Proceedings Paper:
Studený, Milan and Cussens, James orcid.org/0000-0002-1363-2336 (2016) The Chordal Graph Polytope for Learning Decomposable Models. In: Antonucci, Alessandro, Corani, Giorgio and Polpo de Campos, Cassio, (eds.) Proceedings of the Eighth International Conference on Probabilistic Graphical Models. Journal of Machine Learning Research: Workshop and Conference Proceedings . , pp. 499-510.

JMLR: Workshop and Conference Proceedings vol 52, 499-510, 2016 PGM 2016

The Chordal Graph Polytope for Learning Decomposable Models

Milan Studený STUDENY@UTIA.CAS.CZ
Department of Decision-Making Theory
Institute of Information Theory and Automation of the CAS
Prague, 18208 Pod Vodárenskou věží 4, Czech Republic

James Cussens JAMES.CUSSENS@YORK.AC.UK
Department of Computer Science & York Centre for Complex Systems Analysis
University of York
York, YO10 5GH, United Kingdom

Abstract

This theoretical paper is inspired by an *integer linear programming* (ILP) approach to learning the structure of *decomposable models*. We intend to represent decomposable models by special zero-one vectors, named *characteristic imsets*. Our approach leads to the study of a special polytope, defined as the convex hull of all characteristic imsets for chordal graphs, named the *chordal graph polytope*. We introduce a class of *clutter* inequalities and show that all of them are valid for (the vectors in) the polytope. In fact, these inequalities are even facet-defining for the polytope and we dare to conjecture that they lead to a complete polyhedral description of the polytope. Finally, we propose an LP method to solve the *separation problem* with these inequalities for use in a cutting plane approach.

Keywords: Learning decomposable models; integer linear programming; characteristic imset; chordal graph polytope; clutter inequalities; separation problem.

1. Introduction: Motivation

Decomposable models are fundamental graphical models (Lauritzen, 1996). A well-known fact is that elegant mathematical properties of these statistical models form the theoretical basis of the famous method of local computation (Cowell et al., 1999). Decomposable models, which are described by *chordal undirected graphs*, can be viewed as special cases of Bayesian network models (Pearl, 1988), described by directed acyclic graphs.

There are various methods for learning the structure of decomposable models, most of them being specializations of the methods for learning Bayesian networks (Neapolitan, 2004). There are methods based on statistical conditional independence tests like the PC algorithm (Spirtes et al., 1993) or MCMC simulations (Giudici and Green, 1999). This paper is, however, motivated by a *score-based approach*, where the task is to maximize some additively decomposable *score*, like the BIC score (Schwarz, 1978) or the BDeu score (Heckerman et al., 1995). We are interested in the *integer linear programming* (ILP) approach to structural learning (of decomposable models).

The idea behind this approach is to encode graphical models by certain vectors with integer components in such a way that the usual scores become affine/linear functions of the vector representatives. There are several ways to encode Bayesian network models. The most successful one seems to be to encode them by *family-variable* vectors as used by Jaakkola et al. (2010); Cussens (2011) and Bartlett and Cussens (2013). However, the approach discussed in this paper is based on encoding the models by *characteristic imsets* which were introduced by Hemmecke et al. (2012)

and tested by Studený and Haws (2014). This mode of representation leads to an elegant way of encoding decomposable models which we believe to be particularly suitable for structural learning.

Note that two recent conference papers have also been devoted to ILP-based learning of decomposable models, but they used a different binary coding of the models/graphs. More specifically, Sesh Kumar and Bach (2013) used special codes for junction trees of the graph, while Pérez et al. (2014) encoded certain special coarsenings of maximal hyper-trees. Moreover, restricted learning was the goal in both these papers unlike in this contribution: we aim here at a general ILP method for learning decomposable models.

Two other recent papers devoted to learning decomposable models also used encodings of junction trees. Corander et al. (2013) expressed the search space in terms of logical constraints and used constraint satisfaction solvers. Even better running times have been achieved by Kangas et al. (2014), who applied the idea of decomposing junction trees into subtrees, which allowed them to use the method of dynamic programming. Note that the junction tree representation is closely related to the Möbius inversion of the characteristic imset we mention in § 4.2.

Our approach leads to the study of the geometry of a polytope defined as the convex hull of all characteristic imsets for chordal graphs (over a fixed set of variables N), with possible modification that a clique size limit is given. This polytope has already been dealt with by Lindner (2012) in her thesis, where she derived some basic observations on the polytope. For example, she mentioned that a complete facet description of the polytope with cliques size limit two, which corresponds to learning *undirected forests*, can be derived. She also identified some non-trivial inequalities for the polytope with no clique size limit. Being inspired by Lindner (2012) we name this polytope the "chordal graph characteristic imset polytope", but abbreviate this to *chordal graph polytope*.

In this paper we present a facet description of the polytope where $|N| \leq 4$ and describe the situation for the case $|N| = 5$, where the facet description is also available. We have succeeded in classifying all facet-defining inequalities for this polytope in these cases. What we found out is that, with the exception of a natural *lower bound inequality*, there is a one-to-one correspondence between the facet-defining inequalities and the *clutters* (= antichains = Sperner families) of subsets of the variable set N containing at least one singleton, so we call these *clutter inequalities*.

This establishes a sensible *conjecture* about the complete polyhedral description of the polytope (with no clique size limit). We prove that every clutter inequality is valid for the polytope. We also have a proof that every such inequality is facet-defining for the polytope, which we omit due to lack of space. Finally, we tackle an important *separation problem*: that is, given a non-integer solution to an LP relaxation problem, find a clutter inequality which (most) violates the current solution. We discuss preliminary empirical work on exact and approximate solving of this problem.

