(X;Y|Z) - I(X;Y|Z)\right|\right] = 0\,. \tag{10}$$

Theorem 3. *Assume that, in addition to the requirements of the k-NN limit, $\frac{[k\cdot\ln(n)]^2}{n} \to 0$ as $n \to \infty$. Then, our CMI estimator is L^2-consistent, that is,*

$$\lim_{n\to\infty} Var\left[\hat{I}^{0-\infty}(X;Y|Z)\right] = 0\,. \tag{11}$$

In particular, our estimator is asymptotically unbiased and converges in probability to the true CMI. We prove that the $\text{CIT}_{0-\infty}$ defined in Definition 1, which uses our estimator combined with a local permutation scheme (as in Zan et al. (2022) and Huegle et al. (2023)), can control type I error (Thm. 4) while still maintaining power (Thm. 5), under the assm. of the k-NN estimators and Asm. 6.

Definition 1 (CI test $CIT_{0-\infty}$)**.** *We define the conditional independence test using the $MS_{0-\infty}$ estimator together with the k-NN local permutation scheme described in Sec. 2.3 as:*

$$CIT_{0-\infty} := \mathbf{1}\{p \leq \alpha\} \tag{12}$$

where α is the selected significance level and p is the p-value.

Theorem 4. *Under Asm. 1, 2, 4 and 6, when H_0 is true, the following holds for the $CIT_{0-\infty}$ from Def. 1 for any nominal value $\alpha \in (0, 1]$*

$$\mathbb{E}_{P_{XYZ}}[CIT_{0-\infty}] \leq \alpha. \tag{13}$$

Theorem 5. *Under Asm. 1, 2, 4 and 6, when the alternative H_1 is true, $CIT_{0-\infty}$ controls type II error, i.e., for a sufficiently high number of permutations B:*

$$\lim_{n\to\infty} \mathbb{E}_{P_{XYZ}}[1 - CIT_{0-\infty}] = 0. \tag{14}$$

The full proofs, which closely follow the proofs of Huegle et al. (2023), can be found in App. B.4.

4. Numerical evaluation of the CMI estimators

Experimental setup We compare the bias and variance of the MS, ZMADG and $\text{MS}_{0-\infty}$ estimators on four synthetic models, some of which are taken from Mesner and Shalizi (2021) and Zan et al. (2022) for reproducibility. Similar to Mesner and Shalizi (2021) and Zan et al. (2022), we do not apply any preprocessing to the continuous variables. We evaluate the mean and variances of the estimates on 100 realizations qualitatively using violin plots, and quantitatively using statistical tests. As follows, we present and study one model. Further models, results and an evaluation of runtimes are presented in App. C.

"Independent Z" (Mesner and Shalizi, 2021): X is discrete uniform $X \sim \mathcal{U}(\{0,\dots,c\})$ with $1 \leq c \in \mathbb{N}$, $Y \sim \mathcal{U}([X, X+2])$ is continuous uniform, and $Z = (Z_1,\dots,Z_d)$ is discrete with $Z_i \sim Ber(0.5)$ for $1 \leq i \leq d$. The ground truth is $I(X;Y|Z) = \ln c - \frac{c-1}{c}\cdot\ln 2$.

We set $c = 5$, vary the sample size $n \in \{300, 600, 1000, 2000\}$ and $d \in \{1, 3\}$. We compute k as $k = k_c \cdot n$ for MS; $k = \lfloor k_c \cdot n_{cl,\min} \rfloor$ for $\text{MS}_{0-\infty}$; $k = n_{subset} \cdot k_c$ for ZMADG, using $k_c \in \{0.01, 0.1, 0.2, 0.3\}$ to vary k. In their experiments, MS and ZMADG set $k_c = 0.1$, but we expect better performance for $\text{MS}_{0-\infty}$ with larger k_c, especially for small $k_{cl,\min}$.

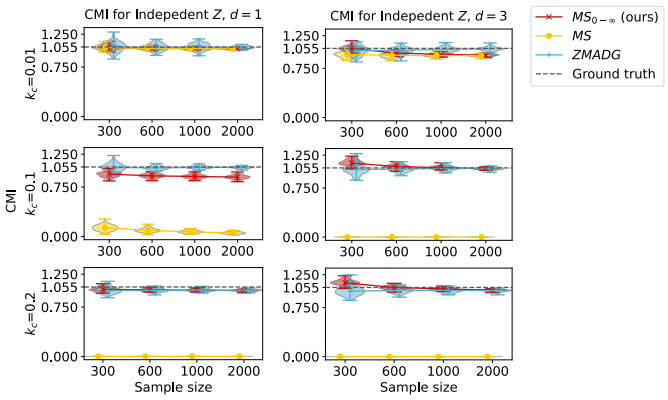

Figure 1: Distribution of the CMI estimates for the "Independent Z" model. Each row shows the results for different k_c.

Results For the "Independent Z" model (Fig. 1), we observe that for both $d = 1$ and $d = 3$ dimensions of the discrete variable Z, all estimators perform well only for $k_c = 0.01$. For $k_c > 0.01$, MS estimates are biased toward zero, unlike the $\text{MS}_{0-\infty}$ estimates. The ZMADG estimator performs well but has the highest variance, especially for small k_c and n, with a small bias for $k_c > 0.1$. Our approach shows slight bias and higher variance for smaller n or k_c, but it has lower bias than MS and lower variance than ZMADG. Across all models, the ZMADG estimator has low bias but high variance, especially for small k and n, while MS has low variance but biases towards zero for larger k and more discrete variables. A statistical comparison (see App. C) confirms that $\text{MS}_{0-\infty}$ strikes a balance between the strengths and weaknesses of the MS and ZMADG estimator, reducing bias towards zero at the cost of slightly higher bias for small n and k_c. $\text{MS}_{0-\infty}$ also handles mixture variables more robustly across k_c values, as shown in App. D.

5. Evaluation of the CITs

We study whether the CIT using $\text{MS}_{0-\infty}$ controls the false positive rate (FPR) and retains statistical power, measured by the true positive rate (TPR) at a fixed significance level $\alpha = 0.05$. We systematically compare our $\text{MS}_{0-\infty}$ with the MS and ZMADG CITs in various synthetic mixed-type data setups with known ground truth. For a broader comparison, we also compare against the kernel-based CIT for mixed-type data (Handhayani and Cussens, 2020) and the generalized covariance measure CIT (GCM, Shah and Peters (2020)). To showcase the applicability of our $\text{MS}_{0-\infty}$ CIT, we also compare CIT performances on a real-world dataset as done in Zan et al. (2022).

5.1. Synthetic data

Experimental setup To recreate a scenario close to the real world, where difficulties such as weak dependence occur, we design four data-generating models with different causal structures inspired by the post-nonlinear model of Zhang and Hyvärinen (2009). Here, we present two models where Z is a confounder. Two further models where Z is part of a chain structure or is independent of X and Y are presented in App. G. In all models, the coefficients $\beta_\cdot$ are randomly drawn as $\beta_\cdot \sim \mathcal{U}([-1, 1])$. We control the conditional dependence of X and Y given Z using an additional noise term η_w that influences both X and Y, where $\eta_w \sim \mathcal{N}(0, 1)$. The coupling factor w defines the dependence strength: For independence, $w = 0$, and for dependence $w > 0$. The random

variable $Z = (Z_1, \dots, Z_m)$ is mixed-type with dim_c continuous and $dim_d = m - dim_c$ discrete components, each with n_c categories. The noise terms η_x, η_y of X and Y follow $\mathcal{N}(0, 1)$. Motivated by the fact that Zan et al. (2022) use rank transformations on the continuous variables, which can put the MS estimator at a disadvantage due to scaling, we evaluate the CITs using standardization, re-scaling to $(0, 1)$, and rank transformation of the continuous variables.

"Confounder" model: The discrete components $Z_1, \dots, Z_{dim_d}$ follow $Z_i \sim Bin(n_c - 1, 0.5)$ and the continuous components $Z_{dim_d+1}, \dots, Z_m$ follow $Z_j \sim \mathcal{N}(0, 1)$. The continuous variable $V \in \{X, Y\}$ is computed using the following formula, where $l^{-1}(x) = \frac{e^x}{1+e^x}$ is the inverse logit function:

$$V = \sum_{j \in 1,\dots,dim_d} \beta_i \cdot l^{-1}(Z_j) + \sum_{j \in dim_c+1,\dots,m} \beta_j \cdot Z_j + \eta_v + w \cdot \eta_w \tag{15}$$

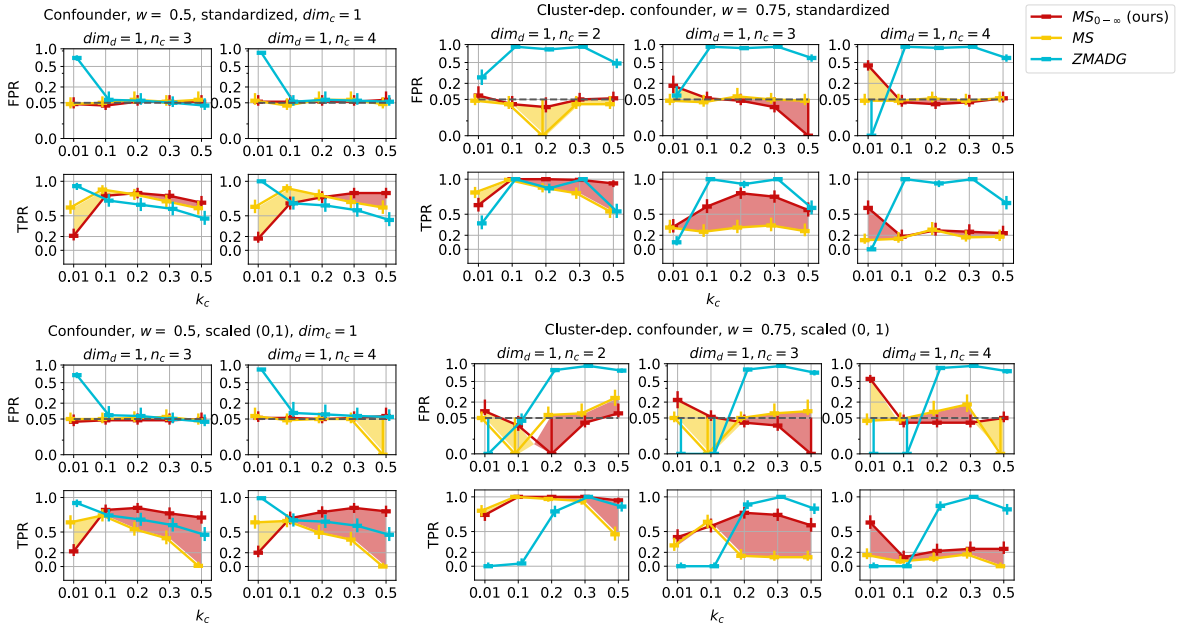

Figure 2: False positive rate (FPR, ideally under 0.05, log-scale) and true positive rate (TPR, higher is better, with 1 best) with standard error bars as the 95% confidence interval (see App. G) for the "Confounder" (left) and "Cluster-dependent confounder" (right) models with standardization (upper two rows) and scaling to $(0, 1)$ (bottom two rows) for the continuous variables and coupling factor w as depicted. The yellow areas indicate an advantage of the MS estimator. The red areas indicate an advantage of the $\text{MS}_{0-\infty}$ estimator.

"Cluster-dependent confounder" model: Here, Z is discrete univariate $Z \sim Bin(n_c - 1, 0.5)$, and X and Y are continuous univariate. For $Z = 0$, X and Y follow the "Confounder" model with coupling factor $w > 0$. For $Z \neq 0$, the "Confounder" model with $w = 0$ is used, thus, X and Y are conditionally dependent only in the cluster where $Z = 0$.

Results We present results for sample size $n = 1000$. The coupling factor for all models is $w = 0.5$, except for "Cluster-dependent confounder" where $w = 0.75$. The number of classes is $n_c \in \{3, 4\}$ for the "Confounder" model, and $n_c \in \{2, 3, 4\}$ for "Cluster-dependent confounder".

We set k as in Sec. 4 using $k_c \in \{0.01, 0.1, 0.2, 0.3, 0.5\}$. For "Confounder" we vary $dim_d \in \{1, 2\}$. We generate p-values with 300 permuted surrogates using $k_{perm} = 5$ and repeat each experiment 100 times. Here, we present results with standardization and scaling to $(0, 1)$ for the configurations displayed in Fig. 2, and postpone the results with rank transformation, and further results to App. G.

"Confounder" model: For the model where Z has one continuous component ($dim_c = 1$) and one discrete component ($dim_c = 1$), and standardized continuous variables, MS and $\text{MS}_{0-\infty}$ perform best (Fig. 2, upper left). When continuous variables are scaled to $(0, 1)$, the scaling-related issue of MS leads to a decrease in TPR as k_c increases (Fig. 2, bottom left). We observe another effect of this problem for $n_c = 4$ and $k_c = 0.5$: concurrently with the low TPR, the FPR of MS suddenly drops to 0 due to the observed and permutation statistics both being equal to 0, which results in a p-value equal to or close to 1. Our $\text{MS}_{0-\infty}$ CIT performs robustly even in this case. ZMADG gives satisfying results yet has lower TPR than MS and $\text{MS}_{0-\infty}$ and does not control FPR for $k_c = 0.01$.

"Cluster-dependent confounder" model: The results for this model (Fig. 2, right-most plots) show how the performance of MS suffers if data distributions differ between clusters. In this case, $\text{MS}_{0-\infty}$ consistently identifies dependence more accurately than MS while controlling the FPR (with some exceptions), especially for $n_c = 3$. ZMADG again suffers from either high FPR or low TPR. For comparison, in the models with dependence in each cluster, MS and $\text{MS}_{0-\infty}$ show similar performance for the standardized and rank transformation case, even as k_c increases.

In summary, MS and $\text{MS}_{0-\infty}$ perform best among the three estimators. However, $\text{MS}_{0-\infty}$ proves to be the most robust across the different models and preprocessing types. MS slightly outperforms $\text{MS}_{0-\infty}$ when dependence holds in all clusters. Yet when the continuous variables are scaled to $(0, 1)$ or data distributions differ between the clusters, the scaling-related problems and violation of local constancy of MS lead to underperformance. Surprisingly, the ZMADG CIT test obtains satisfying results only for the "Confounder" model, and does not control false positives for the other data models, possibly due to observed negative bias which might be specific to weaker dependencies.

Hyperparameter k In our numerical experiments, each estimator seems to have an optimal k_c. We consider $k_c = 0.1$ optimal for MS and ZMADG, as this value has been used in their experiments. We observe that these estimators are sensitive to higher k_c values, which lead to lower TPR or higher FPR. Our approach benefits from higher k_c, and thus higher k, since the "local" heuristic computes k by multiplying k_c with the number of samples in the smallest cluster. Thus, higher k_c can lead to a reduction in variance. With the "local" heuristic, we recommend $k_c > 0.2$, and our numerical results indicate that $k_c = 0.3$ typically works best. See App. F for further discussions on the choice of k_c.

Impact of increasing discrete dimensionality We study the impact of increasing $dim_d = 2$ for the "Confounder" model, and present results in App. G.7. For sample size $n = 1000$, $\text{MS}_{0-\infty}$ outperforms ZMADG but suffers from the curse of dimensionality, and performs slightly worse than MS. It also needs a higher k_c due to the heuristic for setting k. However, a higher sample size $n = 2000$ brings our estimator's performance on par with MS.

5.2. Comparison with other CITs

We also evaluate the kernel-based (Handhayani and Cussens, 2020) and GCM (Shah and Peters, 2020) CITs on the synthetic datasets described in Sec. 5.1. Due to space limitations, we present the details and the results of the comparison in App. I, and only shortly discuss them here. We observe that the GCM CIT consistently obtains a TPR of 1.0 for most of the models, however, the FPR is sometimes slightly higher than 0.05. However, for the "Cluster-dependent model", we observe

one disadvantage of the GCM CIT: there is high variability in the results depending on the used regressor. The kernel-based CIT does not obtain satisfying results, and most probably needs a broader hyperparameter search, which we consider out of scope here.

