DOI: 10.48011/asba.v2i1.1663

Fast Markov Blanket Discovery Without Causal Sufficiency

Pedro V. B. Jeronymo * **Carlos D. Maciel** *

* *Signal Processing Laboratory, Department of Electrical and Computing Engineering, University of São Paulo, São Carlos, Brazil (e-mail: pedro.jeronymo@usp.br, maciel@sc.usp.br)*

Abstract: Faster feature selection algorithms become a necessity as Big Data dictates the *zeitgeist*. An important class of feature selectors are Markov Blanket (MB) learning algorithms. They are Causal Discovery algorithms that learn the local causal structure of a target variable. A common assumption in their theoretical basis, yet often violated in practice, is causal sufficiency: the requirement that all common causes of the measured variables in the dataset are also in the dataset. Recently, Yu et al. (2018) proposed the M3B algorithm, the first to directly learn the MB without demanding causal sufficiency. The main drawback of M3B is that it is time inefficient, being intractable for high-dimensional inputs. In this paper, we derive the Fast Markov Blanket Discovery Algorithm (FMMB). Empirical results that compare FMMB to M3B on the structural learning task show that FMMB outperforms M3B in terms of time efficiency while preserving structural accuracy. Five real-world datasets where used to contrast both algorithms as feature selectors. Applying NB and SVM classifiers, FMMB achieved a competitive outcome. This method mitigates the curse of dimensionality and inspires the development of local-to-global algorithms.

Keywords: Big Data. Causal Discovery. Feature Selection. Markov Blanket.

1. INTRODUCTION

In the Big Data era, as datasets grow in size and complexity, feature selection plays a key role in dimensionality reduction (Bolón-Canedo et al., 2015). Markov Blanket (MB) learning algorithms, originally developed as a method for Feature Selection, are an important class of local structural learning algorithms, with applications in local-to-global structural learning and Causal Discovery (Aliferis et al. (2010a); Aliferis et al. (2010b)). The concept of a MB was introduced by Pearl (1988) as part of his work on reasoning under uncertainty, in which the theory of Bayesian Networks (BNs) is founded. BNs are probabilistic graphical models that factor the joint probability distribution of the variables in a system into a directed acyclic graph (DAG) (Koller and Friedman, 2009). The MB of a target variable T is a set of variables that shield T from the influence of all other variables the system. Fig. 1(a) displays an example of a MB. The target T has adjacent variables A, B and D. A and B are parents of T. D is a child of T. Variables L and M are spouses of T. In the context of DAG MBs, the set of adjacent variables is called the PC set of T. DAG MB learning algorithms are theoretically dependent on Causal Sufficiency (Def. 1). However, this assumption is often violated in practice (Spirtes et al., 2000). Efforts have been made to model latent variables in DAGs (Spirtes et al., 2000), yet these models present computational and theoretical difficulties that can be avoided by adopting Maximal Ancestral Graphs (MAGs), developed as an extension of DAG causal models to accommodate the lack of causal sufficiency (Richardson et al., 2002). Instead of directly representing latent variables, unaccounted common causes induce a bidirected edge between variables. For example, Fig. 1(b) shows the induced MAG Markov Blanket (MMB) of T when hiding L. Because L is a common cause of D and N, a bidirected edge appears between D and N when L can no longer be observed.

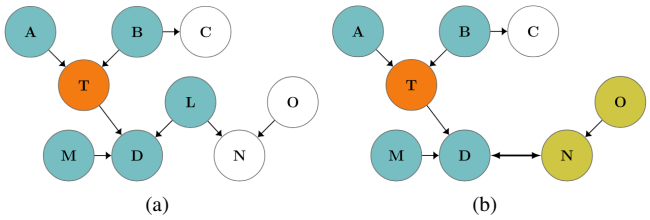

Figure 1. (a) DAG Markov Blanket of T shaded in blue and (b) induced MAG Markov Blanket of T when hiding L, composed of blue and green shaded nodes. Because L can no longer be observed, N and O influence T through D.

The MB was established as the optimal subset of features in the seminal paper of Koller and Sahami (1996), gathering interest in the Feature Selection community. Estimating the DAG MB of a target became feasible after Margaritis and Thrun (2000) proposed the first tractable solution, the Grow-Shrink (GS) algorithm, using conditional independence tests (CITs) as a method for variable selection. First, it grows a candidate set of features by testing association with the target. Then, it shrinks the candidates removing false positives. Since GS, IAMB (Tsamardinos and Aliferis, 2003) and variants (Tsamardinos et al. (2003b); Yaramakala and Margaritis (2005); Yang et al. (2019) among others) improved on these ideas, forming the family of simultaneous algorithms. Pena et al. (2007) showed that simultaneous algorithms are not data efficient. The amount of data required to maintain the power

⋆ This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

of the CITs can sometimes surpass the size of the dataset, as the conditioning set grows proportional to the candidate set. To overcome this, divide-and-conquer algorithms were proposed, the pioneer being the MMMB algorithm (Tsamardinos et al., 2003a). They first learn the adjacent variables to the target (i.e. PC set), then learn spouses by identifying the adjacent variables to each member of PC. This strategy reduces the sample size requirements at the expense of a greater time complexity. A theoretical framework and broad experimental evaluation for this family of algorithms was given on Aliferis et al. (2010a) and Aliferis et al. (2010b). The latest development in this line of research is the EEMB algorithm (Wang et al., 2020). It combines the strategies of simultaneous and divide-and-conquer algorithms to derive a method capable of simultaneously learning the MB while distinguishing PC from spouses. Yu et al. (2019) is an extensive review on DAG MB learning algorithms.

