Learning tractable Bayesian networks in the space of elimination orders

Marco Benjumeda*, Concha Bielza, Pedro Larrañaga

**Computational Intelligence Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Spain**

**ARTICLE INFO**

*Article history:*  
Received 10 January 2018  
Received in revised form 18 August 2018  
Accepted 19 November 2018  
Available online 12 February 2019

**Keywords:**  
Learning tractable models  
Inference complexity  
Bayesian networks  
Treewidth estimation  
Machine learning

**ABSTRACT**

The computational complexity of inference is now one of the most relevant topics in the field of Bayesian networks. Although the literature contains approaches that learn Bayesian networks from high dimensional datasets, traditional methods do not bound the inference complexity of the learned models, often producing models where exact inference is intractable. This paper focuses on learning tractable Bayesian networks from data. To address this problem, we propose strategies for learning Bayesian networks in the space of elimination orders. In this manner, we can efficiently bound the inference complexity of the networks during the learning process. Searching in the combined space of directed acyclic graphs and elimination orders can be extremely computationally demanding. We demonstrate that one type of elimination trees, which we define as valid, can be used as an equivalence class of directed acyclic graphs and elimination orders, removing redundancy. We propose methods for incrementally compiling local changes made to directed acyclic graphs in elimination trees and for searching for elimination trees of low width. Using these methods, we can move through the space of valid elimination trees in polynomial time with respect to the number of network variables and in linear time with respect to treewidth. Experimental results show that our approach successfully bounds the inference complexity of the learned models, while it is competitive with other state-of-the-art methods in terms of fitting to data.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Bayesian networks (BNs) [1,2] concisely model probability distributions over a set of random variables. They are self-explanatory and easy to understand, and they are well suited for representing causal relationships. Some applications of BNs are supervised classification [3] and clustering [4,5]. Each BN $\mathcal{B}$ over a set of variables $\mathcal{X} = \{X_1, \ldots, X_n\}$ is composed of:

- A directed acyclic graph (DAG) $\mathcal{G}$ that represents the conditional independences among triplets of variables in $\mathcal{X}$.
- A set of parameters $P(X_i | \text{Pa}_\mathcal{G}(X_i))$ that represent the conditional probability distributions of each $X_i \in \mathcal{X}$ conditional on its parents $\text{Pa}$ in $\mathcal{G}$.

* Corresponding author.

E-mail addresses: marco.benjumeda.barquita@upm.es (M. Benjumeda), mcbielza@fi.upm.es (C. Bielza), pedro.larranaga@fi.upm.es (P. Larrañaga).

https://doi.org/10.1016/j.artint.2018.11.007  
0004-3702/© 2019 Elsevier B.V. All rights reserved.
A common approach for learning BNs from data is to perform a search process optimizing a scoring function that measures the quality of each structure. Two types of scores are usually used. Bayesian metrics maximize the posterior probability of the network conditional on the data given a prior distribution over all the possible network structures, while information theory metrics try to maximize the data compression achieved by each network. Well-known scoring functions such as Bayesian Dirichlet equivalent (BDe) [6], K2 [7,8], Akaike information criterion (AIC) [9], Bayesian information criterion (BIC) [10] or minimum description length (MDL) [11,12] implicitly or explicitly penalize the number of network parameters. Note that BIC, that is based on the Schwarz Information Criterion [10], is equivalent to MDL as a BN scoring function. The representation complexity, which is given by the number of network parameters, does not place an upper bound on the inference complexity of the models, and a model with a low representation complexity can have a high inference complexity. Thus, more precise estimations of the inference complexity are required to ensure the tractability of models during the learning process.

A good indicator of the inference complexity of a BN $\mathcal{B}$ with structure $\mathcal{G}$ is the treewidth of $\mathcal{G}$ ($\text{tw}(\mathcal{G})$), given that the most widely used exact inference methods for BNs, like variable elimination (VE) or message passing in junction trees (JTs), can be computed in exponential time in $\text{tw}(\mathcal{G})$. Intuitively, the treewidth of a graph $\mathcal{G}$ can be understood as a measure of how similar $\mathcal{G}$ is to a tree (e.g., a tree has treewidth one). It is NP-hard to determine the treewidth of a graph [13], and there are no efficient exact methods for solving this problem. Many heuristics have been proposed for treewidth estimation (see Section 2.2), but most are very computationally demanding. This is especially important when BNs have to be learned from data, since we have to compute the treewidth of each candidate during the learning process to ensure tractability.

VE is one of the simplest methods for inference in BNs. It consists of successively eliminating the variables of a network until it yield the answer to a given query. The elimination of a variable $X_i$ consists of outputting the product of all the factors containing $X_i$, and marginalizing the result over $X_i$. The order in which the variables are removed is called elimination order (EO). The computational cost of VE is upper bounded by the width of the chosen EO $\pi$, which is the number of variables in the biggest factor induced by VE minus one.

The treewidth of a graph $\mathcal{G}$ can also be expressed as the width of the optimal EO $\pi_{\text{opt}}$ for graph $\mathcal{G}$. This means that obtaining an optimal EO of $\mathcal{G}$ is equivalent to obtaining the treewidth of $\mathcal{G}$ [13], and it is also an NP-hard problem. Hence, one way of getting an accurate estimation of $\text{tw}(\mathcal{G})$ is to find a good EO for $\mathcal{G}$. It would often be intractable to get a good EO from scratch for each candidate network during the structure search. As most structure learning methods perform local changes in $\mathcal{G}$ during the learning process, a more efficient solution to this problem is to incrementally update the EOs for each local change performed in $\mathcal{G}$. There are usually multiple equivalent EOs for $\mathcal{G}$ (see Section 2.1). This means that the combined space of DAGs and EOs is highly redundant, and it would be extremely computationally demanding to search for low complexity structures in this space. In this paper, we define a type of elimination trees (ETs) [14], which we call valid ETs, that avoid this redundancy. A single valid ET can be used to represent all the EOs that are equivalent (i.e., induce the same factors during VE) for any graph $\mathcal{G}$. We propose methods for efficiently compiling each possible local change that could be applied in $\mathcal{G}$ (i.e., arc additions, removals or reversals), and provide a framework for learning valid ETs from data using the above methods.

This paper is organized as follows. Section 2 introduces inference complexity in BNs and reviews previous work on bounding the treewidth of BNs and learning models of low inference complexity. Section 3 contains our proposal. We show the relation between ETs and EOs, and the way the former can be used as an equivalence class of EOs and DAGs. Section 4 describes the proposed compilation and optimization methods, and it shows how to use ETs to learn tractable BNs in the space of EOs. Section 5 reports the experimental results. Section 6 outlines the concluding remarks and future research lines.

The software of the proposed method is available at https://github.com/marcobb8/et-learn.

2. Background

2.1. Treewidth and elimination orders

To give a formal definition of treewidth, we must first introduce moral graphs.

**Definition 1. (Moral graph)** The moral graph $\mathcal{G}_M$ of a directed graph $\mathcal{G}$ with nodes $\mathcal{X}$ is the result of:

1. Adding an undirected link between each pair of nodes $X_i, X_j \in \mathcal{X}$ that have a common child in $\mathcal{G}$ and are not connected.
2. Converting every directed arc into an undirected link.

Next, we define the tree decomposition of a graph, also known as jointree or junction tree.

**Definition 2. (Tree decomposition graph)** Let $\mathcal{G}_M$ be the moral graph of a directed graph $\mathcal{G}$ with nodes $\mathcal{X}$. A tree decomposition of $\mathcal{G}$ is a tree $\mathcal{T}$ with a set of clusters $\mathcal{C}$, where each cluster $C_i \in \mathcal{C}$ is a node of $\mathcal{T}$, that satisfy:

- Each cluster $C_i$ of $\mathcal{T}$ is a subset of $\mathcal{X}$.
• For all edges $X_i - X_j$ in $G_M$, there is a cluster $C_k \in C$ such that $X_i, X_j \in C_k$.
• If a node $X_i$ appears in two clusters $C_i$ and $C_j$, it must also appear in every cluster $C_k$ on the path connecting $C_i$ and $C_j$ in the tree decomposition (the running intersection property).

The width of a tree decomposition $T$ is $\max_{C_i \in C} |C_i| - 1$ (i.e., the size of its biggest cluster minus 1). The treewidth of a graph $\mathcal{G}$ is the minimum width across all its decompositions.

An EO of a set of variables $\mathcal{X} = \{X_1, \ldots, X_n\}$ is a permutation $\pi = (\pi(X_1), \ldots, \pi(X_n))$ of $\mathcal{X}$. We use $(X_i < X_j)_\pi$ to denote that $X_i$ must be eliminated before $X_j$ given $\pi$.

The treewidth of a graph $\mathcal{G}$ can also be expressed as the width of its optimal EO. As it is NP-hard to find an optimal EO, several heuristics are used to find EOs that are satisfactory in practice. In Section 2.2 we discuss some options.

The combined space of EOs and DAGs is redundant. This means that there may be multiple EOs that induce the same factors (using VE) for the same BN. We define the equivalence of two EOs as:

**Definition 3. (Equivalence of EOs)** Let $B$ be a BN over $\mathcal{X}$, and $\pi_1$ and $\pi_2$ two EOs of $\mathcal{X}$. Let $\text{Cls}_{\pi_1}(X_i)$ and $\text{Cls}_{\pi_2}(X_i)$ be the clusters induced by visiting node $X_i$ during VE using the EOs $\pi_1$ and $\pi_2$, respectively. $\pi_1$ and $\pi_2$ are equivalent for $B$ if, for each $X_i \in \mathcal{X}$, $\text{Cls}_{\pi_1}(X_i) = \text{Cls}_{\pi_2}(X_i)$.

The completeness of a set of EOs $\mathcal{S}$ for $B$ ensures that if an $\pi_i$ belongs to $\mathcal{S}$ all the EOs that are equivalent to $\pi_i$ for $B$ also belong to $\mathcal{S}$. Note that the completeness of $\mathcal{S}$ does not imply that all the nodes in $\mathcal{S}$ are equivalent for $B$.

**Definition 4. (Completeness of a set of EOs)** A set of EOs $\mathcal{S}$ is complete for $B$ if there are no two equivalent EOs $\pi_i, \pi_j$, with $\pi_i \in \mathcal{S}$ and $\pi_j \notin \mathcal{S}$, for $B$.

For example, assume a network $B$ over variables $\mathcal{X} = \{X_1, \ldots, X_n\}$ that represents the product of marginals $P(X_1, \ldots, X_n) = P(X_1)P(X_2)\cdots P(X_n)$. Given $B$, VE induces the same factors for any EO of $X_1, \ldots, X_n$. Hence, all the $n!$ possible EOs are equivalent for $B$, and there is a single complete set of EOs that contains all the permutations of $\mathcal{X}$.

### 2.2. Treewidth estimation

It is NP-hard to exactly compute the treewidth of a BN [13]. There are many approaches whose time complexity is exponential in the number of network variables [15−18]. In practice, heuristics are most often used. As the treewidth of a graph $\mathcal{G}$ is given by the width of its optimal EO, some well-known heuristics estimate the treewidth of $\mathcal{G}$ by searching for good EOs for $\mathcal{G}$. The list below includes some popular approaches:

• Greedy search methods: Two widely used strategies are to eliminate, at each iteration, the smallest degree node (i.e., the node with fewest neighbors) in the graph [19] or the node that produces the minimum number of fill-in (min-fill) edges [20]. In practice, the min-fill algorithm performance is generally slightly better, but its computational cost is higher.
• Graph recognition techniques: The lexicographic breadth-first search algorithm (LEX) [21] and the maximum cardinality search (MCS) algorithm [22] return an optimal EO only if the input graph is chordal. The chordality assumption is very restrictive in practice, but there are two variants of these methods, respectively called LEX-M [21] and MCS-M [23], that also search for a good EO if the graph is not chordal.
• Local search and evolutionary techniques: Some well-known heuristics like simulated annealing [24], genetic algorithms [25], or tabu search [26] have been used to find good EOs.

Another approach focuses on finding the best graph separators, recursively splitting the clusters of an initial tree decomposition into smaller components [27]. Most methods using this strategy give theoretical guarantees of the treewidth upper bound. Bodlaender and Koster [28] provide an overview of the different heuristics used for computing upper bounds for graph treewidth, including the above methods. Their experiments suggest that greedy search methods outperform graph recognition techniques and approaches that use separators.

Sometimes it is sufficient to check that the treewidth does not exceed a constant $k$ rather than exactly computing the treewidth of $\mathcal{G}$; for instance by learning models with a treewidth less than or equal to $k$ (see Section 2.4). Although this is an NP-complete problem [13], it can be computed in linear time in the number of variables for a fixed $k$. Nevertheless, the time complexity for solving this inequality is super-exponential in the treewidth of $\mathcal{G}$ [29], which means that it may be intractable unless $k$ is very small.

---

1 A graph is chordal if all cycles of four or more nodes have an edge that connects two nodes of the cycle but is not part of the cycle.
2.3. Inference complexity in Bayesian networks

Probabilistic inference can be used to refer to multiple problems in BNs. Some well-known inference problems are: evidence propagation, finding the maximum a posteriori (MAP) hypothesis, and computing the most probable explanation (MPE). Evidence propagation entails finding the posterior probability \( P(X|e) \) of a set of query variables \( Q \) conditional on evidence \( e \). It can be used for some key tasks such as prediction and diagnosis. Finding the MAP consists of searching the most probable configuration of a set of variables in a BN for given evidence. The MPE is a special case of the MAP that involves searching the most probable configuration of all variables not instantiated in a BN for given evidence. Kwisthout [30] provides a thorough overview of the complexity of many MPE and MAP variants.