2. Basic Concepts

Let N be a finite set of *variables*; assume $n := |N| \geq 2$ to avoid the trivial case. In the statistical context, the elements of N correspond to random variables, while in the graphical context they correspond to nodes of graphs.

2.1 Chordal graphs

An undirected graph G *over* N has N as the set of nodes. It is *chordal* if every cycle of length at least 4 has a chord, that is, an edge connecting non-consecutive nodes in the cycle. A set $S \subseteq N$ is complete if every two distinct nodes in S are connected by an edge. Maximal complete sets with

respect to inclusion are called the *cliques* (of G). A well-known equivalent definition of a chordal graph is that the collection of its cliques can be ordered into a sequence $C_1, \ldots, C_m$, $m \geq 1$, satisfying the *running intersection property* (RIP):

$$\forall\, i \geq 2 \;\; \exists\, j < i \quad \text{such that} \;\; S_i := C_i \cap \left(\bigcup_{\ell < i} C_\ell \right) \subseteq C_j \,.$$

The sets $S_i = C_i \cap (\bigcup_{\ell<i} C_\ell)$, $i = 2, \ldots, m$ are the respective separators. The multiplicity of a separator S is the number of indices $2 \leq i \leq m$ such that $S = S_i$; the separators and their multiplicities are known not to depend on the choice of the ordering satisfying the RIP, see (Studený, 2005, Lemma 7.2). Each chordal graph defines the respective *decomposable model*; see (Lauritzen, 1996, § 4.4).

2.2 Learning

The *score-based* approach to structural learning of graphical models is based on maximizing some *scoring criterion* which is a bivariate real function $(G, D) \mapsto \mathcal{Q}(G, D)$ of the graph G and the (observed) database D. In the context of learning Bayesian networks, a crucial technical assumption (Chickering, 2002) is that $\mathcal{Q}$ should be additively *decomposable*, which means, it has the form

$$\mathcal{Q}(G, D) \,=\, \sum_{a \in N} q_D(a \,|\, \mathrm{pa}_G(a)) \,, \quad \text{where the summands } q_D(* \,|\, *) \text{ are called } \textit{local scores}$$

and $\mathrm{pa}_G(a)$ denotes the set of parents of a node a in G. All criteria used in practice satisfy this requirement, as long as the data contain no missing values. Another typical assumption is that $\mathcal{Q}$ is *score equivalent* (Bouckaert, 1995), which means that Markov equivalent graphs yield the same score.

2.3 Characteristic imset

The concept of a *characteristic imset* was introduced by Hemmecke et al. (2012). Each characteristic imset is an element of the vector space $\mathbb{R}^\Lambda$ where $\Lambda := \{\, S \subseteq N \,:\, |S| \geq 2 \,\}$. A fundamental fact is that every additively decomposable and score equivalent scoring criterion turns out to be an affine function (= a linear function plus a constant) of the characteristic imset encoding of a graph. For the current paper we only need the definition of the characteristic imset for a chordal graph. Specifically, given a chordal graph G over N, the *characteristic imset* of G is a zero-one vector c_G with components indexed by subsets S from Λ:

$$\mathsf{c}_G(S) = \begin{cases} 1 & \text{if } S \text{ is a complete set in } G,\ |S| \geq 2, \\ 0 & \text{for remaining } S \subseteq N,\ |S| \geq 2. \end{cases}$$

We also adopt the convention that $\mathsf{c}_G(L) = 1$ for any graph G over N and $L \subseteq N$ where $|L| = 1$; a conventional value for $\mathsf{c}_G(\emptyset)$ does not play substantial role. Since decomposable models induced by chordal undirected graphs can be viewed as special cases of Bayesian network models each sensible scoring criterion is an affine function of the characteristic imset. Specifically, Lemma 3 in (Studený and Haws, 2014) says that $\mathcal{Q}(G, D) \,=\, k + \sum_{S \in \Lambda} \, r^{\mathcal{Q}}_D(S) \cdot \mathsf{c}_G(S)$ where k is a constant and, for any $S \in \Lambda$, $r^{\mathcal{Q}}_D(S) = \sum_{K \subseteq R} (-1)^{|R \setminus K|} \cdot q_D(a \,|\, R)$, with arbitrary $a \in S$ and $R = S \setminus \{a\}$.

2.4 Chordal graph polytope

Let us introduce the *chordal graph polytope* over N with *cliques size limit* k, $2 \leq k \leq n = |N|$:

$$D_N^k \;:=\; \mathrm{conv}\,(\,\{\mathsf{c}_G \,:\, G \text{ chordal graph over } N \text{ with clique size at most } k\}\,)\,,$$

where conv $(\cdot)$ is used to denote the convex hull. The dimension of D_N^k is $\sum_{\ell=2}^{k} \binom{n}{\ell}$. In particular, for the unrestricted polytope $D_N := D_N^n$ one has dim $(D_N) = 2^n - n - 1$, while the most restricted polytope for learning *undirected forests* has dim $(D_N^2) = \binom{n}{2}$.

