Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 9(8): 22743–22793.
DOI: 10.3934/math.20241108
Received: 19 December 2023
Revised: 17 June 2024
Accepted: 1 July 2024
Published: 23 July 2024

Research article

Local discovery in Bayesian networks by information-connecting

Jianying Rong[1,*] **and Xuqing Liu**[2,*]

[1] Department of Quality Education, Jiangsu Vocational College of Electronics and Information, Huai'an 223003, China

[2] Faculty of Mathematics and Physics, Huaiyin Institute of Technology, Huai'an 223003, China

* **Correspondence:** Email: rjy98@163.com, liuxuqing@hyit.edu.cn.

Abstract: Local discovery plays an important role in Bayesian networks (BNs), mainly addressing PC (parents and children) discovery and MB (Markov boundary) discovery. In this paper, we considered the problem of large local discovery. First, we focused on an assumption about conditional independence (CI) tests: We explained why it was unreasonable to assume all CI tests were reliable in large local discovery, studied how the power and reliability of CI tests changed with the data size and the number of degrees of freedom, and then modified the assumption about CI tests in a more reasonable way. Second, we concentrated on improving local discovery algorithms: We posed the problem of premature termination of the forward search, analyze why it arose frequently in large local discovery when implementing the existing local discovery algorithms, put forward an idea of preventing the premature termination of forward search called information connection (IC), and used IC to build a novel algorithm called `ICPC`; the theoretical basis of `ICPC` was detailedly presented. In addition, a more steady incremental algorithm as the subroutine of `ICPC` was proposed. Third, the way of breaking ties among equal associations was considered and optimized. Finally, we conducted a benchmarking study by means of six synthetic BNs from various domains. The experimental results revealed the applicability and superiority of `ICPC` in solving the problem of premature termination of the forward search that arose frequently in large local discovery.

Keywords: Bayesian network; local discovery; PC discovery; MB discovery; information connection
Mathematics Subject Classification: 68T20, 94A15

1. Introduction

Bayesian networks (BNs) are graphical structures used to represent the probabilistic relations among a number of variables [1, 2]. In recent years, BNs are becoming one of the most powerful tools in encoding uncertain knowledge in expert systems [3, 4]; they have been widely used in many actual domains such as medical diagnosis, financial analysis, bioinformatics, and industrial applications [5].

There are three components in a BN denoted by $(\mathbb{G}, \mathbb{P})$: (a) Graphical component: $\mathbb{G}$ is a directed acyclic graph (DAG); (b) probabilistic component: $\mathbb{P}$ is a set of conditional probability distributions with respect to every node conditioned on its parents; (c) Markovianity: $\mathbb{G}$ and $\mathbb{P}$ are supposed to satisfy the Markov condition: Every node is conditionally independent of its nondescendants given its parents. This means that structure learning and parameter learning are two primary subtasks of capturing a complete BN from data. This paper focuses on local structure learning.

Local structure learning mainly addresses two types of local discovery for a target variable, T: One is to discover the set of parents and children (PC) of T, and the other is to find a Markov boundary (MB) of T. Here, an MB of T is a minimal variable set that renders T independent of all other variables. Under the faithfulness condition, all the PC and spouses of T constitute its unique MB. This paper mainly focuses on large PC and MB discovery.

PC discovery is the most critical technique used for the divide-and-conquer local-to-global strategy for learning BNs [6–9], while MB discovery plays a central role in feature selection [10, 11] as well as in the local-to-global strategy for learning Markov networks or moralized BNs [12]. Pearl [1] showed that the conditional probability for T given other variables coincides with the one with an MB as the conditional set. Pellet and Elisseeff [13] proved an MB is the theoretically optimal set of features under the faithfulness condition. Further, under certain assumptions about the learner and the loss function, MB is the solution to the feature selection problem [14–16]. This is why local discovery techniques are receiving more and more attention in recent years [17–19].

In the literature, there have been lots of independence-based (or called constraint-based) approaches for local discovery. Each of these algorithms requires a number of conditional independence (CI) tests to identify the members of the PC or MB. When the PC or MB of a target is not large, these algorithms are often enough for practitioners. However, in the case of large PCs or MBs, the existing local discovery algorithms may not effectively return the expected results due to the unreliability of some CI tests. This paper addresses how to effectively deal with large local discovery problems.

The remainder of this paper is organized as follows. Section 2 poses three problems as the motivation of this paper. Section 3 addresses an assumption about the reliability of CI tests and provides a more reasonable modification. In Section 4, a novel information connection (IC) based algorithm called `ICPC` is proposed to overcome the problem of premature termination of the forward search. Section 5 presents a new way of breaking the possible ties. A benchmarking study is conducted in Section 6. Section 7 concludes this paper and makes some discussions. The appendices provide proofs for theoretical results and also the list of all acronyms.

2. Motivation

This section poses three problems (denoted by $\mathscr{P}_i$ for $i = 1, 2, 3$, respectively) around the shortcomings of independence-based local discovery algorithms. They are the motivation of this paper.

Assumption 1. *Assume all CI tests are reliable. Denote this assumption by $\mathscr{A}_1$.*

Problem $\mathscr{P}_1$ concerns the unreliability of CI tests with large conditional sets, especially when the data size is comparatively small. This well-known problem is common to all independence-based algorithms [10], meaning that the commonly used assumption, $\mathscr{A}_1$, is unreasonable. For a local discovery algorithm, when there are some unreliable CI tests involved and thus $\mathscr{A}_1$ is violated, some true positives (TPs) may become false negatives (FNs), while some true negatives (TNs) may become false positives (FPs). Note that the unreliability of CI tests may lead to spurious *information equivalence* [16, 20, 21].

The above analysis motivates us: (a) To modify the assumption $\mathscr{A}_1$ in a reasonable way; (b) to build a more efficient local discovery algorithm that can output as many TPs as possible and as few FPs as possible under the modified assumption. Note that the quality of statistical decisions is not fixed by the correctness of an independence-based algorithm. This inspires us to take the quality of statistical decisions into account when dealing with (a) and (b).

Problem $\mathscr{P}_2$ concerns the premature termination of the forward search; it is an inherent consequence of the first problem. To be intuitive, we provide the following Example 1, by which we find the detection of true dependencies becomes harder and harder as the conditional set size increases. In Section 3, we will explain why this phenomenon happens.

Example 1. *Consider the six BNs used in Section 6. For each BN and for every case of the conditional set size (denoted by q for convenience), we randomly select 30 true dependencies with nearly the same theoretical degrees of freedom; Figure 1 presents the results, in which each value is averaged over the corresponding 30 true dependencies for every case of q. By the figure, the power of CI tests declines sharply with q, almost closing to zero when $q \geqslant 9$ for any case.*

This example indicates that *a seemingly very large data size may be not large enough* for detecting true dependencies. As a consequence of data insufficiency, a local discovery algorithm may not include all TPs and thus cannot effectively exclude all FPs. This motivates us to seek a feasible method to prevent or alleviate the premature termination of the forward search.

Problem $\mathscr{P}_3$ concerns the way of breaking ties. As we know, in the growing phase of a local discovery algorithm, there is usually a re-ordering procedure by means of an *association* function, $f_{\mathfrak{D}}$. Here, the most widely used selection for $f_{\mathfrak{D}}$ is the *negative p-value* in conjunction with Pearson's χ^2 test or the log-likelihood ratio G^2 test [7–9, 16, 22]. This paper uses the G^2 test. The use of $f_{\mathfrak{D}}$ is an efficient dynamic heuristic. In the meanwhile, it may also lead to some ties in the sense that two or more variables have the largest association with the target mainly because (i) the test statistics are very large such that all the related association values are set to be 0, or (ii) the dataset is insufficient such that these association values happen to be identical. In the literature, the ties are often simply broken at random [10, 23]. However, this way of breaking ties does not consider the case that the selected variable may be an FP; if this is the case, it will lower the quality of the subsequent CI tests. Therefore, it is meaningful to seek some heuristic or optimized criterion, and then use it to guide the way of breaking ties rather than simply breaking ties at random.

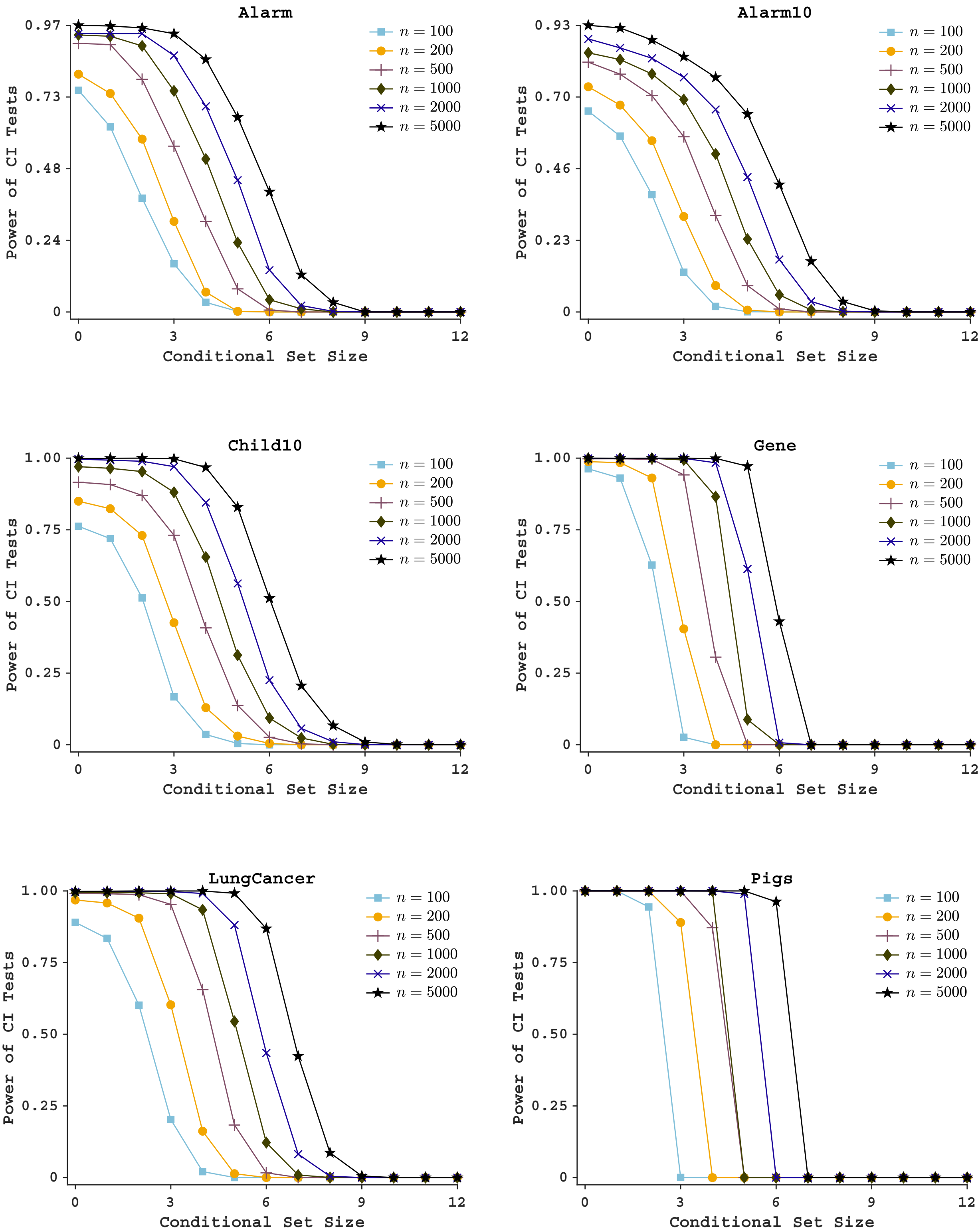


Figure 1. An illustration on the power of CI tests versus the conditional set size.

3. Two algorithmic aspects

This section addresses a part of the problem $\mathscr{P}_1$ posed in Section 2: How to modify the assumption $\mathscr{A}_1$ in a reasonable way. We use the G^2 test in this paper. In addition, the negative p-value as an association function will be briefly discussed in this section.

For convenience, we collect the main symbols used in this paper and list them in Table 1.

Table 1. Main symbols with descriptions.

Symbol	Description
$(\mathbb{G}, \mathbb{P})$	A BN with $\mathbb{G}$ and $\mathbb{P}$ as its graphical and probabilistic components
$\mathscr{P}_i$ $(i = 1, 2, 3)$	Three problems constituting the motivation of this paper
$\mathscr{A}_i$ $(i = 1, 2, 3)$	Three assumptions presented in Section 2 and Section 3
$\mathcal{D}$	A data set containing n data instances
$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$	$\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent given $\boldsymbol{Z}$
$\boldsymbol{X} \perp\!\!\!\perp_{\mathcal{D}} \boldsymbol{Y} \mid \boldsymbol{Z}$	$\boldsymbol{X}$ and $\boldsymbol{Y}$ are deemed to be conditionally independent given $\boldsymbol{Z}$ based on $\mathcal{D}$
$\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$	$\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally dependent given $\boldsymbol{Z}$
$\boldsymbol{X} \not\perp\!\!\!\perp_{\mathcal{D}} \boldsymbol{Y} \mid \boldsymbol{Z}$	$\boldsymbol{X}$ and $\boldsymbol{Y}$ are deemed to be conditionally dependent given $\boldsymbol{Z}$ based on $\mathcal{D}$
$\mathrm{I}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$	Conditional mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$ assumed to be a random variable with $$\mathrm{I}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z}) \sim g(\tau) \triangleq \begin{cases} g_+(\tau), & \tau > 0 \\ \delta(\tau/g_0) = g_0 \cdot \delta(\tau), & \tau = 0 \end{cases}$$ where $g_+(\tau)$ is a nonnegative integrable function on $\tau \in (0, +\infty)$; $g_0 = 1 - \int_0^{+\infty} g_+(\tau)\mathrm{d}\tau \in (0, 1)$; $\delta(\tau)$ is the Dirac δ-function
$\mathrm{I}_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$	Empirical estimate of $\mathrm{I}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$ based on $\mathcal{D}$
$G^2_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$	G^2 statistic defined as $G^2_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z}) \triangleq 2n \cdot \mathrm{I}_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$
$p_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$	p-value defined as $p_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z}) \triangleq \mathrm{P}\{\chi^2(r) \geqslant G^2_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})\}$
$f_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$	Association function taken as the negative p-value: $f_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z}) \triangleq -\mathrm{P}\{\chi^2(r) \geqslant G^2_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})\}$
$\chi^2(r)$	Central χ^2-variate with r degrees of freedom
$f_r(x)$	Probability density function of $\chi^2(r)$
$F_r(x)$	Cumulative distribution function of $\chi^2(r)$
$\chi^2_\alpha(r)$	Upper α-quantile of $\chi^2(r)$
$\chi^2(r, \delta)$	Noncentral χ^2-variate with r degrees of freedom and the noncentrality parameter δ
$f_{r,\delta}(x)$	Probability density function of $\chi^2(r, \delta)$
$F_{r,\delta}(x)$	Cumulative distribution function of $\chi^2(r, \delta)$
α	Significance level used to making CI tests, taken as 0.001 in the experiment of this paper
r	Number of the theoretical degrees of freedom of a G^2 statistic
δ	Noncentrality parameter of the G^2 statistic, $G^2_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$, defined as $\delta \triangleq 2n \cdot \mathrm{I}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$
r_n	Number of the valid degrees of freedom based on the data $\mathcal{D}$
$\langle \boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z} \rangle$	Random variable in the sense of $\langle \boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z} \rangle = \begin{cases} 1, & \text{if } \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \\ 0, & \text{if } \boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \end{cases}$
$\mathrm{E}_{\perp\!\!\!\perp}$	True independence defined as $\mathrm{E}_{\perp\!\!\!\perp} \triangleq$ "$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$" = "$\langle \boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z} \rangle = 1$"
$\mathrm{E}_{\not\perp\!\!\!\perp}$	True dependence defined as $\mathrm{E}_{\not\perp\!\!\!\perp} \triangleq$ "$\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$" = "$\langle \boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z} \rangle = 0$"
$\langle \boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z} \rangle_{\mathcal{D}}$	Random variable in the sense of $\langle \boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z} \rangle_{\mathcal{D}} = \begin{cases} 1, & \text{if } \boldsymbol{X} \perp\!\!\!\perp_{\mathcal{D}} \boldsymbol{Y} \mid \boldsymbol{Z} \\ 0, & \text{if } \boldsymbol{X} \not\perp\!\!\!\perp_{\mathcal{D}} \boldsymbol{Y} \mid \boldsymbol{Z} \end{cases}$
$\mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}}$	Tested independence defined as $\mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}} \triangleq$ "$\boldsymbol{X} \perp\!\!\!\perp_{\mathcal{D}} \boldsymbol{Y} \mid \boldsymbol{Z}$" = "$p_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z}) > \alpha$" = "$\langle \boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z} \rangle_{\mathcal{D}} = 1$"
$\mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}}$	Tested dependence defined as $\mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}} \triangleq$ "$\boldsymbol{X} \perp\!\!\!\perp_{\mathcal{D}} \boldsymbol{Y} \mid \boldsymbol{Z}$" = "$p_{\mathcal{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z}) \leqslant \alpha$" = "$\langle \boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z} \rangle_{\mathcal{D}} = 0$"
$k_{\max}$	A parameter of GLL used to place an absolute limit on the conditional set size, taken as 3 in this paper
PA_T	Parents of T
CH_T	Children of T
PC_T	Parents and children of T: $\mathit{PC}_T \triangleq \mathit{PA}_T \cup \mathit{CH}_T$
SP_T	Spouses of T
MB_T	MB of T with $\mathit{MB}_T = \mathit{PC}_T \cup \mathit{SP}_T$ under the faithfulness condition
TPC_T	Any available tentative PC of T that is a superset of PC_T
EPC_T	Extended PC_T defined by Aliferis et al. [8] as $\mathit{EPC}_T \triangleq \mathit{PC}_T \cup \{X \in \boldsymbol{V} \setminus \mathit{PC}_T \setminus \{T\} : T \not\perp\!\!\!\perp X \mid \boldsymbol{Z}, \forall \boldsymbol{Z} \subseteq \mathit{PC}_T\}$
$\mathit{MB}_T^{(Y)}$	Y-EMB of T in the sense that it is an MB of T in $\boldsymbol{V} \setminus \{Y\}$
$\mathscr{M}_i$ $(i = 1, 2)$	Information flow metaphor of Cheng et al. [24] and our extended information flow metaphor
$\boldsymbol{A} \times \boldsymbol{B}$	Cartesian product of $\boldsymbol{A}$ and $\boldsymbol{B}$ employed in Eq. (5.1)
$\Gamma(\cdot)$	Gamma function defined as $\Gamma(\alpha) \triangleq \int_0^{+\infty} \mathrm{e}^{-x} x^{\alpha-1} \mathrm{d}x$

3.1. CI test

Denote now $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ (*resp.*, $\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$) if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent (*resp.*, dependent) given $\boldsymbol{Z}$, and denote the conditional mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$ by $\mathrm{I}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z})$. It is well-known that $\mathrm{I}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) \geqslant 0$, with equality holding if and only if $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$. For a practical problem, we cannot access to the true value of $\mathrm{I}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z})$; instead, we use its empirical estimate, namely, $\mathrm{I}_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z})$, based on the data $\mathcal{D}$ as Cheng et al. did [24]. Note that $\mathrm{I}_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) \geqslant 0$ also holds for any $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$.

For $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$, G^2 test tries to determine if the null hypothesis, $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$, holds for the significance level α (taken as 0.001 in the experiment of this paper). Let n be the data size. Then, the G^2 statistic is defined as $G^2_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \triangleq 2n \cdot \mathrm{I}_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$, which approximates to a noncentral χ^2-variate with $r \triangleq (r_X - 1)(r_Y - 1)r_Z$ theoretical degrees of freedom and the noncentrality parameter $\delta \triangleq 2n \cdot \mathrm{I}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$. Here, $r_{\boldsymbol{\xi}}$ is the number of configurations for $\boldsymbol{\xi}$ [25–27]. That is, $G^2_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \dot{\sim} \chi^2(r,\delta)$. If the null hypothesis holds, $G^2_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \dot{\sim} \chi^2(r)$. Denote the corresponding p-value by $p_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) \triangleq \mathrm{P}\{\chi^2(r) \geqslant G^2_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z})\}$. Then, the G^2 test concludes $\boldsymbol{X} \perp\!\!\!\perp_{\mathcal{D}} \boldsymbol{Y} \mid \boldsymbol{Z}$ if $p_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) > \alpha$, and asserts $\boldsymbol{X} \not\perp\!\!\!\perp_{\mathcal{D}} \boldsymbol{Y} \mid \boldsymbol{Z}$ if $p_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) \leqslant \alpha$. Accordingly, the negative p-value is used as the association function, $f_{\mathcal{D}}$. That is, $f_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) \triangleq -\mathrm{P}\{\chi^2(r) \geqslant G^2_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z})\}$.

In practical situations, $\mathcal{D}$ may not be large enough for testing $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$ in the sense that there are some invalid cells (low expected counts) in the associated contingency table, as Cochran [28, p. 420] recommended about the working rules for the G^2 test. For this case, many authors have considered some improvements by adjusting G^2 [29–32]. Brin et al. [33] and Silverstein et al. [34] used two heuristic "solutions" as follows: (i) Simply ignore the invalid cells when calculating G^2; and (ii) use the *contingency table support*.

Let the number of valid degrees of freedom based on $\mathcal{D}$ be r_n. Although r_n is actually unknown, it is clear that $r_n \leqslant r$, with inequality holding when $\mathcal{D}$ is insufficient for this CI test. In what follows, we assume r_n is increasing with n in the probabilistic sense.

Denote the upper α-quantile of $\chi^2(r)$ by $\chi^2_\alpha(r)$, and

$$\begin{aligned}
&\mathrm{E}_{\perp\!\!\!\perp} \triangleq \text{“}\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}\text{''}, \quad \mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}} \triangleq \text{“}\boldsymbol{X} \perp\!\!\!\perp_{\mathcal{D}} \boldsymbol{Y}|\boldsymbol{Z}\text{''} = \text{“}G^2_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \leqslant \chi^2_\alpha(r)\text{''} = \text{“}p_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) > \alpha\text{''},\\
&\mathrm{E}_{\not\perp\!\!\!\perp} \triangleq \text{“}\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}\text{''}, \quad \mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}} \triangleq \text{“}\boldsymbol{X} \not\perp\!\!\!\perp_{\mathcal{D}} \boldsymbol{Y}|\boldsymbol{Z}\text{''} = \text{“}G^2_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) > \chi^2_\alpha(r)\text{''} = \text{“}p_{\mathcal{D}}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) \leqslant \alpha\text{''}.
\end{aligned} \tag{3.1}$$

Note that we have treated the truth of the hypothesis "$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$" as a binary random variable located in a meta-space representing all possible independencies in the domain, as Bromberg and Margaritis [22, p. 305] did. Also, this treatment coincides with the viewpoint of Aliferis et al. [9, p. 249] that statistical reliability of a single test is a misleading concept in the context of complex independence-based algorithms. With these notations, we show the following theorem in Appendix A.1.

Theorem 1. *[Power and Reliability of CI Tests] Assume $\mathcal{D}$ is an insufficient dataset. Then, we have*

a) $\mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}} | \mathrm{E}_{\perp\!\!\!\perp}, \mathcal{D})$ *is decreasing with n and increasing with r.*
b) $\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}} | \mathrm{E}_{\not\perp\!\!\!\perp}, \mathcal{D})$ *is increasing with n and decreasing with r.*
c) $\mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp} | \mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}}, \mathcal{D})$ *is increasing with n and decreasing with r.*
d) $\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp} | \mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}}, \mathcal{D})$ *is decreasing with n and increasing with r.* □

In what follows, we discuss Theorem 1 and then modify the assumption $\mathscr{A}_1$ in a reasonable way. For convenience, we call a $\mathcal{D}$-based CI test with the null hypothesis (i.e., "$\perp\!\!\!\perp$") as its decision to be a "$\perp\!\!\!\perp_{\mathcal{D}}$-test", and call a test with the alternative hypothesis (i.e., "$\not\perp\!\!\!\perp$") as its decision a "$\not\perp\!\!\!\perp_{\mathcal{D}}$-test". According to

Theorem 1, there are two factors influencing the power and reliability of CI tests: One is the data size, n, and the other is the number of theoretical degrees of freedom, r. Further, Theorem 1 in conjunction with Lemma 2 indicates the following conclusions:

- *Power of CI tests:* On the one hand, the type-I error is not larger than α since

$$\begin{aligned} r_n \leqslant r \quad &\Rightarrow \quad \mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}} \,|\, \mathrm{E}_{\perp\!\!\!\perp}, \mathcal{D}) = 1 - F_{r_n}(\chi^2_\alpha(r)) \leqslant 1 - F_r(\chi^2_\alpha(r)) = \alpha \\ &\Rightarrow \quad \mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}} \,|\, \mathrm{E}_{\perp\!\!\!\perp}, \mathcal{D}) \geqslant 1 - \alpha. \end{aligned}$$

  This means almost all *true independencies* can be correctly detected in any case of the data size. On the other hand, the type-II error $\mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}} \,|\, \mathrm{E}_{\not\perp\!\!\!\perp}, \mathcal{D})$ increases when n decreases or r increases. This explains the phenomenon shown in Example 1 that the detection of *true dependencies* becomes harder and harder when (i) the data size decreases, or (ii) the conditional set size increases. This also explains why non-PC-based algorithms such as `InterIAPC` and `InterIAMB` become inefficient and even invalid when used for large local discovery. In comparison, the PC-based algorithms such as `GLL` and `PCMB` possess better performance in resisting this kind of violation but still inevitably become invalid if the PCs or MBs are large enough.
- *Reliability of CI tests:* Note that $\lim_{n\to\infty} F_{r_n,2n\tau}(\chi^2_\alpha(r)) = 0$ and $\lim_{n\to\infty} F_{r_n}(\chi^2_\alpha(r)) = 1 - \alpha$. Employing (d) of Theorem 1 and Eq (A.7) of Appendix A.1, we have

$$\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp} \,|\, \mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}}, \mathcal{D}) \geqslant \lim_{n\to\infty} \left[1 + \left(\frac{1}{g_0}\int_0^{+\infty} g_+(\tau)\frac{1 - F_{r_n,2n\tau}(\chi^2_\alpha(r))}{1 - F_{r_n}(\chi^2_\alpha(r))}\mathrm{d}\tau\right)^{-1}\right]^{-1} = \frac{1 - g_0}{1 - (1-\alpha)g_0}.$$

  Here, this lower bound of $\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp} \,|\, \mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}}, \mathcal{D})$ is near to 1, revealing that most of $\not\perp\!\!\!\perp_{\mathcal{D}}$-tests are reliable in the probabilistic sense, especially when the data size is small. Consequently, it is reasonable to assume "all $\not\perp\!\!\!\perp_{\mathcal{D}}$-tests are reliable".

  Conversely, (c) of Theorem 1 implies the reliability of $\perp\!\!\!\perp_{\mathcal{D}}$-tests decreases when n decreases or r increases. Moreover, noting $\lim_{\tau\to 0^+} \chi^2_\alpha(r_n, 2n\tau) < \chi^2_\alpha(r)$ and $\lim_{\tau\to+\infty} \chi^2_\alpha(r_n, 2n\tau) > \chi^2_\alpha(r)$ since $r > r_n$, there must be some $\tau_{r,n} > 0$ such that $\chi^2_\alpha(r_n, 2n\tau_{r,n}) = \chi^2_\alpha(r)$. Then, $F_{r_n,2n\tau}(\chi^2_\alpha(r)) > 1 - \alpha$ holds for any $\tau \in (0, \tau_{r,n})$. Putting $g_{r,n} \triangleq \int_0^{\tau_{r,n}} g_+(\tau)\mathrm{d}\tau$, it follows from Eq (A.6) that

$$\mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp} \,|\, \mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}}, \mathcal{D}) < \left(1 + \frac{1}{g_0}\int_0^{\tau_{r,n}} g_+(\tau)\frac{F_{r_n,2n\tau}(\chi^2_\alpha(r))}{F_{r_n}(\chi^2_\alpha(r))}\mathrm{d}\tau\right)^{-1} < \left(1 + \frac{(1-\alpha)g_{r,n}}{g_0}\right)^{-1},$$

  in which the upper bound will approximate to g_0 if r is large enough (corresponds to the case that all CI-tests become $\perp\!\!\!\perp_{\mathcal{D}}$-tests). Consequently, it is unreasonable to assume "all $\perp\!\!\!\perp_{\mathcal{D}}$-tests are reliable". This is the key of modifying $\mathscr{A}_1$.

