ASSESSING TIME-SERIES CAUSAL DISCOVERY WITHOUT GROUND TRUTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Assessing the relevance of a causal graph inferred from observational data remains limited when no reference graph is available, since standard structural and interventional metrics require ground-truth causal relations. We extend the Causal Relevance Index (CRI) to lagged causal discovery by aggregating edge-specific p-value bounds over a fixed set of candidate lagged relations. The main graph-level properties of the resulting index are shown to follow from a marginal condition on the conditional independence tests: for every absent edge, a super-uniform p-value obtained for at least one tested separating set provides a sufficient condition. This result requires no assumptions on the joint dependence structure among the edge-specific quantities. Sufficient conditions for asymptotic super-uniformity under temporal dependence are derived for Fisher-calibrated partial-correlation tests. An empirical normalization based on iterative amplitude-adjusted Fourier transform surrogates further accounts for configuration-specific null levels. The resulting criteria are evaluated on nonlinear synthetic time series spanning several graph structures, sample sizes, causal discovery algorithms, and conditional independence tests, as well as on measurements from the Causal Chambers wind-tunnel system. In the synthetic experiments, normalized CRI shows a positive association with recall and F_1 for PCMat-HSIC across graph structures and sample sizes. The index summarizes statistical support from the underlying tests without requiring a reference graph.

1 INTRODUCTION

Causal graphs for multivariate time series represent causal relations at different temporal resolutions (Runge et al., 2023). Window graphs restrict the representation to a finite range of lags, whereas summary graphs omit lag information (Assaad et al., 2022; Gong et al., 2024; Assaad et al., 2024). The relevance of an inferred causal graph can be assessed through agreement with a reference graph at the selected temporal resolution. Precision, recall, F_1, false positive and false negative rates, and the Matthews correlation coefficient characterize missing and spurious edges (Raghu et al., 2018; Moraffah et al., 2021; Cheng et al., 2022). Orientation accuracy complements adjacency recovery (Cheng et al., 2022). Structural Hamming distance (SHD) quantifies discrepancies through the edge operations required to transform the estimated graph into the reference graph (Tsamardinos et al., 2006), with separate formulations for DAGs and CPDAGs (Rehak et al., 2024). Window- and summary-graph scores assess different target structures and require separate interpretations (Assaad et al., 2022).

Edge recovery criteria and SHD quantify structural discrepancies without describing their consequences for causal effect identification. Structural intervention distance (SID), introduced by Peters & Bühlmann (2015), addresses this limitation by counting ordered pairs (i, j) associated with intervention distributions incorrectly inferred from the estimated DAG H relative to the reference DAG G through parent adjustment. The adjustment identification distance (AID) framework introduced by Henckel et al. (2024) generalizes this principle by counting, for a specified identification strategy, ordered treatment and outcome pairs for which the identifying formula derived from the estimated graph is incorrect relative to the reference graph. SID applies to DAGs, while AID covers DAGs and CPDAGs. Temporal representations satisfying the corresponding graph conditions fall within these definitions, independently of temporal dependence in the observations. Cyclic summary causal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

graphs fall outside their original definitions, whereas applications to acyclic summary or extended summary graphs require intervention semantics for temporally aggregated nodes.

SID and AID assess the consequences of graph discrepancies for causal effect identification, whereas separation-based measures compare conditional independence implications. The s/c metric of Wahl & Runge (2025) defines a symmetric distance between graphs with identical node sets by aggregating disagreements in separation and connection statements. A reference graph also permits computation of false positive and false negative rates for separation statements. Separation distances based on predefined strategies have low polynomial complexity, whereas the number of statements in the full s/c metric grows exponentially with graph size.

When causal ground truth is unavailable, the preceding measures cannot assess graph accuracy. Faller et al. (2024) introduce self compatibility as a falsification procedure that compares causal graphs learned from different subsets of variables. Detected incompatibilities indicate violated assumptions or finite sample effects, whereas passing the compatibility checks constitutes only a necessary condition and does not establish the correctness of the learned graph. The original framework assumes i.i.d. observations and does not address time series.

Restrictions on graph classes and sampling assumptions limit the applicability of these approaches to time-series causal discovery without ground truth. This paper introduces a graph-level criterion that aggregates test evidence over candidate lagged relations without comparison with a reference graph. The proposed criterion extends to time series the Causal Relevance Index introduced by Kunitomo-Jacquin et al. (2024) and draws on the edge specific p values developed for PC-p by Strobl et al. (2019). PC-p, a modified version of the PC algorithm (Spirtes & Glymour, 1991), combines the p values returned by conditional independence tests to obtain upper bounds on the p-values of edge-specific hypothesis tests and subsequently estimates and controls the false discovery rate. The criterion proposed here instead aggregates edge-level quantities over a predefined set of candidate lagged edges.

2 Related Work

Constraint-based causal discovery combines multiple conditional independence (CI) tests to infer a graph, making edge- and graph-level uncertainty non-trivial. For PC, Strobl et al. (2019) derive edge-specific p-value bounds from the Boolean structure of the corresponding compound hypotheses. For individual p-values $p_1, \ldots, p_m$, a disjunctive alternative yields the union bound

$$p_{\vee} \leq \sum_{i=1}^{m} p_i, \tag{1}$$

where $p_{\vee}$ denotes the p-value of the corresponding compound test. Conversely, when the compound alternative is a conjunction and the compound null states that at least one individual null is true, the intersection bound gives

$$p_{\wedge} \leq \max_{i=1,\ldots,m} p_i. \tag{2}$$

The latter bound is particularly important for skeleton discovery.

