Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information

Abstract

Conditional independence testing is a fundamental problem underlying causal discovery and a particularly challenging task in the presence of nonlinear and high-dimensional dependencies. Here a fully non-parametric shuffle test based on conditional mutual information is presented. Through a nearest neighbor scheme it efficiently adapts to highly heterogeneous distributions due to strongly nonlinear dependencies. Numerical experiments demonstrate that the test reliably simulates the null distribution even for small sample sizes and with high-dimensional conditioning sets. Especially for sample sizes below 2000 the test is better calibrated than kernel-based tests and reaches the same or higher power levels. While the conditional mutual information estimator scales more favorably with sample size than kernel-based approaches, a drawback of the test is its computational expensive shuffle scheme making more theoretical research to analytically approximate the null distribution desirable.

1 Introduction

Conditional independence testing lies at the heart of causal discovery (Spirtes et al., 2000) and at the same time is one of its most challenging tasks. For observed random variables X, Y, Z, measuring that X and Y are independent given Z, denoted as $X \perp\!\!\!\perp Y | Z$, implies that no causal link can exist between X and Y under the relatively weak assumption of *faithfulness* (Spirtes et al., 2000). In most real applications, a finding of conditional independence is, thus, much more trustworthy than the finding of *dependence* from which a causal link only follows under stronger assumptions (Spirtes et al., 2000).

Here we focus on the difficult case of continuous variables. While various conditional independence (CI) tests exist if assumptions such as linearity or additivity are justified (for a numerical comparison see Ramsey (2014)), here we focus on the general definition of CI implying that the conditional joint density factorizes: $p(X, Y|Z) = p(X|Z)p(Y|Z)$. Note that wrong assumptions can lead to incorrectly detecting CI (type II error, false negative), but also to wrongly concluding on conditional dependence (type I error, false positive).

Recent research has focused on the general case without assuming a functional form of the dependencies as well as the data distributions. One approach is to discretize the variable Z and make use of easier unconditional independence tests $X \perp\!\!\!\perp Y | Z = z$ (Margaritis, 2005; Huang, 2010). However, this method suffers from the curse of dimensionality for high-dimensional conditioning sets Z.

On the other hand, kernel-based methods are known for their capability to deal with high dimensions. A popular test is the Kernel Conditional Independence Test (KCIT) (Zhang et al., 2011) which essentially tests for zero Hilbert-Schmidt norm of the partial cross-covariance operator, or the Permutation CI test (Doran et al., 2014) which solves an optimization problem to generate a permutation surrogate on which kernel two sample testing can be applied. Kernel methods suffer from high computational complexity since large kernel matrices have to be computed. Strobl et al. (2017) present an orders of magnitude faster CI test based on approximating kernel methods using *random Fourier features*, called Randomized Conditional Correlation Test (RCoT). Last, Wang et al. (2015) proposed a *conditional distance correlation* (CDC) test based on the correlation of distance matrices between X, Y, Z which have been linked to kernel-based approaches (Sejdinovic et al., 2013).

Kernel and distance methods in general require carefully adjusted bandwidth parameters that characterize the length scales between samples in the different subspaces of X, Y, Z. These bandwidths are *global* in each subspace in the sense that they are applied on the whole range of values for X, Y, Z, respectively. Additionally, the theoretical null distributions derived for RCoT (Strobl et al., 2017) and CDC (Wang et al., 2015) require potentially violated assumptions for their finite sample approximations.

Our approach to testing CI is founded in an information-theoretic framework. The conditional mutual information is zero if and only if $X \perp\!\!\!\perp Y|Z$. Our test combines the well-established Kozachenko-Leonenko k-nearest neighbor estimator (Kozachenko and Leonenko, 1987; Kraskov et al., 2004; Frenzel and Pompe, 2007; Vejmelka and Paluš, 2008) with a nearest-neighbor permutation shuffle test. Their main advantage is that nearest-neighbor statistics are *locally adaptive*: The hypercubes around each sample point are smaller where more samples are available. Unfortunately, few theoretical results are available for the complex mutual information estimator. While the Kozachenko-Leonenko estimator is asymptotically unbiased and consistent (Kozachenko and Leonenko, 1987; Leonenko et al., 2008), the variance and finite sample convergence rates are unknown. Hence, our approach relies on a local permutation test that is also based on nearest neighbors and, hence, data-adaptive.

Our numerical experiments comparing our test with KCIT, RCoT, and CDC show that the test reliably simulates the null distribution even for small sample sizes and with high dimensional conditioning sets. It yields a better calibrated test than asymptotics-based kernel tests such as KCIT or RCoT while reaching the same or higher power levels. While the conditional mutual information estimator scales more favorably with sample size than kernel-based approaches or CDC by making use of KD-tree neighbor search methods, a major drawback is its computationally expensive permutation scheme making more theoretical research to analytically approximate the null distribution desirable.

2 Conditional independence test

2.1 Conditional mutual information

Conditional mutual information (CMI) for continuous and possibly multivariate random variables X, Y, Z is defined as

$$\begin{aligned} I_{X;Y|Z} & \\ &= \iiint dxdydz\; p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)\cdot p(y|z)} \quad (1)\\ &= H_{XZ} + H_{YZ} - H_Z - H_{XYZ}\,, \quad (2)\end{aligned}$$

where H denotes the Shannon entropy. We wish to test the following hypotheses:

$$H_0: \quad X \perp\!\!\!\perp Y \mid Z \quad (3)$$

$$H_1: \quad X \not\perp\!\!\!\perp Y \mid Z \quad (4)$$

From the definition of CMI it is immediately clear that $I_{X;Y|Z} = 0$ if and only if $X \perp\!\!\!\perp Y|Z$. Shannon-type conditional mutual information is theoretically well-founded and its value is well interpretable as the shared information between X and Y not contained in Z. While this does not immediately matter for a conditional independence test's p-value, causal discovery algorithms often make use of the test statistic's value, for example to sort the order in which conditions are tested. CMI here readily allows for an interpretation in terms of the relative importance of one condition over another.

