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

Jakob Runge

German Aerospace Center
Institute of Data Science
07745 Jena, Germany

Grantham Institute
Imperial College
London SW7 2AZ, United Kingdom

Abstract

Conditional independence testing is a fundamental problem in causal discovery and a particularly challenging task in the presence of nonlinear dependencies. Here a fully non-parametric test for continuous data based on conditional mutual information combined with a local permutation scheme is presented. Numerical experiments covering sample sizes from 50 to 2,000 and dimensions up to 10 demonstrate that the test reliably generates the null distribution. For smooth nonlinear dependencies, the test has higher power than kernel-based tests in lower dimensions and similar power in higher dimensions. For highly non-smooth densities the data-adaptive nearest neighbor approach is particularly well-suited while kernel methods yield much lower power. The experiments also show that kernel methods utilizing an analytical approximation of the null distribution are not well-calibrated for sample sizes below 1,000. Combining the local permutation scheme with these kernel tests leads to better calibration but lower power. For smaller sample sizes and lower dimensions, the proposed test is faster than random fourier feature-based kernel tests if (embarrassingly) parallelized, but the runtime increases more sharply with sample size and dimensionality. Thus, more theoretical research to analytically approximate the null distribution and speed up estimation is desirable. As illustrated on real data, the test is ideally suited in combination with causal discovery algorithms.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

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). A finding of conditional independence is then more pertinent to causal discovery than a finding of (conditional) *dependence* from which a causal link only follows under stronger assumptions (Spirtes et al., 2000).

Here the focus is on the difficult case of continuous variables (Bergsma, 2004). Various conditional independence (CI) tests exist if assumptions such as linearity or additivity (Daudin, 1980; Peters et al., 2013) are justified (for a numerical comparison see Ramsey (2014)). However, 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, that is, the goal is the general definition of CI implying that the conditional joint density factorizes: $p(X, Y|Z) = p(X|Z)p(Y|Z)$.

One approach is to discretize or cluster 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 clustering high-dimensional conditioning sets Z. On the other hand, kernel-based methods are known for their capability to deal with nonlinearity and high dimensions (Fukumizu et al., 2008). 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 two types of orders of magnitude faster CI tests based on approximating kernels using *random Fourier features*, called Randomized Conditional Correlation Test (RCoT) and Randomized Conditional Independence Test (RCIT). RCoT can be related to kernelized two-step conditional independence testing (Zhang et al., 2017). 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). Last, very recent work proposes to apply deep learning classifiers in combination with permutation tests (Sen et al., 2017).

Testing for independence requires access to the null distribution under CI. Strobl et al. (2017) and Wang et al. (2015) derived asymptotic approximations of the theoretical null distributions, but such approximations only hold for larger sample sizes. The alternative are permutation-based approaches, where the null-distribution is generated by computing the test statistic from permuted samples.

In the present paper, the proposed approach to testing CI is founded in an information-theoretic framework. The conditional mutual information (CMI) is zero if and only if $X \perp\!\!\!\perp Y|Z$. While some kernel-based measures can also be related to information-theoretic quantities (see, e.g., Fukumizu et al. (2008)), the idea here is to *directly* estimate CMI by combining 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; Póczos and Schneider, 2012; Gao et al., 2017) with a nearest-neighbor local permutation scheme.

Their main advantage is that nearest-neighbor statistics are *locally adaptive* (Fig. 1A): The hypercubes around each sample point are smaller where more samples are available. Kernel methods, on the other hand, 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.

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, the present approach relies on a local permutation test that is also based on nearest neighbors and data-adaptive.

2 Conditional independence test

2.1 Conditional mutual information

CMI for continuous and possibly multivariate random variables X, Y, Z is defined as

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

where H denotes the Shannon entropy and where we assume that the densities $p(\cdot)$ exist. The task is to test the conditional independence hypothesis versus the general alternative:

$$H_0: \quad X \perp\!\!\!\perp Y \mid Z \qquad (3)$$
$$H_1: \quad X \not\perp\!\!\!\perp Y \mid Z\,. \qquad (4)$$

From the definition of CMI it is immediately clear that $I_{X;Y|Z} = 0$ if and only if $X \perp\!\!\!\perp Y|Z$, provided that the densities are well-defined. 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, causal discovery algorithms often make use of the test statistic's value, for example to sort the order in which conditions are tested (Spirtes et al., 2000). CMI here readily allows for an interpretation in terms of the relative importance of one condition over another. Note that the test statistic values of kernel-based tests typically depend on the chosen kernel.