5.3. Real-world data

Experimental setup We evaluate whether the CITs find two CI relations on a dataset containing phenotypic data from children with attention deficit disorder (Bellec et al., 2016). The original dataset contains 23 variables. We focus here on the four variables used in Zan et al. (2022): gender (binary), attention deficit level (continuous), hyperactivity level (continuous) and medication status (binary). The two known CI relations are: **(Case 1)** gender is independent of the hyperactivity level given the attention deficit level (Bauermeister et al., 2007; Willcutt et al., 2000), and **(Case 2)** the hyperactivity level is independent of the medication status given the attention deficit level (Cui et al., 2016). We run each CIT 50 times using the three different preprocessing types for continuous variables. We present the accuracy of each CIT to find the CI relation, computed as the number of times that the respective CIT correctly accepts the null hypothesis over the total number of runs. Ideally, each CIT would have an accuracy of 1.0, independent of the preprocessing type.

Results The results in Table 4 of App. H indicate that $\text{MS}_{0-\infty}$ performs most robustly across the two CI cases for $k_c = 0.3$, where it delivers high accuracy for all three preprocessing types. However, performance drops for $k_c = 0.1$ and $k_c = 0.2$. For MS, we observe the best accuracy for $k_c = 0.3$ as well. For its optimal $k_c = 0.1$, MS fails to find the CI relation in Case 1 when rank preprocessing is applied. For $k_c \geq 0.2$, MS results vary widely, and the fact that accuracy decreases with $k_c = 0.2$, then increases again for $k_c = 0.3$, might indicate bias-related problems. ZMADG performs consistently well across preprocessing types only in Case 1, while for Case 2 it does not consistently find the CI relations across preprocessing types independent of the k_c value.

6. Discussion and conclusion

Understanding the performance of CIT on heterogeneous data is pivotal for causal discovery and machine learning, and is relevant across many applications, such as telemetry or Earth Sciences. In this work, we evaluated the k-NN CMI estimators of Mesner and Shalizi (2021) (MS) and Zan et al. (2022) (ZMADG) for mixed-type data. We discussed and evaluated the effect of their challenges on the bias/variance of the CMI estimation. As a sum of up to 8 estimators computed on subsets of the samples, the ZMADG estimator has low bias but suffers from high variance, particularly for small k and n. The MS estimator treats categorical variables as numeric via one-hot encoding, leading to the conceptual problem of mixing categories. Thus, the MS estimator is not invariant to scaling, and suffers from bias towards zero for larger k and more discrete variables. For CIT, high variance can lead to increased error rates, while bias towards zero can lead to false conclusions of independence.

We propose the novel estimator $\text{MS}_{0-\infty}$, a modification of the MS estimator which combines the advantages of the MS and ZMADG estimators: lower bias and reduced variance without the conceptual problem of mixing categories. The CIT using our estimator has the most robust performance for mixed-type data across various data distributions and preprocessing types. Surprisingly, in our experiments the ZMADG CIT obtains satisfying results only for the "Confounder" model, and does not control false positives for other models. The MS CIT slightly outperforms our CIT when dependence holds in all clusters, but the scaling-related problems and bias towards zero of MS can lead to false claims of conditional independence. When data distributions differ between the

clusters, our method has superior performance compared to MS and ZMADG. This is highly relevant in the context of mixed-type data, where distributions can vary across categories, e.g., across weather regimes in Earth science.

Thus, from a theoretical and an empirical perspective, we recommend $\text{MS}_{0-\infty}$ as the most robust estimator and CIT for mixed-type data when no parametric assumptions about the underlying data distributions can be met. However, we advise users to choose parametric over non-parametric methods whenever possible.

Limitations Our assumptions of a locally constant distribution in the k-neighborhoods can be violated if not enough samples per cluster are available, especially when k becomes small, as experiments in App. G.8 demonstrate. While we cannot ensure that Assumptions 4 and 5 always hold in practice, we have looked at models where these assumptions are violated (e.g. Poisson distribution for the discrete variables, App. C), and still observe good results. We thus only consider these assumptions necessary for the theoretical guarantees. k-NN methods rely on an appropriate k value that ensures local constancy. Furthermore, using the maximum norm can lead to focus on only one dimension of the data. This limitation is inherent to the distance metric used and thus can affect all estimators, i.e., MS and ZMADG as well, as all estimators rely on the maximum norm. To account for this problem, we apply standardization as a transformation in our synthetic data experiments. We observe that each method has an optimal k, and none of the methods consistently has good performance across all k_c values. We discuss optimal k_c values in Sec. 5 and App. F, but an extensive analysis of optimal k values is out of scope. The advantages of our estimator come with a higher runtime than MS and ZMADG for the "local" heuristic (see App. E), yet the "global" heuristic has a computational runtime close to MS and numerical experiments indicate that the performance decrease is minimal. Experiments with a higher number of discrete variables (see App. G.7) show that $\text{MS}_{0-\infty}$ also suffers from the curse of dimensionality. Users must also consider that CIT results additionally depend on the chosen number of permutations. While our analysis covers a range of scenarios, an evaluation of causal discovery is beyond our scope and is left for future work.

Reproducibility All code for the evaluation of the CITs is available at `https://github.com/oanaipopescu/cmiknnmixed`. The details necessary for replicating our experiments, such as random seeds, can be found in the code as well. Furthermore, the implementation of the estimators and the associated CITs are now also part of the Tigramite package at `https://github.com/jakobrunge/tigramite`. Further remarks on reproducibility can be read in App. K.

Acknowledgments

We thank Tom Hochsprung for his helpful comments on the draft. J.R. has received funding from the European Research Council (ERC) Starting Grant CausalEarth under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 948112). J.R., O.I.P., and M.R. have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101003469 (XAIDA). O.I.P has received funding from the German Aerospace Center (DLR). This work was supported by the German Federal Ministry of Education and Research (BMBF, SCADS22B) and the Saxon State Ministry for Science, Culture and Tourism (SMWK) by funding the competence center for Big Data and AI "ScaDS.AI Dresden/Leipzig". The authors gratefully acknowledge the GWK support for funding this project by providing computing time through the Center for Information Services and HPC (ZIH) at TU Dresden.

References

José J. Bauermeister, Patrick E. Shrout, Ligia M. Chavez, Maritza Rubio-Stipec, Rafael R. Ramírez, L Padilla, Adrianne Anderson, Pedro García, and Glorisa J. Canino. Adhd and gender: are risks and sequela of adhd the same for boys and girls? *Journal of child psychology and psychiatry, and allied disciplines*, 48 8:831–9, 2007.

Pierre Bellec, Carlton Chu, François Chouinard-Decorte, Yassine Benhajali, Daniel S. Margulies, and R. Cameron Craddock. The neuro bureau adhd-200 preprocessed repository. *bioRxiv*, 2016. doi: 10.1101/037044.

Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Thomas B. Berrett, Richard J. Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288–318, February 2019. ISSN 0090-5364. doi: 10.1214/18-AOS1688.

Thomas B. Berrett, Yi Wang, Rina Foygel Barber, and Richard J. Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, February 2020. ISSN 1369-7412. doi: 10.1111/rssb.12340.

Carlo E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60. Tipografia del Senato, 1935.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Ruifei Cui, Perry Groot, and Tom Heskes. Copula pc algorithm for causal discovery from mixed data. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken, editors, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages 377–392. Springer, Cham, 2016.

Bao Duong and Thin Nguyen. Normalizing flows for conditional independence testing. *Knowl. Inf. Syst.*, 66(1):357–380, August 2023. ISSN 0219-1377. doi: 10.1007/s10115-023-01964-w.

Stefan Frenzel and Bernd Pompe. Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.*, 99:204101, Nov 2007. doi: 10.1103/PhysRevLett.99.204101.

Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 5967–5978, Long Beach, CA, USA, 2017. Curran Associates, Inc.

Robert M Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

Teny Handhayani and James Cussens. Kernel-based approach for learning causal graphs from mixed data. In Manfred Jaeger and Thomas Dyhre Nielsen, editors, *Proceedings of the 10th International Conference on Probabilistic Graphical Models*, volume 138 of *Proceedings of Machine Learning Research*, pages 221–232. PMLR, 23–25 Sep 2020.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1551–1560, New York, NY, USA, 2018. Association for Computing Machinery. doi: 10.1145/3219819.3220104.

Johannes Huegle, Christopher Hagedorn, and Rainer Schlosser. A knn-based non-parametric conditional independence test for mixed data and application in causal discovery. In *Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part I*, page 541–558, Berlin, Heidelberg, 2023. Springer-Verlag. doi: 10.1007/978-3-031-43412-9_32.

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Local permutation tests for conditional independence. *The Annals of Statistics*, 50, 12 2022. doi: 10.1214/22-AOS2233.

L. F. Kozachenko and N. N. Leonenko. Sample estimate of entropy of a random vector. *Problems of Information Transmission*, (23):95–101, 1987.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138.

Howard Levene. Robust test for equality of variances. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling (I. Olkin ed.)*, pages 278–292, 1960.

Chun Li and Xiaodan Fan. On nonparametric conditional independence tests for continuous variables. *WIREs Comput. Stat.*, 12(3), April 2020. doi: 10.1002/wics.1489.

Octavio César Mesner and Cosma Rohilla Shalizi. Conditional mutual information estimation for mixed, discrete and continuous data. *IEEE Trans. Inf. Theor.*, 67(1):464–484, January 2021. doi: 10.1109/TIT.2020.3024886.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Arman Rahimzamani, Himanshu Asnani, Pramod Viswanath, and Sreeram Kannan. Estimators for multivariate information measures in general probability spaces. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In Amos Storkey and Fernando Perez-Cruz, editors, *International Conference*

on Artificial Intelligence and Statistics, volume 84 of *Proceedings of Machine Learning Research*, pages 938–947. PMLR, 2018.

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514 – 1538, 2020. doi: 10.1214/19-AOS1857.

Chengchun Shi, Tianlin Xu, Wicher Bergsma, and Lexin Li. Double generative adversarial networks for conditional independence testing. *J. Mach. Learn. Res.*, 22(1), January 2021.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Michail Tsagris, Giorgos Borboudakis, Vincenzo Lagani, and I. Tsamardinos. Constraint-based causal discovery with mixed data. *International Journal of Data Science and Analytics*, 6:19 – 30, 2018.

Larry Wasserman. *All of statistics : a concise course in statistical inference*. Springer, New York, 2010. ISBN 9781441923226 1441923225.

Frank Wilcoxon. *Individual Comparisons by Ranking Methods*, pages 196–202. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_16.

Erik G. Willcutt, Bruce F. Pennington, and John C. Defries. Etiology of inattention and hyperactivity/impulsivity in a community sample of twins with learning difficulties. *Journal of Abnormal Child Psychology*, 28:149–159, 2000.

Lei Zan, Anouar Meynaoui, Charles K. Assaad, Emilie Devijver, and Eric Gaussier. A conditional mutual information estimator for mixed data and an associated conditional independence test. *Entropy*, 24(9), 2022. doi: 10.3390/e24091234.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In David McAllester, editor, *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 647–655, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In Avi Pfeffer Fabio Cozman, editor, *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, Arlington, Virginia, USA, 2011. AUAI Press.

Appendix A. Further background and related work

A.1. Example of mixed and mixture-type variables

We exemplify cases 2 and 3 presented in the Preliminaries of Sec. 2 to highlight the difference between mixed and mixture-type variables and show how they can be split into discrete and continuous components. First, given (possibly multivariate) variables X, Y, and Z, we denote with (X, Y, Z) the combined variable. A component of (X, Y, Z) is a component of the joint variable, which is a component of X, Y, or Z.

An example of the second case is: $X = (X_1, X_2)$ with discrete component X_1 and continuous component X_2, continuous $Y = (Y_1, Y_2)$ and $Z = \emptyset$. Here, the continuous component of the joint vector is $(X^d, Y^d, Z^d) = (X_2, Y_1, Y_2)$ and the discrete component is $(X^c, Y^c, Z^c) = (X_1)$.

An example of the third case is: $X = (X_1)$ with X_1 a mixture, $Y = (Y_1)$ with discrete Y_1 and $Z = \emptyset$.

An example of a mixture is $V \in A \cup B$, with $A, B \subset \mathbb{R}$, where V takes continuous values on A and discrete values on B. A more concrete numerical example can be found in App. D.

A.2. Further entropy, MI and CMI estimators

A.2.1. ESTIMATION IN THE FULLY CONTINUOUS CASE

Due to space limitations in the main paper, we introduce the entropy, MI, and CMI k-NN estimators for the fully continuous case, which lay the foundation of the MS and ZMADG CMI estimators here.

KL estimator for differential entropy Let $W : \Omega \to \mathcal{W}$ be a (vector of) continuous random variables with $\mathcal{W} = \mathbb{R}^{m_W}$ and let $w_1, \ldots, w_n$ be *iid* observations of W. The Kozachenko and Leonenko (1987) (KL) estimator of the differential entropy $H(W)$ is the sample average

$$\hat{H}^{KL}(W) = -\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_W(w_i). \tag{16}$$

The local density estimates $\hat{f}_W(w_i)$ are calculated under the assumption that f_W is locally constant within an L^p-ball $B(w_i, \rho_i)$ of radius ρ_i around w_i where ρ_i is the L^p-distance of w_i to its k-th nearest neighbor (not counting w_i itself) for some positive integer k. Since W is continuous, this k-nearest neighbor (k-NN) is unique with probability one. **The local constancy assumption** implies that the probability P_i of the event $w \in B(w_i, \rho_i)$ is $P_i \equiv p_w(w_i) \cdot V_{m_W,p} \cdot \rho_i^{m_W}$, where $V_{m_W,p}$ is the volume of the unit-ball in the L^p-metric. Using that $\mathbb{E}[\log P_i] = \psi(k) - \psi(n)$ with the Digamma function $\psi(x)$, see Kraskov et al. (2004), and approximating $E[\log \rho_i]$ by a sample average, eq. (16) then takes the form

$$\hat{H}^{KL}(W) = \psi(n) - \psi(k) + \log V_{m_W,p} + \frac{m_W}{n}\sum_{i=1}^{n} \log \rho_i \,. \tag{17}$$

KSG estimator for mutual information Kraskov et al. (2004) estimate the MI $I(X;Y) = H(X) + H(Y) - H(X,Y)$ by estimating the three individual entropies with the KL estimator (17). The authors heuristically argue that the errors incurred by the local constancy assumptions approximately cancel out in the combined estimator if all entropy estimates use the same local length scales ρ_i. Thus, they equip $\mathcal{X} \times \mathcal{Y}$ with the maximum metric $d_{\mathcal{X}\times\mathcal{Y}}(\cdot,\cdot) = \max\{d_{\mathcal{X}}(\cdot,\cdot), d_{\mathcal{Y}}(\cdot,\cdot)\}$ and

define ρ_i as in the KL estimate (17) of $H(X,Y)$, and use the same radii ρ_i for estimating $H(X)$ and $H(Y)$. The estimator takes the form

$$\hat{I}^{KSG}(X;Y) = \frac{1}{n}\sum_{i=1}^{n}[\psi(k) + \psi(n) - \psi(k_{X,i}+1) - \psi(k_{Y,i}+1)]\,, \tag{18}$$

where $k_{X,i}$ and $k_{Y,i}$ are defined by (W is placeholder for X and Y and w is placeholder for x and y)

$$k_{W,i} = |\{w_j \mid \|w_j - w_i\| < \rho_i,\, j \neq i\}|\,, \tag{19}$$

that is, as the number of points $x_j \neq x_i$ (resp. $y_j \neq y_i$) within the *open* ball $B(x_i, \rho_i)$ (resp. $B(y_i, \rho_i)$) in $\mathcal{X}$ (resp. $\mathcal{Y}$). The terms with $\log \rho_i$ cancel out due to using the same radii ρ_i in all three entropy estimates. Since $\mathcal{X} \times \mathcal{Y}$ is equipped with the maximum metric, the volume terms cancel out too.

