Biometrika, *pp*. 1–26

Printed in Great Britain

Identifiability assumptions for directed graphical models with feedback

BY GUNWOONG PARK
Department of Statistics, University of Wisconsin-Madison, U.S.A.
parkg@stat.wisc.edu

AND GARVESH RASKUTTI
Department of Statistics, University of Wisconsin-Madison, U.S.A.
Department of Computer Science, University of Wisconsin-Madison, U.S.A.
Wisconsin Institute for Discovery, Optimization Group, U.S.A.
raskutti@stat.wisc.edu

SUMMARY

Directed graphical models provide a useful framework for modeling causal or directional relationships for multivariate data. Prior work has largely focused on identifiability and search algorithms for directed acyclic graphical (DAG) models. In many applications, feedback naturally arises and directed graphical models that permit cycles arise. However theory and methodology for directed graphical models with feedback are considerably less developed since graphs with cycles pose a number of additional challenges. In this paper we address the issue of identifiability for general directed cyclic graphical (DCG) models satisfying only the Markov assumption. In particular, in addition to the faithfulness assumption which has already been introduced for cyclic models, we introduce two new identifiability assumptions, one based on selecting the model with the fewest edges and the other based on selecting the directed cyclic graphical model that entails the maximum d-separation rules. We provide theoretical results comparing these assumptions which shows that: (1) selecting models with the largest number of d-separation rules is strictly weaker than the faithfulness assumption; (2) unlike for directed acyclic graphical modelss, selecting models with the fewest edges do not necessarily result in a milder assumption than the faithfulness assumption. We also provide connections between our two new principles and minimality assumptions which lead to a ranking of how strong and weak various identifiability and minimality assumptions are for both directed acyclic graph and directed cyclic graphcial models. We use our identifiability assumptions to develop search algorithms for small-scale directed cyclic graphcial models. Our simulations results using our search algorithms support our theoretical results, showing that our two new principles generally out-perform the faithfulness assumption in terms of selecting the true skeleton for directed cyclic graphcial models.

Some key words: Directed graphical Models, Identifiability, Faithfulness, Feedback loops.

1. INTRODUCTION

A fundamental goal in many scientific problems is to determine causal or directional relationships between variables in a system. A well-known framework for representing causal or directional relationships are directed graphical models. Most prior work on directed graphical models

has focused on directed acyclic graphical (DAG) models. directed acyclic graphical modelss, also referred to as Bayesian networks are directed graphical models with no directed cycles. One of the core problems in directed acyclic graphs is determining the underlying directed acyclic graph G given the data-generating distribution $\mathbb{P}$.

The directed acyclic graph framework is based on the Markov assumption, which relates conditional independence (CI) statements for a probability distribution $\mathbb{P}$ to so-called *d-separation* rules entailed by a directed graph G (cyclic or acyclic) (see e.g. (Lauritzen, 1996; Spirtes et al., 2000)). While the Markov assumption is fundamental, in order for the directed graph G to be identifiable based on the distribution $\mathbb{P}$, additional identifiability or minimality assumptions are required. For directed acyclic graphical modelss, a number of identifiability and minimality assumptions have been introduced (Glymour et al., 1987; Spirtes et al., 2000) and the connections between them have been discussed (Zhang, 2012). In particular, one of the most widely used assumptions for directed acyclic graphical modelss is the causal faithfulness condition (CFC) which is sufficient for many search algorithms. However the casual faithfulness condition has been shown to be extremely restrictive, especially in the limited data setting (Uhler et al., 2013). In addition two minimality assumptions, the P-minimality and SGS-minimality assumptions have been introduced. These conditions are weaker than the casual faithfulness condition but do not guarantee model identifiability (Zhang, 2012). On the other hand, the recently introduced sparsest Markov representation (SMR) and frugality assumptions introduced in (Forster et al., 2015; Raskutti & Uhler, 2013) provide an alternative that is milder than the casual faithfulness condition and is sufficient to ensure identifiability. The main downside of the frugality and sparsest Markov representation assumptions relative to the casual faithfulness condition is that the frugality and sparsest Markov representation assumptions are sufficient conditions for model identifiability only when exhaustive searches over the directed acyclic graph space are possible (Raskutti & Uhler, 2013), while the casual faithfulness condition is sufficient for polynomial-time algorithms (Glymour et al., 1987; Spirtes & Glymour, 1991).

While the directed acyclic graph framework is useful in many applications in psychology, economics, biology, and other disciplines, the directed acyclic graph framework is limited since feedback loops are known to often exist (see e.g. (Richardson, 1996a,b)). Hence *directed cyclic graphs* (DCGs) (Spirtes et al., 2000) are more appropriate to model such feedback. However learning directed cyclic graphs from data is considerably more challenging than learning directed acyclic graphs (Richardson, 1996a,b) since the presence of cycles poses a number of additional challenges and introduces additional non-identifiability. Consequently there has been considerable less work focusing on directed graphs with feedback both in terms of identifiability assumptions and search algorithms. Spirtes et al. (Spirtes, 1995) discuss the Markov conditions, and Richardson (Richardson, 1996a,b) discusses the causal faithfulness condition (CFC) for directed cyclic graphcial models and introduce the polynomial-time cyclic causal discovery (CCD) algorithm Richardson (1996a). As with directed acyclic graphical modelss, the faithfulness assumption for cyclic models is extremely restrictive (since it is by definition more restrictive than the casual faithfulness condition for directed acyclic graphical modelss). This raises the question of whether a milder identifiability assumption can be imposed for learning directed graphical models with feedback.

In this paper, we address this question in a number of steps. Firstly we adapt the sparsest Markov representation and frugality assumptions developed for directed acyclic graphical modelss to directed cyclic graphcial models. Next we show that unlike for directed acyclic graphical modelss, the adapted sparsest Markov representation and frugality assumptions are not strictly weaker than the faithfulness assumption. Hence we consider a new identifiability assumption based on finding the Markovian directed cyclic graph satisfying the *maximum d-separation rules*

(MDR) which we prove is strictly weaker than the faithfulness assumption and infers the true Markov equivalence class for directed cyclic graphcial models more consistently than the casual faithfulness condition. We also provide a ranking in terms of strength between the maximum d-separation rule assumption, the sparsest Markov representation and frugality assumptions as well as the minimality assumptions for both directed acyclic graph and directed cyclic graphcial models. Finally we use the sparsest Markov representation and maximum d-separation rule identifiability assumptions to develop search algorithms for small-scale directed cyclic graphcial models. Our simulation study supports our theoretical results by showing that the algorithms induced by both the sparsest Markov representation and maximum d-separation rule assumptions have higher recovery rates than the algorithm induced by the faithfulness assumption. We point out that the search algorithms that result from our identifiability assumptions require exhaustive searches and are not computationally feasible for large directed cyclic graphcial models. However the focus of this paper is to develop the weakest possible identifiability assumption.

The remainder of the paper is organized as follows: Section 2 provides the background and prior work for identifiability assumptions for both directed acyclic graph and directed cyclic graphcial models. In Section 3 we adapt the sparsest Markov representation and frugality assumptions to directed cyclic graphcial models and provide a comparison between the sparsest Markov representation assumption, the faithfulness assumption, and the minimality assumptions. In Section 4 we introduce our new maximum d-separation rule principle, finding the Markovian directed cyclic graph that satisfies the maximum number of d-separation rules and provide a comparison of our new principle to the faithfulness, minimality, sparsest Markov representation and frugality assumptions. Finally in Section 5, we use our identifiability assumptions to develop a search algorithm for learning small-scale directed cyclic graphcial models, and show that our simulations support our theoretical results.

2. PRIOR WORK ON DIRECTED GRAPHICAL MODELS

In this section, we introduce the basic concepts of directed acyclic graphs and directed cyclic graphs pertaining to model identifiability. A directed graph $G = (V, E)$ consists of a set of vertices V and a set of directed edges E. Suppose that $V = \{1, 2, \ldots, p\}$ and there exists a random vector $(X_1, X_2, ..., X_p)$ with probability distribution $\mathbb{P}$ over the vertices in G. A directed edge from vertex j to k is denoted by (j, k) or $j \to k$. The set $\text{pa}(k)$ of *parents* of a vertex k consists of all nodes j such that $(j, k) \in E$. If there is a directed path $j \to \cdots \to k$, then k is called a *descendant* of j and j is an *ancestor* of k. The set $\text{de}(k)$ denotes the set of all descendants of a node k. The *non-descendants* of a node k are $\text{nd}(k) = V \setminus (\{k\} \cup \text{de}(k))$. For a subset $S \subset V$, we define $\text{an}(S)$ to be the set of nodes k that are in S or are ancestors of some node in S. Two nodes that are connected by an edge are called *adjacent*. A triple of nodes (j, k, ℓ) is an *unshielded triple* if j and k are adjacent to ℓ but j and k are not adjacent. An unshielded triple (j, k, ℓ) forms a *v-structure* if $j \to \ell$ and $k \to \ell$. In this case ℓ is called a *collider*. Furthermore, an undirected path π from j to k *d-connects* j and k given $S \subset V \setminus \{j, k\}$ if every collider on π is in $\text{an}(S)$ and every non-collider on π is not in S. If G has no path that d-connects j and k given a subset S, then j and k are *d-separated* given S. Finally, let $X_j \perp\!\!\!\perp X_k \mid X_S$ with $S \subset V \setminus \{j, k\}$ denoting the conditional independence (CI) statement that X_j is conditionally independent (as determined by $\mathbb{P}$) of X_k given the set of variables $X_S = \{X_\ell \mid \ell \in S\}$, and let $X_j \not\perp\!\!\!\perp X_k \mid X_S$ denote conditional dependence. The *Causal Markov condition* associates conditional indenpendence relations of $\mathbb{P}$ with a directed graph G:

DEFINITION 1 (CAUSAL MARKOV CONDITION (CMC) (SPIRTES ET AL., 2000)). *A probability distribution $\mathbb{P}$ over a set of vertices V satisfies the* Causal Markov condition *with respect to a (acyclic or cyclic) graphical model $G = (V, E)$ if for all (j, k, S), j is d-separated from k conditioned on $S \subset \{1, 2, ..., p\} \setminus \{j, k\}$ in G, then*

$$X_j \perp\!\!\!\perp X_k \mid X_S \quad \textit{according to } \mathbb{P}.$$

The causal Markov condition applies to both acyclic and cyclic graphs (see e.g. Spirtes et al. (2000)). However not all directed graphical models satisfy the causal Markov condition. In order for a directed cyclic graphical model to satisfy the causal Markov condition, the joint distribution of a directed cyclic graphical model should be defined by the *generalized factorization* (Lauritzen et al., 1990).

DEFINITION 2 (GENERALIZED FACTORIZATION (LAURITZEN ET AL., 1990)). *The joint distribution of X, $f(X)$* factors according to directed graph *G with vertices V if and only if for every subset X of V,*

$$f(an(X)) = \prod_{V \in an(X)} g_V(V, pa(V))$$

where g_V is a non-negative function.

Spirtes et al. (Spirtes, 1995) showed that the generalized factorization is a necessary and sufficient condition for directed graphical models to satisfy the causal Markov condition. For directed acyclic graphical modelss, $g_V(.)$'s must always correspond to a probability distribution function whereas for graphs with cycles, $g_V(.)$'s need only be non-negative functions. As shown by Spirtes et al. (Spirtes, 1995), a concrete example of a cyclic graph that satisfies the factorization above is structural linear directed cyclic graph equation models with additive independent errors. We will later use linear directed cyclic graphcial models in our simulation study.

In general, there are many directed graphs entailing the same d-separation rules. These graphs are *Markov equivalent* and the set of Markov equivalent graphs is called a *Markov equivalence class* (MEC) (Spirtes et al., 2000). For example, consider two 2-node graphs, $G_1 : X_1 \rightarrow X_2$ and $G_2 : X_1 \leftarrow X_2$. Then both graphs are Markov equivalent because they both entail no d-separation rules. Hence, G_1 and G_2 are in the same Markov equivalence class and hence it is impossible to distinguish two graphs by d-separation rules. The precise definition of Markov equivalence class is provided here.

