Noname manuscript No.
(will be inserted by the editor)

Markov blanket discovery using minimum message length

First Author · Second Author · Third Author

Received: date / Accepted: date

Abstract **Keywords** Markov blanket · Markov boundary · Feature selection · Causal discovery · Bayesian network · Minimum message length

1 Introduction

The assumptions made throughout this paper (some of these are made by the MML principle):

1. The dataset is complete, with no hidden variables, discrete and *i.i.d.*.
2. The parameters are independent and follow symmetric Dirichlet distributions.

Local structre and local model refer to structure within a CPT. We may need to change the term to substructure or something.

F. Author
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: fauthor@example.com

S. Author
second address

2 Related work

The concept Markov boundary (a.k.a., Markov blanket, although mistakenly used) of a target variable was introduced by [Pearl, 1988] as the smallest subset of variables, conditioning on which the target is independent to the rest of the variables. Given it carries sufficient information about targets, Markov blankets are the optimal feature subsets for prediction [Koller and Sahami, 1996], [Cooper *et al.*, 1997], [Cheng *et al.*, 2001]. Assuming a set of variables can be modelled by a Bayesian network, then a variable's Markov blanket contains its direct neighbours and spouses (i.e., children's other parents). Because of this, researchers has been trying to reduce the complexity of learning a full Bayesian network by independently learning local structures within Markov blankets then stitching them together. A decent review is in [Aliferis *et al.*, 2010b] with a framework describes high-level steps of such local-to-global methodology.

A natural way of solving this problem, emerged from the conditional independence definition of Markov blankets, is testing dependences between a target and everything else given each subset. It is, however, infeasible to go through all subsets in practice because the number is exponential to the total number of variables. An early work on learning Markov blankets using heuristic and conditional independence test was done by [Margaritis and Thrun, 1999]. This work laid the foundation for the constraint-based Markov blanket discovery approach that typically consists of an admission and deletion phases. Ranking potential candidates according to conditional mutual information with the target, [Tsamardinos *et al.*, 2003b] improved the previous work by admitting variables with high dependences in advance. In despite of using the same statistical test and heuristic algorithm, another strategy that learns direct neighbours and spouses separately had proven to be superior, hence were widely adopted by later constraint-based methods [Aliferis *et al.*, 2003], [Tsamardinos *et al.*, 2003a], [Peña *et al.*, 2007], [Fu and Desmarais, 2008], [Aliferis *et al.*, 2010a],

[de Morais and Aussem, 2010], [Liu and Liu, 2016]. In particular, this strategy learns the target's distance one and two neighbours separately. The distance two neighbours are then filtered to remove false any positives. At the end of each neighbour learning process, the learned variables are enforced to satisfy the symmetry condition of Markov blankets (Proposition 1 in Section 3) in order to further remove false positives.

Since there had been extensive studies on how to search for an optimal Bayesian network from observational data, Markov blankets had also been learned with metric-based approach. The difference from learning a full Bayesian network is that the search space was restricted to a space of local (sub-) structures around a target variable without having to worry about unrelated adjacencies [Cooper *et al.*, 1997], [Madden, 2002], [Acid *et al.*, 2013]. Given most real models are sparse, Markov blankets of these models are considerable smaller. Hence, exact algorithms for learning small Bayesian networks could be applied to find optimal local structures independently. [Niinimaki and Parviainen, 2012] published the first exact Markov blanket learning algorithm and applied it to scale up exact Bayesian network learning. The method was primarily relied on a sub-routine made of dynamic programming and BDeu metric to find optimal local DAGs. Its symmetric enforcement was relaxed by [Gao and Ji, 2017] to reduce the time complexity.

Other than the approaches mentioned above, Markov blankets could also be learned in a similar fashion as wrapper feature selection methods. That is, potential Markov blankets were scored using predictive models such as decision tree [Frey *et al.*, 2003], linear causal model with LASSO estimator [Li *et al.*, 2004] and ridge regularized linear model [Strobl and Visweswaran, 2016].

3 Markov blanket

Firstly, we shall review some important concepts that will be used later on to define a Markov blanket. A *directed acyclic graph (DAG)* is a directed graph with no cycles (Figure!). We use $G = (X, E)$ to denote a DAG over a variable set $X = \{X_1, \ldots, X_n\}$ with a directed edge set E. We say X_i is a *parent* of X_j and X_j is a *child* of X_i if there is an arc $X_i \rightarrow X_j$ going from X_i to X_j. In addition, X_k is a *descendent* of X_i and X_i is an *ascendent* of X_k if there is a directed path from X_i to X_k.

Definition 1 Let P be a joint probability distribution of the random variables in X, and $G = (X, E)$ be a directed acyclic graph. We say (G, P) satisfies the *Markov condition* if for every variable $X_i \in X$, it is conditionally independent of its non-descendants ND_i given its parents set Π_i. That is,

$$X_i \perp\!\!\!\perp_P ND_i \mid \Pi_i.$$

Definition 2 Let P be a joint probability distribution of the random variables in X, and $G = (X, E)$ be a directed acyclic graph. We say $< G, P >$ forms a *Bayesian network* if it satisfies the Markov condition.

Definition 3 Let P be a joint probability distribution of the random variables in X, and $G = (X, E)$ be a directed acyclic graph. We say G *entails* the conditional independence $X_i \perp\!\!\!\perp_P X_j \mid X_k$, if for every joint probability distribution P such that (G, P) satisfies the Markov condition, $X_i \perp\!\!\!\perp_P X_j \mid X_k$ holds.

In practice a DAG may or may not entail all the conditional independences in a joint distribution, so the following two definitions are introduced.

Definition 4 A directed acyclic graph $G = (X, E)$ is called an *independence-map (or I-map)* of a joint probability distribution P, if G entails all the conditional independences in P.

Definition 5 A joint probability distribution P is said to be *faithful* to a directed acyclic graph $G = (X, E)$ if G entails all and only the conditional independences in P.

All Bayesian networks discussed in this paper are assumed to satisfy the faithfulness condition.

Definition 6 Let $G_1 = (X, E_1)$ and $G_2 = (X, E_2)$ be two directed acyclic graphs. Then G_1 and G_2 are *Markov equivalent* if and only if they entail the same conditional independences.

Definition 7 Let $< G = (X, E), P >$ be a Bayesian network. The *Markov blanket* of a variable X_i, denoted by MB_i, is the minimum subset of variables such that the following hold:

$$X_i \perp\!\!\!\perp_P X \setminus \{X_i, MB_i\} \mid MB_i$$

Assuming faithfulness, being the smallest conditioning set ensures the uniqueness of Markov blanket. Given a Bayesian network structure $< G = (X, E), P >$, a variable X_i's Markov blanket consists of its parents, children, and children's other parents (a.k.a., spouses). We use MB_i^G to emphasize the Markov blanket of X_i in the DAG G which is faithful to the joint distribution P.

Proposition 1 *Let $< G = (X, E), P >$ be a Bayesian network. For two distinct variables X_i and X_j, the following is satisfied*

$$X_j \in MB_i \Leftrightarrow X_i \in MB_j.$$

4 Minimum message length

Minimum message length (MML) was devised by [Wallace and Boulton, 1968] as a way of balancing the complexity of a statistical model H against the fit of the model

to a given dataset D. It relies on Bayes' theorem

$$p(H|D) = \frac{p(H,D)}{p(D)} = \frac{p(H) \times p(D|H)}{p(D)},$$

where $p(H)$ is the prior probability distribution of a model, $p(D|H)$ is the likelihood of a dataset given this model. In addition, it employs Shannon's information theory

$$I(E) = -log(p(E))$$

to measure the cost or information content (in nits if the log is natural) for stating an event of probability $p(E)$. Putting these together, the cost for stating a model and a dataset is a two-part message length

$$I(H,D) = I(H) + I(D|H). \tag{1}$$

The first part $I(H)$ measures the message length for stating a model (i.e. its structure and parameters up to a certain precision). The second part $I(D|H)$ measures how well the model compresses the given dataset. The aim in MML inference is to find the model having the shortest two-part message length. Throughout this paper, we use the natural log to calculate the MML score unless stated otherwise.

A feasible approximate method for calculating the total message length is known as *MML87* [Wallace and Freeman, 1987]. It approximates the two parts as follows:

$$I(H) = -ln(p(\boldsymbol{\theta})) + \frac{1}{2}ln(F(\boldsymbol{\theta})) + \frac{|\boldsymbol{\theta}|}{2}ln(\kappa_{|\boldsymbol{\theta}|}), \tag{2}$$

$$I(D|H) = -ln(p(D|H)) + \frac{|\boldsymbol{\theta}|}{2}. \tag{3}$$

For a given model with a parameter set $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$ specifies the parameter prior. The other terms in $I(H)$ give the precision of $\boldsymbol{\theta}$, where $F(\boldsymbol{\theta})$ is the determinant of the expected Fisher information matrix and $\kappa_{|\boldsymbol{\theta}|}$ are lattice constants [Wallace, 2005]. The $\frac{|\boldsymbol{\theta}|}{2}$ term in $I(D|H)$ is the extra cost of using an estimate with optimal limited

precision. (Note that a continuous datum, d, can only ever be measured to limited accuracy, $\pm\frac{\epsilon}{2}$, so it has not just a probability density, $f(d)$, but a proper probability, $f(d) \cdot \epsilon$, assuming that the pdf varies slowly around d.)

From equations 2 and 3, one is able to calculate the total message length if the determinant of the expected Fisher information matrix is calculable, and in particular one is interested in knowing the MML estimates of the parameters. Assuming a dataset D of N *i.i.d.* samples of a random variable comes from a multi-state distribution, the total message length to state the hypothesis and dataset can be calculated efficiently by

$$I(H, D) = \ln\left(\frac{(N + r - 1)!}{(r - 1)! \times \prod_{i=1}^{r} n_i!}\right). \tag{4}$$

It was presented in [Boulton and Wallace, 1969] as the factorial form of the multi-state MML, where the random variable takes r states and each state appears n_i times in D. Equation 4 was theoretically justified to be shorter than the *MML87* message length by a constant difference $\ln\frac{\pi e}{6}$ for each parameter, because it does not state the MML estimated parameters.

Definition 8 Let D be a dataset of N *i.i.d.* records sampled from a Bayesian network $< G = (X, E), P >$. A metric $I : \mathcal{G} \times \mathcal{D} \rightarrow \mathbb{R}^+$ is *decomposable* if it can be written as a sum of scores for each variable X_i given its parents set Π_i. That is,

$$I(G, D) = \sum_{X_i \in X} I(X_i | \Pi_i, D).$$

The benefit of being a decomposable metric makes it convenient to calculate the network structure score without having to deal with the full joint distribution. The second part of MML is the likelihood of a model which can be factorised into a product of individual variable's likelihood score. This is the same for other metrics like BDe, MDL, K2, etc. For Bayesian networks over discrete variables, MML assumes the parameters are independent and follows a uniform distribution (which is generalised

to symmetric Dirichlet distribution in the next section), so the parameter prior can be dealt individually for each variable. The structure prior is an overall assumption on all possible structures so has nothing to do with the joint distribution.