Under the global directed Markov property and d-separation faithfulness, absence of an adjacency corresponds to the existence of a separating set, whereas presence requires conditional dependence for every relevant conditioning set. Let $\mathcal{S}_{AB}$ denote the corresponding family of conditioning sets. Since the edge alternative requires all dependence statements to hold, the intersection bound gives

$$p^{\text{edge}}_{A-B} \leq \max_{S \in \mathcal{S}_{AB}} p_{A \perp\!\!\!\perp B|S} =: p^{\max}_{A-B}. \tag{3}$$

Thus, $p^{\max}_{A-B}$ provides a conservative upper bound on the edge-specific p-value.

For orientations, Strobl et al. (2019) combine the same union and intersection arguments with the hypotheses supporting unshielded colliders. For a v-structure $A \to C \leftarrow B$, the bound incorporates the two adjacencies together with $p_{\gamma_{AB|C}}$, the bound associated with the collider-supporting condition:

$$p^{\text{edge}}_{A \to C \leftarrow B} \leq \max \left\{ p^{\text{edge}}_{A-C}, p^{\text{edge}}_{B-C}, p_{\gamma_{AB|C}} \right\}. \tag{4}$$

Multiple possible v-structures additionally introduce union bounds, and analogous recursive constructions apply to the Meek orientation rules. Consequently, PC-p aggregates the CI evidence required to support both adjacency and orientation.

Type II errors in CI tests can remove true edges and prevent subsequent tests needed to compute the edge-specific bounds, in the PC algorithm. The initial theoretical analysis assumes a zero Type II error rate for the CI tests. To reduce the effect of violations of this assumption, Strobl et al. (2019) incorporate the order-independent skeleton search of Colombo & Maathuis (2014), together with modifications of the orientation stage that preserve conflicting information rather than allowing the result to depend on the order in which orientations are processed. PC-p then applies the Benjamini–Yekutieli procedure to the resulting edge-specific p-value bounds in order to control the false discovery rate under arbitrary dependence (Benjamini & Yekutieli, 2001).

Building on these edge-specific bounds, Kunitomo-Jacquin et al. (2024) proposed a graph-level criterion that does not require a reference causal graph. With CI significance level α and edge-presence threshold $\beta = 1 - \alpha$, their linear calibration approximates edge relevance by $1 - p^{\text{edge}}_{A-B}$. Averaging this quantity over all possible undirected edges yields the Causal Relevance Index

$$\text{CRI}(G) = 1 - \frac{2}{|V|^2 - |V|} \sum_{\{A,B\}} p^{\text{edge}}_{A-B}. \tag{5}$$

Kunitomo-Jacquin et al. (2024) used CRI to assess discretization before PC, observing agreement with mean true-positive-rate trends. Their construction assumes i.i.d. observations and valid edge-specific p-values. Extending such graph-level criteria beyond the i.i.d. setting requires valid edge-specific p-values. A p-value P satisfies super-uniformity under its null hypothesis when

$$\Pr(P \leq u \mid H_0) \leq u, \qquad \forall u \in [0, 1]. \tag{6}$$

This condition weakens exact uniformity and permits conservative p-values. Super-uniformity also appears in multiple-testing theory, where it supports error-rate control under conservative null distributions (Döhler et al., 2024). In contrast, the present work uses marginal super-uniformity as a validity condition for the edge-specific quantities entering a graph-level score. False-discovery-rate procedures such as Benjamini & Yekutieli (2001) concern thresholded collections of dependent tests and therefore depend on their joint dependence structure, whereas expectations of averages of edge-wise quantities require only their marginal distributions. This distinction motivates the extension of the CRI to temporally dependent observations.

3 The Causal Relevance Index for Time-Series Causal Discovery

3.1 Motivation and definition

For a candidate edge e, a constraint-based procedure tests a finite family $\mathcal{S}_e^{\text{test}}$ of conditioning sets and returns p-values $p_e(S)$. For this tested family, the compound null states that at least one conditional independence null is true, whereas the alternative requires all tested relations to be dependent. Let p_e^{edge} denote the p-value of this compound edge-specific hypothesis test. The intersection bound of Strobl et al. (2019) gives

$$p_e^{\text{edge}} \;\leq\; \max_{S \in \mathcal{S}_e^{\text{test}}} p_e(S) \;=:\; p_e^{\max}. \tag{7}$$

We use this computable upper bound to define, for a finite edge set $\mathcal{A}$,

$$\text{CRI}(\mathcal{A}) \;=\; 1 - \frac{1}{|\mathcal{A}|} \sum_{e \in \mathcal{A}} p_e^{\max}. \tag{8}$$

The index is therefore one minus the average conservative upper bound on the compound edge-specific p-values (Kunitomo-Jacquin et al., 2024). Its scale rests on a single fact: at a genuine separating set a calibrated test returns a *super-uniform* p-value, $\mathbb{P}(p \leq u) \leq u$, whose mean is therefore at least $1/2$. A configuration specifies the discovery algorithm, CI test and its hyperparameters, conditioning strategy, maximum lag, significance threshold, and p-value aggregation rule. Equation 8 evaluates the resulting test evidence over a candidate set $\mathcal{E}$ fixed before discovery. The score therefore depends on the configuration and its test outcomes, rather than solely on the final graph.

3.2 THE LAGGED SETTING AND ITS ASSUMPTIONS

Let $(X_t)_{t\in\mathbb{Z}}$ be $\mathbb{R}^d$-valued and fix a maximum lag $\tau_{\max}$. Candidate edges are lag-indexed and directed, $\mathcal{E} = \{(i \to j, \tau) : i \neq j,\ 1 \leq \tau \leq \tau_{\max}\}$. For $e = (i \to j, \tau)$, let $H_{0,e}$ denote absence of the corresponding lagged edge in the full-time causal graph. Under (A3), this implies the existence of a valid separating set; condition (R1) below specifies when at least one such set belongs to $\mathcal{S}_e^{\text{test}}$. We define $\text{CRI}_{\text{TS}} := \text{CRI}(\mathcal{E})$ using Eq. 8 on this fixed candidate set.