Recently, Yu et al. (2018) presented the M3B algorithm to subdue the causal sufficiency assumption. In contrast to previous MAG learning algorithms such as FCI (Spirtes et al., 2000) and RFCI (Colombo et al., 2012) that required learning the complete MAG before extracting the MB, it directly learns the MMB of a target. M3B uses a strategy similar to divide-and-conquer DAG MB algorithms. It discovers members of the MMB by learning the adjacent set of T, then recursively finds other members of the MMB using adjacency. This entails the same time efficiency limitations encountered in the family of DAG divide-and-conquer algorithms. Here, motivated by the need of faster algorithms imposed by the growing number of high-dimensional datasets, we propose the Fast MAG Markov Blanket (FMMB) algorithm. The key difference from M3B is a new strategy for finding non-adjacent members of the MMB that reduces the search space by rapidly eliminating descendants of members of the MMB, thus reducing the amount of CITs needed to learn the MMB and consequently improving time efficiency.

The rest of this paper is organised as follows. Sec. 2 contains the relevant theory. The proposed algorithm is introduced in sec. 3. Experiments comparing the algorithm with the state-of-the-art are reported in sec. 4. Finally, sec. 5 concludes the paper pointing to feature research possibilities.

2. NOTATION AND THEORY

First, the theory of BNs and DAG MBs is introduced via a series of definitions. Then, these notions are extended to MAGs and MMBs. Finally, concepts relevant to the development of FMMB conclude this section.

Definition 1. (Causal Sufficiency). A set of variables $\mathbf{V}$ is *causally sufficient iff* every common cause of two or more variables in $\mathbf{V}$ is in $\mathbf{V}$. (Spirtes et al., 2000)

Definition 2. (Conditional Independence). Two variables X and Y are conditionally independent given Z, denoted $X \perp\!\!\!\perp Y \mid Z$, *iff* $P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z)$ for all instantiations x, y, z of X, Y, Z.

Definition 3. (Bayesian Network). Let $\mathbf{V}$ be a set of variables, J a joint probability distribution over $\mathbf{V}$ and G a DAG that encodes the conditional independence relations between all the variables in $\mathbf{V}$. We call $\langle \mathbf{V}, G, J \rangle$ a Bayesian Network (BN) if for every node $X \in \mathbf{V}$, X is independent of all non-descendants conditioned on its parents ($Pa_G(X)$). (Koller and Friedman, 2009)

Definition 4. (Path). A *path* is a sequence of variables $\mu = [X_1, ..., X_{i-1}, X_i, X_{i+1}, ..., X_n]$ such that $\forall i \in \{1, 2, ..., n-1\}$, X_i is adjacent to X_{i+1}.

Definition 5. (Collider (DAG)). A node X_i is a *collider* on a path μ if $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, that is, X_i has two incoming arrows. A variable X that is not a collider in μ is a *noncollider* in μ. (Spirtes et al., 2000)

Definition 6. (Active and Blocked Paths (DAG)). A path μ between $X \in \mathbf{V}$ and $Y \in \mathbf{V}$ is *active* given $\mathbf{Z}$ if two conditions hold:

(1) every collider in μ is in $\mathbf{Z}$ or has a descendant in $\mathbf{Z}$ and
(2) every noncollider node in μ is not in $\mathbf{Z}$.

If a path is not active given $\mathbf{Z}$ its said to be *blocked* by $\mathbf{Z}$. (Pearl, 1988)

Definition 7. (D-Separation). Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be disjoint subsets of nodes in a DAG. If every path between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ is blocked by $\mathbf{Z}$ then $\mathbf{Z}$ *d-separates* $\mathbf{X}$ and $\mathbf{Y}$. (Pearl, 1988)

Definition 8. (Sepset (DAG)). A set that d-separates X from T is called a *sepset* of X with respect to T, denoted $\mathbf{Sep}_T[X]$. Spirtes et al. (2000)

For example, in Fig. 1(a), the node D is a collider in the path $[T, D, L, N]$. The path is active when D is observed (i.e. $D \in \mathbf{Z}$) and L is unobserved, but can be blocked by observing L, d-separating N from T. In other words, the set $\mathbf{Z} = \{D, L\}$ is a sepset of N with respect to T.