2.2 Nearest-neighbor CMI estimator

Inspired by Dobrushin (1958), Kozachenko and Leonenko (1987) introduced a class of differential entropy estimators that can be generalized to estimators of conditional mutual information. This class is based on nearest neighbor statistics as further discussed in Kozachenko and Leonenko (1987); Frenzel and Pompe (2007). For a D_X-dimensional random variable X the nearest neighbor entropy estimate is defined as

$$\widehat{H}_X = \psi(n) + \frac{1}{n}\sum_{i=1}^{n}\left[-\psi(k_{X,i}) + \log(\epsilon_i^{D_X})\right] + \log(V_{D_X}) \quad (5)$$

with the Digamma function as the logarithmic derivative of the Gamma function $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$, sample length n, volume element V depending on the chosen metric, i.e., $V_{D_X} = 2^{D_X}$ for the maximum metric, $V_{D_X} = \pi^{D_X/2}/\Gamma(D_X/2+1)$ for euclidean metric with Gamma function Γ. For every sample with index i, the integer $k_{X,i}$ is the number of points in the D_X-dimensional ball with radius ϵ_i. Formula (5) holds for any ϵ_i and the corresponding $k_{X,i}$, which will be used in the following definition of a CMI estimator. Based on this entropy estimator, Kraskov et al. (2004) derived an estimator for mutual information where the epsilon balls with radius ϵ_i are hypercubes. This estimator was generalized to an estimator for CMI first by Frenzel and Pompe (2007) and independently by Vejmelka and Paluš (2008). The CMI estimator is obtained by inserting the entropy estimator Eq. (5) for the different entropies in the definition of CMI in Eq. (2). For all entropy terms H_{XZ}, H_{YZ}, H_Z, H_{XYZ} in Eq. (2), we use the maximum norm and choose as the side length $2\epsilon_i$ of the hypercube the distance ϵ_i to the $k = k_{XYZ,i}$-nearest neighbor in the joint space $X \oplus Y \oplus Z$. The CMI estimate then is

$$\widehat{I}_{XY|Z} = \psi(k) + \frac{1}{n}\sum_{i=1}^{n}\left[\psi(k_{Z,i}) - \psi(k_{XZ,i}) - \psi(k_{YZ,i})\right]. \tag{6}$$

The only free parameter k is the number of nearest neighbors in the joint space of $X \oplus Y \oplus Z$ and $k_{xz,i}$, $k_{yz,i}$ and $k_{z,i}$ are computed as follows for every sample point indexed by i:

1. Determine (here in maximum norm) the distance ϵ_i to its k-th nearest neighbor (excluding the reference point which is not a neighbor of itself) in the joint space of $X \oplus Y \oplus Z$.

2. Count the number of points with distance strictly smaller than ϵ_i (including the reference point at i) in the subspace $X \oplus Z$ to get $k_{xz,i}$, in the subspace $Y \oplus Z$ to get $k_{yz,i}$, and in the subspace Z to get $k_{z,i}$.

Similar estimators, but for the more general class of Rényi entropies and divergences, were developed in Wang et al. (2009); Schneider and Póczos (2012). Estimator (6) uses the approximation that the densities are constant within the epsilon environment. Therefore, the estimator's bias will grow with k since larger k lead to larger ϵ-balls where the assumption of constant density is more likely violated. The variance, on the other hand, is the more important quantity in conditional independence testing and it becomes smaller for larger k because fluctuations in the ϵ-balls average out. The decisive advantage of this estimator compared to fixed bandwidth approaches is its *data-adaptiveness*.

The Kozachenko-Leonenko estimator is asymptotically unbiased and consistent (Kozachenko and Leonenko, 1987; Leonenko et al., 2008). Unfortunately, at present there are no results, neither exact nor asymptotically, on the distribution of the estimator as needed to derive analytical significance bounds. In Goria and Leonenko (2005), some numerical experiments indicate that for many distributions of X, Y the asymptotic distribution of MI is Gaussian. But the important finite size dependence on the dimensions D_X, D_Y, D_Z, the sample length n and the parameter k are unknown.

Some notes on the implementation: Before estimating CMI, we rank-transform the samples individually in each dimension: Firstly, to avoid points with equal distance, small amplitude random noise is added to break ties. Then, for all n values $x_1, \ldots, x_n$, we replace x_i with the transformed value r, where r is defined such that x_i is the rth largest among all x values. The main computational cost comes from searching nearest neighbors in the high dimensional subspaces which we speed up using *KD-tree* neighbor search. Hence, the computational complexity will typically scale less than quadratically with the sample size. Kernel methods, on the other hand, typically scale worse than quadratically in sample size (Strobl et al., 2017). Further, the CMI estimator scales roughly linearly in k and D, the total dimension of X, Y, Z.

2.3 Nearest-neighbor permutation test

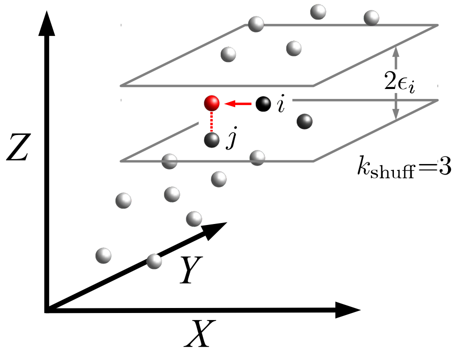

Figure 1: Schematic of local permutation scheme. Each sample point i's x-value is mapped randomly to one of its k_{shuff}-nearest neighbors in subspace Z. The hypercubes with length scale ϵ_i locally adapt to the density making this scheme more data efficient than fixed bandwidth techniques. By keeping track of already 'used' indices j, we approximately achieve a random draw *without* replacement, see Algorithm 1.