2.2 Nearest-neighbor CMI estimator

From the nearest-neighbor entropy estimator by Kozachenko and Leonenko (1987), Kraskov et al. (2004) developed an estimator for mutual information that was generalized to CMI (Frenzel and Pompe, 2007; Vejmelka and Paluš, 2008) based on the CMI definition in Eq. (2):

$$\widehat{I}_{XY|Z} = \psi(k) + \frac{1}{n}\sum_{i=1}^{n} [\psi(k_i^z) - \psi(k_i^{xz}) - \psi(k_i^{yz})] \quad (5)$$

with the Digamma function as the logarithmic derivative of the Gamma function $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$ and sample length n[1]. The only free parameter k is the number

[1]The estimator is, for example, implemented in Python in `https://github.com/jakobrunge/tigramite`.

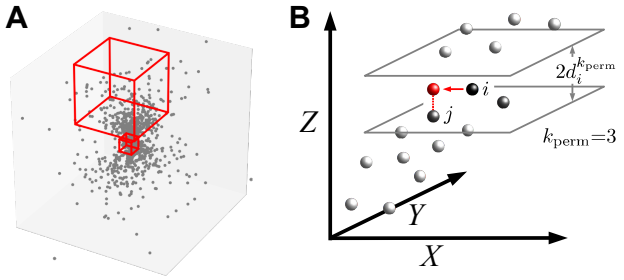

Figure 1: (A) The CMIknn estimator and the local permutation test are data-adaptive making them more data efficient than fixed bandwidth techniques: The hypercubes around each sample point are smaller where more samples are available. (B) Schematic of local permutation scheme. To destroy the dependencies between x and y, but preserve those between x and z, each sample point i's x-value is mapped randomly to one of its k_{perm}-nearest neighbors (as measured in subspace $\mathcal{Z}$). By keeping track of already 'used' indices j, we approximately achieve a random draw *without* replacement, see Algorithm 1.

of nearest neighbors in the joint space of $\mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{Z}$ which defines the local length scale (in maximum norm) ϵ_i around each sample point i. Then k_i^{xz}, k_i^{yz} and k_i^z are computed by counting the number of points with distance strictly smaller than ϵ_i (including the reference point i) in the subspace $\mathcal{X} \otimes \mathcal{Z}$ to get k_i^{xz}, in the subspace $\mathcal{Y} \otimes \mathcal{Z}$ to get k_i^{yz}, and in the subspace $\mathcal{Z}$ to get k_i^z. The decisive advantage of this estimator compared to fixed global bandwidth approaches is its local *data-adaptiveness* (Fig. 1A).

Similar estimators, but for the more general class of Rényi entropies and divergences, were developed in Wang et al. (2009); Póczos and Schneider (2012). 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. Estimator (5) 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.

Some notes on the implementation: Before estimating CMI, in our implementation the samples are rank-transformed 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 $x_1, \ldots, x_n$, x_i is replaced with the transformed value r, where r is defined such that x_i is the rth largest among all x values. This approach gave good results in our tests, other implementations may only standardize the data beforehand.

The main computational cost comes from searching nearest neighbors in the high dimensional subspaces which is $\mathcal{O}(n^2)$ in the worst case, but can be speed up using *KD-tree* neighbor search (Maneewongvatana and Mount, 1999). 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 if they are not based on kernel approximations such as via random Fourier features (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

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

Typically in CMI-based independence testing, CMI-surrogates to simulate independence are generated by randomly permuting *all* x-values in the data $\{x_i, y_i, z_i\}_{i=1}^n$. 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, here a local permutation test utilizing nearest-neighbor search is proposed. To avoid confusion, the CMI-estimation parameter is denoted as k_{CMI} and the permutation-parameter as k_{perm}.

The goal of the permutation scheme in Algorithm 1 (Fig. 1B) is to create a sample $\{x_i^*, y_i, z_i\}_{i=1}^n$, where x_i^* are drawn such that (1) marginals are preserved (drawn without replacement) and (2) x_i is replaced by x_j only if $z_i \approx z_j$ (local permutation). The proposed scheme implements this idea in a straightforward manner that is computationally not much slower than a total permutation (see Supplementary Material Fig. S1): x_i^* are drawn among nearest neighbors of the point i in subspace $\mathcal{Z}$ and a list keeps track of already 'used' samples. However, we cannot always assure draws without replacement, e.g., in the extreme case that one neighbor is shared between two points and $k_{\text{perm}} = 1$. As described in Algorithm 1, the null distribution is

Algorithm 1 Conditional independence test based on nearest-neighbor permutation (Fig. 1B).