In this paper, we use inference complexity to refer to the complexity of evidence propagation in BNs. Exact inference in BNs is generally NP-hard [31], and approximate inference is commonly used when exact inference is intractable. Approximate inference in BNs is also NP-hard [32], and, although it has been useful for solving some otherwise intractable problems, it has some major drawbacks. It degrades the responses output by the model, and hardly any of these algorithms offer any guarantees of convergence.

The message-passing (MP) algorithm [33,34] can perform exact linear time inference in the number of variables of any BN \( B \) when its topology is a polytree. However, there are many situations where polytrees are not representative enough, and this restriction is therefore too strict in practice. Inference in BNs with loops is far from straightforward, and we cannot use MP to perform exact inference in this type of networks. Although MP has been adapted to deal with loops, the resultant method, called loopy belief propagation [11], provides only approximate results. Most exact inference methods for graphs with loops are based on variable elimination [35,36], conditioning [37,38] and clustering [39,40]. For any BN \( B \), the above methods are exponential in the treewidth of \( B \). Thus, \( tw(B) \) is a good estimator of the inference complexity of \( B \).

The literature also includes approaches that are not always exponential in the treewidth. In this case, tractable exact inference does not necessarily call for models with a low treewidth. These methods exploit the local network structures [41,42], or the exchangeability between the model variables [43]. Nevertheless, it is extremely challenging to consider the above properties during the learning process.

2.4. Previous work on learning low inference complexity Bayesian networks

Most approaches that address the problem of inference complexity during the learning process use a bound \( k \) on the model treewidth (i.e., bounded treewidth models). They reject any candidate \( G \) for which \( tw(G) > k \). Learning bounded treewidth BNs is an NP-hard problem [44]. The literature contains exact methods for this problem that reduce the problem to either a weighted maximum satisfiability problem [45] or mixed-integer linear programming formulations [46,47]. These methods scale poorly with respect to the number of model variables and model treewidth.

Elidan and Gould [48] proposed a method that uses an incremental triangulation of BNs during the structure search to output bounded treewidth models. Their method is treewidth-friendly (i.e., each update of the triangulation does not increase its width by more than one), and it basically applies the best chain of arc additions in each iteration given a topological ordering of the variables. Its main limitation is that the method is restricted to a single topological ordering of the variables in each iteration.

Nie et al. [46] proposed an efficient approach that samples \( k \)-trees randomly and selects the best BN structure whose moral graph is the sampled \( k \)-tree. As the convergence of the sampling process can be a problem when the number of variables is not small, Nie et al. [49] also provided a strategy for moving in the space of \( k \)-trees and proposed a score (1-score) to measure how well a \( k \)-tree fits the data. The authors showed that this measure is correlated with the BDeu score of the learned networks.

Scanagatta et al. [50] proposed a method (called \( k \)-greedy) for learning bounded treewidth BNs from very large datasets. Before performing the structure search, \( k \)-greedy initializes a cache of candidate parent sets for each node using the approach of Scanagatta et al. [51]. Then, it samples the space of orderings of variables, performing the next steps for each order. First, an initial structure with the first \( k + 1 \) variables in the order is learned. Depending on the value of \( k \), \( k \)-greedy uses either an exact [52] or an approximate [51] structure learning method. Second, the structure incrementally grows according to the chosen order, ensuring that at each step the moral graph of the structure is a partial \( k \)-tree. This process is repeated until the maximum allowed execution time is met. Very recently, Scanagatta et al. [53] improved \( k \)-greedy by proposing a heuristic score for choosing the order in which the variables are visited. This heuristic ranks the variables by comparing the highest-scoring parent set with the lowest scoring parent set that do not exceed the treewidth bound. The resultant method is called \( k \)-MAX. As the former, \( k \)-MAX requires predefining a maximum execution time to explore the space of solutions. Extensive experiments showed that both approaches consistently outperform some of the above methods [46,47,49] for learning bounded treewidth BNs. A limitation of \( k \)-greedy and \( k \)-MAX is that they only learn BNs whose reverse topological order, when used as an EQ, has at most width \( k \).

There are also several approaches that learn JTJs with bounded treewidth, usually called thin junction trees (TJTs) [54]. This problem is NP-complete when the bound on the treewidth \( k \) is greater than one [55]. Chechetka and Guestrin [56] proposed a method that learns TJTs with probably approximately correct (PAC) guarantees in time \( O(n^k) \), which is intractable when \( k \) is not very small. Shahaf and Guestrin [57] used the graph cuts algorithm [58] to pick the best separator in each iteration during the learning process, requiring polynomial time in both \( n \) and \( k \). As mentioned above, heuristics that use
separators usually perform worse in practice than heuristics that search for good EOs for estimating the treewidth of the models.

Some approaches use a penalization in the inference complexity instead of a hard constraint. Lowd and Domingos [59] proposed the first method (LearnAC) to learn arithmetic circuits (ACs) directly from data. This method penalizes the size of each circuit exploiting the local structures of the models to learn networks that can be tractable even for high treewidths. Moves in the space of ACs can be extremely computationally demanding, as circuit structure can be huge. LearnAC uses a greedy approach to address these difficulties, where the best split (i.e., conditioning the conditional probability distribution of a variable to an instance of another variable) is applied at each iteration. Like EOs, the order of splits can have a major effect on network size, and this type of search process is not capable of reconfiguring the split ordering during the learning process. Benjumeda et al. [60] used topological EOs to learn tractable BNs, penalizing each candidate with the width of the EO for the network structure. This method provides for a flexible learning process, accounting for arc additions, removals and reversals. Its main drawback is that the upper bound on the inference complexity provided by topological EOs is not usually as tight as the bound provided by other representations.

3. Elimination trees

This paper addresses the problem of learning bounded treewidth BNs. We focus on choosing a compact representation of the combined space of DAGs and EOs and a set of operators that allow efficiently moving in this space for the next reasons: First, this search space does not put any restrictions on the structure beyond the treewidth bound. Second, given addition, removal and reversal operators, most score+search BN learning methods can be easily adapted to learn bounded treewidth BNs.

In a BN \( \mathcal{B} \) over \( \mathcal{X} = \{X_1, \ldots, X_n\} \), there are \( n! \) different EOs of \( \mathcal{X} \), although many are usually equivalent (Definition 3) for the structure \( \mathcal{G} \) of \( \mathcal{B} \), especially when \( \mathcal{G} \) is not densely connected. We need to avoid this redundancy to reduce the size of the search space during the learning process. Next, we define elimination trees (ETs), a representation that is especially well suited for this purpose. ETs are based on the representation proposed by Grant and Horsch [14] for recursive conditioning, which we adapt to represent a set \( \mathcal{S} \) of EOs for \( \mathcal{B} \).

**Definition 5. (Elimination tree)** Let \( \mathcal{B} \) be a BN over \( \mathcal{X} = \{X_1, \ldots, X_n\} \). An elimination tree \( \mathcal{E}_{\mathcal{B}} \) over \( \mathcal{X} \) is composed of:

- A set of factors or potentials \( \phi_{X_1}, \ldots, \phi_{X_n} \) that represent the parameters of \( \mathcal{B} \) of each node \( X_1, \ldots, X_n \).
- A tree \( T \) composed of a root node, \( \ast \), an inner node (node with parent and children) for each variable \( X_i \in \mathcal{X} \), and a leaf node labeled \( \phi_{X_i} \) for each potential \( \phi_{X_i} \). The nodes are connected by undirected edges.

Assuming that we use VE over an ET \( \mathcal{E}_{\mathcal{B}} \) to perform inference, the topology of the tree shows the orders in which each variable \( X_i \in \mathcal{X} \) should be eliminated from the factors of the model. If an inner node \( X_i \) is the predecessor (this precedence must be read from the root node to the leaves) of another inner node \( X_j \), then \( X_i \) is eliminated after \( X_j \).

**Definition 6. (ET representation of an EO)** Let \( \mathcal{B} \) be a BN over \( \mathcal{X} = \{X_1, \ldots, X_n\} \). An elimination tree \( \mathcal{E}_{\mathcal{B}} \) represents an EO \( \pi \) for \( \mathcal{B} \) if, for each \( X_i, X_j \in \mathcal{X} \) (with \( X_i < X_j \) implies \( X_j \notin \text{Desc}_{\mathcal{B}}(X_i) \), \( \mathcal{E}_{\mathcal{B}} \) represents a set of EOs \( \mathcal{S} \) for \( \mathcal{B} \) if it represents each \( \pi_i \in \mathcal{S} \) for \( \mathcal{B} \).

Fig. 1 shows an ET \( \mathcal{E}_{\mathcal{B}} \) that represents the set of EOs \( \mathcal{S} \) for the probability distribution \( P(X_1, X_2, X_3) = \phi_{X_1}(X_1, X_2) \cdot \phi_{X_2}(X_2) \cdot \phi_{X_3}(X_1, X_3) \). As \( X_1 \) is a predecessor in \( \mathcal{E}_{\mathcal{B}} \) of \( X_2 \) and \( X_3 \), \( \mathcal{E}_{\mathcal{B}} \) represents each EO \( \pi \) such that \( X_2 < X_1 \) and \( X_3 < X_1 \), that is, \( \pi = (X_2, X_3, X_1) \) and \( \pi = (X_3, X_2, X_1) \).

Let us also consider the product of marginals. If we have a BN \( \mathcal{B} \) over \( \mathcal{X} = \{X_1, \ldots, X_n\} \) that represents the probability distribution \( P(X_1, \ldots, X_n) = \phi_{X_1}(X_1) \cdots \phi_{X_n}(X_n) \), all the EOs of \( X_1, \ldots, X_n \) are equivalent for \( \mathcal{B} \). This can be represented by a single ET, as shown in Fig. 2.

Inference in ETs is straightforward. Given an ET \( \mathcal{E}_{\mathcal{B}} \) that represents a set of EOs \( \mathcal{S} \) for \( \mathcal{B} \), we could use any EO \( \pi_i \in \mathcal{S} \) to perform VE, or to efficiently compile \( \mathcal{B} \) into a JT or an AC.

3.1. Properties of elimination trees

In this section, we introduce some terms that we use in the rest of the paper. Let \( \mathcal{E}_{\mathcal{B}} \) be an ET over \( \mathcal{X} = \{X_1, \ldots, X_n\} \). We use \( \text{Pa}_{\mathcal{E}_{\mathcal{B}}}(X_i) \) and \( \text{Ch}_{\mathcal{E}_{\mathcal{B}}}(X_i) \) to refer to the parent and the children of node \( X_i \) in \( \mathcal{E}_{\mathcal{B}} \). \( \text{Pred}_{\mathcal{E}_{\mathcal{B}}}(X_i) \) and \( \text{Desc}_{\mathcal{E}_{\mathcal{B}}}(X_i) \) refer to the set of predecessors and descendant nodes of \( X_i \) in \( \mathcal{E}_{\mathcal{B}} \), respectively. For example, \( \text{Desc}_{\mathcal{E}_{\mathcal{B}}}(X_1) = \{X_2, X_3, \phi_{X_1}, \phi_{X_2}, \phi_{X_1}\} \) and \( \text{Pred}_{\mathcal{E}_{\mathcal{B}}}(X_1) = \{\ast\} \) in the ET shown in Fig. 3.

---

2 ACs are DAGs in which the inner nodes are addition and multiplication nodes and the leaves are numeric variables or constants. They have been adapted to perform inference in BNs [42].
Fig. 1. Structures of a BN $B$ (left) and an ET $\mathcal{E}_B$ (right). In $\mathcal{E}_B$, $*$ is the root node, $X_1$, $X_2$, and $X_3$ are the inner nodes, and $\phi_{X_1}$, $\phi_{X_2}$, and $\phi_{X_3}$ are the leaves and potentials. The domain of the potential of each leaf node $\phi_{X_i}$ is illustrated below the respective node. $\mathcal{E}_B$ represents the EOs $(X_2, X_3, X_1)$ and $(X_3, X_2, X_1)$ for $B$.

Fig. 2. Structure of an ET that represents the product of marginals. Below each leaf node $\phi_{X_i}$, the domain of its corresponding potential is shown.

Fig. 3. Structure of an ET. The clusters of the ET are shown near to their corresponding nodes.

Given a factor $\phi_X(X_{i(1)}, \ldots, X_{i(n_i)})$, $\text{Dom}(\phi_{X_i})$ represents its domain, that is, the set of nodes $\{X_{i(1)}, \ldots, X_{i(n_i)}\}$, where $X_{i(1)}, \ldots, X_{i(n_i)} \in \mathcal{X}$ and $n_i$ is the cardinality of $\text{Dom}(\phi_{X_i})$. We use $\text{Leaves}(\mathcal{E}_B)$ to refer to the set of leaf nodes in $\mathcal{E}_B$.

ETs closely resemble dtrees, a representation used for recursive conditioning [38]. Unlike ETs, dtrees are full binary trees (i.e., trees in which any inner node has two children), and their inner nodes are labeled with a set of variables instead of being labeled with a single variable. There follows a definition of clusters in ETs, which is analogous to the definition of clusters given by Darwiche [61] for dtrees.