Since no theory on finite sample behavior of the CMI estimator is available, we resort to a permutation-based generation of the distribution under H_0.

Typically in CMI-based independence testing, CMI-surrogates to simulate independence are generated by randomly shuffling *all* values in X. The problem is, that this approach not only destroys the dependence between X and Y, as desired, but also destroys all dependence between X and Z. Hence, this approach does not actually test $X \perp\!\!\!\perp Y \mid Z$. In order to preserve the dependence between X and Z, we propose a local permutation test utilizing nearest-neighbor search. To avoid confusion, we denote the CMI-estimation parameter as $k_{\rm CMI}$ and the shuffle-parameter as $k_{\rm shuff}$.

As illustrated in Fig. 1, we first identify the $k_{\rm shuff}$-nearest neighbors around each sample point i (here including the point itself) in the subspace of Z using the maximum norm. With Algorithm 1 we generate a permutation mapping $\pi : \{1, \ldots, n\} \to \{\pi(1), \ldots, \pi(n)\}$ which tries to achieve draws *without* replacement. Since this is not always possible, some values might occur more than once, i.e., they were drawn *with* replacement as in a bootstrap. In Paparoditis and Politis (2000) a bootstrap scheme that *always* draws with replacement is described which is used for the CDC independence test. Since we rank-transform the data in the CMI estimation, which is also based on nearest-neighbor distances, we try to avoid tied samples as much as possible to preserve the conditional marginals.

Algorithm 1 Algorithm to generate a nearest-neighbor permutation $\pi(\cdot)$ of $\{0, 1, \ldots, n\}$.

```
Denote by $d_i^{k_{\rm shuff}}$ the distance of sample point $z_i$ to its $k_{\rm shuff}$-nearest neighbor (including $i$ itself, i.e., $d_i^{k_{\rm shuff}=1} = 0$)
Compute list of nearest neighbors for each sample point: $\mathcal{N}_i = \{l \in \{0, \ldots, n\} : \|z_l - z_i\| \leq d_i^{k_{\rm shuff}}\}$ with KD-tree algorithm in maximum norm of subspace $Z$
Shuffle $\mathcal{N}_i$ for each $i$
Initialize empty list $\mathcal{U} = \{\}$ of used indices
for all $i \in$ random permutation of $\{1, \ldots, n\}$ do
    $j = \mathcal{N}_i(0)$
    $m = 0$
    while $j \in \mathcal{U}$ and $m < k_{\rm shuff} - 1$ do
        $m = m + 1$
        $j = \mathcal{N}_i(m)$
    $\pi(i) = j$
    Add $j$ to $\mathcal{U}$
return $\{\pi(1), \ldots, \pi(n)\}$
```

The permutation test is then as follows:

1. Generate random permutation $x^* = \{x_{\pi(1)}, \ldots, x_{\pi(n)}\}$ with Algorithm 1
2. Compute surrogate CMI $\widehat{I}(x^*; y|z)$ via Eq. (6)
3. Repeat steps (1) and (2) B times, sort the surrogate values $\widehat{I}_b$ and obtain p-value by

$$p = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\widehat{I}_b \geq \widehat{I}(x;y|z)}\,, \qquad (7)$$

   where $\mathbb{1}$ denotes the indicator function.

The CMI estimator holds for arbitrary dimensions of all arguments X, Y, Z and also the local permutation scheme can be used to jointly shuffle all of X's dimensions. In the following numerical experiments, we focus on the case of univariate X and Y and uni- or multivariate Z.

3 Experiments

3.1 Choosing $k_{\rm CMI}$ and $k_{\rm shuff}$

Our approach has two free parameters $k_{\rm CMI}$ and $k_{\rm shuff}$. The following numerical experiments indicate that restricting $k_{\rm shuff}$ to only very few nearest neighbors already suffices to reliably simulate the null distribution in most cases while for $k_{\rm CMI}$ we derive a rule-of-thumb based on the sample size n.

Figure 2 illustrates the effect of different $k_{\rm shuff}$. If $k_{\rm shuff}$ is too large or even $k_{\rm shuff} = n$, the shuffle distribution under independence (red) is negatively biased. As illustrated by the red markers, this would lead to an increase of false positives (type-I error). On the other hand, for the dependent case, if $k_{\rm shuff} = 1..3$, the shuffle distribution is positively biased yielding lower power (type-II errors). For a range of optimal values of $k_{\rm shuff}$, the shuffled distribution perfectly simulates the true null distribution.

To evaluate the effect of $k_{\rm CMI}$ and $k_{\rm shuff}$ numerically, we followed the post-nonlinear noise model described in Zhang et al. (2011); Strobl et al. (2017) given by $X = g_X(\epsilon_X + \frac{1}{D_Z}\sum_i^{D_Z} Z_i)$, $Y = g_Y(\epsilon_Y + \frac{1}{D_Z}\sum_i^{D_Z} Z_i)$, where $Z_i, \epsilon_X, \epsilon_Y$ have jointly independent standard Gaussian distributions, and g_X, g_Y denote smooth functions uniformly chosen from $(\cdot), (\cdot)^2, (\cdot)^3, \tanh(\cdot), \exp(||\cdot||^2)$. Thus, we have $X \perp\!\!\!\perp Y \mid Z = (Z_1, Z_2, \ldots)$ in any case. To simulate dependent X and Y, we used $X = g_X(c\epsilon_b + \epsilon_X)$,

