Increasing Effect Sizes of Pairwise Conditional Independence Tests between Random Vectors

Abstract

A simple approach to test for conditional independence of two random vectors given a third random vector is to simultaneously test for conditional independence of every pair of components of the two random vectors given the third random vector. In this work, we show that conditioning on additional components of the two random vectors that are independent given the third one increases the tests' effect sizes while leaving the validity of the overall approach unchanged. Up to the effective reduction of the sample size due to enlarging the conditioning sets, these larger effect sizes lead to higher statistical power. We leverage this result to derive a practical pairwise testing algorithm that first chooses tests with a relatively large effect size and then does the actual testing. In simulations, our algorithm outperforms standard pairwise independence testing and other existing methods if the dependence within the two random vectors is sufficiently high.

1 INTRODUCTION

Let $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ be real-valued random vectors. We are interested in testing whether $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent given $\boldsymbol{Z}$. This task arises in numerous research areas such as ecology [Legendre and Legendre, 2012], genetics [Piepho, 2005], Earth sciences [Runge et al., 2019] or causal discovery [Spirtes et al., 2000, Peters et al., 2017] and is statistically much more difficult than unconditional independence testing [Bergsma, 2004, Shah and Peters, 2020].

In this paper, we consider the case where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are multivariate. This case is of high practical relevance and occurs, for instance, when a researcher assorts variables to different groups based on semantic proximity. For example, she might want to determine whether two sectors in the economy behave independently on the stock market given certain external influences. We also envision that this multivariate case becomes relevant for vector-valued causal inference, where one is interested in causal relations between groups of variables and not between individual variables [Wahl* et al., 2022]. There are several conditional independence tests that work in this multivariate setting (for an overview on such tests see Chatterjee [2022], Josse and Holmes [2016] and Li and Fan [2020]). Generally speaking, one can split such tests into two groups: First, tests that directly incorporate the multivariate nature of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Second, tests that are based on aggregating the univariate test statistics corresponding to the pairs of components X_i and Y_j.

A relatively old representative of the first group is the partial Mantel test [Smouse et al., 1986], whose underlying test statistic is the partial correlation between the vectorized distance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ controlled for the vectorized distance matrix of $\boldsymbol{Z}$. A more recent example is the partial distance correlation test from Székely and Rizzo [2014]. This test is based on projecting (suitably centered) distance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ onto the orthogonal complement of the (suitably centered) distance matrix of $\boldsymbol{Z}$ and then calculating a certain scalar product with respect to both these projections.

Other representatives of the first group measure the distance between conditional distributions or quantities derived therefrom. For example, some representatives use the conditional mutual information [Runge, 2018], the Hellinger distance [Su and White, 2008] or the smoothed empirical likelihood ratio [Su and White, 2014]; another approach employs conditional characteristic functions [Su and White, 2007].

Kernel-based approaches constitute another important class of examples in the first group. Fukumizu et al. [2007] suggest to use the Hilbert-Schmidt norm of the normalized conditional cross-covariance operator. Zhang et al. [2011] propose a simple test based on the kernel matrices of $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ which they call the kernel-based conditional independence test (KCIT). Strobl et al. [2019] propose speed-ups

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

of the KCIT.

For tests in the second group, that is, tests that are based on aggregating univariate test statistics, two representatives are the generalized covariance measure [Shah and Peters, 2020] and its weighted extension[Scheidegger et al., 2022]. In both instances, the authors first introduce the respective dependence measures for univariate X and Y. The main idea behind both univariate dependence measures is to first regress X onto $\boldsymbol{Z}$ and Y onto $\boldsymbol{Z}$ using a user-defined regression method and to then calculate a covariance-like measure between the residuals of both regressions. For multivariate $\boldsymbol{X}$ and $\boldsymbol{Y}$, the authors then propose to aggregate the respective measures for every pairs of components X_i and Y_j conditioned on $\boldsymbol{Z}$.

The pairwise approach employed by the second group of tests has several advantages: Firstly, it allows to easily construct multivariate tests using univariate test only; in particular, one can use classical ideas from the multiple testing literature to control the probability of a false positive. Secondly, the pairwise approach is flexible and allows for a wide variety of univariate test statistics. Thirdly, pairwise testing is fast if the employed univariate tests are fast.

In this paper, we further investigate the pairwise approach. We propose a new pairwise conditional independence testing procedure in which the conditioning vectors $\boldsymbol{Z}$ are enlarged by components of $\boldsymbol{X}$ and $\boldsymbol{Y}$ that are conditionally independent given $\boldsymbol{Z}$. We show that this new approach yields larger effect sizes, that is, larger underlying dependence measures than simple pairwise conditional independence testing and hence, if the dimensionality of the enlarged conditioning sets is small in comparison to the sample size, more statistical power.

We structure the paper as follows. In Section 2 we review the notion of conditional independence. In Section 3 we discuss several pairwise approaches including our novel approach. In Section 4 we give a theoretical justification for our new approach. In Section 5 we present numerical experiments and in Section 6 we provide a short summary and outlook.

2 PRELIMINARIES

In this section, we introduce our notation and review the definition of conditional independence. Moreover, we review some elementary properties of conditional independence.

2.1 NOTATION

Let $\boldsymbol{X} := (X_1, \dots, X_{d_X})$, $\boldsymbol{Y} := (Y_1, \dots, Y_{d_Y})$ and $\boldsymbol{Z} := (Z_1, \dots, Z_{d_Z})$ denote d_X-, d_Y-, and d_Z-dimensional real-valued random vectors, respectively. For any set of indices $A \subseteq \{1, \dots, d_X\}$, we write A^c to denote the complement of A in $\{1, \dots, d_X\}$ and $|A|$ to denote the number of indices in A. Moreover, we write $\boldsymbol{X}_A$ to denote the vector that only consists of components of $\boldsymbol{X}$ whose indices are contained in A; we use similar notations for $\boldsymbol{Y}$ and $\boldsymbol{Z}$.

