Causal Discovery with Ancestral Graphs

Zhongyi Hu
Keble College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
September 2023

Abstract

Graphical models serve as a visual representation that captures the underlying conditional independence relationships within distributions, employing either directed or undirected graphs. In this thesis, we explore *maximal ancestral graphs* (MAGs), which is an extension to the conventional *directed acyclic graphs* (DAGs). While DAGs excel in illustrating causal relationships, they fail to capture all the conditional independences on the margin in the absence of latent confounders and selection bias. MAGs provide a more comprehensive depiction of complex dependencies by encompassing both direct causal connections and indirect influences stemming from latent variables and selection bias.

The scalability and accuracy of MAG learning algorithms have been some problems due to the complexity of the space of *Markov equivalence classes* (MECs) of MAGs and instability of scoring criteria. We first use the concept of heads, tails and parametrizing sets to characterize Markov equivalent MAGs. Then we study imsets of MAGs to address the above issues.

The framework of imsets (Studený, 2006) is an algebraic approach to represent conditional independences. Given the remarkable success of standard imsets within DAGs, where they efficiently represent MECs and offer reliable scoring criteria, we endeavor to extend this framework to MAGs. Through an exploration of 0-1 imsets defined by parametrizing sets, we show under which conditions does this extended 'standard imset' of MAGs define the correct model. Consequently, we refine the ordered local Markov property of MAGs (Richardson, 2003), demonstrating that the newly proposed *refined Markov property* can be constructed in polynomial time if we bound maximal head size.

Finally, we apply the above results to develop novel score-based learning algorithms for MAGs. To efficiently traverse between MECs of MAGs, we identify some important graphical features within MAGs whose independence models are subsets of others. Leveraging the imsets derived

from the refined Markov property, we establish a consistent scoring criterion, offering an alternative to BIC by relying solely on estimates of entropy over subsets of variables. Empirical experiments show promising results when compared to state-of-the-art algorithms.

Contents

List of Figures

List of Tables

Preface

This thesis studies maximal ancestral graphs (MAGs), which are a class of graphical models and strictly extend DAGs models. We try to fit it into the framework of imsets and our results can address several issues in existing graph learning algorithms.

Chapter 1 begins with detailed introduction to MAGs and imsets, then follows with formal definitions. The first main result is in Chapter 2, which gives a graphical characterization of MECs of MAGs. This non-parametric characterization uses the concept of parametrizing sets and results in polynomial time algorithms to

(i) verify equivalence between two MAGs, and

(ii) project an acyclic directed mixed graph (ADMGs) to a Markov equivalent MAG.

Then in Chapter 3, we present the following results:

(i) the formula of the 'standard' imset $u_{\mathcal{G}}$ using parametrizing sets, and its properties;

(ii) when the MAGs are *simple* (no head of size three or more), the imset does define the right model;

(iii) a proof that $\mathcal{I}_{u_{\mathcal{G}}} \subseteq \mathcal{I}_{\mathcal{G}}$ for all MAGs $\mathcal{G}$ (provided that $\mathcal{I}_{u_{\mathcal{G}}}$ is well-defined), where $\mathcal{I}_{u_{\mathcal{G}}}$ and $\mathcal{I}_{\mathcal{G}}$ are lists of conditional independences implied by the imset $u_{\mathcal{G}}$ or graph $\mathcal{G}$,

If $\mathcal{I}_{u_{\mathcal{G}}} \subsetneq \mathcal{I}_{\mathcal{G}}$, this means that when the imset we defined does not include all the conditional independences implied by the graph. See full definitions in Section 1.2. Then we introduce the idea of *power DAGs*[1], inspired by the decomposition of $u_{\mathcal{G}}$ of general MAGs. Using the power DAGs, we introduce a new Markov property, simpler than the local Markov property and can be constructed in polynomial time under mild assumptions on the corresponding graphs.

[1]The name of our power DAG is inspired by the *intrinsic power DAG* in Richardson et al. (2017), though we have developed this idea independently.

Later in Chapter 3, we focus on imsets of bidirected graphs and show that for a subclass of bidirected graphs, the imset defines the right model. Also we show that if some specific patterns are shown in subgraphs induced by ancestral sets, then either $\mathcal{I}_{u_\mathcal{G}}$ is not defined, or $\mathcal{I}_{u_\mathcal{G}} \neq \mathcal{I}_\mathcal{G}$. In Section 3.6, we will also report our experimental results on which graphs are perfectly Markovian ($\mathcal{I}_\mathcal{G} = \mathcal{I}_{u_\mathcal{G}}$) when $|V| \leq 7$.

Moreover, we present applications of the above results to MAG learning algorithms in Chapter 4. Finally, discussion is given in Chapter 5.

In this thesis, our main emphasis is on *directed* MAGs, specifically those without any undirected components. Although Section 2.5 extends some findings from previous Sections in Chapter 2 to MAGs incorporating undirected components, it's essential to note that the remaining portion of the thesis solely discusses directed MAGs.

Acknowledgement

I would like to extend my heartfelt gratitude to the individuals who have played pivotal roles in the completion of this DPhil thesis.

First and foremost, I am immensely thankful to my supervisor, Robin Evans, for his unwavering guidance, invaluable insights, and continuous support throughout this journey. His expertise and rigorous attitude towards research have been instrumental in shaping the direction of my research.

I am deeply indebted to my parents for their emotional and financial support. Their unending love, encouragement, and sacrifices have enabled me to pursue my academic aspirations. Their belief in my abilities has been a constant source of motivation.

I extend my appreciation to the other esteemed academic professionals who have contributed to my growth, including Peter Keevash for unveiling the wonderful world of graphs, Derek Goldrei, Colin Please and Dino Sejdinovic for providing advice on my application to a PhD, Geoff Nicholls and Frank Windmeijer for their comments on my confirmation of status. I also appreciate Bryan Andrews for the enlightening discussions and perspectives. Finally I thank my external examiner, Tom Claassen, a distinguished authority in the field of graphical models. His expertise and published papers have been a source of inspiration for some of my work. Moreover, his invaluable feedback have significantly shaped my thesis draft.

A special note of gratitude goes to ChatGPT, whose grammar checks and writing suggestions have significantly improved the clarity and coherence of my writing.

To my partner, Xiaoyun Wang, I owe a debt of gratitude for the unconditional emotional support, patience, and understanding. Her belief in me, even during the most challenging times, has been a driving force.

In closing, I acknowledge the collective contributions of each person mentioned above, which have collectively shaped the outcome of this thesis. While words may fall short of expressing the depth of my appreciation, please accept my heartfelt thanks for being integral parts of this journey.

Chapter 1

Preliminary

1.1 Graphical Models: MAGs

Maximal ancestral graphs (MAGs) (Richardson and Spirtes, 2002) are used to model distributions via *conditional independence* (CI) relations. They are an extension of *directed acyclic graphs* (DAGs) as MAGs remove the assumption of no latent confounders, and allow data arising from distributions with more general independence structure. These graphs have been proven to be useful in various scenarios, for example, to infer causal effects from observational data. Therefore learning the best graph associated with the data is an important task.

There are three classic types of learning algorithms for graphical models: there are constraint-based and score-based methods, and then hybrid methods that combine the first two approaches. The canonical constraint-based method for learning DAGs/MAGs is the PC/FCI algorithm (Spirtes et al., 2000). Briefly speaking, this type of method tests for conditional independences in the empirical distribution, and uses the results to reconstruct the graph. The problem of constraint-based methods is that when the group of variables is large, it is likely that a mistake in testing a conditional independence will be made; this error can be propagated through the algorithm, and the resulting graph will not reflect the true independence structure generating the data (see, e.g. Ramsey et al., 2006; Evans, 2020).

Score-based methods tend to be more robust provided an appropriate parametric model. The main idea is that they go through many graphs and select those with the highest score. There now exist several score-based methods for MAGs (Triantafillou and Tsamardinos, 2016; Rantanen et al., 2021; Chen et al., 2021; Claassen and Bucur, 2022). There are two main problems with these approaches. First, the score used for DAGs is generally the BIC score (Chickering, 2002; Jaakkola et al., 2010), which is also used by the above methods for MAGs. However, although methods for fitting Gaussian or discrete MAG models via maximum likelihood have been given

by Drton et al. (2009) and Evans and Richardson (2010, 2014) (therefore the corresponding BIC score can be obtained), MLEs are not available in closed-form, and therefore they need to be computed iteratively using a numerical method. Moreover, the factorisation of distributions in MAG models is complicated (Richardson, 2009), and the scores are only decomposable with respected to the components connected by bidirected paths, also known as districts or c-components. In contrast, for DAGs the MLE is available in closed form and the BIC score can be decomposed in terms of individual variables and their parent sets.

Another problem is how to reduce the number of graphs visited. Logically we could score every graph, but this is obviously hopelessly inefficient. Graphs that represent the same CI relations are said to be in the same *Markov equivalence class* (MEC). The concept of MECs is important because graphs in the same MEC have the same score and it is a waste of time to score graphs in the visited MEC again. Among the previous works mentioned, only Claassen and Bucur (2022) explore graphs in the space of MECs. However in the paper, they use the BIC score which is computationally inefficient for the reasons mentioned above. In principal their search procedure can be equipped with any consistent scoring criteria.

In this thesis, we explore MAGs and study the concept of heads and tails arising from factorization of MAG models (Richardson, 2009) and discrete parametrizing of MAG models (Evans and Richardson, 2010). We will see that the framework of imsets (Studený, 2006) combined with another representation of MECs (Hu and Evans, 2020), which Chapter 2 describes, provides a solution to address the above two issues at the same time.

1.2 Definitions

A *graph* $\mathcal{G}$ consists of a vertex set $\mathcal{V}$ and an edge set $\mathcal{E}$ of pairs of distinct vertices. We consider mixed graphs with two types of edge: *directed* ($\rightarrow$) and *bidirected* ($\leftrightarrow$). For an edge in $\mathcal{E}$ connecting vertices a and b, we say these two vertices are the *endpoints* of the edge and the two vertices are *adjacent* (if there is no edge between a and b, they are *nonadjacent*).

A *path* of length k is an alternating sequence of $k+1$ distinct vertices v_i, with an edge connecting v_i and v_{i+1} for $i = 0, \ldots, k-1$. Note the graphs we consider are all *simple*, so there is at most one edge between any pair of vertices. A path is *directed* if its edges are all directed and point from v_i to v_{i+1}. A *directed cycle* is a directed path of length k plus the edge $v_k \rightarrow v_0$, and a graph $\mathcal{G}$ is *acyclic* if it has

no directed cycle. A *graph* $\mathcal{G}$ is called an *acyclic directed mixed graph* (ADMG) if it is *acyclic* and contains only directed and bidirected edges.

For a vertex v in an ADMG $\mathcal{G}$, we define the following sets:

$$\begin{aligned}
\mathrm{pa}_{\mathcal{G}}(v) &= \{w : w \to v \text{ in } \mathcal{G}\} \\
\mathrm{sib}_{\mathcal{G}}(v) &= \{w : w \leftrightarrow v \text{ in } \mathcal{G}\} \\
\mathrm{an}_{\mathcal{G}}(v) &= \{w : w \to \cdots \to v \text{ in } \mathcal{G} \text{ or } w = v\} \\
\mathrm{de}_{\mathcal{G}}(v) &= \{w : v \to \cdots \to w \text{ in } \mathcal{G} \text{ or } w = v\} \\
\mathrm{dis}_{\mathcal{G}}(v) &= \{w : w \leftrightarrow \cdots \leftrightarrow v \text{ in } \mathcal{G} \text{ or } w = v\}.
\end{aligned}$$

They are known as the *parents*, *siblings*, *ancestors*, *descendants* and *district* of v, respectively. These operators are also defined disjunctively for a set of vertices $W \subseteq \mathcal{V}$. For example $\mathrm{pa}_{\mathcal{G}}(W) = \bigcup_{w \in W} \mathrm{pa}_{\mathcal{G}}(w)$. Vertices in the same district are connected by a bidirected path and this is an equivalence relation, so we can partition $\mathcal{V}$ and denote the *districts* of a *graph* $\mathcal{G}$ by $\mathcal{D}(\mathcal{G})$. We sometimes ignore the subscript if the graph we refer to is clear, for example $\mathrm{an}(v)$ instead of $\mathrm{an}_{\mathcal{G}}(v)$.

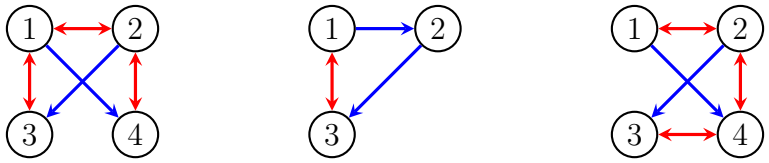
A topological ordering is a ordering on the vertices such that if $w \in \mathrm{an}_{\mathcal{G}}(v)$ then w preceeds v in the ordering. There might be several topological orderings for any single graph.

For an ADMG $\mathcal{G}$, given a subset $W \subseteq \mathcal{V}$, the *induced subgraph* $\mathcal{G}_W$ is defined as the graph with vertex set W and edges in $\mathcal{G}$ whose endpoints are both in W. Also for the district of a vertex v in an induced subgraph $\mathcal{G}_W$, we may denote it by $\mathrm{dis}_W(v) := \mathrm{dis}_{\mathcal{G}_W}(v)$.

Graphs are associated with conditional independence relations via a separation criterion; in the case of ADMGs, we use *m-separation*.

For a path π with vertices v_i, $0 \leq i \leq k$ we call v_0 and v_k the *endpoints* of π and any other vertices the *nonendpoints* of π. For a nonendpoint w in π, it is a *collider* if $? \!\!\to w \leftarrow \!\!?$ on π and a *noncollider* otherwise (an edge $? \!\!\to$ is either $\to$ or $\leftrightarrow$). For two vertices a, b and a set of vertices C ($a, b \notin C$) in $\mathcal{G}$ (C might be empty), a path π is *m-connecting* a, b given C if (i) a, b are endpoints of π, (ii) every noncollider is not in C and (iii) every collider is in $\mathrm{an}_{\mathcal{G}}(C)$. A *collider path* is a path where all the nonendpoints are colliders.

Definition 1.2.1. For three disjoint sets A, B and C (A, B are non-empty), A and B are *m-separated* by C in $\mathcal{G}$ if there is no m-connecting path between any $a \in A$ and any $b \in B$ given C. We denote m-separation by $A \perp_m B \mid C$.



(i) (ii) (iii)

Figure 1.1: (i) An ancestral graph that is not maximal. (ii) A maximal graph that is not ancestral. (iii) A maximal ancestral graph.

Definition 1.2.2. A distribution $P(X_V)$ is said to satisfy the *global Markov property* with respect to an ADMG $\mathcal{G}$ if whenever $A \perp_m B \mid C$ in $\mathcal{G}$, we have $X_A \perp\!\!\!\perp X_B \mid X_C$ under P.

An alternative Markov property that associates a distribution with a graph is called the *local Markov property* introduced by Richardson (2003). It is *equivalent* to the global Markov property in the sense that any conditional independence in the global Markov property can be deduced using semi-graphoids (Richardson, 2003) from the local Markov property. We will employ the local property later because it contains fewer conditional independence statements than the global property, and hence it is easier to show that a distribution is Markov with respect to a given graph. To define the local Markov property, we need a few more definitions.

Definition 1.2.3. If a set A satisfies $\mathrm{an}_{\mathcal{G}}(A) = A$, then it is an *ancestral set*.

If A is an ancestral set in an ADMG $\mathcal{G}$, and v is a vertex in A such that $\mathrm{ch}_{\mathcal{G}}(v) \cap A = \emptyset$, then the *Markov blanket* of v with respect to A is defined as:

$$\mathrm{mb}_{\mathcal{G}}(v, A) = \mathrm{pa}_{\mathcal{G}_A}(\mathrm{dis}_{\mathcal{G}_A}(v)) \cup \mathrm{dis}_{\mathcal{G}_A}(v) \setminus \{v\}.$$

Definition 1.2.4. A distribution $P(X_V)$ is said to satisfy the *local Markov property* with respect to an ADMG $\mathcal{G}$ if for any ancestral set A and a childless vertex v in A,

$$X_v \perp\!\!\!\perp X_{A \setminus (\mathrm{mb}(v,A) \cup \{v\})} \mid X_{\mathrm{mb}(v,A)} \quad [P].$$

1.2.1 MAGs

Definition 1.2.5. An ADMG $\mathcal{G}$ is called a *maximal ancestral graph* (MAG), if:

(i) for every pair of nonadjacent vertices a and b, there exists some set C such that a, b are m-separated given C in $\mathcal{G}$ (*maximality*);

(ii) for every $v \in \mathcal{V}$, $\mathrm{sib}_{\mathcal{G}}(v) \cap \mathrm{an}_{\mathcal{G}}(v) = \emptyset$ (*ancestrality*).

For example, the graph in Figure 1.1(i) is not maximal because 3 and 4 are not adjacent, but no subset of $\{1, 2\}$ will m-separate them. (ii) is not ancestral as 1 is a sibling of 3, which is also one of its descendants. (iii) is a MAG in which the only conditional independence is $X_1 \perp\!\!\!\perp X_3 \mid X_4$.

Definition 1.2.6. Two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ with the same vertex sets, are said to be *Markov equivalent* if any m-separation holds in $\mathcal{G}_1$ if and only if it holds in $\mathcal{G}_2$.

For example, Figure 1.2 (i) and (ii) are Markov equivalent.

For every ADMG $\mathcal{G}$, we can project it to a MAG $\mathcal{G}^m$ such that $\mathcal{G}$ is Markov equivalent to $\mathcal{G}^m$, and $\mathcal{G}^m$ preserves the ancestral relations in $\mathcal{G}$ (Richardson and Spirtes, 2002). Moreover, Hu and Evans (2020) show that the heads and tails defined below (and so the parametrizing sets) are preserved through the projection. Hence in this thesis, we will only consider MAGs.

1.2.2 Heads and Tails

A head is a subset of vertices with a corresponding tail. The concept of heads and tails originates from Richardson (2009), which provides a factorization theorem for the ADMGs. Intuitively, heads together with any subset of its tail are the subsets of vertices such that between any two vertices in the set, conditioning on the remaining vertices and any other vertex outside the head, they are always m-connected.

Definition 1.2.7. For a vertex set $W \subseteq \mathcal{V}$, we define the *barren subset* of W as:

$$\text{barren}_{\mathcal{G}}(W) = \{w \in W : \text{de}_{\mathcal{G}}(w) \cap W = \{w\}\}.$$

A vertex set H is called a *head* if:

(i) $\text{barren}_{\mathcal{G}}(H) = H$;

(ii) H is contained in a single district in $\mathcal{G}_{\text{an}(H)}$.

For an ADMG $\mathcal{G}$, we denote the set of all heads in $\mathcal{G}$ by $\mathcal{H}(\mathcal{G})$.

The *tail* of a head is defined as:

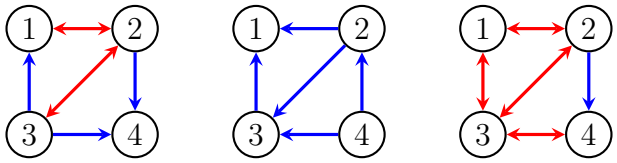
$$\text{tail}_{\mathcal{G}}(H) = (\text{dis}_{\text{an}(H)}(H) \setminus H) \cup \text{pa}_{\mathcal{G}}(\text{dis}_{\text{an}(H)}(H)).$$

The *parametrizing sets* of $\mathcal{G}$, denoted by $\mathcal{S}(\mathcal{G})$ are defined as:

$$\mathcal{S}(\mathcal{G}) = \{H \cup A : H \in \mathcal{H}(\mathcal{G}) \text{ and } \emptyset \subseteq A \subseteq \text{tail}_{\mathcal{G}}(H)\}.$$

We further define $\mathcal{S}_k(\mathcal{G})$ as:

$$\mathcal{S}_k(\mathcal{G}) = \{S \in \mathcal{S}(\mathcal{G}) : |S| = k\}$$



(i) (ii) (iii)

Figure 1.2: Three MAGs where (i) and (ii) are Markov equivalent but (iii) is not.

In particular, we are interested in:

$$\tilde{\mathcal{S}}_3(\mathcal{G}) = \{S \in \mathcal{S}_3(\mathcal{G}) : \text{there are 1 or 2 adjacencies among the vertices in } S\}.$$

We write $\mathcal{S}, \mathcal{S}_k, \tilde{\mathcal{S}}_3$ if the graph $\mathcal{G}$ we are referring to is clear.

Remark 1. Note that for a head H and each $h \in H$, the set $(H \setminus \{h\}) \cup \text{tail}_{\mathcal{G}}(H)$ is the Markov blanket of h within $\text{an}_{\mathcal{G}}(H)$.

We are not considering any singleton sets in $\mathcal{S}_k(\mathcal{G})$ or $\tilde{\mathcal{S}}_3(\mathcal{G})$; these are just all vertices because $\{v\}$ is trivially a head. For a MAG $\mathcal{G}$, a pair of vertices are in $\mathcal{S}(\mathcal{G})$ if and only if the two vertices are adjacent, which is easy to prove).

We give an example to illustrate what the sets defined above are. Consider the three MAGs in Figure 1.2, Table 1.1 lists their heads and tails, Table 1.2 lists their parametrizing sets $\mathcal{S}$ and Table 1.3 lists their $\mathcal{S}_3$ and $\tilde{\mathcal{S}}_3$.

Table 1.1: Heads and tails of graphs in Figure 1.2

Figure	heads	tails	Figure	heads	tails
1.2(i)	1	3	1.2(iii)	1	$\emptyset$
	2	$\emptyset$		2	$\emptyset$
	3	$\emptyset$		3	$\emptyset$
	4	2,3		4	2
	1,2	3		1,2	$\emptyset$
	2,3	$\emptyset$		1,3	$\emptyset$
1.2(ii)	1	2,3		2,3	$\emptyset$
	2	4		3,4	2
	3	2,4		1,2,3	$\emptyset$
	4	$\emptyset$		1,3,4	2

In Figure 1.2, (i) is Markov equivalent to (ii) and they also have the same parametrizing sets; however, (iii) has a different parametrizing set and is not Markov equivalent to either (i) or (ii). In Figure 1.2(i) and (ii), $1 \perp_m 4 \mid 2, 3$ is the only

Table 1.2: Parametrizing set of graphs in Figure 1.2

Figure	parametrizing sets	missing sets
1.2(i)(ii)	$\{1\},\{2\},\{3\},\{4\}$ $\{1,2\},\{1,3\},\{2,3\}$ $\{2,4\},\{3,4\}$ $\{1,2,3\},\{2,3,4\}$	$\{1,4\}$ $\{1,2,4\}$ $\{1,3,4\}$ $\{1,2,3,4\}$
1.2(iii)	$\{1\},\{2\},\{3\},\{4\}$ $\{1,2\},\{1,3\},\{2,3\}$ $\{2,4\},\{3,4\}$ $\{1,2,3\},\{1,3,4\},\{2,3,4\}$ $\{1,2,3,4\}$	$\{1,4\}$ $\{1,2,4\}$

Table 1.3: $\mathcal{S}_3$ and $\tilde{\mathcal{S}}_3$ graphs in Figure 1.2

Figure	$\mathcal{S}_3$	$\tilde{\mathcal{S}}_3$
1.2(i)(ii)	$\{1,2\},\{1,3\},\{2,3\}$ $\{2,4\},\{3,4\}$ $\{1,2,3\},\{2,3,4\}$	$\{1,2\},\{1,3\}$ $\{2,3\},\{2,4\}$ $\{3,4\}$
1.2(iii)	$\{1,2\},\{1,3\},\{2,3\}$ $\{2,4\},\{3,4\}$ $\{1,2,3\},\{2,3,4\}$ $\{1,3,4\}$	$\{1,2\},\{1,3\}$ $\{2,3\},\{2,4\}$ $\{3,4\}$ $\{1,3,4\}$

m-separation while Figure 1.2(iii) encodes $1 \perp_m 4 \mid 2$. Note that these conditional independences correspond precisely to the missing sets which are in the form $\{a,b\} \cup C'$ where $a \perp_m b \mid C$ and $C' \subseteq C$. Thus it is reasonable to conjecture that equivalent graphs should have the same parametrizing sets. It turns out that not only is this true, but in fact equivalence conditions can be refined even further and it is sufficient to consider $\mathcal{S}_3$ or $\tilde{\mathcal{S}}_3$.

For a conditional independence $I = \{X_A \perp\!\!\!\perp X_B \,|\, X_C\}$, let

$$\bar{\mathcal{S}}(I) = \{A' \cup B' \cup C' : \emptyset \subset A' \subseteq A, \emptyset \subset B' \subseteq B, \emptyset \subseteq C' \subseteq C\}.$$

One can think of $\bar{\mathcal{S}}(I)$ as the constrained set 'explained' by I, and we say $\bar{\mathcal{S}}(I)$ are the sets associated with the conditional independence I. Proposition 2.2.2 proves that a set S is *not* in $\mathcal{S}(\mathcal{G})$ if and only if it is associated with an independence entailed by the graph.

Chapter 2

Parametrizing sets of MAGs

In this chapter, we give a graphical condition for when two MAGs are Markov equivalent. There have been several graphical characterizations that give necessary and sufficient conditions for when two MAGs are equivalent. Among those criteria, Ali et al. (2009) firstly provide a polynomial time algorithm to verify Markov equivalence. More recently Claassen and Bucur (2022) and Wienöbst et al. (2022) show that equivalence between MAGs can be verified in $O(n^3)$ in general and even $O(nd^2)$ for sparse graphs with maximal degree d. Zhao et al. (2005) characterize MAGs by *minimal collider paths* (MCPs). The criterion of Spirtes and Richardson (1997) uses *discriminating paths*, which we will define in Section 3 (we will employ them in our proofs). This chapter gives a new characterization and it lead to a faster algorithm to test equivalence compared to existing ones. Also we show a similar equivalence criterion for wider classes of acyclic graphs, ADMGs.

In Section 2.1, we present some past work that are employed in our proof. The main result for Markov equivalence between MAGs is in Section 2.2, which is then extended to ADMGs in Section 2.3. Algorithms for checking Markov equivalence and analysis of their complexity are given in Section 2.4. We further extend our criterion to summary graphs (Wermuth, 2011) in Section 2.5. We also show how to deduce conditional independences (not all of them) from parametrizing sets in Section 2.6. In the final section, we demonstrate how to construct the *partial ancestral graphs* (PAGs) given the parametrizing sets.

2.1 Previous Work

The first theorem on Markov equivalence of MAGs is from Spirtes and Richardson (1997).

Theorem 2.1.1. *Two MAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if (i) $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same adjacencies, (ii) $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same unshielded colliders*

and (iii) if π forms a discriminating path for b in $\mathcal{G}_1$ and $\mathcal{G}_2$, then b is a collider on the path π in $\mathcal{G}_1$ if and only it is a collider on the path π in $\mathcal{G}_2$.

For x and y nonadjacent, a *discriminating path* $\pi = \langle x, q_1, \ldots, q_m, b, y\rangle$, $m \geq 1$ for b, is a subgraph comprised of a collection of paths:

$$x \,?\!\!\rightarrow q_1 \leftrightarrow \cdots \leftrightarrow q_i \rightarrow y, \qquad 1 \leq i \leq m;$$
$$x \,?\!\!\rightarrow q_1 \leftrightarrow \cdots \leftrightarrow q_m \leftarrow\!\!? \, b \,?\!\!\rightarrow y.$$

For example, $\langle 1, 2, 3, 4\rangle$ forms a discriminating path for 3 in both Figure 1.2(i) and (iii), but not (ii). The vertex 3 is a collider on the path in (iii) but not (i), so (i) and (iii) are not equivalent; however (i) and (ii) are equivalent. In general, the cost of identifying all the discriminating paths is not polynomial in the number of vertices and edges. However, we will make use of Theorem 2.1.1 in later proofs.

2.2 Markov Equivalence of MAGs

We now present the main result of this chapter, which is also published in Hu and Evans (2020).

Theorem 2.2.1. *Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two MAGs. Then $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if $\mathcal{S}(\mathcal{G}_1) = \mathcal{S}(\mathcal{G}_2)$.*

Theorem 2.2.1 already provides a method to find equivalence between two MAGs by searching all the heads and corresponding tails, however, the number of heads is not polynomial in the size of the graph.

Corollary 2.2.1.1. *Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two MAGs. Then $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if $\mathcal{S}_3(\mathcal{G}_1) = \mathcal{S}_3(\mathcal{G}_2)$. This in turn occurs if and only if $\tilde{\mathcal{S}}_3(\mathcal{G}_1) = \tilde{\mathcal{S}}_3(\mathcal{G}_2)$.*

The motivation for defining $\tilde{\mathcal{S}}_3(\mathcal{G})$ is that we cannot obtain the same complexity if we allow triangles to be included, as in $\mathcal{S}_3$. To see this, consider a complete bidirected graph with e edges: this will require $O(e^3)$ operations to list all the triangles (which are all heads). Note we do not care about triples with three or zero adjacencies. Theorem 2.1.1 tells us that apart from adjacencies between pairs of vertices, and unshielded triples which lack one adjacency, we only need to find that for a discriminating path $\pi = \langle x, q_1, \ldots, q_m, b, y\rangle$, whether b is a collider on the path or not. Later we will show that $\{x, b, y\} \in \mathcal{S}(\mathcal{G})$ if and only if b is a collider on π, and note that b, y are adjacent but x, y are not.

Corollary 2.2.1.1 is particularly important for identifying Markov equivalence. It not only allows the algorithm to run in polynomial time as we only need to check heads with size at most 3, but also accelerates it further as we do not need to find triples with full adjacencies or no adjacencies, nor to store lots of triangles from the dense part of the graph.

To prove Theorem 2.2.1 and Corollary 2.2.1.1, we first prove the following propositions.

Proposition 2.2.2. *Let $\mathcal{G}$ be a MAG with vertex set $\mathcal{V}$. For a set $W \subseteq \mathcal{V}$, $W \notin \mathcal{S}(\mathcal{G})$ if and only if there are two vertices a, b in W such that we can m-separate them by a set C such that $a, b \notin C$ with $W \subseteq C \cup \{a, b\}$.*

Proof. We prove an equivalent statement of this proposition, that is: $W \in \mathcal{S}(\mathcal{G})$ if and only if for any two vertices a, b in W we cannot m-separate them by a set C such that $a, b \notin C$ with $W \subseteq C \cup \{a, b\}$.

To prove $\Rightarrow$: if $W \in \mathcal{S}(\mathcal{G})$, then there is a nonempty subset $W' \subseteq W$ such that W' is a head and $W \subseteq W' \cup \text{tail}(W')$. Because $\text{tail}(W') \subseteq \text{an}(W')$, we have $W' \cup \text{tail}(W') \subseteq \text{an}(W')$. By definition of the heads and tails, any two vertices a, b in $W \subseteq W' \cup \text{tail}(W')$ are connected by a collider path π where all the colliders are in $\text{an}(W') \subseteq \text{an}(C \cup \{a, b\})$. Let d_i, $1 \leq i \leq n$ be intermediate vertices in the path. Now if all of d_i are ancestors of C then this path m-connects a and b. So some of d_i are only ancestors of a, b.

Suppose there exists some $d_i \in \text{an}_{\mathcal{G}}(a) \backslash \text{an}_{\mathcal{G}}(C)$, let d_j be the furthest one on path π from a, so there exists a directed path $\pi' : a \leftarrow \cdots \leftarrow d_j$ such that none of vertices in π' after a is an ancestor of C and hence not in C. If all d_k after d_j belong to $\text{an}_{\mathcal{G}}(C)$ then we find a m-connecting path between a and b: $a \leftarrow \cdots \leftarrow d_j \leftrightarrow \cdots \leftrightarrow d_n \leftarrow\!? \, b$. If not, let d_m be the first one after d_j such that $d_m \in \text{an}_{\mathcal{G}}(b) \setminus \text{an}_{\mathcal{G}}(C)$ then again we find a m-connecting path between a and b: $a \leftarrow \cdots \leftarrow d_j \leftrightarrow \cdots \leftrightarrow d_m \rightarrow \cdots \rightarrow b$.

If all $d_i \notin \text{an}_{\mathcal{G}}(C)$ are ancestors of b then let d_j be the closest one to a in path π which also leads to a m-connecting path between a and b: $a \,?\!\rightarrow d_1 \leftrightarrow \cdots \leftrightarrow d_j \rightarrow \cdots \rightarrow b$. Hence in all cases any a, b in W are not m-separated given any $C \supseteq W \setminus \{a, b\}$.

To prove $\Leftarrow$: define $W' = \text{barren}(W)$. We claim that it is a head. Suppose it is not a head, by the definitions of a barren set and a head, W' does not lie in a single district in $\mathcal{G}_{\text{an}(W')}$. Let $D_i \subset W'$ index bidirected-connected components of W' in $\text{an}_{\mathcal{G}}(W')$ where $1 \leq i \leq m$. Clearly by assumption $m > 1$, and now consider D_1 and D_2. For any edge in $\mathcal{G}_{\text{an}(W')}$ which has an endpoint $a \in W'$, it is of the form $a \leftarrow\!?$ by definition of a barren set, so if there is a collider path between D_1 and D_2,

it would be a bidirected path which is a contradiction to the definition of D_1 and D_2. This means that any path in $\text{an}_{\mathcal{G}}(W')$ between D_1 and D_2 contains at least one non-collider which is not in W' and hence it is in $\text{an}_{\mathcal{G}}(W') \setminus W'$. Thus for any two vertices in D_1 and D_2, given $\text{an}_{\mathcal{G}}(W') \setminus W'$, they are m-separated in $\text{an}_{\mathcal{G}}(W')$. Since $\text{an}_{\mathcal{G}}(W')$ is ancestral, the m-separation also holds in the whole graph. Thus W' is a head.

By Remark 4.14 in Evans and Richardson (2014), for any head H we have $H \perp_m \text{an}_{\mathcal{G}}(H) \setminus (H \cup \text{tail}(H)) \mid \text{tail}(H)$. Thus if $(W \setminus W')$ is not in $\text{tail}(W')$, we can m-separate a vertex in $(W \setminus W') \setminus \text{tail}(W')$ and a vertex in W' given the remaining vertices in $\text{an}_{\mathcal{G}}(W')$, which is a contradiction. □

Proposition 2.2.3. *For a MAG $\mathcal{G}$, we have (i) any two vertices a and b are adjacent in $\mathcal{G}$ if and only if $\{a, b\} \in \mathcal{S}(\mathcal{G})$; (ii) for any unshielded triple (a, b, c) in $\mathcal{G}$, $\{a, b, c\} \in \mathcal{S}(\mathcal{G})$ if and only if b is a collider on the triple (a, b, c); (iii) if π forms a discriminating path for b with two end vertices x and y in $\mathcal{G}$ then $\{x, b, y\} \in \mathcal{S}(\mathcal{G})$ if and only if b is a collider on the path π.*

Proof. For (i), by maximality, any two vertices a and b are adjacent in a MAG if and only if we can not m-separate them by a set C, hence by Proposition 2.2.2 if and only if $\{a, b\} \in \mathcal{S}(\mathcal{G})$.

For (ii), the only nonadjacent pair of vertices are a, c, for any set C that m-seperates them, $b \notin C$ if and only if b is a collider on the triple (a, b, c), hence by Proposition 2.2.2 if and only if $\{a, b, c\} \in \mathcal{S}(\mathcal{G})$.

For (iii), if x, b are not adjacent, then for any set that m-separates them, y is not in the set, as the path $x \,?\!\!\rightarrow q_1 \leftrightarrow \cdots \leftrightarrow q_m \leftarrow\!\!? \, b$ would be m-connecting x and b. Since x, y are not adjacent, there exists some set C such that $x \perp_m y \mid C$. From page 11 in Ali et al. (2009), we know that for any such C, $q_i \in C$ for all $i \leq n$ and b is a collider if and only if $b \notin C$, hence by Proposition 2.2.2 if and only if $\{x, b, y\} \in \mathcal{S}(\mathcal{G})$. □

Now we are able to prove Theorem 2.2.1 and Corollary 2.2.1.1

Proof of Theorem 2.2.1. ($\Rightarrow$) Proposition 2.2.2 ensures that missing sets in $\mathcal{S}(\mathcal{G})$ are only due to m-separations in graphs. But as Markov equivalence is characterized by m-separations, $\mathcal{S}(\mathcal{G}_1)$ and $\mathcal{S}(\mathcal{G}_2)$ in two equivalent MAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are the same. ($\Leftarrow$) Proposition 2.2.3 implies that any violation of conditions in Theorem 2.1.1 result in different $\mathcal{S}(\mathcal{G}_1)$ and $\mathcal{S}(\mathcal{G}_2)$. Hence if $\mathcal{S}(\mathcal{G}_1) = \mathcal{S}(\mathcal{G}_2)$, $\mathcal{G}_1$ is Markov equivalent to $\mathcal{G}_2$. □

Proof of Corollary 2.2.1.1. ($\Rightarrow$) This follows from Theorem 2.2.1 and the fact that Markov equivalent MAGs have the same adjacencies. ($\Leftarrow$) This follows from the fact that in the proof the 'if' part of Theorem 2.2.1, we only consider sets in $\tilde{\mathcal{S}}_3(\mathcal{G}_1)$ and $\tilde{\mathcal{S}}_3(\mathcal{G}_2)$. □

Frydenberg (1990) gives conditions for when two DAGs are equivalent, i.e. if and only if they have the same adjacencies and unshielded colliders. DAGs are a subclass of MAGs so Corollary 2.2.1.1 also applies to them. When $\mathcal{G}$ is just a DAG, $\tilde{\mathcal{S}}_3(\mathcal{G})$ (and indeed $\mathcal{S}_3(\mathcal{G})$) contains the exact information of $\mathcal{G}$'s adjacencies and unshielded colliders. By Proposition 2.2.3, $\{a, b\} \in \tilde{\mathcal{S}}_3(\mathcal{G})$ if and only if a, b are adjacent. And a triple is in $\tilde{\mathcal{S}}_3(\mathcal{G})$ if and only if it is an unshielded collider; this is because in DAGs, heads are precisely the individual vertices, and the corresponding tails are their parent sets.

2.3 ADMGs

Richardson and Spirtes (2002) give a projection that projects a DAG $\mathcal{G}$ with latent variables L to a Markov equivalent MAG $\mathcal{G}^m$: (i) every pair of vertices $a, b \in \mathcal{V}$ in $\mathcal{G}$ that are connected by an *inducing path* becomes adjacent in $\mathcal{G}^m$; (ii) an edge connecting a, b in $\mathcal{G}^m$ is oriented as follows: if $a \in \text{an}_{\mathcal{G}}(b)$ then $a \to b$; if $b \in \text{an}_{\mathcal{G}}(a)$ then $b \to a$; if neither is the case, then $a \leftrightarrow b$. An *inducing path* between a, b is a path such that every collider in the path is in $\text{an}(\{a, b\})$, and every noncollider is in L. Note if we already have an ADMG $\mathcal{G}$, we can apply the projection to $\mathcal{G}$ with no latent variable to construct the corresponding $\mathcal{G}^m$, so an inducing path in this case is just a collider path with every collider in $\text{an}(\{a, b\})$. In addition, the projection preserves ancestral relations from the original graph.

To extend previous theorems to $\mathcal{G}$ we need following lemmas to link $\mathcal{G}$ and $\mathcal{G}^m$.

Lemma 2.3.1. *If v, w are connected by a collider path π_1 in an ADMG $\mathcal{G}$ then they are connected by a collider path π_2 in $\mathcal{G}^m$ where π_2 uses a subset of the internal vertices of π_1. Also, if π_1 starts with $v \to$, so does π_2.*

Proof. Any adjacent pair in $\mathcal{G}$ is also adjacent in $\mathcal{G}^m$ as any edge is a trivial collider path. So the path π_1 is still present in $\mathcal{G}^m$ however it may not be a collider path (if it is then we are done) and we aim to find a collider path π_2.

Suppose a is an internal vertex and is a noncollider in π_1 in $\mathcal{G}^m$ where $a \leftrightarrow b$ in $\mathcal{G}$ is changed to $a \to b$ in $\mathcal{G}^m$. This is because $a \in \text{an}_{\mathcal{G}}(b)$. Consider the vertex c on the other side of a, suppose it is $c \leftrightarrow a$ in $\mathcal{G}$. Then $b \leftrightarrow a \leftrightarrow c$ is a collider path where $a \in \text{an}_{\mathcal{G}}(\{b, c\})$ so b, c becomes adjacent in $\mathcal{G}^m$ and we can remove a from the

path. If $c \to a$, i.e. c is one of end vertices, then in the projected graph we have $c \to a$. We can do this repeatedly until it terminates and the final path is a collider path in $\mathcal{G}^m$ that connects v, w. □

Lemma 2.3.1 is in analogue to Lemma 23 in Shpitser et al. (2018). Now we show heads and tails are preserved through the projection.

Proposition 2.3.2. *If $\mathcal{G}$ is an ADMG, $\mathcal{H}(\mathcal{G}) = \mathcal{H}(\mathcal{G}^m)$ and for every $H \in \mathcal{H}(\mathcal{G})$, $\text{tail}_{\mathcal{G}}(H) = \text{tail}_{\mathcal{G}^m}(H)$.*

Proof. Suppose H is a head in $\mathcal{G}$. Then it is bidirected-connected in $\mathcal{G}_{\text{an}(H)}$, so by Lemma 2.3.1 each bidirected path connecting vertices in H is preserved as a collider path in $\mathcal{G}^m_{\text{an}(H)}$. Further as the projection preserves ancestral relation and $H = \text{barren}(\text{an}(H))$, each path is bidirected. Hence any head H in $\mathcal{G}$ is a head in $\mathcal{G}^m$. By similar argument, we can see that for a head H in $\mathcal{G}$, any $w \in \text{tail}_{\mathcal{G}}(H)$ is in $\text{tail}_{\mathcal{G}^m}(H)$.

Suppose H is a head in $\mathcal{G}^m$ so it is bidirected-connected in $\text{an}(H)$ in $\mathcal{G}^m$. But each bidirected edge in $\mathcal{G}^m$ corresponds to a collider path in $\mathcal{G}$ with intermediate colliders in ancestors of endpoints; hence as the projection preserves ancestral relations, the path is bidirected. Therefore H is also a head in $\mathcal{G}$. Note in general for any $v \leftrightarrow w$ in $\mathcal{G}^m$, there is a bidirected path between them in $\mathcal{G}$.

Let $z \in \text{tail}_{\mathcal{G}^m}(H)$ so there is a collider path π between z and $h \in H$ in $\mathcal{G}^m$ ending $\cdots \leftrightarrow h$. We know every bidirected edge in the path π corresponds to a bidirected path in $\text{an}(H)$ in $\mathcal{G}$. If the path π begins with $z \leftrightarrow$ then z is bidirected-connected to h in $\text{an}(H)$ so $z \in \text{tail}_{\mathcal{G}}(H)$. If the path π begins with $z \to w_1$ then in $\mathcal{G}$ we have a collider path between z and w_1 in $\text{an}(H)$, which ends with $\leftrightarrow w_1$. Thus z is also in $\text{tail}_{\mathcal{G}}(H)$. □

Definitions of heads and tails are closely related to the projection of ADMGs. The next lemma allows us to project an ADMG to a Markov equivalent MAG in polynomial time. The algorithm is shown in next section. Let $\mathcal{G}$ be a ADMG and $\mathcal{G}^m$ be its projected MAG.

Lemma 2.3.3. *Let v, w be two vertices then (i) $v \to w$ in $\mathcal{G}^m$ if and only if $v \in \text{tail}_{\mathcal{G}}(w)$ and (ii) $v \leftrightarrow w$ in $\mathcal{G}^m$ if and only if $\{v, w\} \in \mathcal{H}(\mathcal{G})$.*

Proof. For (i), if $v \to w$ in $\mathcal{G}^m$ then $v \in \text{an}_{\mathcal{G}}(w)$ and in $\mathcal{G}$ there is an inducing path between v and w (a collider path). If $v \to w$ in $\mathcal{G}$ then we are done. Otherwise any intermediate vertex on the path is in $\text{an}_{\mathcal{G}}(\{v, w\}) = \text{an}_{\mathcal{G}}(w)$ hence $v \,?\!\to \ldots \leftrightarrow w$. Therefore $v \in \text{tail}_{\mathcal{G}}(w)$. Conversely, $v \in \text{tail}_{\mathcal{G}}(w)$ implies that $v \in \text{an}_{\mathcal{G}}(w)$ and there

is a collider path between v and w with any intermediate vertex in $\text{an}_{\mathcal{G}}(w)$ hence the path is an inducing path and $v \to w$ in $\mathcal{G}^m$.

For (ii), if $v \leftrightarrow w$ in $\mathcal{G}^m$ then there is an inducing path between v and w (a collider path) in $\mathcal{G}$ and v, w are not ancestors to each other. Also any intermediate vertex on the path is in $\text{an}_{\mathcal{G}}(\{v, w\})$ which suggests that the path is a bidirected path. Therefore, $\{v, w\}$ forms a head. On the other hand, if $\{v, w\}$ is a head in $\mathcal{G}$ then they are not ancestors to each other and there is a bidirected path between them with any intermediate vertex in $\text{an}_{\mathcal{G}}(\{v, w\})$ so this path is an inducing path and $v \leftrightarrow w$ in $\mathcal{G}^m$. □

Since there is at most one edge between any two vertices in a MAG, if we know the tails of every vertex in $\mathcal{G}^d$ and every head of size 2, this is sufficient to construct $\mathcal{G}^m$.

Consider Figure 1.2.5(i), this is an ADMG but not a MAG. Tails of 1, 2, 3, 4 are $\emptyset$, $\emptyset$, $\{2\}$, $\{1\}$, respectively. Heads of size 2 are $\{1, 2\}$, $\{1, 3\}$, $\{2, 4\}$, $\{3, 4\}$, hence a Markov equivalent MAG of Figure 1.2.5(i) preserves all the original edges and adds one edge $3 \leftrightarrow 4$.

2.3.1 Markov Equivalence of ADMGs

We now show that Theorem 2.2.1 and Corollary 2.2.1.1 can be extended to ADMGs. Note that in general two Markov equivalent ADMGs do not necessarily have the same adjacencies defined with respect to edges; thus we need to redefine adjacencies in terms of m-separations.

Definition 2.3.1. For a ADMG $\mathcal{G}$ and two vertices v, w in $\mathcal{G}$, v and w are *adjacent* if and only if there is no set C such that $v \perp_m w \mid C$ with $v, w \notin C$.

Two vertices that are connected by an edge are clearly adjacent, we are excluding pairs that do not share any edges and yet have no conditional independence. In maximal graphs, these two definitions are equivalent.

Theorem 2.3.4. *For two ADMGs $\mathcal{G}_1$ and $\mathcal{G}_2$, they are Markov equivalent if and only $\mathcal{S}(\mathcal{G}_1) = \mathcal{S}(\mathcal{G}_2)$.*

Proof. This follows from Proposition 2.3.2 and Theorem 2.2.1. □

Corollary 2.3.4.1. *Two ADMGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if $\mathcal{S}_3(\mathcal{G}_1) = \mathcal{S}_3(\mathcal{G}_2)$, and this occurs if and only if $\tilde{\mathcal{S}}_3(\mathcal{G}_1) = \tilde{\mathcal{S}}_3(\mathcal{G}_2)$.*

Proof. By Proposition 2.3.2, $\mathcal{S}_3$ are preserved in $(\mathcal{G}_1)^m$ and $(\mathcal{G}_2)^m$, and with the new definition of adjacencies, the outputs of $\tilde{\mathcal{S}}_3$ are also preserved. Hence the statement follows from Corollary 2.2.1.1. □

2.4 Algorithm

In this section, n, e denote number of vertices and total edges, respectively.

Input: A MAG $\mathcal{G}(\mathcal{V}, \mathcal{E})$
Result: $\tilde{\mathcal{S}}_3(\mathcal{G})$

$S \leftarrow \emptyset$;
for *each* $v \in \mathcal{V}$ **do**
  obtain $\text{an}_{\mathcal{G}}(v) = \{v\} \cup \text{an}_{\mathcal{G}}(\text{pa}_{\mathcal{G}}(v))$;
  for *each* $w \in \text{pa}_{\mathcal{G}}(v)$ **do**
    $S \leftarrow S \cup \{v, w\}$;
  end
  for *each* $z, w \in \text{pa}_{\mathcal{G}}(v)$ *with* $z \neq w$ *and* z *not adjacent to* w **do**
    $S \leftarrow S \cup \{v, w, z\}$;
  end
end
for *each* $v \leftrightarrow w$ **do**
  $S \leftarrow S \cup \{v, w\}$;
  $\text{tail}(\{v, w\}) \leftarrow \text{dis}_{\text{an}(\{v,w\})}(v) \cup \text{pa}_{\mathcal{G}}(\text{dis}_{\text{an}(\{v,w\})}(v)) \setminus \{v, w\}$;
  for *each* $z \in \text{tail}(\{v, w\})$ *with* z *not adjacent to both* v *and* w **do**
    $S \leftarrow S \cup \{v, w, z\}$;
  end
  for *each* $z \in \text{sib}_{\mathcal{G}}(\text{an}_{\mathcal{G}}(\{v, w\})) \cap \text{dis}_{\mathcal{G}}(v) \setminus (\text{an}_{\mathcal{G}}(\{v, w\}) \cup \text{de}_{\mathcal{G}}(\{v, w\}))$ *and not adjacent to* v, w **do**
    obtain $\text{dis}_{\text{an}(\{v,w,z\})}(v)$;
    if $z \in \text{dis}_{\text{an}(\{v,w,z\})}(v)$ **then**
      $S \leftarrow S \cup \{v, w, z\}$;
    end
  end
end
return S;

Algorithm 1: Obtain $\tilde{\mathcal{S}}_3(\mathcal{G})$ for a MAG $\mathcal{G}$

2.4.1 Complexity of algorithms for MAGs

We assume that $n = O(e)$, since otherwise the graph will be disconnected. Firstly, we propose an algorithm to identify $\tilde{\mathcal{S}}_3(\mathcal{G})$ of a given MAG $\mathcal{G}$ and show that it runs in polynomial time ($O(ne^2)$). To test equivalence of two MAGs, it is sufficient to compare their $\tilde{\mathcal{S}}_3$, by Corollary 2.2.1.1. Vertices are assumed to be in topological order. If not, this can be achieved with an $O(n+e)$ sort. We assume we have access to $\text{pa}_{\mathcal{G}}(v)$ and $\text{sib}_{\mathcal{G}}(v)$ for each $v \in \mathcal{V}$.

