Journal of Machine Learning Research 19 (2018) 1-50 Submitted 1/14; Revised 7/18; Published 9/18

Markov Blanket and Markov Boundary of Multiple Variables

Xu-Qing Liu LIUXUQING688@163.COM
State Key Laboratory of Mechanics and Control of Mechanical Structures
Institute of Nano Science and Department of Mathematics
Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
Faculty of Mathematics and Physics
Huaiyin Institute of Technology, Huai'an 223003, China

Xin-Sheng Liu* XSLIU@NUAA.EDU.CN
State Key Laboratory of Mechanics and Control of Mechanical Structures
Institute of Nano Science and Department of Mathematics
Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Editors: Marina Meila; Kevin Murphy; Joris Mooij

Abstract

Markov blanket (Mb) and Markov boundary (MB) are two key concepts in Bayesian networks (BNs). In this paper, we study the problem of Mb and MB for multiple variables. First, we show that Mb possesses the additivity property under the local intersection assumption, that is, an Mb of multiple targets can be constructed by simply taking the union of Mbs of the individual targets and removing the targets themselves. MB is also proven to have additivity under the local intersection assumption. Second, we analyze the cases of violating additivity of Mb and MB and then put forward the notions of Markov blanket supplementary (MbS) and Markov boundary supplementary (MBS). The properties of MbS and MBS are studied in detail. Third, we build two MB discovery algorithms and prove their correctness under the local composition assumption. We also discuss the ways of practically doing conditional independence tests and analyze the complexities of the algorithms. Finally, we make a benchmarking study based on six synthetic BNs and then apply MB discovery to multi-class prediction based on a real data set. The experimental results reveal our algorithms have higher accuracies and lower complexities than existing algorithms.

Keywords: Markov blanket, Markov boundary, Markov blanket supplementary, Markov boundary supplementary, Bayesian network

1. Introduction

Bayesian networks (BNs) are graphical structures used to represent the probabilistic relations among a large number of variables and to make the associated probabilistic inferences (Neapolitan, 2004; Pearl, 1988). In recent years, BNs have become one of the most powerful tools in encoding uncertain expert knowledge in expert systems (Daly et al., 2011; Parviainen and Koivisto, 2013) and also deeply influenced on many other actual domains such as medical diagnosis, financial analysis, bioinformatics, and industrial applications (Zhang and Guo, 2006).

*. Corresponding Author.

As two important concepts in BNs, Markov blanket (Mb) and Markov boundary (MB) play a key role in feature selection (FS; Fu and Desmarais, 2010; Pellet and Elisseeff, 2008; Aliferis et al., 2010a,b). Mathematically, Pearl (1988, pp. 218–221) showed the conditional probability for the target given other variables can be replaced by the MB as the conditional set. Pellet and Elisseeff (2008, pp. 1299, 1302) proved that an MB is the theoretically optimal set of features. Further, under certain assumptions about the learner and the loss function, MB is the solution to the variable selection problem (Tsamardinos and Aliferis, 2003; Statnikov et al., 2013).

So far most authors have focused on the problem of Mb or MB for a single variable. In this paper, we consider the problem of Mb and MB for multiple variables. This occurs if, for example, one wants to compute the joint probability of two or more variables conditioned on all other variables. The basic question for Mb of multiple variables is whether the additivity property holds, that is, can an Mb of multiple variables be constructed by simply taking the union of the Mbs of the individual variables and removing the target variables themselves? The same question is for MB. Further, if the additivity property is violated in some situation, how can we do it?

In the literature, there have been lots of MB discovery algorithms, such as the Koller-Sahami (`KS`) algorithm (Koller and Sahami, 1996), the grow-shrink (`GS`) algorithm (Margaritis and Thrun, 1999, 2000), the incremental association Markov boundary (`IAMB`) algorithm (Tsamardinos et al., 2003) and its several variants, the `HITON` algorithm (Aliferis et al., 2003), the max-min Markov boundary (`MMMB`) algorithm (Tsamardinos et al., 2006), the parents and children based Markov boundary (`PCMB`) algorithm and `KIAMB` algorithms (Peña et al., 2007), the `BFMB` algorithm (Fu and Desmarais, 2007), the algorithmic framework called generalized local learning (`GLL`, Aliferis et al., 2010a), and some others (Fu and Desmarais, 2010; Schlüter, 2014). For a single target variable, most of these algorithms are efficient to seek an approximate MB; for multiple target variables, if simply regarding them as a multivariate variable, these algorithms seem to be feasible. However, this will lead to low accuracies and high computational complexities. Hence, it is necessary to design more efficient MB discovery algorithms for multiple variables.

The remainder of this paper is organized as follows. Section 2 presents necessary preliminaries and the motivations of this paper. Subsection 3.1 shows additivity of Mb and MB under the local intersection assumption. In Subsection 3.2, we first analyze when additivity is violated and then put forward the notions of Markov blanket supplementary (MbS) and Markov boundary supplementary (MBS). The properties of MbS and MBS are studied detailedly. In Section 4, we design two MB discovery algorithms for multiple variables, and prove their correctness under the local composition assumption. In addition, we discuss the ways of practically doing conditional independence (CI) tests and analyze the complexities of the algorithms. Section 5 makes a benchmarking study based on six synthetic BNs, and Section 6 considers a practical application. The experimental results show the superiority of our algorithms with higher accuracies and lower complexities than existing algorithms. Section 7 concludes this paper and presents three remarks.

2. Preliminaries and Motivations

In the paper, we denote a variable and its value by upper-case and lower-case letters in italics (e.g., X, x), a set of variables and its value by upper-case and lower-case bold letters in italics (e.g., $\boldsymbol{X}$, $\boldsymbol{x}$). The difference between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is denoted by $\boldsymbol{X} \setminus \boldsymbol{Y}$. For brevity, we write $(\boldsymbol{X} \setminus \boldsymbol{Y}) \setminus \boldsymbol{Z}$ as $\boldsymbol{X} \setminus \boldsymbol{Y} \setminus \boldsymbol{Z}$. In addition, we use $|\boldsymbol{X}|$ to denote the number of variables involved in $\boldsymbol{X}$.

2.1 Preliminaries

Suppose we have a joint probability distribution $\mathbb{P}$ over $\boldsymbol{V} \triangleq \{X_1, \cdots, X_p\}$ and a directed acyclic graph (DAG) $\mathbb{G}$ with the variables in $\boldsymbol{V}$ as its nodes. We say $(\mathbb{G}, \mathbb{P})$ satisfies the Markov condition if every $X \in \boldsymbol{V}$ is conditionally independent of its nondescendants given its parents; Further, $(\mathbb{G}, \mathbb{P})$ is called a Bayesian network (BN) if it satisfies the Markov condition; Furthermore, $(\mathbb{G}, \mathbb{P})$ satisfies the faithfulness condition if, based on the Markov condition, $\mathbb{G}$ entails all and only conditional independences (CIs) in $\mathbb{P}$ (Pearl, 1988; Neapolitan, 2004).

We write $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$ ($\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$), if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent (dependent) given $\boldsymbol{Z}$ with respect to $\mathbb{P}$. The following properties describe the relations among CI statements (Pearl, 1988; Peña et al., 2007; Statnikov et al., 2013). For any $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W} \subseteq \boldsymbol{V}$, we have (i) *symmetry*: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$ is equivalent to $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{Z}$; (ii) *decomposition*: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \cup \boldsymbol{W} \,|\, \boldsymbol{Z}$ implies $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$ and $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{W} \,|\, \boldsymbol{Z}$; (iii) *weak union*: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \cup \boldsymbol{W} \,|\, \boldsymbol{Z}$ implies $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z} \cup \boldsymbol{W}$; (iv) *contraction*: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z} \cup \boldsymbol{W}$ and $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{W} \,|\, \boldsymbol{Z}$ imply $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \cup \boldsymbol{W} \,|\, \boldsymbol{Z}$; (v) *self-conditioning*: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Y} \cup \boldsymbol{Z}$. Further, if $\mathbb{P}$ is strictly positive, then besides (i)∼(v) we also have (vi) *intersection*: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z} \cup \boldsymbol{W}$ and $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{W} \,|\, \boldsymbol{Z} \cup \boldsymbol{Y}$ imply $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \cup \boldsymbol{W} \,|\, \boldsymbol{Z}$. Furthermore, if $\mathbb{P}$ is faithful to a DAG $\mathbb{G}$, then besides (i)∼(vi) we also have (vii) *composition*: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$ and $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{W} \,|\, \boldsymbol{Z}$ imply $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \cup \boldsymbol{W} \,|\, \boldsymbol{Z}$.

Among these properties, intersection and composition are two global ones. Statnikov et al. (2013, p. 504) provided a relaxed version for composition called *local composition*: one says $\boldsymbol{T} \subseteq \boldsymbol{V}$ satisfies the local composition property, if $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{Z}$ and $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$ imply $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{X} \cup \boldsymbol{Y} \,|\, \boldsymbol{Z}$ for any $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \boldsymbol{T}$. We will provide a relaxed version for the intersection property.

Conditional mutual information (CMI) is one of the basic tools for testing CIs. Denote the CMI between $\boldsymbol{X}$ and $\boldsymbol{Y}$ conditioned on $\boldsymbol{Z}$ by $\mathbb{I}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$. Then $\mathbb{I}(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) \geqslant 0$, with equality holding if and only if $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$ (Zhang and Guo, 2006). For a practical problem, we cannot access to the true CMI; instead, we use its empirical estimate, denoted by $\mathbb{I}_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$, based on the data $\boldsymbol{D}$ (Cheng et al., 2002). Note that $\mathbb{I}_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z}) \geqslant 0$ also holds for any $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{V}$.

The *chain rule* for CMI (Cover and Thomas, 2006) is useful to prove the main results of this paper: $\mathbb{I}(\boldsymbol{X}; \boldsymbol{Y}_1 \cup \boldsymbol{Y}_2 \,|\, \boldsymbol{Z}) = \mathbb{I}(\boldsymbol{X}; \boldsymbol{Y}_1 \,|\, \boldsymbol{Z}) + \mathbb{I}(\boldsymbol{X}; \boldsymbol{Y}_2 \,|\, \boldsymbol{Z} \cup \boldsymbol{Y}_1)$ holds for any four sets of variables $\boldsymbol{X}$, $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$, and $\boldsymbol{Z}$ from $\boldsymbol{V}$.

Another notion closely related to CI is d-separation (Pearl, 1988, p. 117). 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 $\boldsymbol{X} \perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$. Here, $\boldsymbol{Z}$ blocking a path $\mathbb{c}$ means that $\mathbb{c}$ has a head-to-tail node or a tail-to-tail node belonging to $\boldsymbol{Z}$, or that $\mathbb{c}$ has a head-to-head node C such that C and its all descendants are not in $\boldsymbol{Z}$. As well known, $\boldsymbol{X} \perp \boldsymbol{Y} \,|\, \boldsymbol{Z} \Rightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$, if $(\mathbb{G}, \mathbb{P})$ is a BN (Neapolitan, 2004, p. 74). This implication provides a convenient way of identifying CIs.

For example, consider a BN with the graph presented in Figure 1 as its DAG. It follows that: X_2 and X_8 are d-separated by $\{X_4, X_5\}$, meaning $X_2 \perp X_8 \,|\, \{X_4, X_5\}$ and thus $X_2 \perp\!\!\!\perp X_8 \,|\, \{X_4, X_5\}$; X_3 and X_4 are d-separated by $\emptyset$, meaning $X_3 \perp X_4$, so $X_3 \perp\!\!\!\perp X_4$. Note that these two probabilistic CIs can not be directly derived from the Markov condition.

In what follows, the concepts of Mb and MB are presented. They are a direct extension of Mb and MB for a single target variable (Pearl, 1988, p. 97; Neapolitan, 2004, pp. 108–109): an Mb of $\boldsymbol{T}$ is a set of variables shielding $\boldsymbol{T}$ from all other variables, so it carries all information of $\boldsymbol{T}$ that cannot be obtained from other variables, while an MB is a minimal Mb.

Definition 1 *Let $\boldsymbol{T} \subseteq \boldsymbol{V}$ and $\boldsymbol{M} \subseteq \boldsymbol{V} \setminus \boldsymbol{T}$. We call $\boldsymbol{M}$ a Markov blanket (Mb) of $\boldsymbol{T}$ if $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M} \setminus \boldsymbol{T} \,|\, \boldsymbol{M}$. Further, a Markov boundary (MB) of $\boldsymbol{T}$ is any Mb such that none of its proper subsets is an Mb.* ■

When $|\boldsymbol{T}| = 1$, the following results are well known in the literature (Pearl, 1988; Neapolitan, 2004; Statnikov et al., 2013): (a) if $(\mathbb{G}, \mathbb{P})$ is a BN, then for $T \in \boldsymbol{V}$ the set of its all parents, children, and spouses is an Mb of T (denoted by $\boldsymbol{M}_T$); (b) if $\mathbb{P}$ satisfies the intersection property, then T has a unique MB; (c) if $(\mathbb{G}, \mathbb{P})$ satisfies the faithfulness condition, then $\boldsymbol{M}_T$ is the unique MB of T.

Consider again the BN with the graph presented in Figure 1 as its DAG. In this BN, it is seen that $\boldsymbol{M}_{X_4} \triangleq \{X_2, X_6, X_3\}$ is an Mb of X_4; further, $\boldsymbol{M}_{X_4}$ is the unique MB of X_4 if the faithfulness condition is satisfied. Similarly, $\boldsymbol{M}_{X_2} \triangleq \{X_4, X_5\}$ is the unique MB of X_2 under the faithfulness condition.

The above result (b) points out that if the uniqueness of MB is violated, then the intersection property must be violated. Lemeire (2007) provided a case of violating intersection called *information equivalence*: $\boldsymbol{X}$ and $\boldsymbol{Y}$ are called information equivalent with respect to $\boldsymbol{T}$ if $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{X}$, $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{Y}$, $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{Y}$, and $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{X}$. A related notion is *conditional information equivalence* (Lemeire et al., 2012; Statnikov et al., 2013): $\boldsymbol{X}$ and $\boldsymbol{Y}$ are called to be conditionally information equivalent with respect to $\boldsymbol{T}$ given $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \boldsymbol{X} \setminus \boldsymbol{Y} \setminus \boldsymbol{T}$, if $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{Z}$, $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$, $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{Y} \cup \boldsymbol{Z}$, and $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{X} \cup \boldsymbol{Z}$. Lemeire et al. (2012, pp. 1309–1311) showed that (conditional) information equivalence is one of the two major cases in which *adjacency faithfulness* is violated. Here, the adjacency faithfulness condition (Ramsey et al., 2006; Lemeire et al., 2012) is defined as: if X and Y are adjacent, then $X \not\perp\!\!\!\perp Y \,|\, \boldsymbol{Z}$ for any $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{X, Y\}$. Statnikov et al. (2013, p. 503) provided a local version for adjacency faithfulness by focusing on a specific variable.

Here, we employ the information flow metaphor (Cheng et al., 2002) to intuitively explain information equivalence: we can view a BN as a network of information channels, where each node is a *valve* that is either active or inactive; the valves are connected by *information channels*; information can flow through an active valve but not an inactive one; instantiating a node means this valve becomes inactive. We extend this metaphor by viewing a clique of one or more nodes as a valve. In this sense, a CI relation $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$ means all the channels between $\boldsymbol{X}$ and $\boldsymbol{Y}$ are cut off by $\boldsymbol{Z}$ and thus the information between $\boldsymbol{X}$ and $\boldsymbol{Y}$ can not flow once $\boldsymbol{Z}$ becomes inactive. When information equivalence occurs, we further extend this information flow metaphor as follows: if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are information equivalent with respect to $\boldsymbol{T}$ given $\boldsymbol{Z}$, then there exists an *information equivalent valve*, denoted by $\delta_{\boldsymbol{X},\boldsymbol{Y};\boldsymbol{T}|\boldsymbol{Z}}$, which connects $\boldsymbol{T}$ and $\boldsymbol{X}$ and connects $\boldsymbol{T}$ and $\boldsymbol{Y}$; $\delta_{\boldsymbol{X},\boldsymbol{Y};\boldsymbol{T}|\boldsymbol{Z}}$ *is active when and only when both* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are active*. Then, the relation "$\boldsymbol{X}$ and $\boldsymbol{Y}$ are information equivalent with respect to $\boldsymbol{T}$ given $\boldsymbol{Z}$" can be presented in Figure 2.

Information equivalence represents all the possible situations leading to the nonuniqueness of MB. In fact, we can show the following result, which indicates that violating the uniqueness of MB implies the presence of information equivalence. The proof is presented in Section B.

Lemma 1 *The intersection property holds if and only if no information equivalence occurs.* ■

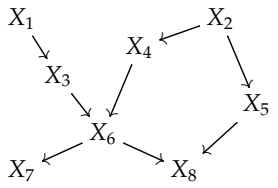

Figure 1: A simple DAG used to illustrate d-separation.

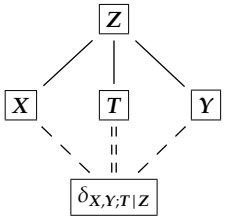

Figure 2: An intuitive illustration for information equivalence.

By analyzing the proof of Lemma 1, we present a relaxed version for the intersection property called *local intersection* as follows.

Definition 2 (Local Intersection) *Letting $\boldsymbol{T} \subseteq \boldsymbol{V}$, we say $\boldsymbol{T}$ satisfies the local intersection property, if the following two types of local conditions hold simultaneously:* (i) *type-I local condition: in the case of $|\boldsymbol{T}| \geqslant 2$, for any disjoint $\boldsymbol{T}_1, \boldsymbol{T}_2 \subseteq \boldsymbol{T}$, there are no disjoint $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V} \setminus \boldsymbol{T}$ such that $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are information equivalent with respect to $\boldsymbol{X}$ conditioned on $\boldsymbol{Y}$; and* (ii) *type-II local condition: there are no disjoint $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \boldsymbol{T}$ such that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are information equivalent with respect to $\boldsymbol{T}$ conditioned on $\boldsymbol{Z}$.* ■

Clearly, *intersection* implies *local intersection* but not vice versa, because the former requires no any information equivalence while the latter only requires no information equivalence between the targets and the remaining variables. Here, we give a lemma concerning the uniqueness of MB under the local intersection assumption. The proof is presented in Appendix B.

Lemma 2 *For $\boldsymbol{T} \subseteq \boldsymbol{V}$, assume the type-II local condition defined in Definition 2 holds. Then $\boldsymbol{T}$ has a unique MB.* ■

To facilitate the identification of information equivalence, Lemeire (2007) ever introduced the notions of target partition (T-partition) and equivalent partition (E-partition), and then provided a relation among information equivalence, T-partition, and E-partition.

- T-partition: The domain, $\boldsymbol{X}_{\text{dom}}$, of $\boldsymbol{X}$ can be partitioned into disjoint subsets $\boldsymbol{X}_{\text{dom}}^{(k)}$ for which $\mathbb{P}(\boldsymbol{T} \,|\, \boldsymbol{x})$ is the same for all $\boldsymbol{x} \in \boldsymbol{X}_{\text{dom}}^{(k)}$. This is called the T-partition of $\boldsymbol{X}_{\text{dom}}$ with respect to $\boldsymbol{T}$.
- E-partition: A relation $\mathfrak{R} \subset \boldsymbol{X} \otimes \boldsymbol{Y}$ defines an E-partition in $\boldsymbol{Y}_{\text{dom}}$ to a partition of $\boldsymbol{X}_{\text{dom}}$, if: (i) $\neg(\boldsymbol{x}_2 \mathfrak{R} \boldsymbol{y}_1)$ holds for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \boldsymbol{X}_{\text{dom}}$ belonging to different partitions and for any $\boldsymbol{y}_1 \in \boldsymbol{Y}_{\text{dom}}$ with $\boldsymbol{x}_1 \mathfrak{R} \boldsymbol{y}_1$; and (ii) for every $\boldsymbol{X}_{\text{dom}}^{(k)}$, there exist $\boldsymbol{x}_1 \in \boldsymbol{X}_{\text{dom}}^{(k)}$ and $y_1 \in \boldsymbol{Y}_{\text{dom}}$ such that $\boldsymbol{x}_1 \mathfrak{R} \boldsymbol{y}_1$.
- Relation among *information equivalence*, *T-partition*, and *E-partition*: If $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{X}$ and $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{X}$, then $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{Y}$ (meaning $\boldsymbol{X}$ and $\boldsymbol{Y}$ are information equivalent with respect to $\boldsymbol{T}$) if and only if the relation $\boldsymbol{x} \mathfrak{R} \boldsymbol{y}$ defined by $\mathbb{P}(\boldsymbol{x}, \boldsymbol{y}) > 0$ with $\boldsymbol{x} \in \boldsymbol{X}_{\text{dom}}$ and $y \in \boldsymbol{Y}_{\text{dom}}$ defines an E-partition in $\boldsymbol{Y}_{\text{dom}}$ to the T-partition of $\boldsymbol{X}_{\text{dom}}$ with respect to $\boldsymbol{T}$.

The graph shown in Figure 3, originally presented by Statnikov and Aliferis (2010), makes an intuitive illustration on T-partition and E-partition. As seen, $\{1, 2\}$ and $\{3\}$ constitute the T-partition of $A_{\text{dom}} \triangleq \{1, 2, 3\}$ with respect to C; $\{1, 2\}$ and $\{3\}$ are the E-partition of $B_{\text{dom}} \triangleq \{1, 2, 3\}$ to the T-partition of A_{dom}, Therefore, A and B are information equivalent with respect to C if $C \not\perp\!\!\!\perp A$, since $C \perp\!\!\!\perp B \,|\, A$ holds inherently because of the Markov condition.

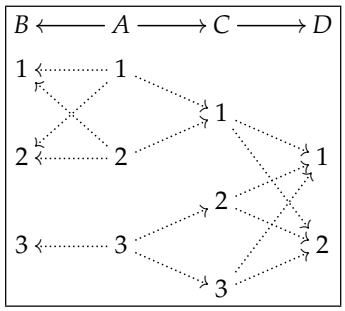

Figure 3: An illustration on T-partition and E-partition: in the DAG "$B \leftarrow A \rightarrow C \rightarrow D$", all variables take $\{1, 2, 3\}$ except for D taking $\{1, 2\}$, and dotted arrows denote all non-zero conditional probabilities of each variable given its parents.

Finally, the notion of *context-independent information equivalence* given by Statnikov et al. (2013) will be used in Example 2. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are called context-independent information equivalent with respect to $\boldsymbol{T}$, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are information equivalent with respect to $\boldsymbol{T}$ given any $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \boldsymbol{X} \setminus \boldsymbol{Y} \setminus \boldsymbol{T}$. For this notion, Statnikov et al. (2013) proved the following conclusion: if $\boldsymbol{M}$ is an Mb of $\boldsymbol{T}$ with $\boldsymbol{X} \subseteq \boldsymbol{M}$, and there is some $\boldsymbol{Y} \subseteq \boldsymbol{V} \setminus \boldsymbol{M} \setminus \boldsymbol{T}$ such that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are context-independent information equivalent with respect to $\boldsymbol{T}$, then $(\boldsymbol{M} \setminus \boldsymbol{X}) \cup \boldsymbol{Y}$ is also an Mb of $\boldsymbol{T}$.

2.2 Two Typical Algorithms: `IAMB` and `KIAMB`

This subsection concisely presents two typical MB discovery algorithms: `IAMB` (Tsamardinos et al., 2003) and `KIAMB` (Peña et al., 2007). We select them because of their high adaptability and time efficiency: (i) correctness of `IAMB` and `KIAMB` requires only the local composition assumption (Statnikov et al., 2013), while the correctness of the parents and children based algorithms, such as `PCMB` and the algorithms in the `GLL` framework, usually requires the faithfulness condition (Peña et al., 2007, Theorem 6; and Aliferis et al., 2010a, Theorem 1); (ii) `IAMB` and `KIAMB` are time efficient and thus suitable for the problem of MB for multiple variables, while the parents and children based algorithms have exponential complexities (Aliferis et al., 2010a, pp. 199–200), so they are hard to work when too many variables are involved, such as the problem of MB discovery for multiple variables.

`IAMB` is an enhanced variant of `GS`. In 2003, Tsamardinos et al. pointed out that `GS` uses a static and potentially inefficient heuristic in the growing phase, and then proposed `IAMB` by employing a dynamic heuristic. Tsamardinos et al. (2003) showed the correctness of `IAMB` under the faithfulness condition; Peña et al. (2007) relaxed the condition to the composition assumption; Statnikov et al. (2013) further relaxed the condition to the local composition assumption. Algorithm 3 describes the pseudo code for `IAMB`. See Appendix A for details.

In the algorithm, there is a function $f_{\boldsymbol{D}}$ (Line 3 of `IAMB` in Algorithm 3) denoting a heuristic used to measure the association between variables (Tsamardinos et al., 2003; Peña et al., 2007). Two widely used selections for $f_{\boldsymbol{D}}$ are CMI (Cheng et al., 2002; Tsamardinos et al., 2003) and the negative p-value (Tsamardinos et al., 2006; Aliferis et al., 2010a,b; Statnikov et al., 2013). Also, Yaramakala (2004, p. 41) suggested an equivalent version of the negative p-value. Subsection 4.3 will make a discussion about the ways of practically doing CI tests and the selections for $f_{\boldsymbol{D}}$.

KIAMB is a stochastic extension of IAMB. It embeds a randomization parameter $K \in [0, 1]$ which specifies the trade-off between greediness and randomness. If taking $K = 1$, KIAMB reduces to IAMB. Peña et al. (2007) proved the correctness of KIAMB under the composition assumption. By the proof, the local composition assumption is sufficient for KIAMB to be correct. Algorithm 3 describes the pseudo code for KIAMB.

For the case of $|\boldsymbol{T}| \geqslant 2$, IAMB and KIAMB can remain correct if strengthening the precondition. We present the correctness of them as follows, without presenting the proof since it is similar to that of the original IAMB and KIAMB (Tsamardinos et al., 2003; Peña et al., 2007; Statnikov et al., 2013). In what follows, we say a CI test for a hypothesis is correct if the statistical decision is correctly made by using a testing method. Subsection 4.3 gives a further discussion on this issue.

Theorem 1 (Correctness of IAMB and KIAMB) *Assume $\boldsymbol{T}$ satisfies the local composition property, and all CI tests are correct. Then* (i) *IAMB outputs an MB of $\boldsymbol{T}$;* (ii) *KIAMB outputs an MB of $\boldsymbol{T}$ for any $K \in [0, 1]$.* ■

2.3 Motivations

This subsection provides three motivations of this paper.

Let $\boldsymbol{M}$ be an MB of T. Then $\mathbb{P}(T \mid \boldsymbol{V} \setminus \{T\}) = \mathbb{P}(T \mid \boldsymbol{M})$. In other words, all information for predicting T is carried by $\boldsymbol{M}$. Further, $\boldsymbol{M}$ is a solution to the FS problem, if the algorithm that constructs the prediction model can learn any probability distribution, and the performance metric is strictly decreasing with the mean-squared loss with a preference for smaller subsets (Tsamardinos and Aliferis, 2003, Proposition 3). For this reason, MB for a single variable is sufficient.

However, there are the situations where MB for multiple variables is preferred. This occurs if we need the probability distribution of more than one variables given all the others. Let $\boldsymbol{M}_i$ be an MB of T_i for $i = 1, 2$. Denoting $\boldsymbol{T} = \{T_1, T_2\}$, it follows that

$$\mathbb{P}(\boldsymbol{T} \mid \boldsymbol{V} \setminus \boldsymbol{T}) = \begin{cases} \mathbb{P}(T_1 \mid \boldsymbol{M}_1)\mathbb{P}(T_2 \mid \boldsymbol{M}_2) & \text{if } T_1 \notin \boldsymbol{M}_2 \text{ or } T_2 \notin \boldsymbol{M}_1 \\ \mathbb{P}(T_1, T_2, \boldsymbol{V} \setminus \boldsymbol{T}) \,/\, \sum_{t_1, t_2} \mathbb{P}(t_1, t_2, \boldsymbol{V} \setminus \boldsymbol{T}) & \text{if } T_1 \in \boldsymbol{M}_2 \text{ and } T_2 \in \boldsymbol{M}_1 \end{cases}$$

As seen, in the case of $T_1 \in \boldsymbol{M}_2$ and $T_2 \in \boldsymbol{M}_1$, the computation is intractable, especially when the dimension is high. Nevertheless, if we have an MB for $\boldsymbol{T}$, denoted by $\boldsymbol{M}$, then $\mathbb{P}(\boldsymbol{T} \mid \boldsymbol{V} \setminus \boldsymbol{T}) = \mathbb{P}(\boldsymbol{T} \mid \boldsymbol{M})$ follows immediately, so the problem is simplified greatly. In this sense, it is meaningful to consider the problem of MB for multiple variables.

