PEC: IMPROVING POSTERIOR EDGE CONFIDENCE FOR CONSTRAINT-BASED CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery has attracted growing attention across scientific fields, as most causal downstream analyses rely on the recovered causal graph. In practice, however, finite samples, incomplete data, and violated assumptions frequently yield erroneous edges, and this problem is aggravated in constraint-based algorithms, where errors in conditional independence (CI) tests can propagate through the subsequent decisions. In response, several methods also provide edge-wise scores that indicate how reliable each edge is, which helps verify the recovered causal graph. Yet these scores typically quantify the strength of causal influence or provide conservative bounds for statistical error control over the set of discovered edges, rather than the posterior probability of edge existence. We introduce Posterior Edge Confidence (PEC), a framework that estimates posterior edge probabilities for constraint-based causal discovery. PEC models the distributions of the test statistic under both hypotheses and combines them by probability rules derived from the logical structure of the skeleton search, rather than taking a single representative statistic. In experiments on synthetic datasets, we show PEC ranks edges more accurately than an existing representative-statistic baseline and improves the baseline as well.

1 INTRODUCTION

Inferring causal relations rather than associations is central to the sciences, since identifying interventional effects from associations alone requires randomized experiments (Pearl, 2009; Peters et al., 2017), which are often infeasible for practical or ethical reasons. Causal discovery, the task of learning causal structure from observational data, has therefore attracted growing attention across scientific fields (Spirtes et al., 2000; Glymour et al., 2019). It produces a causal graph over the observed variables, which downstream analyses, such as causal effect estimation (Maathuis et al., 2009), root cause analysis (Budhathoki et al., 2022), and counterfactual reasoning (Shpitser & Pearl, 2008), typically require this graph, so their conclusions are only as reliable as it is.

In practice, however, observed data are finite and often too limited to determine the causal structure reliably (Uhler et al., 2013). Moreover, data may contain missing values, selection bias, or influences from unobserved confounders, which violate the structural assumptions that discovery methods rely on and thus yield an erroneous causal graph (Spirtes et al., 1995; Colombo et al., 2012; Tu et al., 2019). It is aggravated in constraint-based algorithms: a single incorrect conditional independence (CI) test outcome affects not only the decision about the corresponding edge but also the separating sets available to subsequent steps, propagating through collider detection and orientation rules, so the recovered graph may be much different from the true one (Colombo & Maathuis, 2014).

These problems have motivated a line of work that returns confidence scores along with the recovered causal graph. While several approaches yield a graph-wise confidence score (Strieder et al., 2021; Wang et al., 2023), others assign scores to individual edges, providing more information for verification.

One line of work measures the stability of each edge under perturbation of the data, running the discovery algorithm on bootstrap or subsampled replicates and reporting how often the edge appears (Friedman et al., 1999; Stekhoven et al., 2012). These procedures are algorithm-agnostic and simple, but the resulting frequencies indicate how often the algorithm recovers an edge rather than how probable the edge is, and obtaining them requires many repeated runs of the discovery algorithm.

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.

A second line of work quantifies the strength of an edge rather than the certainty of its existence. Janzing et al. (2013) define the strength of an edge as the distributional difference induced by cutting that edge. Such quantities describe how much influence flows along an edge, not how certain we are that the edge is there; a weak but well-identified edge is indistinguishable from a strong but poorly supported one.

A third line of work derives edge scores from the statistical significance of the tests performed during discovery, providing a guarantee on the error rate of the recovered structure. Tsamardinos & Brown (2008) assign to each adjacency the maximum p-value over the associated CI tests and apply an FDR-controlling procedure, later refined for skeleton identification by Armen & Tsamardinos (2014); Strobl et al. (2019) extend this construction to the entire PC algorithm, propagating the bounds through the orientation phase as well. These methods provide valid p-value bounds for individual edges, but the bounds are conservative by construction and are not the probability that an edge exists.

Bayesian structure learning instead infers a posterior distribution over graphs, yielding edge marginal probabilities. Koivisto & Sood (2004) compute these probabilities exactly for small problems by dynamic programming, Friedman & Koller (2003) and Kuipers & Moffa (2017) approximate them by MCMC over orderings and partitions, and Lorch et al. (2021); Deleu et al. (2022); Lorch et al. (2022) replace sampling with variational or amortized inference. However, these methods require a parametric score over graphs and do not apply to the constraint-based setting.