Require: Data $\{x_i, y_i, z_i\}_{i=1}^n$, k_{perm}-nearest neighbor parameter, number of permutation surrogates B, conditional mutual information estimator (or other conditional dependence measure) $\widehat{I}(x;y|z)$

Estimate $\widehat{I}(x;y|z)$ of original data
Compute lists of nearest neighbors for each sample point i: $\mathcal{N}_i = \{l \in \{1,\dots,n\} : \|z_l - z_i\| \leq d_i^{k_{\text{perm}}}\}$ in maximum norm of subspace of z, where $d_i^{k_{\text{perm}}}$ denotes the distance of sample point z_i to its k_{perm}-nearest neighbor (including i itself, i.e., $d_i^{k_{\text{perm}}=1} = 0$); each list is then of length k_{perm}
for all $b \in \{1,\dots,B\}$ **do**
    Initialize empty list $\mathcal{U} = \{\}$ of used indices
    Initialize empty array x^* of length n
    Shuffle lists $\mathcal{N}_i$ separately for each i
    Create random permutation π of $\{1,\dots,n\}$
    for all $i \in \pi$ **do**
        $j = \mathcal{N}_i(0)$
        $m = 0$
        while $j \in \mathcal{U}$ and $m < k_{\text{perm}} - 1$ **do**
            $m = m + 1$
            $j = \mathcal{N}_i(m)$
        $x_i^* = x_j$
        Add j to $\mathcal{U}$
    Compute $\widehat{I}_b = \widehat{I}(x^*;y|z)$
Compute p-value by $p = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}[\widehat{I}_b \geq \widehat{I}(x;y|z)]$ where $\mathbf{1}$ denotes the indicator function
return p and test statistic value $\widehat{I}(x;y|z)$

estimated by applying the CMI estimator on the permutation surrogates and the p-value is derived as the fraction of surrogate CMIs larger or equal than the CMI of the original data.

The CMI estimator holds for arbitrary dimensions of all arguments X, Y, Z and also the local permutation scheme can be used to jointly permute all of X's dimensions. In the following numerical experiments, the focus is on the case of univariate X and Y and uni- or multivariate Z. Further, the permutation scheme can also be utilized with other conditional dependence measures, in the Supplementary Material we show numerical results for random-fourier-based kernel tests (Strobl et al., 2017).

As for the CMI estimator, the worst case computational complexity in sample size of this scheme, especially for very high dimensions, is $\mathcal{O}(n^2)$ for finding the nearest neighbors. The subsequent step of creating the local permutation is $\mathcal{O}(n)$ and the CMI estimate again $\mathcal{O}(n^2)$. Hence the computational complexity of the whole approach is quadratically in sample size in the worst case, but we numerically saw a more moderate scaling thanks to the use of *KD-tree* neighbor search (Maneewongvatana and Mount, 1999). The computational time mostly depends on whether the nearest neighbor search and the B-loop in Algorithm 1 are (embarrassingly) parallelized, for example on GPUs.

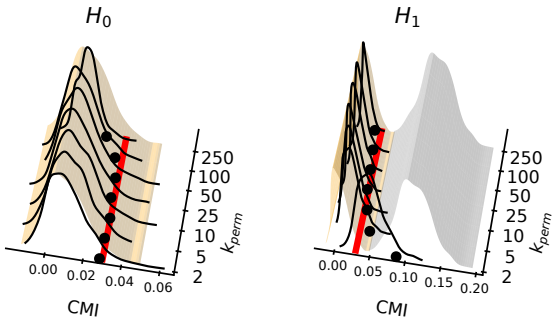

Figure 2: Illustrative simulation of multivariate Gaussian to demonstrate the effect of the nearest-neighbor permutation parameter k_{perm}. The model (6) is described in the text (here only linear dependencies were used and $D_Z = 1$, $n = 250$, $c = 1$, $k_{\text{CMI}} = 50$). (Left panel) Under H_0 the true null distribution of CMI is depicted as the orange surface (same for all k_{perm}) with the 95% quantile marked by a red straight line. The black distributions and markers give the permuted null distributions and their 95% quantiles for different k_{perm}. The sample size is $n = 250$ such that $k_{\text{perm}} = 250$ corresponds to a full non-local permutation. (Right panel) Here the true distribution under H_1 is depicted as the grey surface next to the null distribution and the permuted distributions are generated from the dependent data.