Definition 9 Let D be a dataset of N *i.i.d.* records sampled from a joint probability distribution P over a variable set X. Assuming $G = (X, E_1)$ and $G_2 = (X, E_2)$ are two different directed acyclic graphs. A metric $I : \mathcal{G} \times \mathcal{D} \rightarrow \mathbb{R}^+$ measures the information content for stating a model and the given dataset is *consistent* if the following hold: (what about equivalence class?)

1. if G_1 is an I-map of P and G_2 is not, then $\lim_{n\to\infty} I(G_1, D) < \lim_{n\to\infty} I(G_2, D)$,
2. if G_1 and G_2 are both I-maps of P and G_1 has less number of parameters than G_2, then $\lim_{n\to\infty} I(G_1, D) < \lim_{n\to\infty} I(G_2, D)$.

Proposition 2 *Under the assumptions stated in Section 1, MML is a consistent scoring function.*

Proof Given the models considered in this paper are discrete and have no hidden variables, they belong to the curved exponential family[Geiger *et al.*, 2001]. According to equation 2 and equation 3, the total message length can be expressed as

$$I(H, D) = -\left(ln(p(D|H)) - \frac{|\boldsymbol{\theta}|}{2} a_n\right), \text{ where}$$
$$a_n = 1 - \frac{2ln(p(\boldsymbol{\theta}))}{|\boldsymbol{\theta}|} + \frac{1}{|\boldsymbol{\theta}|} ln(F(\boldsymbol{\theta})) + ln(\kappa_{|\boldsymbol{\theta}|})$$

The only term in a_n is a function of n is the determinant of the expected Fisher information matrix. Each entry in the Fisher information matrix is the second derivative of the negative log likelihood, which grows linearly as $n \rightarrow \infty$. Hence, its determinant grows approximately in $|\theta| log n$. Consequently, $a_n \rightarrow \infty$ and $a_n/n \rightarrow 0$ as $n \rightarrow \infty$. By [Haughton, 1988], MML must be a consistent scoring function. □

Remark 1 [Haughton, 1988]'s result of consistent scoring function applies to both linear and curved exponential families. The linear exponential family contains undi-

rected graphical models that have no hidden variables [Geiger *et al.*, 2001]. The curved exponential family contains directed acyclic graphical models and chain graphs that have no hidden variables and with several families of local models (decision tree, noisy or, not sure about LR) to approximate a full CPT. For grahical acyclic models with hidden variables (e.g. Naive Bayes models???), [Geiger *et al.*, 2001] defined them as stratified exponential family and emphasized that [Haughton, 1988]'s argument does not apply for this family because some assumptions made by Haughton are not necessarily true in this family. Further investigation is needed to prove consistent when applying to this family.

By consistent and decomposable, one can prove that MML is a local consistent scoring function. This allows MML to find the optimal Markov blanket in the limit of infinite data.

Definition 10 Let D be a dataset of N *i.i.d.* records sampled from a probability distribution P over a variable set X. Assuming $G_1 = (X, E_1)$ and $G_2 = (X, E_2)$ are any two directed acyclic graphs such that $E_1 \cup \{X_i \rightarrow X_j\} = E_2$. A consistent metric $I : \mathcal{G} \times \mathcal{D} \rightarrow \mathbb{R}^+$ measures the information content for stating a model and the given dataset is *locally consistent* if the following hold:

1. if $X_i \not\perp\!\!\!\perp_P X_j \mid \Pi_j^{G_1}$, then $\lim_{n\to\infty} I(G_2, D) < \lim_{n\to\infty} I(G_1, D)$,
2. if $X_i \perp\!\!\!\perp_P X_j \mid \Pi_j^{G_1}$, then $\lim_{n\to\infty} I(G_2, D) > \lim_{n\to\infty} I(G_1, D)$,

where $\Pi_j^{G_1}$ is the parents set of X_j in G_1.

Proposition 3 *Assumptions must be satisfied. MML is a locally consistent scoring function.*

Proof Since G_1 is any DAG over X, there must exist a DAG $G_1' = (X, E_1')$ such that $\Pi_j^{G_1} = \Pi_j^{G_1'}$ and $G_1^c = (X, E_1' \cup \{X_i \rightarrow X_j\})$ is a complete DAG. If $X_i \not\perp\!\!\!\perp_P X_j \mid \Pi_j^{G_1}$, then G_1^c must be an I-map of P whilst G_1' is not. Being a decomposable and consistent

metric, it implies $\lim_{n\to\infty} I(G_1, D) - \lim_{n\to\infty} I(G_2, D) = \lim_{n\to\infty} I(G_1', D) - \lim_{n\to\infty} I(G_1^c, D) = d > 0$.

If $X_i \perp\!\!\!\perp_P X_j \mid \Pi_j^{G_1}$, G_1 may or may not be an I-map of P. If it is not, the above argument applies. If G_1 is an I-map of P, the proposition is a consequence of MML being consistent assuming both models' parameters are stated to the same precision. □

5 LEARNING Markov blanket using MML

Assuming one is interested in finding a set of predictors in order to forecast the future value of a target variable. There could be non-unique solutions to this problem, so certain assumptions need to be made to enable variable subsets can be ranked according to a metric. MML assumes Occam's razor - an inductive bias - which prefers a simpler hypothesis among all competing hypotheses that give the same answer. Since there are exponential number of subsets, one could heuristically searching for optimal predictors according to the MML score of a model. The following sections discuss three different model classes, namely the conditional probability table (CPT) model class, the Naive Bayes (NB) model class and the class of Markov blanket polytrees (MBPs).

5.1 MML for conditional probability table model

For a variable $X_i \in X$, its probability density function conditioning on the full joint distribution of its parents set Π_i can be expressed by a $r_i \times r_{\Pi_i}$ conditional probability table (CPT), where r_i and r_{Π_i} are the number of states of X_i and Π_i respectively. We use a CPT model to describe the relation between a target and some input variables by treating those variables as if they are all parents of the target although it is not claiming they actually are all parents. Being the most general model, a full CPT can

capture any pairwise interactions between the inputs as long as there are enough data to support the exponential growth of its total parameters. There are work on learning restricted local models instead of the full CPT [Neil *et al.*, 1999] but they are not under the consideration of this paper. We use $\phi_i(S)$ to denote the CPT model of X_i with a subset $S \subseteq X$ being the hypothetical parents set of X_i.

The parents instantiations partition X_i into several multi-state distributions. By the parameter independent assumption, the message length of a CPT model is a sum of the message length of each multi-state distribution over all r_{Π_i} partitions. Assuming the parameters follow symmetric Dirichlet distributions, the multi-state MML can be generalised to adapte a variety of prior. Hence, the total message length for stating a CPT model $\phi_i(S)$ and the given dataset D over ϕ_i is

$$I(\phi_i(S), D_{\phi_i}) = \sum_{j=1}^{r_{\Pi_i}} \ln\left(\frac{(n_j + \alpha_0 - 1)! \prod_{k=1}^{r_i}(\alpha_k - 1)!}{(\alpha_0 - 1)! \prod_{k=1}^{r_i}(n_{jk} + \alpha_k - 1)!}\right) + \frac{r_{\Pi_i}(r_i - 1)}{2} \ln\frac{\pi e}{6}, \quad (5)$$

where $\boldsymbol{\alpha}$ is the symmetric Dirichlet concentration parameters for X_i such that $\alpha_0 = \sum \boldsymbol{\alpha}$ and $\alpha_k \in \boldsymbol{\alpha}$ corresponds to the k^{th} state of X_i, n_{jk} being the count for Π_i in state j and X_i in state k, and $n_j = \sum_{k=1}^{r_i} n_{jk}$. Noticing the last term in equation 5 may be omitted if one does not MML estimate of the model paremters.

Assuming only a CPT model $\phi_i(S)$ is used to transmit a dataset, the next proposition proves that the shortest encoding length by MML principle occurs when $S = MB_i$ in the limit of infinite data.

Proposition 4 *Let D be a dataset with N i.i.d. records sampled from a joint probability distribution P over variables $X = \{X_1, \ldots, X_n\}$. The MML score for stating a CPT model of X_i and the given dataset must satisfy the following:*

$$\lim_{n\to\infty} I(\phi_i(MB_i), D_{\phi_i}) < \lim_{n\to\infty} I(\phi_i(S), D_{\phi_i}), \forall S \subseteq X \text{ s.t. } S \neq MB_i.$$

Proof Assuming there exists $S \subseteq X$ such that $S \neq MB_i$ and $\lim_{n\to\infty} I(\phi_i(S), D_{\phi_i}) < \lim_{n\to\infty} I(\phi_i(MB_i), D_{\phi_i})$. Let $G_1 = (X, E_1)$ and $G_2 = (X, E_2)$ be two DAGs such that all variables have the same parents sets except X_i such that $\Pi_i^{G_1} = MB_i$ and $\Pi_i^{G_2} = S$.

The local consistent of MML entails that any arc addition and deletion that introduces pairwise dependence and independence which does not exist in the joint distribution P will increase the total message length. Hence, $\lim_{n\to\infty} I(G_2, D) > \lim_{n\to\infty} I(G_1, D)$. Because all other variables have the same parents set, by decomposable we have $\lim_{n\to\infty} I(G_2, D_{\phi_i}) > \lim_{n\to\infty} I(G_1, D_{\phi_i})$ which contradicts the assumption. □

5.2 MML for Naïve Bayes model

Naïve Bayes allows input variables to have marginal dependences, but they become independent from each other once the target variable is given. It is often drawn as a graphical model with the target variable being the only parent of all the input variables (Figure!!!). Similar as a CPT model, a NB model treats input variables as if they are all children of the target although it is not claiming that they actually are all children. With the conditional independence assumption, Naïve Bayes' parameters increase almost linearly which gives it a great scalability in large samples. Another benefit of having this assumption is the factorization of the joint density $p(X_i, S)$ into a product of conditional probabilities

$$p(X_i|S) = \frac{p(X_i)\prod_{X_j\in S} p(X_j|X_i)}{\sum_{x_i=1}^{r_i} p(x_i)\prod_{X_j\in S} p(X_j|x_i)}, \tag{6}$$

where $p(x_i)$ is a short hand for $p(X_i = x_i)$. Each of $p(X_j|X_i)$ and $p(X_i)$ can then be calculated using the adaptive code approach. Hence, the total message length for stating a Naïve Bayes model and the given dataset over it is

$$I(\phi_i(S), D_{\phi_i}) = -\sum_{D_{\phi_i}}\Big[\ln p(X_i) + \sum_{X_j\in S} \ln p(X_j|X_i) - \ln \sum_{x_i=1}^{r_i} p(x_i)\prod_{X_j\in S} p(X_j|x_i)\Big] \tag{7}$$

Noticing this again is a message length without transmitting the MML estimate of the parameters. It is unknown at this stage what the difference for a Naïve Bayes model would be between the adaptive and the *MML87* approaches. Ideally such a difference could be obtained by comparing equation 7 with the *MML87* message length. This is left as a future research problem. A drawback of assuming a Naïve Bayes is the model's inability to represent any interactions among input variables, such as an exclusive-or (XOR).

