Article:

Cussens, James orcid.org/0000-0002-1363-2336, Haws, David and Studený, Milan (2017) Polyhedral aspects of score equivalence in Bayesian network structure learning. Mathematical Programming. pp. 285-324. ISSN 0025-5610

https://doi.org/10.1007/s10107-016-1087-2

Elsevier Editorial System(tm) for
International Journal of Approximate Reasoning
Manuscript Draft

Manuscript Number:

Title: Towards using the chordal graph polytope in learning decomposable models

Article Type: VSI: PGM '16

Section/Category:

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

Corresponding Author: Dr. Milan Studený, DrSc.

Corresponding Author's Institution: Institute of Information Theory and Automation of the ASCR

First Author: Milan Studený, DrSc.

Order of Authors: Milan Studený, DrSc.; James Cussens, Prof.

Manuscript Region of Origin:

Abstract: The motivation for this paper is the integer linear programming approach to learning the structure of a decomposable graphical model. We have chosen to represent decomposable models by means of 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. In this theoretical paper, we introduce a class of clutter inequalities (valid for the vectors in the polytope) and show that all of them are facet-defining for the polytope. We dare to conjecture that they lead to a complete polyhedral description of the polytope. Finally, we propose a linear programming method to solve the separation problem with these inequalities for the use in a cutting plane approach.

*Cover Letter

Cover letter for 2016 IJAR submission

Towards using the chordal graph polytope in learning decomposable models

by M. Studený and J. Cussens

Prague, November 22, 2016.

Dear editors,

we were pleased by an offer to submit an extended version of our PGM 2016 contribution

> *The chordal graph polytope for learning decomposable models*,
> JMLR Workshop and Conference Proceedings 52: PGM 2016,
> (A. Antonucci, G. Corani, and C.P. de Campos eds.), pp. 499–510,

to a special issue of the IJAR journal following PGM'16. In comparison with the original proceedings paper, this extended journal version contains the proof of the main result that all clutter inequalities are facet-defining for the polytope. This technical proof makes the paper relatively long, but we hope it still fits in usual page limits. Note that to make this theoretical paper reader-friendly we moved the substantial proofs to the appendix.

We hope that the paper will be found suitable for the special issue.

All the best

Milan Studený and James Cussens

Towards using the chordal graph polytope in learning decomposable models

Milan Studený[a,*], James Cussens[b]

[a] *Institute of Information Theory and Automation of the CAS Prague, Pod Vodárenskou věží 4, 18208, Czech Republic*
[b] *Dept of Computer Science & York Centre for Complex Systems Analysis University of York, Deramore Lane, York, YO10 5GH, United Kingdom*

Abstract

The motivation for this paper is the *integer linear programming* approach to learning the structure of a *decomposable graphical model*. We have chosen to represent decomposable models by means of 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*. In this theoretical paper, we introduce a class of *clutter inequalities* (valid for the vectors in the polytope) and show that all of them are facet-defining for the polytope. We dare to conjecture that they lead to a complete polyhedral description of the polytope. Finally, we propose a linear programming method to solve the *separation problem* with these inequalities for the use in a cutting plane approach.

Keywords: learning decomposable models, integer linear programming, characteristic imset, chordal graph polytope, clutter inequalities, separation problem

1. Introduction: explaining the motivation

Decomposable models are fundamental probabilistic graphical models [16]. A well-known fact is that elegant mathematical properties of these structural

*Corresponding author
Email addresses: `studeny@utia.cas.cz` (Milan Studený), `james.cussens@york.ac.uk` (James Cussens)

models form the theoretical basis of the famous method of local computation [6]. Decomposable models, which are described by *chordal undirected graphs*, can be viewed as special cases of Bayesian network models [19], which are described by directed acyclic graphs.

Two traditionally separate disciplines in probabilistic graphical models are learning and inference. *Structure learning* is determining the graphical model, represented by a graph, on the basis of observed statistical data. *Inference* in Bayesian network models has two phases. The first one is transformation of the (learned) directed acyclic graph into a *junction tree*, which can be viewed as a representative of a decomposable model. The second phase in inference is proper local computation (of conditional probabilities) in a junction tree. The motivation for the present paper is the idea to merge structural learning with the junction tree construction in one step, which basically means direct learning the *structure of a decomposable model* on basis of data.

There are various methods for learning decomposable model structure, most of them being specializations of the methods for learning Bayesian network structure [18]. There are methods based on statistical conditional independence tests like the PC algorithm [23] or MCMC simulations [11]. Nevertheless, this paper deals with a *score-based approach*, where the task is to maximize some additively decomposable score, like the BIC score [21] or the BDeu score [12]. There are some arguments in favour of this approach in comparison with the methods based on statistical tests [29].

More specifically, 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 in [14, 7, 1]. However, the approach discussed in this paper is based on encoding the models by *characteristic imsets*, which were applied in [13, 26]. This mode of representation leads to an elegant way of encoding decomposable models which we believe is particularly suitable for structural learning of these models.

Note that two recent conference papers have also been devoted to ILP-based learning decomposable models, but they used different binary encodings of the models/graphs. More specifically, Sesh Kumar and Bach [22] used special codes for junction trees of the graphs, while Pérez *et al.* [20] encoded certain special coarsenings of maximal hyper-trees. Moreover, restricted learning was the goal in both these papers unlike in this theoretical paper, which we hope to be the first step towards a general ILP method for learning decomposable models.

Two other recent papers devoted to structural learning decomposable models also used encodings of junction trees. Corander *et al.* [5] 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.* [15], 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 (superset) Möbius inversion of the characteristic imset we mention in § 6.1.

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 nodes N), with the possible modification that a clique size limit is given. This polytope has already been dealt with by Lindner [17] 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 we name this polytope the "chordal graph characteristic imset polytope", but abbreviate this to *chordal graph polytope*.

In this paper, which is an extended version of a proceedings paper [25], we assume that the reader is familiar with basic concepts of polyhedral geometry, as presented in numerous textbooks on this topic; for example in [2, 28]. We present a complete facet description of the polytope where $|N| \leq 4$ and mention 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 (= node) set N containing at least one singleton, so we call these *clutter inequalities*.

This establishes a sensible *conjecture* about a complete polyhedral description of the polytope (with no clique size limit). We prove that every clutter inequality is both valid and facet-defining for the polytope. We also tackle an important *separation problem*: that is, given a non-integer solution to a linear programming (LP) relaxation problem, find a clutter inequality which (most) violates the current solution.

2. Basic concepts

Let N be a finite set of *variables*; assume that $n := |N| \geq 2$ 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. Some conventional notation and terminology

The symbol $\subseteq$ will be used to denote non-strict set inclusion of unlike $\subset$, which will serve to denote strict inclusion: $S \subset T$ means $S \subseteq T$ and $S \neq T$. The *power set* of N will be denoted by $\mathcal{P}(N) := \{S \,:\, S \subseteq N\}$.

We are going to use the term *clutter* to name any class $\mathcal{L}$ of subsets of N that are inclusion incomparable, that is, $L, R \in \mathcal{L}$ and $L \subseteq R$ implies $L = R$. Such classes are alternatively named Sperner families or antichains in the mathematical literature. Occasionally, we will abbreviate notation for the union of sets in a clutter: $\bigcup \mathcal{L} := \bigcup_{L \in \mathcal{L}} L$. Given a clutter $\mathcal{L}$ of subsets of N such that $\emptyset \neq \bigcup \mathcal{L}$, we introduce notation $\mathcal{L}^{\uparrow}$ for the *filter* generated by $\mathcal{L}$, by which is meant a class of subsets of N closed under supersets:

$$\mathcal{L}^{\uparrow} := \{\, T \subseteq N \;:\; \exists\, L \in \mathcal{L} \quad L \subseteq T \,\}\,.$$

Moreover, we are going to use special notation for the zero-one indicator of a predicate/statement $\star\star$:

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

The abbreviation LHS will mean "left-hand side" (of an inequality), while RHS will be a shorthand for "right-hand side". The symbol

$$\Upsilon := \{\, S \subseteq N \,:\, |S| \geq 2 \,\}$$

will be our notation for the class of non-empty non-singletons, used as a standard index set for components of our vectors.

2.2. Chordal undirected graphs

We say that a graph G is *over* N if G has N as the set of nodes and it is *undirected* if every its edge is undirected. An undirected graph is *chordal* if every cycle of the 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 [24, Lemma 7.2]. Each chordal graph defines the respective statistical *decomposable model*; see [16, § 4.4].

2.3. Learning graphical models

The *score-based* approach to structural learning of graphical models is based on maximizing some *scoring criterion*, briefly called a *score*, 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, that is, graphical models described by *directed acyclic graphs* H, a crucial technical assumption [4] is that $\mathcal{Q}$ should be *additively decomposable*, which means, it has the form

$$\mathcal{Q}(H,D) \;=\; \sum_{a\in N} \; q_D(a\,|\,\mathrm{pa}_H(a))$$

where the summands $q_D(*\,|\,*)$ are called *local scores*, and

$\mathrm{pa}_H(a) := \{b \in N \;:\; b \to a\,\}$ is the set of *parents* of a node a in H.

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* [3], which means that Markov equivalent (directed acyclic) graphs yield the same score. However, in this paper, we are going to adopt this approach to learning *chordal* undirected graphical models.

2.4. Characteristic imset

The concept of a *characteristic imset* (for a directed acyclic graph) was introduced in [13]. Each characteristic imset is an element of the real vector space $\mathbb{R}^{\Upsilon}$ where $\Upsilon = \{S \subseteq N : |S| \geq 2\}$ is the class of non-empty non-singletons. 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 the graph. These special zero-one vectors describe uniquely the equivalence classes of directed acyclic graphs.

Nevertheless, in the sub-frame of *decomposable models*, that is, in the frame of graphical models described by chordal undirected graphs, models are in one-to-one correspondence with (chordal undirected) graphs and the next simpler definition can be used; see [13, Corollary 4].

Definition 1 (characteristic imset)**.**
Given a chordal undirected graph G over N, the **characteristic imset** *of G is*

a zero-one vector c_G *with components indexed by subsets* S *of* N *with* $|S| \geq 2$*:*

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

An implicit convention is used that $\mathsf{c}_G(L) := 1$ *for any graph* G *over* N *and any singleton* $L \subseteq N$, $|L| = 1$.

A conventional value $\mathsf{c}_G(\emptyset)$ for the empty set plays no substantial role in the theory because it does not occur in basic versions of our inequalities from § 4. Nonetheless, we accept the convention $\mathsf{c}_G(\emptyset) := 1$ in this paper for it leads to an elegant Möbius inversion formula (see Lemma 3 in § 6.1). In particular, c_G can be viewed as a vector in $\mathbb{R}^{\mathcal{P}(N)}$.

As 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, [26, Lemma 3] implies that additively decomposable and score equivalent criterion $\mathcal{Q}$ has the form

$$\mathcal{Q}(G, D) \;=\; k + \sum_{S \in \Upsilon} r_D^{\mathcal{Q}}(S) \cdot \mathsf{c}_G(S)\,, \quad \text{where } k \text{ is a constant and, for any } S \in \Upsilon,$$

$$r_D^{\mathcal{Q}}(S) = \sum_{K \subseteq R} (-1)^{|R \setminus K|} \cdot q_D(a \,|\, R), \quad \text{with arbitrary } a \in S \text{ and } R = S \setminus \{a\},$$

where $q_D(* \,|\, *)$ are the respective local scores.

