A Bayesian Nonparametric Conditional Two-sample Test with an Application to Local Causal Discovery

Philip A. Boeken[1] Joris M. Mooij[1]

[1]Korteweg-de Vries Institute for Mathematics, University of Amsterdam

Abstract

For a continuous random variable Z, testing conditional independence $X \perp\!\!\!\perp Y | Z$ is known to be a particularly hard problem. It constitutes a key ingredient of many constraint-based causal discovery algorithms. These algorithms are often applied to datasets containing binary variables, which indicate the 'context' of the observations, e.g. a control or treatment group within an experiment. In these settings, conditional independence testing with X or Y binary (and the other continuous) is paramount to the performance of the causal discovery algorithm. To our knowledge no nonparametric 'mixed' conditional independence test currently exists, and in practice tests that assume all variables to be continuous are used instead. In this paper we aim to fill this gap, as we combine elements of Holmes et al. [2015] and Teymur and Filippi [2020] to propose a novel Bayesian nonparametric conditional two-sample test. Applied to the Local Causal Discovery algorithm, we investigate its performance on both synthetic and real-world data, and compare with state-of-the-art conditional independence tests.

1 INTRODUCTION

Conditional independence testing is a fundamental ingredient of many causal inference algorithms such as the PC algorithm [Spirtes et al., 1993], FCI [Spirtes et al., 1999], and the Local Causal Discovery algorithm [Cooper, 1997]. These algorithms can be proven to be complete, sound, or have other desired properties, but these proofs often invoke the use of an 'oracle' for determining conditional independence between variables. In practice, the applicability and performance of the algorithm heavily relies on the reliability of the conditional independence test that is being used. Consequently, incorporating any prior knowledge of the variables involved into the choice of conditional independence test can be desirable.

One way of incorporating prior knowledge is by tailoring the conditional independence tests for $X \perp\!\!\!\perp Y | Z$ on whether the variables involved are discrete or continuous. In the case that the conditioning variable Z is continuous, conditional independence testing is known to be a particularly hard problem [Shah and Peters, 2020] and further specifying whether X and Y are continuous or discrete can be beneficial. For the parametric setting multiple 'mixed' tests are available [Scutari, 2010, Andrews et al., 2018, Sedgewick et al., 2019]. For the nonparametric setting, recent literature proposes multiple tests where X and Y are both assumed to be discrete or both continuous, but to our knowledge no nonparametric test for either X or Y discrete (and the other continuous) currently exists.

Such a 'mixed' conditional independence test has a particularly important role in constraint-based causal discovery algorithms that are applied to datasets which are formed by merging datasets from different contexts [Mooij et al., 2020]. Such a context may for example be whether certain chemicals have been added to a system of proteins (as in Section 3.3), or may be the country of residence of a respondent in an international survey. When certain features of interest (*system variables*) have been measured in different contexts, these measurements can be gathered into a single dataset by adding one or several (often discrete) *context variables* to the dataset, encoding the context that the observation originates from. Merging datasets in this manner may render certain causal relations identifiable, and may improve the reliability of the conditional independence tests due to an increasing sample size [Mooij et al., 2020].

Among the continuous conditional independence tests is a recently proposed Bayesian nonparametric test by Teymur and Filippi [2020] which extends a continuous marginal independence test [Filippi and Holmes, 2017] by utilising *conditional optional Pólya tree priors* [Ma, 2017]. Although

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

this conditional independence test performs well on data originating from continuous distributions, the prior is misspecified in the case of combinations of discrete and continuous variables. Subsequently, the test has close to zero recall when applied to certain datasets consisting of combinations of discrete and continuous variables.

In this paper we focus on the simplified case of testing $X \perp\!\!\!\perp Y | Z$, where Z and either X or Y is continuous, and the other is binary. We propose a Bayesian nonparametric conditional two-sample test by combining elements of the two-sample test by Holmes et al. [2015] and the continuous conditional independence test by Teymur and Filippi [2020]. The two-sample test [Holmes et al., 2015], independence test [Filippi and Holmes, 2017] and our novel conditional two-sample test are empirically compared to both classical and state-of-the-art frequentist (conditional) independence tests when testing for a single (conditional) independence, and when simultaneously testing for multiple (conditional) independences as required by the constraint-based causal discovery algorithm *Local Causal Discovery* (LCD) [Cooper, 1997].[1] Since p-values do not, unlike Bayes factors, reflect any evidence in favour of the null hypothesis, the comparison of Bayesian and frequentist tests in the LCD setting is not straightforward. We propose a measure which allows comparison of the LCD algorithm when using tests from both paradigms, and use it for the comparison of the ensemble of Pólya tree tests with frequentist tests. We observe that LCD with the ensemble of Pólya tree tests outperforms other state-of-the-art (conditional) independence tests, while computation time is substantially lower compared to the competing tests.

We apply the LCD algorithm with the Pólya tree tests to protein expression data from Sachs et al. [2005], and conclude that this implementation provides a result that is more likely to resemble the true model than the output of LCD with the often used partial correlation test.

2 INDEPENDENCE TESTING USING PÓLYA TREE PRIORS

If we let $X : \Omega \to \mathcal{X}$ be a random variable with distribution P and let $\mathcal{M}$ be the space of all probability distributions on $\mathcal{X}$, then for subsets $\mathcal{M}_0 \subset \mathcal{M}$ and $\mathcal{M}_1 \subset \mathcal{M}$ we may test the hypotheses $H_0 : P \in \mathcal{M}_0$ and $H_1 : P \in \mathcal{M}_1$ by considering *random measures* $\mathcal{P}_0$ and $\mathcal{P}_1$ with distributions Π_0 and Π_1 such that $\mathcal{P}_0 \in \mathcal{M}_0$ Π_0-a.s. and $\mathcal{P}_1 \in \mathcal{M}_1$ Π_1-a.s. If the posterior distribution of either $\mathcal{P}_0$ or $\mathcal{P}_1$ is consistent (depending on whether $P \in \mathcal{M}_0$ or $P \in \mathcal{M}_1$) and both models $\mathcal{M}_0$ and $\mathcal{M}_1$ are absolutely continuous with respect to some dominating measure, then we may equivalently state the hypotheses as $H_0 : X \sim \mathcal{P}_0$ and $H_1 : X \sim \mathcal{P}_1$, and test these hypotheses by computing the Bayes factor

$$\mathrm{BF}_{01} = \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} \frac{\int_{\mathcal{M}} \prod_{i=1}^n p(X_i) d\Pi_0(P)}{\int_{\mathcal{M}} \prod_{i=1}^n p(X_i) d\Pi_1(P)}, \tag{1}$$

where $\mathbb{P}(H_j)$ is the prior probability of hypothesis H_j, p is the Radon-Nikodym derivative of P with respect to the dominating measure, and the integral $\int_{\mathcal{M}} \prod_{i=1}^n p(X_i) d\Pi_j(P)$ is the marginal likelihood of the sample $X_1, ..., X_n$ with respect to hypothesis H_j. In this work we will use the *Pólya tree* as a random measure which, under certain assumptions, has a closed form expression for the marginal likelihood of a sample of observations. This is a major benefit compared to e.g. the Dirichlet process, as the Dirichlet process often requires costly MCMC sampling to calculate the marginal likelihood.

