Falsifying Causal Graphs With Outlier Events

William Roy Orchard[*1] **Philipp M. Faller**[*2,3] **Dominik Janzing**[3]

[1]University of Cambridge, Cambridge, UK
[2]Karlsruhe Institute of Technology, Karlsruhe, Germany
[3]Amazon Research, Tübingen, Germany

Abstract

True causal relationships are rarely known, and inferring causal graphs from data is hard. A fundamental challenge is how to assess whether a given causal graph is good in the absence of a ground truth. We propose falsifying candidate causal graphs based on whether they can explain the propagation of an outlier event. Our approach leverages a key principle: weak outliers rarely cause strong ones. While this principle has previously been used in root cause analysis to identify root causes without prior knowledge of the graph, we turn it on its head and use it to falsify candidate causal graphs whose implied outlier propagation is inconsistent with the data. To this end, we present the first statistical tests for the hypothesis that a candidate graph is the true causal graph, and show they have false positive control, power guarantees against incorrect causal graphs, and can operate with a single outlier sample.

1 INTRODUCTION

Reasoning about causal relationships lies at the core of understanding and analysing complex systems. Unlike purely correlational methods, causal methods aim to leverage the underlying generative mechanisms that govern data, enabling robust prediction, counterfactual reasoning, and principled decision-making [Pearl, 2000, Spirtes et al., 2000]. Despite significant progress on learning causal models from observational [Kalisch and Bühlmann, 2007, Chickering, 2002, Shimizu et al., 2006, Zheng et al., 2018, Lam et al., 2022] and interventional data [Heinze-Deml et al., 2018, Brouillard et al., 2020] and combinations of statistical data and expert knowledge [Ankan and Textor, 2025, Hiremath et al., 2025], key challenges remain to reliably obtain a causal graph. Moreover, in practical settings it is often not clear how reliable a causal graph is, either because the given expert knowledge might be incomplete, or due to violated assumptions or finite-sample issues in a learned graph. In light of this, there has been a recent surge in interest for methods to evaluate or falsify estimated causal quantities or models [Peters et al., 2016, Canonne et al., 2017, Heinze-Deml et al., 2018, Acharya et al., 2018, Eigenmann et al., 2020, Choo et al., 2022, Bhattacharya and Nabi, 2022, Schkoda et al., 2025, Eulig et al., 2025, Faller and Janzing, 2025, Faltenbacher et al., 2025]. In this work we present a number of statistical tests for the null hypothesis that a given causal graph is the true causal graph. To this end, we will leverage tools developed for another causal downstream task: root cause analysis.

The goal of root cause analysis is to pinpoint the cause of some rare or anomalous event, e.g. in monitored cloud services [Hardt et al., 2024] or manufacturing [Göbler et al., 2024]. The main intuition behind our test is that, in a certain sense, weak outliers rarely cause strong outliers [Orchard et al., 2025]. Orchard et al. [2025] have proposed to use this to find root causes without prior knowledge about the underlying causal graph. In this work we turn this principle on its head: we propose to falsify causal graphs if the observed propagation of outliers does not match this principle.

As far as we are aware, we present the first statistical tests for the null hypothesis that a given causal graph is the true causal graph that have:

- false positive control;
- power against a broad class of incorrect causal graphs;
- and can operate with only a single outlier sample.

The remainder of the paper is organised as follows. In section 2 we discuss related work. In section 3 we give the preliminaries to causal models and introduce information-theoretic outlier scores, upon which our tests are based, before formalising the problem setting. In section 4 we present our proposed tests along with their theoretical guarantees. We start with tests based on conditional outlier

[*]Authors contributed equally.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5081–5110.

scores and show they have false positive control when the causal graph is a DAG. We are able to adapt some of the proposed tests to the setting where only *marginal* outlier scores are available, and show they retain false positive control when the causal graph is a polytree. By deriving lower bounds on the rate of decay of outlier scores along causal paths, we then provide power guarantees for our tests. Finally, in section 5 we empirically compare our methods to several baselines on synthetic and real-world data.

2 RELATED WORK

In what follows, we model outliers as being the result of a mechanism change. As such, they play a similar role to 'natural experiments' [Angrist et al., 1996] and we can leverage invariances over different environments to facilitate falsification. The idea that different components of a causal model change independently has been used under different names and in different contexts, e.g., 'the principle of independent mechanisms' [Haavelmo, 1944, Aldrich, 1989, Hoover, 1990, Pearl, 2000, Dawid and Didelez, 2010, Schölkopf et al., 2012, Peters et al., 2017]. It is well-known that data from different environments or regimes can render causal quantities identifiable that are unidentifiable from i.i.d. data alone [Tian and Pearl, 2013, Ghassami et al., 2018, Guo et al., 2023, Huang et al., 2017, 2020, Lagani et al., 2012, Mooij et al., 2020, Perry et al., 2022, Ke et al., 2023, Schkoda and Janzing, 2026], especially when the regimes correspond to interventions with known intervention targets [Eberhardt and Scheines, 2007, Hauser and Bühlmann, 2012, Yang et al., 2018]. Further, such data can also improve statistical estimation of causal quantities as data from several regimes can be leveraged [Peters et al., 2016, Heinze-Deml et al., 2018, Pfister et al., 2019] or even estimate quantities for regimes that have not been observed [Shpitser and Pearl, 2008, Bareinboim and Pearl, 2012, 2013, Pearl and Bareinboim, 2011]. Using interventional data is often considered the gold-standard for graph evaluation. Accordingly, previous work has studied the minimum number of interventions needed to falsify a graph [Hauser and Bühlmann, 2014, He and Geng, 2008, Shanmugam et al., 2015, Choo et al., 2022]. However, these methods assume access to a conditional independence oracle, whereas we demonstrate that already a single sample from a different regime can constitute enough evidence to reject a causal hypothesis. In applications such as cloud computing [Hardt et al., 2024] or manufacturing [Göbler et al., 2024] it can be prohibitively expensive to gather multiple samples from an anomalous event, and, in personalised medicine [Li et al., 2025] it may not be possible.

There has also been considerable work on falsifying causal models using observational data alone. In [Faller et al., 2024], a causal discovery algorithm is run on different subsets of variables and the resulting graphs are checked for mutual consistency. Incompatibilities can indicate violated model assumptions or finite-sample errors. In [Schkoda et al., 2025] causal models learned on variable subsets are used to predict dependencies between variables that were omitted during training; large prediction errors suggest incorrect causal structure. Faltenbacher et al. [2025], Faller and Janzing [2025] focus on constraint-based methods and use either additional or already conducted tests to evaluate a graph. None of these methods directly entails a valid statistical test. In contrast, Eulig et al. [2025] propose a permutation-based procedure that enables formal testing against a random-baseline null. Canonne et al. [2017], Schkoda and Drton [2025] develop goodness-of-fit tests for Bayesian networks; the latter considers linear models with non-Gaussian noise and tests whether the implied distribution matches the data. Finally, Li and Wang [2009], Strobl et al. [2019] study how edge-wise false positive control can be established in causal discovery, but they do not provide control for directions or absence of edges.

3 PRELIMINARIES

We assume the causal relationships between $n \in \mathbb{N}$ variables $X_1, \dots, X_n$ are given by a Causal Bayesian Network (CBN) [Pearl, 2000], so that their joint distribution factorises into conditionals according to an underlying DAG $\mathcal{G}$:

$$P_{X_1,\dots,X_n} = \prod_{i=1}^{n} P_{X_i|\mathrm{PA}_i^{\mathcal{G}}}, \tag{1}$$

where each $P_{X_i|\mathrm{PA}_i^{\mathcal{G}}}$ corresponds to the causal mechanism of variable X_i given its parents $\mathrm{PA}_i^{\mathcal{G}}$ in $\mathcal{G}$.

Throughout this work, we will deal with outliers. Formally, we think of outliers as being the result of a soft-intervention [Eberhardt and Scheines, 2007]: let $\mathcal{R} \subseteq \{1, \dots, n\}$ be a non-empty index set, we say outlier sample $(x_1, \dots, x_n)$ is drawn from the 'corrupted' distribution,

$$\tilde{P}_{X_1,\dots,X_n} = \prod_{j\in\mathcal{R}} \tilde{P}_{X_j|\mathrm{PA}_j^{\mathcal{G}}} \prod_{i\notin\mathcal{R}} P_{X_i|\mathrm{PA}_i^{\mathcal{G}}}, \tag{2}$$

where for each of the so-called *root cause* variables X_j, with index $j \in \mathcal{R}$, we assume the 'ordinary' mechanism $P_{X_j|\mathrm{PA}_j^{\mathcal{G}}}$ has been replaced for a 'corrupted' one $\tilde{P}_{X_j|\mathrm{PA}_j^{\mathcal{G}}}$. Note that a root *cause* need not be a root *node*. Next, we define *information theoretic (IT) outlier scores* as proposed by Budhathoki et al. [2022].

Let $\tau_i : \mathcal{X}_i \to \mathbb{R}$ be a feature map, for example, any existing real-valued anomaly scoring function. Define the event

$$E_{x_i} := \{\tau_i(X_i) \geq \tau_i(x_i)\}, \tag{3}$$

of a sample being at least as extreme as the observed event x_i w.r.t. τ. The *marginal IT outlier score* is defined by

$$S_{X_i}^{\tau_i}(x_i) := -\log P(\tau_i(X_i) \geq \tau_i(x_i)) = -\log P(E_{x_i}).$$

Note that each variable may have a different feature map, but we usually drop X_i and τ_i from $S^{\tau_i}_{X_i}$ whenever it is clear from context which variable and feature map we refer to.

As pointed out by Orchard et al. [2025], we can interpret $P(E_{x_i}) = e^{-S(x_i)}$ as a p-value for the null hypothesis that x_i was drawn from its ordinary marginal distribution. We also define the *conditional* IT outlier score,

$$S(x_i \mid \mathrm{pa}_i^{\mathcal{G}}) := -\log P(\tau_i(X_i) \geq \tau_i(x_i) \mid \mathrm{PA}_i^{\mathcal{G}} = \mathrm{pa}_i^{\mathcal{G}}).$$

Similarly, $P(E_{x_i} \mid \mathrm{PA}_i^{\mathcal{G}} = \mathrm{pa}_i^{\mathcal{G}}) = e^{-S(x_i|\mathrm{pa}_i^{\mathcal{G}})}$ is a p-value for the null hypothesis that x_i was drawn from its ordinary conditional distribution given the values of its parents $\mathrm{pa}_i^{\mathcal{G}}$ in $\mathcal{G}$. A large conditional outlier score corresponds to a small p-value and so we favour rejecting the null in such cases.

Problem statement: given an estimated causal graph $\hat{\mathcal{G}}$, test the null hypothesis $H_0 : \hat{\mathcal{G}} = \mathcal{G}$, using the outlier sample $(x_1, \ldots, x_n)$, by leveraging theory for the propagation of outliers in causal systems.

4 FALSIFYING CAUSAL GRAPHS WITH IT OUTLIER SCORES

In this section we propose tests for the null hypothesis H_0. The tests are stated under the assumption that we *know* the root cause(s) of the outlier event, but we will close this section by showing this assumption may be dropped in the case one knows an upper bound on the number of root causes. In each instance we will state a test statistic $T : \mathcal{X} \to \mathbb{R}$ and corresponding p-value p_T for H_0. It should be understood that the corresponding level-α test $\phi_T : \mathcal{X} \to \{0, 1\}$ is defined in the usual way: $\phi_T(x) := \mathbb{I}[p_{T(x)} \leq \alpha]$. Recalling that a p-value for H_0 is a random variable p that satisfies $P_{H_0}(p \leq \alpha) \leq \alpha$, the immediate consequence of identifying a p-value for H_0 is that the corresponding test controls the false positive rate. As such, we will not restate this implication for each p-value we derive.

All proofs, unless otherwise indicated, are given in section A of the appendix.

4.1 FALSIFYING CAUSAL GRAPHS WITH CONDITIONAL OUTLIER SCORES

To test our primary null hypothesis H_0 we propose testing another hypothesis that it implies:

$H_0^{\mathcal{R}}$: for outlier sample $(x_1, \ldots, x_n)$, all x_i with $i \notin \mathcal{R}$ were drawn according to $P(X_i \mid \mathrm{PA}_i^{\hat{\mathcal{G}}} = \mathrm{pa}_i^{\hat{\mathcal{G}}})$.

In other words, the outlier sample, excluding the root causes, was drawn according to the causal conditionals implied by the *estimated* graph $\hat{\mathcal{G}}$. If H_0 is true, the causal conditionals implied by $\hat{\mathcal{G}}$ coincide with the true causal mechanisms, and so $H_0^{\mathcal{R}}$ is true by definition. Rejecting $H_0^{\mathcal{R}}$ is therefore sufficient to reject H_0.[1]

To define our test statistics, we will need the assumption:

1. **Continuity:** every $P(\tau_i(X_i) \mid \mathrm{PA}_i^{\mathcal{G}} = \mathrm{pa}_i^{\mathcal{G}})$ is absolutely continuous with respect to the Lebesgue measure.

Conditional IT scores with variable inputs $X_i, \mathrm{PA}_i^{\mathcal{G}}$ define random variables $S(X_i \mid \mathrm{PA}_i^{\mathcal{G}})$. We then have (see also lemma 3.7 of Orchard et al. [2025]),

Lemma 1 (Conditional scores are independent standard exponentials). *Subject to continuity, for variables X_i distributed according to P, with $i \in \{1, \ldots, n\}$, $S(X_i \mid \mathrm{PA}_i^{\mathcal{G}})$ are independent,* $\mathrm{Exp}(1)$ *distributed random variables.*

This property underpins each of the tests we propose based on conditional scores. With this in place, the first three of our tests are based on the same idea: high conditional outlier scores with respect to $\hat{\mathcal{G}}$ are evidence for rejecting $H_0^{\mathcal{R}}$ and therefore for rejecting H_0. As the number of non-root cause variables will appear in many equations below, we will denote it $L := n - |\mathcal{R}|$ to make presentation less cumbersome.

The first test statistic is the sum of conditional outlier scores of the non-root causes $S^{\hat{\mathcal{G}}}_{\mathrm{sum}} := \sum_{i \notin \mathcal{R}} S(X_i \mid \mathrm{PA}_i^{\hat{\mathcal{G}}})$. The sum of m independent standard exponential random variables is Erlang distributed with scale parameter 1 and shape parameter m [Ibe, 2013]. As such, we can show that

Proposition 2 (Fisher-based test). H_0 *can be rejected with p-value*

$$p_{S^{\hat{\mathcal{G}}}_{\mathrm{sum}}} = e^{-S^{\hat{\mathcal{G}}}_{\mathrm{sum}}} \cdot \sum_{l=0}^{L-1} \frac{(S^{\hat{\mathcal{G}}}_{\mathrm{sum}})^l}{l!}. \tag{4}$$

This corresponds to using Fisher's method [Fisher, 1925] for the combination of the independent p-values $e^{-S(X_i|\mathrm{PA}_i^{\hat{\mathcal{G}}})}$.

The second test statistic simply uses the maximum conditional outlier score from among the non-root causes, $S^{\hat{\mathcal{G}}}_{\max} := \max_{i \notin \mathcal{R}} S(X_i \mid \mathrm{PA}_i^{\hat{\mathcal{G}}})$. This too has a simple distribution under the null so that

Proposition 3 (Tippett-based test). H_0 *can be rejected with p-value*

$$p_{S^{\hat{\mathcal{G}}}_{\max}} = 1 - (1 - e^{-S^{\hat{\mathcal{G}}}_{\max}})^L. \tag{5}$$

This corresponds to using Tippett's method [Tippett, 1931] for the combination of independent p-values.

