Markov Blanket Discovery in AMP Chain Graphs

Mohammad Ali Javidian JAVIDIAN@EMAIL.SC.EDU

Marco Valtorta MGV@CSE.SC.EDU

Pooyan Jamshidi PJAMSHID@CSE.SC.EDU

Department of Computer Science & Engineering, University of South Carolina, Columbia, SC, 29201, USA.

Abstract

This paper provides a graphical characterization of Markov blankets in Andersson–Madigan–Perlman chain graphs (AMP CGs). The characterization is different from the well-known one for Bayesian networks and generalizes it. We prove that the Grow-Shrink algorithm, the IAMB algorithm, and its variants are still correct for Markov blanket discovery in AMP CGs under the same assumptions as for Bayesian networks. We provide a sound and scalable constraint-based framework for learning the structure of AMP CGs from faithful causally sufficient data and prove its correctness when the Markov blanket discovery algorithms in this paper are used. On a large scale experimentation, we show that the proposed algorithms greatly improve on (stable) PC-like (Peña, 2012; Javidian et al., 2020a) in all comparisons. Our proposed algorithms compare positively/competitively against the state-of-the-art LCD-AMP algorithm (Javidian et al., 2020a), depending on the algorithm that is used for Markov blanket discovery. Our proposed algorithms make a broad range of inference/learning problems computationally tractable and more reliable because they exploit locality.

Keywords: AMP chain graph; Markov blanket, conditional independence, decomposition, graphical model, Markov equivalent, structural learning.

1. Introduction

Probabilistic graphical models (PGMs) use graphs, either undirected, directed, or mixed, to represent possible dependencies among the variables of a multivariate probability distribution. PGMs, such as Bayesian networks (BNs) and Markov networks, are now widely accepted as a powerful and mature framework for reasoning and decision making under uncertainty in knowledge-based systems. With the increase of their popularity, the range of graphical models being investigated and used has also expanded. Several types of graphs with different conditional independence interpretations - also known as Markov properties - have been proposed and used in graphical models.

The graphical structure of a Bayesian network has the form of a directed acyclic graph (DAG), which has the advantage of supporting an interpretation of the graph in terms of cause-effect relationships. However, a limitation is that only asymmetric relationships, such as cause and effect relationships, can be modeled between variables in a DAG. Chain graphs, which admit both directed and undirected edges, can be used to overcome this limitation. Today there exist three main different interpretations of chain graphs in the literature: the Lauritzen-Wermuth-Frydenberg (LWF), the Andersson-Madigan-Perlman (AMP), and the multivariate regression (MVR) interpretations.

This paper deals with chain graphs under the alternative Andersson-Madigan-Perlman (AMP) interpretation (Andersson et al., 1996, 2001). AMP CGs are useful when we have a set of variables for which the internal relations has no causal ordering, so the relations should be modelled as a Markov network, but also a second set of variables which can be seen as causes for some of these

variables in the first set. The internal structure of the first set of variables can then be modelled as a Markov network, creating a chain component in an AMP CG, and the causes as parents of some of the variables in the chain component. Note that for AMP CGs the parents only affect the direct children in the chain component, not all the nodes in the chain component as in the case of LWF CGs. An example in medicine (Sonntag and Peña, 2015) when such a model might be appropriate is when we are modelling pain levels on different areas on the body of a patient. The pain levels can then be seen as correlated "geographically" over the body, and hence be modelled as a Markov network. Certain other factors do, however, exist that alters the pain levels locally at some of these areas, such as the type of body part the area is located on or if local anaesthetic has been administered in that area and so on. These outside factors can then be modelled as parents affecting the pain levels locally.

From the *causality* point of view, every AMP CG is Markov equivalent to some DAG with error and selection nodes under marginalization of the error (deterministic) nodes and conditioning of the selection nodes (Peña, 2014). In fact, AMP CGs have been shown to be suitable for representing causal linear models with additive Gaussian noise (Peña, 2016). AMP chain graphs widely studied in different areas from applications in biology (Sonntag and Peña, 2015), to more advanced theoretical investigations (Levitz et al., 2001; Roverato, 2005; Roverato and Rocca, 2006; Drton, 2009; Studený et al., 2009; Peña, 2014, 2015; Sonntag and Peña, 2015; Peña and Gómez-Olmedo, 2016; Peña, 2018a,b).

One important and challenging aspect of PGMs is the possibility of learning the structure of models directly from sampled data. Three *constraint-based* learning algorithms, that use a statistical analysis to test the presence of a conditional independency, exist for learning AMP chain graphs: (1) the (stable) PC-like algorithm (Peña, 2012; Peña and Gómez-Olmedo, 2016; Javidian et al., 2020a), (2) the answer set programming (ASP) algorithm (Peña, 2016), and a decomposition-based algorithm, called AMP-LCD (Javidian et al., 2020a).

In a DAG G with node set V, each local distribution depends only on a single node $v \in V$ and on its parents (i.e., the nodes $u \neq v$ such that $u \longrightarrow v$, here denoted $pa(v)$). Then the overall joint density is simply $p(x) = \Pi_{v \in V}^{n} p(x_v | x_{pa(v)})$. The key advantage of the decomposition in this equation is to make *local computations* possible for most tasks, using just a few variables at a time regardless of the magnitude of $|V| = n$. In Bayesian networks, the concept that enables us to take advantage of local computation is *Markov blanket*. The Markov blanket (*Markov boundary* in Pearl's terminology) of each node v, defined as the set $\mathbf{Mb}(v)$ of nodes that separates v from all other nodes $V \setminus \{v, \mathbf{Mb}(v)\}$. Markov blankets can be used for variable selection for classification, for causal discovery, and for Bayesian network learning (Tsamardinos et al., 2003a).

Markov blanket discovery has attracted a lot of attention in the context of Bayesian network structure learning (see section 2). It is surprising, however, how little attention (if any) it has attracted in the context of learning AMP chain graphs. In this paper, we focus on addressing the problem of Markov blanket discovery for structure learning of AMP chain graphs. For this purpose, we extend the concept of Markov blankets to AMP CGs. We prove that Grow-Shrink Markov Blanket (GSMB) (Margaritis and Thrun, 1999), IAMB, and its variants (Tsamardinos et al., 2003a; Yaramakala and Margaritis, 2005) (that are mainly designed for Markov blanket recovery in Bayesian networks) are still correct for Markov blanket discovery in AMP CGs under the faithfulness and causal sufficiency assumptions.