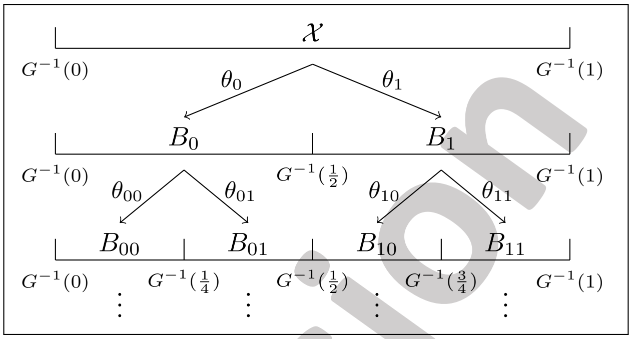


Figure 1: Construction of a one-dimensional Pólya tree based on canonical partitions.

To construct a Pólya tree on $\mathcal{X} \subseteq \mathbb{R}$ we consider the set of *canonical partitions of* $\mathcal{X}$, which is defined as the recursive set of partitions

$$\mathcal{T} = \{\mathcal{X}, \{B_0, B_1\}, \{B_{00}, B_{01}, B_{10}, B_{11}\}, ...\} \tag{2}$$

formed by mapping the family of dyadic partitions of $[0, 1]$ through the inverse of a cumulative distribution function $G : \mathcal{X} \to [0, 1]$ [Ghosal and van der Vaart, 2017]. This results in a family of partitions of $\mathcal{X}$, where for level j we have $\mathcal{X} = \bigcup_{\kappa \in \{0,1\}^j} B_\kappa$, with

$$B_\kappa := \big[G^{-1}(\tfrac{k-1}{2^j}), G^{-1}(\tfrac{k}{2^j})\big), \tag{3}$$

and k denoting the natural number corresponding with the bit string $\kappa \in \{0, 1\}^j$. A schematic depiction of this binary tree of partitions is shown in Figure 1. If we define the index set $K := \{\{0, 1\}^j : j \in \mathbb{N}\}$, then the random measure $\mathcal{P}$ is constructed by letting $\mathcal{P}(\mathcal{X}) := 1$ and recursively assigning random probabilities to $B_\kappa \in \mathcal{T}$ by splitting from the mass that is assigned to B_κ a fraction $\theta_{\kappa 0}$ to $B_{\kappa 0}$ and a fraction $\theta_{\kappa 1}$ to $B_{\kappa 1}$, where we let $(\theta_{\kappa 0}, \theta_{\kappa 1}) \sim \mathrm{Dir}(\alpha_{\kappa 0}, \alpha_{\kappa 1})$. This construction yields a random Borel measure $\mathcal{P}$ on $\mathcal{X}$ [Ghosal and van der Vaart, 2017] which adheres to the following definition:

[1] Code for the (conditional) independence tests, simulations and results on real world data is publicly available at `https://github.com/philipboeken/PTTests`.

Definition 2.1 (Lavine, 1992) *A random probability measure $\mathcal{P}$ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is said to have a* Pólya tree *distribution with parameter $(\mathcal{T}, \mathcal{A})$, written $\mathcal{P} \sim \mathrm{PT}(\mathcal{T}, \mathcal{A})$, if there exist nonnegative numbers $\mathcal{A} = \{(\alpha_{\kappa 0}, \alpha_{\kappa 1}) : \kappa \in K\}$ and random variables $\Theta = \{(\theta_{\kappa 0}, \theta_{\kappa 1}) : \kappa \in K\}$ such that the following hold:*

1. *all the random variables in Θ are independent;*
2. *for every $\kappa \in K$, we have $(\theta_{\kappa 0}, \theta_{\kappa 1}) \sim \mathrm{Dir}(\alpha_{\kappa 0}, \alpha_{\kappa 1})$;*
3. *for every $j \in \mathbb{N}$ and every $\kappa \in \{0,1\}^j$ we have $\mathcal{P}(B_\kappa) = \prod_{i=1}^{j} \theta_{\kappa_1 \ldots \kappa_i}$.*

Let X be a continuous random variable and consider the Pólya tree $\mathcal{P} \sim \mathrm{PT}(\mathcal{T}, \mathcal{A})$. Drawing a distribution from $\mathcal{P}$ is done by drawing from each of the random variables in Θ. If we let $X_1, ..., X_n$ be a sample from X, then the likelihood of that sample with respect to a sampled distribution Θ from the Pólya tree $\mathrm{PT}(\mathcal{T}, \mathcal{A})$ is

$$p(X_{1:n} | \Theta, \mathcal{T}, \mathcal{A}) = \prod_{\kappa \in K} \theta_\kappa^{n_\kappa}, \tag{4}$$

where n_κ denotes the number of observations lying in B_κ, i.e. $n_\kappa := |X_{1:n} \cap B_\kappa|$. If we integrate out Θ we obtain the marginal likelihood

$$p(X_{1:n} | \mathcal{T}, \mathcal{A}) = \prod_{\kappa \in K} \frac{\mathrm{B}(\alpha_{\kappa 0} + n_{\kappa 0}, \alpha_{\kappa 1} + n_{\kappa 1})}{\mathrm{B}(\alpha_{\kappa 0}, \alpha_{\kappa 1})}, \tag{5}$$

where $\mathrm{B}(\cdot)$ denotes the Beta function.

The choice of $\mathcal{T}$ and $\mathcal{A}$ influences certain characteristics of samples from the Pólya tree. For example, if we let $\alpha_{\kappa 0} = \alpha_{\kappa 1}$ for all $\kappa \in K$ then the Pólya tree is centred on the base distribution with cumulative distribution function G, i.e. $\mathbb{E}[\mathcal{P}(B_\kappa)] = \int_{B_\kappa} G'(x) dx$. Kraft [1964] provides sufficient conditions on $\mathcal{A}$ for the Pólya tree to be dominated by Lebesgue measure. These conditions are satisfied if for each $\kappa \in \{0,1\}^j$ we take $\alpha_\kappa = |\kappa|^2$ with $|\kappa| := j$. The choice of the parameter $\mathcal{A}$ is analysed in Section 4.