Following Kim et al. [2022] and Neykov et al. [2021], let $P_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}}$ denote the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. Similarly, let $P_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z}}$ denote the conditional distribution of $(\boldsymbol{X}, \boldsymbol{Y}) \mid \boldsymbol{Z} = \boldsymbol{z}$, and let $P_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P_{\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z}}$ stand for the conditional distributions of $\boldsymbol{X} \mid \boldsymbol{Z} = \boldsymbol{z}$ and $\boldsymbol{Y} \mid \boldsymbol{Z} = \boldsymbol{z}$ respectively. Furthermore, let $P_{\boldsymbol{X}}, P_{\boldsymbol{Y}}$ and $P_{\boldsymbol{Z}}$ denote the marginal distributions of $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ respectively. We assume that $P_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}}$ is absolutely continuous with respect to the Lebesgue measure, and we write $p_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}}$ for the corresponding density; we denote the densities corresponding to the other distributions in an analogous way.

2.2 MULTIVARIATE CONDITIONAL INDEPENDENCE

We say that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent given $\boldsymbol{Z}$ and denote this fact by

$$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \tag{1}$$

if and only if

$$p_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{x}, \boldsymbol{y}) = p_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{x}) \cdot p_{\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{y}) \tag{2}$$

for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ such that $p_{\boldsymbol{Z}}(\boldsymbol{z}) > 0$ [Dawid, 1979]. To express the negation of statement (1), we write $\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$.

In the following, we review some properties of conditional independence which are useful in the context of this work (see e.g., Pearl [2009]). For any set of indices $B \subseteq \{1, \dots, d_Y\}$, the following properties are valid:

- **Decomposition:** $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \Longrightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}_B \mid \boldsymbol{Z}$.
- **Contraction:** $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}_B \mid \boldsymbol{Z}$ & $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}_{B^c} \mid (\boldsymbol{Z}, \boldsymbol{Y}_B) \Longrightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$.
- **Weak Union:** $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \Longrightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}_{B^c} \mid (\boldsymbol{Z}, \boldsymbol{Y}_B)$

These three properties allow us to decompose the multivariate conditional independence statement from (1) into several univariate conditional independence statements. For example, applying the decomposition property to statement (1) always gives

$$X_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Z} \qquad \forall i \in \{1, \dots, d_X\},\ \forall j \in \{1, \dots, d_Y\}. \tag{3}$$

It is well known that the reverse implication does not necessarily hold (that is, statement (3) does not necessarily imply statement (1)). However, there are assumptions under which the reverse implication does indeed hold. We discuss two such assumptions in Sections A.2 and A.3 of the Supplementary Material (SM).

3 PAIRWISE INDEPENDENCE TESTING WITH INCREASED EFFECT SIZES

In this section, we first present the classical pairwise independence testing approach as, for example, used by Shah and Peters [2020], and then, we introduce our novel approach. We discuss our approach both with and without the assumption that some conditional independencies $X_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Z}$ are known a priori.

3.1 STANDARD PAIRWISE INDEPENDENCE TESTING

In the finite sample setting, we assume to have n independent observations $(\boldsymbol{X}^{(1)}, \boldsymbol{Y}^{(1)}, \boldsymbol{Z}^{(1)}), \ldots, (\boldsymbol{X}^{(n)}, \boldsymbol{Y}^{(n)}, \boldsymbol{Z}^{(n)})$, where each observation is distributed according to the unknown distribution $P_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}}$. Our goal is to statistically test whether $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ is true or false. That is, we perform the hypothesis test

$$\mathcal{H}_0 : \ \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \ \text{ vs. } \ \mathcal{H}_1 : \ \boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \tag{4}$$

using the n observations. Here, $\mathcal{H}_0$ is the null hypothesis and $\mathcal{H}_1$ is the alternative hypothesis. To execute this hypothesis test, one can first do the similar hypothesis tests

$$\begin{aligned} &\mathcal{H}_0' : \ \forall i, j : \ X_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Z} \ \text{ vs.} \\ &\mathcal{H}_1' : \ \exists i, j : \ X_i \not\perp\!\!\!\perp Y_j \mid \boldsymbol{Z}. \end{aligned} \tag{5}$$

and reject $\mathcal{H}_0$ if and only if one rejects $\mathcal{H}_0'$.

This "induced" test for $\mathcal{H}_0$ has valid level $\alpha \in (0, 1)$ if the original test has valid level α.

Lemma 1. *If the test corresponding to $\mathcal{H}_0'$ has valid level $\alpha \in (0, 1)$ at sample size n, then the induced test for $\mathcal{H}_0$ has valid level α at sample size n. This result is also true in a pointwise asymptotic and uniformly asymptotic sense (see Section C.1 of the SM for more precise formulations of these two notions).*

Proof. The intuition is as follows (see Section C.1 of the SM for the details): If $\mathcal{H}_0$ is true but rejected, then $\mathcal{H}_0'$ is true (by the discussion in Section 2.2) and had been rejected (by the definition of the "induced" test). Thus, every type I error with respect to $\mathcal{H}_0$ is a type I error with respect to $\mathcal{H}_0'$. □

To obtain a test for $\mathcal{H}_0'$ that has valid level, one can first calculate and then aggregate all univariate test statistics T_{ij} with $i \in \{1, \ldots, d_X\}$ and $j \in \{1, \ldots, d_Y\}$ that correspond to the null hypotheses

$$\mathcal{H}_0^{(ij)} : X_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Z}.$$

To aggregate these test statistics one can use ideas from the multiple testing literature. For example, one can apply the Bonferroni method to control the familywise error rate of all the tests induced by the T_{ij}'s. One can then reject $\mathcal{H}_0'$ if at least one of the tests induced by the T_{ij}'s has been rejected at the adjusted significance level. This test for $\mathcal{H}_0'$ has valid level α if the familywise error rate of the tests induced by the T_{ij}'s has been bounded by α. Instead of the Bonferroni method, one can also define a meta test statistic by taking the maximum of the absolute values of the T_{ij}'s. To control the probability of false positives, one can, for instance, use analytical results [Nadarajah et al., 2019] or a multiplier bootstrap [Chernozhukov et al., 2013, Shah and Peters, 2020].