Definition 9. (Faithfulness). A distribution J is *faithful* to a DAG G if, whenever $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ holds in J, then $\mathbf{Z}$ d-separates X and Y in G. (Koller and Friedman, 2009)

Faithfulness is a key assumption for structural learning because it assures that CITs and d-separation work as intended. In an unfaithful BN, there might be associations that change depending on the state of variables, violating the uniqueness of the MB of a target. (Statnikov et al., 2013)

Definition 10. (Markov Blanket (DAG)). The Markov Blanket of a target variable T is the minimum set of variables $\mathbf{MB}(T)$ such that $\mathbf{S} \perp\!\!\!\perp T \mid \mathbf{MB}(T)$ for all $\mathbf{S} \subseteq \mathbf{V} \setminus \mathbf{MB}(T) \setminus \{T\}$. (Pearl, 1988)

Theorem 11. In a faithful BN $\langle \mathbf{V}, G, J \rangle$, $\forall T \in \mathbf{V}$ the Markov Blanket of T is unique and formed by the parents of T ($\mathbf{pa}(T)$), children of T ($\mathbf{ch}$(T)) and other parents of children of T, called spouses of T ($\mathbf{sp}(T)$). (Pearl, 1988)

Thm. 12 is the tool used by EEMB to search for spouses.

Theorem 12. Let $\langle \mathbf{V}, G, J \rangle, \forall T \in \mathbf{V}$ be a faithful BN, $S, C, T \in \mathbf{V}$ form a collider $S \rightarrow C \leftarrow T$ with S not adjacent to T. If $S \not\perp\!\!\!\perp T \mid \mathbf{Sep}_T[S] \cup \{C\}$, then S is a spouse of T. (Spirtes et al., 2000)

As exemplified in Fig. 1, when causal sufficiency is violated, there may be variables that can no longer be d-separated from the target. A DAG MB algorithm applied to the network in Fig. 2 may detect E as part of PC and F, G, N and M as spouses. However, J, H and O are invisible. To be able to represent and detect the sophisticated relationships between T and J, H and O, MAGs and a MMBs are needed. The following definitions construct these concepts.

Definition 13. (Mixed Graphs). Mixed Graphs are graphs that may contain three types of edges: undirected (–), directed ($\rightarrow$ and $\leftarrow$) and bidirected ($\leftrightarrow$). (Richardson et al., 2002)

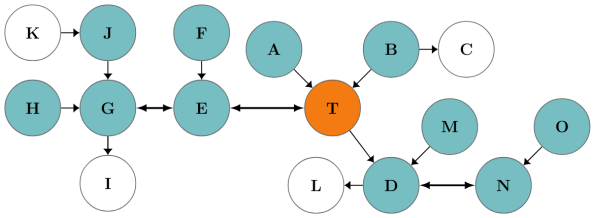

Figure 2. Example of MAG with MMB of a target T shaded in blue. Variables J, H and O are invisible to DAG MB algorithms, but can be detected by MMB algorithms.

Definition 14. (Ancestor and Directed Path). A node Y is said to be an *ancestor* of X if there is a directed path from Y to X or $Y = X$. The set of ancestors of X is denoted $\mathbf{an}(X)$. A path μ is *directed* if there are only directed edges in μ. (Richardson et al., 2002)

Definition 15. (Anterior and Anterior Path). A node Y is said to be an *anterior* of X if $Y = X$ or there is a path μ between Y and X such that all arrows point towards X. That is, if nodes W and Z are adjacent in μ and W precedes Z, $W - Z$ and $W \rightarrow Z$ are allowed, whereas $W \leftarrow Z$ and $W \leftrightarrow Z$ are not. A path satisfying these conditions is called a *anterior path*. The set of all nodes anterior to X is denoted $\mathbf{ant}(X)$. (Richardson et al., 2002)

Definition 16. (Directed cycle). A *directed cycle* occurs when there is a directed path from X to Y and an edge $Y \rightarrow X$. (Richardson et al., 2002)

Definition 17. (Partially directed cycle). A *partially directed cycle* occurs when there is an anterior path from X to Y and $Y \rightarrow X$. (Richardson et al., 2002)

Definition 18. (Ancestral Graph). An *Ancestral Graph* is a mixed graph without directed of partially directed cycles. (Richardson et al., 2002)

Definition 19. (Collider (MAG)). A nonendpoint node X_i on a path μ is a *collider* on μ if $X_{i-1} * \rightarrow X_i \leftarrow * X_{i+1}$, where $*$ denotes the presence or absence of a arrowhead. (Richardson et al., 2002)

For example, the path $[H,G,E,T,D,M]$ in Fig. 2 has three colliders: G, E and D.