DEFINITION 3 (MARKOV EQUIVALENCE (RICHARDSON, 1996A)). *Directed graphs G_1 and G_2 are* Markov equivalent *if any distribution which satisfies the causal Markov condition with respect to one graph satisfies it with respect to the other, and vice versa. The set of graphs which are Markov Equivalent to G is denoted by $\mathcal{M}(G)$.*

The characterization of Markov equivalence classes is different for directed acyclic graphs and directed cyclic graphs. For directed acyclic graphs, Chickering Chickering et al. (1995) developed an elegant characterization of Markov equivalence classes defined by the *skeleton* and v-structures. The skeleton of a directed acyclic graphical model consists of the edges without directions:

THEOREM 1 (THEOREM IN CHICKERING CHICKERING ET AL. (1995)). *Two directed acyclic graphs G_1 and G_2 belong to the same Markov equivalence class if and only if they have the same skeleton and v-structures.*

However for directed cyclic graphs, the presence of feedback means the characterization of the Markov equivalence class for directed cyclic graphs is considerably more involved. Richardson provides a characterization in (Richardson, 1996a). The presence of directed cycles changes the notion of adjacency between pairs. In particular there are *real* adjacencies that are a result of directed edges in the directed cyclic graph and *virtual* adjacencies which are edges that do not exist in the data-generating directed cyclic graph but can not be recognized as a non-edge from the data. The precise definition of real and virtual edges are:

DEFINITION 4 (ADJACENCY (RICHARDSON, 1996B)). *Consider a directed graph* $G = (V, E)$.

(a) For any $j, k \in V$*,* j *and* k *are* really adjacent *in* G *if* $j \to k$ *or* $j \leftarrow k$.
(b) For any $j, k \in V$*,* j *and* k *are* virtually adjacent *if* j *and* k *have a common child* ℓ*, such that* ℓ *is an ancestor of* j *or* k.

Note that a virtual edge can only occur if there is a cycle in the graph. Hence, directed acyclic graphs have only real edges while directed cyclic graphs can have both real edges and virtual edges. Figure 1 show an example of a directed cyclic graph with a virtual edge. In Figure 1, a pair of nodes $(1, 4)$ has a virtual edge (dotted line) because the triple $(1, 4, 2)$ forms a v-structure and the common child 2 is an ancestor of 1. This virtual edge is created by the cycle, $1 \to 2 \to 3 \to 1$.

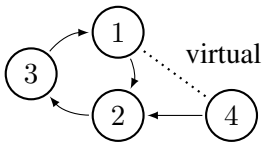


Fig. 1: 4-node example for a virtual edge

Virtual edges generate different types of relationships involving unshielded triples: (1) an unshielded triple of (j, k, ℓ) (that is $j - \ell - k$) is called a *conductor* if ℓ is an ancestor of j or k; (2) an unshielded triple of (j, k, ℓ) is called a *perfect non-conductor* if ℓ is a descendant of the common child of j and k. Furthermore (3) an unshielded triple of (j, k, ℓ) is called an *imperfect non-conductor* if the triple is not a conductor or a perfect non-conductor.

Intuitively, the concept of (1) a conductor is analogous to the notion of non-collider in directed acyclic graphs because for example suppose that (j, k, ℓ) is a conductor, then j is d-connected from k given any set S which does not contain ℓ. Moreover, (2) a perfect non-conductor is analogous to a v-structure because suppose that (j, k, ℓ) is a perfect non-conductor, then j is d-connected from k given any set S which contains ℓ. However, there is no analogous notion of an imperfect non-conductor for directed acyclic graphical modelss. Hence an imperfect non-conductor is another significant difference between directed acyclic graphs and directed cyclic graphs and as we see throughout this paper that this difference creates a major challenge in inferring directed cyclic graphcial models from the underlying distribution $\mathbb{P}$. As shown in Richardson et al. Richardson (1996b), a necessary (but not sufficient) condition for two directed cyclic graphs to belong to the same Markov equivalence class is that they share the same real plus virtual edges and the same conductors, perfect non-conductors and imperfect non-conductors. However unlike for directed acyclic graphs this condition is not sufficient for Markov equivalence since local properties do not completely characterize the global directed cyclic graph structure. A com-

plete characterization of Markov equivalence is provided in Richardson et al. Richardson (1996b) and since it is quite involved, we do not include it here.

Even if we weaken the goal to inferring a Markov equivalence class for a directed acyclic graph or directed cyclic graph, the causal Markov condition is insufficient for discovering the true Markov equivalence class $\mathcal{M}(G^*)$ because there are many graphs satisfying the causal Markov condition, which do not belong to $\mathcal{M}(G^*)$. For example, any fully-connected graph always satisfies the causal Markov condition because it does not entail any d-separation rules. Hence, in order to identify the true Markov equivalence class given the distribution $\mathbb{P}$ stronger identifiability assumptions which encourage the removal of edges are required.

2·1. *Faithfulness and minimality assumptions*

For a directed graph G, let $E(G)$ denote the set of directed edges, $S(G)$ denote the set of edges without directions, also referred to as the skeleton and $D_{sep}(G)$ denote the set of d-separation rules entailed by G. Further let $CI(\mathbb{P})$, denote the conditional independence (CI) statements corresponding to the multivariate distribution $\mathbb{P}$. In this section we discuss identifiability assumptions. To make the notion of identifiability and our assumptions precise, we need to define the notion of a true data-generating graph G^*. All we observe is the distribution $\mathbb{P}$ and we know the pair $(G^*, \mathbb{P})$ satisfies the causal Markov condition. The pair $(G^*, \mathbb{P})$ is *identifiable* if the graph Markov equivalence class for $\mathcal{M}(G^*)$, can be uniquely determined by conditional independence statements $CI(\mathbb{P})$.

One of the most widely imposed identifiability assumptions for both directed acyclic graphs and directed cyclic graphs is the *causal faithfulness condition* (CFC) (Spirtes et al., 2000) also referred to as the stability condition in (Pearl, 2014). A directed graph is *faithful* to a probability distribution if there is no probabilistic independence in the distribution that is not entailed by the causal Markov condition. The casual faithfulness condition states that the graphical model is faithful to the true probability distribution.

DEFINITION 5 (CAUSAL FAITHFULNESS CONDITION (CFC) (SPIRTES ET AL., 2000)). *Consider a directed graphical model $(G^*, \mathbb{P})$. A directed graph G^* is* faithful *to $\mathbb{P}$ if for any $j, k \in V$ and any subset $S \subset V \setminus \{j, k\}$,*

$$j \text{ d-separated from } k \mid S \qquad \Longleftrightarrow \qquad X_j \perp\!\!\!\perp X_k \mid X_S \text{ according to } \mathbb{P}.$$

While the casual faithfulness condition is sufficient to guarantee identifiability and leads to polynomial-time search algorithms (Glymour et al., 1987; Spirtes et al., 2000), it is known to be a very strong assumption (see e.g. (Forster et al., 2015; Raskutti & Uhler, 2013)) that is often not satisfied in practice. Hence, milder assumptions have been considered.

Minimality assumptions, notably the *P-minimality* (Pearl, 2003) and SGS-minimality (Glymour et al., 1987) assumptions are two such assumptions. The P-minimality condition asserts that for directed graphical models satisfying the causal Markov condition, models that satisfy more d-separation rules are preferred. For example, suppose that there are two graphs G_1 and G_2 which are not Markov equivalent. G_1 is *strictly preferred* to G_2 if $D_{sep}(G_2) \subsetneq D_{sep}(G_1)$. The P-minimality condition asserts that no graph is strictly preferred to the true graph G^*. The SGS-minimality condition asserts that there exists no proper sub-graph of G^* that satisfies the causal Markov condition with respect to the probability distribution $\mathbb{P}$. To define the term sub-graph precisely, G_1 is a sub-graph of G_2 if $E(G_1) \subsetneq E(G_2)$. Zhang (Zhang, 2012) proved that the SGS-minimality assumption is weaker than the P-minimality assumption which is weaker than the casual faithfulness condition (for both directed acyclic graphs and directed cyclic graphs).

THEOREM 2 (ZHANG (2013) (ZHANG, 2012)). *If a directed acyclic graphical model* $(G^*, \mathbb{P})$ *satisfies the causal Markov condition and*

(a) the casual faithfulness condition, it satisfies the P-minimality condition.
(b) the P-minimality condition, it satisfies the SGS-minimality condition.

While Zhang (Zhang, 2012) states the results for directed acyclic graphs, the result easily extends to directed cyclic graphs.

2·2. *Sparsest Markov Representation (SMR) for directed acyclic graphs*

While the minimality assumptions are milder than the casual faithfulness condition, neither the P-minimality nor SGS-minimality assumptions imply identifiability of the Markov equivalence class for G^*. Recent work by Raskutti and Uhler developed the *sparsest Markov representation* sparsest Markov representation assumption (Raskutti & Uhler, 2013) and a slightly weaker version later referred to as *frugality* Forster et al. (2015) which applies to directed acyclic graphical modelss. The sparsest Markov representation assumption which we refer to here as the identifiable sparsest Markov representation assumption states that the true directed acyclic graphical model is the directed acyclic graph satisfying the causal Markov condition with the fewest edges. Here we say that a directed acyclic graph G_1 is *strictly sparser* than a directed acyclic graph G_2 if G_1 has *fewer* edges than G_2.

DEFINITION 6 (IDENTIFIABLE SPARSEST MARKOV REPRESENTATION (RASKUTTI & UHLER, 2013)). *A directed acyclic graphical model* $(G^*, \mathbb{P})$ *satisfies the identifiable sparsest Markov representation assumption if* $(G^*, \mathbb{P})$ *satisfies causal Markov condition and* $|S(G^*)| < |S(G)|$ *for every directed acyclic graph* G *such that* $(G, \mathbb{P})$ *satisfies the causal Markov condition and* $G \notin \mathcal{M}(G^*)$.

The identifiable sparsest Markov representation assumption is strictly weaker than the casual faithfulness condition while also ensuring a score-based method known as the SP algorithm (Raskutti & Uhler, 2013) recovers the true Markov equivalence class. Hence the identifiable sparsest Markov representation assumption guarantees identifiability of the Markov equivalence class for directed acyclic graphical modelss. A slightly weaker notion which we refer to as the weak sparsest Markov representation assumption does not guarantee model identifiability.

DEFINITION 7 (WEAK SPARSEST MARKOV REPRESENTATION (FRUGALITY) (FORSTER ET AL., 2015)). *A directed acyclic graphical model* $(G^*, \mathbb{P})$ *satisfies the weak sparsest Markov representation condition if* $(G^*, \mathbb{P})$ *satisfies the causal Markov condition and* $|S(G^*)| \leq |S(G)|$ *for every directed acyclic graph* G *such that* $(G, \mathbb{P})$ *satisfies the causal Markov condition and* $G \notin \mathcal{M}(G^*)$.

A comparison of sparsest Markov representation/frugality to the minimality assumptions and casual faithfulness condition for directed acyclic graphs is provided in Raskutti and Uhler Raskutti & Uhler (2013) and Forster et al. Forster et al. (2015).

THEOREM 3 (RASKUTTI AND UHLER (2013) (RASKUTTI & UHLER, 2013)). *If a directed acyclic graphical model* $(G^*, \mathbb{P})$ *satisfies*

(a) the casual faithfulness condition, it satisfies the identifiable sparsest Markov representation and consequently weak sparsest Markov representation assumptions.
(b) the weak sparsest Markov representation assumption, it satisfies the P-minimality and consequently the SGS-minimality conditions.

(c) the identifiable sparsest Markov representation assumption, G^* is identifiable up to Markov equivalence class.

It is unclear whether the sparsest Markov representation/frugality assumptions apply naturally to directed cyclic graphcial models since the success of the sparsest Markov representation assumption relies on the *local* Markov property which is known to hold for directed acyclic graphical modelss but not directed cyclic graphcial models Richardson (1994). In this paper, we investigate the extent to which these identifiability conditions apply to directed cyclic graphcial models and provide a new principle for learning directed cyclic graphcial models.