3.2 NOVEL PAIRWISE APPROACH WITH A PRIORI KNOWN CONDITIONAL INDEPENDENCIES

In this section, we present our novel approach to multivariate independence testing which is based on a modified version of $\mathcal{H}_0'$. The main idea behind this modification is to enlarge the conditioning sets by those components of $\boldsymbol{X}$ and $\boldsymbol{Y}$ that we know to be independent given $\boldsymbol{Z}$. The rationale behind this idea is that these extra conditions increase the effect sizes of the remaining tests. We defer this result to Section 4 and here only provide a glimpse of it in Example 1.

To fix notation, let $S(X_i)$ contain all indices corresponding to the components of $\boldsymbol{Y}$ that are independent of X_i given $\boldsymbol{Z}$, i.e.[1],

$$S(X_i) := \{j \in \{1, \ldots, d_Y\} : X_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Z}\}.$$

Similarly,

$$S(Y_j) := \{i \in \{1, \ldots, d_X\} : X_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Z}\}.$$

Furthermore, assume that we have a priori knowledge of arbitrary but fixed subsets $Q_i \subseteq S(X_i)$ and $Q_j' \subseteq S(Y_j)$ for all $i \in \{1, \ldots, d_X\}$ and $j \in \{1, \ldots, d_Y\}$. These subsets are allowed to be empty. However, if all of them are empty, then our proposed approach is the same as the one from Section 3.1.

As Propositions 1 and 2 show, additionally conditioning on $\boldsymbol{Y}_{Q_i \setminus \{j\}}$ or $\boldsymbol{X}_{Q_j' \setminus \{i\}}$ increases the effect size of the test $X_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Z}$. Because of this result, we propose to replace the hypothesis test in (5) with

$$\begin{aligned} &\mathcal{H}_0'' : \ \forall i, j : \ X_i \perp\!\!\!\perp Y_j \mid (\boldsymbol{Z}, \boldsymbol{S_{ij}}) \ \text{ vs.} \\ &\mathcal{H}_1'' : \ \exists i, j : \ X_i \not\perp\!\!\!\perp Y_j \mid (\boldsymbol{Z}, \boldsymbol{S_{ij}}), \end{aligned} \tag{6}$$

where $\boldsymbol{S_{ij}}$ is depending on the user's choice either equal

[1] For simplified notation, we do not include $\boldsymbol{Z}$ and $\boldsymbol{Y}$ (or $\boldsymbol{X}$) in the notation "$S(X_i)$" (respectively "$S(Y_j)$").

to $\boldsymbol{Y}_{Q_i \setminus \{j\}}$ or $\boldsymbol{X}_{Q'_j \setminus \{i\}}$.[2]. To choose between $\boldsymbol{Y}_{Q_i \setminus \{j\}}$ or $\boldsymbol{X}_{Q'_j \setminus \{i\}}$, we suggest to take the vector with the larger number of components.

As before, we propose to reject $\mathcal{H}_0$ if and only if we reject $\mathcal{H}''_0$. This new "induced" test for $\mathcal{H}_0$ again has valid level $\alpha \in (0,1)$ if the test for $\mathcal{H}''_0$ has valid level α.

Lemma 2. *If the test corresponding to $\mathcal{H}''_0$ has valid level $\alpha \in (0,1)$ at sample size n, then the induced test for $\mathcal{H}_0$ has valid level α at sample size n. This result is again true in a pointwise asymptotic and uniformly asymptotic sense.*

Proof. The proof is similar to the one of Lemma 1. We just need to show that $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ implies $X_i \perp\!\!\!\perp Y_j \mid (\boldsymbol{Z}, \boldsymbol{S_{ij}})$ for all $i \in \{1, \dots, d_X\}$ and $j \in \{1, \dots, d_Y\}$. For that, let $i \in \{1, \dots, d_X\}$ and $j \in \{1, \dots, d_Y\}$ be arbitrary but fixed indices. Without loss of generality, let $\boldsymbol{S_{ij}} = \boldsymbol{Y}_{Q_i \setminus \{j\}}$. Now, rewriting $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ to $\boldsymbol{X} \perp\!\!\!\perp (\boldsymbol{Y}_{Q_i \setminus \{j\}}, \boldsymbol{Y}_{(Q_i \setminus \{j\})^c}) \mid \boldsymbol{Z}$, we can use the weak-union property (see Section 2.2) to infer that $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}_{(Q_i \setminus \{j\})^c} \mid (\boldsymbol{Z}, \boldsymbol{Y}_{Q_i \setminus \{j\}})$, which by the decomposition property implies that $X_i \perp\!\!\!\perp \boldsymbol{Y}_{(Q_i \setminus \{j\})^c} \mid (\boldsymbol{Z}, \boldsymbol{Y}_{Q_i \setminus \{j\}})$. Because $j \in (Q_i \setminus \{j\})^c$, we can apply the decomposition property again and obtain that $X_i \perp\!\!\!\perp Y_j \mid (\boldsymbol{Z}, \boldsymbol{Y}_{Q_i \setminus \{j\}})$. As i and j were arbitrary, we obtain the result. (For more details, see Section C.2 in the SM). □

To obtain a test with valid level $\alpha \in (0,1)$ for $\mathcal{H}''_0$, we suggest to use the same techniques that we discussed in Section 3.1; so again, one may use the Bonferroni method or the maximum absolute test statistic (or something else).

In the following example, we illustrate our new approach and sketch why it leads to larger effect sizes.