2.5. Chordal graph polytope

Now, we introduce the polytope(s) to be studied. To be flexible, we consider the situation where a maximal *clique size limit* k is given, $2 \leq k \leq n = |N|$. Taking $k = n$ gives the general (unrestricted) case while taking $k = 2$ leads to a well-known special case of *undirected forests*.

Definition 2 (chordal graph polytope)**.**
Let us introduce the **chordal graph polytope** *over* N *with clique size limit* k*, where* $2 \leq k \leq n = |N|$*, as follows:*

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

where conv (−) *is used to denote the convex hull.*

The dimension of the polytope D_N^k is $\sum_{\ell=2}^{k}\binom{n}{\ell}$. Thus, for the *unrestricted* polytope $D_N := D_N^n$, one has $\dim(D_N) = \sum_{\ell=2}^{n}\binom{n}{\ell} = 2^n - n - 1$, while one has $\dim(D_N^2) = \binom{n}{2}$ for the polytope for learning *undirected forests*.

In fact, one can decide to be even more general and consider the polytope

$$D_N^{\mathcal{K}} \;:=\; \mathrm{conv}\,(\,\{\mathsf{c}_G \,:\, G \text{ chordal graph over } N \text{ with complete sets in } \mathcal{K}\}\,)$$

for a class $\mathcal{K} \subseteq \mathcal{P}(N)$ of subsets of N closed under subsets and containing (all) singletons in N. Our long-termed strategic goal is to get the facet description of $D_N^{\mathcal{K}}$ for any such class $\mathcal{K}$ and utilize such result in learning decomposable models by ILP methods. The idea is that $\mathcal{K}$ will be obtained as a result of a *pruning procedure* to be developed, which ensures that every optimal chordal graph has complete sets in $\mathcal{K}$. The point is that $\dim(D_N^{\mathcal{K}}) = |\mathcal{K}|-n-1$ can be considerably smaller than $\dim(D_N)$. Note that the assumption on $\mathcal{K}$ being closed under subsets is not restrictive because, given a general class $\mathcal{K}$ containing sets $L \subseteq N$, $|L| \leq 1$, one has $D_N^{\mathcal{K}} = D_N^{\mathcal{K}'}$ with $\mathcal{K}' = \{S \subseteq N \,:\, T \in \mathcal{K} \text{ for any } T \subseteq S\}$; this follows from the fact that, for any $\mathsf{c} \in D_N$ and $\emptyset \neq T \subseteq S \subseteq N$, $\mathsf{c}(T) = 0$ implies $\mathsf{c}(S) = 0$.

3. Example: the cases of a low number of variables

For small values of $n = |N|$ we have been able to use the `cdd` program [10] to compute a facet description of the *chordal graph polytope* D_N. In the case $n = 3$, D_N has 8 vertices, encoding 8 chordal graphs, and 8 facet-defining inequalities, decomposing into 4 permutation types. With $N = \{a, b, c\}$, these are:

lower bound: $0 \leq \mathsf{c}(\{a, b, c\})$ (1 inequality),

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

upper bounds: $\mathsf{c}(\{a, b\}) \leq 1$ (3 inequalities),

cluster inequality for a 3-element set:

$$\mathsf{c}(\{a, b\}) + \mathsf{c}(\{a, c\}) + \mathsf{c}(\{b, c\}) \leq 2 + \mathsf{c}(\{a, b, c\}) \quad \text{(1 inequality).}$$

Note that the *cluster inequalities* (formulated in terms of family variables) have earlier occurred in the context of learning Bayesian networks [14, 1, 26]; see Example 3 in § 6.2. The restricted polytope D_N^2 only has 7 vertices, encoding 7 undirected forests, and it is specified by 1 equality constraint and 7 inequalities: these are obtained from the above ones by the substitution $\mathsf{c}(\{a,b,c\}) = 0$. In this subcase, the 2-to-3 monotonicity inequalities turn into the lower bounds.

In the case $n = |N| = 4$, the unrestricted polytope D_N has 61 vertices, encoding 61 chordal graphs. The number of facets is only 50, decomposing into 9 permutation types. The list of these types is given in § 5.1, where we also mention the restricted polytopes D_N^3 and D_N^2 with $|N| = 4$.

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. Thus, there is little hope of computing facets directly in case $n = 6$.

An interesting observation is as follows: in the cases $n = |N| \leq 5$, with the exception of the lower bound $0 \leq \mathsf{c}(N)$, all facet-defining inequalities for D_N can be written in the following *generalized monotonicity form*:

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

where γ is a distinguished element of N and the coefficients $\kappa(S)$ are integers. Indeed, the 2-to-3 monotonicity inequalities for $n = 3$ have this form: here $\gamma = c$, $\kappa(\{a,b\}) = 1$ and $\kappa(S) = 0$ for $S \subset \{a,b\}$. The 3-element cluster inequality for $n = 3$ can be re-written in this form in three alternative ways: the choice $\gamma = c$ gives $\mathsf{c}(\{a,c\}) + \mathsf{c}(\{b,c\}) - \mathsf{c}(\{a,b,c\}) \leq \mathsf{c}(\{a\}) + \mathsf{c}(\{b\}) - \mathsf{c}(\{a,b\})$ because of the convention $\mathsf{c}(\{a\}) = 1 = \mathsf{c}(\{b\})$.

4. Clutter inequalities

A deeper observation is that the discussed inequalities can be interpreted as inequalities induced by certain *clutters* of subsets of N.

Definition 3 (clutter inequality).

Let $\mathcal{L}$ be a clutter of subsets of N satisfying $\emptyset \neq \bigcup \mathcal{L}$. The **clutter inequality** *induced by $\mathcal{L}$ is a linear constraint on $\mathsf{c} \in \mathbb{R}^{\mathcal{P}(N)}$ of the form*

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

In this context, recall the convention $\mathsf{c}(L) = 1$ for any $L \subseteq N$, $|L| = 1$. We have formally introduced the inequality for any non-trivial clutter $\mathcal{L}$, which appears to be convenient. Nonetheless, note that (1) is a valid constraint for any extended $\mathsf{c} \in D_N$ only when $\mathcal{L}$ contains a singleton; see § 7. If $\mathcal{L}$ consists of a sole singleton then (1) follows from conventional equality constraints.

One can write the clutter inequality in various forms. In this section we describe a simple way to compute the coefficients with sets in (1), give its non-redundant form in the proper space $\mathbb{R}^{\Upsilon}$ for the polytope D_N and explain its generalized monotonicity interpretation. Later, in § 6, we re-write the clutter inequality (1) in terms of other vector representatives of chordal graphs.

Lemma 1 (basic re-writings of the clutter inequality).

Let $\mathcal{L}$ be a clutter of subsets of N such that $\emptyset \neq \bigcup \mathcal{L}$. Given $\mathsf{c} \in \mathbb{R}^{\mathcal{P}(N)}$, the value $v(\mathsf{c}, \mathcal{L})$ from (1) can be expressed as follows:

$$1 \leq v(\mathsf{c}, \mathcal{L}) = \sum_{S \subseteq N} \kappa_{\mathcal{L}}(S) \cdot \mathsf{c}(S) \quad \text{where} \quad \kappa_{\mathcal{L}}(S) := \sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{L}:\, \bigcup \mathcal{B} = S} (-1)^{|\mathcal{B}|+1} \tag{2}$$

for any $S \subseteq N$. The coefficients $\kappa_{\mathcal{L}}(-)$ in (2) vanish outside the class

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

Within this class, they can be computed recursively using the formula

$$\kappa_{\mathcal{L}}(S) \quad = \quad 1 - \sum_{T \in \mathcal{U}(\mathcal{L}):\, T \subset S} \kappa_{\mathcal{L}}(T) \qquad \text{for any } S \in \mathcal{U}(\mathcal{L}), \tag{3}$$

which implicitly says that $\kappa_{\mathcal{L}}(L) = 1$ for $L \in \mathcal{L}$. Moreover, the formula (2) gets its unique non-redundant form

$$1 - |\mathcal{L} \setminus \Upsilon| \;\leq\; \sum_{S \in \Upsilon} \kappa_{\mathcal{L}}(S) \cdot \mathsf{c}(S) \tag{4}$$

in the proper linear space $\mathbb{R}^{\Upsilon}$, where the polytope D_N is full-dimensional.

Proof. We re-arrange the terms in (1) after the sets $S = \bigcup \mathcal{B}$ and get

$$v(\mathsf{c}, \mathcal{L}) \;=\; \sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{L}} (-1)^{|\mathcal{B}|+1} \cdot \mathsf{c}(\bigcup \mathcal{B}) \;=\; \sum_{S \subseteq N} \mathsf{c}(S) \cdot \underbrace{\sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{L} :\, \bigcup \mathcal{B} = S} (-1)^{|\mathcal{B}|+1}}_{\kappa_{\mathcal{L}}(S)},$$

which gives (2). It is immediate from this that $\kappa_{\mathcal{L}}(S) = 0$ once $S \notin \mathcal{U}(\mathcal{L})$. Having fixed $S \in \mathcal{U}(\mathcal{L})$ observe that the class $\mathcal{L}_S := \{ L \in \mathcal{L} \,:\, L \subseteq S \}$ is non-empty, which allows us to write:

$$\sum_{T \in \mathcal{U}(\mathcal{L}) :\, T \subseteq S} \kappa_{\mathcal{L}}(T) = \sum_{T \subseteq S} \kappa_{\mathcal{L}}(T) = \sum_{T \subseteq S} \sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{L} :\, \bigcup \mathcal{B} = T} (-1)^{|\mathcal{B}|+1} =$$
$$= \sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{L} :\, \bigcup \mathcal{B} \subseteq S} (-1)^{|\mathcal{B}|+1} = 1 - \sum_{\mathcal{B} \subseteq \mathcal{L} :\, \bigcup \mathcal{B} \subseteq S} (-1)^{|\mathcal{B}|} = 1 - \sum_{\mathcal{B} \subseteq \mathcal{L}_S} (-1)^{|\mathcal{B}|} = 1\,,$$

which gives (3). To transform (2) into (4) note that $\kappa_{\mathcal{L}}(\emptyset) = 0$ and, for $L \subseteq N$, $|L| = 1$, one has $\mathsf{c}(L) = 1$ while $\kappa_{\mathcal{L}}(L)$ is either 1 or 0, depending on whether $L \in \mathcal{L}$ or not. Of course, the number of singletons in $\mathcal{L}$ is just $|\mathcal{L} \setminus \Upsilon|$. $\square$

Let us illustrate Lemma 1 by an example; it hopefully indicates that, for small clutters $\mathcal{L}$, the respective coefficient vector $\kappa_{\mathcal{L}} \in \mathbb{R}^{\Upsilon}$ in the non-redundant inequality (4) is sparse because $|\mathcal{U}(\mathcal{L})|$ is small.