5.3 MML for Markov blanket polytree model

This section presents an ensemble method for Markov blanket discovery. It operates by sampling as many as possible local structures as possible, then outputs a weighted average message length over all samples. This way the input variables can be tested under different network structures, so that the number of model parameters on average is less than exponential while the interactions between input variables can still be modelled.

To reduce the super-exponential DAG space, we focus on restricted local structures that assume each input variable is related to a target by being either its parent, child or spouse. The restricted local structures are called Markov blanket polytrees (MBPs). A polytree is a DAG such that its underlying graph is a tree.

Definition 11 Let $< G = (X, E), P >$ be a Bayesian network. A *Markov blanket polytree* T_i of a target variable X_i is a polytree over the variables $\{X_i\} \cup MB_i$ such that

$$MB^{T_i}(X_i) = MB^G(X_i).$$

The next proposition presents a recursive formula for counting the number of labelled Markov blanket polytrees (MBPs) over a set of n input variables.

Proposition 5 *Let Y be a variable whose Markov blanket contains $n \in [1, \infty)$ variables. The number of labelled Markov blanket polytrees of Y can be computed by the following recursive equation*

$$f(n) = \sum_{i=0}^{n} \binom{n}{i} + \sum_{m=1}^{\lfloor \frac{n}{2} \rfloor} \sum_{k=1}^{n-2m+1} g(n, m, k), \tag{8}$$

where

$$g(n, m, k) = \binom{n}{k+1} (k+1) \sum_{k'=1}^{\min\{k, n-k-2(m-1)\}} \frac{q}{m} \cdot g(n-k-1, m-1, k').$$

Proof It is trivial to bound the number of colliders $m \in \left[0, \left\lfloor \frac{n}{2} \right\rfloor\right]$.

Case 1: When $m = 0$

MB_i contains only parents and/or children. There are $\binom{n}{i}$ ways of selecting $i \in [0, n]$ children from n labelled nodes. The order of these parents or children does not matter in a polytree. Therefore, the number of labelled MBPs when $m = 0$ is

$$\sum_{i=0}^{n} \binom{n}{i}. \tag{9}$$

Case 2: When $m > 0$

Each of Y's children and its spouses (if there are any) forms a branch. The largest branch with k spouses can be enumerated in

$$\binom{n}{k+1} (k+1) \tag{10}$$

ways, where $k \in [1, n-2m+1]$. There are $\binom{n}{k+1}$ ways selecting $k+1$ nodes to form the largest branch. And each one of the $k+1$ nodes needs to be a common child once to fully enumerate all cases. k's upper bound is obtained if each of the other $m-1$ branches contains only a collider and a spouse, in which case $n - 2(m-1) - 1 = n - 2m + 1$. Hence, when $m > 0$ the number of MBPs can be obtained by multiplying

equation (10) with the total enumeration of the remaining $n-k-1$ nodes. The subgraph over the remaining nodes can be counted by the same approach. By doing this recursively, we will end up with a subgraph in which Y has no spouse. It can then be enumerated by equation (9). Therefore, the total enumeration of MBPs when $m > 0$ is

$$\sum_{m=1}^{\lfloor \frac{n}{2} \rfloor} \sum_{k=1}^{n-2m+1} g(n,m,k), \tag{11}$$

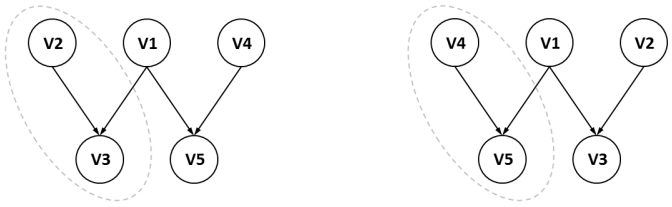
where

$$g(n,m,k) = \binom{n}{k+1}(k+1) \sum_{k'=1}^{\min\{k,n-k-2(m-1)\}} \frac{q}{m} \cdot g(n-k-1, m-1, k'), \tag{12}$$

where $q = 1$ if $k = k'$ and m otherwise. The maximum number of spouses k' in a subgraph is bounded above by the minimum between the maximum number $n-k-2(m-1)$ of available nodes and k from its supergraph.

As the largest branch is enumerated independently from the remaining nodes, we double count the case when $k = k'$. For example, we obtain Figure 1a when labelling the largest branch (i.e., left/right) with $\{V2, V3\}$, and Figure 1b when labelling the largest branch (i.e., left/right) with $\{V4, V5\}$. The resulting two labelled graphs, however, are identical and hence we divide the total number by $\frac{1}{2}$. For general cases, the total number needs to be divided by $\frac{1}{m}$, which is why we have the $\frac{q}{m}$ in equation (12). □

The total number of Markov blanket polytrees (MBPs) is dramatically reduced comparing with DAGs as shown in Table 1.



(a) 2a (b) 2b

Fig. 1: Example of two duplicated Markov blanket polytrees when enumerating.

Table 1: The number of labelled DAGs and MBPs on $n \in [0, 7]$ nodes.

# nodes	# DAGs	# MBPTs
1	1	1
2	3	2
3	25	6
4	543	23
5	29281	104
6	3781503	537
7	1138779265	3100

The message length for transmitting data using a MBP model is calculated as the natural log of the conditional probability

$$p(X_i|S) = \frac{p(X_i \mid \Pi_i^{T_i}) \prod_{X_j \in S} p(X_j|\Pi_j^{T_i})}{\sum_{x_i=1}^{r_i} p(x_i \mid \Pi_i^{T_i}) \prod_{X_j \in S} p(X_j|\Pi_j^{T_i})}$$

is factorized into a product of each variable's probability conditioning on its parents set in a Markov blanket polytree T_i, which can be estimated from data using the adaptive code method. Hence, the total message length has the form

$$\begin{aligned} I(\phi_i(S), D_{\phi_i}) = -\sum_{D_{\phi_i}} \Big[\ln p(X_i \mid \Pi_i^{T_i}) + \sum_{X_j \in S} \ln p(X_j|\Pi_j^{T_i}) - \\ \ln \sum_{x_i=1}^{r_i} p(x_i \mid \Pi_i^{T_i}) \prod_{X_j \in S} p(X_j|\Pi_j^{T_i}) \Big]. \end{aligned} \tag{13}$$

To calculate a weighted average score, we uniformly average the conditional probabilities $p(X_i \mid S)$ over all possible MBPs $\mathcal{T}_i$ containing the same variables $\{X_i\} \cup MB_i$,

then take the negative log of the expected probability. The uniform prior can be replaced by any reasonable prior over the possible Markov blanket polytrees.

5.4 Pseudo-code of the MBMML algorithm

This section presents the pseudo-code of two types of the MBMML algorithm that learns Markov blanket of a target variable using either a fixed local structure (i.e., CPT or NB) or an ensemble of random local structures (i.e., MBPs). Both algorithms use a greedy search starting with empty Markov blanket and iteratively adding the highest ranked candidate into the Markov blanket to reduce the total message length calculated by a MML metric (equation 5 or 7, or 13). Both algorithms stop and output a learned Markov blanket if no scores can be increased by adding more candidates. Algorithms 1 and 2 outline steps of the fixed and ensemble methods.

Algorithm 1 MB discovery using MBMML+CPT/NB

procedure $MBMML(X_i, X, D, \phi_i)$, where X_i is the target variable, X is the set of all variables, D is a given dataset and ϕ_i is fixed to be either a CPT or NB model.
  $S = X \setminus X_i$ ▷ unchecked variables
  $Z = \emptyset$ ▷ learned MB
  $L = I(\phi_i(\emptyset), D_{\phi_i})$ ▷ empty model score
  while $S \neq \emptyset$ **do**
    $X_k = \arg\min_{X_j} I(\phi_i(Z \cup \{X_j\}), D_{\phi_i}), \forall X_j \in S$ ▷ best candidate
    $L' = I(\phi_i(Z \cup \{X_k\}), D_{\phi_i})$ ▷ current best score
    if $L' < L$ **then** ▷ admit when score reduces
      $Z = Z \cup \{X_k\}$
      $S = S \setminus \{X_k\}$
      $L = L'$ ▷ update best score
    else
      Stop
    end if
  end while
  Output Z
end procedure

Algorithm 2 MB discovery using MBMML+ENSEMBLE

procedure *MBMML*(X_i, X, D, ϕ_i, K), where X_i is the target variable, X is the set of all variables, D is a given dataset, ϕ_i is a MBP model, K is the number of randomly sampled MBPs.
  $S = X \setminus X_i$ ▷ unchecked variables
  $Z = \emptyset$ ▷ learned MB
  $L = I(\phi_i(\emptyset), D_{\phi_i})$ ▷ empty model score
  while $S \neq \emptyset$ **do**
    if $f(|Z| + 1) \leq K$ **then** ▷ number of MBPs by equation 8
      $\mathcal{T}_i :=$ {all MBPs over $Z \cup \{X_j\}$} ▷ all MBPs
    else
      $\mathcal{T}_i =$ {K random MBPs over $Z \cup \{X_j\}$} ▷ randomly sampled MBPs
    end if
    $X_k = \arg\min_{X_j} E_{\mathcal{T}_i}(I(\phi_i(Z \cup \{X_j\}), D_{\phi_i})), \forall X_j \in S$ ▷ best candidate
    $L' = E_{\mathcal{T}_i}(I(\phi_i(Z \cup \{X_k\}), D_{\phi_i}))$ ▷ current best expected score
    if $L' < L$ **then** ▷ admit when score reduces
      $Z = Z \cup \{X_k\}$
      $S = S \setminus \{X_k\}$
      $L = L'$ ▷ update best score
    else
      Stop
    end if
  end while
  Output Z
end procedure

To ensure there is no conflict among the learned Markov blankets so that they can be used later for structure learning, we enforced outputs from both MBMML+CPT/N and MBMML+ENSEMBLE algorithms to satisfy the symmetry property as shown in section 3. There are two deterministic enforcements, by union or intersection between two learned Markov blankets. The process of the symmetry enforcement is shown in Algorithm 3. Throughout these experiments we applied the UNION enforcement to MBMML+CPT, because a CPT model's precision converges to 1 as sample size increases. So its exponential increase in parameters is likely to result in more false negatives than false positives. The INTERSECTION enforcement was conducted on MBMML+NB, because a Naïve Bayes model intend to produce more false positives than a CPT model due to its lack of explanation power, but less false

negatives because of its simplicity. It is not clear which enforcement is a better option for MBMML+ENSEMBLE, so we chose the UNION enforcement throughout the experiments.