A theoretical proof that this scheme asymptotically samples from the null distribution (Type I validity) is challenging. Recent work (Sen et al., 2017) demonstrates a proof for the bootstrap case and $k_{\text{perm}} = 1$, but an extension to the permutation case with larger k_{perm} is not straightforward. Typically permutations are preferred for significance testing for smaller sample sizes which motivated our choice, especially since the CMI nearest-neighbor estimator requires non-tied samples. Additionally, larger k_{perm} yield more power as illustrated in Figs. 2,3.

3 Experiments

3.1 Choosing k_{CMI} and k_{perm}

The approach has two free parameters, k_{CMI} and k_{perm}. The following numerical experiments indicate that restricting k_{perm} to only very few nearest neighbors already suffices to reliably simulate the null distribution

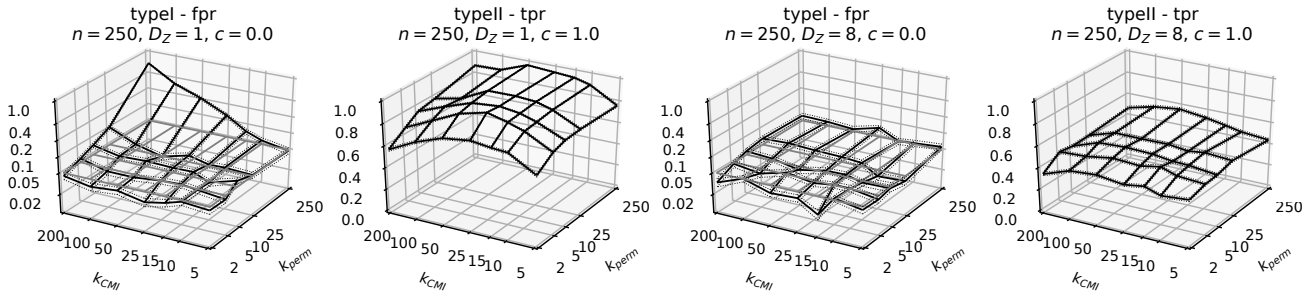

Figure 3: Choice of k_{CMI} and k_{perm} based on experiments with post-nonlinear noise model (6). The sample size is $n = 250$ and 1,000 realizations were generated to evaluate false positives (FPR) and true positives (TPR) for $c = 1$ at a 5% significance level. Shown are FPR and TPR for $D_Z = 1$ (left panels) and $D_Z = 8$ (right panels). False positives are well-calibrated for small k_{perm} and a wide range of k_{CMI} and power is also relatively robust. More sample sizes shown in Supplementary Figs. S2,S3.

while for k_{CMI} a rule-of-thumb $k_{\mathrm{CMI}} \approx 0.1..0.2n$ is proposed.

We study a post-nonlinear noise model similar to Zhang et al. (2011); Strobl et al. (2017) given by

$$X = g_X(c\epsilon_b + \epsilon_X + \tfrac{1}{D_Z}\sum\nolimits_{i=1}^{D_Z} Z_i)$$
$$Y = g_Y(c\epsilon_b + \epsilon_Y + \tfrac{1}{D_Z}\sum\nolimits_{i=1}^{D_Z} Z_i)\,, \qquad (6)$$

where $Z_i, \epsilon_b, \epsilon_X, \epsilon_Y$ 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 $X \perp\!\!\!\perp Y \mid Z = (Z_1, Z_2, \ldots)$ for $c = 0$ (H_0 true) and the dependent case $X \not\perp\!\!\!\perp Y \mid Z$ for $c \neq 0$ (H_1 true). Zhang et al. (2011); Strobl et al. (2017) used a setup where Z is independent of X and Y in the dependent case. Supplementary Fig. S6 gives results for this model setup, but in causal discovery applications these are typically dependent. Model (6) is evaluated for different sample sizes n and dimensions D_Z of the conditioning set. The null distribution for the CMI test was generated with $B = 1,000$ surrogates in all experiments and 1,000 realizations were used to evaluate performance metrics.

Figure 2 illustrates the effect of different k_{perm} for a linear-only version of the model. If k_{perm} is too large or even $k_{\mathrm{perm}} = n$, i.e., a full non-local permutation, the permuted distribution under H_0 is narrower than the true null distribution. As illustrated by the black markers, this would lead to an increase of false positives (type-I errors). On the other hand, for the dependent case under H_1, if k_{perm} is very small, the permuted distribution is positively biased yielding lower power (more type-II errors). For a range of values of k_{perm}, the permuted distribution perfectly generates the true null distribution.

Figure 3 depicts false and true positives for different k_{CMI} and k_{perm} for a sample size $n = 250$. The results indicate that a value $k_{\mathrm{perm}} \approx 5..10$ yields well-calibrated tests while not affecting power much. This holds for a wide range of sample sizes as shown in Supplementary Figs. S2,S3. Sen et al. (2017) proved type-I validity for the $k_{\mathrm{perm}} = 1$ bootstrap case in their scheme which may also serve as a very conservative option with lower power for the CMI test.