Assumption 1. **(A1)** (X_t) is strictly stationary. **(A2)** (X_t) is α-mixing with $\alpha(k) = O(k^{-\beta})$, $\beta > 2$. **(A3)** The process is causally stationary and causally sufficient, the induced distribution is d-separation faithful to the full-time graph, there are no instantaneous cycles, and the covariance matrix of every tested subvector is non-degenerate.

Under standard regularity conditions, stable ARMA models and contractive nonlinear autoregressive processes satisfy (A1) and (A2); these assumptions support the central limit argument of Section 3.4. Assumption (A3) links $H_{0,e}$ to the causal graph and ensures non-degeneracy of the covariance matrices entering the partial-correlation asymptotics. We restrict to $\tau \geq 1$ so that temporal precedence fixes orientation: the lagged problem then reduces to skeleton recovery, and Eq. 7 requires only the intersection bound, without the v-structure or Meek orientation machinery. Allowing $\tau = 0$ would reintroduce an orientation phase with its own bounds and calibration (Runge, 2020).

3.3 REDUCTION TO A MARGINAL PROPERTY

Three ingredients underlie every claim below, and only the first is probabilistic. **(R1)** *Marginal super-uniformity*: for every truly absent edge, some *tested* separating set S^* satisfies $\mathbb{P}(p_e(S^*) \leq u \mid H_{0,e}) \leq u$ on $[0,1]$. **(R2)** *Monotone max-aggregation*: $p_e^{\max} \geq p_e(S^*)$ pointwise. **(R3)** *Linearity of expectation*.

(R2) and (R3) are algebraic, so all probabilistic content sits in (R1). For the present CRI results, (R1) is weaker than the zero-Type-II-error condition used by Strobl et al. (2019): for every absent edge, it is sufficient that one valid separating set S^* is actually tested and yields a super-uniform p-value. Type II errors may therefore occur elsewhere in the discovery procedure without affecting the result, provided that (R1) holds. If needed, $\mathcal{S}_e^{\text{test}}$ can be enlarged with additional conditioning sets selected independently of the adaptive discovery path. Since $p_e^{\max}$ is a maximum, adding such sets can only increase $p_e^{\max}$ and therefore preserves the conservative max-aggregation property.

Theorem 1 (Reduction to marginal super-uniformity)**.** *Let* $\text{CRI}_{\text{TS}} = \text{CRI}(\mathcal{E})$ *and assume (R1)–(R2). Then, with no assumption on the joint law of* $(p_e^{\max})_{e\in\mathcal{E}}$*: (i)* $\text{CRI}_{\text{TS}} \in [0,1]$ *surely; (ii)* $\mathbb{E}[p_e^{\max} \mid H_{0,e}] \geq 1/2$ *and (iii)* $\mathbb{P}(p_e^{\max} \leq \alpha \mid H_{0,e}) \leq \alpha$ *for every absent edge and every* $\alpha \in [0,1]$*; (iv) if every edge is absent,* $\mathbb{E}[\text{CRI}_{\text{TS}}] \leq 1/2$*; (v) if configurations* A, B *satisfy* $\mathbb{E}[p_e^{\max}(B)] \leq \mathbb{E}[p_e^{\max}(A)]$ *for all* e*, strictly somewhere, then* $\mathbb{E}[\text{CRI}_{\text{TS}}(B)] > \mathbb{E}[\text{CRI}_{\text{TS}}(A)]$.

Sketch of proof. Under (R1), $\mathbb{E}[p] = \int_0^1 \mathbb{P}(p > u)\, du$ implies $\mathbb{E}[p_e(S^*) \mid H_{0,e}] \geq 1/2$. Since $p_e^{\max} \geq p_e(S^*)$ pointwise, (R2) establishes (ii), while $\{p_e^{\max} \leq \alpha\} \subseteq \{p_e(S^*) \leq \alpha\}$ establishes (iii). Part (i) follows directly from $0 \leq p_e^{\max} \leq 1$. Linearity of expectation then establishes (iv) and (v). Once (R1) holds, no assumption on the dependence between the resulting edge-specific quantities is required, since the argument relies only on their marginal distributions and linearity of expectation. This differs from FDR control, for which dependence between p-values must be accounted for: the Benjamini–Yekutieli procedure used by Strobl et al. (2019) controls the FDR under arbitrary dependence (Benjamini & Yekutieli, 2001). Dependence induced by adaptive test selection may nevertheless affect the marginal validity required in (R1), and is therefore not covered by this argument alone. Unlike the construction of Strobl et al. (2019), the preceding results do not require a globally zero Type II error rate. That assumption is replaced by (R1), which only requires that one valid separating set be tested for each absent edge. Conditional on this weaker coverage requirement, Type II errors elsewhere in the discovery sequence do not enter the proof. Part (v) then provides the complementary monotonicity result: if one configuration decreases the expected $p_e^{\max}$ edgewise, with a strict decrease for at least one edge, its expected CRI is larger. Further details and the proof are provided in the Supplementary Materials.