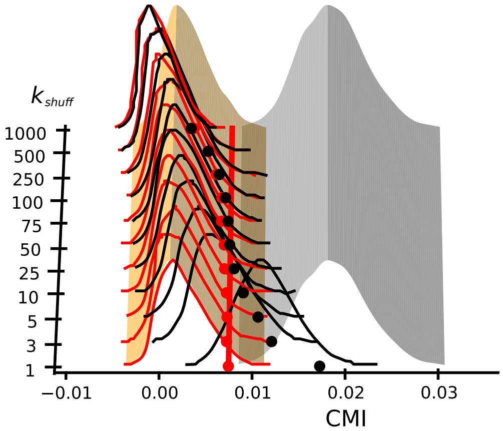

Figure 2: Simulation to illustrate the effect of the nearest-neighbor shuffle parameter k_{shuff}. The true null distribution of CMI is depicted as the orange surface with the 5% quantile marked by a red straight line. The true distribution under dependence is drawn as a grey surface. The red and black distributions and markers give the shuffled null distributions and their 5% quantiles for different k_{shuff} for the independent (red) and dependent (black) case, respectively. Here the sample size is $n = 1000$ such that $k_{\text{shuff}} = 1000$ corresponds to a full non-local permutation.

$Y = g_Y(c\epsilon_b + \epsilon_Y)$ for $c > 0$ and identical Gaussian noise ϵ_b and keep Z independent of X and Y.

In Fig. 3, we show results for sample size $n = 1000$. The null distribution was generated with $B = 1000$ surrogates in all experiments. The results demonstrate that a value $k_{\text{shuff}} \approx 5..10$ yields well-calibrated tests while not affecting power much. This holds for a wide range of sample sizes as shown in Fig. 9.

Larger k_{CMI} yield more power and even for $k_{\text{CMI}} \approx n/2$ the tests are still well calibrated. But power peaks at some value of k_{CMI} and slowly decreases for too large values. Still, the dependency of power on k_{CMI} is relatively robust and we suggest a rule-of-thumb of $k_{\text{CMI}} \approx 0.1..0.2n$. Note that, as shown in Fig. 4, runtime increases linearly with k_{CMI} while k_{shuff} does not impact runtime much.

3.2 Comparison with KCIT and RCoT

In Fig. 5 we show results comparing our CMI test (CMIT) to KCIT and RCoT (Strobl et al., 2017). We used the rule-of-thumb $k_{\text{CMI}} = 0.2n$ and $k_{\text{shuff}} = 5$ with $B = 1000$ permutation surrogates. As a metric for type-I errors, as in Strobl et al. (2017) we evaluated the Kolmogorov-Smirnov (KS) statistic to quantify how uniform the distribution of p-values is. For type-II errors we measure the area under the power curve (AUPC). All metrics were evaluated from 1000 realizations and error bars give the boostrapped standard errors.

Figure 5 demonstrates that CMIT is better calibrated with the lowest KS-values for almost all sample sizes tested. KCIT is especially badly calibrated for smaller sample sizes or higher dimensions D_Z and RCoT better approximates the null distribution only for $n \geq 500$ for $D_Z = 1$ and for $n \geq 1000$ for $D_Z = 8$. Note that this is also expected (Strobl et al., 2017) since the analytical approximation of the null distribution for KCIT and RCoT requires large sample sizes. The power as measured by AUPC is, thus only comparable for $n > 500$ for $D_Z = 1$ and CMIT has the highest power throughout. Also for $D_Z = 8$ and $n \geq 1000$ CMIT has higher power than RCoT. The other side of the story is that the runtime of CMIT is much higher due to the computationally expensive permutation scheme. Note that each single CMI estimate comes at a lower computational complexity compared to KCIT, but not necessarily compared to RCoT whose runtime also depends on the number of random Fourier features used (here the default of 25 for subspace Z and 5 for subspaces X and Y was used). If a permutation scheme is utilized for KCIT and RCoT, their advantage of a faster runtime vanishes.

Another drawback of kernel-based methods is illustrated in Fig. 6 where we consider a multiplicative noise case with the model $X = g_X(0.1\epsilon'_X + \epsilon_X \frac{1}{D_Z}\sum_i^{D_Z} Z_i)$, $Y = g_Y(0.1\epsilon'_Y + \epsilon_Y \frac{1}{D_Z}\sum_i^{D_Z} Z_i)$ with all variables as before and $\epsilon'_{X,Y}$ another independent Gaussian noise term. Even though the density is highly localized in this case, CMIT is still well calibrated for $k_{\text{shuff}} \approx 5$. On the other hand, RCoT (shown with blue markers in Fig. 6) cannot control false positives even if we vary the number of Fourier features to much higher values (which takes much longer) because it doesn't resolve the heterogeneous density.

For an extremely oscillatory sinusoidal dependency like $X = \sin(\lambda Z) + \epsilon_X$ and $Y = \sin(\lambda Z) + \epsilon_Y$, shown in Fig. 7, k_{shuff} needs to be set to a very small value in order to control false positives. Here RCoT does not work at all.