Example 1 (computing coefficients in a clutter inequality)**.** Put $N = \{a, b, c, d\}$ and consider the clutter $\mathcal{L} = \{\ \{a, b\},\ \{a, c\},\ \{b, c\},\ \{d\}\ \}$. Then the fact that $\kappa_{\mathcal{L}}(L) = 1$ for $L \in \mathcal{L}$ gives $\kappa_{\mathcal{L}}(\{a, b\}) = \kappa_{\mathcal{L}}(\{a, c\}) = \kappa_{\mathcal{L}}(\{b, c\}) = \kappa_{\mathcal{L}}(\{d\}) = 1$. The remaining elements in $\mathcal{U}(\mathcal{L})$ are $\{a, b, d\}$, $\{a, c, d\}$, $\{b, c, d\}$, $\{a, b, c\}$ and $\{a, b, c, d\}$. The recursive formula (3) can be applied to $\{a, b, d\}$, whose proper subsets in $\mathcal{U}(\mathcal{L})$ are $\{a, b\}$ and $\{d\}$, which yields

$$\kappa_{\mathcal{L}}(\{a, b, d\}) \;\stackrel{(3)}{=}\; 1 - \kappa_{\mathcal{L}}(\{a, b\}) - \kappa_{\mathcal{L}}(\{d\}) = 1 - 1 - 1 = -1\,.$$

Analogously, $\kappa_{\mathcal{L}}(\{a, c, d\}) = \kappa_{\mathcal{L}}(\{b, c, d\}) = -1$. As concerns $\{a, b, c\}$, it has three proper subsets in $\mathcal{U}(\mathcal{L})$, which leads to

$$\kappa_{\mathcal{L}}(\{a, b, c\}) \;\stackrel{(3)}{=}\; 1 - \kappa_{\mathcal{L}}(\{a, b\}) - \kappa_{\mathcal{L}}(\{a, c\}) - \kappa_{\mathcal{L}}(\{b, c\}) = 1 - 1 - 1 - 1 = -2\,.$$

Finally, $\{a,b,c,d\}$ has all other sets in $\mathcal{U}(\mathcal{L})$ as proper subsets which gives

$$\begin{aligned}\kappa_{\mathcal{L}}(\{a,b,c,d\}) &\overset{(3)}{=} 1-\kappa_{\mathcal{L}}(\{a,b\})-\kappa_{\mathcal{L}}(\{a,c\})-\kappa_{\mathcal{L}}(\{b,c\})-\kappa_{\mathcal{L}}(\{d\})\\ &\quad -\kappa_{\mathcal{L}}(\{a,b,d\})-\kappa_{\mathcal{L}}(\{a,c,d\})-\kappa_{\mathcal{L}}(\{b,c,d\})-\kappa_{\mathcal{L}}(\{a,b,c\})\\ &= 1-1-1-1-1-(-1)-(-1)-(-1)-(-2) \;=\; +2\,.\end{aligned}$$

Because the remaining coefficients $\kappa_{\mathcal{L}}(S)$ vanish and $\mathcal{L}$ contains one singleton, that is, $|\mathcal{L}\setminus\Upsilon|=1$, the non-redundant formula (4) takes the form

$$\begin{aligned}0=1-|\mathcal{L}\setminus\Upsilon| \;\le\; & \mathsf{c}(\{a,b\})+\mathsf{c}(\{a,c\})+\mathsf{c}(\{b,c\})\\ & -\mathsf{c}(\{a,b,d\})-\mathsf{c}(\{a,c,d\})-\mathsf{c}(\{b,c,d\})-2\cdot\mathsf{c}(\{a,b,c\})+2\cdot\mathsf{c}(\{a,b,c,d\})\,.\end{aligned}$$

4.1. Generalized monotonicity interpretation of clutter inequalities

Another interesting observation is that if a clutter contains a singleton then the corresponding inequality can be interpreted as a generalized monotonicity constraint. Indeed, given a clutter $\mathcal{L}\subseteq\mathcal{P}(N)$ containing a singleton $\{\gamma\}$ such that $|\bigcup\mathcal{L}|\ge 2$, let us put $\mathcal{R}:=\mathcal{L}\setminus\{\,\{\gamma\}\,\}$. Then $\bigcup\mathcal{R}\neq\emptyset$ and the formulas (2) and (3) from Lemma 1 allow one to observe that

$$\kappa_{\mathcal{L}}(S)=\begin{cases}\kappa_{\mathcal{R}}(S) & \text{for } S\subseteq N\setminus\{\gamma\},\\ -\kappa_{\mathcal{R}}(S\setminus\{\gamma\}) & \text{for } S\subseteq N \text{ with } \gamma\in S \text{ and } S\setminus\{\gamma\}\neq\emptyset,\\ 1 & \text{for } S=\{\gamma\}.\end{cases} \tag{5}$$

Because of the convention $\mathsf{c}(\{\gamma\})=1$ and the fact $\kappa_{\mathcal{R}}(\emptyset)=0$ the formula (2) can be re-arranged into the following *generalized monotonicity* form:

$$\sum_{S\subseteq N\setminus\{\gamma\}}\kappa_{\mathcal{R}}(S)\cdot\mathsf{c}(S\cup\{\gamma\}) \;\le\; \sum_{S\subseteq N\setminus\{\gamma\}}\kappa_{\mathcal{R}}(S)\cdot\mathsf{c}(S)\,. \tag{6}$$

Observe that if $\mathcal{L}$ contains several singletons then (2) also has several generalized monotonicity re-writings. Let us illustrate the formula (6) by an example.

Example 2 (generalized monotonicity form of a clutter inequality)**.** Consider the same clutter $\mathcal{L}=\{\ \{a,b\},\ \{a,c\},\ \{b,c\},\ \{d\}\ \}$ as in Example 1. Then necessarily $\gamma=d$ and $\mathcal{R}=\{\ \{a,b\},\ \{a,c\},\ \{b,c\}\ \}$ which leads to $\kappa_{\mathcal{R}}(\{a,b\})=$

$\kappa_{\mathcal{R}}(\{a,c\}) = \kappa_{\mathcal{R}}(\{b,c\}) = 1$ and $\kappa_{\mathcal{R}}(\{a,b,c\}) = -2$. Thus, the generalized monotonicity form (6) of the inequality is

$$\begin{aligned} &\mathsf{c}(\{a,b,d\}) + \mathsf{c}(\{a,c,d\}) + \mathsf{c}(\{b,c,d\}) - 2\cdot\mathsf{c}(\{a,b,c,d\}) \\ &\qquad \leq \quad \mathsf{c}(\{a,b\}) + \mathsf{c}(\{a,c\}) + \mathsf{c}(\{b,c\}) - 2\cdot\mathsf{c}(\{a,b,c\}), \end{aligned}$$

which is just a re-writing of the inequality from Example 1.

5. Completeness conjecture

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

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

Recall that the convention $\mathsf{c}(L) = 1$ for $L \subseteq N$, $|L| = 1$, implies that (1) holds with equality provided $|\bigcup\mathcal{L}| = 1$. On the other hand, if a clutter $\mathcal{L}$ with $|\bigcup\mathcal{L}| \geq 2$ *does not contain a singleton* then (1) is *not valid* for $\mathsf{c} \in D_N$ since the characteristic imset of the empty graph produces a RHS of zero in (1).

Conjecture 1 can be viewed as a substantial step towards the solution to a more general problem when a prescribed clique size limit is given.

Conjecture 2. *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 the class $\{L \subseteq N : |L| < k\}$, contain at least one singleton and satisfy $|\bigcup\mathcal{L}| \geq 2$.*

Note that not every inequality from Conjecture 2 is facet-defining for D_N^k (see Example 4); the problem of characterization of facets of D_N^k is more subtle.

5.1. Clutter inequalities in the case of 4 variables

To illustrate Conjecture 1 let us list the 9 types of the 50 facet-defining inequalities for D_N in 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 adopt this abbreviation within this subsection. 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 $1 \leq \mathsf{c}(abc) + \mathsf{c}(d) - \mathsf{c}(abcd)$ and, because of $\mathsf{c}(d) = 1$, one gets $\mathsf{c}(abcd) \leq \mathsf{c}(abc)$ (4 inequalities),

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

The *cluster inequalities*, whose special cases are the upper bounds, correspond to clutters consisting of singletons only (see Example 3 for details):

upper bounds: take $\mathcal{L} = \{a, b\}$, (2) gives $1 \leq \mathsf{c}(a) + \mathsf{c}(b) - \mathsf{c}(ab)$ and, since $\mathsf{c}(a) = \mathsf{c}(b) = 1$, one gets $\mathsf{c}(ab) \leq 1$ (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 a 4-element-set: take $\mathcal{L} = \{a, b, c, d\}$ and (2) leads similarly to

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

Besides 28 "basic" inequalities, which have 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

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

Note that the last inequality is equivalent to the one from Example 2.

In the case $n = 4$ and the clique size limit $k = 3$, the restricted polytope D_N^3 has 60 vertices, encoding 60 chordal graphs in which N is not complete. The polytope is specified by 1 equality constraint and 49 facet-defining inequalities, decomposing into 8 permutation types. These are obtained from the above ones by the substitution $\mathsf{c}(abcd) = 0$. Thus, the number of facets is nearly the same as in the unrestricted case.

However, the polytope D_N^2 with $n = |N| = 4$ and $k = 2$ is considerably simpler: it has 38 vertices, encoding 38 *undirected forests* over four nodes. The polytope is specified by 5 equality constraints of the form $\mathsf{c}(abcd) = 0 = \mathsf{c}(abc)$, and by 17 facet-defining inequalities decomposing into 4 permutation types. These are either the *lower bounds* of form $0 \leq \mathsf{c}(ab)$ or the *cluster inequalities* of 3 types, including the upper bounds $\mathsf{c}(ab) \leq 1$. In particular, some of the clutter inequalities mentioned above are not facet-defining in this subcase.

6. Other versions of clutter inequalities

To prove the validity of the clutter inequalities from Conjecture 1 it is useful to re-write them in terms of alternative vector representatives. In this section, we apply a convenient linear transformation to the vectors $\mathsf{c} \in \mathbb{R}^{\mathcal{P}(N)}$ in (1). Moreover, we re-write (1) in terms of family variable vectors.

6.1. Clutter inequalities in terms of Möbius inversion

A very useful re-writing of the clutter inequality (1) is in terms of a linear transformation of the vector $\mathsf{c} \in \mathbb{R}^{\mathcal{P}(N)}$, known as the Möbius inversion.

Definition 4 (superset Möbius inversion)**.**
Given a vector $\mathsf{c} \in \mathbb{R}^{\mathcal{P}(N)}$*, the* **superset Möbius inversion** *of* c *is the vector* $\mathsf{m} \in \mathbb{R}^{\mathcal{P}(N)}$ *determined by the formula*

$$\mathsf{m}(T) \;:=\; \sum_{S:\, T\subseteq S} (-1)^{|S\setminus T|}\cdot \mathsf{c}(S) \quad \textit{for any} \;\; T\subseteq N, \tag{7}$$

which is equivalent to the condition

$$\mathsf{c}(S) \;=\; \sum_{T:\, S\subseteq T} \mathsf{m}(T) \qquad \textit{for any} \;\; S\subseteq N. \tag{8}$$

Indeed, to verify (7)⇒(8) write for a fixed $S\subseteq N$:

$$\sum_{T:\, S\subseteq T} \mathsf{m}(T) \;\stackrel{(7)}{=}\; \sum_{T:\, S\subseteq T}\,\sum_{L:\, T\subseteq L} (-1)^{|L\setminus T|}\cdot \mathsf{c}(L) = \sum_{L:\, S\subseteq L} \mathsf{c}(L)\cdot \sum_{T:\, S\subseteq T\subseteq L} (-1)^{|L\setminus T|}$$
$$= \sum_{L:\, S\subseteq L} \mathsf{c}(L)\cdot \sum_{B\subseteq L\setminus S} (-1)^{|B|} = \sum_{L:\, S\subseteq L} \mathsf{c}(L)\cdot \delta(\, L\setminus S=\emptyset\,) = \mathsf{c}(S)\,.$$

The proof of the implication (8)⇒(7) is analogous.