Algorithm 3 Symmetry enforcement

procedure Given the learned Markov blankets $\{MB_i\}, \forall X_i \in X$
  for each MB_i **do**
    for each $X_j \in MB_i$ **do**
      if $X_i \notin MB_j$ **then**
        if UNION **then**
          $MB_j = MB_j \cup \{X_i\}$
        else INTERSECTION
          $MB_i = MB_i \setminus \{X_j\}$
        end if
      end if
    end for
  end for
  Output $\{MB_i\}$
end procedure

6 Experiments on Markov blanket discovery

In this section, we present some of the experimental results on comparing different Markov blanket learners. All three algorithms, MBMML+CPT, MBMML+NB, and MBMML+ENSEMBLE were tested against the following algorithms:

- IAMB - a constraint-based algorithm that uses conditional independence test for candidate admission. It starts admitting variables that are dependent with the target conditioning on the current found Markov blanket. It is followed by a backward phase trying to delete any false positives.
- PCMB - a constraint-based algorithm that divides a learning into two sub-tasks. Firstly, it finds the direct neighbours of the target. It then finds the neighbours of each neighbour of the target, and prunes any false positives. This algorithm also relies on conditional independence test.

– SLL - a metric-based algorithm that uses dynamic programming and BDeu score. It is an exact algorithm that searches through the entire space of equivalence classes of local DAGs around a target, then reads off the optimal Markov blanket. Notice SLL does not just learn Markov blankets. It is for scaling up Bayesian network structure learning.

We used the implementations of IAMB and PCMB provided by [Peña *et al.*, 2007] and setted the significant level $\alpha = 0.05$ for conditional independence test. We used SLL's source code provided by [Niinimaki and Parviainen, 2012] and its default equivalent sample size 1 for BDeu. SLL fallbacks to GES algorithm [Chickering, 2002] if it tends to find more than 20 variables for a Markov blanket. The three MML methods assumed uniform parameter prior (i.e., symmetric Dirichlet with concentration parameter $\boldsymbol{\alpha} =< 1 >$). The MBMML+ENSEMBLE algorithm was setted to randomly sample 100 Markov blanket polytrees from the entire space and uniformly averaged their message lengths.

Section 6.1 focuses on testing with real models (Table 2)and standard datasets provided by [1]. The sample size varies in $500, 1000, 5000$ and each size comes with 10 different datasets. These models are often used for testing Markov blanket and causal discovery learners. Section 6.2 extends the experiments to artificial Bayesian networks (2) containing 30 and 50 variables respectively, and the same maximum fan-in and number of states for each variable. For each model specification, we randomly generate 5 different Bayesian networks, each of which was then used to generate 5 different datasets for every one of the sample sizes $100, 500, 2000, 5000$.

The evaluation metrics used are precision, recall, and edit distance. The edit distance between a true Markov blanket and a learned Markov blanket is equal to the sum of false positives and false negatives. The average accuracy over all variables in one structure and all samples with the same size were reported with 95% confidence

[1] `http://www.dsl-lab.org/supplements/mmhc_paper/mmhc_index.html#complete_data`

Table 2: Summary of tested Bayesian networks. 30-5-4-1 and 30-5-4-1 refer to artificial networks with 30 and 50 variables, maximum fan-in 5, maximum number of states 4, and uniform ($\alpha =< 1 >$) parameter prior.

Network	Number of odes	Max fan-in	Mean MB size
CHILD	20	2	3
INSURANCE	27	3	5.19
ALARM	37	4	3.51
BARLEY	48	4	5.25
HAILFINDER	56	4	3.54
30-5-4-1	30	5	8
50-5-4-1	50	5	9.73

intervals. Due to space limit, only selected models were discussed in the main text. The rest of the results are included in the appendix.

6.1 Accuracy on real models

Figures 2 and 3 report the mean edit distance (with confidence intervals) of all algorithms on the CHILD and BARLEY networks. In both cases, IAMB has the lowest accuracy given any sample size. In most cases, the other algorithms show no statistically significant difference from each other, except PCMB and MML+ENSEMBLE are less robust under small samples. Both PCMB and SLL have shown a projection of faster convergence than MML methods and IAMB. Noticing that PCMB failed to run on the BARLEY network possibly due to an implementation error. The edit distance and precision and reacall on all five models are summarized in Table 4 and Table 8 (in appendix) respectively. In general, on these real networks SLL has most of the winnings followed, MBMML+CPT, MBMML+NB, PCMB, MBMML+ENSEMBLE and IAMB.

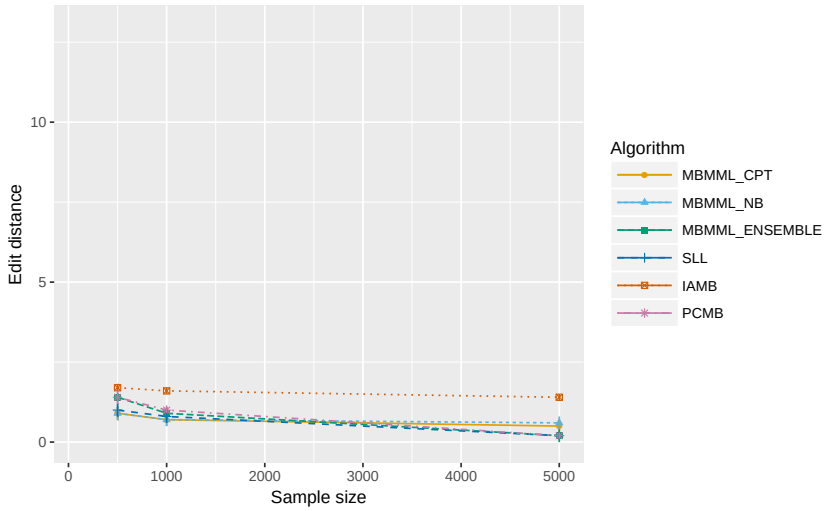


Fig. 2: Edit distance (with 95% confidence intervals) v.s. sample size on CHILD network.

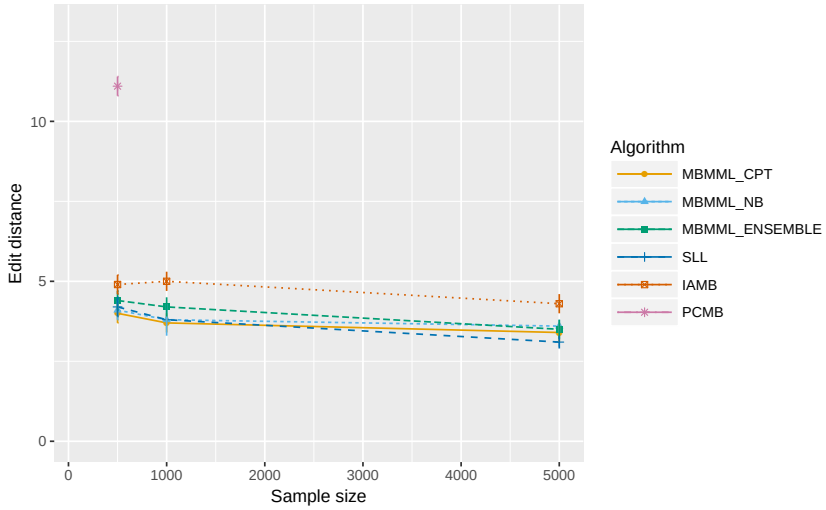


Fig. 3: Edit distance (with 95% confidence intervals) v.s. sample size on BARLEY network. PCMB failed under 1000 and 5000 samples possibly due to an implementation error.

6.2 Accuracy on artificial models

To test all metods on different problems over a number of random networks, we generated 5 different Bayesian networks for each pre-specified model specification. These networks are moderate in the number of nodes, and slightly larger comparing with the real networks used before in maximum fan-in and average Markov blanket size. Their parameters were sampled from uniform distribution which matches the parameter prior used in the multi-state MML metric. This may be an advantage of the MML methods for these particular problems, but it is less risky when nothing is known about the models. Later in this section, it has been shown briefly that this uninformative prior could produce similar accuracy as using the true prior.

Figure 4 and 5 have shown how the tested methods reacted to different sample sizes on artificial networks with specification 30-5-4-1 and 50-5-4-1. The MBMML+NB and IAMB algorithms started well under extreme small samples but could not converge given more samples. This is likely to be caused by the lack of representation power of Naive Bayes models and the unsoundness of IAMB. The MBMML+CPT and MBMML+ENSEMBLE algorithms have shown competitive performances under both small and large samples, and superior low edit distance for moderate samples, but both methods converged slower than PCMB and SLL towards 5000 samples. The main issue is believed to be the exponential number of parameters in both CPT and MBPT models. Both PCMB and SLL converged faster than the others like in the real model cases. The former, however, had the worst accuracy under 100 samples whilst the later are almost no difference from the others.

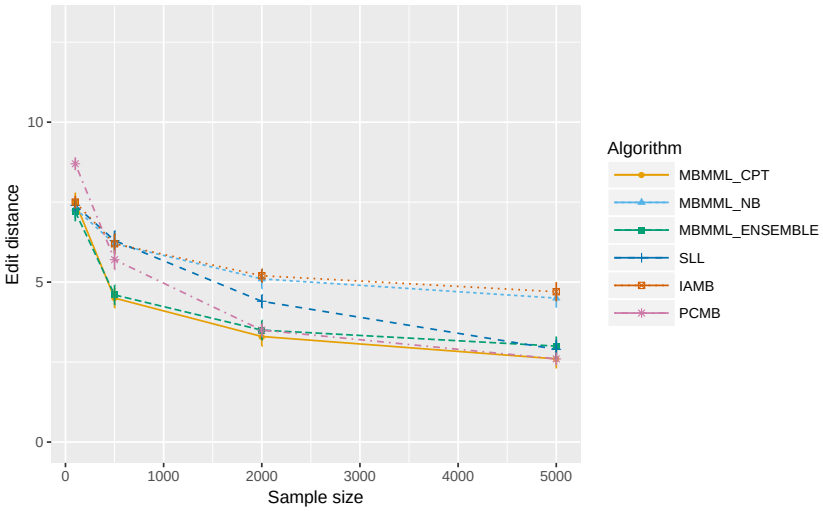


Fig. 4: Edit distance (with 95% confidence intervals) v.s. sample size on artificial Bayesian networks (30-5-4-1) containing 30 variables, maximum 5 parents and maximum 4 states for each variable.