Closest to our aim, a few methods derive probabilities within the constraint-based framework. Claassen & Heskes (2012) yield scores for logical causal statements by summing Bayesian scores over structures on small variable subsets, using the resulting reliabilities to order decisions and resolve conflicts. The reliabilities therefore require a parametric score and a prior over structures, which the constraint-based setting does not otherwise assume. Triantafillou et al. (2014) yield posterior probabilities of edges through modeling the alternative distribution with a beta family but takes the maximum p-value of as a representative for each CI test so the outcome of remaining tests are discarded.

We introduce Posterior Edge Confidence (PEC), a framework that estimates posterior probabilities of edge existence for constraint-based causal discovery. PEC estimates the distribution of the test statistic under the alternative from the statistics observed across the CI tests, and applies probability rules to derive better-calibrated posterior probability of edge presence.

Our contributions are as follows.

1. We formulate the confidence score of an edge existence as its posterior probability for constraint-based causal discovery. Since the posterior is a probability given the observed data, it can be used directly to verify the recovered graph before downstream use.
2. We propose PEC framework, which estimates distribution under the alternative from the observed statistics rather than fixed in advance, and apply probability rules to yield a better-calibrated posterior probability. We also show the effectiveness of the framework by applying the rules to an existing baseline as well.
3. In the experiments on synthetic datasets, we show the improvement of calibration performance of PEC through ranking-based metrics compared to the baseline method PROPeR Triantafillou et al. (2014), and even improves the baseline by applying the probability rules to the method.

2 PRELIMINARIES

2.1 CAUSAL GRAPHS AND NOTATION

Let $\mathbf{X} = \{X_1, \ldots, X_p\}$ be a set of random variables. A causal graph $\mathcal{G} = (\mathbf{X}, \mathcal{E})$ is defined as a directed acyclic graph whose nodes correspond to the variables and in which an edge $X_i \to X_j \in \mathcal{E}$, indicates that X_i is a direct cause of X_j, abbreviated as (i, j).

If $X_i \to X_j$, then X_i is a *parent* of X_j and X_j is a *child* of X_i, denoted by $X_i \in \mathrm{pa}(j)$ and $X_j \in \mathrm{ch}(i)$. Similarly, if $X_i \to \cdots \to X_j$, then X_j is a *descendant* of X_i and X_i an *ancestor* of X_j, denoted by $X_i \in \mathrm{an}(j)$ and $X_j \in \mathrm{de}(i)$. If either $X_i \to X_j$ or $X_j \to X_i$, we say X_i and X_j

are *adjacent*, and write $X_i \in \mathrm{adj}(j)$, equivalently $X_j \in \mathrm{adj}(i)$. The *skeleton* of $\mathcal{G}$ is the undirected graph obtained retaining only the existence of each edge.

A *path* between X_i and X_j is a sequence of distinct variables in which consecutive variables are adjacent. A variable X_k on a path is a *collider* if both incident edges on the path point into it and *unshielded* if in addition its two neighbors on the path are not adjacent to each other.

For a set $\mathbf{S} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$, possibly empty, a path is *blocked* by $\mathbf{S}$ if it contains a non-collider in $\mathbf{S}$, or a collider X_k with $X_k \notin \mathbf{S}$ and $\mathrm{de}(k) \cap \mathbf{S} = \emptyset$. The pair X_i, X_j is *d-separated* by $\mathbf{S}$, written $X_i \perp\!\!\!\perp_{\mathcal{G}} X_j \mid \mathbf{S}$, if every path between them is blocked by $\mathbf{S}$ (Pearl, 2009).

A distribution P over $\mathbf{X}$ is *Markov* with respect to $\mathcal{G}$ if $X_i \perp\!\!\!\perp_{\mathcal{G}} X_j \mid \mathbf{S}$ implies $X_i \perp\!\!\!\perp X_j \mid \mathbf{S}$ in P, and *faithful* to $\mathcal{G}$ if the converse also holds, so that under both conditions d-separation and conditional independence coincide. If every common cause of any two variables $X_i, X_j \in \mathbf{X}$ is itself in $\mathbf{X}$, the variable set is said to be *causally sufficient* (Spirtes et al., 2000). Two causal graphs are *Markov equivalent* if they entail the same conditional independencies; such graphs share the same skeleton and the same set of unshielded colliders. Their equivalence class is called a Markov equivalence class (MEC), represented by a completed partially directed acyclic graph (CPDAG) where an edge is directed if it has that orientation in every member of the class and undirected otherwise (Andersson et al., 1997).