2·3. *Our contributions*

Based on this prior work, a natural question to consider is whether the identifiable and weak sparsest Markov representation assumptions developed for directed acyclic graphs apply to directed cyclic graphs and whether there are similar relationships between the casual faithfulness condition, sparsest Markov representation and minimality assumptions.

In this paper we address this question, by adapting both identifiable and weak sparsest Markov representation assumptions to directed cyclic graphs. One of the challenges we address is dealing with the distinction between real and virtual edges in directed cyclic graphs. We show that unlike for directed acyclic graphical modelss, the identifiable sparsest Markov representation assumption is not necessarily a weaker assumption than the casual faithfulness condition and while our simulations indicate that the identifiable sparsest Markov representation assumption recovers the true Markov equivalence class more frequently than the casual faithfulness condition, there exist no theoretical guarantees.

Consequently, we introduce a new principle which is the maximum d-separation rule (MDR) principle which chooses the directed Markov graph with the greatest number of d-separation rules. We show that our maximum d-separation rule principle is strictly weaker than the casual faithfulness condition and stronger than the P-minimality assumption, while also guaranteeing model identifiability for directed cyclic graphcial models. Our simulation results complement our theoretical results, showing that in general both the identifiable sparsest Markov representation and maximum d-separation rule assumptions are more successful assumptions than the casual faithfulness condition in terms of recovering the Markov equivalence class for directed cyclic graphcial models.

3. SPARSITY AND SPARSEST MARKOV REPRESENTATION FOR DIRECTED CYCLIC GRAPHCIAL MODELS

In this section, we extend notions of sparsity and the sparsest Markov representation assumptions to directed cyclic graphcial models. As mentioned earlier, in contrast to directed acyclic graphs, directed cyclic graphs can have two different types of edges which are real edges and virtual edges. In this paper, we define the *sparsest* directed cyclic graphical model as the model with the fewest *total edges* which are virtual edges without directions plus real edges. The main reason we choose total edges rather than just real edges is that all graphs in the same Markov Equivalence Class (MEC) have the same number of total edges (Richardson, 1994). However, the number of real edges may not be the same amongst the graphs even in the same Markov equivalence class. For example in Figure 2, there are two different Markov equivalence classes and each Markov equivalence class has two graphs: $G_1, G_2 \in \mathcal{M}(G_1)$ and $G_3, G_4 \in \mathcal{M}(G_3)$. G_1 and G_2 have 9 total edges but G_3 and G_4 has 7 total edges. On the other hand, G_1 has 6 real edges, G_2 has 9 real edges, G_3 has 5 real edges, and G_4 has 7 real edges (a bi-directed edge is

counted as 1 total edge). For a directed cyclic graph G, let $S(G)$ denote the *skeleton* of G where $(j,k) \in S(G)$ is a real or virtual edge.

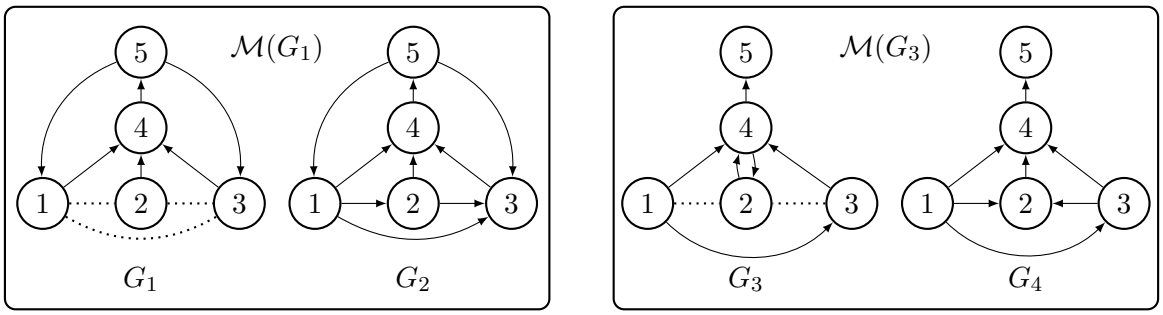


Fig. 2: 5-node examples with the different number of real edges and total edges

Using this definition of skeleton $S(G)$ for a directed cyclic graph, the definitions of the identifiable and weak sparsest Markov representation assumptions carry over from directed acyclic graphs to directed cyclic graphs. For completeness, we re-state the definitions here.

DEFINITION 8 (IDENTIFIABLE SPARSEST MARKOV REPRESENTATION FOR DIRECTED CYCLIC GRAPHS). *A pair $(G^*, \mathbb{P})$ satisfies the identifiable sparsest Markov representation condition if $(G^*, \mathbb{P})$ satisfies the causal Markov condition and $|S(G^*)| < |S(G)|$ for every directed cyclic graph G such that $(G, \mathbb{P})$ satisfies the causal Markov condition and $G \notin \mathcal{M}(G^*)$.*

DEFINITION 9 (WEAK SPARSEST MARKOV REPRESENTATION FOR DIRECTED CYCLIC GRAPHS). *A pair $(G^*, \mathbb{P})$ satisfies the weak sparsest Markov representation condition if $(G^*, \mathbb{P})$ satisfies the causal Markov condition and $|S(G^*)| \leq |S(G)|$ for every directed cyclic graph G such that $(G, \mathbb{P})$ satisfies the causal Markov condition and $G \notin \mathcal{M}(G^*)$.*

Unfortunately as we observe later unlike for directed acyclic graphical modelss, the identifiable sparsest Markov representation assumption is not weaker than the casual faithfulness condition in directed cyclic graphcial models. Hence the identifiable sparsest Markov representation assumption does not guarantee identifiability of Markov equivalence classes for directed cyclic graphcial models. On the other hand, while the weak sparsest Markov representation assumption may not guarantee uniqueness, we prove it is strictly weaker assumption than the casual faithfulness condition. We explore the relationship between the sparsest Markov representation assumptions and the casual faithfulness condition and minimality assumptions in the next section.

3·1. *Comparison of sparsest Markov representation, casual faithfulness condition and minimality assumptions for directed cyclic graphs*

Before presenting our main results of this section, we provide a lemma which highlights the important difference between the sparsest Markov representation assumptions for directed cyclic graphs compared to directed acyclic graphs. Note that the sparsest Markov representation assumptions involves counting the number of edges, whereas the casual faithfulness condition and P-minimality assumptions involve d-separation rules. First, we provide a proof for a fundamental link between the presence of an edge in $S(G)$ and d-separation/connection rules.

LEMMA 1. *For a directed graph G, $(j,k) \in S(G)$ if and only if j is d-connected to k given S for all $S \subset \{1,2,...,p\} \setminus \{j,k\}$.*

Proof. Suppose that (j, k) are adjacent. Then by the definition of d-connection (Richardson, 1994), there is no subset $S \subset \{1, 2, ..., p\} \setminus \{j, k\}$ such that j is d-separated from k given S. Suppose that (j, k) are not adjacent. We now show that there exists an $S \subset \{1, 2, ..., p\} \setminus \{j, k\}$ such that j is d-separated from k given S. Let $S = \text{an}(j) \cup \text{an}(k)$ then X_j is d-separated from X_k given S because the union of ancestors should not have any common child or its descendants otherwise (j, k) are virtually adjacent which is a contradiction. Furthermore, if the ancestors do not contain a common child of (j, k) then there is no path which d-connects j and k conditioned on its ancestors. Hence there is at least one set S d-separating j and k which completes the proof. □

Note that the above statement is true for real or virtual edges and not real edges alone. We now state an important lemma which shows the key difference in comparing the sparsest Markov representation assumptions to other identifiability assumptions for directed cyclic graphs, which does not arise for directed acyclic graphs.

LEMMA 2 *(a) For any two directed graphs G_1 and G_2, $D_{sep}(G_1) \subseteq D_{sep}(G_2)$ implies $S(G_2) \subseteq S(G_1)$.*

(b) There exist two directed graphs G_1 and G_2 such that $S(G_1) = S(G_2)$, but $D_{sep}(G_1) \subsetneq D_{sep}(G_2)$. For directed acyclic graphs, no two such graphs exist.

Proof. We begin with the proof of (a). Suppose that $S(G_1)$ is not a sub-skeleton of $S(G_2)$, meaning that there exists a pair $(j, k) \in S(G_1)$ and $(j, k) \notin S(G_2)$. By Lemma 1, j is d-connected to k given S for all $S \subset \{1, 2, ..., p\} \setminus \{j, k\}$ in G_1 while there exist $S \subset \{1, 2, ..., p\} \setminus \{j, k\}$ d-separating j and k in G_2. Hence it is contradictory that $D_{sep}(G_1) \subset D_{sep}(G_2)$. For (b), we refer to the example in Figure 3.

In Figure 3, the unshielded triple (X_1, X_4, X_2) is a conductor in G_1 and an imperfect non-conductor in G_2 because of a reversed directed edge between (X_4, X_5). By the property of a conductor, in order for (X_1, X_4) to be d-separated in G_1, X_2 should be included in the conditioning set. In contrast, for G_2, X_1 is d-separated from X_4 given the empty set (which does not include X_2).

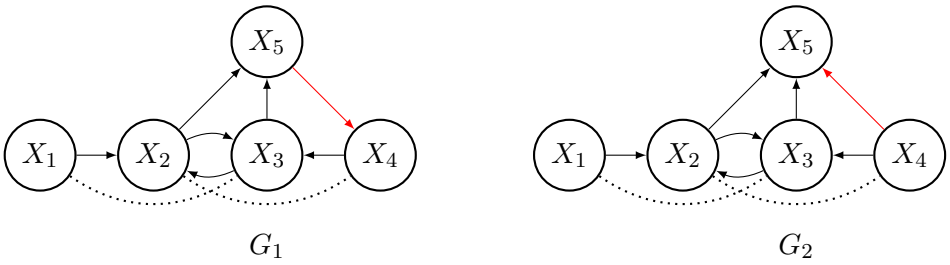


Fig. 3: 5-node examples for Lemma 2 and Theorem 4

Lemma 2 (a) holds for both directed acyclic graph and directed cyclic graphcial models and allows us to conclude a subset-superset relation between edges in the skeleton and d-separation rules in a graph G. Part (b) is where there is a key difference directed acyclic graphs and directed cyclic graphs. Part (b) asserts that there are examples in which the edge set in the skeleton may be totally equivalent, yet one graph entails a strict superset of d-separation rules.

Now we present the main result of this section which compares the identifiable and weak sparsest Markov representation assumptions with the casual faithfulness condition and P-minimality assumption.