Definition 7. (Clusters of ET nodes) The cluster of an inner node $X_i$ in an ET $\mathcal{E}_B$ is defined as:

$$\text{Cls}_{\mathcal{E}_B}(X_i) := \bigcup_{X_j \in \text{Ch}_{\mathcal{E}_B}(X_i)} \text{Cls}_{\mathcal{E}_B}(X_j) \setminus \{X_j\}.$$ 

The cluster of a leaf node $\phi_{X_i}$ in $\mathcal{E}_B$ is defined as:

$$\text{Cls}_{\mathcal{E}_B}(\phi_{X_i}) := \text{Dom}(\phi_{X_i}).$$
When we perform VE in $\mathcal{E}_B$, $\text{Cl}_B(X_i)$ is equivalent to the cluster (domain of the generated factor) induced by eliminating $X_i$. Fig. 3 shows an example of the clusters $\text{Cl}_B(X_i)$ of an ET $\mathcal{E}_B$. The clusters of ETs and the clusters (or cliques) of JTs are also closely related (see Section 3.1.1).

3.1.1. Valid elimination trees

The purpose of using ETs to search for structures with a small treewidth is to reduce the combined space of DAGs and EOs, and consequently allow efficient algorithms for learning bounded treewidth BNs. By the above definition, there are many solutions that are incorrect or redundant. To identify and avoid such ETs during the learning process, we define two new properties: soundness and completeness.

We say that an ET $\mathcal{E}_B$ is sound if all the EOs that it represents are equivalent for $B$.

**Definition 8.** (Sound ETs) Let $\mathcal{E}_B$ be an ET over $\mathcal{X}$. Node $X_i$ is sound for $\mathcal{E}_B$ if $\text{Cl}_B(X_i) \subseteq \text{Pred}_{\mathcal{E}_B}(X_i) \cup \{X_i\}$. A leaf node $\phi_{X_i} \in \text{Leaves}(\mathcal{E}_B)$ is sound for $\mathcal{E}_B$ if $\text{Cl}_{\mathcal{E}_B}(\phi_{X_i}) \subseteq \text{Pred}_{\mathcal{E}_B}(\phi_{X_i})$. $\mathcal{E}_B$ is sound if every node (inner and leaf nodes) is sound for $\mathcal{E}_B$.

Fig. 4 shows the structure of an unsound ET $\mathcal{E}_B$. Given that there are no ancestral relationships between the inner nodes in $\mathcal{E}_B$, it represents all the possible permutations of $\{X_1, X_2, X_3, X_4\}$ as EOs. The clusters of some nodes contain variables (underlined) that are not their predecessors in $\mathcal{E}_B$. For example, $\text{Cl}_B(X_1)$ contains $X_1$ and $X_2$, but $X_2$ is not a predecessor of $X_1$. As there is no ancestral relationship between $X_1$ and $X_2$ in $\mathcal{E}_B$, it equivalent whether $\mathcal{E}_B$ eliminates $X_1$ before or after $X_2$. Unfortunately, this is not true, as eliminating $X_2$ before $X_1$ would induce cluster $\{X_1, X_2\}$. However, this cluster cannot be induced by any EO $\pi_1$ where $(X_1 < X_2)_{\pi_1}$ because $X_1$ will have been eliminated from all factors before $X_2$ has been eliminated. Thus, if there is a variable that belongs to the cluster of a node $X_i$ that is not one of the predecessors of $X_i$ in $\mathcal{E}_B$, then the $\mathcal{E}_B$ is not sound, and it represents EOs that are not equivalent.

The completeness of ETs is analogous to the completeness of a set of EOs.

**Definition 9.** (Complete ETs) Let $\mathcal{E}_B$ be an ET over $\mathcal{X}$. Node $X_i \in \mathcal{X} \setminus \text{Leaves}(\mathcal{E}_B)$ (i.e., $X_i$ is either an inner node from $\mathcal{X}$ or a leaf node from $\text{Leaves}(\mathcal{E}_B)$) is complete for $\mathcal{E}_B$ if $\text{Pa}_{\mathcal{E}_B}(X_i) \in \text{Cl}_{\mathcal{E}_B}(X_i)$ or $\text{Pa}_{\mathcal{E}_B}(X_i) = \ast$. $\mathcal{E}_B$ is complete if every node (inner and leaf nodes) is complete for $\mathcal{E}_B$.

Fig. 5 shows the structure of an incomplete ET $\mathcal{E}_B$. It represents the EOs $S = \{(X_3, X_2, X_1, X_4), (X_3, X_2, X_4, X_1), (X_3, X_4, X_2, X_1), (X_4, X_3, X_2, X_1)\}$, but there are other EOs that are equivalent for $B$ that are not represented by $\mathcal{E}_B$. For example, $(X_2, X_3, X_1, X_4)$ is equivalent to $(X_3, X_2, X_1, X_4)$ given that the clusters induced after eliminating $X_2$ and $X_3$ are $\{X_1, X_2\}$ and $\{X_1, X_3\}$ in both cases.

**Definition 10.** (Valid ETs) An ET $\mathcal{E}_B$ is valid if it is sound and complete.

The ET shown in Fig. 3 is sound (for every node $X_i$, all the variables in its cluster are either its predecessors or $X_i$) and complete (for every node $X_i$ with parent $X_p$, the cluster of $X_i$ contains $X_p$). This means that it is valid. The space of valid ETs does not contain incorrect or redundant solutions.

The process described by Algorithm 1 yields a valid ET $\mathcal{E}_B$, given a BN $B$ and an EO $\pi$.

Algorithm 1 starts with an ET where the parent of every node is the root node $\ast$ (line 1). First, the variables in $\mathcal{X}$ are visited in the order given by $\pi$ (line 2). When variable $X_i$ is visited, node $X_i$ is set as the parent of the nodes whose cluster contains $X_i$ and whose parent is the root node $\ast$ in the ET (lines 3–7). This is analogous to the process of eliminating variable $X_i$ from $B$. The cluster $\text{Cl}_{\mathcal{E}_B}(X_i)$ of $X_i$ in the ET $\mathcal{E}_B$ is output in the same way as the cluster $\text{Cl}_{\pi}(X_i)$ induced by eliminating $X_i$ using $\pi$ in $B$, and they are equal.

Proposition 1 states that given an EO $\pi$ and a BN $B$, Algorithm 1 returns always valid ETs. Hence, there is at least one valid ET for $\pi$ and $B$. 
Proposition 1. Let \( B \) be a BN over \( \mathcal{X} \) and \( \pi \) an EO of \( \mathcal{X} \). Algorithm 1 returns a valid ET that represents \( \pi \) for \( B \).

**Proof.** Algorithm 1 ensures that when any variable \( X_i \) is visited, it is set as the parent of every node whose cluster contains \( X_i \) and whose parent is the root node \( * \). Therefore:

- If the cluster of node \( X_j \) contains \( X_i \) when \( X_i \) is visited, the cluster of each of its predecessors also has \( X_i \), given that node \( X_j \) has no children until \( X_i \) has been visited. When \( X_i \) is visited, it is set as a predecessor of all the nodes whose cluster contains \( X_i \). After visiting \( X_i \), there are no nodes whose clusters contain \( X_i \) that are not their descendants in \( \mathcal{E}_B \). As this applies to each node \( X_i \in \mathcal{X} \), all the nodes in \( \mathcal{E}_B \) must be sound, making \( \mathcal{E}_B \) sound.
- The cluster of every node \( X_j \) that is a child of \( X_i \) contains \( X_i \). Each \( X_j \) is complete for \( \mathcal{E}_B \), making \( \mathcal{E}_B \) complete.
- A node \( X_j \) can only be a descendant of a node \( X_i \) in \( \mathcal{E}_B \) if \( (X_j < X_i) \). Hence, \( \mathcal{E}_B \) represents \( \pi \).

As \( \mathcal{E}_B \) is valid (sound and complete) and represents \( \pi \), there is at least one valid \( \mathcal{E}_B \) for \( B \) and \( \pi \). \( \square \)

Proposition 2 ensures that there is a single valid ET \( \mathcal{E}_B \) that represents an EO \( \pi \) for a BN \( B \).

**Proposition 2.** Let \( B \) be a BN over \( \mathcal{X} \) and \( \pi \) an EO of \( \mathcal{X} \). There is exactly one valid ET \( \mathcal{E}_B \) that represents \( \pi \) for \( B \).

**Proof.** From Proposition 1, we know that there is at least one valid ET for \( B \) and \( \pi \). We prove that there is exactly one by structural induction. We consider two ETs \( \mathcal{E}^1_B \) and \( \mathcal{E}^2_B \) for \( B \) and \( \pi \). We show that if \( \mathcal{E}^1_B \) and \( \mathcal{E}^2_B \) are valid, then, for each node \( X_i \in \mathcal{X} \), \( \text{Pa}_{\mathcal{E}^1_B}(X_i) = \text{Pa}_{\mathcal{E}^2_B}(X_i) \), starting from the leaves (base case). This means that \( \mathcal{E}^1_B \) and \( \mathcal{E}^2_B \) are the same, which implies that there is a single valid ET for \( B \) and \( \pi \).

**Base case:**
The subtrees that have \( \phi_{X_i} \in \text{Leaves}(\mathcal{E}_B) \) as its root in \( \mathcal{E}^1_B \) and \( \mathcal{E}^2_B \) are only composed of node \( \phi_{X_i} \). Hence, they are equal.

**Inductive step:**
Assume that the subtrees hanging from node \( X_i \) in \( \mathcal{E}^1_B \) and \( \mathcal{E}^2_B \) are equal. Let \( X_j = \text{Pa}_{\mathcal{E}^1_B}(X_i) \) (Fig. 6). As \( \mathcal{E}^1_B \) is valid, if \( X_j \neq \text{Pa}_{\mathcal{E}^2_B}(X_i) \), then \( \text{Pa}_{\mathcal{E}^2_B}(X_i) \) is a node \( X_k \) where either:

![Figure 5: Structure of an incomplete ET. The clusters of the ET are shown near to their respective nodes, and the clusters that compromise the completeness of the ET are underlined.](image-url)
Proposition 3. Let $B$ be a BN over $\mathcal{X}$, an EO of $\mathcal{X}$, and $E_B$ a valid ET that represents $\pi_1$ for $B$. For each EO $\pi_2$ equivalent to $\pi_1$ for $B$ (Definition 3), $E_B$ also represents $\pi_2$.

Proof. Let $C_{\pi_1}^1$ and $C_{\pi_2}^2$ be the cluster induced by VE after eliminating $X_i$ from $B$ using the EO $\pi_1$ and $\pi_2$, respectively. As $\pi_1$ and $\pi_2$ are equivalent for $B$, $C_{\pi_1}^1 = C_{\pi_2}^2$.

From Proposition 2 we know that there is a single valid ET $E_B^1$ that represents $\pi_1$ for $B$, and also a single valid ET $E_B^2$ that represents $\pi_2$ for $B$. We prove that $E_B^1$ also represents $\pi_2$ by structural induction. We show that if $E_B^1$ and $E_B^2$ are valid, then, for each node $X_i \in \mathcal{X}$, that represents $\pi_1$ for $B$.

Base case:
The subtrees whose root is $\phi_X \subseteq \text{Leaves}(E_B^1)$ in $E_B^1$ and $E_B^2$ are composed of node $\phi_X$ only. Hence, they are equal.

Inductive step:
Assume that the subtrees hanging from node $X_i$ in $E_B^1$ and $E_B^2$ are equal. Let $X_j = \text{Pa}_{E_B^1}(X_i)$ (Fig. 6). As $E_B^1$ is valid, if $X_j \neq \text{Pa}_{E_B^2}(X_i)$, then $\text{Pa}_{E_B^2}(X_i)$ is a node $X_k$ where either:

a) $X_k \in \text{Pred}_{E_B^1}(X_j)$: then $X_j \notin \text{Cl}_{\pi_1}(X_k)$ given that $X_j \in \text{Desc}_{E_B^1}(X_k)$. Assuming that $\text{Cl}_{\pi_1}(X_i) = \text{Cl}_{\pi_2}(X_i)$, $X_j \in \text{Cl}_{\pi_2}(X_k)$ given that $X_j \in \text{Cl}_{\pi_2}(X_i) = \text{Cl}_{\pi_2}(X_k)$ (the parent of $X_i$ in $E_B^2$) and that $X_k$ is the parent of $X_i$ in $E_B^2$. Hence, $\text{Cl}_{\pi_1}(X_k) \neq \text{Cl}_{\pi_2}(X_k)$.

b) $X_k \notin \text{Pred}_{E_B^1}(X_j)$: then $X_k \notin \text{Cl}_{\pi_1}(X_i)$ given that $X_i \notin \text{Pred}_{E_B^2}(X_k)$, and $X_k \in \text{Cl}_{\pi_2}(X_i)$ given that $\text{Pa}_{E_B^2}(X_i) = X_k$. Thus, $\text{Cl}_{\pi_1}(X_i) \neq \text{Cl}_{\pi_2}(X_i)$.

This means that if $E_B^1$ is valid, then the following condition $\text{Pa}_{E_B^1}(X_i) = \text{Pa}_{E_B^2}(X_i)$ holds. □
Definition 11. (ET width) The width of an ET $\mathcal{E}_B$ is the length of its largest cluster minus one.

Given a BN $B$, the width of the ET $\mathcal{E}_B$ with lowest width that is valid for $B$ is the treewidth of $B$. Therefore, if $\mathcal{E}_B$ is good enough (near-minimum width), its width is an indicator of the inference complexity of the model as it should be close to the treewidth of $B$.

4. Learning elimination trees

4.1. Compiling changes

A naive solution to learn bounded treewidth BNs would be to use a known heuristic to output a good EO (see Section 2.2). The width of the chosen EO is an estimate of the treewidth of the BN candidates. However, it is computationally demanding to search for good EOs from scratch, and it can be intractable if we have to perform this process for each candidate during the structure search. The results shown in Section 3 can be applied to learn tractable BNs in the combined space of DAGs and EOs. Our proposal is to limit the treewidth of the BN by bounding the width of the ET (Definition 11). This strategy requires obtaining a valid ET for each network candidate during the learning process.