The second motivation is that we want to know whether the prediction for $\boldsymbol{T}$ will be affected if the observed values of some variables outside $\boldsymbol{T}$ (in a new observation) are missing. Denote these missing variables by $\boldsymbol{V}_{\mathrm{m}}$. This problem can be considered as follows: find an approximate MB (denoted by $\boldsymbol{M}_{\mathrm{m}}$) of $\boldsymbol{T}$ in $\boldsymbol{V} \setminus \boldsymbol{V}_{\mathrm{m}}$ by means of some method, then check if $\boldsymbol{M}_{\mathrm{m}}$ is an Mb in $\boldsymbol{V}$ via some criterion (e.g., a criterion based on Lemma 2 given by Statnikov et al., 2013); and finally assert $\boldsymbol{T}$ will not be affected if the above checking result is "yes". In this sense, it is also preferred to consider MB for multiple variables.

Figure 4 represents the DAG for the ALARM network (Beinlich et al., 1989), which is well known in the literature. Take $T_1 \triangleq X_{22}$ and $T_2 \triangleq X_{23}$. Then, $\boldsymbol{M}_{T_i}$ is the unique MB of T_i for $i = 1, 2$ under the faithfulness condition, with

$$\boldsymbol{M}_{T_1} \triangleq \{X_1, X_4, X_{15}, X_{21}, X_{23}, X_{27}, X_{29}\} \text{ and } \boldsymbol{M}_{T_2} \triangleq \{X_2, X_{22}, X_{24}, X_{25}, X_{27}, X_{29}\}, \tag{1}$$

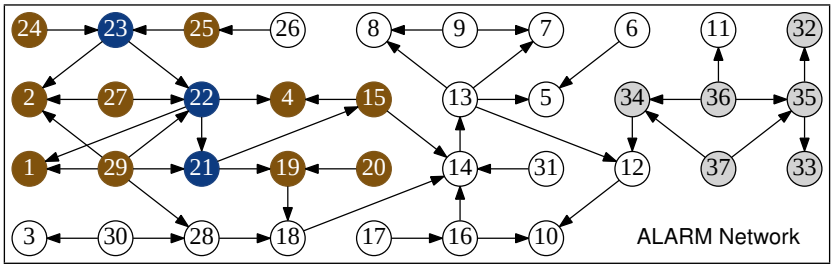

Figure 4: ALARM network (37 nodes and 46 edges): a logical alarm reduction mechanism.

respectively. This leads to intractable computations on the joint probability distribution of T_1 and T_2 given all other variables, consider that $T_1 \in \boldsymbol{M}_{T_2}$ and $T_2 \in \boldsymbol{M}_{T_1}$. Further, if the observed values of some variables (e.g., X_j for $j = 32, 33, \cdots, 37$) in a new observation are missing, can this observation be used any more for predicting T_1 and T_2? Furthermore, we have to face similar problems if three or more target variables are considered.

The third motivation concerns MB discovery algorithms. By Theorem 1, `IAMB` and `KIAMB` can be applied to the problem of MB for multiple variables if simply regarding the targets as a multivariate vector, under the strengthened local composition assumption. However, the assumption of local composition imposed on multiple targets may have more occasions to become invalid than imposed on single targets, due to the synergy effect in the sense that neither X nor Y carries information of T but together they contain some information of T (Rauh et al., 2014).

Here is an illustration: considering the BN with the graph in Figure 5 as its DAG, by direct computations using the FullBNT toolbox (Murphy, 2007), we find that

$$\begin{array}{rrcr} A \perp\!\!\!\perp C, & A \perp\!\!\!\perp D, & \text{and} & A \perp\!\!\!\perp \{C, D\}; \\ B \perp\!\!\!\perp C, & B \perp\!\!\!\perp D, & \text{and} & B \perp\!\!\!\perp \{C, D\}; \\ \{A, B\} \perp\!\!\!\perp C, & \{A, B\} \perp\!\!\!\perp D, & \text{but} & \{A, B\} \not\perp\!\!\!\perp \{C, D\}. \end{array}$$

By this illustration, the idea of applying the existing MB discovery algorithms to multiple targets seems to be practically improper although it is theoretically feasible, because synergy effects may lead to potential inefficiency and even incorrectness. This motivates us to build some algorithms which are resistant to synergy effects and, further, are time efficient.

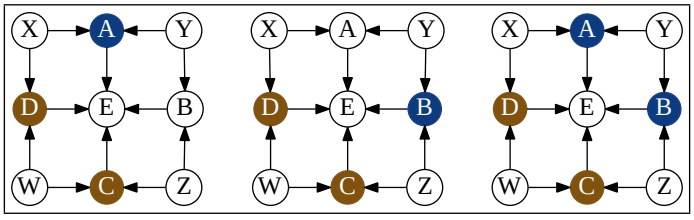

Figure 5: An illustration on synergy effects: each of $\{X, Y, Z, W\}$ takes $\{1, 2\}$ equiprobably; each of $\{A, B, C, D\}$ takes 1 with probabilities p_1, p_2, p_3, p_4 and takes 2 with probabilities $1 - p_1$, $1 - p_2$, $1 - p_3$, $1 - p_4$ given its parents, with $p_4 = p_1 - p_2 + p_3$; E has an arbitrary distribution.

3. Markov Blanket and Markov Boundary for Multiple Variables

This section presents the theoretical results on the problem of Mb and MB for multiple variables when the local intersection property is satisfied and when this property is violated. We study this problem following this way because we are trying to find a suitable approach to transform the problem of Mb and MB from multiple case to single cases, based on which we can build efficient algorithms with high accuracies and low complexities.

3.1 Additivity under Local Intersection

In this subsection, we consider the problem of Mb and MB for multiple variables under the local intersection assumption. We prove Mb and MB possess an ideal property called *additivity*. That is, an Mb of multiple variables can be constructed by simply taking the union of the Mbs of the individual variables and removing the target variables themselves (the same for MB). The results are presented in Theorem 2 and Theorem 3, respectively. Appendix B gives their proofs.

Theorem 2 (Additivity of Mb) *Let* $(\mathbb{G}, \mathbb{P})$ *be a BN over* $\boldsymbol{V}$*. The following two statements hold:*

(i) *Let* $\boldsymbol{M}_i$ *be an Mb of* $\boldsymbol{T}_i \subseteq \boldsymbol{V}$ *for* $i = 1, 2$*, and assume* $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$ *satisfies the local intersection assumption. Then,* $(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus (\boldsymbol{T}_1 \cup \boldsymbol{T}_2)$ *is an Mb of* $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$*.*

(ii) *Let* $\boldsymbol{M}_i$ *be an Mb of* $T_i \in \boldsymbol{V}$ *for* $i = 1, \cdots, k$*, and assume* $\boldsymbol{T} \triangleq \{T_1, \cdots, T_k\}$ *satisfies the local intersection assumption. Then,* $\bigcup_{i=1}^{k} \boldsymbol{M}_i \setminus \boldsymbol{T}$ *is an Mb of* $\boldsymbol{T}$*.* ■

The additivity property of Mb can be intuitively described by the information flow metaphor (Cheng et al., 2002) using Figure 6: $(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus (\boldsymbol{T}_1 \cup \boldsymbol{T}_2)$ is enough to cut off all information channels from $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$ to other valves, when no information equivalence associated with $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$ occurs.

Let $\boldsymbol{T} \subseteq \boldsymbol{V}$ be the set of target variables. As we know, in the case of $|\boldsymbol{T}| = 1$ (denoting $\boldsymbol{T} = \{T\}$), the set $\boldsymbol{M}_T$ composed of the parents, children, and spouses of T is an Mb of it (Pearl, 1988), since $\boldsymbol{M}_T$ d-separates T from all other variables. For the case of $|\boldsymbol{T}| \geqslant 2$ (denoting $\boldsymbol{T} = \{T_1, \cdots, T_k\}$), Theorem 2 indicates that the union of all $\boldsymbol{M}_{T_i}$'s with $T_1, \cdots, T_k$ excluded is an Mb of $\boldsymbol{T}$.

Considering the ALARM network presented in Figure 4, we put $T_1 \triangleq X_{22}$ and $T_2 \triangleq X_{23}$. Then $\boldsymbol{M}_{T_i}$ is an Mb of T_i for $i = 1, 2$, where $\boldsymbol{M}_{T_1}$ and $\boldsymbol{M}_{T_2}$ are defined in (1). Assume $\boldsymbol{T}_{1,2} \triangleq \{T_1, T_2\}$ satisfies the local intersection property. It follows from Theorem 2 that

$$(\boldsymbol{M}_{T_1} \cup \boldsymbol{M}_{T_2}) \setminus \boldsymbol{T}_{1,2} = \{X_1, X_2, X_4, X_{15}, X_{21}, X_{24}, X_{25}, X_{27}, X_{29}\} \triangleq \boldsymbol{M}_{1,2} \tag{2}$$

is an Mb of $\boldsymbol{T}_{1,2}$. Those variables outside $\boldsymbol{M}_{1,2}$ contain no information about $\boldsymbol{T}_{1,2}$ conditioned on $\boldsymbol{M}_{1,2}$ and thereby $\mathbb{P}(\boldsymbol{T}_{1,2} \,|\, \boldsymbol{V} \setminus \boldsymbol{T}_{1,2})$ reduces to $\mathbb{P}(\boldsymbol{T}_{1,2} \,|\, \boldsymbol{M}_{1,2})$. Further, if the observed values of

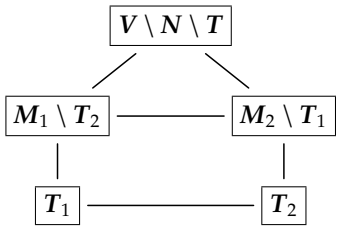

Figure 6: An illustration for additivity of Mb and MB with $\boldsymbol{T} = \boldsymbol{T}_1 \cup \boldsymbol{T}_2$ and $\boldsymbol{N} = (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}$.

X_j for $j = 32, \cdots, 37$ in a new observation are missing, this observation can still be used without affecting the prediction on $\boldsymbol{T}_{1,2}$. Further, $\boldsymbol{M}_{T_3} \triangleq \{X_{15}, X_{19}, X_{20}, X_{22}, X_{29}\}$ is an Mb of $T_3 \triangleq X_{21}$. Assume $\boldsymbol{T}_{1,2,3} \triangleq \{T_1, T_2, T_3\}$ satisfies the local intersection property. Then Theorem 2 shows

$$\boldsymbol{M}_{T_1} \cup \boldsymbol{M}_{T_2} \cup \boldsymbol{M}_{T_3} \setminus \boldsymbol{T}_{1,2,3} = \{X_1, X_2, X_4, X_{15}, X_{19}, X_{20}, X_{24}, X_{25}, X_{27}, X_{29}\} \triangleq \boldsymbol{M}_{1,2,3} \tag{3}$$

is an Mb of $\boldsymbol{T}_{1,2,3}$.

For additivity of Mb shown in (i) of Theorem 2, we have a useful remark (used to simplify our algorithms in Section 4), based on the fact that if $\boldsymbol{M}$ is an Mb of $\boldsymbol{T}$ then $\boldsymbol{M} \cup \boldsymbol{M}_0$ is also an Mb of $\boldsymbol{T}$ for any $\boldsymbol{M}_0 \subseteq \boldsymbol{V} \setminus \boldsymbol{M} \setminus \boldsymbol{T}$. By the remark, the local intersection assumption for additivity of Mb is not required in some special cases. The proof of this remark is given in Appendix B.

Remark 1 *In the case of either $\boldsymbol{T}_1 \subseteq \boldsymbol{V} \setminus \boldsymbol{M}_2$ or $\boldsymbol{T}_2 \subseteq \boldsymbol{V} \setminus \boldsymbol{M}_1$, the conclusion of* (i) *in Theorem 2 holds without requiring the local intersection assumption.* ■

Theorem 2 shows the additivity of Mb. A natural idea is to wonder if additivity is possessed by MB. Theorem 3 affirms this. Appendix B provides the proof. Note that the statements about the uniqueness of MB in this theorem follow from Lemma 2.

Theorem 3 (Additivity of MB) *Let $(\mathbb{G}, \mathbb{P})$ be a BN over $\boldsymbol{V}$. The following two statements hold:*

(i) *Assume $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$ satisfies the local intersection assumption. Let $\boldsymbol{M}_i$ be the unique MB of $\boldsymbol{T}_i$ for $i = 1, 2$. Then, $(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus (\boldsymbol{T}_1 \cup \boldsymbol{T}_2)$ is the unique MB of $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$.*

(ii) *Assume $\boldsymbol{T} \triangleq \{T_1, \cdots, T_k\}$ satisfies the local intersection assumption. Let $\boldsymbol{M}_i$ be the unique MB of T_i for $i = 1, \cdots, k$. Then, $\bigcup_{i=1}^{k} \boldsymbol{M}_i \setminus \boldsymbol{T}$ is the uniqe MB of $\boldsymbol{T}$.* ■

According to Theorem 3, $\boldsymbol{M}_{1,2}$ defined in (2) is not only an Mb but also the unique MB of $\boldsymbol{T}_{1,2}$ in the ALARM network if the faithfulness condition is satisfied. Further, $\boldsymbol{M}_{1,2,3}$ defined in (3) is the unique MB of $\boldsymbol{T}_{1,2,3}$.

3.2 Theoretical Results in the General Case

Let $(\mathbb{G}, \mathbb{P})$ be a BN over $\boldsymbol{V}$, and assume $\boldsymbol{T}_i \subseteq \boldsymbol{V}$ with $|\boldsymbol{T}_i| \geqslant 1$ has an Mb or MB, $\boldsymbol{M}_i$, for $i = 1, 2$. Denote $\boldsymbol{T} = \boldsymbol{T}_1 \cup \boldsymbol{T}_2$ and $\boldsymbol{N} = (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}$. In the case that $\boldsymbol{M}_i$ is an MB of $\boldsymbol{T}_i$, Theorem 3 reveals that $\boldsymbol{N}$ is an MB of $\boldsymbol{T}$ if $\boldsymbol{T}$ satisfies the local intersection assumption. However, when the local intersection assumption does not hold (meaning information equivalence occurs, as Lemma 1 shows), $\boldsymbol{N}$ may be no longer an MB of $\boldsymbol{T}$, due to one of the following reasons: (i) $\boldsymbol{N}$ may be an Mb but it may not possess minimality, as shown by Example 2; (ii) $\boldsymbol{N}$ may be insufficient to shield $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ from all other variables, so it is no longer an Mb in this case, and some extra variables are required to enter into $\boldsymbol{N}$. Example 1 provides an illustration.

For the first case, we need only to optimize $\boldsymbol{N}$ by simply removing redundant variables from $\boldsymbol{N}$; however, for the second case, the additivity property of MB is thoroughly broken, and the problem of constructing an MB for $\boldsymbol{T}$ based on $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ becomes complex. On the one hand, there are some variables in $\boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$ needing to enter into $\boldsymbol{N}$; on the other hand, there may be some variables in $\boldsymbol{N}$ becoming redundant once some new members supplement $\boldsymbol{N}$. What we concern are which variables should enter into $\boldsymbol{N}$ and how we find them.

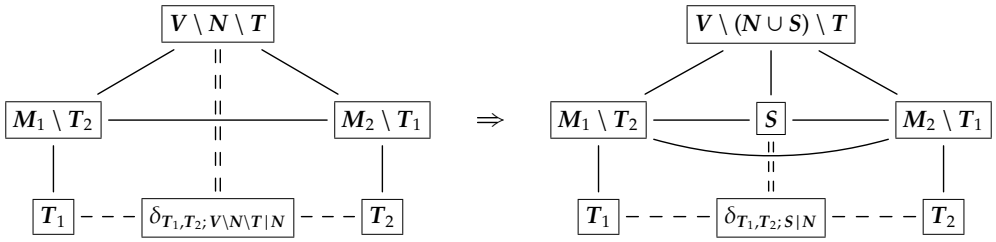

Figure 7: An illustration for the case of violating additivity of Mb and MB, caused by information equivalence.

Assume $\boldsymbol{N}$ is no longer an Mb of $\boldsymbol{T}$. Then, it is easily shown that

$$\boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{T}_1 \,|\, \boldsymbol{N}, \quad \text{but} \quad \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T} \perp\!\!\!\perp \boldsymbol{T}_1 \,|\, \boldsymbol{N} \cup \boldsymbol{T}_2,$$
$$\boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{T}_2 \,|\, \boldsymbol{N}, \quad \text{but} \quad \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T} \perp\!\!\!\perp \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{T}_1.$$

That is, $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ contain equivalent information about $\boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$ given $\boldsymbol{N}$. See Figure 7 for an illustration: the valves $\boldsymbol{M}_1 \setminus \boldsymbol{T}_2$ and $\boldsymbol{M}_2 \setminus \boldsymbol{T}_1$ can not cut off all information channels between $\boldsymbol{T}$ and $\boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$, because some information can flow through $\delta_{\boldsymbol{T}_1,\boldsymbol{T}_2;\boldsymbol{V}\setminus\boldsymbol{N}\setminus\boldsymbol{T}|\boldsymbol{N}}$, an information equivalent valve of $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ with respect to $\boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$ given $\boldsymbol{N}$. In other words, $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ may exchange information directly; besides, they also share the equivalent information about $\boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$. This indicates we should continue to turn off some valves, $\boldsymbol{S} \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$, besides $\boldsymbol{M}_1 \setminus \boldsymbol{T}_2$ and $\boldsymbol{M}_2 \setminus \boldsymbol{T}_1$ such that $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ no longer exchange information through external valves and thus such that $\boldsymbol{T}$ has no information exchange with remaining valves.

This analysis motivates us to give the following definition:

Definition 3 *With the notations above, we call $\boldsymbol{S}$ ($\subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$) a Markov blanket supplementary (MbS) (of $\boldsymbol{T}$ to $\boldsymbol{N}$), if $\boldsymbol{N} \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}$. Further, a Markov boundary supplementary (MBS) is any MbS such that none of its proper subsets is an MbS.* ■

In what follows, we give the properties of MbS and MBS.

Theorem 4 *Assume $\boldsymbol{S} \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$. Then, the following statements are equivalent:*

(i) *$\boldsymbol{S}$ is an MbS;*

(ii) $\mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}) = \min_{\boldsymbol{S}' \subseteq \boldsymbol{V}\setminus\boldsymbol{N}\setminus\boldsymbol{T}} \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}')$*;*

(iii) $\mathbb{I}(\boldsymbol{T}; \boldsymbol{S} \,|\, \boldsymbol{N}) = \max_{\boldsymbol{S}' \subseteq \boldsymbol{V}\setminus\boldsymbol{N}\setminus\boldsymbol{T}} \mathbb{I}(\boldsymbol{T}; \boldsymbol{S}' \,|\, \boldsymbol{N})$*;*

(iv) *$\boldsymbol{N} \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}_1$ in $\boldsymbol{V} \setminus \boldsymbol{T}_2$ (or $\boldsymbol{N} \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}_2$ in $\boldsymbol{V} \setminus \boldsymbol{T}_1$).*

In addition, if $\boldsymbol{S}$ is an MbS, then it is also an MBS if and only if $\boldsymbol{T}_1 \not\perp\!\!\!\perp Y \,|\, \boldsymbol{N} \cup (\boldsymbol{S} \setminus \{Y\})$ or $\boldsymbol{T}_2 \not\perp\!\!\!\perp Y \,|\, \boldsymbol{N} \cup (\boldsymbol{S} \setminus \{Y\})$ holds for any $Y \in \boldsymbol{S}$. ■

The proof of this theorem is presented in Appendix B.

As seen, (ii) and (iii) of Theorem 4 explain the implication of MbS that the information flow metaphor illustrates in Figure 7: finding an MbS is equivalent to turning off some valves such that $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ no longer exchange information through external valves, or equivalent to finding all remaining equivalent information contained by $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$; (iv) and the property of MBS provide a practical way of building MBS discovery algorithms.

Here, we use an example to demonstrate the notions of MbS and MBS and their properties.

Example 1 *Consider the BN* $(\mathbb{G}, \mathbb{P})$ *over* $\boldsymbol{V} = \{A, B, C, D\}$ *presented in Figure 8, in which A, B, and C take* $\{1, 2, 3\}$ *while D takes* $\{1, 2\}$*. Put* $\boldsymbol{T} = \{T_1, T_2\}$*,* $\boldsymbol{N} = (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T} = \emptyset$*, and* $\boldsymbol{S} = \{C\}$*,* $\boldsymbol{S}_0 = \{C, D\}$ *with* $T_1 = A$*,* $T_2 = B$*,* $\boldsymbol{M}_1 = \{B\}$*,* $\boldsymbol{M}_2 = \{A\}$*. Using the theory of information equivalence (Lemeire, 2007), we can show the following results (see Appendix B for the proofs):*

(i) $\boldsymbol{M}_1$ *is an MB of* T_1 *in* $\boldsymbol{V}$*:* $\mathbb{I}(A; C, D \,|\, B) = 0$ *and* $\mathbb{I}(A; C, D) > 0$*;*

(ii) $\boldsymbol{M}_2$ *is an MB of* T_2 *in* $\boldsymbol{V}$*:* $\mathbb{I}(B; C, D \,|\, A) = 0$ *and* $\mathbb{I}(B; C, D) > 0$*;*

(iii) $\boldsymbol{N} \cup \boldsymbol{S}$ *is an Mb of* $\boldsymbol{T}$ *in* $\boldsymbol{V}$*, so* $\boldsymbol{S}$ *is an MbS:* $\mathbb{I}(A, B; D \,|\, C) = 0$*;*

(iv) $\mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}) = \min_{\boldsymbol{S}' \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}} \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}')$*, because of* $\mathbb{I}(A; B \,|\, C) = \mathbb{I}(A; B \,|\, C, D)$*,* $\mathbb{I}(A; B \,|\, C) \leqslant \mathbb{I}(A; B \,|\, D)$*, and* $\mathbb{I}(A; B \,|\, C) \leqslant \mathbb{I}(A; B)$*;*

(v) $\mathbb{I}(\boldsymbol{T}; \boldsymbol{S} \,|\, \boldsymbol{N}) = \max_{\boldsymbol{S}' \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}} \mathbb{I}(\boldsymbol{T}; \boldsymbol{S}' \,|\, \boldsymbol{N})$*;*

(vi) $\boldsymbol{N} \cup \boldsymbol{S}$ *is an MB of* T_1 *in* $\boldsymbol{V} \setminus \{T_2\}$*:* $\mathbb{I}(A; C, D) > 0$ *and* $\mathbb{I}(A; D \,|\, C) = 0$*;*

(vii) $\boldsymbol{N} \cup \boldsymbol{S}$ *is an MB of* T_2 *in* $\boldsymbol{V} \setminus \{T_1\}$*:* $\mathbb{I}(B; C, D) > 0$ *and* $\mathbb{I}(B; D \,|\, C) = 0$*;*

(viii) $\boldsymbol{S}$ *is an MBS;* $\boldsymbol{S}_0$ *is an MbS (not an MBS):* $\mathbb{I}(A, B; C, D) > 0$ *and* $\mathbb{I}(A; B \,|\, C, D) = \mathbb{I}(A; B \,|\, C)$*.* ■

By Example 1, A and B share the equivalent information about C, so turning off the valve A (or B) means cutting off all the channels from B (or A) to C. This is why they can screen off each other from C. However, A and B lose the shield if they are integrated into a whole. In this case, we have to turn C off such that A and B no longer exchange information through external valves. This example reveals that an MBS is a minimal set of variables, $\boldsymbol{S} \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$, such that $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ contain no equivalent information about the remaining variables given $\boldsymbol{N} \cup \boldsymbol{S}$.

When finding an MBS, $\boldsymbol{S}$, and letting the variables in $\boldsymbol{S}$ supplement $\boldsymbol{N}$, there may be some variables in $\boldsymbol{N}$ becoming redundant. In addition, $\boldsymbol{N}$ may be redundant even before supplementing $\boldsymbol{S}$. Example 2 gives an illustration. For both cases, we need to remove the redundant variables.

Example 2 *Consider the BN presented in Figure 9, in which any one variable from* $\{A, B, C\}$ *and another from* $\{D, E, F\}$ *(denoted by X and Y, respectively) contain context-independent equivalent information about G (see Statnikov et al., 2013, Example 3). Then,* $\{X, Y\}$ *is an MB of G. Put now* $\boldsymbol{T}_1 = \{C, F\}$ *and* $\boldsymbol{T}_2 = \{G\}$*, and take* $\boldsymbol{M}_1 = \{B, E\}$ *and* $\boldsymbol{M}_2 = \{B, D\}$*. Note that* $\boldsymbol{T}_1 \subseteq \boldsymbol{V} \setminus \boldsymbol{M}_2$ *(and also* $\boldsymbol{T}_2 \subseteq \boldsymbol{V} \setminus \boldsymbol{M}_1$*). It concludes that* $\boldsymbol{N} = \{B, D, E\}$ *is not an MB but only an Mb of* $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$*, since its proper subset* $\{B, E\}$ *is also an Mb (and also an MB) of* $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$*. This shows why the process of refining* $\boldsymbol{N}$ *is necessary.* ■

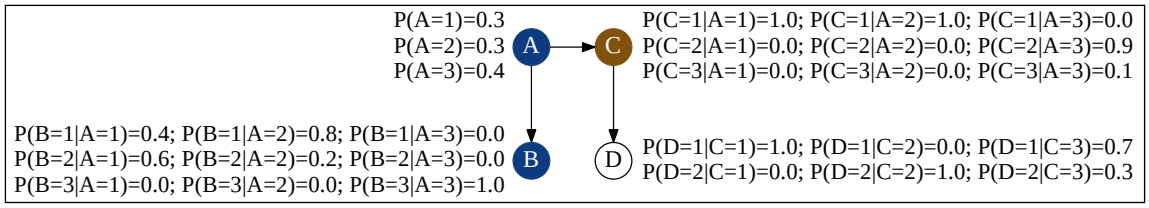

Figure 8: BN $(\mathbb{G}, \mathbb{P})$: $\mathbb{P}$ is a joint probability distribution over $\boldsymbol{V} = \{A, B, C, D\}$ with each variable taking values $\{1, 2, 3\}$ except for D taking $\{1, 2\}$; $\mathbb{G}$ is a DAG over $\boldsymbol{V}$.

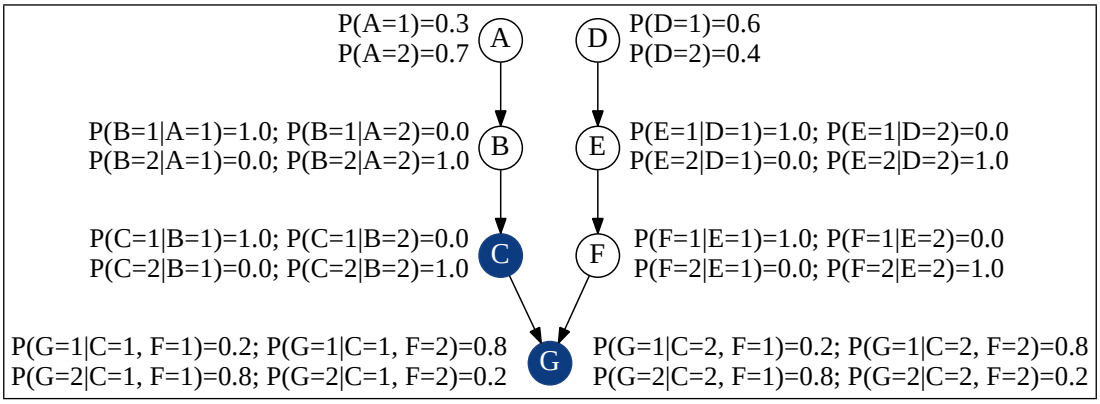

Figure 9: BN $(\mathbb{G}, \mathbb{P})$: $\mathbb{P}$ is a joint probability distribution over $\boldsymbol{V} = \{A, B, C, D, E, F, G\}$ with all variables taking values $\{1, 2\}$, $\mathbb{G}$ is a DAG with the variables in $\boldsymbol{V}$ as its nodes.

3.3 An Alternative Approach

Before building MB discovery algorithms for multiple targets, this subsection concisely presents an alternative approach to the *additivity based* and *MBS based* methods. In Section 5, we will apply this method as an FS strategy in multi-class prediction problems.