THEOREM 4. *For directed cyclic graphcial models,*

(a) the weak sparsest Markov representation assumption is weaker than the casual faithfulness condition.
(b) there exists a directed cyclic graphical model $(G, \mathbb{P})$ *satisfying the casual faithfulness condition that does not satisfy the identifiable sparsest Markov representation assumption.*
(c) the identifiable sparsest Markov representation assumption is stronger than the P-minimality assumption.
(d) there exists a directed cyclic graphical model $(G, \mathbb{P})$ *satisfying the weak sparsest Markov representation assumption that does not satisfy the P-minimality assumption.*

Proof.(a) The proof for (a) follows from Lemma 2 (a). If a directed cyclic graphical model $(G^*, \mathbb{P})$ satisfies the casual faithfulness condition, then for all graphs G such that $(G, \mathbb{P})$ satisfies the causal Markov condition, $D_{sep}(G) \subseteq D_{sep}(G^*)$. Hence based on Lemma 2 (a), $S(G^*) \subseteq S(G)$ and $(G^*, \mathbb{P})$ satisfies the weak sparsest Markov representation assumption.
(b) We refer to the example in Figure 3 where $(G_2, \mathbb{P})$ satisfies the casual faithfulness condition.
(c) The proof for (c) again follows from Lemma 2 (a). Suppose that a directed cyclic graphical model $(G^*, \mathbb{P})$ fails to satisfy the P-minimality assumption. This implies that there exists a directed cyclic graph G such that $(G, \mathbb{P})$ satisfies the causal Markov condition and $D_{sep}(G^*) \subsetneq D_{sep}(G)$. Lemma 2 (a) implies $S(G) \subseteq S(G^*)$. Hence G^* cannot have the fewest skeletons uniquely, therefore $(G^*, \mathbb{P})$ fails to satisfy the identifiable sparsest Markov representation assumption.
(d) We refer to the example in Figure 3 where $(G_1, \mathbb{P})$ satisfies the weak sparsest Markov representation assumption. Further explanation is given in Figure 14 in the Appendix. □

If $(G, \mathbb{P})$ satisfies the casual faithfulness condition, the weak sparsest Markov representation assumption is satisfied whereas the identifiable sparsest Markov representation assumption is not necessarily satisfied. For directed acyclic graphical modelss, the identifiable sparsest Markov representation assumption is strictly weaker than the casual faithfulness condition and the identifiable sparsest Markov representation assumption guarantees identifiability of the true Markov equivalence class. However, Theorem 4 (b) implies that the identifiable sparsest Markov representation assumption is not strictly weaker than the casual faithfulness condition for directed cyclic graphcial models. On the other hand, unlike for directed acyclic graphical modelss, weak sparsest Markov representation assumption does not imply P-minimality assumption for directed cyclic graphcial models, according to (d). In Section 5, we implement an algorithm that uses the identifiable sparsest Markov representation assumption and the results seem to suggest that for most directed cyclic graphcial models, the identifiable sparsest Markov representation assumption is weaker than the casual faithfulness condition.

4. NEW PRINCIPLE: MAXIMUM D-SEPARATION RULES (MDR)

In light of the fact that the identifiable sparsest Markov representation assumption does not lead to a strictly weaker assumption than the casual faithfulness condition, we introduce the Maximum d-separation rules (MDR) assumption. The maximum d-separation rule assumption

asserts that G^* entails more d-separation rules than any other graph satisfying the causal Markov condition according to the given distribution $\mathbb{P}$.

DEFINITION 10 (MAXIMUM D-SEPARATION RULES (MDR) ASSUMPTION). *A directed cyclic graphical model $(G^*, \mathbb{P})$ satisfies the maximum* d-separation *rules (MDR) assumption if $(G^*, \mathbb{P})$ satisfies the causal Markov condition and $|D_{sep}(G)| < |D_{sep}(G^*)|$ for every directed cyclic graph G such that $(G, \mathbb{P})$ satisfies the causal Markov condition and $G \notin \mathcal{M}(G^*)$.*

There is a natural and intuitive connection between the maximum d-separation rule assumption and the P-minimality assumption. Both assumptions encourage directed cyclic graphs to entail more d-separation rules. The key difference between the P-minimality assumption and the maximum d-separation rule assumption is that the P-minimality assumption requires that there is no directed cyclic graphs that entail a *strict superset* of d-separation rules whereas the maximum d-separation rule assumption simply requires that there are no directed cyclic graphs that entail a *greater number* of d-separation rules.

4·1. *Comparison of maximum d-separation rule to casual faithfulness condition and minimality assumptions for directed cyclic graphs*

We provide a comparison of the maximum d-separation rule assumption to the casual faithfulness condition and P-minimality assumptions. For ease of notation, let $\mathcal{G}_M(\mathbb{P})$ and $\mathcal{G}_F(\mathbb{P})$ denote the set of Markovian directed cyclic graphcial models satisfying the maximum d-separation rule assumption and casual faithfulness condition, respectively. In addition, let $\mathcal{G}_P(\mathbb{P})$ denote the set of directed cyclic graphcial models satisfying the P-minimality condition.

THEOREM 5. *Consider a directed cyclic graphical model $(G^*, \mathbb{P})$.*

(a) If $\mathcal{G}_F(\mathbb{P}) \neq \emptyset$, then $\mathcal{G}_F(\mathbb{P}) = \mathcal{G}_M(\mathbb{P})$. Consequently if $(G^*, \mathbb{P})$ satisfies the casual faithfulness condition, then $\mathcal{G}_F(\mathbb{P}) = \mathcal{G}_M(\mathbb{P}) = \mathcal{M}(G^*)$.
(b) There exists $\mathbb{P}$ for which $\mathcal{G}_F(\mathbb{P}) = \emptyset$ while $(G^*, \mathbb{P})$ satisfies the maximum d-separation rule assumption and $\mathcal{G}_M(\mathbb{P}) = \mathcal{M}(G^*)$.
(c) $\mathcal{G}_M(\mathbb{P}) \subset \mathcal{G}_P(\mathbb{P})$.
(d) There exists $\mathbb{P}$ for which $\mathcal{G}_M(\mathbb{P}) = \emptyset$ while $(G^*, \mathbb{P})$ satisfies the P-minimality assumption and $\mathcal{G}_P(\mathbb{P}) \supset \mathcal{M}(G^*)$.

Proof. For (a), suppose that $(G^*, \mathbb{P})$ satisfies the casual faithfulness condition, then $CI(\mathbb{P})$ is the same as the set of d-separation rules entailed by G^*. Note that if $(G, \mathbb{P})$ satisfies the causal Markov condition, then $CI(\mathbb{P})$ is a superset of the set of d-separation rules entailed by G and therefore $D_{sep}(G) \subset D_{sep}(G^*)$. This allows us to conclude that graphs in $\mathcal{M}(G^*)$ should entail the maximum number of d-separation rules among graphs satisfying the causal Markov condition. Furthermore, prior results show that $\mathcal{G}_F(\mathbb{P}) = \mathcal{M}(G^*)$ which completes the proof.

For (c), suppose that $(G^*, \mathbb{P})$ fails to satisfy the P-minimality assumption. By the definition of the P-minimality assumption, there exists $(G, \mathbb{P})$ satisfying the causal Markov condition such that $D_{sep}(G^*) \subsetneq D_{sep}(G)$. Therefore, G^* entails strictly less d-separation rules than G, and $(G^*, \mathbb{P})$ violates the maximum d-separation rule assumption.

For (b) and (d), we refer to the example in Figure 4. First we show that $(G_1, \mathbb{P})$ satisfies the maximum d-separation rule assumption but not the casual faithfulness condition, whereas $(G_2, \mathbb{P})$ satisfies the P-minimality assumption but not the maximum d-separation rule assumption. Suppose that X_1, X_2, X_3, X_4 are random variables with probability distribution $\mathbb{P}$ with the following conditional indenpendence statements:

$$CI(\mathbb{P}) = \{X_1 \perp\!\!\!\perp X_3 \mid X_2,\ X_2 \perp\!\!\!\perp X_4 \mid X_1,\ X_3,\ X_1 \perp\!\!\!\perp X_2 \mid X_4\}. \tag{1}$$

Any graph satisfying the causal Markov condition with respect to $\mathbb{P}$ must only entail a subset of the three d-separation rules: $\{X_1\text{d-sep}X_3 \mid X_2,\ X_2\text{d-sep}X_4 \mid X_1, X_3,\ X_1\text{d-sep}X_2 \mid X_4\}$. Clearly $D_{sep}(G_1) = \{X_1\text{d-sep}X_3 \mid X_2,\ X_2\text{d-sep}X_4 \mid X_1, X_3\}$, therefore $(G_1, \mathbb{P})$ satisfies the causal Markov condition. It can be shown that no graph entails any subset containing two or three of these d-separation rules other than G_1. Hence no graph follows the casual faithfulness condition with respect to $\mathbb{P}$ since there is no graph that entails all three d-separation rules and $(G_1, \mathbb{P})$ satisfies the maximum d-separation rule assumption because no graph that entails more or as many d-separation rules as G_1 entails, and satisfies the causal Markov condition with respect to $\mathbb{P}$.

For (d), note that G_2 entails the sole d-separation rule, $D_{sep}(G_2) = \{X_1\text{d-sep}X_2 \mid X_4\}$ and it is clear to see that $(G_2, \mathbb{P})$ satisfies the causal Markov condition. If $(G_2, \mathbb{P})$ does not satisfy the P-minimality assumption, there exists a graph G such that $(G, \mathbb{P})$ satisfies the causal Markov condition and $D_{sep}(G) \subsetneq D_{sep}(G_2)$. It can be shown that no such graph exists. Therefore, $(G_2, \mathbb{P})$ satisfies the P-minimality assumption. Clearly $(G_2, \mathbb{P})$ fails to satisfy the maximum d-separation rule assumption because G_1 entails more d-separation rules.

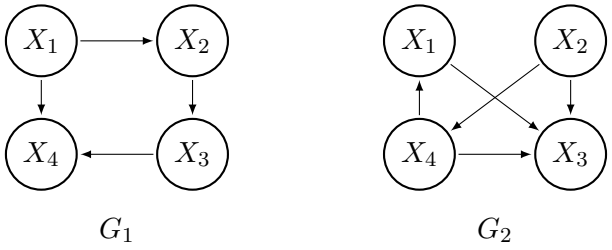


Fig. 4: 4-node examples for Theorem 5

Theorem 5 (a) asserts that whenever the set of directed cyclic graphcial models satisfying the casual faithfulness condition is not empty, it is equivalent to the set of directed cyclic graphcial models satisfying the maximum d-separation rule assumption. Part (b) claims that there exists a distribution in which no directed cyclic graphical model satisfies the casual faithfulness condition, while the set of directed cyclic graphcial models satisfying the maximum d-separation rule assumption consists of its Markov equivalence class. Hence, (a) and (b) show that the maximum d-separation rule assumption is strictly superior to the casual faithfulness condition in terms of recovering the true Markov equivalence class. Theorem 5 (c) claims that any directed cyclic graphcial models satisfying the maximum d-separation rule assumption should lie in the set of directed cyclic graphcial models satisfying the P-minimality assumption. (d) asserts that there are some directed cyclic graphcial models satisfying the P-minimality assumption but violating the maximum d-separation rule assumption. Therefore, (c) and (d) prove that the maximum d-separation rule assumption is strictly stronger than the P-minimality assumption.

4·2. *Comparison between the maximum d-separation rule and sparsest Markov representation assumptions*

Now we show that the maximum d-separation rule assumption is neither weaker nor stronger than the sparsest Markov representation assumptions for both directed acyclic graph and directed cyclic graphcial models.

LEMMA 3*(a) There exists a directed acyclic graphical model that satisfies the identifiable sparsest Markov representation assumption that does not satisfy the maximum d-separation*

rule assumption. Further, there exists a directed acyclic graphical model satisfying the maximum d-separation rule assumption that does not satisfy the weak sparsest Markov representation assumption.

(b) There exists a directed graphical model with cycles that satisfies the same conclusion as (a).

Proof. Our proof for Lemma 3 involves us constructing two sets of examples, one for directed acyclic graphs corresponding to (a) and one for cyclic graphs corresponding to (b). Figure 5 displays two directed acyclic graphs, G_1 and G_2 which are clearly not in the same Markov equivalence class. For clarity, we used red arrows for the difference between graphs. We associate the same distribution $\mathbb{P}$ to each directed acyclic graph, which is provided in Appendix B·1. With this choice of distribution, both $(G_1, \mathbb{P})$ and $(G_2, \mathbb{P})$ satisfy the causal Markov condition (explained in Appendix B·1). The main point of this example is that $(G_2, \mathbb{P})$ satisfies the sparsest Markov representation assumption whereas $(G_1, \mathbb{P})$ does not and $(G_1, \mathbb{P})$ satisfies the maximum d-separation rule assumption whereas $(G_2, \mathbb{P})$ does not. A more detailed proof that $(G_1, \mathbb{P})$ satisfies the maximum d-separation rule assumption whereas $(G_2, \mathbb{P})$ satisfies the sparsest Markov representation assumption is provided in Appendix B·1.