Larger k_{CMI} yield more power and even for $k_{\mathrm{CMI}} \approx n/2$ 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 suggesting a rule-of-thumb of $k_{\mathrm{CMI}} \approx 0.1..0.2n$. Note that, as shown in Supplementary Fig. S1, runtime increases linearly with k_{CMI} while k_{perm} does not impact runtime much. Here the runtime per CMI estimate is computed assuming that the scheme is (embarrassingly) parallelized.

3.2 Comparison with kernel measures and distance correlation

In Fig. 4 we investigate results comparing the CMI test (CMIknn) to a simpler clustering permutation scheme (CMIknnclust), KCIT and the two random-fourier-based approximations RCIT and RCoT introduced in Strobl et al. (2017). The Kolmogorov-Smirnov (KS) statistic as a metric for type-I errors, as in Strobl et al. (2017), is utilized to quantify how uniform the distribution of p-values is. Type-II errors are measured by the area under the power curve (AUPC). All metrics were evaluated from 1,000 realizations of the model with $c = 0.5$ and error bars give the bootstrapped standard errors. Supplementary Figs. S4,S5 depict the results for false and true positive rate metrics. CMIknn was run with $k_{\mathrm{perm}} = 5$, $B = 1,000$ permutation surrogates, and using the rule-of-thumb $k_{\mathrm{CMI}} = 0.1n$ as well as a fixed $k_{\mathrm{CMI}} = 50$.