FP estimator of conditional mutual information Using the same rationale, Frenzel and Pompe (2007) extend the KSG estimator to CMI, with $k_{Z,i}$, $k_{XZ,i}$, $k_{YZ,i}$ as in eq. (19) and ρ_i as in the KL estimate of $H(X,Y,Z)$:

$$\hat{I}^{FP}(X;Y|Z) = \frac{1}{n}\sum_{i=1}^{n}[\psi(k) + \psi(k_{Z,i}+1) - \psi(k_{X,i}+1) - \psi(k_{Y,i}+1)]\,. \tag{20}$$

A.2.2. ESTIMATION IN THE MIXED VARIABLES CASE

GKOV estimator of mutual information Gao et al. (2017) propose an estimator for mixed MI $I(X;Y)$ under the assumption that both $\mathcal{X}$ and $\mathcal{Y}$ are Euclidean spaces, thus implicitly requiring that the discrete variables are either numerical with a semantically meaningful notion of distance or have been mapped to a real space (that is, ignoring the conceptual problem of a semantically non-meaningful L^p-distance). The GKOV estimator builds on KSG and the observation that, in the mixed case, the distance ρ_i of (x_i, y_i) to its k-th nearest neighbor in $\mathcal{X} \times \mathcal{Y}$ can be $\rho_i = 0$ with non-zero probability. Gao et al. (2017) consider the event $\rho_i = 0$ to indicate that point (x_i, y_i) is discrete. Their estimator takes the form

$$\hat{I}^{GKOV}(X;Y) = \frac{1}{n}\sum_{i=1}^{n}[\psi(\tilde{k}'_i) + \log(n) - \log(\tilde{k}_{X,i}+1) - \log(\tilde{k}_{Y,i}+1)]\,, \tag{21}$$

where $\tilde{k}'_i = k$ if $\rho_i > 0$ and $\tilde{k}'_i = \tilde{k}_{XY,i}$ if $\rho_i = 0$ with

$$\tilde{k}_{W,i} = |\{w_j \mid \|w_j - w_i\| \leq \rho_i,\, j \neq i\}|\,. \tag{22}$$

As opposed to eq. (19), eq. (22) uses the non-strict inequality $\|w_j - w_i\| \leq \rho_i$. The combination of $\psi(\cdot)$ and $\log(\cdot)$ terms is ad-hoc and ultimately justified by their consistency proof.

RAVK estimator Rahimzamani et al. (2018) propose a generalization of multiple information theoretic measures as the Kullback-Leibler divergence between the joint distribution of X, Y and Z and their factorization according to a directed acyclic graph. The CMI is a special case of this measure. Further building on the observation from Gao et al. (2017) that, in the mixed case, it is possible that the distance ρ_i of point $w_i = (x_i, y_i, z_i)$ to its k-th NN in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ can be $\rho_i = 0$ if

w_i is fully discrete, the event $\rho_i = 0$ is considered to indicate that w_i is discrete. The estimator is defined as:

$$\hat{I}^{RAVK}(X;Y|Z) = \frac{1}{n} \cdot \sum_{i=1}^{n} \left[\psi(\tilde{k}'_{XYZ,i}) + log(\tilde{k}_{Z,i} + 1) - log(\tilde{k}_{XZ,i} + 1) - log(\tilde{k}_{YZ,i} + 1) \right] \tag{23}$$

where the number of neighbors $\tilde{k}_{W,i}$ are defined as the number of points $w_j \neq w_i$ within the *open* ball $B(w_i, \rho_i)$ in subspace W (i.e., XYZ, XZ, XY and Z, respectively), including boundary points:

$$\tilde{k}_{W,i} = |\{w_j \mid \|w_j - w_i\| \leq \rho_i,\, j \neq i\}|\,, \tag{24}$$

and $\tilde{k}'_i = k$ if $\rho_i > 0$ and $\tilde{k}'_i = \tilde{k}_{XYZ,i}$ if $\rho_i = 0$.

A.3. Nearest-neighbor permutation test

Algorithm 1, adapted from Kim et al. (2022) (see also Wasserman (2010)), describes the procedure for generating a p-value from a set of B permutations $\Pi = \{\pi_j | j = 1, ..., B\}$ of n elements. As described in Sec. 2.3, in our case, each permutation is generated using the nearest-neighbor scheme. For the permutation-based CITs discussed in this paper, the test statistic is the conditional mutual information (CMI), which can be estimated using one of the estimators: MS, $\text{MS}_{0-\infty}$, or ZMADG. We denote the value of the test statistic computed using the chosen estimator $estim$ as $T_{CMI,estim}$. The p-value of the conditional independence test (CIT) of the respective estimator $estim$ as follows:

Algorithm 1 Local Permutation Test

Input: samples $\{(x_i, y_i, z_i) | i = 1, ..., n\}$, permutation set $\Pi = \{\pi_j | j = 1, ..., B\}$, test statistic $T_{CMI,estim}$, nominal level α

Output: p-value

1 For each permutation $\pi_j \in \Pi, j = 1, ..., B$: compute the statistic $T^{\pi_j}_{CMI,estim}$ on the permuted samples

2 Compare the test statistic on the observed samples, $T_{CMI,estim}$ with the permuted test statistics and calculate the p-value as

$$p = \frac{\sum_{\pi_j \in \Pi} \mathbf{1}\{T^{\pi_j}_{CMI,estim} \geq T_{CMI,estim}\} + 1}{B + 1} \tag{25}$$

Appendix B. Proofs

Here, we formally prove the theoretical claims made in Sec. 3 of the main paper.

B.1. Assumptions

As the proof of Mesner and Shalizi (2021) relies on the following assumptions, we make them here too:

Assumption 1. *$P_{XY|Z}$ is non-singular such that $f \equiv \frac{dP_{XY|Z}}{d(P_{X|Z} \times P_{Y|Z})}$ is well defined, and, for some $C > 0$, $f(x, y, z) < C$ for all $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$.*

Assumption 2. $\{(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}\} : P_{XYZ}((x, y, z)) > 0$ *is countable and nowhere dense in* $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$.

Assumption 3. $k \to \infty$ *and* $\frac{k}{n} \to 0$ *as* $n \to \infty$.

Furthermore, our proofs rely on two additional assumptions:

Assumption 4. *There are at most finitely many clusters as defined by the non-numeric components of* XYZ.

Assumption 5. *All numeric components of* XYZ *have a finite range.*

Note that Mesner and Shalizi (2021) also implicitly make Assumption 4 by assuming finite-dimensional random vectors in combination with one-hot encoding of the non-numeric components. As mentioned in the main text, we are confident that the theoretical guarantees of our estimator also hold without Assumption 5 and that, to prove them in this case, only mild adaptions of the corresponding proofs in Mesner and Shalizi (2021) are needed. However, we leave such an adaption of the proofs to future work and here do adopt Assumption 5.

For the proof of Theorem 4, we must make the additional assumption 6 about the number of neighbors k_p used in the permutation scheme described in Sec. 2.3.

Assumption 6. $k_p \to \infty$ *and* $\frac{k_p}{n} \to 0$ *as* $n \to \infty$.

B.2. Proof of Lemma 1

We start by using the triangle inequality to get

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}\left[\left|\hat{I}^{0-\infty}(X;Y|Z) - \hat{I}^{MS}(X';Y'|Z')\right|^q\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_i^{0-\infty}(X;Y|Z) - \frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_i^{MS}(X';Y'|Z')\right|^q\right] \\
&\leq \lim_{n\to\infty} \frac{1}{n^q}\mathbb{E}\left[\left(\sum_{i=1}^{n}\left|\hat{\xi}_i^{0-\infty}(X;Y|Z) - \hat{\xi}_i^{MS}(X';Y'|Z')\right|\right)^q\right] \\
&= \lim_{n\to\infty} \frac{1}{n^q}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_q=1}^{n}\mathbb{E}\left[\prod_{\alpha=1}^{q}\left|\hat{\xi}_{i_\alpha}^{0-\infty}(X;Y|Z) - \hat{\xi}_{i_\alpha}^{MS}(X';Y'|Z')\right|\right].
\end{aligned} \tag{26}$$

Next, let $N_n(C_i)$ be the number of points other than the i-th point that are in the cluster C_i of the i-th point. This random variable $N_n(C_i)$ follows the distribution $Bin(n-1, p_i)$, where $p_i = \mathbb{P}(C_i)$ is the probability that an arbitrary point belongs to the cluster C_i. We have $p_i > 0$ because else the i-th point would not have been in the cluster C_i. The Chernoff bound for the lower tail of the Binomial distribution then gives

$$\mathbb{P}(N_n(C_i) \leq k-1) \leq e^{-\frac{(n-1)\cdot p_i}{2}\cdot\left[1-\frac{k-1}{(n-1)\cdot p_i}\right]^2}. \tag{27}$$

Now let A_{i_α} be the event that $N_n(C_{i_\alpha}) \leq k-1$ and let $A_{i_1,\ldots,i_q} = \cup_{\alpha=1}^{q} A_{i_\alpha}$. The probability of $A_{i_1,\ldots,i_q}$ is upper bounded according to

$$\begin{aligned}\mathbb{P}(A_{i_1,\ldots,i_q}) &\leq \sum_{\alpha=1}^{q} \mathbb{P}(N_n(C_{i_\alpha}) \leq k-1) \\ &\leq \sum_{\alpha=1}^{q} e^{-\frac{(n-1)\cdot p_{i_\alpha}}{2}\cdot\left[1-\frac{k-1}{(n-1)\cdot p_{i_\alpha}}\right]^2}.\end{aligned} \tag{28}$$

Next, we use the law of total expectation to condition the expectation value on the right-hand-side of the last line of ineq. (26) on the events $A_{i_1,\ldots,i_q}$ and $A^c_{i_1,\ldots,i_q}$, which gives

$$\begin{aligned}&\mathbb{E}\left[\prod_{\alpha=1}^{q}\left|\hat{\xi}^{0-\infty}_{i_\alpha}(X;Y|Z) - \hat{\xi}^{MS}_{i_\alpha}(X';Y'|Z')\right|\right] \\ &= \underbrace{\mathbb{E}\left[\prod_{\alpha=1}^{q}\left|\hat{\xi}^{0-\infty}_{i_\alpha}(X;Y|Z) - \hat{\xi}^{MS}_{i_\alpha}(X';Y'|Z')\right| \,\middle|\, A_{i_1,\ldots,i_q}\right]}_{\geq 0}\cdot \underbrace{\mathbb{P}(A_{i_1,\ldots,i_q})}_{\leq \sum_{\alpha=1}^{q}\exp\left\{-\frac{(n-1)\cdot p_{i_\alpha}}{2}\cdot\left[1-\frac{k-1}{(n-1)\cdot p_{i_\alpha}}\right]^2\right\}} \\ &+ \mathbb{E}\left[\underbrace{\prod_{\alpha=1}^{q}\left|\hat{\xi}^{0-\infty}_{i_\alpha}(X;Y|Z) - \hat{\xi}^{MS}_{i_\alpha}(X';Y'|Z')\right|}_{=0} \,\middle|\, A^c_{i_1,\ldots,i_q}\right]\cdot \underbrace{\mathbb{P}(A^c_{i_1,\ldots,i_q})}_{\leq 1} \\ &\leq \mathbb{E}\left[\prod_{\alpha=1}^{q}\left|\hat{\xi}^{0-\infty}_{i_\alpha}(X;Y|Z) - \hat{\xi}^{MS}_{i_\alpha}(X';Y'|Z')\right| \,\middle|\, A_{i_1,\ldots,i_q}\right]\cdot \sum_{\alpha=1}^{q} e^{-\frac{(n-1)\cdot p_{i_\alpha}}{2}\cdot\left[1-\frac{k-1}{(n-1)\cdot p_{i_\alpha}}\right]^2}\end{aligned} \tag{29}$$

Here, we have used $\hat{\xi}^{0-\infty}_{i_\alpha}(X;Y|Z) = \hat{\xi}^{MS}_{i_\alpha}(X';Y'|Z')$ conditioned on the event $A^c_{i_1,\ldots,i_q}$. This equality holds for the following reason: Given $A^c_{i_1,\ldots,i_q}$, for all $1 \leq \alpha \leq q$ at least k points other than the i_α-point are in the cluster C_{i_α} of i_α. Now fix some α with $1 \leq \alpha \leq q$ and let $v_0, v_1, \ldots, v_m$ be the points in cluster C_{i_α}, ordered such that v_0 is the i_α-th point and that $\|v_a - v_0\|_{L^\infty} \leq \|v_b - v_0\|_{L^\infty}$ if $a < b$. Then, $m \geq k$ and the distance ρ_{i_α} used by the estimate $\hat{\xi}^{0-\infty}_{i_\alpha}(X;Y|Z)$, here denoted as $\rho^{0-\infty}_{i_\alpha}$, equals $\|v_k - v_0\|_{L^\infty} < \infty$. Let $h(w) = \lambda\cdot w + c$ with $\lambda \neq 0$ be the non-constant affine function that transforms the numeric components of XYZ to $X'Y'Z'$, and for all $0 \leq a \leq m$ let v'_a denote the transformed version of v_a. Since the same transformation h is applied to all numeric components of XYZ, we get that $\|v'_a - v'_0\|_{L^\infty} \leq \|v'_b - v'_0\|_{L^\infty}$ if $a < b$. Moreover, since the ranges of numeric components of $X'Y'Z'$ are contained within $(0,1)$, we get that $\|v'_m - v'_0\|_{L^\infty} < 1$. Consequently, for the purpose of $\hat{\xi}^{MS}_{i_\alpha}(X';Y'|Z')$ the k-nearest neighbors of v'_0 are $v'_1, \ldots, v'_k$ and the distance ρ_{i_α} used by the estimate $\hat{\xi}^{MS}_{i_\alpha}(X';Y'|Z')$, here denoted as $\rho^{MS}_{i_\alpha}$, equals $\|v'_k - v_0\|_{L^\infty} = \lambda\cdot\rho^{0-\infty}_{i_\alpha} < 1$. We thus see that also $\hat{\xi}^{MS}_{i_\alpha}(X';Y'|Z')$ uses only points within the cluster C_{i_α} of the i_α-th point. Let W be a wildcard for XYZ, XZ, YZ and Z and consider the count $\tilde{k}^{0-\infty}_{W,i_\alpha}$ used by the estimate $\hat{\xi}^{0-\infty}_{i_\alpha}(X;Y|Z)$ as defined in eq. (8) in the main text. By definition, see eqs. (7) in the main text, this count is the number of points other than the i_α-th point v_0 that in W-space have a $0-\infty$ "distance" of at most $\rho^{0-\infty}_{i_\alpha}$ to v_0. Since $\rho^{MS}_{i_\alpha} = \lambda\cdot\rho^{0-\infty}_{i_\alpha} < 1$ and the L^∞-distance in W'-space (where, for example, $W' = X'Y'Z'$ if $W = XYZ$) is λ times the L^∞-distance in W-space, the corresponding

count $\tilde{k}_{W',i_\alpha}$ used by $\hat{\xi}^{MS}_{i_\alpha}(X';Y'|Z')$, see eq. (2) in the main text, equals $\tilde{k}^{0-\infty}_{W,i_\alpha}$. From the equality $\tilde{k}^{0-\infty}_{W,i_\alpha} = \tilde{k}_{W',i_\alpha}$ we conclude that, as claimed, $\hat{\xi}^{0-\infty}_{i_\alpha}(X;Y|Z) = \hat{\xi}^{MS}_{i_\alpha}(X';Y'|Z')$ conditioned on the event $A^c_{i_1,\ldots,i_q}$.