The third test statistic counts the number of conditional outlier scores, from among the non-root causes, that exceed a given threshold t: $N_t := \sum_{i \notin \mathcal{R}} \mathbb{I}[S(X_i \mid \mathrm{PA}_i^{\hat{\mathcal{G}}}) \geq t]$. As

[1] Note that this assumes that the outlier sample is not selected because it is anomalous and then also used for testing, as this would introduce selection bias.

the probability that a standard exponential random variable exceeds a value t is simply e^{-t}, the corresponding p-value is the upper-tail of the Binomial distribution with probability of success e^{-t} and L trials:

Proposition 4 (Binomial test on scores above threshold). *H_0 can be rejected with p-value*

$$p_{N_t} = \sum_{k=N_t}^{L} \binom{L}{k} e^{-kt}(1-e^{-t})^{L-k}. \tag{6}$$

As all three tests have false positive control, a remaining question is when one should prefer one test over another, i.e. when does each test have power. We will devote section 4.5 to a thorough analysis of power, but what is common to all three is that they can only have power when errors in $\hat{\mathcal{G}}$ result in inflated outlier scores. As we will see, certain errors in $\hat{\mathcal{G}}$ do indeed result in inflated outlier scores, but nonetheless we may be interested in a test which has power against more general deviations from the null. As such let $F_{\text{Exp}}(s) := (1-e^{-s})\,\mathbb{I}[s \geq 0]$ denote the CDF of the standard exponential distribution and

$$F_L(s) := \frac{1}{L}\sum_{i \notin \mathcal{R}} \mathbb{I}[S(X_i \mid \text{PA}_i^{\hat{\mathcal{G}}}) \leq s] \tag{7}$$

be the empirical CDF of the conditional outlier scores of the non-root causes. Our fourth test statistic is

$$D_L := \sup_s |F_L(s) - F_{\text{Exp}}(s)|\,. \tag{8}$$

Let K_n be the null distribution function of the Kolmogorov-Smirnov statistic for sample size n [Massey, 1951], then

Proposition 5 (Kolmogorov-Smirnov test). *H_0 can be rejected with p-value*

$$p_{D_L} = 1 - K_L(D_L). \tag{9}$$

Proof. An immediate consequence of lemma 1 is that, under H_0, F_L is the empirical CDF of the standard exponential for sample size L. D_L is therefore the corresponding Kolmogorov-Smirnov statistic and our test is the Kolmogorov-Smirnov test. □

4.2 FALSIFYING CAUSAL GRAPHS WITH MARGINAL OUTLIER SCORES

Inferring the causal conditionals $P(X_i \mid \text{PA}_i^{\hat{\mathcal{G}}})$, and therefore the corresponding conditional outlier scores, is statistically ill-posed, especially when the number of parents is large. One would therefore ideally avoid it. In this section we show how the first three of our tests can be adapted to use only *marginal* outlier scores. However, to do so we need to introduce results on the propagation of outliers.

Let $\mathcal{G}$ be $X \to Y$ and (x, y) be the observed outlier sample. Following Orchard et al. [2025] we give the definition:

Definition 6 (Score typicality). *We say observation (x, y) of (X, Y) satisfies score typicality if*

$$S(y \mid x) \geq |S(y) - S(x)|_+, \tag{10}$$

where $|a|_+ := \max(0, a)$.

The significance of score typicality is that, when satisfied, it implies a key principle of outlier propagation: weak outliers rarely cause strong outliers. To see this, note that subject to score typicality

$$P(S(Y) \geq S(y) \mid X = x) = e^{-S(y|x)} \leq e^{-|S(y)-S(x)|_+}, \tag{11}$$

so that if y was drawn according to its ordinary causal mechanism, $S(y) \gg S(x)$ is exponentially unlikely. That it is justified to refer to this condition as 'typical' is given by the following result from Orchard et al. [2025]:

Lemma 7 (Score typicality probably holds, approximately). *For any given y, and for any $s_X \leq S(y) - 1$, let (possibly multivariate) x be chosen at random according to $P(X \mid S(X) \in [s_X, S(y)])$. Subject to continuity, with probability at least $1 - 2/c$, we obtain an x for which*

$$S(y \mid x) \geq |S(y) - S(x)|_+ - 2\log c. \tag{12}$$

In other words, for any y, although score typicality may not hold for any *particular* x, lemma 7 shows that it is likely to hold approximately for a randomly chosen x. In what follows, we will state certain results *subject to score typicality* for brevity, but in such cases it should be remembered that we could drop this condition at the cost of slightly weaker bounds holding for most samples.

Lemma 7 applies without modification to general DAGs by considering each $(X_i, \text{PA}_i^{\mathcal{G}})$ in turn, in place of (Y, X), though note that for multivariate $\text{PA}_i^{\mathcal{G}}$, the associated feature map for $S(\text{PA}_i^{\mathcal{G}})$ is, correspondingly, multivariate too. When restricting ourselves only to the inference of marginal distributions we add the assumption:

2. **Polytree:** The true causal graph $\mathcal{G}$ is a polytree[2].

A polytree is a DAG whose skeleton contains no undirected cycles. As a consequence, the parents of any given variable are independent, and we may combine their marginal scores, akin to in propositions 2 to 4, to give a joint score. For example, Orchard et al. [2025] show that, subject to continuity, for variable X_i with parents $\text{PA}_i^1, \ldots, \text{PA}_i^k$, and (joint) event pa_i, the following defines a joint IT outlier score:

$$S(\text{pa}_i) := \sum_{j=1}^{k} S(\text{pa}_i^j) - \log \sum_{l=0}^{k-1} \frac{(\sum_{j=1}^{k} S(\text{pa}_i^j))^l}{l!}. \tag{13}$$

[2]See Li et al. [2025], Ebtekar et al. [2025] for why it is not straightforward to drop this assumption for marginal scores.

With this in place we are ready to give the marginal score analogues of our first three tests. Let

$$\hat{S}^{\hat{\mathcal{G}}}_{\text{sum}} := \sum_{i \notin \mathcal{R}} |S(X_i) - S(\text{PA}_i^{\hat{\mathcal{G}}})|_+, \tag{14}$$

$$\hat{S}^{\hat{\mathcal{G}}}_{\max} := \max_{i \notin \mathcal{R}} |S(X_i) - S(\text{PA}_i^{\hat{\mathcal{G}}})|_+, \tag{15}$$

$$\hat{N}_t := \sum_{i \notin \mathcal{R}} \mathbb{I}[|S(X_i) - S(\text{PA}_i^{\hat{\mathcal{G}}})|_+ \geq t], \tag{16}$$

be the marginal score analogues of each of our first three test statistics. Then

Proposition 8 (Marginal score tests). *Subject to score typicality, each of the following are p-values for H_0:*

$$p_{S^{\hat{\mathcal{G}}}_{\text{sum}}} \leq p_{\hat{S}^{\hat{\mathcal{G}}}_{\text{sum}}} = e^{-\hat{S}^{\hat{\mathcal{G}}}_{\text{sum}}} \cdot \sum_{l=0}^{L-1} \frac{(\hat{S}^{\hat{\mathcal{G}}}_{\text{sum}})^l}{l!}, \tag{17}$$

$$p_{S^{\hat{\mathcal{G}}}_{\max}} \leq p_{\hat{S}^{\hat{\mathcal{G}}}_{\max}} = 1 - (1 - e^{-\hat{S}^{\hat{\mathcal{G}}}_{\max}})^L, \tag{18}$$

$$p_{N_t} \leq p_{\hat{N}_t} = \sum_{k=\hat{N}_t}^{L} \binom{L}{k} e^{-kt}(1 - e^{-t})^{L-k}. \tag{19}$$

In each case, the proposition follows from the marginal score test statistics each being stochastically dominated by their conditional outlier score counterparts.

4.3 BOUNDING MARGINAL SCORE JUMPS

When, according to $\hat{\mathcal{G}}$, we observe a weak outlier seemingly causing a strong outlier, there is a marginal score 'jump' from cause to effect. We saw in the previous section that all of our marginal score test statistics increase when there are such jumps, and therefore the corresponding p-values decrease. As such, we expect our marginal score tests to have power precisely when errors in $\hat{\mathcal{G}}$ result in marginal score jumps. In this section, we show when this happens.

To this end, let us add a faithfulness-type assumption [Jaber et al., 2020] that says that the outlier at the root cause(s) $X_{\mathcal{R}}$ induces outliers in all of its descendants:

3. **Effectiveness:** There exists a $\lambda > 0$, such that for all variables X_j among the descendants of $X_{\mathcal{R}}$ in $\mathcal{G}$, $S(X_j) \geq \lambda$.

The causal mechanisms of the non-descendants of the root causes work as normal, and are not influenced by the outlier event, so will generally have small outlier scores. Therefore, when λ is large, if a (true) descendant of the root causes has been assigned non-descendants as parents in $\hat{\mathcal{G}}$ (or no parents), the marginal score jump will be at least $\approx \lambda$.

That we could have power to reject graphs that do not correctly place the descendants of root causes is not surprising when we consider that root causes are equivalently thought of as the targets of a soft intervention. Indeed, if the causal model is deterministic, in the worst case, all descendants of the root cause may receive identical marginal outlier scores and so it will only be possible to distinguish descendants from non-descendants. However, in general, we expect outliers to decay as they propagate further away from root causes. In such cases, marginal score jumps will also be introduced when (true) descendants of root causes are assigned parents, in $\hat{\mathcal{G}}$, that are *further away* from the root causes than they are. To formalise this intuition we need to introduce some concepts from information theory (see Polyanskiy and Wu [2025] for each definition below).

The Kullback-Leibler (KL) divergence for probability distributions P and Q, where $P \ll Q$, is given by

$$D(P \parallel Q) := \mathbb{E}_P \left[\log \frac{dP}{dQ} \right]. \tag{20}$$

For a fixed channel $P_{Y|X}$, we denote by $P_{Y|X} \circ P$ the distribution induced by the push-forward of the distribution P, i.e. the distribution on the output Y when the input X is distributed according to P. The Strong Data Processing Inequality (SDPI) with respect to the KL-divergence is

$$D(P_{Y|X} \circ P \parallel P_{Y|X} \circ Q) \leq \eta(P_{Y|X}, P_X)\, D(P \parallel Q), \tag{21}$$

where $\eta(P_{Y|X}, P_X) \in [0, 1]$ is the *contraction coefficient* for channel $P_{Y|X}$ and fixed input Q, and is defined as:

$$\eta(P_{Y|X}, Q) := \sup_{P : 0 < D(P\|Q) < \infty} \frac{D(P_{Y|X} \circ P \parallel P_{Y|X} \circ Q)}{D(P \parallel Q)}. \tag{22}$$

For brevity we will abbreviate $\eta := \eta(P_{Y|X}, Q)$ when the channel and input are clear from context. As stated, however, the SDPI is not interesting unless $\eta < 1$. There are many channels where this is the case. We give two examples

Proposition 9 (Linear Gaussian model). *For causal model $Y := \beta X + N$, $N \perp\!\!\!\perp X$ with (X, Y) jointly Gaussian and Pearson correlation coefficient ρ,*

$$\eta = \rho^2. \tag{23}$$

Proof. Immediate consequence of Theorem 5 of Makur and Zheng [2020]. □

Proposition 10 (Bounded Gaussian additive noise model). *For causal model $Y := f(X) + N$, $N \perp\!\!\!\perp X$, and there exists a $B \geq 0$ such that $|f(x)| \leq B$ for all x, and $N \sim \mathcal{N}(0, \sigma^2)$,*

$$\eta \leq 2\,\Phi(B/\sigma) - 1, \tag{24}$$

where Φ is the CDF for the standard Gaussian distribution.

We can put the SDPI to use by noting

Lemma 11 (Marginal outlier scores are KL-divergences). *For variable X and outlier event E_x, for $P_X(E_x) > 0$,*

$$S(x) = D(P_{X|E_x} \parallel P_X). \tag{25}$$

Suppose in $\mathcal{G}$ that $X \to Y$ (where X is the possibly multivariate parent set of Y), and we observe outlier sample (x, y) with corresponding outlier events E_x, E_y as in eq. (3). Suppose also that Y is not a root cause. As Y is not a root cause, its mechanism $P_{Y|X}$ worked as normal and so we expect the conditional outlier score for y to be small. Applying lemma 11 to the SDPI, when $S(y \mid E_x) = 0$, we find

Lemma 12 (Lower bound on the rate of outlier decay).

$$\eta S(x) \geq S(y). \tag{26}$$

In other words, when the outlier is propagated by the 'working' causal mechanism, $P_{Y|X}$, it decays by at least a factor of η. Though, for brevity, we have stated the result given $S(y \mid E_x) = 0$, we show in section B of the appendix that, for any given x, for a random y drawn according to $P_{Y|E_x}$, lemma 12 is likely to hold, approximately.

This has immediate consequences for characterising the power of our marginal tests. Suppose in $\hat{\mathcal{G}}$ the relationship between X and Y is *incorrectly* given as $Y \to X$, then when $S(y \mid E_x) = 0$,

Lemma 13 (Anti-causal score jump lower bound).

$$|S(x) - S(y)|_+ \geq (1 - \eta)\, S(x). \tag{27}$$

Proof. An immediate consequence of lemma 12. □

If Y is not a root cause, and X is a true descendant of a root cause, then the score jump will be at least $(1 - \eta)\lambda$. We have therefore established two kinds of errors in $\hat{\mathcal{G}}$ that imply score jumps, both of which concern assigning the wrong parent set to a true descendant of a root cause.

4.4 BOUNDING CONDITIONAL SCORES

We have established the foundation for providing power guarantees for our marginal score tests. We now show we can do the same for conditional outlier scores.

Let $X \to Y$ in $\mathcal{G}$ where X is the (possibly multivariate) parent set of Y, let (x, y) be the outlier sample, and suppose the causal relationship is incorrectly given as $Y \to X$ in $\hat{\mathcal{G}}$. An interesting case is when large marginal score jumps imply large conditional scores, allowing us to simply apply the results from the previous section to conditional scores directly:

Lemma 14 (Monotonicity implies anti-causal conditional score lower bound). *If S_Y is injective, and $P(S(X) \geq S(x) \mid S(Y) = S(y)) \leq P(S(X) \geq S(x) \mid S(Y) \geq S(y))$, then*

$$S(x \mid y) \geq |S(x) - S(y)|_+. \tag{28}$$

This uses a type of monotonicity property. It says that observing a more extreme outlier at Y should only increase our belief that the outlier at X is more extreme too. In this case, whenever our marginal score tests have power, so do our conditional score tests. This monotonicity property holds for a wide class of causal models. For example,

Proposition 15 (Additive log-concave noise model). *For real-valued variables where $Y := f(X) + N$, $N \perp\!\!\!\perp X$, N is log-concave, f is non-decreasing, and S_X, S_Y are strictly increasing then the monotonicity of lemma 14 is satisfied for all (x, y).*

However, inheriting the marginal score results is not the only way to establish when errors in $\hat{\mathcal{G}}$ imply inflated conditional scores, and so we do not need to rely on the monotonicity property. Firstly, there are two trivial cases. One, if a true descendant of a root cause is assigned no parents in $\hat{\mathcal{G}}$, then its conditional score will simply be its marginal score and as such, subject to effectiveness, will be at least λ. Two, if a true descendant of a root cause, call it X_i, is assigned parents $\mathrm{PA}_i^{\hat{\mathcal{G}}}$ in $\hat{\mathcal{G}}$ where $X_i \perp\!\!\!\perp \mathrm{PA}_i^{\hat{\mathcal{G}}}$, then again, subject to effectiveness, $S(x_i \mid \mathrm{pa}_i^{\hat{\mathcal{G}}}) = S(x_i) \geq \lambda$.

The first non-trivial case is when a true descendant of a root cause X_i is assigned parents $\mathrm{PA}_i^{\hat{\mathcal{G}}}$ in $\hat{\mathcal{G}}$ that are non-descendants of root causes (and not root causes themselves), but where $X_i \not\perp\!\!\!\perp \mathrm{PA}_i^{\hat{\mathcal{G}}}$. In this case we find

Lemma 16 (Score lower bound after conditioning on non-descendant). *Subject to effectiveness, let $(x_i, \mathrm{pa}_i^{\hat{\mathcal{G}}})$ be drawn according to the corrupted distribution $\tilde{P}$, then with probability at least $1 - 1/c$*

$$S(x_i \mid \mathrm{pa}_i^{\hat{\mathcal{G}}}) \geq \lambda - \log c. \tag{29}$$

In this case, while $S(x_i \mid \mathrm{pa}_i^{\hat{\mathcal{G}}}) \geq \lambda$ need not hold for all $(x_i, \mathrm{pa}_i^{\hat{\mathcal{G}}})$, it does hold approximately, with high probability, for a random outlier sample.