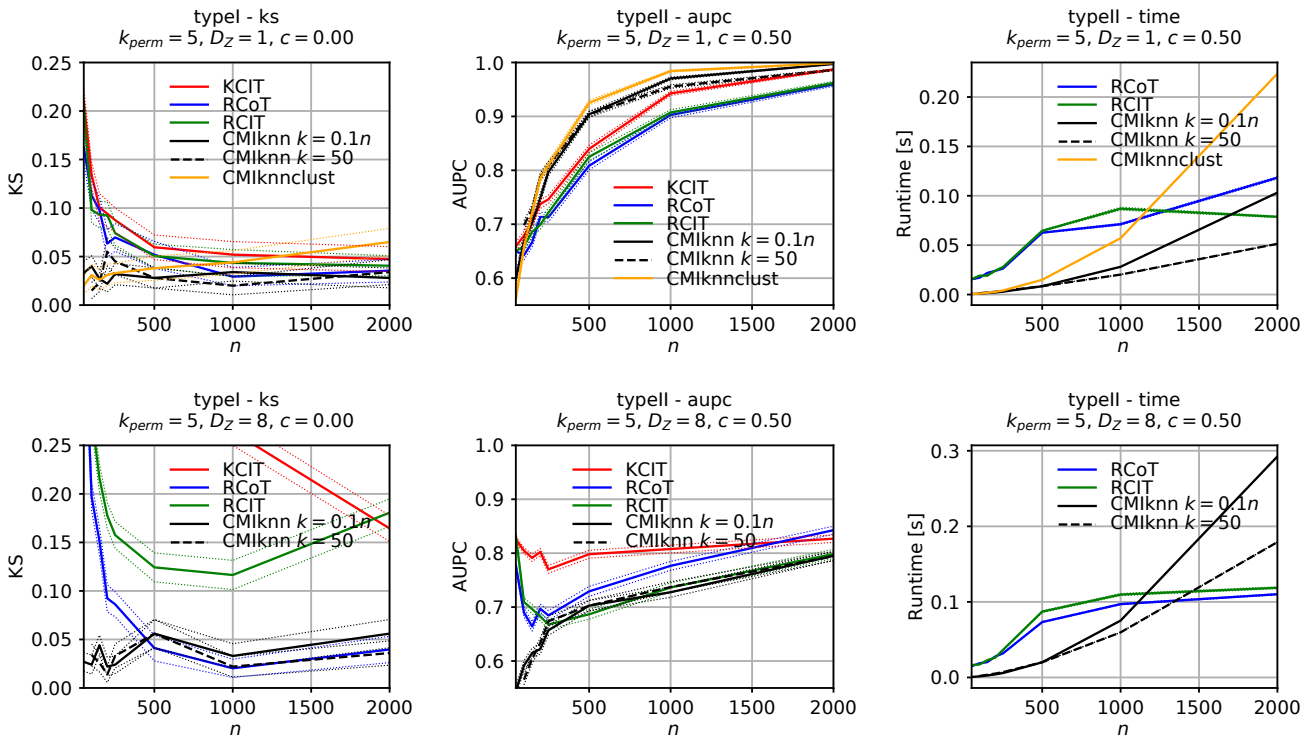

Figure 4: Comparison studies with post-nonlinear noise model (6). Shown are KS (left column), AUPC (center column), and runtime (right column) for a sample size experiment with $D_Z = 1$ (top row) and $D_Z = 8$ (bottom row). In all experiments $k_{\text{perm}} = 5$ is used together with two choices $k_{\text{CMI}} = 0.1n$ and $k_{\text{CMI}} = 50$. Here results for the default $n_{\text{ff}} = 25$ fourier features for RCIT and RCoT are shown, but calibration (false positive control) is relatively sensitive to this parameter (Supplementary Fig. S7). CMIknnclust is based on a simple equi-quantile permutation scheme (only for $D_Z = 1$). Supplementary Fig. S6 gives results for an alternative model with independent Z.

Figure 4 demonstrates that CMIknn is better calibrated with the lowest KS-values for almost all sample sizes tested. A simpler permutation scheme used only for $D_Z = 1$ (orange lines) based on shuffling x within equi-quantile bins of z (here with 20 bins) becomes ill-calibrated for larger n. While the number of bins can be adapted to the sample size, this is difficult for higher dimensional Z. KCIT and RCIT (solid red and green lines) are especially badly calibrated for smaller sample sizes or higher dimensions D_Z and RCoT (solid blue line) better approximates the null distribution only for $n > 500$. Note that this is also expected (Strobl et al., 2017) since the analytical approximation of the null distribution for RCIT and RCoT requires large sample sizes. The power as measured by AUPC is, thus, only comparable for $n > 500$ and CMIknn has the highest power throughout. For $D_Z = 8$, on the other hand, RCoT has slightly higher power than CMIknn. For the model setup with independent Z (Supplementary Fig. S6), CMIknn has again similar or higher power.

If the computationally expensive scheme of CMIknn is (embarrassingly) parallelized, the CMIknn test is faster than RCIT or RCoT for not too large sample sizes due to efficient KD-tree nearest-neighbor search procedures (Maneewongvatana and Mount, 1999), especially for smaller k_{CMI} (right column in Fig. 4). KCI is not shown here because it is orders of magnitude slower. A major computational burden of RCIT and RCoT is the kernel bandwidth computation via the median Euclidean distance heuristic. In `https://github.com/ericstrobl/RCIT` the median is computed from the first 500 samples only, leading to the "kink" in the runtime for RCIT and RCoT. The runtime of RCIT and RCoT depends quadratically on the number of random Fourier features used (here the default of 25 for subspace $\mathcal{Z}$ and 5 for subspaces $\mathcal{X}$ and $\mathcal{Y}$ was used), for more results see Fig. S7. CMIknn's runtime increases more sharply with sample size, especially for $k_{\text{CMI}} = 0.1n$.

Figure 5 explores more cardinalities of the conditioning set. KCI and RCIT are not well-calibrated for higher dimensions while RCoT and CMIknn better approximate the null. The power of CMIknn and RCoT decreases with dimensionality with CMIknn being more power-

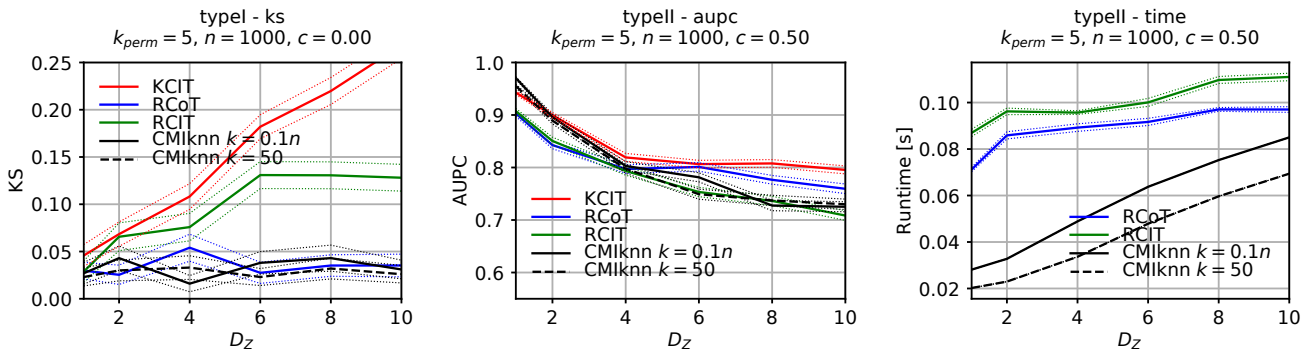

Figure 5: Numerical experiments as in Fig. 4 for different condition dimensions D_Z with fixed $n = 1000$. Supplementary Fig. S6 gives results for an alternative model with independent Z.

ful for smaller dimensions and RCoT more powerful for larger dimensions. For the model setup with independent Z (Supplementary Fig. S6), the power of both approaches is rather insensitive to dimensionality. CMIknn's runtime starts lower, but increases more sharply with D_Z than RCIT and RCoT.