Since constraint-based learning algorithms are sensitive to *error propagation* (Triantafillou et al., 2014), and an erroneous identification of an edge can propagate through the network and lead to er-

roneous edge identifications or conflicting orientations *even in seemingly unrelated parts of the network*, the learned chain graph model will be unreliable. In order to address the problem of reliable structure learning, we present a generic approach (i.e., the algorithm is independent of any particular search strategy for Markov blanket discovery) based on Markov blanket recovery to learn the structure of AMP CGs from a causally sufficient and faithful data. This algorithm first learns the Markov blanket of each node. This preliminary step greatly simplifies the identification of neighbours. This in turn results in a significant reduction in the number of conditional independence tests, and therefore of the overall computational complexity of the learning algorithm. In order to show the effectiveness of this approach, the resulting algorithms are contrasted against (stable) PC-like (Peña, 2012; Javidian et al., 2020a) and LCD-AMP (Javidian et al., 2020a) on simulated data. We report experiments showing that our proposed generic algorithm (via 5 different instantiations) consistently outperform (stable) PC-like, while they provide better/competitive performance against the LCD-AMP algorithm in our Gaussian experimental settings, depending on the approach that is used for Markov blanket discovery. Our proposed approach has an advantage over LCD-AMP because local structural learning in the form of Markov blanket is a theoretically well-motivated and empirically robust learning framework that can serve as a powerful tool in classification and causal discovery (Aliferis et al., 2010). We also note that Markov blankets are useful in their own right, for example in sensor validation and fault analysis (Ibargüengoytla et al., 1996). Code for reproducing our results and its corresponding user manual is available at `https://github.com/majavid/AMPCGs-PGM2020`. Our main theoretical and empirical contributions are as follows:
(1) We extend the concept of Markov blankets to AMP CGs and we prove what variables make up the Markov blanket of a target variable in an AMP CG (Section 4).
(2) We theoretically prove that the Grow-Shrink, IAMB algorithm and its variants are still sound for Markov blanket discovery in AMP chain graphs under the faithfulness and causal sufficiency assumptions (Section 4).
(3) We propose a generic algorithm for structure learning of AMP chain graphs, called *MbAMP*, based on the proposed Markov blanket recovery algorithms in Section 4, and we prove its correctness theoretically (Section 5).
(4) We evaluate the performance of 6 instantiations of the proposed generic algorithm MbAMP with 6 different Markov blanket recovery algorithms on synthetic Gaussian data, and we show that the resulting chain graphs are more accurate and reliable than the state-of-the-art algorithms in the literature (Section 6).

2. Related Work

Markov Blanket Discovery in BNs with Causal Sufficiency Assumption. Margaritis and Thrun (Margaritis and Thrun, 1999) presented the first provably correct algorithm, called Grow-Shrink Markov Blanket (GSMB), that discovers the Markov blanket of a variable from a faithful data under the causal sufficiency assumption. Variants of GSMB were proposed to improve speed and reliability such as the Incremental Association Markov Blanket (IAMB) and its variants (Tsamardinos et al., 2003a), Fast-IAMB (Yaramakala and Margaritis, 2005), and IAMB with false discovery rate control (IAMB-FDR) (Peña, 2008). Since in discrete data the sample size required for high-confidence statistical tests of conditional independence in GSMB and IAMB algorithms grows exponentially in the size of the Markov blanket, several sample-efficient algorithms e.g., HITON-MB (Aliferis et al.,

2010) and Max–Min Markov Blanket (MMMB) (Tsamardinos et al., 2006) were proposed to overcome the data inefficiency of GSMB and IAMB algorithms. One can find alternative computational methods for Markov blanket discovery that were developed in the past two decades in (Peña et al., 2007; Liu and Liu, 2016; Ling et al., 2019), among others.

Markov Blanket Discovery in BNs without Causal Sufficiency Assumption. Gao and Ji (Gao and Ji, 2016) proposed the latent Markov blanket learning with constrained structure EM algorithm (LMB-CSEM) to discover the Markov blankets in BNs in the presence of unmeasured confounders. However, LMB-CSEM was proposed to find the Markov blankets in a DAG and provides no theoretical guarantees for finding all possible unmeasured confounders in the Markov blanket of the target variable. Recently, Yu et. al. (Yu et al., 2018) proposed a new algorithm, called M3B, to mine Markov blankets in BNs in the presence of unmeasured confounders.

Markov Blanket Discovery in LWF CGs. Recently, Javidian et. al. (Javidian et al., 2020b) extended the concept of Markov blankets to LWF CGs and proved what variables make up the Markov blanket of a target variable in an LWF CG. Authors theoretically proved that the Grow-Shrink, IAMB algorithm and its variants are still sound for Markov blanket discovery in LWF chain graphs under the faithfulness and causal sufficiency assumptions (Javidian et al., 2020b, Section 4). In addition, they presented a new algorithm, called MBC-FSP, for learning Markov blankets in LWF chain graphs (Javidian et al., 2020b, Section 4).

In this paper, we extend the concept of Markov blankets to AMP CGs, which is different from Markov blankets defined in DAGs under the causal sufficiency assumption and also is different from Markov blankets defined in maximal ancestral graphs without assuming causal sufficiency.

3. Notation and Key Definitions

In this paper, we consider graphs containing both directed ($\rightarrow$) and undirected ($-$) edges and largely use the terminology of (Andersson et al., 2001), where the reader can also find further details. Below we briefly list some of the central concepts used in this paper.

If $A \subseteq V$ is a subset of the vertex set in a graph $G = (V, E)$, the *induced subgraph* $G_A = (A, E_A)$ is a graph in which the edge set $E_A = E \cap (A \times A)$ is obtained from G by keeping edges with both endpoints in A. If there is an arrow from a pointing towards b, a is said to be a *parent* of b. The set of parents of b is denoted as $pa(b)$. If there is an undirected edge between a and b, a and b are said to be *adjacent* or *neighbors*. The set of neighbors of a vertex a is denoted as $ne(a)$. The expressions $pa(A)$ and $ne(A)$ denote the collection of parents and neighbors of vertices in A that are not themselves elements of A. The *boundary* of a subset of vertices A, $bd(A)$, is the set of vertices in $V \setminus A$ that are parents or neighbors to vertices in A. The *closure* of A is $cl(A) = bd(A) \cup A$.

A *path* of length n from a to b is a sequence $a = a_0, \ldots, a_n = b$ of distinct vertices such that $(a_i, a_{i+1}) \in E$, for all $i = 1, \ldots, n$. A vertex α is said to be an *ancestor* of a vertex β if there is a directed path $\alpha \rightarrow \cdots \rightarrow \beta$ from α to β. We define the *smallest ancestral set* containing A as $An(A) := an(A) \cup A$. A *partially directed cycle* (or semi-directed cycle) in a graph G is a sequence of n distinct vertices $v_1, v_2, \ldots, v_n (n \geq 3)$, and $v_{n+1} \equiv v_1$, such that (a) for all $i (1 \leq i \leq n)$ either $v_i - v_{i+1}$ or $v_i \rightarrow v_{i+1}$, and (b) there exists a $j (1 \leq j \leq n)$ such that $v_j \rightarrow v_{j+1}$.

An *AMP chain graph* is a graph in which there are no partially directed cycles. The *chain components* $\mathcal{T}$ of a chain graph are the connected components of the undirected graph obtained by removing all directed edges from the chain graph. We define the *smallest coherent set* containing A

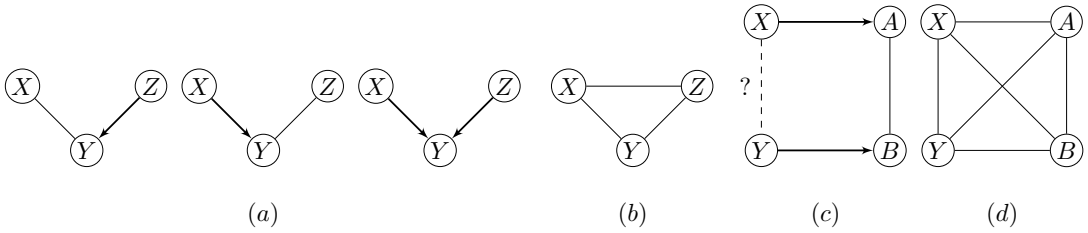

Figure 1: (a) Triplexes, and (b) the corresponding augmented triplex. (c) the four configurations that define the bi-flag, and (d) the corresponding augmented bi-flag. The "?" indicates that either $X - Y \in G$, $X \to Y \in G$, $Y \to X \in G$, or X and Y are not adjacent in G.

as $Co(A) := \cup_\tau \{\tau \in \mathcal{T} | \tau \cap A \neq \emptyset\}$. Let $\overline{G}$ be obtained by deleting all directed edges of G; for $A \subseteq V$ the *extended subgraph* G[A] is defined by $G[A] := G_{An(A)} \cup \overline{G}_{Co(An(A))}$.