3. Example: the Case of a Low Number of Variables

For small values of $n = |N|$ we have been able to use the `cdd` program (Fukuda, 2015) to compute a complete facet description of the unrestricted *chordal graph polytope* D_N. In the case $n = |N| = 3$, D_N has 8 vertices, namely 8 chordal graphs, and 8 facet-defining inequalities, decomposing into 4 permutation types. With $N = \{a, b, c\}$, these are:

lower bound: $0 \leq \mathsf{c}(\{a, b, c\})$,

2-to-3 monotonicity inequalities: $\mathsf{c}(\{a, b, c\}) \leq \mathsf{c}(\{a, b\})$,

upper bounds: $\mathsf{c}(\{a, b\}) \leq 1$,

cluster inequality for 3-element set: $\mathsf{c}(\{a, b\}) + \mathsf{c}(\{a, c\}) + \mathsf{c}(\{b, c\}) \leq 2 + \mathsf{c}(\{a, b, c\})$.

Note that the *cluster inequalities* (formulated in terms of family variables) have already occurred in the context of learning Bayesian networks (Jaakkola et al., 2010; Bartlett and Cussens, 2013; Studený and Haws, 2014).

In the case $n = |N| = 4$, the unrestricted polytope D_N has 61 vertices, i.e. 61 chordal graphs. The number of facets is only 50, decomposing into 9 permutation types. The list of these types is given in § 4.1, where the inequalities are interpreted in terms of clutters.

In the case $n = |N| = 5$, D_N has 822 vertices since there are 822 decomposable models. The number of its facets is again smaller, just 682, and they fall into 29 permutation types. The computation is this case $n = 5$ took more than 24 hours.

4. Clutter Inequalities and a Completeness Conjecture

An interesting observation is this: in the case $n = |N| \leq 5$, with the exception of the lower bound $0 \leq \mathsf{c}(N)$, all facet-defining inequalities for D_N have the form

$$\sum_{S \subseteq N \setminus \{\gamma\}} \kappa(S) \cdot \mathsf{c}(S \cup \{\gamma\}) \;\leq\; \sum_{S \subseteq N \setminus \{\gamma\}} \kappa(S) \cdot \mathsf{c}(S)$$

where γ is a distinguished element of N and the $\kappa(S)$ are integer coefficients. Indeed, the monotonicity inequalities from § 3 have this form: here $\gamma = c$, $\kappa(\{a, b\}) = 1$ and $\kappa(S) = 0$ for $S \subset \{a, b\}$. The cluster inequality from § 3 can also be re-written in the form of a generalized monotonicity inequality: $\mathsf{c}(\{a, c\}) + \mathsf{c}(\{b, c\}) - \mathsf{c}(\{a, b, c\}) \leq \mathsf{c}(\{a\}) + \mathsf{c}(\{b\}) - \mathsf{c}(\{a, b\})$.

A deeper fact is that the inequalities can be interpreted as inequalities induced by certain *clutters* of subsets of N, by which we mean classes of subsets of N that are inclusion incomparable. Such classes are alternatively named *Sperner families* or *antichains*.

Definition 1 *Given a clutter $\mathcal{L}$ of subsets of N which contains at least one singleton and satisfies $|\bigcup \mathcal{L}| \geq 2$, the corresponding* **clutter inequality** *for $\mathsf{c} \in \mathbb{R}^{\Lambda}$ has the form*

$$1 \leq \sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{L}} (-1)^{|\mathcal{B}|+1} \cdot \mathsf{c}\left(\bigcup \mathcal{B}\right) , \tag{1}$$

where a convention is applied that $\mathsf{c}(L) = 1$ whenever $L \subseteq N$, $|L| = 1$.

To re-write (1) in a standard form introduce the notation for the *filter* generated by $\mathcal{L}$:

$$\mathcal{L}^{\uparrow} := \{ T \subseteq N \,:\, \exists\, L \in \mathcal{L} \quad L \subseteq T \}, \qquad \text{to be used in the rest of the paper.}$$

Then, with some minor effort, one can re-write (1) in the following form:

$$1 \;\leq\; \sum_{S \subseteq N} \kappa_{\mathcal{L}}(S) \cdot \mathsf{c}(S) \qquad \text{where } \kappa_{\mathcal{L}}(S) = \sum_{T \subseteq S:\, T \in \mathcal{L}^{\uparrow}} (-1)^{|S \setminus T|} \text{ for any } S \subseteq N. \tag{2}$$

To compute the coefficients $\kappa_{\mathcal{L}}(\cdot)$ from (2) for a given clutter $\mathcal{L}$ a simple recursive procedure can be used. First, it follows from (1) that the coefficients vanish outside the class

$$\mathcal{U}(\mathcal{L}) := \left\{ \bigcup \mathcal{B} \,:\, \emptyset \neq \mathcal{B} \subseteq \mathcal{L} \right\} \qquad \text{of unions of sets from } \mathcal{L}.$$

Second, they can be computed recursively within this class as follows:

$$\kappa_{\mathcal{L}}(S) = 1 - \sum_{T \in \mathcal{U}(\mathcal{L}):\, T \subset S} \kappa_{\mathcal{L}}(T) \qquad \text{for any } S \in \mathcal{U}(\mathcal{L}). \quad \text{In particular, } \kappa_{\mathcal{L}}(L) = 1 \text{ for } L \in \mathcal{L}.$$

Now, the point comes. We have the following conjecture we know is valid in case $|N| \leq 5$.

Conjecture 2 *For any $n = |N| \geq 2$, the set of facet-defining inequalities for $\mathsf{c} \in D_N$ consists of the lower bound $0 \leq \mathsf{c}(N)$ and the inequalities* (1) *for such clutters $\mathcal{L}$ of subsets of N that contain at least one singleton and where $|\bigcup \mathcal{L}| \geq 2$.*

As concerns the case of a prescribed clique size limit k, we conjecture the following.