Let $\boldsymbol{T} \triangleq \{T_1, \cdots, T_k\}$ be the targets of interest and T be $\boldsymbol{T}$'s merged version, taking values $\{1, \cdots, t\}$ with $t \geqslant 3$. This procedure transforms the MB discovery for multiple targets $\boldsymbol{T}$ into the MB discovery for single target T, so all the existing MB discovery algorithms can be employed theoretically if the required conditions are satisfied. However, if t is large, selecting features of $\boldsymbol{T}$ or T directly will be difficult. Subsection 3.1 and Subsection 3.2 provide a way of solving this problem in different situations. In this case, an alternative strategy is to further convert T into a set of dummy variables denoted by $\left\{T_j^{(d)}\right\}_{j=1}^{t}$, where $T_j^{(d)}$ is a 0-1 variable defined as

$$T_j^{(d)} = \begin{cases} 1, & \text{if } T = j \\ 0, & \text{if } T \neq j \end{cases}$$

This transformation produces a multiple-target $\boldsymbol{T}^{(d)} \triangleq \left(T_1^{(d)}, \cdots, T_t^{(d)}\right)$. Clearly, $\boldsymbol{T}$, T, and $\boldsymbol{T}^{(d)}$ have the same MBs. In what follows, we show the MB of $\boldsymbol{T}^{(d)}$ can be derived by simply taking the union of MBs of $T_1^{(d)}, \cdots, T_t^{(d)}$ and then removing the redundant variables in an efficient way. The proof will be given in Appendix B.

Theorem 5 *Let $\boldsymbol{M}_j$ be an MB of $T_j^{(d)}$ in $\boldsymbol{V} \setminus \boldsymbol{T}$ for $j = 1, \cdots, t$. Then, $\boldsymbol{M} \triangleq \cup_{j=1}^{k} \boldsymbol{M}_j$ is an Mb of $\boldsymbol{T}$. Further, $\boldsymbol{M}$ is an MB of $\boldsymbol{T}$ iff for any $X \in \boldsymbol{M}$ there is some j such that $T_j^{(d)} \not\perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\}$.* ■

For why this transformative method is efficient, the fourth concluding remark in Section 7 will make a brief explanation.

4. Algorithms

This section builds MB discovery algorithms for multiple targets, $\{T_1, \cdots, T_k\} \triangleq \boldsymbol{T}$.

Let $\mathbb{A}$ be an MB discovery algorithm, assumed to perform well when used to discover an MB for a single target. In this paper, we employ `IAMB` and `KIAMB` as $\mathbb{A}$. Clearly, $\mathbb{A}$ can be directly used to find an MB for $\boldsymbol{T}$ if simply regarding $\boldsymbol{T}$ as the input of $\mathbb{A}$. Usually, this will lead to low accuracies and high complexities.

By Theorem 4, the MB discovery problem for multiple targets can be translated equivalently into a number of MB discovery problems for single targets, according to the following way: (i) use $\mathbb{A}$ to find an MB of T_i in $\boldsymbol{V}$ for $i = 1, \cdots, k$, denoted by $\boldsymbol{M}_i$; (ii) find an Mb of T_2 in $\boldsymbol{V} \setminus \{T_1\}$ based on $(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \{T_1, T_2\}$, and then get an MB of $\{T_1, T_2\}$, written as $\boldsymbol{M}_{1,2}$; (iii) find an Mb of T_3 in $\boldsymbol{V} \setminus \{T_1, T_2\}$ based on $(\boldsymbol{M}_{1,2} \cup \boldsymbol{M}_3) \setminus \{T_1, T_2, T_3\}$, and then get an MB of $\{T_1, T_2, T_3\}$, written as $\boldsymbol{M}_{1,2,3}$; (iv) the rest can be done in a similar manner. Following this way, the input of $\mathbb{A}$ for each use is a single variable, so this idea successfully avoids assigning an multivariate input to $\mathbb{A}$. Note that in the above process the equivalent information is extracted in a stepwise manner.

4.1 `IAMBS` and `KIAMBS`

Let $(\mathbb{G}, \mathbb{P})$ be a BN over $\boldsymbol{V}$, and assume $\boldsymbol{T}_i \subseteq \boldsymbol{V}$ with $|\boldsymbol{T}_i| \geqslant 1$ has an MB, $\boldsymbol{M}_i$, for $i = 1, 2$. Denote $\boldsymbol{N} = (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus (\boldsymbol{T}_1 \cup \boldsymbol{T}_2)$. This subsection presents the algorithms for discovering an MB of $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$. To design one such algorithm, we note that there may be some variables in $\boldsymbol{N}$ becoming redundant once an MBS, $\boldsymbol{S}$, is supplemented. Therefore, we need to first find $\boldsymbol{S}$ by setting $\boldsymbol{N}$ as a whitelist in $\mathbb{A}$ and then refine $\boldsymbol{N}$.

Applying this idea to `IAMB` and `KIAMB`, we obtain two algorithms called `IAMBS` and `KIAMBS`, in which "S" refers to as "supplementary". Their pseudo codes are presented in Algorithm 1. In order to differentiate these two algorithms, we set K in the `KIAMBS` algorithm as $K \in [0, 1)$. It is mentioned here that $\boldsymbol{S}_2$ is a random subset of $\boldsymbol{S}_1$ with size $\max\{1, \lfloor |\boldsymbol{S}_1| \cdot K \rfloor\}$ in Line 5 of `KIAMBS`. As seen, these two algorithms first find an MbS, $\boldsymbol{S}$, in the growing phase and then refine $\boldsymbol{S}$ and $\boldsymbol{N}$ in sequence in the shrinking phase.

For example, based on a data set drawn from the BN in Example 1, the unique MB, $\{C\}$, of $\{A, B\}$ can be discovered by calling `IAMBS` or `KIAMBS` only once.

Theorem 1 presents the correctness of `IAMB` and `KIAMB` under the assumption that $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$ satisfies the local composition property. The theorem below shows `IAMBS` and `KIAMBS` are correct if $\boldsymbol{T}_2$ (instead of $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$) satisfies the local composition property. Appendix B gives the proof.

Theorem 6 (Correctness of `IAMBS` and `KIAMBS`) *Assume that $\boldsymbol{T}_2$ satisfies the local composition property, and that all CI tests are correct. Then* (i) *`IAMBS` outputs an MB of $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$;* (ii) *`KIAMBS` outputs an MB of $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$ for any $K \in [0, 1)$.* ■

The following remark presents a relation among local intersection, local composition, and the adjacency faithfulness condition, under the *orientation faithfulness condition*. The proof is given in Appendix B. Here, the orientation faithfulness condition (Ramsey et al., 2006; Lemeire et al., 2012) is defined as: for any $X, Y, Z \in \boldsymbol{V}$ such that X and Z are adjacent to Y but X is not adjacent to Z, (i) if $X \to Y \leftarrow Z$, then $X \not\perp\!\!\!\perp Z \,|\, \boldsymbol{W}$ holds for any $\boldsymbol{W} \subseteq \boldsymbol{V} \setminus \{X, Z\}$ with $Y \in \boldsymbol{W}$; (ii) otherwise, $X \not\perp\!\!\!\perp Z \,|\, \boldsymbol{W}$ holds for any $\boldsymbol{W} \subseteq \boldsymbol{V} \setminus \{X, Y, Z\}$.

Remark 2 *The following two statements hold:* (a) *violating local intersection implies violating adjacency faithfulness;* (b) *under the orientation faithfulness condition, violating local composition at the end of the first phase of `IAMB` or `KIAMB` or `IAMBS` or `KIAMBS` means violating adjacency faithfulness.* ■

In addition, the lemma below is useful to explain the succeeding remarks.

Lemma 3 (a) *If there is* $\boldsymbol{P} \subseteq \boldsymbol{M}_1 \setminus \boldsymbol{T}_2$ *such that* $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{P} \,|\, (\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_2$*, then* $(\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_2$ *is an Mb of* $\boldsymbol{T}_1$*;* (b) *If there is* $\boldsymbol{Q} \subseteq \boldsymbol{N} \setminus \boldsymbol{P}$ *such that* $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{Q} \,|\, (\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_2$ *and* $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{Q} \,|\, (\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_1$*, then* $(\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_2$ *is an Mb of* $\boldsymbol{T}_1$*, and* $(\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_1$ *is an Mb of* $\boldsymbol{T}_2$*.* ■

For Algorithm 1, we have three remarks below:

(i) In these two algorithms, the two CI tests for adding members to $\boldsymbol{S}$ and for refining $\boldsymbol{S}$ are based on $\boldsymbol{T}_2$ instead of $\boldsymbol{T}$, while the CI test for refining $\boldsymbol{N}$ is based on $\boldsymbol{T}$ instead of $\boldsymbol{T}_2$. (a) For the first two CI tests, $\boldsymbol{T}_2$ can be replaced with $\boldsymbol{T}$ without affecting the correctness of the algorithms, since $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{N} \cup \boldsymbol{S}' \Leftrightarrow \boldsymbol{T} \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{N} \cup \boldsymbol{S}'$ holds for any $\boldsymbol{S}' \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$ and $\boldsymbol{X} \subseteq \boldsymbol{V} \setminus (\boldsymbol{N} \cup \boldsymbol{S}') \setminus \boldsymbol{T}$. However, if we replace $\boldsymbol{T}_2$ with $\boldsymbol{T}$, the resulting algorithms will need much longer time to run. This is why we use $\boldsymbol{T}_2$ in stead of $\boldsymbol{T}$ in these two places. (b) For the third CI test, $\boldsymbol{T}$ can not be replaced with $\boldsymbol{T}_2$, because of $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{X} \,|\, (\boldsymbol{N} \setminus \boldsymbol{X}) \cup \boldsymbol{S} \nRightarrow \boldsymbol{T} \perp\!\!\!\perp \boldsymbol{X} \,|\, (\boldsymbol{N} \setminus \boldsymbol{X}) \cup \boldsymbol{S}$.

(ii) According to Remark 2, there may be some situations in which both local intersection and local composition are simultaneously violated. In this case, `IAMBS` and `KIAMBS` may not

Algorithm 1: **`IAMBS`** and **`KIAMBS`**

Procedure: $\boldsymbol{M} \leftarrow$ `IAMBS`$(\boldsymbol{D}; \boldsymbol{T}_1, \boldsymbol{T}_2; \boldsymbol{M}_1, \boldsymbol{M}_2)$
Input: a data matrix $\boldsymbol{D}$; two sets of targets $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$; an MB $\boldsymbol{M}_i$ of $\boldsymbol{T}_i$ for $i = 1, 2$.
Output: an MB, $\boldsymbol{M}$, of $\boldsymbol{T} \triangleq \boldsymbol{T}_1 \cup \boldsymbol{T}_2$.

```
//Forward: Growing Phase
S ← ∅
while S has changed do
    M ← N ∪ S
    Y ← arg max_{X∈V\M\T} f_D(T_2; X | M)
    if T_2 ⫫̸ Y | M then
        S ← S ∪ {Y}
    end
end
//Backward: Shrinking Phase
foreach X ∈ S do
    if T_2 ⫫ Y | N ∪ (S \ {Y}) then
        S ← S \ {Y}
    end
end
foreach Y ∈ N do
    if T ⫫ Y | (N \ {Y}) ∪ S then
        N ← N \ {Y}
    end
end
return M ← N ∪ S
```

Procedure: $\boldsymbol{M} \leftarrow$ `KIAMBS`$(\boldsymbol{D}; \boldsymbol{T}_1, \boldsymbol{T}_2; \boldsymbol{M}_1, \boldsymbol{M}_2; K)$
Input: Besides $\{\boldsymbol{D}, \boldsymbol{T}_i, \boldsymbol{M}_i\}$, $K \in [0, 1)$ is a randomization parameter.
Output: an MB, $\boldsymbol{M}$, of $\boldsymbol{T} \triangleq \boldsymbol{T}_1 \cup \boldsymbol{T}_2$.

```
//Forward: Growing Phase
S ← ∅
while S has changed do
    M ← N ∪ S
    if S_1 ← {X∈V\M\T: T_2 ⫫̸ X | M} ≠ ∅ then
        Y ← arg max_{X∈S_2} f_D(T_2; X | M)
        S ← S ∪ {Y}
    end
end
//Backward: Shrinking Phase
foreach X ∈ S do
    if T_2 ⫫ Y | N ∪ (S \ {Y}) then
        S ← S \ {Y}
    end
end
foreach Y ∈ N do
    if T ⫫ Y | (N \ {Y}) ∪ S then
        N ← N \ {Y}
    end
end
return M ← N ∪ S
```

correctly work. Specifically, the *violation of local intersection* means $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ contain equivalent information about $\boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$ given $\boldsymbol{N}$; while the *violation of local composition* indicates not all equivalent information are successfully extracted by $\boldsymbol{N}$. Let $\boldsymbol{P}$ and $\boldsymbol{Q}$ be defined as in Lemma 3, and assume $\boldsymbol{P} \cup \boldsymbol{Q} \neq \emptyset$. Then, it can be shown that $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ contain equivalent information about $\boldsymbol{P} \cup \boldsymbol{Q}$ given $\boldsymbol{N} \setminus (\boldsymbol{P} \cup \boldsymbol{Q})$. This means some equivalent information about $\boldsymbol{P} \cup \boldsymbol{Q}$ shared by $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ conditioned on $\boldsymbol{N} \setminus (\boldsymbol{P} \cup \boldsymbol{Q})$ may mask some equivalent information about $\boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$ contained by $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ conditioned on $\boldsymbol{N}$. This may be why not all equivalent information can be extracted by $\boldsymbol{N}$. According to this analysis, a potential remedy is to run `IAMBS` or `KIAMBS` by replacing $\boldsymbol{N}$ with a superset of $\boldsymbol{N} \setminus (\boldsymbol{P} \cup \boldsymbol{Q})$ that is a subset of $\boldsymbol{N}$.

(iii) By Remark 1, if $\boldsymbol{T}_1 \subseteq \boldsymbol{V} \setminus \boldsymbol{M}_2$ or $\boldsymbol{T}_2 \subseteq \boldsymbol{V} \setminus \boldsymbol{M}_1$, $\boldsymbol{N}$ must be an Mb of $\boldsymbol{T}$, so Lines 2∼14 of `IAMBS` and `KIAMBS` can be omitted. In this case, however, it is still necessary to refine $\boldsymbol{N}$, because $\boldsymbol{N}$ may not possess minimality. Example 2 illustrates this necessity.

In addition, another problem that we concern is whether we can refine $\boldsymbol{N}$ before seeking $\boldsymbol{S}$ and, if this is the case, which variables in $\boldsymbol{N}$ can be removed directly. We consider this problem because any redundant variable in $\boldsymbol{N}$ can lead to unnecessary inaccuracies when using $\boldsymbol{N}$ as a part of the conditional set in practical computations. Lemma 3 indicates we can do like this. However, to avoid the danger of missing the information about $\boldsymbol{P} \cup \boldsymbol{Q}$ (this occurs if the equivalent information involved in $\boldsymbol{P} \cup \boldsymbol{Q}$ given $\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}$ is different in some sense from any part of the equivalent information involved in $\boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$ given $\boldsymbol{N}$), we recommend to first search the members of $\boldsymbol{S}$ in $\boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$ and then check if some variables in $\boldsymbol{P} \cup \boldsymbol{Q}$ are necessary to enter into $\boldsymbol{S}$ when implementing Lines 2∼8 of `IAMBS` and `KIAMBS`. Note that this will increase the total running time.

4.2 `MIAMB` and `MKIAMB`

In this subsection, we present two multivariate Markov boundary discovery algorithms, called `MIAMB` and `MKIAMB`, respectively.

Let $\{T_1, \cdots, T_k\} \in \boldsymbol{V}$ with $\boldsymbol{M}_i$ as its an MB for $i = 1, \cdots, k$. If the local intersection property is satisfied, Theorem 3 shows $\bigcup_{i=1}^{k} \boldsymbol{M}_i \setminus \boldsymbol{T}$ is an MB of $\boldsymbol{T} \triangleq \{T_1, \cdots, T_k\}$. Otherwise, $\boldsymbol{M}$ may be no longer an MB. In this case, we use `MIAMB` or `MKIAMB` to seek an MB for $\boldsymbol{T}$. Given an ordering of $T_1, \cdots, T_k$, saying $\tau \triangleq \{i_1, \cdots, i_k\}$, which determines the priorities of the variables in $\boldsymbol{T}$ entering into the queue whose an MB will be sought in the current step, we denote an MB of $\{T_{i_1}, \cdots, T_{i_\ell}\} \triangleq \boldsymbol{T}_{i_\ell}^*$ by $\boldsymbol{M}_{i_j}^*$.

With these notations, `MIAMB` and `MKIAMB` are pseudo-coded in Algorithm 2. Their correctness, shown by Theorem 7, is a direct consequence of Theorem 1 and Theorem 6. As seen, `MIAMB` or `MKIAMB` uses the following stepwise idea: it first finds an MB of two targets $\{T_{i_1}, T_{i_2}\} = \{T_{i_1}\} \cup \{T_{i_2}\}$, and then finds an MB of three targets $\{T_{i_1}, T_{i_2}, T_{i_3}\} = \{T_{i_1}, T_{i_2}\} \cup \{T_{i_3}\}$; the rest can be done in a similar manner until all the k target variables are considered.

Theorem 7 (Correctness of `MIAMB` and `MKIAMB`) *Assume that T_i satisfies the local composition property for $i = 1, \cdots, k$, and that all CI tests are correct. Denote $\boldsymbol{T} \triangleq \{T_1, \cdots, T_k\}$. Then* (i) *`MIAMB` outputs an MB of $\boldsymbol{T}$;* (ii) *`MKIAMB` outputs an MB of $\boldsymbol{T}$ for any $K \in [0, 1)$.* ■

As we know, for any real data, those preconditions (such as faithfulness or local composition) required by a learning algorithm are hard to hold exactly. However, our algorithms can be seen as an improvement over earlier methods. Specifically, `IAMB`/`KIAMB` algorithms require faithfulness or

local composition for multiple targets when used for MB discovery of multiple targets, while our MIAMB/MKIAMB only need local composition for single targets, which may be more close to real situations than faithfulness or local composition for multiple targets.

For MIAMB or MKIAMB, an ordering τ is set in Algorithm 2 mainly because different orderings may lead to different computational complexities. In Subsection 4.4, we will make a complexity analysis about the algorithms, based on which we present a feasible way of selecting τ, under the expectation that our algorithms should be run as quickly as possible. When $|\boldsymbol{T}| = 2$, however, τ is not necessary.

Besides MIAMB/MKIAMB algorithms (which are *MBS based*), we can consider additivity based (Theorem 3) and dummy variables based (Theorem 5) algorithms: (a) the *additivity based* MIAMB or MKIAMB simply takes the union of outputs of IAMB/KIAMB with respect to all single targets as the output; its correctness requires the conditions in Theorem 7 plus Theorem 3; and (b) the *dummy (variables based)* MIAMB/MKIAMB takes the union of the outputs of IAMB/KIAMB with respect to every dummy variable and removes redundant variables; its correctness requires the same condition as in Theorem 7. Throughout this paper, unless specified, MIAMB/MKIAMB denote the MBS based algorithms.

4.3 A Discussion on CI Test

As argued by Aliferis et al. (2010a, p. 200), the quality of an MB discovery algorithm highly depends on the selected CI testing methods. In this subsection, we discuss the ways of practically doing CI tests. Usually, the Pearson's X^2 test or the log-likelihood ratio G^2 test can be employed for this purpose (Yaramakala, 2004; Bromberg and Margaritis, 2009; Aliferis et al., 2010b; Statnikov et al., 2013). Here, the X^2 statistic and the G^2 statistic have the same asymptotic χ^2 distribution. We can also use some experimental testing methods such as the Akaike information criterion-based test (Cressie and Read, 1989; Scutari, 2010).

Algorithm 2: MIAMB and MKIAMB

Procedure: $\boldsymbol{M} \leftarrow$ MIAMB$(\boldsymbol{D}; \boldsymbol{T}; \tau)$
Input: a data matrix $\boldsymbol{D}$; a target set $\boldsymbol{T} \triangleq \{T_1, \cdots, T_k\}$; and an ordering $\tau \triangleq \{i_1 \cdots, i_k\}$.
Output: an MB, $\boldsymbol{M}$, of $\boldsymbol{T}$.

// MIAMB: $\boldsymbol{M} \leftarrow$ MIAMB$(\boldsymbol{D}; \boldsymbol{T}; \tau)$
for $\ell \leftarrow 1$ **to** k **do**
  $\boldsymbol{M}_{i_\ell} \leftarrow$ IAMB$(\boldsymbol{D}; \{T_{i_\ell}\})$
end
for $\ell \leftarrow 2$ **to** k **do**
  $\boldsymbol{M}^*_{i_\ell} \leftarrow$ IAMBS$(\boldsymbol{D}; \boldsymbol{T}^*_{i_{\ell-1}}, \{T_{i_\ell}\}; \boldsymbol{M}^*_{i_{\ell-1}}, \boldsymbol{M}_{i_\ell})$
end
return $\boldsymbol{M} \leftarrow \boldsymbol{M}^*_{i_k}$

Procedure: $\boldsymbol{M} \leftarrow$ MKIAMB$(\boldsymbol{D}; \boldsymbol{T}; K; \tau)$
Input: a data matrix $\boldsymbol{D}$; a target set $\boldsymbol{T}$; a randomization parameter $K \in [0, 1)$; and an ordering τ.
Output: an MB, $\boldsymbol{M}$, of $\boldsymbol{T}$.

// MKIAMB: $\boldsymbol{M} \leftarrow$ MKIAMB$(\boldsymbol{D}; \boldsymbol{T}; K; \tau)$
for $\ell \leftarrow 1$ **to** k **do**
  $\boldsymbol{M}_{i_\ell} \leftarrow$ KIAMB$(\boldsymbol{D}; \{T_{i_\ell}\}; K)$
end
for $\ell \leftarrow 2$ **to** k **do**
  $\boldsymbol{M}^*_{i_\ell} \leftarrow$ KIAMBS$(\boldsymbol{D}; \boldsymbol{T}^*_{i_{\ell-1}}, \{T_{i_\ell}\}; \boldsymbol{M}^*_{i_{\ell-1}}, \boldsymbol{M}_{i_\ell}; K)$
end
return $\boldsymbol{M} \leftarrow \boldsymbol{M}^*_{i_k}$

Recall that we are dealing with the MB discovery problem for *multiple* target variables. When the target set, namely $\boldsymbol{T}$, contains only a few variables (e.g., 1 or 2), the X^2 test or the G^2 test performs quite well in most situations. Unfortunately, when $\boldsymbol{T}$ contains too many variables (e.g., 5 or 6 or even more), X^2 or G^2 may not work well due to the overmany degrees of freedom. See Appendix C for a detailed discussion. In fact, as Cochran (1954, p. 420) recommended about the working rules for X^2 (also applicable to G^2), these two testing methods are unreliable if more than 20% of the cells in contingency tables have an expected count of less than 5 data points; however, such cases frequently arise in practice (Bromberg and Margaritis, 2009; Yaramakala, 2004).

Many authors have considered improving X^2 and G^2 by adjusting the statistics. Lawley (1956) showed that such tests can be improved by multiplying with a suitable scale factor; Hosmane (1986, 1987, 1990) and the pioneer scholars recommended the following two adjustment procedures (i) replace zero observed counts by a positive constant, leaving nonzero counts intact; and (ii) add a positive constant to all the observed counts. Brin et al. (1997) and Silverstein et al. (1998) used two heuristic "solutions" to the problem of low expected counts as follows: (i) simply ignore these cells when calculating X^2 or G^2; and (ii) use what is called *contingency table support* (CT-support): a set of items S has CT-support s at the t% level if at least t% of the cells in the contingency table for S have value s. Aliferis et al. (2010b) considered a similar heuristic called *heuristic power size*, which denotes the smallest sample size per cell in the contingency table of a reliable CI test.

The above ideas can lead to improvements on X^2 and G^2 to varying degrees if the dimensions are not very high. However, when working on the MB discovery problem for multiple targets, we need more suitable methods to do CI tests. For this reason, we suggest the following *practical operation*: when $|\boldsymbol{T}| \leqslant 2$, we can (i) use X^2 or G^2 or their variants mentioned above to do CI tests; otherwise, we consider the following testing method: (ii) use CMI and an experimental threshold, ε, to make statistical decisions as Cheng et al. (2002) did, in the sense that $\mathbb{I}_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) \geqslant \varepsilon$ asserts $\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$ while $\mathbb{I}_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) < \varepsilon$ concludes $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$, where $\varepsilon \triangleq |\boldsymbol{T}|^{a_1} \cdot \frac{100\, a_2}{n} \cdot \log_2 v$ is related to the the sample size, the average number of values that each variable takes, and the number of targets (denoted by n, v, and $|\boldsymbol{T}|$, respectively), in which a_1 and a_2 are two adjusting factors ($a_1 = 0.5$ and $a_2 \in (0.1, 0.5)$ are recommended). The association function, $f_{\boldsymbol{D}}$, can be selected as

$$f_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) = \mathbb{I}_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) \triangleq f^{(2)}_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}). \tag{4}$$

Besides this experimental method, we can (iii) improve X^2 or G^2 by adjusting the number of the theoretical degrees of freedom.

For the above (iii), to be clear, we consider the G^2 statistic, $G^2(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) \triangleq 2n \cdot \mathbb{I}_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$, which approximates to the chi-square variate with $r \triangleq (r_{\boldsymbol{X}} - 1)(r_{\boldsymbol{Y}} - 1) r_{\boldsymbol{Z}}$ degrees of freedom, namely $\chi^2(r)$, where $r_{\boldsymbol{\xi}}$ represents the number of configurations for $\boldsymbol{\xi}$ (de Campos, 2006, p. 2158). Denote the p-value by

$$p(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) = \mathbb{P}\{\chi^2(r) \geqslant G^2(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z})\}.$$

Then, the G^2 test asserts $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$ if $p(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) > \alpha$ for a significance level α, and concludes $\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$ if $p(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) \leqslant \alpha$. In this paper, α is set to be 0.05. Aliferis et al. (2010a, pp. 200–201) provided a further discussion about this. Accordingly, the *negative p-value* is used as the association function, $f_{\boldsymbol{D}}$, as Tsamardinos et al. (2006), Aliferis et al. (2010a,b), and Statnikov et al. (2013) did:

$$f_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) = -p(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) = -\mathbb{P}\{\chi^2(r) \geqslant G^2(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z})\} \triangleq f^{(1)}_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}). \tag{5}$$

Replace the theoretical value of r in $p(\boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z})$ with its a damped version of the form

$$g_{n,\kappa}(r) \triangleq r\left(1-\mathrm{e}^{-\frac{n}{\kappa r}}\right), \tag{6}$$

where $\kappa > 0$ is a constant, based on which $\frac{n}{\kappa}$ measures the amount of valid cells that n sample instances can support. For convenience, we will call the resulted p-value, denoted by $p_g(\boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z})$ instead of $p(\boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z})$, and the resulted testing method to be the *damped p-value* and the *damped log-likelihood ratio test* (or damped G^2 test). Further, we use the the following association function:

$$f_{\boldsymbol{D}}(\boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z}) = -p_g(\boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z}) = -\mathbb{P}\{\chi^2(g_{n,\kappa}(r)) \geqslant G^2(\boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z})\} \triangleq f^{(3)}_{\boldsymbol{D}}(\boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{Z}). \tag{7}$$

In Appendix C, we will provide the details for this damping procedure, and give some numerical illustrations about its reasonability. Clearly, the damped G^2 test approximately degenerates into the ordinary G^2 test when taking κ as a very small positive number.

4.4 Complexity Analysis

In the following, we analyze the computational complexities of the four algorithms: `IAMB`, `KIAMB`, `MIAMB`, and `MKIAMB`. Usually, the number of CI tests can be employed to measure the complexity of a CI-based MB discovery algorithm (Tsamardinos et al., 2003, 2006; Aliferis et al., 2010a), considering there exists efficient implementations of the CMI-based test or the association computation taking time $O(n\log n)$ if the conditional set is small. However, Aliferis et al. (2010a) also mentioned that the running time, denoted by $t_{n,q}$, for computing per CMI-based statistic is linear to the sample size, n, and exponential to the number, q, of variables in the conditional set. This means we should take $t_{n,q}$ into account, not simply using $O(n\log n)$ to measure the complexity.