The results indicate that the analytical approximations of the null distribution utilized in RCIT and RCoT do not work well for small sample sizes below $n \approx 1000$. The dashed blue and green lines in Supplementary Fig. S6 explore the option to combine the kernel statistics with the nearest-neighbor permutation test. While then RCIT and RCoT loose their computational advantage, the tests are now well-calibrated. Their power is still mostly lower than that of CMIknn, especially for RCIT.

Until now rather smooth dependencies of X and Y on the conditioning variables were considered. Figure 6 depicts a more nonlinear relationship. For an extremely oscillatory sinusoidal dependency like $X = c\epsilon_b + \sin(\lambda Z) + \epsilon_X$ and $Y = c\epsilon_b + \sin(\lambda Z) + \epsilon_Y$ for $\lambda = 30$ (all noise terms $\mathcal{N}(0, 1)$), shown in Fig. 6, k_{perm} needs to be set to a very small value in order to control false positives. CMIknn well detects dependence for this case (true positives equal 1) while the analytical versions of RCIT and RCoT do not work at all (false positives are 1) and the permutation-based versions have much lower power than CMIknn even if we vary the number of Fourier features. Note that higher values take much longer, see Fig. S7.

In Supplementary Tab. S1 the results from Wang et al. (2015) are repeated proposing the conditional distance correlation (CDC) test together with results from RCoT and the 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. CMIknn 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 CMIknn 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 is based on a computationally expensive local bootstrap scheme since the asymptotics break down for small sample sizes.

4 Real data application

CMIknn is well suited for large-scale causal discovery in combination with efficient discovery algorithms (Spirtes et al., 2000) avoiding high-dimensional conditioning by iterative tests. Here we demonstrate CMIknn together with a time series version of the PC causal discovery algorithm (Runge et al., 2017)[2] 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)[3]. The time series were detrended using a Gaussian kernel smoother with bandwidth $\sigma = 1440$ *hours* and the analysis was limited to the first three months of the dataset (2160 samples). After accounting for missing values the effective sample size is $n = 1102$. The CMIknn parameters are $k_{\text{CMI}} = 200$ and $k_{\text{perm}} = 5$ with $B = 1,000$ 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% false discovery-level shown in Fig. 7 indicates that temperature and relative humidity influence Benzene which in turn affects NO2 and CO concentrations. The edge colors depict the CMI test statistic values which demonstrates another advantage of information-theoretic measures: The CMI value

[2]Software is available online under `https://github.com/jakobrunge/tigramite`.

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

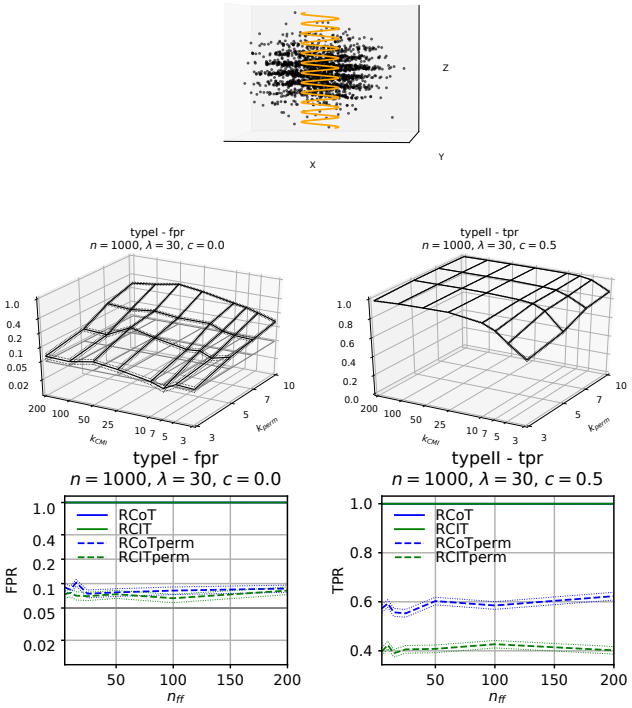

Figure 6: Example of sinusoidal dependence leading to strongly oscillatory structure (top panel for $\lambda = 10$). The two center panels depict FPR and TPR at $\alpha = 0.05$ for CMIknn and the two bottom panels for RCIT and RCoT for different numbers of random fourier features n_{ff}. Here the analytical versions of RCIT and RCoT (solid lines) do not work at all (FPR equal to 1). If RCIT and RCoT are combined with the local permutation test (dashed lines) for $k_{\text{perm}} = 5$, they become better calibrated, but still have much lower power than CMIknn.

