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

**Philip A. Boeken**
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Amsterdam, The Netherlands
philip.boeken@student.uva.nl

**Joris M. Mooij**
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Amsterdam, The Netherlands
j.m.mooij@uva.nl

## Abstract

The performance of constraint-based causal discovery algorithms is prominently determined by the performance of the (conditional) independence tests that are being used. A default choice for the (conditional) independence test is the (partial) correlation test, which can fail in presence of nonlinear relations between the variables. Recent research proposes a Bayesian nonparametric two-sample test (Holmes et al., 2015), an independence test between continuous variables (Filippi and Holmes, 2017), and a conditional independence test between continuous variables (Teymur and Filippi, 2019). We extend this work by proposing a novel Bayesian nonparametric conditional two-sample test. We utilise this conditional two-sample test for testing the conditional independence $C \perp\!\!\!\perp Y | X$ where $C$ denotes a Bernoulli random variable, and $X$ and $Y$ are continuous one-dimensional random variables. This enables a nonparametric implementation of the Local Causal Discovery (LCD) algorithm with binary variables in the experimental setup (e.g. an indicator of treatment/control group). We propose a fair performance measure for comparing frequentist and Bayesian tests in the LCD setting. We utilise this performance measure for comparing our Bayesian ensemble with state-of-the-art frequentist tests, and conclude that the Bayesian ensemble has better performance than its frequentist counterparts. We apply our nonparametric implementation of the LCD algorithm to protein expression data.

## 1 Introduction

Conditional independence testing is a fundamental ingredient of causal inference algorithms (Cooper, 1997; Spirtes et al., 1999). 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 marginal and conditional independence tests that are being used. Conditional independence testing has been proven to be impossible when no additional assumptions are imposed on the distributions involved (Shah and Peters, 2020). Over the years, multiple conditional independence tests have been proposed, each of which require additional assumptions. A class of independence tests that require relatively lenient assumptions are the Bayesian nonparametric independence tests that are based on Pólya tree priors: random measures which have Kullback-Leibler support on the entire space of continuous distributions (Lavine, 1994). Therefore the only assumption is that the data generating process has an absolutely continuous cumulative distribution function.

Among the independence tests that are based on Pólya tree priors is a recently proposed conditional independence test (Teymur and Filippi, 2019) which extends a continuous marginal independence test (Filippi and Holmes, 2017) by utilising conditional optional Pólya trees (Ma, 2017). Although

Preprint. Under review.

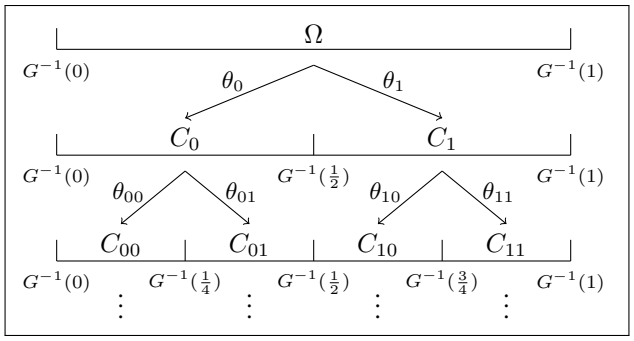



Figure 1: A one-dimensional Pólya tree partitioning scheme

this conditional independence test performs well on synthetic data originating from a continuous distribution, we encountered difficulties when applying it to combinations of discrete and continuous variables. Since causal inference algorithms are often applied to data sets with binary or discrete variables (e.g. indicating certain interventions that have been administered within an experiment (Mooij et al., 2020)), designing and evaluating such tests is a relevant area of research.

In this paper we propose a novel conditional two-sample test by extending a two-sample test based on a Pólya tree prior (Holmes et al., 2015) using a conditional optional Pólya tree prior. We empirically compare the two-sample test (Holmes et al., 2015), independence test (Filippi and Holmes, 2017) and our novel conditional two-sample test 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 Local Causal Discovery (LCD) causal inference algorithm.[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 significantly outperforms other state-of-the-art (conditional) independence tests, while computation time is substantially lower than for competing tests that have been proposed.

We apply the LCD algorithm with the Pólya tree tests on 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

We will test for hypotheses where we assume the observed random variable $X$ to be distributed according to a distribution lying in $\mathcal{M}_0$ or $\mathcal{M}_1$ under the null- and alternative hypothesis respectively. As prior distributions $\mathcal{P}_0$ and $\mathcal{P}_1$ on $\mathcal{M}_0$ and $\mathcal{M}_1$ we take Pólya trees: random measures which, under proper choice of hyperparameters, have Kullback-Leibler support on $\mathcal{M}$ (Lavine, 1994). To construct a Pólya tree random measure on $\Omega \subseteq \mathbb{R}$, we consider a cumulative distribution function $G$ on $\Omega$, and map the family of dyadic partitions of $[0, 1]$ through $G^{-1}$. This results in a family of partitions of $\Omega$, where for level $j$ we have $\Omega = \bigcup_{\kappa \in \{0,1\}^j} C_\kappa$, with

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

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. We define the index set by $K := \{\{0, 1\}^j : j \in \mathbb{N}\}$, so the family of subsets of $\Omega$ that we consider is $\Pi := \{C_\kappa : \kappa \in K\}$. We assign random probabilities to the $C_\kappa \in \Pi$ by splitting from the mass that is assigned to $C_\kappa$ a fraction $\theta_{\kappa 0}$ to $C_{\kappa 0}$ and a fraction $\theta_{\kappa 1}$ to $C_{\kappa 1}$, where we let $(\theta_{\kappa 0}, \theta_{\kappa 1}) \sim \text{Dir}(\alpha_{\kappa 0}, \alpha_{\kappa 1})$. This construction yields the following formal definition:

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

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

1. *all the random variables in $\Theta$ are independent;*
2. *for every $\kappa \in K$, we have $(\theta_{\kappa 0}, \theta_{\kappa 1}) \sim \text{Dir}(\alpha_{\kappa 0}, \alpha_{\kappa 1})$;*
3. *for every $j \in \mathbb{N}$ and every $\kappa \in \{0,1\}^j$ we have $\mathcal{P}(C_\kappa|\Theta) = \prod_{i=1}^{j} \theta_{\kappa_1 \ldots \kappa_{i-1}}$, for $C_\kappa \in \Pi$.*

The support of the Pólya tree is determined by the choice of $\Pi$ and $\mathcal{A}$. We will only consider partitions of the type (1) with $G$ the standard Gaussian cumulative distribution function. Sufficient conditions on $\mathcal{A}$ for the Pólya tree to have support on the continuous distributions are provided by Kraft (1964). These conditions are satisfied if for each $\kappa \in \{0,1\}^j$ we take $\alpha_{\kappa 0} = \alpha_{\kappa 1} = j^2$, which we will use throughout this paper, as it is promoted as a "sensible canonical choice" by Lavine (1992). For this set of parameters $(\Pi, \mathcal{A})$ the Pólya tree is centred on the base distribution with cumulative distribution function $G$, i.e. $\mathbb{E}(\mathcal{P}(C_\kappa)) = \int_{C_\kappa} G'(x)dx$. 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 (Teymur and Filippi, 2019). Hanson and Johnson (2002) provide a 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.

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

$$p(X_{1:n}|\Theta, \Pi, \mathcal{A}) = \prod_{\kappa \in K} \theta_\kappa^{n_{\kappa 0}} (1 - \theta_\kappa)^{n_{\kappa 1}}, \tag{2}$$

where $n_\kappa$ denotes the number of observations lying in $C_\kappa$, i.e. $n_\kappa := \#(X_{1:n} \cap C_\kappa)$. If we integrate out $\Theta$ we obtain the marginal likelihood

$$p(X_{1:n}|\Pi, \mathcal{A}) = \prod_{\kappa \in K} \frac{B(\alpha_{\kappa 0} + n_{\kappa 0}, \alpha_{\kappa 1} + n_{\kappa 1})}{B(\alpha_{\kappa 0}, \alpha_{\kappa 1})}, \tag{3}$$

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

### 2.1 A nonparametric conditional two-sample test

We now propose a novel 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 variable. We combine elements of both the two-sample test proposed by Holmes et al. (2015), and a conditional independence test from Teymur and Filippi (2019). The conditional independence test from Teymur and Filippi (2019) is of the type $X \perp\!\!\!\perp Y|Z$ for $(X, Y, Z)$ a triple of continuous, one-dimensional random variables, and produces no sensible results when testing for $C \perp\!\!\!\perp X|Z$ in our experimental setup of Section 3.1.