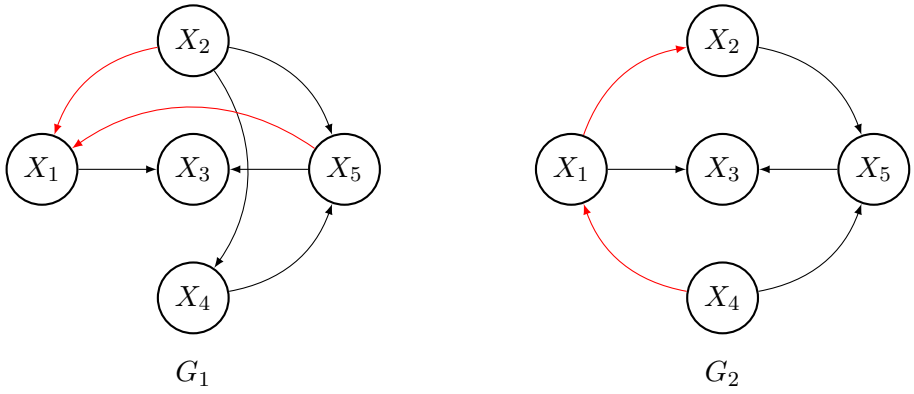


Fig. 5: 5-node examples for Lemma 3.(a)

For Lemma 3 (b), Figure 6 displays two directed cyclic graphs G_1 and G_2 which do not belong to the same Markov equivalence class. Once again red arrows are used to denote the edges (both real and virtual) that are different between the graphs. We associate the same distribution $\mathbb{P}$ to each graph such that both $(G_1, \mathbb{P})$ and $(G_2, \mathbb{P})$ satisfy the causal Markov condition (explained in Appendix C·2). Again, the main idea of this example is that $(G_1, \mathbb{P})$ satisfies the maximum d-separation rule assumption whereas $(G_2, \mathbb{P})$ satisfies the identifiable sparsest Markov representation assumption. A detailed proof that $(G_1, \mathbb{P})$ satisfies the maximum d-separation rule assumption whereas $(G_2, \mathbb{P})$ satisfies the identifiable sparsest Markov representation assumption can be found in Appendix C·2.

Intuitively, the reason why fewer edges does not necessarily translate to satisfying more d-separation rules is that the placement of edges relative to the rest of the graph and what additional paths they allow affects the total number of d-separation rules entailed by the graph.

In summary, the flow chart in Fig. **??** shows how the casual faithfulness condition, sparsest Markov representation, maximum d-separation rule and minimality assumptions are related for both directed acyclic graph and directed cyclic graphcial models:

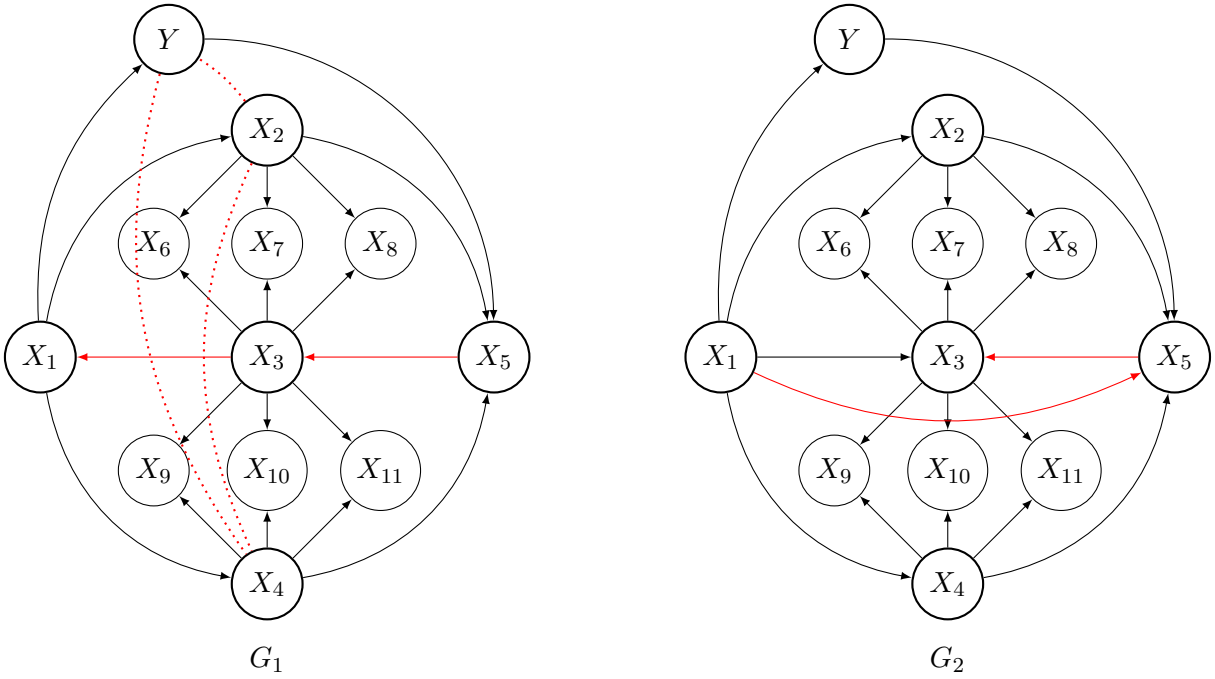
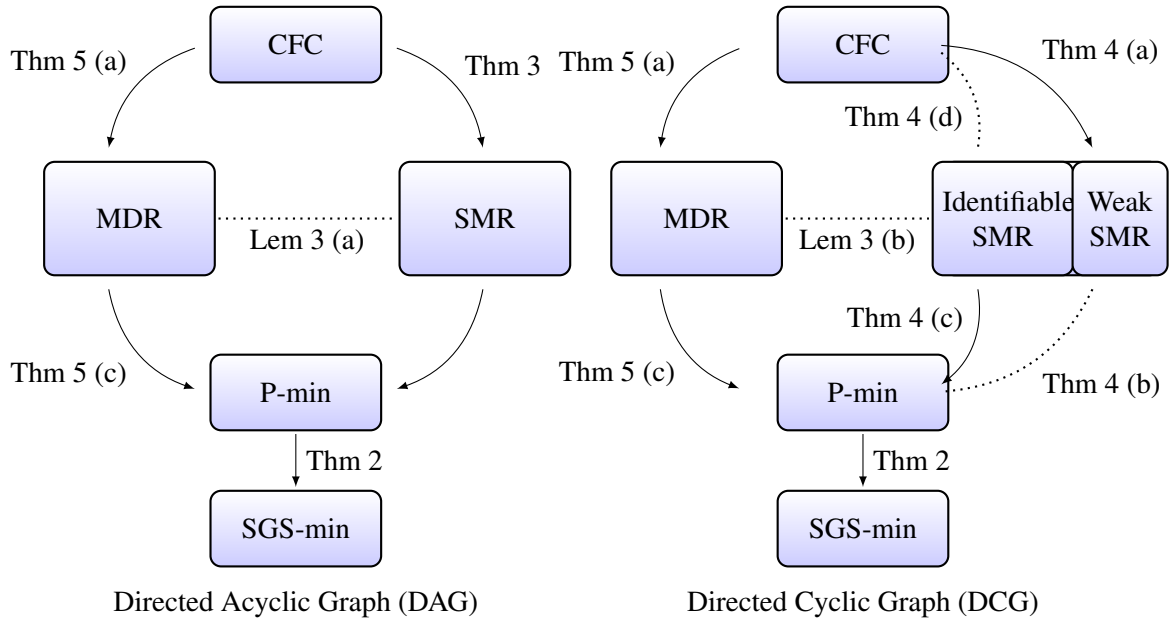


Fig. 6: 12-node examples for Lemma 3.(b)



5. Simulation results

In Section 3 and 4, we showed that the maximum d-separation rule assumption is strictly weaker than the casual faithfulness condition, and the maximum d-separation rule and identifiable sparsest Markov representation assumptions are strictly stronger than the P-minimality as-

sumption for directed cyclic graphcial models. Hence there are more graphical models satisfying the the maximum d-separation rule assumption than the casual faithfulness condition and there are less directed cyclic graphcial models satisfying the maximum d-separation rule assumption or the identifiable sparsest Markov representation assumption than the P-minimality assumption. In this section, we support our theoretical results with numerical experiments on small Gaussian linear directed cyclic graphcial models (see e.g. (Spirtes, 1995)).

The simulation study was conducted using 100 realizations of 5-node random Gaussian linear directed cyclic graphcial models in which distribution $\mathbb{P}$ is defined by the following linear structural equations:

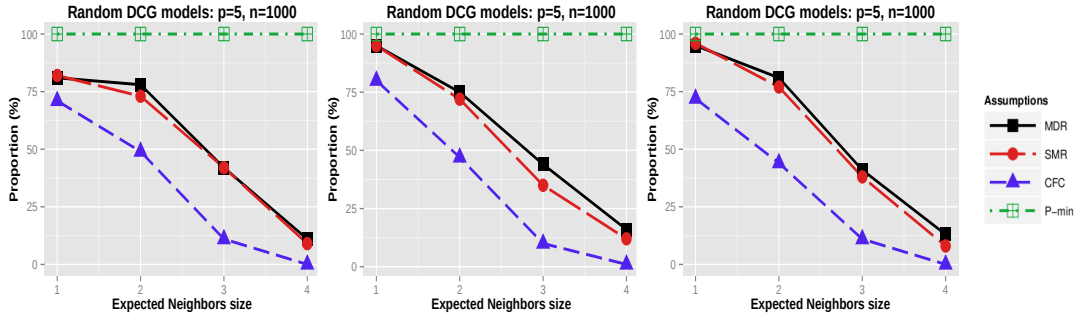
$$(X_1, X_2, ..., X_p)^T = B(X_1, X_2, ..., X_p)^T + \epsilon$$

where $B \in \mathbb{R}^{p\times p}$ is an edge weight matrix with $B_{jk} = \beta_{jk}$ and β_{jk} is a weight of an edge from X_j to X_k and $\epsilon \sim \mathrm{N}(0, I_p)$ where $I_p \in \mathbb{R}^{p\times p}$ is the identity matrix. The matrix B encodes the directed cyclic graph structure since if β_{jk} is non-zero, the pair (X_j, X_k) is *really adjacent* and if there is a set of nodes $S = (s_1, s_2, ..., s_t)$ such that $\beta_{js_1}\beta_{ks_1}\beta_{s_1s_2}...\beta_{s_tj}$ is non-zero, the pair (X_j, X_k) is *virtually adjacent*. The edge weight parameters were chosen uniformly at random in the range $\beta_{jk} \in [-1, -0.25] \cup [0.25, 1]$, ensuring the edge weights are bounded away from 0. Note that if the graph is a directed acyclic graph, we would need to impose the constraint that B is lower triangular however for directed cyclic graphs we impose no such constraints. Further, we impose sparsity by assigning a probability that each coefficient of the matrix B is non-zero and we set the expected neighborhood size range from 1 (sparse directed cyclic graph) to $p - 1$ depending on the edge weight probability.

Subsequently, n samples were drawn from the distribution induced by the Gaussian linear directed cyclic graphical model and we report results for $n \in \{100, 200, 500, 1000\}$. The conditional indenpendence statements were estimated based on Fisher's conditional correlation test where the z-transform with significance levels $\alpha = \{0.01, 0.001, 0.0001\}$. All possible directed graphs satisfying the causal Markov condition are detected from an exhaustive search and we measure two things: (1) what proportion of graphs in the simulation satisfy each assumption (casual faithfulness condition, P-minimality, sparsest Markov representation, maximum d-separation rule); and (2) what proportion of simulations (out of 100) recover the skeleton $S(G)$ for the true graph corresponding to the matrix B according to each assumption. (1) addresses the issue of how strong each assumption is and (2) addresses the issue of how likely each assumption is to recover the true graph.