Proposition 1 (Ground-truth-free lower bound on true edges). *Assume (R1)–(R2), let $\mathcal{E}_1 \subseteq \mathcal{E}$ denote the set of present candidate edges, fix $\alpha \in [0,1)$ and let $R(\alpha) = |\{e \in \mathcal{E} : p_e^{\max} \le \alpha\}|$. Then $|\mathcal{E}_1| \ge \big(\mathbb{E}[R(\alpha)] - \alpha|\mathcal{E}|\big)/(1-\alpha)$.*

Split $\mathbb{E}[R(\alpha)]$ by the truth of the edge, bounding absent edges by (iii) and present ones by one. The resulting inequality bounds the total number of true edges in the candidate set through the expected rejection count; it does not directly certify how many selected links correspond to true edges in a single observed graph. Both $R(\alpha)$ and $\mathrm{CRI}_{\mathrm{TS}}$ summarize the same edge-specific quantities, through thresholding and averaging, respectively. Throughout this work, CRI uses the fixed candidate set $\mathcal{E}$, including candidates omitted from the inferred graph. Consequently, retaining only a few links with small p-values cannot increase the score by excluding the remaining candidates from the average. No claim requires the true graph to maximize the index.

3.4 Super-uniformity under temporal dependence

Only (R1) remains, and temporal dependence affects the calibration of Fisher's z-test. Under i.i.d. Gaussian sampling, the Fisher calibration provides asymptotically uniform null p-values, whereas serial dependence modifies the limiting variance. Let ξ_t denote the first-order influence process of the partial correlation, $\gamma_\xi(k) = \mathrm{Cov}(\xi_t, \xi_{t+k})$ its autocovariance, and $\sigma^2_{\mathrm{LR}} = \sum_{k\in\mathbb{Z}} \gamma_\xi(k) = 2\pi f_\xi(0)$. Let n_{eff} denote the sample-size factor used in the Fisher calibration and define $\kappa = \lim_n (n_{\mathrm{eff}}/n)\sigma^2_{\mathrm{LR}}$, the ratio to the reference variance the Fisher transform assumes.

Theorem 2 (Operational super-uniformity). *Under Assumption 1 and the nonparanormal model, fix a finite, nonempty candidate set $\mathcal{E}$. For each absent edge e, assume that a fixed finite separating set S_e^* belongs to the tested family almost surely for every sample size, with $\rho_{e|S_e^*} = 0$. Assume positive definiteness of the covariance matrix of the population-transformed tested subvector and $\sqrt{n}$-equivalence of the estimated partial correlation to its oracle counterpart.*

Suppose that either (i) the Fisher calibration uses $n_{\mathrm{eff},e}/n \to c_e \in [0,1]$ and $\sum_{k\ge1} \gamma_{\xi_e}(k) \le 0$; or (ii) the statistic $\sqrt{n}\,\mathrm{arctanh}(\widehat{\rho}_{e,n})$ is divided by $\sqrt{\widehat{v}_{e,n}}$, where $\widehat{v}_{e,n} > 0$ and $\widehat{v}_{e,n} \xrightarrow{\mathbb{P}} v_e$ for a finite constant $v_e > 0$ satisfying $v_e \ge \sigma^2_{\mathrm{LR},e}$. Then, for every $u \in (0,1)$,

$$\limsup_{n\to\infty} \mathbb{P}\big(p_{e,n}(S_e^*) \le u \mid H_{0,e}\big) \le u. \tag{9}$$

Under max-aggregation, conclusions (ii)–(iv) of Theorem 1 hold asymptotically:

$$\begin{aligned} \liminf_{n\to\infty} \mathbb{E}[p_{e,n}^{\max} \mid H_{0,e}] &\ge \tfrac{1}{2}, \\ \limsup_{n\to\infty} \mathbb{P}(p_{e,n}^{\max} \le \alpha \mid H_{0,e}) &\le \alpha, \\ \limsup_{n\to\infty} \mathbb{E}_0[\mathrm{CRI}_{\mathrm{TS},n}] &\le \tfrac{1}{2}. \end{aligned} \tag{10}$$

The second inequality holds for every fixed $\alpha \in (0,1)$, and the third assumes the global null. These conclusions do not establish finite-sample condition (R1).

Following the marginal Gaussianization approach of Liu et al. (2009), we estimate partial correlations after transforming the marginal distributions. As specified in the Supplementary Materials, we additionally assume $\sqrt{n}$-equivalence between $\widehat{\rho}_{e,n}$ and the partial-correlation estimator computed using the population marginal transformations under temporal dependence. Under this assumption, the mixing central limit theorem (Ibragimov, 1962; 1975) and the delta method give, under $H_{0,e}$, $\sqrt{n}\,\mathrm{arctanh}(\widehat{\rho}_{e,n}) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\mathrm{LR},e})$, since $\rho_{e|S_e^*} = 0$ and $\mathrm{arctanh}'(0) = 1$. By Slutsky's lemma, the calibrated Fisher statistic $Z_{e,n}$ converges in distribution to $\mathcal{N}(0, \kappa_e)$, with $\kappa_e = c_e\sigma^2_{\mathrm{LR},e}$ in case (i) and $\kappa_e = \sigma^2_{\mathrm{LR},e}/v_e$ in case (ii).

Under the jointly Gaussian transformed model and the null, $\gamma_{\xi_e}(0) = 1$. Thus, in case (i), $\sigma^2_{\mathrm{LR},e} = 1 + 2\sum_{k\ge1}\gamma_{\xi_e}(k) \le 1$ and $c_e \in [0,1]$ imply $\kappa_e \le 1$. In case (ii), the inequality follows from $v_e \ge \sigma^2_{\mathrm{LR},e}$. For $\kappa_e > 0$ and $u \in (0,1)$,

$$\mathbb{P}\big(p_{e,n}(S_e^*) \le u \mid H_{0,e}\big) \longrightarrow 2\left\{1 - \Phi\left(\frac{\Phi^{-1}(1-u/2)}{\sqrt{\kappa_e}}\right)\right\} \le u. \tag{11}$$

For $\kappa_e = 0$, $p_{e,n}(S_e^*)$ converges in probability to one, which also gives asymptotic super-uniformity. The inequality $p_{e,n}^{\max} \geq p_{e,n}(S_e^*)$ transfers the rejection bound to the maximum. The tail-integral identity and Fatou's lemma yield the expectation bounds in Eq. 10. If $\gamma_{\xi_e}(k) = 0$ for every $k \geq 1$, then $\sigma^2_{\mathrm{LR},e} = 1$; in case (i), $p_{e,n}(S_e^*)$ converges to the uniform distribution if and only if $c_e = 1$. This uniform limit need not hold for $p_{e,n}^{\max}$.