2.1 A NONPARAMETRIC CONDITIONAL TWO-SAMPLE TEST

We now propose a conditional independence test of the type $C \perp\!\!\!\perp X | Z$, where X and Z are continuous one-dimensional random variables and C is a binary random variable. Let F be the conditional distribution of $X|Z$, and let the conditional distributions of $X | \{C = 0\}, Z$ and $X | \{C = 1\}, Z$ be $F^{(0)}$ and $F^{(1)}$ respectively. Then we formulate the conditional independence test between C and X given Z as a two-sample test, i.e.

$$\begin{aligned} H_0 &: C \perp\!\!\!\perp X | Z \iff F^{(0)} = F^{(1)} = F \\ H_1 &: C \not\perp\!\!\!\perp X | Z \iff F^{(0)} \neq F^{(1)}. \end{aligned} \tag{6}$$

Following Teymur and Filippi [2020] we will utilise the *conditional optional Pólya tree* (cond-OPT) prior [Ma, 2017] for modelling the conditional distributions F, $F^{(0)}$ and $F^{(1)}$. The cond-OPT is a random conditional probability measure on e.g. $\mathcal{X} \times \mathcal{Z}$, where X is the response variable and Z is the predictor. In order to construct the cond-OPT, we first construct a family of partitions $\mathcal{T}_Z$ of $\mathcal{Z}$ according to the partitioning scheme of the optional Pólya tree (OPT) [Wong and Ma, 2010], which results in a random subset of the canonical partitions $\mathcal{T}$ as constructed by equation (3). This random subset of $\mathcal{T}$ is obtained by first adding $B_\emptyset := \mathcal{Z}$ to $\mathcal{T}_Z$. Then we sample from the random variable $S \sim \mathrm{Bernoulli}(\rho)$; if $S = 1$ we stop the partitioning procedure, and if $S = 0$ we add B_0 and B_1 to $\mathcal{T}_Z$. Then, for both B_0 and B_1 we repeat this procedure; we first draw S from $\mathrm{Bernoulli}(\rho)$ and depending on the outcome we add the children of B_0, then we repeat this to possibly add the children of B_1. This process is iterated, and terminates a.s. when $\rho > 0$.

Having obtained the family $\mathcal{T}_Z$, we construct a 'local' random measure $\mathcal{P}(\cdot | B_\kappa)$ on $\mathcal{X}$ for each $B_\kappa \in \mathcal{T}_Z$ by letting $\mathcal{P}(\cdot | B_\kappa) \sim \mathrm{PT}(\mathcal{T}, \mathcal{A})$, and we define the conditional probability $\mathcal{P}(\cdot | Z = z)$ to be constant and equal to the local Pólya tree $\mathcal{P}(\cdot | B_\kappa)$ on the stopped set $B_\kappa \ni z$. The resulting family of random measures on $\mathcal{X}$ is the conditional optional Pólya tree (cond-OPT) [Ma, 2017]. When using the canonical partitions for both $\mathcal{X}$ and $\mathcal{Z}$ and assuming that all the local Pólya trees are a.s. dominated by Lebesgue measure, Ma [2017] shows that the cond-OPT places positive probability on all L_1 neighbourhoods of any conditional density $f(\cdot|\cdot)$ on $\mathcal{X} \times \mathcal{Z}$.

When we are given n i.i.d. observations $(C_1, X_1, Z_1), ..., (C_n, X_n, Z_n)$, then under the null hypothesis we are interested in the marginal likelihood of a sample $(X_1, Z_1), ..., (X_n, Z_n)$ with respect to the cond-OPT prior. This is obtained by for every $B_\kappa \in \mathcal{T}_Z$ considering the subsample $X(B_\kappa) := \{X_j : Z_j \in B_\kappa\}$. As the cond-OPT prior considers a general Pólya tree prior for this subsample, we simply compute the marginal likelihood

$$p_X(B_\kappa) := p(X(B_\kappa) | \mathcal{T}, \mathcal{A}) \tag{7}$$

using equation (5). If B_κ is a so called leaf-set, i.e. the set contains at most one observation or it has no children in the family of partitions $\mathcal{T}_Z$, then we simply return this marginal likelihood. If B_κ is not a leaf-set, we continue along the children $B_{\kappa 0}$ and $B_{\kappa 1}$. We integrate out the randomness of the random family of partitions by considering the entire family of canonical partitions $\mathcal{T}$ of $\mathcal{Z}$, and incorporating the stopping probabilities S by weighing the elements $B_\kappa \in \mathcal{T}$ of level $|\kappa|$ with $\mathbb{E}(1 - S)^{|\kappa|} = (1 - \rho)^{|\kappa|}$. The recursive mixing formula is given by

$$\Phi_X(B_\kappa) = \begin{cases} p_X(B_\kappa) & \text{if } B_\kappa \text{ is a leaf-set} \\ \rho \cdot p_X(B_\kappa) + \\ (1-\rho) \cdot \Phi_X(B_{\kappa 0}) \Phi_X(B_{\kappa 1}) & \text{otherwise,} \end{cases}$$

and the resulting quantity $\Phi_X(B_\kappa)$ is the marginal likelihood of $\{(X_1, Z_1), ..., (X_n, Z_n)\} \cap \mathcal{X} \times B_\kappa$, with respect to the cond-OPT.

Under the alternative hypothesis we split the sample into sets $X^{(0)} := \{(X_j, Z_j) : C_j = 0\}$ and $X^{(1)} := \{(X_j, Z_j) : C_j = 1\}$, and compute the marginal likelihoods $\Phi_{X^{(0)}}(\mathcal{Z})$ and $\Phi_{X^{(1)}}(\mathcal{Z})$ of these sets with respect to (independent) cond-OPT priors. We finally test the hypothesis by computing the Bayes factor

$$\mathrm{BF}_{01} = \frac{\Phi_X(\mathcal{Z})}{\Phi_{X^{(0)}}(\mathcal{Z})\Phi_{X^{(1)}}(\mathcal{Z})}, \qquad (8)$$

where we have set the prior odds to 1.

We note that when no data is provided for Z and thus $\mathcal{Z}$ constitutes a leaf-set, this test defaults to the two-sample test from Holmes et al. [2015]. An overview of this two-sample test and the continuous independence test by Filippi and Holmes [2017] is provided in the supplement.

3 EXPERIMENTS

Implementing the conditional independence test requires choosing certain hyperparameters. As mentioned earlier, we set $\alpha_\kappa = |\kappa|^2$. As argued by Lavine [1994] we will only consider partitions up to a pre-determined level J, making $\mathcal{P}$ into a *truncated Pólya Tree*. Hanson and Johnson [2002] provide the rule of thumb $J = \lfloor \log_2(n) \rfloor$, which corresponds to on average finding one observation in each element of the partition. We find however that $J = \lfloor \log_4(n) \rfloor$, which corresponds to finding approximately $\sqrt{n}$ observations in each element of the partition, provides similar results and considerably reduces computation time, so we use this maximum depth. Throughout this work we will use the standard Gaussian cdf G to form the canonical partitions. In conjunction with this mean measure, we standardise the data before computing the marginal likelihoods. For computing the marginal likelihood of the cond-OPT we use $\rho = 1/2$ [Ma, 2017]. Similar to the computation of marginal likelihoods of regular Pólya trees, we use a maximum partitioning depth of $\lfloor \log_4(n) \rfloor$, so we consider $B_\kappa \in \mathcal{T}_Z$ to be a leaf-set if it contains at most one value, or if $|\kappa| = \lfloor \log_4(n) \rfloor$.

All experiments are run on a MacBook Pro with a 3.1 GHz CPU and 16GB of RAM, with a parallelised R implementation of the LCD algorithm. Code for the (conditional) independence tests, simulations and results on real world data is publicly available at `https://github.com/philipboeken/PTTests`.