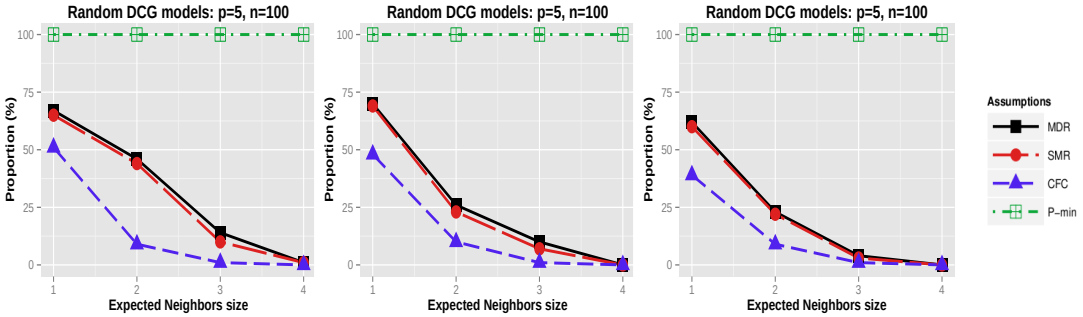
5·1. *Random directed cyclic graphcial models*

In Figure 7, and 8, we simulated how restrictive each identifiability condition (casual faithfulness condition, P-minimality, sparsest Markov representation, maximum d-separation rule) is for random directed cyclic graphcial models with different sample sizes, different significance levels, and different expected neighborhood sizes. As shown in Fig. 7, and 8, there are more directed cyclic graphcial models satisfying the maximum d-separation rule assumption than the casual faithfulness condition and less directed cyclic graphcial models satisfying the maximum d-separation rule assumption than the P-minimality assumption for both large and small sample size cases. We can also see a similar relationships between the casual faithfulness condition, identifiable sparsest Markov representation and P-minimality assumptions. These results support our theoretical result that the maximum d-separation rule assumption is weaker than the casual faithfulness condition but stronger than the P-minimality assumption and the identifiable sparsest Markov representation assumption is stronger then the P-minimality assumption. Although there is no theoretical guarantee that the identifiable sparsest Markov representation condition



(a) $\alpha = 0.01$ (b) $\alpha = 0.001$ (c) $\alpha = 0.0001$

Fig. 7: Proportions of random directed cyclic graphcial models satisfying the casual faithfulness condition, maximum d-separation rule, sparsest Markov representation and P-minimality assumptions with large sample size $n = 1000$, varying expected neighborhood size



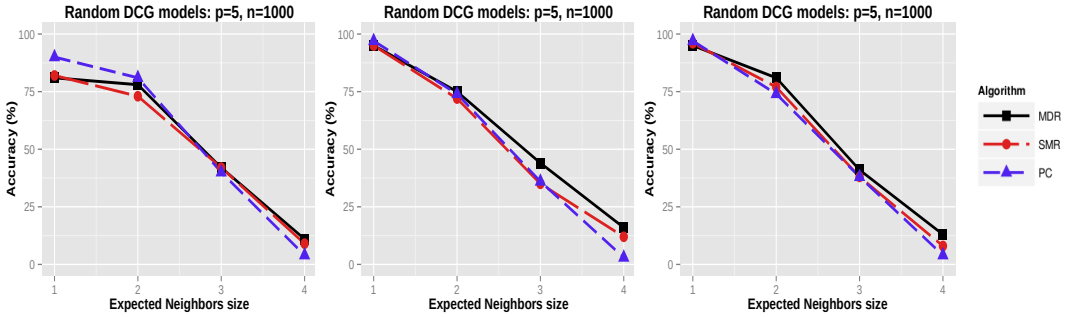
(a) $\alpha = 0.01$ (b) $\alpha = 0.001$ (c) $\alpha = 0.0001$

Fig. 8: Proportions of random directed cyclic graphcial models satisfying the casual faithfulness condition, maximum d-separation rule, sparsest Markov representation and P-minimality assumptions with small sample size $n = 100$, varying expected neighborhood size

is weaker than the casual faithfulness condition, Fig. 7, and 8 represent that the identifiable sparsest Markov representation assumption is substantially weaker than the casual faithfulness condition on average.

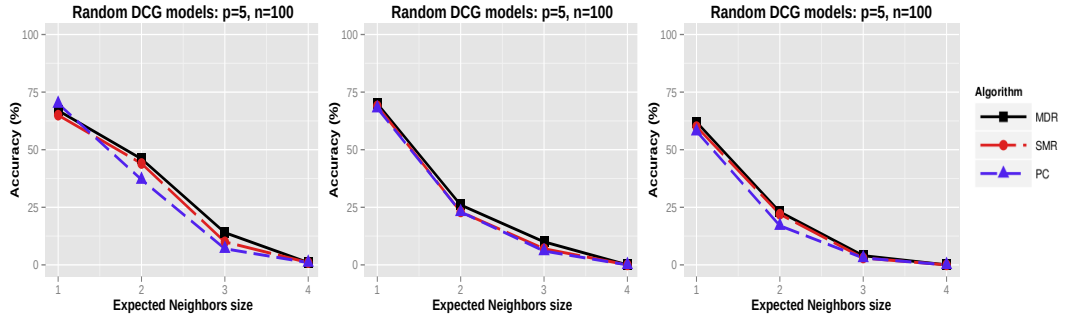
Figure 9, and 10 show recovery rates of skeletons using the maximum d-separation rule and identifiable sparsest Markov representation assumptions, and the PC algorithm (Spirtes & Glymour, 1991) where the weakest known sufficient condition is the casual faithfulness condition. Here we define a success to be that the algorithm recovers the true skeleton of the graph. The reason we use the PC algorithm here which generally applies to directed acyclic graphical modelss instead of the CCD algorithm Richardson (1996a) which applies to directed cyclic graphcial models is that the CCD and PC algorithms are identical in terms of the first step of recovering the skeleton and that is what we are interested in here.

Specifically, the PC algorithm removes an edge between a pair nodes if the pair is conditionally independent given any subset of rest of variables. For the PC algorithm, we used the R package 'pcalg' (Kalisch et al., 2012). We considered the case in which multiple graphs with different skeletons have the most d-separation rules as a failure of our algorithm. Our simulation results



(a) $\alpha = 0.01$ (b) $\alpha = 0.001$ (c) $\alpha = 0.0001$

Fig. 9: Accuracy rates of recovering skeletons of random directed cyclic graphcial models using the maximum d-separation rule and identifiable sparsest Markov representation assumptions, and the PC algorithm with large sample size $n = 1000$, varying expected neighborhood size



(a) $\alpha = 0.01$ (b) $\alpha = 0.001$ (c) $\alpha = 0.0001$

Fig. 10: Accuracy rates of recovering skeletons of random directed cyclic graphcial models using the maximum d-separation rule and identifiable sparsest Markov representation assumptions, and the PC algorithm with small sample size $n = 100$, varying expected neighborhood size

allow us to conclude that the maximum d-separation rule outperforms than the casual faithfulness condition and identifiable sparsest Markov representation assumption on average in terms of recovering of skeletons.

5·2. *Special graph structures*

We also provide a comparison between the maximum d-separation rule, identifiable sparsest Markov representation, P-minimality assumptions and casual faithfulness condition using specific graph structures, namely a tree, bipartite, and cycle. Figure 11 shows examples of skeletons of these special graphs.

We generate these graphs as follows: First, we set the skeleton for our desired graph based on Figure. 11 and then determine the edges weights which are chosen uniformly at random from the set $\beta_{jk} \in [-1, -0.25] \cup [0.25, 1]$, ensuring again that the edge weights are bounded away from 0. Second, we assign a randomly chosen direction to each edge. Therefore, the graphs generated may have cycles and virtual edges. For a cycle graph, we fix the directions of edges in order to

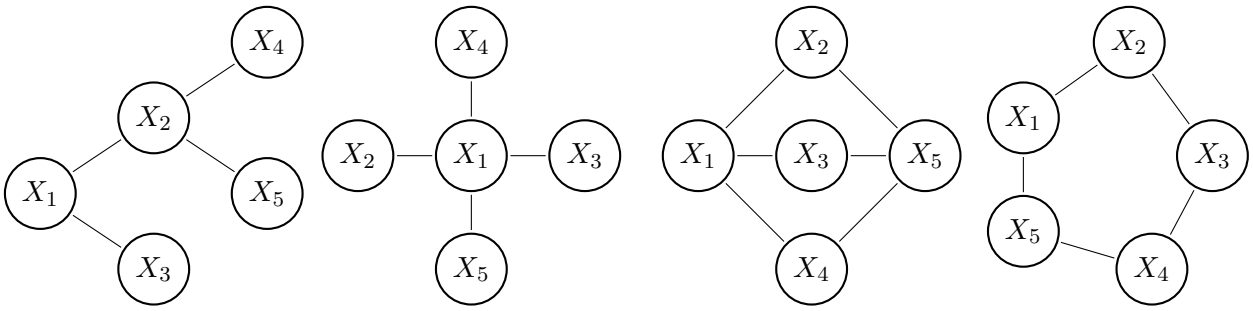
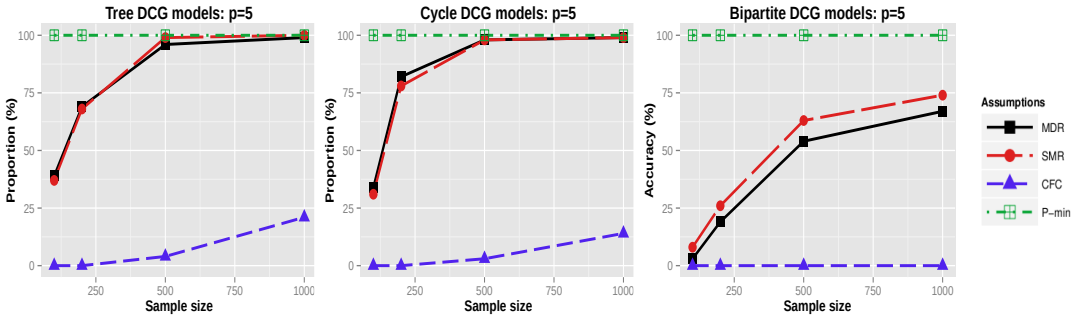


Fig. 11: Skeletons of a tree, bipartite, and cycle graphs

have a cycle $X_1 \to X_2 \to ... \to X_p \to X_1$. We report the results for $n \in \{100, 200, 500, 1000\}$ for the fixed $\alpha = 0.001$ for all experiments.

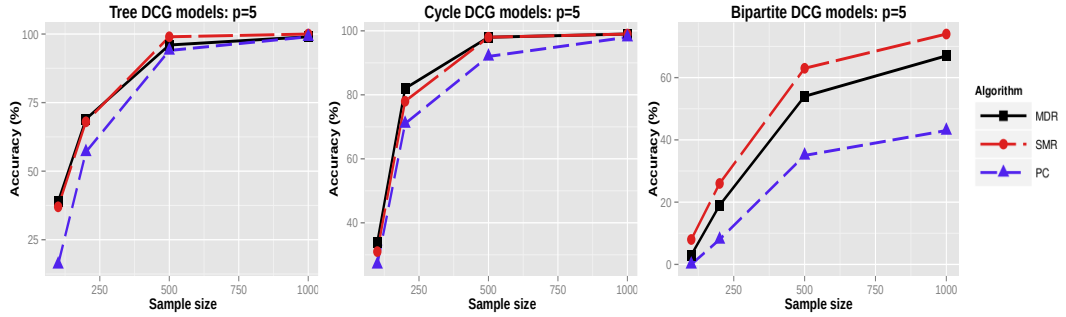


(a) Tree (b) Cycle (c) Bipartite

Fig. 12: Proportions of special types of directed cyclic graphcial models satisfying the casual faithfulness condition, maximum d-separation rule, sparsest Markov representation, and P-minimality assumptions with $\alpha = 0.001$

In Figure 12, we simulated how restrictive the casual faithfulness condition, maximum d-separation rule, identifiable sparsest Markov representation, and P-minimality assumptions are for random tree, bipartite, and cycle directed cyclic graphcial models with different sample sizes. As Fig. 12 shows, there are more directed cyclic graphcial models satisfying the maximum d-separation rule assumption than the casual faithfulness condition and less directed cyclic graphcial models satisfying the maximum d-separation rule assumption than the P-minimality assumption for all tree, bipartite, and cycle graphs. This result is consistent with our theoretical result that the maximum d-separation rule assumption is weaker than the casual faithfulness condition but stronger than the P-minimality assumption. We can see the similar relationships between the casual faithfulness condition, identifiable sparsest Markov representation, and P-minimality assumptions, which supports our main result that the identifiable sparsest Markov representation assumption is stronger than the P-minimality assumption.