By the above analysis, the fewer instances in $\mathcal{D}$ or the more cells in the contingency table, the less reliable $\perp\!\!\!\perp_{\mathcal{D}}$-tests are, and thus the more reliable $\not\perp\!\!\!\perp_{\mathcal{D}}$-tests are. Combined with the idea of *heuristic power size* (hps) employed by Aliferis et al. [8, 9] and the heuristic suggested by Yaramakala [35] that "we add variables as long as the CI tests are reliable enough", we modify $\mathscr{A}_1$ to $\mathscr{A}_2$ as follows:

Assumption 2. *The assumption $\mathscr{A}_2$ contains two parts: (a) All $\not\perp\!\!\!\perp_{\mathcal{D}}$-tests are reliable; (b) all $\perp\!\!\!\perp_{\mathcal{D}}$-tests are reliable except for the following case: If a $\perp\!\!\!\perp_{\mathcal{D}}$-test "$X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{Z}$" with r degrees of freedom is* *incompatible to another ℓ ($\geqslant 1$) $\perp\!\!\!\perp_{\mathcal{D}}$-tests "$X_i \perp\!\!\!\perp_{\mathcal{D}} Y_i \mid \mathbf{Z}_i$" ($i = 1, \cdots, \ell$) with at most r_0 degrees of freedom subject to $r > r_0$, then "$X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{Z}$" is deemed unreliable given no further evidence of independence for it.* □

Consider again the situation where $\mathcal{D}$ is an insufficient dataset. Aliferis et al. [8, 9] recommended an hps-based criterion to deal with this problem in practice: A CI test is reliable if and only if at least hps sample instances per cell in the contingency table are available, and deem an unreliable CI test (but required to make further decisions in the forward or backward searches of an algorithm) to return independence given no evidence of dependence. In their works, hps is set to be 10 in PC-based algorithms and 5 in non-PC-based algorithms. Besides hps, Aliferis et al. [8, 9] provided a second parameter, $k_{\max}$, to place an absolute limit on the conditional set size. The $k_{\max}$-based criterion forces those CI tests with the conditional set sizes larger than $k_{\max}$ not to be performed. Thus, as pointed out by Aliferis et al. [8, p. 201], this criterion participates in the reliability judgment and also restricts the computational complexity of the algorithm involved. In the experimental section of this paper, we set hps and $k_{\max}$ as 10 and 3, respectively.

Under the above criteria based on hps and $k_{\max}$, the following assumption about whether a CI test will be done is then useful for supplementing $\mathscr{A}_1$ and $\mathscr{A}_2$. That is, $\mathscr{A}_1$ or $\mathscr{A}_2$ works under this assumption.

Assumption 3. *$\mathscr{A}_3$ assumes that, for any $T, X \in \mathbf{V}$ and $\mathbf{Z} \subseteq \mathbf{V} \setminus \{T, X\}$, the CI test for T and X conditioned on $\mathbf{Z}$ is done if, and only if, the conditions $(r_T - 1)(r_X - 1)r_{\mathbf{Z}} \cdot hps \leqslant n$ and $|\mathbf{Z}| \leqslant k_{\max}$ are satisfied simultaneously.* □

3.2. Association function

Negative p-value is one of the most widely used association functions [7–9, 16]. This subsection provides a property of this function. For $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ and a dataset $\mathcal{D}$, recall that the theoretical degrees of freedom, the noncentrality parameter, and the valid degrees of freedom based on the data $\mathcal{D}$ are denoted by $r \triangleq (r_{\mathbf{X}} - 1)(r_{\mathbf{Y}} - 1)r_{\mathbf{Z}}$, $\delta \triangleq 2n \cdot \mathrm{I}(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$, and r_n, respectively. The following theorem presents the probabilistic monotonicity of the negative p-value, $f_{\mathcal{D}}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$, with respect to these parameters.

Theorem 2. *Assume $\mathcal{D}$ is an insufficient dataset. Then, the negative p-value $f_{\mathcal{D}}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$ is increasing with δ and n and decreasing with r.*

Proof. Note that $G^2_{\mathcal{D}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z}) \triangleq 2n \cdot \mathrm{I}_{\mathcal{D}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$ is an approximate χ^2-variate with r theoretically degrees of freedom in which only r_n ($< r$) ones are valid. With the notations in Appendix A.1, we have

$$
\begin{aligned}
f_{\mathcal{D}}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z}) &= -\mathrm{P}\{\chi^2(r) \geqslant G^2_{\mathcal{D}}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})\} \\
&= -\mathrm{P}\{G^2_{\mathcal{D}}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z}) \leqslant \chi^2(r)\} = -\int_0^{+\infty} F_{r_n,\delta}(x) f_r(x) \mathrm{d}x \qquad (3.2) \\
&= \mathrm{P}\{\chi^2(r) < G^2_{\mathcal{D}}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})\} - 1 = \int_0^{+\infty} F_r(x) f_{r_n,\delta}(x) \mathrm{d}x - 1. \qquad (3.3)
\end{aligned}
$$

Combined with the conclusion (a) of Lemma 2, Eq (3.2) implies $f_{\mathcal{D}}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$ is increasing with δ and n, while Eq (3.3) indicates $f_{\mathcal{D}}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$ is decreasing with r. □

This theorem reveals how the negative p-value changes with r, δ, and n. Combined with Theorem 1, we find this association function has similar monotonicity to the power and reliability of CI tests. In Section 5, we give a brief discussion about how to improve the way of breaking ties via this theorem.

4. Local discovery algorithms

In this section, we address the problems $\mathscr{P}_2$ posed in Section 2: How to alleviate premature termination of the forward search. We first analyze the existing local discovery algorithms and then present a novel algorithm based on the idea of *information connection*.

4.1. Existing local discovery algorithms

Consider a BN $(\mathbb{G}, \mathbb{P})$ over $\boldsymbol{V} \triangleq \{X_1, \cdots, X_v\}$, assuming $\mathbb{P}$ is faithful to $\mathbb{G}$ and $T \in \boldsymbol{V}$ is the target variable of interest. For convenience, we denote the parents, children, and spouses of T by PA_T, CH_T, and SP_T respectively, and put $\mathit{PC}_T \triangleq \mathit{PA}_T \cup \mathit{CH}_T$ and $\mathit{MB}_T \triangleq \mathit{PC}_T \cup \mathit{SP}_T$.

First, Spirtes et al. [36] showed the following conclusion:

Lemma 1. *Let* $(\mathbb{G}, \mathbb{P})$ *be a BN over* $\boldsymbol{V}$ *satisfying the faithfulness condition. For given* $T, X \in \boldsymbol{V}$*, we have* $X \in \mathit{PC}_T$ *if, and only if,* $T \not\perp\!\!\!\perp X \mid \boldsymbol{Z}$ *holds for any* $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{T, X\}$*.* □

Based on this property, Aliferis et al. [8] analyzed a localized version of SGS [36] and then put forward their GLL-PC algorithmic framework. Their analysis focuses on how to implement the local SGS algorithm more efficiently by reducing the search space of the cut set, $\boldsymbol{Z}$. They first reduced the search space from $\{\boldsymbol{Z} : \boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{T, X\}\}$ to $\{\boldsymbol{Z} : \boldsymbol{Z} \subseteq \mathit{PA}_T \text{ or } \boldsymbol{Z} \subseteq \mathit{PA}_X\}$. This holds if $X \notin \mathit{PC}_T$ because of the Markov condition: $T \perp\!\!\!\perp X \mid \mathit{PA}_T$ if X is a nondescendant of T; and $X \perp\!\!\!\perp T \mid \mathit{PA}_X$ otherwise (in this case, T is a nondescendant of X). However, the parents of a node are practically unknown, so Aliferis et al. [8] made a relaxation as follows:

i) Let TPC_T be any available *tentative PC* of T, which is a superset of PC_T.
ii) For each $X \in \mathit{TPC}_T$, remove it from TPC_T if there is $\boldsymbol{Z} \subseteq \mathit{TPC}_T \setminus \{X\}$ such that $T \perp\!\!\!\perp X \mid \boldsymbol{Z}$.
iii) Repeat (ii) until no such X exists.

This procedure refines TPC_T such that it approximates PC_T quite closely, with $\mathit{PC}_T \subseteq \mathit{TPC}_T \subseteq \mathit{EPC}_T$, where EPC_T was defined by Aliferis et al. [8] as $\mathit{EPC}_T \triangleq \mathit{PC}_T \cup \{X \in \boldsymbol{V} \setminus \mathit{PC}_T \setminus \{T\} : T \not\perp\!\!\!\perp X \mid \boldsymbol{Z}, \forall \boldsymbol{Z} \subseteq \mathit{PC}_T\}$. To avoid the situation where $\mathit{TPC}_T \setminus \mathit{PC}_T \neq \emptyset$ as illustrated in Figure 2, Aliferis et al. [8] used a *pruning procedure* via the AND operator* as Peña et al. [23] did in their PCMB algorithm: (iv) For each $X \in \mathit{TPC}_T$, remove it from TPC_T if $T \notin \mathit{TPC}_X$, where TPC_X is obtained by running the above *refining procedure* (ii)~(iii). These are the main ideas of the GLL-PC framework.

$$
\begin{tikzcd}
& T \arrow[d] & \\
& C \arrow[dl] & \\
X & & S \arrow[ll] \arrow[ul]
\end{tikzcd}
$$

Figure 2. An illustration of the situation where $\mathit{TPC}_T \setminus \mathit{PC}_T \neq \emptyset$: $\mathit{TPC}_T = \mathit{EPC}_T = \{C, X\}$ while $\mathit{PC}_T = \{C\}$. See [8, p. 189] for a more detailed explanation. Aliferis et al. [8] also mentioned that such situations are rare in practice, so in general TPC_T outputted by the refining procedure (ii)~(iii) can approximate PC_T quite closely.

*The AND operator means that a node X is regarded as a PC member of T if, and only if, $X \in \mathit{PC}_T$ and $T \in \mathit{PC}_X$ hold simultaneously.

GLL-PC uses a tentative-PC discovery algorithm $\mathbb{A}_{TPC}$ subject to some admissible rules [8], the data $\mathcal{D}$, a target T as its input, and outputs PC_T. Here, $\mathbb{A}_{TPC}$ contains the following steps: (i) Initialize TPC_T with $\boldsymbol{S} \subseteq \boldsymbol{V} \setminus \{T\}$, and initialize a priority queue ρ for $\boldsymbol{V} \setminus TPC_T \setminus \{T\}$; (ii) Apply the inclusion heuristic function to update TPC_T and ρ; (iii) Refine TPC_T; (iv) Interleave and repeat (ii)~(iii) until the termination criterion is met. Aliferis et al. [8, 9] employed two specified parameters, hps and $k_{\max}$, to reduce the number of CI tests. The pseudo-code of GLL-PC is described by (a) of Algorithm 1.

An alternative method of discovering the PC of a target, T, is to remove all non-PC nodes from the output of an incremental MB discovery algorithm, taking `InterIAPC` [10] for example. `InterIAPC` is pseudo-coded by (b) of Algorithm 1: It first calls `InterIAMB` to get the MB of T and then removes the spouses of T from the output. In the pseudo-code of `InterIAPC`, we suppose this algorithm can start learning with any particular set, $\boldsymbol{S}$, of potential PC nodes, while Morais and Aussem [10] started their `InterIAPC` from an empty set.

For these two different kinds of PC discovery techniques, as Aliferis et al. [8] and Morais and Aussem [10] argued, GLL-PC has an exponential complexity and thus it is time inefficient although it performs relatively well in data efficiency, while `InterIAPC` is data inefficient although it usually runs very fast. However, the assumption $\mathscr{A}_2$ implies that the AND operator used by PCMB and GLL-PC may lead to an over-high threshold for finding PC nodes. Therefore, GLL-PC is not suitable for the discovery of large PCs, just as Aliferis et al. [8, p. 217] pointed out in their paper. As a meta-procedure of PC discovery, the PCOR algorithm of Morais and Aussem [10] successfully applies the OR operator[†] (just like the idea of the local-to-global strategy) to combine the strategy of dividing-and-conquering that GLL-PC uses and the advantage of `InterIAPC`. PCOR is pseudo-coded by (c) of Algorithm 1.

For MB discovery, there are lots of independence-based approaches in the literature. Among them, PCMB [23], BFMB [37], GLL-MB [8], and the algorithm proposed by Khan et al. [38] are divide-and-conquer search techniques; these PC-based algorithms are data efficient. In contrast, those incremental non-PC-based algorithms such as KS [39] and GS [40,41] as well as some variants of GS including IAMB and `InterIAMB` [42] and the `Three-Fast-InterIAMB` [43] are far more time efficient (but also far less data efficient) than the PC-based algorithms. Here, `InterIAMB` is pseudo-coded by the subroutine of (b) in Algorithm 1, whereas GLL-MB is described by (d) of Algorithm 1. When the MB of a target variable T is not large, these algorithms are enough for practitioners to select features for T; when the MB is moderately large, the LRH algorithm [17] performs desirably; In the case of large MBs, the MBOR algorithm put forward by Morais and Aussem [10] is recommended; (e) of Algorithm 1 presents its pseudo-code.

As an extension of PCOR, the MBOR algorithm inherits the merits of the former. It tries to apply an ensemble technique to combine the advantages of both divide-and-conquer and incremental methods to improve accuracy and efficiency, especially on densely connected networks [10, p. 580]. MBOR [10] uses `InterIAMB` and `InterIAPC` as its subroutines; it overwhelmingly outperforms other existing MB discovery algorithms, especially for the case of large MBs. Here, we mention that MBOR may require some longer time to run than other existing algorithms. However, this flaw of MBOR is negligible comparing with its advantage in improving accuracy.

Recently, Liu et al. [44] put forward a novel algorithm called *fast shrinking parents-children learning for Markov blanket-based feature selection* (FSMB). Their simulation study reveals that the accuracy of MBOR and that of FSMB are generally comparable. Considering that MBOR can use

[†]The OR operator means that a node X is regarded as a PC member of T if $X \in PC_T$ or $T \in PC_X$ holds.

different PC discovery algorithms as its subroutine, we will choose MBOR as one of the algorithms for our simulation study.

Algorithm 1: Existing Local Discovery Algorithms

Procedure (a): $PC_T \leftarrow$ GLL-PC($\mathbb{A}_{TPC}, \mathcal{D}, T, \mathcal{S}, \mathcal{L}$)
Input: $\mathbb{A}_{TPC}$ is an algorithm used to find a tentative PC; $\mathcal{D}$ is a data set; T is a target; $\mathcal{S} \triangleq \{\boldsymbol{S}_X$ is a starting set: $X \in \boldsymbol{V}\}$; $\mathcal{L} \triangleq \{\boldsymbol{L}_X$ is a blacklist: $X \in \boldsymbol{V}\}$.
Output: The output is the PC of T.

$TPC_T \leftarrow \mathbb{A}_{TPC}(\mathcal{D}, T, \boldsymbol{S}_T, \boldsymbol{L}_T)$
foreach $X \in TPC_T$ **do**
  if $T \notin \mathbb{A}_{TPC}(\mathcal{D}, X, \boldsymbol{S}_X, \boldsymbol{L}_X)$ **then** $TPC_T \leftarrow TPC_T \setminus \{X\}$;
end
return $PC_T \leftarrow TPC_T$

Procedure (b): $[PC_T, MB_T] \leftarrow$ InterIAPC($\mathcal{D}, T, \boldsymbol{S}, \boldsymbol{L}$)
Input: $\boldsymbol{S}$ is a starting set; $\boldsymbol{L}$ is a blacklist.
Output: The output is the PC and MB of T.

$MB_T \leftarrow$ InterIAMB($\mathcal{D}, T, \boldsymbol{S}, \boldsymbol{L}$) and $TPC_T \leftarrow MB_T$
foreach $X \in TPC_T$ **do**
  if $\exists\, \boldsymbol{Z} \subseteq MB_T$ s.t. $T \perp\!\!\!\perp_{\mathcal{D}} X \mid \boldsymbol{Z}$ **then**
   $TPC_T \leftarrow TPC_T \setminus \{X\}$;
end
return $PC_T \leftarrow TPC_T$ and MB_T

//$MB_T \leftarrow$ InterIAMB($\mathcal{D}, T, \boldsymbol{S}, \boldsymbol{L}$)
$MB_T \leftarrow \boldsymbol{S}$ and $CanMB_T \leftarrow \boldsymbol{V} \setminus MB_T \setminus \{T\} \setminus \boldsymbol{L}$
while $CanMB_T \neq \emptyset$ **do**
  $Y \leftarrow \arg\max_{X \in CanMB} f_{\mathcal{D}}(T; X \mid MB_T)$
  if $T \not\perp\!\!\!\perp_{\mathcal{D}} Y \mid MB_T$ **then**
   $CanMB_T \leftarrow CanMB_T \setminus \{Y\}$ and $MB_T \leftarrow MB_T \cup \{Y\}$
  end
  foreach $X \in MB_T$ **do**
   if $T \perp\!\!\!\perp_{\mathcal{D}} X \mid MB_T \setminus \{X\}$ **then** $MB_T \leftarrow MB_T \setminus \{X\}$;
  end
end
return MB_T

Procedure (c): $PC_T \leftarrow$ PCOR($\mathbb{A}_{PC}, \mathcal{D}, T$)
Input: $\mathbb{A}_{PC}$ is an incremental PC discovery algorithm with the same input as $\mathbb{A}_{TPC}$.
Output: The output is the PC of T.

$PCS_T \leftarrow \boldsymbol{V} \setminus \{T\}$
foreach $X \in PCS_T$ **do**
  if $T \perp\!\!\!\perp_{\mathcal{D}} X$ **then** $PCS_T \leftarrow PCS_T \setminus \{X\}$ and $C_X \leftarrow \emptyset$;
end
foreach $X \in PCS_T$ **do**
  if $\exists\, Y \in PCS_T \setminus \{X\}$ s.t. $T \perp\!\!\!\perp_{\mathcal{D}} X \mid Y$ **then**
   $PCS_T \leftarrow PCS_T \setminus \{X\}$ and $C_X \leftarrow \{Y\}$
  end
end
$SPS_T \leftarrow \emptyset$
foreach $X \in PCS_T$ **do**
  $SPS_X \leftarrow \emptyset$
  foreach $Y \in \boldsymbol{V} \setminus PCS_T \setminus \{T\}$ **do**
   if $T \not\perp\!\!\!\perp_{\mathcal{D}} Y \mid C_Y \cup \{X\}$ **then** $SPS_X \leftarrow SPS_X \cup \{Y\}$;
  end
  foreach $Y \in SPS_X$ **do**
   if $\exists\, Z \in SPS_X \setminus \{Y\}$ s.t. $T \perp\!\!\!\perp_{\mathcal{D}} Y \mid \{X, Z\}$ **then**
    $SPS_X \leftarrow SPS_X \setminus \{Y\}$
   end
  end
  $SPS_T \leftarrow SPS_T \cup SPS_X$
end
$MBS_T \leftarrow PCS_T \cup SPS_T$
$PC_T \leftarrow \mathbb{A}_{PC}(\mathcal{D}, T, \emptyset, \boldsymbol{V} \setminus MBS_T)$
foreach $X \in PCS_T \setminus PC_T$ **do**
  if $T \in \mathbb{A}_{PC}(\mathcal{D}, X, \emptyset, \emptyset)$ **then** $PC_T \leftarrow PC_T \cup \{X\}$;
end
return PC_T

Procedure (d): $MB_T \leftarrow$ GLL-MB($\mathbb{A}_{TPC}, \mathcal{D}, T, \mathcal{S}, \mathcal{L}$)
Input: The same as GLL-PC.
Output: The output is the MB of T.

$PC_T \leftarrow$ GLL-PC($\mathbb{A}_{TPC}, \mathcal{D}, T, \boldsymbol{S}_T, \boldsymbol{L}_T$)
foreach $Y \in PC_T$ **do**
  $PC_Y \leftarrow$ GLL-PC($\mathbb{A}_{TPC}, \mathcal{D}, Y, \boldsymbol{S}_Y, \boldsymbol{L}_Y$)
end
$TMB_T \leftarrow PC_T$ and $TSP_T \leftarrow (\cup_{Y \in PC_T} PC_Y) \setminus PC_T \setminus \{T\}$
foreach $X \in TSP_T$ **do**
  find $\boldsymbol{Z}$ s.t. $T \perp\!\!\!\perp_{\mathcal{D}} X \mid \boldsymbol{Z}$
  foreach $Y \in PC_T$ s.t. $X \in PC_Y$ **do**
   if $T \not\perp\!\!\!\perp_{\mathcal{D}} X \mid \boldsymbol{Z} \cup \{Y\}$ **then** $TMB_T \leftarrow TMB_T \cup \{X\}$;
  end
end
return $MB_T \leftarrow TMB_T$

Procedure (e): $MB_T \leftarrow$ MBOR($\mathbb{A}_{PC}, \mathcal{D}, T$)
Output: The output is the MB of T.

$PC_T \leftarrow$ PCOR($\mathbb{A}_{PC}, \mathcal{D}, T$) and $SP_T \leftarrow \emptyset$
foreach $X \in PC_T$ **do**
  foreach $Y \in \mathbb{A}_{PC}(\mathcal{D}, X, \emptyset, \emptyset) \setminus PC_T \setminus \{T\}$ **do**
   find minimal $\boldsymbol{Z} \subseteq MBS_T \setminus \{T, Y\}$ s.t. $T \perp\!\!\!\perp_{\mathcal{D}} Y \mid \boldsymbol{Z}$
   if $T \not\perp\!\!\!\perp_{\mathcal{D}} Y \mid \boldsymbol{Z} \cup \{X\}$ **then**
    $SP_T \leftarrow SP_T \cup \{Y\}$
   end
  end
end
return $MB_T \leftarrow PC_T \cup SP_T$

4.2. Information connection based method for local discovery

Premature termination of forward search (i.e., $\mathscr{P}_2$) is a potential problem for independence-based algorithms of local discovery (especially for large cases), and it is quite meaningful and challenging to seek an effective solution to this problem. A well-known idea is to use the "divide-and-conquer" strategy (done by `PCMB` and `GLL`) that tries to reduce the conditional set size as much as possible [10]. This idea plays an important role in exploring more new algorithms (such as `PCOR` and `MBOR`) for local discovery and, indeed, leads to great improvements on data efficiency. However, this idea still cannot solve the problem $\mathscr{P}_2$ desirably for large local discovery (see [8, p. 217] for a similar argument), and thus it is necessary to develop new approaches to further improve the learning accuracy.

Before presenting the main idea of our method, we first give an example as follows:

Example 2. *Consider a BN* $(\mathbb{G}, \mathbb{P})$ *over* $\boldsymbol{V} \triangleq \{T,\ X_1, X_2,\ Y_1, \cdots, Y_4,\ Z_1\}$*, in which* $\mathbb{G}$ *is presented in Figure 3. Take* T *as the target with* $PC_T = \{Y_2, Y_3, Z_1\}$ *and* $MB_T = \{Y_2, Y_3, Y_4, Z_1\}$*. Note that there are more than one information channels between* Y_1 *and* T*. Let* $PC_T^{\mathbb{A}} \triangleq \{Y_1, Y_2, Y_3\}$ *and* $MB_T^{\mathbb{A}} \triangleq \{Y_1, \cdots, Y_4\}$ *be the best outputs*[‡] *of the existing local discovery algorithms except* `PCOR` *and* `MBOR`*. Specifically,* `GLL` *finds* $PC_T^{\mathbb{A}}$ *as the PC of* T *and then adds* Y_4 *to* $PC_T^{\mathbb{A}}$ *to return* $MB_T^{\mathbb{A}}$*, while* `InterIAPC` *first discovers* $MB_T^{\mathbb{A}}$ *and then removes* Y_4 *from* $MB_T^{\mathbb{A}}$ *to derive* $PC_T^{\mathbb{A}}$*. The TP,* Z_1*, is excluded because of the premature inclusion of the FP,* Y_1*, and the potential insufficiency of the data* $\mathcal{D}$*. This is the direct consequence of the problem* $\mathscr{P}_2$*. In comparison,* `PCOR` *and* `MBOR` *may detect* Z_1*; however, they may not exclude the FP* Y_1 *due to their partial inefficiency inheriting from the subroutines,* `InterIAMB` *and* `InterIAPC`*.*

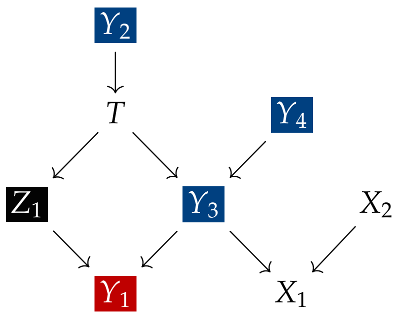


Figure 3. A BN over $\boldsymbol{V} \triangleq \{T,\ X_1, X_2,\ Y_1, \cdots, Y_4,\ Z_1\}$, with $PC_T = \{Y_2, Y_3, Z_1\}$ and $MB_T = \{Y_2, Y_3, Y_4, Z_1\}$.

This example illustrates the following two facts:

- Although the divide-and-conquer strategy usually improves the data efficiency substantially, the resulting algorithms may still suffer from the consequences of the problem $\mathscr{P}_2$.
- As two meta-procedures, `PCOR` and `MBOR` may possess overwhelming advantages over other existing algorithms in local discovery (especially in large local discovery); however, these two algorithms still inherit some shortcomings of their subroutines.

[‡]Specifically, for a certain PC (or MB) algorithm, if after a certain forward search step, the tentative PC of the target T is obtained as $\boldsymbol{U} = \{Y_2, Y_3, Y_1\}$, in which Y_2 and Y_3 are the true PC nodes of T, while Y_1 is not. Y_1 enters $\boldsymbol{U}$ because there are two information channels from T to Y_1 and therefore Y_1 can carry more information about T than Z_1, which is actually a true PC node. Next, to examine if Z_1 can enter $\boldsymbol{U}$, we need to perform the CI test: $T \perp\!\!\!\perp Z_1 \mid \boldsymbol{U}$. At this time, due to the possible insufficiency of the data to allow Z_1 to carry the most of remaining information about T, this CI test may incorrectly give a conclusion of $T \perp\!\!\!\perp_{\mathcal{D}} Z_1 \mid \boldsymbol{U}$. Furthermore, as Z_1 does not enter $\boldsymbol{U}$, the false PC node Y_1 can not be excluded from $\boldsymbol{U}$ in subsequent CI tests until the end of the algorithm.

In brief, it is very attractive to build an effective subroutine for PCOR and MBOR. In the following, we focus on this issue for these two algorithms.

In (c) and (e) of Algorithm 1, we replace InterIAPC in PCOR and MBOR with any particular PC discovery algorithm, $\mathbb{A}_{PC}$. This replacement is theoretically feasible in practice, provided $\mathbb{A}_{PC}$ is also time efficient (so GLL-PC is not suitable for this role); besides InterIAPC, one can choose $\mathbb{A}_{PC}$ by applying the refining procedure of InterIAPC to any other non-PC-based MB discovery algorithms (such as GS and IAMB). Naturally, we expect $\mathbb{A}_{PC}$ can output as few FPs and as many TPs as possible. However, all of its above existing alternatives (even GLL-PC) severely suffer the problem $\mathscr{P}_2$. In what follows, we propose a novel method of solving this problem and use it to enhance $\mathbb{A}_{PC}$ in PCOR and MBOR.

To clearly describe how we deal with the problem $\mathscr{P}_2$, we quote the following *information flow metaphor* that Cheng et al. [24] used: A BN can be viewed as a network of information channels or pipelines, where each node is an in-out valve that is either active (when the corresponding node is not instantiated) or inactive (when instantiated), the valves are connected by noisy information channels (edges), and the information can flow *through* an active valve but not an inactive one.

The metaphor of Cheng et al. [24] says that performing the CI test for the hypothesis "$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$" is equivalent to observing the information flow between $\boldsymbol{X}$ and $\boldsymbol{Y}$ when instantiating the nodes within $\boldsymbol{Z}$ (i.e., inactivating the corresponding valves). However, it works only under the assumption $\mathscr{M}_1$. Now, we follow Theorem 1 to extend this metaphor to the assumption $\mathscr{M}_2$ by regarding each node as a valve that possesses some resistance and adding a *data-driven force* (or called "energy") of propagating information to the network, and assume:

i) Each node is an in-out valve, and each valve has a different resistance. More precisely, the resistance of a valve is increasing with the *number of configurations* (i.e., number of degrees of freedom) for the corresponding node.
ii) The driving force is increasing with the data size.
iii) Inactivating any valve will consume some driving force. More precisely, the amount of consumed energy increases with the resistance of valves that are inactivated. Further, when some valves are inactivated, if the remaining driving force is not sufficient, it may not be possible to further inactivate additional valves.

For convenience, we call this extension to be the *extended information flow metaphor*. In what follows, for convenience, we use $\mathscr{M}_1$ and $\mathscr{M}_2$, respectively, to denote the information flow metaphor of [24] and our extended metaphor.

In our metaphor $\mathscr{M}_2$, the terminology "data-driven force" coincides mathematically with the power of the $\not\perp\!\!\!\perp_{\mathcal{D}}$-tests in some sense. Such an explanation combined with Theorem 1 reveals the reasonability of $\mathscr{M}_2$ as well as the inappropriateness of $\mathscr{M}_1$. In fact, $\mathscr{M}_1$ means

a) information cannot be propagated without driving force, and
b) information can be transmitted to any reachable node if there is driving force,

while $\mathscr{M}_2$ means

a') information cannot be propagated without *sufficient* driving force, and
b') information can be transmitted to any reachable node if there is *sufficient* driving force.