Assume we are seeking an MB for $\boldsymbol{T} \triangleq \{T_1,\cdots,T_k\}$ according to the ordering τ. Without loss of generality, we assume $\tau = \{1,\cdots,k\}$. Consider the case of $k = 2$. Suppose $\boldsymbol{M}_i$ is an MB of T_i with $|\boldsymbol{M}_i| = m_i \geqslant 1$, and $\boldsymbol{S}$ is an MBS for $\boldsymbol{N} \triangleq \boldsymbol{M}_1 \cup \boldsymbol{M}_2 \setminus \{T_1,T_2\}$ with $|\boldsymbol{S}| = s \geqslant 0$. By Remark 1, we assume $T_1 \in \boldsymbol{M}_2$ and $T_2 \in \boldsymbol{M}_1$. Recall that the number of all variables is p. It follows that:

- In view of $|\boldsymbol{N}\cup\boldsymbol{S}| = m_1+m_2+s-2 \triangleq m$, `IAMB` takes time $O[(mp+m)t_{n,m}]$ to finish an execution. Thus, the complexity of `IAMB` is $O(mpt_{n,m})$. `KIAMB` has almost the same complexity.
- For `MIAMB`, it first takes time $O[(m_1p+m_1)t_{n,m_1}+(m_2p+m_2)t_{n,m_2}]$ to find $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$; then it seeks $\boldsymbol{S}$ and refines $\boldsymbol{N}$ taking time $O\{[s(p-m_1-m_2+2)+m]t_{n,m}\}$. Hence, `MIAMB` needs time $O\{(m_1p+m_1)t_{n,m_1}+(m_2p+m_2)t_{n,m_2}+[s(p-m_1-m_2+2)+m]t_{n,m}\}$ to finish an execution, so its complexity is $O(m_1pt_{n,m_1}+m_2pt_{n,m_2}+spt_{n,m})$. `MKIAMB` has almost the same complexity.

By this analysis, the complexity of `MIAMB` or `MKIAMB` is lower than that of `IAMB` or `KIAMB`. In fact, noting $t_{n,q}$ is exponential to q ($\leqslant m$; meaning $t_{n,q} \ll t_{n,m}$ in most situations) for $q = m_1, m_2$, this implies `MIAMB`/`MKIAMB` are expected to need much less time to run than `IAMB`/`KIAMB`, especially when $\boldsymbol{T}$ contains many variables. The evaluation section (Figure 15) confirms this expectation in the case of moderately large sample size.

For the general case, using the notations in Subsection 4.2 with $|\boldsymbol{M}_i| = m_i$ ($i = 1,\cdots,k$), we assume $\boldsymbol{S}_i$ be an MBS for $\boldsymbol{M}^*_{i-1}$ and $\boldsymbol{M}_i$, with $|\boldsymbol{S}_i| = s_i$ ($i = 2,\cdots,k$). Denote $m^*_i \triangleq \sum_{j=1}^{i} m_j + \sum_{j=2}^{i} s_j - i$. Note that, in general, $t_{n,m_a} \ll t_{n,m^*_a} \ll t_{n,m^*_b}$ for $a < b$. Then, the `IAMB` or `KIAMB` algorithm has the complexity $O(m^*_k pt_{n,m^*_k})$, while `MIAMB` or `MKIAMB` has a lower complexity $O(\sum_{i=1}^{k} m_i pt_{n,m_i} + \sum_{i=2}^{k} s_i pt_{n,m^*_i})$.

According to this theoretical result on complexities, we can use the ordering, $\tau \triangleq \{i_1 \cdots, i_k\}$, in MIAMB or MKIAMB such that $m_{i_1} \leqslant \cdots \leqslant m_{i_k}$. This can reduce the complexities to some extent.

Besides, the additivity based MIAMB/MKIAMB algorithms have almost the same complexity as the MBS based MIAMB/MKIAMB, while the dummy MIAMB/MKIAMB have the complexity $O(m r_{\boldsymbol{T}} p t_{n,m})$, where $m = \sum_{j=1}^{k} m_j$, $r_{\boldsymbol{T}} = \prod_{j=1}^{k} r_{T_j}$, r_ξ denotes the number of configurations for ξ. It will be seen from Section 6 that, although the dummy MIAMB/MKIAMB are of high complexity theoretically, they usually perform well in multi-class prediction problems.

5. Benchmarking Study

This section makes a benchmarking study based on the data sets of six synthetic BNs. These data sets, generated by Tsamardinos et al. (2006) and Aliferis et al. (2010a), and the BNs are briefly described in Table 1. As Tsamardinos et al. (2006) and Aliferis et al. (2010a) stated, these BNs are representatives of a wide range of problem domains. Also, these BNs have different complexities (according to the number of nodes, the number of edges, maximal in-degree, maximal out-degree, and domain range). More details about the BNs and the used data sets are provided by Tsamardinos et al. (2006) and Aliferis et al. (2010a).

The following items are clarified before presenting the experimental results:

- *Measurements*: The primary measurement for the performance of an MB discovery algorithm used in our experiment is the weighted accuracy (WA), which is the average of the rate of true members and that of true nonmembers of an MB with respect to the truth. We also compute what we call the weighted precision (WP) as the average of the rate of true members and that of true nonmembers of an MB with respect to the output. In addition, we record the running time (RT) for every data set of each algorithm and for each BN. Here, RT refers to the single CPU time implemented on an Intel i7-3612QM 2.1 GHz and Windows 7 with 64 bits.

BN	Num. Nodes	Num. Edges	Maximal In-degree	Maximal Out-degree	Domain Range	Selected Targets	Sizes of Data Sets	Total RT (Hours)
Child10	200	257	2	7	2 ~ 6	$X_{131}, X_{132}, X_{98}, X_{194},$ $X_{184}, X_{22}, X_{135}, X_{60}$	5×500	3.4297
							1×5000	8.2220
ALARM10	370	570	4	7	2 ~ 4	$X_{341}, X_{48}, X_{37}, X_{249},$ $X_{209}, X_{188}, X_{192}, X_{161}$	5×500	4.6547
							1×5000	6.3721
Pigs	441	592	2	39	3 ~ 3	$X_{390}, X_{357}, X_{180}, X_{400},$ $X_{199}, X_{241}, X_{228}, X_{176}$	5×500	11.7651
							1×5000	15.8443
Link	724	1125	3	14	2 ~ 4	$X_{369}, X_{293}, X_{303}, X_{457},$ $X_{399}, X_{512}, X_{183}, X_{501}$	5×500	9.7932
							1×5000	16.1046
Lung Cancer	800	1476	4	28	2 ~ 3	$X_1, X_{416}, X_{345}, X_{641},$ $X_{513}, X_{198}, X_{78}, X_{746}$	5×500	16.0301
							1×5000	16.3790
Gene	801	977	4	10	3 ~ 5	$X_{801}, X_{301}, X_{569}, X_{317},$ $X_{185}, X_{622}, X_{516}, X_{577}$	5×500	17.2465
							1×5000	35.6790

Table 1: BNs and data sets.

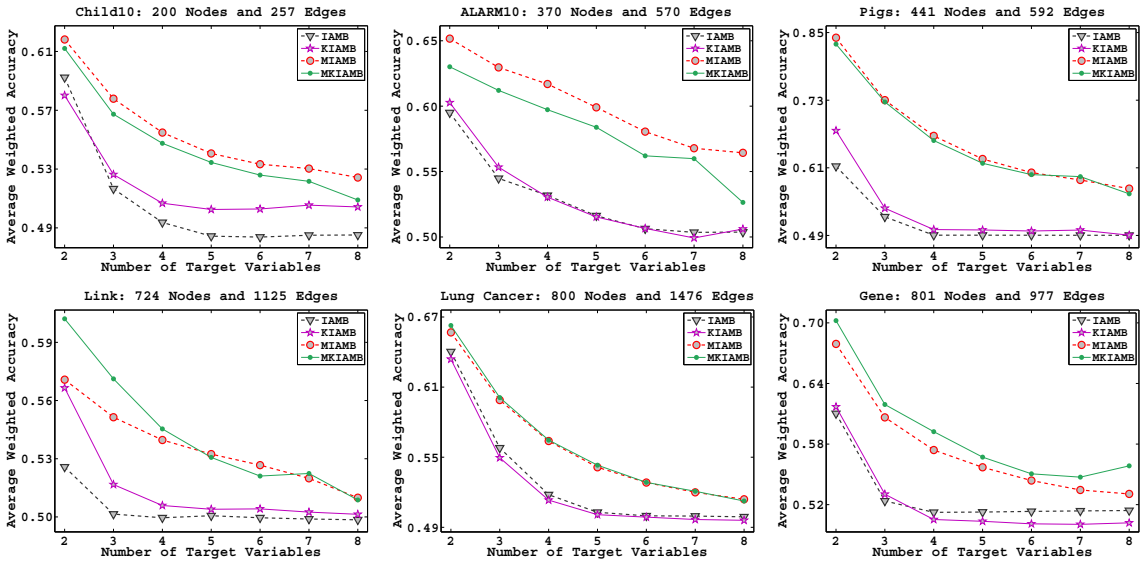

Figure 10: Average WA of the algorithms versus $|\boldsymbol{T}|$ with respect to the data sets of size 500

- *Algorithms*: Four algorithms are used: `IAMB`, `KIAMB`, `MIAMB`, and `MKIAMB`. We take $K = 0.8$ as the randomization parameter in `KIAMB` and `MKIAMB` due to the following two reasons: (i) Peña et al. (2007, p. 227) asserted that $K \in [0.7, 0.9]$ performs best; and (ii) $K = 0.8$ is an appropriate tradeoff between WA (or WP) and RT.
- *Used CI Test*: Following the *practical operation* suggested in Subsection 4.3, we implemented the algorithms via the G^2 test, and found G^2 is suitable for small $|\boldsymbol{T}|$ but is not very suitable and even no longer works for large $|\boldsymbol{T}|$. Then, we used the experimental CMI-based test with a relatively rough $\varepsilon \approx \sqrt{|\boldsymbol{T}|} \cdot \varepsilon_0$, in which $\varepsilon_0 = 0.05$ if $n = 500$ and $\varepsilon_0 = 0.01$ if $n = 5000$; after that, we used the damped G^2 test by setting $\kappa = 5$. The results indicate both alternatives are desirable. Considering the association function, $f^{(2)}_{\boldsymbol{D}}$ defined in (4) corresponding to the CMI-based test, contains no the average number, v, of values that each variable takes, we may need to reselect ε_0 for a BN with a very different v. For these reasons, we eventually decided to use the damped G^2 test for the four algorithms in our experiment.
- *Data*: We use the data sets of sizes 500 and 5000, generated by Tsamardinos et al. (2006) and Aliferis et al. (2010a), which are available at `http://www.nyuinformatics.org/downloads/supplements/JMLR2009/index.html`.
- *Targets*: We employ eight of those variables selected by Aliferis et al. (2010a, p. 226) as the potential targets for each BN. See Table 1 for details. Then, $\boldsymbol{T}$ is any possible combination of k targets for $k = 2, \cdots, 8$.
- *Steps*: For each BN with eight selected targets, the steps of making simulation based on the data set of size 5000 are as follows: (a) for $k = 2, \cdots, 8$, call the four algorithms to obtain four MBs of $\boldsymbol{T} \triangleq \{T_{i_1}, \cdots, T_{i_k}\}$; (b) compute their WAs and WPs, and record the respective RTs; (c) take the average values of these $\binom{8}{k}$ WAs or WPs or RTs for each of the four algorithms. For the five data sets of size 500, each reported WA or WP or RT is the average value of the corresponding five results of an algorithm derived by (a) ~ (c) above.

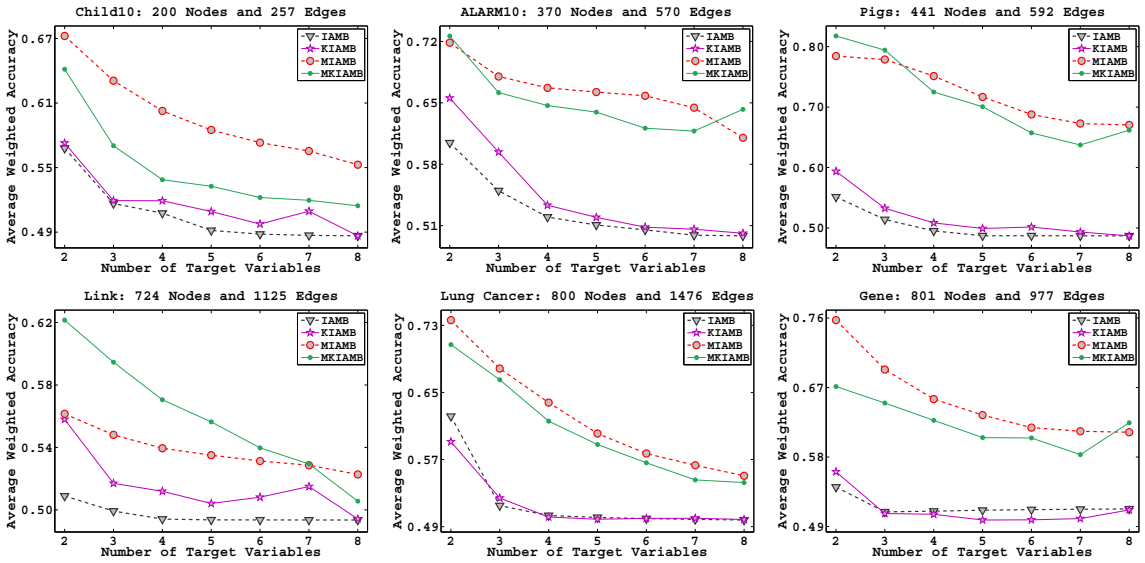

Figure 11: Average WA of the algorithms versus $|\boldsymbol{T}|$ with respect to the data sets of size 5000

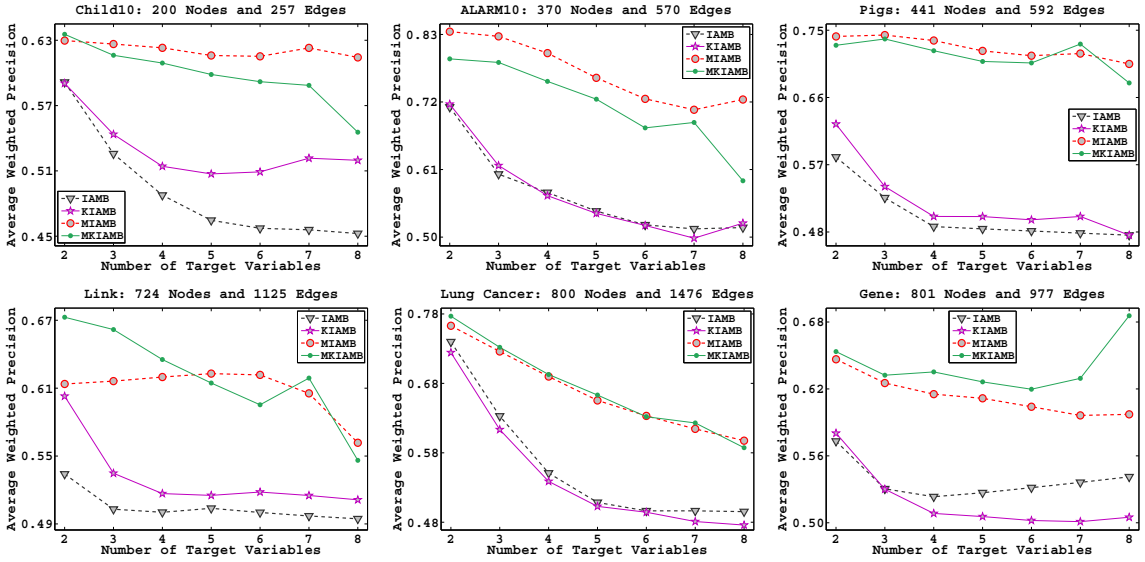

Figure 12: Average WP of the algorithms versus $|\boldsymbol{T}|$ with respect to the data sets of size 500

According to the above description, we make computations with the aid of FullBNT (Murphy, 2007) and MIToolbox (Brown et al., 2012). The results of the WAs are presented in Figure 10 and Figure 11; the results of the WPs are given in Figure 12 and Figure 13; and the results of the RTs are shown in Figure 14, and Figure 15. The total RTs are presented in Table 1. By these figures, it is concluded that, on the whole, our `MIAMB` and `MKIAMB` have higher computational accuracies and lower time complexities than the existing `IAMB` and `KIAMB`.

Specifically, we have:

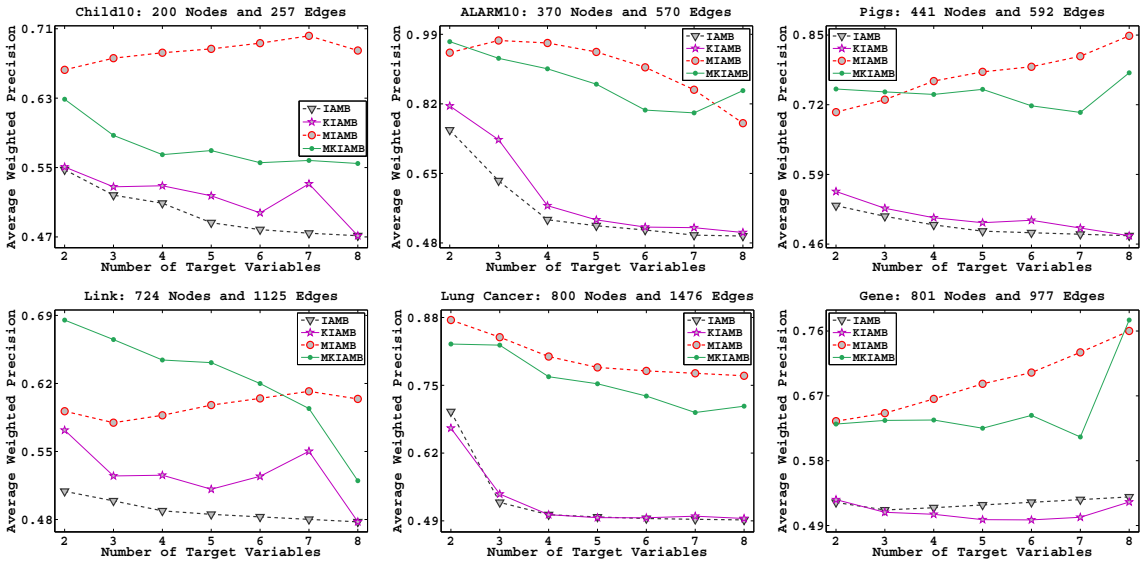

Figure 13: Average WP of the algorithms versus $|\boldsymbol{T}|$ with respect to the data sets of size 5000

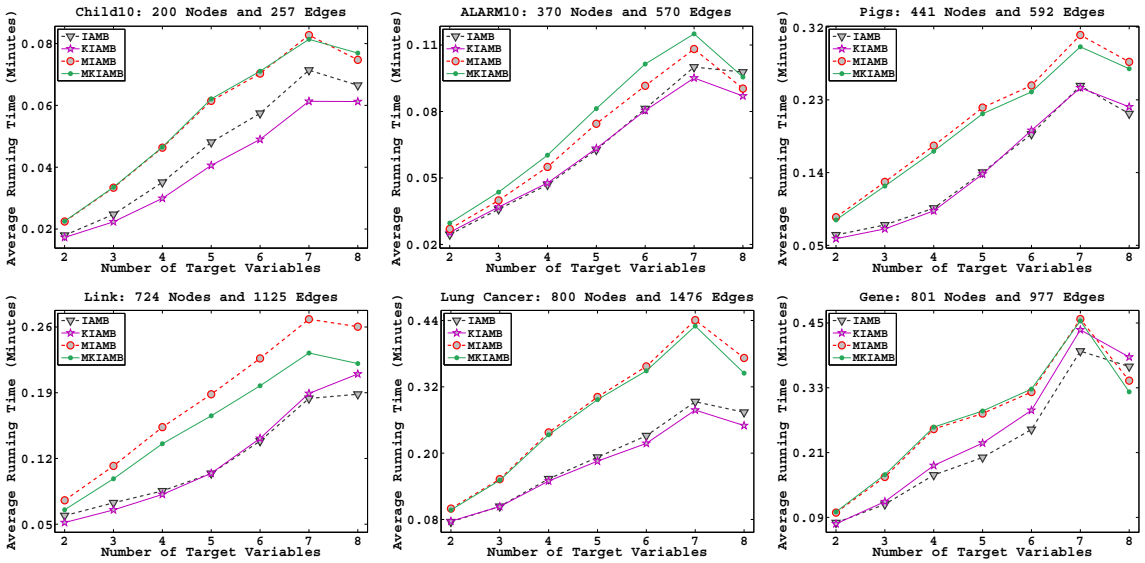

Figure 14: Average RT of the algorithms versus $|\boldsymbol{T}|$ with respect to the data sets of size 500

(i) *Performance on WA*: (a) `MIAMB` and `MKIAMB` have larger WAs than `IAMB` and `KIAMB` for all the six BNs in any case of $|\boldsymbol{T}|$; (b) when $|\boldsymbol{T}|$ increases, WA declines quickly for `IAMB` and `KIAMB`, but it decreases gently for `MIAMB` and `MKIAMB`; and (c) the improvements of `MIAMB` and `MKIAMB` over `IAMB` and `KIAMB` tend to be gradually noticeable and then reduce slightly as $|\boldsymbol{T}|$ increases. The performance degradation along with the increase of $|\boldsymbol{T}|$ can be attributed to two possible aspects: one is that the local composition assumption may be more apt to be violated for a larger $|\boldsymbol{T}|$ because of synergy effects; and the other is that the assumption about

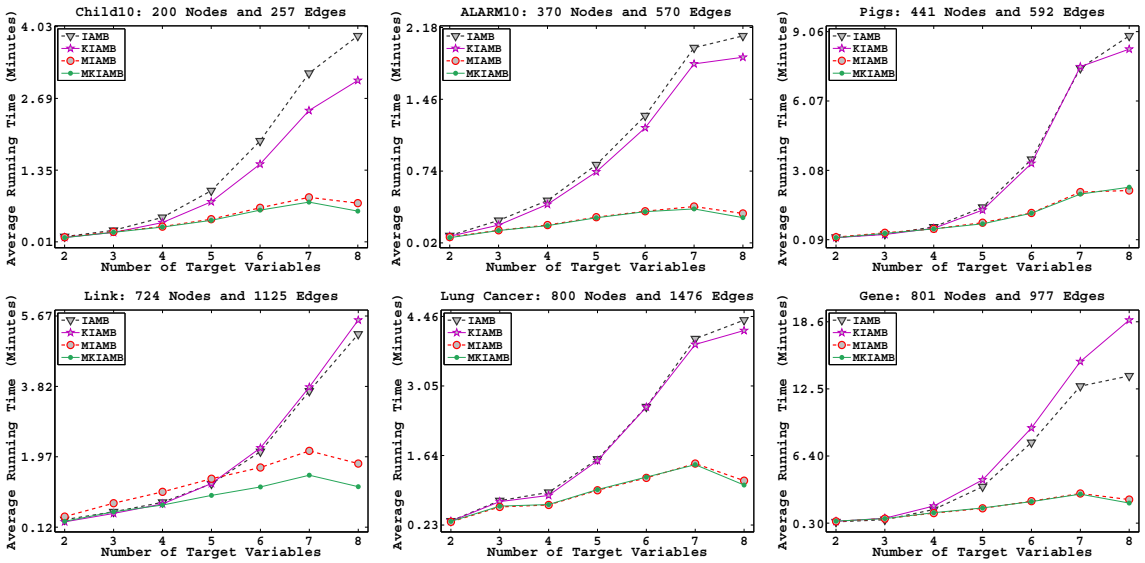

Figure 15: Average RT of the algorithms versus $|\boldsymbol{T}|$ with respect to the data sets of size 5000

the correctness of CI tests may also be more apt to be violated for a larger $|\boldsymbol{T}|$, due to the accumulation and propagation of the cascading errors (Bromberg and Margaritis, 2009). It is mentioned that (a)(b)(c) appear more evidently for the case of $n = 5000$ than for the case of $n = 500$.

(ii) *Performance on WP*: The similar interpretations to (a)(c) of (i) are valid.

(iii) *Performance on RT*: Here, we note that the real RT of an MB discovery algorithm is composed of two parts, in which the *part (I) is for CI tests*, and the *part (II) is for all other computations*. The part (I) is the major part used to measure the complexity of the MB discovery algorithm. Note also that the RT, $t_{n,q}$, of per CI test is linear to the sample size, n, and exponential to the number, q, of variables in the conditional set (see Subsection 4.4 for details).

This means that the part (II) of the real RT may dominate the part (I) if n is not large (for example, $n = 500$).

Let us now observe Figure 14 and Figure 15. First, both figures show the real RT that each algorithm needs is increasing along with the increase of $|\boldsymbol{T}|$. Also, Figure 14 indicates `MIAMB` and `MKIAMB` need slightly longer time to run than `IAMB` and `KIAMB`, because the running for CI tests is dominated by the running for all other computations in the case of a small sample size, while Figure 15 reveals that the real RTs of `IAMB` and `KIAMB` increase sharply as $|\boldsymbol{T}|$ increases and that the real RTs of `MIAMB` and `MKIAMB` increase slowly, just like the theoretical analyses about the complexities of the four algorithms show in Subsection 4.4.

In summary, the existing MB discovery algorithms, `IAMB` and `KIAMB`, can be approximately applied to the problem of MB discovery for multiple target variables when $|\boldsymbol{T}|$ is small, but they will perform poorly if $|\boldsymbol{T}|$ is moderately large. In comparison, our `MIAMB` and `MKIAMB` have higher accuracies and lower complexities for this problem, especially when $|\boldsymbol{T}|$ is large.

6. Application to FS in Multi-Class Prediction Problems

In this section, we apply the MB discovery for multiple targets to FS in multi-class prediction problems based on a real data set, HIVA. This data set is very challenging in WCCI 2006 (`http://www.modelselect.inf.ethz.ch`) and IJCNN 2007 (`http://www.agnostic.inf.ethz.ch`), because it contains many very unbalanced variables.

Let $T \in \boldsymbol{V}$ be a target variable taking values $\{1, \cdots, t\}$, $t \geqslant 3$. The multi-class prediction problem is to select features of T from $\boldsymbol{V} \setminus \{T\}$ such that T can be predicted as accurately as possible based on the chosen features. Let $\boldsymbol{T}^{(d)} \triangleq \left(T_1^{(d)}, \cdots, T_t^{(d)}\right)$ be the dummy version of T. Theoretically, $\boldsymbol{T}^{(d)}$ and T have the same MBs.

With these notations, the experiment is designed as follows:

- *Data*: HIVA contains 4229 data points and 1618 variables.
- *Targets*: In view of the fact that almost all variables in HIVA are binary, we randomly take k 2-class variables ($k = 2, \cdots, 5$) to create a merged 2^k-class target, T. Accordingly, we rearrange the original data to get a data set that is used only for FS of T. Repeat this step $n \triangleq 200$ times. Denote the resulting targets and their dummy versions by $T_1, \cdots, T_n$ and $\boldsymbol{T}_1^{(d)}, \cdots, \boldsymbol{T}_n^{(d)}$, respectively.
- *Algorithms*: We use the following 10 MB discovery algorithms to get the features for each T_j or $\boldsymbol{T}_j^{(d)} \triangleq \left(T_{j1}^{(d)}, \cdots, T_{jt_k}^{(d)}\right)$ with $t_k \triangleq 2^k$, $j = 1, \cdots, n$:
  a) `IAMB/KIAMB`-I: the `IAMB/KIAMB` algorithms working on T_j directly;
  b) `IAMB/KIAMB`-II: the `IAMB/KIAMB` algorithms working on $\boldsymbol{T}_j^{(d)}$ (that is, with $\boldsymbol{T}_j^{(d)}$ as its multiple targets);
  c) `MIAMB/MKIAMB`-I: the additivity based `MIAMB/MKIAMB` algorithms, taking the union of the outputs of `IAMB/KIAMB` with respect to $T_{ji}^{(d)}$ ($i = 1, \cdots, t_k$) as its output;
  d) `MIAMB/MKIAMB`-II: the MBS based `MIAMB/MKIAMB` algorithms, which are pseudo-coded in Algorithm 2;
  e) `MIAMB/MKIAMB`-III: the dummy `MIAMB/MKIAMB` algorithms, which take the union of the outputs of `IAMB/KIAMB` with respect to $T_{ji}^{(d)}$ ($i = 1, \cdots, t_k$) and removing redundant variables.
- *Classifier*: After making a number of preliminary experiments on the six benchmarking BNs, we found that the support vector machines (SVMs; implemented via `LibSVM` v3.22) perform the best in demonstrating the optimality of MBs for FS. This coincides with the assertion of Statnikov et al. (2013). Therefore, we use SVMs for our multi-class prediction problems. All the classifications are performed by 10-fold cross-validation.
- *Measurement* of an algorithm: For each target, the predictive quality of an MB is measured by the *balanced accuracy* defined as $\tau \triangleq \frac{1}{t_k} \sum_{\ell=1}^{t_k} \left(c_{\ell\ell} \big/ \sum_{i=1}^{t_k} c_{i\ell}\right)$, where $\boldsymbol{C} \triangleq \left(c_{i\ell}\right)$ denotes the associated confusion matrix. As seen, τ is equal to one minus the *balanced error rate* used in WCCI 2006 and IJCNN 2007. We choose to use τ (instead of *ordinary accuracy*) because it trades off all values of the target in the sense that any unbalanced value (that the target

Problem	IAMB		MIAMB		
	I	II	I	II	III
4-class	0.9295 ± 0.082	0.9020 ± 0.113	0.9414 ± 0.068	0.9366 ± 0.073	0.9507 ± 0.057
8-class	0.9016 ± 0.102	0.8666 ± 0.133	0.9237 ± 0.091	0.9277 ± 0.087	0.9348 ± 0.077
16-class	0.8878 ± 0.105	0.8461 ± 0.143	0.9131 ± 0.087	0.9167 ± 0.086	0.9256 ± 0.075
32-class	0.8683 ± 0.113	0.8179 ± 0.157	0.9118 ± 0.078	0.9139 ± 0.078	0.9242 ± 0.067

Table 2: Balanced accuracy of IAMB/MIAMB algorithms in the form of "(mean ± std)".