In this section, we propose methods to compile incrementally in ETs the arc additions and removals made to a BN during the learning process, and show that the proposed algorithms always output valid ETs. As the reversal of arc $X_{\text{out}} \rightarrow X_{\text{in}}$ in $B$ can be seen as the removal of arc $X_{\text{out}} \rightarrow X_{\text{in}}$ followed by the addition of the reversed arc $X_{\text{in}} \rightarrow X_{\text{out}}$, we assume that both changes are compiled each time a reversal is made to a BN.

4.1.1. Arc addition

The addition of an arc $X_{\text{out}} \rightarrow X_{\text{in}}$ in $B$ may compromise the soundness of an ET $\mathcal{E}_B$. If $\mathcal{E}_B$ is valid, it represents a complete set of equivalent EOs $S$ (Proposition 2). The addition of $X_{\text{out}} \rightarrow X_{\text{in}}$ in $B$ places a new restriction on the equivalence of the EOs in $S$. After applying this local change, there is at least one factor over both $X_{\text{out}}$ and $X_{\text{in}}$. Therefore, an ET $\mathcal{E}_B'$ can only be valid for the new BN $B'$ if it encodes an ancestral relationship between $X_{\text{out}}$ and $X_{\text{in}}$. Algorithm 2 modifies the structure of $\mathcal{E}_B'$ to meet the new restrictions. The resulting ET $\mathcal{E}_B''$ represents a complete subset of EOs $S' \subseteq S$ (see Fig. 7) that are also equivalent in $B'$.

Algorithm 2 receives a valid ET $\mathcal{E}_B'$, and arc addition $X_{\text{out}} \rightarrow X_{\text{in}}$. In $B'$, variable $X_{\text{out}}$ is added to the domain of $\phi_{X_{\text{in}}}$ (line 2). The clusters of the nodes that are predecessors of $\phi_{X_{\text{in}}}$ and descendants of $X_{\text{out}}$ contain $X_{\text{out}}$ after applying this change. There are three different scenarios that require performing different changes in the ET to ensure its validity.

If $X_{\text{out}}$ is a predecessor of $\phi_{X_{\text{in}}}$ in $\mathcal{E}_B'$, it is not necessary to make any changes in $\mathcal{E}_B''$. Otherwise, some nodes contain $X_{\text{out}}$ in their clusters but not in their predecessors, and $\mathcal{E}_B''$ is not sound. If $X_f = \text{Pa}_{\mathcal{E}_B'}(\phi_{X_{\text{in}}})$ is a predecessor of $X_{\text{out}}$ in $\mathcal{E}_B'$ (line 4), $X_{\text{out}}$ is set as the new parent of $\phi_{X_{\text{in}}}$ in $\mathcal{E}_B''$ (line 5). Thus, $X_{\text{out}}$ is a predecessor of $\phi_{X_{\text{in}}}$ in $\mathcal{E}_B''$.

```
Input: Valid ET $\mathcal{E}_B'$, output node $X_{\text{out}}$, input node $X_{\text{in}}$
Output: Valid ET $\mathcal{E}_B''$
1 let $\mathcal{E}_B''$ be a copy of $\mathcal{E}_B'$;
2 Dom($\phi_{X_{\text{in}}}$) ← Dom($\phi_{X_{\text{in}}}$) ∪ {$X_{\text{out}}$};
3 $X_f$ ← Pa$_{\mathcal{E}_B'}$($\phi_{X_{\text{in}}}$);
4 if $X_f$ ∈ Pred$_{\mathcal{E}_B'}(X_{\text{out}})$ then
5     Pa$_{\mathcal{E}_B''}$($\phi_{X_{\text{in}}}$) ← $X_{\text{out}}$;
6 else if $X_{\text{out}}$ ∈ Pred$_{\mathcal{E}_B'}(\phi_{X_{\text{in}}})$ then
7     let $X_{\text{in}}$ be the deepest node in Pred$_{\mathcal{E}_B'}(X_{\text{out}})$ ∩ Pred$_{\mathcal{E}_B'}(\phi_{X_{\text{in}}})$;
8     let $\mathcal{E}_B^1$ and $\mathcal{E}_B^2$ be two copies of $\mathcal{E}_B'$;
9     $X_k$ ← Ch$_{\mathcal{E}_B'}(X_{\text{in}})$ ∩ Pred$_{\mathcal{E}_B'}(\phi_{X_{\text{in}}})$;
10    Pa$_{\mathcal{E}_B''}$($X_k$) ← $X_{\text{out}}$;
11    $X_k$ ← Ch$_{\mathcal{E}_B'}(X_{\text{in}})$ ∩ Pred$_{\mathcal{E}_B'}(X_{\text{out}})$;
12    Pa$_{\mathcal{E}_B''}$($X_k$) ← $X_f$;
13    Pa$_{\mathcal{E}_B''}$($\phi_{X_{\text{in}}}$) ← $X_{\text{out}}$;
14    if width($\mathcal{E}_B''$) < width($\mathcal{E}_B^1$) then
15        $\mathcal{E}_B''$ ← $\mathcal{E}_B^1$;
16    else
17        $\mathcal{E}_B''$ ← $\mathcal{E}_B^2$;
18 end
19 return $\mathcal{E}_B''$;
```

Algorithm 2: Compilation of the addition of arc $X_{\text{out}} \rightarrow X_{\text{in}}$ (add($\mathcal{E}_B', X_{\text{out}}, X_{\text{in}}$)).
Fig. 7. A set of EOSs $S$ that are equivalent for a BN $B$, and a subset $S'$ of $S$ that are equivalent for the BN $B'$, output after adding an arc in $B$.

Fig. 8. (a) an ET $E_B$, and (b) the ET $E_{B'}$, output after incrementally compiling the arc addition $X_1 \rightarrow X_4$ (i.e., addition of $X_1$ to $\text{Dom}(\phi_{X_4})$) in $E_B$.

If $X_{\text{out}}$ is not a predecessor of $\phi_{X_{in}}$ in $E_B$ and $X_f$ is not a predecessor of $X_{\text{out}}$ in $E_B$ (line 6), the cluster of the node in \{$X_{\text{out}}$\} $\cup$ $\text{Pred}_{E_B}(\phi_{X_{in}}) \cap \text{Desc}_{E_B}(X_{\text{out}})$ contains $X_{\text{out}}$ but the clusters of their predecessors in $E_B$ do not. Algorithm 2 creates two candidate ETs ($E_{B'}^1$ and $E_{B'}^2$). The first is output by setting $X_{\text{out}}$ as the parent of $X_b$ in $E_{B'}^1$ (line 10), that is, the shallowest predecessor of $\phi_{X_{in}}$ in $E_B$ that does not belong to $\text{Pred}_{E_B}(X_{\text{out}})$ (line 9). The second is output by setting $\text{Pa}_{E_B}(\phi_{X_{in}})$ as the parent of $X_b$ in $E_{B'}^2$ (line 12), that is, the shallowest predecessor of $X_{\text{out}}$ in $E_B$ that does not belong to $\text{Pred}_{E_B}(\phi_{X_{in}})$ (line 11), and setting $X_{\text{out}}$ as the new parent of $\phi_{X_{in}}$ in $E_{B'}^2$ (line 13). $E_{B'}^2$ is selected as the ET of smaller width between $E_{B'}^1$ and $E_{B'}^2$ (lines 14–18). Either way, $X_{\text{out}}$ is a predecessor in $E_{B'}$ of the nodes in \{$X_{\text{out}}$\} $\cup$ $\text{Pred}_{E_B}(\phi_{X_{in}}) \cap \text{Desc}_{E_B}(X_{\text{out}})$, and the returned ET $E_{B'}$ is valid (Lemma 1).

Lemma 1. Let $E_B$ be a valid ET that represents $B$ over $X$. Given $X_{\text{out}}$, $X_{in} \in X$, the ET $E_{B'}$ yielded after applying $\text{add}(E_B, X_{\text{out}}, X_{in})$ in Algorithm 2, is also a valid ET representing $B'$ over $X$.

Proof. See Appendix A. □

This process can be performed efficiently in ETs of bounded width (see Theorem 2).

Fig. 8 shows an example of the incremental compilation of the addition of arc $X_1 \rightarrow X_4$ in $X_2$. $E_B$ represents all the permutations of \{$X_1, X_2, X_3, X_4$\} where $X_2$ and $X_3$ are eliminated before $X_1$. After adding $X_1 \rightarrow X_4$ in $B$, there are EOSs represented by $E_B$ that are not equivalent for $B'$ (e.g., \{$X_2, X_3, X_1, X_4$\} and \{$X_2, X_3, X_4, X_1$\}). Using Algorithm 2, $X_{\text{out}} = X_1$ and $X_{in} = X_1$. As $X_f = X_4$ is not a predecessor of $X_1$ in $E_B$ (line 4), and $X_1$ is not a predecessor of $\phi_{X_4}$ (line 6), two ETs $E_{B'}^1$ and $E_{B'}^2$ are created. Assuming that $E_{B'}^1$ is smaller than $E_{B'}^2$, $X_1$ is the new parent of $X_4$ in $E_{B'}$.  

4.1.2. Arc removal

The removal of an arc $X_{\text{out}} \rightarrow X_{in}$ in $B$ may compromise the completeness of an ET $E_{B'}$. Let $S$ be the set of EOSs represented by $E_B$. The removal of $X_{\text{out}} \rightarrow X_{in}$ in $B$ leads to a reduction in the restrictions on EO equivalence in $S$. This means that $E_{B'}$ may not represent all the EOSs that are equivalent to EOSs in $S$. Algorithm 3 yields an ET that represents a complete superset of EOSs $S' \supseteq S$ (see Fig. 9) containing all the EOSs that are equivalent to EOSs in $S$ for $B'$, which is the resulting BN after removing arc $X_{\text{out}} \rightarrow X_{in}$ from $B$.

After removing $X_{\text{out}} \rightarrow X_{in}$ from $B$, the shallowest node in $\text{Pred}_{E_{B'}}(\phi_{X_{in}}) \cup \text{Desc}_{E_{B'}}(X_{\text{out}})$, which we refer to as $X_i$ (line 4), may not be complete. If $X_i$ is not complete (line 5), Algorithm 3 sets the deepest node in $\text{Cls}_{E_{B'}}(X_i)$, which we refer to as $X_j'$, as its new parent (lines 6–7). Note that the idea behind this change is that the new parent of $X_i$ is in its cluster, making $X_i$ complete. After this change, the shallowest node in $\text{Pred}_{E_{B'}}(X_j') \cap \text{Desc}_{E_{B'}}(X_j')$ may not be sound.
Algorithm 3: Compilation of the removal of arc \( X_{\text{out}} \rightarrow X_{\text{in}} \) \((\text{remove}(\mathcal{E}_B, X_{\text{out}}, X_{\text{in}}))\).

Fig. 9. A set of EOs \( S \) that are equivalent for a BN \( B \), and a superset \( S' \) of \( S \) that are equivalent for the BN \( B' \), yielded after removing an arc in \( B \).

Fig. 10. (a) an ET \( \mathcal{E}_B \), and (b) an ET \( \mathcal{E}_B' \) yielded after incrementally compiling the removal of arc \( X_2 \rightarrow X_1 \) (i.e., removal of \( X_2 \) from \( \text{Dom}(\phi_{X_1}) \)) in \( \mathcal{E}_B \).

Thus, Algorithm 3 repeats the same process until every node is complete, guaranteeing the validity of every node in \( \mathcal{E}_B' \) (Lemma 2).

**Lemma 2.** Let \( \mathcal{E}_B \) be a valid ET that represents \( B \) over \( X \). Given \( X_{\text{out}}, X_{\text{in}} \in X \), the ET \( \mathcal{E}_B' \) that represents \( B' \) output after applying \( \text{remove}(\mathcal{E}_B, X_{\text{out}}, X_{\text{in}}) \) in Algorithm 3, is also valid.

**Proof.** See Appendix A. ☐

Fig. 10 shows an example of how Algorithm 3 compiles, in an ET \( \mathcal{E}_B \), the removal of arc \( X_2 \rightarrow X_1 \) in \( B \). Let \( \mathcal{E}_B' \) be a copy of \( \mathcal{E}_B \) where \( X_2 \) has been removed from the domain of \( \phi_{X_1} \). First, \( X'_j \) is set to \( X_{\text{out}} \) (line 3). As \( X_{\text{out}} = X_2 \) and \( X_{\text{in}} = X_1 \), \( X_1 \) is the shallowest node in \( \text{Pred}_{\mathcal{E}_B}(\phi_{X_1}) \cup \{X_1\} \) that is a descendant of \( X_2 \) (line 4), namely \( \phi_{X_1} \).

\( X_{\text{out}} \) is not in the cluster of \( \phi_{X_1} \) in \( \mathcal{E}_B' \) (line 5). Hence, \( X'_j \) is set to \( X_1 \), that is, the deepest node in the cluster of \( \phi_{X_1} \) (line 6). \( X_1 \) is now set as the parent of \( \phi_{X_1} \) in \( \mathcal{E}_B' \), which makes node \( \phi_{X_1} \) complete. The new \( X_1 \) is set to \( X_2 \) (line 8), and the new \( X_j \) is \( X_1 \) (line 9).
As \( X_1 \) is not in the cluster of \( X_2 \) in \( \mathcal{E}_B \), the same process is applied again (lines 6–9). In this case, the root node \( \ast \) is set as the parent of \( X_2 \) in \( \mathcal{E}_B \), and the new \( X_j \) is \( \ast \), ending the loop (line 5) and returning \( \mathcal{E}_B' \).