Figure 13 shows the proportions of recovering skeletons of each type of graph (Tree, Bipartite, Tree) using the maximum d-separation rule and identifiable sparsest Markov representation assumptions, and the PC algorithm. These simulation results show that the maximum d-separation



(a) Tree (b) Cycle (c) Bipartite

Fig. 13: Accuracy rates of recovering skeletons of random directed cyclic graphcial models using the maximum d-separation rule and identifiable sparsest Markov representation assumptions, and PC algorithms with $\alpha = 0.001$

rule and identifiable sparsest Markov representation assumptions are favorable to the casual faithfulness condition on average in terms of recovering of skeletons for all types of graphs. In addition, the large sample size case has higher probability of recovering skeletons than the small sample case because of the small errors of the conditional indenpendence tests.

ACKNOWLEDGEMENT

GP and GR were both supported by NSF DMS-1407028 over the duration of this project.

APPENDIX 1

Examples for Theorem 4 *(d)*

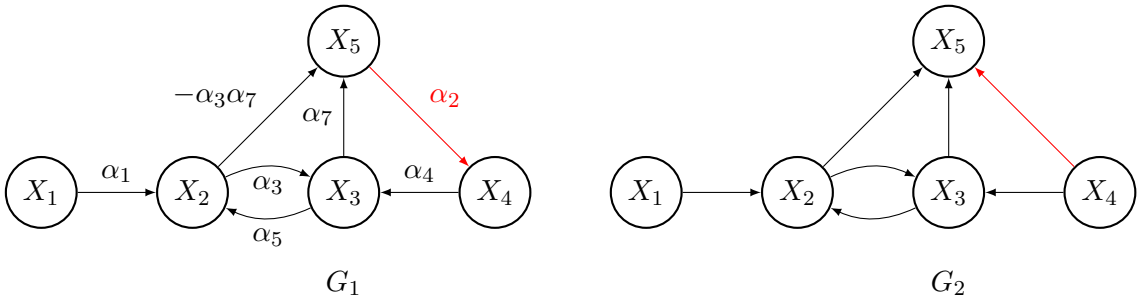


Fig. 14: 5-node examples for Theorem 4 (d)

Suppose that $(G_1, \mathbb{P})$ is a Gaussian linear directed cyclic graphical model with specified edge weights in Figure 14. With this choice of distribution $\mathbb{P}$ based on G_1 in Figure 14, we have a set of conditional indenpendence statements which are the same as the set of d-separation rules entailed by G_1 and an additional set of conditional indenpendence statements, $CI(\mathbb{P}) \supset \{X_1 \perp\!\!\!\perp X_4 | \emptyset$, or$X_5$, $X_1 \perp\!\!\!\perp X_5 | \emptyset$, or $X_4\}$.

It is clear that $(G_2, \mathbb{P})$ satisfies the causal Markov condition and $D_{sep}(G_1) \subsetneq D_{sep}(G_2)$ (explained in Section 3). This implies that $(G_1, \mathbb{P})$ fails to satisfy the P-minimality assumption.

Now we prove that $(G_1, \mathbb{P})$ satisfies the weak sparsest Markov representation assumption. Suppose that $(G_1, \mathbb{P})$ does not satisfy the weak sparsest Markov representation assumption. Then there exists a G such that $(G, \mathbb{P})$ satisfies the causal Markov condition and has fewer edges than G_1. By Lemma 2, if $(G, \mathbb{P})$ satisfies the casual faithfulness condition, G satisfies the weak sparsest Markov representation assumption. Note that G_1 does not have edges between (X_1, X_4) and (X_1, X_5). Since the only additional conditional independence statements that are not entailed by G_1 are $\{X_1 \perp\!\!\!\perp X_4 | \emptyset$, or$X_5$, $X_1 \perp\!\!\!\perp X_5 | \emptyset$, or $X_4\}$, no graph that satisfies the causal Markov condition with respect to $\mathbb{P}$ can have fewer edges than G_1. This leads to a contradiction and hence $(G_1, \mathbb{P})$ satisfies the weak sparsest Markov representation assumption.

APPENDIX 2

B·1. *Proof of Lemma* 3 *(a)*

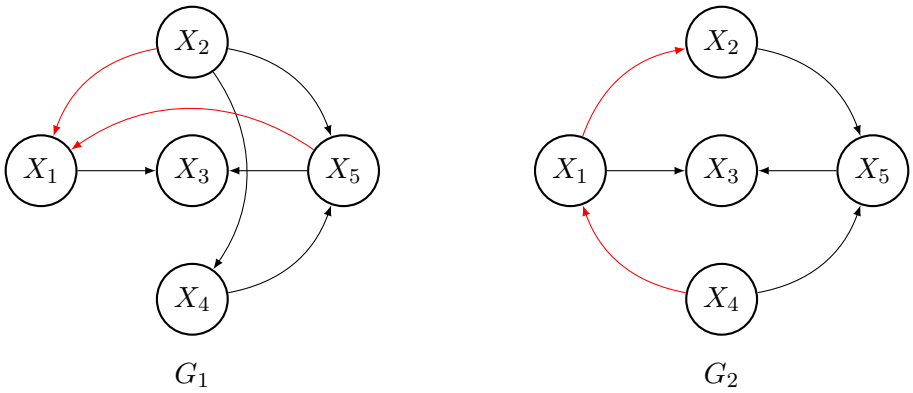


Fig. 15: 5-node examples for Lemma 3.(a)

Proof. Here we show that $(G_1, \mathbb{P})$ satisfies the identifiable sparsest Markov representation assumption and and $(G_2, \mathbb{P})$ satisfies the maximum d-separation rule assumption, where $\mathbb{P}$ has the following conditional indenpendence statements:

$$CI(\mathbb{P}) = \{X_2 \perp\!\!\!\perp X_3 \mid (X_1, X_5) \text{ or } (X_1, X_4, X_5), X_3 \perp\!\!\!\perp X_4 \mid (X_1, X_5), (X_2, X_5), \text{ or } (X_1, X_2, X_5),$$
$$X_1 \perp\!\!\!\perp X_4 \mid (X_2, X_5) \text{ or } (X_2, X_3, X_5), X_2 \perp\!\!\!\perp X_4 \mid X_1, X_1 \perp\!\!\!\perp X_5 \mid (X_2, X_4)\}.$$

Clearly both directed acyclic graphs G_1 and G_2 do not belong to the same Markov equivalence class since they have different skeletons. To be explicit, we state all d-separation rules entailed by G_1 and G_2. Both graphs entail the following sets of d-separation rules:

- X_2 is d-separated from X_3 given (X_1, X_5) or (X_1, X_4, X_5).
- X_3 is d-separated from X_4 given (X_1, X_5) or (X_1, X_2, X_5).

The set of d-separation rules entailed by G_1 which are not entailed by G_2 is as follows:

- X_1 is d-separated from X_4 given (X_2, X_5) or (X_2, X_4, X_5).
- X_3 is d-separated from X_4 given (X_2, X_5).

Furthermore, the set of d-separation rules entailed by G_2 which are not entailed by G_1 is as follows:

- X_1 is d-separated from X_5 given (X_2, X_4).
- X_2 is d-separated from X_4 given X_1. □

With our choice of distribution, both directed acyclic graphical modelss $(G_1, \mathbb{P})$ and $(G_2, \mathbb{P})$ satisfy the causal Markov condition and it is straightforward to see that G_2 has fewer edges than G_1 while G_1 entails more d-separation rules than G_2.

It can be shown from an exhaustive search that there is no graph G such that G is sparser or as sparse as G_2 and $(G, \mathbb{P})$ satisfies the causal Markov condition. Moreover, it can be shown that G_1 entails the maximum d-separation rules amongst graphs satisfying the causal Markov condition with respect to the distribution again through an exhaustive search. Therefore $(G_1, \mathbb{P})$ satisfies the maximum d-separation rule assumption and $(G_2, \mathbb{P})$ satisfies the identifiable sparsest Markov representation assumption.

APPENDIX 3

C·2. *Proof of Lemma* 3 *(b)*

Proof. Suppose that the pair $(G_2, \mathbb{P})$ is a Gaussian linear directed cyclic graphical model with specified edge weights in Figure 16, where the non-specified edge weights can be chosen arbitrarily. Once again to be explicit, we state all d-separation rules entailed by G_1 and G_2. Both graphs entail the following sets of d-separation rules:

(1) For any node $A \in \{X_6, X_7, X_8\}$ and $B \in \{X_1, X_5\}$, A is d-separated from B given $\{X_2, X_3\} \cup C$ for any $C \subset \{X_1, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, Y\} \setminus \{A, B\}$.
(2) For any node $A \in \{X_9, X_{10}, X_{11}\}$ and $B \in \{X_1, X_5\}$, A is d-separated from B given $\{X_3, X_4\} \cup C$ for any $C \subset \{X_1, X_2, X_3, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, Y\} \setminus \{A, B\}$.
(3) For any nodes $A, B \in \{X_6, X_7, X_8\}$, A is d-separated from B given $\{X_2, X_3\} \cup C$ for any $C \subset \{X_1, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, Y\} \setminus \{A, B\}$.
(4) For any nodes $A, B \in \{X_9, X_{10}, X_{11}\}$, A is d-separated from B given $\{X_3, X_4\} \cup C$ for any $C \subset \{X_1, X_2, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, Y\} \setminus \{A, B\}$.

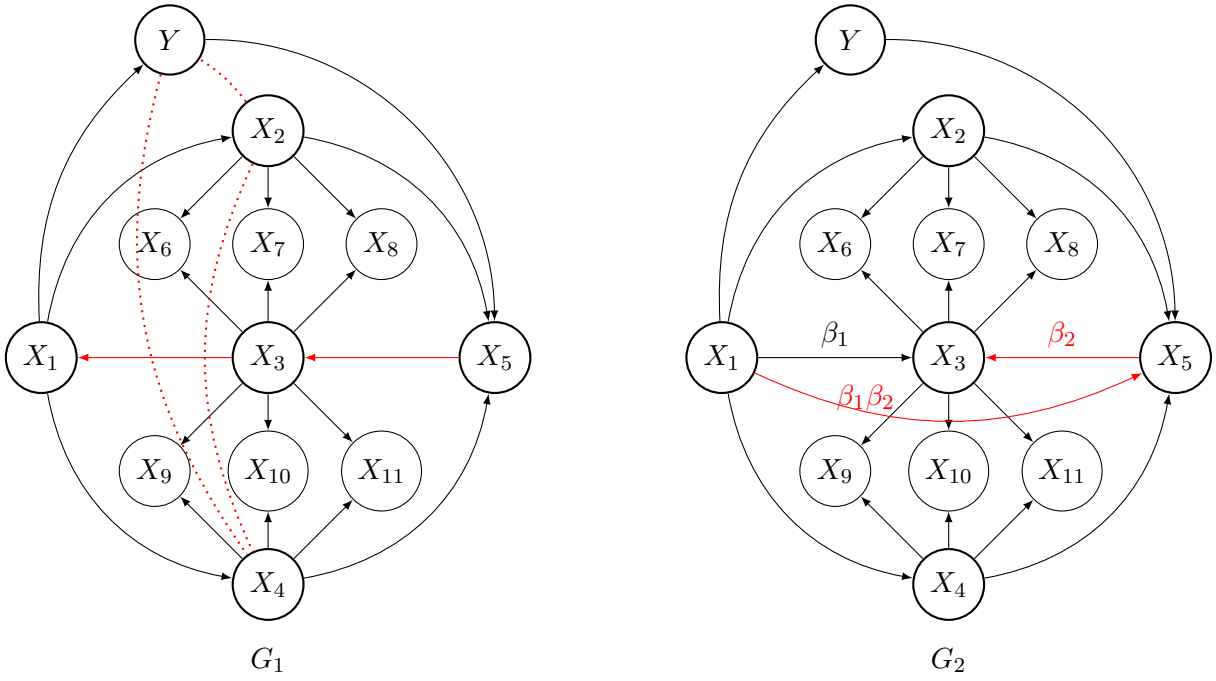


Fig. 16: 12-node examples for Lemma 3.(b)