Problem	KIAMB		MKIAMB		
	I	II	I	II	III
4-class	0.9283 ± 0.085	0.8972 ± 0.124	0.9281 ± 0.092	0.9444 ± 0.066	0.9501 ± 0.058
8-class	0.9007 ± 0.105	0.8631 ± 0.144	0.9245 ± 0.089	0.9280 ± 0.082	0.9340 ± 0.078
16-class	0.8886 ± 0.106	0.8494 ± 0.142	0.9159 ± 0.084	0.9168 ± 0.086	0.9263 ± 0.075
32-class	0.8687 ± 0.114	0.8241 ± 0.151	0.9111 ± 0.081	0.9151 ± 0.076	0.9239 ± 0.068

Table 3: Balanced accuracy of KIAMB/MKIAMB algorithms in the form of "(mean ± std)".

takes) should not impact on the accuracy too much.[1] On the other hand, when two outputs of algorithms have the same total numbers of "true positives + true negatives", the balanced accuracy can identify the output that prefers to protect the scarce class as the better one, while the ordinary accuracy cannot. Finally, we compute the mean and standard deviation (std) of the n values of balanced accuracy, denoting them in the form of "(mean ± std)".

The experiment is then performed following the above procedures. Its results are summarized in Table 2 and Table 3. In these two tables, the backcolor indicates the performance of algorithms with black corresponding to the best while light blue to the worst. By the results, it can be seen that MIAMB/MKIAMB outperform IAMB/KIAMB in most situations. Specifically, we have:

- IAMB/KIAMB algorithms: IAMB/KIAMB-I are much more preferred than IAMB/KIAMB-II.
- MIAMB/MKIAMB algorithms: MKIAMB-I has almost equal performance to KIAMB-I in 4-class problems, and they performs slightly better than IAMB/KIAMB-I in 16- and 32-class problems;

1. For example, consider an unbalanced target T and its classification with the following two confusion matrices (the left is *extremely bad*, while the right is *very good*):

Test \ Truth	$T=1$	$T=2$
$T=1$	948	49
$T=2$	2	1

Test \ Truth	$T=1$	$T=2$
$T=1$	899	0
$T=2$	51	50

Then, we have: (a) for the left *bad* confusion matrix, the ordinary accuracy equals 94.90% (meaning it is impacted deeply by the unbalanced value 1 of T), while its balanced accuracy equals 50.89%; (b) for the right *good* confusion matrix, its ordinary accuracy also equals 94.90%, but its balanced accuracy equals 97.32%. This means balanced accuracy is more reasonable than ordinary accuracy to measure classification performance for a practical problem containing unbalanced variables (note that such problems may frequently occur in practice).

Null hypothesis (H_0)	Problem			
	4-class	8-class	16-class	32-class
MIAMB-I $\preccurlyeq$ IAMB-I	2.0349×10^{-4}	1.3225×10^{-7}	1.3816×10^{-10}	4.0634×10^{-19}
MIAMB-II $\preccurlyeq$ IAMB-I	1.4772×10^{-2}	9.3393×10^{-10}	4.0911×10^{-12}	8.2111×10^{-21}
MIAMB-III $\preccurlyeq$ IAMB-I	2.1276×10^{-10}	9.5876×10^{-16}	1.8800×10^{-20}	9.1061×10^{-30}
MKIAMB-I $\preccurlyeq$ KIAMB-I	0.6221	4.9365×10^{-9}	5.9117×10^{-12}	2.0651×10^{-19}
MKIAMB-I = KIAMB-I	0.7558	——	——	——
MKIAMB-II $\preccurlyeq$ KIAMB-I	2.2839×10^{-6}	5.9538×10^{-11}	5.4564×10^{-12}	2.7021×10^{-22}
MKIAMB-III $\preccurlyeq$ KIAMB-I	4.5857×10^{-10}	1.1261×10^{-15}	4.7802×10^{-20}	6.4206×10^{-30}

Table 4: p-values on paired t-test for comparison between MIAMB/MKIAMB and IAMB/KIAMB. Here, the notations are defined as follows: letting $\mathscr{A}_1$ and $\mathscr{A}_2$ be two algorithms and P be a problem, if $\mathscr{A}_1$ is better (in the sense of possessing higher accuracy) than $\mathscr{A}_2$ when used to solve P, we denote it by $\mathscr{A}_1 \succ \mathscr{A}_2$ (w.r.t. P); otherwise, we write it as $\mathscr{A}_1 \preccurlyeq \mathscr{A}_2$. In addition, we use $\mathscr{A}_1 = \mathscr{A}_2$ to denote $\mathscr{A}_1 \preccurlyeq \mathscr{A}_2$ and $\mathscr{A}_1 \succcurlyeq \mathscr{A}_2$.

MIAMB/MKIAMB-II significantly improve IAMB/KIAMB and even MIAMB/MKIAMB-I in most cases (although MIAMB/MKIAMB-II have larger std values than MIAMB/MKIAMB-I in some cases, the differences are slight). MIAMB/MKIAMB-III perform the best in all situations, with the highest mean values and the smallest std values.

Further, for any two algorithms, denote their balanced accuracy values as n (= 200) paired data points. Then, we can compute the p-values of *paired t-test* of associated hypotheses for one algorithm to be better (in the sense of possessing higher accuracy) than the other. The results are presented in Table 4. This table quantificationally shows the statistical significance of how much MIAMB/MKIAMB improve IAMB/KIAMB: in most cases, the improvement is more and more significant as the classification complex increases.

- The performance of each algorithm degrades with the increase of classification complexity. However, the degenerations of MIAMB/MKIAMB are slower than that of IAMB/KIAMB.

To compare IAMB/KIAMB and MIAMB/MKIAMB detailedly, we take the results of IAMB/KIAMB-I and MIAMB/MKIAMB-III to make a further analysis. For the 4-class prediction problem, denote the results of IAMB-I and MIAMB-III by $\tau_i^{(\mathtt{IAMB})}$ and $\tau_i^{(\mathtt{MIAMB})}$ for $i = 1, \cdots, n$, and draw them in (a) of Figure 16. Put

$$\begin{aligned}
\boldsymbol{I}_1 &= \left\{i \in \{1, \cdots, n\} : \tau_i^{(\mathtt{IAMB})} > \tau_i^{(\mathtt{MIAMB})}\right\}, \\
\boldsymbol{I}_2 &= \left\{i \in \{1, \cdots, n\} : \tau_i^{(\mathtt{IAMB})} = \tau_i^{(\mathtt{MIAMB})}\right\}, \\
\boldsymbol{I}_3 &= \left\{i \in \{1, \cdots, n\} : \tau_i^{(\mathtt{IAMB})} < \tau_i^{(\mathtt{MIAMB})}\right\}.
\end{aligned}$$

Draw the scatters of $\tau_i^{(\mathtt{IAMB})}$ and $\tau_i^{(\mathtt{MIAMB})}$ for $i \in \boldsymbol{I}_j$ in (a_j) of Figure 16. In addition, the information about (mean ± std) of IAMB-I vs that of MIAMB-III is annotated in each title. For other three K-class prediction problems ($K = 8,\ 16,\ 32$), repeat the above steps to get the scatters drawn in the other subplots of Figure 16. Similarly, Figure 17 draws the results of KIAMB-I versus MKIAMB-III.

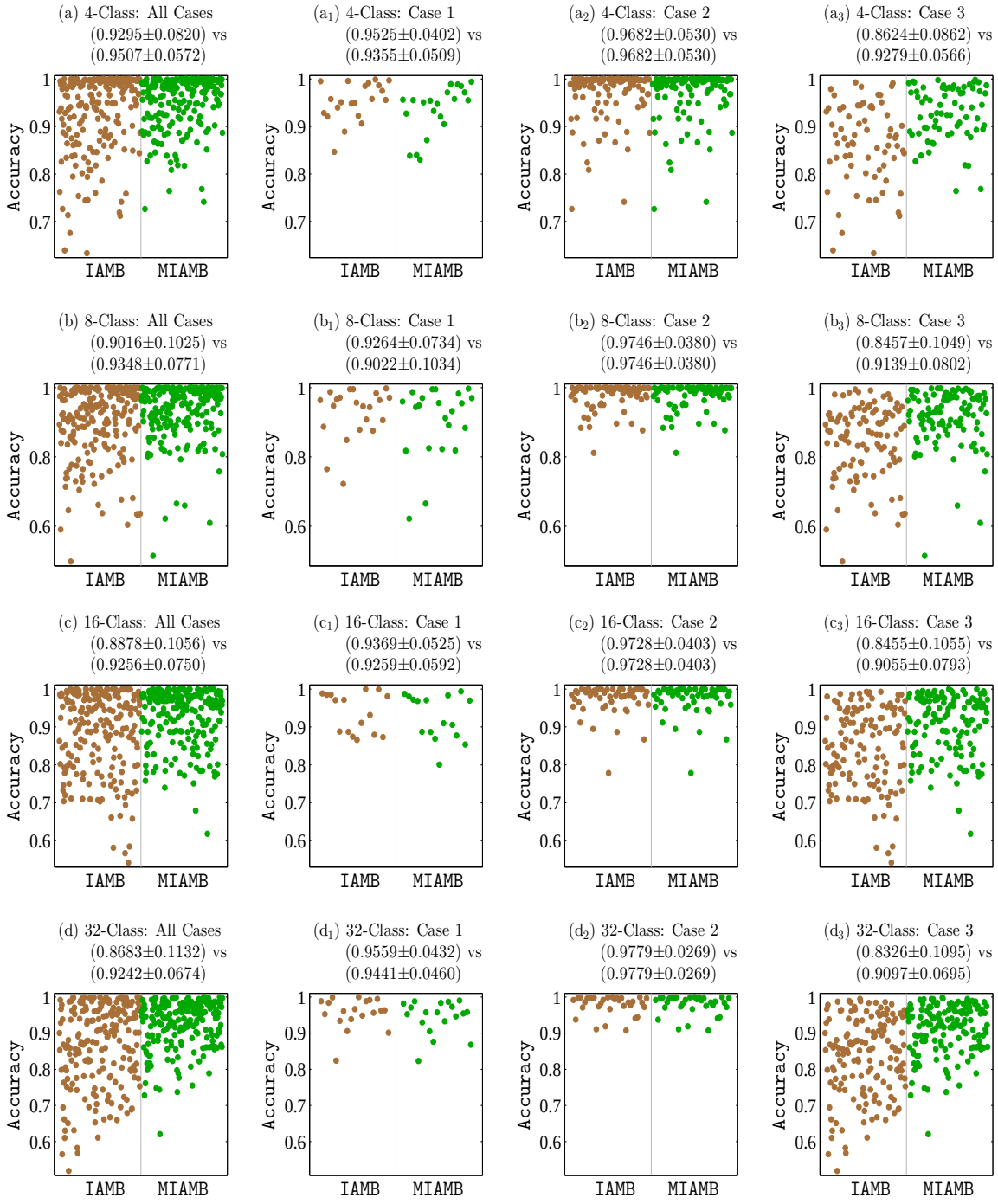

Figure 16: Balanced accuracy results on IAMB/MIAMB algorithms applied to 200 K-class prediction problems (K = 4, 8, 16, 32): the subplots in the first column for all the 200 results; the ones in the second column for the results that IAMB performs better than MIAMB; the ones in the third column for the results that IAMB and MIAMB perform equally well; the ones in the last column for the results that MIAMB performs better than IAMB.

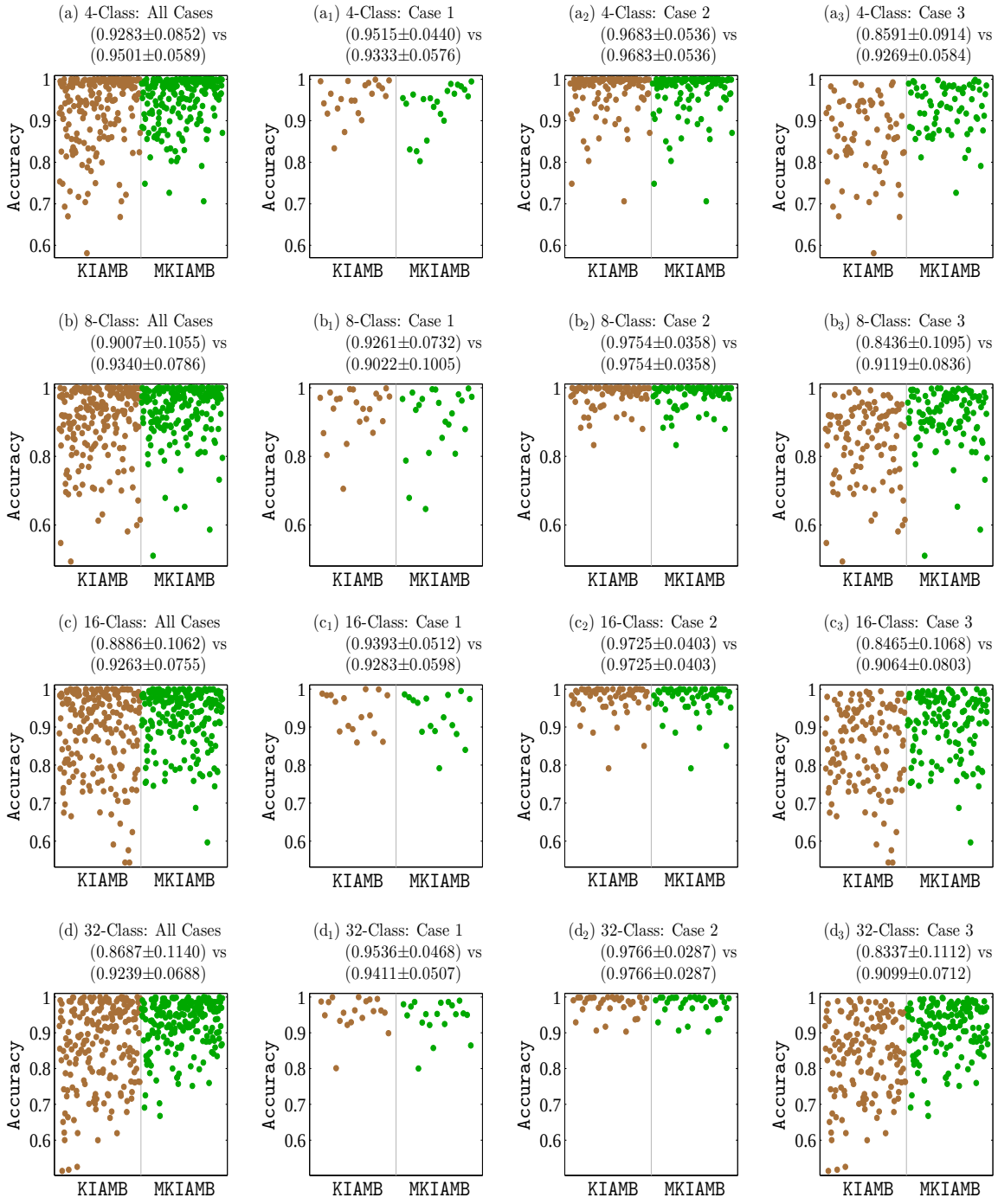

Figure 17: Balanced accuracy results on KIAMB/MKIAMB applied to 200 K-class prediction problems (K = 4, 8, 16, 32): the subplots in the first column for all the 200 results; the ones in the second column for the results that KIAMB performs better than MKIAMB; the ones in the third column for the results that KIAMB and MKIAMB perform equally well; the ones in the last column for the results that MKIAMB performs better than KIAMB.

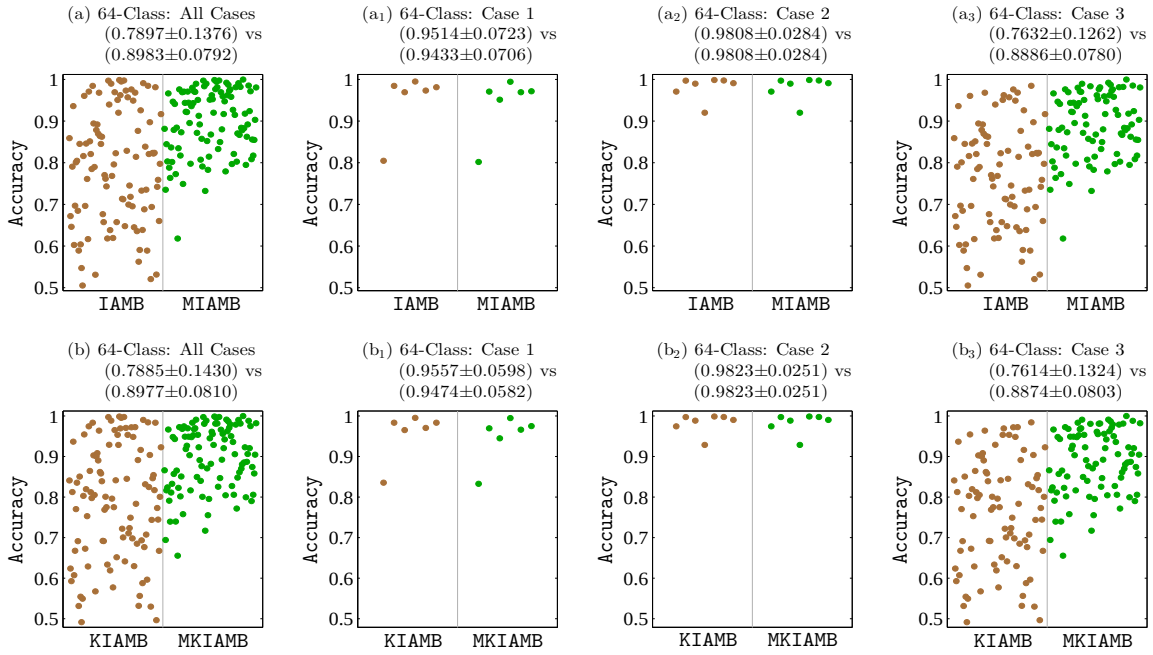

Figure 18: Balanced accuracy results on IAMB/MIAMB and KIAMB/MKIAMB applied to 100 64-class prediction problems: MIAMB/MKIAMB can improve IAMB/KIAMB substantially.

Figure 16 indicates that: (i) In most cases, MIAMB can improve IAMB to various degrees: MIAMB has higher mean values of balanced accuracy and smaller std values as well. Although there are a few of situations in which IAMB performs better than MIAMB, the difference of performance between them is very slight. In addition, there are some situations in which the two algorithms perform equally well (with very high mean and small std). (ii) MIAMB is more resistant to the classification complexity than IAMB: the improvements of MIAMB over IAMB become more and more visible with the increase of K (from 4 to 32). Figure 17 shows similar conclusions.

In brief, an FS problem for multi-class prediction can be transformed into a problem of MB discovery for multiple targets, and then get a more efficient solution. This idea may be particularly useful when the classification complexity is high or very high. To check this imagination, we apply the same procedures to 100 64-class prediction problems (also taken from the HIVA data set). The results are summarized in Figure 18, in which (a) and (a_j) are for IAMB/MIAMB while (b) and (b_j) are for KIAMB/MKIAMB, $j = 1, 2, 3$. By the figure, the improvement (nearly 14% on accuracy) of MIAMB/MKIAMB algorithms over IAMB/KIAMB is really desirable in the case of high classification complexity.

Finally, we apply LibSVM and the random forest (RF) algorithm (Breiman, 2001) to the whole HIVA data without any FS, considering LibSVM is of high classification performance while RF is a state-of-the-art FS algorithm. The results can be served as a baseline to see why FS (or equivalently, MB discovery) is necessary for a complex classification problem. Recall that HIVA contains many unbalanced variables, which enhance the classification complexity. Figure 19 draws the 95% confidence bands of LibSVM and FS, respectively.

By the figure, it follows that:

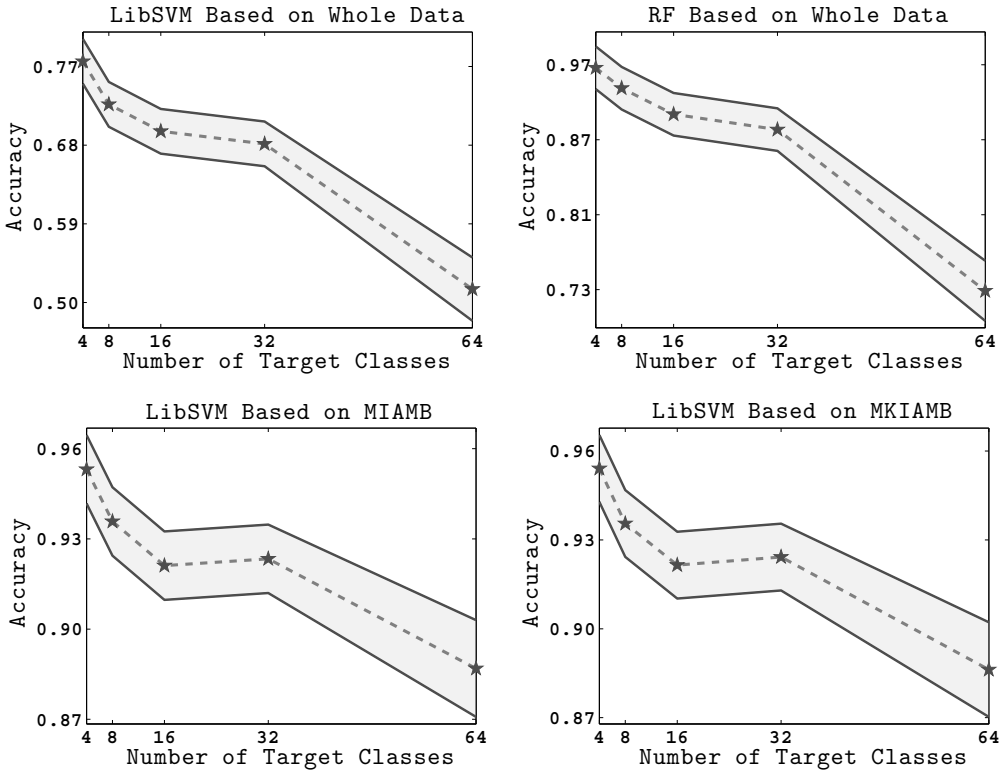

Figure 19: The 95% confidence bands of LibSVM and RF.

- Without any FS, LibSVM performs undesirably in all situations. This may be because too many noisy variables can lead to masking effects upon those unbalanced features such that LibSVM cannot classify targets expectedly. This shows the necessity of FS. In other words, LibSVM may not be suitable for some *high-dimension* problems, especially when there are many unbalanced variables.

- Without any FS, RF performs quite well when the classification complexity is not very high. However, with the increase of classification complexity, the performance of RF decreases gradually and then sharply. In other words, RF may not be suitable for those problems with too *high complexity*, especially when there are many unbalanced variables.

To observe why this happens, we check the results and then randomly take some targets (with extraordinarily low accuracy) to implement LibSVM and RF again by appropriately adjusting the algorithmic setting of LibSVM and increasing the number of trees of RF from 100 to 1000. However, the results change very little.

To make an intuitive comparison, the 95% confidence bands of MIAMB/MKIAMB-based LibSVM are also drawn in Figure 19. As seen, all methods degenerate with the increase of the classification complexity, but our methods degenerate far slower than LibSVM/RF based on the whole data. In a word, MB discovery (or equivalently, FS) is important to make classification, especially when the problem is of *high dimension* and of *high complexity*.

7. Concluding Remarks

In this paper, we considered the problem of Mb and MB of multiple variables. We first addressed their additivity under the local intersection assumption, and then studied this problem in the general case. The two algorithms that we proposed , `MIAMB` and `MKIAMB`, were proven to be correct under the local composition assumption with respect to single targets. The benchmarking study based on six synthetic BNs showed that `MIAMB` and `MKIAMB` have higher accuracies and lower complexities than the existing `IAMB` and `KIAMB`.

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

(i) The first remark concerns a method of using `MIAMB` and `MKIAMB` to find an MB for a single variable. Such an idea is motivated by the following two aspects: (a) the local composition assumption may be violated in practice, and if this is the case, `IAMB` and `KIAMB` may perform not very well in MB discovery even for a single variable; (b) randomness of a data set may result in a violation to the assumption that all the CI tests involved are correct. Naturally, it is useful to take a remedy for these two situations. One remedial strategy is described as follows: letting $T \in \boldsymbol{V}$ be the target variable, and $\boldsymbol{M}_1$ be a potential MB discovered by `IAMB` or `KIAMB`, take $T_0 \triangleq \arg\max_{X \in \boldsymbol{M}_1} f_{\boldsymbol{D}}^{(\ell)}(T; X \mid \boldsymbol{M}_1 \setminus \{X\})$ as a co-target of T for $\ell = 1$ or 2 or 3; then, employ `MIAMB` or `MKIAMB` to find a potential MB for $\{T, T_0\}$, saying $\boldsymbol{M}_2$. Finally, refine $\{T_0\} \cup \boldsymbol{M}_2$ to obtain $\boldsymbol{M}$, by virtue of the shrinking phase of `IAMB` or `KIAMB`, since this phase needs no the local composition precondition.

(ii) Our `MIAMB` and `MKIAMB` contain an ordering τ, which may affect the RT and even the WA or WP. A question arises here: is there an optimal selection of τ such that `MIAMB` or `MKIAMB` has the highest accuracy and the lowest complexity?

(iii) All the considered algorithms (`IAMB`, `KIAMB`, `MIAMB`, and `MKIAMB`) need the local composition assumption to theoretically guarantee their correctness. However, this precondition may be violated in practice and in this case only an approximate MB can be obtained by means of one of the above algorithms. Subsection 4.1 provides a potential remedy. We note that `MIAMB` and `MKIAMB` transform the problem of MB discovery for multiple targets into the ones for single targets. This idea provides a facilitation to use some stochastic optimization methods such as the particle swarm optimization algorithm (Kennedy and Eberhart, 1995, 1997).

(iv) In Subsection 3.3, we provided a method for MB discovery of a *complex* single variable based on an MB discovery of some *simple* multiple variables. Let us now explain why this transformation method is efficient. With the notations used in Subsection 3.3, let

$$\mathrm{MB}_T = \mathrm{MB}_{\boldsymbol{T}^{(d)}} \triangleq \boldsymbol{M}.$$

Then, a variable X can enter and stay in $\boldsymbol{M}$ (in the sense of MB_T) if $T \not\perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\}$. On the other hand, by Theorem 5, X can enter and stay in $\boldsymbol{M}$ (in the sense of $\mathrm{MB}_{\boldsymbol{T}^{(d)}}$) only if $T_j^{(d)} \not\perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\}$ holds for some j. That is, we need to test the following two pairs of hypotheses:

$$\begin{aligned} H_0^{(1)}: \quad & T \perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\} \quad \leftrightarrow \quad H_1^{(1)}: \quad T \not\perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\}; \\ H_0^{(2)}: \quad & T_j^{(d)} \perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\} \quad \leftrightarrow \quad H_1^{(2)}: \quad T_j^{(d)} \not\perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\}. \end{aligned}$$

Clearly, when T is high-dimensional, *the test for* $H_0^{(1)} \leftrightarrow H_1^{(1)}$ requires far more data points than *that for* $H_0^{(2)} \leftrightarrow H_1^{(2)}$, since the test statistic for the first pair of hypotheses contains far more free parameters than that for the second. In addition, the transformation from T to $\boldsymbol{T}^{(d)}$ can be easily made, with almost no running time. This explains why Theorem 5 is useful.

Acknowledgments

The authors are very grateful to the four anonymous reviewers and Prof. Marina Meila and Prof. Joris Mooij for their valuable comments and constructive suggestions which result in the present version. Thanks also to Prof. Kevin Murphy for all of his kind help.

This work was supported by the National Natural Science Foundation of China (61374183, 51472117, 51535005, 51675212), the Research Fund of State Key Laboratory of Mechanics and Control of Mechanical Structures (MCMS-0417G02, MCMS-0417G03), the Fundamental Research Funds for the Central Universities (NP2017101, NC2018001), the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions, and the Open Fund for the Key Laboratory for Traffic and Transportation Security of Jiangsu Province.