Definition 20. (M-Separation). A path μ *m-connects* X and Y given $\mathbf{Z}$ if:

(1) every noncollider on μ is not in $\mathbf{Z}$ and
(2) every collider on μ is in $\mathbf{ant}(X)$

If $\forall X \in \mathbf{X}$ and $\forall Y \in \mathbf{Y}$ all paths between X and Y are *not* m-connected given $\mathbf{Z}$, then $\mathbf{Z}$ *m-separates* $\mathbf{X}$ and $\mathbf{Y}$. (Richardson et al., 2002)

Definition 21. (Sepset (MAG)). If $\mathbf{Z}$ m-separates X and T then $\mathbf{Z}$ is called a *sepset* of X with respect to T, denoted $\mathbf{Sep}_T[X] = \mathbf{Z}$. (Richardson et al., 2002)

Definition 22. (Maximal Ancestral Graph). An ancestral graph G is said to be *maximal* if for any nonadjacent pair of variables X and Y in G there is a set $\mathbf{Z}$ that m-separates them. (Richardson et al., 2002)

Definition 23. (District Set). A node A belongs to the *district set* of X, denoted $dis(X)$, if the path from A to X only contains bidirectional edges. (Yu et al., 2018)

Definition 24. (Collider Path). Two nodes X and Y are *collider connected* if there is a path μ from X to Y such that every node in μ except the endpoints is a collider. Such μ is said to be a *collider path*. (Richardson et al., 2002)

Proposition 25. Let variables X and Y in a MAG be collider connected by path μ. If $\mathbf{Z}$ contains all colliders in μ, then $X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$. (Yu et al., 2018)

Prop. 25 provides a mean to identify the variables invisible to DAG MB algorithms. In Fig. 2, H collider connects to T by $[T,E,G,H]$. If an algorithm wants to detect the influence that H has on T, it must follow this path somehow. The key difference between M3B and FMMB is in their strategy for building collider paths. While M3B walks node by node, identifying adjacent variables and checking if they belong to the path, FMMB tests directly if there is a connection by conditioning on all colliders in the path, e.g. FMMB tests if $H \not\perp\!\!\!\perp T \mid \{E,G\}$ holds.

Definition 26. (MAG Markov Blanket). The Mag Markov Blanket (MMB) of a variable T in a MAG is the minimum set of variables $\mathbf{MMB}(T)$ such that $\forall \mathbf{S} \subseteq \mathbf{V} \setminus \mathbf{MMB}(T) \setminus \{T\}$, $\mathbf{S} \perp\!\!\!\perp T \mid \mathbf{MMB}(T)$. (Yu et al., 2018)

Theorem 27. The $\mathbf{MMB}(T)$ includes:

(1) $\mathbf{pa}(T)$: parents of T
(2) $\mathbf{ch}(T)$: children of T
(3) $\mathbf{sp}(T)$: spouses of T
(4) $\mathbf{dis}(T)$: district set of T
(5) $\mathbf{pa}(\mathbf{dis}(T))$: parents of variables in $\mathbf{dis}(T)$
(6) $\mathbf{pa}(\mathbf{dis}(\mathbf{ch}(T)))$: parents of variables in $\mathbf{dis}(\mathbf{ch}(T))$

(Yu et al., 2018)

Theorem 28. Under causal sufficiency, $\mathbf{MMB}(T)$ equals $\mathbf{MB}(T)$. (Yu et al., 2018)

The subsequent definition and theorem are central to the understanding of the FMMB algorithm and to prove it's correctness. Def. 29 states that variables that are collider connected to the target form the portion of the MMB that is not adjacent to the target. Thm. 30 guarantees that those variables cannot be m-separated from the target. As an example, in Fig. 2 we have $\mathbf{Adj}_T = \{A,B,D,E\}$ and $\mathbf{AS}_T = \{F,G,H,J,M,N,O\}$ as of Def. 29. Thm. 30 assures that H cannot be m-separated from T conditioned on $\{E,G\} \cup \mathbf{Z}$, where $\mathbf{Z}$ may contain any other member of $MMB(T)$.

Definition 29. (Attached Set). Let $\mathbf{Adj}_T$ represent the set of all variables adjacent to T, specifically the union of $\mathbf{PC}_T$ with members of $\mathbf{dis}(T)$ adjacent to T. The set $\mathbf{AS}_T = \mathbf{MMB}(T) \setminus \mathbf{Adj}_T$ is defined as the *attached set* of T and $\mathbf{AS}_T[\mu] \subseteq \mathbf{AS}_T$ is defined as the set of all $X \in \mathbf{AS}_T \setminus \mu$ that are m-connected to T given μ, where $[T] + \mu + [X]$ is a collider path and the path μ is a nonempty subset of $\mathbf{MMB}(T)$.

Theorem 30. If $X \in \mathbf{AS}_T[\mu]$, then $\forall \mathbf{Z} \subseteq \mathbf{Adj}_T \cup \mathbf{AS}_T[\mu] \setminus \{X\}$, $X \not\perp\!\!\!\perp T \mid \mu \cup \mathbf{Z}$.

Proof. $X \in \mathbf{AS}_T[\mu]$ implies X m-connects to T through the collider path $[T] + \mu + [X]$, thus, because μ contains all the colliders, Prop. 25 guaranties $X \not\perp\!\!\!\perp Y \mid \mu \cup \mathbf{Z}, \forall \mathbf{Z} \subseteq \mathbf{Adj}_T \cup \mathbf{AS}_T[\mu] \setminus \{X\}$. ■

3. PROPOSED ALGORITHM

This section starts with an overview of the FMMB algorithm, followed by a derivation of its correctness and time complexity. The algorithm consists of the main routine (`FMMB` — Alg. 1), that returns the $\mathbf{MMB}(T)$ for some target T, and a recursive subroutine (`GetAS` — Alg. 2), that returns all the members of $\mathbf{AS}_T[\mu]$ for some path μ. In the implementation, $\mathbf{Sep}_T$ and $\mathbf{AS}_T$

are lookup tables. The entry $\mathbf{Sep}_T[X]$ represents the sepset of X with respect to T and $X \in \mathbf{Sep}_T$ means that X is the key to the entry. Similar logic applies to $\mathbf{AS}_T$. The function `del` deletes entries from a lookup table.