Note that (a') and (b') are slightly different from (a) and (b). Further, $\mathscr{M}_2$ indicates

c) information can only be transmitted to some (not all) of the reachable nodes if no sufficient driving force is provided.

As seen, these intuitive descriptions about the metaphors $\mathscr{M}_1$ and $\mathscr{M}_2$ coincide with the assumptions $\mathscr{A}_1$ and $\mathscr{A}_2$, respectively. The following remark applies our metaphor $\mathscr{M}_2$ to explain the result of Example 2.

Remark 1. *Consider Example 2 again. If the data $\mathcal{D}$ is large enough, the information from T can flow to every theoretically reachable valve even when inactivating any other valves such that only one information channel is left. However, if $\mathcal{D}$ is not large enough, the data-driven force may not be sufficient for propagating the information to every theoretically reachable valve when inactivating some valves. Specifically, after making $\{Y_1, Y_2, Y_3, Y_4\}$ inactive, the remainder driving force may become insufficient for transmitting the information from T to Z_1. In this case, the result of Example 2 follows.*

This remark explains about the essence of the problem $\mathscr{P}_2$ (i.e., premature termination of the forward search). In the meanwhile, Remark 1 hints two ways of solving $\mathscr{P}_2$ as follows: One is to increase the amount of data instances such that sufficient driving force can be supplied; and the other is to inactivate as few valves as possible such that there is sufficient driving force used to convey the information from the target, T, to a particular valve which is theoretically reachable (from T). In general, the former is impractical while the latter is feasible, so we need only to consider the latter way. Two methods can be employed to achieve the goal of this way: (I) One is the "divide-and-conquer" strategy that has been widely used in the literature [8–10, 23], but it cannot solve the problem $\mathscr{P}_2$ desirably for large local discovery as Example 2 illustrates; and (II) the other is to purposefully cancel to inactivate some of the valves such that more other valves can receive the information from the target T (i.e., more potential TPs can be identified). We will call (II) the *information connection* based method (IC), in which the purpose of cancelling to inactivate some valves is to save some driving force such that the saved driving force can be used to enhance the transmission of information.

The main idea of IC instantiated in PC discovery is as follows:

(i) *Preliminary discovery*: First, we employ a particular PC discovery algorithm, $\mathbb{A}_{PC}$, to obtain a coarse PC of T, denoted by $PC_T^{\mathbb{A}}$. Here, $\mathbb{A}_{PC}$ can be any PC discovery algorithm; however, those with time efficiency and with as high data efficiency as possible are preferred, considering that $\mathbb{A}_{PC}$ is only a subroutine of IC and it may be used repeatedly. Besides `InterIAPC`, Subsection 4.4 will provide an alternative for $\mathbb{A}_{PC}$ by combining the advantages of `GLL-PC` and `InterIAPC`. In this phase, a coarse MB of T denoted by $MB_T^{\mathbb{A}}$ is also outputted.

(ii) *Enhanced forward search*: This phase is the kernel of IC. To detect more theoretically reachable valves (other than those in $PC_T^{\mathbb{A}}$), we may cancel to inactivate some of the valves in $PC_T^{\mathbb{A}}$; these valves can be any subset of $PC_T^{\mathbb{A}}$, but we will take them to be all the single-point subsets considering the time efficiency. Algorithmically, this phase contains three subphases: (ii-a) *Extended forward search*: For each $Y \in PC_T^{\mathbb{A}}$, we use $\mathbb{A}_{PC}$ to get a new PC and a new MB of T, denoted by $PC_T^{(Y)}$ and $MB_T^{(Y)}$, respectively, by moving Y to the blacklist temporarily and starting with $PC_T^{\mathbb{A}} \setminus \{Y\}$. (ii-b) *Refining procedure*: Then, we use Y to refine $MB_T^{(Y)}$ and $PC_T^{(Y)}$. (ii-c) *Remedying procedure*: Finally, interleave a remedying procedure into (ii-b). After that, unite $PC_T^{(Y)}$ and $MB_T^{(Y)}$, respectively, to update $PC_T^{\mathbb{A}}$ and $MB_T^{\mathbb{A}}$.

(iii) *Backward search*: This phase removes those redundant nodes in $\mathtt{PC}_T^{\mathbb{A}}$ based on $\mathtt{MB}_T^{\mathbb{A}}$ without needing the pruning procedure (i.e., symmetry correction; cf. [7]) that is required in PCMB and GLL-PC.

The resulting algorithm, called ICPC, is pseudo-coded in Algorithm 2. In this algorithm, Line 1 carries out the *preliminary discovery*; Line 3, Line 5/Line 11, and Line 7 accomplish the three sub-phases of the *enhanced forward search*; and Line 14 performs the final *backward search*.

In what follows, we give an example to illustrate how our ICPC algorithm works.

Algorithm 2: ICPC

Procedure: $\mathtt{PC}_T \leftarrow \mathtt{ICPC}(\mathbb{A}_{\mathtt{PC}}, \mathcal{D}, T, \boldsymbol{S}, \boldsymbol{L})$

Input: $\mathbb{A}_{\mathtt{PC}}$ is an incremental PC discovery algorithm (the same as in PCOR); $\mathcal{D}$ is a data set; T is a target; $\boldsymbol{S}$ is a starting set of variables; $\boldsymbol{L}$ is a blacklist.

Output: The output is the PC of T.

1 $[\mathtt{PC}_T, \mathtt{MB}_T] \leftarrow \mathbb{A}_{\mathtt{PC}}(\mathcal{D}, T, \boldsymbol{S}, \boldsymbol{L})$ and $\mathcal{X} \leftarrow \emptyset$; // (i) *preliminary discovery*
2 **foreach** $Y \in \mathtt{PC}_T$ **do**
3 $\quad [\mathtt{PC}_T^{(Y)}, \mathtt{MB}_T^{(Y)}] \leftarrow \mathbb{A}_{\mathtt{PC}}(\mathcal{D}, T, \mathtt{PC}_T \setminus \{Y\}, Y)$; // (ii-a) *extended forward search*
4 $\quad$ **while** $\exists\, X \in \mathtt{MB}_T^{(Y)}$ s.t. $T \perp\!\!\!\perp X \mid (\mathtt{MB}_T^{(Y)} \setminus \{X\}) \cup \{Y\}$ **do**
5 $\quad\quad \mathtt{MB}_T^{(Y)} \leftarrow \mathtt{MB}_T^{(Y)} \setminus \{X\};\ \mathtt{PC}_T^{(Y)} \leftarrow \mathtt{PC}_T^{(Y)} \setminus \{X\};$ and $\mathcal{X} \leftarrow \mathcal{X} \cup \{X\}$; // (ii-b) *refining procedure*
6 $\quad\quad$ **while** $\exists\, X' \in \mathcal{X} \setminus \{X\}$ s.t. $T \not\perp\!\!\!\perp X' \mid \mathtt{MB}_T^{(Y)} \cup \{Y\}$ **do**
7 $\quad\quad\quad \mathtt{MB}_T^{(Y)} \leftarrow \mathtt{MB}_T^{(Y)} \cup \{X'\};\ \mathtt{PC}_T^{(Y)} \leftarrow \mathtt{PC}_T^{(Y)} \cup \{X'\};$ and $\mathcal{X} \leftarrow \mathcal{X} \setminus \{X'\}$; // (ii-c) *remedying procedure*
8 $\quad\quad$ **end**
9 $\quad$ **end**
10 **end**
11 $\mathtt{PC}_T \leftarrow [(\cup_{Y \in \mathtt{PC}_T} \mathtt{PC}_T^{(Y)}) \cup \mathtt{PC}_T] \backslash \mathcal{X}$ and $\mathtt{MB}_T \leftarrow [(\cup_{Y \in \mathtt{PC}_T} \mathtt{MB}_T^{(Y)}) \cup \mathtt{MB}_T] \backslash \mathcal{X}$; // (ii-b) *refining procedure* (continued)
12 **foreach** $X \in \mathtt{PC}_T$ **do**
13 $\quad$ **if** $\exists\, \boldsymbol{Z} \subseteq \mathtt{MB}_T$ s.t. $T \perp\!\!\!\perp X \mid \boldsymbol{Z}$ **then**
14 $\quad\quad \mathtt{PC}_T \leftarrow \mathtt{PC}_T \setminus \{X\}$; // (iii) *backward search*
15 $\quad$ **end**
16 **end**
17 **return** $\mathtt{PC}_T$

Example 3. *[Information Connection] Consider the BN in Example 2 again. Let $\mathtt{PC}_T^{\mathbb{A}} \triangleq \{Y_1, Y_2, Y_3\}$ be a coarse PC of T companied by $\mathtt{MB}_T^{\mathbb{A}} \triangleq \{Y_1, \cdots, Y_4\}$ as a coarse MB of T. The TP Z_1 has not been identified because of the insufficiency of data and the fact that the FP Y_1 may collect too much information of T that Z_1 and Y_3 have. Now, we apply IC to every $Y \in \mathtt{PC}_T^{\mathbb{A}}$, taking Y_3 for a typical example. When Y_3 is temporarily deemed hidden (and thus removed from the conditional set), there are two possible consequences:*

(a) Detection of more TPs: Some driving force is saved such that the enhanced propagation of information can reach one or more other TPs than the ones in $\mathtt{PC}_T^{\mathbb{A}} \setminus \{Y_3\}$. Specifically, the TP, Z_1, may be detected and thus enters $\mathtt{PC}_T^{(Y_3)}$. In this case, the problem $\mathscr{P}_2$ gets alleviated to a certain degree.

(b) Addition of redundant variables: Information is propagated to the members of $\mathtt{MB}_{Y_3} \setminus \{T\}$ along the

paths in which Y_3 is a head-to-tail (HT) node or a tail-to-tail (TT) node. Specifically, we have

$$\text{“}T \to Y_3 \to Y_1\text{”} \;\Rightarrow\; \text{“}T \to Y_1\text{”}, \qquad \text{“}Y_4 \to Y_3 \to Y_1\text{”} \;\Rightarrow\; \text{“}Y_4 \to Y_1\text{”},$$
$$\text{“}T \to Y_3 \to X_1\text{”} \;\Rightarrow\; \text{“}T \to X_1\text{”}, \qquad \text{“}Y_4 \to Y_3 \to X_1\text{”} \;\Rightarrow\; \text{“}Y_4 \to X_1\text{”}.$$

In the meanwhile, the collision “$T \to Y_3 \leftarrow Y_4$” is decomposed into “$T \to Y_1 \leftarrow Y_4$” and “$T \to X_1 \leftarrow Y_4$”, meaning that the variables in $\mathtt{CH}_Y$ play the same role as Y_3 in a path where Y_3 is a head-to-head (HH) node. Figure 4 illustrates such an evolution. Thus, X_1 enters $\mathtt{PC}_T^{(Y_3)}$ while X_2 enters $\mathtt{MB}_T^{(Y_3)}$.

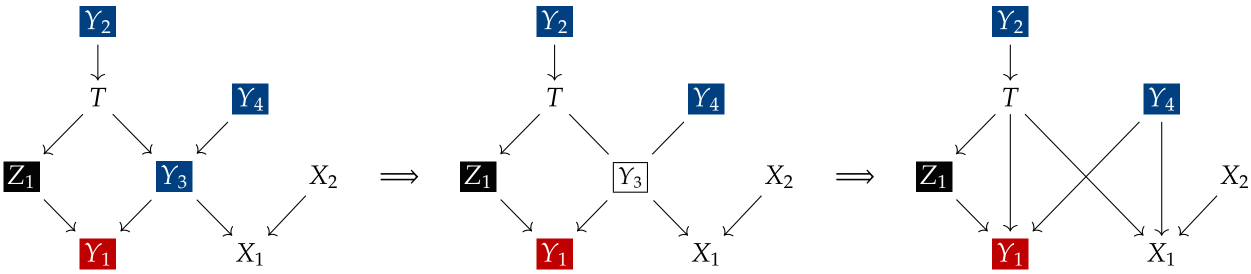


Figure 4. An illustration on the case of addition of some redundant variables.

According to the above analysis, we assume to get

$$\mathtt{PC}_T^{(Y_3)} = \{Y_1, Y_2, Z_1, X_1\} \quad \text{and} \quad \mathtt{MB}_T^{(Y_3)} = \{Y_1, Y_2, Y_4, Z_1, X_1, X_2\}.$$

Next, we use Y_3 to refine $\mathtt{MB}_T^{(Y_3)}$ and $\mathtt{PC}_T^{(Y_3)}$ via Line 5 of Algorithm 2. This procedure is also a key to our $\mathtt{ICPC}$ algorithm. First, the parent node Y_2 is unfortunately excluded due to the insufficiency of data. Second, in view of the fact that “Y_3 d-separates T and $\{X_1, X_2\}$”[§], the two redundant variables, X_1 and X_2, added in the previous step can be excluded timely by associated CI tests. Third, by $T \perp Y_1 \mid \{Z_1, Y_3\}$ (since $\{Z_1, Y_3\}$ d-separates T and Y_1) and, therefore, $T \perp\!\!\!\perp Y_1 \mid \{Z_1, Y_3\}$, the FP, Y_1, having spuriously high association with T can also be identified immediately. Hence, Y_2, X_1, X_2, and Y_1 enter $\mathcal{X}$ after the refining procedure is performed. Here, note that Y_2 is a true PC/MB member of T. Fortunately, the exclusion of $\{X_1, X_2, Y_1\}$ results in sufficiency of data, so the subsequent remedying procedure helps Y_2 re-enter $\mathtt{PC}_T^{(Y_3)}$ and $\mathtt{MB}_T^{(Y_3)}$, and it updates $\mathcal{X}$ to $\{X_1, X_2, Y_1\}$.

In summary, $\mathtt{PC}_T^{(Y_3)} = \{Y_2, Z_1\}$ and $\mathtt{MB}_T^{(Y_3)} = \{Y_2, Z_1, Y_4\}$. As seen, X_1, X_2, and Y_1 are permanently excluded in this process. In a similar fashion, we apply IC to Y_1 and Y_2, assuming to get $\mathtt{PC}_T^{(Y_1)} = \{Y_2, Y_3, Z_1\}$, $\mathtt{MB}_T^{(Y_1)} = \{Y_2, Y_3, Z_1, Y_4\}$; and $\mathtt{PC}_T^{(Y_2)} = \{Y_3, Z_1\}$, $\mathtt{MB}_T^{(Y_2)} = \{Y_3, Z_1, Y_4\}$. Then, using them to update

$$\mathtt{PC}_T^{\mathbb{A}} \leftarrow [(\cup_{Y \in \mathtt{PC}_T^{\mathbb{A}}} \mathtt{PC}_T^{(Y)}) \cup \mathtt{PC}_T^{\mathbb{A}}] \backslash \mathcal{X} = \{Y_2, Y_3, Z_1\} \;\; \text{and}$$

[§]For the concept of *d-separation*, the readers can refer to [1,2] for the details. Appendix A.2 also provides a concise explanation: for a BN defined on a set of nodes, $\boldsymbol{V}$, we say $\boldsymbol{Z}$ d-separates $\boldsymbol{X}$ and $\boldsymbol{Y}$ ($\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subset \boldsymbol{V}$), if $\boldsymbol{Z}$ blocks every path between $\boldsymbol{X}$ and $\boldsymbol{Y}$; and if this is the case we write $\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}$. Here, $\boldsymbol{Z}$ blocking a path $\mathbb{p}$ means that $\mathbb{p}$ has a head-to-tail node or a tail-to-tail node belonging to $\boldsymbol{Z}$, or that $\mathbb{p}$ has a head-to-head node C such that C and its all descendants are not in $\boldsymbol{Z}$. It is well known that $\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} \Rightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ holds in any case, and vice versa if the BN satisfies the faithfulness condition. As seen, Y_3 is a head-to-tail node of the path “$T \to Y_3 \to X_1$” and also a tail-to-tail node of the path “$T \to Z_1 \to Y_1 \leftarrow Y_3 \to X_1$”, indicating $T \perp X_1 \mid Y_3$; and X_1 ($\notin \{Y_3\}$) is a head-to-head node of the two pathes from T to X_2, implying $T \perp X_2 \mid Y_3$.

$$\mathtt{MB}_T^{\mathbb{A}} \leftarrow [(\cup_{Y \in PC_T^{\mathbb{A}}} \mathtt{MB}_T^{(Y)}) \cup \mathtt{MB}_T^{\mathbb{A}}] \backslash \mathcal{X} = \{Y_2, Y_3, Z_1, Y_4\},$$

respectively. Finally, the backward search does not remove any node and returns $\mathtt{PC}_T^{\mathbb{A}} = \{Y_2, Y_3, Z_1\}$*. As seen,* $\mathtt{PC}_T^{\mathbb{A}}$ *contains no redundant variables, so* `ICPC` *correctly outputs the PC of* T*.*

To be more intuitive, we summarize the above process in Table 2, where the operations are listed in pseudo-code order of Algorithm 2.

Table 2. Detailed operations (in pseudo-code order of Algorithm 2) of `ICPC` for discovering the PC, $\{Y_2, Y_3, Z_1\}$, of T for the network presented in Figure 3.

Phase	Result
(i) preliminary discovery	$PC_T^{\mathbb{A}} \leftarrow \{Y_1, Y_2, Y_3\},\ MB_T^{\mathbb{A}} \leftarrow \{Y_1, Y_2, Y_3, Y_4\},\ \mathcal{X} \leftarrow \emptyset$
	$Y = Y_1 \xRightarrow{\text{(ii-a)}} \begin{cases} PC_T^{(Y)} \leftarrow \{Y_2, Y_3, Z_1\} \\ MB_T^{(Y)} \leftarrow \{Y_2, Y_3, Z_1, Y_4\} \\ \mathcal{X} \leftarrow \emptyset \end{cases} \xRightarrow{\text{(ii-b)}} \begin{cases} PC_T^{(Y)} \leftarrow \{Y_2, Y_3, Z_1\} \\ MB_T^{(Y)} \leftarrow \{Y_2, Y_3, Z_1, Y_4\} \\ \mathcal{X} \leftarrow \emptyset \end{cases} \xRightarrow{\text{(ii-c)}} \begin{cases} PC_T^{(Y)} \leftarrow \{Y_2, Y_3, Z_1\} \\ MB_T^{(Y)} \leftarrow \{Y_2, Y_3, Z_1, Y_4\} \\ \mathcal{X} \leftarrow \emptyset \end{cases}$
(ii-a) extended forward search (ii-b) refining procedure (ii-c) remedying procedure	$Y = Y_2 \xRightarrow{\text{(ii-a)}} \begin{cases} PC_T^{(Y)} \leftarrow \{Y_3, Z_1\} \\ MB_T^{(Y)} \leftarrow \{Y_3, Z_1, Y_4\} \\ \mathcal{X} \leftarrow \emptyset \end{cases} \xRightarrow{\text{(ii-b)}} \begin{cases} PC_T^{(Y)} \leftarrow \{Y_3, Z_1\} \\ MB_T^{(Y)} \leftarrow \{Y_3, Z_1, Y_4\} \\ \mathcal{X} \leftarrow \emptyset \end{cases} \xRightarrow{\text{(ii-c)}} \begin{cases} PC_T^{(Y)} \leftarrow \{Y_3, Z_1\} \\ MB_T^{(Y)} \leftarrow \{Y_3, Z_1, Y_4\} \\ \mathcal{X} \leftarrow \emptyset \end{cases}$
	$Y = Y_3 \xRightarrow{\text{(ii-a)}} \begin{cases} PC_T^{(Y)} \leftarrow \{Y_1, Y_2, Z_1, X_1\} \\ MB_T^{(Y)} \leftarrow \{Y_1, Y_2, Y_4, Z_1, X_1, X_2\} \\ \mathcal{X} \leftarrow \emptyset \end{cases} \xRightarrow{\text{(ii-b)}} \begin{cases} PC_T^{(Y)} \leftarrow \{Z_1\} \\ MB_T^{(Y)} \leftarrow \{Y_4, Z_1\} \\ \mathcal{X} \leftarrow \{Y_2, X_1, X_2, Y_1\} \end{cases} \xRightarrow{\text{(ii-c)}} \begin{cases} PC_T^{(Y)} \leftarrow \{Z_1, Y_2\} \\ MB_T^{(Y)} \leftarrow \{Y_4, Z_1, Y_2\} \\ \mathcal{X} \leftarrow \{X_1, X_2, Y_1\} \end{cases}$
(ii-b) refining procedure (continued)	$PC_T^{\mathbb{A}} \leftarrow \left[\left(\cup_{Y \in PC_T^{\mathbb{A}}} PC_T^{(Y)}\right) \cup PC_T^{\mathbb{A}}\right] \Big\backslash \mathcal{X} = \{Y_2, Y_3, Z_1\},\ MB_T^{\mathbb{A}} \leftarrow \left[\left(\cup_{Y \in PC_T^{\mathbb{A}}} MB_T^{(Y)}\right) \cup MB_T^{\mathbb{A}}\right] \Big\backslash \mathcal{X} = \{Y_2, Y_3, Z_1, Y_4\}$
(iii) backward search	$PC_T^{\mathbb{A}} \leftarrow \{Y_2, Y_3, Z_1\}$

This example reveals that the `ICPC` algorithm may capture as many TPs as possible (thus, as much information as possible about the target). Consequently, as few FPs as possible can remain undetected by the end of the *enhanced forward search* (up to Line 11 of Algorithm 2). In this intuitive sense, `ICPC` may be as a desirable selection of $\mathbb{A}_{PC}$ in `PCOR` and `MBOR`. The resulting algorithms are expected to perform well in large local discovery.

4.3. Theoretical basis of `ICPC`

We introduced a novel algorithm, `ICPC`, to deal with the problem $\mathscr{P}_2$ in Subsection 4.2. Its working mechanism is inspired by the extended information flow metaphor, $\mathscr{M}_2$, and thus is intuitively reasonable. In what follows, we explain its theoretical soundness in making the *enhanced forward search*. Specifically, we show: (a) Why the *extended forward search* (Line 3 of Algorithm 2) can capture more TPs; (b) why we need the *refining procedure* (Line 5 and Line 11 of Algorithm 2) and why it can remove FPs effectively; (c) why we need the *remedying procedure* (Line 7 of Algorithm 2).

We first define the output of the enhanced forward search as follows:

Definition 1. *[Information Connection] For* $T, Y \in \boldsymbol{V}$*, we call* $\boldsymbol{M} \subseteq \boldsymbol{V} \setminus \{T, Y\}$ *a Y-extended Markov boundary (Y-EMB) of* T *if it is an MB of* T *in* $\boldsymbol{V} \setminus \{Y\}$*. Denote it by* $\mathtt{MB}_T^{(Y)}$*. That is,* $T \perp\!\!\!\perp (\boldsymbol{V} \setminus \{Y\}) \setminus \mathtt{MB}_T^{(Y)} \setminus \{T\} \mid \mathtt{MB}_T^{(Y)}$*, in which* $\mathtt{MB}_T^{(Y)}$ *cannot be replaced with its any proper subset.* □

The following theorem characterizes the structure of EMB; the proof is given in Appendix A.2.

Theorem 3. *For a BN* $(\mathbb{G}, \mathbb{P})$ *over* $\boldsymbol{V}$ *satisfying the faithfulness condition, the following statements hold:*

- *In one of the following cases:* (i) $Y \in PA_T$*,* (ii) $Y \in CH_T$ *with* $CH_Y \neq \emptyset$*,* (iii) $Y \in SP_T$*, the Y-EMB of T can be expressed as* $MB_T^{(Y)} = (MB_T \cup MB_Y) \setminus \{T, Y\}$*.*
- *If* $Y \in CH_T$ *with* $CH_Y = \emptyset$*, then* $PC_T \setminus \{Y\} \subseteq MB_T^{(Y)} \subseteq MB_T \setminus \{Y\}$*.*
- *If* $Y \notin MB_T$*, then* $MB_T^{(Y)} = MB_T$*.*

This theorem indicates the uniqueness of EMB under the faithfulness condition. By means of this result, we show the following theorem in Appendix A.3:

Theorem 4. *For a BN* $(\mathbb{G}, \mathbb{P})$ *over* $\boldsymbol{V}$ *satisfying the faithfulness condition, let* $MB_T^{(Y)}$ *be the Y-EMB of T, and* $\boldsymbol{M} \subseteq MB_T^{(Y)}$ *subject to* $(MB_T^{(Y)} \setminus \boldsymbol{M}) \cap MB_T = \emptyset$*. Then, for any* $X \in \boldsymbol{M}$*, we have* $X \notin MB_T \Leftrightarrow T \perp\!\!\!\perp X \mid (\boldsymbol{M} \setminus \{X\}) \cup \{Y\}$*.*

By Theorems 3 and 4 in conjunction with the pseudo-code of ICPC, it follows that:

(a) *Why the extended forward search can capture more TPs:* On the one hand, the forward search of every existing PC discovery algorithm, $\mathbb{A}_{PC}$, will be prematurely terminated when the dataset is insufficient, so the coarse PC of T returned by $\mathbb{A}_{PC}$ (pseudo-coded in Line 1 of Algorithm 2) may be undesirable in practice. On the other hand, as illustrated in Remark 1 and as shown by Theorem 3, the extended forward search may lead to the *detection of more TPs*. If so, some PC members swamped by $MB_T^{(Y)}$ due to insufficiency of data will be identified, and thus the problem $\mathscr{P}_2$ about the premature termination of the forward search can get solved or alleviated.

(b) *Why we need the refining procedure* and *why it can remove FPs:* Although the extended forward search can capture more TPs, it may also lead to another consequence: *Addition of redundant variables*. These redundant variables will increase the computational cost of the final backward search to a large extent; they may also increase the possibility of excluding some TPs from PC_T due to the unreliability of some CI tests in practical situations. This explains the importance of doing a refining procedure before making the final backward search. More importantly, Theorem 4 reveals that the refining procedure pseudo-coded in Line 5 and 11 of Algorithm 2 can remove *all and only* redundant variables (in any case) that enter $MB_T^{(Y)}$ and $PC_T^{(Y)}$ in the extended forward search.

In brief, ICPC can detect some more true members without shielding any redundant variables. In other words, while all existing PC discovery algorithms may prematurely terminate the forward search and thus fail to be used for large local discovery, our ICPC algorithm can selectively enhance the forward search and is expected to identify as many TPs as possible and as few FPs as possible.

Finally, we use the following theorem (proven in Appendix A.4) to explain the subsequent issue (c).

Theorem 5. *For* $T \in \boldsymbol{V}$ *and* $\boldsymbol{M} \subseteq \boldsymbol{V} \setminus \{T\}$*, put* $\mathcal{X}_\ell \triangleq \{X_1, \cdots, X_\ell\} \subseteq \boldsymbol{M}$ *and* $\boldsymbol{M}_\ell \triangleq \boldsymbol{M} \setminus \mathcal{X}_\ell$*, in which each* X_ℓ *is subject to the* $\perp\!\!\!\perp_{\mathcal{D}}$*-test "*$T \perp\!\!\!\perp_{\mathcal{D}} X_\ell \mid \boldsymbol{M}_\ell$*",* $\ell = 1, \cdots, k$*. Then, for any* $X_i \in \mathcal{X}_{k-1}$*, the* $\perp\!\!\!\perp_{\mathcal{D}}$*-test "*$T \perp\!\!\!\perp_{\mathcal{D}} X_i \mid \boldsymbol{M}_i$*" is unreliable under the assumption* $\mathscr{A}_2$*, if* $T \not\perp\!\!\!\perp_{\mathcal{D}} X_i \mid \boldsymbol{M}_k$*.*

By this theorem, it follows that:

(c) *Why we need the remedying procedure:* Theorem 4 implies ICPC remains the theoretical correctness of $\mathbb{A}_{PC}$ if $\mathscr{A}_1$ holds. However, just as pointed out by Aliferis et al. [8, p. 216], practical implementations of sound algorithms in the sense of $\mathscr{A}_1$ may be statistically imperfect, because $\mathscr{A}_1$ does not entail any practical feasibility in practice (although it can lead to a convenient proof of correctness). Specifically, the $\perp\!\!\!\perp_{\mathcal{D}}$-tests in the form of $T \perp\!\!\!\perp_{\mathcal{D}} X \mid (MB_T^{(Y)} \setminus \{X\}) \cup \{Y\}$

used in the refining procedure (see Theorem 4 for details) may be unreliable and thus lead to incorrect deletions of TPs from $MB_T^{(Y)}$ and $PC_T^{(Y)}$, so we insert the *remedying procedure* after the refining procedure to avoid such unexpected situations. Theorem 5 shows the reasonability of this procedure under the assumption $\mathscr{A}_2$.

4.4. InterHyPC: Combining GLL-PC and InterIAPC

We put forward ICPC in Subsection 4.2 and provided its theoretical basis to show its superiority in enhancing the forward search in Subsection 4.3. Although ICPC may overcome most of the shortcomings inherited from its subroutine, $\mathbb{A}_{PC}$, we believe a good selection for $\mathbb{A}_{PC}$ may still be more preferred. However, as we argued in Subsection 4.2, any selection of $\mathbb{A}_{PC}$ should be time efficient just like InterIAPC because it may be used repeatedly when implementing ICPC; this narrows the choices of $\mathbb{A}_{PC}$ (in particular, GLL-PC is not suitable for the role of $\mathbb{A}_{PC}$ although it is data efficient). In this subsection, we put forward a new selection for $\mathbb{A}_{PC}$, called InterHyPC, by combining GLL-PC and InterIAPC. Here, "Hy" denotes "hybrid".