Given $n$ draws $\{(C_1, X_1, Z_1), ..., (C_n, X_n, Z_n)\}$ from binary variable $C$ and continuous one-dimensional random variables $X$ and $Z$, define $X^{(0)} := \{X_i : C_i = 0, i \in \{1, .., n\}\}$ and $X^{(1)} := \{X_i : C_i = 1, i \in \{1, .., n\}\}$. If we let $F$ denote the conditional distribution of $X|Z$ and let $F^{(0)}$ (resp. $F^{(1)}$) denote the conditional distribution of $X^{(0)}|Z$ (resp. $X^{(1)}|Z$), then we formulate the conditional independence between $C$ and $X$ given $Z$ as a two-sample test, i.e.

$$H_0 : C \perp\!\!\!\perp X|Z \iff X|Z \sim F \tag{4}$$

$$H_1 : C \not\perp\!\!\!\perp X|Z \iff X^{(0)}|Z \sim F^{(0)} \text{ and } X^{(1)}|Z \sim F^{(1)} \text{ with } F^{(0)} \neq F^{(1)}. \tag{5}$$

Teymur and Filippi (2019) utilise conditional optional Pólya tree (cond-OPT) priors (Ma, 2017) for modelling conditional densities of the form $f_{X|Z}(x|z)$. As we require an expression for the marginal likelihoods of $X|Z, X^{(0)}|Z$ and $X^{(1)}|Z$, we briefly review the construction of the cond-OPT.

We wish to construct a random measure on the space of conditional distributions on $\Omega_X \times \Omega_Z$, where $X$ is the response variable and $Z$ is the predictor. In order to do so, we first construct a family of partitions $\Pi_Z$ on $\Omega_Z$ according to the partitioning scheme of the optional Pólya tree (OPT) (Wong and Ma, 2010), which results in a random subset of the standard family of one-dimensional partitions $\Pi$ as constructed by equation (1). This random subset of $\Pi$ is obtained by first adding $C_\emptyset := \Omega_Z$ to $\Pi_Z$. Then we sample from the random variable $S \sim \text{Bernoulli}(\rho)$; if $S = 1$ we stop the partitioning procedure, and if $S = 0$ we add $C_0$ and $C_1$ to $\Pi_Z$. Then, for both $C_0$ and $C_1$ we repeat this procedure; we draw from $S$ and depending on the outcome we add the children of $C_0$, then we repeat this to possibly add the children of $C_1$. This process is iterated, and is stopped in finite time when $\rho > 0$.

When we have obtained the family $\Pi_Z$, we construct a random measure $\mathcal{P}(\cdot|C_\kappa)$ on $\Omega_X$ for each $C_\kappa \in \Pi_Z$ by letting $\mathcal{P}(\cdot|C_\kappa) \sim \text{PT}(\mathcal{A}, \Pi)$. This family of random measures on $\Omega_X$ is the resulting conditional optional Pólya tree (cond-OPT) (Ma, 2017). In case the family of partitions $\Pi_Z$ generates the Borel sets on $\Omega_Z$, this construction indeed yields a random conditional probability measure on $\Omega_X \times \Omega_Y$. Ma (2017) proves that any conditional density $f(\cdot|\cdot)$ on $\Omega_X \times \Omega_Z$ is in the $L^1$-closure of the support of the cond-OPT.

From the perspective of hypothesis testing, we are interested in the marginal likelihood of a sample $(X_1, Z_1), ..., (X_n, Z_n)$ with respect to a cond-OPT prior. This is obtained by first standardising $Z_{1:n}$, and for every $C_\kappa \in \Pi_Z$ considering the subsample $X(C_\kappa) := \{X_j : Z_j \in C_\kappa\}$, and $X^{(0)}(C_\kappa), X^{(1)}(C_\kappa)$ defined similarly. As the cond-OPT prior considers a general Pólya tree prior for this subsample, we simply compute the marginal likelihood $p(X(C_\kappa)|\Pi, \mathcal{A})$ using equation (3). If $C_\kappa$ 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 $\Pi_Z$, then we simply return this marginal likelihood. If $C_\kappa$ is not a leaf-set, we continue along the children $C_{\kappa 0}$ and $C_{\kappa 1}$. We integrate out the randomness of the random family of partitions by considering the entire family of partitions $\Pi_Z$ of $\Omega_Z$ according to equation (1), and incorporating the stopping probabilities $S$ by weighing the elements $C_\kappa$ of level $j$ with $\mathbb{E}(1-S)^j = (1-\rho)^j$. The recursive mixing formula is given by

$$\Phi(X|C_\kappa) := \begin{cases} p(X(C_\kappa)|\Pi, \mathcal{A}) & \text{if } C_\kappa \text{ is a leaf-set} \\ \rho \cdot p(X(C_\kappa)|\Pi, \mathcal{A}) + (1-\rho) \cdot \Phi(X|C_{\kappa 0})\Phi(X|C_{\kappa 1}) & \text{otherwise.} \end{cases} \tag{6}$$

The quantity $\Phi(X|C_\kappa)$ is the marginal likelihood of $\{(X_1, Z_1), ..., (X_n, Z_n)\} \cap \Omega_X \times C_\kappa$, with respect to the cond-OPT. We repeat this computation for $X^{(0)}$ and $X^{(1)}$, and compute the Bayes factor

$$\text{BF}(H_0, H_1) = \frac{\Phi(X|\Omega_Z)}{\Phi(X^{(0)}|\Omega_Z)\Phi(X^{(1)}|\Omega_Z)}. \tag{7}$$

Throughout this paper we employ $\rho = 1/2$ (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 $C_\kappa$ to be a leaf-set if it contains at most one value, or if the number of digits in $\kappa$ exceeds $\lfloor \log_4(n) \rfloor$. This maximum partitioning depth is also used when computing the Pólya tree marginal likelihoods $p(\cdot|\Pi, \mathcal{A})$.

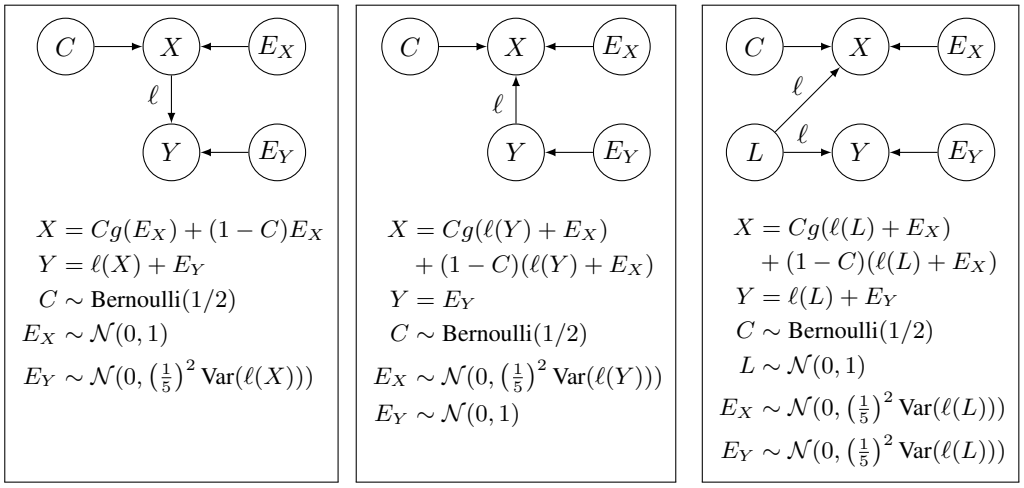
We note that when no data is provided for $Z$ and thus $\Omega_Z$ constitutes a leaf-set, this test defaults to the two-sample test from Holmes et al. (2015). An overview of the two-sample test (Holmes et al., 2015) and the continuous independence test (Filippi and Holmes, 2017) is provided in the supplement. We note that we deviate from the original tests by only considering partitions until a maximum depth of $\lfloor \log_4(n) \rfloor$. In all tests we standardise the data, and use a standard Gaussian base measure for generating the relevant partitions.