Appendix A. Pseudo Codes for `IAMB` and `KIAMB`

This appendix presents the pseudo codes for `IAMB` and `KIAMB`. We mention here that, in Line 4 of `KIAMB`, $\boldsymbol{M}_2$ denotes a random subset of $\boldsymbol{M}_1$ with size $|\boldsymbol{M}_2| = \max\{1, \lfloor |\boldsymbol{M}_1| \cdot K \rfloor\}$.

Algorithm 3: `IAMB` and `KIAMB`

Procedure: $\boldsymbol{M} \leftarrow$ `IAMB`$(\boldsymbol{D}; \boldsymbol{T})$
Input: $\boldsymbol{D}$ is a data matrix; $\boldsymbol{T}$ is a set of target variables.
Output: an MB, $\boldsymbol{M}$, of $\boldsymbol{T}$.

```
//Forward: Growing Phase
1  M ← ∅
2  while M has changed do
3      Y ← arg max_{X∈V\M\T} f_D(T; X | M)
4      if T ⫫̸ Y | M then
5          M ← M ∪ {Y}
6      end
7  end
//Backward: Shrinking Phase
8  foreach X ∈ M do
9      if T ⫫ X | M \ {X} then
10         M ← M \ {X}
11     end
12 end
13 return M
```

Procedure: $\boldsymbol{M} \leftarrow$ `KIAMB`$(\boldsymbol{D}; \boldsymbol{T}; K)$
Input: Besides $\{\boldsymbol{D},\boldsymbol{T}\}$ as in `IAMB`, $K \in [0, 1]$ is a randomization parameter.
Output: an Mb, $\boldsymbol{M}$, of $\boldsymbol{T}$.

```
//Forward: Growing Phase
1  M ← ∅
2  while M has changed do
3      if M_1 ← {X ∈ V \ M \ T : T ⫫̸ X | M} ≠ ∅ then
4          Y ← arg max_{X∈M_2} f_D(T; X | M)
5          M ← M ∪ {Y}
6      end
7  end
//Backward: Shrinking Phase
8  foreach X ∈ M do
9      if T ⫫ X | M \ {X} then
10         M ← M \ {X}
11     end
12 end
13 return M
```

Appendix B. Proofs

In this appendix, we give the proofs of the theoretical results.

Lemma 1 *The intersection property holds if and only if no information equivalence occurs.*

Proof Equivalently, we show that the intersection property is violated if and only if information equivalence occurs. The sufficiency holds clearly. To prove the necessity, we assume the intersection property is violated, that is, there are $\boldsymbol{T}$, $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ such that $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{Z} \cup \boldsymbol{Y}$, and $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z} \cup \boldsymbol{X}$, but $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{X} \cup \boldsymbol{Y} \,|\, \boldsymbol{Z}$. Then, we can show $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{Z}$. In fact, if $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{Z}$, then combined with $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z} \cup \boldsymbol{X}$ and the contraction property, it concludes $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{X} \cup \boldsymbol{Y} \,|\, \boldsymbol{Z}$, which contradicts $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{X} \cup \boldsymbol{Y} \,|\, \boldsymbol{Z}$. Similarly, we can show $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$. Therefore, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are information equivalent with respect to $\boldsymbol{T}$ given $\boldsymbol{Z}$. That is, information equivalence occurs. ■

Lemma 2 *For $\boldsymbol{T} \subseteq \boldsymbol{V}$, assume the type-II local condition holds. Then $\boldsymbol{T}$ has a unique MB.*

Proof Suppose $\boldsymbol{T}$ has two different MBs, $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$. Putting $\boldsymbol{M}_{12} \triangleq \boldsymbol{M}_1 \setminus \boldsymbol{M}_2$, $\boldsymbol{M}_{21} \triangleq \boldsymbol{M}_2 \setminus \boldsymbol{M}_1$, and $\boldsymbol{M} \triangleq \boldsymbol{M}_1 \cap \boldsymbol{M}_2 \subsetneqq \boldsymbol{M}_i$ for $i = 1, 2$, we have

$$\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T} \,|\, \boldsymbol{M}_1 \;\Rightarrow\; \boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T} \,|\, \boldsymbol{M} \cup \boldsymbol{M}_{12}, \tag{8}$$

$$\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M}_2 \setminus \boldsymbol{T} \,|\, \boldsymbol{M}_2 \;\Rightarrow\; \boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M}_2 \setminus \boldsymbol{T} \,|\, \boldsymbol{M} \cup \boldsymbol{M}_{21}. \tag{9}$$

Now we show $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{M}_{12} \,|\, \boldsymbol{M}$. In fact, suppose we have $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{M}_{12} \,|\, \boldsymbol{M}$. This combined with (8) implies $\boldsymbol{T} \perp\!\!\!\perp (\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}) \cup \boldsymbol{M}_{12} \,|\, \boldsymbol{M}$, in view of the contraction property. Equivalently, $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M} \setminus \boldsymbol{T} \,|\, \boldsymbol{M}$, which contradicts the fact that $\boldsymbol{M}_1$ is an MB of $\boldsymbol{T}$, since $\boldsymbol{M} \subsetneqq \boldsymbol{M}_1$. Hence, $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{M}_{12} \,|\, \boldsymbol{M}$. Similarly, we can show $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{M}_{21} \,|\, \boldsymbol{M}$. On the other hand, by the decomposition property, (9) and (8) indicate $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{M}_{12} \,|\, \boldsymbol{M} \cup \boldsymbol{M}_{21}$ and $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{M}_{21} \,|\, \boldsymbol{M} \cup \boldsymbol{M}_{12}$, respectively. Therefore, $\boldsymbol{M}_{12}$ and $\boldsymbol{M}_{21}$ are information equivalent with respect to $\boldsymbol{T}$ conditioned on $\boldsymbol{M}$. This contradicts the precondition. The uniqueness of MB of $\boldsymbol{T}$ is shown under the type-II local condition. ■

Theorem 2 (Additivity of Mb) *Let $(\mathbb{G}, \mathbb{P})$ be a BN over $\boldsymbol{V}$. The following two statements hold:*

(i) *Let $\boldsymbol{M}_i$ be an Mb of $\boldsymbol{T}_i \subseteq \boldsymbol{V}$ for $i = 1, 2$, and assume $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$ satisfies the local intersection assumption. Then, $(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus (\boldsymbol{T}_1 \cup \boldsymbol{T}_2)$ is an Mb of $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$.*

(ii) *Let $\boldsymbol{M}_i$ be an Mb of $T_i \in \boldsymbol{V}$ for $i = 1, \cdots, k$, and assume $\boldsymbol{T} \triangleq \{T_1, \cdots, T_k\}$ satisfies the local intersection assumption. Then, $\bigcup_{i=1}^{k} \boldsymbol{M}_i \setminus \boldsymbol{T}$ is an Mb of $\boldsymbol{T}$.*

Proof It suffices to prove (i), since (ii) is a direct consequence of (i) using induction on the number of variables involved in $\boldsymbol{T}$.

Denote $\boldsymbol{V}_0 = \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$ with $\boldsymbol{N} = \boldsymbol{N}_1 \cup \boldsymbol{N}_2$, in which $\boldsymbol{N}_1 = \boldsymbol{M}_1 \setminus \boldsymbol{T}_2$ and $\boldsymbol{N}_2 = \boldsymbol{M}_2 \setminus \boldsymbol{T}_1$. Note that $\boldsymbol{N}$ can also be expressed as $\boldsymbol{N} = (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}$. First, we prove the following CI relationship, by means of the graphoid properties:

$$\boldsymbol{V}_0 \perp\!\!\!\perp \boldsymbol{T}_1 \,|\, \boldsymbol{N} \cup \boldsymbol{T}_2. \tag{10}$$

In fact, with the above notations, it is readily justified that $\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1 = \boldsymbol{V}_0 \cup [(\boldsymbol{N}_2 \cup \boldsymbol{T}_2) \setminus \boldsymbol{M}_1]$. On the other hand, $\boldsymbol{M}_1$ is an Mb of $\boldsymbol{T}_1$. Therefore, $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1 \,|\, \boldsymbol{M}_1$, and thus we obtain $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V}_0 \cup [(\boldsymbol{N}_2 \cup \boldsymbol{T}_2) \setminus \boldsymbol{M}_1] \,|\, \boldsymbol{M}_1$. By the weak union property, $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V}_0 \,|\, \boldsymbol{M}_1 \cup [(\boldsymbol{N}_2 \cup \boldsymbol{T}_2) \setminus \boldsymbol{M}_1]$. This means (10) holds, since $\boldsymbol{M}_1 \cup [(\boldsymbol{N}_2 \cup \boldsymbol{T}_2) \setminus \boldsymbol{M}_1] = \boldsymbol{N} \cup \boldsymbol{T}_2$.

Similarly, $\boldsymbol{V}_0 \perp\!\!\!\perp \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{T}_1$, which combined with (10) indicates

$$\boldsymbol{V}_0 \perp\!\!\!\perp \boldsymbol{T} \,|\, \boldsymbol{N}$$

by the local intersection assumption. Or equivalently, $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T} \,|\, \boldsymbol{N}$. That is, $(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T} = \boldsymbol{N}$ is an Mb of $\boldsymbol{T}$. The proof is completed. ■

Remark 1 *In the case of either $\boldsymbol{T}_1 \subseteq \boldsymbol{V} \setminus \boldsymbol{M}_2$ or $\boldsymbol{T}_2 \subseteq \boldsymbol{V} \setminus \boldsymbol{M}_1$, the conclusion of* (i) *in Theorem 2 holds without requiring the local intersection assumption.*

Proof If $\boldsymbol{T}_1 \subseteq \boldsymbol{V} \setminus \boldsymbol{M}_2$ but $\boldsymbol{T}_2 \nsubseteq \boldsymbol{V} \setminus \boldsymbol{M}_1$, $\boldsymbol{M}_1 \cup \boldsymbol{M}_2$ is then an Mb of $\boldsymbol{T}_1$ according to the weak union property whereas $(\boldsymbol{M}_1 \setminus \boldsymbol{T}_2) \cup \boldsymbol{M}_2$ is an Mb of $\boldsymbol{T}_2$. Equivalently, we have

$$\begin{aligned}\boldsymbol{T}_1 &\perp\!\!\!\perp \boldsymbol{V} \setminus (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}_1 \,|\, \boldsymbol{M}_1 \cup \boldsymbol{M}_2,\\ \boldsymbol{T}_2 &\perp\!\!\!\perp \boldsymbol{V} \setminus [(\boldsymbol{M}_1 \setminus \boldsymbol{T}_2) \cup \boldsymbol{M}_2] \setminus \boldsymbol{T}_2 \,|\, (\boldsymbol{M}_1 \setminus \boldsymbol{T}_2) \cup \boldsymbol{M}_2.\end{aligned}$$

By means of the contraction property and the decomposition property, it is seen that

$$\begin{aligned}\boldsymbol{V} \setminus [(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}_2] \setminus \boldsymbol{T} &\perp\!\!\!\perp \boldsymbol{T}_1 \,|\, [(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}_2] \cup \boldsymbol{T}_2,\\ \boldsymbol{V} \setminus [(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}_2] \setminus \boldsymbol{T} &\perp\!\!\!\perp \boldsymbol{T}_2 \,|\, (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}_2,\end{aligned}$$

so $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus [(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}_2] \setminus \boldsymbol{T} \,|\, (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}_2$. That is, $(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus (\boldsymbol{T}_1 \cup \boldsymbol{T}_2) = (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}_2$ is an Mb of $\boldsymbol{T}$. If $\boldsymbol{T}_1 \nsubseteq \boldsymbol{V} \setminus \boldsymbol{M}_2$ but $\boldsymbol{T}_2 \subseteq \boldsymbol{V} \setminus \boldsymbol{M}_1$, we can similarly show

$$(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus (\boldsymbol{T}_1 \cup \boldsymbol{T}_2) = (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}_1$$

is an Mb of $\boldsymbol{T}$. Finally, if $\boldsymbol{T}_1 \subseteq \boldsymbol{V} \setminus \boldsymbol{M}_2$ and $\boldsymbol{T}_2 \subseteq \boldsymbol{V} \setminus \boldsymbol{M}_1$, imposing decomposition on

$$\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}_1 \,|\, \boldsymbol{M}_1 \cup \boldsymbol{M}_2$$

and weak union on $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{V} \setminus (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}_2 \,|\, \boldsymbol{M}_1 \cup \boldsymbol{M}_2$, we get

$$\begin{aligned}\boldsymbol{V} \setminus (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T} &\perp\!\!\!\perp \boldsymbol{T}_1 \,|\, \boldsymbol{M}_1 \cup \boldsymbol{M}_2,\\ \boldsymbol{V} \setminus (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T} &\perp\!\!\!\perp \boldsymbol{T}_2 \,|\, (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \cup \boldsymbol{T}_1.\end{aligned}$$

By the contraction property, $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T} \,|\, \boldsymbol{M}_1 \cup \boldsymbol{M}_2$. That is,

$$(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus (\boldsymbol{T}_1 \cup \boldsymbol{T}_2) = \boldsymbol{M}_1 \cup \boldsymbol{M}_2$$

is an Mb of $\boldsymbol{T}$. The conclusion is proved. ■

Theorem 3 (Additivity of MB) *Let $(\mathbb{G}, \mathbb{P})$ be a BN over $\boldsymbol{V}$. The following two statements hold:*

(i) *Assume $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$ satisfies the local intersection assumption. Let $\boldsymbol{M}_i$ be the unique MB of $\boldsymbol{T}_i$ for $i = 1, 2$. Then, $(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus (\boldsymbol{T}_1 \cup \boldsymbol{T}_2)$ is the unique MB of $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$.*

(ii) *Assume $\boldsymbol{T} \triangleq \{T_1, \cdots, T_k\}$ satisfies the local intersection assumption. Let $\boldsymbol{M}_i$ be the unique MB of T_i for $i = 1, \cdots, k$. Then, $\bigcup_{i=1}^{k} \boldsymbol{M}_i \setminus \boldsymbol{T}$ is the uniqe MB of $\boldsymbol{T}$.*

Proof We need only to prove (i), since (ii) is a direct consequence of (i).

Denote $\boldsymbol{N}_1 = \boldsymbol{M}_1 \setminus \boldsymbol{T}_2$ and $\boldsymbol{N}_2 = \boldsymbol{M}_2 \setminus \boldsymbol{T}_1$. By Theorem 2, $(\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T} = \boldsymbol{N}_1 \cup \boldsymbol{N}_2$ is an Mb of $\boldsymbol{T}$. Therefore, it suffices to prove the minimality of $\boldsymbol{N}_1 \cup \boldsymbol{N}_2$, based on Lemma 2. In fact, let $\boldsymbol{N}_0$ be any Mb of $\boldsymbol{T}$ which is a subset of $\boldsymbol{N}_1 \cup \boldsymbol{N}_2$. Note that $\boldsymbol{T}_i \cap \boldsymbol{N}_0 = \emptyset$ for $i = 1, 2$. Denote now $\boldsymbol{M} = \boldsymbol{N}_0 \cup (\boldsymbol{M}_1 \setminus \boldsymbol{N}_1) = \boldsymbol{N}_0 \cup (\boldsymbol{M}_1 \cap \boldsymbol{T}_2)$. It follows that

- $\boldsymbol{M}_1$ is the MB of $\boldsymbol{T}_1$: This implies $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1 \,|\, \boldsymbol{M}_1$, or equivalently, we have

$$\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1 \,|\, (\boldsymbol{M}_1 \cap \boldsymbol{M}) \cup (\boldsymbol{M}_1 \setminus \boldsymbol{M}), \tag{11}$$

  in view of $\boldsymbol{M}_1 = (\boldsymbol{M}_1 \cap \boldsymbol{M}) \cup (\boldsymbol{M}_1 \setminus \boldsymbol{M})$.

- $\boldsymbol{N}_0$ is an Mb of $\boldsymbol{T}$: Equivalently, we have $\boldsymbol{T}_1 \cup \boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{N}_0 \setminus \boldsymbol{T}_2 \setminus \boldsymbol{T}_1 \,|\, \boldsymbol{N}_0$, which gives

$$\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus (\boldsymbol{N}_0 \cup \boldsymbol{T}_2) \setminus \boldsymbol{T}_1 \,|\, \boldsymbol{N}_0 \cup \boldsymbol{T}_2,$$

  according to the weak union property, and thus $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus (\boldsymbol{M} \cup \boldsymbol{T}_2) \setminus \boldsymbol{T}_1 \,|\, \boldsymbol{M} \cup \boldsymbol{T}_2$ in view of $\boldsymbol{N}_0 \cup \boldsymbol{T}_2 = \boldsymbol{M} \cup \boldsymbol{T}_2$, or equivalently we have $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M} \setminus \boldsymbol{T}_1 \setminus \boldsymbol{T}_2 \,|\, \boldsymbol{M} \cup \boldsymbol{T}_2$. By the self-conditioning property, this leads to $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus (\boldsymbol{M}_1 \cap \boldsymbol{M}) \setminus \boldsymbol{T}_1 \,|\, \boldsymbol{M} \cup \boldsymbol{T}_2$, Therefore,

$$\boldsymbol{T}_1 \perp\!\!\!\perp (\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1) \cup (\boldsymbol{M}_1 \setminus \boldsymbol{M}) \,|\, \boldsymbol{M} \cup \boldsymbol{T}_2,$$

  in terms of $\boldsymbol{V} \setminus (\boldsymbol{M}_1 \cap \boldsymbol{M}) \setminus \boldsymbol{T}_1 = (\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1) \cup (\boldsymbol{M}_1 \setminus \boldsymbol{M})$. By the weak union property, this indicates $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{M}_1 \setminus \boldsymbol{M} \,|\, (\boldsymbol{M} \cup \boldsymbol{T}_2) \cup (\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1)$. Consequently,

$$\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{M}_1 \setminus \boldsymbol{M} \,|\, (\boldsymbol{M}_1 \cap \boldsymbol{M}) \cup (\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1), \tag{12}$$

  due to $(\boldsymbol{M} \cup \boldsymbol{T}_2) \cup (\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1) = (\boldsymbol{M}_1 \cap \boldsymbol{M}) \cup (\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1)$.

By the local intersection property, (11)(12) indicate $\boldsymbol{T}_1 \perp\!\!\!\perp (\boldsymbol{M}_1 \setminus \boldsymbol{M}) \cup (\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1) \,|\, \boldsymbol{M}_1 \cap \boldsymbol{M}$, so

$$\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus (\boldsymbol{M}_1 \cap \boldsymbol{M}) \setminus \boldsymbol{T}_1 \,|\, \boldsymbol{M}_1 \cap \boldsymbol{M},$$

since $(\boldsymbol{M}_1 \setminus \boldsymbol{M}) \cup (\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1) = \boldsymbol{V} \setminus (\boldsymbol{M}_1 \cap \boldsymbol{M}) \setminus \boldsymbol{T}_1$. Hence, $\boldsymbol{M}_1 \cap \boldsymbol{M}$ ($\subseteq \boldsymbol{M}_1$) is an Mb of $\boldsymbol{T}_1$. On the other hand, $\boldsymbol{M}_1$ is the MB of $\boldsymbol{T}_1$ and thereby $\boldsymbol{M}_1 \cap \boldsymbol{M} = \boldsymbol{M}_1$, or equivalently,

$$\boldsymbol{N}_1 \cup (\boldsymbol{M}_1 \cap \boldsymbol{T}_2) = \boldsymbol{M}_1 \subseteq \boldsymbol{M} = \boldsymbol{N}_0 \cup (\boldsymbol{M}_1 \cap \boldsymbol{T}_2),$$

which means $\boldsymbol{N}_1 \subseteq \boldsymbol{N}_0$. In a similar fashion, $\boldsymbol{N}_2 \subseteq \boldsymbol{N}_0$. Combined with $\boldsymbol{N}_0 \subseteq \boldsymbol{N}_1 \cup \boldsymbol{N}_2$, the expected relationship $\boldsymbol{N}_0 = \boldsymbol{N}_1 \cup \boldsymbol{N}_2$ follows. This indicates that $\boldsymbol{N}_1 \cup \boldsymbol{N}_2$ is an MB of $\boldsymbol{T}$. The proof is completed, since Lemma 2 shows the uniqueness of MB under the local intersection assumption. ■

Theorem 4 *Assume $\boldsymbol{S} \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$. Then, the following statements are equivalent:*

(i) *$\boldsymbol{S}$ is an MbS;*

(ii) $\mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}) = \min_{\boldsymbol{S}' \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}} \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}')$*;*

(iii) $\mathbb{I}(\boldsymbol{T}; \boldsymbol{S} \,|\, \boldsymbol{N}) = \max_{\boldsymbol{S}' \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}} \mathbb{I}(\boldsymbol{T}; \boldsymbol{S}' \,|\, \boldsymbol{N})$*;*

(iv) *$\boldsymbol{N} \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}_1$ in $\boldsymbol{V} \setminus \boldsymbol{T}_2$ (or $\boldsymbol{N} \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}_2$ in $\boldsymbol{V} \setminus \boldsymbol{T}_1$).*

In addition, if $\boldsymbol{S}$ is an MbS, then it is also an MBS if and only if $\boldsymbol{T}_1 \not\perp\!\!\!\perp Y \,|\, \boldsymbol{N} \cup (\boldsymbol{S} \setminus \{Y\})$ or $\boldsymbol{T}_2 \not\perp\!\!\!\perp Y \,|\, \boldsymbol{N} \cup (\boldsymbol{S} \setminus \{Y\})$ holds for any $Y \in \boldsymbol{S}$.

Proof We first prove (i) $\Leftrightarrow$ (ii). Put $\boldsymbol{Q} = \boldsymbol{V} \setminus (\boldsymbol{N} \cup \boldsymbol{S}) \setminus \boldsymbol{T}$. First, note that $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1 \,|\, \boldsymbol{M}_1$ since $\boldsymbol{M}_1$ is an Mb of $\boldsymbol{T}_1$. In other words, $a \triangleq \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{Q} \cup (\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1 \setminus \boldsymbol{Q}) \,|\, \boldsymbol{M}_1) = 0$. By the chain rule for CMI (Cover and Thomas, 2006), we have $\mathbb{I}(\boldsymbol{T}_1; \boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1 \setminus \boldsymbol{Q} \,|\, \boldsymbol{M}_1) = 0$. It follows that

$$\begin{aligned} a &= \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1 \setminus \boldsymbol{Q} \,|\, \boldsymbol{M}_1) + \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{Q} \,|\, (\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1 \setminus \boldsymbol{Q}) \cup \boldsymbol{M}_1) \\ &= \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{Q} \,|\, (\boldsymbol{V} \setminus \boldsymbol{M}_1 \setminus \boldsymbol{T}_1 \setminus \boldsymbol{Q}) \cup \boldsymbol{M}_1) \\ &= \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{Q} \,|\, \boldsymbol{N} \cup \boldsymbol{S} \cup \boldsymbol{T}_2) \\ &= \mathbb{I}(\boldsymbol{T}; \boldsymbol{Q} \,|\, \boldsymbol{N} \cup \boldsymbol{S}) - \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{Q} \,|\, \boldsymbol{N} \cup \boldsymbol{S}) \\ &\triangleq b - c, \end{aligned} \tag{13}$$

which combined with $a = 0$ gives $b = c$. Observing $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M}_2 \setminus \boldsymbol{T}_2 \,|\, \boldsymbol{M}_2$ since $\boldsymbol{M}_2$ is an Mb of $\boldsymbol{T}_2$, we obtain $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{Q} \,|\, (\boldsymbol{V} \setminus \boldsymbol{M}_2 \setminus \boldsymbol{T}_2 \setminus \boldsymbol{Q}) \cup \boldsymbol{M}_2$ by using the weak union property, or equivalently, $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{Q} \,|\, \boldsymbol{N} \cup \boldsymbol{S} \cup \boldsymbol{T}_1$, so $\mathbb{I}(\boldsymbol{T}_2; \boldsymbol{Q} \,|\, \boldsymbol{N} \cup \boldsymbol{S} \cup \boldsymbol{T}_1) = 0$. This means

$$\begin{aligned} 0 \leqslant c &= \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{Q} \,|\, \boldsymbol{N} \cup \boldsymbol{S}) \\ &= \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{T}_1 \cup \boldsymbol{Q} \,|\, \boldsymbol{N} \cup \boldsymbol{S}) - \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{T}_1 \,|\, \boldsymbol{N} \cup \boldsymbol{S} \cup \boldsymbol{Q}) \\ &= \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{T}_1 \,|\, \boldsymbol{N} \cup \boldsymbol{S}) + \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{Q} \,|\, \boldsymbol{N} \cup \boldsymbol{S} \cup \boldsymbol{T}_1) - \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{T}_1 \,|\, \boldsymbol{N} \cup \boldsymbol{S} \cup \boldsymbol{Q}) \\ &= \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{T}_1 \,|\, \boldsymbol{N} \cup \boldsymbol{S}) - \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{T}_1 \,|\, \boldsymbol{N} \cup \boldsymbol{S} \cup \boldsymbol{Q}). \end{aligned} \tag{14}$$

- (i) $\Leftarrow$ (ii): If $\mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}) \leqslant \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}')$ holds for any $\boldsymbol{S}' \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$, then (14) indicates $c = 0$ since $0 \leqslant c = \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{T}_1 \,|\, \boldsymbol{N} \cup \boldsymbol{S}) - \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{T}_1 \,|\, \boldsymbol{N} \cup \boldsymbol{S} \cup \boldsymbol{Q}) \leqslant 0$. Therefore,

  $$\mathbb{I}(\boldsymbol{T}; \boldsymbol{V} \setminus (\boldsymbol{N} \cup \boldsymbol{S}) \setminus \boldsymbol{T} \,|\, \boldsymbol{N} \cup \boldsymbol{S}) = \mathbb{I}(\boldsymbol{T}; \boldsymbol{Q} \,|\, \boldsymbol{N} \cup \boldsymbol{S}) = b = 0,$$

  because of $b = c$. That is, $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus (\boldsymbol{N} \cup \boldsymbol{S}) \setminus \boldsymbol{T} \,|\, \boldsymbol{N} \cup \boldsymbol{S}$, which means $\boldsymbol{N} \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}$, or equivalently, $\boldsymbol{S}$ is an MbS.
- (i) $\Rightarrow$ (ii): Observe that $\mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{V} \setminus \boldsymbol{T}) = \min_{\boldsymbol{S}' \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}} \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}')$ holds according to (14) holding for any $\boldsymbol{S} \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$ and $\boldsymbol{N} \cup \boldsymbol{S} \cup \boldsymbol{Q} = \boldsymbol{V} \setminus \boldsymbol{T}$. Then,

  $$\mathbb{I}(\boldsymbol{T}; \boldsymbol{Q} \,|\, \boldsymbol{N} \cup \boldsymbol{S}) = \mathbb{I}(\boldsymbol{T}; \boldsymbol{V} \setminus (\boldsymbol{N} \cup \boldsymbol{S}) \setminus \boldsymbol{T} \,|\, \boldsymbol{N} \cup \boldsymbol{S}) = 0$$

  follows immediately if $\boldsymbol{N} \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}$. By (13) and (14), we have

  $$\mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}) = \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{V} \setminus \boldsymbol{T}) = \min_{\boldsymbol{S}' \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}} \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}'), \tag{15}$$

  noting again $\boldsymbol{N} \cup \boldsymbol{S} \cup \boldsymbol{Q} = \boldsymbol{V} \setminus \boldsymbol{T}$. This means $\boldsymbol{N} \cup \boldsymbol{S}$ is an MbS.

To prove the equivalence between (ii) and (iii), we need only to show

$$\mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}) = \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{M}_1) + \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{M}_2) - \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2) - \mathbb{I}(\boldsymbol{T}; \boldsymbol{N} \cup \boldsymbol{S}). \tag{16}$$

In fact, using $\mathbb{I}(\boldsymbol{T}; \boldsymbol{N} \cup \boldsymbol{S}) = \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{N} \cup \boldsymbol{S}) + \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{N} \cup \boldsymbol{S} \,|\, \boldsymbol{T}_1)$, we have

$$\begin{aligned} d &\triangleq \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}) + \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2) + \mathbb{I}(\boldsymbol{T}; \boldsymbol{N} \cup \boldsymbol{S}) \\ &= \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}) + \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{N} \cup \boldsymbol{S}) + \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{T}_1) + \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{N} \cup \boldsymbol{S} \,|\, \boldsymbol{T}_1) \\ &= \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \cup \boldsymbol{N} \cup \boldsymbol{S}) + \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{N} \cup \boldsymbol{S} \cup \boldsymbol{T}_1) \\ &= \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{M}_1) + \mathbb{I}(\boldsymbol{T}_2; \boldsymbol{M}_2), \end{aligned}$$

which is equivalent to (16). This means (ii) $\Leftrightarrow$ (iii).