Within the Gaussian limit considered above, a variance exceeding the reference variance causes asymptotic over-rejection. This calibration issue relates to the longstanding problem of spurious associations between autocorrelated time series (Yule, 1926). Stationarity alone does not ensure super-uniformity. Conditioning on lagged parents, as in the MCI strategy (Runge et al., 2019; Runge, 2020), motivates case (i); however, applying Theorem 2 additionally requires its stated covariance and separating-set coverage conditions. Adding conditioning sets independently of the adaptive discovery path can ensure coverage if the added family contains a fixed valid separator; independence from the path alone does not guarantee this inclusion. Theorem 1 and Proposition 1 retain their finite-sample guarantees under (R1)–(R2), whereas Theorem 2 establishes asymptotic counterparts of Theorem 1(ii)–(iv). Dependence between the aggregated edge-specific quantities does not affect linearity of expectation, but uncertainty quantification requires additional analysis. In particular, replacing $\mathbb{E}[R(\alpha)]$ by the observed $R(\alpha)$ requires a separate probabilistic bound.

3.5 Null-normalized index

Under exact condition (R1) and max-aggregation, Theorem 1 gives $\mathbb{E}_0[\mathrm{CRI}_{\mathrm{TS}}] \leq 1/2$ under the global null. Under the assumptions of Theorem 2, this bound holds only asymptotically. The finite-sample null level can depend on the test, conditioning strategy, sample size, and temporal dependence. We therefore estimate a configuration-specific null reference and report the affine rescaling

$$\widehat{\mathrm{CRI}}_{\mathrm{norm}} = \frac{\mathrm{CRI}(\mathcal{E}) - \widehat{\mu}_0}{1 - \widehat{\mu}_0}, \qquad \widehat{\mu}_0 = \frac{1}{B}\sum_{b=1}^{B} \mathrm{CRI}_0^{(b)}, \tag{12}$$

where each $\mathrm{CRI}_0^{(b)}$ is computed on the same candidate edge set $\mathcal{E}$ by applying the complete discovery procedure to a null replicate generated by per-variable iterative amplitude-adjusted Fourier transform (IAAFT) phase randomization. This preserves each variable's marginal distribution and approximately preserves its spectrum while disrupting cross-variable dependence. Negative normalized values are retained, since they indicate statistical support below the empirical null reference.

4 Experiments

4.1 Experimental setting

Synthetic experiments compare CRI and normalized CRI with reference-based metrics across graph structures, sample sizes, discovery algorithms, and CI tests for time series. The Causal Chambers wind-tunnel dataset provides a complementary physical benchmark comprising 1016 observations of 16 variables, including measurements and actuator signals under continuously varying operating conditions (Gamella et al., 2025). Its experimentally validated reference graph derives from physical design and experimental knowledge but specifies no exact lags. Comparison therefore uses summary graphs obtained by collapsing inferred lagged relations; lag-specific evaluation covers only the synthetic experiments.

We first consider PCMat-HSIC, a constraint-based causal discovery procedure for lagged multivariate time series. The procedure follows a PC-style iterative skeleton search over candidate relations $X^i_{t-\tau} \to X^j_t$, with $\tau \geq 1$. At conditioning order zero, marginal dependence is assessed using the Hilbert–Schmidt independence criterion (HSIC) (Gretton et al., 2007; Chwialkowski & Gretton, 2014). For non-empty conditioning sets, both tested variables are first regressed on the conditioning variables by kernel ridge regression, and HSIC is then applied to the residuals. This residual-based construction follows the same general kernel strategy used to remove dependence on conditioning variables before testing the remaining association (Fukumizu et al., 2007; Zhang et al., 2012), while preserving the iterative conditioning structure used to construct the edge-specific quantities entering the CRI. PCMat-HSIC also includes an optional MCI pruning stage (Runge et al., 2019; Runge,

2020), in which relations retained by the initial PC-style search are retested conditionally on selected target and time-shifted source parents, allowing some retained edges to be removed without introducing new ones.

In the synthetic experiments, PCMat-HSIC uses a Matérn-$1/2$ kernel. Kernel ridge regression and HSIC are approximated using random Fourier features with $D = 100$ components in order to reduce the computational cost of repeated conditional independence testing. Statistical significance uses circular shifts inspired by the Shift HSIC procedure of Chwialkowski & Gretton (2014). The maximum conditioning-set size is two and the significance level is $\alpha = 0.05$. For each candidate edge, all p-values generated during the skeleton search are retained, allowing the edge-specific quantity $p_e^{\max}$ to be computed according to Eq. 7.

The comparison additionally includes PCMCI and PCMCI+ (Runge et al., 2019; Runge, 2020), combined with two conditional independence tests. Partial correlation provides a linear Gaussian baseline, whereas Gaussian-process distance correlation (GPDC) allows nonlinear additive dependencies. This gives four additional configurations: PCMCI-ParCorr, PCMCI-GPDC, PCMCI+-ParCorr, and PCMCI+-GPDC.