The other non-trivial case is akin to that discussed for marginal scores in lemma 13. As before, let $X \to Y$ in $\mathcal{G}$ and suppose that in $\hat{\mathcal{G}}$ the causal relationship is incorrectly given as $Y \to X$, we have

Lemma 17 (Expected anti-causal conditional score bound).

$$\mathbb{E}_{P_{Y|E_x}}[S(x \mid Y)] \geq (1 - \eta)\, S(x). \tag{30}$$

While this only gives a bound on the *expected* conditional score, so long as $S(x \mid Y)$ is concentrated near its expectation under $Y \sim P_{Y|E_x}$, then this implies a bound on

$S(x \mid y)$ for typical y. This is true for a wide class of causal models. A straightforward example is the class of models for which $\mathrm{Var}_{P_{Y|E_x}}[S(x \mid Y)] =: V_x < \infty$, as applying Cantelli's inequality [Ghosh, 2002] we can conclude

Lemma 18 (Anti-causal conditional score lower bound). *For any x, let y be chosen at random according to $P_{Y|E_x}$. With probability at least $1-\delta$, we obtain a y for which*

$$S(x \mid y) \geq (1-\eta)\, S(x) - \sqrt{\frac{1-\delta}{\delta} V_x}. \tag{31}$$

The following lemma gives examples of causal models for which V_x can be bounded so that lemma 18 applies,

Proposition 19 (Bounded additive noise model). *For causal model $Y := f(X) + N$, $N \perp\!\!\!\perp X$, and there exists a $B \geq 0$ such that $|f(x)| \leq B$ for all x,*

1. *If N has a positive density whose logarithm is Lipschitz with constant L_N:*

$$V_x \leq (2 L_N B)^2. \tag{32}$$

   This includes the Cauchy, Laplace, Logistic and Student's t distributions.
2. *If $N \sim \mathcal{N}(0, \sigma^2)$:*

$$V_x \leq \frac{4B^2}{\sigma^4}(B^2+\sigma^2) + \frac{2B^3}{\sigma^4}\sqrt{B^2+\sigma^2} + \frac{B^4}{4\sigma^4}. \tag{33}$$

In the case N has a density that is log-Lipschitz, we actually derive a non-probabilistic bound on $S(x \mid y)$ that in many cases is stronger than that given by lemma 18 (see proof of proposition 19 in the appendix for details). For brevity, we present only the bound above in the main text.

4.5 POWER GUARANTEES

In the previous two subsections, we laid the foundations for providing power guarantees for our tests:

Remark 20 (Detectable errors). *Our conditional score tests, and their marginal score analogues, are sensitive to at least two kinds of errors in $\hat{\mathcal{G}}$:*

1. *assigning to a descendant of a root cause either no parents, or parents that, in $\mathcal{G}$, are not descendants of a root cause. In this case, the marginal score jump (effectiveness) and the conditional score (effectiveness and lemma 14 or lemma 16) are lower bounded by $\approx \lambda$;*
2. *assigning to a descendant of a root cause parents that, in $\mathcal{G}$, are among its descendants. In this case, the marginal score jump (effectiveness and lemma 13) and the conditional score (effectiveness and lemma 14 or lemma 18) are lower bounded by $(1-\eta_{\max})\lambda$,*

where letting η_i be the contraction coefficient for channel $P_{X_i|\mathrm{PA}_i^{\mathcal{G}}}$, then $\eta_{\max} := \max_{i \notin \mathcal{R}} \eta_i$. As $\lambda \geq (1-\eta_{\max})\lambda$, we will refer to $b := (1-\eta_{\max})\lambda$ as the lowest lower bound *on the conditional scores and marginal score jumps that are misspecified in $\hat{\mathcal{G}}$ due to the two kinds of errors just listed.*

To see how each test is sensitive to these errors, let π be the fraction of non-root causes whose parents in $\hat{\mathcal{G}}$ are incorrect in either of the two ways listed above. Informally, π can be thought of as the fraction of $\hat{\mathcal{G}}$ that is 'detectably' incorrect. We have the following bounds on their p-values, where, to avoid writing out each bound twice, we write $p_{\tilde{T}}$ to denote the p-value for *either* the conditional score test with test statistic T or the marginal score test with test statistic $\hat{T}$:

Proposition 21 (Upper bounds on p-values). *Subject to the conditions given in remark 20 and $\pi > 0$,*

$$p_{\tilde{S}^{\hat{\mathcal{G}}}_{\mathrm{sum}}} \leq e^{-\pi L b} \sum_{l=0}^{L-1} \frac{(\pi L b)^l}{l!}, \tag{34}$$

$$p_{\tilde{S}^{\hat{\mathcal{G}}}_{\max}} \leq 1 - (1-e^{-b})^L, \tag{35}$$

$$p_{\tilde{N}_t} \leq \begin{cases} \sum_{k=\pi L}^{L} \binom{L}{k} e^{-kt}(1-e^{-t})^{L-k} & \text{if } b \geq t, \\ 1 & \text{otherwise.} \end{cases} \tag{36}$$

It is important to note that the bounds apply to the marginal score and conditional score test p-values under different conditions (see remark 20). Each of these bounds provides hints for when we expect one test to be preferable to the others. We plot them in section C of the appendix. While the bounds on both the first and second p-values decrease with increasing b, which is generally high when the outlier event is strong (because λ is large), only the first decreases with increasing π. This makes sense as $S^{\hat{\mathcal{G}}}_{\max}$ considers only the maximum score, so that even if there is only a single error in $\hat{\mathcal{G}}$, so long as b is large enough, the p-value will be small. In contrast, the first test may be preferable for more moderate outlier events when one expects $\hat{\mathcal{G}}$ to have a larger proportion of errors. For the third p-value, so long as $b \geq t$, the bound decreases with increasing π but is otherwise insensitive to b. However, the bound tends to decay more sharply with π than for the first p-value, so long as t is not chosen to be too much smaller than b.

So far we have said little about the fourth test (proposition 5) as it is the only without a marginal score analogue, and the only for which the p-value is sensitive to more than increased conditional scores under the alternative. Nonetheless, we can give power guarantees for the test. Consider the alternative

$$H_1\colon\ S(X_i \mid \mathrm{PA}_i^{\hat{\mathcal{G}}}), \text{ for } i \notin \mathcal{R}, \text{ are iid with CDF}$$
$$F(s) := \pi G(s) + (1-\pi)(1-e^{-s}), \tag{37}$$

where $G(s)$ is a CDF such that $G(t) = 0$. This alternative models the case that a proportion π of the conditional

outlier scores are drawn from an unknown distribution, but are bounded away from zero by a value t, while the remaining are standard exponential. This almost matches our case of interest, for $t = b$, but with the additional simplifying assumption that the scores are independent under the alternative. We then have,

Theorem 22 (**Power lower bound on KS test).** *Subject to the conditions given in remark 20*

$$P_{H_1}(D_L \geq D^{\text{crit}}_{L,\alpha}) \geq 1 - 2e^{-2L|\pi(1-e^{-b})-D^{\text{crit}}_{L,\alpha}|^2_+}, \quad (38)$$

where $D^{\text{crit}}_{L,\alpha}$ is the critical value of the KS-statistic for sample size L and significance threshold α, and $t = b$.

Beyond bounds on the power or p-values for each test, it can be instructive to consider for which alternatives each test is optimal, i.e. uniformly most powerful (UMP). In section D of the appendix, we slightly adapt a result from Heard and Rubin-Delanchy [2018] to show in theorem 26 that the $S^{\hat{\mathcal{G}}}_{\text{sum}}$ test is UMP against an alternative wherein the conditional outlier scores remain independent, and exponentially distributed, but with common rate parameter $\nu < 1$. In other words, when each outlier score is stochastically larger than under the null. This is consistent with our observation above that the test is preferable when there is a large proportion of errors in $\hat{\mathcal{G}}$ even when the outlier is only moderately strong. In contrast, theorem 27 shows that the $S^{\hat{\mathcal{G}}}_{\max}$ test is UMP against an alternative where a single score is drawn from a distribution which stochastically dominates the standard exponential, and the remaining are drawn from a standard exponential distribution, truncated to be less extreme than that first score. In this sense, the test is optimal when there is a single, dominating, large score.

For the conditional score binomial test, the alternative is more directly comparable to our setting,

Theorem 23 (**Conditional score binomial test is uniformly most powerful).** *The conditional score binomial test in proposition 4 is uniformly most powerful against the alternative*

$$H_1^{\text{Exp}}\text{: } S(X_i \mid \text{PA}_i^{\hat{\mathcal{G}}}) \text{, for } i \notin \mathcal{R}\text{, are iid with distribution } f(s) = \pi e^{-(s-t)}\mathbb{I}[s \geq t] + (1-\pi)e^{-s}.$$

H_1^{Exp} is a special case of H_1 above where the distribution of $G(s)$ is taken to be the standard exponential distribution conditional on $S \geq t$. This suggests that the N_t test is well suited to our setting so long as t is close to b (but doesn't exceed it). This matches the conclusions drawn from proposition 21.

4.6 FALSIFYING CAUSAL GRAPHS WHEN THE ROOT CAUSE(S) ARE UNKNOWN

So far we have presented our tests under the assumption that the root cause(s) of the outlier event are known. Now we show that this assumption may be dropped in the setting one only knows (an upper bound on) the *number* of root causes.

Note that one cannot simply infer the root cause(s) with the outlier sample $(x_1, \ldots, x_n)$, using a root cause analysis algorithm (e.g. one of the approaches in Orchard et al. [2025], Schkoda and Janzing [2026]), and then apply one of our tests with the inferred root cause(s). This introduces selection bias, so the remaining conditional outlier scores are no longer independent, and our tests no longer valid.

Suppose instead we know an upper bound $k \geq |\mathcal{R}|$ on the number of root causes. We propose the following approach to test H_0. For every $\mathcal{C} \subseteq \{1, \ldots, n\}$ such that $|\mathcal{C}| = k$, let T denote any of the test statistics we have proposed and $p^{\mathcal{C}}_T$ be the corresponding "p-value" after excluding the conditional outlier scores (or marginal score jumps) for variables with index in $\mathcal{C}$. For example, $p^{\mathcal{C}}_{S^{\hat{\mathcal{G}}}_{\text{sum}}}$ is the p-value for the statistic $\sum_{i \notin \mathcal{C}} S(X_i \mid \text{PA}_i^{\hat{\mathcal{G}}})$ given by the (appropriately adjusted) equation in proposition 2. We then have,

Proposition 24 (**Maximum leave-k-out p-value).** *For statistic T, H_0 can be rejected with p-value*

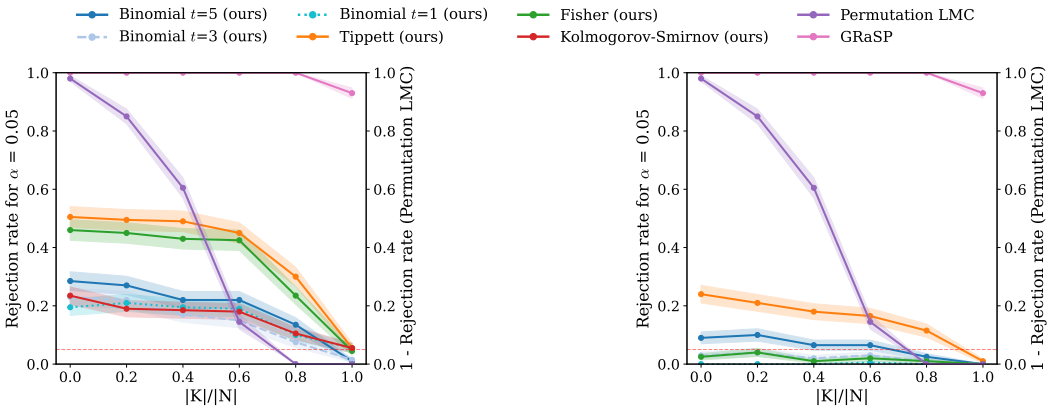
$$p^k_T = \max_{\substack{\mathcal{C} \subseteq \{1,\ldots,n\} \\ \text{s.t. } |\mathcal{C}| = k}} p^{\mathcal{C}}_T. \quad (39)$$

Intuitively: if $p^k_T \leq \alpha$ then $p^{\mathcal{C}}_T \leq \alpha$ for all $\mathcal{C}$, including for all $\mathcal{C} \supseteq \mathcal{R}$. However, for $\mathcal{C} \supseteq \mathcal{R}$, $p^{\mathcal{C}}_T$ is a valid p-value for H_0, so we may reject H_0 at level α.

5 EXPERIMENTS

In this section we compare our proposed tests to two baselines on both synthetic and real-world data. Our first baseline makes use of the GRaSP [Lam et al., 2022] algorithm for causal discovery: we reject the given causal graph if it does not fall into the same Markov equivalence class as the graph empirically estimated by GRaSP. The second is the test proposed by Eulig et al. [2025]. In brief, they test the hypothesis that the given graph is *not* a better match of the conditional independencies inferred in the data than a permutation-based random baseline. As such, their test is rejected when the given graph is *better than random*. Note that a graph may be better than random, but not the true causal graph, and so we expect their test to reject graphs more readily than ours do not reject them.

For full experimental details, and further experiments, see section F. In brief, to generate synthetic data we sample a random DAG of 20 nodes, and assign to each node a randomly initiated linear model of its parents and a Gaussian noise variable, and draw 1000 samples. A root cause is chosen uniformly at random from the nodes and an outlier injected in its noise variable and propagated through the graph for a single additional sample. We adopt the approach of Eulig et al. [2025] for sampling 'expert-provided'



(a) Conditional scores.

(b) Marginal scores.

Figure 1: Average rejection rate for synthetic data and increasing fraction of correctly connected nodes $|K|/|N|$.

graphs to falsify given the ground truth. In brief, a subset $K \subseteq N := \{X_1, \ldots, X_n\}$ of nodes are selected at random. All edges between nodes in K are preserved, while edges between the rest of the nodes are randomly resampled, so that a fraction $|K|/|N|$ of the graph is known. In fig. 1, we investigate how the rejection rate of our tests varies with $|K|/|N|$. As expected, our tests reject fewer graphs the higher $|K|/|N|$, with the first three of our tests having consistently higher rejection rates than the KS-test. Our marginal tests are less powerful than their conditional counterparts as expected. The GRaSP approach is far too sensitive to discrepancies and rejects almost all graphs at every $|K|/|N|$, including the ground truth. The permutation test displays the expected behaviour. In Section F we present several variations of this experiment, e.g., with larger injected outliers we see improvements in the power, and in the setting that the root causes are unknown we see only a slight drop in power.

For the real-world data experiments we consider two datasets. First we used the PetShop dataset from Hardt et al. [2024]. The dataset contains performance metrics for a microservice-based application before and after the introduction of outlier events. We test whether the so-called 'service dependency map' of the application, with its edges reversed, is a good proxy for the causal graph of certain metrics, as has been suggested in the literature [Budhathoki et al., 2022, Li et al., 2022, Gan et al., 2021]. However, Eulig et al. [2025] has questioned this approach. Second we take the light-measurements from the Causal Chambers dataset [Gamella et al., 2025] for which the given graph is expected to be an accurate description of the causal structure. Results are displayed for both in fig. 2. We see the permutation test rejects the hypothesis that the graph is not better than random in all scenarios. In contrast, our tests reject the service graph in PetShop at a higher rate than under the null, especially in the high traffic setting, where the graph is rejected for almost all tested anomalies. This could suggest that while the service graph is better

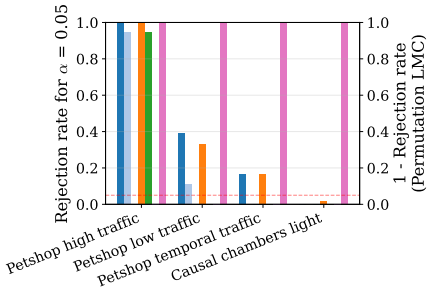


Figure 2: Average rejection rate of our marginal tests on cloud computing and Causal Chambers data.

than random, it is nonetheless not the true causal graph. For Causal Chambers, none of our tests rejects at a higher rate than the significance level, suggesting the given causal graph is likely quite accurate. The GRaSP test always rejects the graph in every setting and for both datasets.

Code: `https://github.com/worchard/rca-falsification`

6 CONCLUSION

In this work, we demonstrate how outlier events can be used to falsify causal graphs. Since rare events propagate along causal paths and rarely occur on their own, they are informative about causal structure, analogously to distribution shifts or interventions. We have shown how to turn these theoretical insights from the root cause analysis literature into a statistical test that rejects false causal graphs. We proved that this test has false positive control and power against certain classes of incorrect graphs even with only a single outlier sample. We have experimentally validated these findings on synthetic and real-world data and have demonstrated that the power of our test is on par with other baselines while controlling for false positives.

Acknowledgements

W.R. Orchard would like to thank Cancer Research UK and Peterhouse, Cambridge for supporting this research. P. M. Faller was supported by a doctoral scholarship of the Studienstiftung des deutschen Volkes (German Academic Scholarship Foundation).

References

Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. *Advances in Neural Information Processing Systems*, 31, 2018.

John Aldrich. Autonomy. *Oxford Economic Papers*, 41(1): 15–34, 1989.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91 (434):444–455, 1996.