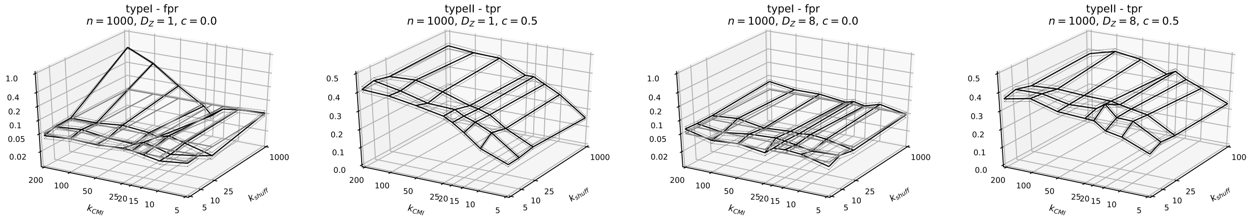

Figure 3: Numerical experiments with post-nonlinear noise model (Zhang et al., 2011; Strobl et al., 2017). The sample size is $n = 1000$ and 1000 realizations were generated to evaluate false positives (fpr) and true positives (tpr) for $c = 0.5$ at the 5% significance level. Shown are fpr and tpr for $D_Z = 1$ (two left panels) and $D_Z = 8$ (two right panels).

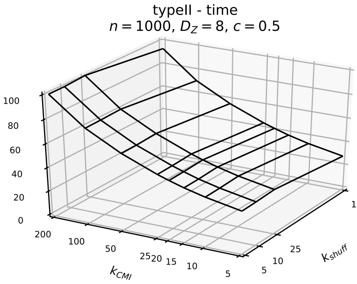

Figure 4: Runtime for the same setup as in the right panel of Fig. 3. For $k_{\text{shuff}} = n$ a computationally cheaper full permutation scheme was used.

3.3 Comparison with CDC

In Tab. 1 we repeat the results from Wang et al. (2015) proposing the CDC test together with results from RCoT and our CMI test. The experiments are described in Wang et al. (2015). Examples 1–4 correspond to conditional independence and Examples 5–8 to dependent cases. CMIT has well-calibrated tests except for Example 4 (as well as Example 8) which is based on discrete Bernoulli random variables while the CMI test is designed for continuous variables. For Examples 5–8 CMIT has competitive power compared to CDC and outperforms KCIT in all and RCoT in all but Example 5 where they reach the same performance. Note that the CDC test also is based on a computationally expensive local permutation scheme since the asymptotics break down for small sample sizes.

4 Real data application

We apply CMIT in a time series version of the PC causal discovery algorithm [reference hidden in review to preserve anonymity] to investigate dependencies between hourly averaged concentrations for carbon monoxide (CO), benzene (C6H6), total nitrogen oxides (NOx), nitrogen dioxide (NO2), as well as temperature (T), relative humidity (RH) and absolute humidity (AH) taken from De Vito et al. (2008)[1]. The time series were smoothed using a Gaussian kernel smoother with bandwidth $\sigma = 1440$ *hours* and we limited the analysis to the first three months of the dataset (2160 samples). After accounting for missing values we obtain an effective sample size of $n = 1102$. As in our numerical experiments, we used the CMIT parameters $k_{\text{CMI}} = 200$ and $k_{\text{shuff}} = 5$ with $B = 1000$ permutation surrogates. The causal discovery algorithm was run including lags from $\tau = 1$ up to $\tau_{\max} = 3$ *hours*. The resulting graph at a 10% FDR-level shown in Fig. 8 indicates that temperature and relative humidity influence Benzene which in turn affects NO2 and CO concentrations.

5 Conclusion

We presented a novel fully non-parametric conditional independence test based on a nearest neighbor estimator of conditional mutual information. Its main advantage lies in the ability to adapt to highly localized densities due to nonlinear dependencies even in higher dimensions. This feature results in well-calibrated tests with reliable false positive rates. We tested setups for sample sizes $n = 50$ to $n = 2000$ and dimensions of the conditional set of $D_Z = 1..10$. The power of CMIT is comparable to advanced kernel based tests such as KCIT or its faster random Fourier feature version RCoT, which, however, are not well-calibrated in the smaller sample limit. CMI has a lower computational complexity than KCIT since efficient nearest-neighbor search schemes can be utilized, but relies on a permutation scheme since no analytics are known for the null distribution. The permutation scheme leads to a higher computational load which, however, can be easily parallelized. Nev-