3.1 LOCAL CAUSAL DISCOVERY

As mentioned earlier, a 'mixed' conditional independence test as proposed in Section 2.1 is specifically needed when applying causal discovery algorithms to datasets containing binary (or discrete) *context variables*, which encode the context that observations of the *system variables* (the variables of interest) originate from. In accordance with Mooij et al. [2020], we regard both the context variables and the system variables as distributed according to the solution of a *Structural Causal Model* (SCM) [Pearl, 2009]. A relatively insightful causal discovery algorithm is the *Local Causal Discovery* (LCD) algorithm [Cooper, 1997]. Although often referred to as an algorithm, it essentially consists of the following proposition:

Proposition 3.1 (LCD, Mooij et al. [2020]) *If the data generating process of the triple of random variables (C, X, Y) has no selection bias, can be modelled by a faithful simple SCM, and X is not a cause of C, then the presence of (in)dependences*

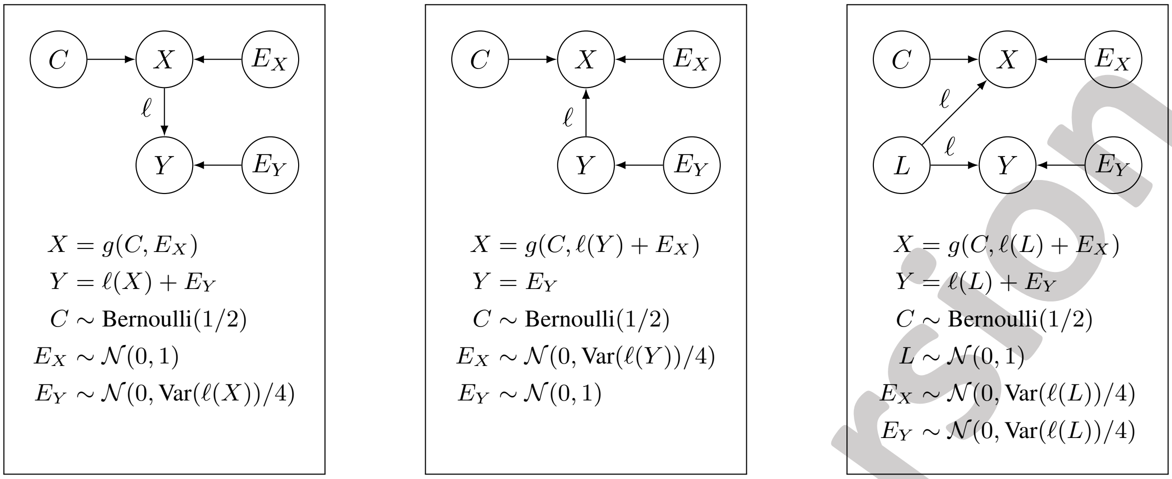
$$C \not\perp\!\!\!\perp X, \quad X \not\perp\!\!\!\perp Y, \quad C \perp\!\!\!\perp Y | X \qquad (9)$$

implies that X is a (possibly indirect) cause of Y. If this is the case, we speak of the 'LCD triple' (C, X, Y).

By repeatedly applying this proposition to different triples of random variables one can (partially) reconstruct the underlying causal graph of the dataset at hand. If we are provided with a dataset consisting of observations of context variables $(C_k)_{k \in \mathcal{K}}$ for some index set $\mathcal{K}$ and system variables $(X_i)_{i \in \mathcal{I}}$ for some index set $\mathcal{I}$ for which we assume that the system variables do not cause the context variables, then we may iteratively apply Proposition 3.1 to all triples $(C_k, X_i, X_{i'})$ where $k \in \mathcal{K}$ and $i \neq i' \in \mathcal{I}$, and provide a directed graph as output where the edges can be interpreted as representing indirect causal effects.

3.2 SIMULATIONS

In our simulations we repeatedly simulate a triple of random variables (C, X, Y). Each time we simulate a set of observations, we test for $C \perp\!\!\!\perp X$, $X \perp\!\!\!\perp Y$ and $C \perp\!\!\!\perp Y | X$ individually, and by combining the output of these three tests we formulate the output of the LCD algorithm. Upon repeating this scheme a number of times we are able to display ROC curves for each of the three test cases, and for the LCD algorithm. To widen the scope of this setup, in each round of simulations we randomly choose one of the graphs of Figure 2, and we randomly pick the relations between C and X and between X and Y from predefined, varying possibilities. More specifically, if we let E be an external factor (possibly depending on Y) we set X equal to $g(C, E)$,



$$\begin{aligned} X &= g(C, E_X) \\ Y &= \ell(X) + E_Y \\ C &\sim \text{Bernoulli}(1/2) \\ E_X &\sim \mathcal{N}(0,1) \\ E_Y &\sim \mathcal{N}(0, \text{Var}(\ell(X))/4) \end{aligned}$$

(a) SCM with $C \perp\!\!\!\perp Y|X$

$$\begin{aligned} X &= g(C, \ell(Y) + E_X) \\ Y &= E_Y \\ C &\sim \text{Bernoulli}(1/2) \\ E_X &\sim \mathcal{N}(0, \text{Var}(\ell(Y))/4) \\ E_Y &\sim \mathcal{N}(0,1) \end{aligned}$$

(b) SCM with $C \not\!\perp\!\!\!\perp Y|X$

$$\begin{aligned} X &= g(C, \ell(L) + E_X) \\ Y &= \ell(L) + E_Y \\ C &\sim \text{Bernoulli}(1/2) \\ L &\sim \mathcal{N}(0,1) \\ E_X &\sim \mathcal{N}(0, \text{Var}(\ell(L))/4) \\ E_Y &\sim \mathcal{N}(0, \text{Var}(\ell(L))/4) \end{aligned}$$

(c) SCM with $C \not\!\perp\!\!\!\perp Y|X$

Figure 2: Three SCMs used for the simulations.

which is randomly chosen from

$$g(c,e) = \begin{cases} e & \text{no intervention} \\ (1-c)e + c(e+\theta) & \text{mean shift} \\ (1-c)e + c(1+\theta)e & \text{variance shift} \\ (1-c)e + c\theta & \text{perfect intervention} \\ (1-c)e + c(e+B) & \text{mean shift mixture,} \end{cases} \tag{10}$$