Now, we show that (i) is equivalent to (iv):

- (i) $\Rightarrow$ (iv): This implication holds clearly due to the decomposition property.
- (i) $\Leftarrow$ (iv): Assume $\boldsymbol{N}\cup\boldsymbol{S}$ is an Mb of $\boldsymbol{T}_1$ in $\boldsymbol{V}\setminus\boldsymbol{T}_2$, that is, $\boldsymbol{T}_1 \perp\!\!\!\perp (\boldsymbol{V}\setminus\boldsymbol{T}_2)\setminus(\boldsymbol{N}\cup\boldsymbol{S})\setminus\boldsymbol{T}_1 \,|\, \boldsymbol{N}\cup\boldsymbol{S}$, or equivalently, $\boldsymbol{V}\setminus(\boldsymbol{N}\cup\boldsymbol{S})\setminus\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{T}_1 \,|\, \boldsymbol{N}\cup\boldsymbol{S}$. On the other hand, $\boldsymbol{M}_2$ is an Mb of $\boldsymbol{T}_2$ in $\boldsymbol{V}$, meaning $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{V}\setminus\boldsymbol{M}_2\setminus\boldsymbol{T}_2 \,|\, \boldsymbol{M}_2$, which combined with the weak union property gives $\boldsymbol{V}\setminus(\boldsymbol{N}\cup\boldsymbol{S})\setminus\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{T}_2 \,|\, \boldsymbol{N}\cup\boldsymbol{S}\cup\boldsymbol{T}_1$. By the contraction property, $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V}\setminus(\boldsymbol{N}\cup\boldsymbol{S})\setminus\boldsymbol{T} \,|\, \boldsymbol{N}\cup\boldsymbol{S}$. This means $\boldsymbol{S}$ is an MbS. Similarly, if $\boldsymbol{N}\cup\boldsymbol{S}$ is an Mb of $\boldsymbol{T}_2$ in $\boldsymbol{V}\setminus\boldsymbol{T}_1$, we can show $\boldsymbol{S}$ is an MbS.

Finally, we prove that an MbS, $\boldsymbol{S}$, is an MBS if and only if $\boldsymbol{T}_2 \not\perp\!\!\!\perp Y \,|\, \boldsymbol{N}\cup(\boldsymbol{S}\setminus\{Y\})$ holds for any $Y\in\boldsymbol{S}$. We first prove the necessity by reductio ad absurdum. Suppose there is some variable $Y\in\boldsymbol{S}$ such that $\boldsymbol{T}_2 \perp\!\!\!\perp Y \,|\, \boldsymbol{N}\cup\boldsymbol{R}$, in which $\boldsymbol{R}\triangleq\boldsymbol{S}\setminus\{Y\}$. Recall that $\boldsymbol{S}$ is an MbS, we get

$$\boldsymbol{T}_2 \perp\!\!\!\perp (\boldsymbol{V}\setminus\boldsymbol{T}_1)\setminus(\boldsymbol{N}\cup\boldsymbol{S})\setminus\boldsymbol{T}_2 \,|\, \boldsymbol{N}\cup\boldsymbol{S}.$$

Equivalently, $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{V}\setminus(\boldsymbol{N}\cup\boldsymbol{R})\setminus\boldsymbol{T}\setminus\{Y\} \,|\, (\boldsymbol{N}\cup\boldsymbol{R})\cup\{Y\}$, which combined with $\boldsymbol{T}_2 \perp\!\!\!\perp Y \,|\, \boldsymbol{N}\cup\boldsymbol{R}$ gives $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{V}\setminus(\boldsymbol{N}\cup\boldsymbol{R})\setminus\boldsymbol{T} \,|\, \boldsymbol{N}\cup\boldsymbol{R}$, in view of the contraction property. That is, $\boldsymbol{T}_2 \perp\!\!\!\perp (\boldsymbol{V}\setminus\boldsymbol{T}_1)\setminus(\boldsymbol{N}\cup\boldsymbol{R})\setminus\boldsymbol{T}_2 \,|\, \boldsymbol{N}\cup\boldsymbol{R}$. Therefore, $\boldsymbol{N}\cup\boldsymbol{R}$ is an Mb of $\boldsymbol{T}_2$ in $\boldsymbol{V}\setminus\boldsymbol{T}_1$, and thus an MbS of $\boldsymbol{T}$ to $\boldsymbol{N}$. This contradicts the condition that $\boldsymbol{S}$ is an MBS of $\boldsymbol{T}$ to $\boldsymbol{N}$, and thus $\boldsymbol{T}_2 \not\perp\!\!\!\perp Y \,|\, \boldsymbol{N}\cup(\boldsymbol{S}\setminus\{Y\})$ holds for any $Y\in\boldsymbol{S}$.

To prove the sufficiency, we suppose $\boldsymbol{S}$ is not a MBS of $\boldsymbol{T}$ to $\boldsymbol{N}$, that is, there is some $\boldsymbol{R}\subsetneq\boldsymbol{S}$ such that $\boldsymbol{R}$ is an MbS of $\boldsymbol{T}$ to $\boldsymbol{N}$. Take any given variable, Y, in $\boldsymbol{S}\setminus\boldsymbol{R}$. Then, $\boldsymbol{N}\cup\boldsymbol{R}$ is an Mb of $\boldsymbol{T}_2$ in $\boldsymbol{V}\setminus\boldsymbol{T}_1$. That is, $\boldsymbol{T}_2 \perp\!\!\!\perp (\boldsymbol{V}\setminus\boldsymbol{T}_1)\setminus(\boldsymbol{N}\cup\boldsymbol{R})\setminus\boldsymbol{T}_2 \,|\, \boldsymbol{N}\cup\boldsymbol{R}$. By the weak union property, we have

$$\boldsymbol{T}_2 \perp\!\!\!\perp (\boldsymbol{V}\setminus\boldsymbol{T}_1)\setminus[\boldsymbol{N}\cup(\boldsymbol{S}\setminus\{Y\})]\setminus\boldsymbol{T}_2 \,|\, \boldsymbol{N}\cup(\boldsymbol{S}\setminus\{Y\}).$$

This combined with the decomposition property means $\boldsymbol{T}_2 \perp\!\!\!\perp Y \,|\, \boldsymbol{N}\cup(\boldsymbol{S}\setminus\{Y\})$, since

$$Y\in(\boldsymbol{V}\setminus\boldsymbol{T}_1)\setminus[\boldsymbol{N}\cup(\boldsymbol{S}\setminus\{Y\})]\setminus\boldsymbol{T}_2,$$

and thus leads to a contradiction to the condition that $\boldsymbol{T}_2 \not\perp\!\!\!\perp Y \,|\, \boldsymbol{N}\cup(\boldsymbol{S}\setminus\{Y\})$ holds for any $Y\in\boldsymbol{S}$.

The proof of Theorem 4 is completed. ■

Example 1 *Consider the BN* $(\mathbb{G},\mathbb{P})$ *over* $\boldsymbol{V}=\{A,B,C,D\}$ *presented in Figure 8, in which A, B, and C take* $\{1,2,3\}$ *while D takes* $\{1,2\}$. *Put* $\boldsymbol{T}=\{T_1,T_2\}$, $\boldsymbol{N}=(\boldsymbol{M}_1\cup\boldsymbol{M}_2)\setminus\boldsymbol{T}=\emptyset$, *and* $\boldsymbol{S}=\{C\}$, $\boldsymbol{S}_0=\{C,D\}$ *with* $T_1=A$, $T_2=B$, $\boldsymbol{M}_1=\{B\}$, $\boldsymbol{M}_2=\{A\}$. *By Figure 3, we can easily conclude that A and B are information equivalent with respect to C. This means*

$$\mathbb{I}(C;A)>0,\ \ \mathbb{I}(C;B\,|\,A)=0;\ \ \textit{and}\ \ \mathbb{I}(C;B)>0,\ \ \mathbb{I}(C;A\,|\,B)=0. \tag{17}$$

It follows from the chain rule for CMI (Cover and Thomas, 2006) that

(i) $\boldsymbol{M}_1$ *is an MB of* T_1 *in* $\boldsymbol{V}$: *By* (17), *we have*

- $\mathbb{I}(A;C,D\,|\,B)=\mathbb{I}(A;C\,|\,B)+\mathbb{I}(A;D\,|\,B,C)=0$, *since* $\{B,C\}$ *d-separates* $\{A\}$ *and* $\{D\}$;
- $\mathbb{I}(A;C,D)\geqslant\mathbb{I}(A;C)>0$.

(ii) $\boldsymbol{M}_2$ *is an MB of* T_2 *in* $\boldsymbol{V}$*: By* (17)*, we have*

- $\mathbb{I}(B; C, D \,|\, A) = \mathbb{I}(B; C \,|\, A) + \mathbb{I}(B; D \,|\, A, C) = 0$*, since* $\{A, C\}$ *d-separates* $\{B\}$ *and* $\{D\}$*;*
- $\mathbb{I}(B; C, D) \geqslant \mathbb{I}(B; C) > 0$*.*

(iii) $\boldsymbol{N} \cup \boldsymbol{S}$ *is an Mb of* $\boldsymbol{T}$ *in* $\boldsymbol{V}$*, so* $\boldsymbol{S}$ *is an MbS: By* (17)*,*

$$\mathbb{I}(A, B; D \,|\, C) = \mathbb{I}(A; D \,|\, C) + \mathbb{I}(B; D \,|\, A, C) = 0,$$

because $\{C\}$ *d-separates* $\{A\}$ *and* $\{D\}$*, while* $\{A, C\}$ *d-separates* $\{B\}$ *and* $\{D\}$*.*

(iv) $\mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}) = \min_{\boldsymbol{S}' \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}} \mathbb{I}(\boldsymbol{T}_1; \boldsymbol{T}_2 \,|\, \boldsymbol{N} \cup \boldsymbol{S}')$*: it suffices to show the following inequalities:*

- $\mathbb{I}(A; B \,|\, C) = \mathbb{I}(A; B \,|\, C, D)$*. In fact,*

$$\begin{aligned} \mathbb{I}(A; B \,|\, C, D) &= \mathbb{I}(A; B, D \,|\, C) - \mathbb{I}(A; D \,|\, C) \\ &= \mathbb{I}(A; B \,|\, C) + \mathbb{I}(A; D \,|\, B, C) - \mathbb{I}(A; D \,|\, C) = \mathbb{I}(A; B \,|\, C), \end{aligned}$$

  since both $\{B, C\}$ *and* $\{C\}$ *d-separate* $\{A\}$ *and* $\{D\}$*;*

- $\mathbb{I}(A; B \,|\, C) \leqslant \mathbb{I}(A; B \,|\, D)$*; In fact,*

$$\begin{aligned} \mathbb{I}(A; B \,|\, C) &= \mathbb{I}(A; B \,|\, C, D) = \mathbb{I}(A; B, C \,|\, D) - \mathbb{I}(A; C \,|\, D) \\ &= \mathbb{I}(A; B \,|\, D) + \mathbb{I}(A; C \,|\, B, D) - \mathbb{I}(A; C \,|\, D) \\ &= \mathbb{I}(A; B \,|\, D) - \mathbb{I}(A; C \,|\, D) \leqslant \mathbb{I}(A; B \,|\, D), \end{aligned}$$

  due to $\mathbb{I}(A; C \,|\, B, D) = 0$*, because of*

$$\begin{aligned} 0 &\leqslant \mathbb{I}(A; C \,|\, B, D) = \mathbb{I}(A; C, D \,|\, B) - \mathbb{I}(A; D \,|\, B) \\ &= \mathbb{I}(A; C \,|\, B) + \mathbb{I}(A; D \,|\, B, C) - \mathbb{I}(A; D \,|\, B) \;=\; -\mathbb{I}(A; D \,|\, B) \leqslant 0, \end{aligned}$$

  since $\mathbb{I}(A; C \,|\, B) = 0$ *(see Equation 17) and* $\{B, C\}$ *d-separates* $\{A\}$ *and* $\{D\}$*;*

- $\mathbb{I}(A; B \,|\, C) \leqslant \mathbb{I}(A; B)$*. In fact, by* (17)*, we have* $\mathbb{I}(A; C \,|\, B) = 0$*. Thus,*

$$\begin{aligned} \mathbb{I}(A; B) &= \mathbb{I}(A; B, C) - \mathbb{I}(A; C \,|\, B) = \mathbb{I}(A; C) + \mathbb{I}(A; B \,|\, C) - \mathbb{I}(A; C \,|\, B) \\ &= \mathbb{I}(A; C) + \mathbb{I}(A; B \,|\, C) \geqslant \mathbb{I}(A; B \,|\, C). \end{aligned}$$

(v) $\mathbb{I}(\boldsymbol{T}; \boldsymbol{S} \,|\, \boldsymbol{N}) = \max_{\boldsymbol{S}' \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}} \mathbb{I}(\boldsymbol{T}; \boldsymbol{S}' \,|\, \boldsymbol{N})$*: the proof is omitted.*

(vi) $\boldsymbol{N} \cup \boldsymbol{S}$ *is an Mb of* T_1 *in* $\boldsymbol{V} \setminus \{T_2\}$*: By* (17)*, we have*

- $\mathbb{I}(A; C, D) \geqslant \mathbb{I}(A; C) > 0$*;*
- $\mathbb{I}(A; D \,|\, C) = 0$*, since* $\{C\}$ *d-separates* $\{A\}$ *and* $\{D\}$*.*

(vii) $\boldsymbol{N} \cup \boldsymbol{S}$ *is an Mb of* T_2 *in* $\boldsymbol{V} \setminus \{T_1\}$*: By* (17)*, we have*

- $\mathbb{I}(B; C, D) \geqslant \mathbb{I}(B; C) > 0$*;*
- $\mathbb{I}(B; D \,|\, C) = 0$*, since* $\{C\}$ *d-separates* $\{B\}$ *and* $\{D\}$*.*

(viii) $\boldsymbol{S}$ *is an MBS;* $\boldsymbol{S}_0$ *is an MbS (but not an MBS):* $\mathbb{I}(A; B \,|\, C, D) = \mathbb{I}(A; B \,|\, C)$*. In fact,*

$$\mathbb{I}(A; B \,|\, C, D) = \mathbb{I}(A; B, D \,|\, C) - \mathbb{I}(A; D \,|\, C) = \mathbb{I}(A; B \,|\, C) + \mathbb{I}(A; D \,|\, B, C) - 0 = \mathbb{I}(A; B \,|\, C),$$

since both $\{B, C\}$ *and* $\{C\}$ *d-separate* $\{A\}$ *and* $\{D\}$*.* ■

Theorem 5 *Let $\boldsymbol{M}_j$ be an MB of $T_j^{(d)}$ in $\boldsymbol{V} \setminus \boldsymbol{T}$ for $j = 1, \cdots, t$. Then, $\boldsymbol{M} \triangleq \cup_{j=1}^{k} \boldsymbol{M}_j$ is an Mb of $\boldsymbol{T}$. Further, $\boldsymbol{M}$ is an MB of $\boldsymbol{T}$ iff for any $X \in \boldsymbol{M}$ there is some j such that $T_j^{(d)} \not\perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\}$.*

Proof Recall that T is the merged version of $\boldsymbol{T}$, while $\boldsymbol{T}^{(d)}$ is the dummy version of T; all of them have the same MBs.

First, we have $T_j^{(d)} \perp\!\!\!\perp (\boldsymbol{V} \setminus \boldsymbol{T}) \setminus \boldsymbol{M}_j \setminus \{T_j^{(d)}\} \mid \boldsymbol{M}_j$ for $j = 1, \cdots, t$. Considering $T_j^{(d)} \notin \boldsymbol{V} \setminus \boldsymbol{T}$, it follows that $T_j^{(d)} \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{T} \setminus \boldsymbol{M}_j \mid \boldsymbol{M}_j$, which combined with the weak union property gives

$$T_j^{(d)} \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{T} \setminus \boldsymbol{M} \mid \boldsymbol{M}, \quad j = 1, \cdots, t, \tag{18}$$

since $\boldsymbol{M}_j \subseteq \boldsymbol{M}$. Putting $\boldsymbol{U} \triangleq \boldsymbol{V} \setminus \boldsymbol{T} \setminus \boldsymbol{M}$, the above independence statements imply

$$\mathbb{P}\big(T_j^{(d)} = 1,\ \boldsymbol{U} = \boldsymbol{u} \,\big|\, \boldsymbol{M}\big) = \mathbb{P}\big(T_j^{(d)} = 1 \,\big|\, \boldsymbol{M}\big)\mathbb{P}\big(\boldsymbol{U} = \boldsymbol{u} \,\big|\, \boldsymbol{M}\big), \quad j = 1, \cdots, t,$$

or equivalently, $\mathbb{P}(T = j,\ \boldsymbol{U} = \boldsymbol{u} \mid \boldsymbol{M}) = \mathbb{P}(T = j \mid \boldsymbol{M})\mathbb{P}(\boldsymbol{U} = \boldsymbol{u} \mid \boldsymbol{M})$, meaning $T \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{T} \setminus \boldsymbol{M} \mid \boldsymbol{M}$, and thus $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{T} \setminus \boldsymbol{M} \mid \boldsymbol{M}$. This shows $\boldsymbol{M}$ is an Mb of $\boldsymbol{T}$.

In what follows, we prove $\boldsymbol{M}$ is an MB of $\boldsymbol{T}$ if and only if, for any $X \in \boldsymbol{M}$, $T_j^{(d)} \not\perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\}$ holds for some j:

"⇒" Assume $\boldsymbol{M}$ is an MB of $\boldsymbol{T}$. Suppose there is some variable X such that $T_j^{(d)} \perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\}$ holds for any j. Then, by (18) and the contraction property, we get

$$T_j^{(d)} \perp\!\!\!\perp (\boldsymbol{V} \setminus \boldsymbol{T}) \setminus (\boldsymbol{M} \setminus \{X\}) \mid \boldsymbol{M} \setminus \{X\}, \quad j = 1, \cdots, t. \tag{19}$$

Similar to the proof of the first conclusion, it can be readily proven that (19) implies

$$\boldsymbol{T} \perp\!\!\!\perp (\boldsymbol{V} \setminus \boldsymbol{T}) \setminus (\boldsymbol{M} \setminus \{X\}) \mid \boldsymbol{M} \setminus \{X\},$$

meaning that $\boldsymbol{M}$ has a proper subset, $\boldsymbol{M} \setminus \{X\}$, which is an Mb of $\boldsymbol{T}$. This contradicts the minimality of $\boldsymbol{M}$, and thus proves the necessity.

"⇐" Suppose $\boldsymbol{M}$ is not an MB of $\boldsymbol{T}$ (or T). Then, there is some $X \in \boldsymbol{M}$ such that $T \perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\}$. It follows that $\mathbb{P}(T = j, X = x \mid \boldsymbol{M} \setminus \{X\}) = \mathbb{P}(T = j \mid \boldsymbol{M} \setminus \{X\})\mathbb{P}(X = x \mid \boldsymbol{M} \setminus \{X\})$ holds for any $j = 1, \cdots, t$. Or equivalently, we have

$$\mathbb{P}\big(T_j^{(d)} = 1, X = x \,\big|\, \boldsymbol{M} \setminus \{X\}\big) = \mathbb{P}\big(T_j^{(d)} = 1 \,\big|\, \boldsymbol{M} \setminus \{X\}\big)\mathbb{P}\big(X = x \,\big|\, \boldsymbol{M} \setminus \{X\}\big). \tag{20}$$

Futher, (20) indicates

$$\begin{aligned}
\mathbb{P}\big(T_j^{(d)} = 0, X = x \,\big|\, \boldsymbol{M} \setminus \{X\}\big) &= \mathbb{P}\big(X = x \,\big|\, \boldsymbol{M} \setminus \{X\}\big) - \mathbb{P}\big(T_j^{(d)} = 1, X = x \,\big|\, \boldsymbol{M} \setminus \{X\}\big) \\
&= \big[1 - \mathbb{P}\big(T_j^{(d)} = 1 \,\big|\, \boldsymbol{M} \setminus \{X\}\big)\big]\mathbb{P}\big(X = x \,\big|\, \boldsymbol{M} \setminus \{X\}\big) \\
&= \mathbb{P}\big(T_j^{(d)} = 0 \,\big|\, \boldsymbol{M} \setminus \{X\}\big)\mathbb{P}\big(X = x \,\big|\, \boldsymbol{M} \setminus \{X\}\big).
\end{aligned} \tag{21}$$

By (20) and (21), we get $T_j^{(d)} \perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\}$, which contradicts $T_j^{(d)} \not\perp\!\!\!\perp X \mid \boldsymbol{M} \setminus \{X\}$. This proves the sufficiency.

The proof is completed. ■

Theorem 6 (Correctness of IAMBS and KIAMBS) *Assume that $\boldsymbol{T}_2$ satisfies the local composition property, and that all CI tests are correct. Then* (i) *IAMBS outputs an MB of $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$;* (ii) *KIAMBS outputs an MB of $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$ for any $K \in [0, 1)$.*

Proof Clearly, $\boldsymbol{N} \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}_2$ in $\boldsymbol{V} \setminus \boldsymbol{T}_1$ at the end of the growing phase of either IAMBS or KIAMBS under the local composition assumption, as in IAMB and KIAMB. Therefore, $\boldsymbol{S}$ is an MbS at the end of this stage. According to the last conclusion of Theorem 4, $\boldsymbol{S}$ is an MBS after it is refined. Finally, as a direct consequence of Lemma 4 (shown below), $\boldsymbol{N} \cup \boldsymbol{S}$ is an MB at the end of the algorithm, considering the process of refining $\boldsymbol{N}$ is similar to that of refining $\boldsymbol{S}$. ■

Remark 2 *The following two statements hold:* (a) *violating local intersection implies violating adjacency faithfulness;* (b) *under the orientation faithfulness condition, violating local composition at the end of the first phase of IAMB or KIAMB or IAMBS or KIAMBS means violating adjacency faithfulness.*

Proof By Lemma 1, the violation of the local intersection property means information equivalence occurs; further, Lemeire et al. (2012) showed that information equivalence is one of the cases of violating adjacency faithfulness. Hence, the violation of local intersection is one of the violations of adjacency faithfulness.

Now, we show that the violation of local composition, which is present at the end of the first phase of IAMB or KIAMB, is also one of the violations of adjacency faithfulness under the *orientation faithfulness condition*.

In fact, let $\boldsymbol{M}$ be the output of the first phase of IAMB or KIAMB, but not an Mb of $\boldsymbol{T}$. Without loss of generality, we assume $|\boldsymbol{T}| = 1$ and $\boldsymbol{T} = \{T\}$. Then, $T \perp\!\!\!\perp X \,|\, \boldsymbol{M}$ holds for any $X \in \boldsymbol{V} \setminus \boldsymbol{M} \setminus \{T\}$ but $T \not\perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M} \setminus \{T\} \,|\, \boldsymbol{M}$. Considering that the set $\boldsymbol{M}_T$ composed of the parents, children, and spouses of T is an Mb of T, we have $\boldsymbol{M} \not\supseteq \boldsymbol{M}_T$. Thus, there is some $X \in \boldsymbol{M}_T$ such that $X \notin \boldsymbol{M}$. If X is a spouse of T, then all the children of T and X are not in $\boldsymbol{M}$ (if not so, $T \not\perp\!\!\!\perp X \,|\, \boldsymbol{M}$ holds immediately following from the orientation faithfulness condition, and thus contradicts $T \perp\!\!\!\perp X \,|\, \boldsymbol{M}$ since $X \notin \boldsymbol{M}$). In this sense, we conclude that there is some node X adjacent to T such that $T \perp\!\!\!\perp X \,|\, \boldsymbol{M}$. This means the adjacency faithfulness condition is violated.

In short words, both the violation of the local composition property (present at the end of the first phase of IAMB or KIAMB) and the violation of the local intersection property are the violations of adjacency faithfulness, under the orientation faithfulness condition. ■

Lemma 3 (a) *If there is $\boldsymbol{P} \subseteq \boldsymbol{M}_1 \setminus \boldsymbol{T}_2$ such that $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{P} \,|\, (\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_2$, then $(\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_2$ is an Mb of $\boldsymbol{T}_1$;* (b) *If there is $\boldsymbol{Q} \subseteq \boldsymbol{N} \setminus \boldsymbol{P}$ such that $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{Q} \,|\, (\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_2$ and $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{Q} \,|\, (\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_1$, then $(\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_2$ is an Mb of $\boldsymbol{T}_1$, and $(\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_1$ is an Mb of $\boldsymbol{T}_2$.*

Proof Considering that $\boldsymbol{N} \cup \boldsymbol{T}_2$ is an Mb of $\boldsymbol{T}_1$, we have $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus (\boldsymbol{N} \cup \boldsymbol{T}_2) \setminus \boldsymbol{T}_1 \,|\, (\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_2 \cup \boldsymbol{P}$, which combined with $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{P} \,|\, (\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_2$ implies $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus [(\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_2] \setminus \boldsymbol{T}_1 \,|\, (\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_2$, in view of the contraction property. The first conclusion is proved.

For convenience, we denote now $\boldsymbol{N}_1 \triangleq \boldsymbol{M}_1 \setminus \boldsymbol{T}_2$. To show the second conclusion, we note that $\boldsymbol{P} \subseteq \boldsymbol{N}_1$, so $(\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_1$ is an Mb of $\boldsymbol{T}_2$. It follows that: (i) $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus [(\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_2] \setminus \boldsymbol{T}_1 \,|\, (\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_2$, which combined with $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{Q} \,|\, (\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_2$ gives $\boldsymbol{T}_1 \perp\!\!\!\perp \boldsymbol{V} \setminus [(\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_2] \setminus \boldsymbol{T}_1 \,|\, (\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_2$; and (ii) $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{V} \setminus [(\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_1] \setminus \boldsymbol{T}_2 \,|\, (\boldsymbol{N} \setminus \boldsymbol{P}) \cup \boldsymbol{T}_1$, which combined with $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{Q} \,|\, (\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_1$ yields $\boldsymbol{T}_2 \perp\!\!\!\perp \boldsymbol{V} \setminus [(\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_1] \setminus \boldsymbol{T}_2 \,|\, (\boldsymbol{N} \setminus \boldsymbol{P} \setminus \boldsymbol{Q}) \cup \boldsymbol{T}_1$. The second conclusion is also proved. ■

Lemma 4 *Let $\boldsymbol{T}_i$ be a subset of $\boldsymbol{V}$ with an Mb $\boldsymbol{M}_i$ for $i = 1, 2$, and $\boldsymbol{S} \subseteq \boldsymbol{V} \setminus \boldsymbol{N} \setminus \boldsymbol{T}$ be an MBS of $\boldsymbol{T}$ to $\boldsymbol{N}$, with $\boldsymbol{T} = \boldsymbol{T}_1 \cup \boldsymbol{T}_2$ and $\boldsymbol{N} = (\boldsymbol{M}_1 \cup \boldsymbol{M}_2) \setminus \boldsymbol{T}$. Assume $\boldsymbol{N}_0$ be a subset of $\boldsymbol{N}$ such that $\boldsymbol{N}_0 \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}$. Then $\boldsymbol{N}_0 \cup \boldsymbol{S}$ is an MB of $\boldsymbol{T}$ if and only if $\boldsymbol{T} \not\perp\!\!\!\perp Y \,|\, (\boldsymbol{N}_0 \setminus \{Y\}) \cup \boldsymbol{S}$ holds for any $Y \in \boldsymbol{N}_0$.*

Proof By the definition of MBS, $\boldsymbol{N} \cup \boldsymbol{R}$ and thus $\boldsymbol{N}_0 \cup \boldsymbol{R}$ will never be an Mb of $\boldsymbol{T}$ for any $\boldsymbol{N}_0 \subseteq \boldsymbol{N}$ and $\boldsymbol{R} \subsetneq \boldsymbol{S}$, in view of the weak union property.