Step one of Alg. 1 is calling the EEMB algorithm to discover $\mathbf{Adj}_T$ and start populating $\mathbf{AS}_T$ and $\mathbf{Sep}_T$. The table $\mathbf{AS}_T$ is populated with entries $C \in \mathbf{ch}(T) \cup \mathbf{dis}(T)$ and $\mathbf{AS}_T[C]$ contains spouses of T, members of $\mathbf{dis}(C)$ adjacent to C and parents of members of $\mathbf{dis}(T)$ adjacent to T. Complementarily, $\mathbf{Sep}_T$ is populated with all other variables not in $\mathbf{Adj}_T$ and currently not in $\mathbf{AS}_T$. Then, for all such C and all members of $Y \in \mathbf{AS}_T[C]$, `FMMB` calls `GetAS` triggering a recursion on all paths μ that start with $[C,Y]$. After all recursions stop, $\mathbf{AS}_T$ is completely populated. Finally, we have $\mathbf{MMB}(T) = \mathbf{Adj}_T \cup \mathbf{AS}_T$.

The subroutine Alg. 2 operates recursively. It populates $\mathbf{AS}_T$ by first starting with an empty set of candidates **Can**. It grows **Can** with all variables currently not in $\mathbf{MMB}(T)$, namely all keys $X \in \mathbf{Sep}_T$ that satisfy Prop. 25 for the argument μ. Then, is shrinks **Can** filtering false positives by conditioning on all $\mathbf{Z}$ such that $\mu \cup \mathbf{Z}$ could m-separate X from T. After shrinking **Can**, only true members of $\mathbf{AS}_T[\mu]$ remain and all entries $X \in \mathbf{AS}_T[\mu]$ in $\mathbf{Sep}_T$ are deleted. At last, for all members X of $\mathbf{AS}_T[\mu]$, `GetAS` recursively calls itself expanding the collider path μ with X. The recursion stops when $\mathbf{AS}_T[\mu]$ is empty for some μ.

Algorithm 1: FMMB

Input : T: target variable; α: significance level
Output: $\mathbf{MMB}(T)$: MAG Markov Blanket of T

$\mathbf{Adj}_T, \mathbf{AS}_T, \mathbf{Sep}_T \leftarrow$ `EEMB`(T, α)
for $C \in \boldsymbol{AS}_T$ **do**
  for $Y \in \boldsymbol{AS}_T[C]$ **do**
    `GetAS`$(T, [C,Y], \alpha)$
  end
end
$\mathbf{MMB}(T) \leftarrow \mathbf{PC}_T \cup \mathbf{AS}_T$

Algorithm 2: GetAS

Input : T: target variable; μ: collider path; α: significance level
Output: $\mathbf{AS}_T[\mu]$

$\mathbf{Can} \leftarrow \emptyset$
for $X \in \boldsymbol{Sep}_T$ **do**
  if $X \not\perp\!\!\!\perp T \mid \mu \cup \boldsymbol{Sep}_T[X]$ **then**
    $\mathbf{Can} \leftarrow \mathbf{Can} \cup \{X\}$
  end
end
for $X \in \boldsymbol{Can}$ **do**
  if $\exists \boldsymbol{Z} \subseteq \boldsymbol{Can} \cup \boldsymbol{Adj}_T \setminus \{X\}$ *such that* $X \perp\!\!\!\perp T \mid \mu \cup \boldsymbol{Z}$ **then**
    $\mathbf{Can} \leftarrow \mathbf{Can} \setminus \{X\}$
  end
end
$\mathbf{AS}_T[\mu] \leftarrow \mathbf{Can}$
for $X \in \boldsymbol{AS}_T[\mu]$ **do**
  `del`$(\boldsymbol{Sep}_T[X])$
end
for $X \in \boldsymbol{AS}_T[\mu]$ **do**
  `GetAS`$(T, \mu + [X], \alpha)$
end

We proceed with a proof for Thm. 31, that formalises the concept of correctness.

Theorem 31. Under faithfulness and assuming the precision of all CITs, FMMB correctly finds the MMB of a given target.