We consider seven directed lagged graph structures: a chain, a fork, a collider, and four increasingly complex structures containing between four and seven variables. Cross-variable causal effects are assigned to lags $\tau \in \{1, 2\}$. For each variable X_j, the data-generating process is

$$X_j(t) = a_j X_j(t-1) + \sum_{(i,j,\tau)\in G} c_{ij,\tau} f_{ij,\tau}(X_i(t-\tau)) + \varepsilon_j(t), \tag{13}$$

where $a_j \in [0.2, 0.6]$, $c_{ij,\tau} \in [0.3, 0.8]$, and the nonlinear mechanisms are selected from hyperbolic tangent, sine, bounded quadratic, radial, and sigmoidal functions. The innovations are independent Gaussian variables with unit standard deviation.

For each structure, ten independent realizations are generated for $T \in \{100, 200, 500, 1000\}$ where T denotes the series length, whereas n denotes the number of observations used in a test after lag alignment. Within each repetition, the causal structure and simulation seed remain fixed across sample sizes. All discovery configurations use the same realizations, enabling paired comparisons.

Null normalization uses per-variable IAAFT surrogates as described in Section 3.5. For synthetic data, surrogate runs across repetitions estimate the null reference separately for each algorithm, graph structure, and sample size. For the wind-tunnel dataset, B independently generated surrogates of the observed series provide the reference for Eq. 12.

4.2 Evaluation criteria

The evaluation is performed at two temporal resolutions: window-graph metrics retain the complete lagged relation (i, j, τ), whereas summary-graph metrics collapse relations over nonzero lags. For metrics requiring a DAG, SID and AID are additionally computed on a finite time-unrolled representation; their summary-graph versions are reported only when the collapsed graph is acyclic.

Graph recovery is assessed using precision, recall, F_1, and the Matthews correlation coefficient (MCC) (Raghu et al., 2018; Moraffah et al., 2021; Cheng et al., 2022); structural discrepancies using SHD (Tsamardinos et al., 2006; Rehak et al., 2024); causal-effect identification using SID and AID (Peters & Bühlmann, 2015; Henckel et al., 2024); and separation properties using the s/c metric (Wahl & Runge, 2025). Normalized versions (nSHD, nSID, and nAID) are also reported. These criteria require a reference graph, whereas CRI and normalized CRI do not.

4.3 Results

Table 1 reports summary-graph results averaged over structures and repetitions. Across configurations, mean recall increases from 0.768 at $T = 100$ to 0.955 at $T = 1000$, alongside improvements in F_1, MCC, and discrepancy-based metrics. Mean normalized CRI increases from 0.137 to 0.174. Window-graph results exhibit similar trends.

For PCMat-HSIC, normalized CRI shows a strong monotonic association with summary-graph recovery across the 28 graph-structure/sample-size conditions after averaging simulation repetitions,

Table 1: Mean summary-graph evaluation results by sample size and causal configuration. Values are averaged over graph structures and 10 independent simulations per structure and sample size. nSID, nAID, and s/c are averaged over runs for which the collapsed summary graph is acyclic.

T	Algorithm	Precision	Recall	F_1	MCC	nSHD	nSID	nAID	s/c	$\widehat{\mathrm{CRI}}_{\mathrm{norm}}$
100	PCMCI+-GPDC	0.911	0.773	**0.817**	**0.772**	0.191	0.212	0.152	0.209	0.139
	PCMCI+-ParCorr	0.920	0.757	0.813	0.770	**0.185**	0.207	0.148	0.220	0.144
	PCMCI-GPDC	0.772	**0.814**	0.771	0.684	0.284	**0.139**	**0.106**	0.154	0.114
	PCMCI-ParCorr	0.791	**0.814**	0.788	0.711	0.248	0.148	0.120	**0.139**	0.084
	PCMat-HSIC	**0.925**	0.684	0.764	0.723	0.227	0.264	0.176	0.298	**0.204**
200	PCMCI+-GPDC	0.957	0.902	**0.921**	**0.899**	**0.086**	0.084	**0.057**	0.077	0.155
	PCMCI+-ParCorr	0.939	0.830	0.869	0.836	0.136	0.158	0.110	0.143	0.158
	PCMCI-GPDC	0.831	**0.917**	0.861	0.807	0.180	**0.079**	0.061	**0.063**	0.129
	PCMCI-ParCorr	0.829	0.858	0.830	0.769	0.198	0.146	0.117	0.119	0.119
	PCMat-HSIC	**0.964**	0.882	0.911	0.889	0.091	0.105	0.073	0.096	**0.236**
500	PCMCI+-GPDC	0.972	0.967	0.966	0.954	0.043	0.031	0.021	0.030	0.165
	PCMCI+-ParCorr	0.954	0.888	0.914	0.887	0.096	0.098	0.072	0.080	0.142
	PCMCI-GPDC	0.844	0.970	0.895	0.853	0.147	0.020	0.012	0.034	0.156
	PCMCI-ParCorr	0.818	0.901	0.848	0.786	0.199	0.078	0.056	0.073	0.124
	PCMat-HSIC	**0.976**	**0.980**	**0.975**	**0.966**	**0.033**	**0.015**	**0.009**	**0.019**	**0.278**
1000	PCMCI+-GPDC	0.949	0.989	0.965	0.951	0.049	0.009	0.006	0.012	0.185
	PCMCI+-ParCorr	0.929	0.888	0.902	0.868	0.113	0.101	0.070	0.080	0.149
	PCMCI-GPDC	0.825	0.992	0.894	0.853	0.153	**0.000**	**0.000**	0.023	0.153
	PCMCI-ParCorr	0.826	0.908	0.857	0.799	0.185	0.064	0.048	0.070	0.106
	PCMat-HSIC	**0.977**	**0.998**	**0.986**	**0.980**	**0.020**	0.001	0.001	**0.003**	**0.277**

Table 2: Causal Chambers wind-tunnel results. Lagged relations are collapsed into summary graphs for comparison with the physical reference graph. $\widehat{\mathcal{E}}$ denotes the edge set of the estimated graph.