Conjecture 3 *For any $2 \leq k \leq n$, a polyhedral description of D_N^k is given by the lower bounds $0 \leq \mathsf{c}(K)$ for $K \subseteq N$, $|K| = k$ and the inequalities* (1) *induced by clutters $\mathcal{L}$ which are subsets of $\{L \subseteq N \,:\, |L| < k\}$, contain at least one singleton and satisfy $|\bigcup \mathcal{L}| \geq 2$.*

Note that not every inequality from Conjecture 3 is facet-defining for D_N^k; the problem of a precise characterization of facets of D_N^k is more subtle.

4.1 Clutter inequalities in the case of 4 variables

To illustrate Conjecture 2 let us list the 9 types of the 50 facet-defining inequalities for D_N in the case $n = |N| = 4$ and interpret them in terms of clutters. An exceptional case, which is not a clutter inequality, is the lower bound:

lower bound: $0 \leq \mathsf{c}(abcd)$ (1 inequality).

Note that we have abbreviated $\{a,b,c,d\}$ to $abcd$; we will adopt this abbreviation from now on. Two types of *monotonicity inequalities* correspond to quite simple clutters, namely to one singleton together with one non-singleton:

3-to-4 monotonicity: take $\mathcal{L} = \{abc, d\}$, (2) gives $\quad 1 \leq \mathsf{c}(abc) + \mathsf{c}(d) - \mathsf{c}(abcd)$
and, because of $\mathsf{c}(d) = 1$, one gets $\quad \mathsf{c}(abcd) \leq \mathsf{c}(abc) \quad$ (4 inequalities),

2-to-3 monotonicity: take $\mathcal{L} = \{ab, c\}$, (2) gives $\quad 1 \leq \mathsf{c}(ab) + \mathsf{c}(c) - \mathsf{c}(abc)$
and, because of $\mathsf{c}(c) = 1$, one gets $\quad \mathsf{c}(abc) \leq \mathsf{c}(ab) \quad$ (12 inequalities).

The *cluster inequalities*, whose special cases are the upper bounds, correspond to clutters consisting of singletons only:

upper bounds: take $\mathcal{L} = \{a, b\}$, (2) gives $\quad 1 \leq \mathsf{c}(a) + \mathsf{c}(b) - \mathsf{c}(ab)$
and, since $\mathsf{c}(a) = \mathsf{c}(b) = 1$, one gets $\quad \mathsf{c}(ab) \leq 1 \quad$ (6 inequalities),

cluster for 3-element-sets: take $\mathcal{L} = \{a, b, c\}$, (2) gives

$$1 \leq \mathsf{c}(a) + \mathsf{c}(b) + \mathsf{c}(c) - \mathsf{c}(ab) - \mathsf{c}(ac) - \mathsf{c}(bc) + \mathsf{c}(abc) \text{ and one gets}$$
$$\mathsf{c}(ab) + \mathsf{c}(ac) + \mathsf{c}(bc) \quad \leq \quad 2 + \mathsf{c}(abc) \qquad \text{(4 inequalities)},$$

cluster for 4-element-set: take $\mathcal{L} = \{a, b, c, d\}$ and (2) leads to

$$\mathsf{c}(ab) + \mathsf{c}(ac) + \mathsf{c}(ad) + \mathsf{c}(bc) + \mathsf{c}(bd) + \mathsf{c}(cd) + \mathsf{c}(abcd)$$
$$\leq \quad 3 + \mathsf{c}(abc) + \mathsf{c}(abd) + \mathsf{c}(acd) + \mathsf{c}(bcd) \qquad \text{(1 inequality)}.$$

Besides 28 "basic" inequalities, which already occurred in the case $n = 3$, there are additionally 22 *non-basic inequalities* decomposing into 3 types; we gave them some auxiliary labels:

one 2-element-set clutter: take $\mathcal{L} = \{ab, c, d\}$ and (2) leads to

$$\mathsf{c}(cd) + \mathsf{c}(abc) + \mathsf{c}(abd) \leq 1 + \mathsf{c}(ab) + \mathsf{c}(abcd) \qquad \text{(6 inequalities)},$$

two 2-element-sets clutter: take $\mathcal{L} = \{ac, bc, d\}$ and (2) leads to

$$\mathsf{c}(abc) + \mathsf{c}(acd) + \mathsf{c}(bcd) \leq \mathsf{c}(ac) + \mathsf{c}(bc) + \mathsf{c}(abcd) \qquad \text{(12 inequalities)},$$

three 2-element-sets clutter: take $\mathcal{L} = \{ac, ac, bc, d\}$ and (2) leads to

$$2 \cdot \mathsf{c}(abc) + \mathsf{c}(abd) + \mathsf{c}(acd) + \mathsf{c}(bcd) \leq \mathsf{c}(ab) + \mathsf{c}(ac) + \mathsf{c}(bc) + 2 \cdot \mathsf{c}(abcd) \quad \text{(4 inequalities)}.$$

4.2 Validity of clutter inequalities

To show the validity of the inequalities (1) for D_N we re-write them in a suitable way. Given $\mathsf{c} \in \mathbb{R}^{\mathcal{P}(N)}$, its (superset) *Möbius inversion* $\mathsf{m} \in \mathbb{R}^{\mathcal{P}(N)}$ is given by the formula

$$\mathsf{m}(T) \;:=\; \sum_{S:\, T \subseteq S} (-1)^{|S \setminus T|} \cdot \mathsf{c}(S) \qquad \text{for } T \subseteq N. \tag{3}$$

Given a chordal graph G over N, the Möbius inversion of its (extended) characteristic imset c_G will be denoted by m_G. There is a one-to-one correspondence between c and its Möbius inversion m; the backward formula to (3) is

$$\mathsf{c}(S) \;=\; \sum_{T:\, S\subseteq T} \mathsf{m}(T) \qquad \text{for } S \subseteq N, \tag{4}$$

which can be verified easily by substituting (3) into (4) and conversely.