Let $A_1(\mathcal{G})$ denote the output of Algorithm 1 when applied to a MAG, $\mathcal{G}$. We will show that for a MAG $\mathcal{G}$, $A_1(\mathcal{G}) = \tilde{\mathcal{S}}_3(\mathcal{G})$

Input: An ADMG $\mathcal{G}(\mathcal{V}, \mathcal{E})$
Result: A Markov equivalent MAG $\mathcal{G}^m(\mathcal{V}, \mathcal{E}^m)$
Start with $\mathcal{G}^m$ that has the same vertices as $\mathcal{G}$ but no adjacencies;
for *each* $v \in \mathcal{V}$ **do**
  obtain $\text{an}_{\mathcal{G}}(v) = \{v\} \cup \text{an}_{\mathcal{G}}(\text{pa}_{\mathcal{G}}(v))$;
  $\text{tail}(v) = \text{dis}_{\text{an}(v)}(v) \cup \text{pa}_{\mathcal{G}}(\text{dis}_{\text{an}(v)}(v)) \setminus \{v\}$;
  add $w \to v \in \mathcal{E}^m$ for each $w \in \text{tail}(v)$;
end
for *each* $v, w \in \mathcal{V}$ *with no ancestral relation and in the same district* **do**
  obtain $\text{dis}_{\text{an}(\{v,w\})}(v)$;
  if $w \in \text{dis}_{\text{an}(\{v,w\})}(v)$ **then**
    add $v \leftrightarrow w \in \mathcal{E}^m$;
  end
end
return $\mathcal{G}^m$;

Algorithm 2: Obtain a MAG $\mathcal{G}^m$ for an ADMG $\mathcal{G}$

2.4.2 Proof that Algorithm 1 outputs $\tilde{\mathcal{S}}_3$

Let $A_1(\mathcal{G})$ be the output of Algorithm 1 and $A_1'(\mathcal{G})$ be the output of Algorithm 1 without checking adjacencies in lines 7, 14 and 17. We also define the following sets for a MAG $\mathcal{G}$:

$$
\begin{aligned}
H_1(\mathcal{G}) &= \{\{v, w, z\} : v \in \mathcal{V} \text{ and } w, z \in \text{pa}_{\mathcal{G}}(v)\} \\
H_2(\mathcal{G}) &= \{\{v, w, z\} : v \leftrightarrow w, z \in \text{tail}(\{v, w\})\} \\
H_3^a(\mathcal{G}) &= \text{all heads of size 3 with some adjacencies} \\
H_3^n(\mathcal{G}) &= \text{all heads of size 3 with no adjacencies} \\
H_3(\mathcal{G}) &= \text{all heads of size 3} = H_3^a(\mathcal{G}) \cup H_3^n(\mathcal{G}) \\
\hat{\mathcal{S}}_3(\mathcal{G}) &= \{S \in \mathcal{S}_3(\mathcal{G}) : \text{there are some adjacencies in } S\} \\
U_3(\mathcal{G}) &= \{S \subseteq \mathcal{V}(\mathcal{G}^u) : |S| = 3 \text{ and } S \text{ is complete}\}.
\end{aligned}
$$

Note $\mathcal{G}^u$ is defined in Definition 2.3.2.

Thus by definition $\tilde{\mathcal{S}}_3(\mathcal{G}) \subseteq \hat{\mathcal{S}}_3(\mathcal{G}) \subseteq \mathcal{S}_3(\mathcal{G})$ and $\mathcal{S}_2(\mathcal{G})$, $H_1(\mathcal{G})$, $H_2(\mathcal{G})$, $H_3^a(\mathcal{G})$, $H_3^n(\mathcal{G})$, $U_3(\mathcal{G})$ are disjoint.

Lemma 2.4.1. *In a MAG $\mathcal{G}$, for any single vertex a,* $\text{tail}(a) = \text{pa}_{\mathcal{G}}(a)$, *and* $\{v, w\}$ *is a head if and only if* $v \leftrightarrow w$.

Proof. If $a \subset \text{dis}_{\text{an}(a)}(a)$ then there is a vertex b such that $b \leftrightarrow a$ and $b \in \text{an}_{\mathcal{G}}(a)$, which contradicts ancestrality. Hence $\text{tail}(a) = \text{pa}_{\mathcal{G}}(a)$.

If $v \leftrightarrow w$ then v, w have no ancestral relation so by definition, it is a head. Suppose $\{v, w\}$ is a head, so $\{v, w\} \in \mathcal{S}(\mathcal{G})$ then they must be adjacent by Proposition 3.4 and the adjacency can not be undirected or directed, thus $v \leftrightarrow w$. □

Thus $H_1(\mathcal{G})$ and $H_2(\mathcal{G})$ are precisely the sets in $\mathcal{S}_3(\mathcal{G})$ that arise from heads of size one and two, respectively.

Lemma 2.4.2. *For a MAG $\mathcal{G}$, we have*

$$\mathcal{S}_3(\mathcal{G}) = \mathcal{S}_2(\mathcal{G}) \cup H_1(\mathcal{G}) \cup H_2(\mathcal{G}) \cup H_3(\mathcal{G}) \cup U_3(\mathcal{G})$$
$$\hat{\mathcal{S}}_3(\mathcal{G}) = \mathcal{S}_2(\mathcal{G}) \cup H_1(\mathcal{G}) \cup H_2(\mathcal{G}) \cup H_3^a(\mathcal{G}) \cup U_3(\mathcal{G}).$$

Proof. Consider the first equality, for $S = \{v, w\} \in \mathcal{S}_3(\mathcal{G})$, by Proposition 2.2.3, v, w are adjacent in $\mathcal{G}$ so $S \in \mathcal{S}_2$; For $S \in \mathcal{S}_3(\mathcal{G})$ and $|S| = 3$, it is a clique in $\mathcal{G}^u$ or it origins from heads of size either 1 or 2 or 3. Thus by Lemma 4.1 and Lemma 4.1, $S \in H_1(\mathcal{G}) \cup H_2(\mathcal{G}) \cup H_3(\mathcal{G}) \cup U_3(\mathcal{G})$; For S in the right hand side, it is in $\mathcal{S}_3(\mathcal{G})$ by definition.

For the second equality, by definition $\hat{\mathcal{S}}_3(\mathcal{G})$ excludes all $S \in \mathcal{S}_3(\mathcal{G})$ that have no adjacencies, but note that all $S \in \mathcal{S}_2(\mathcal{G}) \cup H_1(\mathcal{G}) \cup H_2(\mathcal{G}) \cup U_3(\mathcal{G})$ have some adjacencies. And by definition $H_3^a(\mathcal{G})$ extract all heads of size 3 with some adjacencies. □

Lemma 2.4.3. *For a MAG $\mathcal{G}$, $A_1'(\mathcal{G}) \cup U_3(\mathcal{G}) = \hat{\mathcal{S}}_3(\mathcal{G})$.*

Proof. $S \in A_1'(\mathcal{G})$ obtained at line 5, 8, 12, 15, 20 and 24, correspond to sets in $\mathcal{S}_2(\mathcal{G})$, $H_1(\mathcal{G})$, $\mathcal{S}_2(\mathcal{G})$, $H_2(\mathcal{G})$, $H_3^a(\mathcal{G})$ and $\mathcal{S}_2(\mathcal{G})$, respectively. So by Lemma 2.4.2, $A_1'(\mathcal{G}) \cup U_3(\mathcal{G}) \subseteq \hat{\mathcal{S}}_3(\mathcal{G})$ Conversely, all sets in $\hat{\mathcal{S}}_3(\mathcal{G}) \setminus U_3(\mathcal{G})$ can be obtained at corresponding lines. □

Proposition 2.4.4. *For a MAG $\mathcal{G}$, $A_1(\mathcal{G}) = \tilde{\mathcal{S}}_3(\mathcal{G})$.*

Proof. Compared to $\hat{\mathcal{S}}_3(\mathcal{G})$, $\tilde{\mathcal{S}}_3(\mathcal{G})$ excludes all sets of size 3 that have 3 adjacencies. If the set is clique in $\mathcal{G}^u$ except for edges, it is not added in Algorithm 1. Otherwise note that when sets of size 3 are obtained, lines 7, 14 and 17 check their adjacencies. □

Notice that Algorithm 1 naturally identifies $\hat{\mathcal{S}}_3(\mathcal{G}) \setminus U_3(\mathcal{G})$, but to obtain the full $\hat{\mathcal{S}}_3(\mathcal{G})$ one also needs to identify all triangles in the undirected component; $\tilde{\mathcal{S}}_3(\mathcal{G})$ excludes this set.

Complexity of Algorithm 1

The first loop from line 2 to line 10 runs at most $O(e^2)$ times as the worst case is that one vertex have all others as its parents. There are at most e bidirected edges so the second loop from line 11 to line 23 repeats at most e times. There are three esrial tasks inside the second loop. The first one is line 13 which obtains the tails of $\{v, w\}$. The computation of obtaining tails given parents is $O(n + e)$.

The second task, i.e. the first subloop from line 14 to line 16, is carried at most $n-2$ times as the size of each tail is at most $n-2$. For the third task from line 17 to line 22, there are at most $n-2$ potential candidates for the third member, and obtaining the district costs $O(n+e)$. Thus the overall complexity of Algorithm 1 is $O(e^2 + e((n+e) + n + n(n+e))) = O(ne^2)$.

Note that the number of potential candidates for third member of heads of size 3 depends on sizes of districts. If the number is high then it means districts are large so there are at least as many bidirected edges as potential candidates, so if the graph is sparse we can use e to represent the number of candidates instead of n when computing complexity. There are most $O(e^2)$ sets in $\tilde{\mathcal{S}}_3(\mathcal{G})$, and some graphs achieve this bound, for example, a DAG where one vertex have all others as its parents.

To test ordinary Markov equivalence of two MAGs, it is sufficient to compare their output of Algorithm 1 after a sort of order $O(e^2 \log e^2) = O(e^2 \log e)$. Note that $\log e = O(\log n)$, therefore the complexity of verifying Markov equivalence between two MAGs is still $O(ne^2)$. Thus our algorithm is faster than the one proposed by Ali et al. (2009), which is only $O(ne^4)$.

2.4.3 ADMGs

Algorithm 2 converts an ADMG $\mathcal{G}$ to a Markov equivalent MAG $\mathcal{G}^m$, as proven by Lemma 2.3.3. To test Markov equivalence between two ADMGs, it is sufficient to put their equivalent MAGs in Algorithm 1 to obtain the corresponding sets $\tilde{\mathcal{S}}_3$ and compare the sets.

Complexity Of Algorithm 2

For the first loop from line 2 to 6, it costs $O(n(n+e))$ since there are n vertices and it takes $O(n+e)$ to obtain a district. The second loop from line 7 to 12 is at $O(n^2(n+e))$. Thus the overall complexity is $O(n(n+e)+n^2(n+e)) = O(n^3+n^2e) = O(n^2e)$. The total cost for identifying Markov equivalence between two ADMGs is therefore $O(ne^2)$.

2.4.4 Comparison To Previous Algorithms

Among previous characterizations of MAGs, Ali et al. (2009) provide a polynomial time algorithm to verify Markov equivalence. They consider all triples in a discriminating path; in order to do this, they iterate through (up to) $n-2$ levels; at each

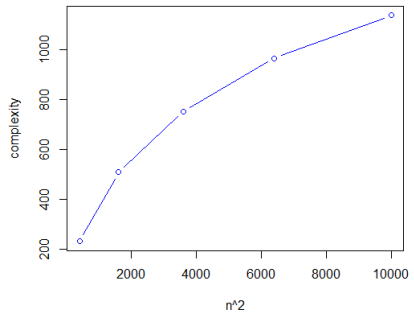


Figure 2.1: Empirical complexity against n^2

level they consider all remaining colliders ($O(e^2)$) and then check each set of reachable edges ($O(e^2)$). Conversely, we ignore any triples for which all three adjacencies are present (since they will trivially always be present).

In addition to the reduction in complexity, if we modify Algorithm 1 to compute $\mathcal{S}_3(\mathcal{G})$, the output contains more information. By Proposition 2.2.2, a set $\{a_1, a_2, a_3\}$ is missing from $\mathcal{S}_3$ if and only if there is a corresponding m-separation between (say) a_1, a_2 conditional on a set that includes a_3. Thus we can view the parametrizing set as a summary of independence information in the graph. This is a novel perspective compared to previous theorems, which characterize graphs by structures like minimal collider paths or colliders with order, and have a more straightforward connection to conditional independence.

2.4.5 Empirical Complexity

An experiment on random graphs shows that empirical complexity of Algorithm 1 is at $O(e^2)$ for sparse graphs ($e = O(n)$). One random graph (ADMG) is generated in the following way. We first fix a topological ordering and the total number of edges ($e = 3n$). Then two vertices become adjacent with uniform probability. Once skeleton is determined, an edge is independently either directed or bidirected with $p = 0.5$. For each $n = 20, 40, 60, 80, 100$, we generate $N = 250$ random graphs then average the empirical complexity. Figure 2.1 is the empirical complexity against n^2.

Suppose directed edges are added independently with probability r/n according to a predetermined topological order, where n is the number of vertices and $r \in \mathbb{R}^+$ is constant. The following proposition bounds the size of the ancestor sets in our sparse random graphs. In particular, the largest average number of ancestors is at most e^r.

Proposition 2.4.5. *Let A_i be the number of ancestors of the vertex i. Then*

$$\mathbb{E}A_i = \left(1 + \frac{r}{n}\right)^{i-1}.$$

In particular,

$$\mathbb{E}A_n = \left(1 + \frac{r}{n}\right)^{n-1} \longrightarrow e^r.$$

Proof. We proceed by induction. The result is trivially true for $A_2 = 1 + \frac{r}{n}$. Suppose the result holds for A_j. Then

$$\begin{aligned}\mathbb{E}A_{j+1} &= 1 + \sum_{i=1}^{j} \mathbb{E}\mathbb{1}_{\{i \to j+1\}} A_i \\ &= 1 + \frac{r}{n} \sum_{i=1}^{j} \left(1 + \frac{r}{n}\right)^{i-1},\end{aligned}$$

using independence of the edge and A_i and the induction hypothesis. Hence

$$\begin{aligned}\mathbb{E}A_{j+1} &= 1 + \sum_{i=1}^{j} \sum_{k=0}^{i-1} \binom{i-1}{k} \left(\frac{r}{n}\right)^{k+1} \\ &= 1 + \sum_{k=0}^{j-1} \left(\frac{r}{n}\right)^{k+1} \sum_{i=k+1}^{j-1} \binom{i-1}{k} \\ &= 1 + \sum_{k=0}^{j-1} \left(\frac{r}{n}\right)^{k+1} \binom{j}{k+1} \\ &= 1 + \sum_{k=1}^{j} \left(\frac{r}{n}\right)^{k} \binom{j}{k}.\end{aligned}$$

by a standard result about binomial coefficients. This gives the result. □

Markov's inequality gives us an easy corollary.

Corollary 2.4.5.1. $\mathbb{P}(A_i \geq k) \leq e^r/k$ *for any* $1 \leq i \leq n$ *and* $k \geq 1$.

Now it is straightforward to show that for sparse graphs, the complexity will be $O(e^2)$. This is because the main contribution of the complexity comes from counting heads of size 3. By bounding the sizes of ancestor sets, line 15 will run in constant time $O(1)$ instead of $O(n + e)$. Thus the overall complexity for sparse graphs is at $O(e^2 + e((n + e) + n + n)) = O(e^2)$.

An example for which the upper bound of complexity of Algorithm 1 is reached is given in Figure 2.2. For every i and j, $\{v_i, w, z_j\}$ forms a head of size 3. If N, M, L are at $O(n)$ then the cost for identifying all these heads is at $O(ne^2)$.

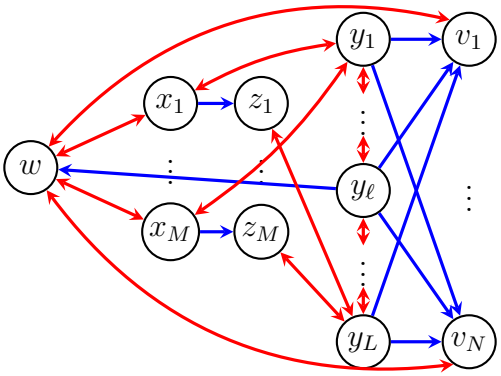


Figure 2.2: A sequence of graphs in which the maximum complexity is achieved by Algorithm 1. Note that y_1 is connected by a bidirected edge to every x_i, and y_L to every z_i.

2.5 Extension to Summary Graphs and MAGs with undirected edges

MAGs defined in Richardson and Spirtes (2002) contain undirected edges which necessitate additional conditions of ancestrality. In addition to the previous condition ($\text{sib}_{\mathcal{G}}(v) \cap \text{an}_{\mathcal{G}}(v) = \emptyset$ and this is referred as condition 1 of ancestrality), one also requires that if an undirected edge is present between two vertices v and w then there is no arrow into v or w. We refer to this as condition 2 of ancestrality.

Definition 2.5.1. A graph $\mathcal{G}$ is *ancestral* if: (1) for every $v \in \mathcal{V}$, $\text{sib}_{\mathcal{G}}(v) \cap \text{an}_{\mathcal{G}}(v) = \emptyset$; (2) if there is an undirected edge $x - y$ then x, y have no parents and no siblings.

A direct consequence of this definition is that vertices with undirected edges are 'at the top' of the graph $\mathcal{G}$. For an acyclic graph $\mathcal{G}$ with three types of edges and only satisfying condition 2 of ancestrality, it can be seen as an ADMG with an undirected component among vertices without parents or siblings and therefore the component is "at the top" of the graph.

Summary graphs defined in Wermuth (2011) are actually the same as ADMGs with undirected components at the top. Graphically, one just needs to change the dashed lines to bidirected edges and they encode the same conditional independence. For simplicity, we will refer to this type of graphs as summary graphs. Among the three graphs in Figure 2.3, (ii) is the only summary graph.

Definition 2.5.2. For a summary graph $\mathcal{G}$, let $U = \{v \in \mathcal{V} : v - w \text{ for some } w \in \mathcal{V}\}$ and $D = \mathcal{V} \setminus U$. Define $\mathcal{G}^u = \mathcal{G}_U$ and $\mathcal{G}^d = \mathcal{G}_D$.

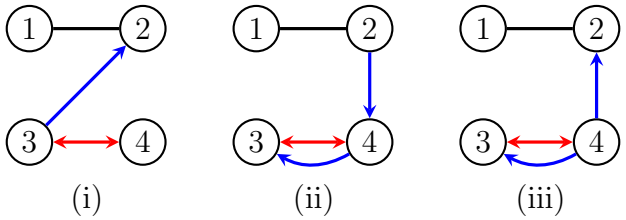


Figure 2.3: (i) A graph that satisfies only condition 1 of ancestrality. (ii) A graph that satisfies only condition 2 of ancestrality. (iii) A graph that does not satisfy either condition 1 or 2 of ancestrality.

It is shown by Richardson and Spirtes (2002) that we can always split a summary graph into two disjoint subgraphs. One is an undirected subgraph $\mathcal{G}^u$ and another one is a subgraph with only directed and bidirected edges $\mathcal{G}^d$. Note that heads and barren sets are only defined in $\mathcal{G}^d$, and tails may include vertices in both $\mathcal{G}^u$ and $\mathcal{G}^d$.

For example, Figure 2.3(ii) can be split as $\mathcal{G}^u = 1 - 2$ and $\mathcal{G}^d = 3 \leftarrow 4, 3 \leftrightarrow 4$. Its heads are $\{3\}$ and $\{4\}$ and the corresponding tails are $\{2, 4\}$ and $\{2\}$.

A vertex a is said to be *anterior* to b if there is a path π on which every edge is either undirected or directed towards b, or if $a = b$. We denote the collection of all vertices anterior to b by $\text{ant}_{\mathcal{G}}(b)$.

An *undirected graph* (UG) is a graph with only undirected edges. A *clique* in an UG is defined as a complete subset of vertices, that is: every pair of vertices is connected by an undirected edge.

For summary graphs, including MAGs, a *clique* is defined in the same manner for vertices in $\mathcal{G}^u$, with completeness referring only to adjacencies by undirected edges.

Remark 2. We extend the definition of parametrizing set by adding all the cliques to the set.

2.5.1 Extension to MAGs With Undirected Edges

We only need to add a few line of argument to extend previous propositions and theorems.

For $\Rightarrow$ of Proposition 3.3: if $W \in \mathcal{S}(\mathcal{G})$, then either W is a clique or there is a nonempty subset $W' \subseteq W$ such that W' is a head and $W \subseteq W' \cup \text{tail}(W')$. The latter case has been proved. For the former case, it clearly implies that we can not m-separate any two vertices in W, given the remaining vertices in W.

For $\Leftarrow$ of Proposition 3.3: For W that does not lie entirely in $\mathcal{G}^u$ we can define $W' = \text{barren}(W)$. For W lying in $\mathcal{G}^u$, if we cannot m-separate any two vertices in W then clearly W is a clique and $W \in \mathcal{S}(\mathcal{G})$.

Proposition 3.4 does not change if we add undirected edges in MAGs, thus Theorem 3.2 and Corollary 3.2.1 hold for MAGs with undirected edges.

2.5.2 Extension to Summary Graphs

The projection described in Section 3.3 can be extended to summary graphs with latent variables L as stated in Richardson and Spirtes (2002). The modified projection is: (i) every pair of vertices $a, b \in \mathcal{V}$ in $\mathcal{G}$ that are connected by an *inducing path* becomes adjacent in $\mathcal{G}^m$; (ii) an edge connecting a, b in $\mathcal{G}^m$ is oriented as follows: if $a \in \text{ant}_{\mathcal{G}}(b)$ then $a \to b$; if $b \in \text{ant}_{\mathcal{G}}(a)$ then $b \to a$; if neither is the case, then $a \leftrightarrow b$; if they are both anterior to one another then the edge is undirected. An *inducing path* between a, b is a path such that every collider in the path is in $\text{an}(\{a, b\})$, and every noncollider is in L. Again, we only consider projections with no latent variables, so an inducing path is just a collider path with every collider in $\text{an}(\{a, b\})$. And the projection still preserves ancestral relations from the original graph. We first show that undirected edges are preserved through projections.

Lemma 2.5.1. *If $\mathcal{G}$ is a summary graph and $\mathcal{G}^m$ is its corresponding projected MAG, then $\mathcal{G}^u = (\mathcal{G}^m)^u$ and $(\mathcal{G}^d)^m = (\mathcal{G}^m)^d$.*

Proof. For the first statement, we can prove it by showing that undirected edges are the same. First of all, notice that all undirected edges in $\mathcal{G}$ are preserved in $\mathcal{G}^m$. Secondly, no additional undirected edges can be added. If a and b are both in $\mathcal{G}^u$ then if they are not adjacent before, they are still nonadjacent since there is no inducing path between them (they are already at the top of the graph). If a and b are both in $\mathcal{G}^d$ then they cannot be anterior to each other, this would violate condition (ii) of ancestrality or the fact that $\mathcal{G}$ is acyclic. If $a \in \mathcal{G}^u$ and $b \in \mathcal{G}^d$ then obviously b cannot be anterior to a.

For the second statement, note the two subgraphs have the same vertices due to the first statement. For vertices in $\mathcal{G}^d$, ancestral relations are the same in $\mathcal{G}$ as there is no directed path passing $\mathcal{G}^u$. Also when we consider inducing paths, any such path would not contain any vertex in $\mathcal{G}^u$. □

We now show that Proposition 2.3.2 also holds for summary graphs, i.e. heads and tails are preserved through projection.

Proof. So we have proved that for ADMGs, heads and tails are preserved through the projection. Now heads are only defined in $\mathcal{G}^d$ and $(\mathcal{G}^m)^d$, thus by Lemma 2.5.1, for a summary graph, heads are preserved in $\mathcal{G}^m$. Also for tails that are in $\mathcal{G}^d$, they are preserved. It remains to show that the result holds when tails are in $\mathcal{G}^u$. For a head

H, let $w \in \mathcal{G}^u$. If $w \in \text{tail}_{\mathcal{G}}(H)$ then we know there is a path $\pi : w \to w_1 \leftrightarrow \cdots \leftrightarrow h$, for $h \in H$ with intermediate vertices in $\text{an}(H)$. Although $w \notin \mathcal{G}^d$, with the same argument in Lemma 3.5, this path is preserved as a collider path in $\text{an}(H)$ in $\mathcal{G}^m$ with $\leftrightarrow h$ (h is in a head) hence $w \in \text{tail}_{\mathcal{G}^m}(H)$. Suppose now $w \in \text{tail}_{\mathcal{G}^m}(H)$, so there is a path $\pi : w \to w_1 \leftrightarrow \cdots \leftrightarrow h$ with intermediate vertices in $\text{an}(H)$, we know every bidirected edge corresponds to a bidirected path in $\text{an}(H)$ in $\mathcal{G}$, and the first directed edge correspond to a path $\pi' : w \to w_1 \leftrightarrow \cdots \leftrightarrow w_2$ in $\mathcal{G}$ with intermediate vertices in $\text{an}(w_2) \subseteq \text{an}(H)$, thus $w \in \text{tail}_{\mathcal{G}}(H)$. □

Since Proposition 3.6 holds for summary graphs, if we change the definition of *adjacencies* in summary graphs in the same manner as ADMGs by referring to m-separations, Theorem 3.8 and Corollary 3.8.1 also hold for summary graphs.

2.5.3 Extension for Algorithms

For Algorithm 1, we only add a line at the end of the algorithm (after line 17) to obtain the connected pairs in $\mathcal{G}^u$ (referred as line 18 in the next section). This costs $O(e)$ and hence does not contribute to the overall complexity.

For Algorithm 2, as shown by Lemma 2.5.1, undirected edges are preserved, it is sufficient to add a line at the end of the algorithm (after line 9) to keep all the undirected edges. This costs $O(e)$ and hence does not contribute to the overall complexity.

2.6 Independence from Parametrizing Sets

This section shows that maximal independences can be deduced from the parametrizing sets directly without explicit knowledge of graphs. We don't have any use of these results yet but this provides some insight.

Lemma 2.6.1. *Let $\mathcal{G}$ be a MAG if $W \notin \mathcal{S}(\mathcal{G})$ then for any $d \in \text{an}(W)$, $\{d\} \cup W = \hat{W}$ is not in $\mathcal{S}(\mathcal{G})$.*

Proof. Since $d \in \text{an}(W)$ we have $\text{barren}_{\mathcal{G}}(W) = \text{barren}_{\mathcal{G}}(\hat{W})$; denote this set by H. If H is not a head then obviously $\hat{W} \notin \mathcal{S}(\mathcal{G})$. If H is a head then the reason that W is not in $\mathcal{S}(\mathcal{G})$ is because there are vertices in W that are in $\text{an}(H)$ but not in $\text{tail}(H)$. Since adding d does not change this fact, $\hat{W}$ is not in $\mathcal{S}(\mathcal{G})$. □

Proposition 2.6.2. *Let $\mathcal{G}$ be a MAG and consider $W \notin \mathcal{S}(\mathcal{G})$ where $W = \{a, b\} \cup C$ such that $a, b \notin C$, if the following conditions hold:*

$$(i) : \textit{for any } S \supset W \textit{ we have } S \in \mathcal{S}(\mathcal{G})$$

$$(ii) : \textit{for any } C' \subseteq C \textit{ we have } \{a, b\} \cup C' = W' \notin \mathcal{S}(\mathcal{G});$$

then $a \perp_m b \mid C$ *in* $\mathcal{G}$

Proof. Suppose there are m-connecting paths between a and b given C. The goal is to find a head H such that $\{a, b\} \cup C' \subseteq H \cup \text{tail}(H)$ for some $C' \subseteq C$ so it contradicts to the condition (ii).

Firstly, we show that any collider on any of these m-connecting paths is in C. Suppose one of them is not in C, denote it by d. Since it is on a m-connecting path, we know $d \in \text{an}(C) \subseteq \text{an}(W)$. By Lemma 2.6.1, $\hat{W} = \{d\} \cup W$ is not in $\mathcal{S}(\mathcal{G})$ which is a contradiction to the condition (i). Hence all the colliders are in C.

Secondly, we show that there is no non-collider on any of these m-connecting paths. Suppose there is a non-collider d, then it is not in W and it must be an ancestor of one of the colliders on the path or endpoints a, b thus $d \in \text{an}(W)$. By Lemma 2.6.1, $\hat{W} = \{d\} \cup W$ is not in $\mathcal{S}(\mathcal{G})$ which is a contradiction to the condition (i).

Now consider any collider path π m-connecting a and b given C. Denote the colliders on the path by $C'' \subseteq C$. Then consider $\hat{W} = \{a, b\} \cup C''$, $H = \text{barren}(\hat{W})$ must be a head because there is a collider path between any two vertices in it. Similarly all the remaining vertices must be in $\text{tail}(H)$. Hence $\hat{W}$ is in $\mathcal{S}(\mathcal{G})$ and this contradicts to the condition (ii). □

Lemma 2.6.3. *Let* $\mathcal{G}$ *be a MAG and consider* $W \notin \mathcal{S}(\mathcal{G})$ *such that for any* $S \supset W$ *we have* $S \in \mathcal{S}(\mathcal{G})$*, then there exists a pair* $\{a, b\}$ *in* W *such that* $a \perp_m b \mid W \setminus \{a, b\}$ *in* $\mathcal{G}$*.*

Proof. By Proposition 2.2.2, there exists a m-separation $a \perp_m b \mid C$ in $\mathcal{G}$, such that $\{a, b\} \subseteq W \subseteq \{a, b\} \cup C$. Since there is no strictly larger set of W that is not in $\mathcal{S}(\mathcal{G})$, we have $W = \{a, b\} \cup C$. □

For a missing set, it is possible that we have to condition on a strictly larger set to m-separate two vertices in it. For example, consider the graph $1 \leftrightarrow 2 \leftarrow 3 \leftarrow 4$. $\{1, 2, 4\}$ is not in the parameterizing set, but in order to m-separate $\{1, 4\}$ or $\{2, 4\}$ we have to include 3 in the conditioning set.

2.7 PAG and the parametrizing set

Given a MAG $\mathcal{G}$, Zhang (2007b) uses its *partial ancestral graph* (PAG) to characterize $[\mathcal{G}]$, which denotes the MEC of $\mathcal{G}$ and captures all the arrowheads and tails that are present in every MAG in $[\mathcal{G}]$. Because the parametrizing set also characterize

$[\mathcal{G}]$, we should be able to construct the PAG given the parametrizing set. The PAG will be useful when we move to Chapter 4 where we use PAGs as the representation of MECs of MAGs, which is more efficient than the parametrizing set.

2.7.1 Definition of PAGs

Given a MAG $\mathcal{G}$, an edge mark in $\mathcal{G}$ is *invariant* if it is present in every graph in $[\mathcal{G}]$.

Definition 2.7.1. Given a MAG $\mathcal{G}$, the *partial ancestral graph* (PAG) for $[\mathcal{G}]$, $\mathcal{P}_\mathcal{G}$, is a graph with three kind of edge marks: arrowheads, tails and circles (six kinds of edges: $-$, $\rightarrow$, $\leftrightarrow$, $\circ\!\!-$, $\circ\!\!-\!\!\circ$, $\circ\!\!\rightarrow$)[1], such that:

- $\mathcal{P}_\mathcal{G}$ has the same adjacencies as any member of $[\mathcal{G}]$;
- a mark of arrowhead is in $\mathcal{P}_\mathcal{G}$ if and only if it is invariant in $[\mathcal{G}]$;
- a mark of tail is in $\mathcal{P}_\mathcal{G}$ if and only if it is invariant in $[\mathcal{G}]$.

Zhang (2007b) present an algorithm, including a set of rules, $\mathcal{R}0$ to $\mathcal{R}10$ to construct the PAG of a given MAG. The algorithm is shown to be sound and complete, it begins with a graph $\mathcal{P}$ that has the same adjacencies as $\mathcal{G}$ and only one kind of edge $\circ\!\!-\!\!\circ$. Then we exhaustively apply the orientation rules until no more edge marks can be changed. The orientation rules are listed here.

The first five rules $\mathcal{R}0$ to $\mathcal{R}4$ developed by Spirtes et al. (2000) is to find all the invariant arrow heads.

$\mathcal{R}0$ For every unshielded triple of vertices (a, b, c), if it is an unshielded collider in $\mathcal{G}$, then orient the triple as $a \ast\!\!\rightarrow b \leftarrow\!\!\ast\, c$. (The mark $\ast$ means we do not care what the original mark is and keep the mark after applying the orientation rules.)

$\mathcal{R}1$ If $a \ast\!\!\rightarrow b \circ\!\!-\!\!\ast\, c$ and a, c are not adjacent, then orient the triple as $a \ast\!\!\rightarrow b \rightarrow c$.

$\mathcal{R}2$ If $a \rightarrow b \ast\!\!\rightarrow c$ or $a \ast\!\!\rightarrow b \rightarrow c$, and $c \circ\!\!-\!\!\ast\, a$, then orient $c \circ\!\!-\!\!\ast\, a$ as $c \leftarrow\!\!\ast\, a$.

$\mathcal{R}3$ If $a \ast\!\!\rightarrow b \leftarrow\!\!\ast\, c, a \ast\!\!-\!\!\circ\, d \circ\!\!-\!\!\ast\, c$, a and c are not adjacent, and $d \ast\!\!-\!\!\circ\, b$, then orient $d \ast\!\!-\!\!\circ\, b$ as $d \ast\!\!\rightarrow b$

$\mathcal{R}4$ If $\pi = \langle d, \ldots, a, b, c\rangle$ is a discriminating path between d and c for b in $\mathcal{P}$, and $b \circ\!\!-\!\!\ast\, c$; then if the edge $b \rightarrow c$ is present in $\mathcal{G}$, orient $b \circ\!\!-\!\!\ast\, c$ as $b \rightarrow c$; otherwise, orient the triple (a, b, c) as $a \leftrightarrow b \leftrightarrow c$.

[1] As we consider only directed MAGs, there are only four kinds of edges

Then Zhang (2007b) developed the remaining rules $\mathcal{R}5$ to $\mathcal{R}10$ to identify all the invariant tails.

We need the following definitions first. A PMG is a mixed graph used during construction of the PAG.

Definition 2.7.2. In a PMG, a path $\pi = \langle v_0, \ldots, v_n \rangle$ is said to be *uncovered* if for every $1 \leq i \leq n-1$, v_{i-1} and v_{i+1} are not adjacent.

Definition 2.7.3. In a PMG, a path $\pi = \langle v_0, \ldots, v_n \rangle$ is said to be *potentially directed* (*p.d.*) from v_0 to v_n if for every $1 \leq i \leq n$, the edge between v_{i-1} and v_i is not into v_{i-1} or out of v_i.

Definition 2.7.4. In a PMG, a path π is a *circle path* if every edge on the path is of the form $\circ\!\!-\!\!\circ$.

The additional rules provided by Zhang (2007b) are:

$\mathcal{R}5$ For every $a \circ\!\!-\!\!\circ b$ if there is an uncovered circle path $\pi = \langle a, c, \ldots, d, b \rangle$ for a, b such that a, d are not adjacent and b, c are not adjacent, then orient $a \circ\!\!-\!\!\circ b$ and all the edges on π as undirected edges;

$\mathcal{R}6$ If $a - b \circ\!\!-\!\!* c$, then orient $b \circ\!\!-\!\!* c$ as $b -\!\!* c$;

$\mathcal{R}7$ If $a -\!\!\circ b \circ\!\!-\!\!* c$, and a, c are not adjacent, then orient $b \circ\!\!-\!\!* c$ as $b -\!\!* c$;

$\mathcal{R}8$ If $a \to b \to c$ or $a -\!\!\circ b \to c$, and $a \circ\!\!\to c$, then orient $a \circ\!\!\to c$ as $a \to c$.

$\mathcal{R}9$ If $a \circ\!\!\to c$, and $\pi = \langle a, b, \ldots, c \rangle$ is an uncovered p.d. path from a to c such that b and c are not adjacent, then orient $a \circ\!\!\to c$ as $a \to c$.

$\mathcal{R}10$ Suppose $a \circ\!\!\to c$, $b \to c \leftarrow d$, π_1 is an uncovered p.d. path from a to b, and π_2 is an uncovered p.d. path from a to d. Let x be the vertex adjacent to a on π_1, and y be the vertex adjacent to a on π_2. If x and y are distinct, and are not adjacent, then orient $a \circ\!\!\to c$ as $a \to c$.

Now we are ready to show how to construct PAG given parametrizing set.

2.7.2 Construct PAG given parametrizing set

We define $[\mathcal{S}]$ to be the set of all MAGs that have the parameterizing set $\mathcal{S}$, so given a MAG $\mathcal{G}$, $[\mathcal{G}] = [\mathcal{S}(\mathcal{G})]$ and naturally we can define $\mathcal{P}_\mathcal{S}$ to denote the PAG that characterizes the Markov equivalent class $[\mathcal{S}]$ in the same manner as Definition 2.7.1. Because the parameterizing sets also characterise $[\mathcal{G}]$, we should be able to compute

$\mathcal{P}_\mathcal{S}$ given a $\mathcal{S}$. Now we demonstrate how to achieve this. The method relies much on Zhang (2007b); Ali et al. (2005).

Given a MAG, the algorithm to construct the PAG begins with a graph $\mathcal{P}$ that has the same adjacencies as $\mathcal{G}$ and only one kind of edge $\circ\!\!-\!\!\circ$. Then exhaustively apply the orientation rules.

Instead of a MAG $\mathcal{G}$, suppose now we are only given a parameterizing set $\mathcal{S}$ (we may not necessarily know $\mathcal{G}$). We will show that with a slight change of the above rules, we are able to identify all the invariant arrow heads in $\mathcal{P}_\mathcal{S}$.

Firstly notice that we can obtain adjacencies from $\mathcal{S}$, so we can construct the initial graph $\mathcal{P}$ as Zhang (2007b) does. Also notice that only $\mathcal{R}0$ and $\mathcal{R}4$ require information from graphs so it is sufficient to construct equivalent rules to replace these two rules. The adapted rules are:

$\mathcal{R}0'$ For every unshielded triple of vertices (a, b, c), if it is in $\mathcal{S}$, then orient the triple as $a \ast\!\!\rightarrow b \leftarrow\!\!\ast\, c$.

$\mathcal{R}4'$ If $\pi = \langle d, \ldots, a, b, c\rangle$ is a discriminating path between d and c for b in $\mathcal{P}$, and $b \circ\!\!-\!\!\ast\, c$; then if the triple (d, b, c) is present in $\mathcal{S}$, orient $b \circ\!\!-\!\!\ast\, c$ as $b \rightarrow c$; otherwise, orient the triple (a, b, c) as $a \leftrightarrow b \leftrightarrow c$.

Proposition 2.7.1. *The orientation rules: $\mathcal{R}0'$, $\mathcal{R}1$, $\mathcal{R}2$, $\mathcal{R}3$, $\mathcal{R}4'$ and $\mathcal{R}5$ to $\mathcal{R}10$ are sound and complete for constructing $\mathcal{P}_\mathcal{S}$ given $\mathcal{S}$. Further if we are only given $\tilde{\mathcal{S}}_3$, these rules are sufficient to construct $\mathcal{P}_\mathcal{S}$.*

Proof. This follows immediately from Proposition 3.4 in Hu and Evans (2020). Note that if a discriminating path is present in $\mathcal{P}_\mathcal{S}$ then it is present in all MAGs in $[\mathcal{S}]$. □

2.7.3 Possible Improvement

The fact that an unshielded triple is in $\mathcal{S}$ if and only it is an unshielded collider contributes to identify two invariant arrowheads. In addition to this, one may notice that apart from unshielded triples, triples with one adjacency in $\mathcal{S}$ also inherit information of invariant arrowheads.

Lemma 2.7.2. *For a triple $\{a, b, c\}$ in $\mathcal{S}$ with one adjacency (WLOG, a and b are adjacent), any MAGs in $[\mathcal{S}]$ has the edge $a \leftrightarrow b$. In other words, $a \leftrightarrow b$ in $\mathcal{P}_\mathcal{S}$.*

Proof. Consider the head of the triple $\{a, b, c\}$. It cannot be a single vertex because $\{a, b, c\}$ has only one adjacency and we know the tail of a single vertex are its parents. If the head is of size 2, it has to be a and b, because we know a pair of vertices $\{a, b\}$ is a head if and only if $a \leftrightarrow b$ and thus $a \leftrightarrow b$. If the head is of size 3,

then the adjacency must be a bidirected edge because there is no ancestral relation inside a head. □

For the arrowheads identified in Step 6, we can recover $7 \leftrightarrow 8$ directly by Lemma 2.7.2. Also we can argue the arrowhead from 6 to 8 by the following: if $6 \rightarrow 8$ then the triple $\{2, 7, 8\}$ would not be in $\tilde{\mathcal{S}}_3$.

Here we give an example on how to recover the PAG given a parametrizing set $\tilde{\mathcal{S}}_3$. Suppose we are given the $\tilde{\mathcal{S}}_3$ in Table 2.1.

Table 2.1: $\tilde{\mathcal{S}}_3$

	adjacencies	unshielded colliders	triples with one adjacency
$\tilde{\mathcal{S}}_3$	$\{1,2\},\{1,3\},\{2,4\}$ $\{3,4\},\{2,5\},\{5,6\}$ $\{5,7\},\{6,7\},\{6,8\}$ $\{7,8\}$	$\{2,5,6\},\{5,6,8\}$ $\{5,7,8\}$	$\{2,7,8\}$

We first identify all the invariant tails. Steps below correspond to figures in Figure 2.4:

Step 1 Begin with a graph with the adjacencies in $\tilde{\mathcal{S}}_3$ and all the edges are $\circ\!\!-\!\!\circ$;

Step 2 Apply $\mathcal{R}0'$ to identify the invariant arrowhead from unshielded triples $\{2,5,6\}$, $\{5,6,8\}$, $\{5,7,8\}$;

Step 3 Apply $\mathcal{R}1$ to $\{2,5,7\}$ so $5 \circ\!\!\rightarrow 7$ becomes $5 \rightarrow 7$;

Step 4 Apply $\mathcal{R}2$ to the triple $\{6,5,7\}$ to recover $6 \circ\!\!\rightarrow 7$;

Step 5 The path $\pi = \langle 2,5,6,7\rangle$ forms a discriminating path for 6 thus by $\mathcal{R}4'$ ($\{2,6,7\}$ is not in $\tilde{\mathcal{S}}_3$), we can recover $6 \rightarrow 7$;

Step 6 The path $\pi = \langle 2,5,6,8,7\rangle$ forms a discriminating path for 8, thus by $\mathcal{R}4'$ ($\{2,7,8\}$ is in $\tilde{\mathcal{S}}_3$), we can recover $6 \leftrightarrow 8 \leftrightarrow 7$;

And no further arrowhead can be identified. We now identify the invariant tails:

Step 7 Apply $\mathcal{R}5$ to $1\circ\!\!-\!\!\circ 2\circ\!\!-\!\!\circ 4\circ\!\!-\!\!\circ 3\circ\!\!-\!\!\circ 1$. So all the circle edges become undirected edges;

Step 8 Apply $\mathcal{R}6$ to $4 - 2 \circ\!\!\rightarrow 5$ to recover $2 \rightarrow 5$.

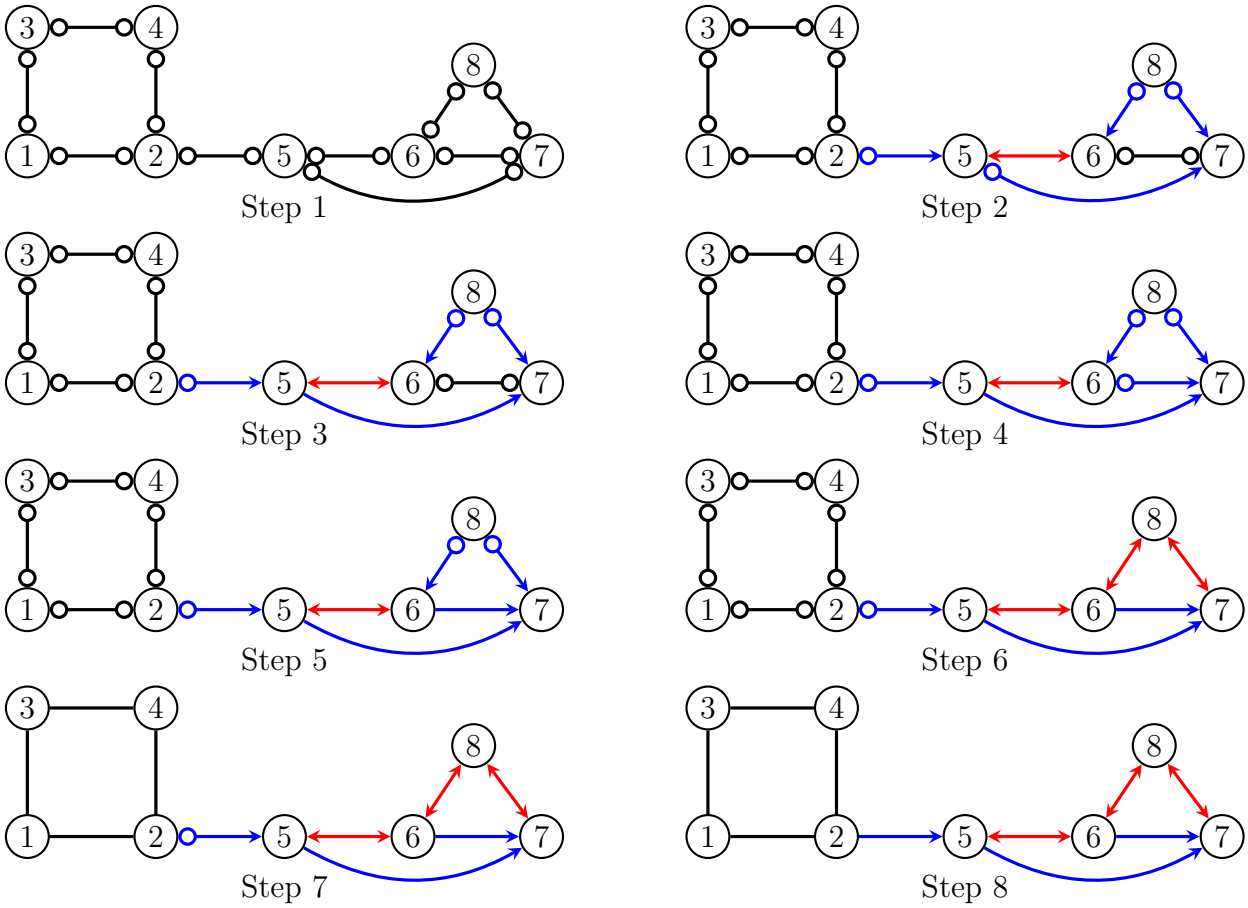


Figure 2.4: Steps for recovering the PAG given the $\tilde{\mathcal{S}}_3$ in Table 2.1

And we can see that there is no circle mark in the graph now so the last figure in Figure 2.4 is the PAG from the parametrizing set $\tilde{\mathcal{S}}_3$ in Table 2.1. Also this is the only MAG that have the corresponding $\tilde{\mathcal{S}}_3$.

From Lemma 2.7.2, we may argue the edge mark by the presence or missingness of certain triples in $\tilde{\mathcal{S}}_3$. For example in Step 5, if we have $6 \leftrightarrow 7$ then the triple $\{2, 6, 7\}$ would be in $\tilde{\mathcal{S}}_3$, which is not true.

Chapter 3

Towards standard imsets of MAGs

3.1 Introduction

Imsets are 'integer valued vectors indexed by subsets of the variables' that can be used to encode arbitrary CI models. Imsets are very useful in the context of graphical models, because DAG models fit into this representation very elegantly. The imset used to represent a DAG model is usually referred as the 'standard imset', meaning that they are the simplest imset that correctly and precisely represent the conditional independences implied by the graphs. For DAGs, there are several advantages to using standard imsets. Firstly, when two DAGs represents the same CI model, in other words they are Markov equivalent, their standard imsets also agree. Moreover, the BIC score of a DAG is an affine function of the standard imset. So we can see that for DAGs, imsets not only provide a representation of the MEC but also lead to a consistent scoring criteria.

We will propose a standard imset formula for MAGs, which though it does not always define the same model as the graphs, it works for large sub-classes of MAGs. Markov equivalent MAGs will, because of the definition we use, always have the same imset.

We use the *parametrizing set* that arise in the discrete parametrization (Richardson, 2009; Hu and Evans, 2020), a factorization theorem (Richardson, 2009) of MAG models and Chapter 2. One of the motivations is that for DAGs, the *characteristic* imset $c_{\mathcal{G}}$ introduced by Studený et al. (2010) that can be obtained by a one-to-one linear transformation from $u_{\mathcal{G}}$ (standard imsets of DAGs), agrees with the *parametrizing set*, thus we work backwards and deduce the formulae for the standard imset. Another motivation is that Chapter 2 (published in Hu and Evans (2020)) show that two MAGs are Markov equivalent if and only if they have the same *parametrizing set*. Thus by construction two MAGs agree on their standard imsets if and only if they are Markov equivalent.

3.2 Definition of Imsets

We now introduce the framework of imsets and conditional independence by Studený (2006).

Let $\mathcal{P}(\mathcal{V})$ be the power set of a finite set of variables $\mathcal{V}$. For any three disjoint sets, $A, B, C \subseteq \mathcal{V}$, we write the triple as $\langle A, B \,|\, C\rangle$ and denote the set of all such triples by $\mathcal{T}(\mathcal{V})$.

The set of conditional independence statements under P is denoted as:

$$\mathcal{I}_P = \{\langle A, B \,|\, C\rangle \in \mathcal{T}(\mathcal{V}) : A \perp\!\!\!\perp B \mid C\,[P]\rangle\},$$

and[1] this is called the *conditional independence model* induced by P.

Definition 3.2.1. An *imset* is an integer-valued function u: $\mathcal{P}(\mathcal{V}) \to \mathbb{Z}$. The *identifier* function δ_A of a set $A \subseteq \mathcal{V}$ is an imset, defined as $\delta_A(B) = 1$ if $B = A$ and otherwise $\delta_A(B) = 0$.

A *semi-elementary imset* $u_{\langle A,B|C\rangle}$ associated with any triple $\langle A, B \,|\, C\rangle \in \mathcal{T}(\mathcal{V})$ is defined as: $u_{\langle A,B|C\rangle} = \delta_{A\cup B\cup C} - \delta_{A\cup C} - \delta_{B\cup C} + \delta_C$. An *elementary* imset corresponds to the case when both A and B are singletons.

An imset u is *combinatorial* if it can be written as a non-negative integer combination of elementary imsets (or equivalently, semi-elementary imsets). We call an imset u *structural* if there exists $n \in \mathbb{N}$ such that $n \cdot u$ is combinatorial.

We also define the *degree* of a structural imset as the number of elementary imsets in the sum for the minimum n that makes $n \cdot u$ combinatorial.

Note that in the case of $|\mathcal{V}| \leq 4$, every structural imset is also combinatorial, but for $|\mathcal{V}| \geq 5$ this is not true (Hemmecke et al., 2008). We will also give an example in Section 3.6.

A triple $\langle A, B \,|\, C\rangle$ is represented in a structural imset u over $\mathcal{V}$, written as $A \perp\!\!\!\perp B \mid C\,[u]$ if there exists $k \in \mathbb{N}$ such that $k \cdot u - u_{\langle A,B|C\rangle}$ is a combinatorial imset over $\mathcal{V}$. The *model induced by* u then is defined as:

$$\mathcal{I}_u = \{\langle A, B \mid C\rangle \in \mathcal{T}(\mathcal{V}) : A \perp\!\!\!\perp B \mid C\,[u]\}.$$

We say that an imset u is *Markovian* with respect to an independence model $\mathcal{I}$ if for every $\langle A, B \mid C\rangle \in \mathcal{I}$, we have that $\langle A, B \mid C\rangle \in \mathcal{I}_u$. If the converse holds, we say it is *faithful*, and if it is both Markovian and faithful we say it is *perfectly Markovian* with respect to $\mathcal{I}$.

[1] Here for simplicity we use $A \perp\!\!\!\perp B \mid C\,[P]$ to represent $X_A \perp\!\!\!\perp X_B \mid X_C$ under P.

For a MAG $\mathcal{G}$, we write $A \perp_m B \mid C\,[\mathcal{G}]$ if A and B are m-separated by C in $\mathcal{G}$. The *model induced by* $\mathcal{G}$ is then defined as:

$$\mathcal{I}_{\mathcal{G}} = \{\langle A, B \mid C\rangle \in \mathcal{T}(\mathcal{V}) : A \perp_m B \mid C\,[\mathcal{G}]\}.$$

Example 3.2.1. Consider the DAG in Figure 3.1; a combinatorial imset u such that $\mathcal{I}_u = \mathcal{I}_{\mathcal{G}}$ is: $u = u_{\langle 1,2\rangle} + u_{\langle 4,12|3\rangle}$, which has entries:

C	$\emptyset$	$\{1\}$	$\{2\}$	$\{1,2\}$	$\{3\}$	$\{3,4\}$	$\{1,2,3\}$	$\{1,2,3,4\}$
$u(C)$	$+1$	-1	-1	$+1$	$+1$	-1	-1	$+1$

.