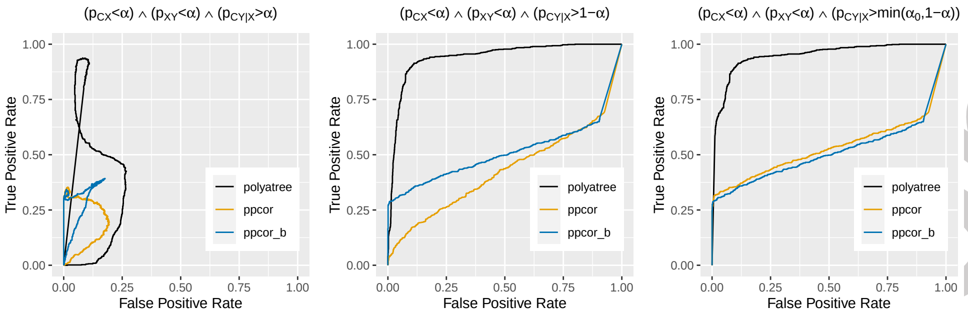
with $\theta \sim \mathcal{U}(\{2,3,4,5,6\})$ independently drawn per round of simulations and $B \sim \mathcal{U}(\{-1,\theta\})$ independently drawn for every observation. These mappings between C and X can be interpreted as setting X equal to the value E in context $C = 0$, and intervening on X in context $C = 1$. If we for example inspect the 'mean shift', then if $C = 1$ we intervene on the distribution of X by shifting the mean of X with the amount θ. When simulating multiple observations, this intervention on X is performed on approximately half of these observations, due to C having a Bernoulli(1/2) distribution. The relation ℓ between X and Y is randomly picked from

$$\ell(x) = \begin{cases} 0 & \text{no link} \\ x & \text{linear} \\ x^2 & \text{parabolic} \\ \sin(12\pi\tilde{x}) & \text{sinusoidal} \end{cases} \tag{11}$$

where $\tilde{x} = x/(\max(x_1,...,x_n) - \min(x_1,...,x_n))$. It depends on which graph from Figure 2 is chosen whether we have $X \overset{\ell}{\to} Y$, $X \overset{\ell}{\leftarrow} Y$ or $X \overset{\ell}{\leftarrow} L \overset{\ell}{\to} Y$, where in the last case the two ℓ's are drawn independently. The possibility of picking $g(c,e) = e$ and $\ell(x) = 0$ ensures the occurrence of $C \perp\!\!\!\perp X$ and $X \perp\!\!\!\perp Y$ respectively, which in turn enables plotting ROC curves of these test cases.

We compare the Pólya tree based ensemble of the two-sample test [Holmes et al., 2015], independence test [Filippi and Holmes, 2017] and conditional two-sample test (Section 2.1), denoted by polyatree, with both classical and recently proposed (conditional) independence tests. The tests that are suitable for mixed testing are mi_mixed and lr_mixed, where the former is based on mutual information and uses the implementation of the bnlearn package Scutari [2010], and where the latter is a likelihood ratio test of linear and logistic regressions [Sedgewick et al., 2019]. Among the more classical continuous tests is the Pearson correlation- and partial correlation test, denoted by ppcor, implemented using the synonymous R-package [Kim, 2015]. Harris and Drton [2013] promote the use of Spearman's (partial) rank correlation test in the context of nonparanormal models, which we denote by spcor. Among the more state-of-the-art continuous tests is the *Generalised Covariance Measure* (GCM) [Shah and Peters, 2020], which can be loosely interpreted as a nonlinear extension of the partial correlation test. The GCM is implemented with penalised regression splines as provided by the R-package GeneralisedCovarianceMeasure, and is denoted by gcm. Departing from the regression-type independence tests, we also consider the *Randomised Conditional Correlation Test* (RCoT) as proposed by Strobl et al. [2019], which closely approximates the Kernel Conditional Independence test by Zhang et al. [2011] at the benefit of significantly lower computation time. For marginal independence testing the RCoT defaults to an approximate version of the Hilbert-Schmidt Independence Criterion [Gretton et al., 2008]. This ensemble is denoted by rcot. Lastly we compare to the *Classifier Conditional Independence Test* (CCIT) [Sen et al., 2017], denoted by ccit, which uses the XGBoost binary classifier to assess presence of conditional independence.

Comparing Bayesian and frequentist tests based on their performance in the LCD algorithm is not straightforward, since the triple of tests for $C \not\!\perp\!\!\!\perp X$, $X \not\!\perp\!\!\!\perp Y$ and $C \perp\!\!\!\perp Y|X$ does not



(a) (b) (c)

Figure 3: ROC curves of different ways of scoring an LCD triple (C, X, Y). See main text for details.

by default output a single confidence score. For each test we output the p-value, or in case of the Bayesian tests the H_0 model evidence $\mathbb{P}(H_0|\text{data})$.[2] We construct ROC curves for testing 'positive' outcomes $C \not\perp\!\!\!\perp X$, $X \not\perp\!\!\!\perp Y$ and $C \not\perp\!\!\!\perp Y|X$ by varying the threshold α, representing the upper bound on the p-value/model evidence for drawing a positive conclusion. The triple (C, X, Y) is given a 'positive' label if the data is generated according to the relation $C \to X \to Y$. Typically, varying the threshold α from 0 to 1 produces an ROC curve between the points $(0, 0)$ and $(1, 1)$. If we denote the frequentist p-values or Bayesian H_0 model evidence for the tests $C \perp\!\!\!\perp X$, $X \perp\!\!\!\perp Y$ and $C \perp\!\!\!\perp Y|X$ with p_{CX}, p_{XY} and $p_{CY|X}$ respectively (with independence under the null hypothesis), and if we were to use the same α as threshold for testing whether $p_{CX} < \alpha$, $p_{XY} < \alpha$ and $p_{CY|X} > \alpha$, then varying α between 0 and 1 does not result in a curve between $(0, 0)$ and $(1, 1)$, as shown in Figure 3a. To assess whether we provide a fair comparison between Bayesian and frequentist tests, we include a Bayesian version of the Pearson (partial) correlation test [Wetzels and Wagenmakers, 2012], denoted by `ppcor_b`. Alternatively we could use α for testing $p_{CX} < \alpha$, $p_{XY} < \alpha$ and $p_{CY|X} > 1 - \alpha$, as shown in Figure 3b. In this case the level α reflects the amount of evidence for the desired conclusions $C \not\perp\!\!\!\perp X$, $X \not\perp\!\!\!\perp Y$ and $C \perp\!\!\!\perp Y|X$. For frequentist tests this would not make sense, as for decreasing α we require more evidence for $H_0 : C \perp\!\!\!\perp Y|X$, and the p-value has a uniform distribution under H_0. This is remedied by, when testing for independence $C \perp\!\!\!\perp Y|X$, only varying α between 0 and a fixed α_0 (Figure 3c). More specifically, for level α the LCD algorithm outputs the score

$$s_{\text{LCD}} = \mathbb{1}_{[0,\alpha]}(p_{CX}) \cdot \mathbb{1}_{[0,\alpha]}(p_{XY}) \cdot \mathbb{1}_{(\alpha_0,1] \cup (1-\alpha,1]}(p_{CY|X}), \quad (12)$$

where we let $\alpha_0 = 0.05$ for frequentist tests and $\alpha_0 = 0.5$ for Bayesian tests. Although this α_0 is quite arbitrarily chosen, the use of this performance measure is corroborated by the observation that in Figure 3c the frequentist partial correlation and Bayesian partial correlation tests have similar performance.