A triple of vertices $\{X, Y, Z\}$ is said to form a *flag* in CG if the induced subgraph $CG_{X \cup Y \cup Z}$ is $X \to Y - Z$ or $X - Y \leftarrow Z$. A triple of vertices $\{X, Y, Z\}$ is said to form a *triplex* in CG if the induced subgraph $CG_{X \cup Y \cup Z}$ is either $X \to Y - Z$, $X \to Y \leftarrow Z$, or $X - Y \leftarrow Z$. A triplex is *augmented* by adding the $X - Z$ edge. A set of four vertices $\{X, A, B, Y\}$ is said to form a *bi-flag* if the edges $X \to A$, $Y \to B$, and $A - B$ are present in the induced subgraph over $\{X, A, B, Y\}$. A bi-flag is augmented by adding the edge $X - Y$. We say that X is a *flag-spouse* of Y and vice versa, and that $fsp(X) = \{Y \in V | \exists$ a triplex or bi-flag between X and $Y\}$. A *minimal complex* (or simply a complex) in a chain graph is an induced subgraph of the form $a \to v_1 - \cdots - v_r \leftarrow b$. The *augmented CG* G^a is the undirected graph formed by augmenting all triplexes and bi-flags in CG and replacing all directed edges with undirected edges (see Figure 1). The *skeleton* (underlying graph) of a CG G is obtained from G by changing all directed edges of G into undirected edges. Vertex Y is an *unshielded collider* (or V-structure) in a DAG G if G contains the induced subsgraph $U \to Y \leftarrow V$.

Definition 1 (Global Markov property for AMP CGs) *For any triple (A, B, S) of disjoint subsets of V such that S separates A from B in $(G[A \cup B \cup S])^a$, in the augmented graph of the extended subgraph of $A \cup B \cup S$, we have $A \perp\!\!\!\perp B | S$ (or $\langle A, B | S \rangle$) i.e., A is independent of B given S.*

An equivalent pathwise separation criterion that identifies all valid conditional independencies under the AMP Markov property was introduced in (Levitz et al., 2001):

Definition 2 (The pathwise p-separation criterion for AMP chain graphs) *A node B in a chain ρ in an AMP CG G is called a* triplex node *in ρ if $A \to B \leftarrow C$, $A \to B - C$, or $A - B \leftarrow C$ is a subchain of ρ. Moreover, ρ is said to be Z-*open *with $Z \subseteq V$ when (i) every triplex node in ρ is in $An(Z)$, and (ii) every non-triplex node B in ρ is outside Z, unless $A - B - C$ is a subchain of ρ and $pa_G(B) \setminus Z \neq \varnothing$.*

Let $X, Y \neq \varnothing$ and Z (may be empty) denote three disjoint subsets of V. When there is no Z-open chain in an AMP CG G between a node in X and a node in Y, we say that X is separated from Y given Z in G and denote it as $X \perp\!\!\!\perp Y | Z$.

Theorem 4.1 in (Levitz et al., 2001) establishes the equivalence of the *p*-separation criterion and the augmentation criterion occurring in the AMP global Markov property for CGs.

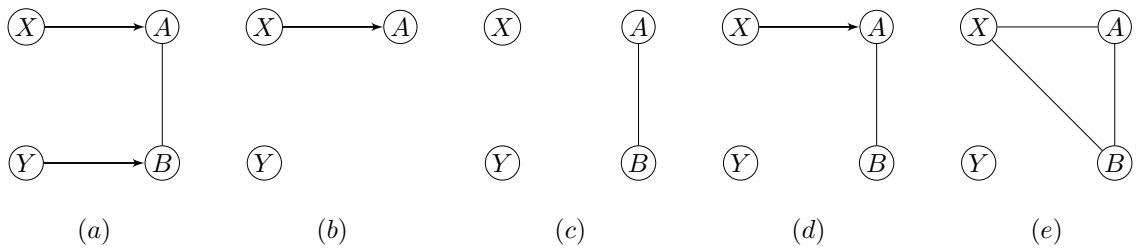

Figure 2: (a) The AMP CG G, (b) $An(X \cup Y \cup A)$, (c) the undirected edges in $Co(An(X \cup Y \cup A))$, (d) $G[X \cup Y \cup A]$, and (e) $(G[X \cup Y \cup A])^a$.

Example 1 *Consider the AMP CG G in Figure 2(a). The global Markov property of AMP chain graphs implies that $X \perp\!\!\!\perp Y|A$ (see Figure 2). There is no A-open chain in the AMP CG G between X and Y because the only chain between X and Y i.e., $X \rightarrow A - B \leftarrow Y$ is blocked at B (B is a triplex node in the chain and $B \notin An(A)$).*

We say that two AMP CGs G and H are *Markov equivalent* or that they are in the same *Markov equivalence class* if they induce the same conditional independence restrictions. Two AMP chain graphs G and H are Markov equivalent if and only if they have the same skeletons and the same triplexes (Andersson et al., 2001). We say that AMP chain graphs G and H belong to the same *strong Markov equivalent class* iff G and H are Markov equivalent and contain the same flags. An AMP CG G^* is said to be the *AMP essential graph* of its Markov equivalence class iff for every directed edge $A \rightarrow B$ that exists in G^* there exists no AMP CG H s.t. G^* and H are Markov equivalent and $A \leftarrow B$ is in H. An AMP CG G^* is said to be the *largest deflagged graph* of its Markov equivalence class iff there exists no other AMP CG H s.t. G^* and H are Markov equivalent and either H contains fewer flags than G^* or G^* and H belong to the same strong Markov equivalence class but H contains more undirected edges. Any largest deflagged graph or AMP essential graph are AMP CGs and both of these have been proven to be unique for the Markov equivalence class they represent (Roverato and Rocca, 2006; Andersson and Perlman, 2006).

The *Markov condition* is said to hold for a DAG $G = (V, E)$ and a probability distribution $P(V)$ if every variable is statistically independent of its graphical non-descendants conditional on its graphical parents in P. Pairs $\langle G, P \rangle$ that satisfy the Markov condition satisfy the implication: $\forall X, Y \in V, \forall Z \subseteq V \setminus \{X, Y\} : (X \perp\!\!\!\perp_d Y|Z \implies X \perp\!\!\!\perp_p Y|Z)$. The *faithfulness condition* states that the only conditional independencies to hold are those specified by the Markov condition: $\forall X, Y \in V, \forall Z \subseteq V \setminus \{X, Y\} : (X \not\perp\!\!\!\perp_d Y|Z \implies X \not\perp\!\!\!\perp_p Y|Z)$. In practice, the Markov condition is used by constraint-based algorithms, which use a statistical analysis to test the presence of a conditional independency, to perform conditional independence tests on the data and build the graph accordingly, and faithfulness is assumed to prove that the graph is correct.