2.2 Conditional independence testing for causal discovery

Under faithfulness and causal sufficiency, d-separation in $\mathcal{G}$ and conditional independence in P coincide, so the structure of $\mathcal{G}$ can be recovered by testing statements of the form $X_i \perp\!\!\!\perp X_j \mid \mathbf{S}$, where $\mathbf{S} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$ is a conditioning set. Constraint-based methods take this approach to identify $\mathcal{G}$ up to its MEC.

Depending on the underlying (or assumed) causal mechanism and the type of the data, the CI test can be chosen differently, but it follows the same hypothesis testing procedure, that is

$$H_0^{ij|\mathbf{S}} : X_i \perp\!\!\!\perp X_j \mid \mathbf{S} \qquad \text{against} \qquad H_1^{ij|\mathbf{S}} : X_i \not\perp\!\!\!\perp X_j \mid \mathbf{S}, \tag{1}$$

where independence is taken as the null since it specifies the distribution of the test statistic while dependence is composite.

One of the most widely considered models is a linear Gaussian model, which follows

$$X_j = \sum_{X_i \in \mathrm{pa}(j)} b_{ji} X_i + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2), \tag{2}$$

where the exogenous variables ε_j are mutually independent, so $\mathbf{X}$ is jointly Gaussian.

Under equation 2, $X_i \perp\!\!\!\perp X_j \mid \mathbf{S}$ holds if and only if the partial correlation $\rho_{ij|\mathbf{S}}$ is zero, so the CI test takes the sample partial correlation $\hat{\rho}_{ij|\mathbf{S}}$. Given n i.i.d. observations, Fisher's z-transform (Fisher, 1915) of $\hat{\rho}_{ij|\mathbf{S}}$, that is $Z_{ij|\mathbf{S}} = \sqrt{n-q-3}\,\operatorname{artanh}(\hat{\rho}_{ij|\mathbf{S}})$, asymptotically satisfies

$$Z_{ij|\mathbf{S}} \overset{a}{\sim} \mathcal{N}\Big(\sqrt{n-q-3}\,\operatorname{artanh}\big(\rho_{ij|\mathbf{S}}\big),\ 1\Big) \tag{3}$$

for arbitrary value of $\rho_{ij|\mathbf{S}}$, where $q = |\mathbf{S}|$. Since the mean becomes zero exactly when $X_i \perp\!\!\!\perp X_j \mid \mathbf{S}$, equation 1 can be tested by comparing the two-sided p-value $p_{ij|\mathbf{S}} = 2\big(1 - \Phi(|Z_{ij|\mathbf{S}}|)\big)$ against a threshold α, where Φ is the standard normal distribution function.

2.3 The PC algorithm

The PC algorithm (Spirtes et al., 2000) is one of the most widely used constraint-based methods, and recovers the CPDAG of $\mathcal{G}$ in two phases.

The first phase recovers the skeleton. The algorithm first assumes the complete undirected graph and then tests the existence of each edge, using the fact that two non-adjacent variables in a DAG are d-separated by some subset of the neighbors of either one, so that for a pair X_i, X_j,

$$X_j \notin \mathrm{adj}(i) \iff \exists\, \mathbf{S} \subseteq \mathrm{adj}(i) \setminus \{X_j\} \text{ such that } X_i \perp\!\!\!\perp X_j \mid \mathbf{S}. \tag{4}$$

the edge between X_i and X_j is removed if CI test with some $\mathbf{S}$ yields independence, and remains if no such $\mathbf{S}$ is found. Every conditioning set that yielded independence is stored as the separating set $\mathrm{sep}(i, j)$, which the next phase uses. In practice, two considerations are typically adopted for feasible use. First, maximum set size $q_{\max}$ of conditioning set is introduced since the size of the conditioning set is maximally 2^{p-2}, making the total number of CI tests grow exponentially in p. Second, the search is confined to the current adjacency sets, justified by equation 4 but means that an incorrect decision on one edge affects the tests afterwards, inducing a dependence between the order in which pairs are visited and the recovered skeleton (Colombo & Maathuis, 2014). To reduce this dependence, the algorithm iterates over the size of the conditioning set in increasing order, deciding the presence of edges after each iteration.