Now, we give the form of clutter inequalities in this context. Note that the transformed coefficient-vector need not be sparse even for small clutters $\mathcal{L}$.

Lemma 2 (clutter inequality in terms of superset Möbius inversion)**.**
Let $\mathcal{L}$ be a clutter of subsets of N such that $\emptyset\neq\bigcup\mathcal{L}$. Then the clutter inequality induced by $\mathcal{L}$ has the following form in terms of superset Möbius inversion m of the vector $\mathsf{c}\in\mathbb{R}^{\mathcal{P}(N)}$:

$$1\leq v(\mathsf{c},\mathcal{L}) \;=\; \sum_{T\subseteq N} \delta(T\in\mathcal{L}^{\uparrow})\cdot \mathsf{m}(T)\,. \tag{9}$$

Moreover, the formula (9) has the following non-redundant form

$$1-|\mathcal{L}\setminus\Upsilon| \;\leq\; \sum_{T\in\Upsilon} \lambda_{\mathcal{L}}(T)\cdot \mathsf{m}(T), \quad \text{where} \tag{10}$$
$$\lambda_{\mathcal{L}}(T) := \delta(T\in\mathcal{L}^{\uparrow}) - \sum_{i\in T} \delta(\{i\}\in\mathcal{L}) \quad \text{for any } T\in\Upsilon.$$

in the proper linear space $\mathbb{R}^{\Upsilon}$.

We moved the proof of Lemma 2 to Appendix A to make the paper smoothly readable. Note that the relation of the coefficients in (4) and in (10) is that $\kappa_{\mathcal{L}}$ is the *subset Möbius inversion* of $\lambda_{\mathcal{L}}$ restricted to Υ:

$$\begin{aligned}
\lambda_{\mathcal{L}}(T) &= \sum_{S\in\Upsilon:\, S\subseteq T} \kappa_{\mathcal{L}}(S) \qquad \text{for } \; T\in\Upsilon, \quad \text{and conversely} \\
\kappa_{\mathcal{L}}(S) &= \sum_{T\in\Upsilon:\, T\subseteq S} (-1)^{|S\setminus T|}\cdot \lambda_{\mathcal{L}}(T) \qquad \text{for } S\in\Upsilon.
\end{aligned}$$

The superset Möbius inversion m_G of the characteristic imset c_G of a chordal graph G can serve as an alternative vector representative of the respective decomposable model. Here is the formula for m_G.

Lemma 3 (superset Möbius inversion of the characteristic imset)**.**
Given a chordal graph G over N, let m_G denote the superset Möbius inversion of its characteristic imset c_G, given by (7) where $\mathsf{c}=\mathsf{c}_G$ and the convention $\mathsf{c}_G(\emptyset):=1$ is accepted. Assume that $\mathcal{C}(G)$ is the class of cliques of G, $\mathcal{S}(G)$ the class of separators in G and let $w_G(S)$ denote the multiplicity of a separator $S\in\mathcal{S}(G)$. Then, for any $T\subseteq N$,

$$\begin{aligned}
\mathsf{m}_G(T) &= \sum_{C\in\mathcal{C}(G)} \delta(T=C) \,-\sum_{S\in\mathcal{S}(G)} w_G(S)\,\cdot\,\delta(T=S) \qquad (11)\\
&= \sum_{j=1}^{m} \delta(T=C_j) \,-\sum_{j=2}^{m} \delta(T=S_j)\,,
\end{aligned}$$

where $C_1,\ldots,C_m$ is an arbitrary ordering of elements of $\mathcal{C}(G)$ satisfying RIP.

The proof of Lemma 3 can be found in Appendix B. It follows from the formula (11) that m_G need not be a zero-one vector because of multiplicities of separators. Nevertheless, in comparison with c_G, its superset Möbius inversion m_G is typically a sparse vector in the sense that most of its components are zeros. The vector m_G is a minor modification of the concept of a *standard imset* treated already in [24, Section 7.2.2] and it is also close to zero-one encodings of junction trees used in [22].

6.2. *Family variable formulation of clutter inequalities*

This subsection requires a reader familiar with details of the ILP-approach to learning Bayesian network structure. Recall from [7] that the *family variable* vector encoding a directed acyclic graph H over N is a zero-one vector η with components indexed by pair (a,B), where $a\in N$ and $B\subseteq N\setminus\{a\}$; let us denote the component of η indexed by such a pair by $\eta_{a\leftarrow B}$. Specifically, $\eta_{a\leftarrow B}=1$ iff $B=\mathrm{pa}_H(a)$ is the set of parents of the node a in H. Thus, every such vector belongs to the polyhedron of vectors η specified by constraints $0\leq \eta_{a\leftarrow B}\leq 1$ for any (a,B) and $\sum_{B\subseteq N\setminus\{a\}}\eta_{a\leftarrow B}=1$ for any $a\in N$, which is a common frame for family variable representatives.

Another possible (non-unique) vector representative of the decomposable model induced by a chordal graph G over N is any family variable vector η encoding a directed acyclic graph H over N inducing the same structural model as G. There is a linear relation between the characteristic imset $\mathsf{c}=\mathsf{c}_G$ and the family variable vector η. Specifically, it was shown in [27, Lemma 3] that

$$\mathsf{c}(S) \;=\; \sum_{a\in S}\;\sum_{B:\,S\setminus\{a\}\subseteq B\subseteq N\setminus\{a\}} \eta_{a\leftarrow B} \qquad \text{for } \emptyset\neq S\subseteq N. \tag{12}$$

Recall in this context that the value $\mathsf{c}(\emptyset)$ for the empty set is irrelevant in (1). The formula (12) allows us to re-formulate the clutter inequality (1) in terms of family variables with zero-one coefficients.

Lemma 4 (clutter inequality in terms of family variable vectors)**.**
Let $\mathcal{L}$ be a clutter of subsets of N such that $\emptyset\neq\bigcup\mathcal{L}$. Then (1), re-written in terms of the family variable vector η inducing c through (12), takes the form

$$1\leq v(\mathsf{c},\mathcal{L}) \;=\; \sum_{a\in\bigcup\mathcal{L}}\;\sum_{B\subseteq N\setminus\{a\}} \rho_{a\leftarrow B}\cdot\eta_{a\leftarrow B}, \quad \text{where} \tag{13}$$

$$\rho_{a\leftarrow B} \;=\; \begin{cases} 1 & \text{if there exists } L\in\mathcal{L} \text{ with } L\subseteq B\cup\{a\} \text{ while}\\ & \quad\text{there is no } R\in\mathcal{L} \text{ with } R\subseteq B,\\ 0 & \text{otherwise.}\end{cases}$$

The proof of Lemma 4 was shifted to Appendix C. Let us illustrate the result by an example.

Example 3 (cluster inequalities)**.** Given a cluster of variables $C \subseteq N$, $|C| \geq 2$, consider the clutter $\mathcal{L} = \{\, \{a\} \;:\; a \in C\}$. Then, in (13), $\rho_{a\leftarrow B} = 1$ iff $a \in C$ and $B \cap C = \emptyset$. In particular, the corresponding clutter inequality has the form

$$1 \leq \sum_{a\in C} \sum_{B\subseteq N\setminus C} \eta_{a\leftarrow B}$$

in family variables. Note that this is a well-known *cluster inequality* mentioned in [14, 7]. One can derive from (4) in Lemma 1 that it has the form

$$1 - |C| \;\leq\; \sum_{S\in\Upsilon:\, S\subseteq C} (-1)^{|S|+1} \cdot \mathsf{c}(S)\,,$$

in terms of the characteristic imset, which also follows from [27, Lemma 7]. The cluster inequalities are known to be facet-defining for the family-variable polytope, defined as the convex hull of all family variable vectors encoding directed acyclic graphs over N; this can be derived from [8, Corollary 4]. Special cases of the cluster inequalities are the upper bounds (see § 5.1) where $|C| = 2$.

7. Validity of clutter inequalities

To show the validity of the clutter inequality (1) for every $\mathsf{c} \in D_N$ we use its re-writing (9) in terms of Möbius inversion from Lemma 2 and the formula (11) for the Möbius inversion of a characteristic imset from Lemma 3.

Corollary 1. Given a chordal graph G over N, let $C_1, \ldots, C_m$, $m \geq 1$, be any ordering of elements of the class $\mathcal{C}(G)$ of (all) cliques of G satisfying the RIP. Given a clutter $\mathcal{L}$ of subsets of N with $\emptyset \neq \bigcup \mathcal{L}$ one has

$$v(\mathsf{c}_G, \mathcal{L}) \;=\; \sum_{j=1}^{m} \delta(C_j \in \mathcal{L}^{\uparrow}) \,-\sum_{j=2}^{m} \delta(S_j \in \mathcal{L}^{\uparrow})\,. \qquad (14)$$

Proof. We write using the formulas (9) and (11):

$$
\begin{aligned}
v(\mathsf{c}_G, \mathcal{L}) &\overset{(9)}{=} \sum_{T \subseteq N} \delta(T \in \mathcal{L}^{\uparrow}) \cdot \mathsf{m}_G(T) \\
&\overset{(11)}{=} \sum_{T \subseteq N} \delta(T \in \mathcal{L}^{\uparrow}) \cdot \left[\sum_{j=1}^{m} \delta(T = C_j) \; - \sum_{j=2}^{m} \delta(T = S_j) \right] \\
&= \sum_{j=1}^{m} \sum_{T \subseteq N} \delta(T = C_j) \cdot \delta(T \in \mathcal{L}^{\uparrow}) \; - \sum_{j=2}^{m} \sum_{T \subseteq N} \delta(T = S_j) \cdot \delta(T \in \mathcal{L}^{\uparrow}) \\
&= \sum_{j=1}^{m} \delta(C_j \in \mathcal{L}^{\uparrow}) \; - \sum_{j=2}^{m} \delta(S_j \in \mathcal{L}^{\uparrow}) \,,
\end{aligned}
$$

which concludes the proof of (14). $\square$

Now, the proof of the validity of (1) is easy.

Theorem 1 (validity of clutter inequalities)**.**
Given a chordal graph G over N, $|N| \geq 2$, all inequalities from Conjecture 1 are valid for the characteristic imset c_G. Hence, they are valid for any $\mathsf{c} \in D_N$.

Proof. The validity of the lower bound $0 \leq \mathsf{c}_G(N)$ is immediate. As concerns (1), 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 [16, Lemma 2.18]. By Corollary 1, one has

$$
v(\mathsf{c}_G, \mathcal{L}) \overset{(14)}{=} \underbrace{\delta(C_1 \in \mathcal{L}^{\uparrow})}_{=1} + \sum_{j=2}^{m} \underbrace{\{ \delta(C_j \in \mathcal{L}^{\uparrow}) - \delta(S_j \in \mathcal{L}^{\uparrow}) \}}_{\geq 0} \geq 1 \,,
$$

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

8. The clutter inequalities define facets

We observe that every inequality induced by a singleton-containing clutter is facet-defining for the unrestricted chordal graph polytope D_N. In fact, we are going to show the next result in the case of a prescribed clique size limit.