- Necessity: Suppose there is some $Y \in \boldsymbol{N}_0$ such that $\boldsymbol{T} \perp\!\!\!\perp Y \,|\, (\boldsymbol{N}_0 \setminus \{Y\}) \cup \boldsymbol{S}$. By the precondition that $\boldsymbol{N}_0 \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}$, we have $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus (\boldsymbol{N}_0 \cup \boldsymbol{S}) \setminus \boldsymbol{T} \,|\, [(\boldsymbol{N}_0 \setminus \{Y\}) \cup \boldsymbol{S}] \cup \{Y\}$. These two relationships combined with the contraction property imply

$$\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus [(\boldsymbol{N}_0 \setminus \{Y\}) \cup \boldsymbol{S}] \setminus \boldsymbol{T} \,|\, (\boldsymbol{N}_0 \setminus \{Y\}) \cup \boldsymbol{S},$$

  or equivalently, $(\boldsymbol{N}_0 \setminus \{Y\}) \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}$. This contradicts that $\boldsymbol{N}_0 \cup \boldsymbol{S}$ is an MB of $\boldsymbol{T}$.
- Sufficiency: Suppose $\boldsymbol{N}_0 \cup \boldsymbol{S}$ is not an MB of $\boldsymbol{T}$, that is, there is some $\boldsymbol{N}_0' \subsetneq \boldsymbol{N}_0$ such that $\boldsymbol{N}_0' \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}$. Take any given variable, Y, in $\boldsymbol{N}_0 \setminus \boldsymbol{N}_0'$. It can be shown that $\boldsymbol{T} \perp\!\!\!\perp Y \,|\, (\boldsymbol{N}_0 \setminus \{Y\}) \cup \boldsymbol{S}$, which leads to a contradiction. Hence, $\boldsymbol{N}_0 \cup \boldsymbol{S}$ is an MB of $\boldsymbol{T}$.

The proof is completed. ■

Appendix C. Improving the Log-Likelihood Ratio Test

In Subsection 4.3, we mentioned that the X^2 or G^2 test is suitable only for cases of small $|\boldsymbol{T}|$, and then summarized some improving methods proposed in the literature (Lawley, 1956; Hosmane, 1986, 1987, 1990; Brin et al., 1997; Silverstein et al., 1998; Aliferis et al., 2010b). However, we need more suitable CI testing methods when working on the MB discovery problem for multiple targets. In this appendix, we discuss a practical way of improving the G^2 test by damping the number of degrees of freedom for the G^2 statistic.

Consider the G^2 statistic, $G^2(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) \triangleq 2n \cdot \mathbb{I}_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$, which approximates to the chi-square variate with $r \triangleq (r_{\boldsymbol{X}} - 1)(r_{\boldsymbol{Y}} - 1) r_{\boldsymbol{Z}}$ degrees of freedom, namely $\chi^2(r)$, where $r_{\boldsymbol{\xi}}$ represents the number of configurations for $\boldsymbol{\xi}$ (de Campos, 2006, p. 2158).

Theoretically, $G^2(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z})$ is a reasonable statistic for testing the hypothesis "$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$" *when n is large enough.* Unfortunately, this precondition is practically hard to be valid in many situations (Cochran, 1954; Yaramakala, 2004; Bromberg and Margaritis, 2009) due to the following reason: Let $\boldsymbol{X} = \{X_{i_1}, \cdots, X_{i_x}\}$, $\boldsymbol{Y} = \{X_{j_1}, \cdots, X_{j_y}\}$, and $\boldsymbol{Z} = \{X_{k_1}, \cdots, X_{k_z}\}$, in which each variable X_ℓ takes r_ℓ values. Then $r = (\prod_{\ell=1}^{x} r_{i_\ell} - 1)(\prod_{\ell=1}^{y} r_{j_\ell} - 1)(\prod_{\ell=1}^{z} r_{k_\ell})$, which is exponential with respect to x, y, and z. On the one hand, by the Wilson-Hilferty approximation for $\chi^2_\alpha(r)$ (de Campos, 2006; Gao, 2005), we obtain $\chi^2_\alpha(r) \approx c_{\alpha,r}\, r$, in which $c_{\alpha,r} \triangleq \left(1 - 2/(9r) + \sqrt{2/(9r)}\, z_\alpha\right)^3$ is a bit larger than 1, with z_α being the upper α-quantile of the standard normal distribution; on the other hand, we can show $\mathbb{I}_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) \leqslant \log_2 r_{\boldsymbol{X},\boldsymbol{Y}}$ with $r_{\boldsymbol{X},\boldsymbol{Y}} \triangleq \min\{\prod_{\ell=1}^{x} r_{i_\ell}, \prod_{\ell=1}^{y} r_{j_\ell}\}$. It follows that

$$p(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) = \mathbb{P}\left\{\chi^2(r) \geqslant 2n \cdot \mathbb{I}_{\boldsymbol{D}}(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z})\right\} \geqslant \mathbb{P}\left\{\chi^2(r) \geqslant 2n \cdot \log_2 r_{\boldsymbol{X},\boldsymbol{Y}}\right\}.$$

Suppose we are doing a G^2 test for the *false hypothesis* "$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$" (i.e., the truth is $\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$). Then, at least $\frac{\chi^2_\alpha(r)}{2 \log_2 r_{\boldsymbol{X},\boldsymbol{Y}}} \approx \frac{c_{\alpha,r}\, r}{2 \log_2 r_{\boldsymbol{X},\boldsymbol{Y}}} = O\left(\frac{r}{\log_2 r_{\boldsymbol{X},\boldsymbol{Y}}}\right)$ instances are required if we expect the statistical

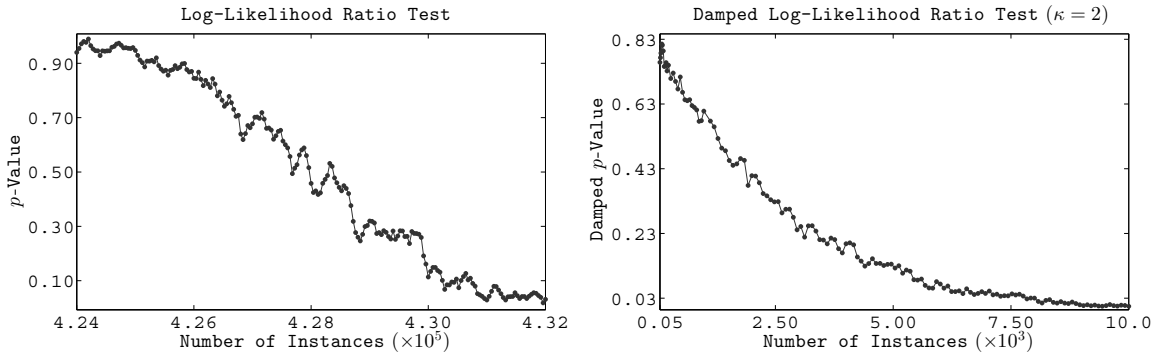

Figure 20: p-value/damped p-value versus the number of instances, n: the left subfigure illustrates why a very large n may still not be "large enough" for making a correct decision about the *false hypothesis* "$\boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{Y}_1 \,|\, \boldsymbol{Z}_1$" based on the G^2 test: at least 4.31×10^5 instances are required; while the right illustrates why the damped G^2 test is suitable for testing the same *false hypothesis*: about 8000 instances are sufficient.

decision can be correctly made with the significance level α (or equivalently, $p(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z}) \leqslant \alpha$). In many practical situations, however, n may be far smaller than the required number of sample instances with the magnitude of at least $O(r/\log_2 r_{\boldsymbol{X},\boldsymbol{Y}})$, recalling that r is exponential with respect to x, y, and z, especially when too many variables are involved. In this case, the statistical decision made for the hypothesis will be wrong.

Taking the ALARM network presented in Figure 4 for example, we put

$$\boldsymbol{X}_1 = \{X_{36}\},\ \boldsymbol{Y}_1 = \{X_{11}, X_{34}, X_{35}, X_{37}\},\ \text{and}\ \boldsymbol{Z}_1 = \{X_4, X_{14}, X_{15}, X_{16}, X_{18}, X_{21}, X_{22}, X_{31}\};$$

then we compute the p-value versus the number of instances from 1000 to 1,000,000. The results are drawn in Figure 20 (averaged over 10 different samples with the same size). Note that the truth is $\boldsymbol{X}_1 \not\perp\!\!\!\perp \boldsymbol{Y}_1 \,|\, \boldsymbol{Z}_1$ since $\boldsymbol{Y}_1$ is an Mb of $\boldsymbol{X}_1$. By the figure, the CI test for the *false hypothesis* "$\boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{Y}_1 \,|\, \boldsymbol{Z}_1$" is not correct unless at least $n_{\min} \approx 4.31 \times 10^5$ sample instances are available. It is mentioned that, in this example, $r/\log_2 r_{\boldsymbol{X}_1,\boldsymbol{Y}_1} \approx 2.42 \times 10^5$.

In short words, the precondition, "*when n is large enough*", for the theoretical assertion that "G^2 is a reasonable statistic for CI testing" may be hard to be guaranteed in practice because the above analysis and the numerical example indicate that a seemingly very large n may still not be "large enough". The problem is then how to improve on the G^2 test.

Observe that, for the G^2 test, the major reason for failing to make a correct statistical decision on CI testing is that the theoretical value of r is far larger than its data-driven value, due to the null cells frequently existing in the multi-contingency tables of $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$ (e.g., Yaramakala, 2004, p. 34). In other words, the linear increase of n is hard to exponentially bring null cells into valid cells. Hence, a feasible way of improving the log-likelihood ratio G^2 test is to damp the increase of r such that the unmatched behaviours of n and r can get alleviated to a certain degree. Mathematically, we replace the theoretical value of r in $p(\boldsymbol{X}; \boldsymbol{Y} \,|\, \boldsymbol{Z})$ with its a damped version, $g_{n,\kappa}(r)$, defined in (6), where $\kappa > 0$ is a constant, based on which $\frac{n}{\kappa}$ measures the amount of valid cells that n sample

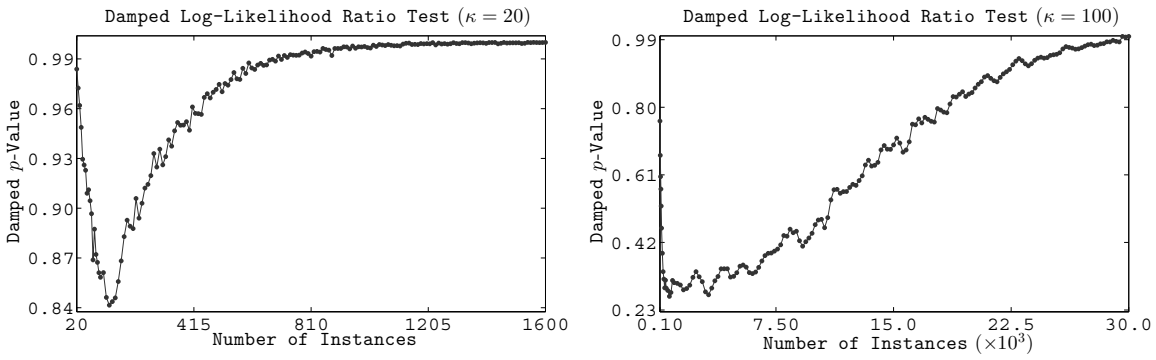

Figure 21: Damped p-value versus the number of instances: κ is taken as 20 and 100, respectively.

instances can support. It is easily seen that $g_{n,\kappa}(\cdot)$ possesses the following properties, which interpret the reasonability of employing such a damping procedure in the G^2 test:

- $g_{n,\kappa}(r)$ is monotonically increasing versus n for given r, and $\lim_{n\to+\infty} g_{n,\kappa}(r) = r$. This means more instances may generate more valid cells in the multi-contingency tables of $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$, and all the theoretical degrees of freedom are valid when n is large enough.
- $g_{n,\kappa}(r)$ is monotonically increasing versus r for given n, and $\lim_{r\to+\infty} g_{n,\kappa}(r) = \frac{n}{\kappa}$. This means a larger r should correspond to a larger $g_{n,\kappa}(r)$, but not exceeding the supporting capacity of the data.
- For sufficient data, the damping function $g_{n,\kappa}(\cdot)$ only plays a little role; while for insufficient data, it trades off the theoretical r and the supporting capacity of the data.

For convenience, we call the resulted p-value, denoted by $p_g(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$ instead of $p(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$, and the resulted testing method to be the *damped p-value* and the *damped log-likelihood ratio test* (or damped G^2 test). Further, we use the negative damped p-value, $f^{(3)}_{\boldsymbol{D}}$ defined in (7), as the association function. It is mentioned here that the damped G^2 test approximately degenerates into the ordinary G^2 test when taking κ as a very small positive number.

The damped G^2 test may be more suitable than the ordinary G^2 test when too many variables are involved in the conditional set. We implement this testing method (by taking κ as $2, 3, \cdots, 10$, respectively) on the *false hypothesis* "$\boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{Y}_1 \mid \boldsymbol{Z}_1$", and find that the correct decision "$\boldsymbol{X}_1 \not\perp\!\!\!\perp \boldsymbol{Y}_1 \mid \boldsymbol{Z}_1$" is always made for $\kappa \geqslant 3$, even when the number of instances is smaller than 1000. For the case of $\kappa = 2$, we present the results in the right subfigure of Figure 20, from which it is seen that, for the damped G^2 test, about 8000 sample instances are sufficient to make the correct decision. Note also that the ordinary G^2 test needs at least 4.31×10^5 instances.

However, the damped G^2 test may also face a potential danger: it may excessively damp the theoretical value of r if a too large κ is inappropriately used. Here, "excessively damping r" means that a too large value of κ will lead to a too small $g_{n,\kappa}(r)$ such that the damped G^2 test incorrectly reject a *true hypothesis* "$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$". To illustrate this explanation, we put

$$\boldsymbol{X}_2 = \{X_{21}\}, \quad \boldsymbol{Y}_2 = \{X_{11}, X_{34}, X_{35}, X_{36}, X_{37}\}, \quad \text{and} \quad \boldsymbol{Z}_2 = \{X_{15}, X_{19}, X_{20}, X_{22}, X_{29}\}$$

from the ALARM network. The truth is $\boldsymbol{X}_2 \perp\!\!\!\perp \boldsymbol{Y}_2 \mid \boldsymbol{Z}_2$ since $\boldsymbol{Z}_2$ is an Mb of $\boldsymbol{X}_2$. Now, use damped G^2 to test the *true hypothesis* "$\boldsymbol{X}_2 \perp\!\!\!\perp \boldsymbol{Y}_2 \mid \boldsymbol{Z}_2$" by taking κ as $3, 4, \cdots, 10, 20, 100$, respectively. All

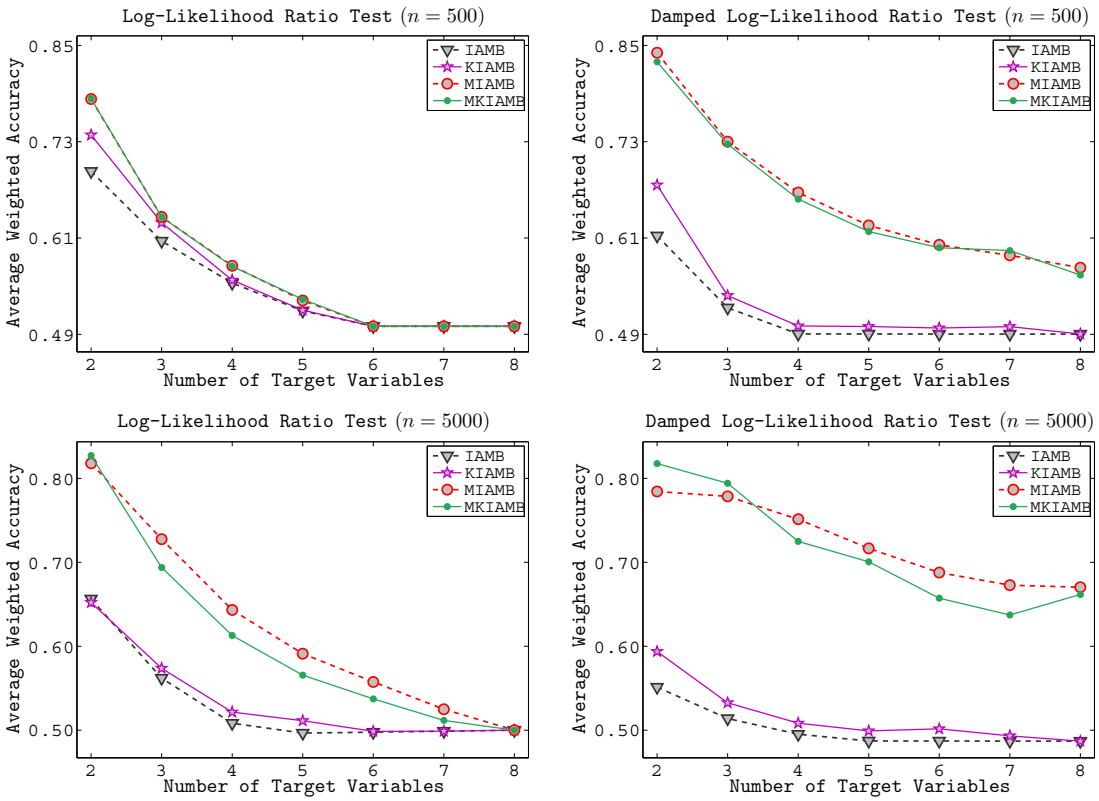

Figure 22: An illustration on why the damped G^2 test is more suitable than the ordinary G^2 test for the problem that involves too many variables, by virtue of the Pigs BN: the G^2 test no longer works when there are too many variables are involved, while the damped G^2 test remains valid in all considered cases.

the decisions are correctly made. However, a too large κ may be more apt to yield a relatively small damped p-value although it is still larger than α, as shown by Figure 21: 0.84 for the case of $\kappa = 20$ and only 0.27 for the case of $\kappa = 100$. Hence, to avoid the potential danger of excessively damping r, we conservatively recommend to take κ from the interval [3, 10] in practice. In our benchmarking study, we employ $\kappa = 5$, which is large enough for testing *false hypotheses* and small enough for testing *true hypotheses*.

To further illustrate why the damped G^2 test is more suitable than the ordinary G^2 when working on a problem that involves too many variables, we make experiments on the six synthetical BNs based on `IAMB`, `KIAMB`, `MIAMB`, and `MKIAMB`.

For each algorithm, the G^2 test and the damped G^2 test are implemented for CI testing. Accordingly, the association functions, $f_{\boldsymbol{D}}^{(1)}$ and $f_{\boldsymbol{D}}^{(3)}$ defined by (5) and (7) are used. Figure 22 presents the results of the Pigs network, in which the left two are based on the ordinary G^2 test while the right two are based on the damped G^2 test. As seen, for the case of $n = 500$ the G^2 test becomes invalid when the target, $\boldsymbol{T}$, contains 6 or more variables, and for the case of $n = 5000$ this method no longer works when $|\boldsymbol{T}| = 8$. In comparison, the damped G^2 test is suitable for all cases. The results of the other five BNs show similar conclusions.

Appendix D. Used Acronyms

BFMB	breadth first search of Markov boundary algorithm (Fu and Desmarais, 2007).
BN	Bayesian network.
CI	conditional independence.
CMI	conditional mutual information.
CT-support	contingency table support: a set of items S has CT-support s at the $t\%$ level if at least $t\%$ of the cells in the contingency table for S have value s (Silverstein et al., 1998).
DAG	directed acyclic graph.
E-partition	equivalent partition (Lemeire, 2007): a relation $\Re \subset \boldsymbol{X} \otimes \boldsymbol{Y}$ defines an E-partition in $\boldsymbol{Y}_{\text{dom}}$ to a partition of $\boldsymbol{X}_{\text{dom}}$, if: (i) $\neg(\boldsymbol{x}_2 \Re \boldsymbol{y}_1)$ holds for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \boldsymbol{X}_{\text{dom}}$ belonging to different partitions and for any $\boldsymbol{y}_1 \in \boldsymbol{Y}_{\text{dom}}$ with $\boldsymbol{x}_1 \Re \boldsymbol{y}_1$; and (ii) for every $\boldsymbol{X}^{(k)}_{\text{dom}}$, there exist $\boldsymbol{x}_1 \in \boldsymbol{X}^{(k)}_{\text{dom}}$ and $y_1 \in \boldsymbol{Y}_{\text{dom}}$ such that $\boldsymbol{x}_1 \Re \boldsymbol{y}_1$.
FS	feature selection.
GLL	generalized local learning: an algorithmic framework for local causal discovery and FS proposed by Aliferis et al. (2010a).
GS	grow-shrink algorithm (Margaritis and Thrun, 1999, 2000).
HITON	an MB discovery algorithm, pronounced hee-tón, from the Greek Χιτών, for "cover", "cloak", or "blanket" (Aliferis et al., 2003).
IAMB	incremental association Markov boundary algorithm (Tsamardinos et al., 2003); see Algorithm 3 for details.
IAMBS	an IAMB-based Markov boundary supplementary algorithm, outputting an MB for multiple targets (Algorithm 1).
KIAMB	a stochastic variant of IAMB (Peña et al., 2007); see Algorithm 3 for details.
KIAMBS	an KIAMB-based Markov boundary supplementary algorithm, outputting an MB for multiple targets (Algorithm 1).
KS	Koller-Sahami algorithm (Koller and Sahami, 1996).
LibSVM	a library for support vector machines contributed by Chang and Lin (2011).
Mb	Markov blanket: we call $\boldsymbol{M}$ an Mb of $\boldsymbol{T}$ if $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \setminus \boldsymbol{M} \setminus \boldsymbol{T} \mid \boldsymbol{M}$ (Definition 1).
MB	Markov boundary: an MB of $\boldsymbol{T}$ is any Mb such that none of its proper subsets is an Mb of $\boldsymbol{T}$ (Definition 1).
MbS	Markov blanket supplementary: we call $\boldsymbol{S}$ an MbS of $\boldsymbol{T}$ to $\boldsymbol{N}$, if $\boldsymbol{N} \cup \boldsymbol{S}$ is an Mb of $\boldsymbol{T}$ (Definition 3).
MBS	Markov boundary supplementary: an MBS is any MbS such that none of its proper subsets is an MbS (Definition 3).
MIAMB	an IAMB and IAMBS-based algorithm, outputting an MB for multiple targets (see Algorithm 2 for details).
MKIAMB	an KIAMB and KIAMBS-based algorithm, outputting an MB for multiple targets (Algorithm 2).
MMMB	max-min Markov boundary algorithm (Tsamardinos et al., 2006).

PCMB	parents and children based Markov boundary algorithm (Peña et al., 2007).
RF	random forest algorithm.
RT	running time: the single CPU time implemented on an Intel i7-3612QM 2.1 GHz and Windows 7 with 64 bits.
SVM	support vector machine (in one-against-one approach).
T-partition	target partition (Lemeire, 2007): the domain, $\boldsymbol{X}_{\text{dom}}$, of $\boldsymbol{X}$ can be partitioned into some disjoint subsets $\boldsymbol{X}_{\text{dom}}^{(k)}$ for which $\mathbb{P}(\boldsymbol{T} \mid \boldsymbol{x})$ is the same for all $\boldsymbol{x} \in \boldsymbol{X}_{\text{dom}}^{(k)}$. This is called the T-partition of $\boldsymbol{X}_{\text{dom}}$ with respect to $\boldsymbol{T}$.
WA	weighted accuracy: WA is the average of the rate of true members and that of true nonmembers of an MB with respect to the truth.
WP	weighted precision: WP is the average of the rate of true members and that of true nonmembers of an MB with respect to the test.

References

Constantin F Aliferis, Ioannis Tsamardinos, and Alexander Statnikov. Hiton: a novel Markov blanket algorithm for optimal variable selection. In *AMIA 2003 Annual Symposium Proceedings*, pages 21–25. American Medical Informatics Association, 2003.

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11:171–234, 2010a.

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: Analysis and extensions. *Journal of Machine Learning Research*, 11:235–284, 2010b.

Ingo A Beinlich, H J Suermondt, R Martin Chavez, and Gregory F Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Second European Conference on Artificial Intelligence in Medicine*, pages 247–256, London, 1989. Springer-Verlag.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL `https://doi.org/10.1023/A:1010933404324`.

Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *SIGMOD'97 Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, volume 26, pages 265–276. ACM, 1997.

Facundo Bromberg and Dimitris Margaritis. Improving the reliability of causal discovery from small data sets using argumentation. *Journal of Machine Learning Research*, 10:301–340, 2009.

Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13:27–66, 2012. (Version: MIToolbox-2.0).

Chih-Chung Chang and Chih-Jen Lin. LibSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell, and Weiru Liu. Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137(1–2):43–90, 2002.

William G Cochran. Some methods for strengthening the common χ^2 tests. *Biometrics*, 10(4): 417–451, 1954.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory (Second Edition)*. John Wiley and Sons, 2006.

Noel Cressie and Timothy RC Read. Pearson's X^2 and the loglikelihood ratio statistic G^2: a comparative review. *International Statistical Review*, 57(1):19–43, 1989.

Rónán Daly, Qiang Shen, and Stuart Aitken. Learning Bayesian networks: Approaches and issues. *The Knowledge Engineering Review*, 26(2):99–157, 2011.

Luis M de Campos. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 7(Oct):2149–2187, 2006.

Shunkai Fu and Michel Desmarais. Local learning algorithm for Markov blanket discovery. In *AI 2007: Advances in Artificial Intelligence*, pages 68–79. Springer Berlin Heidelberg, 2007.

Shunkai Fu and Michel C Desmarais. Markov blanket based feature selection: a review of past decade. In *Proceedings of the World Congress on Engineering*, 2010.

Huixuan Gao. *Statistics Computation*. Peking University Press, Beijing, 2005.

Balakrishna Hosmane. Improved likelihood ratio test for multinomial goodness of fit. *Communications in Statistics-Theory and Methods*, 16(11):3185–3198, 1987.

Balakrishna S Hosmane. Smoothing of likelihood ratio statistic for equiprobable multinomial goodness-of-fit. *Annals of the Institute of Statistical Mathematics*, 42(1):133–147, 1990.

BS Hosmane. Improved likelihood ratio tests and Pearson chi-square tests for independence in two dimensional contingency tables. *Communications in Statistics-Theory and Methods*, 15(6): 1875–1888, 1986.

James Kennedy and Russell C Eberhart. Particle swarm optimization. In *Proceedings of 1995 IEEE International Conference on Neural Networks*, volume 4, pages 1942–1948, Perth, 1995.

James Kennedy and Russell C Eberhart. A discrete binary version of the particle swarm algorithm. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 5, pages 4104–4108, Orlando, 1997.

Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *Thirteen International Conference in Machine Learning*. Stanford InfoLab, 1996.

D.N. Lawley. A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*, 43(3–4):295–303, 1956.

Jan Lemeire. *Learning Causal Models of Multivariate Systems and the Value of it for the Performance Modeling of Computer Programs*. ASP/VUBPRESS/UPA, PhD thesis, 2007.

Jan Lemeire, Stijn Meganck, Francesco Cartella, and Tingting Liu. Conservative independence-based causal structure learning in absence of adjacency faithfulness. *International Journal of Approximate Reasoning*, 53(9):1305–1325, 2012.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. Technical Report CMU-CS-99-134, 1999.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. In *Advances in Neural Information Processing Systems*, volume 12, pages 505–511. Morgan Kaufmann, 2000.

Kevin P. Murphy. *Bayes Net Toolbox for Matlab*, 2007. (Version: FullBNT-1.0.7).

Richard E Neapolitan. *Learning Bayesian Networks*. Upper Saddle River: Prentice Hall, 2004.

Pekka Parviainen and Mikko Koivisto. Finding optimal Bayesian networks using precedence constraints. *Journal of Machine Learning Research*, 14(1):1387–1415, 2013.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann, 1988.

Jean-Philippe Pellet and André Elisseeff. Using Markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9:1295–1342, 2008.

Jose M Peña, Roland Nilsson, Johan Björkegren, and Jesper Tegnér. Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning*, 45(2): 211–232, 2007.

Joseph Ramsey, Jiji Zhang, and Peter L Spirtes. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-2006)*, pages 401–408, 2006.

Johannes Rauh, Nils Bertschinger, Eckehard Olbrich, and Jurgen Jost. Reconsidering unique information: Towards a multivariate information decomposition. In *2014 IEEE International Symposium on Information Theory (ISIT)*, pages 2232–2236. IEEE, 2014.

Federico Schlüter. A survey on independence-based Markov networks learning. *Artificial Intelligence Review*, 42:1069–1093, 2014.

Marco Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35:1–22, 2010.

Craig Silverstein, Sergey Brin, and Rajeev Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2(1):39–68, 1998.

Alexander Statnikov and Constantin F Aliferis. Analysis and computational dissection of molecular signature multiplicity. *PLoS Computational Biology*, 6(5):e1000790, 2010.

Alexander Statnikov, Nikita I. Lytkin, Jan Lemeire, and Constantin F Aliferis. Algorithms for discovery of multiple Markov boundaries. *Journal of Machine Learning Research*, 14(1):499–566, 2013.

Ioannis Tsamardinos and Constantin F Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Algorithms for large scale Markov blanket discovery. In *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 376–381, 2003.

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

Sandeep Yaramakala. *Fast Markov Blanket Discovery*. MS thesis, 2004.

Lianwen Zhang and Haipeng Guo. *Introduction to Bayesian Networks*. Science Press, Beijing, 2006.