[1] `http://archive.ics.uci.edu/ml/datasets/Air+Quality`

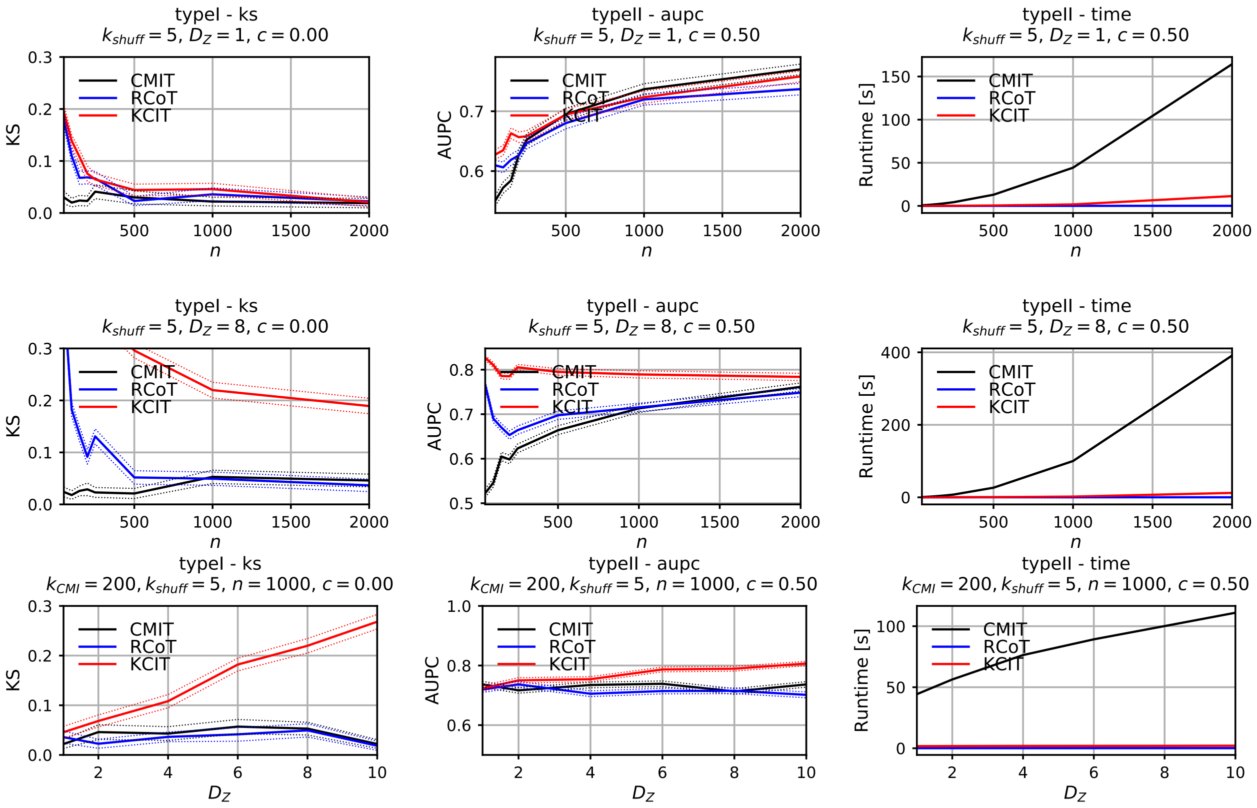

Figure 5: Numerical experiments with post-nonlinear noise model and similar setup as in Strobl et al. (2017). Shown are KS (left column), AUPC (center column), and runtime (right column) for a sample size experiment with $D_Z = 1$ (top row) and $D_Z = 8$ (center row), as well as an experiment for different condition dimensions D_Z with fixed $n = 1000$ (bottom row). In all experiments we set $k_{\text{CMI}} = 0.2n$ and $k_{\text{shuff}} = 5$. CMIT is better calibrated also for small sample sizes and has power on par or higher than RCoT. Note that the higher runtime of CMIT is due to the permutation scheme, each single CMI estimate is much faster than KCIT, but still mostly slower than RCoT depending on RCoT's parameters.

ertheless, more theoretical research is desirable to obtain approximate analytics for the null distribution.

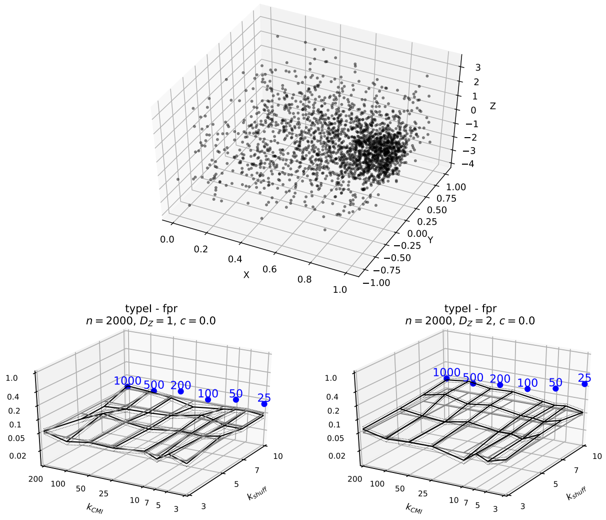

Figure 6: Example of multiplicative dependence of X and Y on Z leading to strongly nonlinear structure (top panel). Here $X \perp\!\!\!\perp Y \mid Z$ and the nearest-neighbor scheme of CMIT can better adapt to the very localized density for $D_Z = 1$ (left) and $D_Z = 2$ (right) with $k_{\text{shuff}} < 7$ while RCoT cannot control false positives for $D_Z = 2$ even if we resolve smaller scales better using a larger number of Fourier features (blue markers) in Z.

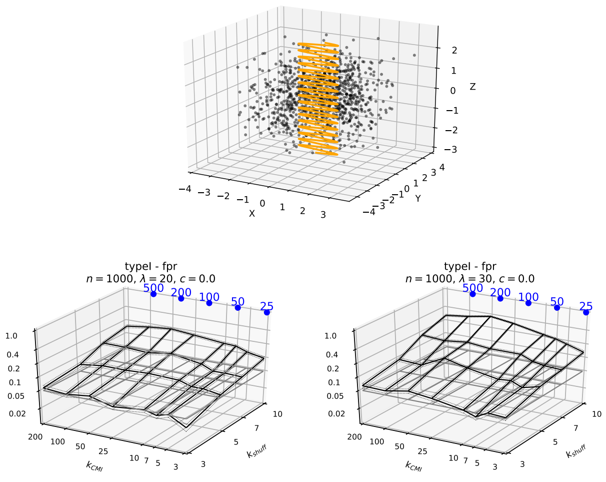

Figure 7: Example of sinusoidal dependence $X = \sin(\lambda Z) + \epsilon_X$ and $Y = \sin(\lambda Z) + \epsilon_Y$ leading to strongly oscillatory structure (top panel for $\lambda = 20$). The bottom shows results for frequencies $\lambda = 20$ (left) and $\lambda = 30$ (right). Here again $X \perp\!\!\!\perp Y \mid Z$ and the nearest-neighbor scheme of CMIT only works for very small $k_{\text{shuff}} = 3$ while RCoT cannot be made to control false positives at all.