Method	$\lvert\widehat{E}\rvert$	TP	FP	FN	Precision	Recall	F_1	MCC	nSHD	CRI	$\widehat{\mathrm{CRI}}_{\mathrm{norm}}$
PCMCI-ParCorr	40	11	29	15	0.275	**0.423**	0.333	0.240	0.358	0.504	0.002
PCMCI-GPDC	37	9	28	17	0.243	0.346	0.286	0.185	0.367	0.519	0.002
PCMCI+-ParCorr	11	6	5	20	0.545	0.231	**0.324**	0.308	0.208	0.533	-0.107
PCMCI+-GPDC	12	6	6	20	0.500	0.231	0.316	0.289	0.217	0.682	-0.017
PCMat-HSIC	15	6	9	20	0.400	0.231	0.293	0.242	0.242	0.478	0.121
PCMat-HSIC-MCI (pruning)	7	5	2	21	**0.714**	0.192	0.303	**0.338**	0.192	0.474	**0.123**

with Spearman correlations of $\rho = 0.675$ with recall and $\rho = 0.653$ with F_1. Its continued variation after recall saturates reflects the additional information carried by the edge-specific p-value bounds, which quantify the strength of statistical evidence beyond the binary recovery of true edges.

At the algorithm level, PCMat-HSIC obtains the largest normalized CRI for all sample sizes, increasing from 0.204 at $T = 100$ to approximately 0.277 at $T = 1000$. At $T = 500$ and $T = 1000$, it also obtains the best or tied-best summary-graph performance for most recovery criteria, reaching $F_1 = 0.986$ and MCC$= 0.980$ at $T = 1000$. At smaller sample sizes, random Fourier feature approximation (D = 100) in kernel ridge regression and HSIC may introduce additional variability into conditional independence testing. Accordingly, the ordering induced by normalized CRI does not systematically coincide with that of every reference-based metric in this regime. For example, PCMCI+-GPDC achieves the highest F_1 and MCC at $T = 100$, whereas PCMCI-GPDC yields the lowest nSID and nAID for several sample sizes. These differences reflect both finite-sample estimation effects and the distinct graph properties captured by the reference-based criteria.

The Causal Chambers wind-tunnel dataset provides a complementary benchmark based on a controlled physical system. In the selected subgraph, the variables `load_in`, `load_out`, `hatch`, `pot_1`, and `pot_2` correspond to manipulable variables, whereas the remaining nodes correspond to sensor measurements (Gamella et al., 2025). The reference relations therefore describe known effects originating from controlled inputs and propagating toward physical measurements. As detailed in Table 2, PCMCI-ParCorr recovers 11 such relations but introduces 29 additional relations relative to the reference graph, including relations among sensor measurements and toward manipulable variables. PCMat-HSIC with MCI pruning produces a substantially sparser summary graph, with 5 true positives and only 2 false positives (precision 0.714, recall 0.192, nSHD$= 0.192$).

The five recovered reference relations with MCI pruning connect `load_in` to `current_in` and `rpm_in`, `load_out` to `current_out` and `rpm_out`, and `pot_1` to

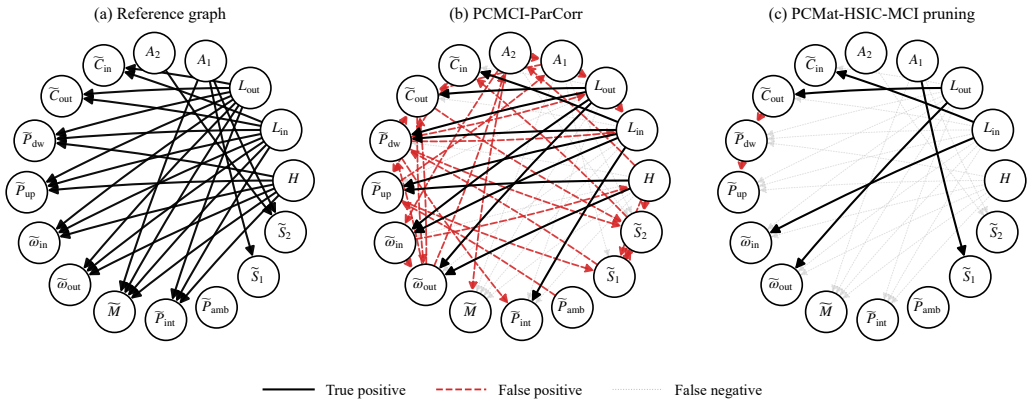


Figure 1: Causal Chambers wind-tunnel summary graphs: (a) physical reference graph, (b) PCMCI-ParCorr estimate, and (c) PCMat-HSIC-MCI pruning estimate. Solid black, dashed red, and dotted gray edges denote true positives, false positives, and false negatives, respectively. Variable notation follows Gamella et al. (2025): H denotes the hatch position; L_{in} and L_{out} the fan loads; A_1 and A_2 the potentiometers; $\widetilde{C}_{\text{in}}$ and $\widetilde{C}_{\text{out}}$ the fan currents; $\widetilde{\omega}_{\text{in}}$ and $\widetilde{\omega}_{\text{out}}$ the fan speeds; $\widetilde{P}_{\text{dw}}$, $\widetilde{P}_{\text{up}}$, $\widetilde{P}_{\text{int}}$, and $\widetilde{P}_{\text{amb}}$ the pressure measurements; $\widetilde{M}$ the microphone output; and $\widetilde{S}_1$, $\widetilde{S}_2$ the speaker-circuit signals.

`signal_1`. The two additional relations, `current_out`→`pressure_downwind` and `pressure_downwind`→`pressure_upwind`, connect sensor measurements and do not introduce incoming relations toward the manipulable variables. Figure 1 illustrates these structural differences relative to PCMCI-ParCorr and the reference graph.