To upper bound the remaining conditional expectation, we use the triangle inequality and the fact that $\ln(a) \geq \psi(a) \geq 0$ for $a \geq 0$ to get

$$\begin{aligned} \left|\hat{\xi}^{0-\infty}_{i}(X;Y|Z)\right| &\leq 2\ln(n) \text{ and} \\ \left|\hat{\xi}^{MS}_{i}(X';Y'|Z')\right| &\leq 2\ln(n)\,, \end{aligned} \tag{30}$$

which implies

$$\left|\hat{\xi}^{0-\infty}_{i}(X;Y|Z) - \hat{\xi}^{MS}_{i}(X';Y'|Z')\right| \leq 4\ln(n)\,. \tag{31}$$

We thus find

$$\mathbb{E}\left[\prod_{\alpha=1}^{q}\left|\hat{\xi}^{0-\infty}_{i_\alpha}(X;Y|Z) - \hat{\xi}^{MS}_{i_\alpha}(X';Y'|Z')\right| \,\middle|\, A_{i_1,\ldots,i_q}\right] \leq 4^q \ln(n)^q\,. \tag{32}$$

By combining the above results, we get

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{E}\left[\left|\hat{I}^{0-\infty}(X;Y|Z) - \hat{I}^{MS}(X';Y'|Z')\right|^q\right] \leq \lim_{n\to\infty} \frac{4^q\cdot\ln(n)^q}{n^q}\cdot \\ &\left(\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_q=1}^{n}\right)\sum_{\alpha=1}^{q} e^{-\frac{(n-1)\cdot p_{i_\alpha}}{2}\cdot\left[1-\frac{k-1}{(n-1)\cdot p_{i_\alpha}}\right]^2}\,. \end{aligned} \tag{33}$$

Let $C^{(1)}, \ldots, C^{(m)}$ be the list of all clusters with positive probability, where $m < \infty$ because according to Assumption 4 there are at most finitely many clusters. Then $p_{min} = \min_{\gamma\in[[1,m]]}\mathbb{P}(C^{(\gamma)})$, which is independent of k and n, exists and $p_{min} > 0$. Noting that $p_{i_\alpha} = \mathbb{P}(C_{i_\alpha}) > 0$ for all i_α because else the i_α-th point would not have been in the cluster C_{i_α}, we get that $1 \geq p_{i_\alpha} \geq p_{min} > 0$ for all i_α and hence

$$\frac{k-1}{(n-1)\cdot p_{i_\alpha}} \leq \frac{k-1}{(n-1)\cdot p_{min}} \tag{34}$$

for all i_α. Since $\frac{k}{n} \to 0$ and $k \to \infty$ in the k-NN limit and since p_{i_α} is independent of k and n, we find that $\frac{k-1}{(n-1)\cdot p_{i_\alpha}} \to 0^+$ for all i_α and $\frac{k-1}{(n-1)\cdot p_{min}} \to 0^+$. Thus, there is a positive integer n_0 such that for all $n \geq n_0$ and for all i_α the bound

$$\left[1 - \frac{k-1}{(n-1)\cdot p_{i_\alpha}}\right]^2 \geq \left[1 - \frac{k-1}{(n-1)\cdot p_{min}}\right]^2 \tag{35}$$

holds. We then get

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{E}\left[\left|\hat{I}^{0-\infty}(X;Y|Z) - \hat{I}^{MS}(X';Y'|Z')\right|^q\right] \leq \lim_{n\to\infty} \frac{4^q\cdot\ln(n)^q}{n^q}\cdot \\ &\left(\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_q=1}^{n}\right)\sum_{\alpha=1}^{q} e^{-\frac{(n-1)\cdot p_{min}}{2}\cdot\left[1-\frac{k-1}{(n-1)\cdot p_{min}}\right]^2}\,. \end{aligned} \tag{36}$$

We can now pull the sum of exponentials (that is, the sum over α) out of the product-sum (that is, the sums over $i_1, \ldots, i_q$) and get

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[\left|\hat{I}^{0-\infty}(X;Y|Z) - \hat{I}^{MS}(X';Y'|Z')\right|^q\right] &\leq \lim_{n\to\infty} \frac{4^q \cdot \ln(n)^q}{n^q} \cdot \left(\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_q=1}^{n}\right) \\
&\quad \sum_{\alpha=1}^{q} e^{-\frac{(n-1)\cdot p_{min}}{2}\cdot\left[1-\frac{k-1}{(n-1)\cdot p_{min}}\right]^2} \\
&= \lim_{n\to\infty} \frac{4^q \cdot \ln(n)^q}{n^q} \cdot \underbrace{\sum_{\alpha=1}^{q} e^{-\frac{(n-1)\cdot p_{min}}{2}\cdot\left[1-\frac{k-1}{(n-1)\cdot p_{min}}\right]^2}}_{=q\cdot\exp\left\{-\frac{(n-1)\cdot p_{min}}{2}\cdot\left[1-\frac{k-1}{(n-1)\cdot p_{min}}\right]^2\right\}} \underbrace{\left(\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_q=1}^{n}\right)}_{n^q} \\
&= \lim_{n\to\infty} 4^q \cdot \ln(n)^q \cdot q \cdot e^{-\frac{(n-1)\cdot p_{min}}{2}\cdot\left[1-\frac{k-1}{(n-1)\cdot p_{min}}\right]^2} \\
&= 0
\end{aligned} \tag{37}$$

The last equality follows because the argument of the exponential goes to $-\infty$ in the k-NN limit. □

B.3. Proofs of Theorems 2 and 3

Proof of Theorem 2. The sequence $X_n = \hat{I}^{0-\infty}(X;Y|Z) - \hat{I}^{MS}(X';Y'|Z')$ converges to $X = 0$ in L^1 according to Lemma 1 with $q = 1$, and the sequence $Y_n = \hat{I}^{MS}(X';Y'|Z') - I(X';Y'|Z')$ converges to $Y = 0$ in L^1 according to the proof of Theorem 3.1 in (Mesner and Shalizi, 2021) [5]. Therefore, the sequence $Z_n = X_n + Y_n = \hat{I}^{0-\infty}(X;Y|Z) - I(X';Y'|Z')$ converges to $X + Y = 0$ in L^1. We conclude the proof by noting the equality $I(X';Y'|Z') = I(X;Y|Z)$, which follows because CMI is invariant under componentwise non-constant affine transformations. □

Proof of Theorem 3. The sequence $X_n = \hat{I}^{0-\infty}(X;Y|Z) - \hat{I}^{MS}(X';Y'|Z')$ converges to $X = 0$ in L^2 according to Lemma 1 with $q = 2$, and the sequence $Y_n = \hat{I}^{MS}(X';Y'|Z') - I(X';Y'|Z')$ converges to $Y = 0$ in L^2 according to Theorem 3.2 in Mesner and Shalizi (2021) in combination with Theorem 3.1 in Mesner and Shalizi (2021). Therefore, the sequence $Z_n = X_n + Y_n = \hat{I}^{0-\infty}(X;Y|Z) - I(X';Y'|Z')$ converges to $X + Y = 0$ in L^2. We conclude the proof by noting the equality $I(X';Y'|Z') = I(X;Y|Z)$, which follows because CMI is invariant under componentwise non-constant affine transformations. □

B.4. Type I and II error

We follow the same approach as Huegle et al. (2023) to prove that our estimator can control type I and II error.

5. That theorem itself only claims asymptotic unbiasedness, which is strictly weaker than L^1-convergence, but the proof actually shows L^1-convergence.

B.4.1. PROOF OF THEOREM 4 (TYPE I ERROR RATE)

We start by bounding the type I error rate as in Huegle et al. (2023), using the total variation distance of $P_{X|Z}$ and the simulated $\tilde{P}_{X|Z}$ using the permutation scheme. Given Asm. 1, $P_{XYZ} \ll P_{X|Z} \times P_{Y|Z} \times P_Z$, such that, under the null hypothesis H_0: $P_{XYZ} \equiv P_{X|Z} \times P_{Y|Z} \times P_Z$. Define the simulated product measure $\tilde{P}_{XYZ} = \tilde{P}_{X|Z} \times \tilde{P}_{Y|Z} \times \tilde{P}_Z$. As in Huegle et al. (2023), we define the finite sample distribution P^n over a sequence $\mathbf{w} = (\mathbf{x}, \mathbf{y}, \mathbf{z})$ of n observed samples, with $\mathbf{x} = (x_1, ..., x_n)$, analogously for $\mathbf{y}$ and $\mathbf{z}$. More specifically, we will work with the conditional distribution $P^n_{X|Z}$.

As in Algorithm 1, we define the set of B permutations $\Pi = \{\pi_b | b = 1, ..., B\}$. The finite sample distribution $\tilde{P}^n_{X|Z,b}$ is the conditional distribution of X given Z after the samples of X have been permuted according to π_b. Finding Π requires an estimate of $P(X|Z)$. We find this estimate using the k-NN approach as described in Sec. 2.3 and also used in Huegle et al. (2023); Zan et al. (2022). However, we use the $0-\infty$ distance to compare samples instead of the L^∞ norm, since we are in the mixed case, and thus only allows neighbors which have the same values for the discrete components and a distance smaller than the ρ_p given by the k_p-NN. After the permutation, for the i-th sample point $w_i = (x_i, y_i, z_i)$, the new value $x_{\pi_b(i)}$ has been assigned to x_i. We denote the assignment after the permutation for all n points of the sample as $\mathbf{x}^{\pi_b} = (x_{\pi_b(i)})^n_{i=1}$, with $(x_{\pi_b(i)})^n_{i=1} \sim \tilde{P}^n_{X|Z=z_i}$. The tuples $(\tilde{\mathbf{x}}, \mathbf{y}, \mathbf{z}), ..., (\tilde{\mathbf{x}}^{\pi_B}, \mathbf{y}, \mathbf{z})$ are i.i.d sampled and thus are exchangeable. As in Huegle et al. (2023); Berrett et al. (2020), we define the rejection region

$$A_\alpha = \{(\mathbf{x}, \mathbf{y}, \mathbf{z}), (\mathbf{x}^{\pi_1}, \mathbf{y}, \mathbf{z}), ..., (\mathbf{x}^{\pi_B}, \mathbf{y}, \mathbf{z}) : p \leq \alpha\} \tag{38}$$

with $p = \frac{\sum_{\pi_j \in \Pi} \mathbf{1}\{\hat{I}^{0-\infty,n}_{\pi_j} \geq I^{0-\infty,n}\}+1}{B+1}$, where $\hat{I}^{0-\infty,n}_{\pi_j}$ is the CMI estimated using our $\text{MS}_{0-\infty}$ estimator on the sample set of n points to which the permutation π_j has been applied, and $I^{0-\infty,n}$ is the value of the CMI estimation using our $\text{MS}_{0-\infty}$ estimator on the original samples. By definition of the rejection region A_α, it holds that

$$\begin{aligned}\mathbb{E}_{P_{XYZ}}[CIT_{0-\infty,n}] &= \mathbb{P}_{P_{XYZ}}((\mathbf{x}, \mathbf{y}, \mathbf{z}), (\mathbf{x}^{\pi_1}, \mathbf{y}, \mathbf{z}), ..., (\mathbf{x}^{\pi_B}, \mathbf{y}, \mathbf{z}) \in A_\alpha) \\ &\leq \mathbb{P}_{P_{XYZ}}((\tilde{\mathbf{x}}, \mathbf{y}, \mathbf{z}), (\tilde{\mathbf{x}}^{\pi_1}, \mathbf{y}, \mathbf{z}), ..., (\tilde{\mathbf{x}}^{\pi_B}, \mathbf{y}, \mathbf{z}) \in A_\alpha) + \mathcal{D}_{TV}(P^n_{XYZ}, \tilde{P}^n_{XYZ})\end{aligned}$$

As shown in Huegle et al. (2023), it holds that

$$\mathbb{E}_{P_{XYZ}}[CIT_{0-\infty,n}] \leq \alpha + \mathcal{D}_{TV}(P^n_{XYZ}, \tilde{P}^n_{XYZ}) \tag{39}$$

where the total variation distance $\mathcal{D}_{TV}(P^n_{XYZ}, \tilde{P}^n_{XYZ}) = \sup_{A \in \mathcal{B}} |P^n_{XYZ}(A) - \tilde{P}^n_{XYZ}(A)|$. We note that for any (U, V) and (U', V'), if $(V|U = u) \equiv (V'|U' = u)$ for any u, then $\mathcal{D}_{TV}((U, V), (U', V')) = \mathcal{D}_{TV}(U, U')$. Given this fact, and since $\tilde{P}_{XYZ} \equiv \tilde{P}_{X|Z} \times P_{Y|Z} \times P_Z$, then we have that

$$\mathcal{D}_{TV}(P^n_{XYZ}, \tilde{P}^n_{XYZ}) = \mathcal{D}_{TV}(P^n_{X|Z}, \tilde{P}^n_{X|Z}) \tag{40}$$

Now we must show that $\mathcal{D}_{TV}(P^n_{X|Z}, \tilde{P}^n_{X|Z}) \to 0$ under the assumptions we made. First, as in Huegle et al. (2023), we relate the total variation distance to the Kullback-Leibler divergence using Pinsker's inequality:

$$(P^n_{X|Z}, \tilde{P}^n_{X|Z}) \leq \sqrt{\frac{1}{2}\mathcal{D}_{KL}(P^n_{X|Z}||\tilde{P}^n_{X|Z})} \tag{41}$$

We assume that $P^n_{X|Z} = P^n_{X|Z=z_1} \times P^n_{X|Z=z_2} \times ... \times P^n_{X|Z=z_n}$ and $\tilde{P}^n_{X|Z} = \tilde{P}^n_{X|Z=z_1} \times \tilde{P}^n_{X|Z=z_2} \times ... \times \tilde{P}^n_{X|Z=z_n}$. As $\frac{k_p}{n} \to \infty$ as $n \to \infty$, we can assume that there are enough samples to approximate the KL divergence well enough, and thus the following holds:

$$\mathcal{D}_{KL}(P^n_{X|Z}||\tilde{P}^n_{X|Z}) = \sum_{i=1}^{n} \mathcal{D}_{KL}(P^n_{X|Z=z_i}||\tilde{P}^n_{X|Z=z_i}) \tag{42}$$

Therefore, it is enough to show that, for $Z = z_i$, $\mathcal{D}_{KL}(P^n_{X|Z=z_i}||\tilde{P}^n_{X|Z=z_i}) \to 0$ under the assumptions made and increasing sample size $n \to \infty$. As in Huegle et al. (2023) we define

$$P^n_{X|Z=z_i}(x, z_i, \rho) = P^n_{X|Z=z_i}(\{x' \in \mathcal{X} : ||(x, z_i) - (x', z_i)||_{0-\infty} \leq \rho\}) \tag{43}$$

$$\tilde{P}^n_{X|Z=z_i}(x, z_i, \rho) = \tilde{P}^n_{X|Z=z_i}(\{x' \in \mathcal{X} : ||(x, z_i) - (x', z_i)||_{0-\infty} \leq \rho\}) \tag{44}$$

Let $f(x, z)$ denote the density $f(x, z) = \frac{dP^n_{X|Z}}{d\tilde{P}^n_{X|Z}}$. As also done in Huegle et al. (2023); Mesner and Shalizi (2021), we partition $\mathcal{X} \times \mathcal{Z}$ into three disjoint sets corresponding to the discrete, continuous and mixed-type cases, as follows:

1. $\Omega_1 = \{(x, z_i) \in \mathcal{X} \times \mathcal{Z} : f(x, z_i) = 0\}$
2. $\Omega_2 = \{(x, z_i) \in \mathcal{X} \times \mathcal{Z} : f(x, z_i) > 0, P^n_{X|Z=z_i}(x, z_i, 0) > 0\}$. For $\rho = 0$, $P(x, z_i, 0) > 0$ is the probability mass of one point, and thus this is the discrete case.
3. $\Omega_3 = \{(x, z_i) \in \mathcal{X} \times \mathcal{Z} : f(x, z_i) > 0, P^n_{X|Z=z_i}(x, z_i, 0) = 0\}$. For $\rho = 0$, $P(x, z_i, 0) = 0$ is the probability mass of one point, and thus this is the continuous or mixed-type case.

Then we can write the KL divergence for each $Z = z_i$ as:

$$\begin{aligned}
\mathcal{D}_{KL}(P^n_{X|Z=z_i}||\tilde{P}^n_{X|Z=z_i}) &= \int \log(f(x, z_i) dP^n_{X|Z=z_i}(x, z_i) && (45)\\
&= \int_{\Omega_1} \log(f(x, z_i) dP^n_{X|Z=z_i}(x, z_i) && (46)\\
&+ \int_{\Omega_2} \log(f(x, z_i) dP^n_{X|Z=z_i}(x, z_i) && (47)\\
&+ \int_{\Omega_3} \log(f(x, z_i) dP^n_{X|Z=z_i}(x, z_i) && (48)
\end{aligned}$$

We now proceed to show that, for each case, we obtain $\mathcal{D}_{KL}(P^n_{X|Z=z_i}||\tilde{P}^n_{X|Z=z_i}) \to 0$ for $n \to \infty \; \forall Z = z_i$ using our $0 - \infty$ distance.

Case 1 Let $(x, z_i) \in \Omega_1$ and $\omega(\Omega_1) = \{(x) : (x, z_i) \in \Omega_1\}$ be the projection of the first coordinate. Using the definition of f as the Radon-Nikodym derivative, we have

$$P^n_{X|Z=z_i}(\omega_X(\Omega_1)) = \int_{\omega_X(\Omega_1)} f dP^n_{X|Z=z_i} = \int_{\omega_X(\Omega_1)} 0 dP^n_{X|Z=z_i} = 0. \tag{49}$$

We use the fact that $f(x, z_i)$ is the density with respect to the measure $dP^n_{X|Z=z_i}(x, z_i)$. Therefore, we can rewrite the integral as:

$$\int_{\Omega_1} \log f(x, z_i) f(x, z_i)\, d\mu(x), \tag{50}$$

where μ is the underlying measure. Now, using the limit $\lim_{x\to 0} f(x, z_i) \log f(x, z_i) = 0$, we obtain

$$\int_{\Omega_1} \log f(x, z_i) dP^n_{X|Z=z_i}(x, z_i) = 0. \tag{51}$$

Case 2: Let $(x, z_i) \in \Omega_2$, which is the partition of discrete points. By Lemma E.8 of Mesner and Shalizi (2021),

$$f(x, z_i) = \frac{P^n_{X|Z=z_i}(x, z_i, 0)}{\tilde{P}^n_{X|Z=z_i}(x, z_i, 0)} \tag{52}$$

We must show that $P^n_{X|Z=z_i}(x, z_i, 0) \equiv \tilde{P}^n_{X|Z=z_i}(x, z_i, 0)$. We define ρ_i the distance from z_i to its k_p-nearest neighbors using the $0 - \infty$ metric. Recall that the $0 - \infty$ distance returns 0 is two discrete values are equal, and ∞ otherwise. Therefore, we can distinguish two cases: $\rho_i = 0$, meaning that there are at least k_p points which fall within the ball with radius $\rho_i = 0$ (as we are in the discrete case), or $\rho_i = \infty$, which indicates that there less than k_p points in the ball with radius ρ_i.