Ankur Ankan and Johannes Textor. Expert-in-the-loop causal discovery: Iterative model refinement using expert knowledge. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Elias Bareinboim and Judea Pearl. Transportability of causal effects: Completeness results. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 698–704, 2012.

Elias Bareinboim and Judea Pearl. Meta-transportability of causal effects: A formal approach. In *Artificial Intelligence and Statistics*, pages 135–143. PMLR, 2013.

Rohit Bhattacharya and Razieh Nabi. On testability of the front-door model via verma constraints. In *Uncertainty in Artificial Intelligence*, pages 202–212. PMLR, 2022.

Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models, 2022.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33: 21865–21877, 2020.

Kailash Budhathoki, Lenon Minorics, Patrick Bloebaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2357–2369. PMLR, 17–23 Jul 2022.

Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing Bayesian networks. In *Conference on Learning Theory*, pages 370–448. PMLR, 2017.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Davin Choo, Kirankumar Shiragur, and Arnab Bhattacharyya. Verification and search algorithms for causal dags. *Advances in Neural Information Processing Systems*, 35:12787–12799, 2022.

A. Philip Dawid and Vanessa Didelez. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4:184–231, 2010. doi: 10.1214/10-SS081.

R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. i. *Theory of Probability & Its Applications*, 1(1):65–80, 1956. doi: 10.1137/1101006.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956. ISSN 00034851, 21688990. URL `http://www.jstor.org/stable/2237374`.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007. ISSN 00318248, 1539767X. URL `http://www.jstor.org/stable/10.1086/525638`.

Aram Ebtekar, Yuhao Wang, and Dominik Janzing. Toward universal laws of outlier propagation. In *Conference on Uncertainty in Artificial Intelligence*, pages 1167–1183. PMLR, 2025.

Marco Eigenmann, Sach Mukherjee, and Marloes Maathuis. Evaluation of causal structure learning algorithms via risk estimation. In *Conference on Uncertainty in Artificial Intelligence*, pages 151–160. PMLR, 2020.

Elias Eulig, Atalanti A Mastakouri, Patrick Blöbaum, Michaela Hardt, and Dominik Janzing. Toward falsifying causal graphs using a permutation-based test. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26778–26786, 2025.

Philipp M Faller and Dominik Janzing. On different notions of redundancy in conditional-independence-based discovery of graphical models. *arXiv preprint arXiv:2502.08531*, 2025.

Philipp M Faller, Leena C Vankadara, Atalanti A Mastakouri, Francesco Locatello, and Dominik Janzing. Self-compatibility: Evaluating causal discovery without ground truth. In *International Conference on Artificial Intelligence and Statistics*, pages 4132–4140. PMLR, 2024.

Sofia Faltenbacher, Jonas Wahl, Rebecca Herman, and Jakob Runge. Internal incoherency scores for constraint-based causal discovery algorithms. *arXiv preprint arXiv:2502.14719*, 2025.

R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.

Juan L Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for ai methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025.

Yu Gan, Mingyu Liang, Sundar Dev, David Lo, and Christina Delimitrou. Sage: practical and scalable ml-driven performance debugging in microservices. In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '21, page 135–151, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383172.

AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. *Advances in neural information processing systems*, 31, 2018.

B. K. Ghosh. Probability inequalities related to Markov's theorem. *The American Statistician*, 56(3):186–190, 2002. ISSN 00031305. URL http://www.jstor.org/stable/3087296.

Nicola Gnecco, Nicolai Meinshausen, Jonas Peters, and Sebastian Engelke. Causal discovery in heavy-tailed models. *The Annals of Statistics*, 49(3):1755 – 1778, 2021. doi: 10.1214/20-AOS2021.

Konstantin Göbler, Tobias Windisch, Mathias Drton, Tim Pychynski, Martin Roth, and Steffen Sonntag. causalAssembly: Generating realistic production data for benchmarking causal discovery. In *Causal Learning and Reasoning*, pages 609–642. PMLR, 2024.

Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. Causal de finetti: On the identification of invariant causal structure in exchangeable data. *Advances in Neural Information Processing Systems*, 36:36463–36475, 2023.

Trygve Haavelmo. The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115, 1944.

Michaela Hardt, William Roy Orchard, Patrick Blöbaum, Elke Kirschbaum, and Shiva Kasiviswanathan. The petshop dataset — finding causes of performance issues across microservices. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 957–978. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/hardt24a.html.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464, 2012.

Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014.

Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(11), 2008.

N A Heard and P Rubin-Delanchy. Choosing between methods of combining p-values. *Biometrika*, 105 (1):239–246, 01 2018. ISSN 0006-3444. doi: 10.1093/biomet/asx076. URL https://doi.org/10.1093/biomet/asx076.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Sujai Hiremath, Dominik Janzing, Philipp Faller, Patrick Blöbaum, Elke Kirschbaum, Shiva Prasad Kasiviswanathan, and Kyra Gan. From guess2graph: When and how can unreliable experts safely boost causal discovery in finite samples? *arXiv preprint arXiv:2510.14488*, 2025.

Kevin D Hoover. The logic of causal inference: Econometrics and the conditional analysis of causation. *Economics & Philosophy*, 6(2):207–234, 1990.

Biwei Huang, Kun Zhang, Jiji Zhang, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Behind distribution shift: Mining driving forces of changes and causal arrows. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 913–918. IEEE, 2017.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Oliver Ibe. *Markov Processes for Stochastic Modelling*. Elsevier, 2nd edition edition, 2013.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9551–9561. Curran Associates, Inc., 2020.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Samuel Karlin. *Total Positivity*, volume 1. Stanford University Press, Stanford, CA, 1968.

Samuel Karlin and Herman Rubin. The theory of decision procedures for distributions with monotone likelihood ratio. *The Annals of Mathematical Statistics*, 27(2):272–299, 1956. ISSN 00034851, 21688990. URL http://www.jstor.org/stable/2236994.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael Curtis Mozer, Christopher Pal, and Yoshua Bengio. Neural causal structure discovery from interventions. *Transactions on Machine Learning Research*, 2023.

Vincenzo Lagani, Ioannis Tsamardinos, and Sofia Triantafillou. Learning from mixture of experimental data: A constraint–based approach. In *Hellenic Conference on Artificial Intelligence*, pages 124–131. Springer, 2012.

Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In *Uncertainty in Artificial Intelligence*, pages 1052–1062. PMLR, 2022.

Jinzhou Li, Benjamin B Chu, Ines F Scheller, Julien Gagneur, and Marloes H Maathuis. Root cause discovery via permutations and cholesky decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf066, 2025.

Junning Li and Z Jane Wang. Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. *Journal of Machine Learning Research*, 10 (2), 2009.

Mingjie Li, Zeyan Li, Kanglin Yin, Xiaohui Nie, Wenchi Zhang, Kaixin Sui, and Dan Pei. Causal inference-based root cause analysis for online service systems with intervention recognition. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3230–3240, 2022.

A. Makur and L. Zheng. Comparison of contraction coefficients for f-divergences. *Problems of Information Transmission*, 56(2):103–156, Apr 2020. ISSN 1608-3253. doi: 10.1134/S0032946020020015. URL https://doi.org/10.1134/S0032946020020015.

Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer New York, 2011. ISBN 9780387682761. doi: 10.1007/978-0-387-68276-1. URL http://dx.doi.org/10.1007/978-0-387-68276-1.

P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, 18(3): 1269–1283, 1990. ISSN 00911798, 2168894X. URL http://www.jstor.org/stable/2244426.

Frank J. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46 (253):68–78, 1951. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2280095.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of machine learning research*, 21(99):1–108, 2020.

Roger B Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer, New York, NY, 2 edition, December 2006.

William Roy Orchard, Nastaran Okati, Sergio Hernan Garrido Mejia, Patrick Blöbaum, and Dominik Janzing. Root cause analysis of outliers with missing structural knowledge. *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2406.05014.

J. Pearl. *Causality*. Cambridge University Press, 2000.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 247–254, 2011.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5): 947–1012, 10 2016. ISSN 1369-7412. doi: 10.1111/rssb.12167. URL https://doi.org/10.1111/rssb.12167.

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

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and Bayesian networks. In Eric Carlen, Mokshay Madiman, and Elisabeth M. Werner, editors, *Convexity and Concentration*, pages 211–249, New York, NY, 2017. Springer New York. ISBN 978-1-4939-7005-6.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

Daniela Schkoda and Mathias Drton. Goodness-of-fit tests for linear non-gaussian structural equation models. *Biometrika*, 112(4):asaf046, 2025.

Daniela Schkoda and Dominik Janzing. Root cause analysis of outliers in unknown cyclic graphs. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=yMy87GlRfE.

Daniela Schkoda, Philipp Michael Faller, Dominik Janzing, and Patrick Blöbaum. Cross-validating causal discovery via leave-one-variable-out. In *Causal Learning and Reasoning*, pages 659–692. PMLR, 2025.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pages 459–466, 2012.

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

Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. *Advances in Neural Information Processing Systems*, 28, 2015.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(64):1941–1979, 2008. URL http://jmlr.org/papers/v9/shpitser08a.html.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Eric V Strobl, Peter L Spirtes, and Shyam Visweswaran. Estimating and controlling the false discovery rate of the pc algorithm using edge-specific p-values. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10 (5):1–37, 2019.

Jin Tian and Judea Pearl. Causal discovery from changes. *arXiv preprint arXiv:1301.2312*, 2013.

L. H. C. Tippett. *The Methods of Statistics: An Introduction Mainly for Workers in the Biological Sciences*. Williams and Norgate Ltd., London, UK, 1931.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pages 5541–5550. PMLR, 2018.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024.

Falsifying Causal Graphs With Outlier Events (Supplementary Material)

William Roy Orchard[*1] **Philipp M. Faller**[*2,3] **Dominik Janzing**[3]

[1]University of Cambridge, Cambridge, UK
[2]Karlsruhe Institute of Technology, Karlsruhe, Germany
[3]Amazon Research, Tübingen, Germany

A PROOFS

A.1 PROOF OF LEMMA 1

See proof of lemma 3.7 in Orchard et al. [2025].

A.2 PROOF OF PROPOSITION 2

By lemma 1, under the null, $S_{\text{sum}}^{\hat{\mathcal{G}}}$ is the sum of L independent, standard exponential random variables. The sum of L independent, standard exponential random variables is Erlang distributed with CDF [Ibe, 2013],

$$F(x) = 1 - e^{-x} \sum_{l=0}^{L-1} \frac{x^l}{l!}. \tag{40}$$

We therefore have,

$$P(S_{\text{sum}}^{\hat{\mathcal{G}}} \geq s) = 1 - F(s) = e^{-s} \sum_{l=0}^{L-1} \frac{s^l}{l!}, \tag{41}$$

so that $1 - F(S_{\text{sum}}^{\hat{\mathcal{G}}})$ is a p-value for H_0.

A.3 PROOF OF PROPOSITION 3

Under the null $\{S(X_i \mid \text{PA}_i^{\hat{\mathcal{G}}})\}_{i \notin \mathcal{R}}$ are independent, standard exponential so that, for each $i \notin \mathcal{R}$, $P(S(X_i \mid \text{PA}_i^{\hat{\mathcal{G}}}) < s) = 1 - e^{-s}$, therefore,

$$P(S_{\max}^{\hat{\mathcal{G}}} < s) = P\left(S(X_i \mid \text{PA}_i^{\hat{\mathcal{G}}}) < s \text{ for all } i \notin \mathcal{R}\right) = \prod_{i \notin \mathcal{R}} P(S(X_i \mid \text{PA}_i^{\hat{\mathcal{G}}}) < s) = (1 - e^{-s})^L. \tag{42}$$

As such, $P(S_{\max}^{\hat{\mathcal{G}}} \geq s) = 1 - (1 - e^{-s})^L$, and we have the result.

A.4 PROOF OF PROPOSITION 4

Proof given in main text immediately preceding the proposition.

[*]Authors contributed equally.

A.5 PROOF OF LEMMA 7

See proof of lemma 3.4 in Orchard et al. [2025].

A.6 PROOF OF PROPOSITION 8

Subject to score typicality we have the following,

$$\hat{s}^{\hat{\mathcal{G}}}_{\text{sum}} := \sum_{i \notin \mathcal{R}} |S(x_i) - S(\text{pa}^{\hat{\mathcal{G}}}_i)|_+ \leq \sum_{i \notin \mathcal{R}} S(x_i \mid \text{pa}^{\hat{\mathcal{G}}}_i) = s^{\mathcal{G}}_{\text{sum}}, \tag{43}$$

$$\hat{s}^{\hat{\mathcal{G}}}_{\max} := \max_{i \notin \mathcal{R}} |S(x_i) - S(\text{pa}^{\hat{\mathcal{G}}}_i)|_+ \leq \max_{i \notin \mathcal{R}} S(x_i \mid \text{pa}^{\hat{\mathcal{G}}}_i) = s^{\hat{\mathcal{G}}}_{\max}, \tag{44}$$

$$\hat{n}_t := \sum_{i \notin \mathcal{R}} \mathbb{I}[|S(x_i) - S(\text{pa}^{\hat{\mathcal{G}}}_i)|_+ \geq t] \leq \sum_{i \notin \mathcal{R}} \mathbb{I}[S(x_i \mid \text{pa}^{\hat{\mathcal{G}}}_i) \geq t] = n_t, \tag{45}$$

so that in each case, the marginal score test statistic is stochastically dominated by the corresponding conditional score test statistic. Substituting the conditional score test statistics for each of the marginal ones, the corresponding outputs are therefore upper bounds for the p-values given in propositions 2 to 4, and therefore are valid p-values for H_0.

A.7 PROOF OF PROPOSITION 10

We need first to introduce some extra concepts from information theory. In what follows, the exposition largely follows Polyanskiy and Wu [2017], however, the original references for the results quoted therein will be given here. First, contraction coefficients can be defined with respect to divergences other than the KL-divergence. For example, defining the total variation for distributions P and Q on the same measurable space, and events E in the corresponding sigma-algebra,

$$d_{\text{TV}}(P, Q) := \sup_E |P(E) - Q(E)|, \tag{46}$$

the contraction coefficient η_{TV} is given by simply replacing D with d_{TV} in eq. (22). Second is that while in eq. (22) we have given the so-called "input-dependent" contraction coefficient, more generally channel contraction coefficients are given without restriction to a certain input:

$$\eta(P_{Y|X}) := \sup_Q \ \eta(P_{Y|X}, Q). \tag{47}$$

With those additional preliminaries in place, we may proceed. From theorem 1 in Polyanskiy and Wu [2017] we have,

$$\eta_{\text{KL}}(P_{Y|X}) \leq \eta_{\text{TV}}(P_{Y|X}), \tag{48}$$

so that a general approach for upper bounding KL contraction coefficients is by inspecting their TV counterparts. To this end, Dobrushin [1956] has shown that,

$$\eta_{\text{TV}}(P_{Y|X}) = \sup_{x,x'} d_{\text{TV}}(P_{Y|X=x}, P_{Y|X=x'}). \tag{49}$$

Noting that in our case, $P(Y \mid X = x) = \mathcal{N}(f(x), \sigma^2)$, we use the fact that for $Q = \mathcal{N}(\mu_1, \sigma^2)$ and $R = \mathcal{N}(\mu_2, \sigma^2)$, $d_{\text{TV}}(Q, R) = 2\Phi(|\mu_1 - \mu_2|/2\sigma) - 1$:

$$\sup_{x,x'} d_{\text{TV}}(P_{Y|X=x}, P_{Y|X=x'}) = \sup_{x,x'} \ 2\Phi(|f(x) - f(x')|/2\sigma) - 1 \leq 2\Phi(B/\sigma) - 1, \tag{50}$$

where we have used the fact that f is bounded, and that Φ is strictly monotonically increasing. Finally, noting that $\eta_{\text{KL}}(P_{Y|X}, Q) \leq \eta_{\text{KL}}(P_{Y|X}) \leq \eta_{\text{TV}}(P_{Y|X})$, we have the result.