Remark 3. Independence models (whether represented by imsets or not) always obey the semi-graphoid axioms, which are listed in Appendix 3.8. In the previous example, the conditional independences $4 \perp\!\!\!\perp 2 \mid 3$ and $4 \perp\!\!\!\perp 1 \mid 2,3$ can both be deduced from $4 \perp\!\!\!\perp 1,2 \mid 3$, and indeed $\mathcal{I}_u$ represents these constraints as well.

Section 3.8 also lists some other rules, but these only apply to probabilistic independence models under some additional assumptions. Note that we restrict to using the semi-graphoids, these additional rules are for discussion with related work in Section 3.3.1.

3.3 Standard imsets

In this section we attempt to define the standard imset of MAGs. The next subsection reviews existing results on imsets for DAGs, which provides the motivation for our definition for MAGs.

3.3.1 Related work

Similar work has been done by Andrews (2021); Andrews et al. (2022), who call the parametrizing set of MAGs the *m-connecting* set. In particular, we have the same initial motivation as them, which is the similarity between the 0-1 characteristic imset (Studený et al., 2010) and parametrizing sets (Hu and Evans, 2020) of DAGs. Our work contains results distinct from those works; we give a more detailed comparison in Section 3.7.

3.3.2 Previous work on DAGs

For a DAG $\mathcal{G}$ over $\mathcal{V}$, its standard imset is defined as:

$$u_{\mathcal{G}} = \delta_{\mathcal{V}} - \delta_{\emptyset} - \sum_{i \in \mathcal{V}} \left\{\delta_{\{i\} \cup \mathrm{pa}_{\mathcal{G}}(i)} - \delta_{\mathrm{pa}_{\mathcal{G}}(i)}\right\}. \tag{3.1}$$

Figure 3.1: (i) A DAG with 4 nodes; (ii) a MAG $\mathcal{G}$ in which there is no topological ordering such that the tail of a head precedes any vertex in the head.

Studený (2006) shows that for a DAG $\mathcal{G}$, $\mathcal{I}_{\mathcal{G}} = \mathcal{I}_{u_{\mathcal{G}}}$. That is, $u_{\mathcal{G}}$ is perfectly Markov with respect to $\mathcal{I}_{\mathcal{G}}$. We will often just say that $u_{\mathcal{G}}$ is perfectly Markovian with respect to $\mathcal{G}$, rather than explicitly invoking the list of independences.

Example 3.3.1. Consider the DAG $\mathcal{G}$ in Figure 3.1 (i), by definition, its standard imset is:

$$\begin{aligned} u_{\mathcal{G}} &= \delta_{1234} - \delta_{\emptyset} - (\delta_{34} - \delta_3) - (\delta_{123} - \delta_{12}) - (\delta_2 - \delta_{\emptyset}) - (\delta_1 - \delta_{\emptyset}) \\ &= (\delta_{1234} - \delta_{123} - \delta_{34} + \delta_3) + (\delta_{12} - \delta_1 - \delta_2 + \delta_{\emptyset}) \\ &= u_{\langle 4,12\,|\,3\rangle} + u_{\langle 1,2\rangle}. \end{aligned}$$

In the last line, the conditional independences of the semi-elementary imsets are the independences required for the local Markov property of $\mathcal{G}$, that is, given the numerical (and also topological) ordering

$$i \perp\!\!\!\perp [i-1] \setminus \mathrm{pa}_{\mathcal{G}}(i) \mid \mathrm{pa}_{\mathcal{G}}(i),$$

where $[i] = \{1, 2, \ldots, i\}$.

For a DAG $\mathcal{G}$, Studený et al. (2010) introduce the *characteristic imset* $c_{\mathcal{G}}$, which can be obtained from the standard imset $u_{\mathcal{G}}$ by a one-to-one linear transformation called the Möbius transform (Lauritzen, 1996):

$$c_{\mathcal{G}}(S) = 1 - \sum_{T:S\subseteq T\subseteq\mathcal{V}} u_{\mathcal{G}}(T) \quad \text{and} \quad u_{\mathcal{G}}(T) = \sum_{S:T\subseteq S\subseteq\mathcal{V}} (-1)^{|V\setminus S|} \{1 - c_{\mathcal{G}}(S)\}.$$

These transforms are just inclusion and exclusion formulae. This gives another representation of the equivalence classes of DAGs, and (Studený et al., 2010), provide the following theorem.

Theorem 3.3.1. *For a DAG $\mathcal{G}$, we have:*

(i) $c_{\mathcal{G}}(S) \in \{0, 1\}$ *for each* $S \subseteq \mathcal{V}$, *and*

(ii) $c_{\mathcal{G}}(S) = 1$ *if and only if either* $S = \emptyset$ *or there exists some* $i \in S$ *with* $S \setminus \{i\} \subseteq \mathrm{pa}_{\mathcal{G}}(i)$.

Inspired by the relation in DAGs between standard imsets and characteristic imsets, and the consistency between characteristic imsets and parametrizing sets, we extend the definition of the characteristic imset to MAGs using the parametrizing set. The previous Möbius transform used on $u_{\mathcal{G}}$ to obtain $c_{\mathcal{G}}$ is linear and thus has an inverse form (we will see later), we then apply this inverse formulae on $1 - c_{\mathcal{G}}$, to obtain the 'standard' imsets, $u_{\mathcal{G}}$. We will show that for many MAGs, these 'standard' imsets are combinatorial and perfectly Markovian with respect to the original graph $\mathcal{G}$.

Example 3.3.2. Consider the graph $\mathcal{G}$ in Figure 3.1 (ii). This is an example from Richardson (2009). There are two absent edges, each of them corresponds to a conditional independence, specifically $1 \perp\!\!\!\perp 3$ and $2 \perp\!\!\!\perp 4 \,|\, 1, 3$. Clearly we want a standard imset $u_{\mathcal{G}}$ such that it is sum of the elementary imsets for these two independences. So we want

$$\begin{aligned} u_{\mathcal{G}} &= u_{\langle 1,3\rangle} + u_{\langle 2,4|13\rangle} \\ &= \delta_{1234} - \delta_{134} - \delta_{123} + 2\delta_{13} - \delta_3 - \delta_1 + \delta_\emptyset. \end{aligned}$$

The fourth term δ_{13} with coefficient 2 is quite interesting. Graphically these vertices are nonadjacent; however, $\{1\}$ is the tail for the head $\{3, 4\}$ that contains 3, and conversely $\{3\}$ is the tail for the head $\{1, 2\}$ that contains 1.

3.3.3 Standard imsets of MAGs

The fact that the parametrizing sets and the characteristic imsets of DAGs agree motivates our definition for the characteristic imsets of MAGs, then we work backwards to deduce the form of the standard imsets of MAGs.

Definition 3.3.1. For a MAG $\mathcal{G}$, define its characteristic imset $c_{\mathcal{G}}$ as $c_{\mathcal{G}}(S) = 1$ if $S \in \mathcal{S}(\mathcal{G})$ and $c_{\mathcal{G}}(S) = 0$ otherwise. Moreover, we define $c_{\mathcal{G}}(\emptyset) = 1$ by convention and in order to preserve the characteristic imset for DAG models.

Theorem 3.3.2. *For a MAG $\mathcal{G}$ with characteristic imset $c_{\mathcal{G}}$, let*

$$u_{\mathcal{G}}(B) = \sum_{A: B \subseteq A \subseteq \mathcal{V}} (-1)^{|A \setminus B|} (1 - c_{\mathcal{G}}(A)).$$

Then

$$u_{\mathcal{G}} = \delta_{\mathcal{V}} - \delta_\emptyset - \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup \mathrm{tail}(H)}.$$

We will refer to $u_{\mathcal{G}}$ as the 'standard' imset of the MAG $\mathcal{G}$, the quotes being in recognition of the fact that it does not always define the model (see Example 3.3.6).

Proof. We can obtain $c_{\mathcal{G}}$ via the following transformation from $u_{\mathcal{G}}$:

$$c_{\mathcal{G}}(S) = 1 - \sum_{T: S \subseteq T \subseteq \mathcal{V}} u_{\mathcal{G}}(T).$$

Then we can prove the theorem by showing that, after substituting

$$u_{\mathcal{G}} = \delta_{\mathcal{V}} - \delta_{\emptyset} - \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup \text{tail}(H)}$$

to the RHS of the previous equality, the result (say $c^*_{\mathcal{G}}$) is the same as $c_{\mathcal{G}}$.

Note that $c^*_{\mathcal{G}}(\emptyset) = 1$, as the sum of coefficients in $u_{\mathcal{G}}$ is 0. Suppose $S \neq \emptyset$ and let $\mathbb{1}_P$ denote an indicator function, taking value 1 if P is true and 0 otherwise. Then

$$\begin{aligned}
c^*_{\mathcal{G}}(S) &= 1 - \sum_{T: S \subseteq T \subseteq \mathcal{V}} \left\{ \delta_{\mathcal{V}}(T) - \delta_{\emptyset}(T) - \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup \text{tail}(H)}(T) \right\} \\
&= \sum_{T: S \subseteq T \subseteq \mathcal{V}} \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup \text{tail}(H)}(T) \\
&= \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{T: S \subseteq T \subseteq \mathcal{V}} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup \text{tail}(H)}(T) \\
&= \sum_{H \in \mathcal{H}(\mathcal{G})} \sum_{T: S \subseteq T \subseteq \mathcal{V}} (-1)^{|H \setminus T|} \mathbb{1}_{\text{tail}(H) \subseteq T \subseteq H \cup \text{tail}(H)} \\
&= \sum_{H \in \mathcal{H}(\mathcal{G})} \mathbb{1}_{S \subseteq H \cup \text{tail}(H)} \sum_{K \subseteq H \setminus S} (-1)^{|K| + |H \setminus S|} \\
&= \sum_{H \in \mathcal{H}(\mathcal{G})} \mathbb{1}_{H \subseteq S \subseteq H \cup \text{tail}(H)}.
\end{aligned}$$

The fifth equality can be seen in the following way: for each S and H we are counting the number of supersets $T \supseteq S$ such that $\text{tail}_{\mathcal{G}}(H) \subseteq T \subseteq H \cup \text{tail}_{\mathcal{G}}(H)$, multiplied by some constant $(-1)^{|H \setminus T|}$. The indicator function comes from the fact that if S is not a subset of $H \cup \text{tail}_{\mathcal{G}}(H)$ then there is no such T. Now S can be partitioned into two sets $S = S_1 \dot{\cup} S_2$ which are subsets of H and $\text{tail}_{\mathcal{G}}(H)$, respectively. For any set $T \supseteq S$ such that $\text{tail}_{\mathcal{G}}(H) \subseteq T \subseteq H \cup \text{tail}_{\mathcal{G}}(H)$, It can be partitioned into three sets: $\text{tail}_{\mathcal{G}}(H) \supseteq S_2$, S_1 and K. The last one K is (any) subset of $H \setminus S$ and the first two sets are deterministic. Moreover, $(-1)^{|H \setminus T|}$ is then equal to $(-1)^{|(H \setminus S) \setminus K|} = (-1)^{|K| + |H \setminus S|}$.

Now the result follows from the proof of Lemma 4.3 in Evans and Richardson (2013) which shows that for each set A there is at most one head H such that $H \subseteq A \subseteq H \cup \text{tail}_{\mathcal{G}}(H)$. □

If $\mathcal{G}$ is a DAG, then $u_{\mathcal{G}}$ in Theorem 3.3.2 agrees with (3.1).

Remark 4. Notice that the form of standard imsets in Theorem 3.3.2 considers a tail with subsets of its head, in an opposite manner to the parametrizing sets where we consider a head with subsets of its tail. One can check that this is how $2\delta_{13}$ is obtained in Example 3.3.2. One δ_{13} comes from the head $\{1,2\}$ with tail $\{3\}$ and the head $\{3,4\}$ with tail $\{1\}$ contributes another δ_{13}.

Corollary 3.3.2.1. *For two MAGs $\mathcal{G}$ and $\mathcal{H}$, they are Markov equivalent if and only if $u_{\mathcal{G}} = u_{\mathcal{H}}$.*

Proof. This follows from Theorem 2.2.1 and the fact that the transformation between the standard imset and the characteristic imset is one-to-one. □

Next, we prove a useful result on 'standard imset' of subgraphs, which says that if the 'standard imset' of a subgraph induced by an ancestral set does not define the correct model then neither does the whole graph.

3.3.4 Forbidden subgraphs

A useful fact about MAGs is that any induced subgraph of a MAG is itself a MAG.

Lemma 3.3.3. *Let $\mathcal{G}$ be a graph that is maximal and ancestral. Then for any subset of the vertices $W \subseteq V$, so is the induced subgraph $\mathcal{G}_W$.*

Proof. The ancestrality follows from Proposition 3.5 of Richardson and Spirtes (2002). For maximality, note that there are no more paths in an induced subgraph than in the original graph, so in particular there cannot be any more inducing paths. Since all bidirected edges between vertices are preserved, this implies that the graph remains maximal. □

Another useful fact will concern conditional independences in induced subgraphs.

Proposition 3.3.4. *Let $\mathcal{G}$ be a MAG and $\mathcal{G}_W$ an induced subgraph. Then for $a, b \in W$, any m-separation $a \perp_m b \mid C$ holding in $\mathcal{G}$ implies that $a \perp_m b \mid E$ holds in $\mathcal{G}_W$, where $E = C \cap W$. Additionally, if W is an ancestral set, then $a \perp_m b \mid E$ in $\mathcal{G}_W$ if and only if it also holds in $\mathcal{G}$.*

Proof. Removing vertices that are not on the path from the conditioning set can only block a path, so the removal of vertices in $C \setminus E$ will not affect the status of any of the paths through W. Hence the result holds. The result for ancestral subgraphs follows, for example, from the results of Richardson (2003). □

The above result is also useful for considering the parametrizing set. Let $\mathcal{P}(W)$ denote the *power set* of W, i.e. the collection of subsets of W.

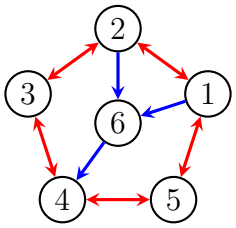


Figure 3.2: A counter example for Prop 3.3.5 when W is not ancestral

Corollary 3.3.4.1. *We have* $\mathcal{S}(\mathcal{G}_W) \subseteq \mathcal{S}(\mathcal{G}) \cap \mathcal{P}(W)$.

Proof. This follows immediately from Propositions 2.2.2 and 3.3.4. □

Note that, of course, the sets of size at most two are always identical in the original MAG and any induced subgraph, since these are just the adjacencies.

Now we are ready to prove the main result of this section that the certain graphs must not appear as induced subgraphs within a MAG, if we want the corresponding 'standard' imset to be perfectly Markovian with respect to it.

Proposition 3.3.5. *Let* $\mathcal{G}$ *be a MAG, and suppose that for some ancestral subset* $A \subset V$ *we have that* $u_{\mathcal{G}_A}$ *is not Markovian with respect to* $\mathcal{G}_A$*. Then the model* $u_{\mathcal{G}}$ *is not perfectly Markovian with respect to* $\mathcal{G}$*.*

Proof. Since the subgraph is ancestral, the independences it encodes are just those from the larger graph, with potentially smaller conditioning sets. Now, suppose that it is not possible to write the independences implied by the smaller graph in such a way as to avoid repeating a set. Then the same problem will clearly arise in the larger graph, since (by Proposition 3.3.4) we must specify isomorphic independences with potentially more restrictions on the conditioning set. □

Remark 5. We know from Proposition 3.3.4 that if we take an ancestral subset $W = A$, the structure of the imset is preserved. Hence, if we marginalize the graph then the imset will just be the induced subimset over the entries that are subsets of A, and so by Proposition 9.3 of Studený (2006), the imset will match the model for the ancestral subgraph. Hence, in this case the result is clear.

The following example shows that Proposition 3.3.5 does not work if a subset W is not ancestral. This is pointed out by one of the reviewers.

Example 3.3.3. One can check that $u_{\langle 4,12|6\rangle} + u_{\langle 6,35|12\rangle} + u_{\langle 1,3|5\rangle} + u_{\langle 2,5|3\rangle} + u_{\langle 3,5\rangle}$ is the standard imset of the MAG in Figure 3.2 and is perfectly Markovian with respect to it. The subgraph induced by $\{1, 2, 3, 4, 5\}$ is the bidirected 5-cycle, however, and hence the 'standard' imset does not work.

For the graph, if we were to interpret it using the nested property (Richardson et al., 2017) then, after fixing 6, we would have two additional constraints: $4 \perp\!\!\!\perp 1 \mid 3$ and $4 \perp\!\!\!\perp 2 \mid 5$. This model *cannot* be rewritten in a manner that avoids overlap, and therefore if the imset represented the nested model, this graph would not have a perfectly Markovian 'standard' imset. Of course, the imset *does not* represent the nested model, and so this is merely an academic point.

Before we proceed, we would like to clarify some of the terms and notations used in this thesis. Originally, the standard imset of a DAG refers to the fact that it is the simplest imset that is perfectly Markovian w.r.t. the graph (Studený, 2006). However for MAGs, the 'standard' imset we defined is not necessarily perfectly Markovian, and we use the quotes to refer to the imset obtained by applying the Möbius inversion formula to the 0-1 imset defined by the parametrizing set.

If we use c with some subscript to denote an imset then it is in the characteristic form, i.e. obtained by applying the Möbius transform to some imset u. If u is structural then it induces a model, and we associate the same model with both c and u.

3.3.5 Choice of the characteristic imset

Obviously, there are many imset representations for an independence model. The reason we choose to firstly work with a 0-1 characteristic imset is because if its corresponding 'standard' imset *is* perfectly Markovian w.r.t. the graph, then it is also the minimal representation. We proceed to prove this fact by studying the characteristic imset form c of any structural imset u that represent the model.

Example 3.3.7 gives a graph for which the 'standard' imset is not structural. Fortunately, such graphs seem to be comparatively unusual as shown in experiment section..

We begin with a simple observation on the linear relationship between structural imset and its characteristic imset.

Lemma 3.3.6. *For a structural imset u and its characteristic imset c, we have that $c(S) = 1$ if and only if u contains no independence with S as a constrained set, and otherwise $c(S) \leq 0$.*

Corollary 3.3.6.1. *Consider a MAG $\mathcal{G}$. For any structural imset u such that $\mathcal{I}_u = \mathcal{I}_\mathcal{G}$, its characteristic imset c must be an integer valued vector and satisfy the following:*

(i) if $S \in \mathcal{S}(\mathcal{G})$, then $c_\mathcal{G}(S) = 1$;

(ii) if $S \notin \mathcal{S}(\mathcal{G})$*, then* $c_{\mathcal{G}}(S) \leq 0$.

We now argue for the choice of the 0-1 characteristic imset, which Corollary 3.3.6.1 shows is unique for any given graph. There is no way to totally order combinatorial imsets, though the degree does provide a partial ordering. It is clear that, if our imset is combinatorial and perfectly Markovian, then it is also an *imset of the smallest degree* (Studený, 2006), because there is exactly one independence for each missing edge. We will see in Example 3.3.6 that some MAG models cannot be represented in this way.

We can define a partial order on imsets defining the same model based on the coefficients of their corresponding characteristic imsets. That is, given two structural imsets u and u' with corresponding characteristic imsets c and c', we say that u *is smaller than* u' if $c(S) \geq c'(S)$ for every $S \subseteq \mathcal{P}(V)$. Lemma 3.3.6 makes clear that this is a useful definition.

By Corollary 3.3.6.1, if the 0-1 imset's standard imset induces the same model as the graph, it is both an imset of smallest degree and the unique minimal such imset according to this partial order.

3.3.6 Simple MAGs

Unlike the standard imset of a DAG, it is often very hard to tell how to decompose this standard imset as sum of semi-elementary imsets, because the size and number of heads are arbitrary. However, if we restrict the size of heads to two, we can show that its standard imset is always both combinatorial and perfectly Markovian with respect to $\mathcal{G}$.

Definition 3.3.2. A MAG $\mathcal{G}$ is said to be *simple* if $\mathcal{G}$ contains no head with size more than two.

Before we prove that the 'standard imset' of any simple MAG is always perfectly Markovian w.r.t. the graph, we show how dense simple MAGs are as a subset of all MAGs by: (i) listing how many Markov equivalence classes contain at least one simple MAG; and (ii) simulating MAGs and counting proportions of them being Markov equivalent to some simple MAGs. Note that simple MAGs can have large districts and not be Markov equivalent to a graph with smaller districts; see Figure 3.3.

Example 3.3.4. We first start by providing an example of a simple MAG with an arbitrarily large district. This illustrates that a search algorithm over simple MAGs is potentially very useful, since it includes a considerable number of causal models

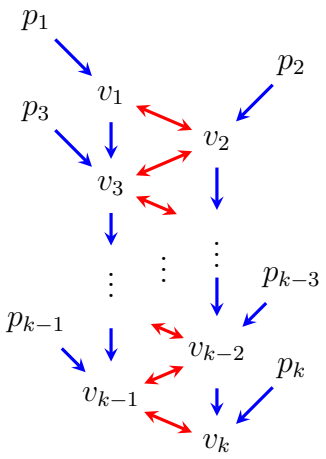


Figure 3.3: Simple MAGs with arbitrarily large districts

than one would obtain by restricting the maximum district size to two or three. This is in contrast with other methods for score-based learning, which generally make this kind of restriction (e.g. Chen et al., 2021).

The graph in Figure 3.3 has a district of size k and that district also has k parents. However, note that the only heads are just the bidirected edges, so this is a simple MAG. In addition, the independences for each of the V_j are of the form

$$V_j \perp\!\!\!\perp P_1, V_1, \ldots, V_{j-3}, P_{j-2}, P_{k-1} \mid P_j, V_{j-2},$$

and (unsurprisingly) there is no way to order the variables so that the independences are nested within one another as would be required by a DAG.

We demonstrate how common simple MAGs are by using the following results. The first column in Table 3.1 is the number vertices. Then the second column list the number of equivalence classes of MAGs with the corresponding number of vertices. The next two columns further count how many equivalence classes that contain at least one simple MAG and one DAG, respectively. In particular, the proportion of equivalence classes that contain simple MAGs decreases but not as sharply as that for DAGs.

Then Figure 3.4, we simulate 1000 random MAGs for number of vertices ranging from five to forty, and plot empirical probabilities that the simulated graphs are Markov equivalent to some simple MAGs. The method we use to simulate MAGs is the same as Claassen and Bucur (2022). We fix the average and maximal degree of each vertex to three and ten respectively. For each edge, the probability of being bidirected is 0.2. We simulate an ADMG first and project it into a Markov equivalent MAG. Then by converting the MAG to a *partially ancestral graph* (PAG),

we finally select a representative MAG from the PAG. The last two steps are from Zhang (2007a). If the representative MAG is simple, we consider the original graph as being Markov equivalent to some simple MAGs. Note that the probability of simple MAGs drops quickly as number of variables increase. This is because as size of graphs grows, it is more likely to have a subgraph that is not simple.

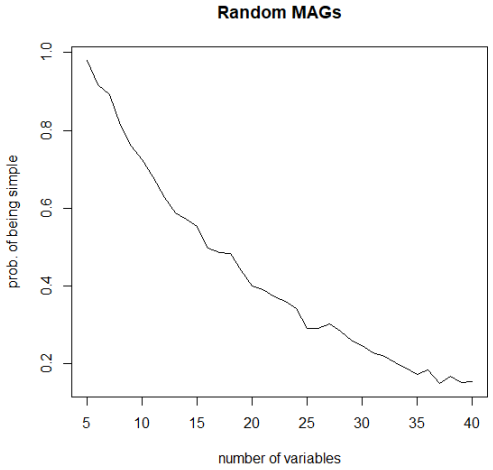


Figure 3.4: A plot for probability of random graphs being Markov equivalent to some simple MAGs

$\lvert V\rvert$	equiv. classes	simple MAGs	DAGs
5	285	205	119
6	13,303	6,278	2,025
7*	1,161,461	331,310	57,661

*having at most 13 or at least 18 edges.

Table 3.1: Number of equivalence classes

It is not hard to show that for a MAG, $\{i, j\}$ is a head if and only if $i \leftrightarrow j$. Moreover, consider any $j \leftrightarrow i \leftrightarrow k$. If the MAG is simple, then there must be ancestral relations between j and k so that $\{i, j, k\}$ does not form a head (this is necessary and sufficient as shown by Lemma 7.3 of Evans and Richardson, 2013). Hence for each vertex i, there is a total ordering on heads that contain i, and we now show that their tails are nested within one another.

Lemma 3.3.7. *Suppose $\mathcal{G}$ is a simple MAG with a given topological ordering. For every vertex i and all heads of size two, $\{i, j_s\}$, with $j_1 < \cdots < j_k \leq i$, we have* $\mathrm{pa}_{\mathcal{G}}(i) = \mathrm{tail}_{\mathcal{G}}(i) \subseteq \mathrm{tail}_{\mathcal{G}}(i, j_1) \subseteq \cdots \subseteq \mathrm{tail}_{\mathcal{G}}(i, j_k)$.

Example 3.3.5. Consider the simple MAG $\mathcal{G}$ in Figure 3.5. The sets $\{7, 8\}$ and $\{6, 8\}$ are the heads of size two associated with the vertex 8. Their tails are $\{1, 2, 3, 4, 5, 6\}$ and $\{1, 3, 4, 5\}$ respectively. Note that $\mathrm{tail}_{\mathcal{G}}(\{6, 8\}) \subset \mathrm{tail}_{\mathcal{G}}(\{7, 8\})$. Also, we have $\mathrm{pa}_{\mathcal{G}}(8) = \{1, 5\}$, which is a subset of both the other tails.

One can check that its standard imset can be written as:

$$\begin{aligned} u_{\mathcal{G}} = {} & u_{\langle 8,2|13456\rangle} + u_{\langle 8,34|15\rangle} + u_{\langle 7,1345|26\rangle} \\ & + u_{\langle 6,125|34\rangle} + u_{\langle 6,3\rangle} + u_{\langle 5,123|4\rangle} \\ & + u_{\langle 4,12|3\rangle} + u_{\langle 3,12\rangle} + u_{\langle 1,2\rangle}. \end{aligned}$$

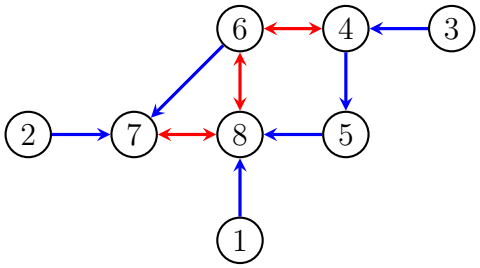


Figure 3.5: A simple MAG

Let us focus on the semi-elementary imsets that contain 8. The conditional independence between vertex 8 and vertex 2 can be seen in this way. After marginalizing $\{7\}$ (the remaining seven vertices still form an ancestral set), $\{2\}$ would be outside of the Markov blanket of vertex 8 which is $\text{tail}_{\mathcal{G}}(\{6,8\}) = \{1,3,4,5,6\}$, and this independence is implied by the local Markov property. Similarly, if one further marginalizes $\{6\}$, then $\{3,4\}$ are not in the Markov blanket of vertex 8, which is just $\text{pa}_{\mathcal{G}}(8) = \{1,5\}$, and this corresponds to $8 \perp\!\!\!\perp 3,4 \mid 1,5$.

Thus for simple MAGs with a given topological ordering, we can obtain conditional independences by sequentially marginalizing heads of size two associated with the last vertex and using local Markov property. See details in the proof of the following theorem.

Theorem 3.3.8. *For a MAG $\mathcal{G}$ and its standard imset $u_{\mathcal{G}}$, if $\mathcal{G}$ is simple then $\mathcal{I}_{u_{\mathcal{G}}} = \mathcal{I}_{\mathcal{G}}$.*

Proof. Given a topological ordering, we can write the standard imset $u_{\mathcal{G}}$ of a simple MAG $\mathcal{G}$ in the following way:

$$
\begin{aligned}
u_{\mathcal{G}} &= \delta_{\mathcal{V}} - \delta_{\emptyset} - \sum_{i\in\mathcal{V}}\left\{\delta_{\{i\}\cup\text{pa}(i)} - \delta_{\text{pa}(i)}\right\} - \sum_{i\leftrightarrow j}\left\{\delta_{\{i,j\}\cup\text{tail}(i,j)} - \delta_{\{i\}\cup\text{tail}(i,j)} - \delta_{\{j\}\cup\text{tail}(i,j)} + \delta_{\text{tail}(i,j)}\right\};\\
&= \sum_{i\in\mathcal{V}}\left\{\delta_{[i]} - \delta_{[i-1]} - \delta_{\{i\}\cup\text{pa}(i)} + \delta_{\text{pa}(i)}\right\} - \sum_{i\leftrightarrow j}\left\{\delta_{\{i,j\}\cup\text{tail}(i,j)} - \delta_{\{i\}\cup\text{tail}(i,j)} - \delta_{\{j\}\cup\text{tail}(i,j)} + \delta_{\text{tail}(i,j)}\right\};\\
&= \sum_{i\in\mathcal{V}}\Bigg\{\delta_{[i]} - \delta_{[i-1]} - \delta_{\{i\}\cup\text{pa}(i)} + \delta_{\text{pa}(i)}\\
&\quad + \sum_{i\leftrightarrow j, i>j} -\delta_{\{i,j\}\cup\text{tail}(i,j)} + \delta_{\{i\}\cup\text{tail}(i,j)} + \delta_{\{j\}\cup\text{tail}(i,j)} - \delta_{\text{tail}(i,j)}\Bigg\}
\end{aligned}
$$

For each vertex i, consider the topological ordering on all the j such that $j < i$ and $i \leftrightarrow j$, where $j_1 < \ldots < j_k < i$. Hence we have:

$$
u_{\mathcal{G}} = \sum_{i\in\mathcal{V}}\Bigg\{\delta_{[i]} - \delta_{[i-1]} - \delta_{\{i\}\cup\text{pa}(i)} + \delta_{\text{pa}(i)}
$$

$$
\begin{aligned}
&+\sum_{l=1}^{k} -\delta_{\{i,j_l\}\cup\mathrm{tail}(i,j_l)} + \delta_{\{i\}\cup\mathrm{tail}(i,j_l)} + \delta_{\{j_l\}\cup\mathrm{tail}(i,j_l)} - \delta_{\mathrm{tail}(i,j_l)}\Bigg\} \\
=&\sum_{i\in\mathcal{V}}\Bigg\{\delta_{[i]} - \delta_{[i-1]} - \delta_{\{i,j_k\}\cup\mathrm{tail}(i,j_k)} + \delta_{\{j_k\}\cup\mathrm{tail}(i,j_k)} \\
&+\sum_{l=1}^{k-1} -\delta_{\{i,j_l\}\cup\mathrm{tail}(i,j_l)} + \delta_{\{i\}\cup\mathrm{tail}(i,j_{l+1})} + \delta_{\{j_l\}\cup\mathrm{tail}(i,j_l)} - \delta_{\mathrm{tail}(i,j_{l+1})}; \\
&+\delta_{\{i\}\cup\mathrm{tail}(i,j_1)} - \delta_{\mathrm{tail}(i,j_1)} - \delta_{\{i\}\cup\mathrm{pa}(i)} + \delta_{\mathrm{pa}(i)}\Bigg\}.
\end{aligned}
$$

Now for each vertex i, consider the following list of conditional independence, denoted by $\mathbb{L}_i$:

$$
\begin{aligned}
&i \perp\!\!\!\perp [i-1] \setminus (\mathrm{tail}(i,j_k)\cup\{j_k\}) \mid \mathrm{tail}(i,j_k)\cup\{j_k\}; \\
&i \perp\!\!\!\perp \mathrm{tail}(i,j_k) \setminus (\mathrm{tail}(i,j_{k-1})\cup\{j_{k-1}\}) \mid \mathrm{tail}(i,j_{k-1})\cup\{j_{k-1}\}; \\
&\quad\vdots \\
&i \perp\!\!\!\perp \mathrm{tail}(i,j_2) \setminus (\mathrm{tail}(i,j_1)\cup\{j_1\}) \mid \mathrm{tail}(i,j_1)\cup\{j_1\}; \\
&i \perp\!\!\!\perp \mathrm{tail}(i,j_1) \setminus \mathrm{pa}(i) \mid \mathrm{pa}(i).
\end{aligned}
$$

It is straightforward to check that $u_{\mathcal{G}}$ is a sum of the semi-elementary imsets corresponding to the conditional independence list $\mathbb{L} = \bigcup_{i\in\mathcal{V}} \mathbb{L}_i$. Notice that if there is no such j for i (k=0) then it reduces to the local Markov property of DAGs:

$$
i \perp\!\!\!\perp [i-1] \setminus \mathrm{pa}(i) \mid \mathrm{pa}(i).
$$

Thus $u_{\mathcal{G}}$ is a combinatorial imset and every independence in $\mathbb{L}$ is in $\mathcal{I}_{u_{\mathcal{G}}}$.

By Theorem 3.4.5, any local Markov property can be deduced from $\mathbb{L} = \bigcup_i \mathbb{L}_i$ using semi-graphoid thus $\mathcal{I}_{\mathcal{G}} \subseteq \mathcal{I}_{u_{\mathcal{G}}}$. Now the faithfulness result in Richardson and Spirtes (2002) shows that for every MAG $\mathcal{G}$ there exists a distribution P such that $\mathcal{I}_P = \mathcal{I}_{\mathcal{G}}$ and Theorem 5.2 in Studený (2006) implies the existence of a structural imset u such that $I_u = I_P$. Now every independence in $\mathbb{L}$ is in $\mathcal{I}_u$ by Theorem 3.4.5, and $u_{\mathcal{G}}$ is sum of the semi-elementary imsets corresponding to the independence in $\mathbb{L}$, so by Lemma 6.1 in Studený (2006), we have $I_{u_{\mathcal{G}}} \subseteq \mathcal{I}_u = \mathcal{I}_{\mathcal{G}}$. Note the idea of this proof is similar to the proof of Lemma 7.1 in Studený (2006). □

Corollary 3.3.8.1. *For a MAG $\mathcal{G}$, if there exists a simple MAG $\mathcal{H}$ which is Markov equivalent to $\mathcal{G}$, then $\mathcal{I}_{u_{\mathcal{G}}} = \mathcal{I}_{\mathcal{G}}$.*

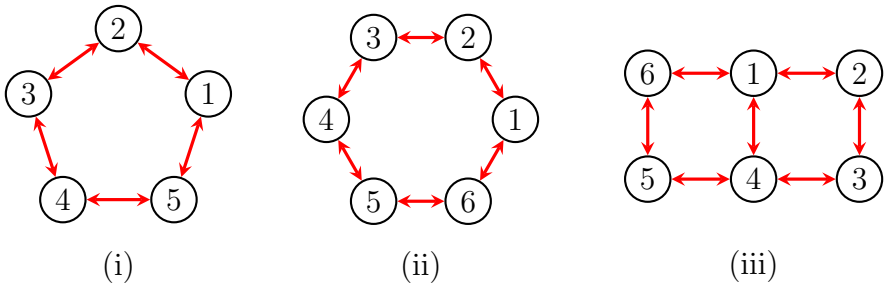


Figure 3.6: (i) a bidirected 5-cycle; (ii) a bidirected 6-cycle; (iii) a bidirected 6-cycle with an additional edge.

3.3.7 Examples where the imset is not perfectly Markovian

In general—as the following examples show—the 'standard' imset we have defined is not structural, and even if it is structural, $u_{\mathcal{G}}$ may not be perfectly Markovian with respect to the model induced by the graph.

Example 3.3.6. Consider the 5-cycle $\mathcal{G}$ with bidirected edges in Figure 3.6. Its 'standard' imset is given by

$$\begin{aligned} u_{\mathcal{G}} &= u_{\langle 1,3|4\rangle} + u_{\langle 2,4|5\rangle} + u_{\langle 3,5|1\rangle} + u_{\langle 4,1|2\rangle} + u_{\langle 5,2|3\rangle} \\ &= u_{\langle 1,3|5\rangle} + u_{\langle 2,4|1\rangle} + u_{\langle 3,5|2\rangle} + u_{\langle 4,1|3\rangle} + u_{\langle 5,2|4\rangle}. \end{aligned}$$

Any (strictly) conditional independence that holds in the graph is contained in $\mathcal{I}_{u_{\mathcal{G}}}$; however, marginal independences such as $1 \perp\!\!\!\perp 3$ are not in the imset.

Note, however, that if we assume the underlying distribution is strictly positive then we are able to deduce that any marginal independence in the graph is in $\mathcal{I}_{u_{\mathcal{G}}}$, by noting that (for example) $1 \perp\!\!\!\perp 3 \mid 4$ and $1 \perp\!\!\!\perp 4 \mid 3$ are both in $\mathcal{I}_{u_{\mathcal{G}}}$, and can be combined if the density is positive to obtain $1 \perp\!\!\!\perp 3, 4$. It turns out that for $n = 5$, this 5-cycle is the only graph such that $u_{\mathcal{G}}$ is not perfectly Markovian with respect to $\mathcal{G}$.

For this 5-cycle, there are 10 sets S such that $c_{\mathcal{G}}(S) = 0$. Among these sets, five of them have size 2 and the other five have size 3. This suggests that for some MAGs, if we want a structural imset $u_{\mathcal{G}}$ that is perfectly Markovian with respect to the graph, it is inevitable that there is some overlap between the sets associated with the independence decomposition of $u_{\mathcal{G}}$. This also shows that the minimum degree of a model defining the 5-cycle is 6, even though there are only 5 pairs of vertices that are not adjacent. Hence there are MAG models which can only be defined by having a set of independences that repeats an independence between some pair of vertices.

Example 3.3.7. Consider the bidirected 6-cycle $\mathcal{G}$ shown in Figure 3.6(ii), and $u_{\mathcal{G}}$, which is too long to list here. Write $u_{\mathcal{G}} = \sum_{A \in \mathcal{P}(\mathcal{V})} \alpha_A \delta_A$, where α_A are coefficients for each identifier imset. The largest disconnected sets are of size 4, and $\sum_{|A|=4} \alpha_A = 9$ and $\sum_{|A|=3} \alpha_A = -22$; It is easy to show that when we add semi-elementary imset to an imset, if the semi-elementary imset contributes to the term with largest set size, we will subtract at most two to the coefficient of the terms with set size exactly one smaller than the largest one. But 22 is larger than $18 = 2 \cdot 9$. Hence this imset cannot be neither combinatorial or structural.

Example 3.3.8. Consider Figure 3.6 (iii). This graph has a structural 'standard' imset that is not perfectly Markovian with respect to $\mathcal{G}$, and it is also not combinatorial. However, it *is* structural, because:

$$\begin{aligned} 2u_{\mathcal{G}} = {} & u_{\langle 1,3\rangle} + u_{\langle 1,3|5,6\rangle} + u_{\langle 1,5\rangle} + u_{\langle 1,5|2,3\rangle} + u_{\langle 2,4\rangle} + u_{\langle 2,4|5,6\rangle} + \\ & + u_{\langle 2,5|1,3\rangle} + u_{\langle 2,5|4,6\rangle} + u_{\langle 2,6\rangle} + u_{\langle 2,6|3,5\rangle} + u_{\langle 3,5\rangle} + u_{\langle 3,5|2,6\rangle} + \\ & + u_{\langle 3,6|2,4\rangle} + u_{\langle 3,6|1,5\rangle} + u_{\langle 4,6\rangle} + u_{\langle 4,6|2,3\rangle}. \end{aligned}$$

This constitutes another example of an imset that is structural but not combinatorial, and it arises in a much more natural way that the one given by Hemmecke et al. (2008).

3.3.8 Relating to scoring criteria

The following inner product notation is defined for scoring purpose.

Definition 3.3.3. Given a function f which takes X_A for any $A \subseteq V$ as input, and an imset u over V, we define

$$\langle u, f \rangle = \sum_{A \subseteq V} u(A) f(x_A).$$

Here we explain how to use imsets to provide a consistent scoring criteria. For DAGs, the maximum likelihood part of the BIC (defined in Section 4.4.2) can be shown to be the empirical entropy of $X_{\mathcal{V}}$, minus the inner product between the standard imset and the empirical entropy vector (take f as the entropy in Definition 3.3.3 and also we demonstrate this in Section 4.4.2 for discrete variables). This inner product is explained in the following.

For random variables $X_{\mathcal{V}}$ with density function p, define the entropy $\mathsf{H}(X_{\mathcal{V}})$ as the expectation of $-\log p(X_{\mathcal{V}})$, i.e. $\mathbb{E}[-\log p(X_{\mathcal{V}})]$. For three random variables X_A, X_B and X_C, the inner product between the semi-elementary imset $u_{\langle A,B|C\rangle}$ and the entropy vector, whose entries correspond to the entropy H of every subset of $\mathcal{V}$, is

the mutual information between X_A, X_B given X_C; that is, $\mathsf{H}(X_{ABC}) - \mathsf{H}(X_{AC}) - \mathsf{H}(X_{BC}) + \mathsf{H}(X_C)$. This quantity is always non-negative, and is zero if and only if the independence $A \perp\!\!\!\perp B \mid C$ holds under p.

Hence for DAGs, BIC scores can be interpreted as the discrepancy for a list of independences from the ordered local Markov property plus penalty terms for model complexity. It follows that we can do something similar for simple MAGs, since if $u_{\mathcal{G}}$ is perfectly Markovian w.r.t. the graph, this inner product (suitably penalized) provides a valid score. In fact, we prove in Section 4.4.3 that if $\mathcal{I}_{u_{\mathcal{G}}} = \mathcal{I}_{\mathcal{G}}$ then, this score is consistent as it approximates the BIC. However, for simple MAGs this score can be obtained much faster as the imsets can be constructed in quadratic time in the number of vertices; in contrast, there is no guarantee on computation time for BIC. Note that consistency of the BIC score only requires that the data are generated from one of the models being scored (and do not coincidentally lie on any other models with at most the same number of parameters). For this score, we also require that the graph that generates the data has a perfectly Markovian standard imset.

Moreover, as we have shown the decomposition (with positive integer coefficients) of simple MAGs, if the CI relations in a distribution can be described by a simple MAG, we always have that this inner product is zero. Empirically we observed that this is true even for general MAGs, in spite of previous examples where our imsets are not perfectly Markovian w.r.t. MAGs or not even structural. This suggests that the standard imsets we defined *can* be expressed as the sum of semi-elementary imsets corresponding to conditional independences advertised by the graph, but some of the coefficients are negative, this will indeed turn out to be the case (Theorem 3.4.18).

3.3.9 Motivations to simplify Markov property

Next we discuss different choices of imsets. Obviously there are many imsets that represent the same model induced by graphs, but they are different in terms of computational complexity and statistical performance. Both the pairwise Markov property (Sadeghi et al., 2014) and the (reduced) ordered local Markov property (Richardson, 2003) can be used to construct imsets by summing semi-elementary imsets corresponding to list of independences[2].

However, the pairwise Markov property in general have more conditioning variables if graphs have complicated ancestral relations, and thus require to estimate

[2]For the pairwise Markov property to be equivalent to the global Markov property it requires that the independence model is a 'compositional graphoid' (see Appendix 3.8). One can show the inner product with entropy is zero if and only if the distribution obeys the pairwise Markov property.

entropy of more variables, which is hard problem. Even if it has the same degree as the standard imsets for simple MAGs, in practice we found it having worse performance in model search algorithms. For imsets corresponding to the (reduced) ordered local Markov property, since it in general contains redundant conditional independences (even for simple MAGs, see the example on next section), it has higher degree compared to the standard imsets, hence are less useful. Therefore for simple MAGs, our standard imsets are faster to compute (polynomial time) and more reliable.

Now for general MAGs, we have seen that the 'standard' imsets obtained from the 0-1 characteristic imsets in general are invalid. To construct valid imsets one can of course use the imsets from the pairwise/reduced ordered local Markov property. However, we aim for imsets that have lower degree and fewer conditioning variables, hence simplifying the Markov property is essential to find the standard (simplest) imsets for general MAGs. In the next section we use a graphical tool called the Power DAGs to achieve this, and we show that our *refined Markov property* is strictly simpler than the ordered local Markov property. The result is optimal for simple MAGs as it gives the list of independences that decomposes the standard imsets, but not optimal in general.

Our refined Markov property results in an imset that is perfectly Markovian to a given MAG and have fewer degree/conditioning variables compared to pairwise/ordered local Markov property. We also show that fixing the maximal head size, the computational time for this imset is polynomial.

3.4 Power DAGs

We aim to give a simpler representation of the conditional independence model induced by MAGs. To describe these relations, we use *power DAGs* of these graphs, which are DAGs over the set of heads.

Our power DAG is completely analogous to the *intrinsic power DAG* in (Richardson et al., 2023), as there is a one-to-one correspondence between the collections of head and *intrinsic sets* used in that paper; indeed, for most MAGs, the two graphs are isomorphic. The approach we give later to simplifying the power DAG is not suitable for intrinsic power DAGs for the *nested Markov* model studied in that paper, because the order in which vertices are *fixed* (marginalized in our case) is important for deriving nested constraints.

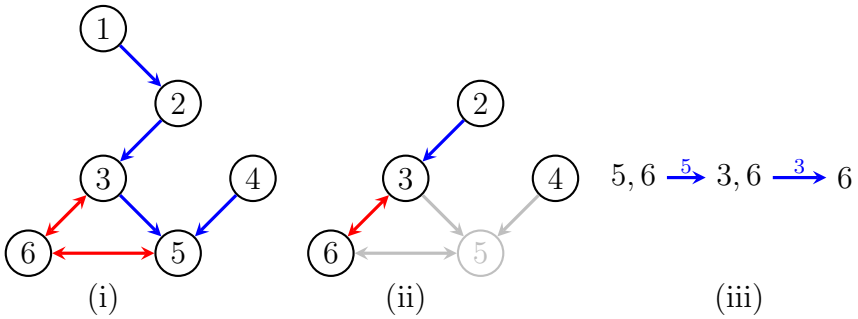


Figure 3.7: (i) A simple MAG $\mathcal{G}$; (ii) the graph after removing 1 and marginalizing 5; (iii) a DAG on heads in $\mathcal{G}$ that contain the vertex 6.

3.4.1 Motivations and examples

Example 3.4.1. Consider the simple MAG in Figure 3.7(i), given the numerical ordering, the reduced ordered local Markov property for the vertex 6 would require:

$$6 \perp\!\!\!\perp 1 \mid 2,3,4,5 \qquad 6 \perp\!\!\!\perp 1,4 \mid 2,3 \qquad 6 \perp\!\!\!\perp 1,2,4.$$

However, these independences are equivalent to:

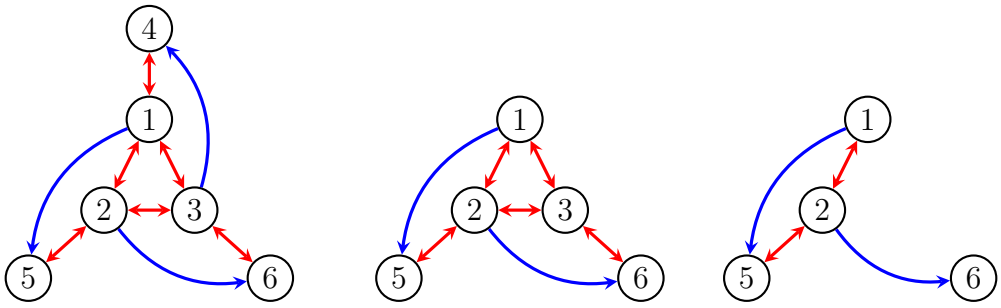
$$6 \perp\!\!\!\perp 1 \mid 2,3,4,5 \qquad 6 \perp\!\!\!\perp 4 \mid 2,3 \qquad 6 \perp\!\!\!\perp 2.$$

This list is the simplest as it comes from decomposition of standard imset $u_{\mathcal{G}}$ of the simple MAG $\mathcal{G}$.

We start with a topological ordering, in this case the numeric ordering of the vertices. Then, for each vertex we consider the graph of its predecessors; for the vertex 6 this is just the graph itself. We see from this that 1 is not in the Markov blanket of 6, deduce that $6 \perp\!\!\!\perp 1 \mid 2,3,4$ holds, so we remove 1 from the graph. Then we must marginalize something other than 6 in the head $\{5,6\}$; for simple MAGs there is only ever one choice, so we obtain the graph in Figure 3.7(ii). The maximal head in this graph is now $\{3,6\}$, and we now see that $\{4\}$ is no longer in the Markov blanket of 6; hence we obtain $6 \perp\!\!\!\perp 4 \mid 2,3$ and remove it from further consideration. Finally we marginalize 3 and see that $6 \perp\!\!\!\perp 2$.

Alongside these operations, in Figure 3.7(iii) we construct the corresponding component of the power DAG for heads that contain 6. We (potentially) associate a single independence with the initial node (so $6 \perp\!\!\!\perp 1 \mid 2,3,4,5$ in our case) and another with each of the transitions in the graph ($6 \perp\!\!\!\perp 4 \mid 2,3$ and $6 \perp\!\!\!\perp 2$).

Specifically, in this example, we reach the head $\{3,6\}$ from the head $\{5,6\}$ by marginalizing $\{5\}$, as illustrated in Figure 3.7. Now $\{4\}$ is no longer in the Markov blanket of 6, so we obtain $6 \perp\!\!\!\perp 4 \mid 2,3$ for it and remove it from consideration. Next by marginalizing $\{3\}$, we reach the head $\{6\}$ and now $\{2\}$ is not in the Markov



(i):$\mathcal{G} = \mathcal{G}_{\text{an}(\{4,5,6\})}$ (ii): Marginalize 4 $\mathcal{G}' = \mathcal{G}_{\text{an}(\{4,5,6\})\setminus\{4\}} = \mathcal{G}_{\text{an}(\{3,5,6\})}$ (iii): Marginalize 3 $\mathcal{G}'' = \mathcal{G}_{\text{an}(\{3,5,6\})\setminus\{3\}}$

Figure 3.8: An example for Definition 3.4.1.

blanket of $\{6\}$ anymore, so we add the independence $6 \perp\!\!\!\perp 2$. For better illustration, we can draw a DAG on heads if one head can reach another by marginalizing vertices in the head, see Figure 3.7 (iii). Then the second and third independences in our list are represented by this DAG.

Now we give formal definitions of power DAGs and a more complicated example on non-simple MAGs will be given.

3.4.2 From one head to another

We assume throughout that the numerical ordering of vertices is also topological. To properly formulate the definition of power DAGs, we need a few more notations and definitions. For a vertex set A and a vertex i, we write $A \leq i$ if i is the *maximal* vertex in A w.r.t. a given topological ordering.

Definition 3.4.1. For a MAG $\mathcal{G}$ and two heads $H, H' \leq i$, we write $H \to^K H'$ $(\emptyset \subset K \subseteq H \setminus \{i\})$ if $\text{barren}_{\mathcal{G}'}(\text{dis}_{\mathcal{G}'}(i)) = H'$, where $\mathcal{G}' = \mathcal{G}_{\text{an}(H)\setminus K}$. We will refer to K as a marginalization set.

Graphically, $H \to^K H'$ means that in the subgraph $\mathcal{G}_{\text{an}(H)}$, the maximal head (i.e. the barren subset of the district) that contains i after marginalizing K (subset of the barren subset) is H'. Moreover, to save space we eschew set notation and union signs and write (e.g.) k for $\{k\}$ and HT for $H \cup T$.