If $\rho_i = 0$, then there are at least k_p neighbors. By our definition of the local permutation, we can permute the x values of all neighboring points without changing the bin, i.e., the points with indices $\{k : ||z_k - z_i||_{0-\infty} = 0, k \neq i\}$. Since all $Z = z_i$ values are equal, we can write that $||(x_{\pi_b(i)}, z_i) - (x_i, z_i)||_{0-\infty} = ||x_{\pi_b(i)} - x_i||$. According to the definition of $P^n_{X|Z=z_i}$ and $\tilde{P}^n_{X|Z=z_i}$ these only differ by such permutations, therefore, $P^n_{X|Z=z_i}(x, z_i, 0) \equiv \tilde{P}^n_{X|Z=z_i}(x, z_i, 0)$. Thus, for $n \to \infty$, $f = 1$, and

$$\lim_{n\to\infty} \int_{\Omega_2} \log f(x, z_i) dP^n_{X|Z=z_i} = \lim_{n\to\infty} \int_{\Omega_2} \log 1 dP^n_{X|Z=z_i} = 0 \tag{53}$$

The second case, $\rho_i = \infty$, occurs when there are less than k_p points available. However, under the premise that $n \to \infty$, and under Asm. 4, we can show that this happens with probability close to 0.

The number of points $|z_i| \sim Binomial(n, P^n(Z = z_i))$, and, by the law of large numbers, $|z_i| = m$ as $n \to \infty$. Now we want to show that the probability that $\mathbb{P}(|z_i| < k_p) \to 0$ as $n \to \infty$.

$$\begin{aligned}\mathbb{P}(\rho_i = \infty) =& \mathbb{P}(|z_i| < k_p) = \\ &= \mathbb{P}(Binomial(n, P^n(Z = z_i)) \\ &= \sum_{0}^{k_p - 1} \binom{n}{k} (P^n(Z = z_i))^k (1 - P^n(Z = z_i)))^{n-k}\end{aligned}$$

Let $m = n \cdot P^n(Z = z_i)$ be the expected number of points with $Z = z_i$. For $n \to \infty$, we can approximate the binomial distribution using the normal distribution as $\mathcal{N}(m, \sqrt{m(1 - P^n(Z = z_i))})$. By standardization, we obtain the variable $NZ = \frac{|z_i| - m}{m(1 - P(Z = z_i))} \sim \mathcal{N}(0, 1)$. Then we obtain that the probability that $|z_i| < k_p$ can be written as

$$\begin{aligned}\mathbb{P}(|z_i| < k_p) = \mathbb{P}(\frac{||z_i| - m|}{m(1 - P^n(Z = z_i)} < \frac{k_p - m}{m(1 - P^n(Z = z_i)}) = \\ \mathbb{P}(NZ < \frac{k_p - n \cdot P^n(Z = z_i))}{n \cdot P^n(Z = z_i)(1 - P^n(Z = z_i))})\end{aligned}$$

Taking the limit as $n \to \infty$, we obtain:

$$\begin{aligned}\lim_{n \to \infty} \frac{k_p - n \cdot P^n(Z = z_i)}{\sqrt{n \cdot P^n(Z = z_i) \cdot (1 - P^n(Z = z_i))}} = \\ = \lim_{n \to \infty} \frac{-n(\frac{k_p}{n} + P^n(Z = z_i))}{\sqrt{n}\sqrt{P^n(Z = z_i) \cdot (1 - P^n(Z = z_i))}} = \\ = -\sqrt{n} \cdot \text{const}\end{aligned}$$

Thus, we obtain that

$$\mathbb{P}(\rho_i = \infty) \to \lim_{C \to \infty} \mathbb{P}(NZ < -C) = 0 \quad \text{as} \quad n \to \infty. \tag{54}$$

Case 3: Let $(x, z_i) \in \Omega_3$, i.e., the partition with continuous or mixed-type points. Recall that, for the permutation scheme described in Sec. 2.3, we are only considering neighbors in the Z subspace.

Therefore, we first have to ensure that there are enough samples such that $|\{(x, z_i) : ||z_i - z_j||_{0-\infty} \leq \rho\}| \to k_p$ almost surely. As we have shown in the proof of Lemma 1, as $n \to \infty$, the k-nearest neighbors are the same when using the L_∞ and $0 - \infty$ distance. Therefore, using our $0 - \infty$ distance, we are in the second case of Lemma E.5 in Mesner and Shalizi (2021) and thus for $n \to \infty$ there are enough samples such that $|\{(x', z_i)||(x, z_i) - (x', z_i)||_{0-\infty} \leq \rho\}| \to k_p$ almost surely.

According to Lemma E.7 in Mesner and Shalizi (2021), for $\epsilon > 0$, we can write

$$\lim_{\rho_i \to 0} \mathbb{P}(|\frac{P^n_{X|Z=z_i}(x, z_i, \rho_i)}{\tilde{P}^n_{X|Z=z_i}(x, z_i, \rho_i)} - f(x, z_i)| \leq \epsilon) = 1 \tag{55}$$

Using our $0-\infty$ distance, it holds that $|(z_j^c, z_j^d) - (z_i^c, z_i^d)|_{0-\infty} = |z_j^c - z_i^c|_\infty$ if $z_i^d = z_j^d$, and $|z_j - z_i|_{0-\infty} = \infty$ otherwise.

The case $|z_j - z_i|_{0-\infty} = \infty$ happens whenever there are less available points than k_p in the ball defined by ρ_i. However, as also shown for case 2, the probability $\mathbb{P}(|z_i| < k_p) \to 0$ as $n \to \infty$.

We are therefore only left with the case $|z_j - z_i|_{0-\infty} \neq \infty$, which means that there are exactly k_p points in the radius defined by ρ_i. We now want to show that, by construction of the permutations to be local in Z, for $\mathcal{X}_i := \{x \in \mathcal{X} : ||z_i - z||_{0-\infty} \leq \rho_i\}$, i.e., all neighbors in the Z subspace within radius ρ_i, it holds that $P^n_{X|Z=z_i}(x, z_i, \rho_i) = P^n_{X|Z=z_i}(\mathcal{X}) = \tilde{P}^{n,b}_{X|Z=z_i}(\mathcal{X}) = \tilde{P}^n_{X|Z=z_i}(x, z_i, \rho_i)$, for all permutation indices $b = 1, \dots, B$.

By construction of the permutations to be local in Z, for $\mathcal{X}_{i,Z} := \{x : ||z_i - z||_{0-\infty} \leq \rho_i\}$, we have $\pi_b(\mathcal{X}_{i,Z}) = \mathcal{X}_{i,Z}$. For the $0-\infty$ distance, it holds that for any $x_i, x_j \in \mathcal{X}_{i,Z}$ if $||(x_i, z_i) - (x_j, z_j)||_{0-\infty} \neq \infty$, then $||(x_i) - (x_j)||_{0-\infty} \leq \sigma_i$, with $0 < \sigma_i < \infty$. Since $\pi_b(\mathcal{X}_{i,Z}) = \mathcal{X}_{i,Z}$, for any pair $(x_j, x_{\pi_b(j)})$ of values for X at index j and the index after applying the permutation π, given that $x_j \in \pi_b(\mathcal{X}_{i,Z}) = \mathcal{X}_{i,Z}$, it holds that $||(x_i) - (x_j)||_{0-\infty} \leq \sigma_i$. By the definitions of $P^n_{X|Z=z_i}$ and $\tilde{P}^n_{X|Z=z_i}$ in Eq. 43 and 44, we can conclude that $P^n_{X|Z=z_i} \equiv \tilde{P}^n_{X|Z=z_i}$. Thus, $\int_{\Omega_3} \log(f(x, z_i) dP^n_{X|Z=z_i}(x, z_i) \to 0$ as $n \to \infty$.

B.4.2. PROOF OF THEOREM 5 (TYPE II ERROR RATE)

Here, we follow the proof of Huegle et al. (2023) and use the results from Theorem 2 which prove the consistency of our $\text{MS}_{0-\infty}$ estimator.

Let $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ with $\mathbf{x} = (x_1, ..., x_n)$ be drawn from P_{XYZ}, and $(\mathbf{x}^{\pi_b}, \mathbf{y}, \mathbf{z})$ with $\mathbf{x}^{\pi_b} = (x_{\pi_b(1)}, ..., x_{\pi_b(n)})$ be drawn from $\tilde{P}_{XYZ}$ using the local permutation scheme described in Sec. 2.3. Under $H_1 : X \not\perp\!\!\!\perp Y|Z$, $I(X;Y|Z) > 0$ (see Sec. 1 and Gray (2011)). As Theorem 2 shows, our $\hat{I}^{0-\infty}(X;Y|Z)$ is L_1 consistent,

$$\lim_{n\to\infty} \mathbb{E}[|\hat{I}^{0-\infty}(X;Y|Z) - I(X;Y|Z)|] = 0. \tag{56}$$

For the approximated null distribution obtained using the local permutation approach presented in Sec. 2.3, i.e., obtained by applying $\pi_b \in \Pi$, it also holds that

$$\mathbb{E}[|\hat{I}^{0-\infty}(X_{\pi_b};Y|Z)|] \to 0. \tag{57}$$

Since convergence in L^1 implies convergence in probability, Eq. 56 and 57 converge in probability for the absolute values, thus for the signed values $x = \hat{I}^{0-\infty}(X;Y|Z) - I(X;Y|Z)$ and $y = \hat{I}^{0-\infty}(X_{\pi_b};Y|Z)$. We can thus apply the continuous mapping theorem with these x, y and the function $g(x,y) = |x-y|$, and obtain that $\forall \epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|\hat{I}^{0-\infty}(X;Y|Z) - \hat{I}^{0-\infty}(X_{\pi_b};Y|Z) - I(X;Y|Z)| > \epsilon) = 0. \tag{58}$$

Let $\beta > 0$ arbitrary. Define the falsely accepted region where $CIT_{0-\infty} = 1$, i.e., the set where $p \leq \beta$ although H_1 is true for a nominal value β, as follows

$$A_\beta = \{(\mathbf{x}, \mathbf{y}, \mathbf{z}), (\mathbf{x}^{\pi_1}, \mathbf{y}, \mathbf{z}), ..., (\mathbf{x}^{\pi_B}, \mathbf{y}, \mathbf{z}) : p \leq \beta\} \tag{59}$$

with $p = \frac{\sum_{\pi_j} \mathbf{1}\{\hat{I}^{0-\infty,n}_{\pi_j} \geq I^{0-\infty,n}\}+1}{B+1}$. Then, by the definition of A_β, it holds that

$$\mathbb{E}_{P_{XYZ}}[1 - CIT_{0-\infty}] = 1 - \mathbb{P}_{P_{XYZ}}((\mathbf{x}, \mathbf{y}, \mathbf{z}), (\mathbf{x}^{\pi_1}, \mathbf{y}, \mathbf{z}), ..., (\mathbf{x}^{\pi_B}, \mathbf{y}, \mathbf{z}) \in A_\beta). \tag{60}$$

However, for $(\mathbf{x}, \mathbf{y}, \mathbf{z}), (\mathbf{x}^{\pi_1}, \mathbf{y}, \mathbf{z}), ..., (\mathbf{x}^{\pi_B}, \mathbf{y}, \mathbf{z}) \in A_\beta$, as shown in Eq. 57, $\mathbb{E}[|\hat{I}^{0-\infty}(X_{\pi_b}; Y|Z)|] \to 0$, while $\lim_{n\to\infty} \mathbb{E}[|\hat{I}^{0-\infty}(X; Y|Z) - I(X; Y|Z)|] = 0$, as shown in Eq. 58, and thus we obtain $\forall \epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\{\frac{\sum_{b=1}^{B} \mathbf{1}\{\hat{I}_{\pi_b}^{0-\infty,n} \geq I^{0-\infty,n}\} + 1}{B+1}\} \leq \beta) = \mathbb{P}(\{\frac{1}{B+1} \leq \beta\}). \tag{61}$$

Therefore, for $B > \frac{1}{\beta} + 1$, we obtain

$$\lim_{n\to\infty} \mathbb{E}_{P_{XYZ}}[1 - CIT_{0-\infty}] = 1 - 1 = 0. \tag{62}$$

Appendix C. Further results of the numerical evaluation of the CMI estimators

Experimental setup We keep the experimental setup as described in Sec. 4 of the main paper and generate the remaining three synthetic datasets according to the following models:

"Chain structure" (Mesner and Shalizi (2021)): Here, $X \sim \exp(10)$, $Z = (Z_1, \ldots, Z_d)$ is multivariate with $Z_1 \sim Poisson(X)$ and $Z_i \sim \mathcal{N}(0, 1)$ for $2 \leq i \leq d$, and $Y \sim Bin(Z_1, 0.5)$. The ground truth is $I(X; Y|Z) = 0$.

"Confounder with Gaussian X and Y" (Zan et al. (2022)): This model describes a confounder structure with normally distributed X and Y, where $X \sim \mathcal{N}(Z, 1)$, $Y \sim \mathcal{N}(Z, 1)$, and $Z \sim \mathcal{U}(\{0, \ldots, m\})$. The ground truth is $I(X, Y|Z) = 0$. As in Zan et al. (2022), $m = 9$.

"Confounder with uniform X and Y": This model describes a confounder structure with uniformly distributed X and Y, where $X \sim \mathcal{U}(0, Z)$, $Y \sim \mathcal{U}(Z, Z + 1)$, and $Z \sim \mathcal{U}(\{0, 1\})$. The ground truth is $I(X, Y|Z) = 0$.

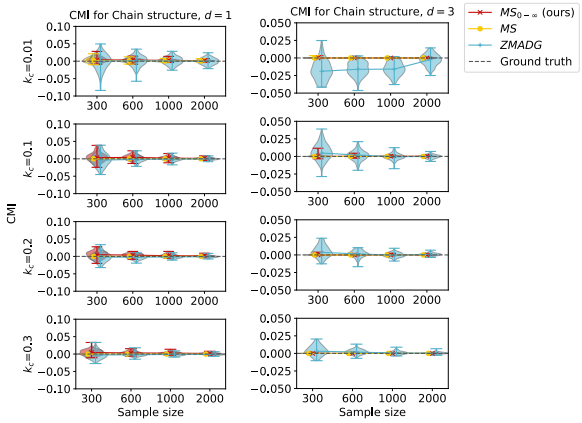

Figure 3: Distribution of the CMI estimates for the "Chain structure" model with $d = 1$ (left) and $d = 3$ (right). Each row shows results for a different k_c value.

Results "Chain structure": For this model (Fig. 3, left) with $d = 1$ the MS, $\text{MS}_{0-\infty}$ and ZMADG estimators perform comparably well. The ZMADG estimator has slightly higher variance

than $\text{MS}_{0-\infty}$ and MS, especially for smaller n. However, for the model with $d = 3$ (Fig. 3, right), ZMADG suffers from significant variance compared to the MS and our estimator. Notably, although this data model violates Assumption 4, our estimator still obtains good results in practice.

"Confounder with Gaussian X and Y": For this model (Fig. 4, left), we observe that $\text{MS}_{0-\infty}$ performs best, while the MS estimator overestimates for $k_c \geq 0.1$. The ZMADG estimator again suffers from higher variance compared to our estimator.

"Confounder with uniform X and Y": For this model (Fig. 4, right), both MS and our estimator perform well. Besides having high variance, ZMADG wrongly finds a strong conditional dependency between X and Y. This observation seems specific to the uniform distribution, as estimates are correct for the same model using normally distributed data.