Proof. Let T be a target. The EEMB algorithm uses Thm. 12 to find spouses in DAGs. When applied in MAGs, during the growth phase, it finds all triples $\langle T, C, S \rangle$ that satisfy Thm. 12 where $C \in \mathbf{Adj}_T$, but using the definition of a collider in the context of MAGs (Def. 19) and m-separation. This is equivalent to discovering all subsets $\mathbf{AS}_T[\mu]$ where $\mu = [C], C \in \mathbf{Adj}_T$. All other paths $\mu \subseteq \mathbf{MMB}(T)$ are exhaustively explored by `GetAS`. During growth, Alg. 2 finds all X that m-connect to T through μ. During shrinking, all false positives are removed by finding a violation of Thm. 30. Let μ be a path in $\mathbf{AS}_T$. FMMB correctly filters all false positives in the candidate set **Can** of $\mathbf{AS}_T[\mu]$ as follows: Let Y and Z be the second to last and last nodes in μ respectively. Since Y and Z are observed, any X that m-connects to T through μ must either form a collider $Y* \rightarrow Z \leftarrow *X$ or have an active path with a node that forms a collider with Y and Z. This eliminates all descendants of Z because they cannot form a collider or m-connect with a node that form a collider. If X m-connects to T through μ but $X \notin \mathbf{AS}_T[\mu]$, there is a $Y \in \mathbf{AS}_T[\mu]$ such that X m-connects to Y when Y is not observed. Because $\mathbf{Sep}_T[X] \subseteq \mathbf{Adj}_T$ and $\mathbf{AS}_T[\mu] \subseteq \mathbf{Can}$, there exists a $\mathbf{Z} \subseteq \mathbf{Can} \cup \mathbf{Adj}_T \setminus \{X\}$ that contains $\{Y\} \cup \mathbf{Sep}_T[X]$, m-separating X from T when observed. Therefore, FMMB correctly finds the $\mathbf{MMB}(T)$. ■

The most time-consuming operations in FMMB and EEMB are the CITs. Therefore, it is natural to express computational complexity in terms of the number of CITs employed during execution. The EEMB algorithm has time complexity $\mathcal{O}(|\mathbf{V}|2^{|\mathbf{Adj}_T|})$ (Wang et al., 2020). FMMB exhaustively considers every path μ between T and $X \in \mathbf{AS}_T$, and for every μ Alg. 2 is executed with time complexity $\mathcal{O}(|\mathbf{V}|2^{|\mathbf{V}|})$. This amounts to a overall complexity of $\mathcal{O}(|\mathbf{V}|2^{|\mathbf{Adj}_T|}) + \mathcal{O}(|\mathbf{AS}_T||\mathbf{V}|2^{|\mathbf{V}|}) = \mathcal{O}(|\mathbf{AS}_T||\mathbf{V}|2^{|\mathbf{V}|})$.

4. EXPERIMENTS

This section presents a comparison of the proposed algorithm, FMMB, *vs* the state-of-the-art algorithm for MMB discovery, M3B. Two experiments were performed. The first measures structural learning accuracy using the benchmark BN Alarm (Fig. 3). The second measures accuracy of classification applying both algorithms for feature selection in five real-world datasets. Algorithms were implemented in Python 3.7 and experiments were conducted on a machine running Linux Kernel version 4.15 with 16 GB of RAM and Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz. The G^2 test of conditional independence (Kullback, 1997) is used.

Structural learning is the primary task of this class of algorithms. Tab. 1 displays the results for the three test cases (TC) used in Yu et al. (2018). The standard metrics (Aliferis et al. (2010b); Yu et al. (2018); Yu et al. (2019)) were applied:

- Precision: the number of true positives in the output of the algorithm divided by the total size of the output of the algorithm. This measures the false positive rate in the output;

Table 1. Comparison on MMB learning accuracy of FMMB *vs* M3B. Three test cases were created based on the Alarm Bayesian Network. Ten sample datasets were generated for each test case and sample size. The statistics are in the format *mean* ± *std*.

Sample Size	Algorithm	Precision	Recall	F1	Time	CITs
TC 1 – Estimation of the MMB of 'LVV' hiding 'HYP'						
500	FMMB	0.80 ±0.22	0.68 ±0.11	0.71 ±0.14	0.78 ±0.09	197.90 ±51.57
	M3B	1.00 ±0.00	0.50 ±0.00	0.67 ±0.00	1.52 ±0.02	144.10 ±2.47
5000	FMMB	1.00 ±0.00	1.00 ±0.00	1.00 ±0.00	0.88 ±0.03	217.80 ±3.03
	M3B	1.00 ±0.00	1.00 ±0.00	1.00 ±0.00	3.09 ±0.46	284.70 ±2.93
TC 2 – Estimation of the MMB of 'VTUB' hiding 'INT'						
500	FMMB	0.92 ±0.08	0.79 ±0.10	0.84 ±0.07	0.93 ±0.06	268.00 ±34.74
	M3B	1.00 ±0.00	0.39 ±0.07	0.55 ±0.07	2.51 ±0.55	249.00 ±27.36
5000	FMMB	1.00 ±0.00	0.70 ±0.04	0.82 ±0.03	1.51 ±0.12	476.20 ±47.26
	M3B	1.00 ±0.00	0.66 ±0.07	0.79 ±0.05	4.52 ±0.73	626.20 ±40.73
TC 3 – Estimation of the MMB of 'VTUB' hiding 'INT' and 'KINK'						
500	FMMB	0.88 ±0.12	0.75 ±0.11	0.80 ±0.09	0.84 ±0.06	229.80 ±32.22
	M3B	1.00 ±0.00	0.45 ±0.08	0.62 ±0.08	2.03 ±0.23	229.60 ±24.54
5000	FMMB	1.00 ±0.00	0.67 ±0.00	0.80 ±0.00	1.18 ±0.03	352.40 ±12.36
	M3B	1.00 ±0.00	0.60 ±0.08	0.75 ±0.07	2.98 ±0.30	500.80 ±39.85