Let a Bayesian network $G = (V, E, P)$ be given. Then, V is a set of random variables, (V, E) is a DAG, and P is a joint probability distribution over V. Let $T \in V$. Then, the *Markov blanket* $\mathbf{Mb}(T)$ is the minimal set of variables such that T is conditionally independent of all the other variables given $\mathbf{Mb}(T)$ for any choice of numerical parameters. That is, $T \perp\!\!\!\perp_p V \setminus (\mathbf{Mb}(T) \cup \{T\})|\mathbf{Mb}(T)$ for any distribution that respects the conditional independencies. Suppose $\langle G, P \rangle$ satisfies the Markov condition. Then for each variable T, the set of all parents of T, children of T, and spouses of T is a Markov blanket of T. If $\langle G, P \rangle$ also satisfies the faithfulness condition, then

for each variable T, the set of all parents of T, children of T, and spouses of T is the unique Markov blanket of T (Tsamardinos et al., 2003a).

4. Markov Blanket Discovery in AMP Chain Graphs

The Markov blanket of the target variable T in an AMP chain graph probabilistically shields T from the rest of the variables. In this section, we first prove which variables make up the Markov blanket of a target variable T in an AMP CG. Then, we prove that GSMB, IAMB and its variants are sound for Markov blanket discovery in AMP CGs under the faithfulness and causal sufficiency assumptions.

Theorem 3 *Suppose (G, P) satisfies the global Markov property for AMP CGs, where $G = (V, E)$ is an AMP chain graph and P is a joint probability distribution over V. Then for each variable T, the set of all parents of T, children of T, neighbors of T, and flag-spouses of T is the Markov blanket of T. Formally, **Mb**$(T) = bd(T) \cup ch(T) \cup fsp(T)$.*

Proof First, we observe that if G is a DAG, then $ne(T) = \varnothing$ and since the only flag-spouses of T are $sp(T)$, $\mathbf{Mb}(T)$ reduces to the set of all parents, children, and spouses of T, as it should. We now prove the general case by distinguishing 3 cases. Let S be a generic variable (node) other than T.
(1) S is directly linked to T in G. Clearly, no set can p-separate T from S in G. So, $S \in \mathbf{Mb}(T)$. In other words, $bd(T) \cup ch(T) \subseteq \mathbf{Mb}(T)$.
(2) S is not directly linked to T, $S \in fsp(T)$, and there is a triplex or bi-flag between T and S. We consider the four following sub-cases: (i) the triplex between T and S is of the form $T \longrightarrow X \longleftarrow S$, (ii) the triplex between T and S is of the form $T \longrightarrow X \longleftarrow S$, (iii) the triplex between T and S is of the form $T \longrightarrow X \longrightarrow S$, and (iv) the bi-flag between T and S is of the form $T \longrightarrow X \longrightarrow Y \longleftarrow S$. As shown in case (1), all children and neighbours of T are in $\mathbf{Mb}(T)$. So, in the augmented graph of the extended subgraph of $T \cup S \cup X$, where $X \in ch(T)$ or $X \in ne(T)$, T and S are linked directly by adding an undirected edge. This means that no set that contains $bd(T) \cup ch(T)$ can p-separate T from S in G in sub-cases (i), (ii), and (iii). In sub-case (iv), since $T \longrightarrow X \longrightarrow Y$ is a triplex, sub-case (iii) implies that $Y \in \mathbf{Mb}(T)$. So, in the augmented graph of the extended subgraph of $T \cup S \cup X \cup Y$, T and S are linked directly by adding an undirected edge. This means that no set that contains $bd(T) \cup ch(T) \cup sp(T)$ can p-separate T from S in G in sub-case (iv).
(3) S is not directly linked to T and there is a path (of length > 1) between T and S, and $S \notin fsp(T)$. From the definition of the augmented graph of the extended subgraphs, it is straightforward to see that S is not linked to T directly by adding an undirected edge in the augmented graph of the extended subgraph of $T \cup S \cup CMB$, where $CMB := bd(T) \cup ch(T) \cup fsp(T)$. This means that $T \perp\!\!\!\perp S | CMB$. From the definition of the Markov condition it follows that every p-separation relation in G implies conditional independence in every joint probability distribution P that satisfies the Markov condition for G. Thus, we have $T \perp\!\!\!\perp_p S | CMB$ in P for every variable $S \in V \setminus CMB \setminus \{T\}$, from which it follows that CMB is a Markov blanket of T.
(1), (2), and (3) together imply that $\mathbf{Mb}(T) = bd(T) \cup ch(T) \cup fsp(T)$. ■

Example 2 *Suppose (G, P) satisfies the Markov condition where G is the AMP chain graph in Figure 3. Then due to Theorem 3, the set of $\{C, F, G, H, K, L\}$ is the Markov blanket of T. Because*

$pa(T) = \{C, G\}, ch(T) = \{K\}, ne(T) = \{F\}, fsp(T) = \{E, L, J, S\}$. Note that if only T's adjacents are instantiated, then T is not p-separated from L, J, and S in G.

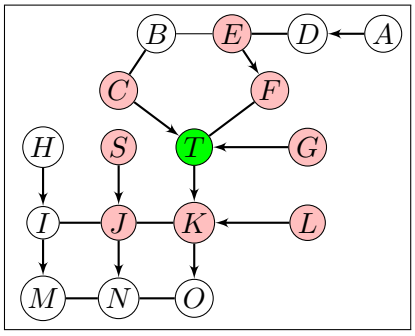

Figure 3: The AMP CG G. The Markov blanket of the target node T is $\mathbf{Mb}(T) = \{C, F, E, G, J, K, L, S\}$.

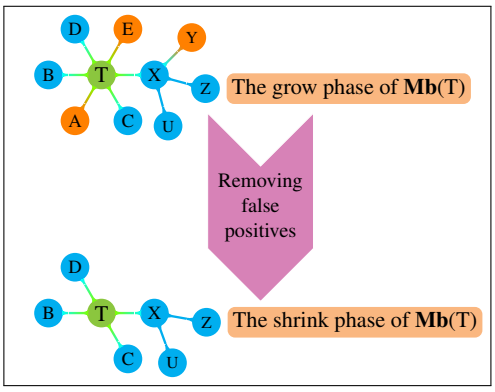

Figure 4: The procedure of Markov blanket recovery in the Grow-Shrink based algorithms.