Example 1. *Let $(X_1, Y_1, Y_2, \boldsymbol{Z})$ have a multivariate normal distribution with univariate components X_1, Y_1, Y_2 and a possibly multivariate $\boldsymbol{Z}$. Assume that $X_1 \perp\!\!\!\perp Y_1 \mid \boldsymbol{Z}$ holds and suppose that we want to test $X_1 \perp\!\!\!\perp (Y_1, Y_2) \mid \boldsymbol{Z}$. The usual pairwise approach would calculate and aggregate test statistics corresponding to*

$$X_1 \perp\!\!\!\perp Y_1 \mid \boldsymbol{Z} \quad \& \quad X_1 \perp\!\!\!\perp Y_2 \mid \boldsymbol{Z},$$

while we propose to use test statistics corresponding to

$$X_1 \perp\!\!\!\perp Y_1 \mid \boldsymbol{Z} \quad \& \quad X_1 \perp\!\!\!\perp Y_2 \mid (\boldsymbol{Z}, Y_1).$$

If one uses test statistics based on the partial correlation, then the corresponding effect sizes are indeed larger in our approach because

$$\begin{aligned} |\rho_{X_1 Y_2 | \boldsymbol{Z}, Y_1}| &= \left| \frac{\rho_{X_1 Y_2 | \boldsymbol{Z}} - \overbrace{\rho_{X_1 Y_1 | \boldsymbol{Z}}}^{=0} \rho_{Y_1 Y_2 | \boldsymbol{Z}}}{\underbrace{\sqrt{1 - \rho^2_{X_1 Y_1 | \boldsymbol{Z}}}}_{=1} \sqrt{1 - \rho^2_{Y_1 Y_2 | \boldsymbol{Z}}}} \right| \\ &= \left| \frac{\rho_{X_1 Y_2 | \boldsymbol{Z}}}{\sqrt{1 - \rho^2_{Y_1 Y_2 | \boldsymbol{Z}}}} \right| \\ &\geq |\rho_{X_1 Y_2 | \boldsymbol{Z}}|. \end{aligned}$$

Note that this increase of the effect size is particularly strong if $\rho_{Y_1 Y_2 | \boldsymbol{Z}}$, i.e., the within-$\boldsymbol{Y}$ correlation, is large. Moreover, note that the sample size for our approach has effectively decreased by just one.

Intuitively, our approach conditions away the dependence between Y_1 and Y_2 given $\boldsymbol{Z}$, which would otherwise overlay the dependence between X_1 and Y_2 given $\boldsymbol{Z}$, and which would hence make the dependence between X_1 and Y_2 given $\boldsymbol{Z}$ harder to detect.

3.3 NOVEL PAIRWISE APPROACH WITHOUT A PRIORI KNOWN CONDITIONAL INDEPENDENCIES

In this section, we extend the idea from Section 3.2 to the case where one does not assume a priori knowledge of subsets $Q_i \subseteq S(X_i)$ and $Q'_j \subseteq S(Y_j)$. In this case, we propose the following two-step procedure:

- Step 1: Estimate $S(X_i)$ and $S(Y_j)$ for all $i \in \{1, \dots, d_X\}$ and for all $j \in \{1, \dots, d_Y\}$. Denote the estimates by $\hat{S}(X_i)$ and $\hat{S}(Y_j)$.
- Step 2: Execute the procedure from Section 3.2 with input $Q_i = \hat{S}(X_i)$ for all $i \in \{1, \dots, d_X\}$ and $Q'_j = \hat{S}(Y_j)$ for all $j \in \{1, \dots, d_Y\}$.

As before, we propose to reject $\mathcal{H}_0$ if and only if we reject $\mathcal{H}''_0$ in Step 2 with the $\hat{S}(X_i)$'s and $\hat{S}(Y_j)$'s as input. We again obtain a result on the level of this new "induced" test.

Lemma 3. *If for each possible input of Q_i's $\subseteq \{1, \dots, d_Y\}$ and Q'_j's $\subseteq \{1, \dots, d_X\}$ for Step 2 the corresponding test has valid level $\alpha \in (0,1)$ for fixed sample size n conditioned on the fact that the Q_i's and Q'_j's have been selected in Step 1 (for a precise notion of this conditioning see Section C.3 of the SM), then the induced test for $\mathcal{H}_0$ has valid level α for fixed sample size n.*

In particular, if one splits the sample between Step 1 and Step 2, and the test in Step 2 based on the second part of the sample has valid level $\alpha \in (0,1)$, then the induced test for $\mathcal{H}_0$ has valid level α for the entire dataset of size n.

Proof. The proof is similar to the one of Lemma 2. First of all, we note that the goodness of the estimates $\hat{S}(X_i)$ and

[2]We do not allow that $\boldsymbol{S_{ij}} = \boldsymbol{Y}_{Q_i \setminus \{j\}} \cup \boldsymbol{X}_{Q'_j \setminus \{i\}}$, as this invalidates our theoretical reasoning behind increasing effect sizes in Section 4

$\hat{S}(Y_j)$ does not matter for controlling the type I error rate (it matters for increasing effect sizes, however). That means, it does not matter whether the $\hat{S}(X_i)$'s and $\hat{S}(Y_j)$'s are indeed subsets of the $S(X_i)$'s respectively $S(Y_j)$'s. To see this relaxation, we just need to realize that the proof of Lemma 2 works for general sets Q_i and Q'_j; the definitions, namely that the Q_i's respectively Q'_j's are subsets of the $S(X_i)$'s respectively $S(Y_j)$'s, were never used in that proof. We defer the other technical details including the part regarding the conditioning to Section C.3 of the SM. □

It is necessary to condition away the fact that particular Q_i's and Q'_j's have been selected in Step 1, because otherwise we would run into a typical example of selective inference. We would then test hypotheses that were already deemed promising, and not adjusting for this selection-effect invalidates classical error bounds. Conditioning away the selection step generally makes the requirements on the second step stricter. A common approach to meet these requirements is sample splitting. There are other, more elaborate approaches than sample splitting, e.g., data carving [Fithian et al., 2014], or approaches based on differential privacy [Dwork et al., 2015]. However, we consider it out of scope to develop these approaches here.