Lemma 5. Given $2 \leq k \leq n = |N|$, let $\mathcal{L}$ be a clutter of subsets of N containing a singleton such that $|\bigcup \mathcal{L}| \geq 2$ and $|L \cup R| \leq k$ for any $L, R \in \mathcal{L}$. Then the inequality (1) induced by $\mathcal{L}$ is facet-defining for D_N^k.

Since the proof of Lemma 5 is very long it is shifted to Appendix D. Note that it provides solely a sufficient condition on a clutter $\mathcal{L}$ to define a facet of D_N^k as the example below shows. However, we believe that the proof from Appendix D works under weaker conditions on $\mathcal{L}$.

Example 4 (non-facet clutter inequality in the restricted case)**.** If $n = 5$ and $k = 3$ then take the clutter $\mathcal{L} = \{\,\{a,b\},\{c,d\},\{e\}\,\}$ with $N = \{a,b,c,d,e\}$. Thus, $\mathcal{L}$ is a subclass of $\{L \subseteq N : |L| < k\}$ mentioned in Conjecture 2 but the condition from Lemma 5 is not fulfilled. By (4), the non-redundant clutter inequality for $\mathcal{L}$ has the next form in this restricted case:

$$0 \leq \mathsf{c}(\{a,b\}) + \mathsf{c}(\{c,d\}) - \mathsf{c}(\{a,b,e\}) - \mathsf{c}(\{c,d,e\}) \qquad \text{for } \mathsf{c} \in D_N^3.$$

This is, however, the sum of the inequalities

$$0 \leq \mathsf{c}(\{a,b\}) - \mathsf{c}(\{a,b,e\}), \quad 0 \leq \mathsf{c}(\{c,d\}) - \mathsf{c}(\{c,d,e\}) \qquad \text{for } \mathsf{c} \in D_N,$$

which are clutter inequalities for $\mathcal{L}_1 = \{\,\{a,b\},\{e\}\,\}$ and $\mathcal{L}_2 = \{\,\{c,d\},\{e\}\,\}$.

Now, the main result follows.

Theorem 2 (clutter inequalities define facets)**.**
For every clutter $\mathcal{L} \subseteq \mathcal{P}(N)$ containing a singleton and satisfying $|\bigcup \mathcal{L}| \geq 2$, the corresponding inequality (1) *is facet-defining for $D_N \equiv D_N^n$.*

Proof. If $k = n$ then the condition on $\mathcal{L}$ from Lemma 5 is fulfilled. $\square$

9. The separation problem in the cutting plane method

The effort to find a complete polyhedral description of the polytope $D_N^{\mathcal{K}}$ from § 2.5 is motivated by the aim to apply a *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 (restricted) chordal graph polytope.

However, since every clutter inequality is facet-defining for D_N (see § 8), 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 *cutting plane method* [7]. 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 $D_N \subseteq \mathsf{P}$, 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: $\mathbb{Z}^\Upsilon \cap \mathsf{P} = \mathbb{Z}^\Upsilon \cap D_N$. Moreover, facet-defining inequalities for D_N appear to be the most useful ones, leading to good overall performance.

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* [28], 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 following 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 $\mathcal{L} \mapsto v(\mathsf{c}^*, \mathcal{L})$ over singleton-containing clutters $\mathcal{L} \subseteq \mathcal{P}(N)$ with $|\bigcup \mathcal{L}| \geq 2$.

Our idea is to re-formulate this task in the form of a few auxiliary LP problems. To this end we fix an 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 separation problem into $n = |N|$ subproblems.

To solve the subproblem with fixed $\gamma \in N$ we denote

$$M := N \setminus \{\gamma\}\,, \quad \mathcal{R} := \mathcal{L} \setminus \{\,\{\gamma\}\,\} \;\text{ for any considered clutter } \mathcal{L}$$

and realize that $\mathcal{R}$ is a clutter of subsets of M with $\emptyset \neq \bigcup \mathcal{R}$. Write using the formulas from § 4 and the convention $\mathsf{c}^*(L) = 1$ for $L \subseteq N$, $|L| = 1$:

$$\begin{aligned} v(\mathsf{c}^*, \mathcal{L}) - 1 &\stackrel{(2)}{=} \sum_{S\subseteq N} \kappa_{\mathcal{L}}(S) \cdot \mathsf{c}^*(S) - 1 \\ &\stackrel{(5)}{=} \sum_{S\subseteq M} \kappa_{\mathcal{R}}(S) \cdot \mathsf{c}^*(S) - \sum_{\emptyset\neq L\subseteq M} \kappa_{\mathcal{R}}(L) \cdot \mathsf{c}^*(L \cup \{\gamma\}) + \underbrace{\mathsf{c}^*(\{\gamma\}) - 1}_{0} \\ &= \sum_{S\subseteq M} \kappa_{\mathcal{R}}(S) \cdot [\, \mathsf{c}^*(S) - \mathsf{c}^*(S \cup \{\gamma\})\,] , \end{aligned}$$

because of $\kappa_{\mathcal{R}}(\emptyset) = 0$. Thus, the subproblem is to minimize

$$\mathcal{R} \;\mapsto\; \sum_{S\subseteq M} \kappa_{\mathcal{R}}(S) \cdot \underbrace{[\, \mathsf{c}^*(S) - \mathsf{c}^*(S \cup \{\gamma\})\,]}_{o^*(S)} \tag{15}$$

over clutters $\mathcal{R} \subseteq \mathcal{P}(M)$ with $\emptyset \neq \bigcup \mathcal{R}$ and it can be re-formulated in the form of an LP problem to minimize a linear objective o^* over the *clutter polytope*

$$\mathsf{Q} := \mathrm{conv}\,(\,\{\, \kappa_{\mathcal{R}} \in \mathbb{R}^{\mathcal{P}(M)} \,:\, \mathcal{R} \subseteq \mathcal{P}(M) \text{ is a clutter with } \bigcup \mathcal{R} \neq \emptyset \,\}\,)\,. \tag{16}$$

Note that the inequality (1) corresponding to $\mathcal{L} = \mathcal{R} \cup \{\, \{\gamma\} \,\}$ is violated by c^* iff the respective value of the objective in (15) is strictly negative. Moreover, provided the monotonicity inequalities (see § 5.1) are involved in the specification of the starting relaxation P the objective vector $o^* \in \mathbb{R}^{\mathcal{P}(M)}$ in (15) has non-negative components. Below we give a polyhedral description of the clutter polytope Q, which is surprisingly simple: if $|M| \geq 3$ then the number of facets of Q is smaller than the number of its vertices.

The proof of our result is based on the following auxiliary observation; recall that a *filter* is a class of sets closed under supersets.

Lemma 6 (polyhedral description of a transformed clutter polytope)**.**
Let M be a non-empty finite set. Given $\mathcal{F} \subseteq \mathcal{P}(M)$, introduce

$$\sigma_{\mathcal{F}}(T) \;:=\; \delta(T \in \mathcal{F}) \qquad \text{for } T \subseteq M$$

the indicator vector of $\mathcal{F}$. Then the filter polytope

$$\mathsf{R} := \mathrm{conv}\,(\,\{\, \sigma_{\mathcal{F}} \in \mathbb{R}^{\mathcal{P}(M)} \,:\, \mathcal{F} \subseteq \mathcal{P}(M) \text{ is a filter with } \emptyset \notin \mathcal{F}, M \in \mathcal{F} \,\}\,) \tag{17}$$

is characterized the following linear constraints:

$$\sigma(\emptyset)=0,\quad \sigma(M)=1,\quad \sigma(B)\leq\sigma(B\cup\{a\})\quad \text{for } a\in M,\ B\subseteq M\setminus\{a\}\,. \qquad (18)$$

The proof of Lemma 6 is in Appendix E. Now, one can show the following.

Theorem 3 (polyhedral description of the clutter polytope)**.**
The clutter polytope Q *from* (16) *is determined by the following linear constraints on* $\kappa\in\mathbb{R}^{\mathcal{P}(M)}$*:*

- $0=\kappa(\emptyset)$, $1=\sum_{S\subseteq M}\kappa(S)$,
- $0\leq\sum_{L\subseteq B}\kappa(L\cup\{a\})$ *for any pair* (a,B) *where* $a\in M$, $B\subseteq M\setminus\{a\}$.

Observe that the inequalities from Theorem 3 imply $0\leq\kappa(\{a\})$ for any $a\in M$. Note that the number of inequalities in Theorem 3 is just the the number of family variables for M, that is, $|M|\cdot 2^{|M|-1}$, or equivalently, the number of edges of the Hasse diagram for the poset $(\mathcal{P}(M),\subseteq)$.

Proof. The idea is to use a suitable linear transformation. Recall from the proof of Lemma 2, formula (A.1), that $\kappa_{\mathcal{R}}$ is the subset Möbius inversion of the indicator of $\mathcal{F}:=\mathcal{R}^{\uparrow}$, the filter generated by $\mathcal{R}$, that is,

$$\sigma_{\mathcal{F}}(T)\ =\ \delta(T\in\mathcal{F})\ =\ \delta(T\in\mathcal{R}^{\uparrow})\ =\ \sum_{S\subseteq T}\kappa_{\mathcal{R}}(S)\quad \text{for any } T\subseteq M,$$

The one-to-one linear mapping $\kappa\leftrightarrow\sigma$ transforms Q to the polytope R defined by (17). It follows from Lemma 6 that R is specified by constraints (18), which turn into the constraints mentioned in Theorem 3 because of the transformation formula $\sigma(T)=\sum_{S\subseteq T}\kappa(S)$ for $T\subseteq N$. $\square$

10. Preliminary computational experiments

We have implemented some methods for solving the separation problem from § 9 by extending the GOBNILP system [7] 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 the cluster inequalities, that is, the clutter inequalities where all clutter members are singletons (see Example 3). Extending this to find the guaranteed best clutter cut for all possible clutters, for example by exploiting Theorem 3, 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.* [15], 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 [15] 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 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.* [15, § 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 is delayed until further progress in pruning methods is achieved.

11. Conclusion: further theoretical results and open tasks

We have achieved several theoretical results on the clutter inequalities. In particular, we have succeeded to show that every inequality from Conjecture 1 is *facet-defining* for the chordal graph polytope D_N.

There are further supporting arguments for the conjectures from § 5. More specifically, we are able to show using a classic matroid theory result by Edmonds [9] that a complete polyhedral description for D_N^2 consists of the lower bounds and the cluster inequalities. Thus, Conjecture 2 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. Nevertheless, to keep the length of this paper within standard limits we decided to postpone the proofs of these two results to a later publication.

The big theoretical challenge remains: to confirm/disprove Conjecture 1. Even if confirmed, another open problem is to characterize clutter inequalities defining facets for D_N^k, $2 \leq k \leq n$.

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 result of such pruning procedure should be a class $\mathcal{K} \subseteq \mathcal{P}(N)$ of sets closed under subsets defining the restricted chordal graph polytope (see § 2.5). 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.

Acknowledgements

The research of Milan Studený has been supported by the grant GAČR number 16-12010S. We are grateful to reviewers of our former PGM-2016 contribution for their comments.

Appendix A. Proof of Lemma 2

Let us recall what we are going to prove.