A.8 PROOF OF LEMMA 11

Noting that for any event E with $P_X(E) > 0$, we have $\frac{dP_{X|E}}{dP_X}(x) = \frac{\mathbf{1}_E(x)}{P_X(E)}$, P_X-a.e.. We then have from the definition of the KL-divergence (eq. (20)) and the definition of marginal IT scores:

$$D(P_{X|E_x} \mid\mid P_X) = \int \log \frac{\mathbf{1}_{E_x}(x')}{P_X(E_x)} dP_{X|E_x}(x') = \log \frac{1}{P_X(E_x)} = S^{\tau}_X(x). \tag{51}$$

A.9 PROOF OF LEMMA 12

From the SPDI in eq. (21) with inputs $P_{X|E_x}$ and P_X and channel $P_{Y|X}$ we have

$$\eta D(P_{X|E_x} \parallel P_X) \geq D(P_{Y|X} \circ P_{X|E_x} \parallel P_{Y|X} \circ P_X),$$

so that, applying lemma 11 along with $P_Y = P_{Y|X} \circ P_X$ and $P_{Y|E_x} = P_{Y|X} \circ P_{X|E_x}$ we have,

$$\eta S_X^\tau(x) \geq D(P_{Y|E_x} \parallel P_Y). \tag{52}$$

For arbitrary distributions $P \ll Q$ and event A such that $Q(A) > 0$, consider the binary channel defined by $Y = \mathbf{1}_A(X)$, then by the data processing inequality

$$D(P \parallel Q) \geq D(\text{Bern}(P(A)) \parallel \text{Bern}(Q(A))) = P(A) \log \frac{P(A)}{Q(A)} + (1 - P(A)) \log \frac{1 - P(A)}{1 - Q(A)}. \tag{53}$$

Taking the outlier event for $E_y := \{\tilde{\tau}(Y) \geq \tilde{\tau}(y)\}$, assume $P_Y(E_y) > 0$, we therefore have,

$$\begin{aligned}
&D(P_{Y|E_x} \parallel P_Y) \\
&\geq P_{Y|E_x}(E_y) \log \frac{P_{Y|E_x}(E_y)}{P_Y(E_y)} + (1 - P_{Y|E_x}(E_y)) \log \frac{1 - P_{Y|E_x}(E_y)}{1 - P_Y(E_y)} \\
&= e^{-S^{\tilde{\tau}}_{Y|E_x}(y)} \log \frac{e^{-S^{\tilde{\tau}}_{Y|E_x}(y)}}{e^{-S^{\tilde{\tau}}_Y(y)}} + (1 - e^{-S^{\tilde{\tau}}_{Y|E_x}(y)}) \log \frac{1 - e^{-S^{\tilde{\tau}}_{Y|E_x}(y)}}{1 - e^{-S^{\tilde{\tau}}_Y(y)}} \\
&= e^{-S^{\tilde{\tau}}_{Y|E_x}(y)} (S^{\tilde{\tau}}_Y(y) - S^{\tilde{\tau}}_{Y|E_x}(y)) + (1 - e^{-S^{\tilde{\tau}}_{Y|E_x}(y)}) \log \frac{1 - e^{-S^{\tilde{\tau}}_{Y|E_x}(y)}}{1 - e^{-S^{\tilde{\tau}}_Y(y)}}.
\end{aligned} \tag{54}$$

Taking $S^{\tilde{\tau}}_{Y|E_x}(y) = 0$, the right hand side of eq. (54) is $S^{\tilde{\tau}}_Y(y)$, and so finally, by combining eq. (52) with eq. (54), we have

$$\eta S_X^\tau(x) \geq S^{\tilde{\tau}}_Y(y). \tag{55}$$

A.10 PROOF OF LEMMA 14

$$\begin{aligned}
P(\tau(X) \geq \tau(x) \mid Y = y) &= P(S(X) \geq S(x) \mid S(Y) = S(y)) &(56)\\
&\leq P(S(X) \geq S(x) \mid S(Y) \geq S(y)) &(57)\\
&\leq \frac{P(S(X) \geq S(x))}{P(S(Y) \geq S(y))}, &(58)
\end{aligned}$$

where the equality follows from injectivity of S_Y and the definition of IT outlier scores, the first inequality from the monotonicity property, and the final inequality because $P(B \mid A) \leq P(B)/P(A)$ for any two events A, B. Finally, taking the negative log-transform of each side, and using the definition of the conditional and marginal IT outlier scores, we reach the result.

A.11 PROOF OF PROPOSITION 15

In order to understand this proof, we need to introduce some terminology from the theory of 'total positivity' (see [Karlin, 1968]). A function $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is called totally positive of order two (TP_2) if for $x_1 \leq x_2$ and $y_1 \leq y_2$,

$$h(x_2, y_1)h(x_1, y_2) \leq h(x_1, y_1)h(x_2, y_2). \tag{59}$$

It is a fundamental result (see, for example, A.10 of [Marshall et al., 2011]) that the function

$$h(x, y) = g(y - x), \tag{60}$$

is TP_2 if and only if g is non-negative and log-concave on $\mathbb{R}$.

In our case, we have that

$$p(y \mid x) = p_N(y - f(x)), \tag{61}$$

where p_N is log-concave by assumption. As such, eq. (59) applies applies to $h(u, y) := p_N(y - u)$ for all $u_1, \leq u_2$ and $y_1 \leq y_2$. Furthermore, as f is non-decreasing, we have $x_1 \leq x_2 \implies f(x_1) \leq f(x_2)$, and so taking $u_i := f(x_i)$, we have that $h(f(x), y) = p(y \mid x)$ is TP_2 as a function of x and y.

Substituting $p(y \mid x)$ into eq. (59) and multiplying both sides by $p(x_1)p(x_2)$, we find that the joint density $p(x, y)$ satisfies the inequality too and is therefore TP_2. It is relatively straightforward to show (see theorem 5.2.19 of [Nelsen, 2006] and note that they refer to TP_2 in this context as PLR) that if $p(x, y)$ is TP_2 then $P(X \geq x \mid Y = y)$ is a non-decreasing function of y for all x. As such, by taking the expectation of $P(X \geq x \mid Y = y)$ with respect to $P(Y \mid Y \geq y)$ this implies

$$P(X \geq x \mid Y = y) \leq P(X \geq x \mid Y \geq y). \tag{62}$$

The final step is to note that, as both S_X, S_Y are strictly increasing we have $S(X) \geq S(x) \iff X \geq x$ and $S(Y) \geq S(y) \iff Y \geq y$, so that eq. (62) can be restated as precisely the monotonicity property of the proposition.

A.12 PROOF OF LEMMA 16

For clarity of exposition, denote X_i as Y and $\text{PA}_i^{\hat{\mathcal{G}}}$ as X. The probability of interest is $\tilde{P}(S(Y \mid X) \geq \lambda - \log c)$, for which we need a lower bound. First note,

$$\begin{aligned}
\tilde{P}(S(Y \mid X) \geq \lambda - \log c) &\geq \tilde{P}(S(Y) \geq \lambda, S(Y \mid X) \geq \lambda - \log c) && (63)\\
&= \tilde{P}(S(Y) \geq \lambda)\tilde{P}(S(Y \mid X) \geq \lambda - \log c \mid S(Y) \geq \lambda). && (64)
\end{aligned}$$

However, by assumption Y is a descendant of a root cause in $\mathcal{G}$, so we have, subject to effectiveness, that $\tilde{P}(S(Y) \geq \lambda) = 1$, and as such we only need to lower bound $\tilde{P}(S(Y \mid X) \geq \lambda - \log c \mid S(Y) \geq \lambda)$. Equivalently we need to upper bound $\tilde{P}(S(Y \mid X) < \lambda - \log c \mid S(Y) \geq \lambda)$.

Note now that, conditional on $S(Y) \geq \lambda$, for any outlier sample (x, y), clearly $S(y) \geq \lambda$, so that $P(S(Y) \geq S(y) \mid X = x) \leq P(S(Y) \geq \lambda \mid X = x)$, and by taking negative log-transforms that $S(y \mid x) \geq S(\lambda \mid x)$. Accordingly, for such samples, $S(y \mid x) < \lambda - \log c \implies S(\lambda \mid x) < \lambda - \log c$, and therefore that,

$$\begin{aligned}
\tilde{P}(S(Y \mid X) < \lambda - \log c \mid S(Y) \geq \lambda) &\leq \tilde{P}(S(\lambda \mid X) < \lambda - \log c \mid S(Y) \geq \lambda) && (65)\\
&\leq \frac{\tilde{P}(S(\lambda \mid X) < \lambda - \log c)}{\tilde{P}(S(Y) \geq \lambda)} && (66)\\
&= \tilde{P}(S(\lambda \mid X) < \lambda - \log c), && (67)
\end{aligned}$$

where the second inequality follows from the fact that $P(B \mid A) \leq P(B)/P(A)$ for any two events A, B, and the equality follows from effectiveness as before.

What is key is that this final expression now depends only on X, and by assumption X is not a root cause or a descendant of a root cause so $\tilde{P}_X = P_X$ and therefore

$$\tilde{P}(S(\lambda \mid X) < \lambda - \log c) = \tilde{P}_X(S(\lambda \mid X) < \lambda - \log c) = P_X(S(\lambda \mid X) < \lambda - \log c). \tag{68}$$

Finally, noting that $\mathbb{E}_{P_X}[P(S(Y) \geq \lambda \mid X)] = P(S(Y) \geq \lambda) \leq e^{-\lambda}$, we may apply Markov's inequality to conclude

$$\begin{aligned}
P_X(S(\lambda \mid X) < \lambda - \log c) &= P_X(P(S(Y) \geq \lambda \mid X) > ce^{-\lambda}) && (69)\\
&\leq \frac{e^{-\lambda}}{ce^{-\lambda}} = 1/c. && (70)
\end{aligned}$$

Putting everything together, we have the result.

A.13 PROOF OF LEMMA 17

By Bayes theorem

$$P_X(E_x \mid Y = y) = P_X(E_x)\frac{dP_{Y|E_x}}{dP_Y}(y), \tag{71}$$

so that by taking a negative log-transform, and applying the definition of IT outlier scores,

$$S(x \mid y) = S(x) - \log \frac{dP_{Y|E_x}}{dP_Y}(y). \tag{72}$$

Next, taking the expectation with respect to $P_{Y|E_x}$, we find

$$\mathbb{E}_{P_{Y|E_x}}[S(x \mid Y)] = S(x) - \mathbb{E}_{P_{Y|E_x}}\left[\log \frac{dP_{Y|E_x}}{dP_Y}(Y)\right]. \tag{73}$$

Note that the last term on the right is exactly $D(P_{Y|E_x} \mid\mid P_Y)$, so that by the SDPI and lemma 11, we have

$$\mathbb{E}_{P_{Y|E_x}}[S(x \mid Y)] \geq S(x) - \eta S(x) = (1-\eta)\, S(x). \tag{74}$$

A.14 PROOF OF LEMMA 18

For real-valued random variable X with finite variance, Cantelli's inequality (sometimes called the one-sided Chebyshev inequality) [Ghosh, 2002] is

$$P\left(X \geq \mathbb{E}[X] - \sqrt{\frac{1-\delta}{\delta}\mathrm{Var[X]}}\right) \geq 1 - \delta, \tag{75}$$

so that, in our case,

$$P_{Y|E_x}\left(S(x \mid Y) \geq \mathbb{E}[S(x \mid Y)] - \sqrt{\frac{1-\delta}{\delta}V_x}\right) \geq 1 - \delta. \tag{76}$$

We then may substitute $\mathbb{E}_{P_{Y|E_x}}[S(x \mid Y)]$ for its lower bound in lemma 17 and V_x for any finite upper bound.

A.15 PROOF OF PROPOSITION 19

By Bayes theorem

$$P_X(E_x \mid Y = y) = P_X(E_X)\frac{p_{Y|E_x}(y)}{p_Y(y)}, \tag{77}$$

so that by taking a negative log-transform, and applying the definition of IT outlier scores,

$$S(x \mid y) = S(x) - \log \frac{p_{Y|E_x}(y)}{p_Y(y)}. \tag{78}$$

Note that only the last term on the right depends on y, so that it is sufficient to bound it in order to bound $\mathrm{Var}_{P_{Y|E_x}}(S(x \mid Y))$ in y for each x. Noting first that $p(y \mid x) = p_N(y - f(x))$, let's first consider the log-Lipschitz case.

As p_N is log-Lipschitz we have that, for any a, b

$$\begin{aligned} |\log p_N(y - f(a)) - \log p_N(y - f(b))| &\leq L_N|(y - f(a)) - (y - f(b))| && (79)\\ &= L_N|f(a) - f(b)| && (80)\\ &\leq 2L_N B, && (81) \end{aligned}$$

where the final inequality is due to f being bounded. As such we have

$$e^{-2L_N B} \leq \frac{p_N(y - f(a))}{p_N(y - f(b))} \leq e^{2L_N B}, \tag{82}$$

so that

$$e^{-2L_N B} p_N(y - f(b)) \leq p_N(y - f(a)) \leq e^{2L_N B} p_N(y - f(b)). \tag{83}$$

Averaging first over a with respect to $P_{X|E_x}$, and then b with respect to P_X we have

$$e^{-2L_N B} p_Y(y) \leq p_{Y|E_x}(y) \leq e^{2L_N B} p_Y(y), \tag{84}$$

so that together, dividing by $p_Y(y)$ and taking log-transforms we have

$$\left| \log \frac{p_{Y|E_x}(y)}{p_Y(y)} \right| \leq 2L_N B. \tag{85}$$

Returning to eq. (78), this then immediately implies

$$S(x) - 2L_N B \leq S(x \mid y) \leq S(x) + 2L_N B, \tag{86}$$

holds for all x and y. Equivalently, we have $|S(x \mid y) - S(x)| \leq 2L_N B$. Finally, we have

$$\begin{aligned}
\mathrm{Var}_{P_{Y|E_x}}[S(x \mid Y)] &= \mathrm{Var}_{P_{Y|E_x}}[S(x \mid Y) - S(x)] && (87)\\
&= \mathbb{E}_{P_{Y|E_x}}[(S(x \mid Y) - S(x))^2] - \mathbb{E}_{P_{Y|E_x}}[S(x \mid Y) - S(x)]^2 && (88)\\
&\leq \mathbb{E}_{P_{Y|E_x}}[(S(x \mid Y) - S(x))^2] && (89)\\
&\leq (2L_N B)^2. && (90)
\end{aligned}$$

This completes the proof for part one of the proposition. Note in addition, however, that if we take one side of eq. (86), we simply have

$$S(x \mid y) \geq S(x) - 2L_N B, \tag{91}$$

and this bound holds uniformly, for all x and y, and is often stronger than that given by Cantelli's inequality in lemma 18.

We now turn our attention to part two of the proposition where $N \sim \mathcal{N}(0, \sigma^2)$. The proof follows a similar technique to part one, but we cannot now rely on log-Lipschitz continuity. First note, for any a and b,

$$\begin{aligned}
|\log p_N(y - f(a)) - \log p_N(y - f(b))| &= \frac{1}{2\sigma^2} \left| (y - f(b))^2 - (y - f(a))^2 \right| && (92)\\
&= \frac{1}{2\sigma^2} \left| 2y(f(a) - f(b)) + f(b)^2 - f(a)^2 \right| && (93)\\
&\leq \frac{1}{2\sigma^2} \left(4|y|B + B^2 \right) && (94)\\
&= \frac{2B}{\sigma^2}|y| + \frac{B^2}{2\sigma^2}, && (95)
\end{aligned}$$

where the first equality follows from p_N being the Gaussian density, and the inequality from $|f(a) - f(b)| \leq 2B$ and $|f(b)^2 - f(a)^2| \leq B^2$, both of which are due to f being bounded.

It follows from exactly the same steps as we followed in eqs. (82) to (85) that

$$\left| \log \frac{p_{Y|E_x}(y)}{p_Y(y)} \right| \leq \frac{2B}{\sigma^2}|y| + \frac{B^2}{2\sigma^2}. \tag{96}$$

For ease of exposition, let us rewrite the bound as

$$C(y) := \alpha|y| + \beta, \tag{97}$$

where $\alpha := 2B/\sigma^2$ and $\beta := B^2/2\sigma^2$, and noting that, as before, we have $\mathrm{Var}_{P_{Y|E_x}}[S(x \mid Y)] \leq \mathbb{E}_{P_{Y|E_x}}[C(Y)^2]$. As such, note

$$\mathbb{E}_{P_{Y|E_x}}[C(Y)^2] = \alpha^2 \mathbb{E}_{P_{Y|E_x}}[Y^2] + 2\alpha\beta \mathbb{E}_{P_{Y|E_x}}[|Y|] + \beta^2. \tag{98}$$

To give the final expression, we need to evaluate the two expectations. Noting that for $Y \sim P_{Y|E_x}$, we have $Y = f(X) + N$ where $X \sim P_{X|E_x}$ and $X \perp\!\!\!\perp N$, we have

$$\begin{aligned}
\mathbb{E}[Y^2] &= \mathbb{E}_{[}(f(X) + N)^2] && (99)\\
&= \mathbb{E}[f(X)^2 + 2f(X)N + N^2] && (100)\\
&= \mathbb{E}[f(X)^2] + \mathbb{E}_{P_{Y|E_x}}[N^2] && (101)\\
&\leq B^2 + \sigma^2 && (102)
\end{aligned}$$

where in the third equality we used the fact that $\mathbb{E}[2f(X)N] = 2\mathbb{E}[f(X)]\mathbb{E}[N] = 0$ as $\mathbb{E}[N] = 0$, and in the inequality that f is bounded.