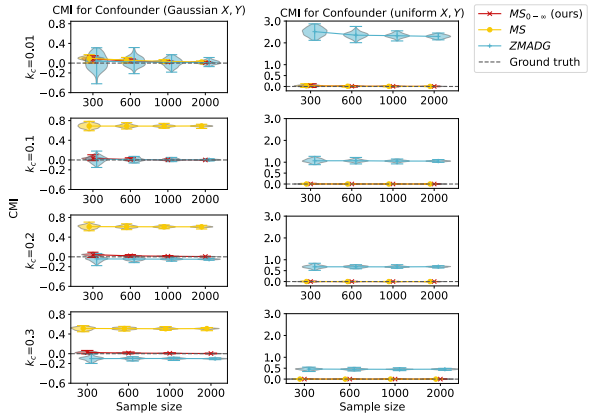

Figure 4: Distribution of the CMI estimates for the "Confounder with Gaussian X and Y" model (left) and "Confounder with uniform X and Y" model (right). Each row shows results for a different k_c value.

C.1. Comparison of the mean and variances of the CMI

As outlined in Sec. 3 of the main paper, our estimator addresses the challenges of the MS and ZMADG estimators. Consequently, we expect that, for most of the models, there will likely be significant differences between the bias of the MS and $\text{MS}_{0-\infty}$ estimators due to our estimator's capacity to reduce the bias towards 0. We also expect significant differences between the variance of the $\text{MS}_{0-\infty}$ estimator and the variance of the ZMADG estimator, as our estimator is not an aggregation of multiple estimators.

Since bias and variance are not scaled metrics, we perform statistical tests to investigate whether there are statistically significant differences among the CMI estimators using the models in Sec. 4 and App. C. Because the MS and ZMADG estimators perform very differently across k_c values, for each model, we identify the k_c value with optimal performance regarding bias and variance across all sample sizes for each model. To compare bias of the MS estimator with the bias of the $\text{MS}_{0-\infty}$ estimator, we compute the mean absolute error (MAE) $MAE_{estim,j}$ for each estimator in $estim \in \{MS, MS_{0-\infty}\}$ and each repetition of the experiment $j \in \{1, \ldots, 100\}$ as the absolute difference between the estimated CMI using estimator $estim$ and the ground truth CMI value

$I(X;Y|Z)$ of the data model:

$$MAE_{estim,j} = |\hat{I}_j^{estim}(X;Y|Z) - I_j(X;Y|Z)|. \tag{63}$$

Then, we compare bias between the estimators using the Wilcoxon-Signed-Rank Test (Wilcoxon, 1992). We perform one-sided tests, as we expect that the MAE of the MS estimator is higher than the MAE of our $\text{MS}_{0-\infty}$ estimator. For each model and each sample size, we formulate the hypothesis as follows:

H_0: *The median of the difference between the* MAE_{MS} *and* $MAE_{MS_{0-\infty}}$ *is negative.* vs.

H_1: *The median of the difference between the* MAE_{MS} *and* $MAE_{MS_{0-\infty}}$ *is non-negative.*

To compare the variance of the estimates $\hat{I}_j^{estim}(X;Y|Z)$ obtained with the $\text{MS}_{0-\infty}$ and ZMADG estimators, we test the following hypothesis of equality of variance using Levene's Test (Levene, 1960) for each model and each sample size:

H_0: *The variance of the* $\text{MS}_{0-\infty}$ *estimator and the variance of the ZMADG estimator are equal.* vs.

H_1: *The variance of the* $\text{MS}_{0-\infty}$ *estimator and the variance of the ZMADG estimator are not equal.*

To account for repeated testing, we apply the Bonferroni correction (Bonferroni, 1935) by dividing the significance level by the number of hypotheses, in our case 48, and thus reject the null hypothesis for p-values under $\frac{0.05}{48} = 0.001$.

The Tables below report the p-values of the statistical tests for each model, measurement, and sample size. We observe that the MAE of the MS estimator is significantly greater than the MAE of our $\text{MS}_{0-\infty}$ estimator for the "Independent Z" with $d = 3$ and the "Confounder with uniform X and Y" models. This aligns with the expectations stated in Sec. 3, namely that our estimator should suffer from less bias towards 0. Furthermore, as expected, for almost all models and all sample sizes, the hypothesis of equality of variances of our estimator and the ZMADG estimator can be rejected. An inspection of the values for the variance shows that, indeed, the variance of ZMADG is higher than that of our estimator, and thus, we can confirm our expectation that our estimator has a lower variance compared to ZMADG.

Table 1: Results of the statistical tests for the "Independent Z" model with $d = 1$ (left table) and $d = 3$ (right table). We select k_c for the individual estimators as follows: $k_{c,MS} = 0.01$, $k_{c,MS_{0-\infty}} = 0.2$ and $k_{c,ZMADG} = 0.1$.

p-values for the "Independent Z" Model with $d = 1$

n	Bias MS vs. $\text{MS}_{0-\infty}$	Var $\text{MS}_{0-\infty}$ vs. ZMADG
300	0.9993	0.0000
600	1.0000	0.0000
1000	1.0000	0.0000
2000	1.0000	0.0000

p-values for the "Independent Z" Model with $d = 3$

n	Bias MS vs. $\text{MS}_{0-\infty}$	Var $\text{MS}_{0-\infty}$ vs. ZMADG
300	0.0146	0.0000
600	0.0000	0.0000
1000	0.0000	0.0000
2000	0.0000	0.0000

Table 2: Results of the statistical tests for the "Chain structure" model with $d = 1$ (left table) and $d = 3$ (right table). We select k_c for the individual estimators as follows: $k_{c,MS} = 0.01$, $k_{c,MS_{0-\infty}} = 0.2$ and $k_{c,ZMADG} = 0.2$.

p-values for the "Chain structure" Model with $d = 1$

n	Bias MS vs. $MS_{0-\infty}$	Var $MS_{0-\infty}$ vs. ZMADG
300	1.0000	0.0000
600	1.0000	0.0000
1000	1.0000	0.0000
2000	1.0000	0.0000

p-values for the "Chain structure" Model with $d = 3$

n	Bias MS vs. $MS_{0-\infty}$	Var $MS_{0-\infty}$ vs. ZMADG
300	0.9996	0.0000
600	1.0000	0.0000
1000	1.0000	0.0000
2000	1.0000	0.0000

Table 3: Left table: Results of the statistical tests for the "Confounder with uniform X and Y" model. We select k_c for the individual estimators as follows: $k_{c,MS} = 0.1$, $k_{c,MS_{0-\infty}} = 0.1$ and $k_{c,ZMADG} = 0.3$. Right table: Results of the statistical tests for the "Confounder with Gaussian X and Y" model. We select k_c for the individual estimators as follows: $k_{c,MS} = 0.01$, $k_{c,MS_{0-\infty}} = 0.3$ and $k_{c,ZMADG} = 0.1$.

p-values for the "Confounder with uniform X, Y" Model

n	Mean MS vs. $MS_{0-\infty}$	Var $MS_{0-\infty}$ vs. ZMADG
300	0.0000	0.0000
600	0.0000	0.0000
1000	0.0000	0.0000
2000	0.0000	0.0000

p-values for the "Confounder with Gaussian X, Y" Model

n	Bias MS vs. $MS_{0-\infty}$	Var $MS_{0-\infty}$ vs. ZMADG
300	0.5000	0.0000
600	0.5000	0.0000
1000	0.5000	0.0000
2000	0.5000	0.0000

Appendix D. Results of the numerical evaluation of the estimators for the mixture-type variable case

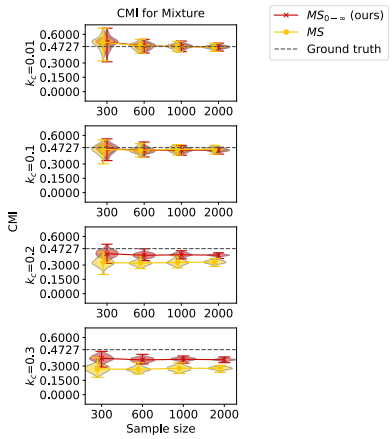

Figure 5: Distribution of CMI estimates for the "Mixture" model. Each row shows the results for different k_c.

As mentioned in the main paper, our study did not focus on the mixture-type variable case. However, the $MS_{0-\infty}$ estimator can be used with mixture-type variables. We present a preliminary evaluation of the CMI estimation for this use case. We do not evaluate the ZMADG estimator, since this method was not designed for use with mixture-type variables. Keeping the experimental setup as for the previous experiments, we evaluate the MS and $MS_{0-\infty}$ estimators on data generated from a model inspired by Mesner and Shalizi (2021), defined as follows:

"Mixture" (adapted from Mesner and Shalizi (2021)): Here, Z is discrete with $Z \sim Bin(1, p = 0.3)$. With probability $1-p$, X and Y are drawn from a multivariate Gaussian with correlation coefficient of 0.6, and with probability p, $X \sim \mathcal{U}(\{0, ..., 4\})$ and $Y \sim \mathcal{U}([X, X + 2])$. The ground truth is $I(X; Y|Z) = -(1-p) \cdot \ln(1 - 0.36) \cdot 0.5 + p \cdot (\ln 5 - \frac{4}{5} \cdot \ln 2) = 0.472$.

Results As Fig. 5 shows, our approach performs best across all k_c values. Both MS and $MS_{0-\infty}$ suffer from bias towards 0 for $k_c \geq 0.2$, however, our estimator is considerably less affected than the MS estimator.

Appendix E. Choice of heuristic for k

In contrast to the theoretical setting where we assume infinite samples are available, we only have access to a finite number of samples in practice. Thus, for large enough k with fixed n, it is probable that some of the k-nearest neighbors in the dataset are at ∞ distance, i.e., originate from a different cluster. Because our estimator does not allow neighbors from different clusters, a heuristic was necessary for setting an adaptive k such that only neighbors from the same cluster are considered. We defined and tested three different heuristics. Here, we describe the other two heuristics besides the "local" heuristic described in Sec. 3. We also motivate our choice for the "local" heuristic, which we based on an empirical comparison of the $\text{MS}_{0-\infty}$ estimator's performance using the different heuristics.

E.1. Two alternative heuristics

Additional to the "local" heuristic presented in Sec. 3, we also developed and investigated two additional heuristics for setting k: the "global" and "cluster-size" heuristics, which we present below.

E.1.1. "GLOBAL" HEURISTIC

The "global" heuristic defines a "global" k as a fraction of the number of samples n. If the distance to the k-nearest neighbor is ∞, then, instead of enforcing k nearest neighbours for all sample points, we allow for the following adaptiveness: If the k-th NN of w_i is at distance ∞ from w_i (that is, if $n_{cl}^i \leq k$ with n_{cl}^i the number of points in the cluster of w_i), then for this i we replace k by $k_i = \lfloor k_c \cdot n_{cl}^i \rfloor$. Explicitly: For all i let $k_i^{0-\infty} = k$ if $k + 1 \leq n_{cl}^i$ and $k_i^{0-\infty} = \lfloor k_c \cdot n_{cl}^i \rfloor$ else. Thus, in effect, all considered nearest neighbours of w_i come from the same cluster as w_i.

E.1.2. "CLUSTER-SIZE" HEURISTIC

This heuristic still uses a "global" k as a fraction of the number of samples n. However, if the distance to the k-th nearest neighbor is ∞, the "cluster-size" heuristic deals with this case by simply setting $k_i = n_{cl}^i$, where n_{cl}^i is the number of samples in the cluster of point i, defined as previously described.

E.2. Numerical evaluation of the heuristics

Experimental setup We run numerical experiments to compare the bias and variance of our $\text{MS}_{0-\infty}$ estimator using the three different heuristics. Alongside, we also compare with the MS and ZMADG estimators, using the same rules for setting k as described in Sec. 4. We keep the same experimental setting as in the experiments described in Sec. 4 and App. C and evaluate results on the "Independent Z" and "Chain structure" models.

Results The violin plots in Figure 6 show the results of the CMI estimation using the "local", "global" and "cluster-size" heuristics presented above. We observe that the "cluster-size" heuristic suffers from bias towards zero. This is expected because, when $k = n_{cl}^i$, the distance to the k-th nearest neighbor is equal to the distance from point i to the farthest point in its respective cluster. Thus, for the subspaces XY, XZ and Z, the number of counted neighbors is equal to n_{cl}^i with high probability, which results in a local estimate equal to or close to 0. The "global" heuristic has the highest variance across the different heuristics, yet still has lower variance compared to the ZMADG estimator. The "global" approach also slightly suffers from bias for higher dimesionality, for example,

for the "Chain structure" model with $d = 3$. The best bias-variance trade-off is obtained using the "local" heuristic, with a slightly higher bias compared to the "global" heuristic, but lower variance. This motivates our choice to use this heuristic for the experiments in the main paper.

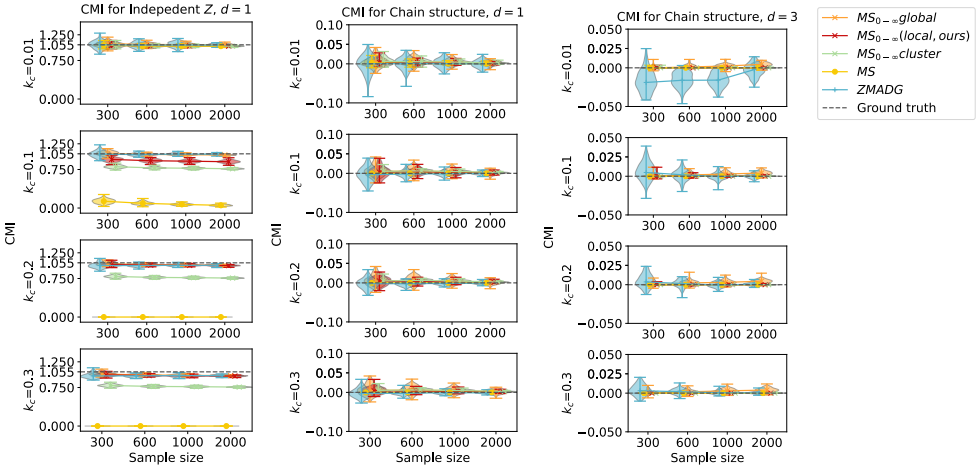

Figure 6: CMI estimation results for the "Independent Z" model with $d = 1$ and "Chain structure" model with $d = 1$ and $d = 3$ using the three different heuristics for setting the hyperparameter k: the "local" heuristic (used in the main paper, $\text{MS}_{0-\infty}$ local), the "global" heuristic ($\text{MS}_{0-\infty}$ global), and the "cluster-size" heuristic ($\text{MS}_{0-\infty}$ cluster). The ground truth CMI values are indicated by the dashed line. The rows show results for different k_c values.

Appendix F. Discussion on the choice of k_c

Generally, for k-NN methods, a small k leads to lower bias at the cost of increased variance, while larger k values lead to low variance but increased bias (Kraskov et al., 2004), and we refer to Berrett et al. (2019) for a comprehensive theoretical discussion. We highlight that the bias scales as $O(\frac{k}{n})$, which explains why smaller values decrease the bias, and the variance scales as $\mathcal{O}(\frac{1}{kn})$, which explains why larger values decrease the variance. Additionally, there is a connection to the local constancy assumption: as k increases, this assumption is more likely to be violated, leading to higher bias. We are not aware of theoretical methods to choose k, and it is unclear how to apply cross-validation or other data-driven approaches for selecting k in practice, where ground-truth is not available. In the machine learning community, a common heuristic for setting k when applying k-NN methods depends on the dataset size, e.g. Chaudhuri and Dasgupta (2014). Our heuristics for setting k, described in detail in Sec. 3 and App. E, depend on the dataset size as well.

As mentioned in Sec. 5.1, we considered $k_c = 0.1$ "optimal" for the MS and ZMADG estimators, as these are the only values that the authors of both estimator methods use for their corresponding numerical experiments. For our estimator, we observe from the CMI estimation and the CIT results that a larger k_c, e.g., $k_c > 0.2$, is beneficial for our method, especially for the case of weak dependence and smaller sample size. This is to be expected due to the "local" heuristic. To exemplify, consider the case where $n = 1000$ and $dim_d = 2$: There are approximate $1000/(2 \cdot 3) \approx 166$ samples per cluster for $n_c = 3$ and approximately $1000/(2 \cdot 4) = 125$ samples for $n_c = 4$. Thus,

with $k_c = 0.1$, we would select at most $k = 13$ samples. Thus, k_c should be high enough such that k is large enough, and this explains why our method performs better for $k_c > 0.2$. As the sample size increases, e.g. $n = 2000$, we observe that a smaller k_c, e.g., $k_c = 0.2$ performs better. Generally, we recommend users to consider both the number of samples and the dimensionality and choose a higher $k_c > 0.1$ when the number of samples in the cluster is small. Alternatively, users could consider the "global" k_c heuristic, especially if the cluster imbalance is high.