To build InterHyPC, let us first recall Example 2, imagining that $Z_1 \in PC_T$ may not be incorrectly excluded any longer due to insufficiency of data if Y_4 ($\notin PC_T$) can be delayed to be included. A feasible heuristic for such an imagination is to partially apply the *elimination strategy* of GLL-PC [8, p. 192] in the sense that all variables conditionally independent of T should be discarded in each iteration and never considered again. This strategy can lead to an improvement on efficiency to a great degree, because the resulting algorithm (pseudo-coded from Line 1 to Line 10 of Algorithm 3; we call it the *TPC-subroutine* for convenience) can avoid a number of disruptive CI tests and hence detect true members of PC_T as early as possible. Nevertheless, as Figure 2 illustrates, only a tentative PC of T, namely, TPC_T, can be returned theoretically in this process. This is why GLL-PC proceeds to employ a *pruning procedure* (pseudo-coded from Line 2 to Line 4 of (a) in Algorithm 1) after the *TPC*-subroutine. PCMB [23] also uses the same procedure to ensure its output. However, the pruning procedure will increase the computational complexity many times and, thus, greatly decrease the time efficiency. A natural way of solving this problem is to implement InterIAPC by starting from TPC_T such that the data efficiency of GLL-PC and the time efficiency of InterIAPC can be appropriately traded off. This is the main idea of our InterHyPC algorithm. As seen, InterHyPC is actually a hybrid of GLL-PC and InterIAPC. We present its pseudo-code in Algorithm 3. Algorithm 3 with its Line 11 replaced by "$MB_T \leftarrow \mathtt{InterIAMB}(\mathcal{D}, T, TPC_T, \boldsymbol{L})$" to be InterHyMB.

It should be mentioned here that, as Aliferis et al. [8, p. 189] argued, although TPC_T may contain some nonmembers of PC_T, such situations are rare in practice so, in general, TPC_T can approximate PC_T quite closely. In other words, the *TPC*-subroutine provides a good start to InterIAPC, so the resulting InterHyPC algorithm is expected to perform better than InterIAPC.

Here, we shortly discuss the time complexity of InterHyPC/InterHyMB. For any independence-based PC or MB algorithm, as Aliferis et al. [8, p. 199] did, we also use the number of CI tests performed (or the associations computed) to measure its complexity. In fact, in Lines 3–6 of Algorithm 3, we need $|CanPC_T|$ tests and only one computation for the association: $|CanPC_T| + 1 = O(|\boldsymbol{V}|)$. In Lines 7–9, we need at most

$$\sum_{i=1}^{|TPC_T|-1} 2^i + 2^{|TPC_T|}(|\boldsymbol{V}| - 2|TPC_T|) = 2^{|TPC_T|}(|\boldsymbol{V}| - 2|TPC_T| + 1) - 2 = O(|\boldsymbol{V}| \cdot 2^{|MB_T|})$$

CI tests. Finally, Line 11 needs further $O(|TPC_T| \cdot |MB_T| + 2^{|MB_T|} \cdot |MB_T|) = O(2^{|MB_T|} \cdot |MB_T|)$ CI tests. In summary, `InterHyPC` is of the complexity $O(|\boldsymbol{V}| + |\boldsymbol{V}| \cdot 2^{|MB_T|} + 2^{|MB_T|} \cdot |MB_T|) = O(|\boldsymbol{V}| \cdot 2^{|MB_T|}) \triangleq f_5(T)$. Similarly, the time complexity of `InterHyMB` is $O(|\boldsymbol{V}|+|\boldsymbol{V}|\cdot 2^{|MB_T|}) = O(|\boldsymbol{V}|\cdot 2^{|MB_T|}) \triangleq g_5(T) = f_5(T)$. The complexities of the other independence-based PC or MB algorithms are presented in Tables 4 and 5.

Algorithm 3: `InterHyPC`

Procedure: $PC_T \leftarrow$ `InterHyPC`$(\mathcal{D}, T, \boldsymbol{S}, \boldsymbol{L})$
Input: $\mathcal{D}$ is a dataset; T is a target; $\boldsymbol{S}$ is a starting set; $\boldsymbol{L}$ is a blacklist.
Output: The output contains the PC of T.

1 $TPC_T \leftarrow \boldsymbol{S}$ and $CanPC_T \leftarrow \boldsymbol{V} \setminus TPC_T \setminus \{T\} \setminus \boldsymbol{L}$
2 **while** $CanPC_T \neq \emptyset$ **do**
3     $Y \leftarrow \arg\max_{X \in CanPC_T} f_{\mathcal{D}}(T; X \mid TPC_T)$
4   **if** $T \not\perp\!\!\!\perp Y \mid TPC_T$ **then**
5         $CanPC_T \leftarrow \{X \in CanPC_T : T \not\perp\!\!\!\perp X \mid TPC_T\} \setminus \{Y\}$ and $TPC_T \leftarrow TPC_T \cup \{Y\}$
6   **end**
7   **foreach** $X \in TPC_T$ **do**
8   **if** $\exists\, \boldsymbol{Z} \subseteq TPC_T \setminus \{X\}$ s.t. $T \perp\!\!\!\perp X \mid \boldsymbol{Z}$ **then** $TPC_T \leftarrow TPC_T \setminus \{X\}$;
9   **end**
10 **end**
11 $PC_T \leftarrow$ `InterIAPC`$(\mathcal{D}, T, TPC_T, \boldsymbol{L})$
12 **return** PC_T and MB_T

5. Breaking ties

This section addresses the problem $\mathscr{P}_3$ about the way of breaking ties among equal negative p-values. We consider this problem because it may arise frequently in large local discovery.

In the literature, the ties are often simply broken at random [10, 23]. However, this way of dealing with $\mathscr{P}_3$ does not consider the possible consequence that the selected variable may be an FP; if this is the case, the quality of the subsequent CI tests will be lowered due to the so-called cascading errors [22]. Besides, Tsamardinos et al. [7] used the G^2 statistic, $G^2_{\mathcal{D}}(\cdot)$, to break ties without giving a reason. In what follows, we analyze why the way of borrowing $G^2_{\mathcal{D}}(\cdot)$ to break ties is theoretically reasonable, then we explain why this way can only be used in rare situations. After that, we present a new method accompanied by an example used to illustrate how the new method works.

5.1. Using the G^2 statistic to break ties

Assume we are trying to choose one from all the ℓ variables, $\{Y_1, \cdots, Y_\ell\} \triangleq \mathcal{Y}$, with equal largest negative p-values given $\boldsymbol{M}$; i.e., $f_{\mathcal{D}}(T; Y_1 \mid \boldsymbol{M}) = \cdots = f_{\mathcal{D}}(T; Y_\ell \mid \boldsymbol{M}) = \max_{X \in \boldsymbol{V} \setminus \boldsymbol{M} \setminus \{T\}} f_{\mathcal{D}}(T; X \mid \boldsymbol{M})$. Further, assume the G^2 statistic, $G^2_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M})$, is an approximate χ^2-variate with r_i theoretically degrees of freedom (in which only $r_{n,i}$ ($< r_i$) ones are valid) and the noncentrality parameter δ_i ($i = 1 \cdots, \ell$). It is mentioned here that if $\mathcal{D}$ is large enough, it is unnecessary to consider the problem $\mathscr{P}_3$; the way of breaking ties at random may have been desirable. In the following, we only consider the case of data insufficiency.

First of all, Theorem 2 reveals $f_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M})$ is increasing with n and δ_i and decreasing with r_i. This means $f_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M})$ will no longer properly measure the association of Y_i with T in the case of data insufficiency, because in this case the value of $G^2_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M})$ only can match $r_{n,i}$ out of the r_i theoretical degrees of freedom. Briefly, data inefficiency is a potential reason for leading to ties, since in this case r_i is spuriously large and may overly decrease the associated negative p-value.

The above analysis implies we can break ties by alleviating the influence of r_i on $f_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M})$. This hint can be just what the way of borrowing $G^2_{\mathcal{D}}(\cdot)$ follows. In fact, by $G^2_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M}) \dot{\sim} \chi^2(r_{n,i}, \delta_i)$, the expectation of $G^2_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M})$ approximates to $r_{n,i} + \delta_i$, meaning that $G^2_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M})$ is increasing with $r_{n,i}$ and δ_i. Here, $r_{n,i}$ is increasing with n and r_i. In the meanwhile, a larger r_i may lead to a larger dispersion of data instances and, thus, more invalid degrees of freedom; that is, $(r_i - r_{n,i})$ is also increasing with r_i. Mathematically, letting $r_{n,i} \approx \gamma(r_i)$, then the derivative of $\gamma(\cdot)$ has the following property: $\gamma'(\cdot) > 0$ and $1 - \gamma'(\cdot) > 0$, or, equivalently, $0 < \gamma'(\cdot) < 1$. In other words, $r_{n,i}$ is increasing with r_i at a slower speed. In summary, $G^2_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M})$ is increasing with n and δ_i; it is also increasing with r_i, but the speed of increase is slow. This explains why the way of borrowing $G^2_{\mathcal{D}}(\cdot)$ to break ties that Tsamardinos et al. [7] used is theoretically reasonable.

Although the G^2 statistic can be used to break ties from the theoretical angle, such situations are actually rare. In fact, by the proof of Theorem 2, it is easily concluded that $f_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M})$ is decreasing with r_i and increasing with $g^2_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M})$, given $G^2_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M}) = g^2_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M})$. This means

$$r_i = r_j \;\Leftrightarrow\; g^2_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M}) = g^2_{\mathcal{D}}(T; Y_j \mid \boldsymbol{M}),$$

under the condition $f_{\mathcal{D}}(T; Y_i \mid \boldsymbol{M}) = f_{\mathcal{D}}(T; Y_j \mid \boldsymbol{M})$. Without loss of generality, we assume $\ell = 2$. It follows that $\xi \triangleq f_{\mathcal{D}}(T; Y_1 \mid \boldsymbol{M}) - f_{\mathcal{D}}(T; Y_2 \mid \boldsymbol{M})$ is a continuous random variable if $r_1 \neq r_2$ and it degenerates to zero otherwise. Hence, we have: (a) In the case of $r_1 = r_2$, $g^2_{\mathcal{D}}(T; Y_1 \mid \boldsymbol{M}) = g^2_{\mathcal{D}}(T; Y_2 \mid \boldsymbol{M})$, so the way of borrowing $G^2_{\mathcal{D}}(\cdot)$ to break ties fails to work; (b) In the case of $r_1 \neq r_2$, $g^2_{\mathcal{D}}(T; Y_1 \mid \boldsymbol{M}) \neq g^2_{\mathcal{D}}(T; Y_2 \mid \boldsymbol{M})$, but

$$\mathrm{P}\{f_{\mathcal{D}}(T; Y_1 \mid \boldsymbol{M}) = f_{\mathcal{D}}(T; Y_2 \mid \boldsymbol{M})\} \;=\; \mathrm{P}(\xi = 0) \;=\; 0.$$

This explains that the way of borrowing $G^2_{\mathcal{D}}(\cdot)$ to break ties can only be used in rare situations. The analysis also implies that the $G^2_{\mathcal{D}}(\cdot)$-based method coincides with choosing the variable in $\mathcal{Y}$ with the largest number of configurations.

5.2. Replacing procedure

We briefly present a more practical way of breaking ties as follows: For $X \in \boldsymbol{M}$ and $Y \in \mathcal{Y}$, we wonder if X has a higher association with T than Y; if not, replace X with Y. Mathematically, use $(\boldsymbol{M} \setminus \{X\}) \cup \{Y\}$ to replace $\boldsymbol{M}$ in the current search, if $f_{\mathcal{D}}(T; X \mid (\boldsymbol{M} \setminus \{X\}) \cup \{Y\}) < f_{\mathcal{D}}(T; Y \mid \boldsymbol{M})$, in which

$$(X, Y) = \mathop{\arg\min}_{(\xi,\eta) \in \boldsymbol{M} \times \mathcal{Y}} f_{\mathcal{D}}(T; \xi \mid (\boldsymbol{M} \setminus \{\xi\}) \cup \{\eta\}). \tag{5.1}$$

If there are ties when determining (X, Y) via (5.1), a pair of X and Y will be selected randomly from the pairs corresponding to the ties. After this operation, we check if there are ties with respect to the updated $\boldsymbol{M}$. If the answer is "yes", we break the ties at random and then proceed to the current search. This is the main idea of our *replacing procedure*.

The following example illustrates how such a procedure works.

Example 4. *Consider the BN in Example 2 again. Assume we have obtained* $\boldsymbol{M} \triangleq \{Y_1, Y_2, Y_3\}$ *in a certain stage with ties over* $\mathcal{Y} \triangleq \{Y_4, Z_1\}$*:* $f_{\mathcal{D}}(T; Y_4 \mid \boldsymbol{M}) = f_{\mathcal{D}}(T; Z_1 \mid \boldsymbol{M})$*. If breaking the ties at random, one may select* Y_4 *entering* $\boldsymbol{M}$*; if it is the case, the consequence of Example 2 follows immediately. Thus, we use the replacing procedure to break ties. Observing* $T \perp\!\!\!\perp Y_1 \mid (\boldsymbol{M} \setminus \{Y_1\}) \cup \{Z_1\}$*, we assume* (Z_1, Y_1) *is the only pair of nodes satisfying* (5.1)*. Following the replacing procedure,* $\boldsymbol{M}$ *is updated with* $(\boldsymbol{M} \setminus \{Y_1\}) \cup \{Z_1\} = \{Z_1, Y_2, Y_3\}$*, conditioned on which no ties exist in the current stage. Note that the updated* $\boldsymbol{M}$ *has optimized its original version, because the FP* Y_1 *is replaced with the TP* Z_1*.*

6. Experimental results

This section makes a benchmarking study based on six synthetic BNs considered in [7, 8]. These BNs are representatives of a wide range of problem domains with different complexities. The details of the six networks are summarized in Table 3. See [7, 8] for more details.

Table 3. Summary of BNs.

BN	Number of Nodes	Number of Edges	$\lvert\mathcal{T}\rvert$	$\max_{T\in\mathcal{T}} \lvert PC_T\rvert$	$\operatorname{mean}_{T\in\mathcal{T}} \lvert PC_T\rvert$	$\min_{T\in\mathcal{T}} \lvert PC_T\rvert$	$\max_{T\in\mathcal{T}} \lvert MB_T\rvert$	$\operatorname{mean}_{T\in\mathcal{T}} \lvert MB_T\rvert$	$\min_{T\in\mathcal{T}} \lvert MB_T\rvert$
Alarm	37	46	18	6	3.80	3	8	5.53	4
Alarm10	370	570	24	9	5.96	4	13	9.42	8
Child10	200	257	20	8	6.50	5	8	8.00	8
Gene	801	977	22	11	7.41	5	15	11.09	10
Lung Cancer	800	1476	24	29	14.54	9	56	27.54	20
Pigs	441	592	18	41	10.39	5	68	16.22	8

The following items are clarified before presenting the experimental results:

- *Data*. For each network, we generate 10 datasets (with high Bayesian information criterion (BIC) scores[¶]) of size n with the aid of `FullBNT` [45], where the data size n is taken as 300, 500, 800, 1000, 2000, 5000; there are in total 360 (= 10×6×6) datasets used in our experiment. To alleviate the randomness of data, the runs of these 10 datasets will be averaged.
- *Targets*. To highlight the topic of this paper, we select a set of about 20 targets, $\mathcal{T}$, having the most PC or MB members for each BN. The details of $\mathcal{T}$ are described in Table 3. Each result is also averaged over the runs of these $|\mathcal{T}|$ selected targets to evaluate the overall performance of an algorithm.
- *Algorithms*. The experiment contains two parts: One part is for PC discovery and the other is for MB discovery. For the former, we consider nine independence-based algorithms including four `InterIAPC`-based ones (including `InterIAPC`), four `InterHyPC`-based ones (including `InterHyPC`), and the `GLL-PC` algorithm; for the latter, we consider seven independence-based algorithms including three `InterIAMB`- or `InterIAPC`-based ones (including `InterIAMB`), three `InterHyMB`- or `InterHyPC`-based ones (including `InterHyMB`), and the `GLL-MB` algorithm. Tables 4 and 5 describe these local discovery algorithms. In addition, all algorithms use the replacing procedure to break ties.

[¶]Specifically, for given $n = 300, 500, 800, 1000, 2000, 5000$ and for every $i = 1, \cdots, 10$, we randomly generated 100 sets of data samples, calculated their BIC scores, and selected the dataset with the highest score as the experimental dataset.

Table 4. Independence-based PC discovery algorithms performed in this section.

Notation	Description	Complexity	Reference(s)
`InterIAPC`	The `InterIAPC` algorithm	$O(\lvert V\rvert \cdot \lvert MB_T\rvert + 2^{\lvert MB_T\rvert} \cdot \lvert MB_T\rvert) \triangleq f_1(T)$	[10, 42]
`ICPC.InterIAPC`	`ICPC` with $\mathbb{A}_{PC} \triangleq$ "`InterIAPC`"	$f_1(T) + \sum_{Y \in PC_T} f_1(Y) \triangleq f_2(T)$	[24, this paper]
`PCOR.InterIAPC`	`PCOR` with $\mathbb{A}_{PC} \triangleq$ "`InterIAPC`"	$O(\lvert V\rvert^2) + f_1(T) + \sum_{X \in PC_T} f_1(X) = O(\lvert V\rvert^2) + f_2(T) \triangleq f_3(T)$	[10]
`PCOR.ICPC.InterIAPC`	`PCOR` with $\mathbb{A}_{PC} \triangleq$ "`ICPC.InterIAPC`"	$O(\lvert V\rvert^2) + f_2(T) + \sum_{X \in PC_T} f_2(X) \triangleq f_4(T)$	[10, this paper]
`InterHyPC`	The `InterHyPC` algorithm	$O(\lvert V\rvert \cdot 2^{\lvert MB_T\rvert}) \triangleq f_5(T)$	[24, this paper]
`ICPC.InterHyPC`	`ICPC` with $\mathbb{A}_{PC} \triangleq$ "`InterHyPC`"	$f_5(T) + \sum_{Y \in PC_T} f_5(Y) \triangleq f_6(T)$	[24, this paper]
`PCOR.InterHyPC`	`PCOR` with $\mathbb{A}_{PC} \triangleq$ "`InterHyPC`"	$O(\lvert V\rvert^2) + f_5(T) + \sum_{X \in PC_T} f_5(X) = O(\lvert V\rvert^2) + f_6(T) \triangleq f_7(T)$	[10, this paper]
`PCOR.ICPC.InterHyPC`	`PCOR` with $\mathbb{A}_{PC} \triangleq$ "`ICPC.InterHyPC`"	$O(\lvert V\rvert^2) + f_6(T) + \sum_{X \in PC_T} f_6(X) \triangleq f_8(T)$	[10, this paper]
`GLL-PC`	The `GLL-PC` algorithm	$O(\lvert V\rvert \cdot \lvert PC_T\rvert \cdot 2^{\lvert PC_T\rvert}) \triangleq f_9(T)$	[8]

Table 5. Independence-based MB discovery algorithms performed in this section.

Notation	Description	Complexity	Reference(s)
`InterIAMB`	The `InterIAMB` algorithm	$O(\lvert V\rvert \cdot \lvert MB_T\rvert) \triangleq g_1(T)$	[42]
`MBOR.InterIAPC`	`MBOR` with $\mathbb{A}_{PC} \triangleq$ "`InterIAPC`"	$f_3(T) + \sum_{X \in PC_T} [f_3(X) + O(\lvert PC_X\rvert \cdot 2^{\lvert PC_T\rvert})] \triangleq g_3(T)$	[10]
`MBOR.ICPC.InterIAPC`	`MBOR` with $\mathbb{A}_{PC} \triangleq$ "`ICPC.InterIAPC`"	$f_4(T) + \sum_{X \in PC_T} [f_4(X) + O(\lvert PC_X\rvert \cdot 2^{\lvert PC_T\rvert})] \triangleq g_4(T)$	[10, this paper]
`InterHyMB`	The `InterHyMB` algorithm	$O(\lvert V\rvert \cdot 2^{\lvert MB_T\rvert}) \triangleq g_5(T)$	[24, this paper]
`MBOR.InterHyPC`	`MBOR` with $\mathbb{A}_{PC} \triangleq$ "`InterHyPC`"	$f_5(T) + \sum_{X \in PC_T} [f_5(X) + O(\lvert PC_X\rvert \cdot 2^{\lvert PC_T\rvert})] \triangleq g_7(T)$	[10, this paper]
`MBOR.ICPC.InterHyPC`	`MBOR` with $\mathbb{A}_{PC} \triangleq$ "`ICPC.InterHyPC`"	$f_6(T) + \sum_{X \in PC_T} [f_6(X) + O(\lvert PC_X\rvert \cdot 2^{\lvert PC_T\rvert})] \triangleq g_8(T)$	[10, this paper]
`GLL-MB`	`GLL-MB`	$f_9(T) + \sum_{Y \in PC_T} f_9(Y) \triangleq g_9(T)$	[8]

Here, motivated by one of the referees, we also compare the score-based algorithm, *score-based simultaneous Markov blanket discovery* ($\mathtt{S^2TMB}$), proposed by Gao and Ji [46] with independence-based algorithms. This algorithm is an improved version of the *score-based local learning* [47, `SLL`]. For `SLL` and $\mathtt{S^2TMB}$, the subroutine of learning the substructures can use any global structure learning algorithm such as the *dynamic programming*-based [48] or the *integer linear programming*-based [49], both of which are score-based methods. By preliminary experiments, we find any of both [48, 49] as the subroutine will need a very long time to run. For this reason, we will employ the independence-based *three-phase dependency analysis* (`TPDA`) algorithm of Cheng et al. [24] to learn the substructures involved in the $\mathtt{S^2TMB}$ algorithm.

- *Measurements*. We use the Euclidean distance from (*precision*, *recall*) to $(1, 1)$ over all selected targets to evaluate the accuracy of an algorithm in the sense that the smaller the better, where *precision* is the number of TPs in the output divided by the number of nodes in the output, while *recall* is the number of TPs in the output divided by the number of TPs in the true network. To observe the mechanism of an algorithm in improving the accuracy, *precision* and *recall* are also separately studied. We also compute their *F-measure* values (or called F1 scores), defined as the harmonic mean of *precision* and *recall*, to measure the PC/MB algorithms.

By the above descriptions, the experiment is done with the aid of `FullBNT` [45] and `MIToolbox` [50]. The results on *precision*, *recall*, *Euclidean distance*, and *F-measure* for PC algorithms are presented in Figures 13, 15, 17, and 19, respectively, while the results for MB algorithms are given in Figures 14, 16, 18, and 20. We also provide the results on *running time* in

Figures 21 and 22. To be more concise, we integrate the runs of the six BNs by further averaging them and present the results in Figures 5–10. From the figures, it is concluded that our methods perform desirably in large local discovery. Specifically, we have

(a) *`InterHyPC` outperforms `InterIAPC`*: (i) `InterHyPC` has larger *precision* and *recall* values (and thus smaller Euclidean distance and larger F-measure values) than `InterIAPC` in any case of data size n. (ii) The *precision* of `InterIAPC` tends to decrease along with the increase of n, meaning that a larger dataset may lead to the inclusion of more FPs for `InterIAPC`; while the *precision* of `InterHyPC` increases steadily with n or remains at a high level, indicating the robustness of `InterHyPC`. (iii) The *recall* of `InterHyPC` grows faster than that of `InterIAPC`.
(b) *Each `InterHyPC`-based algorithm outperforms the corresponding `InterIAPC`-based algorithm* with respect to each of the four measurements (*precision*, *recall*, Euclidean distance, and F-measure). This may be attributed to the inheritance of the performances of `InterHyPC` and `InterIAPC`.
(c) *The $\mathbb{A}_{PC}$-based `ICPC` algorithm performs better than $\mathbb{A}_{PC}$*, as expected when building `ICPC`: (i) This holds true when n is not very small. (ii) When n is very small, the $\mathbb{A}_{PC}$-based `ICPC` algorithm may have smaller *precision* than $\mathbb{A}_{PC}$; even so, `ICPC` still remains its capacity of capturing more information about T in such situations. Section 7 explains why it is the case and how we deal with this possibility. (iii) `ICPC` improves $\mathbb{A}_{PC}$ substantially on *recall*, so the idea of IC (information connection) reaches the goal of detecting more TPs in the true sense.
(d) *`PCOR` and `MBOR` inherit the superiority of `ICPC` specified in (c)*: On the one hand, the $\mathbb{A}_{PC}$-based `PCOR` or `MBOR` algorithm overwhelmingly outperforms $\mathbb{A}_{PC}$; on the other hand, "$\mathbb{A}$.ICPC.$\mathbb{A}_{PC}$" can further improve "$\mathbb{A}.\mathbb{A}_{PC}$", in which $\mathbb{A}$ stands for `PCOR` or `MBOR`, and $\mathbb{A}_{PC}$ denotes `InterIAPC` or `InterHyPC`. This is just what we expected when building `ICPC`. Finally, we mention that "$\mathbb{A}$.ICPC.InterHyPC" possesses more robust performance than "$\mathbb{A}$.ICPC.InterIAPC" in most situations.

In summary, the `InterHyPC` algorithm can be used to replace `InterIAPC` for local discovery due to its more desirable performance on *precision* and *recall*; the $\mathbb{A}_{PC}$-based `ICPC` algorithm can usually lead to a great improvement on $\mathbb{A}_{PC}$; the `ICPC`-based `PCOR` and `MBOR` algorithms can be an ideal solution to the problem of premature termination of the forward search that arises frequently in large local discovery.

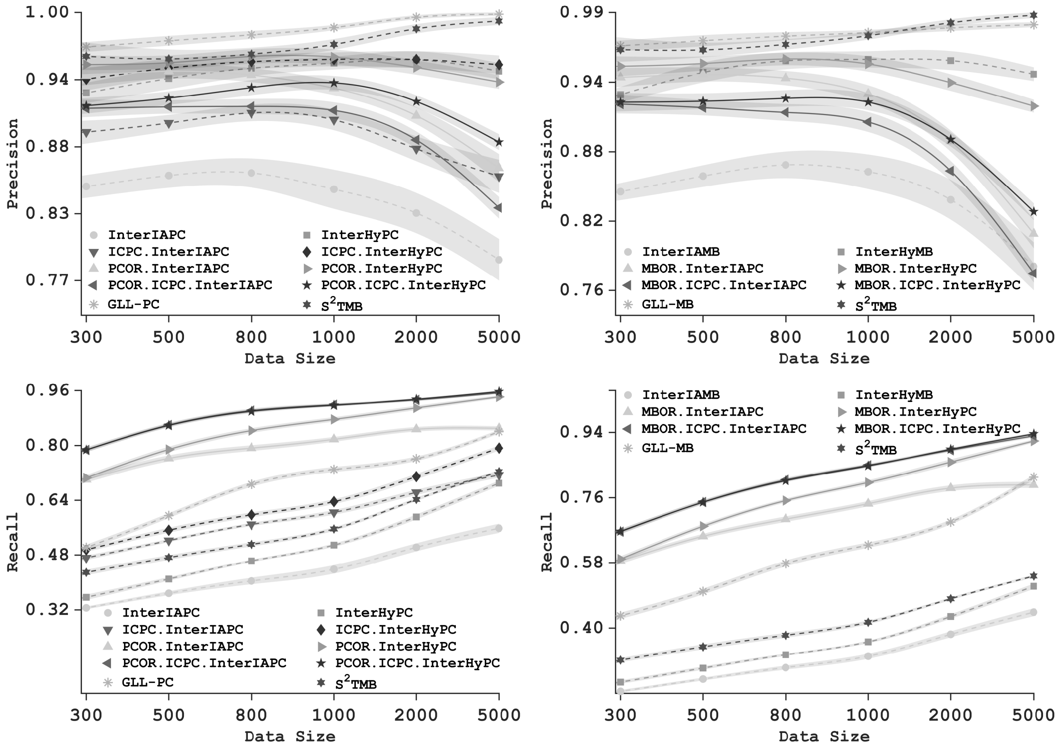


Figure 5. *Precision*, *recall*, and their 95% confidence bands averaged over the six synthetic BNs.