By Cauchy-Schwarz we additionally have $\mathbb{E}[|Y|] \leq \sqrt{\mathbb{E}[Y^2]} \leq \sqrt{B^2 + \sigma^2}$. Finally, by substituting both bounds on the expectations and the definitions of α and β, we have the result.

A.16 PROOF OF PROPOSITION 21

All three bounds follow from substituting lower bounds on each of the corresponding test statistics into each of their p-value functions. In the 'worst-case', all the conditional scores (or marginal score jumps) from among the πL non-root causes that are subject to errors in $\hat{\mathcal{G}}$ are equal to b, and all of the remaining $(1-\pi)L$ conditional scores (or marginal score jumps) are 0. In such a case we have, in the case for marginal score jumps,

$$\hat{s}^{\hat{\mathcal{G}}}_{\text{sum}} = \sum_{i \notin \mathcal{R}} |S(x) - S(\text{pa}_i^{\hat{\mathcal{G}}})|_+ = \pi L b, \tag{103}$$

$$\hat{s}^{\hat{\mathcal{G}}}_{\max} = \max_{i \notin \mathcal{R}} |S(x) - S(\text{pa}_i^{\hat{\mathcal{G}}})|_+ = b, \tag{104}$$

$$\hat{n}_t = \sum_{i \notin \mathcal{R}} \mathbb{I}[S(x_i \mid \text{pa}_i^{\hat{\mathcal{G}}}) \geq t] = \begin{cases} \pi L & \text{if } b \geq t, \\ 0 & \text{otherwise.} \end{cases} \tag{105}$$

With the same bounds applying to the conditional score test p-values with the same reasoning (though under different assumptions).

A.17 PROOF OF THEOREM 22

The following is subject to alternative hypothesis H_1. For convenience, denote the CDF of the standard exponential distribution $F_{\text{Exp}}(s) = (1 - e^{-s})$

$$\begin{aligned} &\sup_s |F(s) - F_{\text{Exp}}(s)| \geq \sup_{s<b} |F(s) - F_{\text{Exp}}(s)| \\ &= \sup_{s<b} |(1-\pi)F_{\text{Exp}}(s) - F_{\text{Exp}}(s)| = \sup_{s<b} |\pi F_{\text{Exp}}(s)| \\ &= \pi F_{\text{Exp}}(b). \end{aligned} \tag{106}$$

Let F_L be the empirical distribution of F for L variables, our secondary test statistic is the Kolmogorov-Smirnov statistic D_L for L samples. We then have,

$$\begin{aligned} D_L &:= \sup_s |F_L(s) - F_{\text{Exp}}(s)| \\ &= \sup_s |F_L(s) - F(s) + F(s) - F_{\text{Exp}}(s)| \\ &\geq \sup_s |F(s) - F_{\text{Exp}}(s)| - \sup_s |F_L(s) - F(s)| \\ &\geq \pi F_{\text{Exp}}(b) - \sup_s |F_L(s) - F(s)|, \end{aligned} \tag{107}$$

where the first inequality follows from the triangle inequality, and the second from Equation 106. For $\pi > 0$, and significant threshold α, the power of the test is $P(D_L \geq D^{\text{crit}}_{L,\alpha})$. If $\pi F_{\text{Exp}}(b) > D^{\text{crit}}_{L,\alpha}$, then for all $\varepsilon > 0$ such that $\pi F_{\text{Exp}}(b) - \varepsilon \geq D^{\text{crit}}_{L,\alpha}$, we have that $\sup_s |F_L(s) - F(s)| \leq \varepsilon$ implies $D_L \geq D^{\text{crit}}_{L,\alpha}$ by Equation 107. Therefore, in such cases,

$$P(D_L \geq D^{\text{crit}}_{L,\alpha}) \geq P(\sup_s |F_L(s) - F(s)| \leq \varepsilon). \tag{108}$$

Finally, by the Dvoretzky–Kiefer–Wolfowitz inequality [Dvoretzky et al., 1956, Massart, 1990], for all $\varepsilon > 0$,

$$P(\sup_s |F_L(s) - F(s)| \leq \varepsilon) \geq 1 - 2e^{-2L\varepsilon^2}, \tag{109}$$

so that picking $\varepsilon = |\pi F_{\text{Exp}}(b) - D^{\text{crit}}_{L,\alpha}|_+$, and combining Equation 108 with Equation 109, we have the result. Note that the theorem trivially holds for $\pi F_{\text{Exp}}(b) \leq D^{\text{crit}}_{L,\alpha}$ too.

A.18 PROOF OF THEOREM 23

Taking the likelihood ratio for a single variable we have,

$$\frac{\pi e^{-(S(x_i|\mathrm{pa}_i^{\hat{\mathcal{G}}})-t)}\mathbb{I}[S(x_i \mid \mathrm{pa}_i^{\hat{\mathcal{G}}}) \geq t] + (1-\pi)e^{-S(x_i|\mathrm{pa}_i^{\hat{\mathcal{G}}})}}{e^{-S(x_i|\mathrm{pa}_i^{\hat{\mathcal{G}}})}} = \pi e^t \mathbb{I}[S(x_i \mid \mathrm{pa}_i^{\hat{\mathcal{G}}}) \geq t] + (1-\pi) \tag{110}$$

So for the full sample, the ratio is

$$\Lambda_b = \prod_{i \notin \mathcal{R}} \left(\pi e^t \mathbb{I}[S(x_i \mid \mathrm{pa}_i^{\hat{\mathcal{G}}}) \geq t] + (1-\pi)\right) \tag{111}$$

$$= \left(\pi e^t + (1-\pi)\right)^{N_t} (1-\pi)^{L-N_t}, \tag{112}$$

Note that, for fixed t, and for every π, likelihood ratio is monotonically increasing with increasing N_t so that a test based on the upper tail of the ratio is equivalent to one on the upper tail of N_t. By the Karlin-Rubin theorem [Karlin and Rubin, 1956], for fixed t, proposition 4 defines a test which is uniformly most powerful against the alternative $H_1^{\pi,t}$.

A.19 PROOF OF PROPOSITION 24

Let $H_0^{\mathcal{C}}$ be the null hypothesis that all x_i with $i \notin \mathcal{C}$ were drawn according to $P(X_i \mid \mathrm{PA}_i^{\hat{\mathcal{G}}} = \mathrm{pa}_i^{\hat{\mathcal{G}}})$. Since $|\mathcal{R}| \leq k$, there exists a set $\mathcal{C}^* \subseteq \{1, \ldots, n\}$ with $|\mathcal{C}^*| = k$ and $\mathcal{R} \subseteq \mathcal{C}^*$, which is included in the maximisation defining p_T^k. As such, note that H_0 implies $H_0^{\mathcal{C}^*}$: if H_0 is true, then after excluding the variables in $\mathcal{C}^*$, all remaining variables are non-root causes, and so are all drawn according to their true causal mechanisms as the causal conditionals implied by $\hat{\mathcal{G}}$ coincide with those in $\mathcal{G}$. Rejecting $H_0^{\mathcal{C}^*}$ is therefore sufficient to reject H_0. As $p_T^{\mathcal{C}^*}$ is a p-value for $H_0^{\mathcal{C}^*}$, it is therefore a p-value for H_0. Finally note,

$$P_{H_0}\left(\max_{\substack{\mathcal{C} \subseteq \{1,\ldots,n\} \\ \text{s.t. } |\mathcal{C}|=k}} p_T^{\mathcal{C}} \leq \alpha\right) = P_{H_0}\left(\bigcap_{\mathcal{C}} \{p_T^{\mathcal{C}} \leq \alpha\}\right) \leq P_{H_0}(p_T^{\mathcal{C}^*} \leq \alpha) \leq \alpha, \tag{113}$$

so p_T^k is a p-value for H_0.

B ADDITIONAL MARGINAL OUTLIER DECAY RESULT

Lemma 12 holds exactly for $S(y \mid E_x) = 0$, however, more generally we have,

Proposition 25. *For any x such that $P(E_x) > 0$, let y be chosen at random according to $P(Y \mid E_x)$. With probability at least $1-\delta$, we obtain a y for which,*

$$\eta S(x) \geq \delta S(y) - \log 2. \tag{114}$$

Proof. Denote for convenience $p(y) := P_{Y|E_x}(E_y)$ and $q(y) = P_Y(E_y)$. We may conclude from eq. (54) in the proof of lemma 12 that

$$\eta S(x) \geq p(y) \log \frac{p(y)}{q(y)} + (1-p(y)) \log \frac{1-p(y)}{1-q(y)} \tag{115}$$

$$\geq p(y) \log \frac{p(y)}{q(y)} + (1-p(y)) \log(1-p(y)) \tag{116}$$

$$= p(y) \log p(y) + (1-p(y)) \log(1-p(y)) - p(y) \log q(y) \tag{117}$$

$$= -H_b(p(y)) - p(y) \log q(y), \tag{118}$$

where the second inequality follows from $-\log(1-q(y)) \geq 0$, and H_b denotes the binary entropy function. Noting that $\max_{u \in [0,1]} H_b(u) = \log 2$, and rewriting $p(y)$ and $q(y)$ we therefore have,

$$\eta S(x) \geq e^{-S(y|E_x)} S(y) - \log 2. \tag{119}$$

By the properties of IT scores $P_{Y|E_x}(S(Y|E_x) \leq c) = 1 - e^{-c}$, so that

$$P_{Y|E_x}\left(\eta S(x) \geq e^{-c} S(Y) - \log 2\right) \geq 1 - e^{-c}. \tag{120}$$

The result is then given by defining $\delta := e^{-c}$.

□

C UPPER BOUNDS ON P-VALUES

In fig. 3 we can see the upper bounds on the p-values of our tests from proposition 21. The bound for $p_{S^{\mathcal{G}}_{\max}}$ is constant with respect to π, and so can be strongly preferable to the other tests when π is small. In contrast $p_{S^{\mathcal{G}}_{\text{sum}}}$ and p_{N_t} are often clearly preferable for relatively low b as π increases. Notably, the tail decays much more sharply with increasing π for p_{N_t} compared to $p_{S^{\mathcal{G}}_{\text{sum}}}$ for suitable settings of t, but it is insensitive, beyond the threshold, to increases in b.

These upper bounds can also help to understand how the power of our tests depend on the size of the graph being tested. Supposing the number of root causes is fixed, then the size of the graph is characterised by L. If we assume π and b are fixed then as L increases, the upper bound on the p-value for the Tippett's method test tends to 1. For the Fisher's method test and the Binomial test, whether the upper bounds on their p-values tends to 0, tends to $1/2$ or tends to 1 depends on the relative values of π and b in the case of the former, and on the relative values of π and t for the latter. If we also consider the lower bound on the power for the Kolmogorov-Smirnov test given in theorem 22, we see it tends to 1 as L grows. As such, the Kolmogorov-Smirnov test becomes more powerful for larger graphs, the Tippett's method test suffers, and the remaining tests may become more powerful, less powerful or remain stable depending on the other factors. In practice, we might expect both the number of root causes, the strength of the outlier among descendants (quantified by b) and the proportion of errors to vary with graph size, and so to give precise characterisation on power in this setting would require making modelling assumptions concerning how and when these parameters vary for larger L.

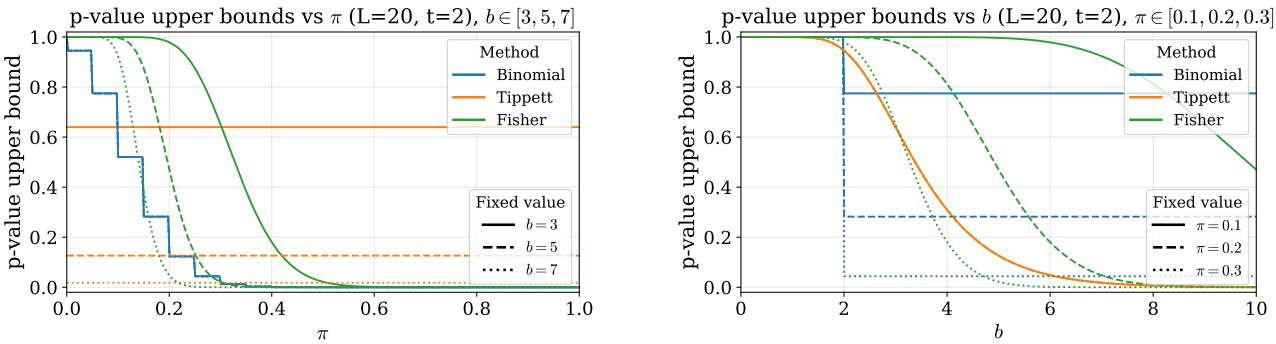
D WHEN IS EACH TEST UNIFORMLY MOST POWERFUL

As the first test is simply an IT score analogue to Fisher's method, as mentioned, we can adapt a result from Heard and Rubin-Delanchy [2018],

Theorem 26 (Optimality of sum of conditional outlier scores test). *The test defined by $p_{S^{\hat{\mathcal{G}}}_{\text{sum}}}$ in proposition 2 is uniformly most powerful against the alternative*

$$H_1\colon\ S(X_i \mid \text{PA}_i^{\hat{\mathcal{G}}}), \text{ for } i \notin \mathcal{R}, \text{ are iid exponential random variables with rate parameter } \nu < 1.$$

Proof. Immediate consequence of lemma 1 and example 1 in [Heard and Rubin-Delanchy, 2018]. □



(a) (b)

Figure 3: Bounds on the p-values of our tests.

Similarly, for the second test,

Theorem 27 (Optimality of the max conditional outlier score test). *The test defined by $p_{S^{\hat{\mathcal{G}}}_{\max}}$ in proposition 3 is uniformly most powerful against the alternative H_1 :*

$$S(X_j \mid \mathrm{PA}_j^{\hat{\mathcal{G}}}) \sim \frac{1}{B(\alpha, L)} e^{-\alpha s}(1-e^{-s})^{L-1}, \quad j \notin \mathcal{R}, \alpha \in (0,1),$$
$$S(X_i \mid \mathrm{PA}_i^{\hat{\mathcal{G}}}) \sim \frac{e^{-s}}{1-e^{-S(x_j \mid \mathrm{pa}_j^{\hat{\mathcal{G}}})}}, \quad i \neq j \notin \mathcal{R}, s \in [0, S(x_j \mid \mathrm{pa}_j^{\hat{\mathcal{G}}})],$$

where α is a shape parameter of the Beta distribution, and B is the beta function.

Proof. Consequence of our lemma 1 and proposition 2 in [Heard and Rubin-Delanchy, 2018] after a change of variables. □

E FURTHER COMMENTS ON RELATED WORK

In section 2 we cite a number of related works that may initially appear to be suitable candidates for comparison with our tests in section 5, and so the question may arise as to why we do not include them. We address first the goodness-of-fit tests developed by Canonne et al. [2017], Schkoda and Drton [2025]. In short, it is not clear how such a comparison would be done. The approach of Schkoda and Drton [2025] tests the hypothesis that some (observational) data is consistent with *any* linear structural equation model, and so does not allow for testing a given or estimated causal graph, as the null hypothesis is agnostic as to the underlying causal graph. [Canonne et al., 2017] is more complex. Firstly, they restrict themselves to binary variable Bayesian networks in the observational setting (whereas we are concerned with the continuous setting with outlier(s)). Secondly, they do not propose a test in the classical hypothesis testing sense, but rather a 'testing algorithm' which rejects if the data distribution is at least a user-specified ε away (in L_1 distance) from any binary Bayesian network with the given graph, with some probability. As such, it is not clear what a suitable ε should be if we were to use their approach as the basis of comparison. Accordingly, it is not clear how we would perform comparisons with their approach either.

Similar comments can be made regarding the approaches of Faller et al. [2024] and Schkoda et al. [2025]. Both approaches falsify, in the observational setting, causal discovery algorithms (relative to a given observational dataset) rather than falsifying a given causal graph, and so a what would constitute an appropriate direct comparison with out tests is not clear.