Take the graph in Figure 3.8(i) , under the numerical topological ordering. Consider the final vertex 6; other barren vertices in its Markov blanket are 4 and 5, which together with 6 form a head. By marginalizing 4, we reach the head $\{3,5,6\}$, hence by definition, $\{4,5,6\} \to^4 \{3,5,6\}$. Then if we marginalize 3, we have $\{3,5,6\} \to^3 \{6\}$. The above two marginalization steps are shown in Figure 3.8 (ii) and (iii).

With each $H \to^K H'$ we associate a conditional independence. By Definition 3.4.1, if $H \to^K H'$ then $\mathrm{an}_{\mathcal{G}}(H) \setminus K = B$, where B is an ancestral set, $T' = \mathrm{tail}_{\mathcal{G}}(H')$ and $\{i\} \cup \mathrm{mb}_{\mathcal{G}}(i, B) = H' \cup T'$. Hence by the ordered local Markov property and marginalizing vertices that lie outside of the Markov blanket of i in $\mathcal{G}_B$, we have

$$i \perp\!\!\!\perp (H \cup T) \setminus (H' \cup T' \cup K) \mid (H' \cup T') \setminus \{i\}.$$

For instance, in Figure 3.8, the conditional independence associated with the edge $\{3,5,6\} \to^3 \{6\}$ is obtained as the following: we have $H = \{3,5,6\}, T = \mathrm{tail}(H) = \{1,2\}, H' = \{6\}, T' = \mathrm{tail}(H') = \{2\}$ and $K = \{3\}$, hence

$$6 \perp\!\!\!\perp (\{3,5,6\} \cup \{1,2\}) \setminus (\{6\} \cup \{2\} \cup \{3\}) \mid \{6\} \cup \{2\} \setminus \{6\} = 6 \perp\!\!\!\perp 1,5 \mid 2.$$

3.4.3 Complete power DAGs

Definition 3.4.2. Consider a MAG $\mathcal{G}$ with a topological ordering. Given a set $S \subseteq \mathcal{V}$ we say that $s \in S$ is a *marginalization vertex* if it is in $\mathrm{barren}_{\mathcal{G}}(S)$ and is not maximal in S.

Define the *complete power DAG* $\mathfrak{I}(\mathcal{G})$ as a graph with vertices $\mathcal{H}(\mathcal{G})$. An edge is added from $H \to H'$ if there is a marginalization vertex $k \in H$ such that $H \to^k H'$. In this case we call H a *parent head* of H'. There is a unique component for each vertex i, which we denote $\mathfrak{I}_i(\mathcal{G})$.

Next we present a few results that justify the nomenclature of power DAGs.

Lemma 3.4.1. *For a MAG $\mathcal{G}$ and two heads H, H' whose maximal vertex is i, if $H \to^K H'$ and $H \to^L H'$ then $H \to^{K \cap L} H'$.*

Proof. Let $T = K \cap L$ and $\dot{K} = K \setminus T$. Suppose $H \to^T H''$. Now after marginalizing T, the vertices in $\dot{K}$ are either not in the same district as i or lie in H''. If all of them are not in the same district with i, then $H' = H''$ because then marginalizing $\dot{K}$ would not change the barren subset of the district containing i, which are H'' and H' before and after the marginalization, respectively.

Let $\tilde{K}$ be those that lie in H'' and similarly we define $\dot{L}, \tilde{L}$ for L, note that they are disjoint by definition, because T is the intersection of K and L, and $\dot{K}, \dot{L}$ are the complement of T in K and L respectively. Also $H \to^{T \cup \tilde{K}} H'$ and $H \to^{T \cup \tilde{L}} H'$.

We firstly consider $H \to^{T \cup \tilde{K}} H'$. This implies that for any bidirected path from any vertex $i \in \tilde{L}$ to any $w \in H' \cup T$, there is a vertex from $\tilde{K}$ on the path, otherwise some vertices of $\tilde{L}$ would be preserved, which is a contradiction. The equivalent statement by swapping $\tilde{K}$ and $\tilde{L}$ also holds. If one considers the first vertex in either $\tilde{K}$ or $\tilde{L}$ on any such path, one of the two statements would be false. Hence the lemma is true. □

Lemma 3.4.2. *For a MAG $\mathcal{G}$ and any i, there is at most one edge between any two heads in $\mathfrak{I}_i(\mathcal{G})$.*

Proof. This is a direct consequence of Lemma 3.4.1. □

Lemma 3.4.3. *For two heads $H, H' \leq i$, we have $H > H'$ if and only if H is an ancestor of H' in $\mathfrak{I}_i(\mathcal{G})$*

Proof. This can be proved by marginalizing vertices in $\mathrm{an}(H) \setminus \mathrm{an}(H)$ step by step. □

A useful fact is that for any i and $\mathfrak{I}_i^{\mathcal{G}}$, there exists a (maximal) head H such that $H \geq H'$ for any $H' \leq i$ and this head is the barren subset of the district of i in $\mathcal{G}_{[i]}$.

Lemma 3.4.4. *For a MAG $\mathcal{G}$ and any vertex i in $\mathcal{G}$, $\mathfrak{I}_i(\mathcal{G})$ is a DAG*

Proof. This is a direct consequence of Lemma 3.4.2 and 3.4.3. □

Here we prove that the list of independences associated with edges in $\mathfrak{I}_i^{\mathcal{G}}$ for every i, combined with the independences $i \perp\!\!\!\perp [i-1] \setminus \mathrm{mb}_{\mathcal{G}}(i,[i]) \mid \mathrm{mb}_{\mathcal{G}}(i,[i])$, are sufficient to deduce the ordered local Markov property.

Definition 3.4.3. For a MAG $\mathcal{G}$ and any vertex i in $\mathcal{G}$, we associate $\mathfrak{I}_i^{\mathcal{G}}$ with a collection of independences $\mathbb{L}_i^{\mathcal{G}}$ that contains:

(*a*) $i \perp\!\!\!\perp [i-1] \setminus \mathrm{mb}_{\mathcal{G}}(i,[i]) \mid \mathrm{mb}_{\mathcal{G}}(i,[i])$, and

(*b*) for every head H (except $\{i\}$) whose maximal element is i:

$$i \perp\!\!\!\perp (H \cup T) \setminus (H' \cup T' \cup k) \mid H' \cup T' \setminus \{i\} \qquad \text{for } k \in H \setminus \{i\},$$

where $H \to^k H'$, and $T = \mathrm{tail}_{\mathcal{G}}(H)$ and $T' = \mathrm{tail}_{\mathcal{G}}(H')$.

Theorem 3.4.5. *For a MAG $\mathcal{G}$, the collection $\mathbb{L}^{\mathcal{G}} = \bigcup_i \mathbb{L}_i^{\mathcal{G}}$ is equivalent to the list of independences implied by the ordered local Markov property for $\mathcal{G}$.*

Proof. ($\Longleftarrow$): notice that $H' \cup T' \setminus \{i\}$ is the Markov blanket of i in the ancestral set $\mathrm{an}(H \cup T \setminus K)$, thus it follows from the ordered local Markov property by marginalizing irrelevant vertices.

($\Longrightarrow$): let $\mathcal{A}^x$ denote the set of all ancestral sets whose maximal elements are x. Further let $\mathcal{A}_r^x = \{A \in \mathcal{A}^x : |A| = r\}$ (note $1 \leq r \leq x$). We will proceed by induction on x from $x = 1$ to $x = n$. To show for every $A \in \mathcal{A}^x$, the corresponding independence implied by the ordered local Markov property is implied by $\mathbb{L}_i$, we will apply a further induction on $\mathcal{A}_r^x$ from $r = x$ to $r = 1$.

For $A \in \mathcal{A}^x$, $A \subseteq [x]$. The base case $x = 1$ is trivial. Suppose the induction hypothesis is true, i.e. for any ancestral set $A \in \mathcal{A}^x, x \leq i-1$ the corresponding conditional independence implied by the ordered local Markov property is in $\mathcal{I}_{u_\mathcal{G}}$.

Now consider $\mathcal{A}^i$, we will then apply induction on $\mathcal{A}^i_r$. For the base case $r = i$, i.e. $A = [i]$, the independence is in $\mathbb{L}_i$, that is, (a). Now suppose the induction hypothesis is true, that is: for any $A \in \mathcal{A}^i_r, s+1 \leq r \leq i$, the corresponding conditional independence from the ordered local Markov property is implied by $\mathbb{L}_i$ and we can use the ordered local Markov property on $\mathcal{G}_{[i-1]}$.

Now consider any $A \in \mathcal{A}^i_s$. There is at least one vertex v in $[i] \setminus A$ such that v is parentless in $[i] \setminus A$, in other words, $A \cup \{v\} = A'$ is ancestral and $A' \in \mathcal{A}^i_{s+1}$. Hence by the second induction hypothesis we have

$$i \perp\!\!\!\perp A' \setminus (\mathrm{mb}(i, A') \cup \{i\}) \mid \mathrm{mb}(i, A'). \tag{3.2}$$

If $\mathrm{mb}(i, A) = \mathrm{mb}(i, A')$, then we can get required conditional independence by marginalizing $v \in A' \setminus (\mathrm{mb}(i, A') \cup \{i\})$.

So now assume $\{v\} \cup \mathrm{mb}(i, A) \subseteq \mathrm{mb}(i, A')$ (v, i are in the same district). Let $A'' = A \cup \{v\} \setminus \{i\} = A' \setminus \{i\}$. We know A'' is ancestral and is in $\mathcal{A}^{i-1}$, thus by the firt induction hypothesis, we can apply the ordered local Markov property to any vertex in $\mathrm{barren}(A'')$ (changing the topological order on $[i-1]$). Moreover, $v \in \mathrm{barren}(A'')$ because if v has any child in A then A is not ancestral. So we have:

$$v \perp\!\!\!\perp A'' \setminus (\mathrm{mb}(v, A'') \cup \{v\}) \mid \mathrm{mb}(v, A'').$$

Next notice that by assumption, $\mathrm{mb}(v, A') \cup \{v\} = \mathrm{mb}(i, A') \cup \{i\}$ (i, v are in the same district in A'). As $A'' \subset A'$, we have $\mathrm{mb}(v, A'') \subset \mathrm{mb}(v, A') = \mathrm{mb}(i, A') \cup \{i\} \setminus \{v\}$. Then because $\mathrm{mb}(v, A'') \subset \mathrm{mb}(i, A') \cup \{i\} \setminus \{v\}$ and $v \in \mathrm{mb}(i, A')$, by semi-graphoids, moving $\mathrm{mb}(i, A') \setminus (\mathrm{mb}(v, A'') \cup \{v\})$ (which is a subset of A'') to the conditioning set, we have:

$$v \perp\!\!\!\perp A'' \setminus \mathrm{mb}(i, A') \mid \mathrm{mb}(i, A') \setminus \{v\}. \tag{3.3}$$

Now (3.2) is equivalent to:

$$i \perp\!\!\!\perp A'' \setminus \mathrm{mb}(i, A') \mid \mathrm{mb}(i, A'). \tag{3.4}$$

Thus by semi-graphoids, we can deduce the following from (3.3) and (3.4):

$$\{i, v\} \perp\!\!\!\perp A'' \setminus \mathrm{mb}(i, A') \mid \mathrm{mb}(i, A') \setminus \{v\}.$$

The next step is to marginalize v so that:

$$\{i\} \perp\!\!\!\perp A'' \setminus \mathrm{mb}(i, A') \mid \mathrm{mb}(i, A') \setminus \{v\}.$$

which is equivalent to:

$$\{i\} \perp\!\!\!\perp A \setminus ((\mathrm{mb}(i, A') \setminus \{v\}) \cup \{i\}) \mid \mathrm{mb}(i, A') \setminus \{v\}. \tag{3.5}$$

Note that $(\mathrm{mb}(i, A') \setminus \{v\}) \subseteq A$. To extract vertices in $(\mathrm{mb}(i, A') \setminus \{v\}) \setminus \mathrm{mb}(i, A)$, the key point is to notice that $\{v, i\} \subseteq \mathrm{barren}(\mathrm{dis}_{A'}(i)) = H$, where H is a head that contains the maximal vertex i. Thus $\{v\}$ can be a marginalizing set K for H with the tail T, and $H \cup T = \mathrm{mb}(i, A')$. Now by marginalizing $\{v\}$ we reach $B = A' \setminus \{v\} = A$, thus the following conditional independence is in $\mathbb{L}_i$:

$$\{i\} \perp\!\!\!\perp \mathrm{mb}(i, A') \setminus ((\mathrm{mb}(i, A) \cup \{v\}) \cup \{i\}) \mid \mathrm{mb}(i, A). \tag{3.6}$$

From (3.5) and (3.6), we can deduce that:

$$i \perp\!\!\!\perp A \setminus (\mathrm{mb}(i, A) \cup \{i\}) \mid \mathrm{mb}(i, A).$$

□

3.4.4 Refined power DAG

We have shown that the list of independences associated with the complete power DAGs is sufficient to deduce the ordered local Markov property (Theorem 3.4.5). However, this leads to many redundant independences, and we will show that it is sufficient to only include a single independence for each head. Consequently we will call these simpler power DAGs *refined*. An example of the power DAGs of the graph in Figure 3.8 is given in Figure 3.9.

The following definitions will help us to characterize the marginalization sets.

Definition 3.4.4. For a MAG $\mathcal{G}$ and a set of vertices W, define the *ceiling* of W as

$$\mathrm{ceil}_{\mathcal{G}}(W) = \{w \in W : W \cap \mathrm{an}_{\mathcal{G}}(w) = w\}.$$

Given a head H we define its *Hamlet*[3] as

$$\mathrm{ham}_{\mathcal{G}}(H) = \mathrm{sib}_{\mathcal{G}}(\mathrm{dis}_{\mathrm{an}(H)}(H)) \setminus \mathrm{dis}_{\mathrm{an}(H)}(H).$$

Lemma 3.4.6. *For a MAG $\mathcal{G}$ with heads $H, H' \leq i$, if $H \to^k H'$, then $k \in \mathrm{ceil}_{\mathcal{G}}(\mathrm{ham}_{\mathcal{G}}(H'))$.*

Proof. This is a direct consequence of Lemma 3.4.10. □

[3]This nomenclature makes sense on understanding that the *Claudius* of H, within a set such that H is barren, is the subset of vertices after strict siblings of H and their descendants are removed. Note that this set that has been removed is precisely the Hamlet of H.

Intuitively, $\text{ham}_{\mathcal{G}}(H)$ serves as the bidirected boundary of H and so must be contained in the marginalization set to reach a graph in which H is the maximal head. Also clearly the last marginalization vertex must be at ceiling of the hamlet, otherwise the barren subset of the district will contain some vertices not in H.

Our definition for a refined power DAG will require a partial ordering on the heads.

Definition 3.4.5. Define a partial ordering on heads by setting $H < H'$ if and only if H and H' share the same maximal vertex, and $\text{an}_{\mathcal{G}}(H) \subseteq \text{an}_{\mathcal{G}}(H')$.

Lemma 3.4.7. *For any MAG $\mathcal{G}$ and topological ordering, the relation $<$ is a partial ordering on $\mathcal{H}(\mathcal{G})$.*

Proof. This is a direct consequence of Lemma 4.8 in Evans and Richardson (2014) as the ordering is a sub-order (in that heads are only comparable if they have the same maximal vertex) of the partial orders defined in that paper. □

Definition 3.4.6. For a MAG $\mathcal{G}$ and a topological order $<$, the *refined power DAG* $\widetilde{\mathfrak{J}}^{\mathcal{G}}_{<}$ for $\mathcal{G}, <$ consists of a component for each vertex i. Denote this by $\widetilde{\mathfrak{J}}^{\mathcal{G}}_{i}$; it has vertices $\{H : H \leq i\}$, and an edge $H' \to^k H''$ where

$$k = \min \text{ceil}_{\mathcal{G}}(\text{ham}_{\mathcal{G}}(H')), \text{ and}$$
$$H'' = \max\{H' : H' \in \text{pa}_{\mathfrak{J}_i(\mathcal{G})}(H'') \text{ and } H' \to^k H''\}.$$

That is, for each head, we only take at most one edge and therefore at most one independence into it.

Taking the maximal set among parent heads of H requires some justification. The following section proves the existence of maximal parents for each head.

3.4.5 Existence of maximal parents in the complete power DAG

Lemma 3.4.8. *Suppose that for two heads $i \geq H, H'$, we have $H \to^K H'$. Then $H \to^L H'$ for $L = H \setminus H'$.*

Proof. Clearly, any marginalization set from H to H' must not contain any element from H'. By existence of the minimal marginalization set, we can just add all irrelevant vertices to the set. □

Proposition 3.4.9. *For a MAG $\mathcal{G}$, suppose that for three heads $i \geq H_1, H_2, H$, we have $H_1 \to^K H$ and $H_2 \to^L H$. Then $H_3 = \text{barren}(H_1 \cup H_2)$ is a head and $H_3 \to^{K'} H$ for $K' = H_3 \setminus H$. This means that in the power DAG for i, if a head has a parent head, then there exists a maximal parent head.*

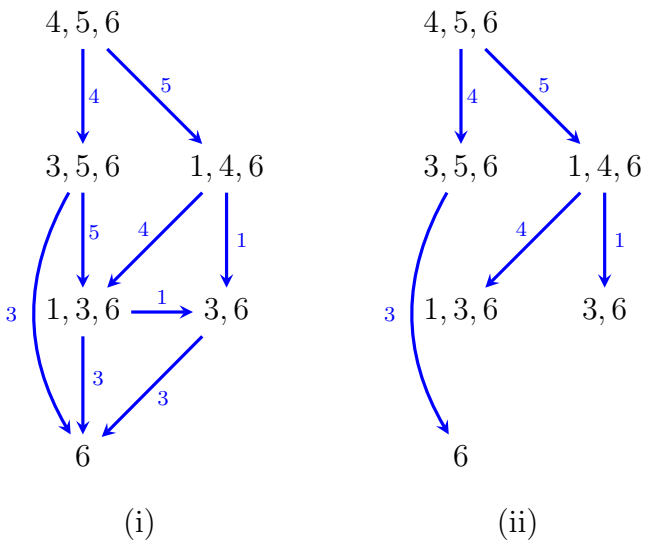


Figure 3.9: (i) The component of the power DAG for the graph in Figure 3.8(i); and (ii) the refined version of the same component. Both are on the heads of $\mathcal{G}$ with maximal vertex 6.

Proof. First of all, we show that H_3 is a head. Since i is the maximal vertex, $i \in H_3$, then for any $j \in H_3$, it is either in H_1 or H_2, which means that j lies in the same district as i in either $\mathcal{G}_{\text{an}(H_1)}$ or $\mathcal{G}_{\text{an}(H_2)}$. This graph is a subgraph of $\mathcal{G}_{\text{an}(H_3)}$, hence j lies in the same district as i in $\mathcal{G}_{\text{an}(H_3)}$.

Now let $K_3 = H_3 \setminus H$. It does not contain i and also it is not empty since $H_3 \neq H$, thus it is a valid marginalization set for H_3. Suppose also that $H_3 \to^{K_3} H'$. We know that $\text{an}(H_3) \supseteq \text{an}(H_1) \supseteq \text{an}(H)$, so $\text{an}(H_3) \setminus K_3 = B = \text{an}(H') \supseteq \text{an}(H)$. Now suppose $H' \neq H$, this means that $\text{dis}_{\text{an}(H')}(i)$ contains some vertices that are not in $\text{dis}_{\text{an}(H)}(i)$. Among those vertices, there are the strict siblings of $\text{dis}_{\text{an}(H)}(i)$; there must exist such vertices because H' is bidirected-connected. Now select one of these siblings, say, j. WLOG, j belongs to $\text{an}(H_1)$. Then to go from H_1 to H, j must be marginalized, thus $j \in H_1$. If j is not in H_3, then this means that there are some descendants of j that are in H_2, but then we cannot go from H_2 to H since j stays in the districts, there is some extra vertex in the barren subset of the district of i, other than H. Hence $j \in H_3$. But then as j is in $\text{an}(H')$, j does not lie in K_3, so $j \in H$, which is a contradiction to the definition of j. □

Lemma 3.4.10. *For a MAG $\mathcal{G}$, if head H is a parent head of H' in the power DAG, then the minimal marginalization set K is $H \cap \text{ceil}_{\mathcal{G}}(\text{ham}_{\mathcal{G}}(H'))$.*

Proposition 3.4.11. *Consider the four heads H, H_k, $k = 1, 2, 3$ with the same setting in Proposition 3.4.9, then the minimal marginalization set for H_3 (to H) is the union of the minimal marginalization sets of H_1, H_2 (to H).*

Proof. Lemma 3.4.10 shows that the minimal marginalization set is the intersection between the parent head and ceil(ham(H)), thus since $\text{an}_{\mathcal{G}}(H_3)$ is just an union of $\text{an}_{\mathcal{G}}(H_1)$ and $\text{an}_{\mathcal{G}}(H_2)$, it will not introduce extra vertices in ceil(ham(H)). □

In Lemma 3.4.10, by construction, $H_3 \geq H_1, H_2$. One can show that if $H \geq H'$ then $H \cup T \supseteq H' \cup T'$, therefore by Lemma 3.4.10, the maximal independence from the H_3 always implies the maximal independences from H_1 and H_2.

Proposition 3.4.12. *Suppose in a power DAG for i, H is a parent head of H', then the independence associated with the minimal marginalization set K^m,*

$$i \perp\!\!\!\perp H \cup T \setminus H' \cup T' \cup K^m \mid H' \cup T' \setminus \{i\},$$

implies all the independences associated with any $H \to^K H'$ and any other marginalization set K.

Proof. The conditioning set is fixed, and the more elements K has, the fewer i is independent from. □

We call this the *maximal independence* associated with $H \to^K H'$

Therefore, a direct consequence of Proposition 3.4.11 is that if $H_1 \to^k H'$ and $H_2 \to^k H'$, then $H_3 := \text{barren}_{\mathcal{G}}(H_1 \cup H_2)$ satisfies $H_3 \to^k H'$. Thus there always exists a maximal parent head in any non-empty $\{H : H \in \text{pa}_{\mathfrak{J}_i(\mathcal{G})}(H')$ and $H \to^k H'\}$.

For simple MAGs, the complete and refined power DAGs are identical and are chains, in the sense that any node has at most one parent head and at most one child. An important motivation inspired by simple MAGs is that the list of conditional independences associated with every edge in the power DAG of a simple MAG decomposes is $u_{\mathcal{G}}$, and $u_{\mathcal{G}}$ is the simplest imset that defines the same model as the graph.

Now we deduce that the list of independences associated with edges in the refined power DAGs are equivalent to the ordered local Markov property.

3.4.6 The refined Markov property from the refined power DAGs

Recall that $[n]$ denotes the set $\{1, \ldots, n\}$.

Definition 3.4.7. For a MAG $\mathcal{G}$ and each i, let $\widetilde{\mathbb{L}}_i^{\mathcal{G}}$ be a list of independences, such that:

(a) it contains $i \perp\!\!\!\perp [i-1] \setminus \text{mb}_{\mathcal{G}}(i, [i]) \mid \text{mb}_{\mathcal{G}}(i, [i])$, and

(*b*) for every head H' other than the maximal one, it contains the independence associated with the unique edge into it in $\widetilde{\mathfrak{J}}_i^{\mathcal{G}}$

We will refer to the collection $\widetilde{\mathbb{L}}^{\mathcal{G}} = \bigcup_i \widetilde{\mathbb{L}}_i^{\mathcal{G}}$ as the *refined (ordered) Markov property.*

Proposition 3.4.13. *For a MAG $\mathcal{G}$, the refined Markov property is equivalent to the ordered local Markov property. Further, it contains fewer and smaller independences compared to the reduced ordered local Markov property.*

By 'smaller', we mean any independence in the refined Markov property can be deduced from the reduced ordered local Markov property by simply marginalizing some vertices.

Proof. In Theorem 3.4.5, we prove that the list of independences associated with the complete power DAGs, denoted as $\mathbb{L}^{\mathcal{G}}$, is equivalent to the ordered local Markov property. Because $\widetilde{\mathbb{L}}^{\mathcal{G}} \subseteq \mathbb{L}^{\mathcal{G}}$, by Theorem 3.4.5, it is sufficient to prove that $\widetilde{\mathbb{L}}^{\mathcal{G}}$ implies $\mathbb{L}^{\mathcal{G}}$.

We proceed by three inductions. The first induction is on the topological ordering of vertices and it is sufficient to show that given $\mathbb{L}_k^{\mathcal{G}}$ is true for $1 \leq k \leq i-1$, $\mathbb{L}_i^{\mathcal{G}}$ is implied by $\widetilde{\mathbb{L}}_i^{\mathcal{G}}$. The base case is trivial.

The second induction is on the topological ordering on heads. We start from the maximal head in $\mathfrak{I}_i(\mathcal{G})$ and proceed downwards to show that the independences in $\mathbb{L}_i^{\mathcal{G}}$ associated with each head and its parent heads are true. Again, the base case is trivial. Now suppose that for a head H', every independence associated with any heads preceding H' is true, in particular, independences in $\mathbb{L}^{\mathcal{G}}$ associated with any parent head of H' hold. We need to show that independences associated with edges $H \to^k H'$ for any parent head H of H' hold.

Now consider the topological ordering on the vertices in $\mathrm{ceil}_{\mathcal{G}}(\mathrm{ham}_{\mathcal{G}}(H'))$, which is made of all the marginalization vertices and we proceed to the third induction on this ordering. The base case is for those parent heads of H' with marginalization vertex $k = \min \mathrm{ceil}_{\mathcal{G}}(\mathrm{ham}_{\mathcal{G}}(H'))$. Clearly the independence from the maximal parent head is in $\widetilde{\mathbb{L}}_i^{\mathcal{G}}$, and this implies all independences from other parent heads H such that $H \to^k H'$.

For the inductive step: consider some parent head H_j of H' with marginalization vertex $l \in \mathrm{ceil}_{\mathcal{G}}(\mathrm{ham}_{\mathcal{G}}(H'))$ and $l > k$. Take some parent head H_1 of H' with marginalization vertex k, then clearly $k \notin H_j$ and $l \notin H_1$. Let

$$H^m = \mathrm{barren}_{\mathcal{G}}(H_j \cup H_1),$$

then by Proposition 3.4.11, $H^m \to^{kl} H'$. Therefore, we have $H^m \to^l H^{m'}$ for some parent head $H^{m'}$ of H' and also we have $H^{m'} \to^k H'$.

By the hypothesis of the second induction, we have

$$i \perp\!\!\!\perp H^m T^m \setminus H^{m'} T_t^m l \mid H^{m'} T^{m'} \setminus \{i\}.$$

By the hypothesis of the third induction, we have

$$i \perp\!\!\!\perp H^{m'} T^{m'} \setminus H' T' k \mid H' T' \setminus \{i\}.$$

Hence, by the contraction semi-graphoid, we have

$$i \perp\!\!\!\perp H^m T^m \setminus H' T' k l \mid H' T' \setminus \{i\}.$$

Now we have $k \notin H_j$ and also, by construction, $H^m T^m \supseteq H_j T_j$. Therefore, we can marginalize irrelevant vertices to get

$$i \perp\!\!\!\perp H_j T_j \setminus H' T' l \mid H' T' \setminus \{i\}.$$

Next we show it contains fewer and smaller independences than the reduced ordered local Markov property. For each maximal ancestral set A and its maximal vertex i, there is a corresponding head by taking barren of the district of i and this relation is one-to-one. Hence the number of independences in the refined Markov property is not more than the number of independences in the refined Markov property. Further every independence in the refined Markov property can be deduced by an independence from the reduced ordered local Markov property by simply marginalizing some vertices, hence the statement is true. □

Now, since the number of independences is bounded by the number of heads, this will greatly reduce the number of independences arising from the reduced ordered local Markov property.

Remark 6. For simple MAGs, the refined Markov property is the simplest possible description of the model and cannot be further reduced; however, for some graphs even less complicated descriptions exist. For example, for the bidirected 5-cycle, adding the semi-elementary imsets corresponding to these independences would give an imset of degree 7. However, in fact one can build an imset that represents the model of 5-cycle of only degree 6, simply by adding the elementary imset corresponding to any valid marginal independence to our 'standard' imset.

Another example is the 5-chain with bidirected edges (see Figure 3.6(i)); its standard imset *is* perfectly Markovian for the graph, but the list of independences that defines the imset model is smaller than our refined Markov property (see Example 3.5.1). In Appendix 3.4.8, we give one more interesting example where there are still redundant independences arising in refined power DAGs.

Remark 7. We use the concept of the 'Hamlet' to characterize minimal marginalization sets and maximal parent heads. This concept originates in Richardson (2003), though he did not use this term. He uses the Hamlet to reduce the ordered local Markov property by only visiting maximal ancestral sets. This is similar to our approach in that one visits each head only once (one can easily show there is a one-to-one correspondence between heads and maximal ancestral sets), however, Richardson only reduces the number of ancestral sets visited, while we also marginalize some unnecessary vertices in the independence statements. In Example 3.5, for instance, $A = \{1, 2, 3, 4, 5, 8\}$ is a maximal ancestral set with $\mathrm{mb}_{\mathcal{G}}(8; A) = \mathrm{pa}_{\mathcal{G}}(8) = \{1, 5\}$ for the head $\{8\}$. The reduced ordered local Markov property gives $8 \perp\!\!\!\perp 2, 3, 4 \mid 1, 5$, but we only need $8 \perp\!\!\!\perp 3, 4 \mid 1, 5$; this is because $8 \perp\!\!\!\perp 2 \mid 1, 3, 4, 5, 6$ has already been obtained for the head $\{6, 8\}$, and we already know that $6 \perp\!\!\!\perp 2 \mid 1, 3, 4, 5$ from variables earlier in the graph. Reducing the number of variables in a independence statement is crucial in practice when trying to obtain a combinatorial imset by adding the semi-elementary imsets corresponding to the list of independences.

3.4.7 Computing refined power DAGs

In this section, we show that $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ can be obtained without computing the complete power DAG and the complexity of the algorithm is polynomial for graphs with restricted head size. We use Algorithm 3 to achieve this and it contains two important ideas:

(i) only marginalizing vertices that are smaller (w.r.t. a topological ordering) than those already marginalized, and

(ii) for each head, only keep one parent of it, which is on the shortest path from the maximal head.

Proposition 3.4.14. *Given a MAG $\mathcal{G}$, Algorithm 3 computes $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ for each i.*

Proof. At line 12, the algorithm proceeds by the partial ordering on heads. Thus for any head H, once we arrive at line 12, there will be no edge added that is into it. So it is sufficient to show that for any head H, once we arrive it at line 12, there is only one edge into it, which corresponds to the edge in $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$. We prove this by induction on the partial ordering on heads. Clearly the algorithm does not add any edge into the maximal head. The base case is for the children of the maximal head and this clearly holds.

For the inductive step, consider all the parent heads of a head H', which exceed H' in the partial ordering and consider H^* and k defined in Definition 3.4.6. Suppose

$H^* \to^k H'$ does not appear and instead we have $H \to^{k'} H'$ for some other parent head of H'.

Firstly if $k' \neq k$ then by Lemma 3.4.6 and definition of k, we must have $k' > k$. Now it is clear that $k \in \text{ceil}_{\mathcal{G}}(\text{ham}_{\mathcal{G}}(H))$, so k must be marginalized at some point to reach H, therefore $k \leq M(H)$ and then line 14 prevents marginalizing of k'.

Then we have $k' = k$, so $H \to^k H'$. Then by definition of H^* we have $H^* > H$, thus from the maximal head, H^* can be reached within fewer steps than H, and line 16 prevents adding $H \to^k H'$. □

Proposition 3.4.15. *Given a MAG $\mathcal{G}$ with n vertices, e edges, and maximal head size k, then the complexity of Algorithm 3 and 4 are $O(kn^k(n+e))$ and $O(n^k(n+e))$, respectively.*

Proof. For Algorithm 1: there are at most $\binom{n}{k}$ heads at line 3 and 12. Then line 14 is of order k. At line 15, it takes $O(n+e)$ to check which new head is reached. Remaining lines all takes constant time or does not contain any loop.

In Algorithm 4, there are at most $\binom{n}{k}$ heads. Then we need to compute the tail for each head and add the corresponding semi-elementary imset. The latter is of constant time and the tail of each head can be computed simultaneously when we visit the head from its parent head, therefore is also of $O(n+e)$ time. □

Remark 8. If the input MAG is given to be simple, then the refined power DAG, which is the same as the complete power DAG, can be obtained very fast as shown in Algorithm 5. In fact, the complexity of Algorithm 5 is linear in the number of edges and vertices, as the structure of the power DAG is inherited completely given the topological ordering.

3.4.8 An example for redundant independences in the refined power DAGs

Example 3.4.2. Consider Figure 3.10(i). The independences associated with the component of its refined power DAG for 6 under a numerical topological ordering are:

$$6 \perp\!\!\!\perp 3 \mid 1,2,4 \qquad 6 \perp\!\!\!\perp 2 \mid 1,3,5 \qquad 6 \perp\!\!\!\perp 2 \mid 1.$$

Whatever the topological order over the other vertices, there is always a third independence that is redundant. For example, with this numerical ordering, one can deduce $6 \perp\!\!\!\perp 2 \mid 1$ from $6 \perp\!\!\!\perp 2 \mid 1,3,5$ and $2 \perp\!\!\!\perp 3,5 \mid 1$. If one adds an edge between 2 and 3 as in Figure 3.10(ii), then only semi-graphoids are insufficient to deduce

Input: A MAG $\mathcal{G}([n], \mathcal{E})$

Result: The refined power DAGs $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ for each i

define $M(H)$ *the minimum vertex marginalized to reach* $H \in \mathcal{H}_i(\mathcal{G})$;
define $SD(H)$ *the shortest path to* $H \in \mathcal{H}_i(\mathcal{G})$;
for $i \in [n]$ **do**
  define VS_i *the set of visited heads for* $H \in \mathcal{H}_i(\mathcal{G})$;
  define UVS_i *the set of not visited heads for* $H \in \mathcal{H}_i(\mathcal{G})$;
  Compute the maximal head $H^* = \text{barren}(\text{mb}(i, [i]))$ and set $VS_i = \{\{i\}\}, UVS_i = \{H^*\}$;
  if $H^* = \{i\}$ **then**
    Next
  end
  Set $M(H^*) = i$, $SD(H^*) = 0, M(\{i\}) = SD(\{i\}) = \infty$;
  Start with a graph $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ with vertex H^*;
  while UVS_i *is not empty* **do**
    Take any $H \in UVS_i$ and **move** H to VS_i;
    for $k \in H \setminus \{i\}$ *and* $k < M(H)$ **do**
      Let $H' : H \to^k H'$;
      if $H' \notin VS_i \cup UVS_i$ **then**
        add H' to $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ and $H \to^k H'$; $UVS_i = UVS_i \cup \{H'\}$; $SD(H') = SD(H) + 1$; $M(H') = k$;
      else
        if $SD(H') > SD(H) + 1$ **then**
          delete any edges into H';
          add $H \to^k H'$;
          $SD(H') = SD(H) + 1$;
          $M(H') = k$;
        end
      end
    end
  end
end
return $(\widetilde{\mathfrak{I}}_1^{\mathcal{G}}, \widetilde{\mathfrak{I}}_2^{\mathcal{G}}, \ldots, \widetilde{\mathfrak{I}}_n^{\mathcal{G}})$

Algorithm 3: Obtain the refined power DAGs for a general MAG

Input: The refined power DAG $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ of a MAG $\mathcal{G}$ for each i
Result: The imset $u_{\mathcal{G}}^r$ that is perfectly Markovian w.r.t. $\mathcal{G}$

Let $u_{\mathcal{G}}^r$ be an empty imset;
for $i \in [n]$ **do**
  for H *in* $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ *along the partial ordering* **do**
    Compute $T = \text{tail}(H)$;
    if $\text{pa}(H) = \emptyset$ **then**
      $u_{\mathcal{G}}^r := u_{\mathcal{G}}^r + u_{\langle i,|[i]\setminus HT|HT\rangle}$
    else
      Let H' be the *only* parent of H and $T' = \text{tail}(H')$;
      $u_{\mathcal{G}}^r := u_{\mathcal{G}}^r + u_{\langle i,H'T'\setminus HT|HT\rangle}$
    end
  end
end
return $u_{\mathcal{G}}^r$

Algorithm 4: Obtain the imset $u_{\mathcal{G}}^r$ from refined Markov property

Input: A simple MAG $\mathcal{G}([n], \mathcal{E})$ ($\mathcal{E}$ stored as adjacencies)
Result: The refined power DAGs $\widetilde{\mathfrak{I}}_i^{\mathcal{G}}$ for each i

for $i \in [n]$ **do**
  Let smaller siblings of i be $j_1, \ldots, j_k$
  Set $\widetilde{\mathfrak{I}}_i = \{\{j_k, i\} \to \ldots \to \{j_1, i\} \to \{i\}\}$
end
return $(\widetilde{\mathfrak{I}}_1^{\mathcal{G}}, \widetilde{\mathfrak{I}}_2^{\mathcal{G}}, \cdots, \widetilde{\mathfrak{I}}_n^{\mathcal{G}})$

Algorithm 5: Obtain the refined power DAGs for a simple MAG

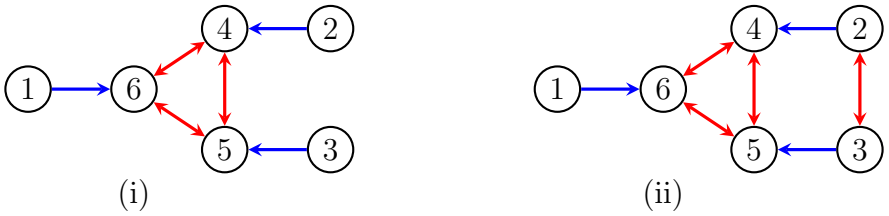


Figure 3.10: Example where refined Markov property is still redundant

$6 \perp\!\!\!\perp 2 \mid 1$ from only first two independences and independences associated with earlier vertices; hence, in this case, we need the third independence.

This example suggests that a minimal list of independences required to define the model (under semi-graphoids) sometimes depends on structures not local to the vertex being considered. We are investigating how to obtain such a list, but conjecture that it may be computationally difficult to do so in general.

In Section 3.4.10, we will present a Theorem on decomposing $u_{\mathcal{G}}$. We need the following results to achieve this.

3.4.9 Some useful results on heads

An implication of Lemma 3.4.1 is that for any $H, H' \leq i$, if H is a parent head of H' on the power DAG for i, there exists a unique *minimal marginalization set of vertices* that leads from H to H'.

Definition 3.4.8. For a MAG $\mathcal{G}$ and two heads $H, H' \leq i$, let K^m be the minimal marginalization set of vertices such that $H \to^{K^m} H'$.

Lemma 3.4.16. *For a MAG $\mathcal{G}$ and two heads $H, H' \leq i$, and $H \to^{K^m} H'$, then $H \to^{K} H'$ if and only if $K = K^m \dot{\cup} B$ where $B \subseteq (H \setminus (H' \cup K^m))$.*

Proof. ($\Rightarrow$): By definition of K^m, $K^m \subseteq K$, and $H \setminus (H' \cup K^m)$ are simply all the vertices we can marginalize if we want to reach the head H'.

($\Leftarrow$): This is because after marginalization of some vertices in the barren subset, the remaining vertices are either outside of the Markov blanket of i or it stays in the barren subset of the district, i.e. the head. Then by the definition of K^m, we may marginalize any other irrelevant vertices in the barren subset after marginalization of K^m as they would be outside of the Markov blanket. □

We only require one more lemma to validate our decomposition of u_G. One key step is to find situations when there is only one marginalization set, which is also the minimal marginalization set. We propose the following definitions.

Lemma 3.4.17. *Consider $H' \to^K H$ where $K = H' \setminus H$ is the minimal marginalization set of vertices, i.e. K is the only marginalization set to reach H from H'. Then this happens if and only if $H' = \text{barren}(K \cup H)$ where $K \subseteq \text{ceil}(\text{ham}(H))$.*

Proof. ($\Rightarrow$): After marginalizing vertices in the barren subset, the remaining vertices of H' either stay in the barren subset or outside of the Markov blanket. Since K is the only marginalization set of vertices, we know the remaining vertices stay in the barren subset (also in H). Hence $H' = \text{barren}(K \cup H)$. We first show that $H' \setminus H \subseteq \text{ham}(H)$. Suppose, for a contradiction, that it is not true; let $K^* = K \setminus (\text{sib}(\text{dis}_{\text{an}(H)}(i)) \setminus \text{dis}_{\text{an}(H)}(i))$. Consider any bidirected path between any $i \in K^*$ and any $k \in \text{dis}_{\text{an}(H)}(i)$, the first vertex x from k that is not in $\text{dis}_{\text{an}(H)}(i)$ lies in $\text{an}(K) \setminus \text{an}(H)$.

If x is in $\text{an}(K) \setminus (\text{an}(H) \cup K)$ we have after marginalizing K then $\text{barren}(\text{dis}_{H \setminus K}(i)) \neq H$, because it would include a descendant of x (possibly x itself) that is *not* an ancestor of H. This is a contradiction to our assumption.

Hence x must lie in K, so by assumption $x \in K \setminus K^*$. Since the choice of the path and vertices are free, it means that if we marginalize $K \setminus K^*$ then K^* would be outside the district of i, so K is not minimal. Hence we reach another contradiction.

Then $K \subseteq \text{ceil}(\text{ham}(H))$ follows, as if it has some ancestors that are also siblings of $\text{dis}_{\text{an}(H)}(i)$ then after marginalizing K, the barren subset of the district would not be H.

($\Leftarrow$): Let K be any subset of $\text{ceil}(\text{ham}(H))$ (but not $\emptyset$) and consider $H' = \text{barren}(K \cup H)$. Clearly H' is a head. Moreover $K \subseteq H'$ as K either has no ancestral relation with H or K are descendent of H; in particular this means that K is a valid marginalization set for H'. We still need to show that (1) after marginalizing K, we reach the head H and (2) K is minimal.

For (1), suppose there is a vertex $t \notin K$ stays in the barren subset of the district but t is not in H. Consider any bidirected path from t to $\text{dis}_{\text{an}(H)}(i)$ in $\mathcal{G}_{\text{an}(H')}$. By definition, on this path there is a vertex in K (not just $\text{an}(K)$) next to some vertex in $\text{dis}_{\text{an}(H)}(i)$, then this path is removed after marginalizing K. Also note that all vertices in $\text{dis}_{\text{an}(H)}(i)$ stay in the graph.

For (2), If we marginalize some subset of K, then the remaining vertices of K stay in the Markov blanket and hence in the barren subset of the district, which then would not be H. □

3.4.10 Decomposition of the 'standard' imset for general MAGs

In this section, we give a decomposition of $u_{\mathcal{G}}$, using all independences arising from $H \to^K H'$ for every possible pair of heads H, H' and every possible marginalization set K. Note that this theorem does not have practical use but it proves why empirically the inner product between the standard imset and entropy is zero for any MAGs. Moreover, it shows that if the standard imset is structural then $\mathcal{I}_{u_{\mathcal{G}}} \subseteq \mathcal{I}_{\mathcal{G}}$. See discussion after the sketch proof.

Theorem 3.4.18. *For a MAG $\mathcal{G}$, with vertices $[n]$ (topologically ordered), we have*

$$u_{\mathcal{G}} = \sum_{i=1}^{n} \Bigg\{ u_{\langle i,[i-1]\setminus \mathrm{mb}(i,[i])|\mathrm{mb}(i,[i])\rangle} + \sum_{\substack{H\in\mathcal{H}(\mathcal{G})\setminus\{i\}:\\ H\leq i}} \sum_{\substack{\emptyset\subset K\subseteq H\setminus\{i\}:\\ H\to^K H'}} (-1)^{|K|+1} u_{\langle i,HT\setminus H'T'K|H'T'\setminus i\rangle} \Bigg\}.$$

Proof. Let $T = \mathrm{tail}_{\mathcal{G}}(H)$. Note that $u_{\mathcal{G}}$ from Theorem 3.3.2 can be rewritten as:

$$\sum_{i=1}^{n} \Bigg\{ \delta_{[i]} - \delta_{[i-1]} - \sum_{\substack{H\in\mathcal{H}(\mathcal{G})\\ i\geq H}} \sum_{W\subseteq H} (-1)^{|H\setminus W|} \delta_{W\cup T} \Bigg\}.$$

By induction, if we restrict to the final vertex of a topological ordering, say n, then all we need to prove is that:

$$\begin{aligned} &\delta_{[n]} - \delta_{[n-1]} - \sum_{\substack{H\in\mathcal{H}(\mathcal{G})\\ n\geq H}} \sum_{W\subseteq H} (-1)^{|H\setminus W|} \delta_{W\cup T} \\ &= u_{\langle n,[n-1]\setminus \mathrm{mb}(n,[n])|\mathrm{mb}(n,[n])\rangle} + \sum_{\substack{H\in\mathcal{H}(\mathcal{G})\setminus\{n\}\\ n\geq H}} \sum_{\substack{\emptyset\subset K\subseteq H\setminus\{i\}:\\ H\to^K H'}} (-1)^{|K|+1} u_{\langle i,HT\setminus H'T'K|H'T'\setminus i\rangle}. \end{aligned}$$

Note that $u_{\langle n,[n-1]\setminus \mathrm{mb}(n,[n])|\mathrm{mb}(n,[n])\rangle} = \delta_{[n]} - \delta_{[n-1]} - \delta_{\{n\}\cup\mathrm{mb}(n)} + \delta_{\mathrm{mb}(n)}$, so we can reduce the equivalence to

$$- \sum_{\substack{H\in\mathcal{H}(\mathcal{G})\\ n\geq H}} \sum_{W\subseteq H} (-1)^{|H\setminus W|} \delta_{W\cup T} \tag{3.7}$$

$$\begin{aligned} &= \delta_{\mathrm{mb}(n,[n])} - \delta_{\{n\}\cup\mathrm{mb}(n,[n])} \\ &\quad + \sum_{\substack{H\in\mathcal{H}(\mathcal{G})\setminus\{n\}\\ n\geq H}} \sum_{\substack{\emptyset\subset K\subseteq H\setminus\{i\}:\\ H\to^K H'}} (-1)^{|K|+1} (\delta_{HT\setminus K} - \delta_{(HT\setminus Kn)} - \delta_{H'T'} + \delta_{H'T'\setminus n}). \end{aligned} \tag{3.8}$$

We can rewrite (3.7) as:

$$
\begin{aligned}
&- \sum_{\substack{H \in \mathcal{H}(\mathcal{G}) \\ n \geq H}} \sum_{W \subseteq H} (-1)^{|H \setminus W|} \delta_{W \cup T} \\
&= - \sum_{\substack{H \in \mathcal{H}(\mathcal{G}) \\ n \geq H}} \sum_{K \subseteq H} (-1)^{|K|} \delta_{HT \setminus K} \\
&= \sum_{\substack{H \in \mathcal{H}(\mathcal{G}) \\ n \geq H}} \sum_{K \subseteq H \setminus \{n\}} (-1)^{|K|+1} (\delta_{HT \setminus K} - \delta_{HT \setminus Kn}). \qquad (3.9)
\end{aligned}
$$

The repeating part is only $-\delta_{HT} + \delta_{HT \setminus n}$ for each H (when $K = \emptyset$). For each H, the two terms do not appear in the summation for H (as we rule out the case when $K = \emptyset$ in Theorem 3.4.18) but they appear when other heads marginalized to H and the two terms are multiplied by some constants.

Our objective is to show that (3.9) and (3.8) are equivalent, for which it is sufficient to prove that for every head H, the coefficient of $-\delta_{HT} + \delta_{HT \setminus i}$ is 1 in (3.8). For the largest head, which is $\text{barren}(\text{dis}_{[n]}(n))$, it is clearly true because there is no head that is 'larger' than it and for this head $-\delta_{HT} + \delta_{HT \setminus i}$ is simply $\delta_{\text{mb}(n,[n])} - \delta_{\{n\} \cup \text{mb}(n,[n])}$, which never appears in the summation.

For any other head H with maximal vertex n, we need to prove:

$$
\sum_{K, H' : H' \to^K H} (-1)^{|K|+1} = 1.
$$

Note that the summation is over both H' and K because for a pair of heads there might exist different marginalization sets that lead one to another.

Now by Lemma 3.4.16, any head H' that can reach H with multiple marginalizing sets and minimal marginalizing set K' contributes $\sum_{B : B \subseteq (H' \setminus (H \cup K'))} (-1)^{|K'|+|B|+1} = 0$ to the coefficient of $-\delta_{H \cup T} + \delta_{(H \setminus \{n\} \cup T)}$.

Thus it is sufficient to consider any head H' that can reach H with the only (minimal) marginalizing set $K = H' \setminus H$. By Lemma 3.4.17, we find all these heads and in total they contribute

$$
\sum_{K : \emptyset \subset K \subseteq \text{ceil}(\text{sib}(\text{dis}_{\text{an}(H)}(i)) \setminus \text{dis}_{\text{an}(H)}(i))} (-1)^{|K|+1} = 1
$$

to the coefficient of $-\delta_{H \cup T} + \delta_{(H \setminus \{n\} \cup T)}$. □

Corollary 3.4.18.1. *For a MAG $\mathcal{G}$, if $u_{\mathcal{G}}$ is structural, then $\mathcal{I}_{u_{\mathcal{G}}} \subseteq \mathcal{I}_{\mathcal{G}}$.*

Proof. By Theorem 3.4.18, the 'standard' imset $u_{\mathcal{G}}$ can be expressed as one combinatorial imset u_p subtracted by another combinatorial imset u_n where u_p are obtained

by adding all semi-elementary imsets with positive coefficients (where $|K|$ is odd), and similarly for u_n but with negative coefficients (where $|K|$ is even).

Let u_p^1 be sum of semi-elementary imsets corresponding to marginalize only one vertex, which are all in u_p. By Theorem 3.4.5, $I_{\mathcal{G}} \subseteq \mathcal{I}_{u_p^1}$. The faithfulness result in Richardson and Spirtes (2002) shows that for every MAG $\mathcal{G}$ there exists a distribution P such that $\mathcal{I}_P = \mathcal{I}_{\mathcal{G}}$ and Theorem 5.2 in Studený (2006) implies the existence of a structural imset u such that $I_u = I_P$. Now every independence in the list of independences that we used to construct u_p^1 is in $\mathcal{I}_u$ by Theorem 3.4.5, so by Lemma 6.1 in Studený (2006), we have $I_{u_p^1} \subseteq \mathcal{I}_u = \mathcal{I}_{\mathcal{G}}$. Hence $\mathcal{I}_{u_p^1} = \mathcal{I}_{\mathcal{G}}$. Now the remaining semi-elementary imsets that we use to construct u_p are those corresponding to marginalize odd number of vertices (more than one), and the independences they correspond to are all in $\mathcal{I}_{\mathcal{G}} = \mathcal{I}_{u_p^1}$. Therefore $\mathcal{I}_{u_p^1} = \mathcal{I}_{u_p}$.

Suppose $u_{\mathcal{G}}$ is structural and take any distribution P that is not Markov to it. Studený (2006) shows that a distribution is Markov to a structural imset if and only if the inner product between the imset and entropy vector of the distribution is zero. Moreover this inner product is non-negative. Therefore, for P and $u_{\mathcal{G}}$, this inner product is positive. Further, as $u_{\mathcal{G}}$ is also $u_p - u_n$ where u_p and u_n are both combinatorial, this means that the inner product between P's entropy vector and u_p are also positive, and hence P is not Markov to u_p. Thus P is not Markov to $I_{\mathcal{G}}$. As a result, $\mathcal{I}_{u_{\mathcal{G}}} \subseteq \mathcal{I}_{\mathcal{G}}$. □

Example 3.4.3. Consider Figure 3.11. The multiple labels on the edges in (ii) means that there are different sets of vertices that can be marginalized to lead from one head to another.