For PCMat-HSIC, MCI pruning removes seven false positives and one true positive, increasing precision from 0.400 to 0.714 and decreasing recall from 0.231 to 0.192 (Table 2). The retained reference relations describe direct effects from manipulable variables to associated measurements, while a substantial fraction of pressure, acoustic, and downstream relations remains absent. Normalized CRI reaches its highest point estimate for PCMat-HSIC with MCI pruning (0.123 versus 0.121 without pruning). This criterion summarizes statistical support for tested conditional dependencies relative to a configuration-specific null reference. Here, its highest value coincides with the highest precision and MCC and the lowest nSHD, despite incomplete reference-graph recovery.

5 Conclusion

The extension of the Causal Relevance Index to lagged causal discovery aggregates edge-specific p-value bounds over a fixed candidate set. The theoretical guarantees follow from marginal super-uniformity at a tested separating set for each absent edge, without assumptions on the joint dependence among the aggregated quantities. Normalization using IAAFT surrogates accounts for configuration-specific null levels. In the synthetic experiments, normalized CRI tracks improvements in edge recovery for PCMat-HSIC, which combines a PC-style search with kernel-based independence tests. On Causal Chambers, PCMat-HSIC with MCI pruning achieves the highest normalized CRI, precision, and MCC, together with the lowest structural discrepancy, although recall remains limited.

These results support CRI as a criterion for assessing statistical support for conditional dependencies without a reference graph, while retaining continuous information beyond thresholded decisions. Under the stated calibration and separating-set conditions, the index provides a basis for comparing discovery configurations through their statistical support relative to a configuration-specific null reference. This perspective complements structural and interventional criteria by enabling assessment when causal ground truth remains unavailable. Further work will quantify uncertainty in the index and its estimated null reference under temporal dependence, to assess whether differences between discovery configurations exceed sampling variability.

AI USE STATEMENT

Generative AI tools were used for English grammar checking and occasional rephrasing to improve readability. The authors have reviewed and approved all suggested revisions and take full responsibility for the final text.

ETHICS STATEMENT

This work uses no human subjects and no personal data.

REFERENCES

Charles Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

Charles K. Assaad, Emilie Devijver, Eric Gaussier, Gregor Goessler, and Anouar Meynaoui. Identifiability of total effects from abstractions of time series causal graphs. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244, pp. 173–185. PMLR, 2024.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

Lu Cheng, Ruocheng Guo, Raha Moraffah, Paras Sheth, K Selçuk Candan, and Huan Liu. Evaluation methods and measures for causal learning algorithms. *IEEE Transactions on Artificial Intelligence*, 3(6):924–943, 2022.

Kacper Chwialkowski and Arthur Gretton. A kernel independence test for random processes. In *International Conference on Machine Learning*, pp. 1422–1430. PMLR, 2014.

Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

Sebastian Döhler, Iqraa Meah, and Etienne Roquain. Online multiple testing with super-uniformity reward. *Electronic Journal of Statistics*, 18(1):1293–1354, 2024.

Philipp M. Faller, Leena C. Vankadara, Atalanti A. Mastakouri, Francesco Locatello, and Dominik Janzing. Self-compatibility: Evaluating causal discovery without ground truth. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4132–4140. PMLR, 2024.

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

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.

Chang Gong, Chuzhe Zhang, Di Yao, Jingping Bi, Wenbin Li, and Yongjun Xu. Causal discovery from temporal data: An overview and new perspectives. *ACM Computing Surveys*, 57(4):1–38, 2024.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

Leonard Henckel, Theo Würtzen, and Sebastian Weichwald. Adjustment identification distance: A gadjid for causal structure learning. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 1569–1598. PMLR, 2024.

I Ibragimov. Independent and stationary sequences of random variables. *Wolters, Noordhoff Pub.*, 1975.

Ildar A Ibragimov. Some limit theorems for stationary processes. *Theory of Probability & Its Applications*, 7(4):349–382, 1962.

Lucie Kunitomo-Jacquin, Aurore Lomet, and Geoffrey Daniel. Assessing causal graph uncertainty and optimized multivariate discretization strategy. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2024)*, volume 1174 of *Lecture Notes in Networks and Systems*, pp. 257–269, Cham, 2024. Springer. doi: 10.1007/978-3-031-74003-9_21.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: semiparametric estimation of high-dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009.

Raha Moraffah, Paras Sheth, Mansooreh Karami, Anchit Bhattacharya, Qianru Wang, Anique Tahir, Adrienne Raglin, and Huan Liu. Causal inference for time series analysis: Problems, methods and evaluation. *Knowledge and Information Systems*, 63(12):3041–3085, 2021.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural Computation*, 27(3):771–799, 2015.

Vineet K. Raghu, Allen Poon, and Panayiotis V. Benos. Evaluation of causal structure learning methods on mixed data types. In Thuc Duy Le, Kun Zhang, Emre Kıcıman, Aapo Hyvärinen, and Lin Liu (eds.), *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, volume 92 of *Proceedings of Machine Learning Research*, pp. 48–65. PMLR, 20 Aug 2018.

Josephine Rehak, Alexander Falkenstein, Frank Doehner, and Jürgen Beyerer. Metrics for the evaluation of learned causal graphs based on ground truth. In *ML4CPS – Machine Learning for Cyber-Physical Systems*. UB HSU, 2024. doi: 10.24405/15305.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1388–1397. PMLR, 2020.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11):eaau4996, 2019.

Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

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*, 10(5):1–37, 2019.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

Jonas Wahl and Jakob Runge. Separation-based distance measures for causal graphs. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3412–3420. PMLR, 2025.

G Udny Yule. Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society*, 89(1):1–63, 1926.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.