Rephrasing Lemma 2: Let $\mathcal{L}$ be a clutter of subsets of N such that $\emptyset \neq \bigcup \mathcal{L}$. Recall that the superset Möbius inversion m of the vector $\mathsf{c} \in \mathbb{R}^{\mathcal{P}(N)}$ satisfies

$$\mathsf{c}(S) \;=\; \sum_{T:\, S\subseteq T} \mathsf{m}(T) \qquad \text{for any} \quad S \subseteq N. \tag{8}$$

Then the clutter inequality (1) induced by $\mathcal{L}$ has the following form in terms m:

$$1 \leq v(\mathsf{c},\mathcal{L}) \;=\; \sum_{T\subseteq N} \delta(T \in \mathcal{L}^{\uparrow})\cdot \mathsf{m}(T)\,. \tag{9}$$

Moreover, the formula (9) has the following non-redundant form

$$1 - |\mathcal{L} \setminus \Upsilon| \;\leq\; \sum_{T\in\Upsilon} \lambda_{\mathcal{L}}(T) \cdot \mathsf{m}(T), \quad \text{where} \tag{10}$$

$$\lambda_{\mathcal{L}}(T) := \delta(T \in \mathcal{L}^{\uparrow}) - \sum_{i\in T} \delta(\{i\} \in \mathcal{L}) \quad \text{for any } T \in \Upsilon.$$

in the proper linear space $\mathbb{R}^{\Upsilon}$.

Proof. The first observation is that the coefficient-vector $\kappa_{\mathcal{L}} \in \mathbb{R}^{\mathcal{P}(N)}$ from (2) is closely related to the indicator of $\mathcal{L}^{\uparrow}$:

$$\delta(T \in \mathcal{L}^{\uparrow}) = \sum_{S\subseteq T} \kappa_{\mathcal{L}}(S) \qquad \text{for any } T \subseteq N. \tag{A.1}$$

To this end fix $T \subseteq N$, denote $\mathcal{L}_T := \{\, L \in \mathcal{L} \,:\, L \subseteq T\}$ and write

$$\begin{aligned}\sum_{S\subseteq T} \kappa_{\mathcal{L}}(S) &\stackrel{(2)}{=} \sum_{S\subseteq T} \; \sum_{\emptyset\neq\mathcal{B}\subseteq\mathcal{L}:\, \bigcup\mathcal{B}=S} (-1)^{|\mathcal{B}|+1} = \sum_{\emptyset\neq\mathcal{B}\subseteq\mathcal{L}:\, \bigcup\mathcal{B}\subseteq T} (-1)^{|\mathcal{B}|+1} \\ &= \; 1 - \sum_{\mathcal{B}\subseteq\mathcal{L}:\, \bigcup\mathcal{B}\subseteq T} (-1)^{|\mathcal{B}|} = 1 - \sum_{\mathcal{B}\subseteq\mathcal{L}_T} (-1)^{|\mathcal{B}|} = \delta(\mathcal{L}_T \neq \emptyset)\end{aligned}$$

and it remains to realize that $\mathcal{L}_T \neq \emptyset$ iff $T \in \mathcal{L}^{\uparrow}$. This allows us to write:

$$\begin{aligned}v(\mathsf{c},\mathcal{L}) &\stackrel{(2)}{=} \sum_{S\subseteq N} \kappa_{\mathcal{L}}(S) \cdot \mathsf{c}(S) \stackrel{(8)}{=} \sum_{S\subseteq N} \kappa_{\mathcal{L}}(S) \cdot \sum_{T:\, S\subseteq T} \mathsf{m}(T) \\ &= \sum_{T\subseteq N} \mathsf{m}(T) \cdot \sum_{S\subseteq T} \kappa_{\mathcal{L}}(S) \stackrel{(A.1)}{=} \sum_{T\subseteq N} \mathsf{m}(T) \cdot \delta(T \in \mathcal{L}^{\uparrow})\,,\end{aligned}$$

which concludes the proof of (9). To derive (10) from (9) note that $\emptyset \notin \mathcal{L}^{\uparrow}$ and, for any $i \in N$, one has $\{i\} \in \mathcal{L}^{\uparrow} \;\Leftrightarrow\; \{i\} \in \mathcal{L}$ and

$$\mathsf{m}(\{i\}) \stackrel{(8)}{=} \mathsf{c}(\{i\}) - \sum_{S\in\Upsilon:\, i\in S} \mathsf{m}(S) = 1 - \sum_{S\in\Upsilon:\, i\in S} \mathsf{m}(S)\,,$$

which allows one to write:

$$
\begin{aligned}
v(\mathsf{c},\mathcal{L}) &\stackrel{(9)}{=} \sum_{T\in\Upsilon} \mathsf{m}(T)\cdot\delta(T\in\mathcal{L}^{\uparrow}) + \sum_{i\in N} \mathsf{m}(\{i\})\cdot\delta(\{i\}\in\mathcal{L}) \\
&= \sum_{T\in\Upsilon} \mathsf{m}(T)\cdot\delta(T\in\mathcal{L}^{\uparrow}) + \sum_{i\in N}\left[1-\sum_{S\in\Upsilon:\, i\in S}\mathsf{m}(S)\right]\cdot\delta(\{i\}\in\mathcal{L}) \\
&= \sum_{i\in N}\delta(\{i\}\in\mathcal{L}) + \sum_{T\in\Upsilon}\mathsf{m}(T)\cdot\delta(T\in\mathcal{L}^{\uparrow}) - \sum_{i\in N}\sum_{S\in\Upsilon:\, i\in S}\mathsf{m}(S)\cdot\delta(\{i\}\in\mathcal{L}) \\
&= |\mathcal{L}\setminus\Upsilon| + \sum_{T\in\Upsilon}\mathsf{m}(T)\cdot\delta(T\in\mathcal{L}^{\uparrow}) - \sum_{S\in\Upsilon}\mathsf{m}(S)\cdot\sum_{i\in S}\delta(\{i\}\in\mathcal{L}) \\
&= |\mathcal{L}\setminus\Upsilon| + \sum_{T\in\Upsilon}\mathsf{m}(T)\cdot\underbrace{\left[\delta(T\in\mathcal{L}^{\uparrow}) - \sum_{i\in T}\delta(\{i\}\in\mathcal{L})\right]}_{\lambda_{\mathcal{L}}(T)}.
\end{aligned}
$$

which concludes the proof of (10). □

Appendix B. Proof of Lemma 3

Let us recall what we are going to prove.

Recalling Lemma 3: Given a chordal graph G over N, let m_G denote the superset Möbius inversion of its characteristic imset c_G, where $\mathsf{c}=\mathsf{c}_G$ and the convention $\mathsf{c}_G(\emptyset)=1$ is accepted. Assume that $\mathcal{C}(G)$ is the class of cliques of G, $\mathcal{S}(G)$ the class of separators in G and let $w_G(S)$ denote the multiplicity of a separator $S\in\mathcal{S}(G)$. Then, for any $T\subseteq N$,

$$
\begin{aligned}
\mathsf{m}_G(T) &= \sum_{C\in\mathcal{C}(G)}\delta(T=C) \;-\sum_{S\in\mathcal{S}(G)} w_G(S)\,\cdot\,\delta(T=S) \qquad (11)\\
&= \sum_{j=1}^{m}\delta(T=C_j) \;-\sum_{j=2}^{m}\delta(T=S_j)\,,
\end{aligned}
$$

where $C_1,\ldots,C_m$ is an arbitrary ordering of elements of $\mathcal{C}(G)$ satisfying RIP.

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

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

Thus, c_G is obtained from m' by the backward formula (8). Hence, since the Möbius inversion is a one-to-one transformation, one has

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

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

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

Indeed, (B.2) can be derived from (B.1) 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

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

because the other terms cancel each other (this follows from the RIP). The above formula then justifies the induction step because $\mathcal{C}(G)\setminus\{C\}$ is also the class of cliques of a chordal graph (over a smaller set of variables).

Since the order of cliques is irrelevant in (B.1), the expression in (B.2) does not depend on the choice of the ordering satisfying the RIP. In particular, (B.2) can be written in the form (11), 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. $\square$

Appendix C. Proof of Lemma 4

Let us recall what we are going to prove.

Rephrasing Lemma 4: Let $\mathcal{L}$ be a clutter of subsets of N such that $\emptyset \neq \bigcup \mathcal{L}$. Recall the formula relating $\mathsf{c} \in \mathbb{R}^{\Upsilon}$ to the family variable vector η:

$$\mathsf{c}(S) \;=\; \sum_{a\in S} \;\sum_{B:\, S\setminus\{a\}\subseteq B\subseteq N\setminus\{a\}} \eta_{\,a\leftarrow B} \qquad \text{for } \emptyset\neq S\subseteq N. \tag{12}$$

Then the clutter inequality (1), re-written in terms of η takes the form

$$1 \leq v(\mathsf{c},\mathcal{L}) \;=\; \sum_{a\in\bigcup\mathcal{L}} \;\sum_{B\subseteq N\setminus\{a\}} \rho_{\,a\leftarrow B}\cdot \eta_{\,a\leftarrow B}, \quad \text{where} \tag{13}$$

$$\rho_{\,a\leftarrow B} \;=\; \begin{cases} 1 & \text{if there exists } L\in\mathcal{L} \text{ with } L\subseteq B\cup\{a\} \text{ while} \\ & \quad \text{there is no } R\in\mathcal{L} \text{ with } R\subseteq B, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Let us substitute (12) into (1) and get

$$\begin{aligned} 1 &\leq \sum_{\emptyset\neq\mathcal{B}\subseteq\mathcal{L}} (-1)^{|\mathcal{B}|+1}\cdot \mathsf{c}(\textstyle\bigcup\mathcal{B}) \\ &\overset{(12)}{=} \sum_{\emptyset\neq\mathcal{B}\subseteq\mathcal{L}} (-1)^{|\mathcal{B}|+1}\cdot \sum_{a\in\bigcup\mathcal{B}} \;\sum_{B\subseteq N\setminus\{a\}\,:\,(\bigcup\mathcal{B})\setminus\{a\}\subseteq B} \eta_{\,a\leftarrow B} \\ &= \sum_{a\in\bigcup\mathcal{L}} \;\sum_{B\subseteq N\setminus\{a\}} \eta_{\,a\leftarrow B}\cdot \underbrace{\sum_{\mathcal{B}\subseteq\mathcal{L}:\, a\in\bigcup\mathcal{B}\subseteq B\cup\{a\}} (-1)^{|\mathcal{B}|+1}}_{\rho_{\,a\leftarrow B}}\,. \end{aligned}$$

To derive a formula for $\rho_{\,a\leftarrow B}$, with fixed $a\in\bigcup\mathcal{L}$ and $B\subseteq N\setminus\{a\}$, we put

$$\begin{aligned} \mathcal{L}[a\leftarrow B] &:= \{\, L\in\mathcal{L} \,:\, a\in L \;\;\&\;\; L\subseteq B\cup\{a\}\,\}, \quad \text{and} \\ \mathcal{L}[B] &:= \{\, R\in\mathcal{L} \,:\, R\subseteq B\,\}\,. \end{aligned}$$

Firstly, show that $\mathcal{L}[B]\neq\emptyset \;\Rightarrow\; \rho_{\,a\leftarrow B}=0$. To this end choose and fix $R\in\mathcal{L}[B]$ and realize that the condition $a\in\bigcup\mathcal{B}\subseteq B\cup\{a\}$ holds for $\mathcal{B}\subseteq\mathcal{L}$ iff it holds for $\mathcal{B}\cup\{R\}$, respectively for $\mathcal{B}\setminus\{R\}$. Thus, the index set in the sum defining