The second phase orients edges, beginning by detecting colliders. Since conditioning on X_k separates X_i and X_j unless X_k is a collider, for X_i, X_j, X_k satisfying $X_i, X_j \in \mathrm{adj}(k)$ and $X_j \notin \mathrm{adj}(i)$,

$$X_k \notin \mathrm{sep}(i, j) \implies X_i \to X_k \leftarrow X_j \tag{5}$$

The remaining edges are then oriented by following rules until none of them applies (Meek, 1995),

$$X_i \to X_j — X_k, \quad X_k \notin \mathrm{adj}(i) \implies X_j \to X_k, \tag{6}$$

$$X_i \to X_k \to X_j, \quad X_i — X_j \implies X_i \to X_j, \tag{7}$$

$$X_i — X_m \to X_j \text{ for } m \in \{k, l\}, \quad X_i — X_j, \quad X_l \notin \mathrm{adj}(k) \implies X_i \to X_j. \tag{8}$$

where $X_i — X_j$ denotes an edge that is present but not yet oriented. Each rule orients an edge whose reverse orientation would either introduce a collider that equation 5 did not detect or create a directed cycle, so no further edge can be oriented from the CI test outcomes alone. Under faithfulness, causal sufficiency and correct test outcomes, the two phases return the CPDAG of $\mathcal{G}$. In practice, however, the orientation step also relies on the recovered skeleton and separating sets, so a single incorrect CI test outcome affects not only the skeleton but the orientations that follow from it, which motivates quantifying how reliable the recovered graph is.

3 The PEC Framework

3.1 Posterior probabilities of the hypotheses

While the appropriate notion of reliability depends on the application, we employ the posterior probability as the confidence score for edge existence, since it indicates the probability given the observed data and is interpretable on its own. To begin, it follows from Bayes' rule that the posterior probabilities of the null and the alternative hypothesis under observed data D decompose into their likelihoods and priors as

$$P(H_0|D) = \frac{P(D|H_0)\,\pi_0}{P(D|H_0)\,\pi_0 + P(D|H_1)\,\pi_1}, \quad P(H_1|D) = \frac{P(D|H_1)\,\pi_1}{P(D|H_0)\,\pi_0 + P(D|H_1)\,\pi_1}, \tag{9}$$

where H_0, H_1 are mutually exclusive and $\pi_0 = P(H_0)$ and $\pi_1 = P(H_1) = 1 - \pi_0$. The posterior is therefore determined once both likelihoods and the prior are available.

In hypothesis testing, the alternative is commonly composite, so its likelihood is not defined by the hypotheses alone. A test is accordingly carried out on the null side only, comparing a p-value against a threshold to choose between H_0 and H_1. Because that decision uses no information about the alternative, it can be wrong when the alternative lies close to the null, since the p-value measures evidence against H_0 rather than the probability that H_1 holds.

The same issue arises for CI tests. The alternative is conditional dependence, which is composite in that it covers every form of dependence, and this remains the case even when the causal mechanism is specified. For the linear Gaussian model of Section 2.2, the alternative is the statement that $\rho_{ij|\mathbf{S}}$ is nonzero, and equation 3 gives one distribution for every value it may take; unless the distribution of $\rho_{ij|\mathbf{S}}$ is known, the likelihood under the alternative remains undefined.

3.2 PEC for skeleton discovery

Nevertheless, test statistics under the alternative do not follow the null and can therefore be expected to form a distinguishable distribution, which is unknown but can be estimated from the observed

statistics. We therefore introduce the Posterior Edge Confidence (PEC) framework: it computes a posterior probability of edge existence from the statistics of the CI tests performed at the skeleton discovery phase, where the alternative is estimated from the observed data. PEC also derive posterior probability by applying the probability rules for entire CI test outcomes, rather than choosing a representative among them such as the maximum (Tsamardinos & Brown, 2008; Triantafillou et al., 2014), which discards the information from the rest.