The following observation, shown in Appendix A, gives an interpretable formula for the value on the right-hand side of (1). We are going to use special notation for the zero-one indicator of a predicate/statement $\star\star$:

$$\delta(\star\star) \;:=\; \begin{cases} 1 & \text{if the statement } \star\star \text{ holds,} \\ 0 & \text{if } \star\star \text{ does not hold.} \end{cases}$$

Lemma 4 Given a chordal graph G over N, let $\mathcal{C}(G)$ denote the collection of cliques of G and $C_1, \ldots, C_m$, $m \geq 1$ is an arbitrary ordering of elements of $\mathcal{C}(G)$ satisfying RIP. Given a clutter $\mathcal{L}$ of subsets of N with $\emptyset \neq \bigcup \mathcal{L}$ one has

$$\sum_{S\subseteq N} \kappa_{\mathcal{L}}(S) \cdot \mathsf{c}_G(S) = \sum_{j=1}^{m} \delta(C_j \in \mathcal{L}^{\uparrow}) \;- \sum_{j=2}^{m} \delta(S_j \in \mathcal{L}^{\uparrow})\,. \tag{5}$$

The proof of Lemma 4 can be found in the appendix. Now, the proof of the validity of (1) is easy.

Corollary 5 Given a chordal graph G over N, $|N| \geq 2$, all inequalities from Conjecture 2 are valid for the characteristic imset c_G.

Proof The validity of the lower bound $0 \leq \mathsf{c}_G(N)$ is immediate. As concerns (1), re-written in the form (2), given a clutter $\mathcal{L}$ of subsets of N containing a singleton $\{\gamma\}$, choose a clique $C \in \mathcal{C}(G)$ containing γ and an ordering $C_1, \ldots, C_m$, $m \geq 1$ of cliques of G satisfying RIP and $C_1 = C$. Such an ordering exists by (Lauritzen, 1996, Lemma 2.18). By Lemma 4, using the formula (5), one has

$$\sum_{S\subseteq N} \kappa_{\mathcal{L}}(S) \cdot \mathsf{c}_G(S) \;=\; \underbrace{\delta(C_1 \in \mathcal{L}^{\uparrow})}_{=1} + \sum_{j=2}^{m} \underbrace{\{\, \delta(C_j \in \mathcal{L}^{\uparrow}) - \delta(S_j \in \mathcal{L}^{\uparrow}) \,\}}_{\geq 0} \geq 1\,,$$

because $\{\gamma\} \in \mathcal{L}$ implies $C_1 \in \mathcal{L}^{\uparrow}$ and, moreover, $S_j \in \mathcal{L}^{\uparrow}$, $S_j \subseteq C_j \;\Rightarrow\; C_j \in \mathcal{L}^{\uparrow}$. ■

5. The Separation Problem in the Cutting Plane Method

The effort to find a complete polyhedral description of D_N^k for $2 \leq k \leq n$ is motivated by the aim to apply an (integer) *linear programming* (LP) approach to learning decomposable models. More specifically, as explained in § 2, the statistical learning task can, in principle, be transformed into an LP problem to maximize a linear function over the chordal graph polytope.

However, since every clutter inequality is facet-defining for D_N (see § 6), the number of inequalities describing D_N is super-exponential in $n = |N|$ and the use of a pure LP approach is not

realistic. Instead, *integer linear programming* (ILP) methods can be applied, specifically the use of *cutting planes*. In this approach, the initial task is to solve an LP problem which is a relaxation of the original problem: namely to maximize the objective over a polyhedron P with $D_N \subseteq P$, where P is specified by a modest number of inequalities. Typically, P is given by some sub-collection of valid inequalities for D_N and there is a requirement that integer vectors in P and D_N coincide: $\mathbb{Z}^\Lambda \cap P = \mathbb{Z}^\Lambda \cap D_N$. Moreover, facet-defining inequalities for D_N appear to be the most useful ones, leading to good overall performace.

In this approach, if the optimal solution c^* to the relaxed problem has only integer components then it is also the optimal solution to the unrelaxed problem. Otherwise, one has to solve the *separation problem* (Wolsey, 1998), which is to find a linear constraint (a *cutting plane*) which separates c^* from D_N. This new constraint is added and the method repeats starting from this new more tightly constrained problem.

If our search is limited to the *clutter inequalities* then it leads to the next task:

> Given $\mathsf{c}^* \notin D_N$ find clutter(s) $\mathcal{L}$ such that the inequality (1) is (most) violated by c^*, in other words, we minimize $\mathcal{L} \mapsto \sum_{S \subseteq N} \kappa_\mathcal{L}(S) \cdot \mathsf{c}^*(S)$ over $\mathcal{L}$.