Even though the goodness of the estimates for the $S(X_i)$'s and $S(Y_j)$'s does not matter for obtaining a test for $\mathcal{H}_0$ with valid level $\alpha \in (0,1)$ (see the proof of Lemma 3), it does matter for increasing the effect sizes. If a particular $\hat{S}(X_i)$ or $\hat{S}(Y_j)$ contains indices that are not an element of the respective $S(X_i)$ or $S(Y_j)$, then the results on increasing the effect size are not necessarily true anymore. From that perspective it is, however, not a problem if there is an $\hat{S}(X_i)$ or $\hat{S}(Y_j)$ that is a strict subset of the respective $S(X_i)$ or $S(Y_j)$ because the framework in Section 3.2 is specifically designed for subsets $Q_i \subseteq S(X_i)$ and $Q'_j \subseteq S(Y_j)$.

For the estimation in Step 1 several approaches are possible. We suggest to estimate the $S(X_i)$'s and $S(Y_j)$'s by testing all conditional independencies $X_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Z}$ on one part of the sample at a rather large significance level α_{pre}. Then, we write all indices corresponding to hypotheses that are not rejected (here really understood as accepted) into the corresponding sets $\hat{S}(X_i)$ and $\hat{S}(Y_j)$. Specifically, if $X_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Z}$ is not rejected for some fixed i and j, then we write j into $\hat{S}(X_i)$ and i into $\hat{S}(Y_j)$.

A large significance level α_{pre} reduces the probability of type II errors (if there is dependence), but, it increases the probability of type I errors (if there is no dependence). However, type I errors are not a problem because they only make the $\hat{S}(X_i)$'s respectively the $\hat{S}(Y_j)$'s strictly smaller than the respective $S(X_i)$'s or the respective $S(Y_j)$'s; in the "worst" case, the $\hat{S}(X_i)$'s and $\hat{S}(Y_j)$'s are empty sets. Type II errors are a problem, though, as they lead to indices being wrongly included in the $\hat{S}(X_i)$'s and $\hat{S}(Y_j)$'s.

To obtain a test for Step 2 that has valid level $\alpha \in (0,1)$ for the second part of the sample, we propose to apply the same techniques as in Section 3.1 and Section 3.2 on the second part of the sample.

4 THEORETICAL JUSTIFICATION

The test from Section 3.2 assumes that certain conditional independencies are known a priori. The variables corresponding to these conditional independencies are then conditioned out in the remaining conditional independence tests. In the following Proposition, we show that doing so increases the respective partial correlations.

Proposition 1. *For any set of indices $Q_i \subseteq S(X_i)$,*

$$|\rho_{X_iY_j|\boldsymbol{Z},\boldsymbol{Y}_{Q_i\setminus\{j\}}}| \geq |\rho_{X_iY_j|\boldsymbol{Z}}|.$$

Similarly, for any set of indices $Q'_j \subseteq S(Y_j)$,

$$|\rho_{X_iY_j|\boldsymbol{Z},\boldsymbol{X}_{Q'_j\setminus\{i\}}}| \geq |\rho_{X_iY_j|\boldsymbol{Z}}|.$$

Proof. If Q_i (or respectively Q'_j) is a singleton, then the proof is the same as the calculation given in Example 1. For a larger Q_i (respectively Q'_j), one can prove the statement using the recursion formula for partial correlations, see Section C.4 of the SM for details. □

The partial correlation is one particular underlying dependence measure and its absolute value is one particular example of an effect size. To generalize Proposition 1 to other dependence measures, we use the notion of conditional mutual information (see Section A.1 of the SM for a review). The conditional mutual information quantifies the entire dependence structure of random vectors, is nonnegative and equal to zero if and only if conditional independence holds. Therefore, an increased conditional mutual information indicates that other well-chosen dependence measures should also increase. In particular, the following results generalize to dependence measures that are monotonically increasing functions of the conditional mutual information.

To be able to proof a generalization of Proposition 1, we make the following assumption.

Assumption 1. *For all $A \subseteq \{1,\dots,d_X\}$ and $B \subseteq \{1,\dots,d_Y\}$, we assume that*

$$\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{Y}_B \mid \boldsymbol{Z}$$

is equivalent to

$$X_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Z} \qquad \forall i \in A,\ \forall j \in B.$$

In Section A.2 and Section A.3 of the SM, we show that Assumption 1 holds if $P_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}}$ is multivariate normal or if $P_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}}$ is faithful and globally Markov with respect to an underlying directed acyclic graph.

Proposition 2. *Let Assumption 1 hold. Then, for any set of indices $Q_i \subseteq S(X_i)$,*

$$I(X_i; Y_j | \boldsymbol{Z}, \boldsymbol{Y}_{Q_i \setminus \{j\}}) \geq I(X_i; Y_j | \boldsymbol{Z}).$$

Similarly, for any set of indices $Q'_j \subseteq S(Y_j)$,

$$I(X_i; Y_j | \boldsymbol{Z}, \boldsymbol{X}_{Q'_j \setminus \{i\}}) \geq I(X_i; Y_j | \boldsymbol{Z}).$$

Proof. We only prove the statement for any arbitrary but fixed $Q_i \subseteq S(X_i)$, the proof for any arbitrary but fixed $Q'_j \subseteq S(Y_j)$ is analogous.
Write $S(X_i) \setminus \{j\} = \{j_1, \ldots, j_m\}$, where m is a natural number such that $1 \leq m \leq d_Y - 1$. Without loss of generality (as we can relabel the elements $j_1, \ldots, j_m$ arbitrarily), we prove the statement for all sets $\{j_1, \ldots, j_k\} \subseteq S(X_i)$ for all $1 \leq k \leq m$. Now, by applying the chain rule for conditional mutual information, we obtain that

$$\begin{aligned} &I(X_i; Y_j, Y_{j_1}, \ldots, Y_{j_k} | \boldsymbol{Z}) \\ &= I(X_i; Y_j | \boldsymbol{Z}) + I(X_i; Y_{j_1} | \boldsymbol{Z}, Y_j) \\ &\quad + \ldots + I(X_i; Y_{j_k} | \boldsymbol{Z}, Y_j, Y_{j_1} \ldots, Y_{j_{k-1}}) \\ &\geq I(X_i; Y_j | \boldsymbol{Z}) \end{aligned} \tag{7}$$

because the conditional mutual information is always non-negative. Similarly, by applying the chain rule the other way round, we obtain