## 3 Experiments

As mentioned earlier, we investigate the performance of the Pólya tree prior based independence tests when implemented as part of the LCD algorithm (Cooper, 1997). To recapitulate, the LCD algorithm is based on the result that if the data generating process of the triple of random variables $(X_1, X_2, X_3)$ has no selection bias, can be modelled by a faithful structural causal model (SCM) (Pearl, 2009), and $X_2$ is not a cause of $X_1$, then the presence of (in)dependences

$$X_1 \not\perp\!\!\!\perp X_2, \quad X_2 \not\perp\!\!\!\perp X_3, \quad X_1 \perp\!\!\!\perp X_3 | X_2 \tag{8}$$

implies that $X_2$ is a (possibly indirect) cause of $X_3$. If this is the case, we speak of the 'LCD triple' $(X_1, X_2, X_3)$. Mooij et al. (2020) have recently shown that LCD is able to deal with cyclic relations.




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

Figure 2: Three SCM's used for the simulations.

The data sets we consider consist of observations of *context* variables $(C_k)_{k \in \mathcal{K}}$ (interventions), and *system* variables $(X_i)_{i \in \mathcal{I}}$. We implement LCD by iterating over all triples $(C_k, X_i, X_{i'})$, where $k \in \mathcal{K}$ and $i \neq i' \in \mathcal{I}$. We emphasise that the LCD algorithm only retrieves ancestral relations, and thus no direct causes. As mentioned earlier, we consider all context variables to be binary variables, indicating whether or not a certain intervention has been applied to the system variables.

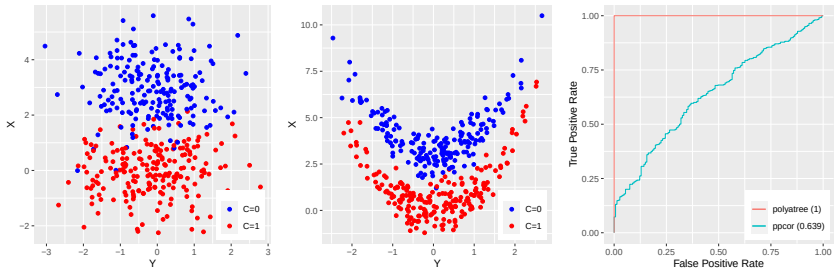
### 3.1 Simulations

We are interested in the performance of the LCD algorithm, applied to a data set which may contain nonlinear relations between the variables. To analyse the performance, we simulate data according to the graphs shown in Figure 2. We let the binary variable $C$ determine whether we perform intervention $g$ on $X$; if $C = 1$ we intervene on $X$, and if $C = 0$ we don't intervene on $X$. The link function $\ell$ and intervention $g$ are randomly chosen from

$$\ell(x) = \begin{cases} 0 \\ x \\ x^2 \\ \sin(12\pi\tilde{x}) \end{cases} \quad \text{and} \quad g(x) = \begin{cases} x & \text{no intervention} \\ x + \theta & \text{mean shift} \\ (1 + \theta)x & \text{variance shift} \\ \theta & \text{perfect intervention} \\ x + B & \text{mean shift mixture,} \end{cases} \tag{9}$$

where $\tilde{x} = x/(\max(x_1, ..., x_n) - \min(x_1, ..., x_n))$, $\theta \sim \mathcal{U}(\{2, 3, 4, 5, 6\})$ (independently drawn per round of simulations) and $B \sim \mathcal{U}(\{-1, \theta\})$ (independently drawn for every $x$). The possibility of picking $g(x) = x$ or $\ell(x) = 0$ ensures the occurrence of $C \perp\!\!\!\perp X$ and $X \perp\!\!\!\perp Y$ respectively. Note that if we pick $\ell(x) = 0$, then in the graphs of Figures 2b and 2c we have $C \perp\!\!\!\perp Y | X$. To have a balanced occurrence of $C \perp\!\!\!\perp X, X \perp\!\!\!\perp Y, C \perp\!\!\!\perp Y | X$ and the presence of an LCD triple, we pick the graph of Figure 2a with probability $3/5$, and the graphs of Figures 2b and 2c with probability $1/5$. We choose 'no intervention' for $g$ and $\ell(x) = 0$ (no connection between the variables) independently with probability $1/5$.

A canonical choice for testing conditional independence is the partial correlation test. Despite its ubiquity, its assumptions do not comply with nonlinear data. To illustrate this we generate 400 samples from the graph of Figure 2b 2000 times, where we only consider the intervention $g(x) = x + 3$ (mean shift) and link functions $\ell(x) = 0$ (Figure 3a) or $\ell(x) = x^2$ (Figure 3b). Depending on the choice of $\ell$ we either have $C \perp\!\!\!\perp X | Z$ or $C \not\perp\!\!\!\perp X | Z$. We compare the novel conditional two-sample test from Section 2.1 (denoted by `polyatree`) with Pearson's partial correlation test (denoted by `ppcor`) based on their ROC curve, as shown in Figure 3c. We see that partial correlations performs just a little better than random guessing, and that the Bayesian conditional two-sample test performs well.



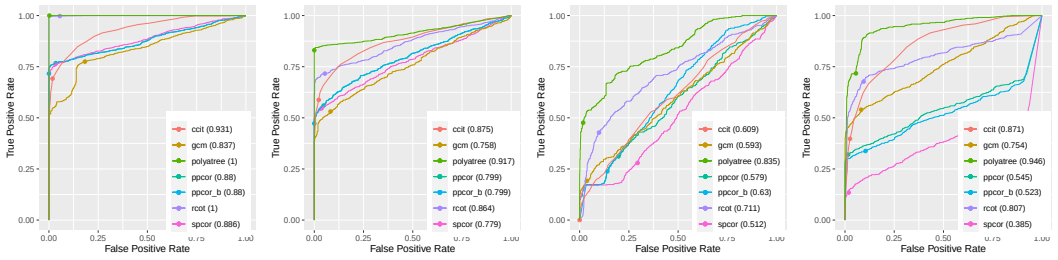

(a) No link; $\ell(x) = 0$ (b) Parabolic link; $\ell(x) = x^2$ (c) Testing $C \not\perp\!\!\!\perp Y | X$

Figure 3: An example of failure of the partial correlation test.

We compare our ensemble of Pólya tree based nonparametric tests (being 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. As mentioned earlier, we compare with the ensemble consisting of the Pearson correlation- and partial correlation test, denoted by `ppcor`. We also include a Bayesian version of the Pearson (partial) correlation test (Wetzels and Wagenmakers, 2012), denoted by `ppcor_b`. Harris and Drton (2013) propose the use of Spearman's (partial) rank correlation test for nonparanormal models, which we denote by `spcor`. We compare with a nonlinear extension of the partial correlation test, the Generalised Covariance Measure (GCM) (Shah and Peters, 2020), implemented with penalised regression splines as provided by the R-package `GeneralisedCovarianceMeasure`. We denote this (conditional) independence test 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). For marginal independence testing, this test defaults to an approximate version of the Hilbert-Schmidt Independence Criterion (Gretton et al., 2008). This ensemble is denoted by `rcot`. Finally we compare to the Classifier Conditional Independence Test (CCIT) (Sen et al., 2017), denoted by `ccit`.

We do 2000 rounds of simulations. In each round we select a graph from Figure 2, select link function $\ell$ and intervention $g$, and simulate 400 observations from the resulting graph. Then we apply each of the test ensembles to the two-sample test $C \perp\!\!\!\perp X$, the independence test $X \perp\!\!\!\perp Y$, and the conditional two-sample test $C \perp\!\!\!\perp Y | X$. 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 $\alpha$, representing the upper bound on the p-value/model evidence for drawing a positive conclusion. These results can be seen in Figures 4a, 4b and 4c. On the ROC curves we have marked the reference points $\alpha = 0.05$ and $\alpha = 1/2$ for respectively frequentist and Bayesian tests. The areas under the ROC curves (auc) are shown in the legends of the plots.

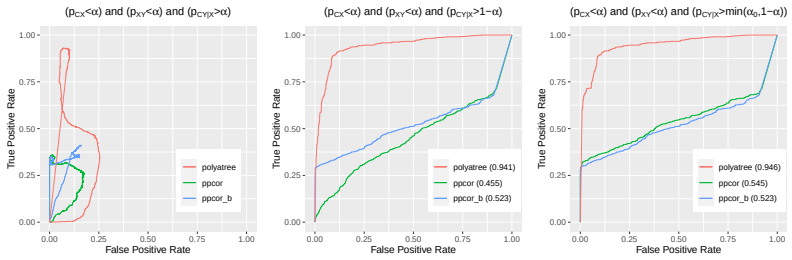
Comparing Bayesian and frequentist tests based on their performance in the LCD algorithm is not straightforward, since the triple of tests does not by default output a confidence score with respect




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

Figure 4: ROC results for simulated data for individual tests (a–c) and for the LCD test ensemble (d).