Our idea is to re-formulate this in the form of a few auxiliary LP problems. To this end we fix a distinguished element $\gamma \in N$ and limit our search to clutters $\mathcal{L}$ with $\{\gamma\} \in \mathcal{L}$ and $(\bigcup \mathcal{L}) \setminus \{\gamma\} \neq \emptyset$. Thus, we decompose the whole problem into $n = |N|$ subproblems. To solve the subproblem we realize that, with $M := N \setminus \{\gamma\}$ and $\mathcal{R} := \mathcal{L} \setminus \{\{\gamma\}\}$, one has $\kappa_\mathcal{L}(S) = -\kappa_\mathcal{L}(S \cup \{\gamma\}) = \kappa_\mathcal{R}(S)$ for any $S \subseteq M$. Because $\kappa_\mathcal{L}(\{\gamma\}) = 1$ while $\kappa_\mathcal{L}(\emptyset) = \kappa_\mathcal{R}(\emptyset) = 0$, one can re-formulate the subproblem as an LP problem to minimize the above objective given by c^* over the polytope

$$\mathrm{conv}\,(\,\{\,[\kappa_\mathcal{R}(S)]_{\emptyset \neq S \subseteq M}\ :\ \mathcal{R} \text{ is a clutter of subsets of } M \text{ with } \textstyle\bigcup \mathcal{R} \neq \emptyset\,\}\,)\,. \tag{6}$$

This is indeed possible owing to the following observation whose proof we skip due to lack of space.

Lemma 6 The polytope (6) has the following polyhedral description: (1) $1 = \sum_{\emptyset \neq S \subseteq M} \kappa(S)$, (2) $0 \leq \kappa(\{i\})$ for $i \in M$ and (3) $0 \leq \sum_{L \subseteq K} \kappa(L \cup \{i\})$ for any pair (i, K) where $i \in M$, $\emptyset \neq K \subseteq M \setminus \{i\}$.

We have implemented some methods for solving this separation problem by extending the GOBNILP system (Cussens, 2011) for learning Bayesian networks. This was done by adding a *constraint handler* for chordal graph learning to the development version of GOBNILP which can be found at `https://bitbucket.org/jamescussens/gobnilp`.

GOBNILP already looks for the deepest-cutting cutting planes which are cluster inequalities, i.e. clutter inequalities where all clutter members are singletons. Extending this to find the guaranteed best clutter cut for all possible clutters, for example by exploiting Lemma 6, has proved (so far) to be too slow. Instead preliminary results indicate that an approximate approach is superior: monotonicity inequalities ($|\mathcal{L}| = 2$) are added initially and then the separation problem is solved approximately by searching only for clutters where $|\mathcal{L}| \in \{3, 4\}$. With this approach, GOBNILP can find the optimal chordal graph for the BRIDGES UCU dataset (12 variables, 108 datapoints) in 230s. In contrast, as shown by Kangas et al. (2014), the current stable version of GOBNILP, which learns chordal graphs by simply ruling out immoralities, cannot solve this problem even when given several hours. This is a clear improvement, however, when there is no limit on clique size, performance remains far behind that of the `Junctor` algorithm (Kangas et al., 2014) which, for example, can solve the BRIDGES learning problem in only a few seconds.

Interestingly, with the separation algorithm turned off and no monotonicity inequalities added (development-version) GOBNILP could still not solve this problem after 59,820s (at which point we aborted since GOBNILP was using 12Gb of memory!). This shows the practical importance of using the (facet-defining) clutter inequalities in an ILP approach to chordal graph learning.

Our conclusion from the preliminary empirical experiments is that the present poor performance is mainly caused by the large number of ILP variables one has to create. This is because one cannot apply the normal pruning for Bayesian network learning, as has already been noted by Kangas et al. (2014, § 4). Given our present state of knowledge, only when one restricts the maximal clique size (= treewidth) is there hope for reasonable performance. Thus, more extensive experimentation should be delayed until further progress in pruning methods is achieved.

6. Conclusion: Further Theoretical Results and Open Tasks

We achieved several theoretical results on the clutter inequalities (whose proofs are not included due to lack of space). In particular, we are able to show that every inequality from Conjecture 2 is *facet-defining* for the chordal graph polytope D_N.

There are further supporting arguments for the conjectures from § 4. Specifically, we have established—from a classic matroid theory result by Edmonds (1970) as reported in (Schrijver, 2003, chapter 40)—that a complete polyhedral description for D_N^2 consists of the lower bounds and the cluster inequalities. (Proof omitted due to lack of space.) Thus, Conjecture 3 is true in case $k = 2$.

We also have a promising ILP formulation for chordal graph learning using a subset of the facet-defining inequalities of D_N as constraints. Specifically, we have proved (proof omitted) that the vertices of D_N coincide with the integer vectors from the polyhedron specified by the *lower bound*, the *monotonicity inequalities*, which are the inequalities induced by clutters of the form $\mathcal{L} = \{ \{\gamma\}, L \}$, the *cluster inequalities*, whose clutters consist of singletons, that is, they have the form $\mathcal{L} = \{ \{c\} : c \in C\}$ where $C \subseteq N$, $|C| \geq 2$, and

- the inequalities induced by clutters of the form $\mathcal{L} = \{\{\gamma\}\} \cup \{L : L \subseteq R, |L| = |R| - 1\}$, where $R \subseteq N \setminus \{\gamma\}$ and $|R| \geq 3$.

Note that an example of an inequality of the last type was given in § 4.1, under the label "three 2-element-sets clutter". Considering solely lower bounds, monotonicity and cluster inequalities does not lead to a polytope whose integer vectors coincide with the vertices of D_N.

The big theoretical challenge remains: to confirm/disprove Conjecture 2. Even if confirmed, a further open problem is to characterize facet-defining inequalities for D_N^k, $2 \leq k \leq n$, within the clutter ones. The preliminary empirical experiments indicate that a further theoretical goal should be to develop special pruning methods under the assumption that the optimal chordal graph is the learning goal. The subsequent goal, based on the result of pruning, can be to modify the proposed LP methods for solving the separation problem to become more efficient.