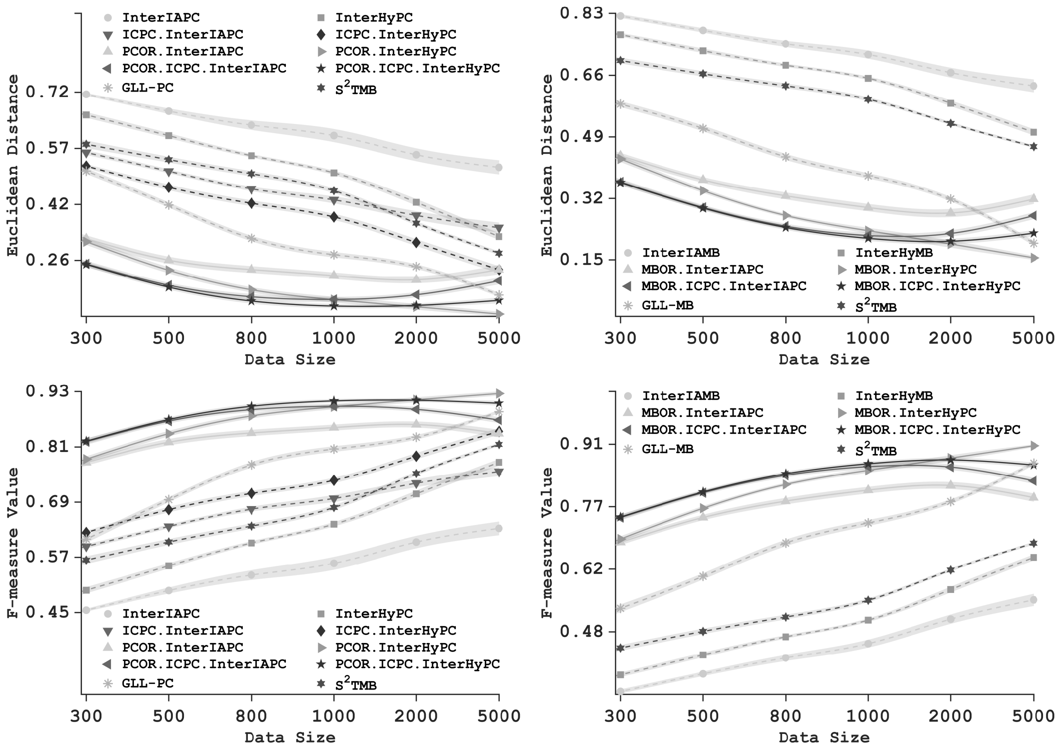


Figure 6. *Euclidean distance*, *F-measure*, and their 95% confidence bands averaged over the six synthetic BNs.

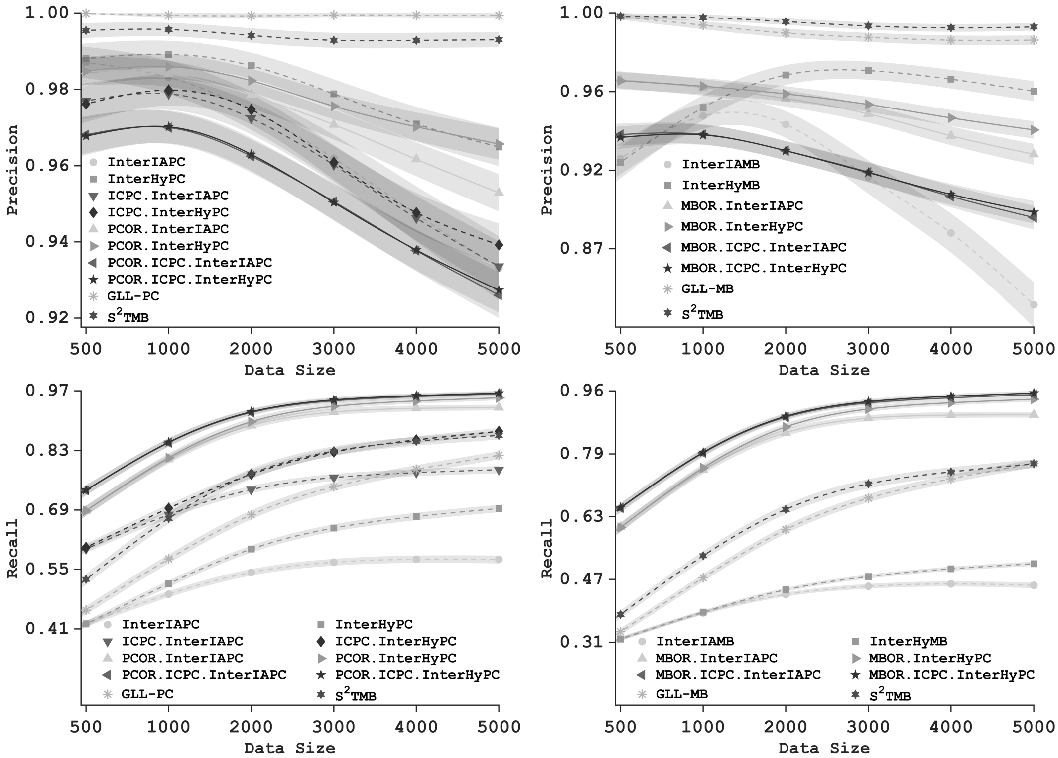


Figure 7. *Precision*, *recall*, and 95% confidence bands averaged over the 1000 random BNs.

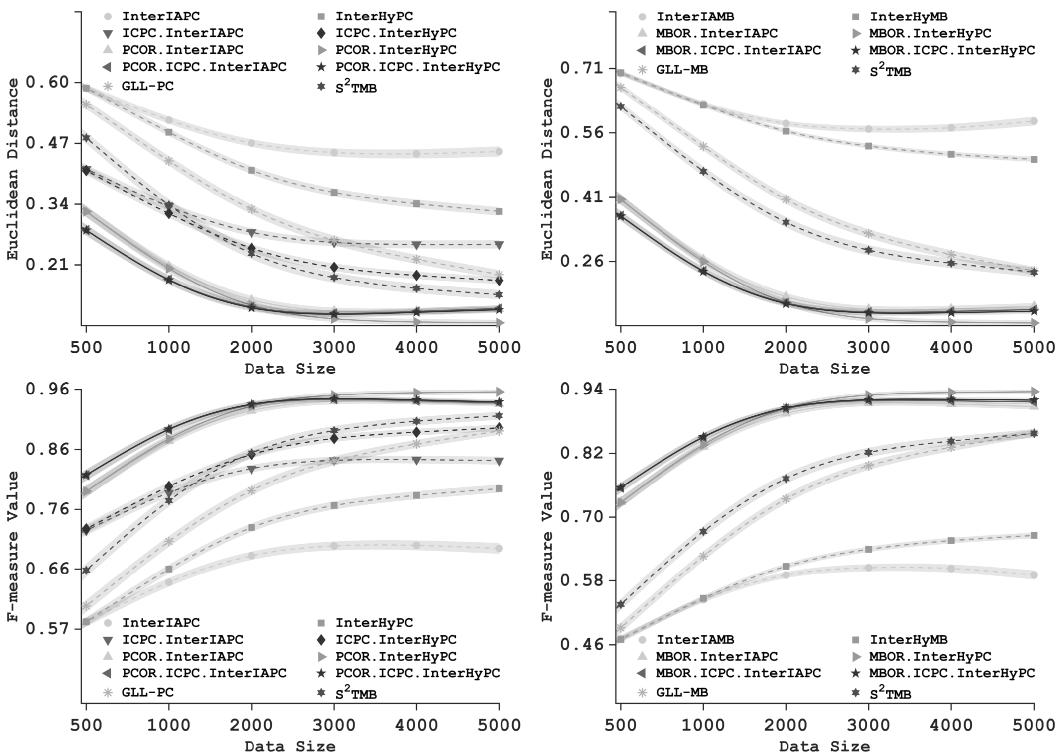


Figure 8. *Euclidean distance*, *F-measure*, and 95% confidence bands averaged over the 1000 random BNs.

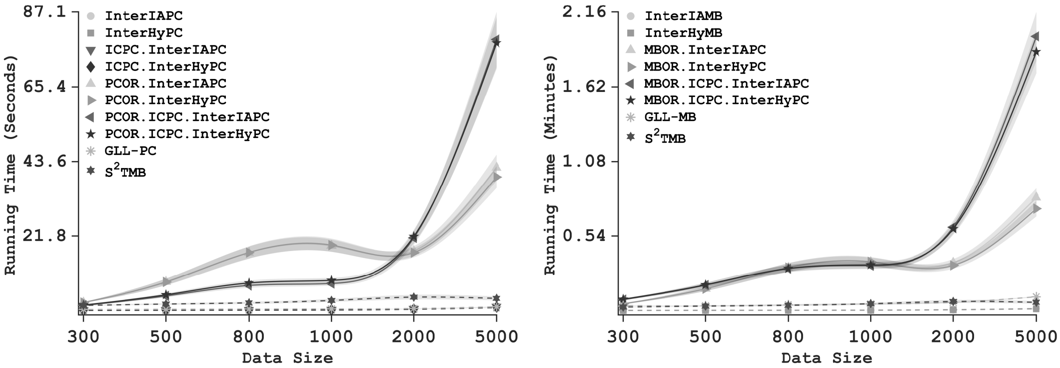


Figure 9. *Running time* and 95% confidence bands averaged over the six synthetic BNs.

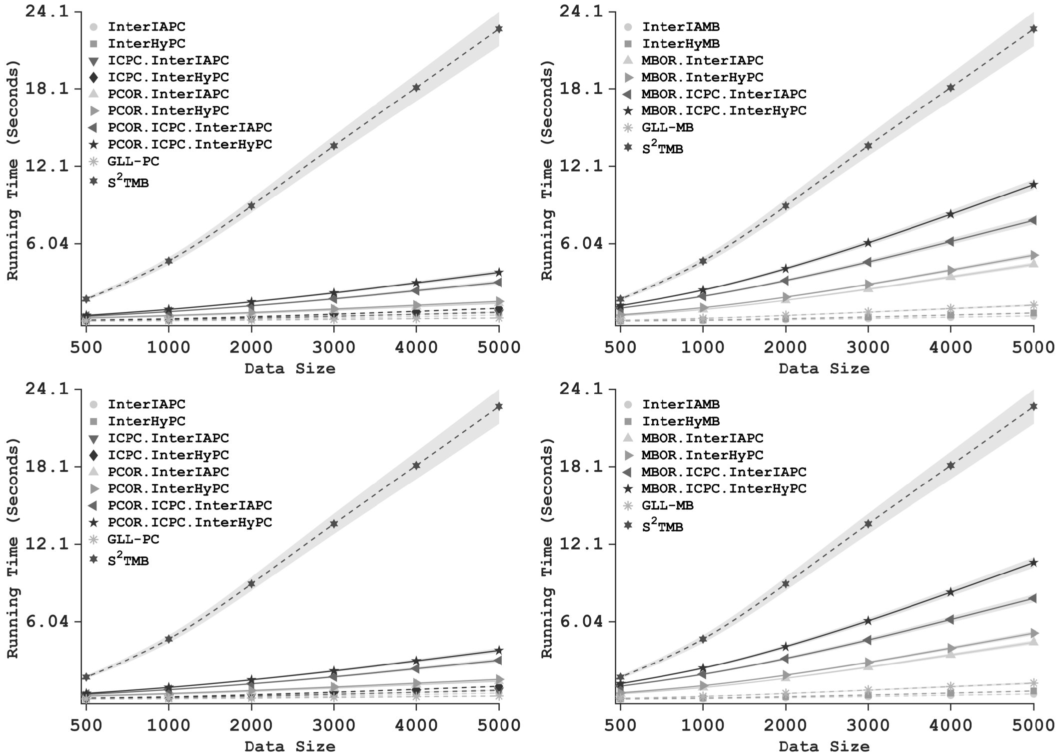


Figure 10. *Running time* and 95% confidence bands averaged over the 1000 random BNs.

7. Conclusions and Discussion

In this paper, we address the challenges of local discovery in three aspects: i) Examining the reliability of CI tests and proposing a more realistic approach; ii) enhancing existing local discovery algorithms to prevent premature termination of forward search, introducing the concept of information connection and a novel algorithm; and iii) optimizing the method for breaking ties among equal associations. Specifically, as motivated by the three problems, we studied how to modify the assumption $\mathscr{A}_1$ more reasonably, put forward the ICPC algorithm based on the idea of *information connection* and the extended information flow metaphor by providing detailed theoretical backgrounds, and presented a new way of breaking ties among equal negative p-values. By discussing the impact of data size on the reliability of CI tests and defining the concept of *extended Markov boundary*, we theoretically proved the correctness of `ICPC`. As demonstrated, compared to the existing state of the art algorithms, the `PCOR` and `MBOR` algorithms based on `ICPC` perform better in most cases and thus can be deemed to be a desirable solution when the PC or MB of the target contains too many nodes.

Before ending this paper, we present two concluding remarks as follows:

- *Complexity of `ICPC`, `PCOR`, and `MBOR`.* By Algorithm 2, `ICPC` calls $\mathbb{A}_{PC}$ repeatedly when the enhanced forward search is done. However, every such a calling may be finished very rapidly because it starts from $PC_T \setminus \{Y\}$ instead of an empty set. Therefore, the complexity of `ICPC` is only slightly higher than that of $\mathbb{A}_{PC}$[‖]. As a consequence, the complexities of "`PCOR.ICPC.`$\mathbb{A}_{PC}$" and "`MBOR.ICPC.`$\mathbb{A}_{PC}$" are only a bit higher than that of "`PCOR.`$\mathbb{A}_{PC}$" and "`MBOR.`$\mathbb{A}_{PC}$", respectively.
- *Measurements used in the experiment.* In the benchmarking study, we used the Euclidean distance and F-measure based on *precision* and *recall* to evaluate the accuracy of an algorithm. Denote the true and discovered PCs or MBs of T by $\boldsymbol{M}$ and $\boldsymbol{M}_{\mathbb{A}}$, respectively. Then, *precision* and *recall* are defined as follows:

$$precision \triangleq \frac{|\boldsymbol{M} \cap \boldsymbol{M}_{\mathbb{A}}|}{|\boldsymbol{M}_{\mathbb{A}}|} \quad \text{and} \quad recall \triangleq \frac{|\boldsymbol{M} \cap \boldsymbol{M}_{\mathbb{A}}|}{|\boldsymbol{M}|}.$$

These definitions are applicable to the cases that $|\boldsymbol{M}|$ and $|\boldsymbol{M}_{\mathbb{A}}|$ are not very small. However, when $|\boldsymbol{M}|$ or $|\boldsymbol{M}_{\mathbb{A}}|$ is very small, *precision* and *recall* cannot measure the accuracy of an algorithm very well; we provide an illustration on this assertion in Example 5. This example implies that we can weigh *precision* and *recall* with their opposites to alleviate the phenomenon that Example 5 presents. Following this hint, we denote $\boldsymbol{M}^{\mathrm{C}} = (\boldsymbol{V} \setminus \{T\}) \setminus \boldsymbol{M}$, $\boldsymbol{M}_{\mathbb{A}}^{\mathrm{C}} = (\boldsymbol{V} \setminus \{T\}) \setminus \boldsymbol{M}_{\mathbb{A}}$, and put

$$\begin{aligned} precision_{\mathrm{w}} &\triangleq \frac{1}{2}\left(\frac{|\boldsymbol{M} \cap \boldsymbol{M}_{\mathbb{A}}|}{|\boldsymbol{M}_{\mathbb{A}}|} + \frac{|\boldsymbol{M}^{\mathrm{C}} \cap \boldsymbol{M}_{\mathbb{A}}^{\mathrm{C}}|}{|\boldsymbol{M}_{\mathbb{A}}^{\mathrm{C}}|}\right) = \frac{1}{2}\left(\frac{|\boldsymbol{M} \cap \boldsymbol{M}_{\mathbb{A}}|}{|\boldsymbol{M}_{\mathbb{A}}|} + \frac{|(\boldsymbol{V} \setminus \{T\}) \setminus (\boldsymbol{M} \cup \boldsymbol{M}_{\mathbb{A}})|}{|\boldsymbol{V} \setminus \{T\} \setminus \boldsymbol{M}_{\mathbb{A}}|}\right), \\ recall_{\mathrm{w}} &\triangleq \frac{1}{2}\left(\frac{|\boldsymbol{M} \cap \boldsymbol{M}_{\mathbb{A}}|}{|\boldsymbol{M}|} + \frac{|\boldsymbol{M}^{\mathrm{C}} \cap \boldsymbol{M}_{\mathbb{A}}^{\mathrm{C}}|}{|\boldsymbol{M}^{\mathrm{C}}|}\right) = \frac{1}{2}\left(\frac{|\boldsymbol{M} \cap \boldsymbol{M}_{\mathbb{A}}|}{|\boldsymbol{M}|} + \frac{|(\boldsymbol{V} \setminus \{T\}) \setminus (\boldsymbol{M} \cup \boldsymbol{M}_{\mathbb{A}})|}{|\boldsymbol{V} \setminus \{T\} \setminus \boldsymbol{M}|}\right). \end{aligned}$$

When $|\boldsymbol{M}|$ or $|\boldsymbol{M}_{\mathbb{A}}|$ is very small, $precision_{\mathrm{w}}$ and $recall_{\mathrm{w}}$ may be more suitable than *precision* and *recall* for measuring the accuracy of an algorithm, as do the corresponding Euclidean distance and F-measure. Example 5 illustrates this idea.

Example 5. *Let $\boldsymbol{V}$ be a set of 20 variables, $\mathcal{D}$ be a dataset over $\boldsymbol{M}$, $\mathbb{A}_1$ and $\mathbb{A}_2$ be two local discovery algorithms. For $T \in \boldsymbol{V}$, denote its PC (or MB) by $\boldsymbol{M}$ with $|\boldsymbol{M}| = 9$. Assume $\boldsymbol{M}_i$ is the output of $\mathbb{A}_i$, with*

$$|\boldsymbol{M}_1| = 1, \ \ |\boldsymbol{M} \cap \boldsymbol{M}_1| = 1; \quad |\boldsymbol{M}_2| = 8, \ \ |\boldsymbol{M} \cap \boldsymbol{M}_2| = 7.$$

Then we have

$$\begin{aligned} &precision^{(1)} = 1, && recall^{(1)} = 0.1111; && precision_{\mathrm{w}}^{(1)} = 0.7778. && recall_{\mathrm{w}}^{(1)} = 0.5556, \\ &precision^{(2)} = 0.8750. && recall^{(2)} = 0.7778; && precision_{\mathrm{w}}^{(2)} = 0.8466, && recall_{\mathrm{w}}^{(2)} = 0.8389. \end{aligned}$$

Hence, the weighted version of precision and recall can measure the algorithms in a more suitable way.

[‖]According to the simulation study, we find that this is very applicable for the results of 1000 random BNs. However, for the six real-world BNs, it is only suitable when the sample size is not very large (e.g., not larger than 1000). Note that, when the data size is large, the running time increases sharply, which may be related to the original intention (or motivation) of our algorithms. Our motivation is to *solve the problem of low efficiency of local learning when the sample size is not sufficient*. In fact, when the sample size is large, more nodes will stay in the $MB_T^{\mathbb{A}}$ at Line 11 of Algorithm 2, greatly increasing the running cost of *backward search*. We are planning to undertake how to solve this problem in the near future, but this may be a long process.

A. Proofs

This appendix provides the proofs of some theoretical results.

A.1. Proof of Theorem 1

Let $f_{r,\delta}(x)$ and $F_{r,\delta}(x) \triangleq \int_0^x f_{r,\delta}(t)\mathrm{d}t$ be the probability density function and the cumulative distribution function, respectively, for the χ^2-variate with r degrees of freedom and the noncentrality parameter δ (namely, $\chi^2(r,\delta)$), where

$$f_{r,\delta}(x) = \frac{\mathrm{e}^{-(x+\delta)/2}x^{r/2-1}}{2^{r/2}} \sum_{k=0}^{\infty} \frac{(\delta/2)^k (x/2)^k}{k!\Gamma(k+r/2)},$$

if $x > 0$ and $f_{r,\delta}(x) = 0$ otherwise. For distinction, we slightly abuse these notations and use $f_r(x)$ and $F_r(x)$ as shorthand for $f_{r,0}(x)$ and $F_{r,0}(x)$, respectively. By direct calculations, it concludes that:

$$\frac{\partial f_{r,\delta}(x)}{\partial x} = \tfrac{1}{2}[f_{r-2,\delta}(x) - f_{r,\delta}(x)], \tag{A.1}$$

$$\frac{\partial f_{r,\delta}(x)}{\partial \delta} = \tfrac{1}{2}[f_{r+2,\delta}(x) - f_{r,\delta}(x)]. \tag{A.2}$$

Before presenting the proof of Theorem 1, we first prove a lemma.

Lemma 2. *For any $\triangle\delta > 0$, the following statements hold:*

a) *$F_{r,\delta}(x)$ is increasing with x and decreasing with r or δ.*
b) *$f_{r+2,\delta}(x)/f_{r,\delta}(x)$ is increasing with x.*
c) *For any $\triangle\delta > 0$, $F_{r,\delta+\triangle\delta}(x)/F_{r,\delta}(x)$ and $[1 - F_{r,\delta+\triangle\delta}(x)]/[1 - F_{r,\delta}(x)]$ are increasing with x and decreasing with r; specifically, $F_{r,\delta}(x)/F_r(x)$ and $[1 - F_{r,\delta}(x)]/[1 - F_r(x)]$ are increasing with x and decreasing with r.*

Proof. The first statement is shown in [51, 52].

To prove (b), we denote

$$a_{i,j} = \frac{(\delta/2)^{i+j}(x/2)^{i+j}}{i!\,j!}\left[\frac{1}{\Gamma(i+r/2)\Gamma(j+r/2)} - \frac{1}{\Gamma(i+r/2+1)\Gamma(j+r/2-1)}\right].$$

In view of Eq (A.1), it follows that

$$\begin{aligned}
\frac{\partial}{\partial x}\left(\frac{f_{r+2,\delta}(x)}{f_{r,\delta}(x)}\right) &= \frac{[\partial f_{r+2,\delta}(x)/\partial x]\, f_{r,\delta}(x) - [\partial f_{r,\delta}(x)/\partial x]\, f_{r+2,\delta}(x)}{f_{r,\delta}^2(x)} \\
&= \frac{\frac{1}{2}\,[f_{r,\delta}(x) - f_{r+2,\delta}(x)]\, f_{r,\delta}(x) - \frac{1}{2}\,[f_{r-2,\delta}(x) - f_{r,\delta}(x)]\, f_{r+2,\delta}(x)}{f_{r,\delta}^2(x)} \\
&= \frac{g_{r,\delta}(x)}{2f_{r,\delta}^2(x)},
\end{aligned} \tag{A.3}$$

with $g_{r,\delta}(x) \triangleq f_{r,\delta}^2(x) - f_{r+2,\delta}(x)f_{r-2,\delta}(x) = 2^{-n}\mathrm{e}^{-(x+\delta)}x^{n-2}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} a_{i,j}$. For any $i \geqslant 0$ and $j \geqslant 1$, we have

$$a_{i,0} = \frac{(\delta/2)^i (x/2)^i}{i!}\left[\frac{1}{\Gamma(i+r/2)\Gamma(r/2)} - \frac{1}{\Gamma(i+r/2+1)\Gamma(r/2-1)}\right]$$

$$
\begin{aligned}
&= \frac{(\delta/2)^i (x/2)^i}{i!\Gamma(i+r/2+1)\Gamma(r/2)}\Big[\Big(i+\frac{r}{2}\Big)-\Big(\frac{r}{2}-1\Big)\Big] \\
&> 0, \\
a_{i,i+1} &= \frac{(\delta/2)^{2i+1}(x/2)^{2i+1}}{i!(i+1)!}\Big[\frac{1}{\Gamma(i+r/2)\Gamma(i+1+r/2)} - \frac{1}{\Gamma(i+r/2+1)\Gamma(i+1+r/2-1)}\Big] \\
&= 0.
\end{aligned}
$$

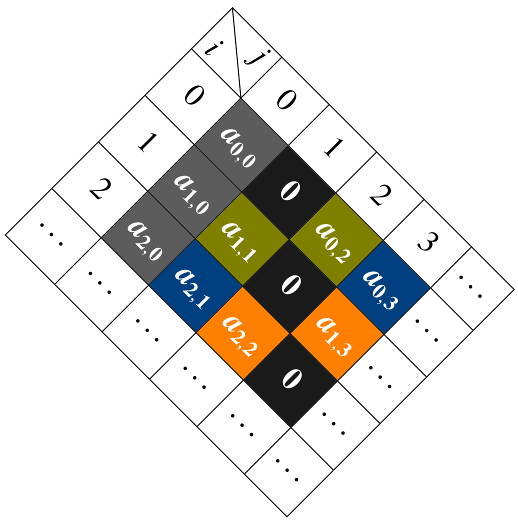
Furthermore, for any $i \geqslant 0$ and $j \geqslant 1$ with $j \neq i+1$, we obtain

$$
\begin{aligned}
a_{i,j} + a_{j-1,i+1} &= \frac{(\delta/2)^{i+j}(x/2)^{i+j}}{i!\,j!}\Big[\frac{1}{\Gamma(i+r/2)\Gamma(j+r/2)} - \frac{1}{\Gamma(i+r/2+1)\Gamma(j+r/2-1)}\Big] \\
&\quad + \frac{1}{(j-1)!(i+1)!}\Big[\frac{(\delta/2)^{j-1+i+1}(x/2)^{j-1+i+1}}{\Gamma(j-1+\frac{r}{2})\Gamma(i+1+\frac{r}{2})} - \frac{(\delta/2)^{j-1+i+1}(x/2)^{j-1+i+1}}{\Gamma(j-1+\frac{r}{2}+1)\Gamma(i+1+\frac{r}{2}-1)}\Big] \\
&= \frac{(\delta/2)^{i+j}(x/2)^{i+j}[(i+1)(i+\frac{r}{2})-(i+1)(j+\frac{r}{2}-1)+j(j+\frac{r}{2}-1)-j(i+\frac{r}{2})]}{(i+1)!\,j!\Gamma(i+1+r/2)\Gamma(j+r/2)} \\
&= \frac{(\delta/2)^{i+j}(x/2)^{i+j}(i-j+1)^2}{(i+1)!\,j!\Gamma(i+1+r/2)\Gamma(j+r/2)} \\
&> 0.
\end{aligned}
$$

According to the hint of Table 6, it is concluded that

$$
\begin{aligned}
g_{r,\delta}(x) &= 2^{-r}\mathrm{e}^{-(x+\delta)}x^{r-2}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} a_{i,j} \\
&= 2^{-r}\mathrm{e}^{-(x+\delta)}x^{r-2}\Big[\sum_{i=0}^{\infty} a_{i,0} + \sum_{i=0}^{\infty} a_{i,i+1} + \sum_{i=1}^{\infty}\sum_{j=1}^{i}(a_{i,j}+a_{j-1,i+1})\Big] \\
&> 0.
\end{aligned}
$$

Table 6. Explanation for the following equality: $\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} a_{i,j} = \sum_{i=0}^{\infty} a_{i,0} + \sum_{i=0}^{\infty} a_{i,i+1} + \sum_{i=1}^{\infty}\sum_{j=1}^{i}(a_{i,j}+a_{j-1,i+1})$, in which the first part is tabulated with dark grey as the back color, the second part is with black, and the third part is multicolored.



This combined with Eq (A.3) implies

$$\frac{\partial}{\partial x}\left(\frac{f_{r+2,\delta}(x)}{f_{r,\delta}(x)}\right)=\frac{g_{r,\delta}(x)}{2f_{r,\delta}^2(x)}>0,$$

so $f_{r+2,\delta}(x)/f_{r,\delta}(x)$ is increasing with respect to x.