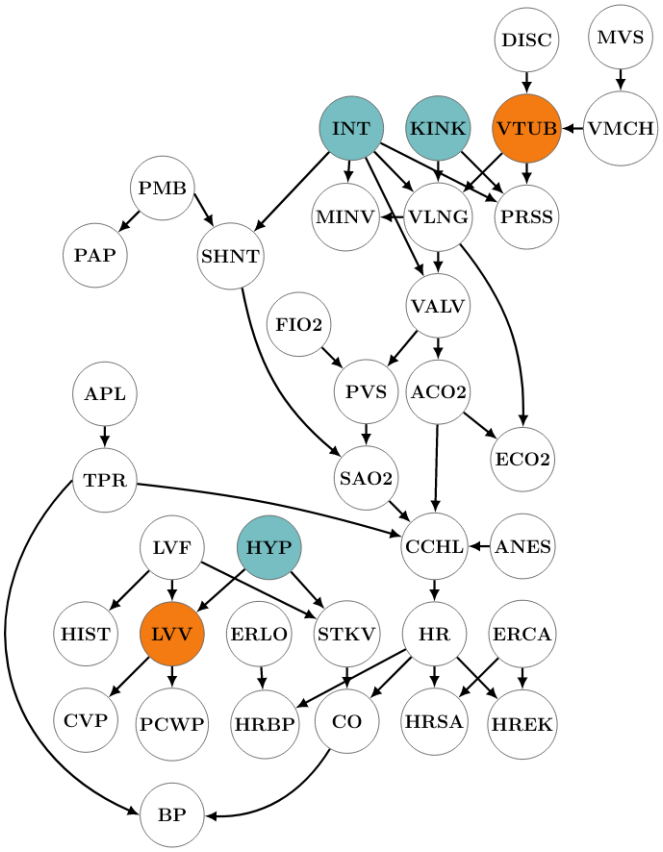

Figure 3. Standard Benchmark Bayesian Network "Alarm". Highlighted nodes are used to construct test cases. The target nodes are shaded in orange and the hidden nodes are shaded in blue.

- Recall: the number of true positives in the output of the algorithm divided by the size of the true MMB of the target. This reports the true positive rate in the output;
- F1: $f1 = 2 * precision * recall/(precison + recall)$ is the harmonic mean of precision and recall.
- CITs: total number of CITs conducted;
- Time: elapsed time of execution.

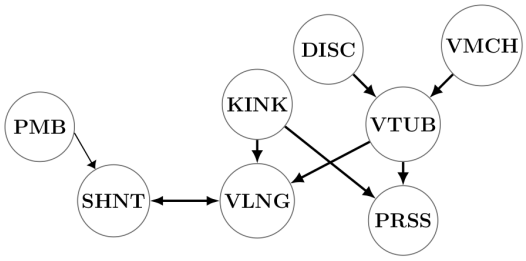

Figure 4. Test Case Number 2: MAG Markov Blanket of "VTUB" when "INT" is hidden. Despite having "INT" as a common cause, "MINV" and "VALV" do not belong to the MMB of "VTUB" because they can be m-separated from "VTUB" by "VLNG".

For small sample sizes, FMMB is unstable. Due to it's strategy for building a collider path between T and members of $\mathbf{AS}_T$, the conditioning sets increase in size, decreasing the accuracy of the G^2 tests. This explains greater recall, but lower precision, when contrasted with M3B. Although the strategy used by M3B is data efficient, leading to precise tests, the connection strengths between T and some members of $\mathbf{MMB}(T)$ are not strong enough to be consistently detected at 500 samples, leading to less CITs than FMMB because it stopped searching prematurely. This contributes to M3B's lower recall. For example, in TC 2 (displayed in Fig. 4), "SHNT" has a weak connection to "VLNG", making it harder for the algorithms to detect "SHNT" and "PMB". When sample size is sufficiently large (i.e. 5000), FMMB outperforms M3B in all test cases with better recall and less CITs. This results show that, given a large enough sample size, FMMB's strategy leads to faster and better results regarding structural learning.

Table 2. Summary of Five Real-World Datasets

Dataset	Features	Samples
mushroom	22	8124
spect	22	267
spectf	44	267
chess	36	3196
splice	60	3190

Table 3. 10-fold Cross-Validation Prediction Accuracy Comparison of FMMB against M3B on Five Real-World Datasets.