The complexity of this process is polynomial in the number of variables and the width of the ET (see Theorem 2).

### 4.2. Optimization

The methods described above adapt an ET \( \mathcal{E}_B \) (resulting in \( \mathcal{E}_B' \)) to a local change in a BN \( B \) (resulting in \( B' \)). The objective of these methods is to make \( \mathcal{E}_B \) valid for the new BN. In addition to the incremental compilation of ETs, we also propose a strategy to search in the space of ETs given a BN \( B \). The purpose of this procedure is to reduce the width of an ET without modifying \( B \). We use a simple and efficient heuristic to address this problem. In this paper, we use the optimization process to refine (reduce the width) of the ETs returned by Algorithms 2–3.

Algorithm 4 swaps the position in \( \mathcal{E}_B \) of node \( X_i \) with the position of its parent \( X_p \), also changing the parents of any children of \( X_i \) in \( \mathcal{E}_B \) whose validity is compromised by the swap. Algorithm 4 guarantees the validity of the resulting ETs. Note that after each swap only the clusters of \( X_i \) and \( X_p \) may change.

The sets of ETs represented by \( \mathcal{E}_B \) and the new ET \( \mathcal{E}_B' \), which we refer to as \( S \) and \( S' \) respectively, are disjoint (see Fig. 11). In \( \mathcal{E}_B \), \( (X_p < X_i)_\pi \) for any EO \( \pi \in S \), whereas \( (X_i < X_p)_{\pi'} \) in \( \mathcal{E}_B' \) for any EO \( \pi' \in S' \).

Algorithm 4 proceeds as follows: First, a copy \( \mathcal{E}_B' \) of \( \mathcal{E}_B \) is created (line 1), and \( X_j \) is set to the parent of \( X_i \) in \( \mathcal{E}_B \) (line 2). Then, the positions of \( X_i \) and \( X_j \) are swapped in \( \mathcal{E}_B' \) (lines 3 and 4). After that, the children of \( X_i \) whose cluster contained \( X_j \) set their parent in \( \mathcal{E}_B' \) to \( X_j \) (lines 5–9), because otherwise these nodes would not be sound.

**Lemma 3.** Let \( \mathcal{E}_B \) be a valid ET that represents \( B \) over \( X \). Given \( X_i \in X \), the ET \( \mathcal{E}_B' \) representing \( B' \) yielded after applying swap(\( \mathcal{E}_B \), \( X_i \)) in Algorithm 4, is also valid.

**Proof.** See Appendix A. 

Fig. 12 shows the result of applying Algorithm 4 to an ET \( \mathcal{E}_B \). In this example, the positions of \( X_3 \) and \( X_1 \) (parent of \( X_3 \) in \( \mathcal{E}_B \)) in the resulting ET \( \mathcal{E}_B' \) are swapped. \( \mathcal{E}_B \) represents the EOws \( \pi \) of \( \{X_1, X_2, X_3, X_4\} \) where \( X_2 < X_1 \pi \) and \( X_3 < X_1 \pi \) (e.g., \( X_2, X_3, X_1, X_4 \), \( X_4, X_3, X_2, X_1 \), ...), while \( \mathcal{E}_B' \) represents the EOws \( \pi' \) of \( \{X_1, X_2, X_3, X_4\} \) where \( X_2 < X_1 \pi' \) and \( X_3 < X_2 \pi' \) (e.g., \( X_4, X_2, X_1, X_3 \), \( X_2, X_4, X_1, X_3 \), ...).

We use a greedy heuristic (see Algorithm 5), which, given an ET \( \mathcal{E}_B \) and a set of nodes for optimization \( (X_{\text{opt}}) \), visits each node \( X_i \in X_{\text{opt}} \) from the shallowest to the deepest, checking at each step whether swapping the position of \( X_i \) and its parent reduces the width of the ET.
Algorithm 5: Optimization of an ET \( \text{optimize}(E', X_{\text{opt}}) \).

Orderings that are good for one BN \( B^0 \) (i.e., their width is close to the treewidth of \( B^0 \)) may not be good for the BN \( B \) yielded after applying a local change in \( B^0 \). We perform the optimization process after the compilation of each local change. For efficiency, we select the nodes that may have a different cluster after compiling the local change to initialize \( X_{\text{opt}} \), given that they are more likely to produce relevant changes in the width of the ET. Next, we show the set of nodes selected for optimization \( (X_{\text{opt}}) \) after compiling arc additions and removals. Each bullet point describes the assignment to \( X_{\text{opt}} \) at a possible scenario. We also explain the reason why the cluster of any of the nodes in \( X_{\text{opt}} \) may have changed.

- **Addition of arc** \( X_{\text{out}} \rightarrow X_{\text{in}} \):
  - If \( X_{\text{out}} \in \text{Pred}_{E_O}(\phi_{X_{\text{in}}}) \), \( X_{\text{opt}} = \text{Pred}_{E_O}(\phi_{X_{\text{in}}}) \cap \text{Desc}_{E_O}(X_{\text{out}}) \):
    The width of the clusters in \( \text{Pred}_{E_O}(\phi_{X_{\text{in}}}) \) grows, given that they now contain \( X_{\text{out}} \).
  - Else, if \( \text{Pa}_{E_O}(\phi_{X_{\text{in}}}) \in \text{Pred}_{E_O}(X_{\text{out}}) \), \( X_{\text{opt}} = (\text{Pred}_{E_O}(X_{\text{out}}) \cup \{X_{\text{out}}\}) \cap \text{Desc}_{E_O}(\text{Pa}_{E_O}(\phi_{X_{\text{in}}})) \):
    The width of the clusters in \( \text{ Pred}_{E_O}(X_{\text{out}}) \) grows, given that they now contain \( \text{Dom}(\phi_{X_{\text{in}}}) \).
  - Otherwise, \( X_{\text{opt}} = (\text{Pred}_{E_O}(\phi_{X_{\text{in}}}) \cup \text{Pred}_{E_O}(X_{\text{out}}) \cup \{X_{\text{out}}\}) \setminus (\text{Pred}_{E_O}(\phi_{X_{\text{in}}}) \cap \text{Pred}_{E_O}(X_{\text{out}})) \):
    Node \( \phi_{X_{\text{in}}} \) is set as a descendant of \( X_{\text{out}} \), and the nodes in \( \text{Pred}_{E_O}(\phi_{X_{\text{in}}}) \) are either predecessors or descendants of \( X_{\text{out}} \), and \( \text{Pred}_{E_O}(X_{\text{out}}) \) are either predecessors or descendants of \( X_{\text{out}} \).
    \( \text{Pred}_{E_O}(\phi_{X_{\text{in}}}) \cap \text{Pred}_{E_O}(X_{\text{out}}) \) in the new ET.

- **Removal of arc** \( X_{\text{in}} \rightarrow X_{\text{out}} \), \( X_{\text{opt}} = \text{Pred}_{E_O}(\phi_{X_{\text{in}}}) \cap \text{Desc}_{E_O}(X_{\text{out}}) \):
  Let \( X_j \) be the last node that had a new child \( X_i \) assigned by Algorithm 3. The nodes in \( \text{Pred}_{E_O}(\phi_{X_{\text{in}}}) \cap \text{Desc}_{E_O}(X_i) \) may have smaller clusters in the new ET.

The optimization of an ET takes polynomial time in the number of variables and in the width of the ET (see Theorem 2).
4.3. Learning elimination trees from data

Using the incremental compilation and optimization methods described above, it is rather straightforward to learn ETs from a dataset \( \mathcal{D} \) in combination with any score+search BN learning method that applies local changes during the search. Learning low inference complexity BNs with this approach is also easily derived. It can be achieved by bounding the width of each ET during the learning process (which bounds the treewidth of their corresponding BNs).

Theorem 1 ensures that any algorithm that uses the above strategy will always produce valid ETs.

**Theorem 1.** Let \( \mathcal{E}_B \) be a valid ET over \( \mathcal{X} \), and \( \mathcal{E}'_B \), the result of incrementally compiling on \( \mathcal{E}_B \) any local change in \( B \) using Algorithms 2 and 3 and optimizing the resulting ET using Algorithm 5. Then \( \mathcal{E}'_B \) is a valid ET.

**Proof.** See Appendix A. \( \square \)

As we apply the incremental compilation and optimization methods to each candidate during the learning process, efficiency is a critical issue. Theorem 2 bounds the computational time complexity of the incremental compilation and optimization methods proposed above.

**Theorem 2.** Let \( \mathcal{E}_B \) be a valid ET over a set of variables \( \mathcal{X} = \{X_1, \ldots, X_n\} \). The process described in Theorem 1 to output \( \mathcal{E}'_B \), can be performed in time \( O(n^2 \cdot \text{width}(\mathcal{E}_B)) \).

**Proof.** See Appendix B. \( \square \)

Let \( \mathcal{A} \) be any algorithm that learns the structure of a BN using only local changes in the structure of the network during the learning process. \( \mathcal{A} \) can be adapted to learn low inference complexity BNs compiling and optimizing (Algorithms 2–5) all the local changes that are applied to the BN \( B \) to its respective ET \( \mathcal{E}_B \). Thus, \( \mathcal{E}_B \) can be used to bound the treewidth of \( B \) using Definition 11. Note that the input for the adaptation of \( \mathcal{A} \) should be an ET \( \mathcal{E}'_{B_0} \) valid for the initial BN \( B^0 \).

5. Experimental results

In this section, we empirically analyzed the performance of the proposed framework in terms of fitting and computational complexity. Although our approach could be used with most score+search BN learning methods, in the experiments we combine the incremental compilation and optimization methods proposed in Section 4 with a greedy hill-climbing for the structure search. We call the resulting method hc-ET. We compared hc-ET with k-greedy and k-MAX to highlight the advantages and drawbacks of using our approach to learn bounded treewidth BNs. We also tested a polynomial version of hc-ET that only considers arc additions during the structure search. We call this method hc-ET-poly.

To perform the experiments, we used 22 real-world datasets. These datasets were previously used in several papers [53,62–64], and can be found at https://github.com/UCLA-StarAI/Density-Estimation-Datasets. Additionally, we generated synthetic data from 12 real-world BNs. These BNs were obtained from the bnlearn BN repository http://www.bnlearn.com/bnrepository/, and are cited therein. Table 1 briefly describes the basic properties of each dataset.

For each dataset we learned three BNs with each of the compared methods, using different treewidth bounds (3, 5 and 7). In all cases, the score function to maximize was BIC, k-greedy and k-MAX require to fix a maximum execution time, which we set to \( n \) seconds (i.e., a second for each variable) to compute the cache of best parent sets and \( n/10 \) seconds for the structure search. These values were used by Scanagatta et al. [53] in their experiments. To compare the results we used the following performance measures: the BIC score and the log-likelihood (LL) of the models in the training dataset, the learning time, and the treewidth of the returned models.

We analyzed the significance of the differences found for each performance measure in all the datasets and for all the treewidth bounds using the Friedman test with \( \alpha = 0.05 \) and Holm’s [65] and Shaffer’s [66] post-hoc procedures. Both Holm’s and Shaffer’s procedures associate pairwise comparisons with a set of hypotheses and perform a step-down process with the corresponding set of ordered p-values to adjust the value of \( \alpha \) [67].

Experiments were performed on a computer with an Intel Core i7-6700K CPU at 4.00 GHz with 16 GB main memory, running Ubuntu 16.04 LTS. hc-ET and hc-ET-poly were written in Python 2.7.12 and C++11 (version 5.4.0), while k-greedy and k-MAX were downloaded from http://ipg.idsia.ch/software/blip and are written in Java.

5.1. Comparison

Tables 2–4 give an overview of the results obtained using the treewidth bounds 3, 5 and 7, respectively. For each performance measure, the mean rank \( \pm \) the standard deviation of each method over all the datasets is shown. The ranking of the methods is given by their average performance (BIC, LL and time) compared to the rest (i.e., the best is ranked the first and the worst is ranked the fourth). The mean treewidth \( \pm \) the standard deviation is also shown. The detailed results are supplied as Supplementary Material.
Each experiments, treewidth followed significant in all datasets. Results presented in Figs. 13–15 graphically represent these findings using Shaffer’s procedure [68]. Each figure presents both procedures, given that the significant differences observed by Shaffer’s procedure are identical to those observed by Holm’s procedure.