Appendix G. Further results of the numerical evaluation of the CIT

Due to space limitations for the main paper, we present further results of the CIT evaluation here. Before presenting the results, we describe how we compute true positive and false positive rates, and the error bars of the CIT plots. Then, we continue with the results with rank preprocessing for the models presented in Sec. 5, and the results for two other synthetic data models. Lastly, we presents results when the dimension of the discrete variables increases.

G.1. Computation of the TPR / FPR

To compute the true positive rates (TPR) and the false positive rates (FPR), we first calculated the proportion of tests that rejected the null hypothesis H_0 given a significance level α. Specifically, for a given set of tests, we define the positive rate pp as:

$$pp = \frac{\#\text{tests with } p \leq \alpha}{n_{rep}} \tag{64}$$

The pp metric, when computed over a set of tests where the alternative hypothesis is true (i.e., when the coupling factor $w > 0$.), reflects the TPR. When evaluated on tests under the null condition (i.e., when the coupling factor $w = 0$), the same metric represents the FPR.

G.2. Computation of the confidence intervals for the TPR/FPR plots

The error bars of the CIT plots represent the 95% confidence interval of the FPR and TPR. The confidence intervals are obtained by modeling the false and true positives as distributed according to the binomial distribution. We describe the computation of the confidence interval for the FPR, and obtain the confidence interval for the TPR analogously.

For a given model and a set of values of the CIT parameters, the probability of obtaining a false positive in the $n_{rep} = 100$ repetitions of our experiments is p_{FP} (ideally, $p_{FP} = \alpha$). Under the assumption that the repetitions are independent, the random variable that describes the number of false positives, FP, is distributed according to the binomial distribution:

$$FP \sim Bin(n_{rep}, p_{FP}). \tag{65}$$

We can estimate p_{FP} as the empirical fraction of false positives that we have obtained in our repetitions: $\hat{p}_{FP} = \frac{\#FP}{n_{rep}} = FPR$. We now wish to obtain a confidence interval for $\hat{p}_{FP}$, i.e., find the lower and upper bounds p_L, p_U of the confidence interval such that $P(p_L(FP) < \hat{p}_{FP} < p_U(FP)) = 1 - \alpha$. Since $FP \sim Bin(n_{rep}, p_{FP})$, we obtain p_L and p_U by numerically solving the following two equations:

$$1 - CDF(\hat{p}_{FP}, n_{rep}, p_L) = \frac{1 - \alpha}{2} \tag{66}$$

$$CDF(\hat{p}_{FP}, n_{rep}, p_U) = \frac{1-\alpha}{2} \tag{67}$$

Here, CDF is the cumulative distribution function of the binomial distribution.

G.3. Results with rank preprocessing for the models presented in Sec. 5

We now present the results for the models presented in Sec. 5 using rank preprocessing for the continuous variables. For the "Confounder" model with $dim_c = 1$ and $dim_d = 1$ and sample size $n = 1000$, the CITs with rank preprocessing (Fig. 7, left) behave similarly to the CITs with standardization. Considering their optimal k_c values, the CIT using our estimator performs similarly to the CIT using the MS estimator. For the "Cluster-dependent confounder" model (Fig. 7, right) the CITs with rank preprocessing also behave similarly to the CITs using standardization, but there is a slightly more significant performance gap between MS and our $\text{MS}_{0-\infty}$, with our method showcasing superior performance, especially for $n_c = 2$ and $n_c = 3$.

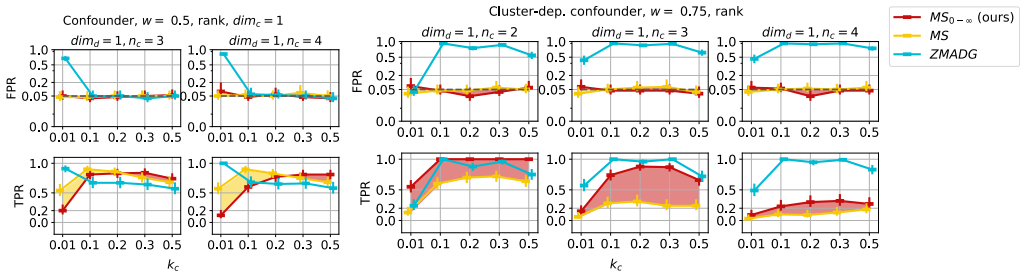

Figure 7: False positive rate (FPR, ideally under 0.05) and true positive rate (TPR, higher is better, with 1 best) for the "Confounder" model where Z has $dim_c = 1, dim_d = 1$ (left) and the "Cluster-dependent confounder" model where Z has $dim_d = 1$ (right). The "Confounder" model has coupling factor in the dependent case $w = 0.5$. The coupling factor in the dependent case for the "Cluster-dependent confounder" model is $w = 0.75$. All models have rank preprocessing for the continuous variables and sample size $n = 1000$.

G.4. Distribution of CMI values for the "Confounder" model

To better understand the outcome of the CIT evaluation, we inspected the distribution of the estimated CMI values for each estimator individually. We present here plots of the true null and the permuted CMI distributions for the different k_c values for the "Confounder" model. For clarity reasons, we refrain from presenting all plots here.

The true null and the permuted CMI distribution plots allow us to investigate whether the FPR/TPR reflects the desired behavior of the tests: The true null and the permuted distributions should have CMI values distributed around 0 and should be similar to each other. In contrast, in the dependent case, estimated CMI values should be larger than 0, and their distribution should have minimal overlap if the null hypothesis does not hold. A closer look at the distributions of the CMI values in Fig. 8 for $n_c = 3$ and Fig. 9 for $n_c = 4$, reveals that the MS and $\text{MS}_{0-\infty}$ estimators behave similarly for $k_c > 0.1$, while ZMADG suffers from slight negative bias.

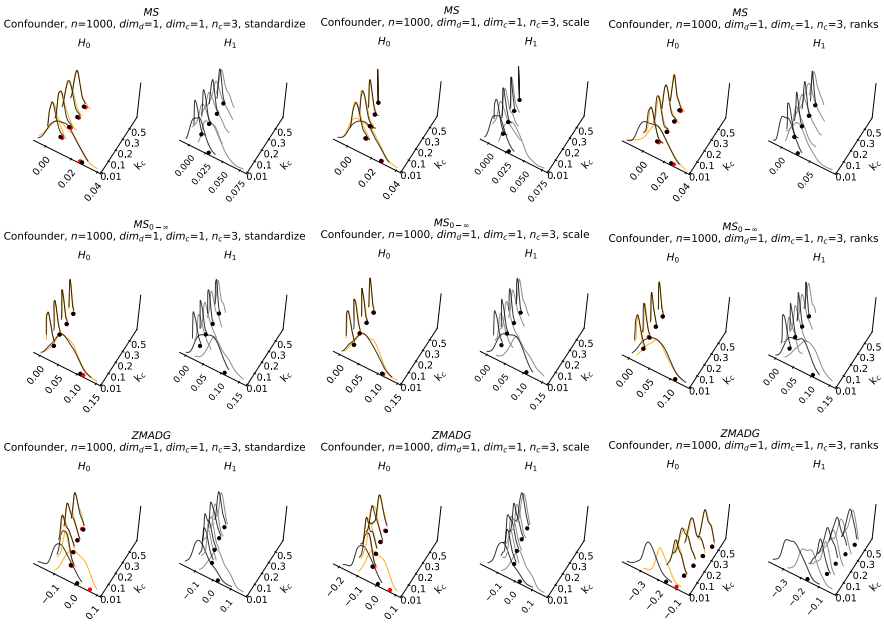

Figure 8: The true null and permuted distributions of the CMI estimated with the MS, $\text{MS}_{0-\infty}$ and ZMADG estimators for the "Confounder" model with $n = 1000$ where Z has $dim_c = 1, dim_d = 1$ and $n_c = 3$ with standardization (left column), scaling to $(0, 1)$ (center column) and rank preprocessing (right column). The left figure of each plot pair shows the true null distribution under H_0 as the orange line with the red dot indicating the 95% quantile, and the permuted null distribution is shown as the black line with the black dot indicating the 95% quantile. The right figure of the plot pair shows the true distribution under H_1 as the grey line, and the permuted distributions as the black line with the black dot indicating the 95% quantile.

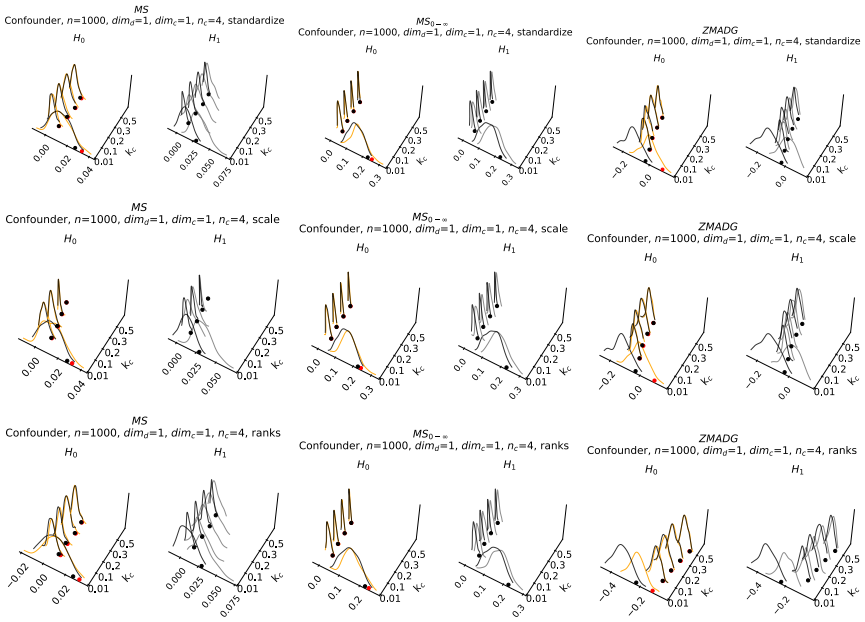

Figure 9: The true null and permuted distributions of CMI estimated with the MS, $\text{MS}_{0-\infty}$ and ZMADG estimators for the "Confounder" model with $n = 1000$ where Z has $dim_c = 1, dim_d = 1$ and $n_c = 4$ with standardization (left column), scaling to $(0, 1)$ (center column) and rank preprocessing (right column). The distributions are depicted as described in Fig. 8.

G.5. Results for the "Independent Z" model

"Independent Z": For this model, Z has discrete components $Z_1, \ldots, Z_{dim_d=m}$ where $Z_j \sim \overline{Bin(n_c - 1, 0.5)}$. Both X and Y are continous univariate and computed as:

$$X = \eta_x + w \cdot \eta_w, \quad Y = \eta_y + w \cdot \eta_w. \tag{68}$$

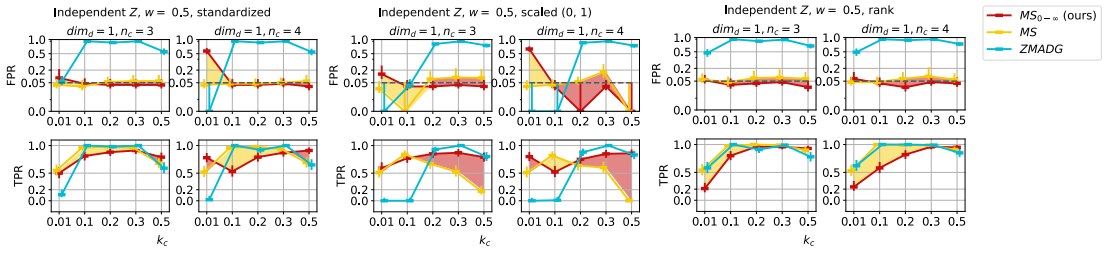

Figure 10: False positive rate (FPR, ideally under 0.05) and true positive rate (TPR, higher is better, with 1 best) for the "Independent Z" model, with standardization (left), scaling to $(0, 1)$ (center) and rank preprocessing (right) for the continuous variables and coupling factor in the dependent case $w = 0.5$.

Here, we present results for sample size $n = 1000$, with coupling factor $w = 0.5$. The number of classes is $n_c \in \{3, 4\}$. We set k as in Sec. 4 using $k_c \in \{0.01, 0.1, 0.2, 0.3, 0.5\}$. We vary $dim_d \in \{1, 2\}$. As for the experiments in Sec. 4, we generate p-values with 300 permuted surrogates using $k_{perm} = 5$ and repeat each experiment 100 times. As we can observe in Fig. 10, ZMADG has either low TPR or high FPR. MS and $\text{MS}_{0-\infty}$ perform better for both preprocessing types, occasionally having higher FPR. The scaling-related issues of MS persist for this model with scaling to $(0, 1)$ and $k_c > 0.1$. $\text{MS}_{0-\infty}$ performs well, except for $k_c = 0.01$, where FPR is not controlled. The FPR of $\text{MS}_{0-\infty}$ drops to 0 for $k_c = 0.2$ and $k_c = 0.5$, yet the error bars and the high TPR values indicate that these drops do not stem from bias towards 0.

G.6. Results for the "Chain" model

"Chain": X and Y are continuous, and Z is discrete, and all variables are univariate. The model is defined as follows:

$$\begin{aligned} X &= \eta_x + w \cdot \eta_w, \\ Z &= [\sigma_\sim(\beta_x \cdot X, (n_c - 1)) + \eta_z] \mod (n_c - 1), \\ Y &= \beta_y \cdot l^{-1}(Z) + \eta_y + w \cdot \eta_w \end{aligned} \tag{69}$$

Here, $\sigma_\sim(\beta_x \cdot X, n_c - 1)$ denotes sampling from the multinomial distribution with $n_c - 1$ categories where the a-th category (starting the count at 0) has probability $\frac{e^{a \cdot x}}{\sum_{a'=0}^{n_c-2} e^{a' \cdot x}}$. The noise η_z follows $\eta_z \sim Bin(2, 0.7)$.

We present results for sample size $n = 1000$, with coupling factor $w = 0.5$. As for the "Independent Z" model, the number of classes is $n_c \in \{3, 4\}$. We set k as in Sec. 4 using $k_c \in \{0.01, 0.1, 0.2, 0.3, 0.5\}$. We generate p-values with 300 permuted surrogates using $k_{perm} = 5$ and repeat each experiment 100 times. For the "Chain" model (Fig. 11), ZMADG consistently suffers from high FPR. In contrast, MS and $\text{MS}_{0-\infty}$ demonstrate good performance across varying k_c values and dimensionalities regarding TPR. The MS and $\text{MS}_{0-\infty}$ CITs perform similarly with standardization and rank preprocessing. In cases where $n_c = 4$, our approach has slightly lower TPR. Nonetheless, both CITs control FPR effectively, except when $k_c = 0.01$. The scaling-related problems of MS lead to performance issues when variables are scaled to $(0, 1)$ and $k_c > 0.1$: an increase in k_c leads to elevated FPR and decreased TPR. Our approach demonstrates robust performance across varying k_c values.

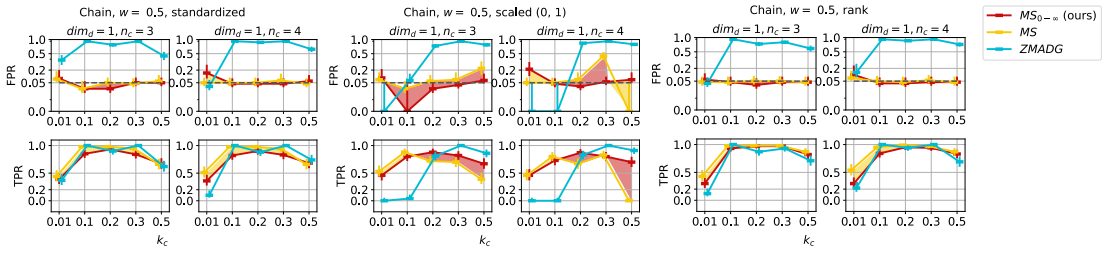

Figure 11: False positive rate (FPR, ideally under 0.05) and true positive rate (TPR, higher is better, with 1 best) for the "Chain" model, with standardization (left), scaling to $(0, 1)$ (center) and rank preprocessing (right) for the continuous variables and coupling factor in the dependent case $w = 0.5$.