Dataset	FMMB	M3B
	NB	
mushroom	0.92 ±0.11	0.94 ±0.10
spect	0.79 ±0.02	0.73 ±0.13
spectf	0.61 ±0.10	0.61 ±0.10
chess	0.82 ±0.10	1.00 ±0.00
splice	0.96 ±0.01	0.95 ±0.01
	SVM	
mushroom	0.82 ±0.16	0.86 ±0.19
spect	0.74 ±0.13	0.73 ±0.17
spectf	0.60 ±0.21	0.60 ±0.21
chess	0.90 ±0.08	1.00 ±0.00
splice	0.85 ±0.02	0.86 ±0.02

As this class of algorithms were originally developed as tools for feature selection, a classification comparison after applying FMMB and M3B in Five real-world datasets is presented. This datasets where selected from the UCI Machine Learning Repository (Dua and Graff, 2017) and the KEEL dataset repository (Alcalá-Fdez, 2011). Table 2 summarises their the characteristics. The classification accuracy for each dataset, obtained by Naive Bayes (NB) and Support Vector Machine (SVM) classifiers with 10-fold cross-validation, after feature selection using FMMB and M3B is shown in table 3. The algorithms tied using the NB classifier, whereas M3B was better in three out of five datasets using SVM. Although unfaithful relationships are present, degrading the performance of both algorithms, FMMB achieved competitive accuracy when compared to M3B.

5. CONCLUSION

We revisited the problem of MB discovery without assuming causal sufficiency by introducing a MAG Markov Blanket Learning Algorithm with a faster approach that mitigates the curse of dimensionality. The key idea is the strategy of non-adjacent member discovery, that improved the state-of-the-art in terms of time efficiency by reducing the number of CITs needed to find the MMB of a target. The main drawback of FMMB is the requirement of a large enough sample size, although data efficiency should not be an issue in Big Data applications. In Feature Selection, FMMB proved to be a viable alternative to M3B, achieving comparable classification accuracy. In Causal Discovery, with enhanced structural learning accuracy, FMMB inspires the development of local-to-global MAG learning algorithms. The next steps towards better MMB discovery are to study the behaviour of algorithms when connection strength is varied and to increase robustness to unfaithful relationships.

ACKNOWLEDGEMENTS

We thank Vitor P. Ribeiro for helping us with the TikZ Picture of the Alarm Network.

REFERENCES

Alcalá-Fdez, J.e.a. (2011). Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17: 2-3, 255–287. URL `https://sci2s.ugr.es/keel/index.php`.

Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X.D. (2010a). Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(Jan), 171–234.

Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X.D. (2010b). Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: Analysis and extensions. *Journal of Machine Learning Research*, 11(Jan), 235–284.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-Based Systems*, 86, 33–45.

Colombo, D., Maathuis, M.H., Kalisch, M., and Richardson, T.S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 294–321.

Dua, D. and Graff, C. (2017). UCI machine learning repository. URL `http://archive.ics.uci.edu/ml`.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Koller, D. and Sahami, M. (1996). Toward optimal feature selection. Technical report, Stanford InfoLab.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Margaritis, D. and Thrun, S. (2000). Bayesian network induction via local neighborhoods. In *Advances in neural information processing systems*, 505–511.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco: Morgan Kaufmann.

Pena, J.M., Nilsson, R., Björkegren, J., and Tegnér, J. (2007). Towards scalable and data efficient learning of markov boundaries. *International Journal of Approximate Reasoning*, 45(2), 211–232.

Richardson, T., Spirtes, P., et al. (2002). Ancestral graph markov models. *The Annals of Statistics*, 30(4), 962–1030.

Spirtes, P., Glymour, C.N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Statnikov, A., Lytkin, N.I., Lemeire, J., and Aliferis, C.F. (2013). Algorithms for discovery of multiple markov boundaries. *Journal of Machine Learning Research*, 14(Feb), 499–566.

Tsamardinos, I. and Aliferis, C.F. (2003). Towards principled feature selection: relevancy, filters and wrappers. In *AISTATS*.

Tsamardinos, I., Aliferis, C.F., and Statnikov, A. (2003a). 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*, 673–678.

Tsamardinos, I., Aliferis, C.F., Statnikov, A.R., and Statnikov, E. (2003b). Algorithms for large scale markov blanket discovery. In *FLAIRS conference*, volume 2, 376–380.

Wang, H., Ling, Z., Yu, K., and Wu, X. (2020). Towards efficient and effective discovery of markov blankets for feature selection. *Information Sciences*, 509, 227 – 242. doi:https://doi.org/10.1016/j.ins.2019.09.010. URL `http://www.sciencedirect.com/science/article/pii/S0020025519308552`.

Yang, X., Wang, Y., Ou, Y., and Tong, Y. (2019). Three-fast-inter incremental association markov blanket learning

algorithm. *Pattern Recognition Letters*, 122, 73–78.

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

Yu, K., Guo, X., Liu, L., Li, J., Wang, H., Ling, Z., and Wu, X. (2019). Causality-based feature selection: Methods and evaluations.

Yu, K., Liu, L., Li, J., and Chen, H. (2018). Mining markov blankets without causal sufficiency. *IEEE transactions on neural networks and learning systems*, 29(12), 6333–6347.