At the skeleton discovery phase, the existence of an edge is determined by a logical combination of the associated CI statements. Let H_0^{ij} and H_1^{ij} denote the hypotheses that the edge between X_i and X_j is absent and present, respectively. Using the notation $H_0^{ij|\mathbf{S}}$ and $H_1^{ij|\mathbf{S}}$ from equation 1, we have

$$H_0^{ij} \iff \bigcup_{\mathbf{S}} H_0^{ij|\mathbf{S}}, \quad H_1^{ij} \iff \bigcap_{\mathbf{S}} H_1^{ij|\mathbf{S}}, \tag{10}$$

where $\mathbf{S} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$ and subscripts $\{0, 1\}$ denotes null and alternative, respectively. We take the absence of an edge as the null H_0^{ij}, since it is composed of the independence hypotheses $H_0^{ij|\mathbf{S}}$.

To derive the posterior probability of edge existence from the above, we first propose equations that combine the likelihoods of the individual hypotheses, following the approach of Strobl et al. (2019).

Let $\{H_i\}_{i=1}^m$ be hypotheses and $\{Z_i\}_{i=1}^m$ be test statistics corresponding to the hypotheses. Then the law of total probability yields that the joint density of $\mathbf{Z} = (Z_1, ..., Z_m)$ is represented as

$$P(\mathbf{Z}) = P(\mathbf{Z}|\bigvee_{i=1}^m H_i)P(\bigvee_{i=1}^m H_i) + P(\mathbf{Z}|\bigwedge_{i=1}^m \neg H_i)P(\bigwedge_{i=1}^m \neg H_i). \tag{11}$$

Assuming that the $\{H_i\}_{i=1}^m$ are mutually independent and that each Z_i depends only on H_i, we have

$$\prod_{i=1}^m P(Z_i) = P\Big(\mathbf{Z}\Big|\bigvee_{i=1}^m H_i\Big)\Big(1 - \prod_{i=1}^m P(\neg H_i)\Big) + \prod_{i=1}^m P(Z_i|\neg H_i)\prod_{i=1}^m P(\neg H_i) \tag{12}$$

Furthermore, it follows that $P(Z_i) = P(Z_i|H_i)P(H_i) + P(Z_i|\neg H_i)P(\neg H_i)$ so the likelihood term of the union of hypotheses can be represented by the priors of the hypotheses and their individual likelihoods. The intersection term likewise factorizes into individual likelihoods. Solving equation 12 for the union term therefore gives the two rules

$$P\Big(\mathbf{Z}\Big|\bigvee_{i=1}^m H_i\Big) = \frac{\prod_{i=1}^m P(Z_i) - \prod_{i=1}^m P(Z_i|\neg H_i)\prod_{i=1}^m P(\neg H_i)}{1 - \prod_{i=1}^m P(\neg H_i)}, \tag{13}$$

$$P\Big(\mathbf{Z}\Big|\bigwedge_{i=1}^m H_i\Big) = \prod_{i=1}^m P(Z_i|H_i). \tag{14}$$

In the case of skeleton discovery, the likelihood of non-edge side H_0^{ij} can be derived by applying the union rule to the $\{H_0^{ij|\mathbf{S}_i}\}_{i=1}^m$, and that of edge side H_1^{ij} can be derived by applying the intersection rule to the $\{H_1^{ij|\mathbf{S}_i}\}_{i=1}^m$, where m is the number of conditioning sets tested for the pair.

Since the priors $\pi_0^{ij|\mathbf{S}}, \pi_1^{ij|\mathbf{S}}$ and the likelihood under $H_1^{ij|\mathbf{S}}$ are not supplied by the CI tests, we estimate these terms from the observed data. In this work we consider the priors and the likelihoods are shared across all variable pairs X_i, X_j and estimate them with a mixture of normal distributions which is a simple and natural choice for a density, but it can be chosen arbitrarily within the PEC framework. For the estimation, we collect the absolute values $|Z_{ij|\mathbf{S}}|$ of the statistics of all CI tests performed during the skeleton search and fit a two-component mixture to them, where the component concentrated near zero is taken as the null and the other as the alternative.

Once we have the priors π_0, π_1 and the likelihoods of H_1 from the estimation step, the two rules give both likelihoods and priors at the edge level, and the posterior probability of edge existence follows from equation 9 as

$$P(H_1|\mathbf{Z}) = \frac{P(\mathbf{Z}|H_1)\,\pi_1}{P(\mathbf{Z}|H_0)\,\pi_0 + P(\mathbf{Z}|H_1)\,\pi_1}. \tag{15}$$

Algorithm 1 summarizes the procedure.