(5) For any nodes $A \in \{X_6, X_7, X_8\}$ and $B \in \{X_4\}$, A is d-separated from B given $\{X_2, X_3\} \cup C$ for any
$C \subset \{X_1, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, Y\} \setminus \{A, B\}$, or given $\{X_1, X_2, X_5\} \cup D$ for any
$D \subset \{X_4, X_6, X_7, X_8, Y\} \setminus \{A, B\}$

(6) For any nodes $A \in \{X_6, X_7, X_8\}$ and $B \in \{Y\}$, A is d-separated from B given $\{X_2, X_3\} \cup C$ for any
$C \subset \{X_1, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, Y\} \setminus \{A, B\}$, or given $\{X_1, X_2, X_5\} \cup D$ for any
$D \subset \{X_4, X_6, X_7, X_8, , X_9, X_{10}, X_{11}, Y\} \setminus \{A, B\}$

(7) For any nodes $A \in \{X_9, X_{10}, X_{11}\}$ and $B \in \{X_2\}$, A is d-separated from B given $\{X_3, X_4\} \cup C$ for any
$C \subset \{X_1, X_2, X_5, X_9, X_{10}, X_{11}, Y\} \setminus \{A, B\}$, or given $\{X_1, X_4, X_5\} \cup D$ for any
$D \subset \{X_2, X_9, X_{10}, X_{11}, Y\} \setminus \{A, B\}$.

(8) For any nodes $A \in \{X_9, X_{10}, X_{11}\}$ and $B \in \{Y\}$, A is d-separated from B given $\{X_3, X_4\} \cup C$ for any
$C \subset \{X_1, X_2, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, Y\} \setminus \{A, B\}$, or given $\{X_1, X_4, X_5\} \cup D$ for any
$D \subset \{X_2, X_6, X_7, X_8, X_9, X_{10}, X_{11}, Y\} \setminus \{A, B\}$.

(9) For any nodes $A \in \{X_6, X_7, X_8\}$, $B \in \{X_9, X_{10}, X_{11}\}$, A is d-separated from B given $\{X_3\} \cup C \cup D$ for $C \subset \{X_1, X_2, X_4\}$, $C \neq \emptyset$ and $D \subset \{X_1, X_2, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, Y\} \setminus \{A, B, C\}$.

(10) X_2 is d-separated from X_3 given $\{X_1, X_5\} \cup C$ for any $C \subset \{X_1, X_4, X_5, X_9, X_{10}, X_{11}, Y\}$.

(11) X_3 is d-separated from X_4 given $\{X_1, X_5\} \cup C$ for any $C \subset \{X_1, X_4, X_5, X_6, X_7, X_8, Y\}$.

(12) X_3 is d-separated from Y given $\{X_1, X_5\} \cup C$ for any $C \subset \{X_1, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}\}$.

(13) X_2 is d-separated from X_3 given $\{X_1, X_5\} \cup C$ for any $C \subset \{X_4, X_9, X_{10}, X_{11}, Y\}$.

(14) X_4 is d-separated from X_3 given $\{X_1, X_5\} \cup C$ for any $C \subset \{X_2, X_6, X_7, X_8, Y\}$.

(15) Y is d-separated from X_3 given $\{X_1, X_5\} \cup C$ for any $C \subset \{X_2, X_6, X_7, X_8, X_4, X_9, X_{10}, X_{11}\}$.

The set of d-separation rules entailed by G_1 that is not entailed by G_2 is as follows:

(a) X_1 is d-separated from X_5 given $\{X_2, X_3, X_4, Y\} \cup C$ for any $C \subset \{X_6, X_7, X_8, X_9, X_{10}, X_{11}\}$.

Furthermore, the set of d-separation rules entailed by G_2 that is not entailed by G_1 is as follows:

(b) X_2 is d-separated from X_4 given X_1 or $\{X_1, Y\}$.

(c) X_2 is d-separated from Y given X_1 or $\{X_1, X_4\}$.

(d) X_4 is d-separated from Y given X_1 or $\{X_1, X_2\}$. □

It can then be shown that by using the co-efficients specified for G_2 in Figure 16, $CI(\mathbb{P})$ is the union of the conditional indenpendence statements implied by the sets of d-separation rules entailed by both G_1 and G_2. Therefore $(G_1, \mathbb{P})$ and $(G_2, \mathbb{P})$ satisfy the causal Markov condition. It is straightforward to see that G_2 is sparser than G_1 while G_1 entails more d-separation rules than G_2.

Now we prove that $(G_1, \mathbb{P})$ satisfies the maximum d-separation rule assumption and $(G_2, \mathbb{P})$ satisfies the identifiable sparsest Markov representation assumption. First we prove that $(G_2, \mathbb{P})$ satisfies the identifiable sparsest Markov representation assumption. Suppose that $(G_2, \mathbb{P})$ does not satisfy the identifiable sparsest Markov representation assumption. Then there exists a G such that $(G, \mathbb{P})$ satisfies the causal Markov condition and G has the same number of edges as G_2 or fewer edges than G_2. Since the only additional CI statements that are not implied by the d-separation rules of G_2 are $X_1 \perp\!\!\!\perp X_5 \mid \{X_2, X_3, X_4, Y\} \cup C$ for any $C \subset \{X_6, X_7, X_8, X_9, X_{10}, X_{11}\}$ and $(G, \mathbb{P})$ satisfies the causal Markov condition, we can consider two graphs, one with an edge between (X_1, X_5) and another without an edge between (X_1, X_5). We firstly consider a graph without an edge between (X_1, X_5). Since G does not have an edge between (X_1, X_5) and by Lemma 1, G should entail at least one d-separation rule from (a) X_1 is d-separated from X_5 given $\{X_2, X_3, X_4, Y\} \cup C$ for any $C \subset \{X_6, X_7, X_8, X_9, X_{10}, X_{11}\}$. If G does not have an edge between (X_2, X_3), by Lemma 1 G should entail at least one d-separation rule from (10)

X_2 is d-separated from X_3 given $\{X_1, X_5\} \cup C$ for any $C \subset \{X_1, X_4, X_5, X_9, X_{10}, X_{11}, Y\}$. These two sets of d-separation rules can exist only if a cycle $X_1 \to X_2 \to X_5 \to X_3 \to X_1$ or $X_1 \leftarrow X_2 \leftarrow X_5 \leftarrow X_3 \leftarrow X_1$ exists. In the same way, if G does not have edges between (X_3, X_4) and (X_3, Y), there should be cycles which are $X_1 \to A \to X_5 \to X_3 \to X_1$ or $X_1 \leftarrow A \leftarrow X_5 \leftarrow X_3 \leftarrow X_1$ for any $A \in \{X_4, Y\}$ as occurs in G_1. However these cycles create virtual edges between $(X_2, X_4), (X_2, Y)$ or (X_4, Y) as occurs in G_1. Therefore G should have at least 3 edges either real or virtual edges. This leads to a contradiction that G has the same number of edges of G_2 or fewer edges than G_2.

Secondly, we consider a graph G with an edge between (X_1, X_5) such that $(G, \mathbb{P})$ satisfies the causal Markov condition and G has fewer edges than G_2. Note that G_1 entails the maximum number of d-separation rules amongst graphs with an edge between (X_1, X_5) satisfying the causal Markov condition because $CI(\mathbb{P}) \setminus \{X_1 \perp\!\!\!\perp X_5 \mid \{X_2, X_3, X_4, Y\} \cup C$ for any $C \subset \{X_6, X_7, X_8, X_9, X_{10}, X_{11}\}\}$ is exactly matched to the d-separation rules entailed by G_1. This leads to $D_{sep}(G) \subsetneq D_{sep}(G_1)$. By Lemma 2, G cannot contain fewer edges than G_1. However since G_2 has fewer edges than G_1, it is contradictory that G has the same number of edges of G_2 or fewer edges than G_2. Therefore, $(G_2, \mathbb{P})$ satisfies the identifiable sparsest Markov representation assumption.

Now we prove that $(G_1, \mathbb{P})$ satisfies the maximum d-separation rule assumption. Suppose that $(G_1, \mathbb{P})$ fails to satisfy the maximum d-separation rule assumption. Then, there is a graph G such that $(G, \mathbb{P})$ satisfies the causal Markov condition and G entails more d-separation rules than G_1 or as many d-separation rules as G_1. Since $(G, \mathbb{P})$ satisfies the causal Markov condition, in order for G to entail at least the same number of d-separation rules entailed by G_1, G should entail at least one d-separation rule from (b) X_2 is d-separated from X_4 given X_1 or $\{X_1, Y\}$, (c) X_2 is d-separated from Y given X_1 or $\{X_1, X_4\}$ and (d) X_4 is d-separated from Y given X_1 or $\{X_1, X_2\}$. This implies that G does not have an edge between (X_2, X_4), (X_2, Y) or (X_4, Y) by Lemma 1. As we discussed, there is no graph satisfying the causal Markov condition without edges (X_2, X_4), (X_2, Y), (X_4, Y), and (X_1, X_5) unless G has additional edges as occurs in G_1. Note that the graph G entails at most six d-separation rules than G_1 (the total number of d-separation rules of (b), (c), and (d)). However, adding any edge in the graph G generates more than six more d-separation rules because by Lemma 1, G loses an entire set of d-separation rules from the sets (1) to (15) which each contain more than six d-separation rules. This leads to a contradiction that G entails more d-separation rules than G_1 or as many d-separation rules as G_1.

REFERENCES

CHICKERING, D., GEIGER, D. & HECKERMAN, D. (1995). Learning bayesian networks: Search methods and experimental results. In *proceedings of fifth conference on artificial intelligence and statistics*.

FORSTER, M., RASKUTTI, G., STERN, R. & WEINBERGER, N. (2015). The frugal inference of causal relations. *British Journal for the Philosophy of Science* .

GLYMOUR, C., SCHEINES, R., SPIRTES, P. & KELLY, K. (1987). Discovering causal structure: Artificial intelligence. *Philosophy of science, and Statistical Modeling* , 205–212.

KALISCH, M., MÄCHLER, M., COLOMBO, D., MAATHUIS, M. H. & BÜHLMANN, P. (2012). Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software* **47**, 1–26.

LAURITZEN, S. L. (1996). *Graphical models*. Clarendon Press.

LAURITZEN, S. L., DAWID, A. P., LARSEN, B. N. & LEIMER, H.-G. (1990). Independence properties of directed markov fields. *Networks* **20**, 491–505.

PEARL, J. (2003). Causality: models, reasoning and inference. *Economet. Theor* **19**, 675–685.

PEARL, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

RASKUTTI, G. & UHLER, C. (2013). Learning directed acyclic graphs based on sparsest permutations. *arXiv preprint arXiv:1307.0366* .

RICHARDSON, T. (1994). Properties of cyclic graphical models. *MS ThesisCarnegie Mellon Univ* .

RICHARDSON, T. (1996a). A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc.

RICHARDSON, T. (1996b). A polynomial-time algorithm for deciding markov equivalence of directed cyclic graphical models. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc.

SPIRTES, P. (1995). Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc.

SPIRTES, P. & GLYMOUR, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review* **9**, 62–72.

SPIRTES, P., GLYMOUR, C. N. & SCHEINES, R. (2000). *Causation, prediction, and search*. MIT press.

UHLER, C., RASKUTTI, G., BÜHLMANN, P., YU, B. et al. (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics* **41**, 436–463.

ZHANG, J. (2012). A comparison of three occam's razors for markovian causal models. *The British Journal for the Philosophy of Science* , axs005.

[*Received April* 2012. *Revised September* 2012]