Theorem 4 *Given the Markov assumption, the faithfulness assumption, a graphical model represented by an AMP CG, and i.i.d. sampling, in the large sample limit, the Markov blanket recovery algorithms GS (Margaritis and Thrun, 1999), IAMB (Tsamardinos et al., 2003a), fastIAMB (Yaramakala and Margaritis, 2005), Interleaved Incremental Association (interIAMB) (Tsamardinos et al., 2003a), and fdrIAMB (Peña, 2008) correctly identify all Markov blankets for each variable. (Note that* Causal Sufficiency *is assumed i.e., all causes of more than one variable are observed.)*

Proof [Sketch of proof] If a variable belongs to $\mathbf{Mb}(T)$, then it will be admitted in the first step (Grow phase) at some point, since it will be dependent on T given the candidate set of $\mathbf{Mb}(T)$. This holds because of the faithfulness and because the set $\mathbf{Mb}(T)$ is the minimal set with that property. If $X \notin \mathbf{Mb}(T)$, then conditioned on $\mathbf{Mb}(T) \setminus \{X\}$, it will be independent of T and thus will be removed from the candidate set of $\mathbf{Mb}(T)$ in the second phase (Shrink phase) because the Markov condition entails that independencies in the distribution are represented in the graph. Since the faithfulness condition entails dependencies in the distribution from the graph, we never remove any variable X from the candidate set of $\mathbf{Mb}(T)$ if $X \in \mathbf{Mb}(T)$. Using this argument inductively we will end up with the $\mathbf{Mb}(T)$. ■

The algorithms listed in Theorem 4 are structurally similar to the standard Markov blanket discovery algorithms and follow the same two-phase *grow-shrink* structure as shown in the Figure 4. An estimate of the $\mathbf{Mb}(T)$ is kept in the set CMB. In the grow phase all variables that belong in $\mathbf{Mb}(T)$ and possibly more (*false positives*) enter CMB while in the shrink phase the false positives are identified and removed so that $CMB = \mathbf{Mb}(T)$ in the end.

5. Learning AMP Chain Graphs via Markov Blankets

Any sound algorithm for learning Markov blankets of AMP CGs can be employed and extended to a full AMP CG learning algorithm, as originally suggested in (Margaritis and Thrun, 1999) for Grow-Shrink Markov blanket algorithm (for Bayesian networks). Thanks to the proposed Markov blanket discovery algorithms listed in Theorem 4, we can now present a generic algorithm for learning AMP CGs. Algorithm 1 lists pseudocode for the three main phases of this approach.

Algorithm 1: MbAMP: Learning AMP CGs via Markov blanket discovery

Input: a set V of nodes and a probability distribution p faithful to an unknown AMP chain graph $G = (V, E)$.
Output: An AMP CG H that is triplex equivalent to G.

/* **Phase 1: Learning Markov blankets** */
For each variable $X_i \in V$, learn its Markov blanket $\mathbf{Mb}(X_i)$;
Check whether the Markov blankets are symmetric, e.g., $X_i \in \mathbf{Mb}(X_j) \leftrightarrow X_j \in \mathbf{Mb}(X_i)$. Assume that nodes for whom symmetry does not hold are false positives and drop them from each other's Markov blankets;
Set $\mathbf{Sepset}(X_i, X_j) = \mathbf{Sepset}(X_j, X_i)$ to the smallest of $\mathbf{Mb}(X_i)$ and $\mathbf{Mb}(X_j)$ if $X_i \notin \mathbf{Mb}(X_j)$ and $X_j \notin \mathbf{Mb}(X_i)$;
/* **Phase 2: Skeleton Recovery** */
Construct the undirected graph $H = (V, E)$, where $E = \{X_i \text{ —— } X_j | X_j \in \mathbf{Mb}(X_i) \text{ and } X_i \in \mathbf{Mb}(X_j)\}$;
for $i \leftarrow 0$ **to** $|V_H| - 2$ **do**
  while *possible* **do**
    Select any ordered pair of nodes u and v in H such that $u \in ad_H(v)$ and $|[ad_H(u) \cup ad_H(ad_H(u))] \setminus \{u, v\}| \geq i$, using order($V$);
    /* $ad_H(x) := \{y \in V | x \longrightarrow y, y \longrightarrow x, \text{ or } x \text{ —— } y\}$ */
    if *there exists* $S \subseteq ([ad_H(u) \cup ad_H(ad_H(u))] \setminus \{u, v\})$ *s.t.* $|S| = i$ *and* $u \perp\!\!\!\perp_p v | S$ **then**
      Set $\mathbf{Sepset}(u, v) = \mathbf{Sepset}(v, u) = S$;
      Remove the edge u —— v from H;
    end
  end
end
/* **Phase 3: Orientation phase:** */
while *possible* **do**
  Apply the rules R1-R4 in the Figure 5 to H.
  /* A block is represented by a perpendicular line at the edge end such as in ⟝ or ⟞⟝, and it means that the edge cannot be a directed edge pointing in the direction of the block. Note that ⟞⟝ means that the edge must be undirected. The ends of some of the edges in the rules are labeled with a circle such as in o— or o–o. The circle represents an unspecified end, i.e. a block or nothing. */
end
Replace every edge ⟝ (⟞⟝) in H with $\longrightarrow$ (——);

Phase 1: Learning Markov blankets. This phase consists of learning the Markov blanket of each variable with feature selection to reduce the number of candidate structures early on. Any algorithm in Theorem 4 can be plugged in Step 1. Once all Markov blankets have been learned,

they are checked for consistency (Step 2) using their symmetry; by definition $X_i \in \mathbf{Mb}(X_j) \leftrightarrow X_j \in \mathbf{Mb}(X_i)$. Asymmetries are corrected by treating them as false positives and removing those variables from each other's Markov blankets. At the end of this phase, separator sets of X and Y set to the smallest of $\mathbf{Mb}(X)$ and $\mathbf{Mb}(Y)$ if $X \notin \mathbf{Mb}(Y)$ and $Y \notin \mathbf{Mb}(X)$.

Phase 2: Skeleton Recovery. First, we construct the augmented graph of the extended graph of the AMP CG G that is an undirected graph in which each node of the original G is now connected to its Markov blanket (line 4 of Algorithm 1). Lines 5-13 learn the *skeleton* of the AMP CG by removing the spurious edges. In fact, we remove the added undirected edge(s) between each variable T and its flag-spouses due to the fact that $fsp(T) \subseteq \mathbf{Mb}(T)$. Separation sets are updated correspondingly.

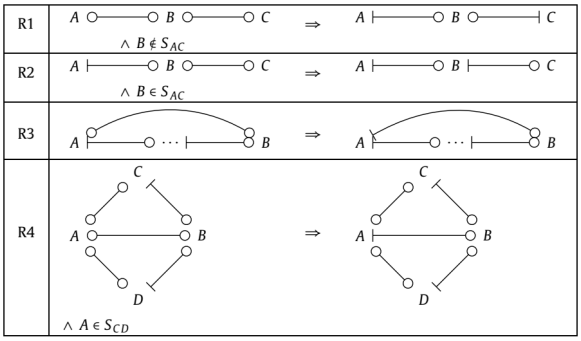

Figure 5: Rules R1-R4 (Peña, 2012)

Phase 3: Orientation Recovery. We use the resulting undirected graph obtained in the previous step to orient undirected edges via rules R1-R4 in Figure 5 (Peña, 2012; Peña and Gómez-Olmedo, 2016). This process is formally described in Algorithm 1.