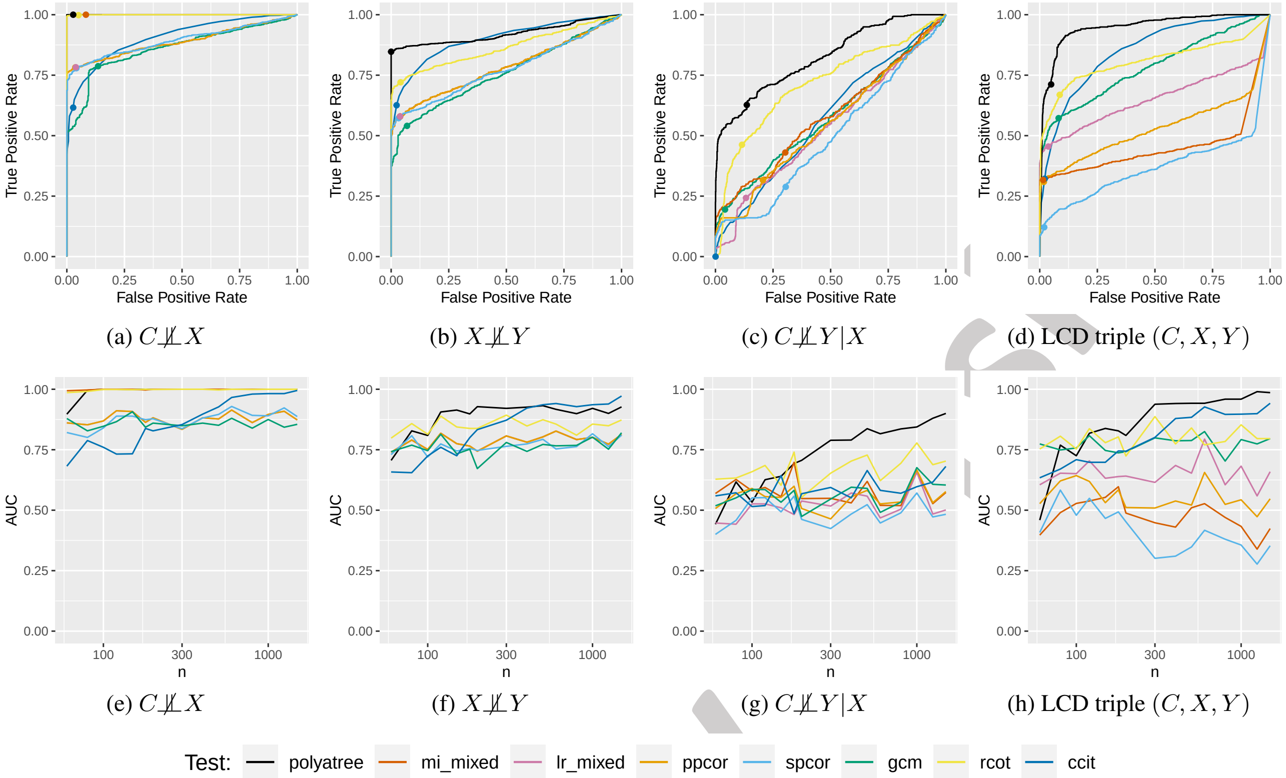
Figures 4 (a–d) show the results of 2000 rounds of simulations, where in each round we simulate 400 observations. On the ROC curves we have marked the reference points $\alpha = 0.05$ and $\alpha = 0.5$ for respectively frequentist and Bayesian tests. Figures 4 (e–h) generalise these results, as they show the areas under the ROC curves (AUC) for varying sample sizes. We note that for conditional independence testing (Figure 4c and 4g), the Pólya tree test from Section 2.1 and the RCoT perform relatively well. It is interesting to see that the other tests have performance close to random guessing. It is however unclear whether this is due to the nonlinearity ℓ, the intervention g or the fact that C is binary instead of continuous. From Figures 4d and 4h we see that the high performance of the Pólya tree tests accumulates into good performance of the LCD algorithm. Interestingly, the CCIT also performs quite well, despite its weak performance in conditional independence testing.

In Figure 5 we display for each independence test the computation times of the three test cases, accumulated over 2000 rounds of simulation at a sample size of $n = 400$, as performed for generating Figures 4 (a–c). The reader should be aware that for the GCM the difference in runtime between marginal and conditional independence testing is due to the fact that for conditional independence testing two nonlinear regressions are performed, and for marginal testing a statistic similar to partial correlation is computed. The CCIT has relatively high computation time due to costly training of the XGBoost classifier for each round of simulations, which makes it rather impractical to use. The partial correlation tests clearly perform best in terms of runtime. Overall, we conclude that the Pólya tree tests provide a very good trade-off between performance and computation time.

3.3 PROTEIN EXPRESSION DATA

We apply the LCD algorithm, implemented with the Bayesian ensemble of independence tests, to protein expres-

[2] Recall that $\mathbb{P}(H_0|\text{data}) = 1 - (1 + \text{BF}_{01})^{-1}$.



(a) $C \not\perp\!\!\!\perp X$ (b) $X \not\perp\!\!\!\perp Y$ (c) $C \not\perp\!\!\!\perp Y | X$ (d) LCD triple (C, X, Y)

(e) $C \not\perp\!\!\!\perp X$ (f) $X \not\perp\!\!\!\perp Y$ (g) $C \not\perp\!\!\!\perp Y | X$ (h) LCD triple (C, X, Y)

Figure 4: ROC and AUC results for simulated data. The first row depicts ROC curves for individual tests (a–c) and for the LCD algorithm (d) over 2000 rounds of simulations at sample size $n = 400$. The second row depicts the median AUC for varying sample size (ranging from 60 to 1500) for individual tests (e–g) and for the LCD algorithm (h) over 200 rounds of simulations.

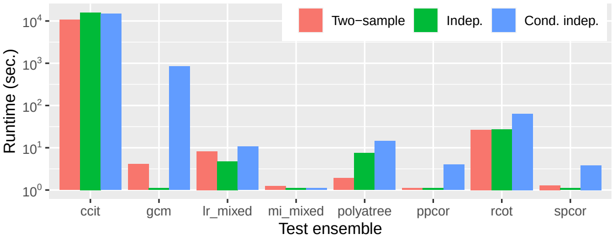


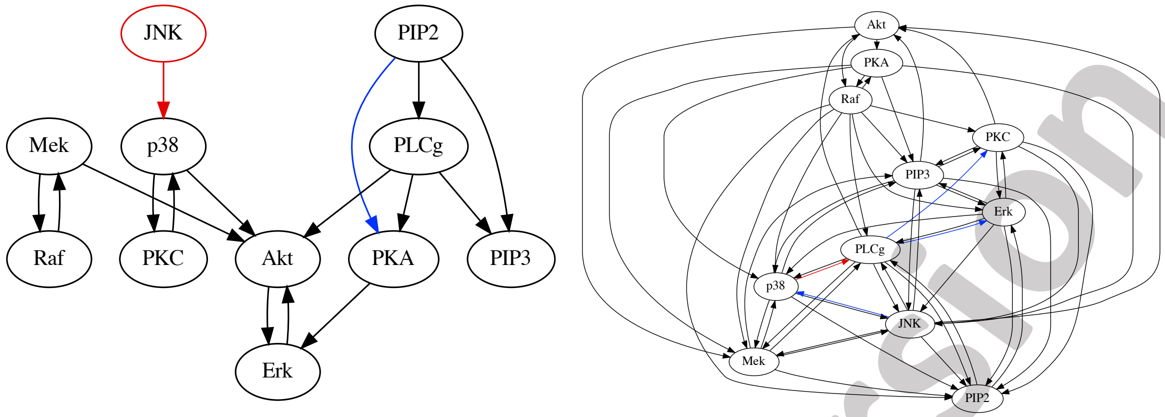
Figure 5: Runtimes of the different tests on 2000 rounds of simulations at $n = 400$.

sion data [Sachs et al., 2005]. Sachs et al. provide an ‘expert network’, depicting the consensus (at that time) among biologists on the true network of signals between 11 proteins and phospholipids, and 10 reagents that are added to the cells. They estimate a causal graph which deviates from the expert network by some edges, refraining from claiming whether these edges should be added to the true network. For a detailed description of the data set and a depiction of the expert network we refer to the supplement.