Fig. 13 shows significant differences between the BIC score achieved by all methods; hc-ET performs the best overall, followed by hc-ET-poly, k-MAX, and k-greedy. Moreover, Tables 2–4 show that similar results can be found for all the tested treewidth bounds. The detailed results also show that hc-ET performs better than k-MAX and k-greedy in over 94% of the experiments, and performs better than hc-ET-poly in around 80% of the experiments. The treewidth of the models output by each method suggests that one of the reasons why our proposal manages to optimize better the BIC score is that it allows a tighter fitting to the treewidth bound.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N. vars</th>
<th>N. inst</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLCS</td>
<td>16</td>
<td>16,181</td>
</tr>
<tr>
<td>MSNDBC1</td>
<td>17</td>
<td>291,326</td>
</tr>
<tr>
<td>KDDCup</td>
<td>65</td>
<td>180,092</td>
</tr>
<tr>
<td>Plants</td>
<td>69</td>
<td>17,412</td>
</tr>
<tr>
<td>Audio</td>
<td>100</td>
<td>15,000</td>
</tr>
<tr>
<td>Jester</td>
<td>100</td>
<td>9,000</td>
</tr>
<tr>
<td>Netflix</td>
<td>100</td>
<td>15,000</td>
</tr>
<tr>
<td>Accidents</td>
<td>111</td>
<td>12,758</td>
</tr>
<tr>
<td>Retail</td>
<td>135</td>
<td>22,041</td>
</tr>
<tr>
<td>Pumsb-star</td>
<td>163</td>
<td>12,262</td>
</tr>
<tr>
<td>DNA</td>
<td>180</td>
<td>1,600</td>
</tr>
<tr>
<td>Kosarek</td>
<td>190</td>
<td>33,375</td>
</tr>
<tr>
<td>MSWeb</td>
<td>294</td>
<td>29,441</td>
</tr>
<tr>
<td>Book</td>
<td>500</td>
<td>8,700</td>
</tr>
<tr>
<td>EachMovie</td>
<td>500</td>
<td>4,525</td>
</tr>
<tr>
<td>WebKB</td>
<td>839</td>
<td>2,803</td>
</tr>
<tr>
<td>Reuters-521</td>
<td>889</td>
<td>6,532</td>
</tr>
<tr>
<td>20 NewsGroup</td>
<td>910</td>
<td>11,293</td>
</tr>
<tr>
<td>Movie reviews</td>
<td>1,001</td>
<td>1,600</td>
</tr>
<tr>
<td>BRC</td>
<td>1,058</td>
<td>1,670</td>
</tr>
<tr>
<td>Voting</td>
<td>1,359</td>
<td>1,214</td>
</tr>
<tr>
<td>Ad</td>
<td>1,556</td>
<td>2,461</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N. vars</th>
<th>N. inst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hailfnder</td>
<td>56</td>
<td>5,000</td>
</tr>
<tr>
<td>Hepar II</td>
<td>70</td>
<td>5,000</td>
</tr>
<tr>
<td>Win95pts</td>
<td>76</td>
<td>5,000</td>
</tr>
<tr>
<td>Pathfinder</td>
<td>135</td>
<td>5,000</td>
</tr>
<tr>
<td>Munin1</td>
<td>186</td>
<td>5,000</td>
</tr>
<tr>
<td>Andes</td>
<td>223</td>
<td>5,000</td>
</tr>
<tr>
<td>Diabetes</td>
<td>413</td>
<td>5,000</td>
</tr>
<tr>
<td>Pigs</td>
<td>413</td>
<td>5,000</td>
</tr>
<tr>
<td>Link</td>
<td>724</td>
<td>5,000</td>
</tr>
<tr>
<td>Munin2</td>
<td>1,003</td>
<td>5,000</td>
</tr>
<tr>
<td>Munin3</td>
<td>1,041</td>
<td>5,000</td>
</tr>
<tr>
<td>Munin4</td>
<td>1,038</td>
<td>5,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>hc-ET</th>
<th>hc-ET-poly</th>
<th>k-greedy</th>
<th>k-MAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean rank BIC</td>
<td>1.32 ± 0.53</td>
<td>1.79 ± 0.54</td>
<td>3.79 ± 0.41</td>
<td>3.09 ± 0.62</td>
</tr>
<tr>
<td>Mean rank LL</td>
<td>1.38 ± 0.55</td>
<td>1.68 ± 0.47</td>
<td>3.79 ± 0.41</td>
<td>3.15 ± 0.56</td>
</tr>
<tr>
<td>Mean rank time</td>
<td>1.88 ± 0.33</td>
<td>1.12 ± 0.33</td>
<td>3.56 ± 0.5</td>
<td>3.44 ± 0.5</td>
</tr>
<tr>
<td>Mean treewidth</td>
<td>3 ± 0</td>
<td>3 ± 0</td>
<td>2.85 ± 0.66</td>
<td>2.68 ± 0.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>hc-ET</th>
<th>hc-ET-poly</th>
<th>k-greedy</th>
<th>k-MAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean rank BIC</td>
<td>1.35 ± 0.69</td>
<td>1.91 ± 0.57</td>
<td>3.76 ± 0.61</td>
<td>2.97 ± 0.63</td>
</tr>
<tr>
<td>Mean rank LL</td>
<td>1.47 ± 0.71</td>
<td>1.71 ± 0.52</td>
<td>3.71 ± 0.63</td>
<td>3.12 ± 0.59</td>
</tr>
<tr>
<td>Mean rank time</td>
<td>1.88 ± 0.33</td>
<td>1.12 ± 0.33</td>
<td>3.56 ± 0.5</td>
<td>3.44 ± 0.5</td>
</tr>
<tr>
<td>Mean treewidth</td>
<td>4.88 ± 0.48</td>
<td>4.91 ± 0.38</td>
<td>3.56 ± 1.13</td>
<td>3.41 ± 1.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>hc-ET</th>
<th>hc-ET-poly</th>
<th>k-greedy</th>
<th>k-MAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean rank BIC</td>
<td>1.35 ± 0.69</td>
<td>1.97 ± 0.58</td>
<td>3.79 ± 0.59</td>
<td>2.88 ± 0.69</td>
</tr>
<tr>
<td>Mean rank LL</td>
<td>1.29 ± 0.46</td>
<td>1.82 ± 0.52</td>
<td>3.74 ± 0.62</td>
<td>3.15 ± 0.56</td>
</tr>
<tr>
<td>Mean rank time</td>
<td>1.91 ± 0.29</td>
<td>1.09 ± 0.29</td>
<td>3.41 ± 0.5</td>
<td>3.59 ± 0.5</td>
</tr>
<tr>
<td>Mean treewidth</td>
<td>6.74 ± 0.96</td>
<td>6.74 ± 0.86</td>
<td>3.91 ± 1.31</td>
<td>4 ± 1.44</td>
</tr>
</tbody>
</table>
The comparison of the log-likelihood in Tables 2–4 and Fig. 14 leads us to conclusions that are similar to those drawn for the BIC score. Nevertheless, in this case no significant differences were found between hc-ET and hc-ET-poly. This suggests that hc-ET-poly requires a higher number of parameters to obtain a similar fitting to hc-ET in terms of log-likelihood.

As shown in Fig. 15, we observed significant differences between the learning time of all the methods with the exception of k-MAX and k-greedy. The latter result was expected, given that the imposed limit in execution time of both approaches is the same. hc-ET-poly is the fastest method in all cases, followed by hc-et. However, we must be cautious when interpreting these results. First, these methods are implemented in different programming languages. Second, the bound in execution time set for k-greedy and k-MAX compels their learning time to scale linearly. Therefore, the difference in learning time between our method and k-MAX and k-greedy is clearly higher in the smaller datasets. Finally, although hc-et takes slightly more time than hc-ET-poly in all the experiments, we think that the improvement in BIC score is worthwhile in most situations.

6. Conclusions and future research

Traditional methods for learning BNs usually output models where exact inference is intractable. In this paper, we provide a novel framework for learning tractable BNs. We defined valid ETs, and proposed compilation methods for adapting valid ETs to any local change that may be applied to a BN (i.e., arc addition, removal, and reversal). We proved that the proposed methods always return valid ETs in polynomial time (Theorems 1 and 2). Our approach can be easily combined with any score+search BN learning method that uses only local changes in the network during the structure search. Valid ETs can be used to search in the combined space of DAGs and EOs, avoiding redundant solutions (i.e., all the EOs that are equivalent for a BN are represented by the same valid ET). Hence, we used this representation to efficiently bound the inference complexity of each BN during the learning process.

Experimental results showed that our approach places a tight upper bound on the inference complexity of the networks. The models learned with the proposed methods were competitive with other state-of-the-art methods, performing better in terms of BIC score and log-likelihood in most cases.

Future research will focus on adapting this framework to learning tractable multidimensional BN classifiers (MBCs) [69, 70]. Most probable explanations can be computed in polynomial time in the treewidth of the pruned graph of an MBC [71], that is, a transformation of its structure that entails moralizing the respective DAG and removing the feature variables from the resultant graph. Thus, the application of the methods proposed in this paper to bound the treewidth of the pruned graph of an MBC should lead to tractable MBCs.

Also, we aim to study the relationship between the density of DAGs and the number of equivalent EOs. This would clarify the situations in which it is better to use ETs during the learning process.

Existing methods for learning BNs from incomplete datasets, such as the structural expectation–maximization algorithm [72], require inference during the learning process. We plan to bound the inference complexity of the models to make the structure search tractable when there are missing values or latent variables. Scanagatta et al. [53] successfully introduced the k-MAX algorithm in the maximization step of the structural expectation–maximization algorithm. Thus, we think that adapting our proposal to this problem could lead to promising results.
Fig. 16. ET $E_G'$ yielded after compiling an arc addition $X_{out} \rightarrow X_{in}$ in $E_B$ when $X_{out} \in \text{Pred}_{E_B}(\phi_{X_{in}})$. The value of the cluster of each node in $E_B'$ is shown near to the respective node, and the changes in the clusters with respect to their value in $E_B$ are underlined.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science, Innovation and Universities through the Cajal Blue Brain (C080020-09; the Spanish partner of the Blue Brain initiative from EPFL) and TIN2016-79684-P projects, by the Regional Government of Madrid through the S2013/ICE-2845-CASI-CAM-CM project, and by Fundación BBVA grants to Scientific Research Teams in Big Data 2016. This project has received funding from the European Union’s Horizon 2020 Framework Programme for Research and Innovation under Specific Grant Agreement No. 785907 (HBP SGA2). M. Benjumeda is supported by a predoctoral contract for the formation of doctors from the Spanish Ministry of Science, Innovation and Universities (BES-2014-068637).

Appendix A. Proof of Theorem 1

We use the following lemmas to prove that the compilation and optimization methods proposed in this paper always return valid ETs.

Lemma 1. Let $E_B$ be a valid ET that represents $B$ over $\mathcal{X}$. Given $X_{out}, X_{in} \in \mathcal{X}$, the ET $E_B'$ yielded after applying add($E_B, X_{out}, X_{in}$) in Algorithm 2, is also a valid ET representing $B'$ over $\mathcal{X}$.

Proof. By cases:

We prove that, for each possible arc addition scenario, Algorithm 2 always outputs valid ETs. We show that, in each case, all the nodes are complete (i.e., for each node its cluster in $E_B'$ contains its parents (Definition 9)) and sound (i.e., for each node its cluster in $E_B'$ is a subset of its predecessors and itself (Definition 8)).

• **Case 1**: $X_{out} \in \text{Pred}_{E_B}(\phi_{X_{in}})$.
  This occurs when neither of the conditions in lines 4 and 6 (Algorithm 1) are fulfilled. Algorithm 1 does not produce any change in the structure of $E_B$ (see Fig. 16). Hence, neither the parents nor the predecessors of each node in $E_B$ change. For each node $X_i \in (\text{Pred}_{E_B'}(\phi_{X_{in}}) \cap \text{Desc}_{E_B'}(X_{out})) \cup \{\phi_{X_{in}}\}$, the cluster of $X_i$ now contains $X_{out}$, but $X_{out} \in \text{Pred}_{E_B}(X_i)$.

  Hence, each $X_i$ is sound. As $\text{Cls}_{E_B'}(X_i) \supseteq \text{Cls}_{E_B}(X_i)$, each $X_i$ is complete, and therefore valid. There are no changes in the clusters of the other nodes. Hence, they are valid.

• **Case 2**: $X_f = \text{Pa}_{E_B}(\phi_{X_{in}}) \in \text{Pred}_{E_B}(X_{out})$ (line 4 of Algorithm 2).
  Here, Algorithm 1 sets $\text{Pa}_{E_B'}(\phi_{X_{in}})$ to $X_{out}$ (line 5), and the predecessors and parents of the other nodes are unchanged (see Fig. 17).

  - For each node $X_i \in \text{Pred}_{E_B'}(\phi_{X_{in}}) \cap \text{Desc}_{E_B'}(X_f)$, we have that $\text{Cls}_{E_B'}(X_i) = \text{Cls}_{E_B}(X_i) \cup \text{Cls}_{E_B}(\phi_{X_{in}})$. First, $\text{Cls}_{E_B}(X_i) \subseteq \text{Pred}_{E_B}(X_i) = \text{Pred}_{E_B'}(X_i)$. Also, $\text{Cls}_{E_B}(\phi_{X_{in}}) \subseteq \text{Pred}_{E_B}(\phi_{X_{in}}) = \text{Pred}_{E_B}(X_f) \cup \{X_f\} \subseteq \text{Pred}_{E_B}(X_i) = \text{Pred}_{E_B'}(X_i)$.

    Therefore, $\text{Cls}_{E_B'}(X_i) \subseteq \text{Pred}_{E_B'}(X_i)$, and $X_i$ is sound in $E_B'$. As $\text{Cls}_{E_B'}(\phi_{X_{in}}) = \text{Cls}_{E_B}(\phi_{X_{in}}) \cup \{X_{out}\}$ and $\text{Pred}_{E_B'}(\phi_{X_{in}}) \supseteq \text{Pred}_{E_B}(\phi_{X_{in}}) \cup \{X_{out}\}$, $\phi_{X_{in}}$ is sound. The rest of the nodes are sound given that there are no changes in their clusters.

    - Node $\phi_{X_{in}}$ is complete given that $X_{out} = \text{Pa}_{E_B'}(\phi_{X_{in}})$ and $X_{out} \in \text{Cl}_{E_B'}(\phi_{X_{in}})$. The rest of the nodes are complete given that $\text{Cl}_{E_B'}(X_i) \supseteq \text{Cl}_{E_B}(X_i)$ and $\text{Pa}_{E_B'}(X_i) = \text{Pa}_{E_B}(X_i)$ for each $X_i \in (\mathcal{X} \setminus \text{Leaves}(E_B)) \setminus \{\phi_{X_{in}}\}$. 