Acknowledgments

The research of Milan Studený has been supported by the grant GAČR number 16-12010S.

Appendix A. Proof of Lemma 4

In the proof we use some deeper facts from the theory of chordal graphs, as presented in (Lauritzen, 1996, § 2.1.2 - 2.1.3) or in (Studený, 2005, § 7.2). We are going to prove the following result:

Proposition Given a chordal graph G over N, let $\mathcal{C}(G)$ denote the collection of cliques of G, $\mathcal{S}(G)$ the collection of separators in G and $w_G(S)$ the multiplicity of a separator $S \in \mathcal{S}(G)$. Then, for any $T \subseteq N$,

$$\mathsf{m}_G(T) = \sum_{C \in \mathcal{C}(G)} \delta(T = C) \, - \sum_{S \in \mathcal{S}(G)} w_G(S) \, \cdot \, \delta(T = S) \, = \sum_{j=1}^{m} \delta(T = C_j) \, - \sum_{j=2}^{m} \delta(T = S_j) \, , \qquad (7)$$

where $C_1, \ldots, C_m$ is an arbitrary ordering of elements of $\mathcal{C}(G)$ satisfying RIP. In particular, if $\mathcal{L}$ is a clutter of subsets of N with $\emptyset \neq \bigcup \mathcal{L}$ then

$$\begin{aligned} \sum_{S \subseteq N} \kappa_{\mathcal{L}}(S) \cdot \mathsf{c}_G(S) \quad &= \quad \sum_{C \in \mathcal{C}(G)} \delta(C \in \mathcal{L}^{\uparrow}) - \sum_{S \in \mathcal{S}(G)} w_G(S) \cdot \delta(S \in \mathcal{L}^{\uparrow}) \qquad (8) \\ &= \quad \sum_{j=1}^{m} \delta(C_j \in \mathcal{L}^{\uparrow}) \, - \sum_{j=2}^{m} \delta(S_j \in \mathcal{L}^{\uparrow}) \, , \quad \text{which is the form given in (5).} \end{aligned}$$

Proof Let us put

$$\mathsf{m}'(T) := \sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{C}(G)} (-1)^{|\mathcal{B}|+1} \cdot \delta(T = \bigcap \mathcal{B}) \quad \text{for any } T \subseteq N;$$

the aim to show $\mathsf{m}' = \mathsf{m}_G$. Thus, we denote

$$\mathcal{C}(G, S) := \{C \in \mathcal{C}(G) \, : \, S \subseteq C\} \quad \text{for any fixed } S \subseteq N,$$

and write

$$\begin{aligned} \sum_{T \colon S \subseteq T} \mathsf{m}'(T) &= \sum_{T \colon S \subseteq T} \sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{C}(G)} (-1)^{|\mathcal{B}|+1} \cdot \delta(T = \bigcap \mathcal{B}) \\ &= \sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{C}(G)} (-1)^{|\mathcal{B}|+1} \cdot \sum_{T \colon S \subseteq T} \delta(T = \bigcap \mathcal{B}) = \sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{C}(G)} (-1)^{|\mathcal{B}|+1} \cdot \delta(S \subseteq \bigcap \mathcal{B}) \\ &= \sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{C}(G,S)} (-1)^{|\mathcal{B}|+1} = 1 + \sum_{\mathcal{B} \subseteq \mathcal{C}(G,S)} (-1)^{|\mathcal{B}|+1} \, = \, \delta(\, \mathcal{C}(G,S) \neq \emptyset \,) = \mathsf{c}_G(S) \, . \end{aligned}$$

Thus, c_G is obtained from m' by the backward formula (4). Hence, since the Möbius inversion is one-to-one transformation, one has

$$\mathsf{m}_G(T) \, = \sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{C}(G)} (-1)^{|\mathcal{B}|+1} \cdot \delta(T = \bigcap \mathcal{B}) \qquad \text{for any } T \subseteq N. \qquad (9)$$

The formula (9) can be re-written: given (any) ordering $C_1, \ldots, C_m$, $m \geq 1$ of (all) cliques of G satisfying RIP and the respective separators $S_i = C_i \cap (\bigcup_{\ell < i} C_\ell)$, $i = 2, \ldots m$, one has

$$\mathsf{m}_G(T) \, = \, \delta(T = C_1) + \sum_{j=2}^{m} \, \{ \, \delta(T = C_j) - \delta(T = S_j) \, \} \qquad \text{for } \, T \subseteq N. \qquad (10)$$

Indeed, (10) can be derived from (9) by induction on m: if $C = C_m$, $m \geq 2$ then a preceding clique $K = C_j$, $j < m$ exists with $S_m = C \cap K$ and one has

$$\sum_{\mathcal{B} \subseteq \mathcal{C}(G) \colon C \in \mathcal{B}} (-1)^{|\mathcal{B}|+1} \cdot \delta(T = \bigcap \mathcal{B}) \, = \, \delta(T = C) - \delta(T = C \cap K) \, ,$$

because the other terms cancel each other. Since the order of cliques is irrelevant in (9), the expression in (10) does not depend on the choice of an ordering satisfying RIP. In particular, (10) can be written in the form (7), where $w_G(S)$ is the number of $2 \leq j \leq m$ with $S = S_j$ for $S \in \mathcal{S}(G)$, which is the multiplicity of the separator S. Given a clutter $\mathcal{L}$ with $\emptyset \neq \bigcup \mathcal{L}$ and $T \subseteq N$ we write