The edge $\{4,5,6\} \rightarrow^{45} \{1,3,6\}$ indicates that when we compute the semi-elementary imset for $\{4,5,6\}$ by marginalizing vertices in $\{4,5\}$, we only reach the head $\{1,3,6\}$ once. Hence the head $\{4,5,6\}$ contributes $(-1)^{2+1} = -1$ to the coefficient of $-\delta_{1236} + \delta_{123}$ for the head $\{1,3,6\}$. Similarly the edge from $\{3,5,6\}$ to $\{6\}$ means that we may reach to the head $\{6\}$ from the head $\{3,5,6\}$ by marginalizing either $\{3\}$ or $\{3,5\}$. So the head $\{3,5,6\}$ contributes $(-1)^{1+1} + (-1)^{2+1} = 0$ to the coefficient of $-\delta_{26} + \delta_2$ for the head $\{6\}$.

Remark 9. Note also that we can separate the independences out by districts, by replacing the first sum in Theorem 3.4.18 by a sum over districts, and then pushing in the summations over vertices in that district.

Definition 3.4.9. Let $\mathcal{G}$ be a MAG containing a district D. Then by $\mathcal{G}^{|D}$ denote the induced subgraph over $D \cup \mathrm{pa}_{\mathcal{G}}(D)$, but where all vertices within $\mathrm{pa}_{\mathcal{G}}(D) \setminus D$ have been joined by bidirected edges.

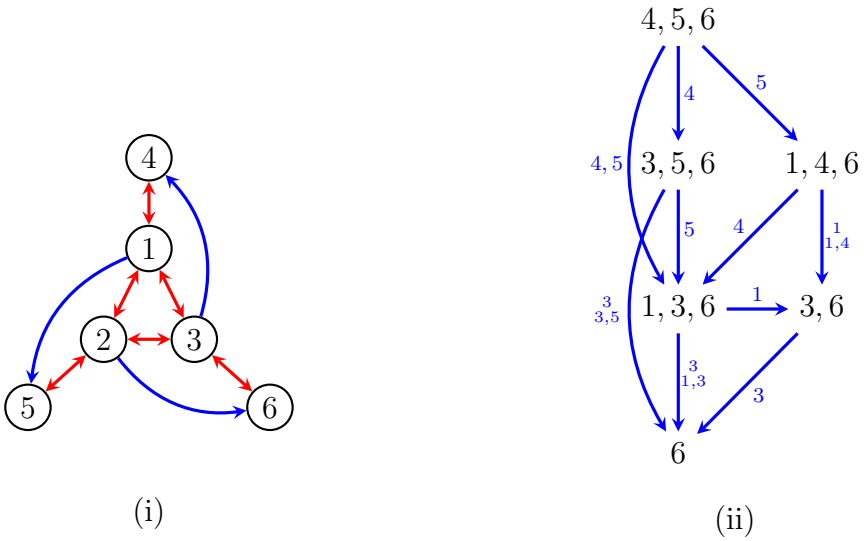


Figure 3.11: (i) A MAG $\mathcal{G}$ (ii) A complete power DAG on the heads of $\mathcal{G}$ with maximal vertex 6, under the numerical topological ordering.

Corollary 3.4.18.2. *Let $\mathcal{G}$ be a MAG. The 'standard' imset $u_{\mathcal{G}}$ is perfectly Markovian with respect to $\mathcal{G}$ if the 'standard' imsets $u_{\mathcal{G}^{|D}}$ are perfectly Markovian with respect to $\mathcal{G}^{|D}$.*

Proof. Take the expression for the standard imset given in Then note that, for a district D, each summand other than the very first term only contains independences between an element of the district, and other elements of the district and its parents, and the collection of all first terms gives a DAG model. Furthermore, the subimset defined by the second inner summation represents the model in which all the vertices in $\mathrm{pa}_{\mathcal{G}}(D) \setminus D$ have been joined by an edge. Since we know that any MAG can be defined by independences only within the district (plus the initial independence in the expression above), this shows that a 'standard' imset for a MAG $\mathcal{G}$ will be perfectly Markovian with respect to the graph if and only if the associated standard imsets for each district and its parents are perfectly Markovian with respect to the induced subgraph obtained after filling in any missing edges between parents. □

We also study our imsets of bidirected graphs and show that for a large class of bidirected graphs, the imset defines the right model. The condition we give is proved to be sufficient and empirically we checked that it is also necessary for graphs with at most seven vertices. The condition, however, is complicated, thus we conjecture that it is combinatorically difficult to obtain a minimal list of independences that define the model. We also give a list of forbidden induced subgraphs; that is, a motif that graphs cannot contain if $\mathcal{I}_{u_{\mathcal{G}}}$ is to be valid and be perfectly Markovian with respect to $\mathcal{G}$.

3.5 Bidirected graphs

For general MAGs, the problem of characterizing the conditions for when $\mathcal{I}_{u_\mathcal{G}}$ is well defined and $\mathcal{I}_{u_\mathcal{G}} = \mathcal{I}_\mathcal{G}$, seems hard in general. To reduce the difficulty, we focus on bidirected graphs in this section. We will give a condition such that our proposed 'standard' imsets $u_\mathcal{G}$ for *bidirected graphs* are always combinatorial and perfectly Markovian with respect to $\mathcal{G}$. We have computationally verified that this condition is also necessary for $|\mathcal{V}| \leq 7$, and not found any graphs for which it is not.

Definition 3.5.1. Let $\mathcal{G}$ be a bidirected graph, and define its undirected *dual graph*, $\overline{\mathcal{G}}$, by $i - j \in \overline{\mathcal{G}}$ if and only if $i \not\leftrightarrow j$ in $\mathcal{G}$.

The dual graph is a powerful tool when we analyse the bidirected graphs; see Example 3.5.1.

3.5.1 How does the characteristic imset help?

There are many MAGs that are not simple, but still have combinatorial standard imsets that are perfectly Markovian with respect to the graph; one example is the bidirected 4-cycle, which has a standard imset consisting of the elementary imsets for its two marginal independences. For bidirected graphs, we found that since the vertices can be given any topological order and there are many heads (any connected subgraph), it is difficult to directly decompose the standard imset or prove the validity of decomposition; however, it turns out to be easier if we work with the characteristic imset. In this section, we consider the relationship between the semi-elementary imset decomposition of combinatorial standard imsets and the characteristic imset.

Recall that two conditional independences I_1, I_2 overlap if $\bar{\mathcal{S}}(I_1) \cap \bar{\mathcal{S}}(I_2) \neq \emptyset$.

Proposition 3.5.1. *A 'standard' imset $u_\mathcal{G}$ is combinatorial if and only if there is a list of non-overlapping conditional independences $\mathbb{L}$ such that $\cup_{I\in\mathbb{L}}\bar{\mathcal{S}}(I) = \mathcal{P}(\mathcal{V}) \setminus \mathcal{S}(\mathcal{G})$.*

Proof. This follows from Lemma 3.3.6, Corollary 3.3.6.1 and the arguments used in their proofs. □

For example, if $\mathcal{G}$ is the bidirected 4-cycle $1 \leftrightarrow 2 \leftrightarrow 3 \leftrightarrow 4 \leftrightarrow 1$ the independences are $1 \perp\!\!\!\perp 3$ and $2 \perp\!\!\!\perp 4$. The corresponding sets to be constrained are $\{1, 3\}$ and $\{2, 4\}$, and these are clearly disjoint. Therefore the standard imset is perfectly Markovian and is simply $u_\mathcal{G} = u_{\langle 1,3\rangle} + u_{\langle 2,4\rangle}$.

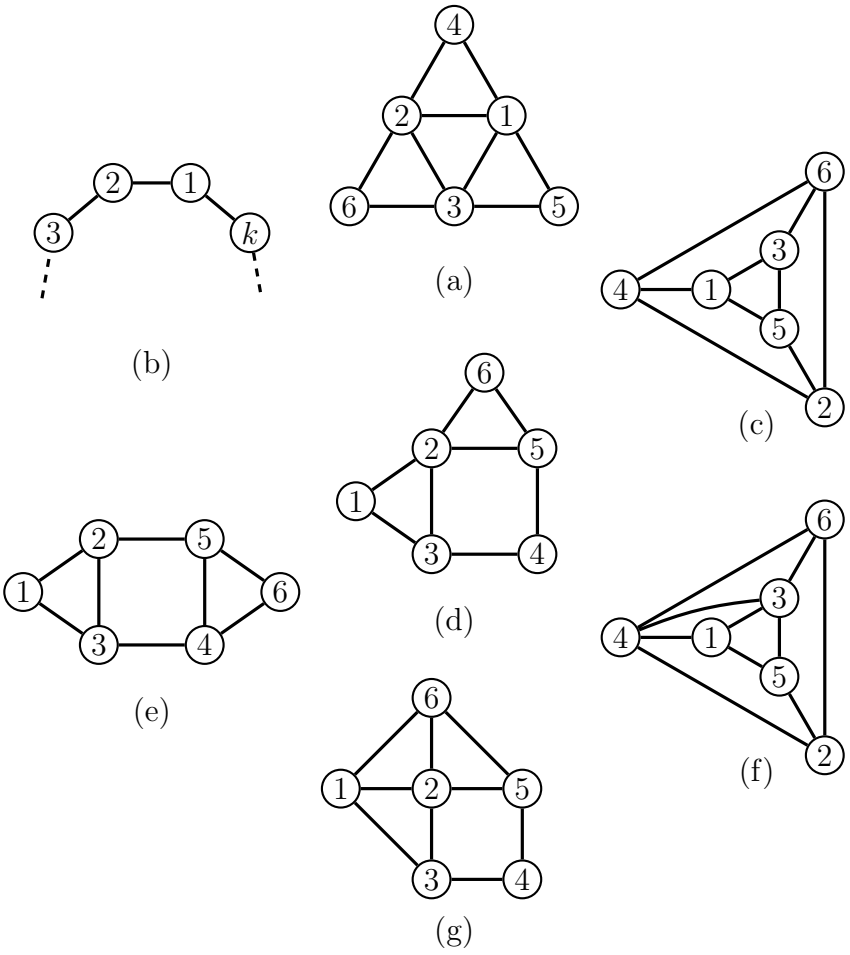


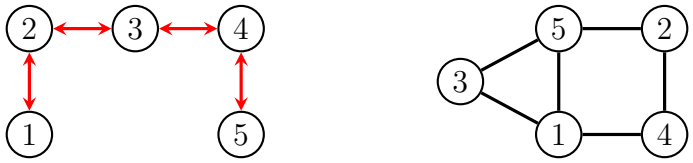
Figure 3.12: Forbidden dual subgraphs: (a) triangles; (b) a k-cycle, for $k \geq 5$; (c) dual to the 6-cycle; (f) dual to the 6-chain; (d), (e), (g) other graphs with at least one chordless 4-cycle.

Proposition 3.3.5 allows us to quickly determine if a standard imset is not perfectly Markovian w.r.t. the graph by checking if the graph contains certain ancestral structures.

Remark 10. In Figure 3.12, we display all dual graphs with at most 6 vertices that must not appear as induced subgraphs to the dual of a bidirected graph for the standard imset to be perfectly Markovian with respect to the graph. This is primarily a consequence of Proposition 3.3.5.

3.5.2 For which bidirected graphs do we get standard imsets that are perfectly Markovian?

We will present a theorem which gives sufficient criteria for bidirected graphs to have combinatorial standard imsets and induce the same model as the graph. For



(i) (ii)

Figure 3.13: (i) The bidirected 5-chain and (ii) its dual graph.

$|\mathcal{V}| \leq 7$ we have empirically verified that this is also necessary. Before we present the theorem, we give a motivating example.

Example 3.5.1. Consider the bidirected 5-chain and its dual graph in Figure 3.13. Consider the following list of conditional independences:

$$\left\{ \begin{array}{l} 4 \perp\!\!\!\perp 2 \,|\, 1,5 \\ 4 \perp\!\!\!\perp 1 \,|\, 3,5 \end{array} \right\} \qquad 2 \perp\!\!\!\perp 5 \,|\, 1,3 \qquad 5 \perp\!\!\!\perp 1,3 \qquad 1 \perp\!\!\!\perp 3.$$

One can check that the sum of semi-elementary imsets corresponding to the above list of independences is the same as the standard imset for the bidirected 5-chain. In addition the standard imset is perfectly Markovian with respect to the graph.

This decomposition starts with 4, and by symmetry it could also start with 2; however, none of the other vertices will work, and we cannot obtain a decomposition in this manner if we try to do so. In particular, if we take the imset for the graph where all edges for 1, 3, 5 are added and subtract from it the standard imset for the 5-chain, what remains is not a structural imset. This shows that for bidirected graphs, even though the topological order can be arbitrary, in order to properly decompose the standard imset further restrictions are required.

Here are some observations. In the dual graph, 4 has neighbours 1 and 2, these vertices share one common neighbour 5, and 1 has one more neighbour 3. However for 3, two of its neighbours have distinct neighbours. This suggests that perhaps we need the neighbours of neighbours to be nested within one another.

For a vertex v, we may also want to treat its neighbours which have the same neighbours as a block, since any path to any one of them also lead to any other vertex in the block. Then any edge within one block does not have any effect on any path to v.

Suppose we have a vertex v and that it has neighbours A in the dual graph. Now partition its neighbours $A^j, 1 \leq j \leq m$ such that the neighbours of each $w \in A^j$ (outside $\{v\} \cup A$) are precisely the set N^j, and such that $N^1 \subseteq N^2 \subseteq \cdots \subseteq N^m$.

Now the question is how the edges connects the blocks of neighbours of v so that v is a valid ending vertex. One observation from the 5-chain is that for two blocks A_i and A_j, if there is one edge between A_i and A_j, then we want A_i and A_j to be fully connected. For example, consider vertex 1 which has neighbours $\{3, 4, 5\}$. Vertices 4 and 5 share the same neighbours outside A, so they are in the same block, while 3 is connected only to 5.

Another question is how A_i and A_j are connected, i.e. the cross-block edges. Suppose v has three blocks of neighbours, A_1, A_2 and A_3. There are 8 possible ways to connect the three blocks. We cannot write out a proper decomposition for three of them. They are: $A_1 - A_3$, $A_2 - A_3$ and $A_1 - A_2 - A_3$. Inspired by the 5-chain example, we have the following definition. For convenience, we introduce the notation $A^{[i,j]}$ to denote $\bigcup_{k=i}^{j}\{A^k\}$ and each A^k is referred as a block (of vertices).

Definition 3.5.2. For a vertex v and an ascending partition of its neighbours $A^{[j_l,j_h]}$, we say the partition is *rooted* if:

(i) it is empty; or

(ii) there exists a block A^j (the root) for some $j_l \leq j \leq j_h$ and $H^j := A^{[j+1,j_h]}$ and $L^j := A^{[j_l,j-1]}$ (respectively empty if $j = j_h$ or j_l) such that A^j and each set in H^j are fully connected to each set in L^j, and for every set in H^j it is entirely disconnected from A^j. In addition, H^j and L^j must themselves be rooted.

One can check that for three blocks and 8 possible ways to connect them, the three ways that do not lead to a proper decomposition also do not satisfy the above definition, but that the other five ways do.

Definition 3.5.3. For a sequence of blocks $A^{[j_l,j_h]}$ rooted at A^j, we define T^j to be the subset of $A^{[1,j_l-1]}$ such that A^i is in T^j if and only if it is connected to A^j.

Lemma 3.5.2. *Let $A^{[j_l,j_h]}$ be a collection of neighbours of v that is rooted. Then any subset of it is also rooted.*

Lemma 3.5.3. *For a vertex v and a collection of its consecutive neighbours, if it is rooted then the root is unique.*

Note that H^j or L^j and $A^{[i,j]}$ are written as sets of sets, but when we use them in certain set expressions or in any independence, we just think of it as the union of the blocks contained in that set, i.e. sets of variables. The context should prevent any confusion.

Lemma 3.5.4. *Suppose a block $A^{[j_l,j_h]} = H^j \cup L^j \cup \{A^j\}$ is rooted at A^j, where H^j and L^j are rooted at A^k and A^t respectively (non-empty), then $T^k = T^j \cup L^j$ and $T^t = T^j$. Moreover $k_l = j + 1$ and $t_l = j_l$.*

Theorem 3.5.5. *For a bidirected graph, if there exists an ordering of the vertices such that for each i in $\overline{\mathcal{G}}_{[i]}$, its neighbours A can be partitioned into blocks $\{A^j, 1 \leq j \leq m\}$ such that*

(i) the neighbours of each $w \in A^j$ (outside $\{v\} \cup A$) in $\overline{\mathcal{G}}$ are the same set N^j;

(ii) $N^1 \subseteq N^2 \subseteq \cdots \subseteq N^m$;

(iii) for any two blocks, either there is no edge between them or they are fully connected;

(iv) $\{A^j, 1 \leq j \leq m\}$ are rooted.

Then $u_{\mathcal{G}}$ is combinatorial and $\mathcal{I}_{u_{\mathcal{G}}} = \mathcal{I}_{\mathcal{G}}$.

The key to the proof of Theorem 3.5.5 is to construct a proper independence decomposition of $u_{\mathcal{G}}$ with the aid of Proposition 3.5.1, and show that it is equivalent to the global Markov property. The following example will give some intuition on how the independences are constructed.

Example 3.5.2. Consider a representation of the dual graph $\overline{\mathcal{G}}$ in Figure 3.14. The vertex i has neighbours partitioned into A^1, A^2, A^3 which have neighbours $N^1 \subseteq N^2 \subseteq N^3$ respectively.

To construct a proper decomposition of $u_{\mathcal{G}}$, we first look at the following conditional independences:

$$i \perp\!\!\!\perp \widetilde{A} \,|\, (\cap_{w \in \widetilde{A}} \mathrm{nb}(w)) \setminus \{i\}, \tag{3.10}$$

for every $\widetilde{A} \subseteq A = \mathrm{nb}(i)$. These independences are definitely implied by the graph as none of the conditioning variables are siblings of $\widetilde{A}$, so no paths are open. Moreover, one can see that each disconnected set containing i is associated with one of the independences. However, the independences also overlap. Proposition 3.5.1 suggests that we do not need all of them, but just a subset of independences (that may require further modification), such that:

(i) the associated constrained sets do not overlap; and

(ii) the independences associated with any disconnected set can be deduced from them using the semi-graphoid axioms.

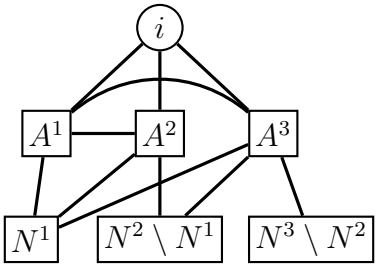


Figure 3.14: Example for Theorem 3.5.5

Let's consider the following independences from (3.10):

$$i \perp\!\!\!\perp A^3, A^2, A^1 \mid N^1 \tag{3.11}$$

$$i \perp\!\!\!\perp A^3, A^2 \mid N^2, A^1 \tag{3.12}$$

$$i \perp\!\!\!\perp A^3, A^1 \mid N^1 \tag{3.13}$$

$$i \perp\!\!\!\perp A^3 \mid N^3, A^1 \tag{3.14}$$

$$i \perp\!\!\!\perp A^2 \mid N^2, A^1 \tag{3.15}$$

$$i \perp\!\!\!\perp A^1 \mid N^1, A^2, A^3. \tag{3.16}$$

(3.13), (3.15) and (3.16) are each implied by (3.11), (3.12) and (3.11) respectively, so we can ignore them. Then looking at (3.11), (3.12) and (3.14) we see that there are some overlaps. The set $\{i\} \cup A^3 \cup A^1$ is associated with both (3.11) and (3.14), but the latter cannot be further simplified so we marginalize A^3 in (3.11).

The set $\{i\} \cup A^3 \cup A^2 \cup A^1$ only appears in (3.12) so we want to keep that, but $\{i\} \cup A^3 \cup A^1$ is associated with both (3.12) and (3.14). Move A^3 into the conditioning set for (3.12) to resolve this. Also $\{i\} \cup A^2 \cup A^1$ is associated with both (3.11) and (3.12), so we marginalize A^2 in (3.11). This gives the following:

$$\begin{aligned}
(3.11) \quad &\rightarrow \quad i \perp\!\!\!\perp A^1 \mid N^1 \\
(3.12) \quad &\rightarrow \quad i \perp\!\!\!\perp A^2 \mid A^3, A^1, N^2 \\
(3.14) \quad &\rightarrow \quad \text{keep (3.14).}
\end{aligned}$$

One can check that any independence involving i can be deduced from the above three independences, together with other independences that do not involve i from the induction hypothesis. In addition, these cover all the disconnected sets containing i.

Proof of Theorem 3.5.5. The main idea of this proof is to construct a list of independences $\mathbb{L}$, where the sum of semi-elementary imsets corresponding to the independences in $\mathbb{L}$ is $u_{\mathcal{G}}$, then using the similar ideas in the proof of Theorem 3.3.8, we just need to show that $\mathbb{L}$ holds in $\mathcal{G}$ so $I_{u_{\mathcal{G}}} \subseteq I_{\mathcal{G}}$, and $\mathbb{L}$ implies the global Markov

property of $\mathcal{G}$ so $I_{\mathcal{G}} \subseteq I_{u_{\mathcal{G}}}$ or the disconnected Markov property from Drton and Richardson (2008).

We will proceed by induction on the ordering of vertices. For a vertex v, suppose its neighbours in $\overline{\mathcal{G}}$ are partitioned into $A^j, 1 \leq j \leq m$, each of which have neighbours N^j (outside of $\{v\} \cup \mathrm{nb}_{\overline{\mathcal{G}}}(v)$) such that $N^1 \subset N^2 \subset \cdots \subset N^m$.

For each A^j, it will be a root for exactly one collection of consecutive sets $\hat{A}^j = A^{[j_l, j_h]}$, and we consider the following list of independences $\mathbb{L}$:

$$v \perp\!\!\!\perp A^j \mid H^j, L^j, N^j, T^j, \qquad j \in 1, \ldots, m.$$

We will prove that for each j, the following independence holds for each j from the above list of independences:

$$v \perp\!\!\!\perp \hat{A}^j \mid N^{j_l}, T^j.$$

We will proceed by two inductions, an outer induction on the number of vertices in $\mathcal{G}$, and an inner induction on the lengths of $\hat{A}^j$. The base case for the outer induction is trivial, since a graph with one vertex has no independences. For the inner induction, the base case is $|\hat{A}^j| = 1$, so $\hat{A}^j = \{A^j\}$, $j_l = j$ and $H^j = L^j = \emptyset$, hence it is in the given list of independences.

For the inner induction step, suppose $\hat{A}^j = H^j \cup L^j \cup A^j$, where H^j and L^j are rooted at k and t (non-empty), respectively. Then by the induction hypothesis, we have:

$$v \perp\!\!\!\perp H^j \mid N^{k_l}, T^k \qquad\qquad v \perp\!\!\!\perp L^j \mid N^{t_l}, T^t,$$

but by Lemma 3.5.4, this is equivalent to:

$$v \perp\!\!\!\perp H^j \mid N^{j+1}, L^j, T^j \tag{3.17}$$

$$v \perp\!\!\!\perp L^j \mid N^{j_l}, T^j. \tag{3.18}$$

Since v is the last vertex, any independence not involving v will hold by the induction hypothesis, hence we have:

$$H^j \perp\!\!\!\perp L^j, N^{j+1} \mid T^j \tag{3.19}$$

$$H^j, A^j \perp\!\!\!\perp L^j, N^j \mid T^j. \tag{3.20}$$

Combining (3.19) and (3.17), we have:

$$H^j \perp\!\!\!\perp \{v\}, L^j, N^{j+1} \mid T^j. \tag{3.21}$$

Then marginalizing N^{j+1} to N^j and moving $L^j \cup N^j$ to conditioning variables, we have:

$$v \perp\!\!\!\perp H^j \mid L^j, N^j, T^j. \tag{3.22}$$

Now from the given list of independences we have:

$$v \perp\!\!\!\perp A^j \mid H^j, L^j, N^j, T^j.$$

Putting this with (3.22) we obtain:

$$v \perp\!\!\!\perp H^j, A^j \mid L^j, N^j, T^j. \tag{3.23}$$

Using (3.20), (3.23) then changes to:

$$\{v\}, L^j, N^j \perp\!\!\!\perp H^j, A^j \mid T^j. \tag{3.24}$$

marginalizing N^{j_l+1} to N^j and moving $L^j \cup N^{j_l}$ to the conditioning variables, we have:

$$v \perp\!\!\!\perp H^j, A^j \mid L^j, N^{j_l}, T^j. \tag{3.25}$$

Finally, combining (3.25) and (3.18), we obtained the required independence:

$$v \perp\!\!\!\perp H^j, A^j, L^j \mid N^{j_l}, T^j. \tag{3.26}$$

Next we show that for any disconnected sets involving v, we can deduce the associated independence, that is, the global Markov property. We use the independence (3.24) that appears in the previous deduction and combine it with an independence from the induction hypothesis, $T^j \perp\!\!\!\perp H^j, A^j$, to get:

$$\{v\}, L^j, T^j, N^j \perp\!\!\!\perp H^j, A^j. \tag{3.27}$$

Based on this, we prove by induction that for every j, we have:

$$\{v\}, L^j, T^j, N^j \perp\!\!\!\perp A^{[j,m]}. \tag{3.28}$$

Before the induction, one should notice that for any $i > j$, $L^i \cup T^i \supseteq L^j \cup T^j$.

An order can be given to blocks based on the collection of blocks in which they are rooted. If H^j, L^j are rooted at A^k, A^t respectively, then we say that j is precedes both k and t, and apply the induction on this order. The base case is for the root of $A^{[1,m]}$, say A^j. In this case, $H^j \cup A^j = A^{[j,m]}$, and it is true.

Suppose for a block A^j, any other block A^i such that i precedes j in the above order so we have $\{v\} \cup L^i \cup T^i \cup N^i \perp\!\!\!\perp A^{[i,m]}$. If $H^j \cup A^j \neq A^{[j,m]}$, then A^{j_h+1} must be a block that also precedes A^j in the order, and so we also have $\{v\} \cup L^{j_h+1} \cup T^{j_h+1} \cup N^{j_h+1} \perp\!\!\!\perp A^{[j_h+1,m]}$ and (more importantly) $L^{j_h+1} \cup T^{j_h+1} \supseteq A^{[j,j_h]} \cup L^j \cup T^j$ (it is possible these sets are equal). now put $A^{[j,j_h]} = H^j \cup A^j$ into the conditioning variables and marginalize N^{j_h+1} to N^{j-1} and other irrelevant variables, to obtain:

$$\{v\} \cup L^j \cup T^j \cup N^j \perp\!\!\!\perp A^{[j_h,m]} \mid H^j \cup A^j. \tag{3.29}$$

Then combining this with (3.24), the result holds for all j.

Now we prove that for every disconnected set C involving v, the associated independence $\{v\} \cup D \perp\!\!\!\perp A$ is true, where $\{v\} \cup D$ is the district for $\{v\}$ in the subgraph induced by C. We only need to consider vertices in $\{v\} \cup A^{[1,m]} \cup N^m \cup \cdots \cup N^1$, as other vertices will definitely link v and any of its neighbours. This A must be a subset of $A^{[1,m]}$. Suppose A^j is the least block that have non-empty intersection with A, so $A \subseteq A^{[j,m]}$.

Consider the vertices in D. Firstly notice that it cannot contain any vertices in N^i or A^i where $i > j$. The first one is clear, suppose the second is not true, consider the bidirected path from A^i to v, the last vertex before v is a sibling of v and it must belong to some of the N^k. Now if $k > j$, then as A^j is connected to N^k, this contradicts to the assumption of j. so $k < j$ but then as A^i is a neighbour of N^k in the dual graph, they are disconnected, so there must be other blocks lower than A^j involved. however as we have $L^i \cup T^i \supseteq L^j \cup T^j$, this means that any block connect to A^i must also connect to A^j in the bidirected graphs and then A^j lies in the same districts as v, contradiction. Thus D can only contain vertices in $L^j \cup T^j \cup N^j$, and we have the required independence.

We are left to prove that the list of independence $v \perp\!\!\!\perp A^j \mid H^j \cup L^j \cup T^j \cup N^j$ are non-overlapping and associated with every disconnected set. For the disconnected set M, suppose $D \cup \{v\} = \mathrm{dis}_M(i)$ and we have this independence $\{v\} \cup D \perp\!\!\!\perp A$ with A^j the least block. Then consider the root of $A^m, \ldots, A^j$ (by Lemma 3.5.2), say A^i, and also consider $\hat{A}^j$ then it is clear that the sets associates with the independence $v \perp\!\!\!\perp A^i \mid H^i \cup L^i \cup T^i \cup N^i$ contains this disconnected set.

To show there is no overlap, assume that $i < j$ (in the numerical sense). Then A^i will appear in the conditioning set for j only if it is in $L^j \cup T^j$, which by definition implies that A^j and A^i are fully connected. However A^j will appear in the conditioning set for A^i only if $A^j \subseteq H^i$, which implies that they are completely disconnected. Since at most one of these conditions can be true, there is no overlap in the sets generated by these independences. □

3.6 Experimental results

We went through graphs with $|\mathcal{V}| \leq 7$ nodes, checking their standard imset from Theorem 3.3.2. For $|\mathcal{V}| \leq 4$, all MAGs have combinatorial standard imsets that are perfectly Markovian with respect to the graph. We summarize the information in Table 3.2 for sizes of graph between 4 and 7. For $|\mathcal{V}| = 5$, the only failure is the bidirected 5-cycle mentioned in Figure 3.6 (i). For $|\mathcal{V}| = 6$, all the imsets

which are perfectly Markovian with respect to the graph are also combinatorial. For those models where the imset is not perfectly Markovian, only two of them are not combinatorial—one of these imsets *is* structural, and both graphs are shown in Figures 3.6(ii–iii) in Section 3.3.

When number of variables are small, the proportion of 'SNPM' is very low, so one might argue we can use it in practice if the size of districts is small, or less than five (Andrews, 2021). Though we can see it rises quickly from $n = 6$ to $n = 7$, therefore the 'standard' imset is not useful for graphs with large district size.

$\|V\|$	equiv. classes	PM	SNPM	NS
4	19	19	0	0
5	285	284	1	0
6	13,303	13,248	54	1
7*	1,161,461	1,146,501	14,562	8

Table 3.2: Number of equivalence classes of connected maximal ancestral graphs for various numbers of nodes (for 7 nodes we only include graphs having at most 13 or at least 18 edges.) PM represents models that are Perfectly Markovian, SNPM those where the imset is Structural but represents a strict subset of the independences (so is Not Perfectly Markovian), and NS where the imset is Not Structural.

3.7 Discussion

3.7.1 Relation to the work in Andrews

The main result of Andrews et al. (2022) is an algorithm that computes an imset that is guaranteed to define the model induced by an ADMG $\mathcal{G}$, but its complexity is exponential in the number of vertices even restricting the maximum head size. In comparison, our imset $u^r_{\mathcal{G}}$ in Algorithm 4 can be constructed in polynomial time if the maximum head size is bounded. In addition, the imset that Andrews et al. gives has much higher degree than $u^r_{\mathcal{G}}$ for most graphs.

A focus of both works is on consistent scoring and searching for an optimal model, and they are the first to introduce $\mathrm{BIC}_{\mathrm{MF}}$, the inner product between empirical entropy and the standard imset from the 0-1 characteristic imset, to approximate the actual BIC. They conduct a brute force search for Gaussian MAG models with five variables by scoring all MECs and selecting the best one. The result is better than FCI (Spirtes et al., 2000) and GFCI (Ogarrio et al., 2016), and comparable to scoring by the true BIC (Drton et al., 2009).

Since, as we have seen, for any MAG with at most five vertices, $u_{\mathcal{G}}$ is perfectly Markovian w.r.t. the graph, this application of $\mathrm{BIC}_{\mathrm{MF}}$ is valid. Note that Andrews

et al. (2022) finesses this by assuming the model is a graphoid, and that therefore the imset for the bidirected 5-cycle is also perfectly Markovian as discussed in Example 3.3.6. They also express this slightly differently by saying the imset is valid if every set in the parametrizing set has cardinality at most five (see their Theorem 3.1).

For general MAGs, our algorithmic advantage in constructing $u^r_{\mathcal{G}}$ comes from the refined power DAGs, which consider only one parent of each head and marginalize one vertex at each time. The approach of Andrews et al. is to take intersections between all subsets of every Markov blanket which gives the exponential complexity. As well as looking all graphs up to a certain size, we present theoretical proofs that for simple MAGs and a class of bidirected graphs, the standard imset $u_{\mathcal{G}}$ is valid.

Some extra assumptions are made by Andrews in his Theorem 3.1. In particular he makes an assumption that one can use the *intersection axiom* (see Section 3.8), which does not generally hold for conditional independence. This means that our 'standard' imsets work for every MAG with size less than or equal to five. We emphasize that we make no such assumption, and use only properties that hold for all distributions, i.e. the semi-graphoid axioms. Stronger rules can change the simplest representation (standard imset) of an independence models. For example, if one assumes positivity of the distribution, then we have already seen that the simplest imset representation for the 5-cycles has degree five, whereas the minimum degree needed without this condition is six. We have verified that using compositional graphoids is sufficient for 51 of the 54 graphs with five or six vertices such that $\mathcal{I}_{u_{\mathcal{G}}} \neq \mathcal{I}_{\mathcal{G}}$, to define the model. However, there are three other graphs that require additional axioms. Specifically, the 'standard' imsets for these graphs will define the model if we can assume *ordered downward stability* (see Section 3.8). For further details, see Example 3.8.1 in Section 3.8.

3.8 Graphoids

Graphoids are rules for logical implications between conditional independences. There is no finite axiomatization of conditional independences (Studeny, 1992), but there are some rules that hold for any distribution, for example, the *semi-graphoids*:

(1) *symmetry*: $X \perp\!\!\!\perp Y \mid Z \iff Y \perp\!\!\!\perp X \mid Z$;

(2) *decomposition*: $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp Y \mid Z$;

(3) *weak union*: $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp W \mid Y, Z$;

(4) *contraction*: $(X \perp\!\!\!\perp Y \mid Z)$ and $(X \perp\!\!\!\perp W \mid Y, Z) \implies X \perp\!\!\!\perp Y, W \mid Z$.

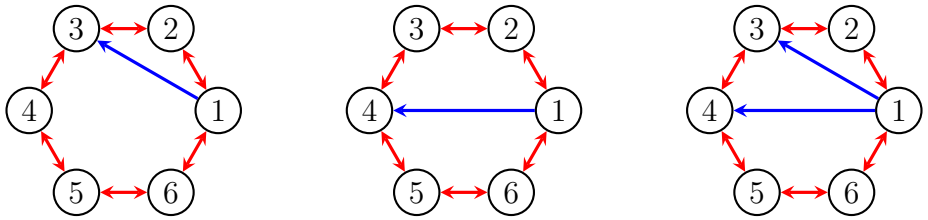


Figure 3.15: Three graphs that do not define the model without assuming rules beyond composition and intersection.

Under the assumption that the distribution is positive, there is an additional property that is guaranteed to hold:

(5) *intersection*: $(X \perp\!\!\!\perp Y \mid W, Z)$ and $(X \perp\!\!\!\perp W \mid Y, Z) \implies X \perp\!\!\!\perp Y, W \mid Z$.

The semi-graphoids with (5) are called *graphoids*. Also if one assumes that the distribution is Markov w.r.t. a graph, then the following axiom also holds:

(6) *composition*: $(X \perp\!\!\!\perp Y \mid Z)$ and $(X \perp\!\!\!\perp W \mid Z) \implies X \perp\!\!\!\perp Y, W \mid Z$.

Graphoids with (6) are called *compositional graphoids*,

Sadeghi (2017) defines three mode graphoid-like rules; two of these require a 'pre-order' $\prec$, which for directed mixed graphs is essentially the partial order implied by the directed edges.

(7) *singleton transitivity*: $(X_i \perp\!\!\!\perp X_j \mid Y)$ and $(X_i \perp\!\!\!\perp X_j \mid Y, X_k) \implies (X_i \perp\!\!\!\perp X_k \mid Y)$ or $(X_j \perp\!\!\!\perp X_k \mid Y)$.

(8) *ordered upward stability*: $(X_i \perp\!\!\!\perp X_j \mid Y) \implies (X_i \perp\!\!\!\perp X_j \mid Y, X_k)$ for any $k \prec i$ or $k \prec j$.

(9) *ordered downward stability*: $(X_i \perp\!\!\!\perp X_j \mid Y, X_k) \implies (X_i \perp\!\!\!\perp X_j \mid Y)$ for any k that is either larger than or incomparable to i, j and every element of Y.

Here is an example where ordered downward stability is needed to deduce all the independences in the graph.

Example 3.8.1. Consider the bidirected 6-cycle with one additional directed edge in Figure 3.15(i). The imset $u_{\mathcal{G}}$ of this graph $\mathcal{G}$ is structural and $\mathcal{I}_{u_{\mathcal{G}}}$ contains the following independences:

$6 \perp\!\!\!\perp 2$	$6 \perp\!\!\!\perp 4 \mid 1, 3$	$2 \perp\!\!\!\perp 4 \mid 1, 6$	$2 \perp\!\!\!\perp 5 \mid 4, 6$
$1 \perp\!\!\!\perp 4$	$1 \perp\!\!\!\perp 5 \mid 2, 4$	$6 \perp\!\!\!\perp 3 \mid 1, 5$	$3 \perp\!\!\!\perp 5 \mid 1, 2.$

In particular, even with the intersection and compositional graphoids, we do not get any joint independences, when the graph says we should have (e.g.) $2 \perp\!\!\!\perp 4, 5, 6$. In the list of the local Markov property, it does not satisfy $4 \perp\!\!\!\perp 1, 2$ or $5 \perp\!\!\!\perp 1, 2$ or $4 \perp\!\!\!\perp 2, 6$. However, ordered downward stability does help. The only pair that is ordered is $1 \to 3$, with all other pairs being incomparable. Then we can deduce $2 \perp\!\!\!\perp 5$ by removing 4 and 6 from $2 \perp\!\!\!\perp 5 \mid 4, 6$, and use composition to obtain $2 \perp\!\!\!\perp 5, 6$; we can then obtain $2 \perp\!\!\!\perp 4, 5, 6$. Similarly we can remove first 3 and then 1 from $6 \perp\!\!\!\perp 4 \mid 1, 3$ to obtain the marginal independence, and get $4 \perp\!\!\!\perp 2, 6$ using composition (or just contraction). Also, we can remove $2, 4$ from $5 \perp\!\!\!\perp 1 \mid 2, 4$ to get $5 \perp\!\!\!\perp 1$, then combined with $5 \perp\!\!\!\perp 2$, we have $5 \perp\!\!\!\perp 1, 2$. On the other hand, we can obtain $4 \perp\!\!\!\perp 2$ by removing 6 and 1 from $4 \perp\!\!\!\perp 2 \mid 1, 6$, then use composition to get $4 \perp\!\!\!\perp 1, 2$, as $4 \perp\!\!\!\perp 1$ is already in $\mathcal{I}_{u\mathcal{G}}$.

3.9 Future work

Chapter 3 leaves open several interesting questions. First, if we are to use an algorithm to search for the optimal MAG using the $\mathrm{BIC}_{\mathrm{MF}}$-score, what moves should it propose and make? The 0-1 imset formulation makes it easy to search, as—provided that Meek's conjecture (Meek, 1997) for MAGs is true—we may search by first adding, and later deleting sets from the parametrizing set.

Another still open problem is to provide a complete solution to obtaining standard imsets of MAGs. This is analogous to undirected graphs, where Kashimura and Takemura (2015) provide a solution for the standard imset of non-decomposable undirected graphs; previously graphs with chordless cycles were not covered by the theory. It seems clear from the example of the bidirected 5-cycle that we will have to choose between having the lowest possible degree and symmetry of the standard imset. We hope that this can expand the class of graphs that can be scored further, and that it may lead to algorithms that are more accurate or more efficient than the current state-of-the-art.

Chapter 4

Search algorithm

In this chapter, we develop a search algorithm for MAGs. The setting is only given observational data, assuming generated from a distribution that is faithful to some MAG and we aim to find the optimal MAG. This chapter is organized as follows: we start with an introduction in Section 4.1; then in Section 4.2, we define extra terminologies; in Section 4.3, we demonstrate how to move between Markov equivalence classes of MAGs and present some results to speed the procedure up; in Section 4.4, we show how to use the frame work of imsets (Studený, 2006) and the reduced Markov property for MAGs (Hu and Evans, 2022) to construct a new scoring criteria for MAGs and prove its consistency; in Section 4.5, we propose our new algorithm by combining results in previous two sections; then finally in Section 4.6, we conduct a simulated experiment and show superior performance to existing MAG learning algorithms.

4.1 Introduction

Causal discovery is an essential part of causal inference(Spirtes et al., 2000; Peters et al., 2017). Estimating causal effect is challenging if the underlying causal graph is unknown. Algorithms for learning causal graphs vary based on different parametric assumptions and class of graphical models used (Spirtes et al., 2000; Kaltenpoth and Vreeken, 2023; Claassen and Bucur, 2022; Nowzohour et al., 2017; Zhang and Hyvarinen, 2009; Peters et al., 2017). In this chapter, we consider the assumption that the conditional independences of distributions can be represented by graphs. One prominent graphical model in causal inference is the *directed acyclic graphs*, also known as DAGs. They offer a clear interpretation and are straightforward to conduct inference. DAGs are associated with distributions by encoding conditional independence. However, for many distributions, DAGs have limitations in fully expressing all conditional independences. For instance, in the presence of hidden

variables, DAGs can not capture all independences over the observed variables. To address this issue, the *maximal ancestral graphs* (MAGs) were developed by Richardson and Spirtes (2002). MAGs provide a more comprehensive representation, overcoming the limitations of DAGs.

Classic graph learning methods for DAGs and MAGs are mainly of three types: constrained-based, scored-based and hybrid which combines features of the first two. Constrained-based methods are known for fast speed. But they can have low accuracy when number of variables grows (Evans, 2020; Ramsey et al., 2006) as empirical mistakes propagate through the algorithm. Classic DAG/MAG learning algorithms are PC/FCI(Spirtes et al., 2000). Some variation of these algorithms are developed to accelerate and increase precision, for example, RFCI/FCI+(Colombo et al., 2012; Claassen et al., 2013). On the other hand, score-based methods search through many graphs and compute a score for each one, then selects the graph with the highest score. Therefore in general they are more robust (require stronger assumption in the parametric models) but slower than those constrained-based methods. GES (Chickering, 2002) perhaps is the most well-known scored-based DAG learning algorithm, which proves greedy learning procedure will output global optimal in the limit of infinite sample size. This was originally known as 'Meek's conjecture' (Meek, 1997). We will prove some result (Proposition 4.3.3) for MAGs by assuming the MAG version of the conjecture to speed up our algorithm.

4.1.1 An overview of past work on score-based method

There are two key components to such score-based algorithms: the score, and the search procedure.

Existing score-based algorithms for MAGs (Triantafillou and Tsamardinos, 2016; Rantanen et al., 2021; Chen et al., 2021; Claassen and Bucur, 2022) all use the *Bayesian information criteria* (BIC). Although Drton et al. (2009) and Evans and Richardson (2010, 2014) have provided methods for fitting Gaussian and discrete MAG models using maximum likelihood, which allows for obtaining the corresponding BIC score, maximum likelihood estimates cannot be obtained in a closed-form, and therefore require iterative computation using numerical methods. Moreover, the optimization function is not generally convex if the model is not a DAG, which means that the such algorithms may converge to a non-globally optimal point. Additionally, the factorization of distributions in MAG models is complex, and the scores are only decomposable with respect to the components connected by bidirected paths, also known as districts or c-components; this makes search methods for MAGs computationally intensive. In this chapter, we use a score from Hu and

Evans (2022), based on work of Andrews et al. (2022), in the framework of imsets (Studený, 2006); it essentially measures the discrepancy in the data from a list of independences implied by the graph. This list of independences is equivalent to but generally simpler than the (reduced) ordered local Markov property (Richardson, 2003; Hu and Evans, 2022).

The search procedure is also very important. Clearly scoring every MAG is inefficient, since the number of MAGs grows super-exponentially as the number of vertices increases. The above mentioned algorithms all search in a greedy manner by only considering neighbouring MAGs; these are different from the current MAG by only a difference in an edge or edge mark (see details later). Among them, only Claassen and Bucur (2022) search through *Markov equivalence class* (MEC) of MAGs, and thus avoid repeatedly scoring graphs which, if the distributions are assumed to be discrete or multivariate Gaussian, always have the same BIC. However, Claassen and Bucur's method still has some inefficiencies. We address these issues and provide a new method; we show that for sparse graphs, under some other mild assumptions, our new algorithm runs in polynomial time.

4.2 Preliminary

The scoring criteria we construct later essentially are measuring the list of independence in some Markov property by using mutual information as a continuous score, in addition to some model complexity. To estimate mutual information, it is sufficient to compute the entropy given a set of variables. Therefore our scoring criteria is not restricted to discrete or Gaussian model. As long as one can estimate entropy, this scoring criteria would be consistent in the limit of infinite sample size.

Definition 4.2.1. For a real-valued and continuous variable X with probability density function $f(x)$, its *entropy* is defined as:

$$\mathsf{H}(X) = -\int_{-\infty}^{\infty} f(x)\log(f(x))\,dx$$

For discrete variables, one replaces the integral sign and probability density function with summation sign and probability mass function respectively.

In this thesis we run simulated experiments with Gaussian variables. The plug-in estimator of Gaussian entropy uses sample mean and variance, which is known to produce underestimation (Basharin, 1959). If the mean of Gaussian distribution is known to be zero, then Ahmed and Gokhale (1989) finds an unbiased estimator with minimal variance, i.e. UMVUE estimator. Further Misra et al. (2005) present

a UMVUE estimator for the general case. We ran our algorithm for a variety of estimators mentioned later, and they all produced very similar final result. Entropy estimation is a widely studied topic and are not focus of this paper so we only briefly discuss these estimators here.

Recall the inner product notation from Definition 3.3.3

Definition 4.2.2. Given a function f which takes X_A for any $A \subseteq V$ as input, and an imset u over V, we define

$$\langle u, f \rangle = \sum_{A \subseteq V} u(A) f(x_A).$$

4.2.1 Meek's conjecture for MAGs

Chickering (2002) proves the Meek's conjecture for DAGs and we state its analogue version of MAGs here.

Theorem 4.2.1. *Let $\mathcal{G}$ and $\mathcal{H}$ be any pair of MAGs such that $\mathcal{I}(\mathcal{G}) \supseteq \mathcal{I}(\mathcal{H})$. Let r be the number of edges in $\mathcal{H}$ that are different to the edges in $\mathcal{G}$, and let m be the number of edges in $\mathcal{H}$ that do not exist in $\mathcal{G}$. There exists a sequence of at most $2r + m$ edge mark change and edge additions in $\mathcal{G}$ with the following properties:*

- *after each edge mark change or edge addition, $\mathcal{G}$ is a MAG and $\mathcal{I}(\mathcal{G}) \supseteq \mathcal{I}(\mathcal{H})$;*
- *after all edge mark changes and edge additions, $\mathcal{G} = \mathcal{H}$.*

This theorem has been proven for DAGs and therefore it guarantees that greedy learning can output the optimal solution in the limit of infinite sample size for DAG models. While this theorem has not been proven for MAG models, many scored-based algorithms for MAGs implicitly assume it and search greedily, see Claassen and Bucur (2022), Triantafillou and Tsamardinos (2016), and Rantanen et al. (2021). Zhang and Spirtes (2005) show that for Markov equivalent MAGs, there exists sequence of single edge mark changes for reaching from one MAG to another while staying in the same MEC, but there has been little progress since then. Throughout this chapter we will assume that Meek's conjecture holds for MAG models and derive some useful facts that accelerate the searching procedure.

4.3 Moving between Markov equivalence classes

Recall in Section 2.7, *partial ancestral graph* (PAG) is defined to characterize $[\mathcal{G}]$, which captures all the arrowheads and tails that are present in every MAG in $[\mathcal{G}]$. In

this section, we describe how we move between MECs by using PAGs as a representation of the MECs. In Ali et al. (2009), they use skeleton and *colliders with order* to represent the MEC and Claassen and Bucur (2022) further simplifies this to linear complexity for sparse MAG equivalence by looking at both collider and noncollider triples with order. To visit other MECs, they perform graphical operations including adding or deleting adjacencies, or altering orientation of colliders with order. After the modification, they compute the PAG of the resulting MEC and check that it is valid. We show that this procedure can be simplified and improved using the orientation rules of PAGs and using PAGs as representation of MECs directly.

4.3.1 PAGs

Recall the definition of PAGs. Given a MAG $\mathcal{G}$, an edge mark in $\mathcal{G}$ is *invariant* if it is present in every graph in $[\mathcal{G}]$.

Definition 4.3.1 (Definition 2.7.1)**.** Given a MAG $\mathcal{G}$, the *partial ancestral graph* (PAG) for $[\mathcal{G}]$, $\mathcal{P}_\mathcal{G}$, is a graph with three kind of edge marks: arrowheads, tails and circles (six kinds of edges: $-$, $\rightarrow$, $\leftrightarrow$, $\circ\!\!-$, $\circ\!\!-\!\!\circ$, $\circ\!\!\rightarrow$)[1], such that:

- $\mathcal{P}_\mathcal{G}$ has the same adjacencies as any maximal member of $[\mathcal{G}]$;
- a mark of arrowhead is in $\mathcal{P}_\mathcal{G}$ if and only if it is invariant in $[\mathcal{G}]$;
- a mark of tail is in $\mathcal{P}_\mathcal{G}$ if and only if it is invariant in $[\mathcal{G}]$.

Recall that Zhang (2007b) present an algorithm, including a sound and complete set of rules, $\mathcal{R}0$ to $\mathcal{R}10$, which are listed in Section 2.7, to construct the PAG of a given MAG.

A direct approach to score a PAG is to construct a MAG represented by the PAG (Zhang, 2007b) and fit the MAG to the data as Claassen and Bucur (2022) did. We show that a representative MAG can be constructed by only an *arrow complete* PAG and thus save the computational cost of orienting the invariant tails.

Remark 11. We should point it out that this is not new as the proof of Ali et al. (2005)'s result on characterizing the MEC by arrow complete PAGs partly relies on it. But we did not find any formal statement of this result in the literature so here we formulate it properly.

[1] As we consider only directed MAGs, there are only four kinds of edges

4.3.2 Representative MAG

Algorithm `PAG-to-MAG` explicitly describe the steps needed for constructing a representative MAG. In fact, just to represent the MEC it is sufficient to only apply rules $\mathcal{R}0$–$\mathcal{R}4$, which obtain all the invariant arrowheads; this *arrow complete* PAG (Ali et al., 2005) is sufficient to characterize the MEC. The remaining rules $\mathcal{R}5$–$\mathcal{R}10$ correspond to finding invariant tails, and generally have higher computation cost than $\mathcal{R}0$–$\mathcal{R}4$.

Now we show that how to obtain the representative MAG from only arrow complete PAGs instead of fully oriented PAGs. This comes from the following observations:

- $\mathcal{R}5$ and $\mathcal{R}6$ will not be called if the MEC contains a directed MAG, as pointed out by Zhang (2007b);
- $\circ\!\!-$ is produced only by $\mathcal{R}6$;
- $\mathcal{R}7$ is called only if there is $\circ\!\!-$, which does not exist for a MEC that contains a directed MAG;
- finally, $\mathcal{R}8$–$\mathcal{R}10$ only change $\circ\!\!\rightarrow$ to $\rightarrow$, and we can always do this without loss of generality.