Remark 5 *One can apply Algorithm 3 in (Roverato and Rocca, 2006) to to the resulting chain graph of Algorithm 1 to obtain the largest deflagged graph. Also, one can apply Algorithm 1 in (Sonntag and Peña, 2015) to the resulting chain graph of Algorithm 1 to obtain the AMP essential graph.*

Computational Complexity Analysis of Algorithm 1 Assume that the "learning Markov blankets" phase uses the grow-shrink (GSMB) approach and $n = |V|$, $m = |E|$, where $G = (V, E)$ is the true AMP CG. Since the Markov blanket algorithm involves $O(n)$ conditional independence (CI) tests, Phase 1 (learning Markov blankets) involves $O(n^2)$ tests. If $b = max_X \mathbf{Mb}(X)$, the skeleton recovery (line 5-13) does $O(nb2^b)$ CI tests. In the worst case, i.e. when $b = O(n)$ and $m = O(n^2)$ i.e. the original graph is dense, the total complexity for these 2 phases becomes $O(n^2 + nb2^b)$ or $O(n^2 2^n)$. Under the assumption that b is bounded by a constant (the sparseness assumption), the complexity of Phase 1 and 2 together is $O(n^2)$ in the number of CI tests. Since the orientation phase (phase 3) does not generate new CI tests, the total number of CI tests for the entire algorithm is therefore $O(n^2 + nb2^b)$. Under the assumption that b is bounded by a constant, this algorithm is $O(n^2)$ in the number of CI tests.

6. Experimental Evaluation

We performed a large set of experiments on simulated data for contrasting our proposed structure learning algorithms (GSAMP, IAMBAMP, interIAMBAMP, fastIAMBAMP, and fdrIAMBAMP) against the state-of-the-art algorithms (stable) PC-like and LCD-AMP for AMP CG recovery. We implemented all algorithms in R by extending code from the bnlearn (Scutari, 2010) and pcalg (Kalisch et al., 2012) packages to AMP CGs. We run our algorithms and the algorithm (stable) PC-like and LCD-AMP on randomly generated AMP CGs and we compare the results and report summary error measures.

6.1 Performance Evaluation Metrics

We evaluate the performance of our proposed generic algorithm MbAMP (via 5 different instantiations: GSAMP, fastIAMBAMP, fdrIAMBAMP, interIAMBAMP, and IAMBAMP) in terms of six measurements: (a) the true positive rate (TPR) (also known as **recall**), (b) the false positive rate (FPR), (c) the true discovery rate (TDR) (also known as **precision**), (d) accuracy (ACC) for the skeleton, (e) the structural Hamming distance (SHD)[1], and (f) run-time for the CG recovery algorithms. In short, TPR is the ratio of the number of correctly identified edges over total number of edges (in true graph), FPR is the ratio of the number of incorrectly identified edges over total number of gaps, TDR is the ratio of the number of correctly identified edges over total number of edges (both in estimated graph), ACC is the ratio of the number of correctly identified edges plus the number of correctly identified gaps over the number of real positive cases plus the number of real negative cases, and SHD is the number of legitimate operations needed to change the current resulting graph to the true CG, where legitimate operations are: (i) add or delete an edge and (ii) insert, delete or reverse an edge orientation. In principle, a large TDR, TPR and ACC, a small FPR and SHD indicate good performance.

6.2 Data Generation Procedure

Now, we explain the way in which the random AMP chain graphs and random samples are generated. Given a vertex set V , let $p = |V|$ and N denote the average degree of edges (including undirected and pointing out and pointing in) for each vertex. We generate a random AMP chain graph on V as follows:

- Order the p vertices and initialize a $p \times p$ adjacency matrix A with zeros;
- For each element in the lower triangle part of A, set it to be a random number generated from a Bernoulli distribution with probability of occurrence $s = N/(p-1)$;
- Symmetrize A according to its lower triangle;
- Select an integer k randomly from $\{1, \ldots, p\}$ as the number of chain components;
- Split the interval $[1, p]$ into k equal-length subintervals $I_1, \ldots, I_k$ so that the set of variables falling into each subinterval I_m forms a chain component C_m;
- Set $A_{ij} = 0$ for any (i, j) pair such that $i \in I_l, j \in I_m$ with $l > m$.

This procedure yields an adjacency matrix A for a chain graph with $(A_{ij} = A_{ji} = 1)$ representing an undirected edge between V_i and V_j and $(A_{ij} = 1, A_{ji} = 0)$ representing a directed edge from V_i to V_j. Moreover, it is not difficult to see that $\mathbb{E}[\text{vertex degree}] = N$, where an adjacent vertex can be linked by either an undirected or a directed edge. In order to sample from the artificial CGs, we first transformed them into DAGs and then sampled from these DAGs under marginalization and conditioning as indicated in (Peña, 2014). The transformation of an AMP CG G into a DAG H is as follows: First, every node X in G gets a new parent ϵ^X representing an error term, which by definition is never observed. Then, every undirected edge X — Y in G is replaced by

1. This is the metric described in (Tsamardinos et al., 2006) to compare the structure of the learned and the original graphs.

$\epsilon^X \longrightarrow S_{XY} \longleftarrow \epsilon^Y$ where S_{XY} denotes a selection bias node, i.e. a node that is always observed. Given a randomly generated chain graph G with ordered chain components $C_1, \ldots, C_k$, we generate a Gaussian distribution on the corresponding transformed DAG H using the Hugin API. Note that the probability distributions of samples are likely to satisfy the faithfulness assumption, but there is no guarantee i.e., samples can have additional independencies that cannot be represented by the CG G.

6.3 Results and their Implications

In our simulation, we change three parameters p (the number of vertices), n (sample size) and N (expected number of adjacent vertices) as follows:

- $p \in \{10, 20, 30, 40, 50\}$,
- $n \in \{500, 5000\}$, and
- $N \in \{2, 3\}$.

For each (p, N) combination, we first generate 30 random AMP CGs. We then generate a random Gaussian distribution based on each graph and draw an identically independently distributed (i.i.d.) sample of size n from this distribution for each possible n. For each sample, three different significance levels $(\alpha = 0.005, 0.05)$ are used to perform the hypothesis tests. The *null hypothesis* H_0 is "two variables u and v are conditionally independent given a set C of variables" and alternative H_1 is that H_0 may not hold. We then compare the results to access the influence of the significance testing level on the performance of our algorithms.

Some highlights for AMP CGs recovery: (1) As shown in our experimental results (see Figure 6), except for GSAMP, our proposed Markov blanket based algorithm, MbAMP, (via 4 different instantiations: fastIAMBAMP, fdrIAMBAMP, interIAMBAMP, and IAMBAMP) is (slightly) better than PCAMP and is as good as or even (slightly) better than LCDAMP in many settings. The reason is that both LCDAMP and MbLWF algorithms take advantage of local computations that make them equally robust against the choice of learning parameters. (2) While our Markov blanket based algorithms (except for GSAMP) have better precision and FPR, the LCDAMP algorithm enjoys (slightly) better recall. The reason for this may be that the faithfulness assumption makes the LCDAMP algorithm search for a CG that represents all the independencies that are detected in the sample set. However, such a CG may also represent many other independencies. Therefore, the LCDAMP algorithm trades precision for recall. In other words, it seems that the faithfulness assumption makes the LCDAMP algorithm overconfident and aggressive, whereas under this assumption MbAMP algorithms are more cautious, conservative, and more importantly more precise than the LCDAMP algorithm. (3) The performance of the LCDAMP algorithm based on ACC is slightly better than other algorithms. Except for the GSAMP algorithm and the fdrIAMB algorithm in small sample size, there is no meaningful difference among the performance of the Markov blanket based algorithms based on ACC. (4) The best SHD belongs to LCDAMP in small sample size settings, and to fastIAMBAMP, interIAMBAMP, IAMBAMP, and LCD in large sample size settings. This is consistent with the results obtained in (Tsamardinos et al., 2003b) for Bayesian networks. Because IAMB and its variants are not data-efficient (Peña, 2008). (5) Except for the FPR, the p-value (α parameter) has a very small impact on the performance of algorithms. (6) Markov blankets of different variables can be learned independently from each other, and later merged and reconciled to

produce a coherent AMP CG. This allows the parallel implementations for scaling up the task of learning chain graphs from data containing more than hundreds of variables, which is crucial for big data analysis tools. In fact, our proposed structure learning algorithms can be parallelized following (Scutari, 2017); see `https://github.com/majavid/AMPCGs-PGM2020` for a detailed example.