$$\begin{aligned} &I(X_i; Y_j, Y_{j_1}, \ldots, Y_{j_k} | \boldsymbol{Z}) \\ &= I(X_i; Y_{j_1} | \boldsymbol{Z}) + I(X_i; Y_{j_2} | \boldsymbol{Z}, Y_{j_1}) \\ &\quad + \ldots + I(X_i; Y_j | \boldsymbol{Z}, Y_{j_1}, \ldots, Y_{j_k}) \\ &= I(X_i; Y_{j_1}, \ldots, Y_{j_k} | \boldsymbol{Z}) + I(X_i; Y_j | \boldsymbol{Z}, Y_{j_1}, \ldots, Y_{j_k}) \\ &= I(X_i; Y_j | \boldsymbol{Z}, Y_{j_1}, \ldots, Y_{j_k}). \end{aligned} \tag{8}$$

because $X_i \perp\!\!\!\perp Y_{j_l} \mid \boldsymbol{Z}$ for all $l \in \{1, \ldots, k\}$ and hence by Assumption 1 that $X_i \perp\!\!\!\perp Y_{j_1}, \ldots, Y_{j_k} \mid \boldsymbol{Z}$ and thus $I(X_i; Y_{j_1}, \ldots, Y_{j_k} | \boldsymbol{Z}) = 0$. Combining inequality 7 and equation 8 yields the result. □

Larger effect sizes are a key ingredient for more statistical power; other key ingredients are the sizes of the conditioning sets and the significance level. In our setting, we enlarge the conditioning sets to increase the effect sizes. These enlargements of the conditioning sets, however, also reduce statistical power as estimation is more difficult and imprecise for larger conditioning sets. For example, for the multivariate normal distribution, the sample size for estimating the partial correlation effectively reduces by the size of the conditioning set (see e.g., Anderson [2003], p. 143).

In settings where the enlargements of the conditioning sets are small in comparison to the sample size, conditioning on extra variables might yield large increases in the effect size while only negligibly reducing estimation accuracy. In practice, one can choose a threshold based on the effective reduction of the sample size for the partial correlation and the multivariate normal distribution. In particular, one may choose some $\kappa \in (0, 1)$ and require that all Q_i's (analogously all Q'_j's) satisfy

$$\frac{n - d_Z - 2 - |Q_i|}{n - 2 - d_Z} \geq \kappa. \tag{9}$$

If there is some Q_i (or Q'_j) for which (9) is not satisfied, then one can randomly throw out some elements of the respective Q_i (analogously the respective Q'_j). For the algorithm with sample splitting (Section 3.3), one needs to replace n with the sample size for the second step of the procedure.

5 NUMERICAL EXPERIMENTS

To empirically compare our novel approach to the baseline approaches, we employ a slightly modified version of the model considered in Shi et al. [2022]. In our modified model, $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. We restrict our attention to the case where $\boldsymbol{Z}$ is one-dimensional (henceforth denoted as Z), and where $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same number of components, i.e., $d_X = d_Y$. Regarding the covariance matrix $\boldsymbol{\Sigma}$, we consider two cases, which we label $\boldsymbol{\Sigma}^{(1)}$ and $\boldsymbol{\Sigma}^{(2)}$. In the first case, $\boldsymbol{\Sigma}^{(1)}$ takes the form

$$\Sigma^{(1)}_{ij} = \Sigma^{(1)}_{ji} = \begin{cases} \tau^{|i-j|}, & \text{for } i, j \in \{1, \ldots, d_X\}, \\ \tau^{|i-j|}, & \text{for } i, j \in \{d_X + 1, \ldots, \\ & \qquad\qquad d_X + d_Y\}, \\ 1, & \text{for } i = j = d_X + d_Y + d_Z, \\ \rho, & \text{for } i = 1,\ j = d_X + 1, \\ 0, & \text{otherwise.} \end{cases}$$

The zero entries of this matrix imply that the dependence between the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ is solely due to dependence between their components X_1 and Y_1. To define $\boldsymbol{\Sigma}^{(2)}$, we start with $\boldsymbol{\Sigma}^{(1)}$, then we randomly choose 16 entries $\Sigma^{(1)}_{ij} = \Sigma^{(1)}_{ji}$ with $i \in \{1, \ldots, d_X\}$ and $j \in \{d_X + 1, \ldots, d_X + d_Y\}$ (excluding $i = 1, j = d_X + 1$) and make them nonzero by setting them to $\rho/16$.

For both $\boldsymbol{\Sigma}^{(1)}$ and $\boldsymbol{\Sigma}^{(2)}$, each component of $(\boldsymbol{X}, \boldsymbol{Y}, Z)$ has unit variance. The parameter τ characterizes the within-group correlation, and the parameter ρ characterizes the between-group correlation. We look at the cases $\tau = 0$ (no within-$\boldsymbol{X}$ and within-$\boldsymbol{Y}$ dependence), $\tau = 0.5$ (medium within-$\boldsymbol{X}$ and within-$\boldsymbol{Y}$ dependence) and $\tau = 0.9$ (high within-$\boldsymbol{X}$ and within-$\boldsymbol{Y}$ dependence); we vary $\rho \in \{0, 0.005, \ldots, 0.15\}$, consider the sample sizes $n \in \{216, 432, 864\}$ and the dimensions $d_X = d_Y \in \{5, 7\}$. For each of these parameter settings, we do 100 replications and plot the mean rejection rate over these replications with 1 standard error.

For the three different pairwise approaches from Sections 3.1, 3.2 and 3.3, we evaluate pairwise dependence by the partial correlation coefficient (c.f. Anderson [2003], p.121 and p.143). That is, we employ the test statistics

$$T_{ij} := \sqrt{n-2-d_Z}\frac{\hat{\rho}_{X_iY_j|Z}}{\sqrt{1-\hat{\rho}^2_{X_iY_j|Z}}}$$

for testing the hypotheses $H_0^{(ij)}$ and proceed analogously for the pairwise approaches from Section 3.2 and Section 3.3. Here, $\hat{\rho}_{X_iY_j|Z}$ is the sample partial correlation between X_i and Y_j given Z. It is known that T_{ij} is t-distributed with $n-2-d_Z$ degrees of freedom (analogously for the versions from Section 3.2 and Section 3.3). To aggregate the univariate tests, we use the Bonferroni method.