Algorithm 1 PEC for skeleton discovery

Require: data over $\mathbf{X}$, significance level α, maximum conditioning set size $q_{\max}$
Ensure: skeleton and posterior probabilities $P(H_1^{ij}|\mathbf{Z})$

initialise the complete undirected graph and $\mathcal{Z}_{ij} \leftarrow \emptyset$ for every pair
for $q = 0, 1, \ldots, q_{\max}$ **do**
  $\mathcal{A}(i) \leftarrow \mathrm{adj}(i)$ for every X_i; $\mathcal{R} \leftarrow \emptyset$
  for each adjacent pair X_i, X_j with $|\mathcal{A}(i) \setminus \{X_j\}| \geq q$ **do**
    for each $\mathbf{S} \subseteq \mathcal{A}(i) \setminus \{X_j\}$ with $|\mathbf{S}| = q$ **do**
      compute $Z_{ij|\mathbf{S}}$ and append it to $\mathcal{Z}_{ij}$
      if $p_{ij|\mathbf{S}} \geq \alpha$ **then**
        $\mathcal{R} \leftarrow \mathcal{R} \cup \{(i, j)\}$ and **break**
      end if
    end for
  end for
  remove every edge in $\mathcal{R}$
end for
fit a two-component mixture to $\{|z| : z \in \bigcup_{ij} \mathcal{Z}_{ij}\}$, to derive the priors $\{\pi_0, \pi_1\}$ and the density of alternatives H_1
for each pair X_i, X_j **do**
  evaluate the per-test likelihoods at the statistics in $\mathcal{Z}_{ij}$
  compute the posterior from equation 13 , equation 14, and equation 15
end for

4 Experiments

4.1 Experiential setup

We generate 30 random DAGs over $p \in \{10, 15, 20\}$ variables with dp edges, where $d \in \{0.75, 1.00, 1.25, 1.50\}$ is the number of edges per variable, so that the average degree $2|E|/p$ ranges from 1.5 to 3. From each DAG we construct datasets of $n \in \{100, 300, 1000, 3000\}$ samples from the linear Gaussian model equation 2, in which root variables follow $\mathcal{N}(0, 1)$ and the coefficients b_{ij} are drawn uniformly from $[0.3, 1] \cup [-1, -0.3]$ and then rescaled so that every variable has unit variance. We take the same threshold α for all methods but choose differently for each experiment as it is sensitive to the density of edges, chosen as the value the two Gaussian cluster H_0, H_1 meet each other. We set the maximum conditioning set size $q_{\max} = 3$ for all experiments.

Baselines. We compare against PROPeR (Triantafillou et al., 2014), one of the closest approaches to the our method which derives the posterior probability of edge presence. It takes the maximum p-value of the tests attempted for a pair as a representative and models its alternative distribution with a beta family. Specifically, the method estimates π_0 by the spline method of Storey & Tibshirani (2003), which frequently saturates at the boundary, addressed in Triantafillou et al. (2014), since the representative p-values are neither i.i.d. nor uniform under the null: independence is accepted only when $p \geq \alpha$. We therefore also take a variant that fits π_0 and the beta exponent jointly by maximum likelihood, which is the stronger baseline and the one we compare against.

Ablation. To show the effectiveness of applying the probability rules we proposed, we also consider the "rule-applied" version of above two baselines, becomes 4 in total. We denotes "MLE" to methods that take the maximum likelihood fitting and "rules" to methods that applies our probability rules.

Metrics. We evaluated the derived posterior probability with following metric for each adjacency: the Brier score (Brier) (Brier, 1950) which is the mean squared difference between the posterior and the indicator of edge existence, the expected calibration error (ECE) (Naeini et al., 2015; Guo et al., 2017) which bins the posteriors and averages the absolute gap between the mean posterior and the observed frequency within each bin, weighted by bin size, AUROC (Fawcett, 2006) and AUPRC (Davis & Goadrich, 2006) which measure the ability to distinguish the edges from non-edges, implying how well the posteriors order existing edges above absent ones.