G.7. Results with increased dimensionality of the discrete variable

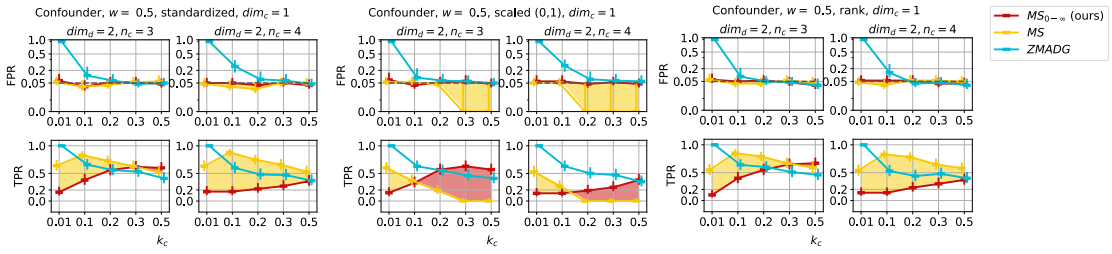

Figure 12: False positive rate (FPR, ideally under 0.05) and true positive rate (TPR, higher is better, with 1 best) for the "Confounder" model with $n = 1000$ where Z has $dim_c = 1, dim_d = 2$, with standardization (left), scaling to $(0, 1)$ (center) and rank preprocessing (right) for the continuous variables and coupling factor in the dependent case $w = 0.5$.

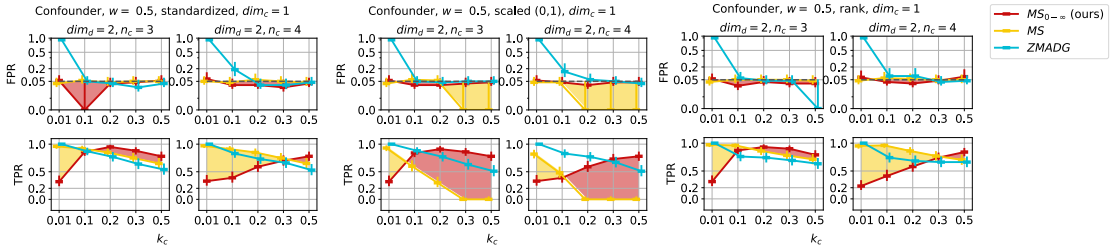

Figure 13: False positive rate (FPR, ideally under 0.05) and true positive rate (TPR, higher is better, with 1 best) for the "Confounder" model with $n = 2000$, where Z has $dim_c = 1, dim_d = 2$, with standardization (left), scaling to $(0, 1)$ (center) and rank preprocessing (right) for the continuous variables and coupling factor in the dependent case $w = 0.5$. We note that for these results, the p-value was computed as $p = \frac{\sum_{\pi_j \in \Pi} \mathbf{1}\{T^{\pi_j}_{CMI,estim} \geq T_{CMI,estim}\}}{B}$ (see App. A.3 for the difference to the other results). This different computation stems from a previous version of the paper and we refrained from redoing the time-consuming computations.

We inspect how the CIT performance changes as the dimensionality of the discrete variable increases to $dim_d = 2$ for the "Confounder" model under the different preprocessing types. The results, presented in Fig. 12, indicate that our approach needs higher k_c, e.g., $k_c \geq 0.3$ for good performance. This is most likely due to the "local" heuristic, where k is computed as the product between k_c and the smallest cluster size in the data (see also App. F). As the dimensionality increases, clusters become smaller, and k decreases as well. The distribution of CMI values indicates that this leads to positive bias. Considering both FPR and TPR for $n_c = 3$ with standardization and rank preprocessing, our approach performs slightly worse than MS but slightly better than ZMADG. For $n_c = 4$, our approach performs worse than MS and slightly worse than ZMADG. The MS estimator performs well in the standardized and rank preprocessing cases despite points from other clusters being considered

neighbors because, for this model, dependence holds in every cluster. When continuous variables are scaled to $(0, 1)$, the scaling-related problems of MS lead to a rapid decline in performance as k_c increases, and our estimator slightly outperforms MS.

However, results for the same model with a larger sample size of $n = 2000$ (see Fig. 13) indicate that the performance of our approach considerably increases given enough samples, and its performance becomes on-par with MS for $n_c = 3$, and only slightly worse for $n_c = 4$. Notably, in the case of a larger sample size, a smaller $k_c = 0.2$ gives the optimal results.

G.8. Results for smaller sample sizes

To inspect the effect of varying sample size n, we evaluate the performance of the CITs when $n \in \{400, 600, 800\}$ and present these in Fig. 14. We observe that the MS estimator performs best with lower sample sizes with all preprocessing types, except for scaling to $(0, 1)$. The lower performance of our $\text{MS}_{0-\infty}$ CIT compared to the MS CIT can be due to the "local" heuristic, which sets k as the fraction k_c of the size of the smallest cluster in the data. Especially as the number of classes for the discrete dimensions increases, this can lead to setting a smaller k even for clusters where more data is available.

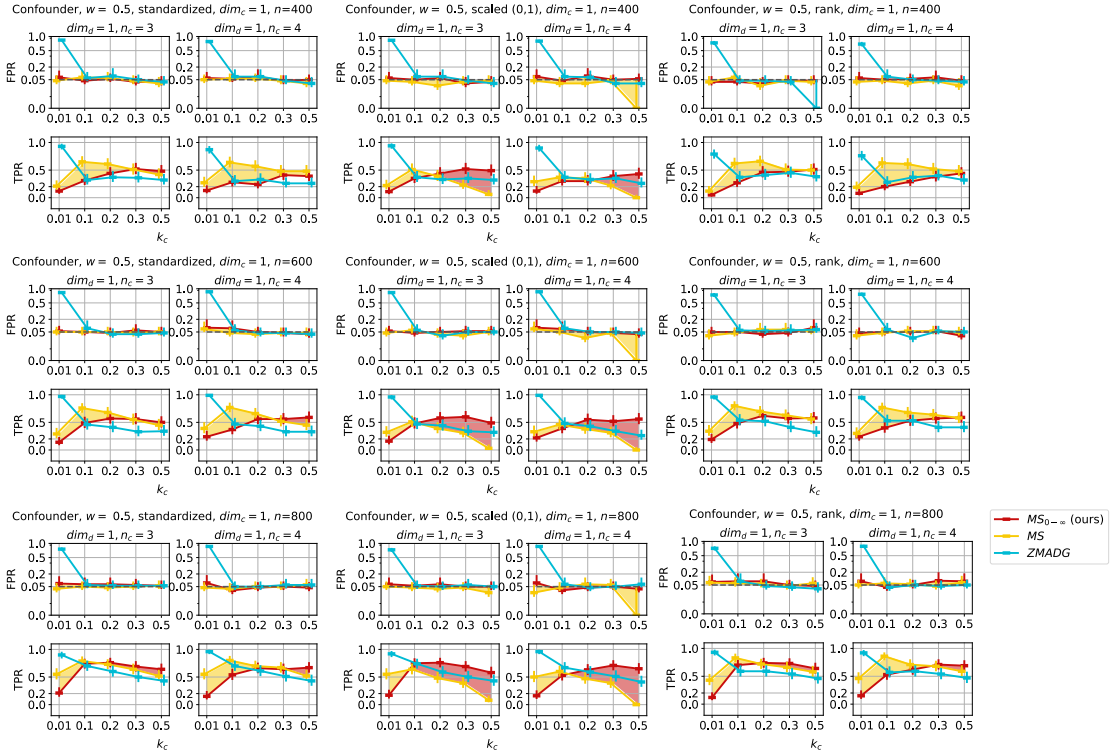

Figure 14: False positive rate (FPR, ideally under 0.05) and true positive rate (TPR, higher is better, with 1 best) for the "Confounder" model with $n \in \{400, 600, 800\}$ where Z has $dim_c = 1, dim_d = 1$, with standardization (left), scaling to $(0, 1)$ (center) and rank preprocessing (right) for the continuous variables and coupling factor in the dependent case $w = 0.5$.

Appendix H. Results on real-world data

Here, we present the numerical results for the experiments described in Sec. 5.3 on the ADHD dataset Bellec et al. (2016). We report the accuracies in Table 4, and refer the reader to a discussion on the results to Sec. 5.3.

Table 4: The accuracy of the $\text{MS}_{0-\infty}$, MS and ZMADG CITs on finding two CI relations (Case 1 and 2) from the ADHD dataset (higher is better, with 1.0 best). Highlighted in bold are the accuracy values for what we consider the optimal k_c. Transformations abbreviated as "rk" (rank), "std" (standardization), "sc" (scale).

		Case 1			Case 2		
k_c	Trsf.	$\text{MS}_{0-\infty}$	MS	ZMADG	$\text{MS}_{0-\infty}$	MS	ZMADG
0.1	rk	0.22	**0.0**	**1.0**	0.94	**0.92**	**1.0**
0.1	std	0.74	**1.0**	**0.96**	0.0	**1.0**	**1.0**
0.1	sc	0.98	**1.0**	**0.98**	0.0	**1.0**	**0.0**
0.2	rk	1.0	0.0	1.0	0.8	0.94	1.0
0.2	std	0.4	0.0	1.0	0.22	0.0	0.0
0.2	sc	0.94	0.12	0.94	0.44	0.0	0.0
0.3	rk	**1.0**	0.6	1.0	**1.0**	1.0	1.0
0.3	std	**1.0**	1.0	1.0	**0.96**	1.0	0.0
0.3	sc	**1.0**	1.0	1.0	**0.94**	1.0	0.0

Appendix I. Comparison with two other CITs

Experimental setup As mentioned in Sec. 5, we also compare against two others CITs, namely the GCM CIT of Shah and Peters (2020) and the kernel-based CIT of Zhang et al. (2011). We use the same synthetic models and hyperparameters described in Sec. 5 and App. G.

For the GCM CIT, we use two different regressors: a k-NN regressor (we use the scikit-learn implementation, Pedregosa et al. (2011)) and a random forest regressor (Breiman, 2001) (using the scikit-learn implementation as well, Pedregosa et al. (2011)). To accommodate for mixed-type data, we apply one-hot encoding to the data. For the k-NN regressor, we set $k = k_c \cdot n$ with $k_c = 0.1$. For the random forest, we set the number of trees to 100. For the kernel-based CIT, we set $\sigma = \{0.001, 0.01, 0.1, 1\}$ and $\theta = 0.5$, following the recommendations of Zhang et al. (2011). We repeat each experiment 100 times.

Results We present the results for the GCM CIT in Table 5. We observe that GCM obtains the best results with the k-NN regressor, and consistently obtains a TPR of 1.0. However, for some of the models, for example for the "Confounder" and "Independent Z" models, the FPR is higher than the desired 0.05 level. The results for the "Cluster-dependent confounder" show one possible disadvantage of the GCM CIT: the wrong choice of regressor can lead to high variability in the results. While the k-NN regressor obtains good results, especially for $n_c = 2$ and $n_c = 3$, the random forest has very high FPR. This is only specific to this model, as for the other models, where dependency holds across clusters, this behaviour cannot be observed.

We do not present any results for the kernel-based CIT as this CIT has either both an FPR and TPR equal to 0.0 or equal to 1.0. Most probably, a further hyperparameter search would be beneficial. However, we see this as out of scope for this work, and leave a comprehensive comparison of different CITs for future work.

Table 5: The false positive rate (FPR, the lower the better, ideally under 0.05) and true positive rate (TPR, the higher the better, ideally 1) of the GCM CIT on the four models described in Sec. 5 and App. G.

Model	**Confounder**				**Independent Z**				**Chain**				**Cluster-dep. confounder**			
Regression type	**k-NN**		**Random forest**		**k-NN**		**Random forest**		**k-NN**		**Random forest**		**k-NN**		**Random forest**	
Measure	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**
$n_c = 2$	-	-	-	-	-	-	-	-	-	-	-	-				
$n_c = 3$	0.08	1.0	1.0	1.0	0.06	1.0	0.07	1.0	0.05	1.0	0.03	1.0	0.03	0.95	0.63	0.84
$n_c = 4$	0.07	1.0	1.0	1.0	0.07	1.0	0.06	1.0	0.07	1.0	0.04	1.0	0.023	0.66	0.63	0.84

Appendix J. Comparison of computational runtimes

We report the computational runtimes for each model in the CMI estimation experiments. The experiments were performed on an Intel(R) Core(TM) i7-6600U CPU. We recorded the runtimes for each run of the 100 repetitions of the CMI experiments described in Sec. 4 using the different k_c values and sample sizes n. In the Tables 6 to 8, we present the average runtime for each sample size n in seconds, averaged over the 100 runs and all k_c values. We also include the average runtime of the different heuristics for our approach: "local", "global", and "cluster" described in App. E. We note that in the tables below, $\text{MS}_{0-\infty}$ stands for the $\text{MS}_{0-\infty}$ in combination with the "$\text{MS}_{0-\infty}$ local" heuristic, and "$\text{MS}_{0-\infty}$ cluster" stands for our estimator in combination with the "cluster-size" heuristic.

Across all estimators, we can observe an increase in runtime with larger sample sizes. $\text{MS}_{0-\infty}$ with the "local" heuristic consistently shows the highest runtimes, indicating that it is the most computationally intensive method among the estimators. This is due to the additional computational steps for the heuristic, as, in comparison, the "global" heuristic has a computational runtime close to the runtime of MS. Thus, users might consider switching to the "global" heuristic for larger experiments, since the numerical results in App. E indicate that the decrease in performance is minimal. ZMADG is the fastest among the three estimators, since conditioning on the discrete variables considerably decreases sample size.

Table 6: Computational runtimes (in seconds) for the "Independent Z" model with $d = 1$ and $d = 3$.

n	**$\text{MS}_{0-\infty}$**	**$\text{MS}_{0-\infty}$ global**	**$\text{MS}_{0-\infty}$ cluster**	**MS**	**ZMADG**	n	**$\text{MS}_{0-\infty}$**	**$\text{MS}_{0-\infty}$ global**	**$\text{MS}_{0-\infty}$ cluster**	**MS**	**ZMADG**
300	0.013	0.016	0.013	0.029	0.036	300	0.021	0.018	0.02	0.025	0.059
600	0.033	0.023	0.034	0.03	0.046	600	0.052	0.043	0.051	0.037	0.087
1000	0.069	0.046	0.076	0.05	0.046	1000	0.107	0.081	0.111	0.061	0.118
2000	0.249	0.142	0.271	0.143	0.05	2000	0.327	0.215	0.333	0.165	0.164

Table 7: Computational runtimes (in seconds) for the "Chain structure" model with $d = 1$ and $d = 3$.

n	**$\text{MS}_{0-\infty}$**	**$\text{MS}_{0-\infty}$ global**	**$\text{MS}_{0-\infty}$ cluster**	**MS**	**ZMADG**		n	$\text{MS}_{0-\infty}$	$\text{MS}_{0-\infty}$ global	$\text{MS}_{0-\infty}$ cluster	MS	ZMADG
300	0.012	0.015	0.012	0.029	0.014	0	300	0.019	0.011	0.019	0.024	0.027
600	0.036	0.023	0.035	0.033	0.021	1	600	0.057	0.024	0.055	0.035	0.041
1000	0.082	0.05	0.08	0.053	0.027	2	1000	0.116	0.041	0.113	0.052	0.054
2000	0.332	0.171	0.306	0.136	0.039	3	2000	0.433	0.097	0.399	0.131	0.079

Table 8: Computational runtimes (in seconds) for the "Confounder with uniform X and Y" model and "Confounder with Gaussian X and Y" model.

n	**MS**$_{0-\infty}$	**MS**$_{0-\infty}$ **global**	**MS**$_{0-\infty}$ **cluster**	**MS**	**ZMADG**	n	**MS**$_{0-\infty}$	**MS**$_{0-\infty}$ **global**	**MS**$_{0-\infty}$ **cluster**	**MS**	**ZMADG**
300	0.01	0.012	0.01	0.025	0.023	300	0.009	0.013	0.009	0.027	0.022
600	0.03	0.023	0.029	0.027	0.023	600	0.027	0.021	0.026	0.029	0.023
1000	0.068	0.051	0.068	0.042	0.026	1000	0.06	0.046	0.059	0.043	0.028
2000	0.253	0.183	0.252	0.115	0.037	2000	0.221	0.156	0.216	0.098	0.04

Appendix K. Remarks on reproducibility

We intentionally reduced the number of plots and tables in this appendix for length reasons. However, the CMI distribution plots, all code to obtain the CIT plots, the measurements for the statistical significance tests, the results for the other two CITs, and the computational time reports can be found in the repository.