With the use of our generic algorithm (Algorithm 1), the problem of structure learning is reduced to finding an efficient algorithm for Markov blanket discovery in AMP CGs. This greatly simplifies the structure-learning task and makes a wide range of inference/learning problems computationally tractable because they exploit locality.

7. Discussion and Conclusion

An important novelty of local methods in general and Markov blanket recovery algorithms in particular for structure learning is circumventing non-uniform graph connectivity. A chain graph may be non-uniformly dense/sparse. In a global learning framework, if a region is particularly dense, that region cannot be discovered quickly and many errors will result when learning with a small sample. These errors propagate to remote regions in the chain graph including those that are learnable accurately and fast with local methods. In contrast, local methods such as Markov blanket discovery algorithms are fast and accurate in the less dense regions. In addition, when the dataset has tens or hundreds of thousands of variables, applying global discovery algorithms that learn the full chain graph becomes impractical. In those cases, Markov blanket based approaches that take advantage of local computations can be used for learning full AMP CGs. For this purpose, we extended the concept of Markov blankets to AMP CGs. We proved that GSMB and IAMB and its variants are still sound for Markov blanket discovery in AMP CGs under the faithfulness and causal sufficiency assumptions. This, in turn, enabled us to extend these algorithms to a new family of global structure learning algorithms based on Markov blanket discovery. As we have shown for MbAMP algorithm (via 5 different instantiations) improves the quality of the learned AMP CG. One interesting direction for future work is answering the following question: Can we relax the faithfulness assumption and develop a correct, scalable, and data efficient algorithm for learning Markov blankets in AMP CGs?

Acknowledgments

We are grateful to Professor Jose M. Peña and Dr. Dag Sonntag for providing us with code that helped in the data generating procedure. This work has been partially supported by AFRL and DARPA (FA8750-16-2-0042).

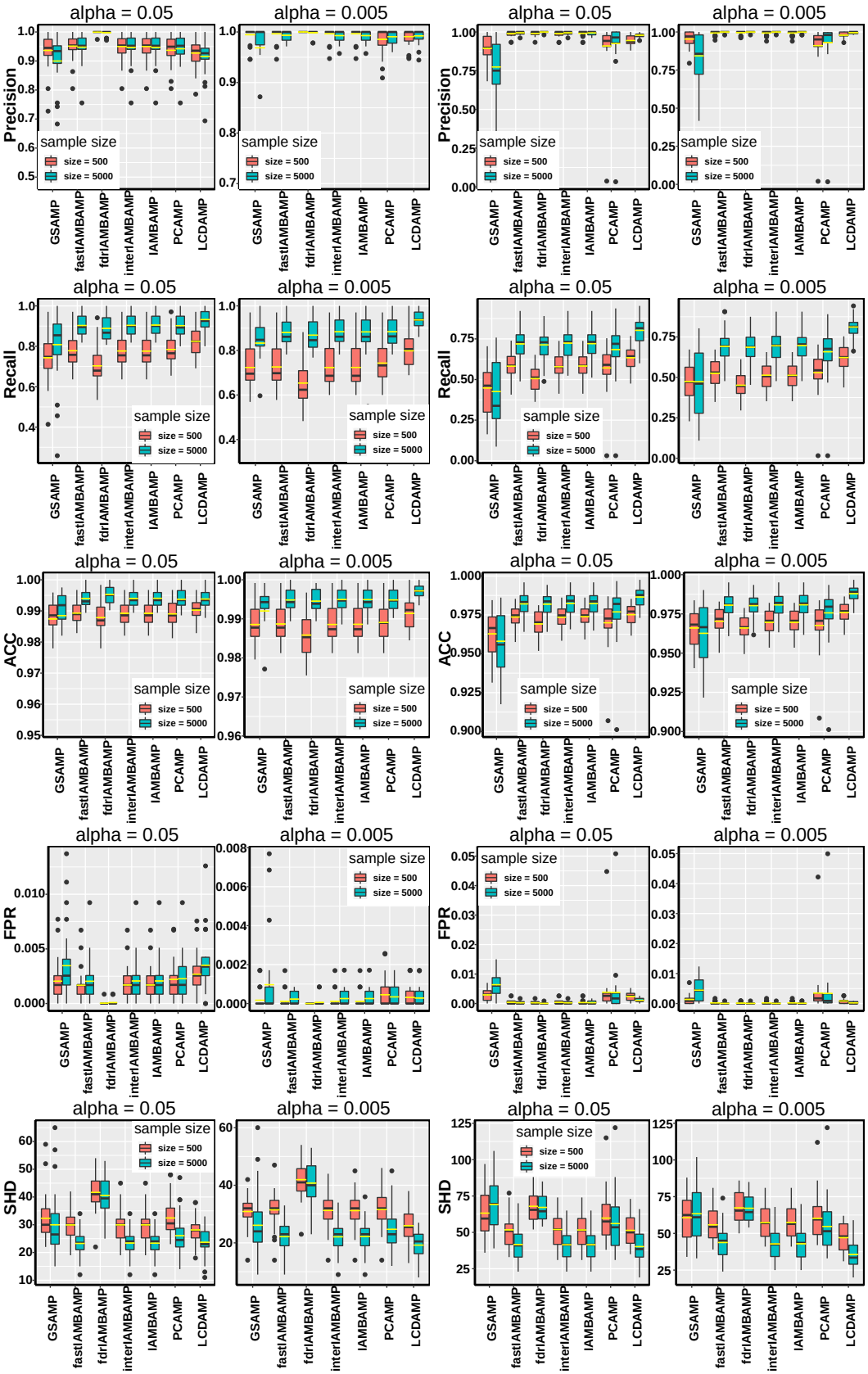

Figure 6: Performance of MbAMP, PC-like, and LCD-AMP algorithms for randomly generated Gaussian AMP CGss: over 30 repetitions with 50 variables correspond to N = 2 (the first two columns) and 3 (the last two columns). The yellow line in a box indicates the mean of that group.