$\rho_{a\leftarrow B}$ decomposes into pairs $\mathcal{B}\cup\{R\} \leftrightarrow \mathcal{B}\setminus\{R\}$ and one can write:

$$\begin{aligned}
\rho_{a\leftarrow B} &= \sum_{\mathcal{B}\subseteq\mathcal{L}\colon a\in\bigcup\mathcal{B}\subseteq B\cup\{a\}} (-1)^{|\mathcal{B}|+1} \\
&= \sum_{\mathcal{B}\subseteq\mathcal{L}\colon a\in\bigcup\mathcal{B}\subseteq B\cup\{a\}\ \&\ R\notin\mathcal{B}} (-1)^{|\mathcal{B}|+1} + \sum_{\mathcal{B}\subseteq\mathcal{L}\colon a\in\bigcup\mathcal{B}\subseteq B\cup\{a\}\ \&\ R\in\mathcal{B}} (-1)^{|\mathcal{B}|+1} \\
&= \sum_{\mathcal{B}\subseteq\mathcal{L}\colon a\in\bigcup\mathcal{B}\subseteq B\cup\{a\}\ \&\ R\notin\mathcal{B}} \underbrace{\left[(-1)^{|\mathcal{B}|+1} + (-1)^{|\mathcal{B}\cup\{R\}|+1}\right]}_{0} \;=\; 0\,.
\end{aligned}$$

Secondly, assume that $\mathcal{L}[B]=\emptyset$, that is, $\forall\, L\in\mathcal{L} \;\; L\subseteq\{a\}\cup B \;\Rightarrow\; a\in L$, and observe that then, for any $\mathcal{B}\subseteq\mathcal{L}$, one has

$$\left[\; a\in\bigcup\mathcal{B}\subseteq B\cup\{a\}\;\right] \quad\Leftrightarrow\quad \emptyset\neq\mathcal{B}\subseteq\mathcal{L}[a\leftarrow B]\,.$$

This allows one to write in the case $\mathcal{L}[B]=\emptyset$:

$$\begin{aligned}
\rho_{a\leftarrow B} = \sum_{\mathcal{B}\subseteq\mathcal{L}\colon a\in\bigcup\mathcal{B}\subseteq B\cup\{a\}} (-1)^{|\mathcal{B}|+1} &= \sum_{\emptyset\neq\mathcal{B}\subseteq\mathcal{L}[a\leftarrow B]} (-1)^{|\mathcal{B}|+1} \\
= \; 1+\sum_{\mathcal{B}\subseteq\mathcal{L}[a\leftarrow B]} (-1)^{|\mathcal{B}|+1} \;&=\; \delta(\,\mathcal{L}[a\leftarrow B]\neq\emptyset\,)\,.
\end{aligned}$$

Hence, $\rho_{a\leftarrow B} = \delta(\,\mathcal{L}[B]=\emptyset\,)\cdot\delta(\,\mathcal{L}[a\leftarrow B]\neq\emptyset\,)$, which gives (13) because in case $\mathcal{L}[B]=\emptyset$ every $L\in\mathcal{L}$, $L\subseteq B\cup\{a\}$ contains $\{a\}$ and belongs to $\mathcal{L}[a\leftarrow B]$. □

Appendix D. Proof of Lemma 5

We base our proof on the following lemma, which is a kind of re-formulation of the method from [28, Approach 2 to Problem 1 in § 9.2.3].

Lemma 7. Let P be a *full-dimensional* polytope in $\mathbb{R}^s$, $s\geq 1$, and

$$\lambda_0 \;\leq\; \sum_{i=1}^{s} \lambda_i\cdot x_i \quad \text{for } x\equiv[x_1,\ldots,x_s]\in\mathbb{R}^s \quad (\text{where } \lambda_0,\lambda_1,\ldots,\lambda_s\in\mathbb{R}) \tag{D.1}$$

a valid inequality for any $x\in\mathsf{P}$, with at least one non-zero coefficient from $\lambda_1,\ldots,\lambda_s\in\mathbb{R}$. Assume that there exist vectors $x^1,\ldots,x^r$, $r\geq s$, on the respective face, that is, vectors from P satisfying (D.1) with equality, such that

every real solution $\mu_0, \mu_1, \ldots, \mu_s$ of the equations

$$\forall\, j = 1, \ldots, r \qquad \mu_0 = \sum_{i=1}^{s} \mu_i \cdot x_i^j \tag{D.2}$$

is a multiple of $\lambda_0, \lambda_1, \ldots, \lambda_s$, i.e. $\exists\, \alpha \in \mathbb{R} \quad \mu_i = \alpha \cdot \lambda_i$ for $i = 0, 1, \ldots, s$.

Then the inequality (D.1) is facet-defining for P. In case $r = s$ the vectors $x^1, \ldots, x^s$ satisfying (D.2) are necessarily affinely independent.

Note that at least one of the coefficients $\lambda_1, \ldots, \lambda_s \in \mathbb{R}$ is assumed to be non-zero since otherwise the existence of $x^1, \ldots, x^r$ implies $\lambda_0 = 0$ and (D.1) is valid with equality for any $x \in \mathbb{R}^s$ and, therefore, it is not facet-defining.

Proof. Firstly, observe that the condition (D.2) implies that the affine hull of $\{x^1, \ldots, x^r\}$ is an affine subspace of $\mathbb{R}^s$ given by $\lambda_0 = \langle \lambda, x \rangle := \sum_{i=1}^{s} \lambda_i \cdot x_i$. Indeed, $x \in \mathbb{R}^s$ belongs to the affine hull iff the corresponding extended vector $\tilde{x} := (1, x) \equiv (1, x_1, \ldots, x_s) \in \mathbb{R}^{s+1}$ is in the linear hull of extended vectors $\tilde{x}^1, \ldots, \tilde{x}^r \in \mathbb{R}^{s+1}$: this is because for $\beta_j \in \mathbb{R}$, $j = 1, \ldots, r$ one has

$$(1, x) = \sum_{j=1}^{r} \beta_j \cdot (1, x^j) \quad \Leftrightarrow \quad \left[\, \sum_{j=1}^{r} \beta_j = 1 \ \ \& \ \ x = \sum_{j=1}^{r} \beta_j \cdot x^j \,\right].$$

Thus, it is enough to show that (D.2) implies

$$\mathrm{Lin}\,(\{\tilde{x}^1, \ldots, \tilde{x}^r\}) = \underbrace{\{\, (y_0, \ldots, y_s) \in \mathbb{R}^{s+1} \ : \ -\lambda_0 \cdot y_0 + \sum_{i=1}^{s} \lambda_i \cdot y_i = 0 \,\}}_{L},$$

where $\mathrm{Lin}\,(-)$ denotes the linear hull and L the linear space specified by the constraint given by the coefficients $-\lambda_0, \lambda_1, \ldots, \lambda_s$; note that, for $x \in \mathbb{R}^s$, one has $\tilde{x} = (1, x) \in L$ iff x satisfies $\lambda_0 = \langle \lambda, x \rangle$.

The inclusion $\mathrm{Lin}\,(\{\tilde{x}^1, \ldots, \tilde{x}^r\}) \subseteq L$ is evident because vectors $x^1, \ldots, x^r$ are assumed to belong to the face given by the respective inequality in (D.1). The other inclusion $L \subseteq \mathrm{Lin}\,(\{\tilde{x}^1, \ldots, \tilde{x}^r\})$ is equivalent to the converse inclusion of their orthogonal complements $\mathrm{Lin}\,(\{\tilde{x}^1, \ldots, \tilde{x}^r\})^{\perp} \subseteq L^{\perp}$. But this is exactly what the condition (D.2) requires: whenever $\tilde{\mu} = (-\mu_0, \mu_1, \ldots, \mu_s) \in \mathrm{Lin}\,(\{\tilde{x}^1, \ldots, \tilde{x}^r\})^{\perp}$ then $\tilde{\mu} \in \mathrm{Lin}\,(\,\{(-\lambda_0, \lambda_1, \ldots, \lambda_s)\}\,) = L^{\perp}$.

Thus, provided (D.2) holds, the affine hull of $\{x^1, \ldots, x^r\}$ is determined by just one equality constraint in $\mathbb{R}^s$ and has the dimension $s-1$, because $\lambda_1, \ldots, \lambda_s$ are non-vanishing. In particular, the inequality (D.1) defines a face of P of the dimension $s-1$, that is, a facet.

The conclusion that in case $r = s$ the vectors $x^1, \ldots, x^s$ satisfying (D.2) are affinely independent can be derived as follows. In this case, the linear hull of $\tilde{x}^1, \ldots, \tilde{x}^s \in \mathbb{R}^{s+1}$ is the space L of the dimension s. But every set of s vectors linearly generating an s-dimensional subspace must be linearly independent. The linear independence of $\tilde{x}^1, \ldots, \tilde{x}^s$ implies for $\gamma_j \in \mathbb{R}$, $j = 1, \ldots, s$, that

$$\left[\sum_{j=1}^{s} \gamma_j = 0 \;\;\&\;\; \sum_{j=1}^{s} \gamma_j \cdot x^j = 0 \in \mathbb{R}^s\right]$$
$$\Rightarrow \;\; \sum_{j=1}^{s} \gamma_j \cdot \tilde{x}^j = 0 \in \mathbb{R}^{s+1} \quad \Rightarrow \quad [\; \gamma_j = 0 \quad \text{for } j = 1, \ldots, s \;],$$

that is, $x^1, \ldots, x^s \in \mathbb{R}^s$ are affinely independent. $\square$

Let us recall Lemma 5 in more appropriate form before giving its proof.

Rephrasing of Lemma 5: Given $2 \leq k \leq n = |N|$, let us denote

$$\mathcal{K} \;:=\; \{\, S \subseteq N \;:\; |S| \leq k \,\}\,.$$

Let $\mathcal{L}$ be a clutter of subsets of N containing a singleton such that $|\bigcup \mathcal{L}| \geq 2$ and $L \cup R \in \mathcal{K}$ for any $L, R \in \mathcal{L}$. Then the inequality (1) induced by $\mathcal{L}$ is facet-defining for D_N^k.

Proof. The proof is more transparent if we transform the polytope $D_N^k \subseteq D_N$ by the superset Möbius inversion (7) $\mathsf{c} \mapsto \mathsf{m}$ to the polytope

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

and rewrite (1) accordingly. The dimension of P is $\sum_{\ell=2}^{k}\binom{n}{\ell}$, the same like the one of D_N^k; the affine hull of P is

$$\begin{aligned}\mathsf{A} \;=\; \{\, \mathsf{m}\in\mathbb{R}^{\mathcal{P}(N)} :\quad & \mathsf{m}(T)=0 \quad \text{for}\ \ T\notin\mathcal{K},\ \text{while}\\ & \sum_{T\subseteq N}\mathsf{m}(T)=1 \ \ \text{and} \ \ \sum_{T\subseteq N:\, a\in T}\mathsf{m}(T)=1 \ \ \text{for any}\ a\in N\,\},\end{aligned}$$

where we use the fact that $\mathcal{P}(N)\setminus\mathcal{K}$ is a filter. Elements of $\mathsf{P}\subseteq\mathsf{A}$ can be identified with vectors in $\mathbb{R}^{\mathcal{K}\cap\Upsilon}$ where

$$\Upsilon=\{S\subseteq N \,:\, |S|\geq 2\}$$

is the class of non-empty non-singletons. This is because the restriction of $\mathsf{m}\in\mathsf{A}$ to components in $\mathcal{K}\cap\Upsilon$ determines affinely the values $\mathsf{m}(T)$ for $T\subseteq N$ outside $\mathcal{K}\cap\Upsilon$. Moreover, P is a full-dimensional polytope in $\mathbb{R}^{\mathcal{K}\cap\Upsilon}$, which fact can be derived from Lemma 3.