[2]Recall that $\mathbb{P}(H_0|\text{data}) = 1 - (1 + \text{BF}(H_0, H_1))^{-1}$.




(a) (b) (c)

Figure 5: ROC curves of different ways of scoring an LCD triple $(C, X, Y)$.

to the hypothesis. Typically, varying the threshold $\alpha$ 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 $\alpha$ as threshold for testing whether $p_{CX} < \alpha$, $p_{XY} < \alpha$ and $p_{CY|X} > \alpha$, then varying $\alpha$ between 0 and 1 does not result in a curve between $(0, 0)$ and $(1, 1)$, as shown in Figure 5a. Alternatively we could use $\alpha$ for testing $p_{CX} < \alpha$, $p_{XY} < \alpha$ and $p_{CY|X} > 1 - \alpha$, as shown in Figure 5b. In this case the level $\alpha$ 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 $\alpha$ 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 $\alpha$ between 0 and a fixed $\alpha_0$ (Figure 5c). More specifically, for level $\alpha$ the LCD algorithm outputs

$$p_{\text{LCD}} = \mathbb{1}_{[0,\alpha]}(p_{CX}) \cdot \mathbb{1}_{[0,\alpha]}(p_{XY}) \cdot \mathbb{1}_{(\alpha_0,1]\cup(1-\alpha,1]}(p_{CY|X}), \tag{10}$$

where we let $\alpha_0 = 0.05$ for frequentist tests and $\alpha_0 = 1/2$ for Bayesian tests. The triple $(C, X, Y)$ is given a 'positive' label if the data is generated according to the relation $C \to X \to Y$. The use of this performance measure is corroborated by the observation that in Figure 5c the frequentist partial correlation and Bayesian partial correlation tests have similar performance. We finally compare the performance of the different LCD implementations using this performance measure, as shown in Figure 4d.

We compare the computation times of the different tests in Figure 6. We note that the Pólya tree tests provide a very good trade-off between ROC performance (Figure 4) and computation time (Figure 6).

### 3.2 Protein expression data

We apply the LCD algorithm, implemented with the Bayesian ensemble of independence tests, to protein expression data (Sachs et al., 2005). For a detailed description of the data set we refer to the supplement. Sachs et al. (2005) provide an 'expert network', depicting the consensus (at that time) among biologist literature 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. 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

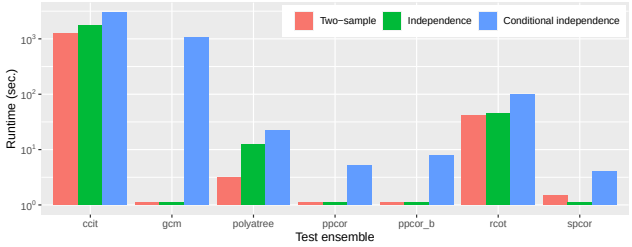



Figure 6: Runtimes of the different tests ensembles on the entire batch of simulations.

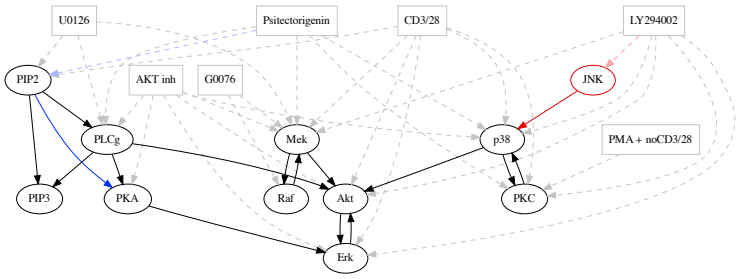



Figure 7: The output of LCD on the Sachs data. Edges indicate (indirect) causal effects between the nodes. We report relations for a Bayes Factor threshold of $k = 10$ (strong evidence, depicted in black), $k = 4$ (substantial evidence, depicted in red) and $k = 1$ (weak evidence, depicted in blue) (Kass and Raftery, 1995). Interventions on their (indirect) causal effects are indicated with light-coloured nodes and edges.

(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.

The output of the LCD algorithm, implemented with the Bayesian ensemble of tests, is shown in Figure 7. We report the output of the LCD algorithm for multiple thresholds for the statistical tests. We accept $H_1 : C \not\perp\!\!\!\perp X$ and $H_1 : X \not\perp\!\!\!\perp Y$ when $\text{BF}(H_1, H_0) = 1/\text{BF}(H_0, H_1) > k$ and accept $H_0 : C \perp\!\!\!\perp Y | X$ when $\text{BF}(H_0, H_1) > k$, and we compare the results for multiple values of $k$.

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. For this frequentist implementation we consider the threshold $\alpha = 0.01$ for testing $C \not\perp\!\!\!\perp X$, $X \not\perp\!\!\!\perp Y$ and $C \perp\!\!\!\perp Y | X$. The results of this implementation are provided in the supplement. First we note that the output of LCD differs among the use of different statistical tests, corroborating the premise that the performance of the algorithm highly depends on the choice of statistical test. We also note that LCD with partial correlations produces a very dense causal graph, whereas LCD with Pólya tree tests produces a graph which is more likely to resemble the true causal model.

## 4 Discussion & Conclusion

We note that, although the Pólya tree ensemble of independence tests provides good results in our setup, we have made some assumptions that might be reconsidered when using these tests in practice. First we note that the choice of $\mathcal{A}$ may highly influence the suitability of the test. Even when $\mathcal{A}$ satisfies the conditions such that samples from the Pólya tree are continuous distributions, there is a wide variety of parameters to be considered. Walker and Mallick (1999) for example consider $\alpha_j = cj^2$ for $c > 0$, and propose placing a prior on the parameter $c$. We note that choosing $c$ between 1 and 10 is in general a good choice (Holmes et al., 2015), where lower values of $c$ correspond with higher variance of the Pólya tree, and thus less fixation on the mean $G$ (Hanson, 2006). Another consideration is the choice of the family of partitions $\Pi$. We have quite arbitrarily picked $G$ to be standard Gaussian. As we have chosen $\alpha_j = j^2$ (causing relatively low dependence on $G$) and pre-process the data by standardising, we believe that our results are not solely valid for the data sets we considered. We also note that the maximum partitioning depth $J = \lfloor \log_4(n) \rfloor$ is quite arbitrarily chosen, and may be reconsidered when presented with either very small or very large data sets.

As many constraint-based causal inference algorithms (other than LCD) require conditional independence testing of the form $C \perp\!\!\!\perp X | Z$ for multidimensional $Z$, further research should look into how this can be achieved.

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.

## Acknowledgments and Disclosure of Funding

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

## References

Cooper, G. F. (1997). A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2(4):615–629.

Filippi, S. and Holmes, C. C. (2017). A Bayesian nonparametric approach to testing for dependence between random variables. *Bayesian Analysis*, 12(4):919–938.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2008). A kernel statistical test of independence. In *Advances in neural information processing systems 20*, pages 585–592, Red Hook, NY, USA. Max-Planck-Gesellschaft, Curran.

Hanson, T. and Johnson, W. O. (2002). Modeling regression error with a mixture of Pólya trees. *Journal of the American Statistical Association*, 97(460):1020–1033.

Hanson, T. E. (2006). Inference for mixtures of finite Pólya tree models. *Journal of the American Statistical Association*, 101(476):1548–1565.

Harris, N. and Drton, M. (2013). PC algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(1):3365–3383.

Holmes, C. C., Caron, F., Griffin, J. E., and Stephens, D. A. (2015). Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Analysis*, 10(2):297–320.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kraft, C. H. (1964). A class of distribution function processes which have derivatives. *Journal of Applied Probability*, 1(2):385–388.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, 20(3):1222–1235.

Lavine, M. (1994). More aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, 22(3):1161–1176.

Ma, L. (2017). Recursive partitioning and multi-scale modeling on conditional densities. *Electronic Journal of Statistics*, 11(1):1297–1325.

Mooij, J. M., Magliacane, S., and Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

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

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):529–528.

Sen, R., Suresh, A. T., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. (2017). Model-powered conditional independence test. In *Advances in neural information processing systems*, pages 2951–2961.

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

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

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

Teymur, O. and Filippi, S. (2019). A Bayesian nonparametric test for conditional independence. *arXiv.org preprint*, arxiv:1910.11219 [stat.ME].

Walker, S. and Mallick, B. K. (1999). A Bayesian semiparametric accelerated failure time model. *Biometrics*, 55(2):477–483.