is well-interpretable as the magnitude of information flow between components of the system.

5 Conclusion

The paper presents 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. This feature results in well-calibrated tests with reliable false positive rates. Numerical experiments for sample sizes $n = 50$ to $n = 2,000$ and dimensions of the conditional set of $D_Z = 1..10$ were evaluated. The experiments suggest that a permutation nearest-neighbor parameter of $k_{\text{perm}} = 5$ provides well-calibrated tests while not affecting power much, but smaller values can serve as

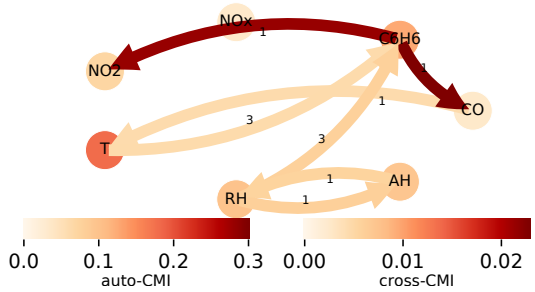

Figure 7: Causal discovery with CMIknn and the algorithm proposed in Runge et al. (2017) on time series of air pollutants and various weather variables. The node color gives the strength of auto-CMI, i.e., the lag-1 CMI of a variable with itself, and the edge color the cross-CMI with the link labels denoting the time lag in *hours*. Significance based on 10% false discovery-level.

a more conservative option. The power of CMIknn was found higher than advanced kernel based tests such as KCIT or its faster random Fourier feature versions RCIT and RCoT for smaller dimensions and similar for higher dimensions. CMIknn is preferable especially for highly non-smooth densities due to its local data-adaptiveness. For not too large sample sizes CMIknn has a shorter runtime since efficient nearest-neighbor search schemes can be utilized, but its runtime increases more sharply with sample size and dimensionality than the fourier-feature based kernel tests. Here approximate nearest-neighbor techniques could speed up computation. The permutation scheme leads to a higher computational load which, however, can be easily parallelized, for example on GPUs. Nevertheless, more theoretical research is desirable to obtain approximate analytics for the null distribution in the large sample limit. For small sample sizes below $n \approx 1,000$ our results demonstrate that a permutation-approach is inevitable also for kernel-based approaches.

Acknowledgments

The author thanks Eric Strobl for kindly providing R-code and Dino Sejdinovic for many helpful comments. The author gratefully acknowledges the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research and the Land Brandenburg for providing resources on the high-performance computer system at the Potsdam Institute for Climate Impact Research.

References

Bergsma, W. (2004). Testing conditional independence for continuous random variables. *Eurandom technical report*, 48:1–19.

Daudin, J. J. (1980). Partial association and an measures to qualitative application regression. *Biometrika*, 67(3):581–590.

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.

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.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008). Kernel Measures of Conditional Dependence. *Advances in Neural Information Processing Systems 21 (2008)*, 20:1–13.

Gao, W., Oh, S., and Viswanath, P. (2017). Demystifying fixed k-nearest neighbor information estimators. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1267–1271. IEEE.

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):16.

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.

Maneewongvatana, S. and Mount, D. (1999). It's okay to be skinny, if your friends are fat. *Center for Geometric Computing 4th Annual Workshop on Computational Geometry*, pages 1–8.

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

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2013). Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research*, 15(1):2009–2053.

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

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

Runge, J., Sejdinovic, D., and Flaxman, S. (2017). Detecting causal associations in large nonlinear time series datasets. *http://arxiv.org/abs/1702.07007*.

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

Sen, R., Suresh, A. T., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. (2017). Model-Powered Conditional Independence Test. In *Advances in Neural Information Processing Systems 30 (2017)*, pages 2955–2965.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. 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. *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813.

Zhang, Q., Filippi, S., Flaxman, S., and Sejdinovic, D. (2017). Feature-to-feature regression for a two-step conditional independence test. In *Proceedings*

of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017).