Table 1: Posterior quality averaged over all 48 settings, each the mean of 30 repetitions. The right-hand block counts the settings in which a method is best on that metric. Lower is better for Brier and ECE, higher for AUROC and AUPRC.

	mean over settings				settings best (of 48)			
method	Brier ↓	ECE ↓	AUROC ↑	AUPRC ↑	Brier	ECE	AUROC	AUPRC
PROPeR	0.055	0.058	0.978	0.960	0	0	1	1
PROPeR (MLE)	**0.034**	**0.035**	0.978	0.960	15	**24**	0	0
+ rules	0.035	0.038	**0.994**	**0.984**	0	0	15	17
+ rules (MLE)	0.034	0.037	0.994	0.984	**20**	15	4	6
PEC	0.036	0.039	0.994	0.984	13	9	**28**	**24**

Table 2: AUROC and Brier score by graph density, averaged over p and n. $d = |E|/p$ is the number of edges per variable, so the average degree is $2d$. The gap between the representative-statistic baselines and the rule-based methods widens as the graph gets denser.

	PROPeR		PROPeR (MLE)		+ rules		+ rules (MLE)		PEC	
d	AUROC	Brier	AUROC	Brier	AUROC	Brier	AUROC	Brier	AUROC	Brier
0.75	0.994	0.015	0.994	0.007	0.999	0.006	0.999	0.006	0.997	0.007
1.00	0.986	0.038	0.986	0.020	0.998	0.020	0.998	0.019	0.998	0.021
1.25	0.978	0.063	0.978	0.037	0.995	0.039	0.995	0.038	0.995	0.039
1.50	0.955	0.104	0.955	0.073	0.985	0.076	0.985	0.074	0.985	0.076

4.2 Results

Table 1 reports the aggregate, Table 2 the dependence on density, and Figure 1 the dependence on sample size at the densest setting. We have mainly three observations to follow.

Combining every test improves the ranking of edges. Averaged over the 48 settings, the rule-based methods raise AUROC from 0.978 to 0.994 and AUPRC from 0.960 to 0.984, and they are best on those two metrics in total 48 settings respectively. Paired over repetitions, applying the rules to the beta model gains 0.016 AUROC and 0.024 AUPRC on average. The gain grows with density: comparing PEC against the stronger PROPeR variant, AUROC improves by about 0.01 at $d = 0.75$ and by 0.02 to 0.045 at $d = 1.5$, and it shrinks as n grows, which is what one expects if the extra tests matter most where individual decisions are least reliable.

The estimated alternative matters little once the rules are applied. PEC and the beta variant with the same rules agree to within 0.001 AUROC in almost every setting, and the calibration metrics differ by about 0.002 in favour of the beta variant. PEC is best on AUROC in 28 settings and on AUPRC in 24, more often than any other method, but the margin over the beta variant is smaller than the standard deviation across the 30 repetitions. We therefore attribute the improvement to the combination rules rather than to the estimated alternative.

The prior estimator of the published method is the weaker part. Replacing the spline estimator of π_0 with a joint fit lowers the Brier score by 0.021 and the ECE by 0.023 on average, while leaving AUROC and AUPRC unchanged: the prior enters monotonically, so it rescales the posteriors without reordering them. Calibration therefore separates the two variants and discrimination does not, which is why both are reported. On the calibration metrics the MLE baseline is the best method in 15 and 24 settings, more often than PEC; the rule-based methods pay for their sharper ranking with a slightly larger calibration error at the smallest sample sizes.

5 Conclusion

We introduced PEC, a framework that estimates posterior edge probabilities for constraint-based causal discovery by combining every CI test performed for a pair, and showed that the rules improve