Fig. 17. (a) $\mathcal{E}_{B^1}$ is the ET yielded after adding $X_{out} \rightarrow X_m$ in $B$ when $X_f \in \text{Pred}_{\mathcal{E}_B}(X_{out})$ before the compilation of the arc addition; (b) $\mathcal{E}_{B^0}$ is the result of compiling the arc addition $X_{out} \rightarrow X_m$ in $\mathcal{E}_B$. The value of the cluster of each node in $\mathcal{E}_{B^0}$ is shown near to the respective node, and the changes in the clusters with respect to their value in $\mathcal{E}_B$ are underlined. The changes that compromise the validity of the ET are highlighted in red. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

• Case 3: $X_{out} \notin \text{Pred}_{\mathcal{E}_B}(\phi_{X_m})$ and $X_f \notin \text{Pred}_{\mathcal{E}_B^0}(X_{out})$ (line 6 of Algorithm 2).

   In this case, there are two possible output ETs, $\mathcal{E}_{B^1}^1$ and $\mathcal{E}_{B^2}^2$ (line 8).

   1. In $\mathcal{E}_{B^1}^1$ (see Fig. 18b), $Pa_{\mathcal{E}_{B^1}^1}(X_k)$ is set to $X_{out}$ (line 10):

      - Each node $X_i$ that is not in $(\text{Pred}_{\mathcal{E}_{B^1}^1}(\phi_{X_m}) \cap \text{Desc}_{\mathcal{E}_{B^1}^1}(X_m)) \cup \{\phi_{X_m}\}$ has the same parents and clusters in $\mathcal{E}_B$ and $\mathcal{E}_{B^1}^1$, and $\text{Pred}_{\mathcal{E}_{B^1}^1}(X_k) \ni \text{Pred}_{\mathcal{E}_B}(X_i)$. Hence, it is valid.

      - Each node $X_i$ in $\text{Pred}_{\mathcal{E}_{B^1}^1}(X_k) \cap \text{Desc}_{\mathcal{E}_{B^1}^1}(X_m)$ has the same predecessors and parents in $\mathcal{E}_B$ and $\mathcal{E}_{B^1}^1$, and $\text{Cls}_{\mathcal{E}_{B^1}^1}(X_i) = \text{Cls}_{\mathcal{E}_B}(X_i) \cup \text{Cls}_{\mathcal{E}_B}(X_k) \setminus \{X_k\}$ (making $X_i$ complete). As $\text{Cls}_{\mathcal{E}_B}(X_k) \setminus \{X_k\} \subseteq \text{Pred}_{\mathcal{E}_B}(X_k) \subseteq \text{Pred}_{\mathcal{E}_{B^1}^1}(X_k)$, each $X_i$ is also sound. Thus, each $X_i$ is valid.

      - For each $X_i \in (\text{Pred}_{\mathcal{E}_{B^1}^1}(\phi_{X_m}) \cap \text{Desc}_{\mathcal{E}_{B^1}^1}(X_m)) \cup \{\phi_{X_m}\}$, $\text{Cls}_{\mathcal{E}_{B^1}^1}(X_i) = \text{Cls}_{\mathcal{E}_B}(X_i) \cup \{X_m\}$, $\text{Pred}_{\mathcal{E}_{B^1}^1}(X_i) \cup \{X_m\} \ni \text{Pred}_{\mathcal{E}_B}(X_i)$, and $\text{Pred}_{\mathcal{E}_{B^1}^1}(X_i) \cup \{X_m\}$, $\text{Cls}_{\mathcal{E}_B}(X_i) \cup \{X_m\} \ni \text{Pred}_{\mathcal{E}_B}(X_i)$, $Pa_{\mathcal{E}_{B^1}^1}(X_i) = Pa_{\mathcal{E}_B}(X_i)$ if $X_i \neq X_k$, and $Pa_{\mathcal{E}_{B^1}^1}(X_k) = X_{out}$. Hence, $X_i$ is valid.

   As each node in $\mathcal{E}_{B^1}^1$ is valid, $\mathcal{E}_{B^2}^1$ is valid.

   2. In $\mathcal{E}_{B^2}^2$ (see Fig. 18c), $Pa_{\mathcal{E}_{B^2}^2}(X_k)$ is set to $X_f$ (line 12).

      - Each node has the same parent in $\mathcal{E}_{B^2}^2$ and $\mathcal{E}_B$, with the exception of $\phi_{X_m}$ and $X_h$, where $Pa_{\mathcal{E}_{B^2}^2}(\phi_{X_m}) = X_{out}$ (line 13) and $Pa_{\mathcal{E}_{B^2}^2}(X_h) = X_f$. All the nodes are complete, given that $X_{out} \in \text{Cls}_{\mathcal{E}_{B^2}^2}(\phi_{X_m})$, $\text{Cls}_{\mathcal{E}_{B^2}^2}(X_i) \ni \text{Cls}_{\mathcal{E}_B}(\phi_{X_m}) \cup \{X_f\}$, and for each other node $X_i$, $\text{Cls}_{\mathcal{E}_{B^2}^2}(X_i) \ni \text{Cls}_{\mathcal{E}_B}(X_i)$.

      - For each node $X_i$ not in $(\text{Pred}_{\mathcal{E}_{B^2}^2}(\phi_{X_m}) \cap \text{Desc}_{\mathcal{E}_{B^2}^2}(X_m)) \cup \{\phi_{X_m}\}$, the clusters of $X_i$ are the same in $\mathcal{E}_{B^2}^2$ and in $\mathcal{E}_B$ and $\text{Pred}_{\mathcal{E}_{B^2}^2}(X_i) \ni \text{Pred}_{\mathcal{E}_B}(X_i)$. Hence each $X_i$ is sound.

      - For each $X_i \in (\text{Pred}_{\mathcal{E}_{B^2}^2}(\phi_{X_m}) \cap \text{Desc}_{\mathcal{E}_{B^2}^2}(X_m)) \cup \{\phi_{X_m}\}$, $\text{Cls}_{\mathcal{E}_{B^2}^2}(X_i) = \text{Cls}_{\mathcal{E}_B}(X_i) \cup \text{Cls}_{\mathcal{E}_B}(X_h) \setminus \{X_h\}$ and $\text{Cls}_{\mathcal{E}_B}(X_h) \setminus \{X_h\} \subseteq \text{Pred}_{\mathcal{E}_B}(X_h) \ni \text{Pred}_{\mathcal{E}_{B^2}^2}(X_h)$, and $\text{Cls}_{\mathcal{E}_B}(X_h) \setminus \{X_h\} \ni \text{Pred}_{\mathcal{E}_{B^2}^2}(X_h)$, $Pa_{\mathcal{E}_{B^2}^2}(X_i) = Pa_{\mathcal{E}_B}(X_i)$ if $X_i \neq X_k$, and $Pa_{\mathcal{E}_{B^2}^2}(X_k) = X_{out}$. Hence, $X_i$ is sound.

      - Node $\phi_{X_m}$ contains $X_{out}$ in its cluster and predecessors in $\mathcal{E}_{B^2}^2$. Hence, $\text{Cls}_{\mathcal{E}_{B^2}^2}(\phi_{X_m}) = \text{Cls}_{\mathcal{E}_{B^2}^2}(\phi_{X_m}) \cup \{X_{out}\} \ni \text{Pred}_{\mathcal{E}_{B^2}^2}(\phi_{X_m}) \cup \{X_{out}\}$, making $\phi_{X_m}$ sound.

   As every node in $\mathcal{E}_{B^2}^2$ is sound and complete, $\mathcal{E}_{B^2}^2$ is valid. □

**Lemma 2.** Let $\mathcal{E}_B$ be a valid ET that represents $B$ over $X$. Given $X_{out}, X_m \in X$, the ET $\mathcal{E}_{B^0}$ that represents $B^0$ output after applying remove($\mathcal{E}_B, X_{out}, X_m$) in Algorithm 3, is also valid.
Fig. 18. $E_{B'}$ (a) is the ET yielded after adding $X_{out} \rightarrow X_{in}$ in $B$ when $X_{out} \notin \text{Pred}_{E_B}(\phi_{X_{in}})$ and $X_{f} \notin \text{Pred}_{E_B}(X_{out})$ before the compilation of the arc addition. $E_{B'}^1$ (b) and $E_{B'}^2$ (c) correspond to the two possible outcomes of compiling the arc addition. The value of the cluster of each node in $E_{B'}$ is shown near to the respective node, and the changes in the clusters with respect to their value in $E_B$ are underlined. The changes that compromise the validity of the ET are highlighted in red.

**Proof.** By induction. We show that after removing an arc from $E_{B'}$ only one node $X_i$ may not be complete (base case). In each iteration $\text{iter}$ of Algorithm 3, the completeness of $X_i$ is amended and $E_{B'}^{\text{iter}}$ is built, and only one other node $X'_i$, which was a predecessor of $X_i$ in the previous ET, may not be complete after the change. It is evident that eventually node $X'_i$ will be complete (e.g., when the parent of $X'_i$ is the root node $*$).

**Base case:**
Given a valid ET $E_{B'}$, removing arc $X_{out} \rightarrow X_{in}$ from $B$ will produce an ET $E_{B'}$. For each $X_h \in \mathcal{X} \cup \text{Leaves}(E_B)$, $\text{Pred}_{E_{B'}}(X_h) = \text{Pred}_{E_B}(X_h)$, $\text{Pa}_{E_{B'}}(X_h) = \text{Pa}_{E_B}(X_h)$, $\text{Cls}_{E_{B'}}(X_h) \supseteq \text{Cls}_{E_B}(X_h) \supseteq \text{Cls}_{E_B}(X_h) \setminus \{X_{out}\}$ if $X_h \in \text{Desc}_{E_{B'}}(X_{out}) \cap \text{Pred}_{E_{B'}}(\phi_{X_{in}}) \cup \{\phi_{X_{in}}\}$, and $\text{Cls}_{E_{B'}}(X_h) = \text{Cls}_{E_B}(X_h)$ otherwise. Therefore, each node in $E_{B'}$ is sound, and only one node $X_i$ such that $X_i \in \text{Ch}_{E_{B'}}(X_{out}) \cap \text{Pred}_{E_{B'}}(\phi_{X_{in}}) \cup \{\phi_{X_{in}}\}$ may not be complete.

**Iterative step:**
Assume that $E_{B'}^{\text{iter}}$ (Fig. 19a) is sound and only node $X_i$ is not complete. Algorithm 3 sets $\text{Pa}_{E_{B'}}(X_i) = X_j$ (line 7), that is, the deepest node in $E_{B'}^{\text{iter}}$ belonging to $\text{Cls}_{E_{B'}}(X_i) \setminus \{X_i\}$ (line 6). Hence, all $X_i$ and their descendants are sound. Thus, node
Theorem 1

We have shown that the resulting ET $E_B'$ in (Fig. 19b) satisfies Lemma 1 and Lemma 2. Therefore, $E_B'$ is complete.

Proof. We prove that $E_B'$ is complete by showing that for each node $X_j$ in $E_B'$, there is a valid ET $E_B$ such that $E_B'$ is the result of incrementally compiling $E_B$. We use Lemma 1 and Lemma 2 to show that $E_B'$ is complete.

Let $E_B$ be a valid ET over $\mathcal{X}$. We construct $E_B'$ from $E_B$ by performing a sequence of arc removals and optimizing the ET using Algorithm 5. For each arc removal, we ensure that the resulting ET $E_B'$ is valid. Therefore, $E_B'$ is complete.

Finally, Theorem 1 can be proved using the above lemmas.

**Theorem 1.** Let $E_B$ be a valid ET over $\mathcal{X}$, and $E_B'$ be the result of incrementally compiling on $E_B$ any local change in $B$ using Algorithms 2 and 3 and optimizing the resulting ET using Algorithm 5. Then $E_B'$ is a valid ET.

Proof. By Lemmas 1 and 2 we know that if $E_B$ is valid, the tree returned after compiling a local change in $E_B$ is also valid. Hence, if the input for Algorithm 5 is a valid ET, we know, by Lemma 3 (the optimization is composed of a sequence of swaps), that it will also return a valid ET.
Appendix B. Proof of Theorem 2

The following lemmas are used later to prove Theorem 2. First, we need to know the computational cost of outputting the cluster of a node in an ET.

**Lemma 4.** Let \( \mathcal{E}_B \) be an ET over \( \mathcal{X} = \{X_1, \ldots, X_n\} \). The cluster of a node \( X_i \in \mathcal{X} \) can be computed in time \( O(|\text{Ch}_{\mathcal{E}_B}(X_i)| \cdot \text{width}(\mathcal{E}_B)) \) given the clusters of the nodes in \( \text{Ch}_{\mathcal{E}_B}(X_i) \).