Wetzels, R. and Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, 19(6):1057–1064.

Wong, W. H. and Ma, L. (2010). Optional Pólya tree and Bayesian inference. *The Annals of Statistics*, 38(3):1433–1459.

## Appendix A Hypothesis testing with Pólya tree priors

In general, our setup for independence testing will assume availability of independent samples $X_1, ..., X_n$ of a continuous random variable $X$. We denote the domain of $X$ with $\Omega_X$, and the space of distributions with continuous cumulative distribution functions on $\Omega_X$ with $\mathcal{M}$. Our hypotheses will be of the form

$$H_0 : X \sim f \text{ with } f \in \mathcal{M}_0, \quad H_1 : X \sim f \text{ with } f \in \mathcal{M}_1, \tag{11}$$

where $\mathcal{M}_0, \mathcal{M}_1 \subset \mathcal{M}$, and $\mathcal{M}_0 \cap \mathcal{M}_1 = \emptyset$. Since we wish to device a Bayesian test, we will define prior distributions $\mathcal{P}_0$ and $\mathcal{P}_1$ on $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively. Then we compare the evidence of the models given the data via the Bayes factor, i.e.

$$\mathrm{BF}(H_0, H_1) = \frac{\mathbb{P}(H_0|X_{1:n})}{\mathbb{P}(H_1|X_{1:n})} = \frac{p(X_{1:n}|H_0)}{p(X_{1:n}|H_1)} \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} = \frac{\int_{\mathcal{M}_0} \prod_{i=1}^n f(X_i) d\mathcal{P}_0(f)}{\int_{\mathcal{M}_1} \prod_{i=1}^n f(X_i) d\mathcal{P}_1(f)} \tag{12}$$

where we have placed equal prior weights on $H_0$ and $H_1$, so $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$.

A canonical choice for a prior on a space of probability distributions is the Dirichlet Process. However, the support of the Dirichlet process contains only discrete distributions, and since we wish $\mathcal{M}$ to be contained in the support of our prior, the Dirichlet Process is not a suitable choice for our setup. The Pólya tree prior does not suffer from this characteristic (Ferguson, 1974), and is thus a suitable prior on $\mathcal{M}$. Since the elements of $\mathcal{M}$ have support on $\Omega_X$, we will speak of a Pólya tree on $\Omega_X$. We will first construct a Pólya tree on $\Omega_X \subseteq \mathbb{R}$, and then extend this definition to a Pólya tree on $\Omega_X \times \Omega_Y \subseteq \mathbb{R}^2$.

First we recall the construction of the one-dimensional Pólya tree as described in the main paper. In particular, we construct a Pólya tree on $(\Omega, \mathcal{B}(\Omega))$, where $\Omega \subseteq \mathbb{R}$, and $\mathcal{B}(\Omega)$ denotes the Borel sigma-algebra on $\Omega$. In order to construct a random measure on $\mathcal{B}(\Omega)$, we will assign random probabilities to a family of subsets $\Pi$ of $\Omega$ which generates the Borel sets. The family of subsets that we consider are the dyadic partitions of $[0, 1]$, mapped under the inverse of some cumulative distribution function $G$ on $\Omega$. This results in a family of partitions of $\Omega$, where for level $j$ we have $\Omega = \bigcup_{\kappa \in \{0,1\}^j} C_\kappa$, with

$$C_\kappa := [G^{-1}(\tfrac{k-1}{2^j}), G^{-1}(\tfrac{k}{2^j})), \tag{13}$$

and $k$ is the natural number corresponding to the bit string $\kappa \in \{0, 1\}^j$. A schematic depiction of this binary tree of partitions is shown in Figure 8. We define the index set by $K := \{\{0, 1\}^j : j \in \mathbb{N}\}$, so the family of subsets of $\Omega$ that we consider is $\Pi := \{C_\kappa : \kappa \in K\}$. From basic measure theory we know that $\Pi$ indeed generates $\mathcal{B}(\Omega)$. We assign random probabilities to the elements of $\Pi$ by first assigning random probabilities to $C_0$ and $C_1$, and randomly subdividing these masses among the children of $C_0$ and $C_1$. In particular, for the first level of the partition we assign the random probabilities $\mathcal{P}(C_0) = \theta_0$ and $\mathcal{P}(C_1) = \theta_1$ with $(\theta_0, \theta_1) \sim \mathrm{Dir}(\alpha_0, \alpha_1)$, for some hyper-parameters $\alpha_0$ and $\alpha_1$. Then, for every $C_\kappa \in \Pi$ we split the mass that is assigned to $C_\kappa$ by assigning a fraction $\theta_{\kappa 0}$ to $C_{\kappa 0}$ and a fraction $\theta_{\kappa 1}$ to $C_{\kappa 1}$, where we let $(\theta_{\kappa 0}, \theta_{\kappa 1}) \sim \mathrm{Dir}(\alpha_{\kappa 0}, \alpha_{\kappa 1})$. This construction yields a Pólya tree on $\Omega$, which is a random measure on $\Pi$ and thus on $\mathcal{B}(\Omega)$, which we formalise as follows:

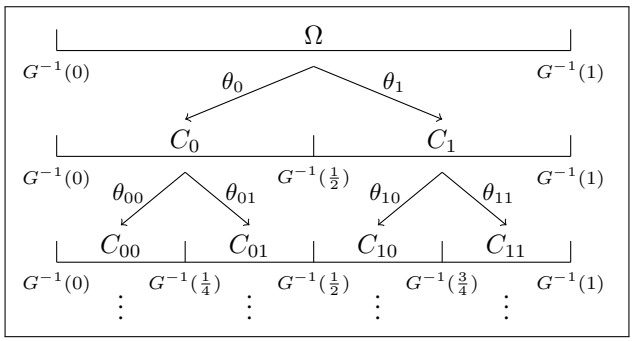



Figure 8: A one-dimensional Pólya tree partitioning scheme.

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

1. *all the random variables in $\Theta$ are independent;*
2. *for every $\kappa \in K$, we have $(\theta_{\kappa 0}, \theta_{\kappa 1}) \sim \text{Dir}(\alpha_{\kappa 0}, \alpha_{\kappa 1})$;*
3. *for every $j \in \mathbb{N}$ and every $\kappa \in \{0,1\}^j$ we have $\mathcal{P}(C_\kappa|\Theta) = \prod_{i=1}^{j} \theta_{\kappa_1 \ldots \kappa_{i-1}}$, for $C_\kappa \in \Pi$.*

The support of the Pólya tree is determined by the choice of $\Pi$ and $\mathcal{A}$. In general, any separating binary tree of partitions of $\Omega$ can be considered. In this paper we only consider partitions of the type of equation (13), where we choose $G$ to be the standard Gaussian cumulative distribution function. Ferguson (1974) shows that the Pólya tree is a Dirichlet process if $\alpha_\kappa = \alpha_{\kappa 0} + \alpha_{\kappa 1}$. The parameter of this Dirichlet process is $mG_0$, where $m = \alpha_\emptyset$ and $G_0$ is the mean of the Pólya tree, i.e. $G_0(C_\kappa) = \mathbb{E}(\mathcal{P}(C_\kappa))$ (Lavine, 1994). This implies that for this choice of $\mathcal{A}$, the support of the Pólya tree is contained in the space of discrete distributions. Sufficient conditions on $\mathcal{A}$ for the Pólya tree to have support on the continuous distributions is given by the following theorem:

**Theorem A.1 (Kraft, 1964)** *Let $\bar{\sigma}_j := \sup\{Var(\theta_\kappa) : \kappa \in \{0,1\}^j\}$. If $\mathbb{E}(\theta_\kappa) = 1/2$ for all $\kappa \in K$ and $\sum_{j=1}^{\infty} \bar{\sigma}_j < \infty$, then with probability one, samples from $\mathcal{P}$ are absolutely continuous with respect to Lebesgue measure.*

This condition is satisfied if for each $\kappa \in \{0,1\}^j$ we take $\alpha_{\kappa 0} = \alpha_{\kappa 1} = j^2$, which we will use throughout this paper, as it is promoted as a 'sensible canonical choice' by Lavine (1992). In this case we indeed have $\mathbb{E}(\theta_\kappa) = 1/2$, and thus for every $j \in \mathbb{N}$, the mass is (in expectation) split uniformly over the $C_\kappa$ for all $\kappa \in \{0,1\}^j$. As a consequence the Pólya tree is centred on the base distribution with cumulative distribution function $G$, i.e. $\mathbb{E}(\mathcal{P}(C_\kappa)) = \int_{C_\kappa} G'(x)dx$. As mentioned in the main paper we only consider partitions up to a pre-determined level $J = \lfloor \log_4(n) \rfloor$.