Lemma 4.3.1. *Let $\mathcal{P}$ and $\mathcal{P}'$ be a fully oriented PAG and an arrow complete PAG, respectively. Suppose they represent the same MEC that contains at least one directed MAG, then the outputs of Algorithm* `PAG-to-MAG` *are the same for $\mathcal{P}$ and $\mathcal{P}'$.*

Therefore, we will use the arrow complete PAGs for scoring MECs, since they are easier to compute. Note that in practice when we are not certain that the incomplete PAG will be standing for a MEC that contains some directed MAGs, we call $\mathcal{R}5$ once and if there is no undirected edge found then the premise of Lemma 4.3.1 is satisfied.

Input: A PAG or an arrow complete PAG $\mathcal{P}$
Result: A MAG $\mathcal{G}$ such that $\mathcal{P}_{\mathcal{G}} = \mathcal{G}$
1 Let $\mathcal{G} = \mathcal{P}$;
2 **Change** every $-\!\!\circ$, $\circ\!\!\rightarrow$ in $\mathcal{G}$ into $\rightarrow$;
3 **Orient** $\circ\!\!-\!\!\circ$ component in $\mathcal{G}$ into a DAG with no unshielded collider;
4 **return** $\mathcal{G}$

Algorithm `PAG-to-MAG`**:**

4.3.2.1 Consistent invariant edge marks

Here we show a result that follows from assuming Meek's conjecture for MAGs. The result will accelerate our greedy learning algorithms.

We say two PAGs are *inconsistent* at an edge mark if it is an invariant arrow head in one PAG and an invariant tail in the other one.

First we need the following lemma.

Lemma 4.3.2. *For any two MAGs $\mathcal{G}$ and $\mathcal{G}'$, that have the same adjacencies but are not Markov equivalent, it is neither $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(\mathcal{G}')$ nor $\mathcal{I}(\mathcal{G}') \subseteq \mathcal{I}(\mathcal{G})$.*

Proof. Consider any unshielded triples $a \ast\!\!-\!\!\ast\, b \ast\!\!-\!\!\ast\, c$ in $\mathcal{G}$ and $\mathcal{G}'$. If it is an unshielded collider triple in $\mathcal{G}$, then there is an independence $a \perp\!\!\!\perp c \mid B$ in $\mathcal{I}(\mathcal{G})$ such that $b \notin B$. If it is an unshielded non-collider triple in $\mathcal{G}'$, then there is an independence $a \perp\!\!\!\perp c \mid B$ in $\mathcal{I}(\mathcal{G}')$ such that $b \in B$. Therefore if $\mathcal{G}$ and $\mathcal{G}'$ have any unshielded triple that is oriented differently, then we are done.

Now suppose $\mathcal{G}$ and $\mathcal{G}'$ have the same skeleton and unshielded collider triples. The remaining piece of their MECs are orientation of discriminating paths when we orient their PAGs. Since they have the same skeleton and unshielded collider triples, any discriminating path arising when orienting one PAG will also appear in another PAG. For similar reason, these discriminating paths must have the same orientation(Richardson and Spirtes, 2002), otherwise we are done. But then since the two MAGs have the same skeleton, unshielded collider triples and orientation of discriminating path when orienting PAGs, they must be Markov equivalent. □

Proposition 4.3.3. *Assuming Meek's conjecture for MAGs holds. Let $\mathcal{P}$ and $\mathcal{P}'$ be two PAGs such that $\mathcal{I}(\mathcal{P}) \supseteq \mathcal{I}(\mathcal{P}')$, and $\mathcal{P}'$ has one more edge $\{i, j\}$ than $\mathcal{P}$. Then there is no inconsistent edge mark between $\mathcal{P}$ and $\mathcal{P}'$.*

Proof. Suppose $b \longleftarrow\!\!\ast\, c$ in $\mathcal{P}$. By Meek's conjecture, for some MAG $\mathcal{G}$ represented by $\mathcal{P}$, there exists a sequence of graph operations, consisting of either adding edge or changing of edge mark, that leads to some MAG $\mathcal{G}'$ represented by $\mathcal{P}'$. There must be only one edge addition. Consider any change of edge mark before the adding edge operation, it cannot change the arrowhead at $b \longleftarrow\!\!\ast\, c$, because this edge mark is invariant, so changing it would lead to another MEC $\mathcal{P}''$ with the same skeleton and such that neither $\mathcal{I}(\mathcal{P}) \subseteq \mathcal{I}(\mathcal{P}'')$ nor $\mathcal{I}(\mathcal{P}'') \subseteq \mathcal{I}(\mathcal{P})$. Therefore after the edge addition operation, $b \longleftarrow\!\!\ast\, c$ remains. Now by Lemma 4.3.2, any change of edge mark later will not change the MEC as skeleton remains the same, so it is always represented by $\mathcal{P}'$. Now since $b \longleftarrow\!\!\ast\, c$ is in some MAG represented by $\mathcal{P}'$, the edge mark then cannot be an invariant tail in $\mathcal{P}'$. □

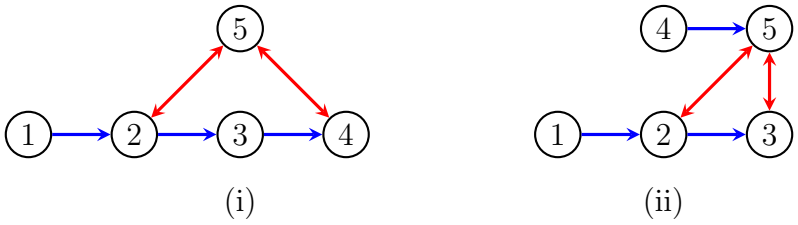


Figure 4.1: Examples for redundant triples

In some cases, Proposition 4.3.3 help us to orient new unshielded triples or discriminating path when we visit a new MEC so we do not need to consider different orientations of them, and saves computational costs. We will show these in details later.

4.3.3 Equivalence classes

To construct the PAG of a MEC, we need the following information: (i) the skeleton; (ii) the unshielded colliders (for $\mathcal{R}0$); and (iii) the orientation of discriminating paths for which $\mathcal{R}4$ is called. Any characterization of MECs should contain this information, and so the algorithm of Zhang (2007b) can be adapted to construct the PAG based upon it. In Section 2.7.2, we showed how to construct a PAG by using the parametrizing set. Claassen and Bucur (2022) show how to do the same using colliders with order. Both the parametrizing set and collider with order have the same information about (i) and (ii), but they contain different triples for (iii); both characterizations may contain *redundant* triples.

Example 4.3.1. In Figure 4.1(i), $\{1, 4, 5\} \in \mathcal{S}_3(\mathcal{G})$ is unnecessary as when orienting the PAG, $\mathcal{R}4$ will not be called and only skeleton and unshielded colliders are needed. Similarly, in Figure 4.1(ii), $\{2, 3, 5\}$ is a collider with order, but again the PAG is completely determined by the skeleton and unshielded colliders.

Claassen and Bucur (2022) would change the MEC of Figure 4.1(ii) by modifying both unshielded colliders and colliders with order, and hence visit the same MEC twice by changing $\{3, 4, 5\}$ or $\{2, 3, 5\}$ to non-colliders. Even though the authors report that, empirically, 95% of the proposed changes result in a valid MEC, they do not discuss how many classes are repeatedly visited. We do not fully address the issue here but since we are only changing the orientation of unshielded tripless, it is clear that this leads fewer repetitions.

In the next few subsections, we present how our algorithm moves between MECs, overcoming the above issue, together with some observations that improve overall efficiency compared to Claassen and Bucur (2022). We also have the problem of

repeatedly visiting the same MECs but our experiment suggests that the overall efficiency is improved. For each step, Claassen and Bucur (2022) consider every possible move including adding one adjacency, deleting one adjacency and altering orientation of one triple with order. We choose to mimic the procedure in Hauser and Bühlmann (2012) that firstly only adds adjacencies, then only deletes adjacencies, and finally alters the orientation of colliders. This will reduce the number of possible moves and is still consistent, provided that Meek's conjecture is true for MAGs.

4.3.4 Adding adjacencies

4.3.4.1 Determine unshielded collider triples

When we add an adjacency, we need to investigate what happens to the three objects we use to characterize equivalence. First, given which adjacency we are trying to add, the new skeleton is clear. For unshielded triples, if it remains unshielded after adding the adjacency, we keep its orientation status, i.e. collider or non-collider, as justified by the following lemma.

Lemma 4.3.4. *Let $\mathcal{G}$ and $\mathcal{H}$ be two MAGs such that $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(\mathcal{H})$. If a triple $\{i, j, k\}$ is unshielded in both $\mathcal{G}$ and $\mathcal{H}$, then $\{i, j, k\}$ is an unshielded collider triple in $\mathcal{G}$ if and only if it is an unshielded collider triple in $\mathcal{H}$.*

If an unshielded triple becomes a full triple after adding the adjacency, then clearly we remove it from consideration; the difficulty here what happens when there are new unshielded triples. By simply going through each possible orientation of these triples and restricting maximal degree of each node to d, we would go through up to 2^{2d} combinations.

Proposition 4.3.3 would help to reduce the complexity. Let $\mathcal{P}$ and $\mathcal{P}'$ be two PAGs such that $\mathcal{I}(\mathcal{P}) \supseteq \mathcal{I}(\mathcal{P}')$, and $\mathcal{P}'$ has one more edge $\{i, j\}$ than $\mathcal{P}$. Let $i \ast\!\!-\!\!\ast\, j \ast\!\!-\!\!\ast\, k$ be an unshielded triple in $\mathcal{P}'$. If $j -\!\!\ast\, k$ in $\mathcal{P}$ then $\{i, j, k\}$ is an unshielded non-collider triple in $\mathcal{P}'$. Thus we only consider the cases for $j \circ\!\!-\!\!\ast\, k$ and $j \leftarrow\!\!\ast\, k$.

We show a trick to simplify the situation by imagining the edge mark of the new edge $i \ast\!\!-\!\!\ast\, j$ at j in the PAG $\mathcal{P}'$ of the new MEC. If it is an arrowhead, i.e. $i \ast\!\!\rightarrow j$, then in the case of $j \leftarrow\!\!\ast k$, $\{i, j, k\}$ is definitely an unshielded collider triple in $\mathcal{P}'$. We use UC_j^d to denote all such triples. In the case of $j \circ\!\!-\!\!\ast k$ in $\mathcal{P}$, $\{i, j, k\}$ may be an unshielded collider or non collider triple in $\mathcal{P}'$. We use UC_j^p to denote all such triples and we need to go through each combination. If it is $i \ast\!\!-\!\!\circ\, j$ or $i \ast\!\!-\, j$ in $\mathcal{P}'$, then $\{i, j, k\}$ cannot be an unshielded collider triple in $\mathcal{P}'$.

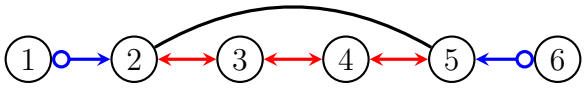


Figure 4.2: A PAG with a new adjacency

Given a PAG $\mathcal{P}$ and an adjacency i, j to add, Algorithm `UC-triples-add` summarize the above procedure and output $UC_i^d, UC_j^d, UC_i^p,\ UC_j^p$. Example 4.3.2 demonstrates the usefulness of this trick. Algorithm `UC-triples-add` also gives an incomplete PAG that are only applied $\mathcal{R}0$ by considering all triples that are both unshielded in $\mathcal{P}$ and $\mathcal{P}'$, and are colliders triples in $\mathcal{P}$. The PAG of any MEC in the next iteration will be oriented by starting at this incomplete PAG.

Example 4.3.2. Suppose we have a PAG $1 \circ\!\!\rightarrow 2 \leftrightarrow 3 \leftrightarrow 4 \leftrightarrow 5 \leftarrow\!\!\circ 6$ and we wish to add the adjacency $\{2, 5\}$ as illustrated by Figure 4.2. There are four new unshielded triples and naively going through them would go through 16 combinations. But Algorithm `UC-triples-add` outputs $UC_2^d = \{\{1, 2, 5\}, \{2, 3, 5\}\}, UC_5^d = \{\{2, 5, 6\}, \{2, 4, 5\}\}$ and $UC_2^p = UC_5^p = \emptyset$. Hence we only need to go through the four cases when adding $UC_2^d \cup UC_5^d$, UC_2^d, UC_5^d or $\emptyset$ as additional unshielded collider triples.

The method we described for proposing possible orientation of new unshielded triples are by no mean sound and complete. Future work can focus on efficient, sound and complete algorithms for orienting these new unshielded triples.

The main improvement we made is described in the following section.

4.3.4.2 Creating branches for $\mathcal{R}4$

The efficiency of our approach comes to the fact that we only determine orientation of discriminating paths when $\mathcal{R}4$ is called, which is the remaining uncertain piece for the new MEC. Claassen and Bucur (2022) determine the new MEC by pre-setting the skeleton, unshielded colliders and colliders with orders, where the latter may contain redundant information as we have seen in Example 4.3.1.

Our idea is straightforward: when $\mathcal{R}4$ is called, we create two branches. For one branch, we orient the triple in the discriminating path as non-collider and for the other one, the triple is oriented as collider. Then we keep orienting each incomplete PAG and whenever $\mathcal{R}4$ is called, we perform the same procedure until graphs are completely oriented.

In some cases, it is unnecessary to create branches. Suppose we are constructing a new PAG $\mathcal{P}''$ from an edge addition to $\mathcal{P}$. Suppose $\mathcal{R}4$ is called for the discriminating path $\pi = \{d, \cdots, a, b, c\}$, if $b \leftarrow\!\!* c$ or $b \circ\!\!-\!\!* c$ in $\mathcal{P}$ then by Proposition 4.3.3, we can

Input: A PAG $\mathcal{P}$, $\{i, j\}$
Result: A incomplete PAG $\mathcal{P}'$,UC_i^d, UC_j^d, UC_i^p and UC_j^p

```
Initialize P' with only ∘−∘ and the same skeleton as P;
Add i ∘−∘ j to P';
Let UC_i^d = UC_j^d = UC_i^p = UC_j^p = ∅;
Apply R0 to P' by considering all triples that are both unshielded in P and
  P', and are colliders in P;
for a ∈ {i, j} do
    let b = {i, j} \ a;
    let UC_a be the set of new unshielded triples that centred at a;
    for {a, b, k} ∈ UC_a do
        if a −∗k in P then
            next
        else
            if a ←∗k in P then
                add {a, b, k} to UC_a^d
            else
                add {a, b, k} to UC_a^p
            end
        end
    end
end
return P', UC_i^d, UC_j^d, UC_i^p, UC_j^p
```

Algorithm `UC-triples-add`**:**

just orient the discriminating path by the edge mark at b in $\mathcal{P}$. Essentially, we only need to create branches for the discriminating path if $b \circ\!\!-\!\!* c$ in $\mathcal{P}$. See Algorithm `Branch-for-`$\mathcal{R}4$`-add` that summarizes the above procedure. Now we are ready to present the full algorithm for adding adjacencies.

Input: A PAG $\mathcal{P}$ and an incomplete PAG $\mathcal{P}'$
Result: An arrow complete PAG $\mathcal{P}'$ or two incomplete PAGs ($\mathcal{P}'_c$, $\mathcal{P}'_n$)
1 Exhausitively apply $\mathcal{R}1 - \mathcal{R}4$ to $\mathcal{P}'$;
2 **if** $\mathcal{R}4$ *is called for an edge* $b \circ\!\!-\!\!* c$ **then**
3 **if** $\{d, b, c\} \in \mathcal{S}(\mathcal{P})$ *or* $b \leftarrow\!\!* c$ *or* $b -\!\!* c$ *in* $\mathcal{P}$ **then**
4 orient b as collider or non-collider in $\mathcal{P}'$, respectively;
5 keep orienting;
6 **else**
7 orient b as collider and non-collider, and let the resulting two incomplete PAGs be $\mathcal{P}'_c$ and $\mathcal{P}'_n$, respectively;
8 **return** *(*$\mathcal{P}'_c$, $\mathcal{P}'_n$*)*
9 **end**
10 **end**
11 **return** $\mathcal{P}'$

Algorithm `Branch-for-`$\mathcal{R}4$`-add`**:**

4.3.4.3 Algorithm for adding adjacency

Algorithm `Add-adj` combines previous algorithms with an additional section that runs dynamically. When Algorithm `Add-adj` proceeds to Line 14, the set S consist of incomplete PAGs that are determined by the same skeleton but different sets of unshielded collider triples. To visit a new MEC, we need to keep applying these orientation rules and decide orientation of discriminating path when $\mathcal{R}4$ is called.

In Appendix, we also list algorithms for deleting adjacency and changing direction of colliders. It is similar to Algorithm `Add-adj`, so we omit them here.

4.3.5 Deleting adjacency

Similar to the algorithm for adding adjacency, we need to think about what happens to skeleton, unshielded collider triples, and orientation of discriminating path when $\mathcal{R}4$ is called. Again the skeleton is clear, given which edge to delete. Then, by Lemma 4.3.4, we would also like to keep unshielded collider triples that remain unshielded after deleting the edge. For those full triples that became unshielded collider triples after deleting an adjacency if the previous PAG contains invariant edge marks that help to orient them, we then keep the orientation of these triples as we did by Proposition 4.3.3 before so we do not need to consider different orientation

Input: A complete PAG $\mathcal{P}$ and an adjacency $\{i, j\}$ to add
Result: A set of arrow complete PAGs

$\mathcal{P}'$,UC_i^d, UC_j^d, UC_i^p, UC_j^p = `UC-triples-add`($\mathcal{P}$,$\{i, j\}$) ;
$S = \{\mathcal{P}'\}$;
for $UC \subseteq UC_i^p$ **do**
  Apply $\mathcal{R}0$ to $\mathcal{P}'$ with additional unshielded triples $UC \cup UC_i^d$;
  add the resulting incomplete PAG to S.
end
for $UC \subseteq UC_j^p$ **do**
  Apply $\mathcal{R}0$ to $\mathcal{P}'$ with additional unshielded triples $UC \cup UC_j^d$;
  add the resulting incomplete PAG to S.
end
for $UC \subseteq UC_i^p \cup UC_j^p$ **do**
  Apply $\mathcal{R}0$ to $\mathcal{P}'$ with additional unshielded triples $UC \cup UC_i^d \cup UC_j^d$;
  add the resulting incomplete PAG to S.
end
$O = \emptyset$;
for $\mathcal{P}' \in S$ **do**
  K = `Branch-for-`$\mathcal{R}4$`-add` $(\mathcal{P}, \mathcal{P}')$;
  while $|K| > 0$ **do**
    Let $\mathcal{P}'' \in K$; K' = `Branch-for-`$\mathcal{R}4$`-add` $(\mathcal{P}, \mathcal{P}'')$;
    if $|K'| = 1$ **then**
      add K' to O; delete $\mathcal{P}'$ from K
    else
      add K' to K
    end
  end
end
return O

Algorithm `Add-adj`**:**

of each triples. Similarly if $\mathcal{R}4$ is called for some discriminating paths and previous PAG helps to orient them then we do not need to create branches.

Let the UC_{ij}^{p} denote the remaining uncertain unshielded triples. We need to enumerate MECs by exploring different orientations of triples in UC_{ij}^{p} and creating branches for new discriminating paths.

Input: A PAG $\mathcal{P}$, $\{i, j\}$
Result: A incomplete PAG $\mathcal{P}'$, UC_{ij}^{p}
Initialize $\mathcal{P}'$ with only $\circ\!-\!\circ$ and the same skeleton as $\mathcal{P}$;
Delete $i \circ\!-\!\circ j$ to $\mathcal{P}'$;
Let $UC_{ij}^{p} = \emptyset$;
Apply $\mathcal{R}0$ to $\mathcal{P}'$ by considering all triples that are both unshielded in $\mathcal{P}$ and $\mathcal{P}'$, and are colliders in $\mathcal{P}$;
Let A be sets of nodes that are adjacent to i, j in $\mathcal{P}$;
for $a \in A$ **do**
    if $i* \!\rightarrow a \leftarrow\! *j$ *in* $\mathcal{P}$ **then**
        orient $i* \!\rightarrow a \leftarrow\! *j$ in $\mathcal{P}'$;
    else
        if $i *\!-\! a *\!-\! * j$ *and* $i *\!-\! * a -\! * j$ *not in* $\mathcal{P}$ **then**
            Add $\{i, j, a\}$ to UC_{ij}^{p};
        end
    end
end
return $\mathcal{P}'$,UC_{ij}^{p}

Algorithm `UC-triples-delete`**:**

See Algorithm `UC-triples-delete`, `Branch-for-`$\mathcal{R}4$`-delete` and `Delete-adj`. They are similar to `UC-triples-add`, `Branch-for-`$\mathcal{R}4$`-add` and `Add-adj`.

4.3.6 Turning phase

Unlike the previous two phases for adding and deleting adjacency, Meek's conjecture does not allow us to change the status of unshielded triples. In fact, such a change between an unshielded collider or non-collider triple would result in a new MEC, which cannot still be an $\mathcal{I}$-map of the true distribution. The turning phase introduced by Hauser and Bühlmann (2012) that changes unshielded triples in DAG models is used to correct mistakes made earlier due to finite sample sizes. We mimic their procedure, and generalize it for MAG models.

We briefly describe what we did here. Suppose we have a PAG from previous two phases. For the turning phase, we would like to keep the skeleton. Then we set a parameter t for how many unshielded triples that we can change orientation at once, which is usually set to one. Once orientation of every unshielded triples is

Input: A PAG $\mathcal{P}$ and an incomplete PAG $\mathcal{P}'$
Result: An arrow complete PAG $\mathcal{P}'$ or two incomplete PAGs ($\mathcal{P}'_c$, $\mathcal{P}'_n$)

Exhaustively apply $\mathcal{R}1 - \mathcal{R}4$ to $\mathcal{P}'$;
if $\mathcal{R}4$ *is called for an edge* $b \circ\!\!- * c$ **then**
  if $\{d, b, c\} \in \mathcal{S}(\mathcal{P})$ *or* $b - * c$ *in* $\mathcal{P}$ **then**
    orient b as collider or non-collider in $\mathcal{P}'$, respectively;
    keep orienting;
  else
    orient b as collider and non-collider, and let the resulting two incomplete PAGs be $\mathcal{P}'_c$ and $\mathcal{P}'_n$, respectively;
    return ($\mathcal{P}'_c$, $\mathcal{P}'_n$)
  end
end
return $\mathcal{P}'$

Algorithm `Branch-for-`$\mathcal{R}$`4-delete`**:**

Input: A PAG $\mathcal{P}$ and an adjacency $\{i, j\}$ to delete
Result: A set of arrow complete PAGs

$\mathcal{P}'$,UC_{ij}^p = `UC-triples-delete`($\mathcal{P}$,$\{i, j\}$) ;
$S = \{\mathcal{P}'\}$;
for $UC \subseteq UC_{ij}^p$ **do**
  Apply $\mathcal{R}0$ to $\mathcal{P}'$ with additional unshielded triples UC;
  Add the resulting incomplete PAG to S.
end
$O = \emptyset$;
for $\mathcal{P}' \in S$ **do**
  $K =$ `Branch-for-`$\mathcal{R}$`4-delete` $(\mathcal{P}, \mathcal{P}')$;
  while $|K| > 0$ **do**
    Let $\mathcal{P}' \in K$; $K' =$ `Branch-for-`$\mathcal{R}$`4-delete` $(\mathcal{P}, \mathcal{P}')$;
    if $|K'| = 1$ **then**
      $O = O \cup K'$; $K = K \setminus \{\mathcal{P}'\}$
    else
      $K = K \cup K'$
    end
  end
end
return O

Algorithm `Delete-adj`**:**

decided, we further orient the new PAG and whenever $\mathcal{R}4$ is called, we create two branches and do not consider its edge mark on previous MEC. This is implemented in Algorithm `Turning` and `Branch-for-`$\mathcal{R}4$`-turning`.

Input: An incomplete PAG $\mathcal{P}'$
Result: An arrow complete PAG $\mathcal{P}'$ or two incomplete PAGs ($\mathcal{P}'_c$, $\mathcal{P}'_n$)
Exhaustively apply $\mathcal{R}1 - \mathcal{R}4$ to $\mathcal{P}'$;
if $\mathcal{R}4$ *is called for an edge* $b \circ\!\!-\!\!* c$ **then**
  orient b as collider and non-collider, and let the resulting two incomplete PAGs be $\mathcal{P}'_c$ and $\mathcal{P}'_n$, respectively;
  return *(*$\mathcal{P}'_c$, $\mathcal{P}'_n$*)*
end
return $\mathcal{P}'$

Algorithm `Branch-for-`$\mathcal{R}4$`-turning`**:**

Input: An arrow complete PAG $\mathcal{P}$, max changes t
Result: A set of arrow complete PAGs
Let UT be the set of unshielded triples in $\mathcal{P}$;
$S = \{\mathcal{P}'\}$;
for $UT_{turn} \subseteq UT$ *and* $|UT_{turn}| \leq t$ **do**
  Change the orientation status of triples in UT_{turn} in $\mathcal{P}'$;
  Add the resulting incomplete PAG to S.
end
$O = \emptyset$;
for $\mathcal{P}' \in S$ **do**
  $K =$`Branch-for-`$\mathcal{R}4$`-turning` $(\mathcal{P}, \mathcal{P}')$;
  while $|K| > 0$ **do**
    Let $\mathcal{P}' \in K$; $K' =$ `Branch-for-`$\mathcal{R}4$`-turning` $(\mathcal{P}, \mathcal{P}')$;
    if $|K'| = 1$ **then**
      $O = O \cup K'$; $K = K \setminus \{\mathcal{P}'\}$
    else
      $K = K \cup K'$
    end
  end
end
return O

Algorithm `Turning`**:**

There are various methods to jump to new MECs that are not $\mathcal{I}$-maps to the previous MEC. Again working with DAGs, Linusson et al. (2023) give a geometric interpretation and generalize the turning phase in Hauser and Bühlmann (2012). Their method can turn more than one unshielded triple at the same time, which is similar to what we did here. One future work is to extend the work in Linusson et al. (2023) to MAG models and design a more robust and efficient turning phase.

4.4 Scoring Criteria

The BIC (Schwarz, 1978) is a consistent (defined below) scoring criterion. For discrete models, Evans and Richardson (2010) provide procedures for fitting ADMGs, obtaining MLE and thus we can compute the BIC.

Let ℓ be the log-likelihood and q^θ to denote the family of distributions that are Markov to the fitted graph $\mathcal{G}$, with parameter θ, which achieve maximum of ℓ at $\hat{\theta}$. And let $d = |\mathcal{S}(\mathcal{G})|$ be the dimension of the discrete model (Evans and Richardson, 2014). Also let N and $N(x_\mathcal{V})$ be the number of samples and the number of samples such that $X_\mathcal{V} = x_\mathcal{V}$, respectively. Then the BIC for fitting $\mathcal{G}$ is:

$$-2\hat{\ell} + d \log N,$$

where $\hat{\ell} = \sum_{x_\mathcal{V}} N(x_\mathcal{V}) \log q_{\hat{\theta}}(x_\mathcal{V})$.

However this score is not very suitable for a greedy learning algorithm for MAGs, as we need to re-fit the whole graph when we consider new models and often unstable. Thus in this chapter we aim to develop a scoring criterion that is decomposable with respect to the parametrizing set $\mathcal{S}(\mathcal{G})$. The motivation is that equivalent MAGs have the same BIC and the parametrizing set, and every time we move between Markov equivalence classes of MAGs, we simply change the 0-1 vector of the characteristic imset, and it would save a lot computations if for those sets remaining in the parametrizing set, we do not need to compute their corresponding scores again.

We propose a new scoring criterion: $-2N\langle u^r_\mathcal{G}, \hat{\mathsf{H}}\rangle + d \log N$, where $\hat{\mathsf{H}}$ is the plug-in estimate of entropy defined below and $u^r_\mathcal{G}$ is an imset from the refined Markov property. We can also show that if we use the 'standard' imset $u_\mathcal{G}$ in the inner product and $\mathcal{I}(\mathcal{G}) = \mathcal{I}(u_\mathcal{G})$, the inner product approximates $\hat{\ell}/N$.

4.4.1 Entropy and interactive information of discrete variables

In this subsection, we present some useful results on discrete variables to help readers better understand the idea about scoring using imsets and entropy.

Definition 4.4.1. The *entropy* of a set of discrete variables $X_\mathcal{V}$, is defined as:

$$\mathsf{H}(X_\mathcal{V}) = \sum_{x_\mathcal{V}} P(X_\mathcal{V} = x_\mathcal{V}) \log P(X_\mathcal{V} = x_\mathcal{V}).$$

Definition 4.4.2. The interaction information of a set of discrete variables $X_\mathcal{V}$, is defined as:

$$\mathsf{I}(X_\mathcal{V}) = \sum_{S \subseteq \mathcal{V}} (-1)^{|V \setminus S|} \mathsf{H}(X_S).$$

We have the following identity:

$$\mathsf{H}(X_{\mathcal{V}}) = \sum_{T \subseteq V} \sum_{S \subseteq T} (-1)^{|T \setminus S|} \mathsf{H}(X_S) = \sum_{T \subseteq V} \mathsf{I}(X_T).$$

The plug-in estimate of entropy of $X_{\mathcal{V}}$, is

$$\hat{\mathsf{H}}(X_{\mathcal{V}}) = \sum_{x_{\mathcal{V}}} \frac{N(x_{\mathcal{V}})}{N} \log \frac{N(x_{\mathcal{V}})}{N}.$$

The plug-in estimate of interaction information of $X_{\mathcal{V}}$, is

$$\hat{\mathsf{I}}(X_{\mathcal{V}}) = \sum_{S \subseteq V} (-1)^{|V \setminus S|} \hat{\mathsf{H}}(X_S).$$

Theorem 3.4.18 allows us to decompose the entropy in terms of the interaction information over the parametrizing set.

Let H, I be the entropy function and the interaction information function, respectively.

Proposition 4.4.1. *For a MAG $\mathcal{G}$ with vertex set V and a distribution p that is Markov to $\mathcal{G}$, we have*

$$\mathsf{H}(X_{\mathcal{V}}) = \sum_{T \in \mathcal{S}(\mathcal{G})} \mathsf{I}(X_T) = \langle c_{\mathcal{G}}, \mathsf{I} \rangle = \langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, \mathsf{H} \rangle.$$

Proof. Recall that $\mathcal{P}(S)$ denotes the power-set of S. By Theorem 3.4.18, it is sufficient to show that $\langle u_{\mathcal{G}}, \mathsf{H} \rangle = \sum_{T \notin \mathcal{S}(\mathcal{G})} \mathsf{I}(X_T) = 0$.

$$\begin{aligned}
\langle u_{\mathcal{G}}, \mathsf{H} \rangle &= \sum_{S \in \mathcal{P}(\mathcal{V})} \sum_{\substack{T \in \mathcal{P}(\mathcal{V}) \\ S \subseteq T \subseteq \mathcal{V}}} (-1)^{|T \setminus S|} (1 - c_{\mathcal{G}}(T)) \mathsf{H}(X_S) \\
&= \sum_{S \in \mathcal{P}(\mathcal{V})} \sum_{\substack{T \in \mathcal{P}(\mathcal{V}) \\ S \subseteq T \subseteq \mathcal{V}}} (-1)^{|T \setminus S|} \delta_{T \notin \mathcal{S}(\mathcal{G})} \mathsf{H}(X_S) \\
&= \sum_{S \in \mathcal{P}(\mathcal{V})} \sum_{\substack{T \notin \mathcal{S}(\mathcal{G}) \\ S \subseteq T \subseteq \mathcal{V}}} (-1)^{|T \setminus S|} \mathsf{H}(X_S) \\
&= \sum_{T \notin \mathcal{S}(\mathcal{G})} \sum_{S \subseteq T} (-1)^{|T \setminus S|} \mathsf{H}(X_S) \\
&= \sum_{T \notin \mathcal{S}(\mathcal{G})} \mathsf{I}(X_T).
\end{aligned}$$

The last equality in the statement then holds, as $\langle \delta_{\mathcal{V}}, \mathsf{H} \rangle = \mathsf{H}(X_{\mathcal{V}})$. □

Remark 12. We will not use interactive information in our algorithms for scoring as we are scoring by taking the inner product between entropy and the imset constructed by summing semi-elementary imsets from the refined Markov property. But the equivalence between the inner product of 'standard imset' and entropy, and the inner product of 0-1 characteristic imset and the interactive information is interesting. If one further develops an efficient method to traverse between MECs of MAGs by using 0-1 characteristic imset as representation of MECs, then scoring by interactive information may be more efficient.

To help readers better understand the inner product score, we first show the equivalence between BIC and the score for discrete Bayesian networks.

4.4.2 Scoring discrete Bayesian networks

Recall that for DAGs, the maximum likelihood estimate (MLE) of $P(x_v \mid x_{\mathrm{pa}_v})$ is $N(x_v, x_{\mathrm{pa}_v})/N(x_{\mathrm{pa}_v})$. Given N i.i.d. samples, the log-likelihood can be expressed as:

$$\begin{aligned}
\ell(P; N) &= \sum_{x_{\mathcal{V}}} N(x_{\mathcal{V}}) \log P(x_{\mathcal{V}}) \\
&= \sum_{x_{\mathcal{V}}} N(x_{\mathcal{V}}) \sum_{v} \log P(x_v \mid x_{\mathrm{pa}(v)}) \\
&= \sum_{v} \sum_{x_v, x_{\mathrm{pa}(v)}} N(x_v, x_{\mathrm{pa}(v)}) \log P(x_v \mid x_{\mathrm{pa}(v)}) \\
&= \sum_{v} \sum_{x_{\mathrm{pa}(v)}} \sum_{x_v} N(x_v, x_{\mathrm{pa}(v)}) \log P(x_v \mid x_{\mathrm{pa}(v)}).
\end{aligned}$$

Replacing $P(x_v \mid x_{\mathrm{pa}(v)})$ by the MLE $N(x_v, x_{\mathrm{pa}_v})/N(x_{\mathrm{pa}_v})$, the log-likelihood is then

$$\begin{aligned}
\hat{\ell} &= \sum_{v} \sum_{\mathrm{pa}(v)} \sum_{x_v} N(x_v, x_{\mathrm{pa}(v)}) \log \frac{N(x_v, x_{\mathrm{pa}(v)})}{N(x_{\mathrm{pa}(v)})} \\
&= \sum_{v} \sum_{\mathrm{pa}(v)} \left\{ \sum_{x_v} N(x_v, x_{\mathrm{pa}(v)}) \log \frac{N(x_v, x_{\mathrm{pa}(v)})}{N} + \right. \\
&\qquad \left. - \sum_{x_v} N(x_v, x_{\mathrm{pa}(v)}) \log \frac{N(x_{\mathrm{pa}(v)})}{N} \right\} \\
&= \sum_{v} \sum_{\mathrm{pa}(v)} \sum_{x_v} N(x_v, x_{\mathrm{pa}(v)}) \log \frac{N(x_v, x_{\mathrm{pa}(v)})}{N} + \\
&\qquad - \sum_{v} \sum_{\mathrm{pa}(v)} N(x_{\mathrm{pa}(v)}) \log \frac{N(x_{\mathrm{pa}(v)})}{N} \\
&= N \langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, \hat{\mathsf{H}} \rangle
\end{aligned}$$

$$= N\langle c_{\mathcal{G}}, \hat{\mathsf{I}} \rangle.$$

The above calculations have appeared in the literature before (Studený, 2006) but we demonstrate it in our notation to make other proofs clear.

4.4.3 Consistency for MAGs when $u_{\mathcal{G}}$ is perfectly Markovian

In this section, we show that when fitted MAGs have 'standard imsets' that are perfectly Markovian w.r.t. the graphs, the scoring criterion is *consistent*. First, we define what is the requirement for a score to be consistent.

Definition 4.4.3. Let P be the true distribution. A score $S(\mathcal{G})$ is said to be consistent, if in the limit of infinite sample size, the following is true:

(i) if $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$ but $\mathcal{I}(\mathcal{G}') \not\subseteq \mathcal{I}(P)$, then $S(\mathcal{G}) < S(\mathcal{G}')$;

(ii) if $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$ and $\mathcal{I}(\mathcal{G}') \subseteq \mathcal{I}(P)'$ but $\mathcal{G}$ has less dimension than $\mathcal{G}'$, then $S(\mathcal{G}) < S(\mathcal{G}')$.

Next, we introduce the well-studied Kullback-Leibler divergence.

Definition 4.4.4. For two distributions P and Q defined over $X_{\mathcal{V}}$, and two disjoint subsets $A, C \subseteq V$, define the *Kullback-Leibler* (KL) divergence between $P_{A|C}$ and $Q_{A|C}$ as

$$KL(P_{A|C} \parallel Q_{A|C}) = \mathbf{E}_{P_{A|C}}[\log \frac{P_{A|C}}{Q_{A|C}}].$$

Two nice properties of KL divergence are that it is non-negative, and that it is zero if and only if $P = Q$ almost surely.

Proposition 4.4.2. *For any two distributions P and Q and any semi-elementary imset $u_{\langle A,B|C\rangle}$, provided that Q satisfies $X_A \perp\!\!\!\perp X_B \mid X_C$, we have*

$$\langle u_{\langle A,B|C\rangle}, KL(P \parallel Q)\rangle \geq 0,$$

with equality holding if and only if P also satisfies $X_A \perp\!\!\!\perp X_B \mid X_C$.

Proof.

$$\begin{aligned}
&\langle u_{\langle A,B|C\rangle}, KL(P \parallel Q)\rangle \\
&= \mathbf{E}_{P_C}[KL(P_{AB|C} \parallel Q_{AB|C}) - KL(P_{A|C} \parallel Q_{A|C}) - KL(P_{B|C} \parallel Q_{B|C})] \\
&= \mathbf{E}_{P_C}\left[\mathsf{I}(A; B \mid C) - \mathbf{E}_{P_{AB|C}} \log \frac{Q_{AB|C}}{Q_{A|C}Q_{B|C}}\right] \\
&= \mathbf{E}_{P_C}[\mathsf{I}(A; B \mid C)] \qquad \text{(provided } Q \text{ satisfies } A \perp\!\!\!\perp B \mid C)
\end{aligned}$$

$$\geq 0,$$

where the last inequality comes from the fact that mutual information is always non-negative. □

Next we show that the score is consistent.

Proposition 4.4.3. *The score* $-2N\langle c_{\mathcal{G}}, \hat{\mathsf{I}}\rangle + d\log N$, *is consistent when* $\mathcal{I}_{u_{\mathcal{G}}} = \mathcal{I}_{\mathcal{G}}$, *where* $\mathcal{G}$ *is the fitting MAG.*

Proof. Let us first consider any general MAG $\mathcal{G}$ and q^θ with the parameters $\theta = \{q(x_H \mid x_{\text{tail}(H)}) : H \in \mathcal{H}(\mathcal{G})\}$ (Evans and Richardson, 2014). Note that q^θ can be factorized by Theorem 3.4.18 and the log-likelihood can be rewritten as the inner product between $\delta_{\mathcal{V}} - u_{\mathcal{G}}$ and $\log q^\theta$ where the inner product is taken over all subsets of V:

$$\begin{aligned}\ell &= \sum_{x_{\mathcal{V}}} N(x_{\mathcal{V}})\langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, \log q^\theta\rangle \\ &= N\langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, \sum_{x_{\mathcal{V}}} \frac{N(x_{\mathcal{V}})}{N} \log q^\theta\rangle.\end{aligned}$$

The difference between $\hat{\ell}$ and $N\langle c_{\mathcal{G}}, \hat{\mathsf{I}}\rangle$ is then

$$\begin{aligned}\hat{\ell} - N\langle c_{\mathcal{G}}, \hat{\mathsf{I}}\rangle &= N\langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, \sum_{x_{\mathcal{V}}} \frac{N(x_{\mathcal{V}})}{N} \log q^{\hat{\theta}} - \hat{\mathsf{H}}\rangle \\ &= -N\langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, KL(\hat{P} \parallel q^{\hat{\theta}})\rangle.\end{aligned}$$

Hence the difference from the BIC of the fitted model $q^{\hat{\theta}}$ with graph $\mathcal{G}$ and the new score is

$$D_1 = 2N \times \langle \delta_{\mathcal{V}} - u_{\mathcal{G}}, KL(\hat{P} \parallel q^{\hat{\theta}})\rangle.$$

Moreover the difference between the fitted model $q^{\hat{\theta}}$ with graph $\mathcal{G}$ and the BIC of the true model $p^{\hat{\theta}'}$ is:

$$\begin{aligned}D_2 &= -2\sum_{x_{\mathcal{V}}} N(x_{\mathcal{V}}) \log q^{\hat{\theta}}(x_{\mathcal{V}}) + d_2 \log N + 2\sum_{x_{\mathcal{V}}} N(x_{\mathcal{V}}) \log p^{\hat{\theta}'}(x_{\mathcal{V}}) - d_1 \log N \\ &= 2N \sum_{x_{\mathcal{V}}} \hat{P}(x_{\mathcal{V}}) \log \frac{p^{\hat{\theta}'}(x_{\mathcal{V}})}{q^{\hat{\theta}}(x_{\mathcal{V}})} + (d_2 - d_1)\log N \\ &= 2N \sum_{x_{\mathcal{V}}} \hat{P}(x_{\mathcal{V}}) \log \left[\frac{\hat{P}(x_{\mathcal{V}})}{q^{\hat{\theta}}(x_{\mathcal{V}})} \frac{p^{\hat{\theta}'}(x_{\mathcal{V}})}{\hat{P}(x_{\mathcal{V}})}\right] + (d_2 - d_1)\log N \\ &= 2N \times KL(\hat{P} \parallel q^{\hat{\theta}}) - 2N \times KL(\hat{P} \parallel p^{\hat{\theta}'}) + (d_2 - d_1)\log N,\end{aligned}$$

where d_1, d_2 denote the dimension of the true model and fitted model, respectively.

To show that the score is consistent, it is sufficient to show that the following conditions are satisfied:

(i) when $D_2 = 0$, i.e. the fitted model is the true model, $D_1 = O_p(1)$;

(ii) when $D_2 > 0$, i.e. the fitted model is not the true model, we have $D_2 - D_1$ diverges at rate at least $O_p(\log N)$.

Consider condition (i) when we are fitting the true model. For any subset $A \subseteq V$, the term

$$2N \times KL(\hat{P}(X_A) \parallel q^{\hat{\theta}}(X_A))$$

has χ^2 distribution with some degree of freedom. Since $\delta_{\mathcal{V}} - u_{\mathcal{G}}$ has fixed number of terms, $D_1 = O_p(1)$ on all these as N tends to infinity.

Consider condition (ii), we have:

$$D_2 - D_1 = 2N \times \langle u_{\mathcal{G}}, KL(\hat{P} \parallel q^{\hat{\theta}}) \rangle - 2N \times KL(\hat{P} \parallel p^{\hat{\theta}'}) + (d_2 - d_1) \log N.$$

The second term $2N \times KL(\hat{P} \parallel p^{\hat{\theta}'})$ has $\chi^2_{d_1}$ distribution, so it is $O_p(1)$. Moreover, by Proposition 4.4.2 and given that $u = \sum u_{\langle A,B|C \rangle}$ is combinatorial, we have:

$$2N \times \langle u_{\mathcal{G}}, KL(\hat{P} \parallel q^{\hat{\theta}}) \rangle = 2N \sum \mathbb{E}_{\hat{P}_C}[\mathsf{I}(A; B \mid C)]$$

Now we need to split into two cases. The first case is if the fitted model contains the true model, but has higher dimension and contains more parameters then $d_2 > d_1$ so the third term $(d_2 - d_1) \log N$ is at $O(\log N)$. For the first term, because the fitted model contains the true model, it satisfies all the conditional independence corresponding to the semi-elementary imsets in u_G. Treating each $2N \times \mathbb{E}_{\hat{P}_C}[\mathsf{I}(A; B \mid C)]$ as a likelihood ratio test, it has chi-squared distribution, so in total the first two terms are at $O_p(1)$. Thus in this case, $D_2 - D_1 = O_p(\log N)$.

For the second case when the fitted model is wrong and does not contain the true model, given that $u_{\mathcal{G}}$ is combinatorial and is perfectly Markovian with respect to the graph, there exists at least one conditional independence in $u_{\mathcal{G}} = \sum u_{\langle A,B|C \rangle}$ such that $\mathbb{E}_{P_C}[\mathsf{I}(A; B \mid C)] = \lambda > 0$. Empirically, when N tends to infinity, $\mathbb{E}_{\hat{P}_C}[\mathsf{I}(A; B \mid C)]$ will be close to λ, and hence the first term in $D_2 - D_1$ grows at $O_p(N)$. Even if the third term $(d_2 - d_1) \log N$ is negative, it grows at $O(\log N)$, which is slower than the first term, hence in the second case, $D_2 - D_1 = O_p(N)$. □

4.4.4 Scoring MAGs using the refined Markov property

Given a MAG $\mathcal{G}$, let $u^r_{\mathcal{G}}$ denote the imset from the refined Markov property in Hu and Evans (2022). We propose to use the following score:

$$S^r_{\mathcal{G}} = -2N\langle \delta_{\mathcal{V}} - u^r_{\mathcal{G}}, \hat{\mathsf{H}} \rangle + d \log N,$$

where $\hat{\mathsf{H}}$ is the vector of empirical estimates of entropy over every subset of V, d is the dimension of the model and N is the sample size.

The idea of scoring MAGs with inner product between imsets and empirical entropy originated from Andrews et al. (2022). They used imsets constructed from their new Markov property which does not have polynomial bound on number of independences unlike the refined Markov property.

The BIC of DAGs and MAGs are known to be *score-equivalent*, that is, Markov equivalent graphs have the same score. This unfortunately does not hold for $S^r_{\mathcal{G}}$ since Markov equivalent MAGs may have the different list of conditional independences in the refined Markov property, which, though, are still equivalent under semi-graphoids.

For learning algorithms searching in the space of MECs, score-equivalent is not a necessary property as long as the scores are *consistent*. Next we show that the score $S^r_{\mathcal{G}}$ is consistent.

Proposition 4.4.4. *The score $S^r_{\mathcal{G}}$ is a consistent score.*

Proof. Suppose $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$ but $\mathcal{I}(\mathcal{G}') \not\subseteq \mathcal{I}(P)$, then there is at least one independence $I = A \perp\!\!\!\perp B \mid C$ in the refined Markov property of $\mathcal{G}'$ that is not satisfied by P. Then $\langle u_I, \hat{\mathsf{H}} \rangle$ will converge to the true mutual information $c > 0$ of I, hence $S^r_{\mathcal{G}'}$ grows at $O(N)$. On the other hand, $N\langle u^r_{\mathcal{G}}, \hat{\mathsf{H}} \rangle$ grows at $O(1)$ as all the independences are satisfied, so $S^r_{\mathcal{G}}$ grows at $O(\log N)$.

Suppose $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$ and $\mathcal{I}(\mathcal{G}') \subseteq \mathcal{I}(P)'$. Then both $S^r_{\mathcal{G}}$ and $S^r_{\mathcal{G}'}$ grow at $O(\log N)$ but since $\mathcal{G}$ has less dimension than $\mathcal{G}'$, $S(\mathcal{G}) < S(\mathcal{G}')$ in the infinite sample size limit. □

In principal, any Markov property can be used to construct an imset for scoring. The reason to use the refined Markov property is that, Hu and Evans (2022) showed that if the maximal head size is k, then the imset can be constructed in $O(kn^k(n+e))$ time while there is no polynomial bound on computing the global Markov property or the ordered local Markov property Richardson (2003).

If one assumes additional graphoids, then the *pairwise* Markov property (Sadeghi et al., 2014) is shown to be equivalent to the global Markov property and hence can

be used for scoring. It can be constructed in polynomial time. However, as we will show by simulation, since the pairwise Markov property requires to condition on ancestors of non-adjacent pair of nodes, its performance is worse than the refined Markov property if the ancestral relations are complicated. This is because it would require to estimate entropy of large set of variables.

Having all the theories being introduced we now are able to describe the full algorithm to score a MEC represented by an arrow complete PAG.

Suppose we have a $n \times p$ data matrix $\mathcal{D}$ where $\mathcal{D}_{ij}$ is the ith observation of jth variable. Algorithm `SCORE` computes the score of $\mathcal{P}$ by using the imset from the refined Markov property.

Input: An arrow complete PAG $\mathcal{P}$, a $N \times n$ data matrix $\mathcal{D}$
Result: A score from the refined Markov property
Let $\mathcal{G} =$ `PAG-to-MAG`$(\mathcal{P})$;
if $\mathcal{G}$ *is not a MAG* **then**
  return $False$
end
Compute $u^r_{\mathcal{G}}$ and dimension d of the model of $\mathcal{G}$;
return $-2N\langle \delta_{[}n] - u^r_{\mathcal{G}}, \hat{\mathsf{H}} \rangle + d \log N$

Algorithm `SCORE`**:**

4.5 Greedy learning algorithm

We describe our MAG learning algorithm here. The Algorithm `GESMAG` essentially explores every possible edge to add or delete and score every PAGs returned by `Add-adj`, `Delete-adj` and `Turning`. If there is a reduction on the score then update the score and PAG. We only list the adding phase here as the two other phases are similar.

4.5.1 $\mathcal{I}$-maps given maximal head size

In our package for `GESMAG`, we also implement a choice for searching with restricted maximal head size. It has a few practical advantages compared to no such restriction as we will shown by experiment. But first, let us justify such restriction.

Proposition 4.5.1. *Given a MAG $\mathcal{G}$ with maximal head size $k \geq 2$, then for any integer k', where $1 \leq k' < k$, there exists a MAG $\mathcal{G}'$ with maximal head size k' such that $\mathcal{I}(\mathcal{G}') \subseteq \mathcal{I}(\mathcal{G})$.*

Input: A $n \times p$ data matrix $\mathcal{D}$
Result: A PAG $\mathcal{P}$
Initialize $\mathcal{P}$ as an empty graph with n nodes;
Score = $\texttt{SCORE}(\mathcal{P}, \mathcal{D})$ and Move = $\{\{i, j\} \mid 1 \leq i \neq j \leq n\}$;
Update = $True$;
while *Update* **do**
    Update = $False$;
    $\mathcal{P}_{prev} = \mathcal{P}$;
    for $\{i, j\} \in Move$ **do**
        $\mathcal{O} = \texttt{Add-adj}(\mathcal{P}_{prev}, \{i, j\})$;
        for $\mathcal{P}' \in \mathcal{O}$ **do**
            $\text{Score}_{new} = \texttt{SCORE}(\mathcal{P}', \mathcal{D})$;
            if $Score_{new} \neq False$ *and* $Score_{new} < Score$ **then**
                Score = Score_{new} and $\mathcal{P} = \mathcal{P}'$;
                Update = $True$;
            end
        end
    end
    orient tail of $\mathcal{P}$;
end
return $\mathcal{P}$

Algorithm `GESMAG`**:** Adding phase

Proof. It is sufficient to prove that there exists such a MAG $\mathcal{G}'$ with maximal head size $k-1$. Consider any head H in $\mathcal{G}$ with size k. Let $x, y \in H$ and $x \neq y$. We add $x \to y$ to $\mathcal{G}$ and let the resulting ADMG be $\mathcal{G}'$. Then by Proposition 3.6 in Hu and Evans (2020), which says that the ADMG to MAG projection preserves heads and tails, it is sufficient to prove that there is no new head in $\mathcal{G}'$ with size greater or equal to k and there is a head in $\mathcal{G}'$ with size $k-1$.

Let $H' := \text{barren}(H)$, since H is in the same district in $\mathcal{G}$ clearly H' lies in the same district in $\mathcal{G}'$. Hence by definition, H' is a head and its size is clearly $k-1$.