Another work that is relevant to discuss is that of Gnecco et al. [2021] as, at a high-level their work also uses the properties of tail events to infer causal directions (this work for falsification, theirs for causal discovery). However, there are key differences. Firstly, their model is parametric. Secondly, for them, extreme events are expected since the distribution is long-tailed, while we model extreme events as observations that are not drawn from the hypothetical distribution. This second difference is actually just a change of perspective, after all, we could also consider a mixture of the distribution considered in the null hypothesis with one that describes the outlier. The first difference, however, is more significant because their model predicts that, if $X \to Y$, then extreme values in X are *likely* to cause extreme values in Y, while this need not be the case in our non-parametric setting.

F ADDITIONAL EXPERIMENTS AND FURTHER DETAILS

F.1 EXPERIMENTAL DETAILS

For the experiment in figs. 1 and 5 we generated 200 DAGs randomly according to an Erdős-Rényi model with 20 nodes and three expected edges per node. We sample coefficients for linear structural equations uniformly from $[-1, 3]$[1]. We then add standard Gaussian noise, while root nodes are with equal probability either Gaussian or uniform noise, or are a mixture of two to four Gaussian distributions. The means are picked uniformly from $[-1, 1]$ and also the number of components is picked uniformly. For each dataset, we generate 1000 samples. We then uniformly pick a node as root cause and introduce an anomaly, by increasing (or decreasing) the noise value by three standard deviations, with probability 1/2, respectively for fig. 1 and six standard deviations for fig. 5. We reject an anomalous sample if no node (not even the root cause) has a

[1]We observed empirically that an asymmetric interval allows the anomalies to propagate further through the graphs.

marginal anomaly score of at least 3. Unless stated otherwise, we always assume to only have access to a single anomalous sample.

In order to create the node-expert graphs with knowledge fraction $|K|/|N|$, where $K \subseteq N := \{X_1, \dots, X_n\}$ is the set of nodes where the expert is correct, we start by uniformly at random picking these $|K|$ nodes. We then pick a causal ordering uniformly at random, such that the relative ordering between the nodes in K is preserved. We keep the edges between nodes in K. We then calculate the size of the graph theoretic cut of K, i.e. $c := |\{(X_i, X_j) \in \mathcal{G} \mid X_i \notin K \vee X_j \notin K\}|$. Finally, we uniformly sample c new edges from the set of possible edges (between nodes in K and $N \setminus K$) that respect our new causal order.

In order to estimate the marginal outlier scores, we leverage methods from the Python package `DoWhy GCM` [Blöbaum et al., 2022]. Specifically, we use the rescaled median CDF quantile scorer. This empirically estimates the score

$$S(X_i) = -\log\left(2\min\left\{P(X_i > x_i) + \frac{P(X_i = x_i)}{2}, P(X_i < x_i) + \frac{P(X_i = x_i)}{2}\right\}\right),$$

by counting samples that are larger or smaller than x_i in the given non-anomalous dataset, respectively, and taking two times the minimum of these numbers. The score is then rescaled with the negative natural logarithm. It is worth noting that while our theoretical results hold for any real-valued τ (so long as we satisfy continuity), in practice one should choose a feature map suited to their particular application, as it must be large when the corresponding variable should be considered 'extreme' in that application. We do not opt for the median rescaled CDF quantile scorer for any special reason other than it is a broadly applicable, non-parametric anomaly scorer: the z-score would have also been suitable, but in general we leave this up to the user.

For the conditional scores, we first estimate causal mechanisms using the `auto` module from `DoWhy`. This means, for each node X_i, we fit an additive noise model f_i for the conditional distribution using several linear and non-linear estimators from the `scikit-learn` package [Pedregosa et al., 2011], where we use the `better` parameter from `DoWhy`. The best model is picked via five-fold cross-validation. Then synthetic samples from the conditional of X_i given pa_i can be generated by predicting $f(\mathrm{pa}_i)$ and drawing noise uniformly from the empirical residuals $x_i^j - f_i(\mathrm{pa}_i^j)$, where x_i^j and pa_i^j denote the values of X_i and its parents in the given non-anomalous samples for $j = 1, \dots, m \in \mathbb{N}$. Finally, the conditional anomaly scores $S(x_i \mid \mathrm{pa}_i)$ are estimated using the rescaled median CDF quantile scorer on a synthetic sample of the conditional with 10000 samples.

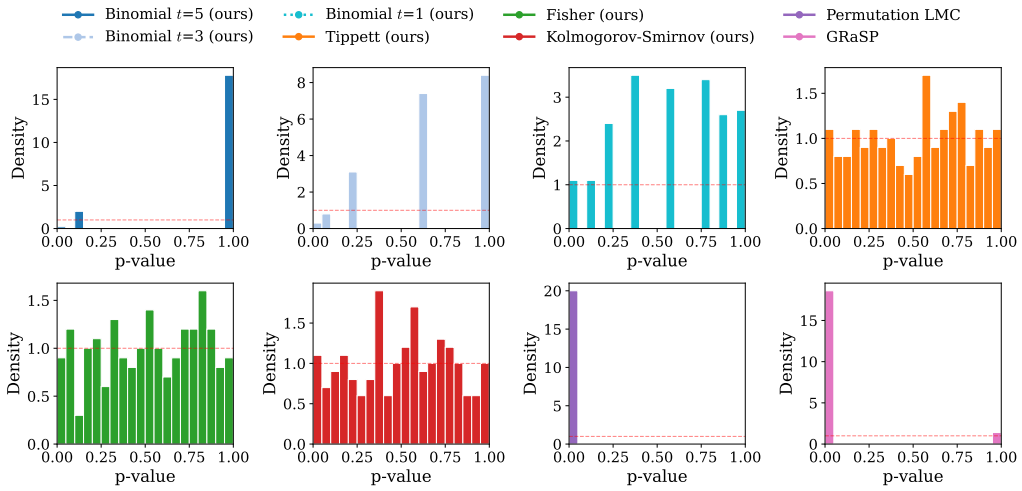
For the test proposed by Eulig et al. [2025], we use the implementation from `DoWhy`, but with the Fisher Z test, as implemented in the `causal-learn` package [Zheng et al., 2024]. We use 200 permutations and set the significance threshold of the independence tests to $\alpha = 0.05$. For GRaSP, we also use an implementation from `causal-learn` and use default parameters. Since GRaSP only outputs a CPDAG, representing the Markov-equivalence class of the ground truth, we cannot calculate the structural Hamming distance directly.

Concerning the datasets in figs. 2 and 11, we only focus on the latency data from PetShop, drop metrics with fewer than ten unique values and use mean-imputation for missing values. The dataset provides three different traffic scenarios and we consider 18 anomaly events. As preprocessing, we take a single sample and conduct our test with this sample and the provided non-anomalous data for each event. fig. 2 shows the average rejection rate over these runs with the marginal version of our test and the baselines. For Causal Chambers, we use the `lt_interventions_standard_v1` dataset provided on the website. We drop constant columns and intervention indicators from the dataset and ground truth graph, and also take a single sample from one of the 58 provided interventional scenarios.

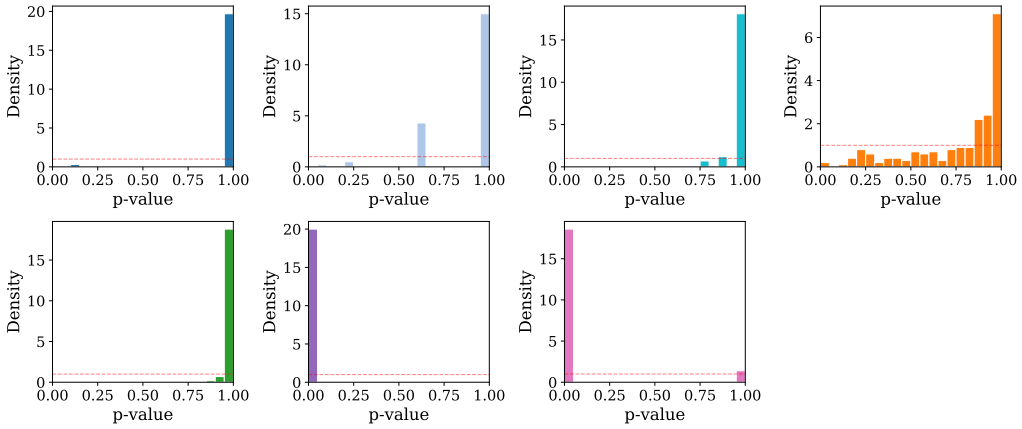
Each experiment can be run within 48 hours on an EC2 instance of type `c5.4xlarge`.

F.2 HISTOGRAMS OF P-VALUES

In fig. 4 we show the p-values of all tests from fig. 1 for $|K|/|N| = 1$, i.e. under the null hypothesis. We can see that the p-values of our conditional tests are roughly uniformly distributed, except for the Binomial test for larger values of t where the tests are conservative even for the conditional score tests. The marginal score tests are, as expected, consistently more conservative than their conditional score counterparts (and indeed seem to be super-uniformly distributed).



(a) Conditional scores.



(b) Marginal scores.

Figure 4: Histograms of the p-values of our tests under the null hypothesis.

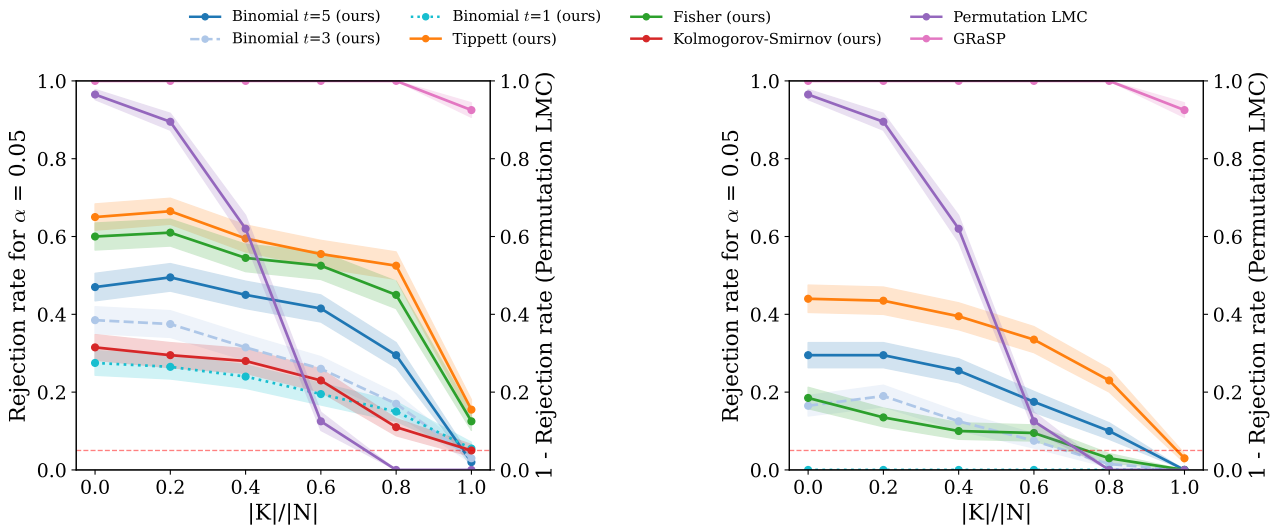
F.3 STRONGER ANOMALIES

In fig. 5 we repeated the experiment from fig. 1 with the only difference being that we inject stronger anomalies. In particular, we increase or decrease the noise of the root cause by six instead of three standard deviations. We can see that this indeed increases the power of our test.

F.4 MORE ANOMALIES

It is a direct consequence of the proofs of propositions 2 to 5 and 8 that we can also conduct the tests with more than a single sample, even if these anomalies correspond to different root cause events. In order to do this, we simply need to drop all root cause scores (or score gaps) and combine the p-values of the remaining nodes for all samples. Specifically, for proposition 2 this means summing over all of them to get $S^{\mathcal{G}}_{\text{sum}}$. For proposition 3 this means taking the max over all of them for $S^{\mathcal{G}}_{\text{max}}$. For proposition 4 we threshold and sum over all nodes and for proposition 5 we use all these scores to estimate the empirical distribution of the conditional scores.

In fig. 6 we repeated the experiments from figs. 1 and 5 with 100 anomalous samples (with identical root cause) per anomaly event. As we can see, especially the tests based on conditional scores gain power this way. The gains for the marginal tests are not as pronounced, potentially because the multiple-testing burden overwhelms the weak signal in the data in many cases. Interestingly, we can see that in this scenario the test based on Fisher's method and marginal scores seems to outperform Tippett's version, which is the opposite behaviour of the experiments in figs. 1 and 5. This is in line with our theory, as



(a) Conditional scores.

(b) Marginal scores.

Figure 5: Average rejection rate for synthetic data with larger anomalies and increasing fraction of correctly connected nodes $|K|/|N|$. We observe a similar behaviour as with the smaller anomalies, except that our tests seem to have even more power.

Tippett is expected to have optimal power if there is a high outlier score and Fisher when many of the combined scores are elevated.

F.5 NO GIVEN ROOT CAUSE

In fig. 7 we repeated the experiments from figs. 1 and 5 without assuming to know the root cause. We can see that the proposed approach loses very little power in our experiments.

F.6 GRAPHS FROM CAUSAL DISCOVERY

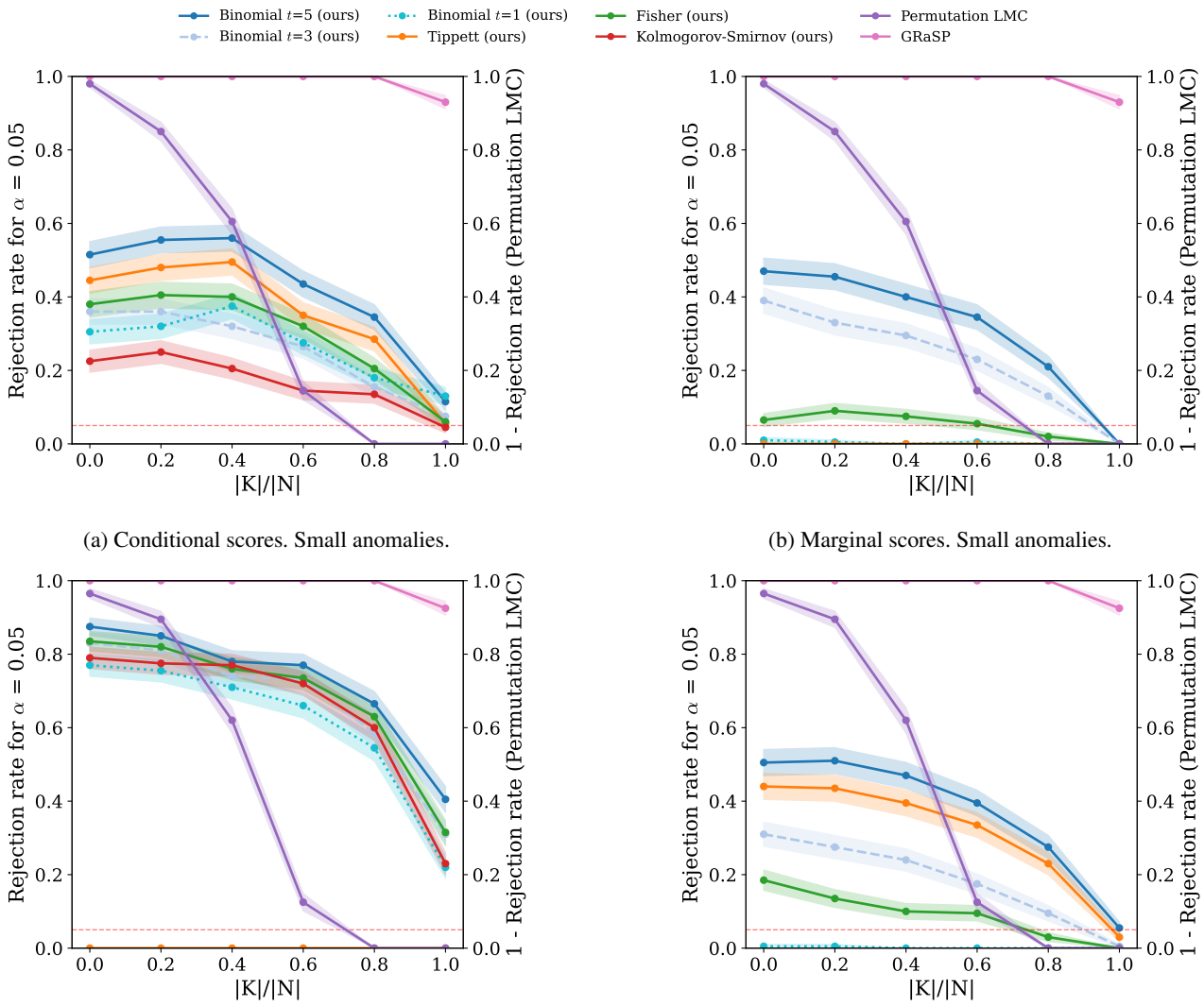
We further wanted to investigate the performance of our tests on graphs that have been estimated using a causal discovery method. We generate the data analogously to before, except that we draw all noise terms uniformly from $[-1, 1)$. This renders the data amenable for the DirectLinGAM algorithm [Shimizu et al., 2011]. We estimate 200 graphs on samples of size 50, 100, 200, 400, 800 and 1600, respectively. In fig. 8 we sort each resulting graph in one of five bins of equal size with respect to the structural Hamming distance to the ground truth graph, i.e. the number of edges that would need to change in order to match the ground truth graph. As we can see, the rejection rate of our tests increases as SHD gets larger. In contrast to previous experiments, the permutation test baseline rejects most graphs (i.e. considers them *better* than random). This could be due to the fact that the graphs in this experiment are learned from the given data and therefore indeed represent the conditional independence structure of the data better than a random graph.