Now, we prove (c) by means of (b). The proof is divided into the following four parts:

c_1) *Proving $F_{r,\delta+\triangle\delta}(x)/F_{r,\delta}(x)$ is increasing with x:* In fact, by (b), $f_{r+2,\delta}(y)/f_{r,\delta}(y) < f_{r+2,\delta}(x)/f_{r,\delta}(x)$ holds for any x and y with $0<y<x$, or equivalently, we have $f_{r+2,\delta}(y)f_{r,\delta}(x) < f_{r+2,\delta}(x)f_{r,\delta}(y)$. By means of Eq (A.2), this gives

$$\begin{aligned}\frac{\partial}{\partial\delta}\left(\frac{f_{r,\delta}(y)}{f_{r,\delta}(x)}\right) &= \frac{[\partial f_{r,\delta}(y)/\partial\delta]\,f_{r,\delta}(x)-[\partial f_{r,\delta}(x)/\partial\delta]\,f_{r,\delta}(y)}{f_{r,\delta}^2(x)}\\ &= \frac{\frac{1}{2}\left[f_{r+2,\delta}(y)-f_{r,\delta}(y)\right]f_{r,\delta}(x)-\frac{1}{2}\left[f_{r+2,\delta}(x)-f_{r,\delta}(x)\right]f_{r,\delta}(y)}{f_{r,\delta}^2(x)}\\ &= \frac{f_{r+2,\delta}(y)f_{r,\delta}(x)-f_{r+2,\delta}(x)f_{r,\delta}(y)}{2f_{r,\delta}^2(x)}\\ &< 0,\end{aligned}$$

which further indicates $f_{r,\delta+\triangle\delta}(y)/f_{r,\delta+\triangle\delta}(x) < f_{r,\delta}(y)/f_{r,\delta}(x)$. Note that $f_{r,\delta+\triangle\delta}(y)/f_{r,\delta+\triangle\delta}(x) - f_{r,\delta}(y)/f_{r,\delta}(x)$ is a continuous function of y in the closed interval $[0,x]$. It follows that

$$\begin{aligned}\frac{f_{r,\delta+\triangle\delta}(y)}{f_{r,\delta+\triangle\delta}(x)}<\frac{f_{r,\delta}(y)}{f_{r,\delta}(x)} &\Rightarrow \frac{F_{r,\delta+\triangle\delta}(x)}{f_{r,\delta+\triangle\delta}(x)}=\int_0^x\frac{f_{r,\delta+\triangle\delta}(y)}{f_{r,\delta+\triangle\delta}(x)}\mathrm{d}y<\int_0^x\frac{f_{r,\delta}(y)}{f_{r,\delta}(x)}\mathrm{d}y=\frac{F_{r,\delta}(x)}{f_{r,\delta}(x)}\\ &\Rightarrow f_{r,\delta+\triangle\delta}(x)F_{r,\delta}(x)>f_{r,\delta}(x)F_{r,\delta+\triangle\delta}(x)\\ &\Rightarrow \frac{\partial}{\partial x}\left(\frac{F_{r,\delta+\triangle\delta}(x)}{F_{r,\delta}(x)}\right)=\frac{f_{r,\delta+\triangle\delta}(x)F_{r,\delta}(x)-f_{r,\delta}(x)F_{r,\delta+\triangle\delta}(x)}{F_{r,\delta}^2(x)}>0.\end{aligned}$$

This means $F_{r,\delta+\triangle\delta}(x)/F_{r,\delta}(x)$ is increasing with x.

c_2) *Proving* $[1-F_{r,\delta+\triangle\delta}(x)]/[1-F_{r,\delta}(x)]$ *is increasing with respect to x:* According to the proof of (c_1), we have $f_{r,\delta+\triangle\delta}(y)/f_{r,\delta+\triangle\delta}(x) > f_{r,\delta}(y)/f_{r,\delta}(x)$ for any x and y with $y>x>0$. For any z ($>x$), noting $f_{r,\delta+\triangle\delta}(y)/f_{r,\delta+\triangle\delta}(x) - f_{r,\delta}(y)/f_{r,\delta}(x)$ is a continuous function of $y\in[x,z]$, it follows that

$$\begin{aligned}\frac{f_{r,\delta+\triangle\delta}(y)}{f_{r,\delta+\triangle\delta}(x)}>\frac{f_{r,\delta}(y)}{f_{r,\delta}(x)} &\Rightarrow \int_x^z\frac{f_{r,\delta+\triangle\delta}(y)}{f_{r,\delta+\triangle\delta}(x)}\mathrm{d}y>\int_x^z\frac{f_{r,\delta}(y)}{f_{r,\delta}(x)}\mathrm{d}y \text{ and } \int_z^{+\infty}\frac{f_{r,\delta+\triangle\delta}(y)}{f_{r,\delta+\triangle\delta}(x)}\mathrm{d}y\geqslant\int_z^{+\infty}\frac{f_{r,\delta}(y)}{f_{r,\delta}(x)}\mathrm{d}y\\ &\Rightarrow \frac{1-F_{r,\delta+\triangle\delta}(x)}{f_{r,\delta+\triangle\delta}(x)}=\int_x^{+\infty}\frac{f_{r,\delta+\triangle\delta}(y)}{f_{r,\delta+\triangle\delta}(x)}\mathrm{d}y>\int_x^{+\infty}\frac{f_{r,\delta}(y)}{f_{r,\delta}(x)}\mathrm{d}y=\frac{1-F_{r,\delta}(x)}{f_{r,\delta}(x)}\\ &\Rightarrow [1-F_{r,\delta+\triangle\delta}(x)]f_{r,\delta}(x)>[1-F_{r,\delta}(x)]f_{r,\delta+\triangle\delta}(x)\\ &\Rightarrow \frac{\partial}{\partial x}\left(\frac{1-F_{r,\delta+\triangle\delta}(x)}{1-F_{r,\delta}(x)}\right)=\frac{-f_{r,\delta+\triangle\delta}(x)[1-F_{r,\delta}(x)]+f_{r,\delta}(x)[1-F_{r,\delta+\triangle\delta}(x)]}{[1-F_{r,\delta}(x)]^2}>0.\end{aligned}$$

This means $[1-F_{r,\delta+\triangle\delta}(x)]/[1-F_{r,\delta}(x)]$ is increasing with respect to x.

c$_3$) *Proving $F_{r,\delta+\triangle\delta}(x)/F_{r,\delta}(x)$ is dereasing with r:* Let $\xi \sim \chi^2(r-2)$, $\eta \sim \chi^2(1)$, $\zeta \sim \chi^2(1, \delta+\triangle\delta)$, and $\tau \sim \chi^2(1, \delta)$ be four independent χ^2 variables. Then,

$$\begin{aligned} F_{r,\delta+\triangle\delta}(x) &= \mathrm{P}(\xi+\eta+\zeta \leqslant x) = \int_0^x f_1(y)\mathrm{P}(\xi+\zeta \leqslant x-y)\mathrm{d}y = \int_0^x f_1(y)F_{r-1,\delta+\triangle\delta}(x-y)\mathrm{d}y, \\ F_{r,\delta}(x) &= \mathrm{P}(\xi+\eta+\tau \leqslant x) = \int_0^x f_1(y)\mathrm{P}(\xi+\tau \leqslant x-y)\mathrm{d}y = \int_0^x f_1(y)F_{r-1,\delta}(x-y)\mathrm{d}y. \end{aligned}$$

According to (c$_1$), $F_{r-1,\delta+\triangle\delta}(x-y) < F_{r-1,\delta}(x-y)F_{r-1,\delta+\triangle\delta}(x)/F_{r-1,\delta}(x)$ holds for any $y \in (0, x)$. Therefore,

$$\begin{aligned} \frac{F_{r,\delta+\triangle\delta}(x)}{F_{r,\delta}(x)} = \frac{\int_0^x f_1(y)F_{r-1,\delta+\triangle\delta}(x-y)\mathrm{d}y}{\int_0^x f_1(y)F_{r-1,\delta}(x-y)\mathrm{d}y} &< \frac{\int_0^x f_1(y)F_{r-1,\delta}(x-y)F_{r-1,\delta+\triangle\delta}(x)/F_{r-1,\delta}(x)\mathrm{d}y}{\int_0^x f_1(y)F_{r-1,\delta}(x-y)\mathrm{d}y} \\ &= \frac{F_{r-1,\delta+\triangle\delta}(x)}{F_{r-1,\delta}(x)}. \end{aligned}$$

This implies that $F_{r,\delta+\triangle\delta}(x)/F_{r,\delta}(x)$ is decreasing with respect to r.

c$_4$) *Proving* $[1 - F_{r,\delta+\triangle\delta}(x)]/[1 - F_{r,\delta}(x)]$ *is dereasing with r:* First, using (c$_2$), it concludes that

$$1 - F_{r-1,\delta+\triangle\delta}(x-y) < [1 - F_{r-1,\delta}(x-y)] \cdot \frac{1 - F_{r-1,\delta+\triangle\delta}(x)}{1 - F_{r-1,\delta}(x)}, \tag{A.4}$$

holds for any $y \in (0, x)$. Inserting (A.4), we have

$$\frac{1 - F_{r,\delta+\triangle\delta}(x)}{1 - F_{r,\delta}(x)} = \frac{\int_x^{+\infty} f_1(y)[1 - F_{r-1,\delta+\triangle\delta}(x-y)]\mathrm{d}y}{\int_x^{+\infty} f_1(y)[1 - F_{r-1,\delta}(x-y)]\mathrm{d}y} < \frac{1 - F_{r-1,\delta+\triangle\delta}(x)}{1 - F_{r-1,\delta}(x)}.$$

Hence, $[1 - F_{r,\delta+\triangle\delta}(x)]/[1 - F_{r,\delta}(x)]$ is decreasing with respect to r.

The proof is completed. □

Using this lemma, we prove the following theorem:

Theorem 1 (Power and Reliability of CI Tests)**.** *Assume $\mathcal{D}$ is an insufficient dataset. Then, we have*

a) $\mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}} \,|\, \mathrm{E}_{\perp\!\!\!\perp}, \mathcal{D})$ *is decreasing with n and increasing with r.*
b) $\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}} \,|\, \mathrm{E}_{\not\perp\!\!\!\perp}, \mathcal{D})$ *is increasing with n and decreasing with r.*
c) $\mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp} \,|\, \mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}}, \mathcal{D})$ *is increasing with n and decreasing with r.*
d) $\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp} \,|\, \mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}}, \mathcal{D})$ *is decreasing with n and increasing with r.*

Proof. First, it is easily seen that $\mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}} \,|\, \mathrm{E}_{\perp\!\!\!\perp}, \mathcal{D}) = F_{r_n}(\chi^2_\alpha(r))$. This indicates (a) of Theorem 1 is just a direct consequence of (a) of Lemma 2, since r_n is increasing with n.

To compute $\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}} \,|\, \mathrm{E}_{\not\perp\!\!\!\perp}, \mathcal{D})$, we let $\langle \boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z} \rangle$ and $\langle \boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z} \rangle_{\mathcal{D}}$ denote two random variables in the sense of

$$\langle \boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z} \rangle = \begin{cases} 1, & \text{if } \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z} \\ 0, & \text{if } \boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z} \end{cases} \quad \text{and} \quad \langle \boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z} \rangle_{\mathcal{D}} = \begin{cases} 1, & \text{if } \boldsymbol{X} \perp\!\!\!\perp_{\mathcal{D}} \boldsymbol{Y} \,|\, \boldsymbol{Z}, \\ 0, & \text{if } \boldsymbol{X} \not\perp\!\!\!\perp_{\mathcal{D}} \boldsymbol{Y} \,|\, \boldsymbol{Z}, \end{cases}$$

respectively. These two notations are inspired by the notion of "meta-space" [22] representing all possible independencies in the domain. Similarly, we also regard $\mathrm{I}(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z})$ as a random variable with

$$\mathrm{I}(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) \sim g(\tau) = \begin{cases} g_+(\tau), & \tau > 0, \\ \delta(\tau/g_0) = g_0 \cdot \delta(\tau), & \tau = 0, \end{cases}$$

where $g_+(\tau)$ is a nonnegative integrable function on $\tau \in (0, +\infty)$; $g_0 = 1 - \int_0^{+\infty} g_+(\tau)\mathrm{d}\tau \in (0, 1)$; $\delta(\tau)$ is the Dirac δ-function. Figure 11 presents a meta-BN for the relationship among $\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z})$, r, $\mathcal{D}$, $\langle \boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z}\rangle$, and $\langle \boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z}\rangle_{\mathcal{D}}$. Using these notions, we have

$$
\begin{aligned}
\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}} \big| \mathrm{E}_{\not\perp\!\!\!\perp}, \mathcal{D}) &= \mathrm{P}(\langle \boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z}\rangle_{\mathcal{D}} = 0 \big| \mathrm{I}(\boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z}) > 0, \mathcal{D}) \\
&= \frac{\int_0^{+\infty} g_+(\tau)\mathrm{P}(\langle \boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z}\rangle_{\mathcal{D}} = 0 \big| \mathrm{I}(\boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z}) = \tau, \mathcal{D})\mathrm{d}\tau}{\int_0^{+\infty} g_+(\tau)\mathrm{d}\tau} \\
&= \frac{\int_0^{+\infty} g_+(\tau)[1 - F_{r_n, 2n\tau}(\chi^2_\alpha(r))]\mathrm{d}\tau}{1 - g_0}. \qquad \text{(A.5)}
\end{aligned}
$$

This means that (b) of Theorem 1 is also a direct consequence of (a) of Lemma 2, since r_n and $2n\tau$ are increasing with n.

$$
\begin{tikzcd}
\mathrm{I}(X;Y|Z) \arrow[d, Rightarrow] \arrow[dr] & r \arrow[d] & \mathcal{D} \arrow[dl] \\
\langle X;Y|Z\rangle & \langle X;Y|Z\rangle_{\mathcal{D}} &
\end{tikzcd}
$$

Figure 11. A meta-BN for the relationship among $\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z})$, r, $\mathcal{D}$, $\langle \boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z}\rangle$, and $\langle \boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z}\rangle_{\mathcal{D}}$.

Next, we prove (c) and (d). In fact, similar to the computation of (A.5), we have

$$
\begin{aligned}
\frac{1}{\mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp}\,|\,\mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}}, \mathcal{D})} &= 1 + \frac{\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp})}{\mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp})} \cdot \frac{\mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}}\,|\,\mathrm{E}_{\not\perp\!\!\!\perp}, \mathcal{D})}{\mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp_{\mathcal{D}}}\,|\,\mathrm{E}_{\perp\!\!\!\perp}, \mathcal{D})} = 1 + \frac{1}{g_0}\int_0^{+\infty} g_+(\tau)\frac{F_{r_n, 2n\tau}(\chi^2_\alpha(r))}{F_{r_n}(\chi^2_\alpha(r))}\mathrm{d}\tau, \quad \text{and} \qquad \text{(A.6)} \\
\frac{1}{\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp}\,|\,\mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}}, \mathcal{D})} &= 1 + \frac{\mathrm{P}(\mathrm{E}_{\perp\!\!\!\perp})}{\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp})} \cdot \frac{\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}}\,|\,\mathrm{E}_{\perp\!\!\!\perp}, \mathcal{D})}{\mathrm{P}(\mathrm{E}_{\not\perp\!\!\!\perp_{\mathcal{D}}}\,|\,\mathrm{E}_{\not\perp\!\!\!\perp}, \mathcal{D})} = 1 + \left(\frac{1}{g_0}\int_0^{+\infty} g_+(\tau)\frac{1 - F_{r_n, 2n\tau}(\chi^2_\alpha(r))}{1 - F_{r_n}(\chi^2_\alpha(r))}\mathrm{d}\tau\right)^{-1} \qquad \text{(A.7)}
\end{aligned}
$$

Therefore, (c) and (d) of Theorem 1 follow directly from (c) of Lemma 2. This completes the proof. □

A.2. Proof of Theorem 3

This appendix provides the proof of Theorem 3.

Before characterizing the property of EMB, the notion of *d-separation* [1,2] is briefly presented as follows. For a DAG $\mathbb{G}$ over $\boldsymbol{V}$, letting $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{V}$ be disjoint, we say $\boldsymbol{Z}$ d-separates $\boldsymbol{X}$ and $\boldsymbol{Y}$ if it blocks every path between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and if this is the case we write $\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}$. Here, $\boldsymbol{Z}$ blocking a path $\mathbb{p}$ means that $\mathbb{p}$ has an HT node or a TT node belonging to $\boldsymbol{Z}$, or that $\mathbb{p}$ has an HH node C such that C and its all descendants are not in $\boldsymbol{Z}$. As is well-known, $\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} \Leftrightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$, if the BN $(\mathbb{G}, \mathbb{P})$ satisfies the faithfulness condition [2]. This implication provides a convenient way of identifying CI relationships. For example, consider a BN with the graph presented in Figure 12 as its DAG. Then, X_2 and X_8 are d-separated by $\{X_4, X_5\}$, meaning $X_2 \perp X_8 \mid \{X_4, X_5\}$ and, thus, $X_2 \perp\!\!\!\perp X_8 \mid \{X_4, X_5\}$; X_3 and X_4 are d-separated by $\emptyset$, indicating $X_3 \perp X_4$, so $X_3 \perp\!\!\!\perp X_4$.

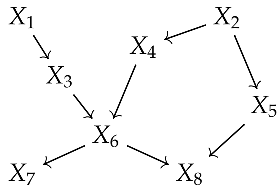


Figure 12. The DAG of the ASIA network used to illustrate the notions of d-separation.

In addition, the following properties are helpful [1, 16]. For any $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W} \subseteq \boldsymbol{V}$, we have (i) *decomposition*: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \cup \boldsymbol{W} \mid \boldsymbol{Z}$ implies $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ and $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{Z}$; (ii) *weak union*: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \cup \boldsymbol{W} \mid \boldsymbol{Z}$ implies $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \cup \boldsymbol{W}$; (iii) *contraction*: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \cup \boldsymbol{W}$ and $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{Z}$ imply $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \cup \boldsymbol{W} \mid \boldsymbol{Z}$. Further, under the faithfulness condition, besides (i)~(iii), we also have (iv) *intersection*: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \cup \boldsymbol{W}$ and $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{Z} \cup \boldsymbol{Y}$ imply $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \cup \boldsymbol{W} \mid \boldsymbol{Z}$; (v) *composition*: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ and $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{Z}$ imply $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \cup \boldsymbol{W} \mid \boldsymbol{Z}$.

Lemma 3. *Let $\mathbb{G}$ be a DAG over $\boldsymbol{V}$. The statements below hold* [2]*:*

- *For given $T, X \in \boldsymbol{V}$, we have $X \in \mathtt{PC}_T$ if, and only if, $T \not\perp\!\!\!\perp X \mid \boldsymbol{Z}$ holds for any $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{T, X\}$.*
- *For an uncoupled meeting "$T — Y — X$" (i.e., T and X are not adjacent), the following are equivalent:* (a) *Y is a collider of T and X;* (b) *there is a set of nodes not containing Y and its descendants that d-separates T and X;* (c) *any set of nodes containing Y or its a descendant does not d-separate T and X.* □

Lemma 4. *For a BN $(\mathbb{G}, \mathbb{P})$ over $\boldsymbol{V}$ satisfying the faithfulness condition, we have $\mathtt{MB}_T^{(Y)} \subseteq (\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}$.*

Proof. By the uniqueness of MB (under the faithfulness condition), it suffices to prove

$$T \perp\!\!\!\perp (\boldsymbol{V} \setminus \{Y\}) \setminus [(\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}] \setminus \{T\} \mid (\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}. \tag{A.8}$$

In fact, according to the definition of MB, we have

$$T \perp\!\!\!\perp \boldsymbol{V} \setminus \mathtt{MB}_T \setminus \{T\} \mid \mathtt{MB}_T, \tag{A.9}$$
$$Y \perp\!\!\!\perp \boldsymbol{V} \setminus \mathtt{MB}_Y \setminus \{Y\} \mid \mathtt{MB}_Y. \tag{A.10}$$

Putting now $\boldsymbol{M}_{TY} \triangleq (\mathtt{MB}_Y \setminus \mathtt{MB}_T \setminus \{T\}) \cup (\{Y\} \setminus \mathtt{MB}_T)$ and $\boldsymbol{M}_{YT} \triangleq (\mathtt{MB}_T \setminus \mathtt{MB}_Y \setminus \{Y\}) \cup (\{T\} \setminus \mathtt{MB}_Y)$, we employ the weak union property to obtain the following implications:

$$\begin{aligned}
\text{(A.9)} \quad &\Rightarrow \quad (\boldsymbol{V} \setminus \mathtt{MB}_T \setminus \{T\}) \setminus \boldsymbol{M}_{TY} \perp\!\!\!\perp T \mid \mathtt{MB}_T \cup \boldsymbol{M}_{TY} \\
&\Rightarrow \quad \boldsymbol{V} \setminus [(\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}] \setminus \{T, Y\} \perp\!\!\!\perp T \mid [(\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}] \cup \{Y\}, \quad \text{(A.11)} \\
\text{(A.10)} \quad &\Rightarrow \quad (\boldsymbol{V} \setminus \mathtt{MB}_Y \setminus \{Y\}) \setminus \boldsymbol{M}_{YT} \perp\!\!\!\perp Y \mid \mathtt{MB}_Y \cup \boldsymbol{M}_{YT} \\
&\Rightarrow \quad \boldsymbol{V} \setminus [(\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}] \setminus \{T, Y\} \perp\!\!\!\perp Y \mid [(\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}] \cup \{T\}. \quad \text{(A.12)}
\end{aligned}$$

Using the intersection property, (A.11) and (A.12) imply $\boldsymbol{V} \setminus [(\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}] \setminus \{T, Y\} \perp\!\!\!\perp \{T, Y\} \mid (\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}$, or, equivalently, $\{T, Y\} \perp\!\!\!\perp (\boldsymbol{V} \setminus \{Y\}) \setminus [(\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}] \setminus \{T\} \mid (\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}$. By the decomposition property, (A.8) follows immediately. The proof is completed. □

Theorem 3. *For a BN $(\mathbb{G}, \mathbb{P})$ over $\boldsymbol{V}$ satisfying the faithfulness condition, the following statements hold:*

- *In one of the following cases:* (i) $Y \in \mathtt{PA}_T$, (ii) $Y \in \mathtt{CH}_T$ *with* $\mathtt{CH}_Y \neq \emptyset$, (iii) $Y \in \mathtt{SP}_T$, *the* Y*-EMB of* T *can be expressed as* $\mathtt{MB}_T^{(Y)} = (\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}$.
- *If* $Y \in \mathtt{CH}_T$ *with* $\mathtt{CH}_Y = \emptyset$, *then* $\mathtt{PC}_T \setminus \{Y\} \subseteq \mathtt{MB}_T^{(Y)} \subseteq \mathtt{MB}_T \setminus \{Y\}$.
- *If* $Y \notin \mathtt{MB}_T$, *then* $\mathtt{MB}_T^{(Y)} = \mathtt{MB}_T$.

Proof. First, by Definition 1 and Lemma 1, it is readily seen that $\mathtt{PC}_T \setminus \{Y\} \subseteq \mathtt{MB}_T^{(Y)}$ holds in any case. Here, we recall that $\mathtt{SP}_T$ denotes the set of the spouses of T, excluding its parents and children. Based on the d-separation theory (under the faithfulness condition) and Lemma 3, we have

- **Case 1.** $Y \in \mathtt{PA}_T$. Clearly, $\mathtt{SP}_T \subseteq \mathtt{MB}_T^{(Y)}$ because $Y \notin \mathtt{CH}_T \subseteq \mathtt{MB}_T^{(Y)}$, in view of the acyclicity of a DAG.

  Now, we show $\mathtt{MB}_Y \setminus \{T\} \subseteq \mathtt{MB}_T^{(Y)}$. In fact, Y is the only HT node in the path "$P \to Y \to T$", while it is the only TT node in the path "$C \leftarrow Y \to T$", where $P \in \mathtt{PA}_Y$ and $C \in \mathtt{CH}_Y$. Hence, any node set without Y cannot d-separate T and $\mathtt{PC}_Y \setminus \{T\}$, so $\mathtt{PC}_Y \setminus \{T\} \subseteq \mathtt{MB}_T^{(Y)}$. Further, any $C \in \mathtt{CH}_Y \setminus \{T\}$ is an HH node in the path "$S \to C \leftarrow Y \to T$" for $S \in \mathtt{SP}_Y\{T\}$; also, Y is the only TT node in this path. This indicates that any set of nodes containing $\mathtt{CH}_Y \setminus \{T\}$ but not containing Y cannot d-separate T and $\mathtt{SP}_Y$. Therefore, $\mathtt{SP}_Y \subseteq \mathtt{MB}_T^{(Y)}$.

  In summary, $\mathtt{MB}_T^{(Y)} \supseteq (\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}$. On the other hand, $\mathtt{MB}_T^{(Y)} \subseteq (\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}$ follows from Lemma 4, so we have shown $\mathtt{MB}_T^{(Y)} = (\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}$ in this case.
- **Case 2.** $Y \in \mathtt{CH}_T$ **with** $\mathtt{CH}_Y \neq \emptyset$. In this case, Y is the only HT node in the path "$T \to Y \to C$" for $C \in \mathtt{CH}_Y$. Hence, any set of nodes without Y cannot d-separate T and $\mathtt{CH}_Y$, so $\mathtt{CH}_Y \subseteq \mathtt{MB}_T^{(Y)}$. Further, C is an HH node in the path "$T \to Y \to C \leftarrow S$" for $S \in \mathtt{SP}_Y \setminus \{T\}$, while Y is the only HT node in this path. Thus, any set of nodes containing $\mathtt{CH}_Y$ but not containing Y cannot d-separate T and $\mathtt{SP}_Y$, indicating $\mathtt{SP}_Y \subseteq \mathtt{MB}_T^{(Y)}$. Moreover, it can be readily concluded that

$$\left.\begin{array}{lll} \mathtt{CH}_T \setminus \{Y\} \subseteq \mathtt{MB}_T^{(Y)} & \Rightarrow & \mathtt{SP}_T \setminus \mathtt{PA}_Y \subseteq \mathtt{MB}_T^{(Y)} \\ \mathtt{CH}_Y \neq \emptyset & \Rightarrow & \mathtt{PA}_Y \setminus \{T\} \subseteq \mathtt{MB}_T^{(Y)} \end{array}\right\} \Rightarrow \mathtt{SP}_T \cup (\mathtt{PA}_Y \setminus \{T\}) \subseteq \mathtt{MB}_T^{(Y)}.$$

  The above analysis combined with Lemma 4 means $\mathtt{MB}_T^{(Y)} = (\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}$ in this case.
- **Case 3.** $Y \in \mathtt{CH}_T$ **with** $\mathtt{CH}_Y = \emptyset$. By Lemma 4, $\mathtt{MB}_T^{(Y)} \subseteq (\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}$. On the other hand, $\mathtt{CH}_Y = \emptyset$ implies $\mathtt{SP}_Y = \emptyset$ and, thus, $\mathtt{MB}_Y = \mathtt{PA}_Y \subseteq \mathtt{SP}_T \cup \{T\}$, since $Y \in \mathtt{CH}_T$. Consequently, $\mathtt{MB}_T^{(Y)} \subseteq \mathtt{MB}_T \setminus \{Y\}$.
- **Case 4.** $Y \in \mathtt{SP}_T$. In this case, T and Y have a common child, C. That is, $T \to C \leftarrow Y$; C is an HH node. Similar to the case of $Y \in \mathtt{PA}_T$, it can be easily shown that $\mathtt{MB}_T^{(Y)} = (\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}$, in view of $C \in \mathtt{PC}_T = \mathtt{PC}_T \setminus \{Y\} \subseteq \mathtt{MB}_T^{(Y)}$.
- **Case 5.** $Y \notin \mathtt{MB}_T$. In this case, it is not hard to see $\mathtt{MB}_T \subseteq \mathtt{MB}_T^{(Y)}$. Using the definition of MB, we obtain $T \perp\!\!\!\perp \boldsymbol{V} \setminus \mathtt{MB}_T \setminus \{T\} \mid \mathtt{MB}_T$, which combined with decomposition gives $T \perp\!\!\!\perp (\boldsymbol{V} \setminus \{Y\}) \setminus \mathtt{MB}_T \setminus \{T\} \mid \mathtt{MB}_T$ since $Y \notin \mathtt{MB}_T$, indicating $\mathtt{MB}_T^{(Y)} \subseteq \mathtt{MB}_T$. Thus, $\mathtt{MB}_T^{(Y)} = \mathtt{MB}_T$.

The proof is completed. □

A.3. Proof of Theorem 4

This appendix provides the proof of Theorem 4.

Theorem 4. *For a BN* $(\mathbb{G}, \mathbb{P})$ *over* $\boldsymbol{V}$ *satisfying the faithfulness condition, let* $\mathtt{MB}_T^{(Y)}$ *be the* Y*-EMB of* T*, and* $\boldsymbol{M} \subseteq \mathtt{MB}_T^{(Y)}$ *subject to* $(\mathtt{MB}_T^{(Y)} \setminus \boldsymbol{M}) \cap \mathtt{MB}_T = \emptyset$. *Then, for any* $X \in \boldsymbol{M}$, *we have* $X \notin \mathtt{MB}_T \Leftrightarrow T \perp\!\!\!\perp X \mid (\boldsymbol{M} \setminus \{X\}) \cup \{Y\}$.

Proof. We first show the sufficiency. Clearly, $X \notin \mathtt{PC}_T$ in view of Lemma 3. Suppose $X \in \mathtt{SP}_T$, meaning there is $Z \in \mathtt{CH}_T$ such that T, Z, and X constitute a collision "$T \rightarrow Z \leftarrow X$". Then, it follows from Lemma 3 that any set, $\boldsymbol{N}$, of nodes containing Z cannot d-separates T and X. That is, $T \not\perp X \mid \boldsymbol{N}$, so $T \not\perp\!\!\!\perp X \mid \boldsymbol{N}$. On the other hand, according to Theorem 3, $\mathtt{MB}_T^{(Y)} \supseteq \mathtt{PC}_T \backslash \{Y\}$, so $(\boldsymbol{M} \backslash \{X\}) \cup \{Y\} \supseteq \mathtt{PC}_T \supseteq \{Z\}$ since $(\mathtt{MB}_T^{(Y)} \setminus \boldsymbol{M}) \cap \mathtt{MB}_T = \emptyset$ and $X \notin \mathtt{PC}_T$. Consequently, $T \not\perp\!\!\!\perp X \mid (\boldsymbol{M} \setminus \{X\}) \cup \{Y\}$. This contradicts $T \perp\!\!\!\perp X \mid (\boldsymbol{M} \setminus \{X\}) \cup \{Y\}$, implying $X \notin \mathtt{SP}_T$. This combined with $X \notin \mathtt{PC}_T$ shows $X \notin \mathtt{MB}_T$.