$$\sum_{S \subseteq T} \kappa_{\mathcal{L}}(S) \overset{(2)}{=} \sum_{S \subseteq T} \sum_{A \subseteq S\colon A \in \mathcal{L}^{\uparrow}} (-1)^{|S \setminus A|} = \sum_{A \subseteq T\colon A \in \mathcal{L}^{\uparrow}} \sum_{S\colon A \subseteq S \subseteq T} (-1)^{|S \setminus A|}$$
$$= \sum_{A \subseteq T\colon A \in \mathcal{L}^{\uparrow}} \sum_{B \subseteq T \setminus A} (-1)^{|B|} = \sum_{A \subseteq T\colon A \in \mathcal{L}^{\uparrow}} \delta(\, T \setminus A = \emptyset\,) \;=\; \delta(T \in \mathcal{L}^{\uparrow})\,. \tag{11}$$

Therefore, one can observe

$$\sum_{S \subseteq N} \kappa_{\mathcal{L}}(S) \cdot \mathsf{c}_G(S) \overset{(4)}{=} \sum_{S \subseteq N} \kappa_{\mathcal{L}}(S) \cdot \sum_{T\colon S \subseteq T} \mathsf{m}_G(T) = \sum_{T \subseteq N} \mathsf{m}_G(T) \cdot \sum_{S \subseteq T} \kappa_{\mathcal{L}}(S)$$
$$\overset{(11)}{=} \sum_{T \subseteq N} \mathsf{m}_G(T) \cdot \delta(T \in \mathcal{L}^{\uparrow}) \overset{(7)}{=} \sum_{C \in \mathcal{C}(G)} \delta(C \in \mathcal{L}^{\uparrow}) - \sum_{S \in \mathcal{S}(G)} w_G(S) \cdot \delta(S \in \mathcal{L}^{\uparrow})\,,$$

which concludes the proof of (8), including its detailed form (5). ■

References

M. Bartlett and J. Cussens. Advances in Bayesian network learning using integer programming. In A. Nicholson and P. Smyth, editors, *Uncertainty in Artificial Intelligence 29*, pages 182–191. AUAI Press, 2013.

R. R. Bouckaert. *Bayesian belief networks: from construction to evidence*. PhD thesis, University of Utrecht, 1995.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

J. Corander, T. Janhunen, J. Rintanen, H. Nyman, and J. Pensar. Learning chordal Markov networks by constraint satisfaction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weiberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1349–1357. Curran Associates, 2013.

R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, 1999.

J. Cussens. Bayesian network learning with cutting planes. In F. Cozman and A. Pfeffer, editors, *Uncertainty in Artificial Intelligence 27*, pages 153–160. AUAI Press, 2011.

J. Edmonds. Submodular functions, matroids, and certain polyhedra. In R. Guy, H. Hanani, N. Sauer, and J. Schönheim, editors, *Combinatorial Structures and Their Applications*, pages 69–87. Gordon and Breach, 1970.

K. Fukuda. cdd and ccdplus homepage, May 2015. `https://www.inf.ethz.ch/personal/fukudak/cdd_home/`.

P. Giudici and P. J. Green. Decomposable graphical Gaussian model determination. *Biometrika*, 86: 785–801, 1999.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 20:194–243, 1995.

R. Hemmecke, S. Lindner, and M. Studený. Characteristic imsets for learning Bayesian network structure. *International Journal of Approximate Reasoning*, 53:1336–1349, 2012.

T. Jaakkola, D. Sontag, A. Globerson, and M. Meila. Learning Bayesian network structure using LP relaxations. In Y. W. Teh and M. Titterington, editors, *JMLR Workshop and Conference Proceedings 9: AISTATS 2010*, pages 358–365, 2010.

K. Kangas, T. Niinimäki, and M. Koivisto. Learning chordal Markov networks by dynamic programming. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2357–2365. Curran Associates, 2014.

S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

S. Lindner. *Discrete optimization in machine learning: learning Bayesian network structures and conditional independence implication*. PhD thesis, TU Munich, 2012.

R. E. Neapolitan. *Learning Bayesian Networks*. Pearson Prentice Hall, Upper Saddle River, 2004.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman, San Mateo, 1988.

A. Pérez, C. Blum, and J. A. Lozano. Learning maximum weighted $(k+1)$-order decomposable graphs by integer linear programming. In L. C. van der Gaag and A. J. Feelders, editors, *Lecture Notes in AI 8754: PGM 2014*, pages 396–408, 2014.

A. Schrijver. *Combinatorial Optimization - Polyhedra and Eficience, volume B*. Springer, Berlin, 2003.

G. E. Schwarz. Estimation of the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

K. S. Sesh Kumar and F. Bach. Convex relaxations for learning bounded-treewidth decomposable graphs. In S. Dasgupta and D. McAlester, editors, *JMLR Workshop and Conference Proceedings 28: ICML 2013*, volume 1, pages 525–533, 2013.

P. Spirtes, C. Glymour, and R. Scheines. *Causality, Prediction and Search*. Springer, New York, 1993.

M. Studený. *Probabilistic Conditional Independence Structures*. Springer, London, 2005.

M. Studený and D. Haws. Learning Bayesian network structure: towards the essential graph by integer linear programming tools. *International Journal of Approximate Reasoning*, 55:1043–1071, 2014.

L. A. Wolsey. *Integer Programming*. John Wiley, New York, 1998.