Now suppose there is a new head H' in $\mathcal{G}'$ which has size greater or equal to k and is not a head in $\mathcal{G}$. Then the reason for this must be that the vertices in H' do not lie in the same district in $\mathcal{G}$. Consider $H'' := \text{barren}_{\mathcal{G}}(\text{an}_{\mathcal{G}'}(H'))$. By construction, H'' is a head in $\mathcal{G}$ and $H' \subseteq H''$. If $H' = H''$ then H' would be a head in $\mathcal{G}$. Hence $H' \subset H''$ and H'' has size larger than k. This contradicts the assumption. □

Our proof is constructive but not for constructing a minimal $\mathcal{I}$-map.

Note that Proposition 4.5.1 essentially generalize the result in Ogarrio et al. (2016), which shows that the skeleton of output of GES will consistently contain the skeleton of underlying true MAG in the limit of infinite sample size.

Hence assuming Meek's conjecture, we can start to only add adjacency from the skeleton of output of GES. This would reduce complexity.

We should also point out that although Proposition 4.5.1 ensures that with restricted head size, in the limit of infinite sample size, the independence model of the output of `GESMAG` is contained in the true model, its PAG may contain additional invariant edge marks that has no causal meaning.

4.5.2 Bounding the complexity

We show here that under some assumption of graph structure, the complexity of `GESMAG` can be bounded in polynomial time for sparse graphs in terms of number of variables, maximal degree, maximal head size and maximal number of discriminating paths. This is similar to the result in Claassen et al. (2013), which shows that the constrained-based approach FCI+ is of polynomial time by considering sparse graphs. We prove similar result for our score-based approach, for which in nature is more complicated and time consuming than constrained-based algorithms and therefore require further assumptions.

Proposition 4.5.2. *The complexities of the adding and deleting phase of* `GESMAG` *are polynomial, if the following are restricted: maximal degree, maximal head size, and maximal number of discriminating path.*

Proof. It is sufficient to prove that the complexity of adding phase is bounded. In Algorithm `GESMAG`. The first and second loop at Line 4 and 7 repeats at most $n(n-1)/2$ times. Because number of maximal degree and maximal number of discriminating path are bounded, the third loop at Line 9 is bounded. Now if we fix the maximal head size, Hu and Evans (2022) showed that the imsets from the refined Markov property can be constructed in polynomial time. Hence scoring at Line 10 is also polynomial. □

We have bounded the complexity of the first two phases of `GESMAG`. The turning phase is potentially exponential even if we change orientation of only one unshielded triple at each time. But in practice, the time spent on turning phase usually does not exceed *one third* of the time spent on the first two phases. One can also restrict the number of iteration in turning phase.

4.6 Experiments

We conduct experiments on simulated data, run `GESMAG` and make comparison to GPS (Claassen and Bucur, 2022) (base and hybrid version), GFCI (Ogarrio et al.,

2016) and the classic FCI (Spirtes et al., 2000). GPS is the only existing purely scored-based algorithm that searches in the space of MECs. GFCI is a hybrid learning algorithm that first performs GES and then FCI on the skeleton of GES output. FCI is a purely constraint-based algorithm. For other score-based algorithm that explores in the space of MAGs, GPS in Claassen and Bucur (2022) shows superior performance than them so we do not include them in the experiment section.

4.6.1 Simulate MAGs

For each $n \in \{5, 10, 15, 20\}$ and $p_d \in \{0.8, 0.6, 0.4\}$, we randomly generate 100 ADMGs such that the average degree is 3. For each edge, the probabilities of being directed is p_d and otherwise bidirected. Then we project each ADMG to a Markov equivalent MAG(Richardson and Spirtes, 2002) and we simulate a linear Gaussian MAG graphical model such that the coefficients of directed and bidirected edges are drawn uniformly from $\pm[0.1, 1]$.

Most previous score-based algorithms simulated graphs with small districts size (two or three) (Chen et al., 2021) or low probability of bidirected edges (Claassen and Bucur, 2022) and hence relatively small maximal head size (around 30, out of 100, MAGs are simple for $n = 20, p_d = 0.8$). Part of the reason for this is that BIC does not perform well when districts are large. We will show by experiments that the imset score performs better than BIC, and not only when head size is small.

Before we proceed, we'd like to have an empirical study of how p_d effects head size and this will be helpful for the analysis of performance of algorithms later.

4.6.1.1 Different maximal head size

By maximal head size, we mean the size of largest head in a MAG. For each simulated MAG, we compute its maximal head size, then we use histograms in Figure 4.3 to illustrate different frequencies of maximal head size under different probabilities of directed edges and each $n \in \{10, 15, 20\}$ (for $n = 5$, there is not much difference and we put in Appendix). This is important as they can partly explain the variation of performance of algorithms under different probabilities of directed edges.

Clearly, as p_d decreases, it is more likely to have larger heads. In particular, the largest maximal head appear when $n = 20$ and $p_d = 0.4$, which also most double the largest maximal head size when $n = 20$ and $p_d = 0.6$.

As there are many variation of our algorithms, we decide that firstly, we compare the performance of `GESMAG` under different turning stages, then we compare `GESMAG` to other algorithms with the number of simultaneous turns to consider t being set to one.

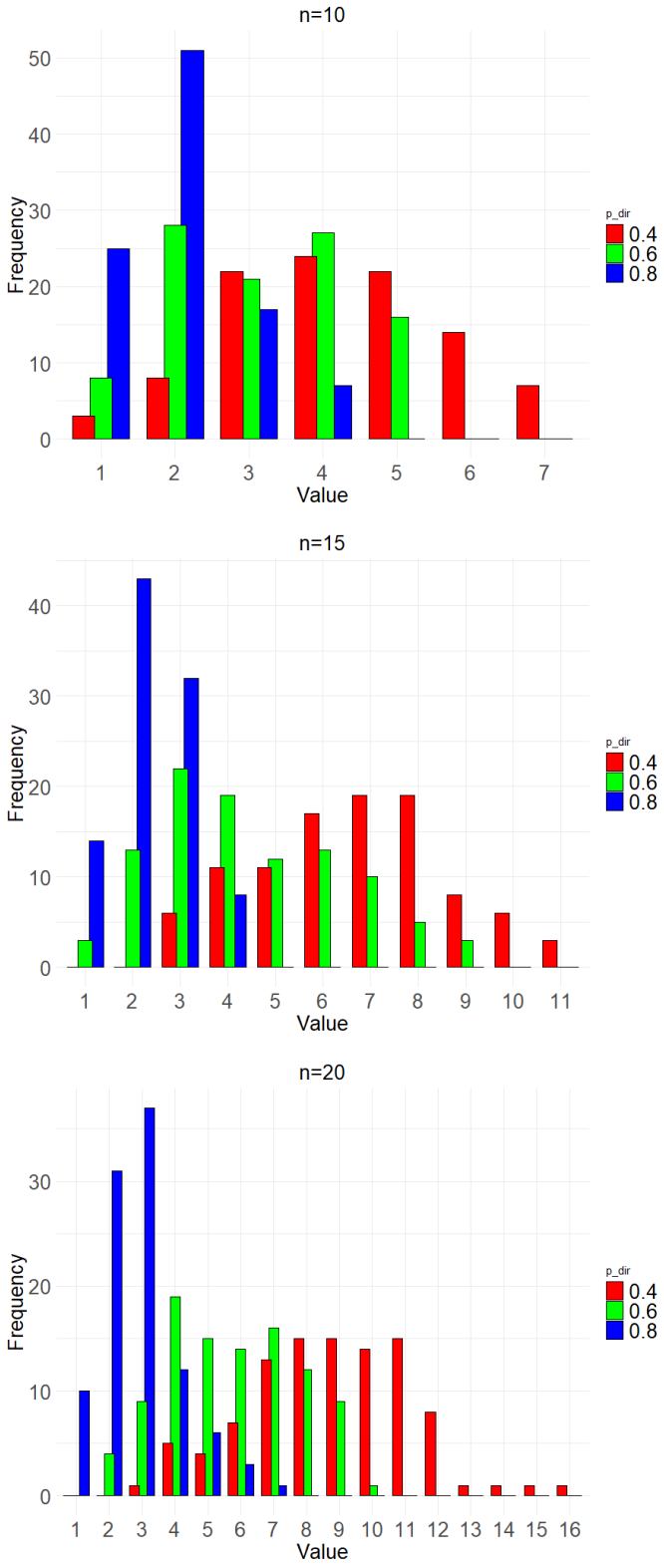


Figure 4.3: Histograms of maximal head size for $n = 10, 15, 20$

4.6.2 Metrics for performance

A common approach to evaluate the performance of structure learning algorithms is the *accuracy* of edge marks by comparing the edge marks on the output PAG with the ground truth PAG. In addition to this, we also include TP (*true positive rate*), FP (*false positive rate*) for each kind of edge in Appendix .1.

Another metric we use is the logarithm (for scale purpose) of difference between BIC of true model and BIC of estimated model. The lower it is, the closer the estimated model is to the true model.

4.6.3 Performance of algorithms

Notice that the baseline version of GPS consider new triple with order as non-collider by default, whereas we explore both options and hence our algorithm should be compared to hybrid or extended version of GPS. The extended version of GPS showed similar performance compared to its hybrid version in terms of accuracy and average BIC, and the computational time is longer than the hybrid version, hence we omit it in the plot.

Note that ROMP(i,j) stands for scoring by refined (ordinary) Markov property, searching with restricted head size i and turning_phase = j, and anc(j) stands for scoring by pairwise Markov property (Sadeghi et al., 2014). Also, ROMP(j) stands for scoring by refined (ordinary) Markov property, searching with unrestricted head size and turning_phase = j.

4.6.3.1 Comparison of GESMAG with different hyper parameters

In Figure 4.4, we plot accuracy of our algorithms with different restricted head size against number of variables. There are three plots corresponding to each $p_d \in \{0.8, 0.6, 0.4\}$.

Despite of a few fluctuation, there is a tendency for increasing accuracy as number of variables grows. This is because we fixed the average degree to three, therefore, as graphs grow, they become sparser and hence easier to determine edge mark orientations. The performance of ROMP(i,0) for any i is always worse than the corresponding ROMP(i,1) at costs of more computation time, for which we will analysis later. Moreover, one can observe that for different p_d, the best performance of ROMP(i,j) is of different restricted head size i. This can be explained by the following. Suppose the maximal head size of underlying true MAG is i, and if we search by not restricting head size or restricting to larger head size, then we explore more MECs and empirically this means that it is more likely to fall into

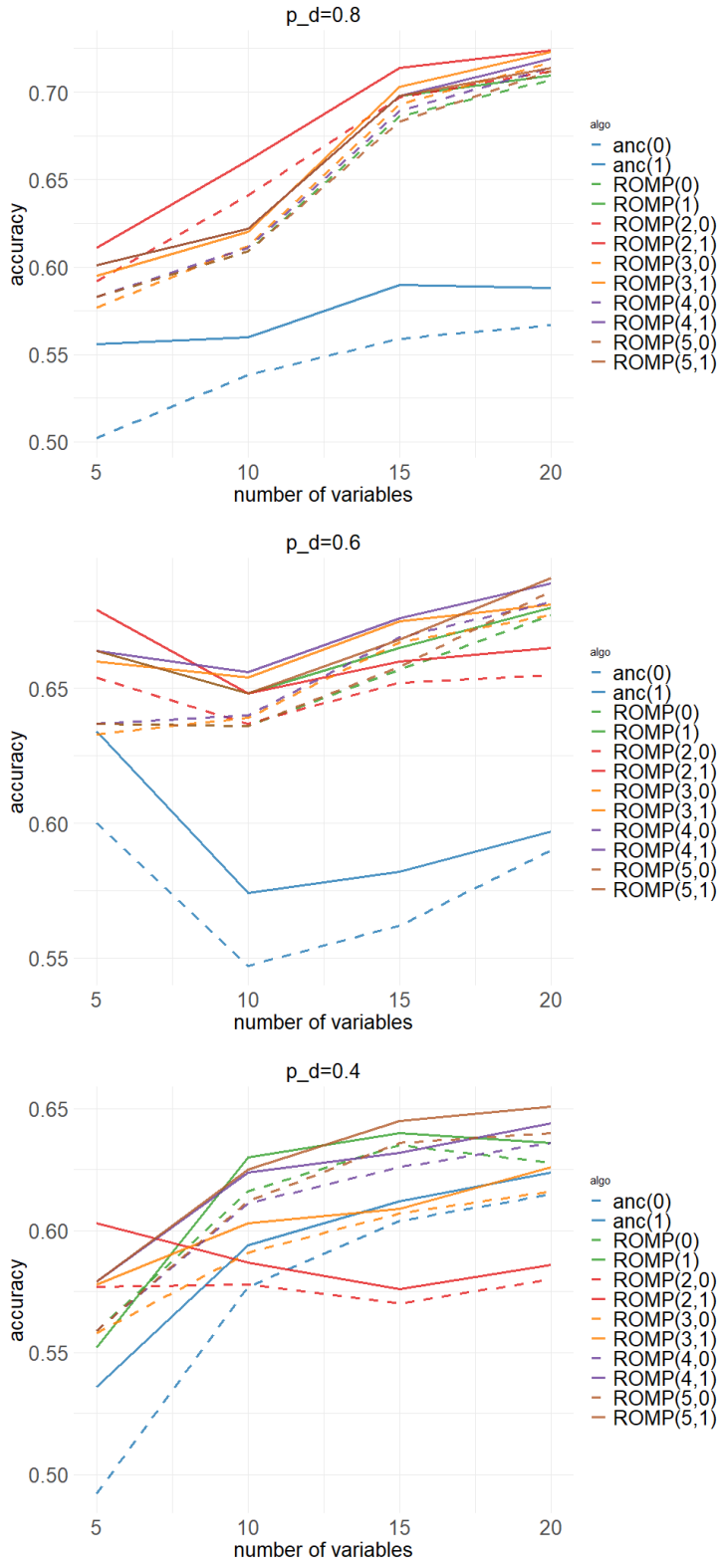


Figure 4.4: Accuracy of algorithms that score by using imsets

local optimum or make false decision at each step. Hence we suggest that if one has prior knowledge about size of head or district, then restricting search space can output more robust results.

Similarly, one can observe the above phenomenon for logarithm of difference between true BIC and BIC of estimated PAG. See Figure 4.5.

Next we plot logarithm of computation time of each variation of `GESMAG` in Figure 4.6

There are two key observations here. Firstly, scoring by taking imsets from pairwise Markov property in general are more time consuming, where the computational cost grows faster than imsets from refined Markov property. The complexity of computing pairwise Markov property, though, can be bounded in polynomial time. Secondly, if we do not restrict head size, ROMP(j) spent much longer time than others apart from when $p_d = 0.8$. This is because we expect much larger head size when $n = 20$ and $p_d = 0.6/0.4$ as we have seen in Figure 4.3, and also that the refined Markov property is computed iteratively, and its computational time goes exponentially as size of heads grows.

4.6.3.2 Comparison of `GESMAG` and other algorithms

Now we compare our algorithms to other MAG learning algorithms.

In Figure 4.7 and Figure 4.8, we compare different variation of `GESMAG`, baseline/hybrid version of GPS and FCI/GFCI. We have the accuracy plots and the plots of logarithm of average difference in BIC, respectively.

Clearly, one can see that variations of `GESMAG` outperforms other algorithms. Compared to baseline/hybrid GPS, FCI/GFCI show better performance in terms of edge mark accuracy but much worse performance in terms of BIC. This is not surprised as the objective of FCI/GFCI is not minimizing BIC but GPS's objective is.

For computational time, FCI/GFCI all spend around 1.2 seconds for each data set regardless of number of variables. This is because of the well designed package (*rcausal* in R) that supports the algorithms and the nature of FCI/GFCI (constrained based). They explore significantly fewer number of MECs than score-based approaches. On the other hand, one can clearly see that the time required for hybrid version of GPS grows much faster than time for `GESMAG`. While the computational time of the base version of GPS is close to `GESMAG`, this baseline version has some fairly basic flaws in nature. In brief, it sets any new triple with orders to non-collider by default and does not explore both options and hence easily becomes stuck in the local optimum.

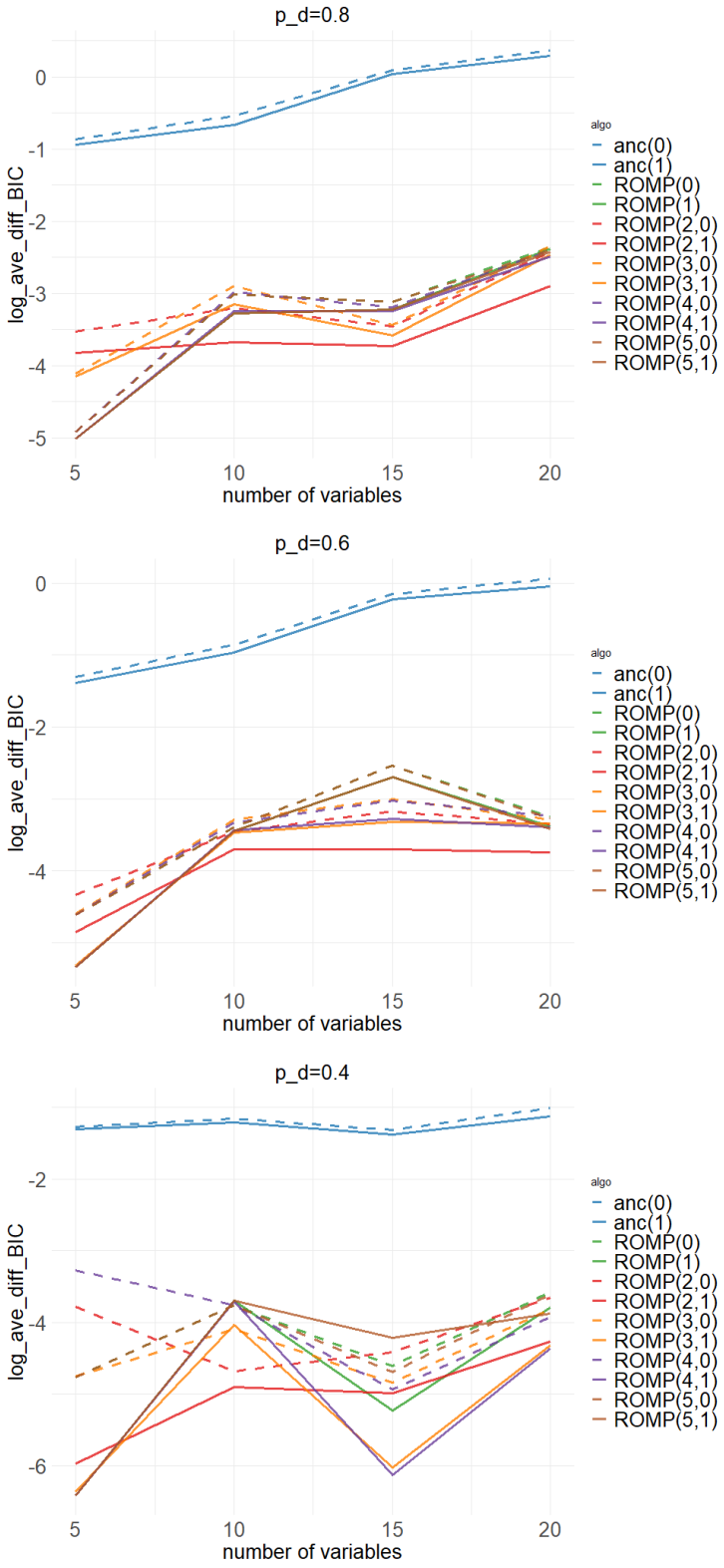


Figure 4.5: Log of average difference in BIC of algorithms that score by using imsets

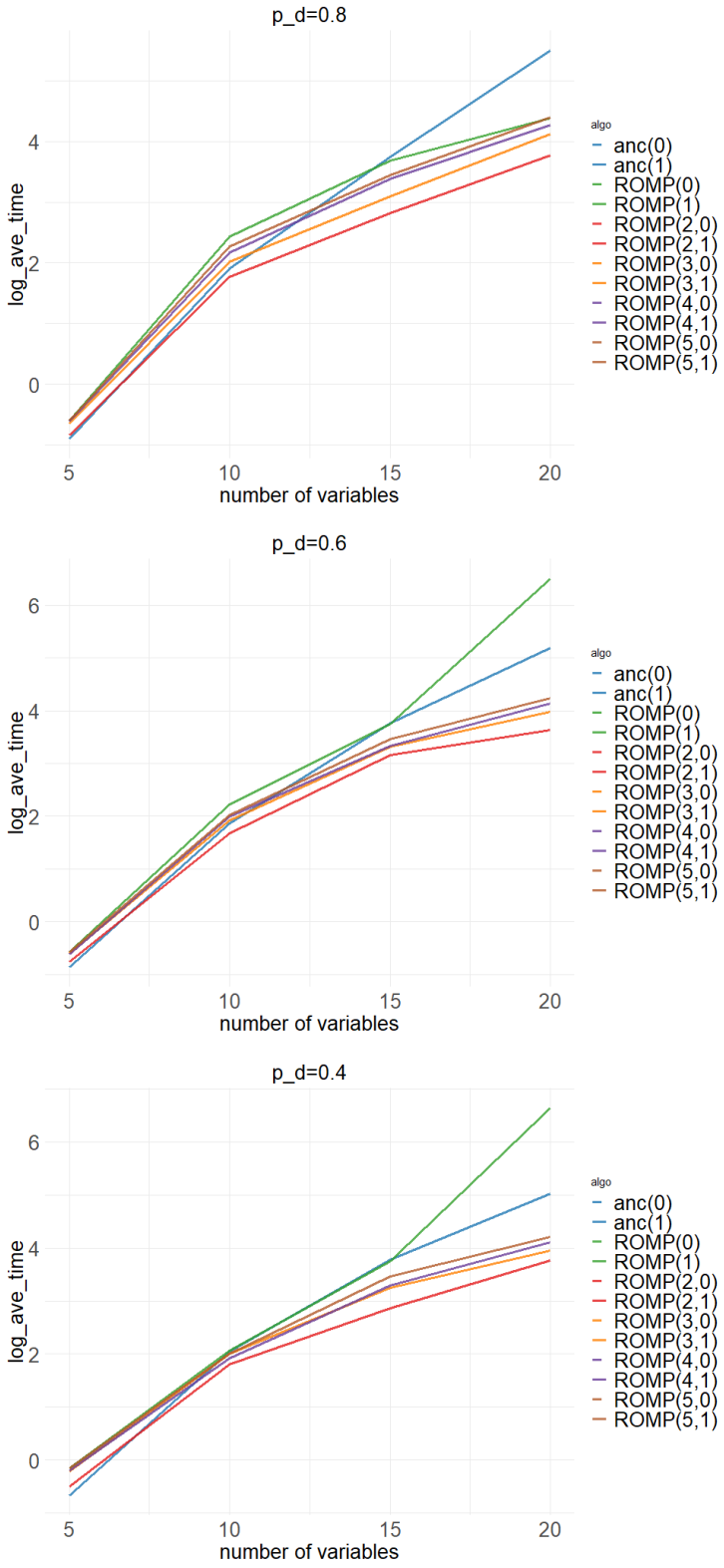


Figure 4.6: logarithm of computation time of algorithms that score by using imsets

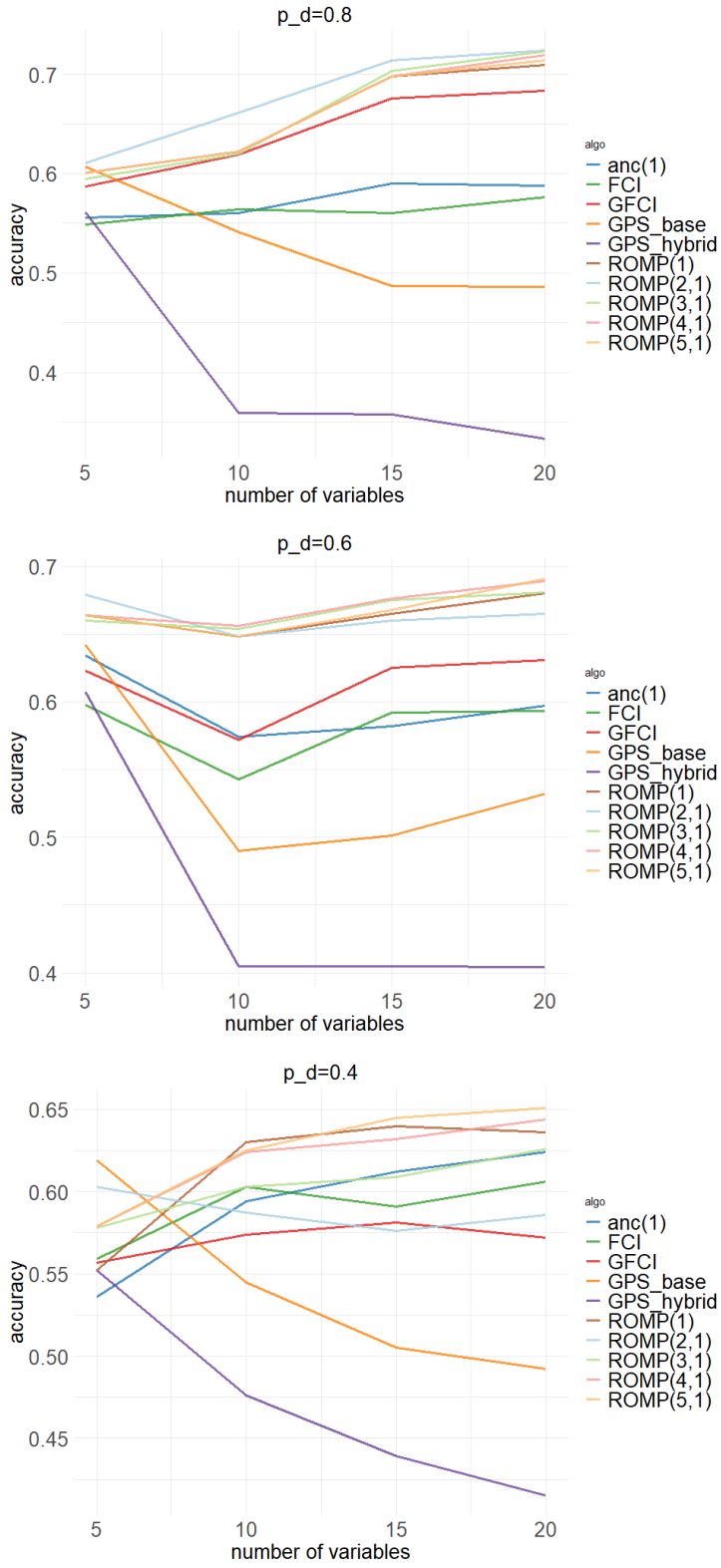


Figure 4.7: Accuracy of different algorithms

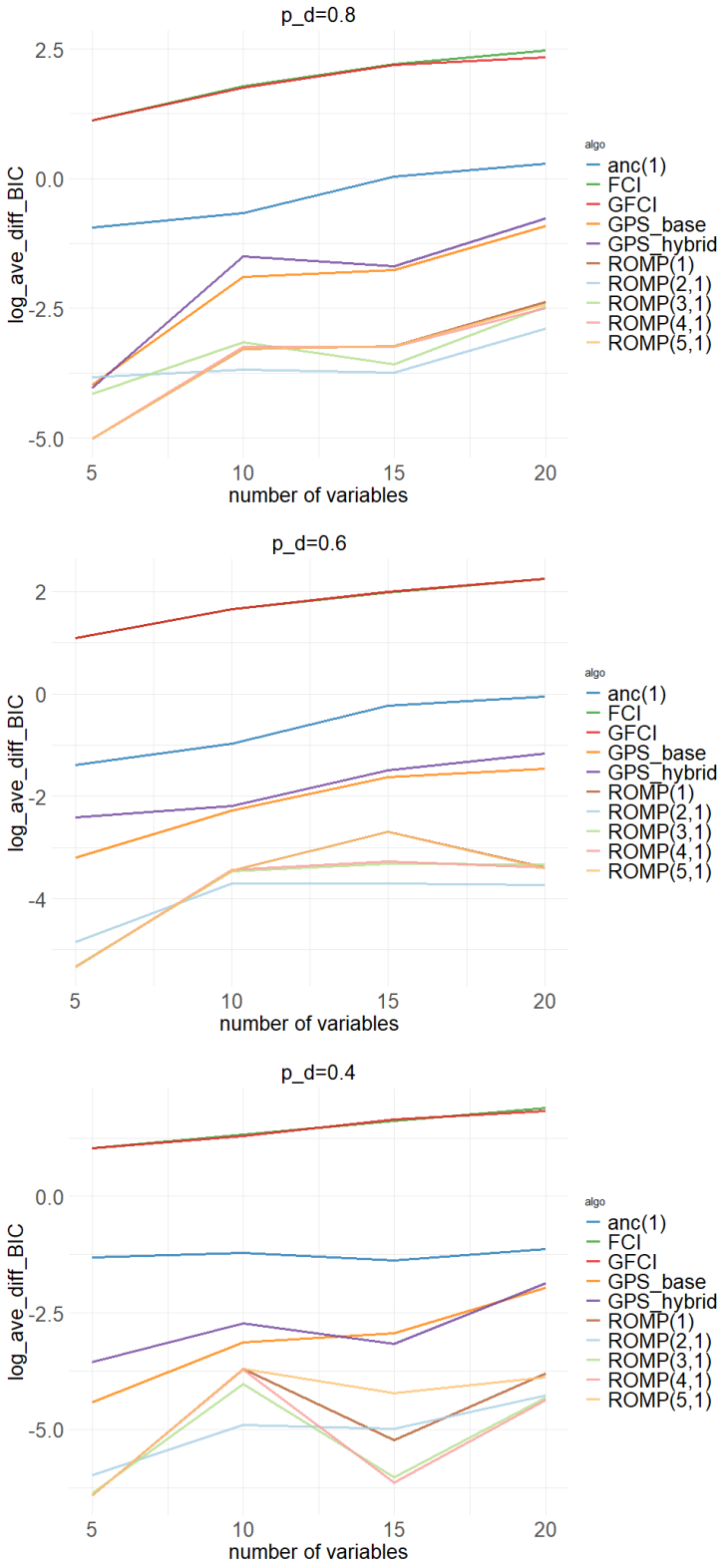


Figure 4.8: Log of average difference in BIC of different algorithms

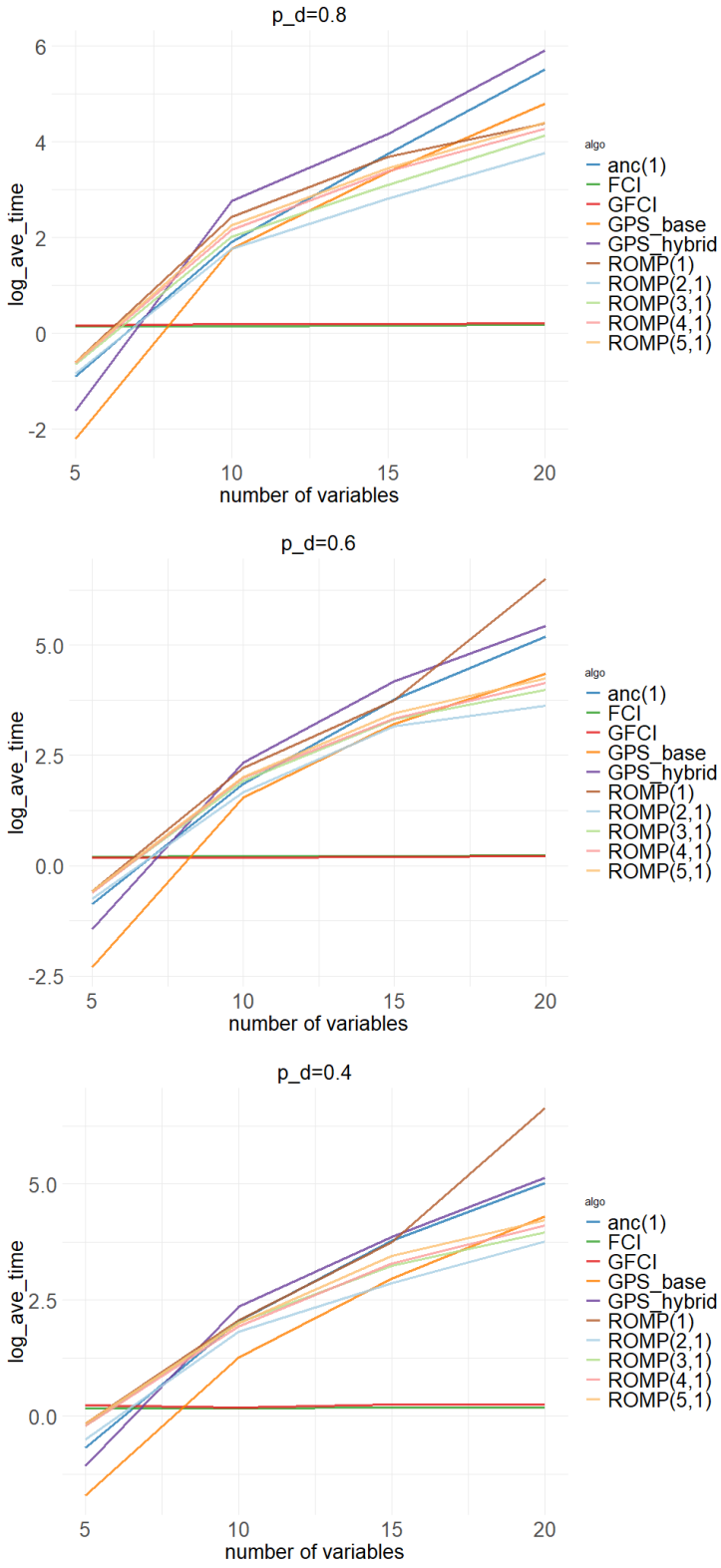


Figure 4.9: logarithm of computational time of different algorithms

Moreover, as p_d decreases, the running time of hybrid version of GPS increases for fixed number of variables, this suggests that using BIC as score are more easily to fall into local optimum if district/head size is large. On the other hand, although `GESMAG` without restricting head size requires more time if sizes of district or head grow large, the algorithms still remain high accuracy.

We split our contribution into two parts. The average percentage of time spent on scoring ranges from around 40% to 60% as head size varies from two to five while GPS usually spent around $40\% - 50\%$ on scoring. As the overall computational time is improved, we conclude that the revised search strategy improves search efficiency compared to GPS. This improvement is however not significant and our main contribution is to propose scoring by imsets from various Markov property, in particular the refined Markov property clearly shows best performance in terms of both edge mark accuracy and BIC.

There are still more issues to be sorted so `GESMAG` can be further accelerated, see discussion at Chapter 5.

Chapter 5

Discussion

In this Chapter, we summarize our contribution, then discuss several possible future work.

5.1 Summary of contribution

In Chapter 2, a new characterization for MECs of MAGs or ADMGs is proposed based on concept of heads, tails and parametrizing sets, together with efficient polynomial time algorithms to construct such representation. Then by spotting the similarity between parametrizing sets and characteristic imsets of DAG models, we propose a formulae for the 'standard' imset of MAG models. We explore the formulae in Chapter 3 and provide theoretical proof for when the imsets are valid for a wide range sub-models of MAG models. Further, based on decomposition of the 'standard' imset, we proposed a novel tool called power DAGs that help reducing the Markov property of MAGs and we show that the new refined Markov property can be constructed in polynomial time if we restrict maximal head size.

Next, in the Chapter 4, we apply the results in Chapter 3 to propose a new MAG search algorithm with scoring criteria that is different from BIC. At the same time, we also present several theoretical results about MAGs whose independence models contain another (Proposition 4.3.3 and Proposition 4.5.1). The empirical experiment shows very promising results. In terms of accuracy and BIC, `GESMAG` beats both constrained-based and score-based methods. On the other hand, for complexity, compared to the only other score based method (Claassen and Bucur, 2022) that search in space of MECs, our algorithm's computational time costs less and we also provide a polynomial bound on the complexity of algorithms under the assumption of maximal head size and sparsity of graphs.

5.2 Future work

We first discuss issues about using the parametrizing set or characterizing imset as representation of MECs. In this thesis, the result of parametrizing set characterization of MECs is only used in Chapter 4, where we use it to derive that Markov equivalent MAGs have the same 'standard' imset. In the beginning of development of `GESMAG`, we use the parametrizing set as representation of MECs. The problem is that it is in general difficult to determine what are the possible new sets after we add a new edge. The unshielded triples are easy to locate but there are many other possible sets that are non-local in the graph; indeed, the number of possible combination grows exponentially and we do not have enough theoretical tools to narrow them down. Moreover, we have to verify if the new parametrizing set actually corresponds to any MAG. We do not have direct approach to do so. Instead, we construct PAG based on the new parametrizing set and we check if the representative graph of the PAG is actually a MAG. In this way, the computation cost is certainly more than directly using PAG as representation of MECs. However, if one finds an efficient way to traverse between MECs that are represented by the parametrizing set, algorithms could be accelerated by using the parametrizing set as representation of MECs.

In Chapter 4, we present a method to traverse between MECs represented by PAGs. We apply orientation rules to construct new PAGs from beginning for every time we visit a new MEC. This procedure could be improved if one finds a method that can output the new PAG by directly operating on the previous PAG. This is in analogue to Chickering (2002), where he presents theorems about necessary and sufficient conditions for when such a modification is needed.

For DAG models, linear programming techniques have been used for graph learning algorithms (Jaakkola et al., 2010; Cussens, 2020), the BIC of DAG models can be expressed as inner product between the standard imsets and empirical entropy (or characteristic imsets and empirical interaction information), in addition with some penalty terms for model complexity. For MAG models, we suggest the following to investigate.

Although we know that the 0/1 characteristic imset or 'standard' imset does not necessarily induce the correct model, we did prove that for simple MAGs, which is a meaningful generalization of DAG models, these imsets can be used for scoring. The problem of applying LP approaches for learning simple MAGs is that it is hard to express the conditions of being simple MAGs into linear constraints. The acyclicity constraint is the same as DAG models, and Chen et al. (2021) shows how to put 'ancestral' into linear constraints. The issue lies on 'maximal' and 'simple', which

are related. If one finds a linear constraint for maximality, then it is easy to check if the MAG is simple. This is because, in Section 3.3.6, we have discussed that given the graph is a MAG, a necessary and sufficient condition for being simple is that for any two siblings of a node, the two siblings must have ancestral relation (Evans and Richardson, 2013), which is a similar constraint to being ancestral. On the other hand, it is not necessary to constrain the graph to be a MAG as we have shown in Chapter 2, any projection from ADMGs to MAGs preserves the heads and tails structure, so if provided that the maximal head size of an ADMG is two then we can still use the imset for scoring. However it is then not straightforward to restrict maximal head size by linear constraints.

We also have the following conjecture. Suppose we have a Lebesgue continuous prior on the distributions of a true model, and we use the 'standard' imset of a MAG for scoring. If the 'standard' imset is structural but not perfectly Markovian and the distribution is not Markov to the MAG, the probability of giving a zero inner product is zero. Similarly if the imset is not structural, by Theorem 3.4.18, we know the inner product is zero if the distribution is Markov to the model, and we conjecture again that the probability of getting an exact zero inner product is also zero if the distribution is not Markov to the model. This would be useful because we can then use the 'standard' imsets for scoring and it would be a score-equivalent scoring criteria.

For the refined Markov property, there are two possible directions to dig into. As we have discussed in Chapter 3, it can be further simplified and we have provided examples. Secondly, when we use imsets from the refined Markov property for scoring, we compute the imset from scratch, ignoring the list of independences from previous MEC. This certainly can be improved if one finds a method to utilize past information. In the case of the BIC, for example, it can be decomposed into districts. Therefore, if the district has not changed, as well as the parents of the district, the local score of this district then would not be changed. This may hold for the refined Markov property. Consider $1 \leftrightarrow 3 \leftrightarrow 4 \leftarrow 2 \leftarrow 1$ with numerical ordering. If we remove $4 \leftarrow 1$, the procedure for computing the refined Markov property for 4 would be changed since previously $\{1, 3, 4\}$ was a head but not after anymore, the list of independences, though, remain the same.

Lastly, the power DAG approach to simplify Markov property is intriguing and one may try to develop similar graphical tools on sets of nodes for other graphical models.

Appendices

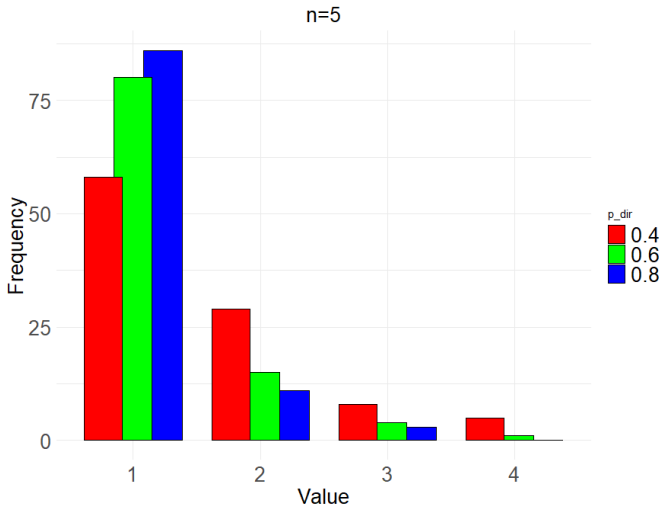


Figure 1: Histogram of maximal head size for n=5

.1 Extra plots

Figure 1 is the histogram plot of maximal head size for $n = 5$.

In addition, we provide extra plots for comparison between variations of `GESMAG` and other MAG learning algorithms, in terms of *accuracy*, *true positive rate* (TPR) and *false positive rate* (FPR) of adjacencies and each kind of edges in PAGs, which are, directed edge ($\rightarrow$), bidirected edge ($\leftrightarrow$), partially directed edge ($\circ\!\!\rightarrow$) and not directed edge ($\circ\!\!-\!\!\circ$)

Because the accuracy is computed by dividing possible number of edges, which is large compared to the number of edges that are actually present in graphs, we suggest that TPR and FPR plots reflect more information.

For plots of adjacencies, `GESMAG` are better than others. The low TPR value of FCI and GFCI suggests that the confidence level should be increased. The baseline and hybrid version of GPS shows poor performance in edge FPR plot, suggesting that the algorithms add wrong edges more often than others.

For directed or bidirected edges, although FCI and GFCI shows better or close performance compared to variations of `GESMAG` in the accuracy plots, `GESMAG` still win in terms of TPR. Once again, GPS shows poor performance in terms of FPR, which means it often gives false directed or bidirected edges. We argue that this may result from the instability of BIC. When there are more arrows in the PAG, it is more likely to have large districts. On the opposite of this, GPS performs poorly in terms of TPR for both partially directed and not directed edges as it tends to orient triples with orders as colliders. Hence we also believe that it is the reason for why GPS performs best in terms of FPR of partially directed and not directed edges.

In summary, BIC performs poorly in terms of both adjacency and edge orientation and becomes unstable especially when the sizes of districts are large.