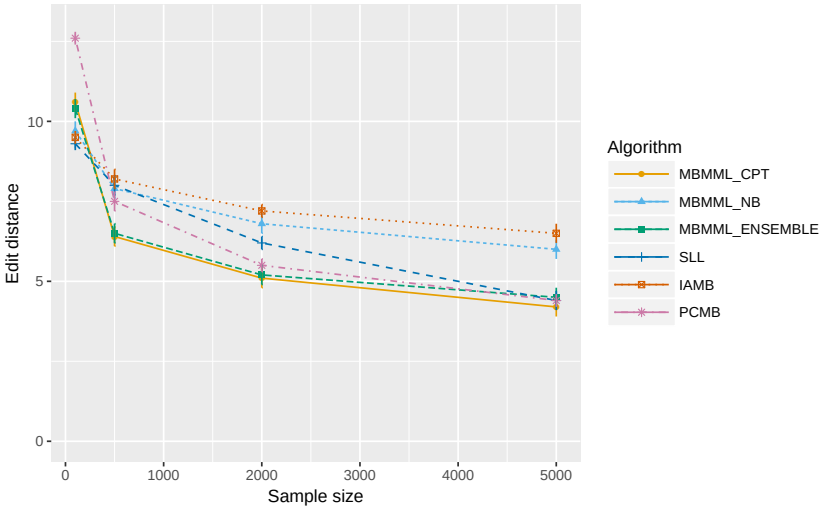


Fig. 5: Edit distance (with 95% confidence intervals) v.s. sample size on artificial Bayesian networks (50-5-4-1) containing 50 variables, maximum 5 parents and maximum 4 states for each variable.

So far, we have investigated the overall performance of Markov blanket learners given some networks. Now, we group together problems having similar complexity and look for the trend when increasing Markov blanket size. Figure 6 and 7 are edit distances of all methods for different Markov blanket size given 500 and 5000 samples respectively. The rest of the results are contained in the appendix.

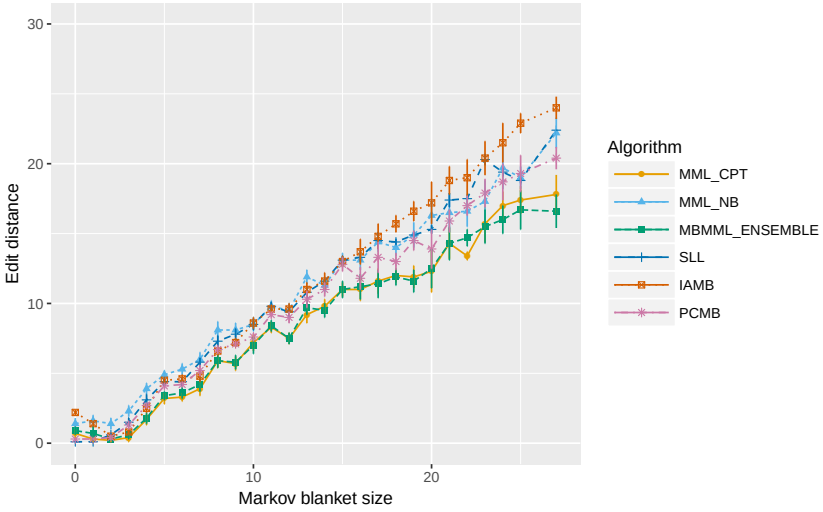


Fig. 6: Edit distance against Markov blanket size on 50-5-4-1 models with 500 samples.

Given 500 samples (Figure 6) for 50 variables networks, PCMB and SLL did well by identifying unconnected variables, whilst the others more or less had some false positives. As Markov blanket size was increased, the MBMML+CPT and MBMML+ENSEMBLE algorithms produced the fewest false findings all the way up to the maximum Markov blankets. It is worth noticing that for Markov blankets over 20 variables, it is GES's performance instead of SLL. On averaged, MBMML+CPT and MBMML+ENSEMBLE have the lowest edit distance which is consistent with the ranking in Figure 5 for 500 samples.

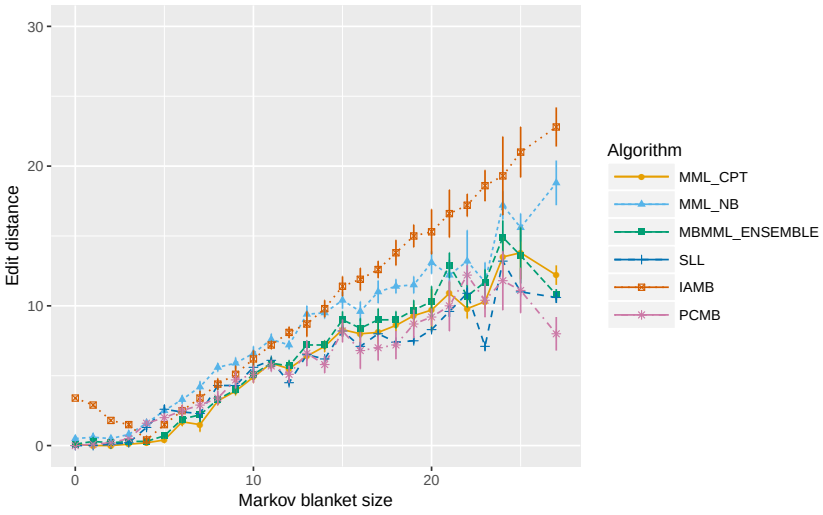


Fig. 7: Edit distance against Markov blanket size on 50-5-4-1 models with 5000 samples.

Given 5000 samples (Figure 7), most methods did well for Markov blanket size smaller than 4 except IAMB, who had a wired U-shape where the bottom is at 4 variables. MBMML+CPT and MBMML+ENSEMBLE kept the lowest edit distance for medium Markov blankets, but were unable to keep it low as the learning problems became more complex so being outperformed occasionally by PCMB and SLL. In summary, the two MML methods except MBMML+NB have lowest error rate for medium-sized learning problems given moderate samples.

There may be a concern that if the generating models used a non-uniform prior, the MML methods could produce different results. In principle, if the true prior is not uniform, using uniform prior would give no better results than uing the true prior. In practice, however, it is also depends on the quality and size of samples. To show the impact of using uninformative prior, MBMML+CPT was given both the true prior and uniform prior then tested on a 30-5-4-1 network whose parameters were sampled from a symmetric Dirichlet distribution with different concentration

parameter $\alpha \in \{0.1, 0.4, 0.7, 1, 10, 40, 70, 100\}$. The experiments were completed for 500 and 5000 samples.

Figure 8 and 9 have shown insiginificant difference between the use of true prior and uniform prior for the case when the concentration parameters $\alpha \leq 1$. This is because adding small α values to parameter estimations almost do not matter when there exists at last some data for each parameter. When $\alpha > 1$, uniform prior produced slightly less edit distance than the true prior. By looking into the number of learned Markov blanket size, we notice that as α increases the learned Markov blankets also increase in size. It is believed that this is caused by the same reason as the sensitivity of BDeu as shown by [Silander *et al.*, 2007]. The MML metric for CPT model with symmetric Dirichlet prior is similar as the BDeu metric, except the former includes costs for stating estimated parameters and used uniform Dirichlet prior over all model parameters whilst the later does not state parameters and only used uniform Dirichlet prior over parameters of each node. But both metrics penalise model complexity using a function of α, which decreases as α increases. Hence, under large α the MML methods discovered larger Markov blankets which could contain a large propotion of false positives especially under small samples. This is also why when using 5000 samples, the true prior produced similar edit distance as uniform prior.

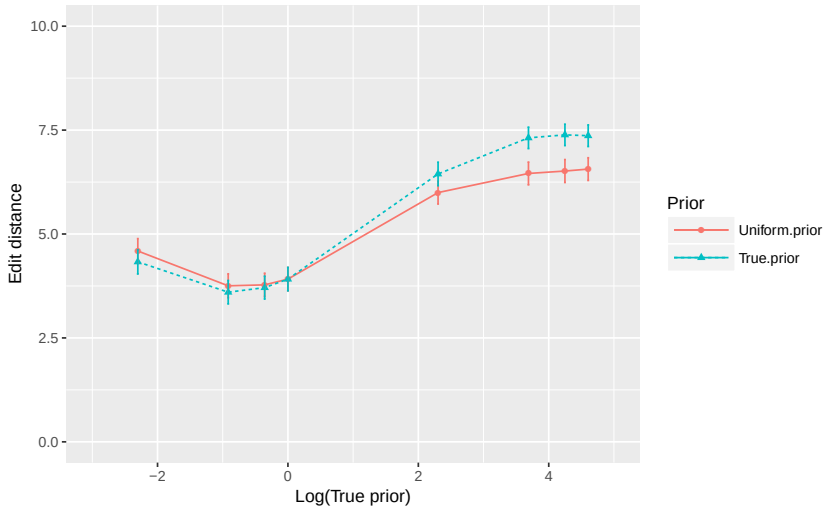


Fig. 8: MBMML+CPT's edit distances using the true prior and uniform prior on a 30-5-4-1 model with 500 samples. The X-axis is the natural log scale of the true symmetric Dirichlet concentration parameter $\alpha = \{0.1, 0.4, 0.7, 1, 10, 40, 70, 100\}$.

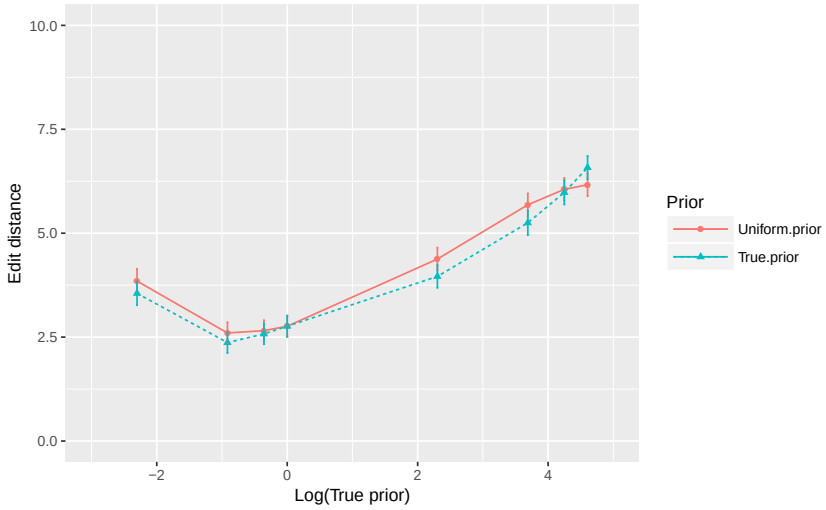


Fig. 9: MBMML+CPT's edit distances using the true prior and uniform prior on a 30-5-4-1 model with 5000 samples. The X-axis is the natural log scale of the true symmetric Dirichlet concentration parameter $\alpha = \{0.1, 0.4, 0.7, 1, 10, 40, 70, 100\}$.

Table 3: Algorithm's computational complexity in big O notation.