For the approach from Section 3.2, we assume that *all* possible conditional independencies are known a priori. For the sample-splitting approach we set $\alpha_{pre} = 0.5$ (see Section B.1 of the SM for other choices) and consider two different sample splits in which, respectively, 20% and 50% of the samples are used for Step 1. Because we consider sample sizes that are large in comparison to the total number of variables, we do not cut off the conditioning sets as explained at the end of Section 4.

As a baseline method, we use the partial distance correlation from Székely and Rizzo [2014] that directly incorporates the multivariate nature of $\boldsymbol{X}$ and $\boldsymbol{Y}$. For this method we use 1000 permutations to approximate the null distribution.

For all of the above approaches, we set the significance level to 0.05. We implemented all simulations in R [R Core Team, 2020], using the implementation of the partial distance correlation in the *energy*-package [Rizzo and Szekely, 2022] and the *ggplot2*-package [Wickham, 2016] for plotting.

Figure 1 displays the results. We observe that both the pairwise approach which assumes a conditional independence oracle (Section 3.2) and the pairwise approaches with sample splitting (Section 3.3) outperform the simple pairwise approach (Section 3.1) in case of strong within-$\boldsymbol{X}$ and within-$\boldsymbol{Y}$ correlation ($\tau = 0.9$) for both $\boldsymbol{\Sigma}^{(1)}$ and $\boldsymbol{\Sigma}^{(2)}$. If the within-$\boldsymbol{X}$ and within-$\boldsymbol{Y}$ correlation is medium-sized ($\tau = 0.5$), then the algorithm that assumes a conditional independence oracle slightly outperforms the other pairwise algorithms, which are on par with each other. For no within-$\boldsymbol{X}$ and within-$\boldsymbol{Y}$ correlation ($\tau = 0$), the algorithm that assumes a conditional independence oracle and the simple pairwise approach perform similarly well, and the algorithms that learn conditional independencies in the first step perform slightly worse. The sample split with 50% for the first step usually performs worse than the other sample split, however, that effect is also not very strong. We also observe that our novel algorithms perform slightly better for $\boldsymbol{\Sigma}_1$, however, the results for $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are still very similar.

The partial distance correlation performs worse than all pairwise approaches for both covariance matrices. The comparably low performance of the partial distance correlation might be due to the fact that it is a rather generally applicable criterion that is not specifically adapted to the considered example (whereas the pairwise approaches are adapted, because we here combine them with a partial correlation test).

These empirical results are in line with the theory. Looking at the proof of Proposition 1 (or Proposition 2, or Example 1), we see that larger within-$\boldsymbol{X}$ or within-$\boldsymbol{Y}$ correlation leads to higher increases of the effect sizes. If the within-$\boldsymbol{X}$ or within-$\boldsymbol{Y}$ correlations are low, then there is not much (or nothing) to be gained from conditioning on extra variables because within-group dependencies only weakly overlay between-group dependencies; and conditioning out these overlaying-effects were the basis for increasing the effect sizes. The algorithm that uses some part of its sample to learn conditional independencies (Section 3.3) generally trades off sample size for larger effect sizes. Thus, if there is no effect size to be gained ($\tau = 0$), this algorithm is expected to perform worse; and if there is a lot of effect size to be gained, then this algorithm is expected to perform better ($\tau = 0.9$). Further numerical experiments (see Section B in the SM) show that these general findings also apply for other experimental setups and for multivariate independence criteria other than the partial distance correlation.

6 DISCUSSION AND OUTLOOK

In this paper, we introduced a new method for testing conditional independence of random vectors. This new method uses already known or learned pairwise conditional independencies to increase the effect sizes of the remaining univariate tests. The **strength** of this approach is that it efficiently utilizes strong dependencies *within* random vectors and sparse dependence structures *between* random vectors. These are often present in applications of conditional independence testing, for example, on variables describing regionally coherent climate phenomena like El Niño [Runge et al., 2019]. Furthermore, our new approach is comparably fast if the univariate test statistics are fast; it is also flexible with respect to the chosen univariate test statistics. Current **weaknesses** are that not knowing conditional independencies a priori requires the sample to be split and that the algorithms for sample splitting only perform better if the within-vector dependence is sufficiently strong. Moreover, Proposition 1 and 2 only incorporate conditional independence statements such as $X_i \perp\!\!\!\perp Y_j \mid \boldsymbol{Z}$. Further and more complex a priori knowledge as, for example, encoded in causal graphs, is not yet included.

These weaknesses can be tackled in future work, for example, by incorporating a priori knowledge of an underlying causal graph or by developing better approaches than sample splitting using ideas from differential privacy [Dwork et al., 2015] or data carving [Fithian et al., 2014].