**Proof.** The cluster of node \( X_i \) can be output by computing the union of the clusters of its children in \( \mathcal{E}_B \) (Definition 7). The union of sets \( S_1, \ldots, S_m \) can be computed in time \( |S_1| + \cdots + |S_m| \). As the size of each cluster in \( \mathcal{E}_B \) is less than or equal to \( \text{width}(\mathcal{E}_B) + 1 \), then \( \sum_{X_j \in \text{Ch}_{\mathcal{E}_B}(X_i)} |\text{Cls}_{\mathcal{E}_B}(X_j)| \leq \sum_{X_j \in \text{Ch}_{\mathcal{E}_B}(X_i)} (\text{width}(\mathcal{E}_B) + 1) = |\text{Ch}_{\mathcal{E}_B}(X_i)| \cdot (\text{width}(\mathcal{E}_B) + 1) \). Hence, the cluster of \( X_i \) can be output in time \( O(|\text{Ch}_{\mathcal{E}_B}(X_i)| \cdot \text{width}(\mathcal{E}_B)) \). \( \square \)

**Lemma 5.** Let \( \mathcal{E}_B \) be an ET over \( \mathcal{X} = \{X_1, \ldots, X_n\} \). All the clusters in \( \mathcal{E}_B \) can be computed in time \( O(n \cdot \text{width}(\mathcal{E}_B)) \).

**Proof.** The cluster of a leaf node is its domain. The cluster of the inner nodes can be output bottom-up such that before computing the cluster of node \( X_i \), the cluster of each child of \( X_i \) in \( \mathcal{E}_B \) is known. From Lemma 4, we know that the cluster of each \( X_i \in \mathcal{X} \) can be output in time \( O(|\text{Ch}_{\mathcal{E}_B}(X_i)| \cdot \text{width}(\mathcal{E}_B)) \). Hence, the clusters of all the nodes in \( \mathcal{X} \) can be computed in time \( O(\sum_{X_i \in \mathcal{X}} |\text{Ch}_{\mathcal{E}_B}(X_i)| \cdot \text{width}(\mathcal{E}_B)) \). As \( \sum_{X_i \in \mathcal{X}} |\text{Ch}_{\mathcal{E}_B}(X_i)| \cdot \text{width}(\mathcal{E}_B) < 2n \cdot \text{width}(\mathcal{E}_B) \) (there are \( n \) inner nodes with only one parent, of which at least one is a child of the root node *, and \( n \) edges including inner and leaf nodes), all the clusters of \( \mathcal{E}_B \) can be computed in time \( O(n \cdot \text{width}(\mathcal{E}_B)) \). \( \square \)

A local change in an ET \( \mathcal{E}_B \) produces changes in the clusters of the tree, which have an influence on the computational complexity of Algorithm 5.

**Lemma 6.** Let \( \mathcal{E}_B \) be a valid ET, and \( \mathcal{E}_B' \) the result of swapping (Algorithm 4) a node \( X_i \) and its parent in \( \mathcal{E}_B \). Then

\[
\text{width}(\mathcal{E}_B') \leq 2 \cdot \text{width}(\mathcal{E}_B).
\]

**Proof.** After swapping \( X_i \) and its parent \( X_p \) in \( \mathcal{E}_B \), only the clusters of \( X_i \) and \( X_p \) may change. On the one hand, \( \text{Cls}_{\mathcal{E}_B'}(X_p) \subseteq \text{Cls}_{\mathcal{E}_B}(X_p) \). Hence, the width of \( X_p \) does not grow. On the other hand, the width of \( X_i \) may grow, but \( \text{Cls}_{\mathcal{E}_B'}(X_i) \subseteq (\text{Cls}_{\mathcal{E}_B}(X_i) \cup \text{Cls}_{\mathcal{E}_B}(X_p)) \setminus \{X_p\} \). Hence, the width of \( \text{Cls}_{\mathcal{E}_B'}(X_i) \) is less than \( |\text{Cls}_{\mathcal{E}_B}(X_i)| + |\text{Cls}_{\mathcal{E}_B}(X_p)| - 1 \leq 2(\text{width}(\mathcal{E}_B)) + 1 \). This means that \( \text{width}(\mathcal{E}_B') \leq 2 \cdot \text{width}(\mathcal{E}_B) \). \( \square \)

Next, we bound the time complexity of the compilation and optimization methods.

**Lemma 7.** Let \( \mathcal{E}_B \) be a valid ET over \( \mathcal{X} = \{X_1, \ldots, X_n\} \). The addition of any arc in \( B \) can be compiled in \( \mathcal{E}_B \) in time \( O(n^2) \) by Algorithm 2.

**Proof.** There are no loops in Algorithm 2, and the only operations that cannot be completed in time \( O(1) \) are:
Therefore, the addition of an arc can be compiled in time $O(n^2)$. □

**Lemma 8.** Let $E_B$ be a valid ET over $\mathcal{X} = \{X_1, \ldots, X_n\}$. The removal of any arc in $B$ can be compiled in $E_B$ in time $O(n^2 \cdot \text{width}(E_B))$ by Algorithm 3.

**Proof.** In each iteration, Algorithm 3 checks if a node $X_i$ contains its current parent in its cluster (line 5); else, the new parent of $X_i$ is set to the deepest node $X_{j'}$, which appears in the cluster of $X_i$ (line 6). Then, the child of $X_{j'}$, which was previously a predecessor of $X_i$, is set as the new $X_i$, and $X_{j'}$ is set as the new $X_j$ for the next iteration (lines 8 and 9 of Algorithm 3). Therefore, node $X_i$ is not visited again in the next iterations. This means that there are fewer than $n$ iterations.

The operations in lines 6-9 of Algorithm 3 can be completed in time $O(n)$. The clusters of several nodes must be output after each iteration. By Lemma 5, we know that the clusters of all the nodes in $\mathcal{X}$ can be computed in time $O(n \cdot \text{width}(E_B))$.

As there are fewer than $n$ iterations, Algorithm 3 can be run in time $O(n^2 \cdot \text{width}(E_B) + n^2) = O(n^2 \cdot \text{width}(E_B))$. □

**Lemma 9.** Let $E_B$ be a valid ET over $\mathcal{X} = \{X_1, \ldots, X_n\}$. Swapping node $X_i$ and its parent $X_p$ in $E_B$ using Algorithm 4 and updating the clusters of $E_B$ can be completed in time $O(\text{width}(E_B)(\text{Ch}_{E_B}(X_i)) + \text{Ch}_{E_B}(X_p)))$.

**Proof.** To swap a node $X_i \in \mathcal{X}$ with its parent in $E_B$, Algorithm 4 assigns a new parent to $X_i$ and to its previous parent $X_p$. It also assigns $X_p$ as the new parent of any children of $X_i$ whose cluster contains $X_p$. This can be completed in time $O(|\text{Ch}_{E_B}(X_i)|)$. Note that we can check if $X_p$ belongs to a cluster in time $O(1)$.

After swapping $X_i$ and $X_p$, only the clusters of $X_i$ and $X_p$ change. By Lemma 4, we know that this can be computed in $O(|\text{Ch}_{E_B}(X_i)|\text{width}(E_B) + |\text{Ch}_{E_B}(X_p)|\text{width}(E_B)) = O(\text{width}(E_B)(|\text{Ch}_{E_B}(X_i)| + |\text{Ch}_{E_B}(X_p)|))$. □

**Lemma 10.** Let $E_B$ be a valid ET over $\mathcal{X} = \{X_1, \ldots, X_n\}$. Algorithm 5 can be computed in time $O(n^2 \cdot \text{width}(E_B))$.

**Proof.** The input of Algorithm 5 is a list of nodes $X_{\text{opt}} = (X_{l(1)}, \ldots, X_{l(m)})$ for optimization. Assuming that these nodes are ordered from the shallowest to the deepest (i.e., the depth of $X_{l(i)}$ is greater than or equal to the depth of $X_{l(i+1)}$), Algorithm 5 starts swapping $X_{l(1)}$ while the width of the ET does not increase, and then it performs the same process with $X_{l(2)}, \ldots, X_{l(m)}$. Thus, the cost of Algorithm 5 is given by the cost of each swap performed during the optimization. According to Lemma 9, Algorithm 5 can be computed in time

$$O\left(\sum_{i=1}^{m} \sum_{j=1}^{k_i} \text{width}(E_{B_{i,j}^{i-1}})(|\text{Ch}_{E_{B_{i,j}^{i-1}}}(X_{l(i)})| + |\text{Ch}_{E_{B_{i,j}^{i-1}}}(\text{Pa}_{E_{B_{i,j}^{i-1}}}(X_{l(i)}))|)\right),$$

where:

- $k_i < n$ is the number of times that node $X_{l(i)}$ is swapped.
- $E_{B_{i,j}^{i-1}}$ is the ET obtained after swapping node $X_{l(i)}$, $j$ times after nodes $X_{l(1)}, \ldots, X_{l(i-1)}$ have been optimized.
- $E_{B_{i,j}^{i-1}} = E_{B_{i,j}^{i-1}}$ if $\text{Pa}_{E_{B_{i,j}^{i-1}}}(X_{l(i-1)}) = ∗$ (i.e., swapping node $X_{l(i-1)}$ always reduces the width of the ET candidates until its parent is the root node) and $E_{B_{i,j}^{i-1}} = E_{B_{i,j}^{i-1}}$ otherwise.
- $E_{B_{i,j}^{i-1}} = E_{B}$. When the width of a candidate $E_{B_{i,j}^{i-1}}$ is bigger than $\text{width}(E_B)$, $E_{B_{i,j}^{i-1}}$ is rejected. Thus, by Lemma 6, $\text{width}(E_{B_{i,j}^{i-1}}) \leq 2 \cdot \text{width}(E_B)$, and

$$\sum_{i=1}^{m} \sum_{j=1}^{k_i} \text{width}(E_{B_{i,j}^{i-1}})(|\text{Ch}_{E_{B_{i,j}^{i-1}}}(X_{l(i)})| + |\text{Ch}_{E_{B_{i,j}^{i-1}}}(\text{Pa}_{E_{B_{i,j}^{i-1}}}(X_{l(i)}))|) \leq 2 \cdot \text{width}(E_B)(\sum_{i=1}^{m} \sum_{j=1}^{k_i} \text{width}(E_{B_{i,j}^{i-1}})(|\text{Ch}_{E_{B_{i,j}^{i-1}}}(X_{l(i)})| + |\text{Ch}_{E_{B_{i,j}^{i-1}}}(\text{Pa}_{E_{B_{i,j}^{i-1}}}(X_{l(i)}))|).$$
The complexity of Algorithm 5 can be output by counting the number of children of each $X_{l(i)}$ and its parent in each iteration.

In any ET $\mathcal{E}_{B}^{L}$, there are less than 2n arcs without counting arcs from the root node. Also, note that after swapping node $X_{l(i)}$ with its parent $X_{p}$ in an ET $\mathcal{E}_{B}^{L}$, $\mathcal{C}_{B}^{L}(X_{p}) \supseteq \mathcal{C}_{B}^{L}(X_{p}) \setminus \{X_{l(i)}\}$, $\mathcal{C}_{B}^{L}(X_{l(i)}) \subseteq \mathcal{C}_{B}^{L}(X_{l(i)}) \setminus \{X_{p}\}$ and $\mathcal{C}_{B}^{L}(X_{l(i)}) \cup \mathcal{C}_{B}^{L}(X_{p}) = \mathcal{C}_{B}^{L}(X_{l(i)}) \cup \mathcal{C}_{B}^{L}(X_{p})$. This implies that if a node $X_{c}$ is the child of a node $X_{b}$ in any $\mathcal{E}_{B}^{L}, \ldots, \mathcal{E}_{B}^{L, k-1}$, it cannot be the child of another node that is not $X_{b}$ or $X_{l(i)}$ in $\mathcal{E}_{B}^{L, 0}, \ldots, \mathcal{E}_{B}^{L, k-1}$. Therefore, it is evident that $\sum_{i=1}^{m} |\mathcal{C}_{B}^{L, i-1}(Pa_{B}^{L, i-1}(X_{l(i)}))| < 2n$.

To bound $\sum_{i=1}^{m} |\mathcal{C}_{B}^{L, i-1}(X_{l(i)}))|$, let us focus on the number of children that each node has when it is swapped. As the nodes in $X_{opt}$ are visited from the shallowest to the deepest, if a node $X_{b}$ is a child of node $X_{l(i)}$ when $X_{l(i)}$ is optimized, it cannot be a child of another node $X_{l(j)} \in X_{opt}$ when $X_{l(j)}$ is optimized. Thus, each node $X_{b}$ can be counted less than n times, and given that there are 2n inner and leaf nodes in $\mathcal{E}_{B}^{L}$, $\sum_{i=1}^{m} |\mathcal{C}_{B}^{L, i-1}(X_{l(i)}))| < 2n^2$.

Finally, $2 \cdot \text{width}(\mathcal{E}_{B}^{L}) \cdot (\sum_{i=1}^{m} |\mathcal{C}_{B}^{L, i-1}(X_{l(i)}))| + \sum_{i=1}^{m} \sum_{j=1}^{k_{i}} |\mathcal{C}_{B}^{L, i-1}(Pa_{B}^{L, i-1}(X_{l(i)}))| < \text{width}(\mathcal{E}_{B}^{L}) \cdot (4n^2 + 4n)$. Therefore, Algorithm 5 can be computed in time $O(n^2 \cdot \text{width}(\mathcal{E}_{B}^{L}))$. □

Theorem 2 can be proved using the lemmas shown above.

**Theorem 2.** Let $\mathcal{E}_{B}^{L}$ be a valid ET over a set of variables $\mathcal{X} = \{X_{1}, \ldots, X_{n}\}$. The process described in Theorem 1 to output $\mathcal{E}_{B}^{L'}$ can be performed in time $O(n^2 \cdot \text{width}(\mathcal{E}_{B}^{L}))$.

**Proof.** By Lemmas 7, 8 and 10, we know that both the compilation and optimization process can be performed in time $O(n^2 \cdot \text{width}(\mathcal{E}_{B}^{L}))$. □

## Appendix C. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artint.2018.11.007.

## References