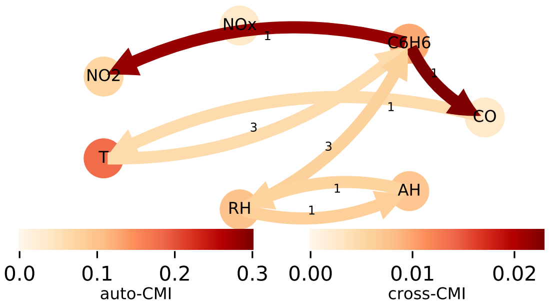

Figure 8: Causal discovery in time series of air pollutants and various weather variables. The node color gives the strength of auto-CMI and the edge color the cross-CMI with the link labels denoting the time lag in *hours*.

Table 1: Results from Wang et al. (2015) together with results from RCoT and our CMI test. The experiments are described in Wang et al. (2015). Examples 1–4 correspond to conditional independence showing false positives and Examples 5–8 to dependent cases showing true positives at the 5% significance level. CMIT was run with $k_{\text{CMI}} = 0.2n$ and $k_{\text{shuff}} = 5, 10$.

	Example 1					Example 2				
Test	50	100	150	200	250	50	100	150	200	250
CDIT	0.035	0.034	0.05	0.057	0.048	0.046	0.053	0.055	0.048	0.058
CI.test	0.041	0.051	0.037	0.054	0.041	0.062	0.046	0.044	0.045	0.039
KCI.test	0.039	0.043	0.041	0.04	0.046	0.035	0.004	0.037	0.047	0.05
Rule-of-thumb	0.017	0.027	0.028	0.033	0.033	0.034	0.052	0.044	0.042	0.045
RCoT	0.074	0.059	0.055	0.043	0.050	0.056	0.056	0.069	0.055	0.073
CMIT ($k_{\text{shuff}} = 5$)	0.064	0.055	0.050	0.053	0.045	0.076	0.060	0.074	0.061	0.065
CMIT ($k_{\text{shuff}} = 10$)	0.058	0.061	0.057	0.058	0.046	0.075	0.066	0.053	0.057	0.071
	Example 3					Example 4				
Test	50	100	150	200	250	50	100	150	200	250
CDIT	0.035	0.048	0.055	0.053	0.043	0.049	0.054	0.051	0.058	0.053
CI.test	0.222	0.363	0.482	0.603	0.677	0.043	0.064	0.066	0.05	0.053
KCI.test	0.058	0.047	0.057	0.061	0.054	0.037	0.035	0.058	0.039	0.049
Rule-of-thumb	0.019	0.038	0.032	0.039	0.039	0.037	0.04	0.055	0.059	0.053
RCoT	0.074	0.047	0.046	0.053	0.054	0.115	0.072	0.066	0.061	0.053
CMIT ($k_{\text{shuff}} = 5$)	0.044	0.043	0.046	0.046	0.054	0.084	0.071	0.067	0.079	0.070
CMIT ($k_{\text{shuff}} = 10$)	0.063	0.065	0.061	0.076	0.067	0.101	0.113	0.106	0.098	0.084
	Example 5					Example 6				
Test	50	100	150	200	250	50	100	150	200	250
CDIT	0.898	0.993	1	1	1	0.752	0.995	1	1	1
CI.test	0.978	1	1	1	1	0.468	0.434	0.467	0.476	0.474
KCI.test	0.158	0.481	0.557	0.602	0.742	0.296	0.862	0.995	1	1
Rule-of-thumb	0.368	0.793	0.927	0.983	0.994	1	1	1	1	1
RCoT	0.817	0.986	0.998	1	1	0.301	0.533	0.679	0.807	0.860
CMIT ($k_{\text{shuff}} = 5$)	0.782	0.981	0.998	1	1	0.806	0.997	0.999	1	1
CMIT ($k_{\text{shuff}} = 10$)	0.855	0.995	1	1	1	0.805	0.995	1	1	1
	Example 7					Example 8				
Test	50	100	150	200	250	50	100	150	200	250
CDIT	0.918	0.998	1	1	1	0.361	0.731	0.949	0.977	0.994
CI.test	0.953	0.984	0.983	0.995	0.987	0.456	0.476	0.464	0.461	0.485
KCI.test	0.574	0.947	0.998	1	1	0.089	0.401	0.685	1	1
Rule-of-thumb	0.073	0.302	0.385	0.514	0.515	0.043	0.233	0.551	0.851	0.972
RCoT	0.594	0.880	0.962	0.985	0.991	0.275	0.392	0.470	0.624	0.654
CMIT ($k_{\text{shuff}} = 5$)	0.753	0.963	0.992	0.997	1	0.302	0.644	0.804	0.916	0.958
CMIT ($k_{\text{shuff}} = 10$)	0.798	0.976	0.999	0.999	0.999	0.323	0.680	0.832	0.920	0.971