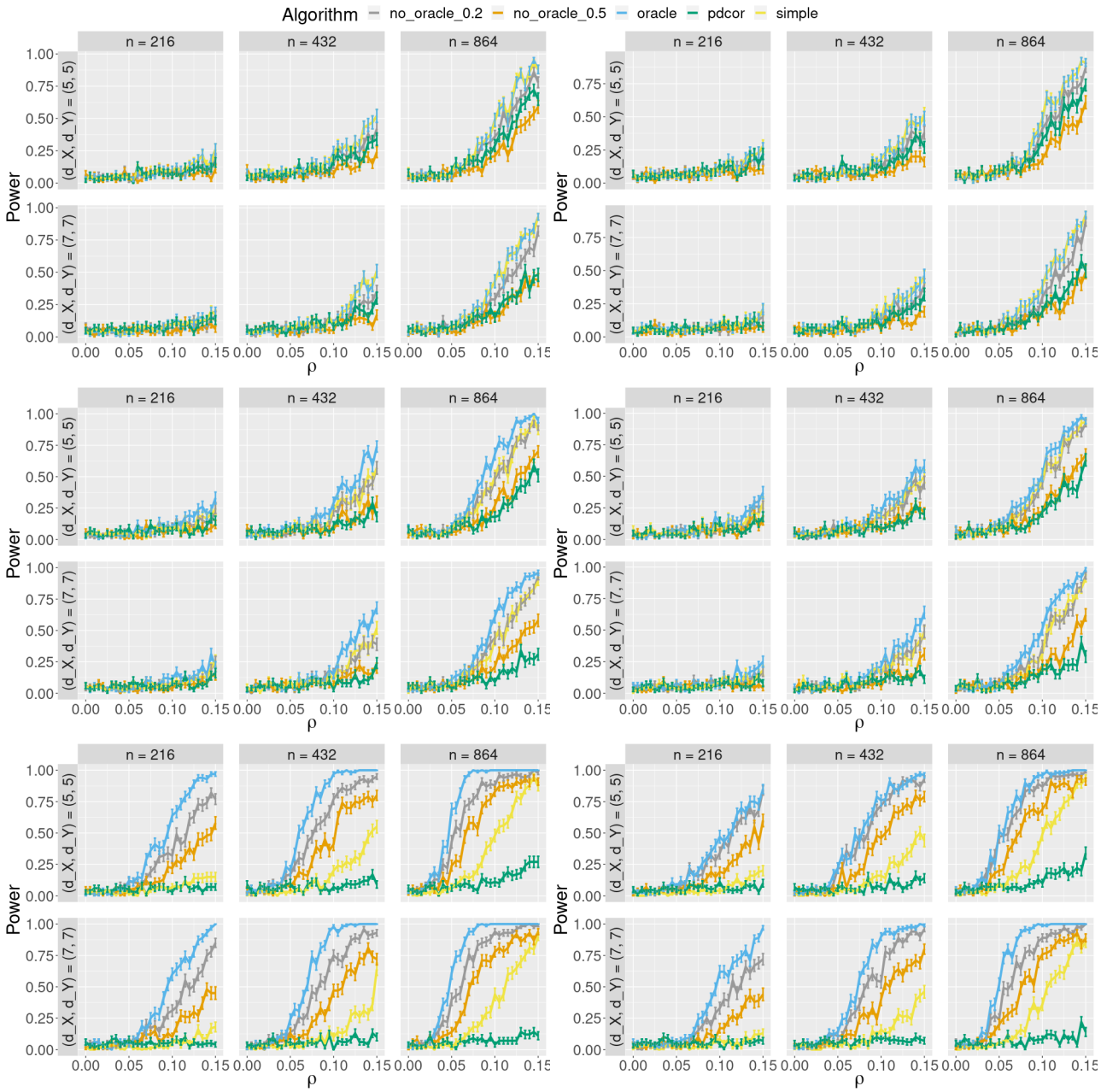

Figure 1: Simulation results for the setting explained in Section 5. The left 3 and the right 3 columns display the results for $\mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma}^{(2)}$ respectively. The first two rows are for $\tau = 0$, the middle two rows for $\tau = 0.5$, and the last two rows for $\tau = 0.9$. The abbreviation *simple* stands for the approach from Section 3.1, *oracle* for the approach from Section 3.2, *no_oracle_0.2* and *no_oracle_0.5* for the sample split approaches from Section3.3 with 20% respectively 50% of the sample used for the first part of the algorithm, and *pdcor* for the partial distance correlation.

References

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey, 3rd edition, 2003.

Wicher Bergsma. Testing conditional independence for continuous random variables. *EURANDOM technical report*, 48:1–19, 2004.

Sourav Chatterjee. A survey of some recent developments in measures of association. *ArXiv*, abs/2211.04702, 2022.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

A. Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B*, 41(1):1–15, 1979.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*, pages 117–126, 2015.

William Fithian, Dennis L. Sun, and Jonathan E. Taylor. Optimal inference after model selection. *arXiv*, abs/2211.04702, 2014.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. *Advances in neural information processing systems*, 20, 2007.

Julie Josse and Susan P. Holmes. Measuring multivariate association and beyond. *Statistics surveys*, 10:132–167, 2016.

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry A. Wasserman. Local permutation tests for conditional independence. *The Annals of Statistics*, 50(6): 3388–3414, 2022.

Pierre Legendre and Louis Legendre. *Numerical Ecology*. Amsterdam, 3rd english edition, 2012.

Chun Li and Xiaodan Fan. On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3): e1489, 2020.

Saralees Nadarajah, Emmanuel Afuecheta, and Stephen Chan. On the distribution of maximum of multivariate normal random vectors. *Communications in Statistics - Theory and Methods*, 48(10):2425–2445, 2019.

Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177, 2021.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK, 2nd edition, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Hans-Peter Piepho. Permutation tests for the correlation among genetic distances and measures of heterosis. *Theoretical and Applied Genetics*, 111:95–99, 2005.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Maria Rizzo and Gabor Szekely. *energy: E-Statistics: Multivariate Inference via the Energy of Data*, 2022. URL `https://CRAN.R-project.org/package=energy`. R package version 1.7-11.

Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947. PMLR, 2018.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13, 2019.

Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann. The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68, 2022.

Rajen Dinesh Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Hongjian Shi, Mathias Drton, and Fang Han. Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, 117(537):395–410, 2022.

Peter E. Smouse, Jeffrey C Long, and Robert R. Sokal. Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic Biology*, 35(4): 627–632, 1986.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2nd edition, 2000.

Eric V. Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1), 2019.

Liangjun Su and Halbert White. A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829–864, 2008.

Liangjun Su and Halbert White. Testing conditional independence via empirical likelihood. *Journal of Econometrics*, 182(1):27–44, 2014.

Liangjun Su and Halbert L. White. A consistent characteristic-function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834, 2007.

Gábor J. Székely and Maria L. Rizzo. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382–2412, 2014.

Jonas Wahl*, Urmi Ninad*, and Jakob Runge. Vector causal inference between two groups of variables. *arXiv preprint arXiv:2209.14283*, 2022.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Kun Zhang, J. Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.