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

$$p(X_{1:n}|\Theta, \Pi, \mathcal{A}) = \prod_{\kappa \in K} \theta_\kappa^{n_{\kappa 0}} (1 - \theta_\kappa)^{n_{\kappa 1}}, \tag{14}$$

where $n_\kappa$ denotes the number of observations lying in $C_\kappa$, i.e. $n_\kappa := \#(\{X_1, ..., X_n\} \cap C_\kappa)$. If we integrate over all possible values of all $\theta_\kappa$, we obtain the marginal likelihood

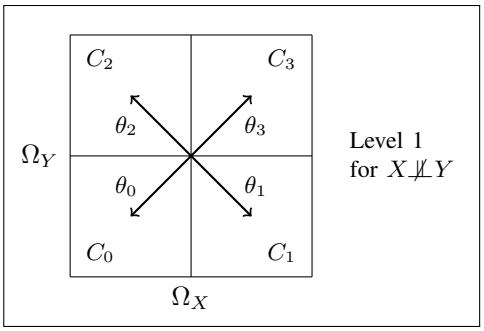
$$p(X_{1:n}|\Pi, \mathcal{A}) = \prod_{\kappa \in K} \frac{B(\alpha_{\kappa 0} + n_{\kappa 0}, \alpha_{\kappa 1} + n_{\kappa 1})}{B(\alpha_{\kappa 0}, \alpha_{\kappa 1})}, \tag{15}$$

where $B(\cdot)$ denotes the Beta function. Note that this quantity corresponds to the marginal likelihood $\int_{\mathcal{M}} \prod_{i=1}^{n} f(X_i) d\mathcal{P}(f)$, a version of which occurs in the numerator and denominator of the right-hand side of equation (12). This marginal likelihood will therefore be a fundamental quantity in the Bayesian tests that we consider.

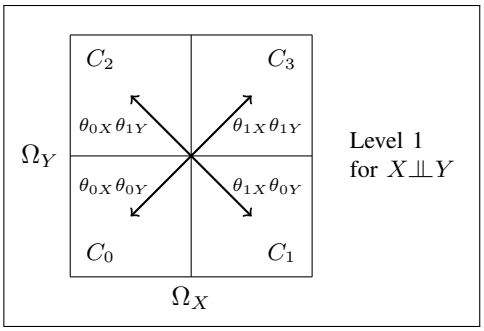
### A.1 A nonparametric two-sample test

In order to use the Pólya tree prior for Bayesian testing, we have to formulate our hypotheses $H_0$ and $H_1$ in terms of the relevant spaces of distributions $\mathcal{M}_0$ and $\mathcal{M}_1$, as suggested by equation (11). This is done by picking a Pólya tree prior $\mathcal{P}_i$ under $H_i$, and defining $\mathcal{M}_i$ to be the support of $\mathcal{P}_i$, for $i = 0, 1$. Given data to test our hypothesis with, we calculate marginal likelihoods via equation (15) for both Pólya trees $\mathcal{P}_0$ and $\mathcal{P}_1$, which are in turn used for calculating the Bayes factor via (12).

We first use this procedure to describe the nonparametric two-sample test, as proposed by Holmes et al. (2015). Given a sample $\{(X_1, C_1), ..., (X_n, C_n)\}$ from binary variable $C$ and continuous variable $X$, define $X^{(0)} := \{X_i : C_i = 0, i = 1, .., n\}$ and $X^{(1)} := \{X_i : C_i = 1, i = 1, .., n\}$. Let




(a) Partitioning scheme for $X \not\perp\!\!\!\perp Y$.




(b) Partitioning scheme for $X \perp\!\!\!\perp Y$.

$F$ denote the distribution of $X$, and let $F^{(0)}$ (resp. $F^{(1)}$) denote the distribution of $X^{(0)}$ (resp. $X^{(1)}$). We formulate the independence between $X$ and $C$ as a two-sample test, i.e.

$$H_0 : X \perp\!\!\!\perp C \iff X^{(0)} \sim F \text{ and } X^{(1)} \sim F \iff X \sim F \tag{16}$$

$$H_1 : X \not\perp\!\!\!\perp C \iff X^{(0)} \sim F^{(0)} \text{ and } X^{(1)} \sim F^{(1)} \text{ with } F^{(0)} \neq F^{(1)}. \tag{17}$$

Under $H_0$ we standardise the sample $X_{1:n}$, and compute its marginal likelihood using equation (15). Under $H_1$, we model $X^{(0)}$ and $X^{(1)}$ as being samples from independent random variables, having different distributions. Since separately normalising $X^{(0)}$ and $X^{(1)}$ may erase distinctive features between the samples, we first standardise $X$, and then subdivide $X$ into $X^{(0)}$ and $X^{(1)}$.

We formulate the Bayes factor as

$$\text{BF}(H_0, H_1) = \frac{p(X_{1:n}|\Pi, \mathcal{A})}{p(X^{(0)}|\Pi, \mathcal{A})p(X^{(1)}|\Pi, \mathcal{A})}. \tag{18}$$

Upon inspection of equation (15) we see that the Bayes factor can be written as an infinite product of fractions, being

$$\text{BF}(H_0, H_1) = \prod_{\kappa \in K} \frac{B(\alpha_{\kappa 0} + n_{X|\kappa 0}, \alpha_{\kappa 1} + n_{X|\kappa 1})B(\alpha_{\kappa 0}, \alpha_{\kappa 1})}{B(\alpha_{\kappa 0} + n_{X^{(0)}|\kappa 0}, \alpha_{\kappa 1} + n_{X^{(0)}|\kappa 1})B(\alpha_{\kappa 0} + n_{X^{(1)}|\kappa 0}, \alpha_{\kappa 1} + n_{X^{(1)}|\kappa 1})}, \tag{19}$$

where $n_{X|\kappa} := \#(X_{1:n} \cap C_\kappa)$. We note that whenever $n_{X|\kappa} \leq 1$ the fraction has a value of 1, so we calculate the marginal likelihoods until we either reach the maximum partitioning depth $\lfloor \log_4(n) \rfloor$, or until $n_{\cdot|\kappa} \leq 1$.

### A.2 Two-dimensional Pólya trees

Now that we have defined a Pólya tree on $(\Omega, \mathcal{B}(\Omega))$ with $\Omega \subseteq \mathbb{R}$, we extend this definition to a Pólya tree on $(\Omega_X \times \Omega_Y, \mathcal{B}(\Omega_X \times \Omega_Y))$ with $\Omega_X \times \Omega_Y \subseteq \mathbb{R}^2$. This construction is done similarly to the construction on $\Omega$. We consider a base measure with cumulative distribution function $G$ on $\Omega_X \cup \Omega_Y$, and partition $\Omega_X \times \Omega_Y$ into the four quadrants $C_0, C_1, C_2$ and $C_3$, where the boundaries of the $C_i$ are determined by $G^{-1}$. We assign random probability $\theta_i$ to quadrant $C_i$ with $(\theta_0, ..., \theta_3) \sim \text{Dir}(\alpha_0, ..., \alpha_3)$. Then we recursively partition $C_\kappa$ into quadrants $C_{\kappa 0}, ..., C_{\kappa 3}$, and split the mass assigned to $C_\kappa$ according to $(\theta_{\kappa 0}, ..., \theta_{\kappa 3}) \sim \text{Dir}(\alpha_{\kappa 0}, ..., \alpha_{\kappa 3})$. This partitioning scheme is shown in Figure 9a. Similar to the one-dimensional case we have that $\Pi_2$ generates the Borel sigma algebra on $\Omega_X \times \Omega_Y$. We will denote this two dimensional partition with $\Pi_2$, the set of parameters $\alpha_\kappa$ with $\mathcal{A}_2$, and the set of splitting variables $\theta_\kappa$ with $\Theta_2$, where the subscript $_2$ emphasises the dimension of the space $\Omega_X \times \Omega_Y$. This leads to the following definition of the two dimensional Pólya tree:

**Definition A.2 (Hanson (2006))** *A random probability measure $\mathcal{P}$ on $(\Omega_X \times \Omega_Y, \mathcal{B}(\Omega_X \times \Omega_Y))$ is said to have a Pólya tree distribution with parameter $(\Pi_2, \mathcal{A}_2)$, written $\mathcal{P} \sim \text{PT}(\Pi_2, \mathcal{A}_2)$, if there exist nonnegative numbers $\mathcal{A}_2 = \{\alpha_\kappa : \kappa \in K_2\}$ and random variables $\Theta_2 = \{\theta_\kappa : \kappa \in K_2\}$ such that the following hold:*