Algorithm	Big O notation
IAMB	$O(n^2)$
MBMML+NB	$O(n^2)$
MBMML+CPT	$O(n2^{n-1})$
MBMML+RANDOM	$O(n2^{n-1})$
PCMB	$O(n^2 2^{n-1})$
SLL	$O(n^4 2^n)$

6.3 Algorithm complexity

Table 3 orders all algorithms by ascending computational complexity. The while loop in Algorithm 1 runs at most $n-1$ times. Each time it runs through all unchecked nodes to find the best candidate according to a MML metric. For a CPT model, there could be at most $n-1$ parents in which case the multi-state MML is summed over all 2^{n-1} parents instantiations. So the computational complexity of the MBMML+CPT algorithm is $O(n2^{n-1})$. For a NB model, the worst case is when all $n-1$ nodes are children of the target, which is linear in n within the *WHILE* loop, so gives a complexity $O(n^2)$. The worst case of MBMML+RANDOM is when a CPT model appears in the sampled Markov blanket polytrees. In general, a random model is slower than a CPT by a constant factor, which is determined by the number of sampled local structures. But it has the same complexity as a CPT model. For PCMB, the total time required is dominated by the process of finding the direct neighbours of the target. This process tries to find a subset of the neighbour set, conditioning on which the target is independent with a candidate. And such a process runs through all variables to ensure the symmetry property. Hence, its complexity in the worst case is $O(n^2 2^{n-1})$. The total time required by IAMB and SLL were published in the associated papers.

7 Conclusion

This paper presented three MML methods for learning Markov blankets of variables. The three methods are all built based on the multi-state MML metric, but associated with different models that are CPT, Naive Bayes and an ensemble of random Markov blanket polytree models. The MBMML+CPT algorithm was proven to be correct under infinite samples, though may not be data efficient for large Markov blankets due to exponential number of parameters. The MBMML+NB algorithm was developed to overcome the problem of exponential model complexity by sacrificing the the modelling of variable interactions. Both methods have no tendency of learning the subgraphs within Markov blankets, although certain model structures are assumed during learning. Realizing it is restricted to consider only a particular model structure for learning Markov blankets, we generalised to the MBMML+ENSEMBLE algorithm that assesses the validity of a variable being in a Markov blanket under random polytrees within Markov blankets. The search algorithms used are greedy search but can be replaced with any heuristics or probabilistic searches. The three MML algorithms were tested against several other Markov blanket algorithms on both real and artificial Bayesian networks given different sample sizes. In general, MBMML+CPT and MBMML+ENSEMBLE show superior results given moderate samples than the others. In particular, they had the lowest edit distance when learning medium-sized Markov blankets. The method with Naive Bayes model occationally appeared to be competitive for very small samples but could not converge given more samples. Comparing with the approximate algorithm PCMB, the two MML methods have advantages in both accuracy and efficiency, especially under small and medium samples. Comparing with the exact learning algorithm SLL, they are competitive in accuracy and faster in computational complexity by a factor of n^3.

Table 4: Summary of edit distance (with 95% confidence intervals) of all Markov blanket discovery algorithms on both real and artificial Bayesian networks. The best results are highlighted in grey. In real networks, SLL wins most of the times followed by MBMML+CPT, MBMML+NB, PCMB, MMLL+ENSEMBLE, IAMB. PCMB failed to learn on BARLEY networks under 1000 and 5000 samples possibly due to an implementation error. In artificial networks, MBMML+CPT and MBMML+ENSEMBLE win most of the times followed by SLL, PCMB, MBMML+NB, IAMB.

Network	SAMPLES	*MBMML +CPT*	*MBMML +NB*	*MBMML +ENSEMBLE*	IAMB	PCMB	SLL
CHILD	500	0.9+-0.2	0.9+-0.2	1.4+-0.2	1.7+-0.2	1.4+-0.2	1+-0.2
	1000	0.7+-0.1	0.7+-0.2	0.9+-0.2	1.6+-0.2	1+-0.2	0.8+-0.1
	5000	0.5+-0.1	0.6+-0.1	0.2+-0.1	1.4+-0.2	0.2+-0.1	0.2+-0.1
INSURANCE	500	3.3+-0.2	3.5+-0.2	3.7+-0.3	3.6+-0.3	3.4+-0.2	3.1+-0.2
	1000	2.9+-0.2	3.3+-0.2	3.1+-0.3	3.7+-0.3	3+-0.2	2.7+-0.2
	5000	2.1+-0.2	2.8+-0.2	2.4+-0.2	2.8+-0.2	1.8+-0.2	2+-0.2
ALARM	500	1.4+-0.1	2.1+-0.2	2.3+-0.2	2.2+-0.2	1.7+-0.2	0.8+-0.1
	1000	1+-0.1	1.8+-0.2	1.9+-0.2	2+-0.2	1.1+-0.1	0.6+-0.1
	5000	0.5+-0.1	1.5+-0.2	1.5+-0.1	1.5+-0.2	0.2+-0.1	0.2+-0
BARLEY	500	4+-0.3	4.1+-0.3	4.4+-0.3	4.9+-0.3	11.1+-0.5	4.2+-0.2
	1000	3.7+-0.3	3.8+-0.3	4.2+-0.3	5+-0.3	NA	3.8+-0.2
	5000	3.4+-0.3	3.6+-0.3	3.5+-0.3	4.3+-0.3	NA	3.1+-0.2
HAILFINDER	500	4.4+-0.3	4.3+-0.2	5.2+-0.3	4.2+-0.2	7.6+-0.5	4.3+-0.3
	1000	4.4+-0.3	4.3+-0.2	5+-0.3	4.5+-0.2	6.7+-0.4	4.1+-0.3
	5000	4.3+-0.3	4.3+-0.2	5.1+-0.3	5.1+-0.2	3.9+-0.2	4+-0.3
30-5-4-1	100	7.5+-0.3	7.3+-0.3	7.2+-0.3	7.5+-0.3	8.7+-0.3	7.4+-0.3
	500	4.5+-0.3	6.3+-0.3	4.6+-0.3	6.2+-0.3	5.7+-0.3	6.3+-0.3
	2000	3.3+-0.2	5.1+-0.3	3.5+-0.2	5.2+-0.3	3.5+-0.2	4.4+-0.3
	5000	2.6+-0.2	4.5+-0.2	3+-0.2	4.7+-0.3	2.6+-0.2	2.9+-0.2
50-5-4-1	100	10.66+-0.3	9.7+-0.3	10.4+-0.3	9.5+-0.3	12.6+-0.3	9.3+-0.3
	500	4.5+-0.3	6.3+-0.3	4.6+-0.3	6.2+-0.3	5.7+-0.3	6.3+-0.3
	2000	5.1+-0.2	6.9+-0.3	5.2+-0.2	7.2+-0.3	5.5+-0.2	6.2+-0.3
	5000	4.2+-0.2	6.1+-0.2	4.5+-0.2	6.5+-0.3	4.4+-0.2	4.4+-0.2

8 appendix

Network	SAMPLES	MBMML+CPT		MBMML+NB		MBMML+RANDOM		IAMB		PCMB		SLL	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
CHILD	500	0.94+-0.03	0.8+-0.04	0.94+-0.03	0.82+-0.04	0.76+-0.04	0.89+-0.03	0.83+-0.04	0.73+-0.04	0.83+-0.04	0.81+-0.04	0.94+-0.03	0.78+-0.04
	1000	0.98+-0.02	0.88+-0.03	0.97+-0.02	0.86+-0.03	0.82+-0.03	0.96+-0.01	0.81+-0.03	0.81+-0.03	0.91+-0.03	0.84+-0.03	0.97+-0.02	0.84+-0.03
	5000	1+-0.01	0.91+-0.02	1+-0.01	0.89+-0.02	0.95+-0.02	1+-0	0.77+-0.04	0.91+-0.02	0.98+-0.01	0.99+-0.01	1+-0	0.97+-0.01
INSURANCE	500	0.82+-0.04	0.48+-0.03	0.8+-0.04	0.42+-0.03	0.7+-0.04	0.6+-0.03	0.86+-0.03	0.44+-0.03	0.75+-0.04	0.52+-0.03	0.83+-0.04	0.51+-0.04
	1000	0.86+-0.03	0.54+-0.03	0.85+-0.04	0.45+-0.03	0.76+-0.03	0.65+-0.03	0.78+-0.03	0.5+-0.03	0.79+-0.04	0.56+-0.03	0.88+-0.03	0.58+-0.03
	5000	0.95+-0.01	0.68+-0.03	0.93+-0.02	0.57+-0.03	0.82+-0.03	0.76+-0.03	0.86+-0.02	0.66+-0.03	0.93+-0.03	0.71+-0.03	0.98+-0.01	0.69+-0.03
ALARM	500	0.85+-0.03	0.77+-0.03	0.79+-0.03	0.67+-0.03	0.66+-0.03	0.87+-0.02	0.81+-0.03	0.65+-0.03	0.84+-0.03	0.71+-0.03	0.92+-0.02	0.89+-0.02
	1000	0.9+-0.02	0.82+-0.03	0.86+-0.02	0.68+-0.03	0.69+-0.03	0.92+-0.02	0.81+-0.02	0.75+-0.03	0.9+-0.02	0.81+-0.03	0.94+-0.01	0.94+-0.01
	5000	0.97+-0.01	0.93+-0.02	0.95+-0.02	0.7+-0.03	0.75+-0.02	0.95+-0.01	0.79+-0.02	0.89+-0.02	1+-0	0.96+-0.01	0.98+-0.01	0.98+-0.01
BARLEY	500	0.74+-0.03	0.37+-0.03	0.74+-0.03	0.35+-0.02	0.66+-0.03	0.51+-0.03	0.7+-0.04	0.17+-0.01	0.25+-0.01	0.59+-0.03	0.63+-0.04	0.25+-0.02
	1000	0.79+-0.03	0.42+-0.03	0.79+-0.03	0.37+-0.02	0.68+-0.03	0.57+-0.03	0.63+-0.04	0.19+-0.02	NA	NA	0.72+-0.04	0.35+-0.03
	5000	0.8+-0.03	0.52+-0.03	0.81+-0.03	0.47+-0.02	0.72+-0.03	0.7+-0.03	0.73+-0.03	0.36+-0.03	NA	NA	0.85+-0.03	0.5+-0.03
HAILFINDER	500	0.3+-0.03	0.18+-0.02	0.25+-0.03	0.12+-0.01	0.27+-0.03	0.2+-0.02	0.31+-0.03	0.17+-0.02	0.28+-0.03	0.19+-0.02	0.28+-0.03	0.14+-0.02
	1000	0.31+-0.03	0.22+-0.02	0.26+-0.03	0.12+-0.01	0.29+-0.03	0.24+-0.02	0.29+-0.03	0.19+-0.02	0.32+-0.03	0.21+-0.02	0.3+-0.03	0.18+-0.02
	5000	0.34+-0.03	0.26+-0.02	0.26+-0.03	0.14+-0.02	0.3+-0.03	0.27+-0.03	0.24+-0.02	0.23+-0.02	0.32+-0.03	0.21+-0.02	0.34+-0.03	0.22+-0.02
30-5-4-1	100	0.56+-0.02	0.36+-0.02	0.6+-0.03	0.23+-0.02	0.58+-0.02	0.36+-0.02	0.69+-0.03	0.16+-0.01	0.43+-0.02	0.37+-0.02	0.5+-0.03	0.17+-0.02
	500	0.91+-0.02	0.56+-0.02	0.86+-0.02	0.35+-0.02	0.86+-0.02	0.56+-0.02	0.84+-0.02	0.37+-0.02	0.87+-0.02	0.41+-0.02	0.79+-0.03	0.3+-0.02
	2000	0.97+-0.01	0.68+-0.02	0.94+-0.01	0.48+-0.02	0.94+-0.01	0.68+-0.02	0.89+-0.02	0.51+-0.02	0.93+-0.01	0.69+-0.02	0.94+-0.02	0.54+-0.02
	5000	0.99+-0	0.76+-0.02	0.96+-0.01	0.57+-0.02	0.96+-0.01	0.73+-0.02	0.87+-0.02	0.58+-0.02	0.91+-0.01	0.82+-0.01	0.98+-0.01	0.7+-0.02
50-5-4-1	100	0.44+-0.02	0.28+-0.01	0.47+-0.02	0.19+-0.01	0.42+-0.02	0.27+-0.01	0.61+-0.02	0.12+-0.01	0.31+-0.01	0.31+-0.01	0.45+-0.03	0.12+-0.01
	500	0.85+-0.02	0.46+-0.02	0.77+-0.02	0.29+-0.02	0.8+-0.02	0.46+-0.02	0.79+-0.02	0.3+-0.02	0.81+-0.02	0.33+-0.02	0.74+-0.02	0.26+-0.02
	2000	0.97+-0.01	0.59+-0.02	0.91+-0.01	0.4+-0.02	0.92+-0.01	0.6+-0.02	0.85+-0.02	0.43+-0.02	0.9+-0.01	0.54+-0.02	0.9+-0.02	0.44+-0.02
	5000	0.99+-0	0.68+-0.01	0.97+-0.01	0.49+-0.02	0.97+-0.01	0.67+-0.01	0.87+-0.01	0.51+-0.02	0.91+-0.01	0.68+-0.01	0.96+-0.01	0.6+-0.02