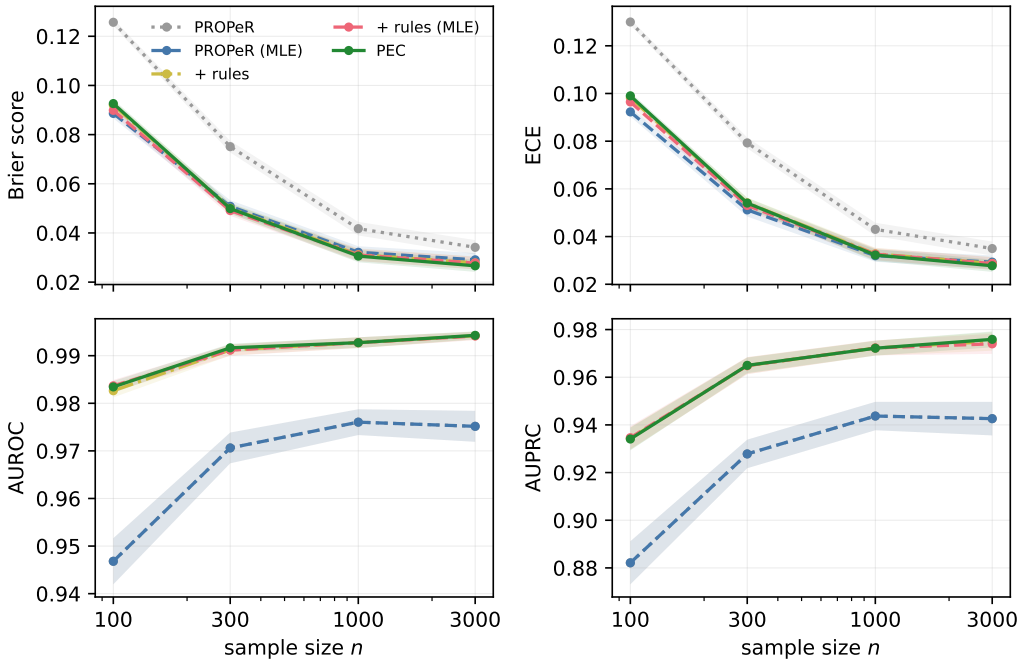


Figure 1: All four metrics against sample size at $p = 20$ with 30 edges (average degree 3), the densest setting. Lower is better for the Brier score and the ECE, higher for AUROC and AUPRC. Lines are means over 30 repetitions and bands one standard error. In the two ranking panels the curves of the two PROPeR variants coincide, since the prior enters the posterior monotonically and does not change the ordering of edges.

both our instantiation and an existing method. We confine the framework to skeleton discovery: expressing the presence of a given orientation as a logical combination of CI outcomes is possible only if the preceding edge decisions are treated as certain, since equation 5 reads off $\mathrm{sep}(i, j)$ and equations 6–8 propagate from the resulting configuration. Conditioning on decisions that are themselves uncertain would defeat the purpose of reporting a posterior, so we leave orientation confidence to future work.

AI USE STATEMENT

In this work, we used generative AI tools for revising the manuscript, tables, and figures.

REFERENCES

Steen A. Andersson, David Madigan, and Michael D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Alexandros P. Armen and Ioannis Tsamardinos. Estimation and control of the false discovery rate of Bayesian network skeleton identification. Technical Report TR-441, University of Crete, 2014.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Kailash Budhathoki, Lenon Minorics, Patrick Bloebaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2357–2369, 2022.

Tom Claassen and Tom Heskes. A Bayesian approach to constraint based causal inference. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 207–216, 2012.

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

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 233–240, 2006.

Tristan Deleu, António Góis, Chris Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio. Bayesian structure learning with generative flow networks. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

Nir Friedman and Daphne Koller. Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50(1–2):95–125, 2003.

Nir Friedman, Moises Goldszmidt, and Abraham Wyner. Data analysis with Bayesian networks: A bootstrap approach. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 196–205, 1999.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

Mikko Koivisto and Kismat Sood. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5:549–573, 2004.

Jack Kuipers and Giusi Moffa. Partition MCMC for inference on acyclic digraphs. *Journal of the American Statistical Association*, 112(517):282–299, 2017.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. DiBS: Differentiable Bayesian structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 403–410, 1995.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 499–506, 1995.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

Daniel J. Stekhoven, Izabel Moraes, Gardar Sveinbjörnsson, Lars Hennig, Marloes H. Maathuis, and Peter Bühlmann. Causal stability ranking. *Bioinformatics*, 28(21):2819–2823, 2012.

John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

David Strieder, Tobias Freidling, Stefan Haffner, and Mathias Drton. Confidence in causal discovery with linear causal models. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1217–1226, 2021.

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.

Sofia Triantafillou, Ioannis Tsamardinos, and Anna Roumpelaki. Learning neighborhoods of high confidence in constraint-based causal discovery. In *Proceedings of the 7th European Workshop on Probabilistic Graphical Models (PGM)*, pp. 487–502, 2014.

Ioannis Tsamardinos and Laura E. Brown. Bounding the false discovery rate in local Bayesian network learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1100–1105, 2008.

Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. Causal discovery in the presence of missing data. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013.

Y. Samuel Wang, Mladen Kolar, and Mathias Drton. Confidence sets for causal orderings. In *arXiv preprint arXiv:2305.14506*, 2023.