1. *all the random variables in $\Theta_2$ are independent;*
2. *for every $\kappa \in K_2$ we have $(\theta_{\kappa 0}, \theta_{\kappa 1}, \theta_{\kappa 2}, \theta_{\kappa 3}) \sim \text{Dir}(\alpha_{\kappa 0}, \alpha_{\kappa 1}, \alpha_{\kappa 2}, \alpha_{\kappa 3})$;*
3. *for every $j \in \mathbb{N}$ and every $\kappa \in \{0,1,2,3\}^j$ we have $\mathcal{P}(C_\kappa|\Theta_2) = \prod_{i=1}^{j} \theta_{\kappa_1 \ldots \kappa_{i-1}}$, for $C_\kappa \in \Pi$.*

Similarly to the one-dimensional case, samples from the Pólya tree $\mathcal{P} \sim \text{PT}(\Pi_2, \mathcal{A}_2)$ are continuous with respect to the two-dimensional Lebesgue measure if we take $\alpha_{\kappa 0} = \alpha_{\kappa 1} = \alpha_{\kappa 2} = \alpha_{\kappa 3} = j^2$, where $j$ denotes the number of bits in the binary number $\kappa \in K$ (Walker and Mallick, 1999). Similar to the one-dimensional case, we only consider partitions up to depth $J = \lfloor \log_4(n) \rfloor$.

When observing a sample $(X_1, Y_1), ..., (X_n, Y_n)$ from continuous random variables $X$ and $Y$ of which the joint distribution lies in the support of the two-dimensional Pólya tree $\mathcal{P}$, we have that the marginal likelihood of that sample is

$$p((X,Y)_{1:n}|\Theta_2, \Pi_2, \mathcal{A}_2) = \prod_{\kappa \in K} \theta_{\kappa 0}^{n_{\kappa 0}} \theta_{\kappa 1}^{n_{\kappa 1}} \theta_{\kappa 2}^{n_{\kappa 2}} \theta_{\kappa 3}^{n_{\kappa 3}}. \tag{20}$$

If we integrate over all possible values of all $\theta_\kappa$, we obtain

$$p((X,Y)_{1:n}|\Pi_2, \mathcal{A}_2) = \prod_{\kappa \in K} \frac{\tilde{B}(n_{\kappa 0} + \alpha_{\kappa 0}, n_{\kappa 1} + \alpha_{\kappa 1}, n_{\kappa 2} + \alpha_{\kappa 2}, n_{\kappa 3} + \alpha_{\kappa 3})}{\tilde{B}(\alpha_{\kappa 0}, \alpha_{\kappa 1}, \alpha_{\kappa 2}, \alpha_{\kappa 3})}, \tag{21}$$

where $\tilde{B}$ denotes the multivariate Beta function.[3] Similar to the one-dimensional case, we note that this quantity corresponds to the marginal likelihood $\int_{\mathcal{M}_2} \prod_{i=1}^{n} f(X_i, Y_i) d\mathcal{P}(f)$.

Under the assumption $X \perp\!\!\!\perp Y$, we construct a prior similar to the two-dimensional Pólya tree. First we note that the two-dimensional family of partitions $\Pi_2$ can be regarded as the per-level Cartesian product of the partitions, i.e.

$$\Pi_2 = \bigcup_{j \in \mathbb{N}} \{C_{X|\kappa} \times C_{Y|\kappa'} : C_{X|\kappa} \in \Pi_X, C_{Y|\kappa'} \in \Pi_Y, \kappa, \kappa' \in \{0,1\}^j\}, \tag{22}$$

where $\Pi_X$ and $\Pi_Y$ are families of one-dimensional partitions over $\Omega_X$ and $\Omega_Y$ respectively. For every level $\kappa$, we first split the mass over the elements of $\Pi_X$ according to $(\theta_{\kappa 0,X}, \theta_{\kappa 1,X}) \sim \text{Dir}(\alpha_{\kappa 0,X}, \alpha_{\kappa 1,X})$, and then independently split the mass over the elements of $\Pi_Y$ according to $(\theta_{\kappa 0,Y}, \theta_{\kappa 1,Y}) \sim \text{Dir}(\alpha_{\kappa 0,Y}, \alpha_{\kappa 1,Y})$. We denote the set of parameters $\alpha_{\kappa,X}$ with $\mathcal{A}_X$, and the parameters $\alpha_{\kappa,Y}$ with $\mathcal{A}_Y$. This prior yields a marginal likelihood of

$$\begin{aligned} p((X,Y)_{1:n}|\Pi_2, \mathcal{A}_X, \mathcal{A}_Y) = \prod_{\kappa \in K} & \frac{B(n_{\kappa 0} + n_{\kappa 2} + \alpha_{\kappa 0,X}, n_{\kappa 1} + n_{\kappa 3} + \alpha_{\kappa 1,X})}{B(\alpha_{\kappa 0,X}, \alpha_{\kappa 1,X})} \\ & \times \frac{B(n_{\kappa 0} + n_{\kappa 1} + \alpha_{\kappa 0,Y}, n_{\kappa 2} + n_{\kappa 3} + \alpha_{\kappa 1,Y})}{B(\alpha_{\kappa 0,Y}, \alpha_{\kappa 1,Y})}, \end{aligned} \tag{23}$$

as shown by Filippi and Holmes (2017). We notice that this equals the product of the marginal likelihoods of $X$ and $Y$ according to independent one-dimensional Pólya tree priors $\mathcal{P}_X \sim \text{PT}(\Pi_X, \mathcal{A}_X)$ on $\Omega_X$ and $\mathcal{P}_Y \sim \text{PT}(\Pi_Y, \mathcal{A}_Y)$ on $\Omega_Y$, i.e.

$$p((X,Y)_{1:n}|\Pi_2, \mathcal{A}_X, \mathcal{A}_Y) = p(X_{1:n}|\Pi_X, \mathcal{A}_X) p(Y_{1:n}|\Pi_Y, \mathcal{A}_Y), \tag{24}$$

where the univariate marginal likelihoods are computed according to equation (15). To ensure that this prior is not biased when considered in conjunction with the two-dimensional Pólya tree, we consider parameters $\alpha_{\kappa 0,X} = \alpha_{\kappa 0} + \alpha_{\kappa 2}$, $\alpha_{\kappa 1,X} = \alpha_{\kappa 1} + \alpha_{\kappa 3}$, $\alpha_{\kappa 0,Y} = \alpha_{\kappa 0} + \alpha_{\kappa 1}$ and $\alpha_{\kappa 1,Y} = \alpha_{\kappa 2} + \alpha_{\kappa 3}$ (Filippi and Holmes, 2017). Since we use the set of standard parameters $\mathcal{A}_2$ for the two-dimensional Pólya tree, we have $\mathcal{A}' := \mathcal{A}_X = \mathcal{A}_Y = \{2j^2 : j \in \mathbb{N}\}$. As families of partitions $\Pi_X$ and $\Pi_Y$ we only consider the standard $\Pi$ as constructed in equation (13) with a standard Gaussian base measure.

### A.3 A nonparametric independence test

A Bayesian independence test that utilises two-dimensional Pólya trees is proposed by Filippi and Holmes (2017). Considering one-dimensional continuous random variables $X$ and $Y$, we test the hypotheses

$$H_0 : X \perp\!\!\!\perp Y \quad H_1 : X \not\perp\!\!\!\perp Y. \tag{25}$$

[3] which is defined as $\tilde{B}(\alpha_1, \alpha_2, \alpha_3, \alpha_4) := \prod_{i=1}^{4} \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^{4} \alpha_i)$

using the Bayes factor

$$\text{BF}(H_0, H_1) = \frac{p(X_{1:n}|\Pi, \mathcal{A}')p(Y_{1:n}|\Pi, \mathcal{A}')}{p((X,Y)_{1:n}|\Pi_2, \mathcal{A}_2)}, \tag{26}$$

where the marginal likelihoods are computed according to equations (15) and (23). As mentioned earlier, we take the partitions $\Pi$ and $\Pi_2$ as given by equations (13) and its two-dimensional equivalent, both with a standard Gaussian base measure. We use the standard parameter sets $\mathcal{A}'$ and $\mathcal{A}_2$.