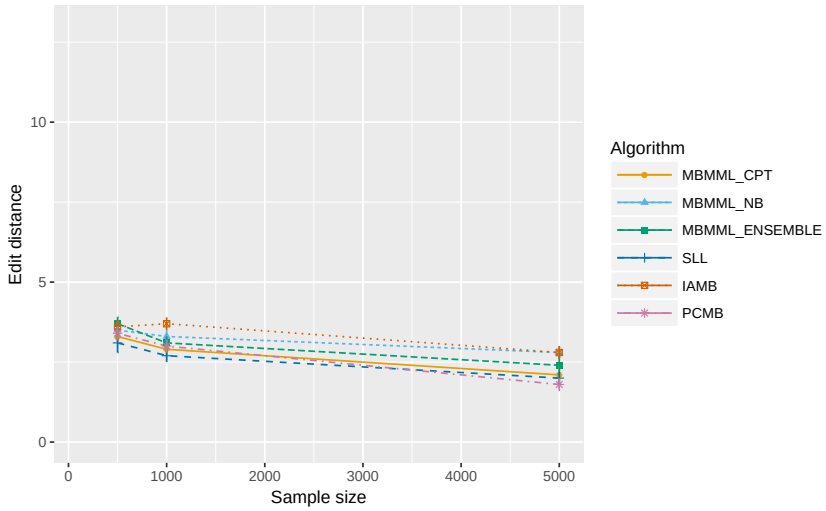


Fig. 10: Edit distance (with 95% confidence intervals) v.s. sample size on INSURANCE network.

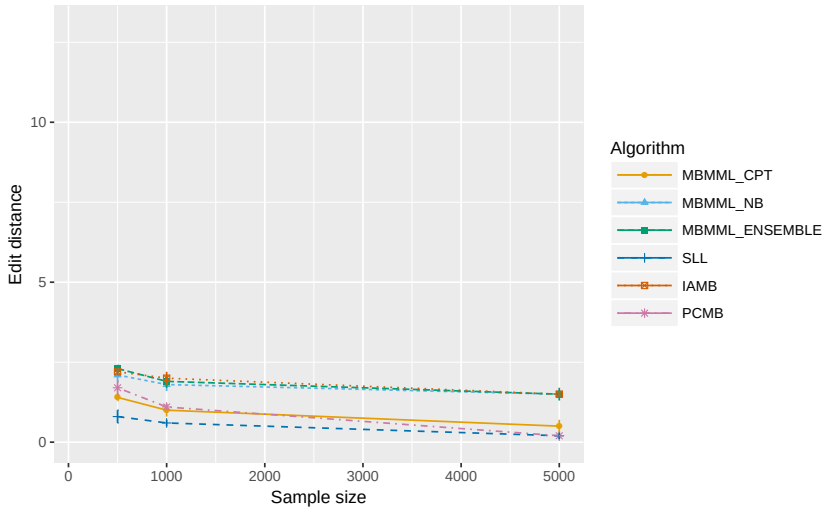


Fig. 11: Edit distance (with 95% confidence intervals) v.s. sample size on ALARM network.

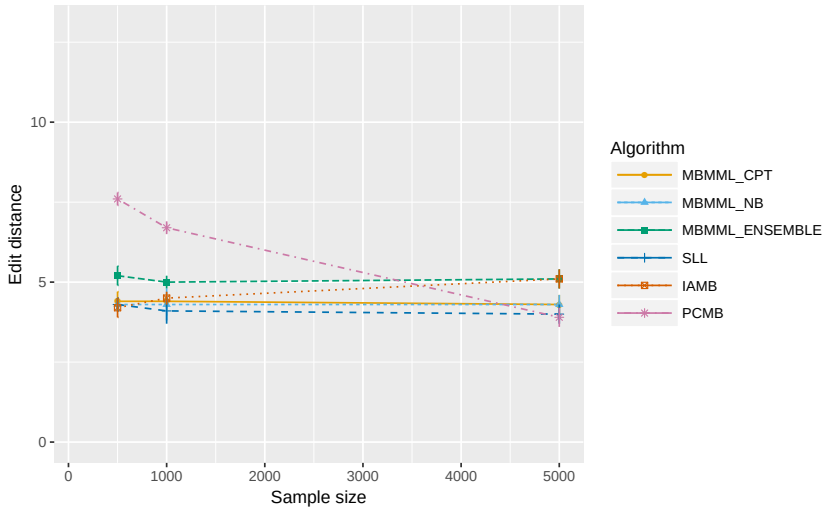


Fig. 12: Edit distance (with 95% confidence intervals) v.s. sample size on HAILFINDER network.

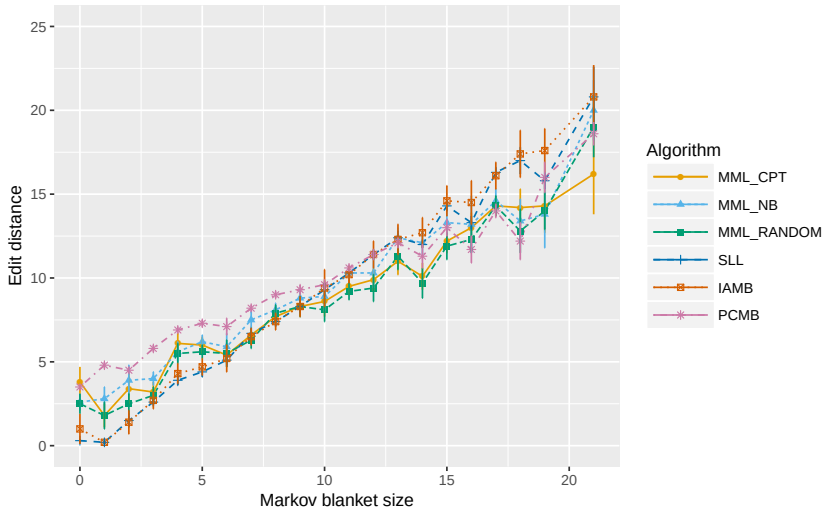


Fig. 13: Edit distance against Markov blanket size on 30-5-4-1 models with 100 samples.

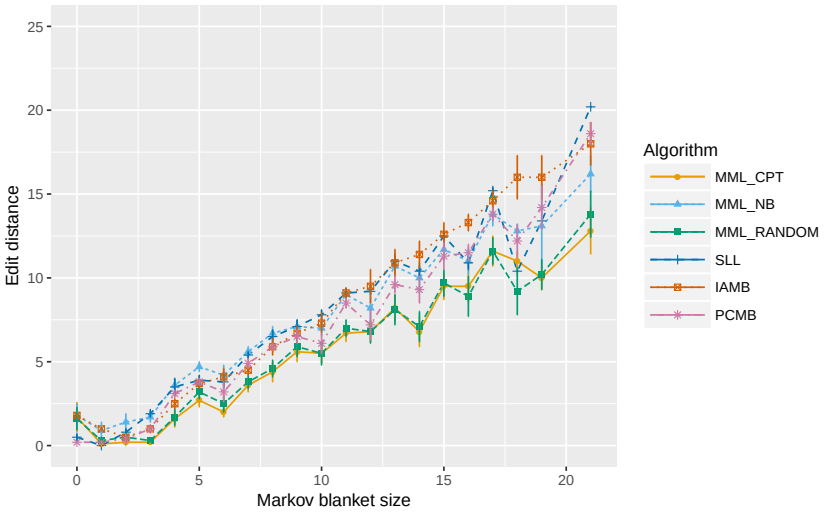


Fig. 14: Edit distance against Markov blanket size on 30-5-4-1 models with 500 samples.

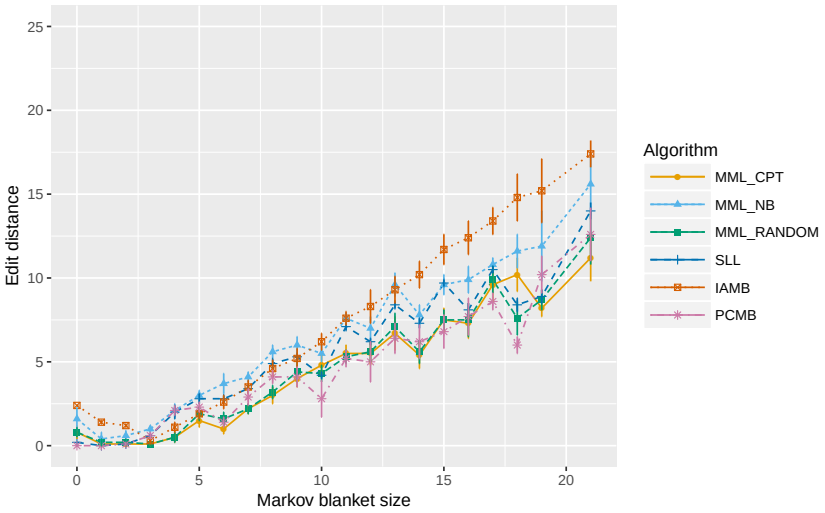


Fig. 15: Edit distance against Markov blanket size on 30-5-4-1 models with 2000 samples.