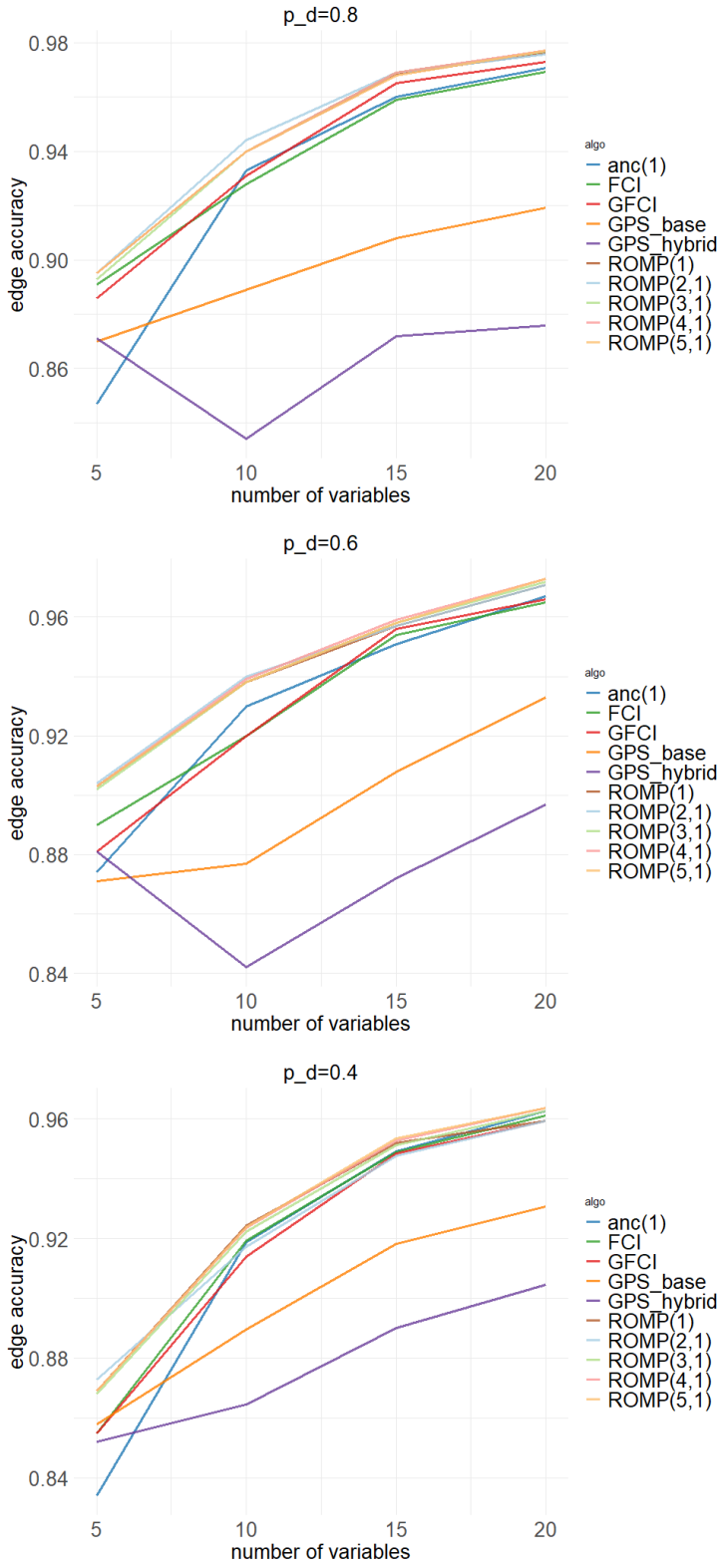


Figure 2: adjacency accuracy plots

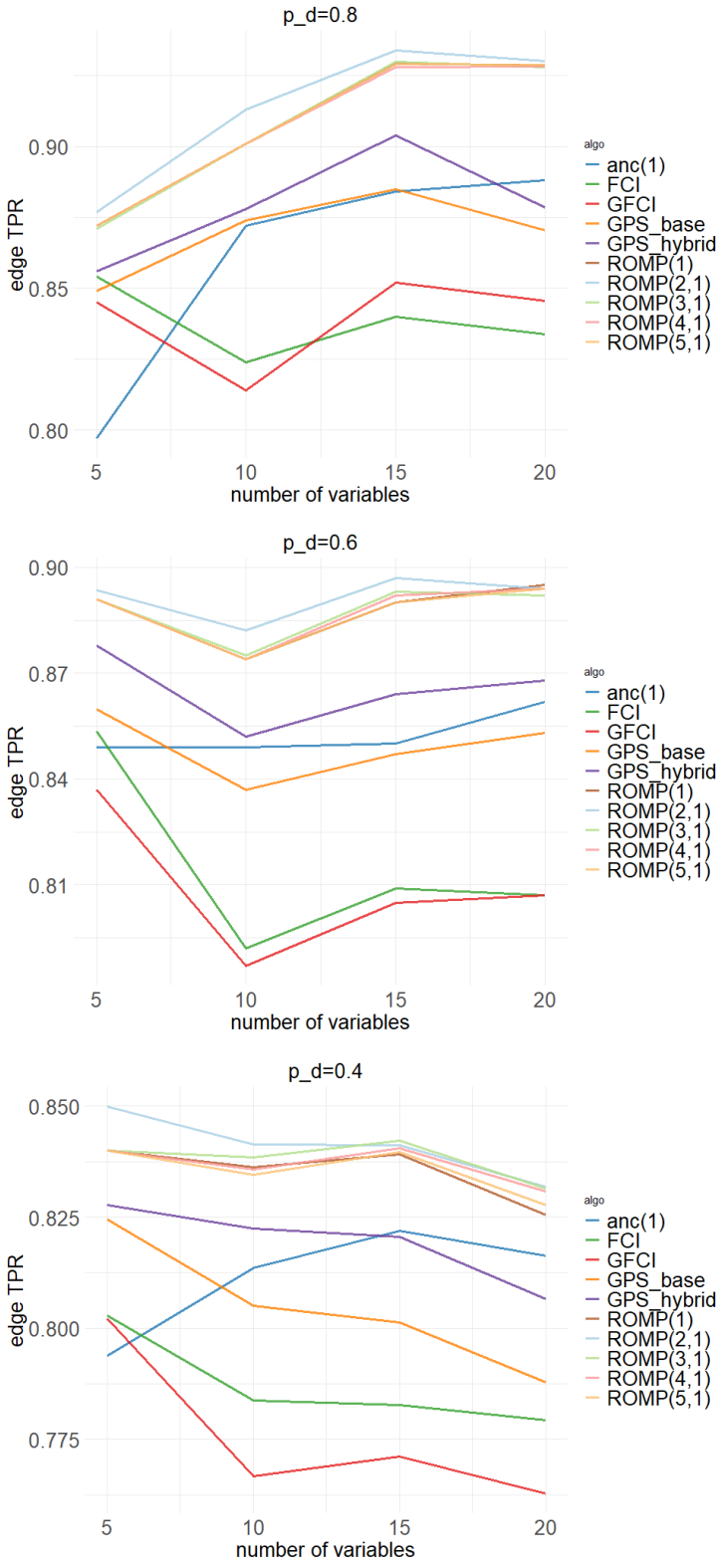


Figure 3: adjacency TPR plots

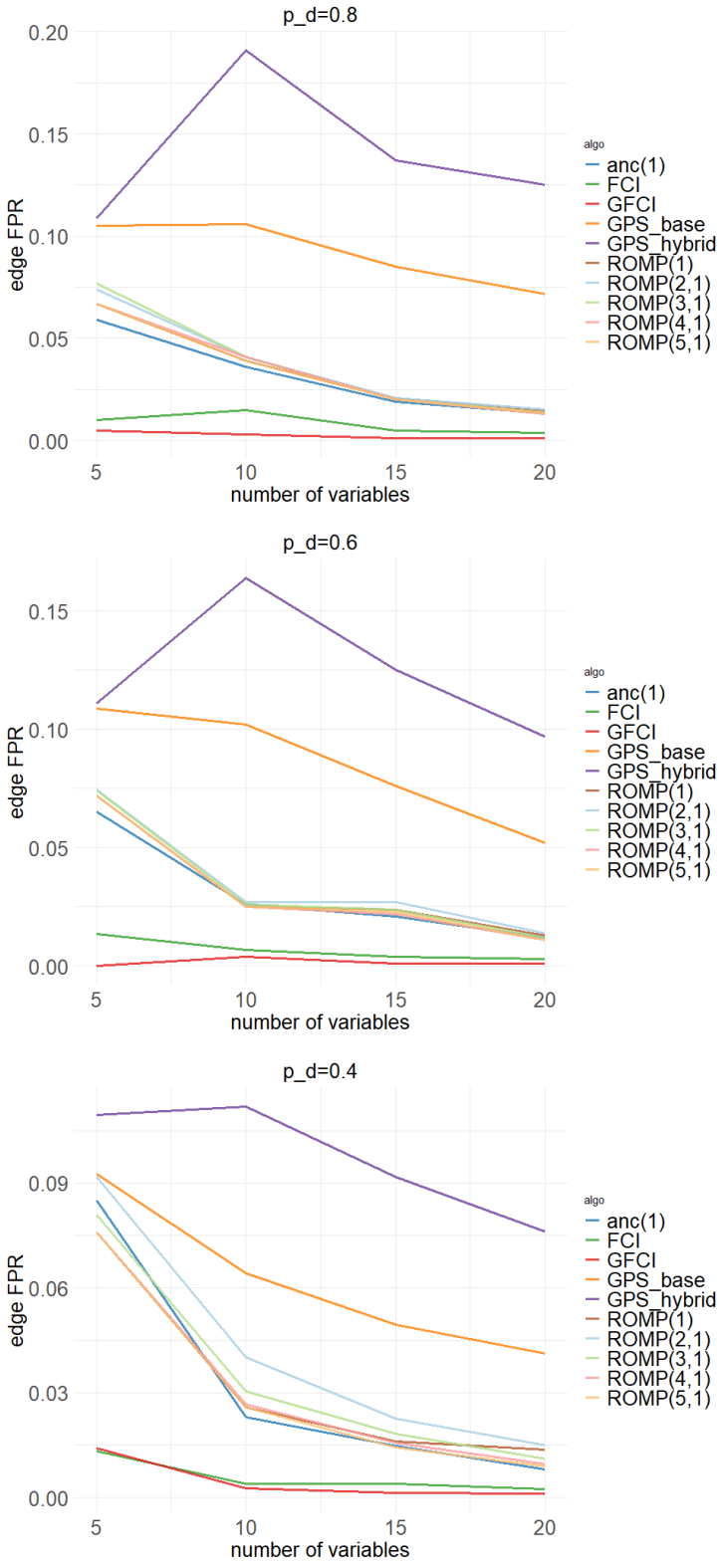


Figure 4: adjacency FPR plots

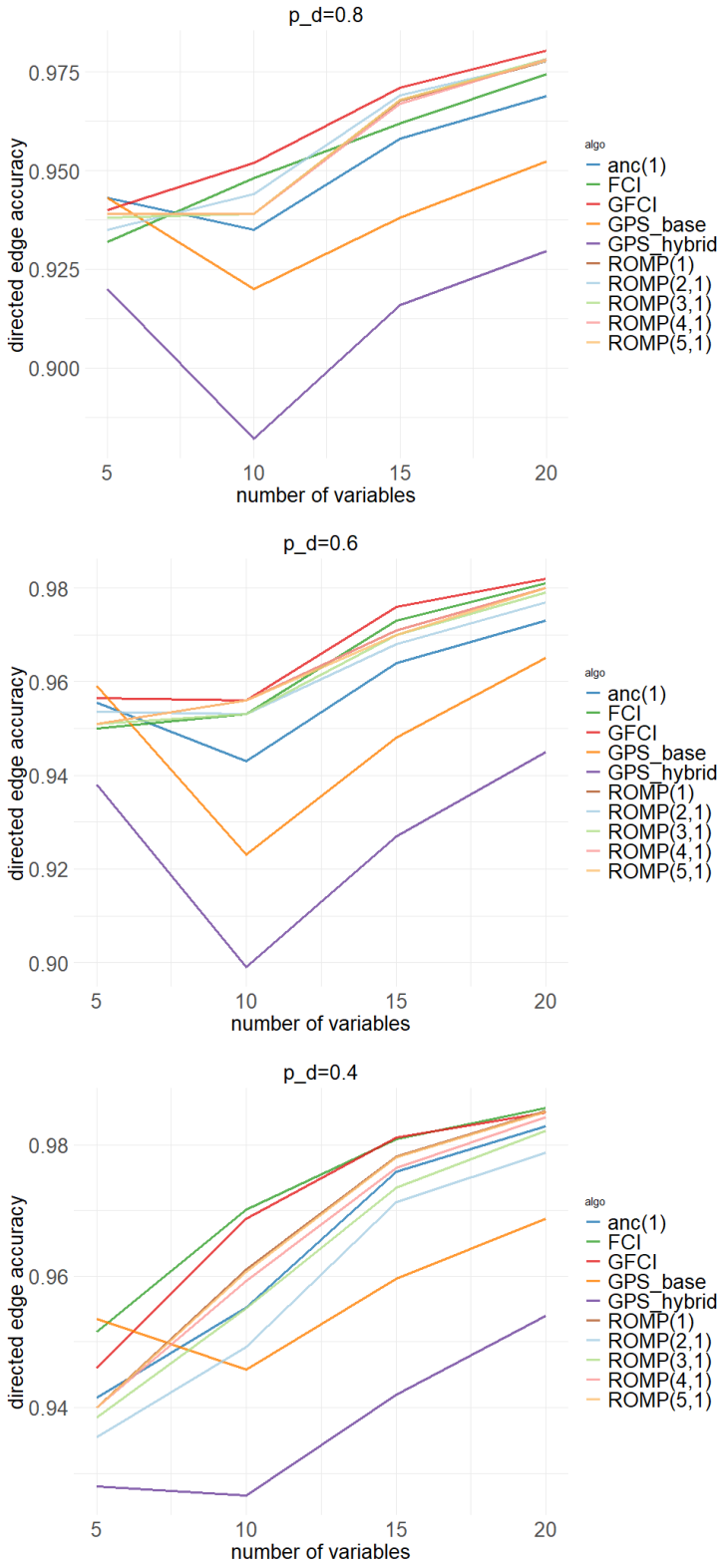


Figure 5: directed edge accuracy plots

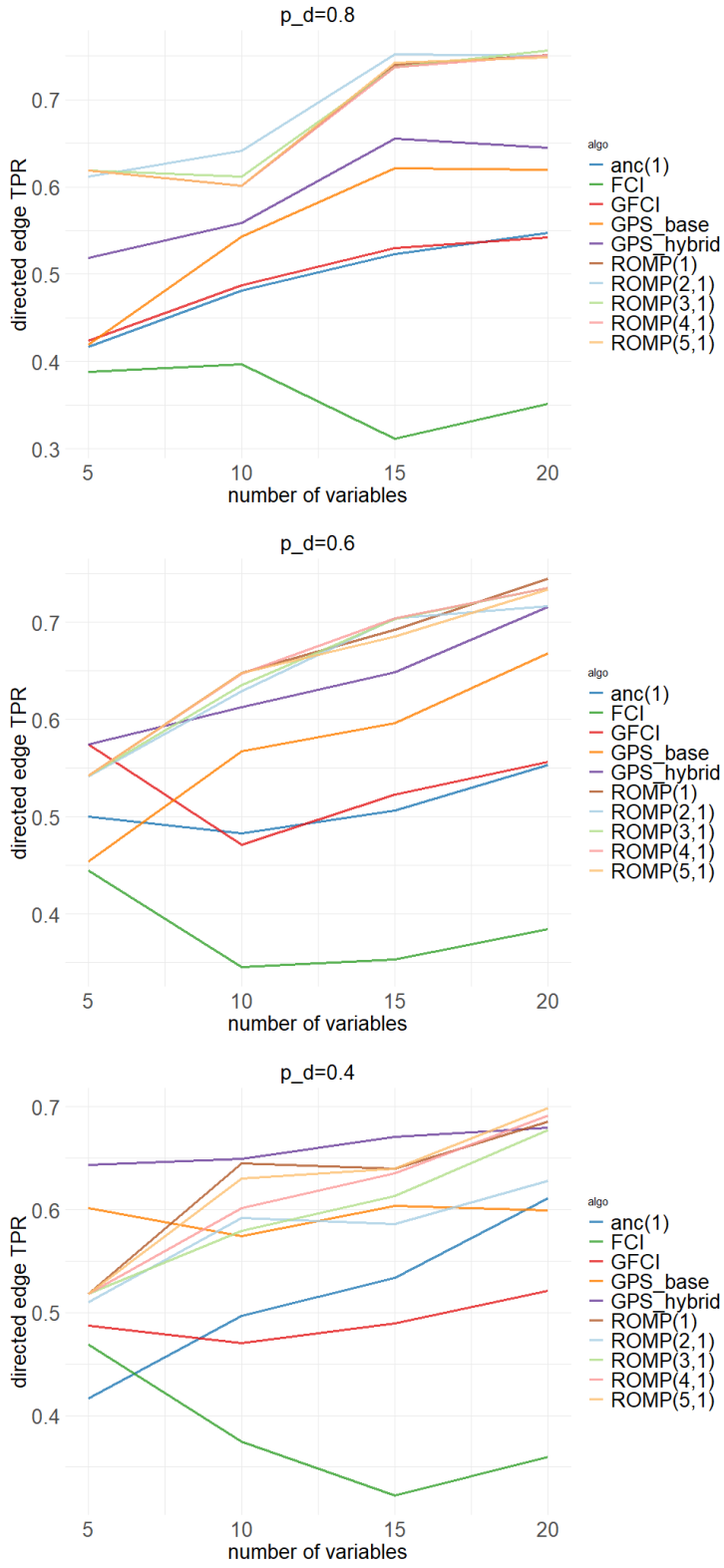


Figure 6: directed edge TPR plots

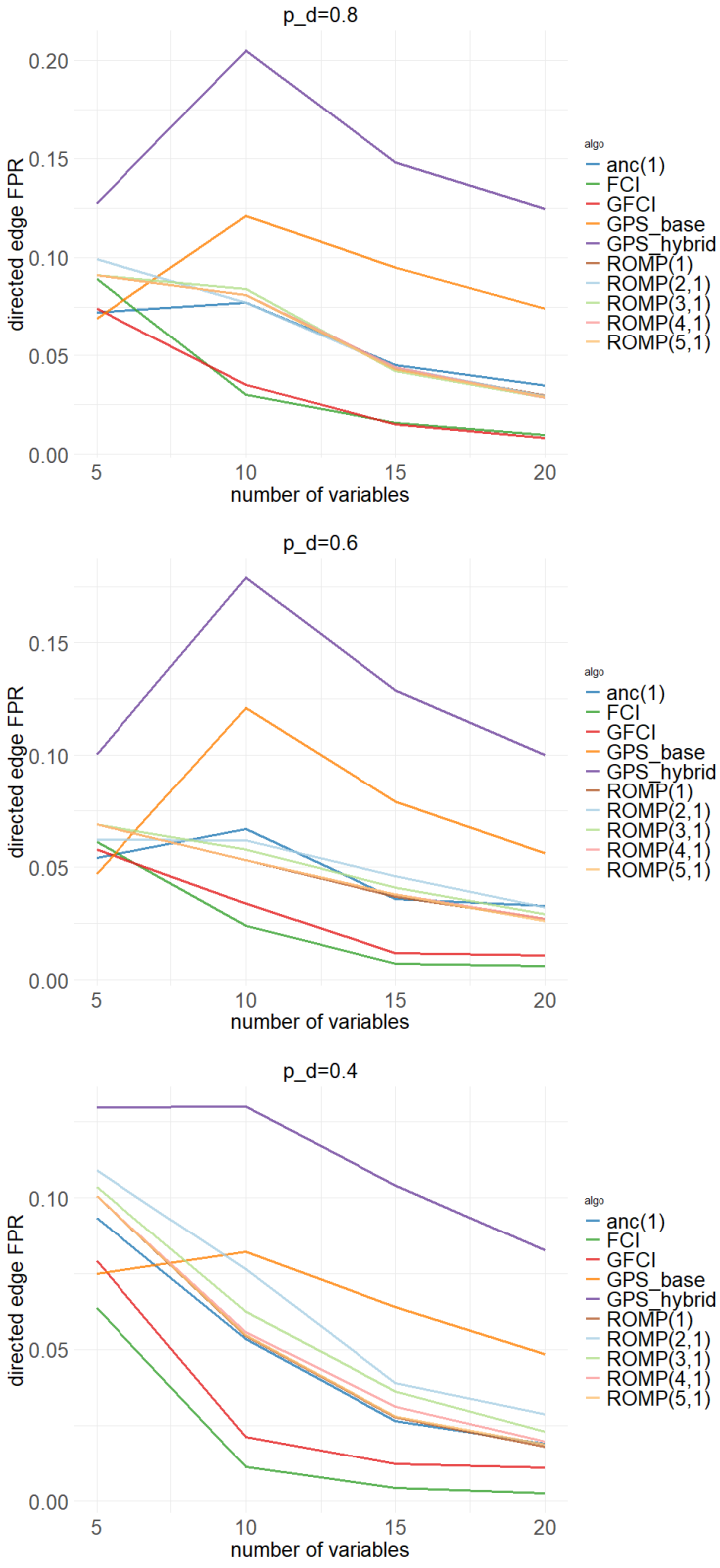


Figure 7: directed edge FPR plots

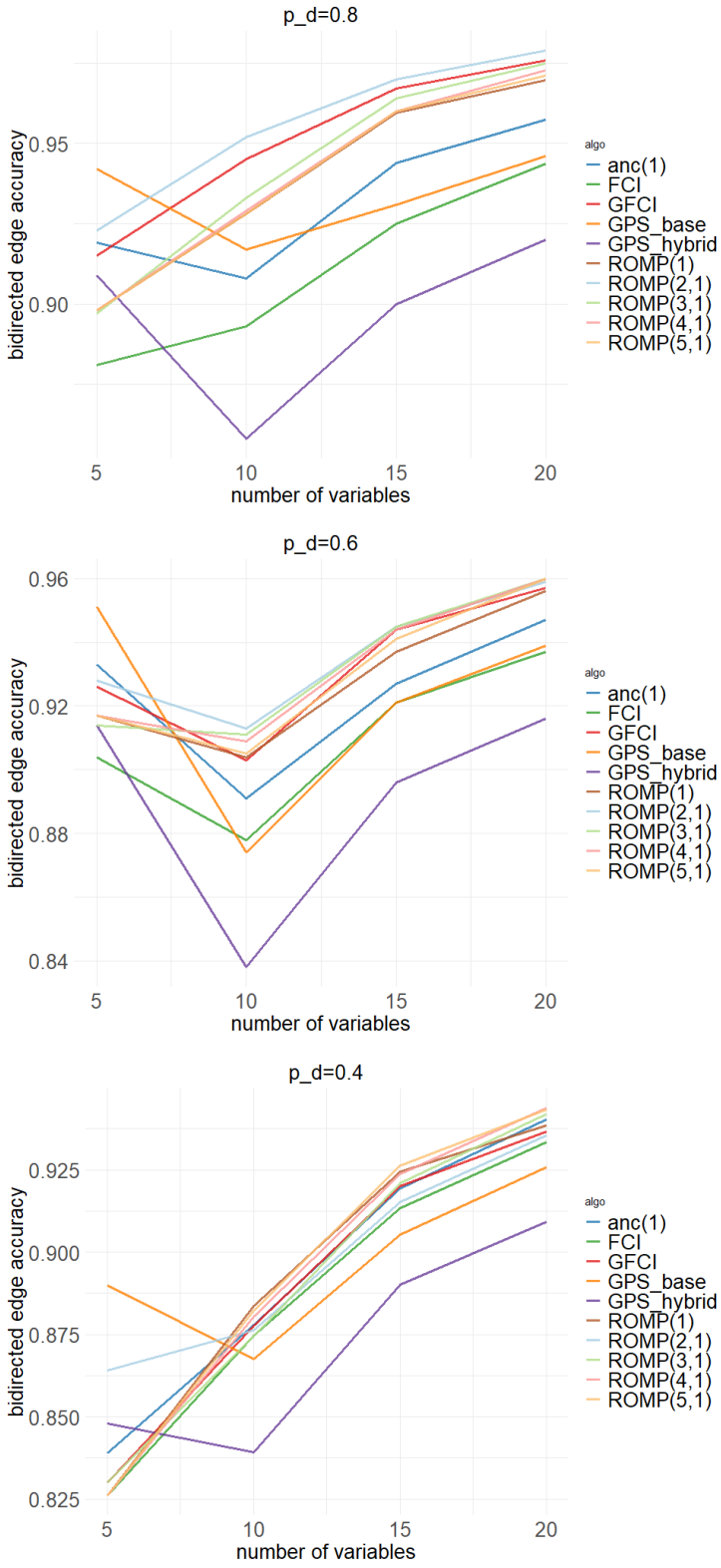


Figure 8: bidirected edge accuracy plots

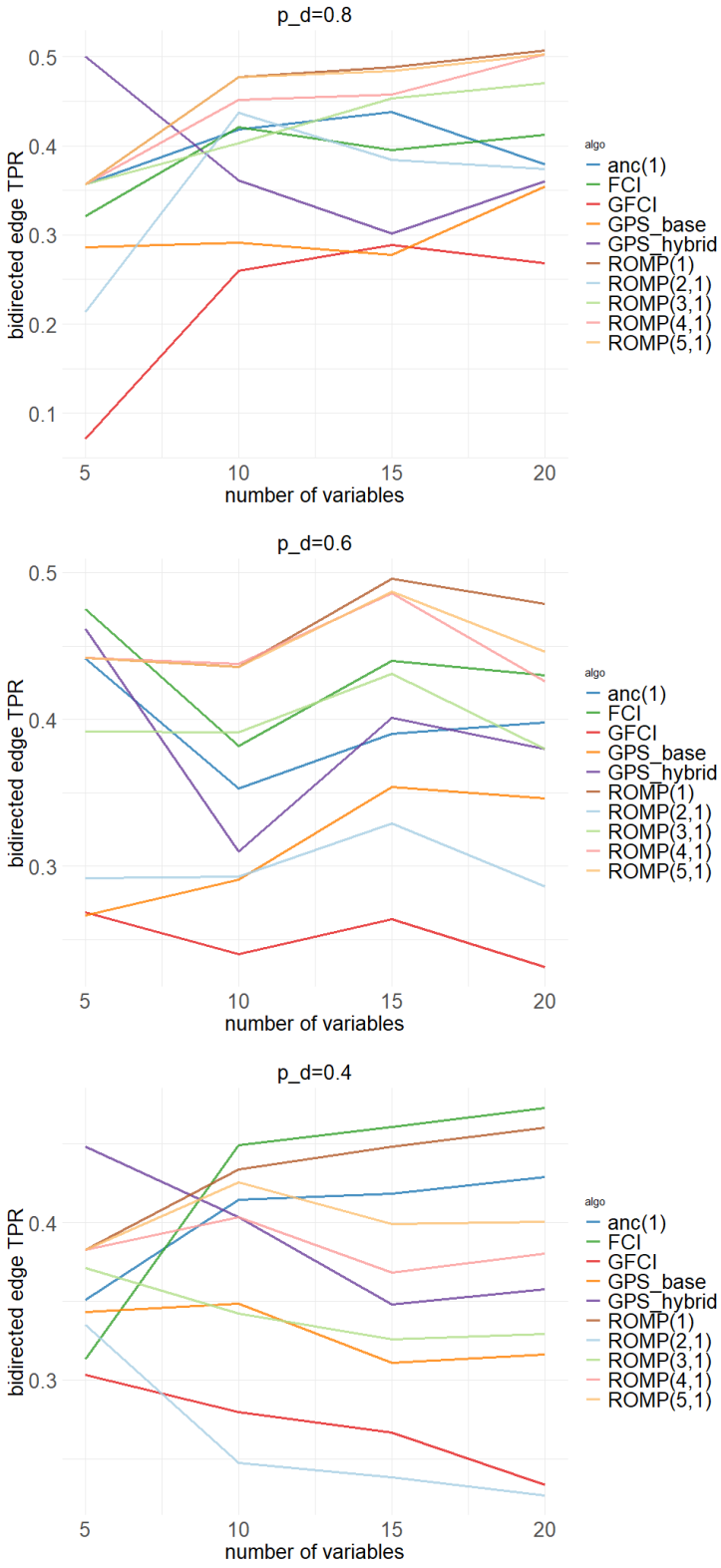


Figure 9: bidirected edge TPR plots

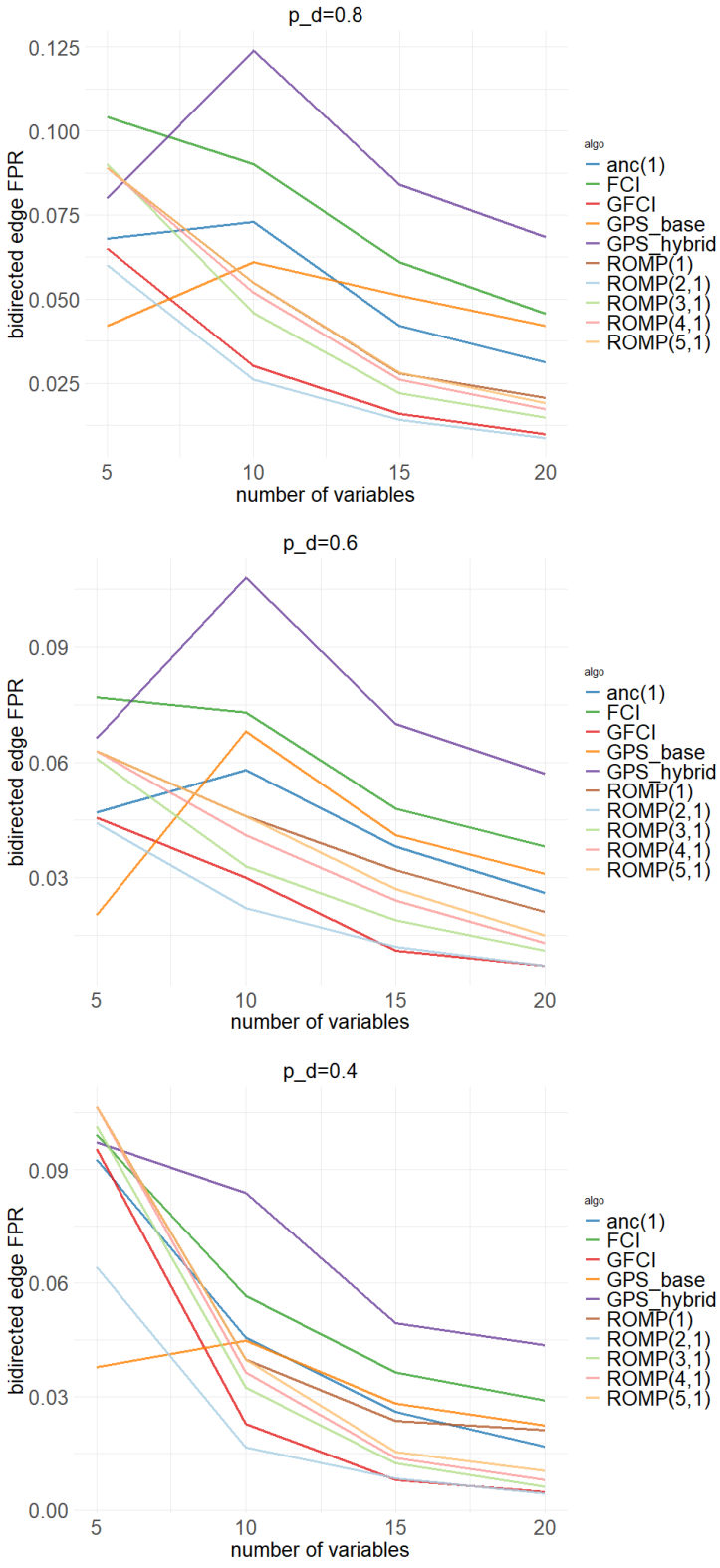


Figure 10: bidirected edge FPR plots

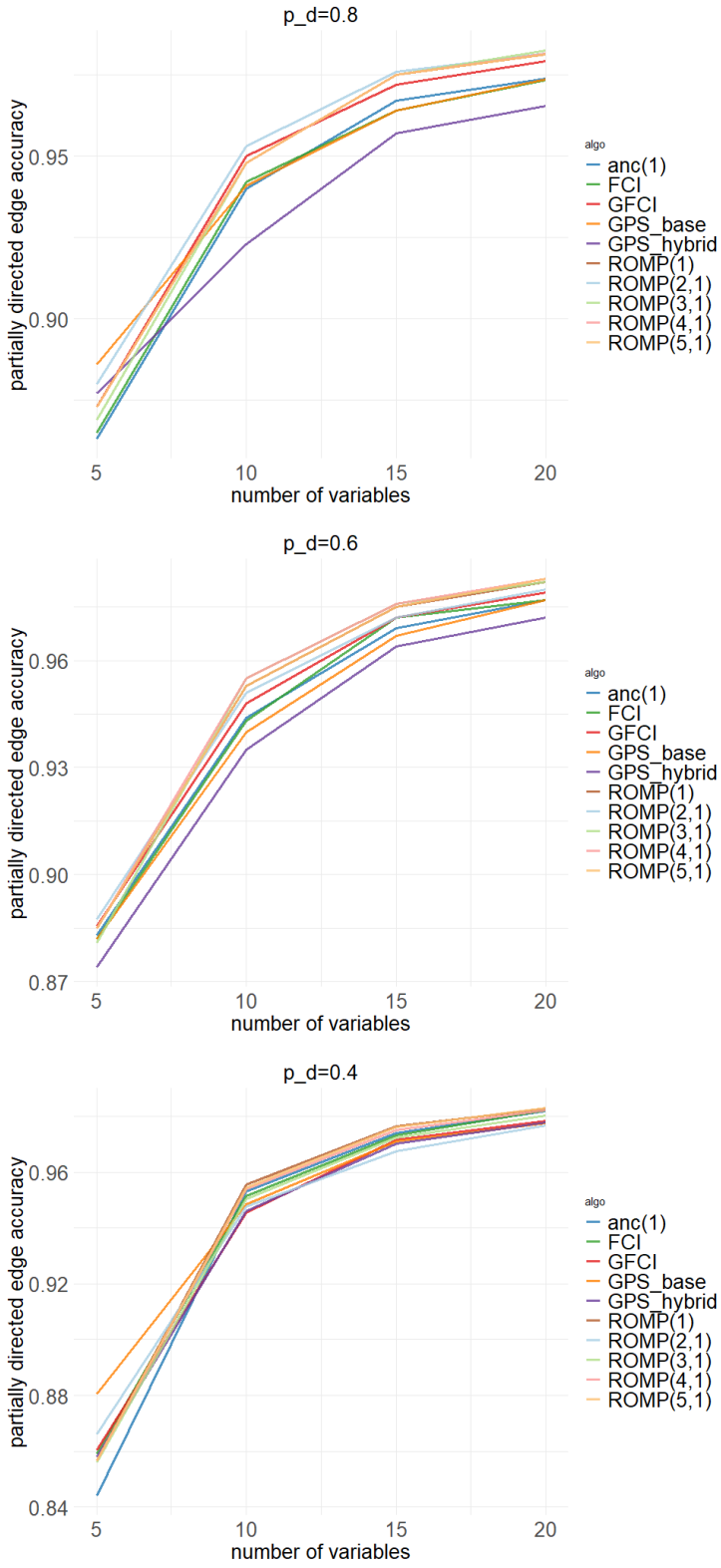


Figure 11: partially directed edge accuracy plots

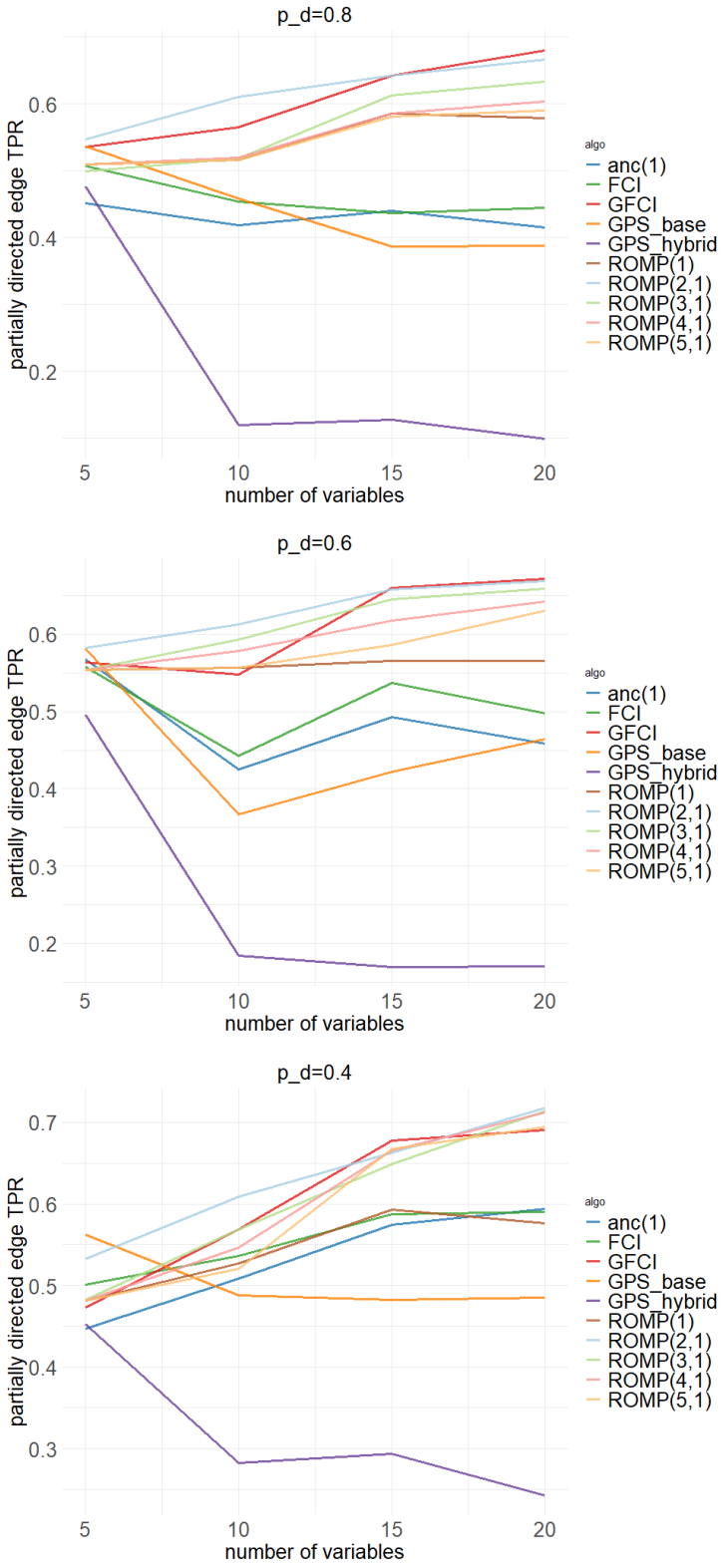


Figure 12: partially directed edge TPR plots

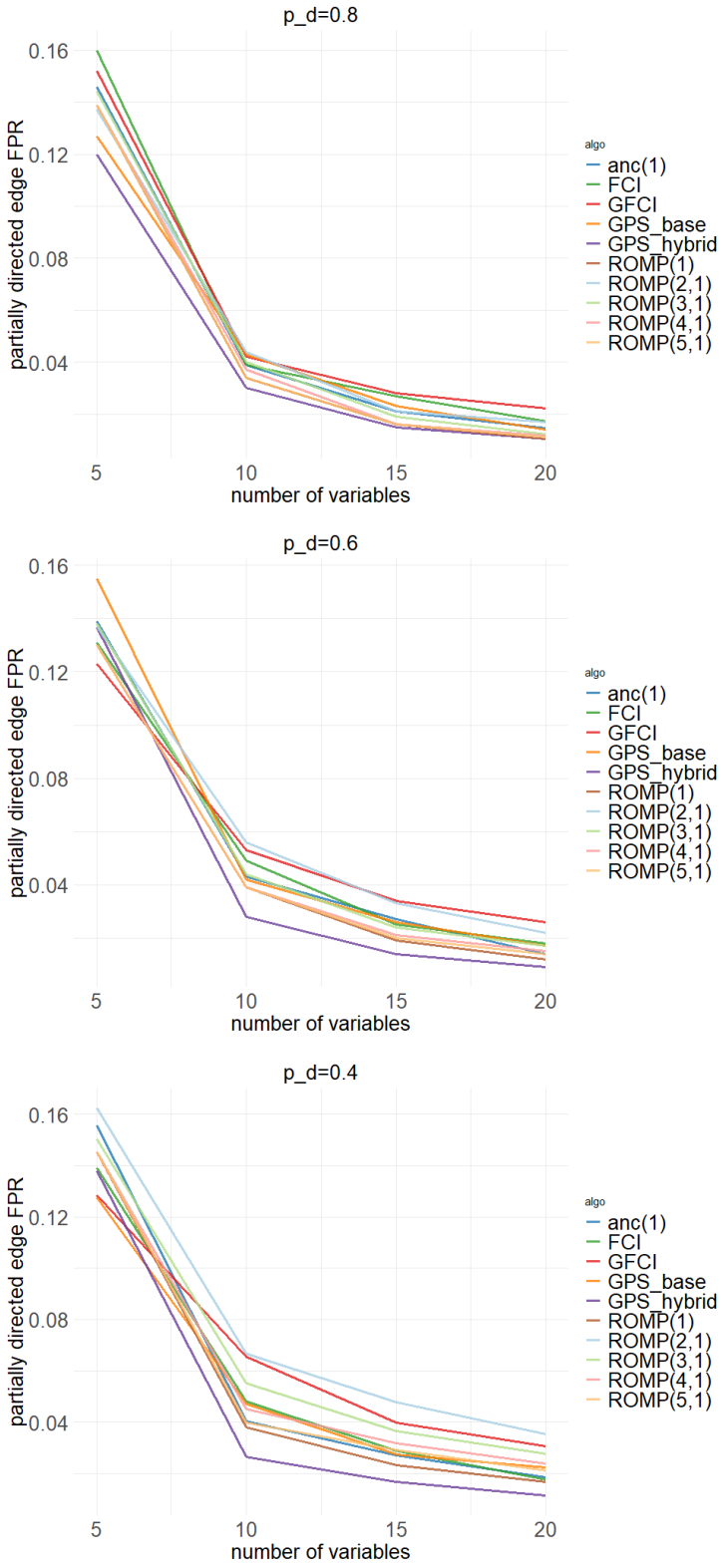


Figure 13: partially directed edge FPR plots

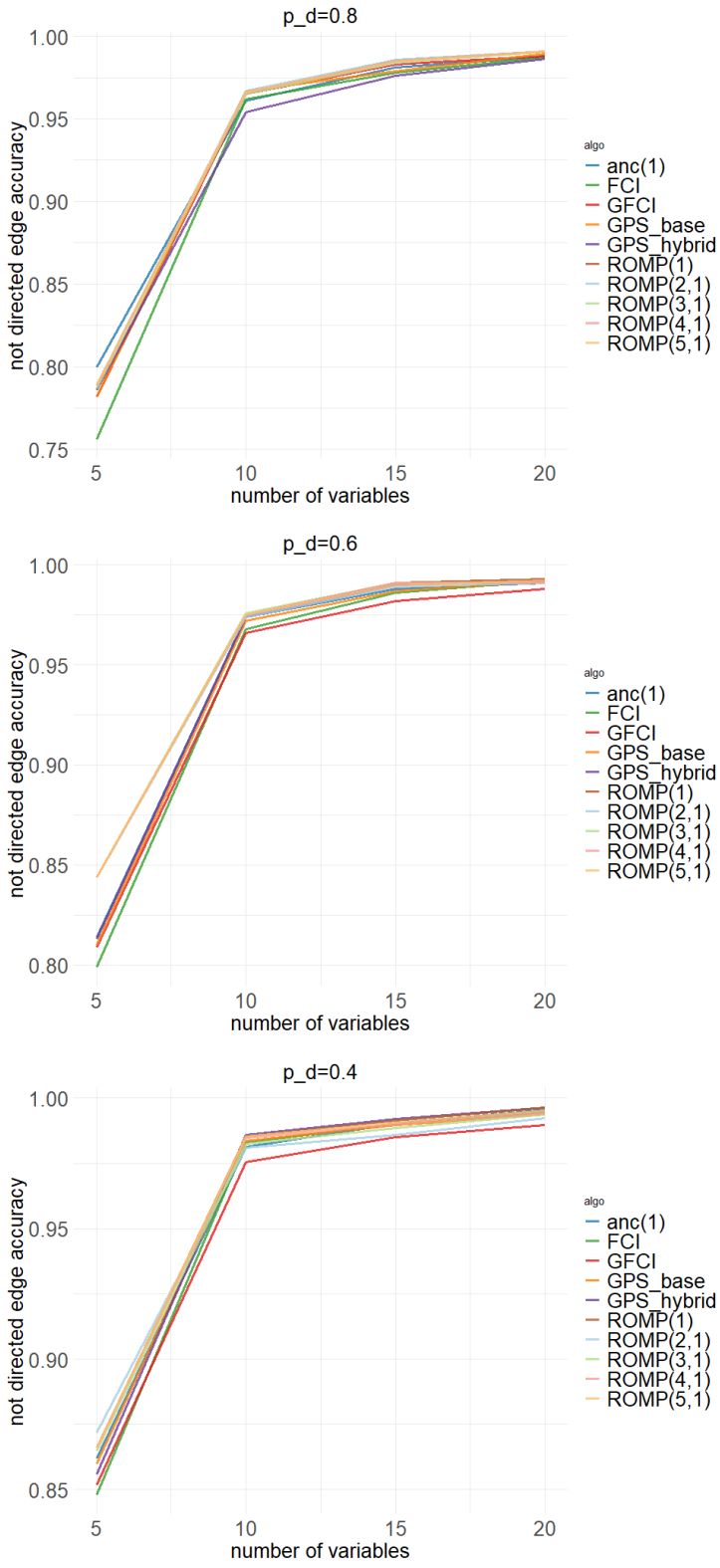


Figure 14: not directed edge accuracy plots

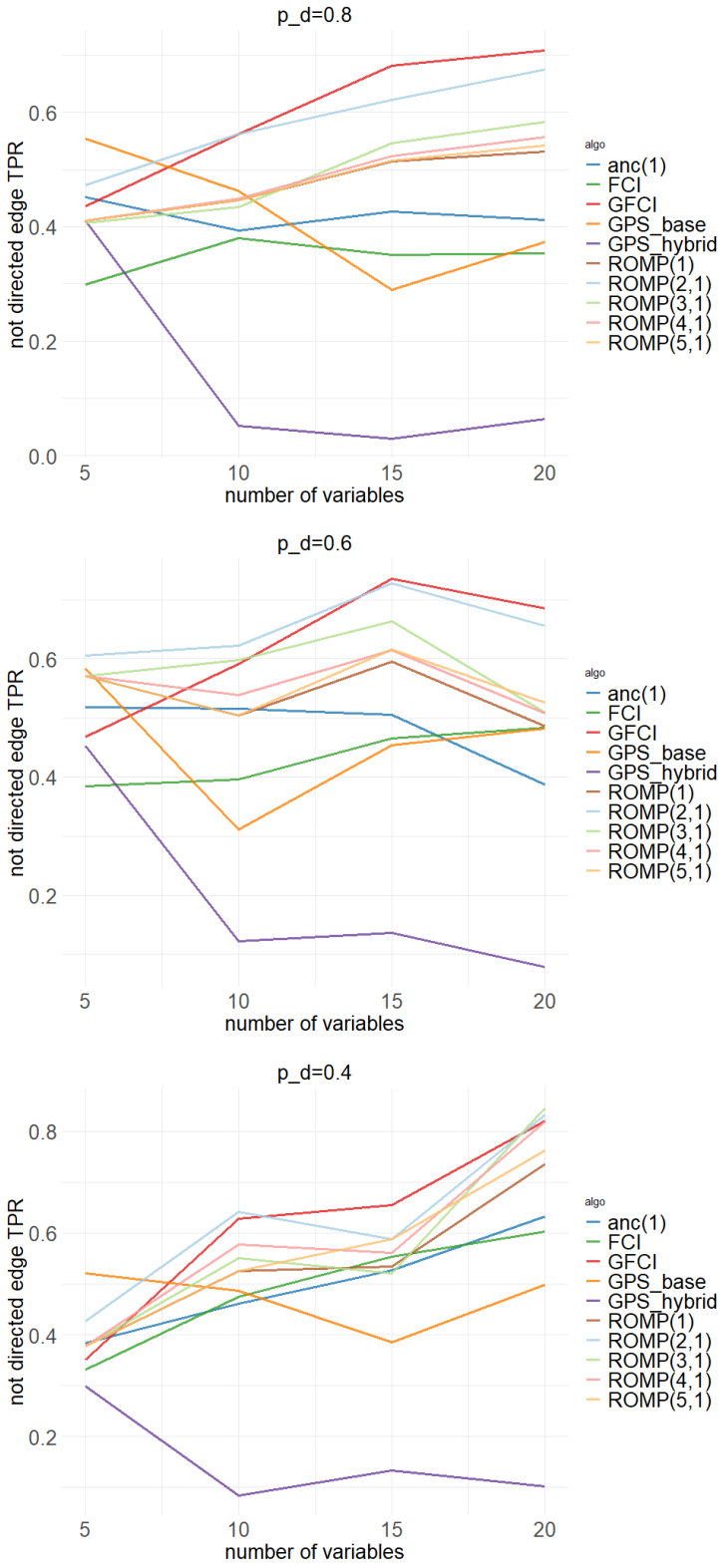


Figure 15: not directed edge TPR plots

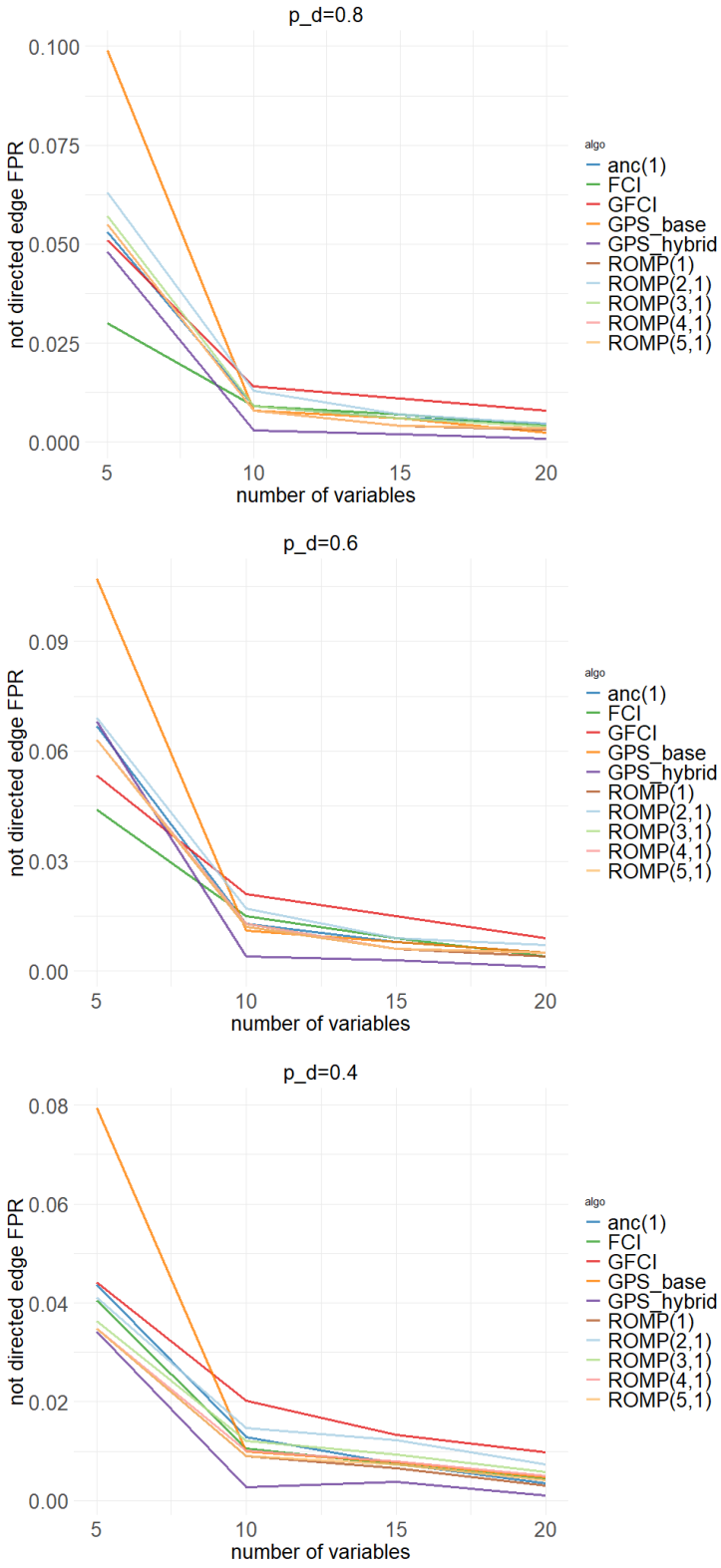


Figure 16: not directed edge FPR plots

Bibliography

Nabil Ali Ahmed and DV Gokhale. Entropy expressions and their estimators for multivariate distributions. *IEEE Transactions on Information Theory*, 35(3):688–692, 1989.

Ayesha R Ali, Thomas S Richardson, Peter L Spirtes, and Jiji Zhang. Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. *arXiv preprint arXiv:1207.1365*, 2005.

R. Ayesha Ali, Thomas S. Richardson, and Peter Spirtes. Markov equivalence for ancestral graphs. *Annals of Statistics*, 37(5B):2808–2837, 10 2009.

Bryan J. Andrews, Gregory F. Cooper, Thomas S. Richardson, and Peter Spirtes. The m-connecting imset and factorization for ADMG models. *arXiv preprint:2207.08963*, 2022.

Bryan James Andrews. *Inducing Sets: A New Perspective for Ancestral Graph Markov Models*. PhD thesis, University of Pittburgh, 2021. URL `http://d-scholarship.pitt.edu/42158/`.

Georgij P Basharin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability & Its Applications*, 4(3): 333–336, 1959.

Rui Chen, Sanjeeb Dash, and Tian Gao. Integer programming for causal structure learning in the presence of latent variables. In *International Conference on Machine Learning*, pages 1550–1560. PMLR, 2021.

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

Tom Claassen and Ioan Gabriel Bucur. Greedy equivalence search in the presence of latent confounders. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI-2022)*. PMLR, 2022.

Tom Claassen, Joris Mooij, and Tom Heskes. Learning sparse causal models is not np-hard. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI-2013)*. PMLR, 2013.

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

James Cussens. Gobnilp: Learning bayesian network structure with integer programming. In *International Conference on Probabilistic Graphical Models*, pages 605–608. PMLR, 2020.

Mathias Drton and Thomas S Richardson. Binary models for marginal independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):287–309, 2008.

Mathias Drton, Michael Eichler, and Thomas S Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(10), 2009.

Robin J Evans. Model selection and local geometry. *Annals of Statistics*, 48(6): 3513–3544, 2020.

Robin J Evans and Thomas S Richardson. Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI-2010)*. PMLR, 2010.

Robin J. Evans and Thomas S. Richardson. Marginal log-linear parameters for graphical Markov models. *Journal of the Royal Statistical Society, Series B*, 75 (4):743–768, Sep 2013.

Robin J. Evans and Thomas S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, 42(4):1452–1482, 2014.

Morten Frydenberg. The chain graph Markov property. *Scandinavian Journal of Statistics*, pages 333–353, 1990.

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

Raymond Hemmecke, Jason Morton, Anne Shiu, Bernd Sturmfels, and Oliver Wienand. Three counter-examples on semi-graphoids. *Combinatorics, Probability and Computing*, 17(2):239–257, 2008.

Zhongyi Hu and Robin Evans. Faster algorithms for Markov equivalence. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI-2020)*. PMLR, 2020.

Zhongyi Hu and Robin Evans. Towards standard imsets for maximal ancestral graphs. *arXiv preprint arXiv:2208.10436*, 2022.

Tommi Jaakkola, David Sontag, Amir Globerson, and Marina Meila. Learning bayesian network structure using lp relaxations. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 358–365. JMLR Workshop and Conference Proceedings, 2010.

David Kaltenpoth and Jilles Vreeken. Causal discovery with hidden confounders using the algorithmic markov condition. In *Uncertainty in Artificial Intelligence*, pages 1016–1026. PMLR, 2023.

Takuya Kashimura and Akimichi Takemura. Standard imsets for undirected and chain graphical models. *Bernoulli*, pages 1467–1493, 2015.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Svante Linusson, Petter Restadh, and Liam Solus. Greedy causal discovery is geometric. *SIAM Journal on Discrete Mathematics*, 37(1):233–252, 2023.

Christopher Meek. *Graphical Models: Selecting causal and statistical models*. PhD thesis, PhD thesis, Carnegie Mellon University, 1997.

Neeraj Misra, Harshinder Singh, and Eugene Demchuk. Estimation of the entropy of a multivariate normal distribution. *Journal of multivariate analysis*, 92(2): 324–342, 2005.

Christopher Nowzohour, Marloes H Maathuis, Robin J Evans, and Peter Bühlmann. Distributional equivalence and structure learning for bow-free acyclic path diagrams. 2017.

Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Proc. 8th Int. Conf. Probabilistic Graph. Models*, pages 368–379. PMLR, 2016.

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

Joe Ramsey, Peter Spirtes, and Jiji Zhang. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI-2006)*. PMLR, 2006.

Kari Rantanen, Antti Hyttinen, and Matti Järvisalo. Maximal ancestral graph structure learning via exact search. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI-2021)*. PMLR, 2021.

Thomas S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

Thomas S Richardson. A factorization criterion for acyclic directed mixed graphs. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI-09)*, 2009.

Thomas S. Richardson and Peter Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 08 2002.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs, 2017.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. *Ann. Statist.*, 51(1):334–361, 2023.

Kayvan Sadeghi. Faithfulness of probability distributions and graphs. *J. Mach. Learn. Res.*, 18(148):1–29, 2017.

Kayvan Sadeghi, Steffen Lauritzen, et al. Markov properties for mixed graphs. *Bernoulli*, 20(2):676–696, 2014.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

Ilya Shpitser, Robin J. Evans, and Thomas S. Richardson. Acyclic linear SEMs obey the nested Markov property. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI-2018)*. PMLR, 2018.

Peter Spirtes and Thomas S. Richardson. A polynomial time algorithm for determining DAG equivalence in the presence of latent variables and selection bias, 1997.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT Press, 2000.

Milan Studeny. Conditional independence relations have no finite complete characterization. *Inf. Theory Statist. Decis. Funct. Random Process. Trans. 11th Prague Conf.*, 1992.

Milan Studený. *Probabilistic conditional independence structures*. Springer Science & Business Media, 2006.

Milan Studený, Raymond Hemmecke, and Silvia Lindner. Characteristic imset: a simple algebraic representative of a bayesian network structure. In *Proceedings of the 5th European workshop on probabilistic graphical models*, pages 257–264. HIIT Publications, 2010.

Sofia Triantafillou and Ioannis Tsamardinos. Score-based vs constraint-based causal learning in the presence of confounders. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI-2016)*. PMLR, 2016.

Nanny Wermuth. Probability distributions with summary graph structure. *Bernoulli*, 17(3):845–879, 08 2011.

Marcel Wienöbst, Max Bannach, and Maciej Liskiewicz. A new constructive criterion for markov equivalence of mags. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-22)*, 2022.

Jiji Zhang. A characterization of Markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 450–457, 2007a.

Jiji Zhang. A characterization of Markov equivalence classes for directed acyclic graphs with latent variables. *arXiv preprint arXiv:1206.5282*, 2007b.

Jiji Zhang and Peter L Spirtes. A transformational characterization of Markov equivalence for directed acyclic graphs with latent variables. *arXiv preprint arXiv:1207.1419*, 2005.

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI-2009)*. PMLR, 2009.

Hui Zhao, Zhongguo Zheng, and Baijun Liu. On the Markov equivalence of maximal ancestral graphs. *Science in China Series A: Mathematics*, 48(4):548–562, Apr 2005.