F.7 NON-LINEAR MECHANISMS

In fig. 9 we repeated the experiment from fig. 1 with the main difference that we generate the synthetic data from a model with non-linear causal mechanisms. Precisely, we generate variable X_i with non-empty parent set via polynomials of degree two in their parents, i.e.

$$x_i = \sum_{X_j, X_k \in \mathrm{PA}_i^{\mathcal{G}}} \alpha_{j,k} x_j x_k + n_i,$$

where n_i is an independent standard Gaussian noise term like before and $\alpha_{i,j} \in \mathbb{R}$. We initially draw $\alpha_{i,j}$ uniformly from $[-1, 1]$ and generate an (unnormalised) sample from these coefficients and the normal-regime samples from the parents. We then divide the coefficients by half the empirical standard deviation of this initial sample and re-generate the sample from



(a) Conditional scores. Small anomalies.

(b) Marginal scores. Small anomalies.

(c) Conditional scores. Large anomalies.

(d) Marginal scores. Large anomalies.

Figure 6: Average rejection rate for synthetic data with 100 anomalous samples and increasing fraction of correctly connected nodes $|K|/|N|$. We observe similar trends as before just with even more power.

the same parent values. This way we get normal data with stable variances but in the anomalous regime outliers can still propagate.

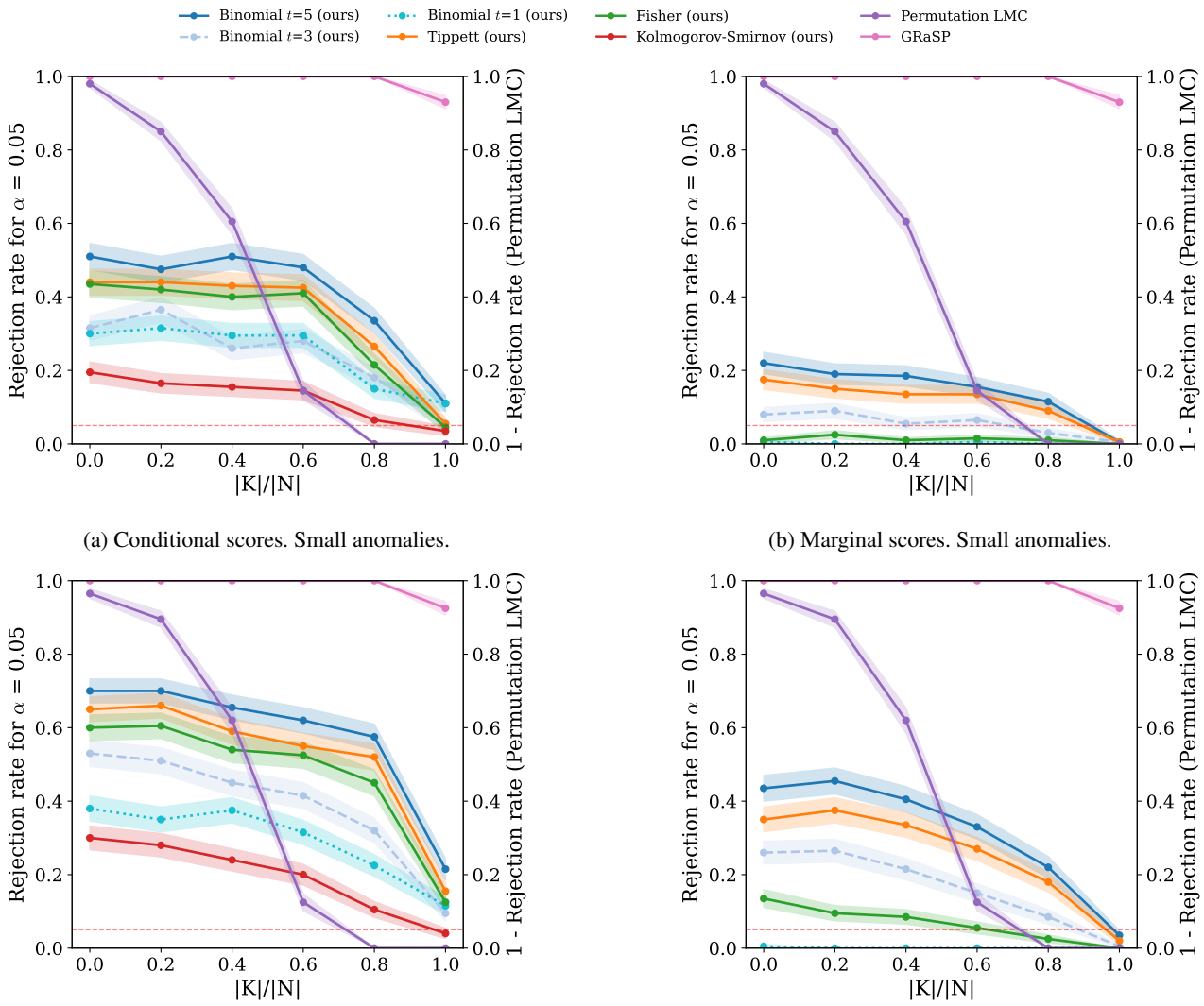
For the permutation test baseline we used the generalised covariance measure [Shah and Peters, 2020] as implemented in `DoWhy GCM` and we use gradient boosting regression as implemented in `scikit-learn` to estimate the conditionals.

For computational reasons we only consider 20 different DAGs in this experiment.

As we can see the tests behave similarly to our previous experiments.

F.8 INTERVENTIONAL SAMPLES FOR PERMUTATION BASELINE

In fig. 10 we repeated the experiment from fig. 1 and added a so-called F-node [Pearl, 2000] to the graph used by the permutation test baseline. More precisely, we concatenate each normal dataset with the outlier sample and add a column F that indicates whether the given row is from the normal or outlier regime. We also add a node F to the estimated causal graph that has a single outgoing edge pointing to the known root cause. We then proceed as before. As expected, the single outlier sample is not enough to substantially change the behaviour of the test.



(a) Conditional scores. Small anomalies.

(b) Marginal scores. Small anomalies.

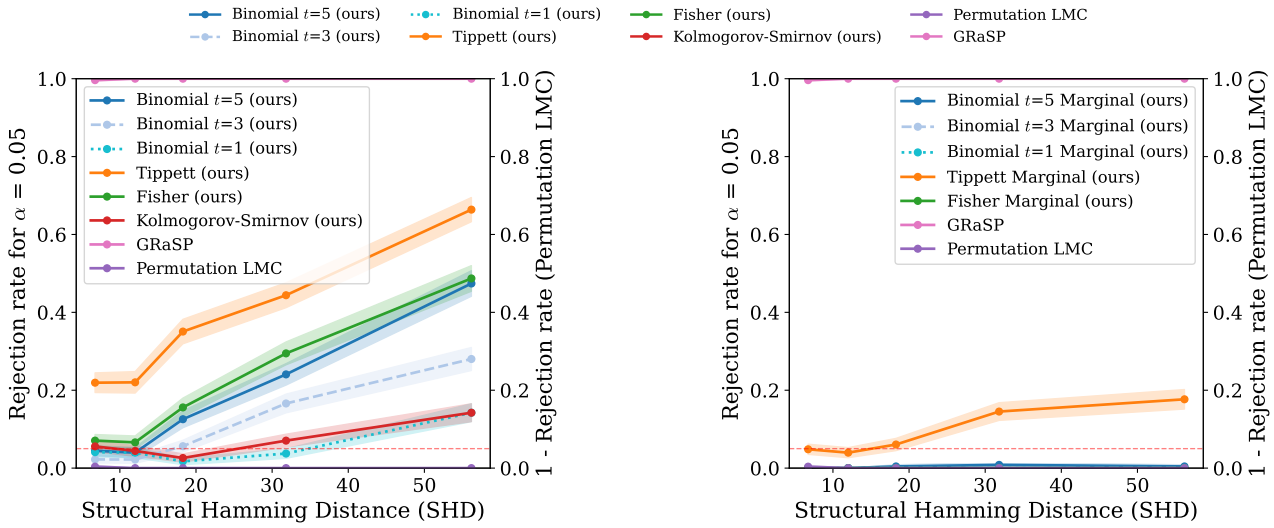
(c) Conditional scores. Large anomalies.

(d) Marginal scores. Large anomalies.

Figure 7: Average rejection rate for synthetic data when no root cause is given and increasing fraction of correctly connected nodes $|K|/|N|$. We observe similar trends as before just slightly less power.

F.9 CONDITIONAL TESTS ON PETSHOP AND CAUSAL CHAMBERS DATA

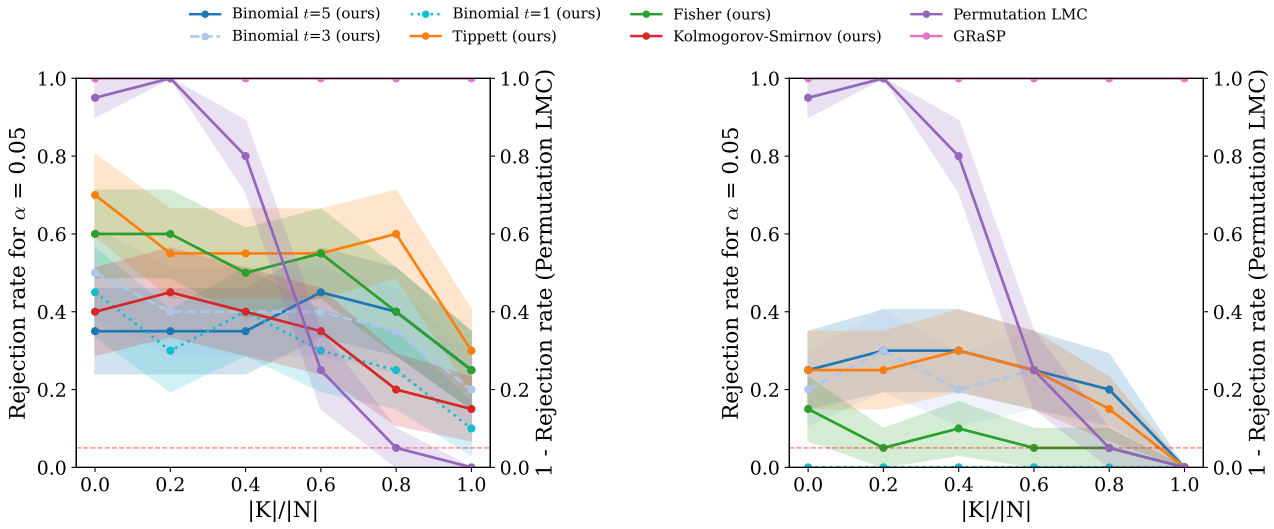
In fig. 11 we repeated the experiments in fig. 2 but with the conditional version of the tests from propositions 2 to 5. We can see that the tests reject even more often than their marginal counterparts. In contrast to their marginal counterparts, we can see that also the Causal Chambers graph gets rejected slightly more often than we would expect from the selected significance threshold. Comparing with the synthetic data in figs. 1a and 5a, this might be due to statistical errors in the estimation of the conditional scores.



(a) Conditional scores.

(b) Marginal scores.

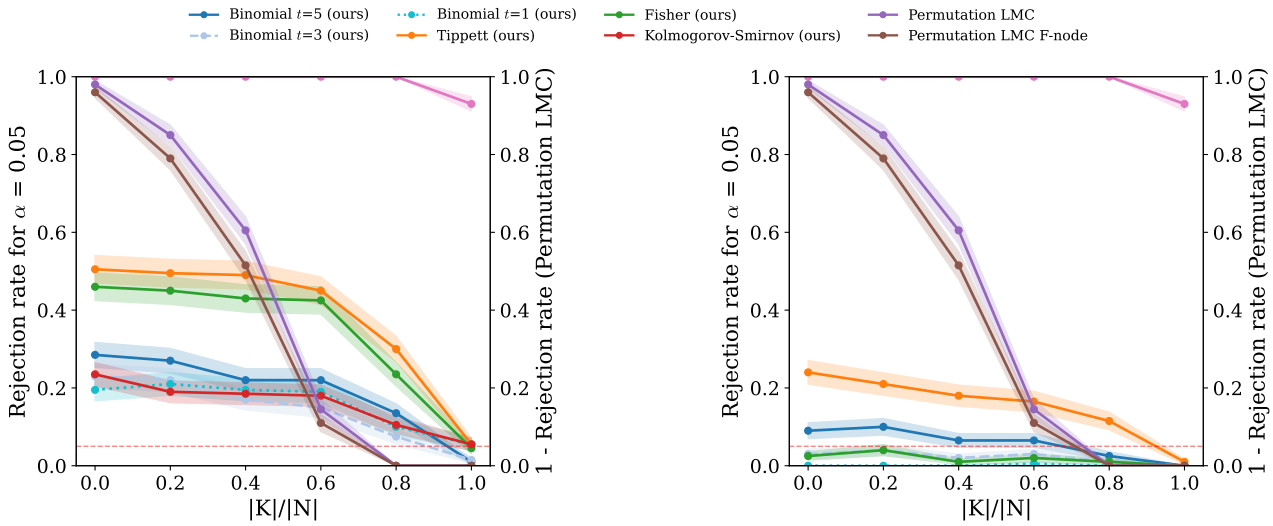
Figure 8: Average rejection rate for synthetic data with linear non-Gaussian data and graphs estimated using DirectLiNGAM [Shimizu et al., 2011].



(a) Conditional scores.

(b) Marginal scores.

Figure 9: Average rejection rate for synthetic data with non-linear causal mechanisms and increasing fraction of correctly connected nodes $|K|/|N|$.



(a) Conditional scores.

(b) Marginal scores.

Figure 10: Average rejection rate for synthetic data with an increasing fraction of correctly connected nodes $|K|/|N|$. We added an F-node to the permutation test baseline.

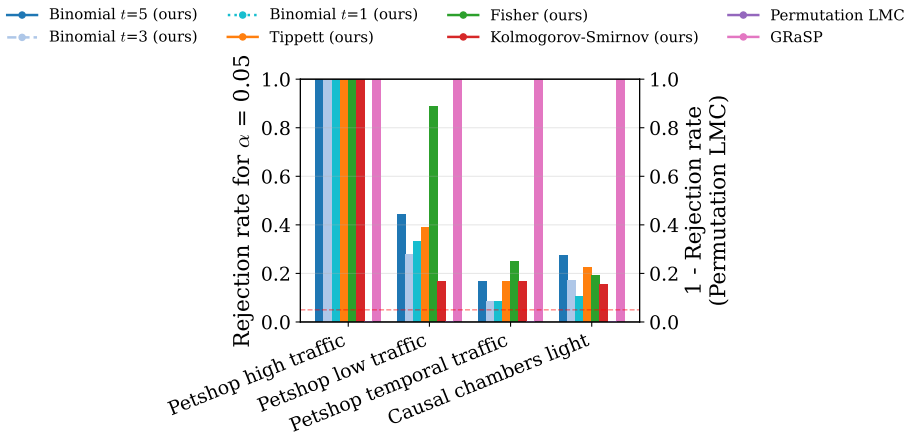


Figure 11: Average rejection rate of our conditional tests on cloud computing and causal chambers data. While the inverted call graph seems to represent the cloud data better than random, our test rejects this graph as ground truth for several anomaly events. The graph from causal chambers gets rejected more rarely but still slightly above the prescribed false positive rate.