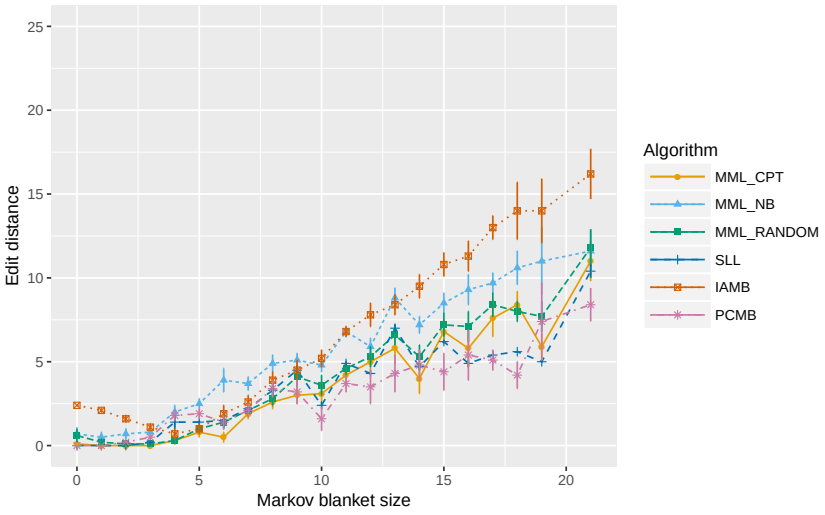


Fig. 16: Edit distance against Markov blanket size on 30-5-4-1 models with 5000 samples.

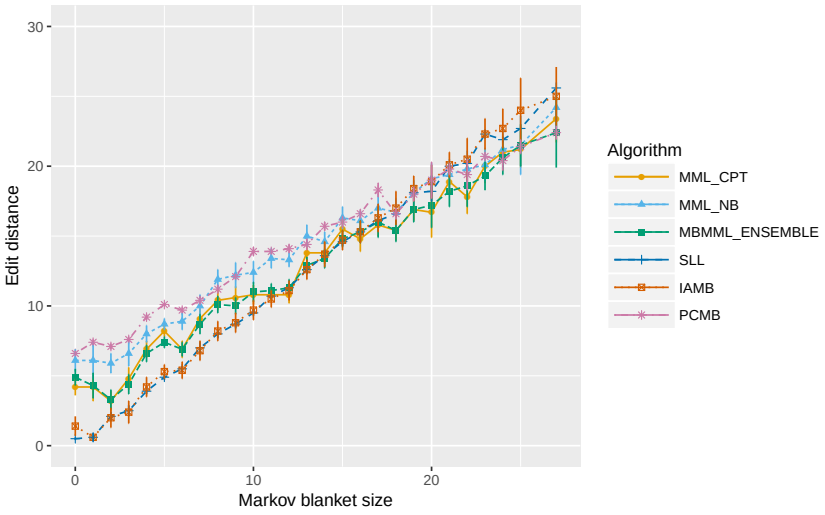


Fig. 17: Edit distance against Markov blanket size on 50-5-4-1 models with 100 samples.

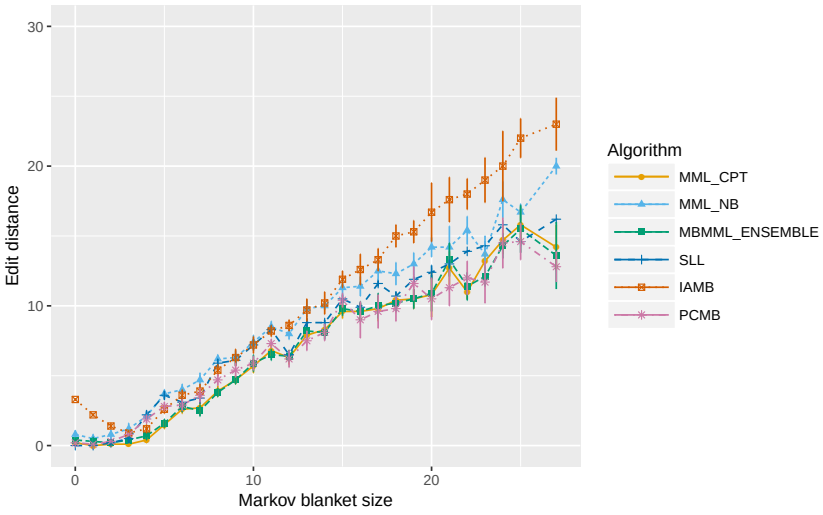


Fig. 18: Edit distance against Markov blanket size on 50-5-4-1 models with 2000 samples.

Table 5: Notations

G	a directed acyclic graph
X	a variable set
E	an edge set
P	a joint distribution
$< G, P >$	a Bayesian network
$\mathcal{H}$	a model class
H	a hypothesis or model
θ	a set model parameters
$\mathcal{D}$	a set of datasets
D	a dataset
N	the number of observations in a dataset
$p(\cdot)$	the probability of an event
$I(\cdot)$	the information content or message length of an event
$\lvert \cdot \rvert$	the cardinality of a set
X_i	the i^{th} variable in X
r_i	the number of states of a variable X_i
Π_i	a parents set of the variable X_i
r_{Π_i}	the total number of parents instantiations
$\boldsymbol{\alpha}$	a vector of symmetric Dirichlet concentration parameters
α_i	the i^{th} parameter in $\boldsymbol{\alpha}$
n_{ijk}	the count of Π_i is in state j and X_i in state k, also known as sufficient statistics
n_{ij}	the count of Π_i in state j
T_i	a Markov blanket polytree of variable X_i
$\mathcal{T}_i$	a set of Markov blanket polytrees containing the same set of variables $\{X_i\} \cup MB_i$

8.1 Notations

References

[Acid *et al.*, 2013] S. Acid, L. M. de Campos, and M. Fernández. Score-based methods for learning Markov boundaries by searching in constrained spaces. *Data Mining and Knowledge Discovery*, 26(1):174–212, 2013.

[Aliferis *et al.*, 2003] C. F. Aliferis, I. Tsamardinos, and A. Statnikov. HITON: a novel Markov blanket algorithm for optimal variable selection. In *AMIA Annual Symposium Proceedings*, pages 21–25. American Medical Informatics Association, 2003.

[Aliferis *et al.*, 2010a] C. F. Aliferis, A. R. Statnikov, I. Tsamardinos, S. Mani, and X. Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation. *The Journal of Machine Learning Research*, 11:171–234, 2010.

[Aliferis *et al.*, 2010b] C. F. Aliferis, A. R. Statnikov, I. Tsamardinos, S. Mani, and X. Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: Analysis and extensions. *The Journal of Machine Learning Research*, 11:235–284, 2010.

[Boulton and Wallace, 1969] D. M. Boulton and C. S. Wallace. The information content of a multistate distribution. *Journal of Theoretical Biology*, 23(2):269–278, 1969.

[Cheng *et al.*, 2001] J. Cheng, C. Hatzis, H. Hayashi, M. A. Krogel, S. Morishita, D. Page, and J. Sese. Kdd cup 2001 report. *ACM SIGKDD Explorations Newsletter*, 3(2):47–64, 2001.

[Chickering, 2002] D. M. Chickering. Learning equivalence classes of Bayesian-network structures. *Journal of machine learning research*, 2(Feb):445–498, 2002.

[Cooper *et al.*, 1997] G. F. Cooper, C. F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa, J. E. Janosky, C. Meek, T. Mitchell, T. Richardson, and P. Spirtes. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9(2):107–138, 1997.

[de Morais and Aussem, 2010] S. R. de Morais and A. Aussem. A novel Markov boundary based feature subset selection algorithm. *Neurocomputing*, 73(4):578–584, 2010.

[Frey *et al.*, 2003] L. Frey, D. Fisher, I. Tsamardinos, C. F. Aliferis, and A. Statnikov. Identifying markov blankets with decision tree induction. In *Proceedings of the 3rd IEEE International Conference on Data Mining, ICDM'03*, 2003.

[Fu and Desmarais, 2008] S. Fu and M. C. Desmarais. Fast Markov blanket discovery algorithm via local learning within single pass. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 96–107. Springer, 2008.

[Gao and Ji, 2017] T. Gao and Q. Ji. Efficient score-based Markov Blanket discovery. *International Journal of Approximate Reasoning*, 80:277–293, 2017.

[Geiger *et al.*, 2001] D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: graphical models and model selection. *The Annals of Statistics*, pages 505–529, 2001.

[Haughton, 1988] D. M. A. Haughton. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16(1):342–355, 1988.

[Koller and Sahami, 1996] D. Koller and M. Sahami. Toward optimal feature selection. In *Proceedings of the 13th conference in Machine Learning*, pages 284–292, 1996.

[Li *et al.*, 2004] G. Li, H. Dai, and Y. Tu. Identifying Markov blankets using lasso estimation. In *Advances in Knowledge Discovery and Data Mining*, number 2004, pages 308–318. Springer, 2004.

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

[Madden, 2002] M. G. Madden. A new Bayesian network structure for classification tasks. In *Irish Conference on Artificial Intelligence and Cognitive Science*, pages 203–208. Springer, 2002.

[Margaritis and Thrun, 1999] D. Margaritis and S. Thrun. Bayesian network induction via local neighborhoods. *In Advances in Neural Information Processing Systems*, 12:505–511, 1999.

[Neil *et al.*, 1999] J. R. Neil, C. S. Wallace, and K. B. Korb. Learning Bayesian networks with restricted causal interactions. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 486–493. Morgan Kaufmann Publishers Inc., 1999.

[Niinimaki and Parviainen, 2012] T. Niinimaki and P. Parviainen. Local structure discovery in Bayesian networks. 2012.

[Pearl, 1988] J. Pearl. *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann San Mateo, CA, 1988.

[Peña *et al.*, 2007] 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.

[Silander *et al.*, 2007] T. Silander, P. Kontkanen, and P. Myllymäki. On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, pages 360–367. AUAI Press, 2007.

[Strobl and Visweswaran, 2016] E. V. Strobl and S. Visweswaran. Markov boundary discovery with ridge regularized linear models. *Journal of Causal Inference*, 4(1):31–48, 2016.

[Tsamardinos *et al.*, 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 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 673–678. ACM, 2003.

[Tsamardinos *et al.*, 2003b] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov. Algorithms for large scale Markov blanket discovery. In *FLAIRS Conference*, pages 376–381, 2003.

[Wallace and Boulton, 1968] C. S. Wallace and D. M. Boulton. An information measure for classification. *The Computer Journal*, 11(2):185–194, 1968.

[Wallace and Freeman, 1987] C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society*, 49(3):240–265, 1987.

[Wallace, 2005] C. S. Wallace. *Statistical and inductive inference by minimum message length.* Springer, 2005.