To prove the necessity, we assume $X \notin \mathtt{MB}_T$. By Theorem 3, we have

- If $Y \in \mathtt{PA}_T \cup \mathtt{SP}_T$ or $Y \in \mathtt{CH}_T$ but $\mathtt{CH}_Y \neq \emptyset$, then $\mathtt{MB}_T^{(Y)} = (\mathtt{MB}_T \cup \mathtt{MB}_Y) \setminus \{T, Y\}$. This combined with $(\mathtt{MB}_T^{(Y)} \setminus \boldsymbol{M}) \cap \mathtt{MB}_T = \emptyset$ and $X \notin \mathtt{MB}_T$ implies $(\boldsymbol{M} \setminus \{X\}) \cup \{Y\} \supseteq \mathtt{MB}_T$; that is, $(\boldsymbol{M} \setminus \{X\}) \cup \{Y\}$ is a Markov blanket of T. By the weak union property and the decomposition property, it is readily concluded that $T \perp\!\!\!\perp X \mid (\boldsymbol{M} \setminus \{X\}) \cup \{Y\}$ in these cases.
- If $Y \notin \mathtt{MB}_T$ or $Y \in \mathtt{CH}_T$ but $\mathtt{CH}_Y = \emptyset$, we have $\boldsymbol{M} \subseteq \mathtt{MB}_T^{(Y)} \subseteq \mathtt{MB}_T$. Therefore, $\boldsymbol{M} \equiv \mathtt{MB}_T^{(Y)}$ and no such an X exists in these two cases, since $(\mathtt{MB}_T^{(Y)} \setminus \boldsymbol{M}) \cap \mathtt{MB}_T = \emptyset$.

The proof of the necessity is also completed. □

A.4. Proof of Theorem 5

Before proving Theorem 5, we first show the following lemma:

Lemma 5. *If $T \not\perp\!\!\!\perp X \mid \boldsymbol{M}$ and $T \perp\!\!\!\perp Y \mid \boldsymbol{M}$, then $T \not\perp\!\!\!\perp X \mid \boldsymbol{M} \cup \{Y\}$.*

Proof. Supposing $T \perp\!\!\!\perp X \mid \boldsymbol{M} \cup \{Y\}$, by the contraction property, this combined with $T \perp\!\!\!\perp Y \mid \boldsymbol{M}$ gives $T \perp\!\!\!\perp \{X, Y\} \mid \boldsymbol{M}$. By the decomposition property, we obtain $T \perp\!\!\!\perp X \mid \boldsymbol{M}$, which contradicts $T \not\perp\!\!\!\perp X \mid \boldsymbol{M}$. □

Theorem 5. *For $T \in \boldsymbol{V}$ and $\boldsymbol{M} \subseteq \boldsymbol{V} \setminus \{T\}$, put $\mathcal{X}_\ell \triangleq \{X_1, \cdots, X_\ell\} \subseteq \boldsymbol{M}$ and $\boldsymbol{M}_\ell \triangleq \boldsymbol{M} \setminus \mathcal{X}_\ell$, in which each X_ℓ is subject to the $\perp\!\!\!\perp_{\mathcal{D}}$-test "$T \perp\!\!\!\perp_{\mathcal{D}} X_\ell \mid \boldsymbol{M}_\ell$", $\ell = 1, \cdots, k$. Then, for any $X_i \in \mathcal{X}_{k-1}$, the $\perp\!\!\!\perp_{\mathcal{D}}$-test "$T \perp\!\!\!\perp_{\mathcal{D}} X_i \mid \boldsymbol{M}_i$" is unreliable under the assumption $\mathscr{A}_2$, if $T \not\perp\!\!\!\perp_{\mathcal{D}} X_i \mid \boldsymbol{M}_k$.*

Proof. First of all, all $\not\perp\!\!\!\perp_{\mathcal{D}}$-tests are deemed reliable in view of $\mathscr{A}_2$, so we have

$$T \not\perp\!\!\!\perp_{\mathcal{D}} X_i \mid \boldsymbol{M}_k \Rightarrow T \not\perp\!\!\!\perp X_i \mid \boldsymbol{M}_k. \tag{A.13}$$

Now we show "$T \perp\!\!\!\perp_{\mathcal{D}} X_i \mid \boldsymbol{M}_i$" is incompatible to the $\perp\!\!\!\perp_{\mathcal{D}}$-tests "$T \not\perp\!\!\!\perp_{\mathcal{D}} X_\ell \mid \boldsymbol{M}_\ell$" for $\ell = i+1, \cdots, k$. In fact, assuming these $(k-i)$ $\perp\!\!\!\perp_{\mathcal{D}}$-tests are reliable, it follows from Lemma 5 and (A.13) that

$$\left.\begin{array}{l} T \not\perp\!\!\!\perp X_i \mid \boldsymbol{M}_k \\ T \perp\!\!\!\perp X_k \mid \boldsymbol{M}_k \end{array}\right\} \begin{array}{l} \Rightarrow \quad T \not\perp\!\!\!\perp X_i \mid \boldsymbol{M}_k \cup \{X_k\} \quad (\text{noting } \boldsymbol{M}_k \cup \{X_k\} = \boldsymbol{M}_{k-1}) \end{array}$$
$$\begin{aligned} &\Rightarrow \quad T \not\perp\!\!\!\perp X_i \mid \boldsymbol{M}_{k-1} \quad (\text{combined with } T \perp\!\!\!\perp X_{k-1} \mid \boldsymbol{M}_{k-1}) \\ &\Rightarrow \quad \cdots \\ &\Rightarrow \quad T \not\perp\!\!\!\perp X_i \mid \boldsymbol{M}_{i+1} \quad (\text{combined with } T \perp\!\!\!\perp X_{i+1} \mid \boldsymbol{M}_{i+1}) \\ &\Rightarrow \quad T \not\perp\!\!\!\perp X_i \mid \boldsymbol{M}_i. \end{aligned}$$

This proves "$T \perp\!\!\!\perp_{\mathcal{D}} X_i \mid \boldsymbol{M}_i$" is incompatible to "$T \not\perp\!\!\!\perp_{\mathcal{D}} X_\ell \mid \boldsymbol{M}_\ell$" ($\ell = i+1, \cdots, k$). Observe that these $(1+k-i)$ $\perp\!\!\!\perp_{\mathcal{D}}$-tests have

$$r_i \triangleq (r_T - 1)(r_{X_i - 1}) r_{\boldsymbol{M}_i} = (r_T - 1)(r_{X_i - 1}) r_{X_{i+1}} r_{\boldsymbol{M}_{i+1}}$$

$$\begin{aligned} r_{i+1} &\triangleq (r_T-1)(r_{X_{i+1}-1})r_{\boldsymbol{M}_{i+1}}\ (<r_i) \\ &\vdots \\ r_k &\triangleq (r_T-1)(r_{X_{k-1}})r_{\boldsymbol{M}_k}\ (<r_{k-1}), \end{aligned}$$

degrees of freedom, respectively. According to the assumtion $\mathscr{A}_2$, the $\perp\!\!\!\perp_{\mathcal{D}}$-test "$T \perp\!\!\!\perp_{\mathcal{D}} X_i \mid \boldsymbol{M}_i$" is deemed unreliable. The proof is completed. $\square$

B. Figures

This appendix displays the figures derived in Section 6.

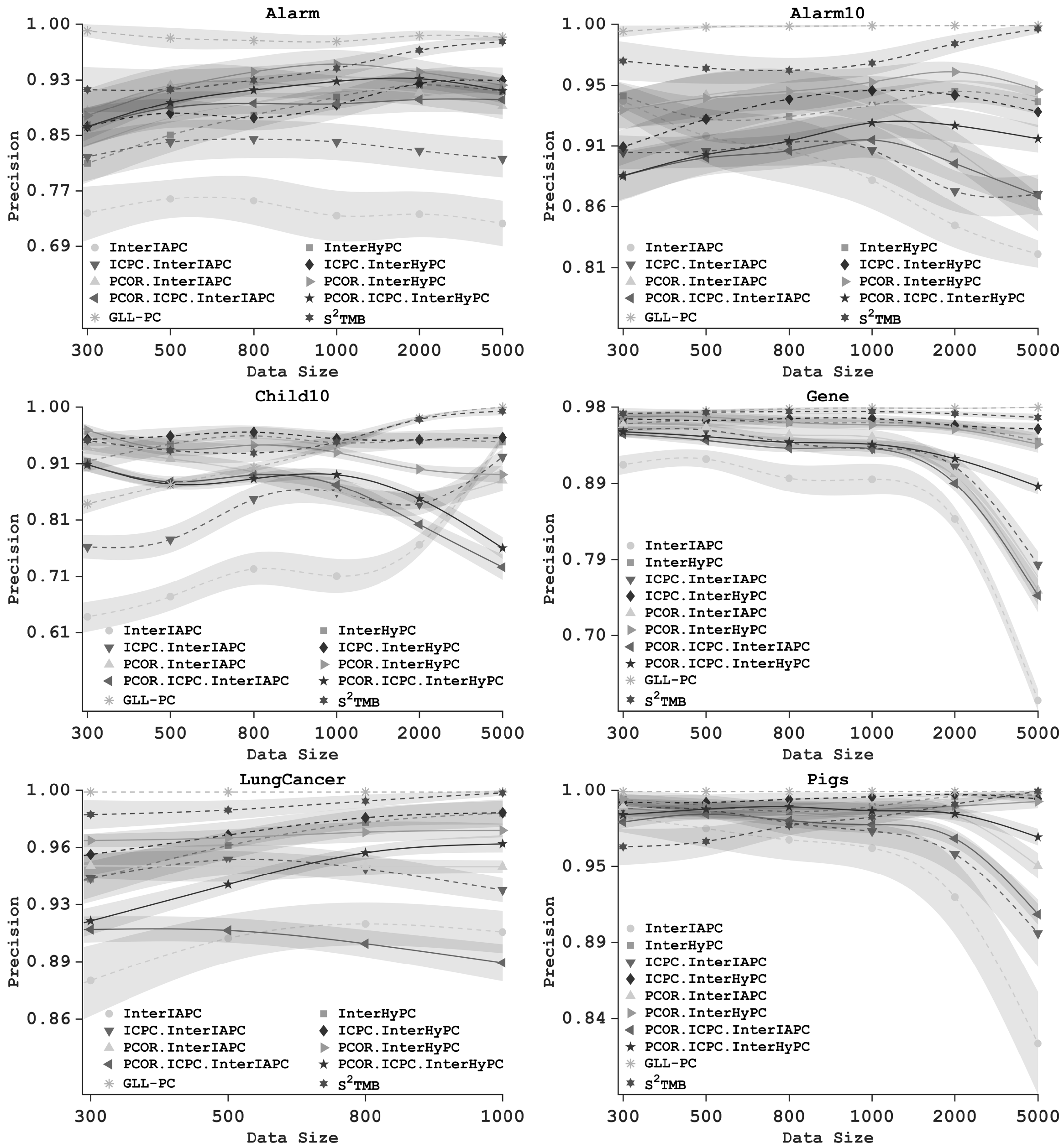


Figure 13. Average precision of PC algorithms versus data size.

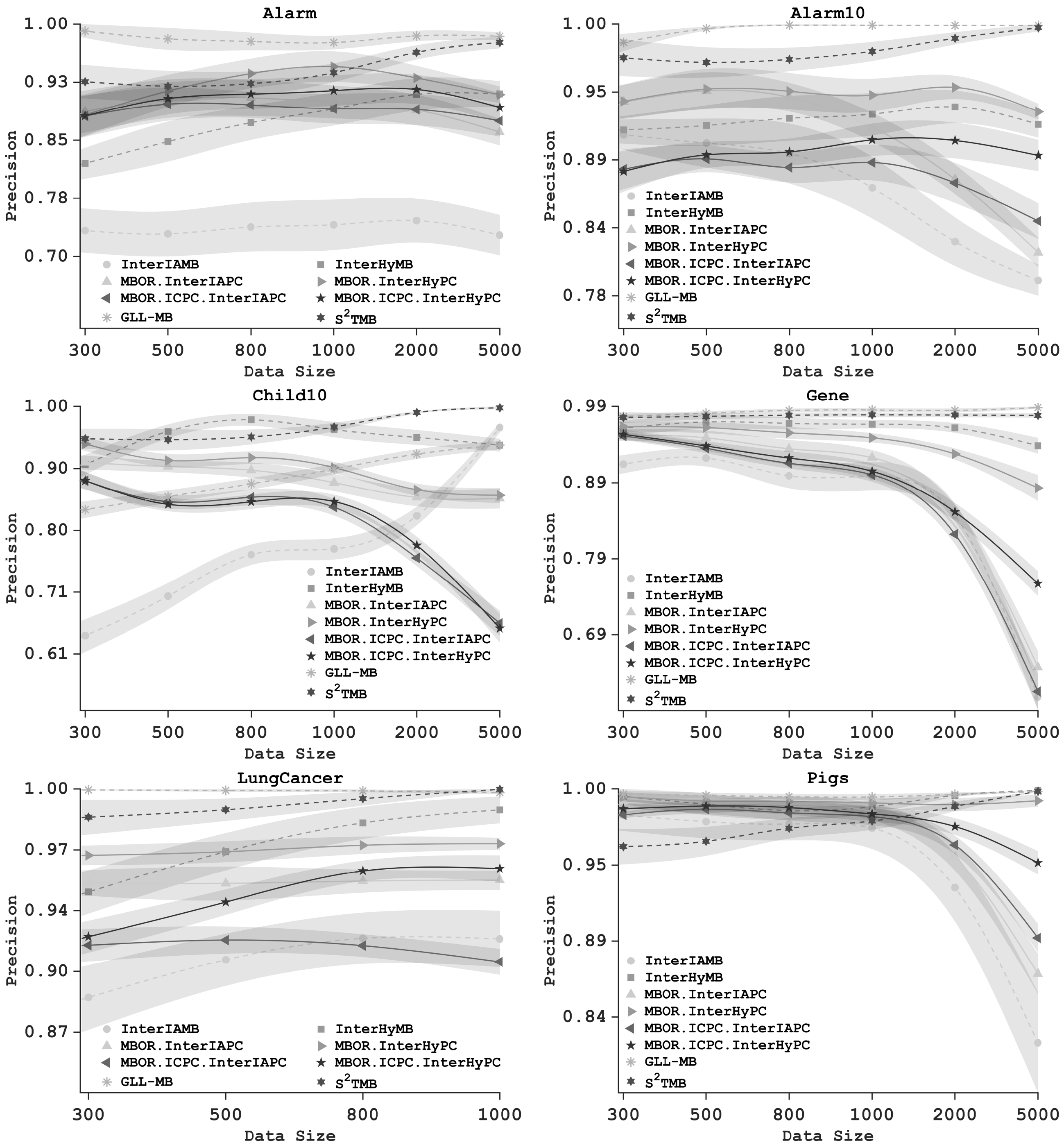


Figure 14. Average precision of MB algorithms versus data size.

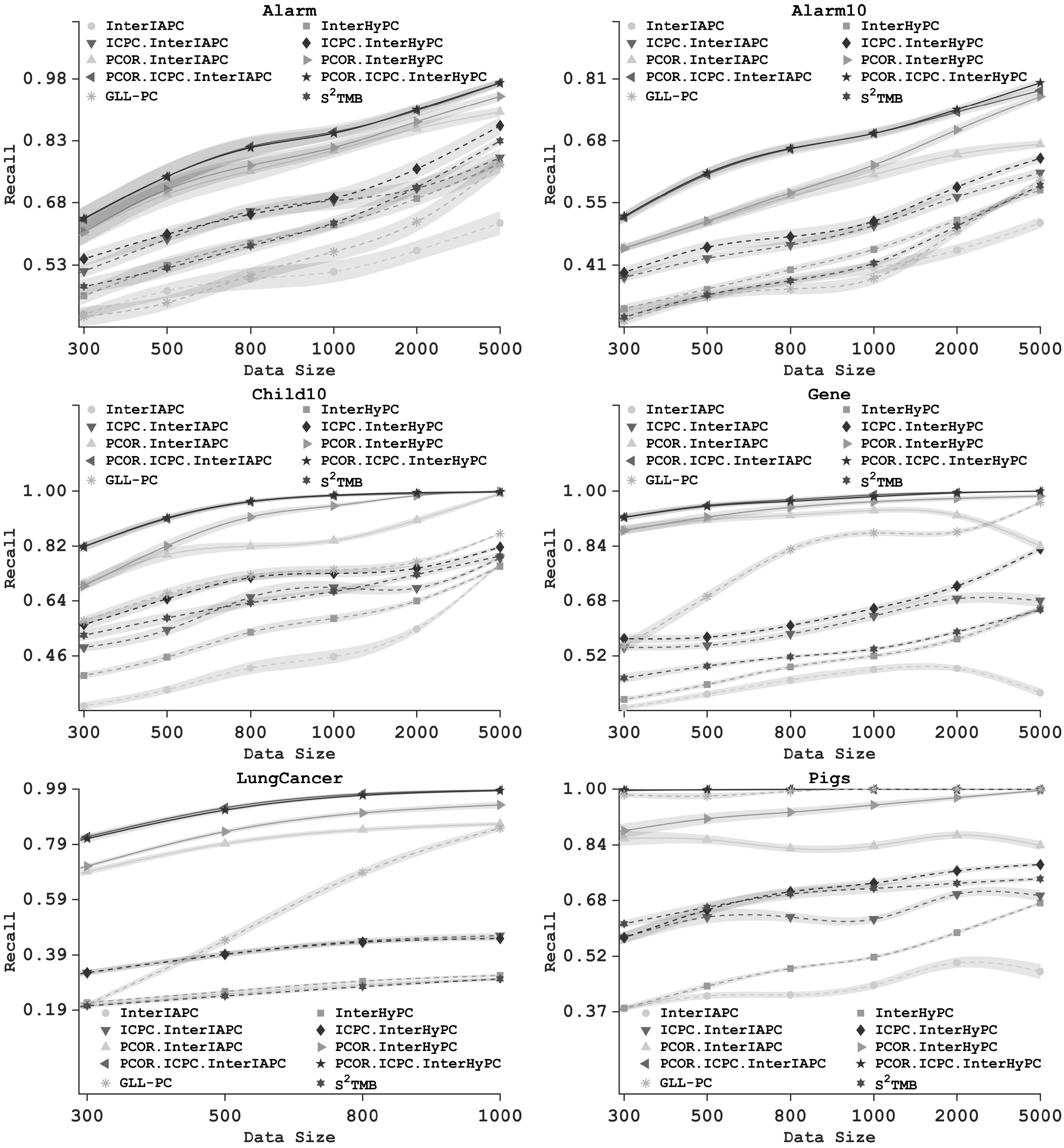


Figure 15. Average recall of PC algorithms versus data size.

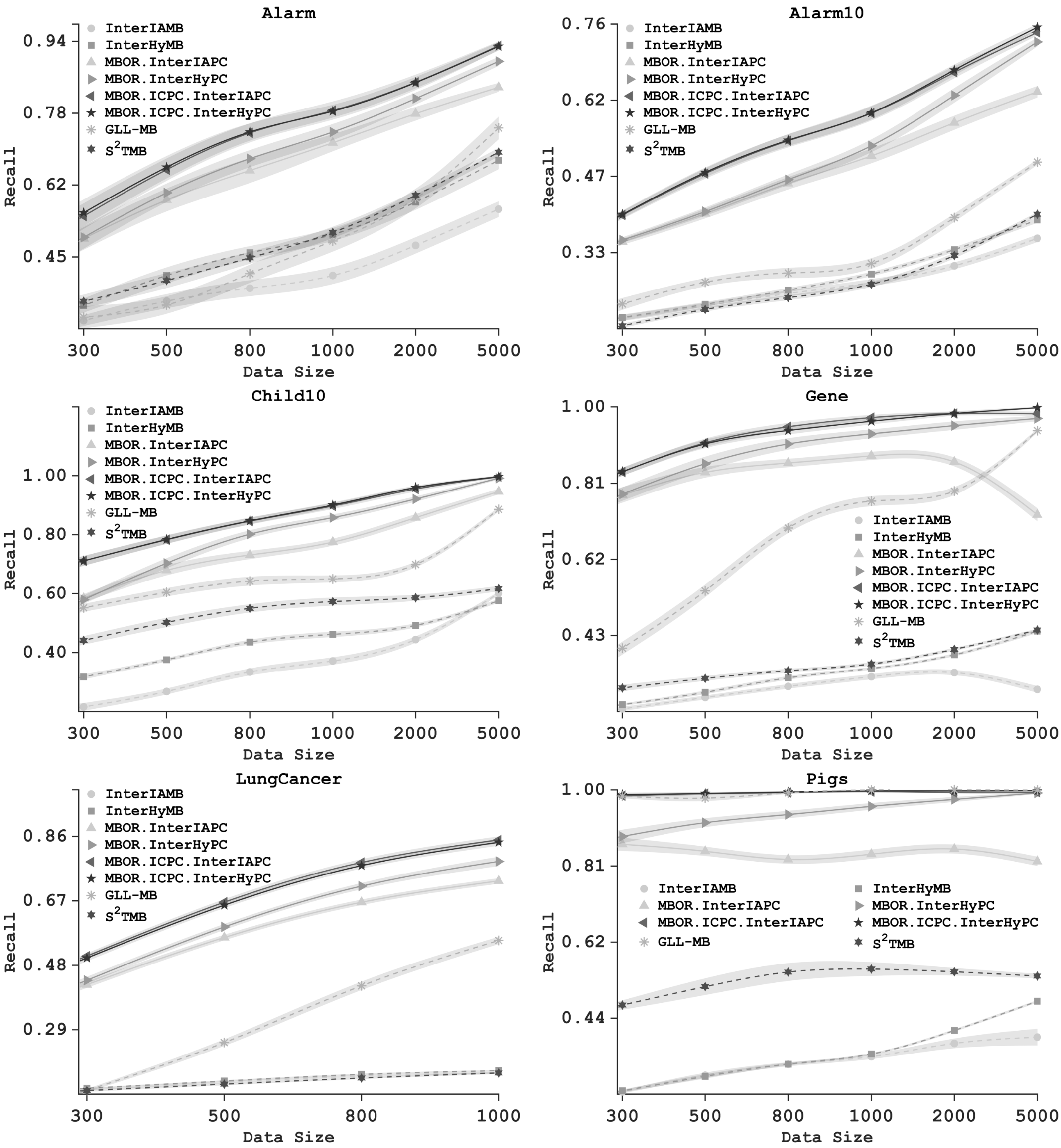


Figure 16. Average recall of MB algorithms versus data size.

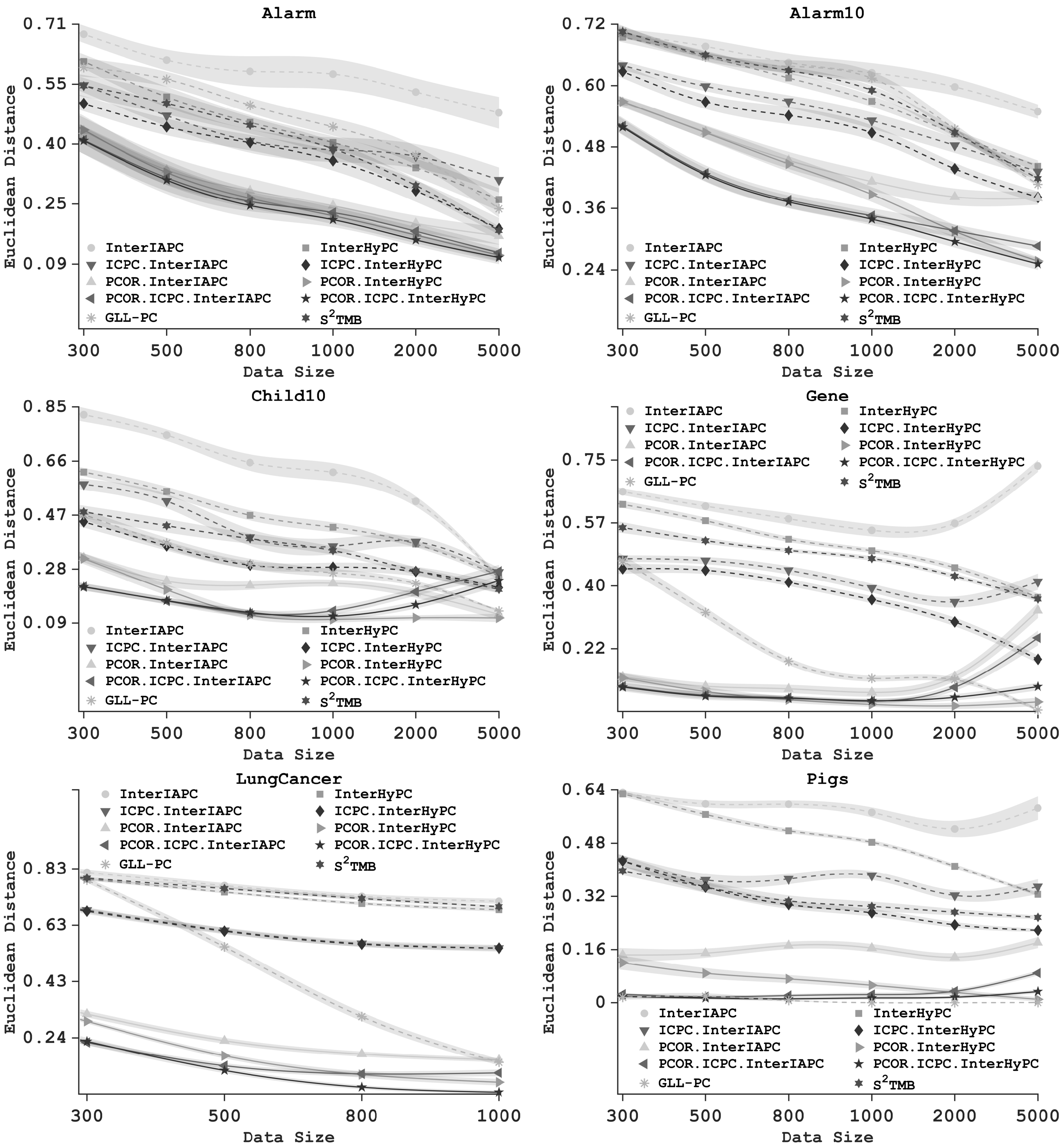


Figure 17. Average Euclidean distance of PC algorithms versus data size.

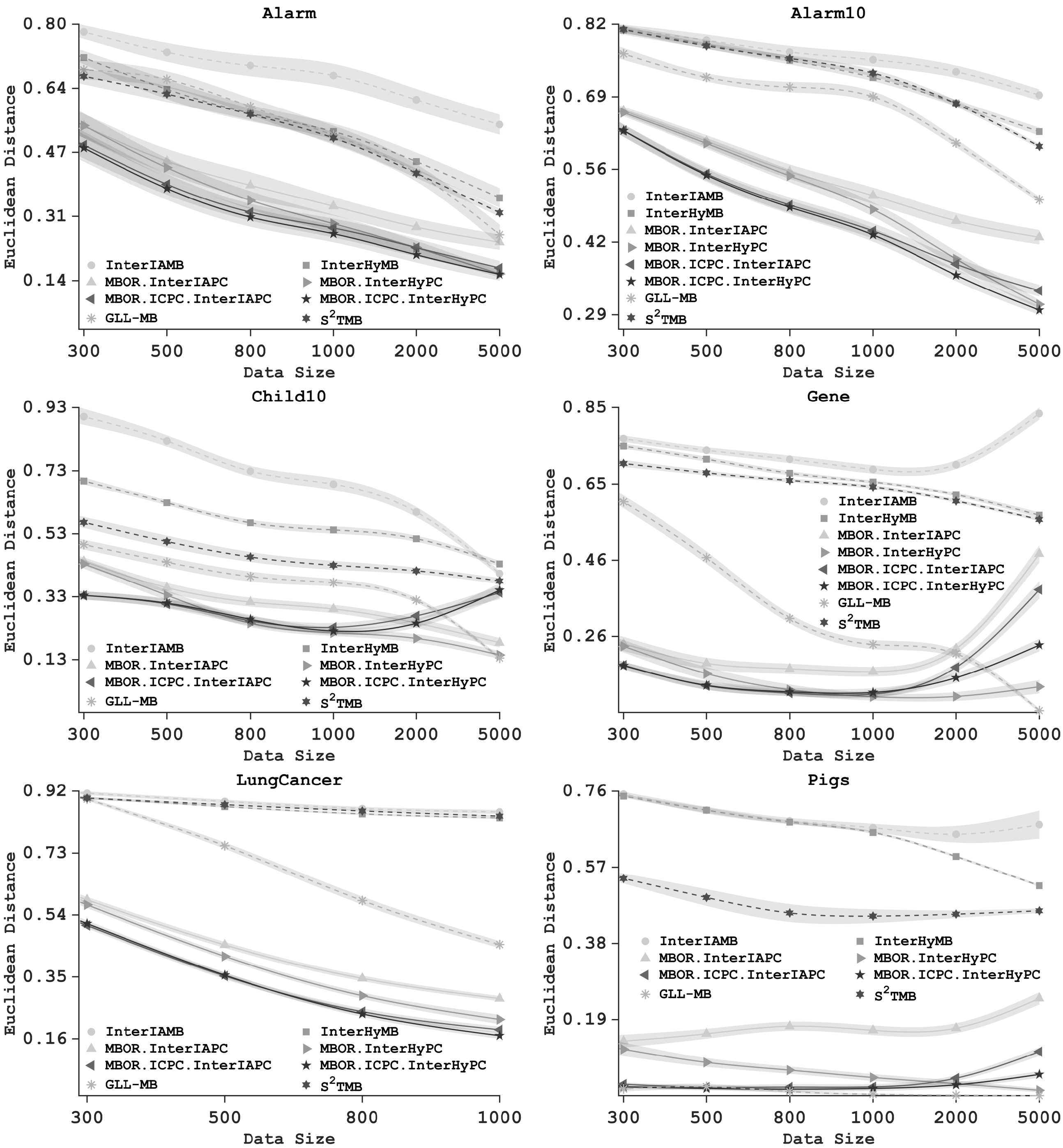


Figure 18. Average Euclidean distance of MB algorithms versus data size.