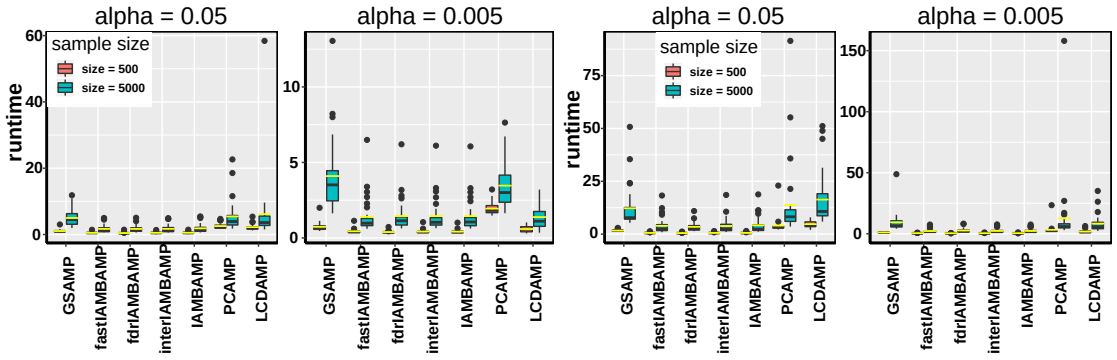

Figure 7: Runtime of MbAMP, PC-like, and LCD-AMP algorithms for randomly generated Gaussian AMP CGss: over 30 repetitions with 50 variables correspond to N = 2 (the first two columns) and 3 (the last two columns). The yellow line in a box indicates the mean of that group.

References

C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *J. Mach. Learn. Res.*, 11:171–234, Mar. 2010.

S. A. Andersson and M. D. Perlman. Characterizing Markov equivalence classes for AMP chain graph models. *The Annals of Statistics*, 34(2):939–972, 04 2006.

S. A. Andersson, D. Madigan, and M. D. Perlman. An alternative Markov property for chain graphs. In E. Horvitz and F. V. Jensen, editors, *Proceedings of the Twelfth Conference on Uncertainty in artificial intelligence*, pages 40–48, 1996.

S. A. Andersson, D. Madigan, and M. D. Perlman. Alternative Markov properties for chain graphs. *Scandinavian Journal of Statistics*, 28(1):33–85, 2001.

M. Drton. Discrete chain graph models. *Bernoulli*, 15(3):736–753, 2009.

T. Gao and Q. Ji. Constrained local latent variable discovery. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 1490–1496, 2016.

P. H. Ibargüengoytla, L. E. Sucar, and S. Vadera. A probabilistic model for sensor validation. In *Proceedings of the UAI'96*, pages 332–339, 1996.

M. A. Javidian, M. Valtorta, and P. Jamshidi. AMP chain graphs: Minimal separators and structure learning algorithms, 2020a. arXiv:2002.10870.

M. A. Javidian, M. Valtorta, and P. Jamshidi. Learning LWF chain graphs: A Markov blanket discovery approach, 2020b. submitted to UAI 2020.

M. Kalisch, M. Mächler, D. Colombo, M. Maathuis, and P. Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software, Articles*, 47(11): 1–26, 2012.

M. Levitz, M. D. Perlman, and D. Madigan. Separation and completeness properties for AMP chain graph Markov models. *The Annals of Statistics*, 29(6):1751–1784, 2001.

Z. Ling, K. Yu, H. Wang, L. Liu, W. Ding, and X. Wu. Bamb: A balanced Markov blanket discovery approach to feature selection. *ACM Trans. Intell. Syst. Technol.*, 10(5), Oct. 2019.

X. Liu and X. Liu. Swamping and masking in Markov boundary discovery. *Machine Learning*, 104 (1):25–54, Jul 2016.

D. Margaritis and S. Thrun. Bayesian network induction via local neighborhoods. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 505–511, Cambridge, MA, USA, 1999. MIT Press.

J. M. Peña. Learning gaussian graphical models of gene networks with false discovery rate control. In E. Marchiori and J. H. Moore, editors, *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 165–176, 2008.

J. M. Peña. Learning AMP chain graphs under faithfulness. In *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models*, pages 251–258, 2012.

J. M. Peña. Marginal AMP chain graphs. *International Journal of Approximate Reasoning*, 55(5): 1185–1206, 2014.

J. M. Peña. Every LWF and AMP chain graph originates from a set of causal models. *Symbolic and quantitative approaches to reasoning with uncertainty, Lecture Notes in Comput. Sci., 9161, Lecture Notes in Artificial Intelligence, Springer, Cham*, pages 325–334, 2015.

J. M. Peña. Alternative Markov and causal properties for acyclic directed mixed graphs. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 577–586, Arlington, Virginia, United States, 2016. AUAI Press.

J. M. Peña. Reasoning with alternative acyclic directed mixed graphs. *Behaviormetrika*, pages 1–34, 2018a.

J. M. Peña. Identification of strong edges in AMP chain graphs. In A. Globerson and R. Silva, editors, *Proceedings of the 34th Conference on Uncertainty in artificial intelligence*, pages 33–42, 2018b.

J. M. Peña and M. Gómez-Olmedo. Learning marginal AMP chain graphs under faithfulness revisited. *International Journal of Approximate Reasoning*, 68:108 – 126, 2016.

J. M. Peña, R. Nilsson, J. Björkegren, and J. Tegnér. Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning*, 45(2):211 – 232, 2007. Eighth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2005).

A. Roverato. A unified approach to the characterization of equivalence classes of dags, chain graphs with no flags and chain graphs. *Scandinavian Journal of Statistics*, 32(2):295–312, 2005.

A. Roverato and L. L. Rocca. On block ordering of variables in graphical modelling. *Scandinavian Journal of Statistics*, 33(1):65–81, 2006.

M. Scutari. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3):1–22, 2010.

M. Scutari. Bayesian network constraint-based structure learning algorithms: Parallel and optimized implementations in the bnlearn r package. *Journal of Statistical Software, Articles*, 77(2):1–20, 2017. ISSN 1548-7660. doi: 10.18637/jss.v077.i02. URL `https://www.jstatsoft.org/v077/i02`.

D. Sonntag and J. M. Peña. Chain graphs and gene networks. In A. Hommersom and P. J. Lucas, editors, *Foundations of Biomedical Knowledge Representation: Methods and Applications*, pages 159–178. Springer, 2015.

D. Sonntag and J. M. Peña. Chain graph interpretations and their relations revisited. *International Journal of Approximate Reasoning*, 58:39 – 56, 2015. Special Issue of the Twelfth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2013).

M. Studený, A. Roverato, and Š. Štěpánová. Two operations of merging and splitting components in a chain graph. *Kybernetika*, 45(2):208–248, 2009.

S. Triantafillou, I. Tsamardinos, and A. Roumpelaki. Learning neighborhoods of high confidence in constraint-based causal discovery. In L. C. van der Gaag and A. J. Feelders, editors, *Probabilistic Graphical Models*, pages 487–502. Springer International Publishing, 2014.

I. Tsamardinos, C. Aliferis, A. Statnikov, and E. Statnikov. Algorithms for large scale Markov blanket discovery. In *In The 16th International FLAIRS Conference, St*, pages 376–380. AAAI Press, 2003a.

I. Tsamardinos, C. F. Aliferis, and A. Statnikov. Time and sample efficient discovery of Markov blankets and direct causal relations. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 673–678, 2003b. doi: 10.1145/956750.956838.

I. Tsamardinos, , L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, Oct 2006.

S. Yaramakala and D. Margaritis. Speculative Markov blanket discovery for optimal feature selection. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 4 pp.–, 2005.

K. Yu, L. Liu, J. Li, and H. Chen. Mining Markov blankets without causal sufficiency. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):6333–6347, Dec 2018.