Using similar arguments as for the two-sample test, the Bayes factor can be denoted as an infinite product, of which the terms are equal to one when $n_{XY|\kappa} \leq 1$. Therefore we compute the marginal likelihoods up to a level in which either the depth is $\lfloor \log_4(n) \rfloor$, or all elements of the partition contain at most one observation.

## Appendix B Protein expression data

In the main paper we apply the LCD algorithm, implemented with the Bayesian ensemble of independence tests, to protein expression data (Sachs et al., 2005). The data set consists of measurements of 11 phosphorylated proteins and phospholipids (Raf, Erk, p38, JNK, Akt, Mek, PKA, PLCg, PKC, PIP2 and PIP3) and 8 indicators of different interventions, performed by adding reagents to the cellular system, see Table 1. The biological details of these proteins, phospholipids, and reagents are described in Sachs et al. (2005). Using flow cytometry, these 11 components are measured from an individual human immune system cell. Flow cytometry allows for simultaneous, independent observation of hundreds of cells, producing a statistically large sample, and thus allowing for the application of causal inference algorithms (Sachs et al., 2005). The 'expert network' from Sachs et al. (2005) is depicted in Figure 10. We note that, as argued in the main paper, we do not accept this network as the true causal graph, but merely display it suggestively.

Table 1: Interventions from the data set of Sachs et al. (2005).

| | Description | Nr. of observations |
|---|---|---|
| 1 | CD3, CD28 | 853 |
| 2 | CD3, CD28, Akt-inhibitor | 911 |
| 3 | CD3, CD28, G0076 | 723 |
| 4 | CD3, CD28, Psitectorigenin | 810 |
| 5 | CD3, CD28, U0126 | 799 |
| 6 | CD3, CD28, LY294002 | 848 |
| 7 | PMA | 913 |
| 8 | $\beta$2CAMP | 707 |

We assume that the intervention variables and system variables are not confounded, so when finding an LCD triple $(C, X, Y)$ we output $C \to X \to Y$. When performing a statistical test we always use the entire set of observations. Note that we may therefore not interpret the arrows from context variables to system variables as an effect of merely performing a single intervention as opposed to a purely observational setting, but as performing an intervention as opposed to performing all the other interventions. See Mooij et al. (2020) for a detailed discussion on the assumption of no confounding between context and system variables, and on combining multiple contexts.

By means of comparison, we show the results of both LCD when implemented with the Pólya tree tests, and an implementation with the partial correlation test. These results are shown in Figure (11). In both cases, we report edges for different thresholds. We stress that we have not 'tuned' these thresholds to yield a sparse graph, and merely report results with the default hyperparameters as discussed in the main paper. The output of LCD with partial correlations complies with equation (10) from the main paper, as we only vary the threshold $\alpha$ for testing $C \perp\!\!\!\perp X$ and $X \perp\!\!\!\perp Y$, and determine $C \perp\!\!\!\perp Y|X$ at a fixed level $\alpha_0 = 0.05$.

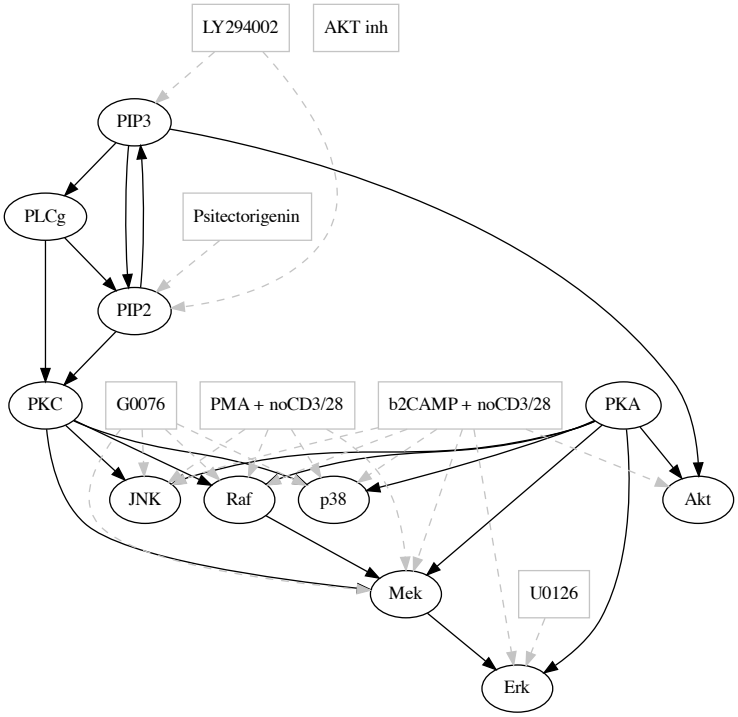
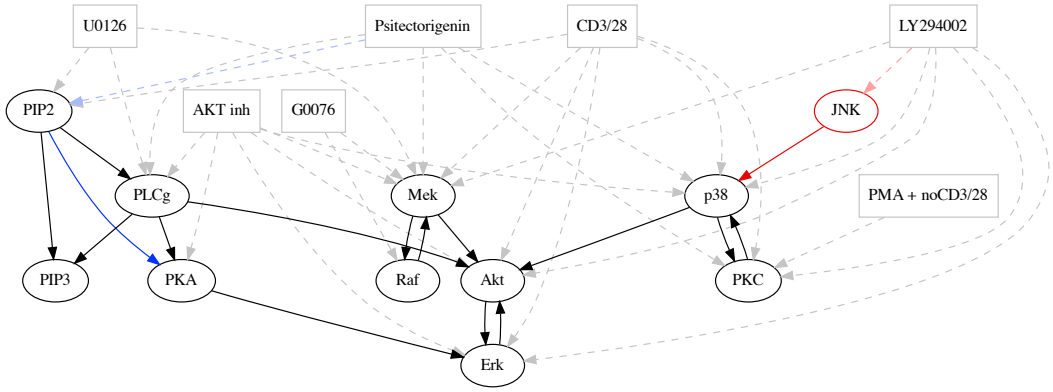
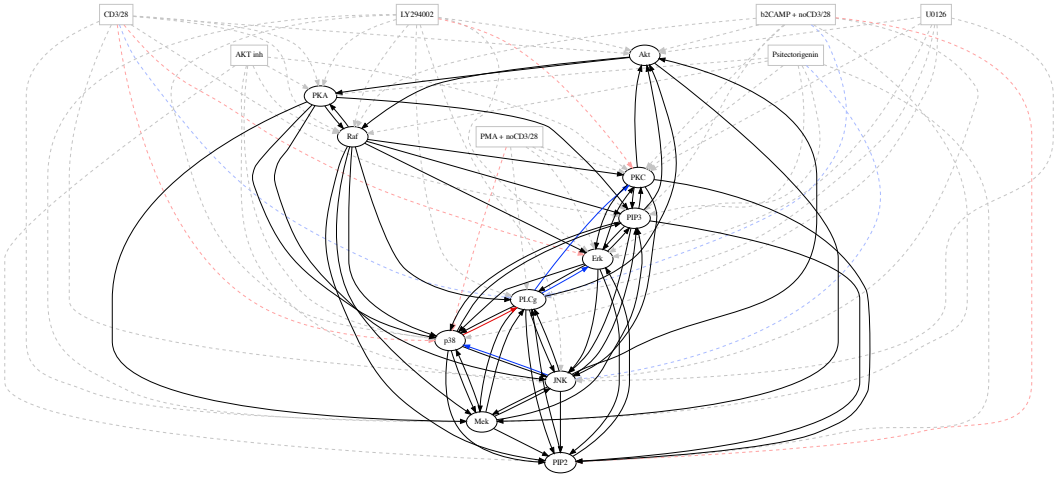



Figure 10: The 'expert network' as provided by Sachs et al. (2005). Edges indicate (indirect) causal effects between the nodes. Interventions and their (indirect) causal effects are indicated with light-coloured and dashed nodes and edges.




(a) The output of LCD with Pólya tree tests on the Sachs data. We report relations for a Bayes factor threshold 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).




(b) The output of LCD with (partial) correlation tests on the Sachs data. We report relations for a p-value threshold of $\alpha = 0.0001$ (depicted in black), $\alpha = 0.005$ (depicted in red) and $\alpha = 0.05$ (depicted in blue).

Figure 11: LCD output on the Sachs data. Edges indicate (indirect) causal effects between the nodes. Interventions and their (indirect) causal effects are indicated with light-coloured and dashed nodes and edges.