Many authors have used this data set for estimating the underlying causal network, of which the graph of the original paper [Sachs et al., 2005] most closely resembles the expert network [Ramsey and Andrews, 2018]. Furthermore, Ramsey and Andrews [2018] and Mooij et al. [2020] provide sufficient grounds for rejecting the expert network as being the true causal graph of the data. As we have no reliable ground truth to compare the output of the LCD algorithm with, we compare the output of LCD with its implementation with partial correlation.

The output of the LCD algorithm implemented with the Bayesian tests and with the partial correlation test is shown in Figure 6. In both cases we report the output of the LCD algorithm for multiple thresholds for the statistical tests. For the Bayesian tests (Figure 6a) we use Bayes factor thresholds of $k = 10$ (strong evidence, depicted in black), $k = 4$ (substantial evidence, depicted in red) and $k = 1$ (weak evidence, depicted in blue) [Kass and Raftery, 1995], and for the partial correlation test (Figure 6b) we report results for the p-value thresholds $\alpha = 0.0001$ (strong evidence, depicted in black), $\alpha = 0.005$ (substantial evidence, depicted in red) and $\alpha = 0.05$ (weak evidence, depicted in blue).

In general, we note that the output of LCD differs strongly among the use of different statistical tests, corroborating



(a) Output of LCD with Pólya tree tests.

(b) Output of LCD with the partial correlation test.

Figure 6: The output of LCD on the Sachs data. Edges indicate (possibly indirect) causal effects between the nodes. Black edges indicate strong evidence, red edges indicate substantial evidence, and blue edges indicate weak evidence.

the premise that the performance of the algorithm highly depends on the choice of statistical test. Since the partial correlation test does not detect nonlinear conditional independencies, it has relatively low recall when compared with the Pólya tree test, as shown in Figure 4c. This causes the LCD algorithm with partial correlations to output more false positives, resulting in a very dense causal graph, whereas LCD with the Pólya tree tests produces a graph which is more likely to resemble the true causal model.

4 SENSITIVITY ANALYSIS

As mentioned earlier, the Pólya tree is parametrised by the set $\mathcal{A}$, where in the previous section we have used $\alpha_\kappa = |\kappa|^2$. In general we can let $\alpha_\kappa := \rho(|\kappa|)$ for any positive function ρ, in which case we have

$$\mathrm{Var}(\mathcal{P}(B_\kappa)) = \frac{1}{4^{|\kappa|}}\left(\prod_{j=1}^{|\kappa|}\frac{2\rho(j)+2}{2\rho(j)+1}-1\right), \qquad (13)$$

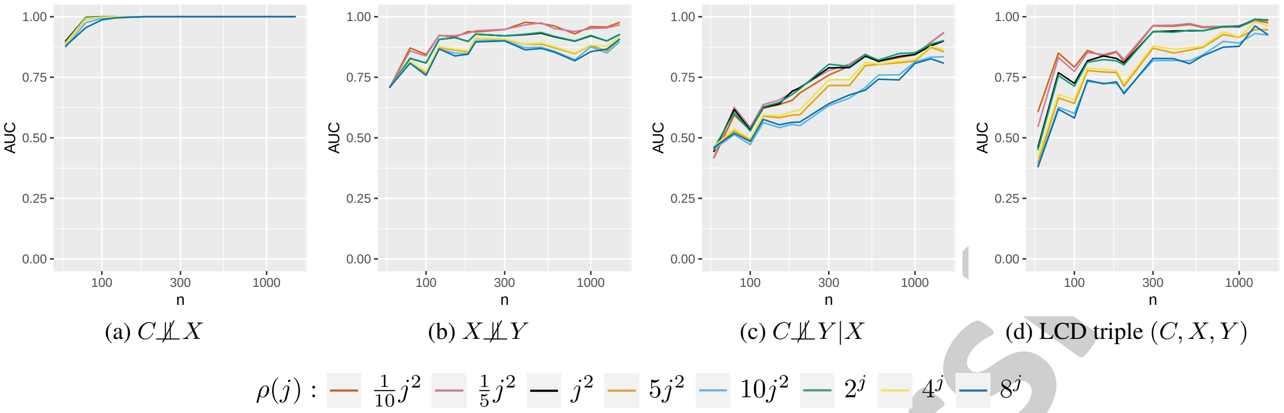
and samples from the Pólya tree are dominated by Lebesgue measure if $\sum_{j=1}^{\infty}\rho(j)^{-1} < \infty$ [Kraft, 1964]. Walker and Mallick propose to use $\rho(j) = cj^2$ for some $c > 0$, in which case decreasing c increases the variance of $\mathcal{P}$, causing $\mathcal{P}$ to be less dependent on the choice of G. We have chosen $c = 1$ as a default value in the previous section as it is promoted as a "sensible canonical choice" by Lavine [1992]. According to Holmes et al. [2015], having c between 1 and 10 is in general a good choice. To obtain an better understanding of the dependency of the Pólya tree on this parameter, we have repeated the experiments of Figure 4 (e–h) for different choices of ρ. More specifically, we have repeated the experiments for parameters $\rho(j) = \frac{1}{10}j^2, \frac{1}{5}j^2, j^2, 5j^2, 10j^2, 2^j, 4^j$ and 8^j [Berger and Guglielmi, 2001]. The results are shown in Figure 7. We note that the performance of the tests is not heavily influenced by the choice of $\mathcal{A}$, and that $\rho(j) = j^2$ seems to be an appropriate default value.

5 DISCUSSION

In this work we have proposed a novel nonparametric conditional two-sample test, which is possibly the first conditional independence test of this type. The test is analysed in its own right and as a subroutine of the Local Causal Discovery algorithm, and in both cases can outperform current state-of-the-art nonparametric continuous conditional independence tests and parametric mixed conditional independence tests. However, we have made some modelling decisions which might be reconsidered when using this test in practice.

First we note that the choice of $\mathcal{A}$ may influence the suitability of the test. Section 4 suggests that $\alpha_\kappa = |\kappa|^2$ is a sensible parametrisation, but this may be reconsidered in applications. Another consideration is the choice of the family of partitions $\mathcal{T}$. Having canonical partitions increases the intelligibility of the Pólya tree, but essentially any recursive partitioning tree suffices. We note that the maximum partitioning depth $J = \lfloor \log_4(n) \rfloor$ is quite arbitrarily chosen to reduce computation time. However, as our choice of α_κ implies relatively low dependence on the base measure G and as we standardise the data to approximately fit the standard Gaussian base measure, we believe that we have chosen sensible default parameters.