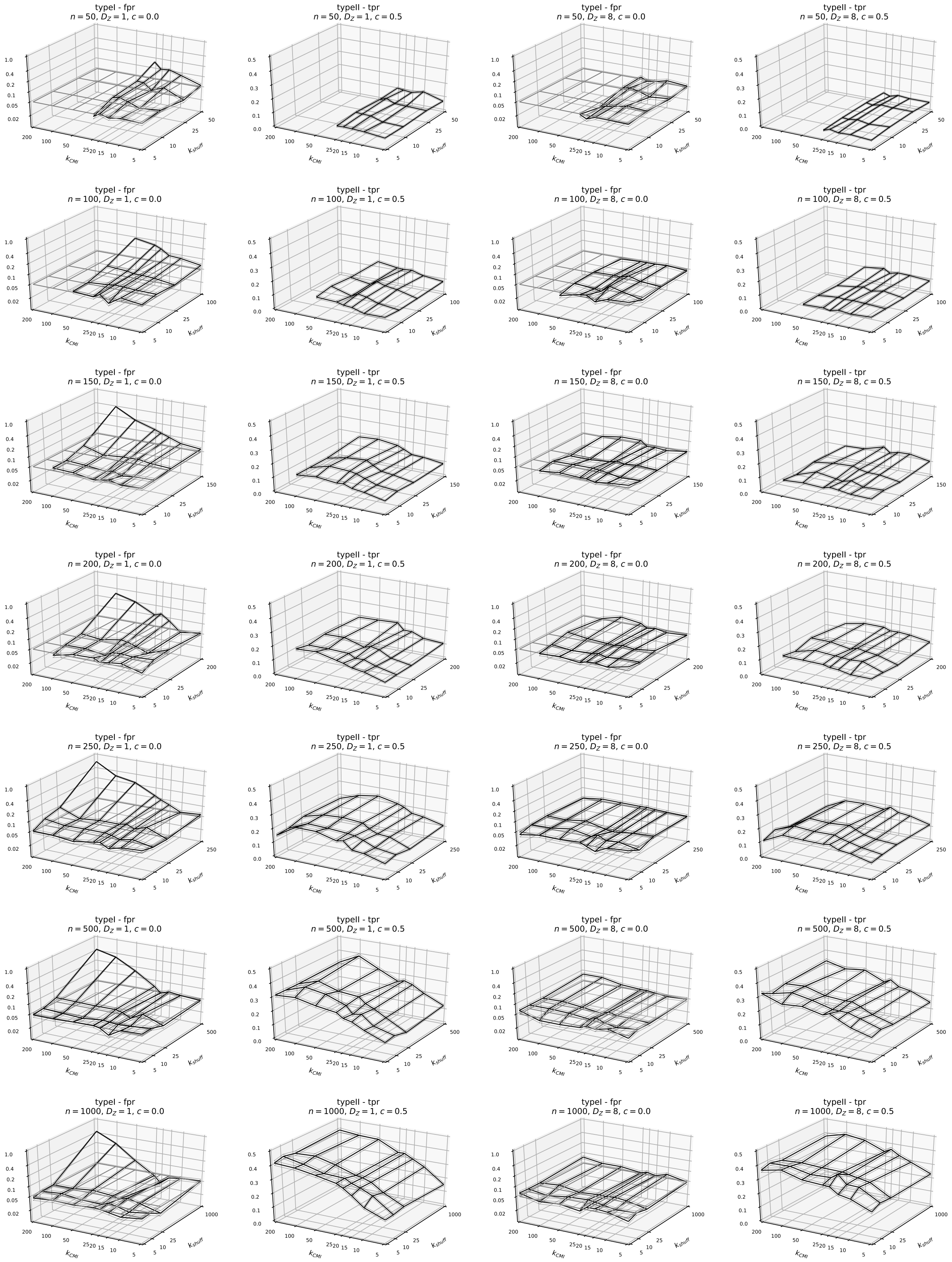

Figure 9: Same as in Fig. 3, but for more sample sizes from $n = 50$ (top) to $n = 1000$ (bottom).

Acknowledgements

We thank Eric Strobl for kindly providing R-code for KCIT and RCoT.

References

De Vito, S., Massera, E., Piga, M., Martinotto, L., and Di Francia, G. (2008). On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators, B: Chemical*, 129(2):750–757.

Dobrushin, R. L. (1958). A simplified method of experimentally evaluating the entropy of a stationary sequence. *Theory of Probability & Its Applications*, 3(4).

Doran, G., Muandet, K., Zhang, K., and Schölkopf, B. (2014). A Permutation-Based Kernel Conditional Independence Test. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 132–141.

Frenzel, S. and Pompe, B. (2007). Partial Mutual Information for Coupling Analysis of Multivariate Time Series. *Physical Review Letters*, 99(20):204101.

Goria, M. N. and Leonenko, N. N. (2005). A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Nonparametric Statistics*, 17(3):277–297.

Huang, T. M. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *Annals of Statistics*, 38(4):2047–2091.

Kozachenko, L. F. and Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6):066138.

Leonenko, N. N., Pronzato, L., and Savani, V. (2008). A class of Rényi information estimators for multidimensional densities. *The Annals of Statistics*, 36(5):2153–2182.

Margaritis, D. (2005). Distribution-free learning of Bayesian network structure in continuous domains. *Proceedings of the National Conference on Artificial Intelligence*, 20(2):825.

Paparoditis, E. and Politis, D. N. (2000). The local bootstrap for kernel estimators under general dependence conditions. *Annals of the Institute of Statistical Mathematics*, 52(1):139–159.

Ramsey, J. D. (2014). A Scalable Conditional Independence Test for Nonlinear, Non-Gaussian Data. *https://arxiv.org/abs/1401.5031*.

Schneider, J. and Póczos, B. (2012). Nonparametric estimation of conditional information and divergences. *15th International Conference on Artificial Intelligence and Statistics*, XX:914–923.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Stat.*, 41(5):2263–2291.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*, volume 81. The MIT Press, Boston.

Strobl, E. V., Zhang, K., and Visweswaran, S. (2017). Approximate Kernel-based Conditional Independence Tests for Fast Non-Parametric Causal Discovery. *http://arxiv.org/abs/1702.03877*.

Vejmelka, M. and Paluš, M. (2008). Inferring the directionality of coupling with conditional mutual information. *Physical Review E*, 77(2):026214.

Wang, Q., Kulkarni, S. R., and Verdú, S. (2009). Divergence Estimation for Multidimensional Densities Via k -Nearest-Neighbor Distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015). Conditional Distance Correlation. *Journal of the American Statistical Association*, 110(512):1726–1734.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based Conditional Independence Test and Application in Causal Discovery. In *UAI*, pages 804–813.