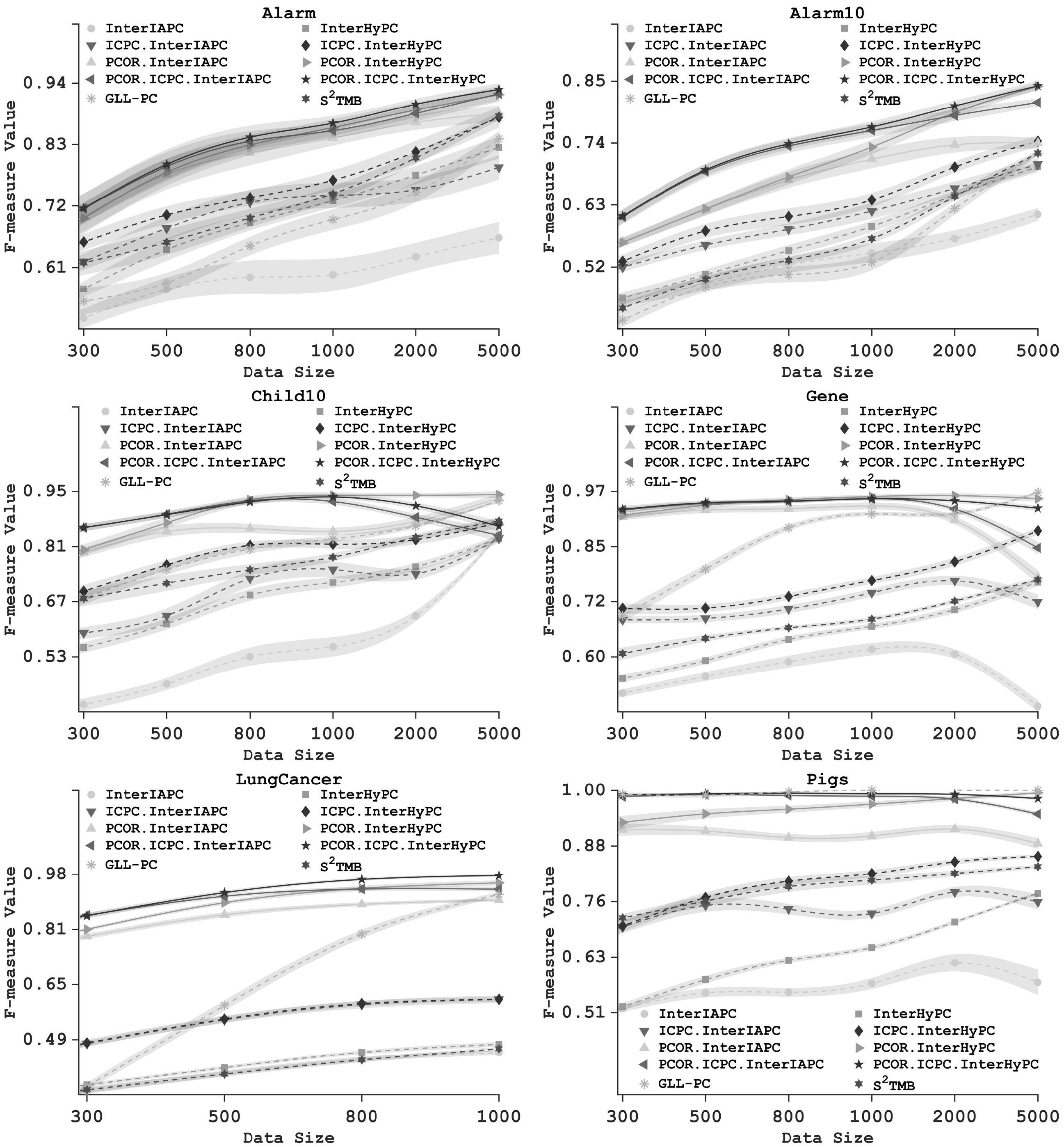


Figure 19. Average F-measure value of PC algorithms versus data size.

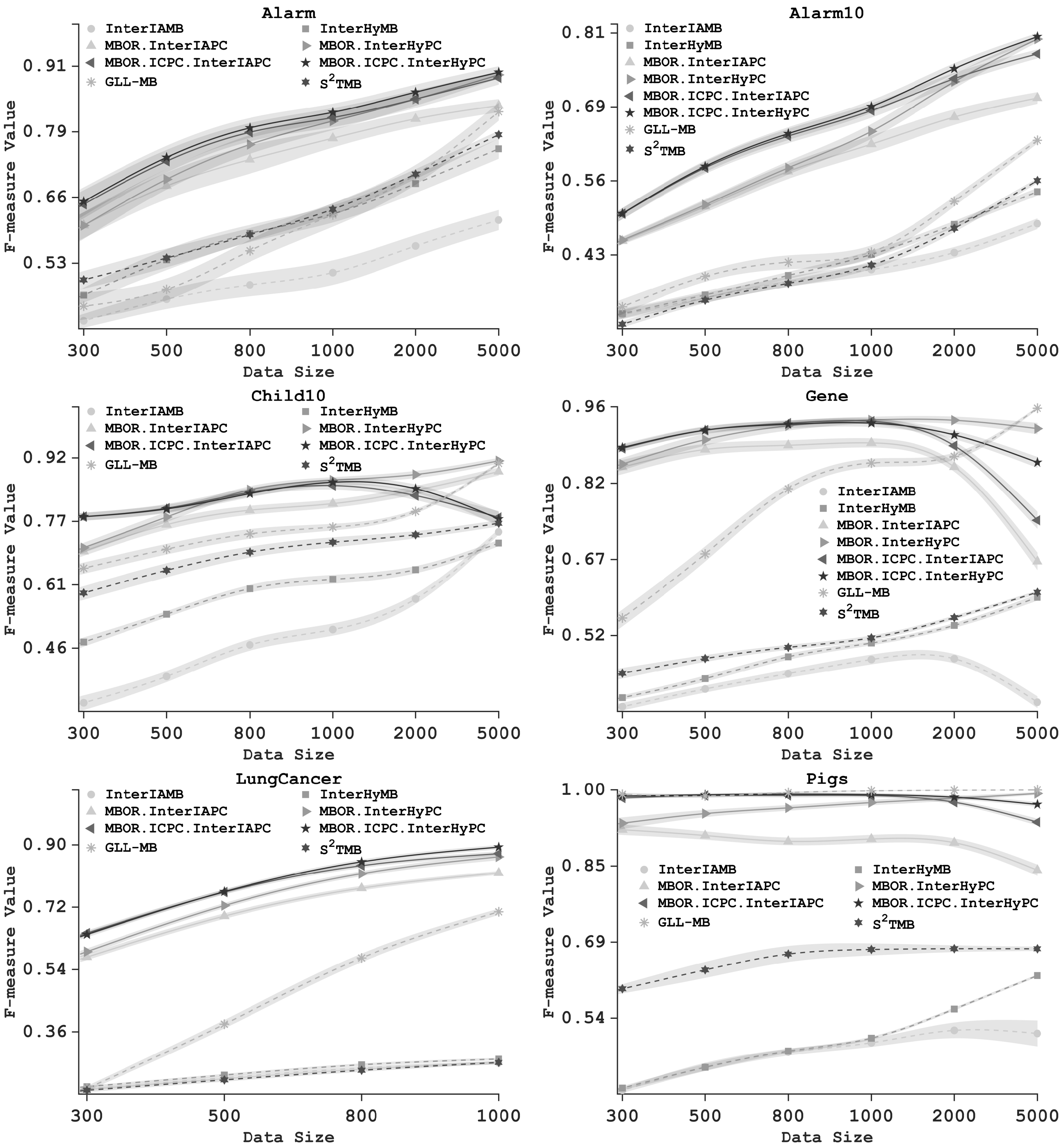


Figure 20. Average F-measure value of MB algorithms versus data size.

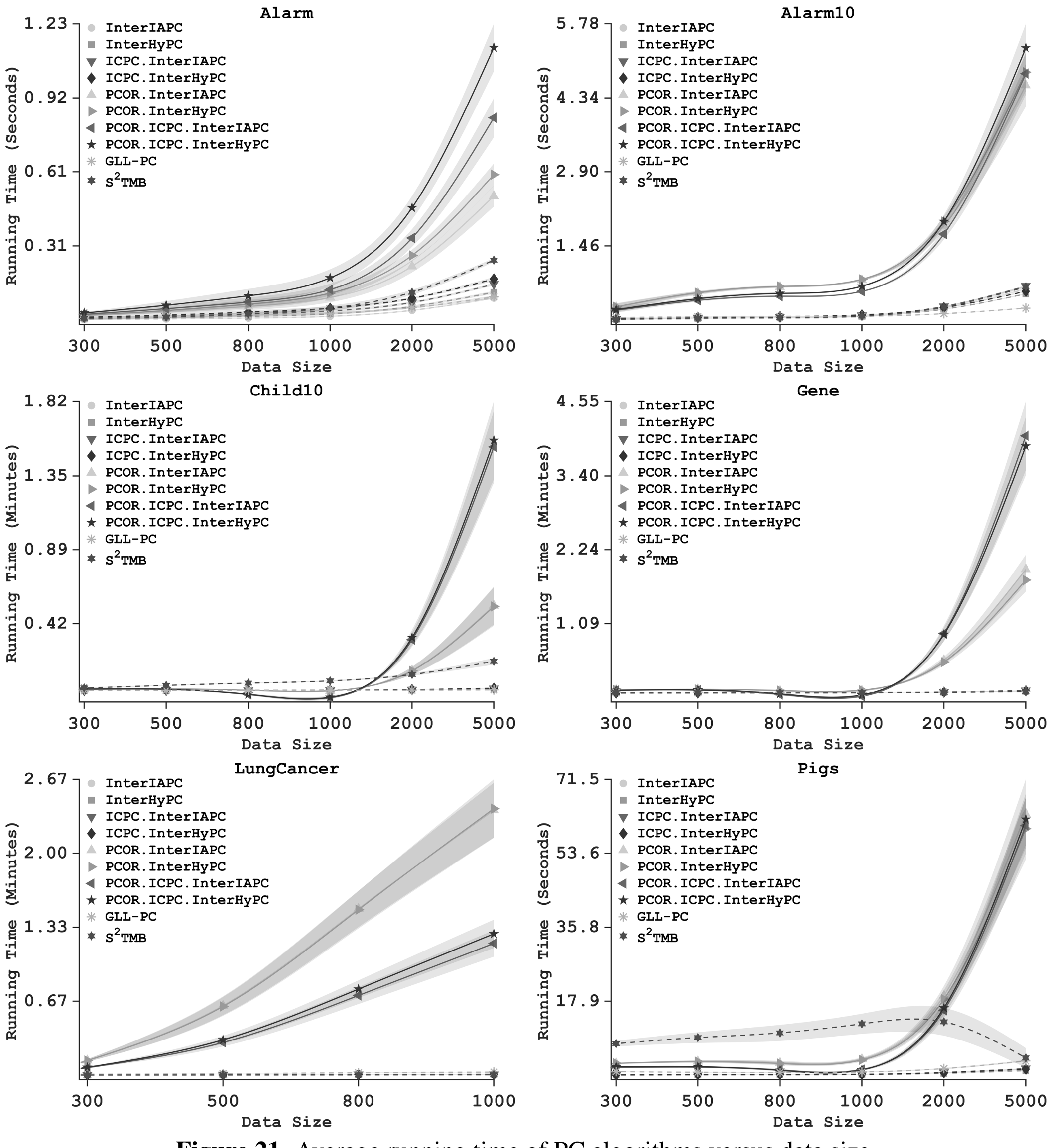


Figure 21. Average running time of PC algorithms versus data size.

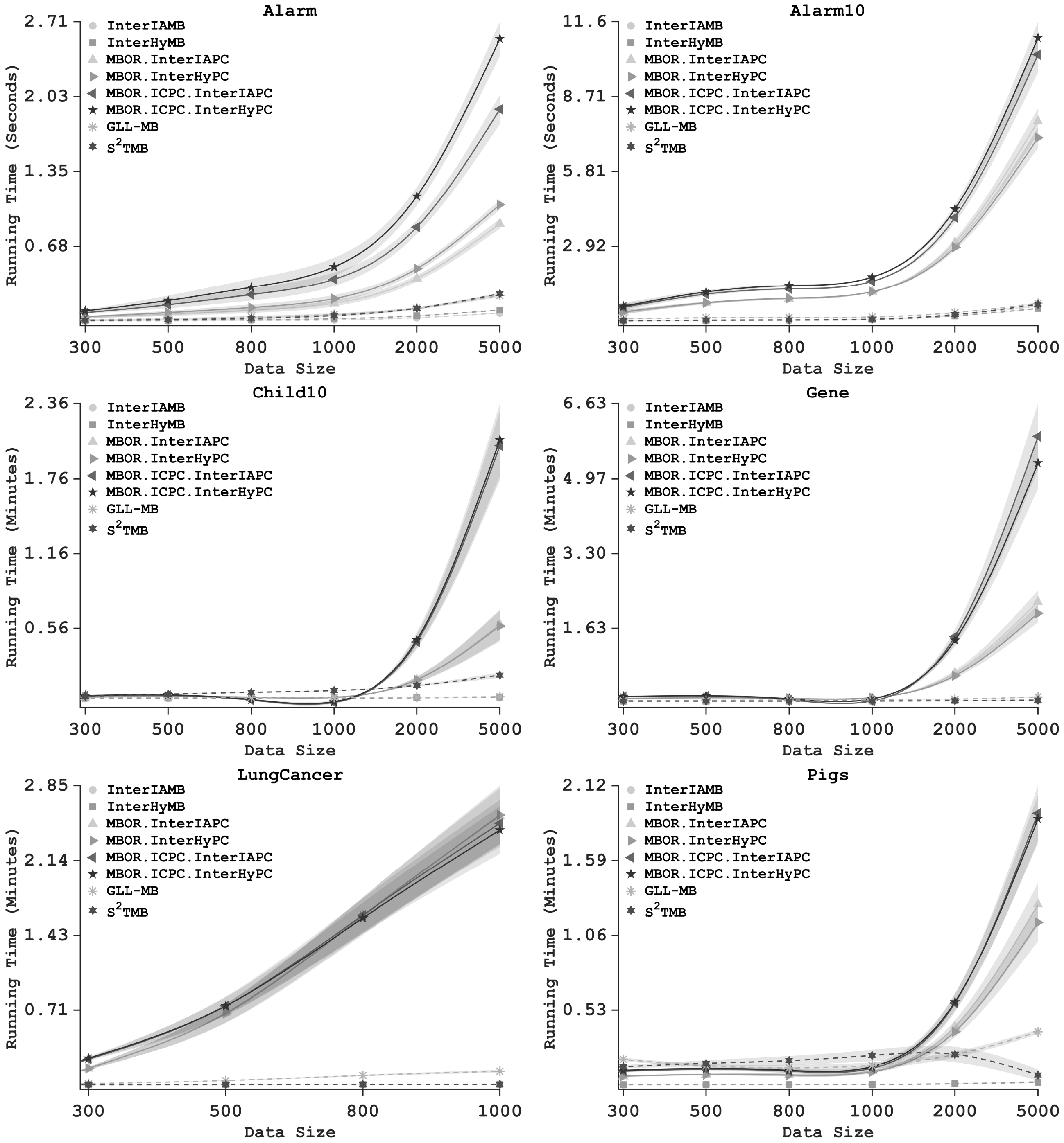


Figure 22. Average running time of MB algorithms versus data size.

C. Acronyms

BFMB	breadth first search of Markov boundary algorithm [37].
BIC	Bayesian information criterion [53].
BN	Bayesian network.
CI	conditional independence.

DAG	directed acyclic graph.
EMB	extended Markov boundary (Definition 1).
FN	false negative.
FP	false positive.
FSMB	fast shrinking parents-children learning for Markov blanket-based feature selection algorithm [44].
FullBNT	BN toolbox created by Prof. Kevin P. Murphy [45]; see http://www.cs.ubc.ca/~murphyk/ for details.
GLL	generalized local learning: an algorithmic framework proposed by Aliferis et al. [8,9], used for local causal discovery and feature selection.
GLL-MB	a sub-framework of GLL used for MB discovery [8,9]; see (d) of Algorithm 1 for details.
GLL-PC	a sub-framework of GLL used for PC discovery [8,9]; see (a) of Algorithm 1 for details.
GS	grow-shrink algorithm [40,41].
HH	head-to-head (converging) node.
hps	heuristic power size as a parameter in GLL of Aliferis et al. [8,9].
HT	head-to-tail (serial) node.
IAMB	incremental association Markov boundary algorithm [42].
IC	information connection based method.
ICPC	our PC discovery algorithm based on the IC strategy (Algorithm 2).
InterHyMB	our MB discovery algorithm as a combination of GLL-PC and InterIAMB; see Algorithm 3 for details.
InterHyPC	our PC discovery algorithm as a combination of GLL-PC and InterIAPC; see Algorithm 3 for details.
InterIAMB	interleaved incremental association MB algorithm [42]; see (b) of Algorithm 1 for details.
InterIAPC	InterIAMB-based PC discovery algorithm.
KS	Koller-Sahami algorithm [39].
LRH	an MB discovery algorithm used to overcome swamping and masking [17].
MB	Markov boundary.
MBOR	an MB discovery algorithm proposed by Morais and Aussem [10]; see (e) of Algorithm 1 for details.
MIToolbox	a toolbox developed by Brown et al. [50] for Shannon's information theory functions; Please refer to http://www.cs.man.ac.uk/~pococka4/MIToolbox.html for the details.
PC	parents and children.
PCMB	parents and children based MB algorithm proposed by Peña et al. [23].
PCOR	an PC discovery algorithm proposed by Morais and Aussem [10]; see (c) of Algorithm 1 for details.
SGS	a PC discovery algorithm proposed by Spirtes, Glymour, and Scheines [36].
S^{2}TMB	score-based simultaneous Markov blanket discovery algorithm [46].
TN	true negative.
TP	true positive.
TPDA	three-phase dependency analysis algorithm [24].
TT	tail-to-tail (diverging) node.

Author Contributions

Jianying Rong: Conceptualization, Methodology, Designing algorithms, Formal analysis, Writing original draft, Making major revisions; Xuqing Liu: Writing programs, Formal analysis, Writing original draft. All authors have read and approved the final version of the manuscript for publication.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors are very grateful to the three anonymous reviewers for their valuable comments and constructive suggestions, which were helpful in improving the paper.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. J. Pearl, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, San Francisco: Morgan Kaufmann, 1988.
2. R. E. Neapolitan, *Learning bayesian networks*, Upper Saddle River: Prentice Hall, 2004.
3. R. Daly, Q. Shen, S. Aitken, Learning bayesian networks: Approaches and issues, *Knowl. Eng. Rev.*, **26** (2011), 99–157. https://doi.org/10.1017/S0269888910000251
4. P. Parviainen, M. Koivisto, Finding optimal bayesian networks using precedence constraints, *J. Mach. Learn. Res.*, **14** (2013), 1387–1415. https://www.jmlr.org/papers/volume14/parviainen13a/parviainen13a.pdf
5. L. W. Zhang, H. P. Guo, *Introduction to bayesian networks*, Beijing: Science Press, 2006.
6. N. Friedman, I. Nachman, D. Peér, Learning Bayesian network structure from massive datasets: The "sparse candidate" algorithm, *arXiv Preprint*, 2013.
7. I. Tsamardinos, L. E. Brown, C. F. Aliferis, The max-min hill-climbing Bayesian network structure learning algorithm, *Mach. Learn.*, **65** (2006), 31–78. https://doi.org/10.1007/s10994-006-6889-7
8. C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, X. D. Koutsoukos, Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation, *J. Mach. Learn. Res.*, **11** (2010), 171–234. https://www.jmlr.org/papers/volume11/aliferis10a/aliferis10a.pdf
9. C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, X. D. Koutsoukos, Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: Analysis and extensions, *J. Mach. Learn. Res.*, **11** (2010), 235–284. https://www.jmlr.org/papers/volume11/aliferis10b/aliferis10b.pdf
10. S. R. de Morais, A. Aussem, A novel Markov boundary based feature subset selection algorithm, *Neurocomputing*, **73** (2010), 578–584. https://doi.org/10.1016/j.neucom.2009.05.018
11. S. Fu, M. C. Desmarais, Markov blanket based feature selection: A review of past decade, *In: Proceedings of the World Congress on Engineering*, 2010, 321–328.

12. F. Schlüter, A survey on independence-based Markov networks learning, *Artif. Intell. Rev.*, **42** (2014), 1069–1093. https://doi.org/10.1007/s10462-012-9346-y
13. J. P. Pellet, A. Elisseeff, Using Markov blankets for causal structure learning, *J. Mach. Learn. Res.*, **9** (2008), 1295–1342. https://www.jmlr.org/papers/volume9/pellet08a/pellet08a.pdf
14. A. R. Masegosa, S. Moral, A Bayesian stochastic search method for discovering markov boundaries, *Knowl.-Based Syst.*, **35** (2012), 211–223. https://doi.org/10.1016/j.knosys.2012.04.028
15. I. Tsamardinos, C. F. Aliferis, Towards principled feature selection: Relevancy, filters and wrappers, *In: International Workshop on Artificial Intelligence and Statistics*, 2003, 300–307.
16. A. Statnikov, N. I. Lytkin, J. Lemeire, C. F. Aliferis, Algorithms for discovery of multiple Markov boundaries, *J. Mach. Learn. Res.*, **14** (2013), 499–566. https://www.jmlr.org/papers/volume14/statnikov13a/statnikov13a.pdf
17. X. Q. Liu, X. S. Liu, Swamping and masking in Markov boundary discovery, *Mach. Learn.*, **104** (2016), 25–54. https://doi.org/10.1007/s10994-016-5545-0
18. X. Q. Liu, X. S. Liu, Markov blanket and markov boundary of multiple variables, *J. Mach. Learn. Res.*, **19** (2018), 1–50. https://www.jmlr.org/papers/volume19/14-033/14-033.pdf
19. N. K. Kitson, A. C. Constantinou, Z. G. Guo, Y. Liu, K. Chobtham, A survey of Bayesian network structure learning, *Artif. Intell. Rev.*, **56** (2023), 8721–8814. https://doi.org/10.1007/s10462-022-10351-w
20. J. Lemeire, *Learning causal models of multivariate systems and the value of it for the performance modeling of computer programs*, ASP/VUBPRESS/UPA, 2007.
21. J. Lemeire, S. Meganck, F. Cartella, T. T. Liu, Conservative independence-based causal structure learning in absence of adjacency faithfulness, *Int. J. Approx. Reason.*, **53** (2012), 1305–1325. https://doi.org/10.1016/j.ijar.2012.06.004
22. F. Bromberg, D. Margaritis, Improving the reliability of causal discovery from small datasets using argumentation, *J. Mach. Learn. Res.*, **10** (2009), 301–340. https://www.jmlr.org/papers/volume10/bromberg09a/bromberg09a.pdf
23. J. M. Peña, R. Nilsson, J. Björkegren, J. Tegnér, Towards scalable and data efficient learning of Markov boundaries, *Int. J. Approx. Reason.*, **45** (2007), 211–232. https://doi.org/10.1016/j.ijar.2006.06.008
24. J. Cheng, R. Greiner, J. Kelly, D. Bell, W. R. Liu, Learning Bayesian networks from data: An information-theory based approach, *Artif. Intell.*, **137** (2002), 43–90. https://doi.org/10.1016/S0004-3702(02)00191-1
25. H. Cramér, *Mathematical methods of statistics*, New Jersey: Princeton University Press, 1999.
26. S. Kullback, *Information theory and statistics*, New York: Dover Publications, 1997.
27. L. M. de Campos, A scoring function for learning Bayesian networks based on mutual information and conditional independence tests, *J. Mach. Learn. Res.*, **7** (2006), 2149–2187. https://www.jmlr.org/papers/volume7/decampos06a/decampos06a.pdf
28. W. G. Cochran, Some methods for strengthening the common χ^2 tests, *Biometrics*, **10** (1954), 417–451. https://doi.org/10.2307/3001616

29. D. N. Lawley, A general method for approximating to the distribution of likelihood ratio criteria, *Biometrika*, **43** (1956), 295–303. https://doi.org/10.2307/2332908
30. B. S. Hosmane, Improved likelihood ratio tests and pearson chi-square tests for independence in two dimensional contingency tables, *Commun. Stat.-Theor. M.*, **15** (1986), 1875–1888. https://doi.org/10.1080/03610928608829224
31. B. S. Hosmane, Improved likelihood ratio test for multinomial goodness of fit, *Commun. Stat.-Theor. M.*, **16** (1987), 3185–3198. https://doi.org/10.1080/03610928708829566
32. B. S. Hosmane, Smoothing of likelihood ratio statistic for equiprobable multinomial goodness-of-fit, *Ann. Inst. Stat. Math.*, **42** (1990), 133–147. https://doi.org/10.1007/BF00050784
33. S. Brin, R. Motwani, C. Silverstein, Beyond market baskets: Generalizing association rules to correlations, *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, **26** (1997), 265–276. https://doi.org/10.1145/253260.253327
34. C. Silverstein, S. Brin, R. Motwani, Beyond market baskets: Generalizing association rules to dependence rules, *Data Min. Knowl. Disc.*, **2** (1998), 39–68. https://doi.org/10.1023/A:1009713703947
35. S. Yaramakala, *Fast Markov blanket discovery*, Iowa State University, 2004.
36. P. Spirtes, C. Glymour, R. Scheines, *Causation, prediction, and search*, Cambridge: MIT Press, 2001.
37. S. K. Fu, M. Desmarais, Local learning algorithm for Markov blanket discovery, *Advances in Artificial Intelligence*, 2007, 68–79.
38. W. Khan, L. F. Kong, S. M. Noman, B. Brekhna, A novel feature selection method via mining Markov blanket, *Appl. Intell.*, **53** (2023), 8232–8255. https://doi.org/10.1007/s10489-022-03863-z
39. D. Koller, M. Sahami, Toward optimal feature selection, *In: Thirteen International Conference in Machine Learning*, Stanford InfoLab, 1996, 284–292.
40. D. Margaritis, S. Thrun, *Bayesian network induction via local neighborhoods*, Carnegie Mellon University, 1999.
41. D. Margaritis, S. Thrun, Bayesian network induction via local neighborhoods, *In: Advances in Neural Information Processing Systems*, Morgan Kaufmann, 1999, 505–511.
42. I. Tsamardinos, C. F. Aliferis, A. Statnikov, Algorithms for large scale Markov blanket discovery, *In: Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 2003, 376–381.
43. X. L. Yang, Y. J. Wang, Y. Ou, Y. H. Tong, Three-fast-inter incremental association Markov blanket learning algorithm, *Pattern Recogn. Lett.*, **122** (2019), 73–78. https://doi.org/10.1016/j.patrec.2019.02.002
44. H. R. Liu, Q. R. Shi, Y. B. Cai, N. T. Wang, L.Y. Zhang, D. Y. Liu, Fast shrinking parents-children learning for markov blanket-based feature selection, *Int. J. Mach. Learn. Cyber.*, **15** (2024), 3553–3566. https://doi.org/10.1007/s13042-024-02108-4
45. K. P. Murphy, *Bayes Net Toolbox for matlab*, Version: FullBNT-1.0.7, 2007. Available from: `https://github.com/bayesnet/bnt`

46. T. Gao, Q. Ji, Efficient score-based Markov blanket discovery, *Int. J. Approx. Reason.*, **80** (2017), 277–293. https://doi.org/10.1016/j.ijar.2016.09.009

47. T. Niinimäki, P. Parviainen, Local structure disocvery in Bayesian network, *arXiv Preprint*, 2012.

48. T. Silander, P. Myllymäki, A simple approach for finding the globally optimal bayesian network structure, *arXiv Preprint*, 2012.

49. J. Cussens, M. Bartlett, E. M. Jones, N. A. Sheehan, Maximum likelihood pedigree reconstruction using integer linear programming, *Genet. Epidemiol.*, **37** (2013), 69–83. https://doi.org/10.1002/gepi.21686

50. G. Brown, A. Pocock, M. J. Zhao, M. Luján, Conditional likelihood maximisation: A unifying framework for information theoretic feature selection, *J. Mach. Learn. Res.*, **13** (2012), 27–66. https://www.jmlr.org/papers/volume13/brown12a/brown12a.pdf

51. K. T. Fang, J. L. Xu, *Statistical distributions*, Beijing: Science Press, 1987.

52. N. L. Johnson, S. Kotz, *Distributions in statistics: Continuous univariate distributions-2*, Boston: John Wiley & Sons, 1970.

53. G. Schwarz, Estimating the dimension of a model, *Ann. Stat.*, **6** (1978), 461–464. https://www.jstor.org/stable/2958889

AIMS Press