In general, it is hard to theoretically analyse for which types of distributions conditional independence tests work properly. For frequentist tests, the asymptotic distribution of the test statistic is often provided, which holds under rather technical assumptions which may be hard to validate against a provided dataset (see Strobl et al. [2019] for an example of



(a) $C \not\perp\!\!\!\perp X$

(b) $X \not\perp\!\!\!\perp Y$

(c) $C \not\perp\!\!\!\perp Y | X$

(d) LCD triple (C, X, Y)

Figure 7: Sensitivity of the performance of the Pólya tree tests with respect to the parameter $\mathcal{A}$.

such assumptions). The same holds for theoretical consistency results of the test statistic under the alternative. Shah and Peters [2020] show that in order to have power against an acceptably large set of alternatives, one should restrict the set of distributions considered under H_0. In a Bayesian setting, consistency of the Bayes Factor is determined by whether the posterior corresponding to the true hypothesis is consistent (i.e. the marginal likelihood is large), and the marginal likelihood remains small under the false hypothesis. Sufficient conditions for posterior convergence are e.g. provided by Doob's Theorem and Schwartz's Theorem [Ghosal and van der Vaart, 2017], but necessary conditions (which could be used to restrict H_0 and H_1) are not available to our knowledge. One should also investigate the behaviour of the posterior likelihood under misspecification to properly determine for which H_0 and H_1 the test works properly.

Many constraint-based causal inference algorithms (other than LCD) require conditional independence testing of the form $C \perp\!\!\!\perp X | Z$ for d-dimensional Z with $d > 1$. Extending our method is straightforward, as the canonical partitions of $\mathcal{Z}$ can be constructed as the per-level cartesian product of d one-dimensional canonical partitions [Hanson, 2006]. However, this extension suffers from the curse of dimensionality, so further research should look into how this problem can be mitigated.

This work only addresses testing $C \perp\!\!\!\perp X | Z$ where C is binary. Although this test is already of high importance to the field of causal discovery, extending this test to discrete C would be of real use and is the subject of current research.

The ensemble of Pólya tree prior based independence tests provides good results when utilised in a causal inference algorithm applied on synthetic data, and produces sensible output on real world data. We therefore believe that it is a promising area of research, which hopefully will improve the robustness and applicability of causal inference algorithms.

Acknowledgements

JMM was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 639466).

References

Bryan Andrews, Joseph Ramsey, and Gregory F Cooper. Scoring bayesian networks of mixed variables. *International Journal of Data Science and Analytics*, 6(1): 3–18, 2018. URL `https://doi.org/10.1007/s41060-017-0085-7`.

James O Berger and Alessandra Guglielmi. Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, 96(453):174–184, 2001. URL `https://doi.org/10.1198/016214501750333045`.

Gregory F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224, 1997. URL `https://doi.org/10.1023/A:1009787925236`.

Sarah Filippi and Chris C. Holmes. A Bayesian nonparametric approach to testing for dependence between random variables. *Bayesian Analysis*, 12(4):919–938, 12 2017. URL `https://doi.org/10.1214/16-BA1027`.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017. URL `https://doi.org/10.1017/9781139029834`.

A. Gretton, K. Fukumizu, CH. Teo, L. Song, B. Schölkopf, and AJ. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*

20, pages 585–592. Max-Planck-Gesellschaft, Curran, September 2008.

Timothy Hanson and Wesley O. Johnson. Modeling regression error with a mixture of Pólya trees. *Journal of the American Statistical Association*, 97(460):1020–1033, 2002. URL http://www.jstor.org/stable/3085827.

Timothy E. Hanson. Inference for mixtures of finite Pólya tree models. *Journal of the American Statistical Association*, 101(476):1548–1565, 2006. URL http://www.jstor.org/stable/27639772.

Naftali Harris and Mathias Drton. PC algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(69):3365–3383, 2013. URL http://jmlr.org/papers/v14/harris13a.html.

Chris C. Holmes, François Caron, Jim E. Griffin, and David A. Stephens. Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Analysis*, 10(2):297–320, 06 2015. URL https://doi.org/10.1214/14-BA914.

Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. URL http://www.jstor.org/stable/2291091.

Seongho Kim. ppcor: An r package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods*, 22(6):665–674, 2015. URL http://doi.org/10.5351/CSAM.2015.22.6.665.

Charles H. Kraft. A class of distribution function processes which have derivatives. *Journal of Applied Probability*, 1(2):385–388, 1964. URL http://www.jstor.org/stable/3211867.

Michael Lavine. Some aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, 20(3):1222–1235, 09 1992. URL https://doi.org/10.1214/aos/1176348767.

Michael Lavine. More aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, 22(3):1161–1176, 1994. URL http://www.jstor.org/stable/2242220.

Li Ma. Recursive partitioning and multi-scale modeling on conditional densities. *Electronic Journal of Statistics*, 11(1):1297–1325, 2017. URL https://doi.org/10.1214/17-EJS1254.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020. URL http://jmlr.org/papers/v21/17-123.html.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

Joseph Ramsey and Bryan Andrews. FASK with interventional knowledge recovers edges from the sachs model. *arXiv.org preprint*, arxiv:1805.03108 [q-bio.MN], 2018.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):529–528, 2005. URL http://www.jstor.org/stable/3841298.

Marco Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software, Articles*, 35(3):1–22, 2010. URL https://www.jstatsoft.org/v035/i03.

Andrew J Sedgewick, Kristina Buschur, Ivy Shi, Joseph D Ramsey, Vineet K Raghu, Dimitris V Manatakis, Yingze Zhang, Jessica Bon, Divay Chandra, Chad Karoleski, Frank C Sciurba, Peter Spirtes, Clark Glymour, and Panayiotis V Benos. Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. *Bioinformatics*, 35(7):1204–1212, 2019. doi: 10.1093/bioinformatics/bty769. URL https://doi.org/10.1093/bioinformatics/bty769.

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, pages 2951–2961, 2017.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48(3):1514–1538, 06 2020. URL https://doi.org/10.1214/19-AOS1857.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. Springer-Verlag, 1993.

Peter Spirtes, Christopher Meek, and Thomas S. Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. In Peter Spirtes, Christopher Meek, and Thomas S. Richardson, editors, *Computation, causation, and discovery*, chapter 6, page 211–252. The MIT Press, Cambridge, Massachusetts, 1999.

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. URL https://doi.org/10.1515/jci-2018-0017.

Onur Teymur and Sarah Filippi. A Bayesian nonparametric test for conditional independence. *Foundations of Data Science*, 2(2):155–172, 2020.

Stephen Walker and Bani K. Mallick. A Bayesian semiparametric accelerated failure time model. *Biometrics*, 55 (2):477–483, 1999. URL http://www.jstor.org/stable/2533795.

Ruud Wetzels and Eric-Jan Wagenmakers. A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, 19(6):1057–1064, 2012. URL https://doi.org/10.3758/s13423-012-0295-x.

Wing H. Wong and Li Ma. Optional Pólya tree and Bayesian inference. *The Annals of Statistics*, 38(3):1433–1459, 06 2010. URL https://doi.org/10.1214/09-AOS755.

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