Given a singleton-containing clutter $\mathcal{L}$ with $L,R\in\mathcal{L} \;\Rightarrow\; L\cup R\in\mathcal{K}$ and $|\bigcup\mathcal{L}|\geq 2$, we need appropriate rewriting of (1) in terms of $\mathbb{R}^{\mathcal{K}\cap\Upsilon}$. This is in Lemma 2, the formula (10), where we include the constraints for $\mathsf{m}\in\mathsf{A}$:

$$\begin{aligned}\lambda_0 \;\leq\; & \sum_{T\in\mathcal{K}\cap\Upsilon}\lambda(T)\cdot\mathsf{m}(T), \quad \text{for}\ \mathsf{m}\in\mathbb{R}^{\mathcal{K}\cap\Upsilon} \quad \text{with} \qquad \text{(D.3)}\\ \lambda_0 \;&=\; 1-|\mathcal{L}\setminus\Upsilon|\\ \lambda(T) \;&=\; \delta(T\in\mathcal{L}^{\uparrow})-\sum_{i\in T}\delta(\{i\}\in\mathcal{L}) \quad \text{for}\ T\in\mathcal{K}\cap\Upsilon.\end{aligned}$$

Observe that $|\bigcup\mathcal{L}|\geq 2$ implies that the coefficients in the RHS of (D.3) are not identically vanishing. One can derive from Theorem 1 that (D.3) is valid for any $\mathsf{m}\in\mathsf{P}$. Thus, we can use the criterion from Lemma 7 with $\mathsf{P}\subseteq\mathbb{R}^{\mathcal{K}\cap\Upsilon}$ and the inequality (D.3).

To apply that criterion one has to construct a class $\mathcal{G}$ of *chordal graphs* G over N with *cliques in* $\mathcal{K}$ that are *tight for the clutter* $\mathcal{L}$, which means that m_G satisfies (D.3) with equality. The vectors m_G for $G\in\mathcal{G}$, viewed as elements in $\mathbb{R}^{\mathcal{K}\cap\Upsilon}$, will serve as the vectors on the face of P given by (D.3); a formula for

m_G in $\mathbb{R}^{\mathcal{K}\cap\Upsilon}$ follows from (11) in Lemma 3:

$$m_G(T) = \sum_{C\in\mathcal{C}(G)\cap\Upsilon} \delta(T=C) \; - \sum_{S\in\mathcal{S}(G)\cap\Upsilon} w_G(S)\cdot\delta(T=S) \quad \text{for } T\in\mathcal{K}\cap\Upsilon.$$

The goal is to construct such class $\mathcal{G}$ that the condition (D.2) from Lemma 7 holds for $\{\mathsf{m}_G \,:\, G\in\mathcal{G}\}$ in place of $x^1,\ldots,x^r$, which means that, *every collection* of real numbers μ_0 and $\mu(T)$, $T\in\mathcal{K}\cap\Upsilon$, satisfying

$$\forall\, G\in\mathcal{G} \qquad \mu_0 \;=\; \sum_{T\in\mathcal{K}\cap\Upsilon} \mu(T)\cdot \mathsf{m}_G(T) \tag{D.4}$$

must be a multiple of the collection λ_0 and $\lambda(T)$, $T\in\mathcal{K}\cap\Upsilon$, from (D.3). If we find such a class $\mathcal{G}$ of graphs then Lemma 7 implies that (D.3) is facet-defining for P. The fact that the superset Möbius inversion (7) is linearly invertible by (8) then implies that (1) is facet-defining for D^k_N.

Roughly, a general principle of the construction of $\mathcal{G}$ is as follows: for every $S\in\mathcal{K}\cap\Upsilon$, we include in $\mathcal{G}$ a pair of graphs G and H such that the validity of (D.4) for m_G and m_H allows one to derive a conclusion on the value of $\mu(S)$. Given a clutter $\mathcal{L}$ containing a singleton and $|\bigcup\mathcal{L}|\geq 2$ we introduce

$$\Lambda \;:=\; \bigcup(\mathcal{L}\setminus\Upsilon) = \bigcup_{\{i\}\in\mathcal{L}} \{i\}\,, \qquad \text{and} \qquad \Gamma \;:=\; N\setminus\Lambda\,,$$

and the details of the construction of graphs included in $\mathcal{G}$ depend on whether

- $\lambda_0<0$, that is, $|\Lambda|\geq 2$, or
- $\lambda_0=0$, that is, $|\Lambda|=1$, in other words, $\mathcal{L}$ only has one singleton.

The sets in $\mathcal{K}\cap\Upsilon$ will be classified into 4 classes (= 4 cases of the construction):

A. (if $|\Lambda|\geq 2$) sets $S\in\mathcal{K}\cap\Upsilon$ such that $S\subseteq\Lambda$,

B. sets $S\in\mathcal{K}\cap\Upsilon$ with $S\cap\Lambda\neq\emptyset\neq S\cap\Gamma$,

C. sets $S\in\mathcal{K}\cap\Upsilon$ with $S\subseteq\Gamma$ and $S\not\in\mathcal{L}^{\uparrow}$,

D. (if $\mathcal{L}\cap\Upsilon\neq\emptyset$) sets $S\in\mathcal{K}\cap\Upsilon$ with $S\subseteq\Gamma$ and $S\in\mathcal{L}^{\uparrow}$.

Moreover, one special graph will be constructed and included in $\mathcal{G}$ in order

E. to derive a conclusion on the constant μ_0.

Now, the specific constructions in the above described cases will be given. Throughout the constructions, the vector $\delta_S \in \mathbb{R}^{\mathcal{K}\cap\Upsilon}$, where $S \subseteq N$, will denote the zero-one identifier of the set S:

$$\delta_S(T) := \begin{cases} 1 & \text{if } T = S, \\ 0 & \text{if } T \neq S, \end{cases} \qquad \text{for } T \in \mathcal{K} \cap \Upsilon.$$

A. If $|\Lambda| \geq 2$ consider the collection of sets

$$\mathcal{S} := \{S \in \mathcal{K} \cap \Upsilon \,:\, S \subseteq \Lambda\}\,,$$

which is non-empty then, and realize that one has $\lambda(S) = 1 - |S| < 0$ for any set $S \in \mathcal{S}$. The whole consideration in this A-case has four steps. All these steps are empty in case $|\Lambda| = 2$ because then $|\mathcal{S}| = 1$; thus, assume $|\Lambda| \geq 3$.

A.1. Verify that $\mu(S) = \mu(T)$ for every pair $S, T \subseteq \Lambda$ with $|S| = |T| = 2$.

To this end, it is enough to verify $\mu(S) = \mu(T)$ under an additional assumption that $|S \cap T| = 1$: this is because in case $S \cap T = \emptyset$ choose $s \in S$, $t \in T$, put $R = \{s, t\}$ and have $R \subseteq \Lambda$ while $|S \cap R| = 1 = |R \cap T|$. Thus, without loss of generality assume $S = \{a, c\}$ and $T = \{b, c\}$ and construct a tree J over $\Lambda \setminus \{a, b\}$ in which c is a leaf (= it has at most one neighbour in the tree J). Then the corresponding construction of two graphs will be as follows:

- the graph G will have cliques $\{a, b\}$, $\{a, c\}$, all two-element cliques of J, and the singletons in Γ,
- the graph H will have cliques $\{a, b\}$, $\{b, c\}$, all two-element cliques of J, and the singletons in Γ.

Since G and H are forests over N, they are chordal graphs having cliques in $\mathcal{K}$. Both graphs also have exactly $|\Lambda| - 1$ two-element cliques; these cliques C are subsets of Λ and one has $\lambda(C) = -1$ for them. Thus, the RHS of (D.3) for

m_G and m_H is $1 - |\Lambda| = 1 - |\mathcal{L} \setminus \Upsilon| = \lambda_0$, which means that G and H are tight for $\mathcal{L}$. Hence, we can include G and H in $\mathcal{G}$. Because, in $\mathbb{R}^{\mathcal{K}\cap\Upsilon}$, one has $\mathsf{m}_G - \mathsf{m}_H = \delta_{\{a,c\}} - \delta_{\{b,c\}}$, it follows from (D.4) that

$$\begin{aligned} 0 = \mu_0 - \mu_0 &\stackrel{(D.4)}{=} \sum_{T\in\mathcal{K}\cap\Upsilon} \mu(T)\cdot \mathsf{m}_G(T) - \sum_{T\in\mathcal{K}\cap\Upsilon} \mu(T)\cdot \mathsf{m}_H(T) \\ &= \sum_{T\in\mathcal{K}\cap\Upsilon} \mu(T)\cdot [\mathsf{m}_G(T) - \mathsf{m}_H(T)] = \mu(\{a,c\}) - \mu(\{b,c\}) = \mu(S) - \mu(T)\,, \end{aligned}$$

which was the goal.

A.2. Denote by μ^* the shared value $\mu(S)$ for $S \subseteq \Lambda$, $|S| = 2$.

A.3. Verify that, for every $S \in \mathcal{S}$, $|S| \geq 3$, one has $\mu(S) = (|S| - 1)\cdot \mu^*$.

To this end, choose a node $c \in S$ and a tree J over S in which c is a leaf. In case $\Lambda \setminus S \neq \emptyset$ also choose a node $d \in \Lambda \setminus S$ and a tree I over $\Lambda \setminus S$ in which d is a leaf. Then the construct

- the graph G which has as cliques S, the singletons in Γ and, optionally in case $\Lambda \setminus S \neq \emptyset$, also $\{c,d\}$ and two-element cliques of I,
- the graph H which has as cliques of those of J, the singletons in Γ and, in case $\Lambda \setminus S \neq \emptyset$, also $\{c,d\}$ and the two-element cliques of I.

It is easy to observe that G and H are chordal graphs over N, and, since $S \in \mathcal{S} \subseteq \mathcal{K}$, their cliques are in $\mathcal{K}$. Because H is a forest, the RHS in (D.3) for m_H is λ_0 for the same reason as mentioned in A.1-case. As concerns G, in $\mathbb{R}^{\mathcal{K}\cap\Upsilon}$, one has $\mathsf{m}_G - \mathsf{m}_H = \delta_S - \sum_{\{u,v\}\in\mathcal{J}} \delta_{\{u,v\}}$, where $\mathcal{J}$ is the set of cliques of J. Thus, because $\lambda(S) = 1 - |S| = \sum_{\{u,v\}\in\mathcal{J}} \lambda(\{u,v\})$, the RHS in (D.3) for m_G is also λ_0. Therefore, we can include both G and H into $\mathcal{G}$. It follows from (D.4) by subtracting the respective equations that

$$0 = \mu(S) - \sum_{\{u,v\}\in\mathcal{J}} \mu(\{u,v\}) = \mu(S) - (|S| - 1)\cdot \mu^*,$$

using the convention A.2.

A.4. Summary: we have constructed and put in $\mathcal{G}$ such graphs that (D.4) implies that there exists μ^* such that $\mu(S) = (|S| - 1)\cdot \mu^*$ for any $S \in \mathcal{S}$.

B. If $S \in \mathcal{K} \cap \Upsilon$ with $S \cap \Lambda \neq \emptyset \neq S \cap \Gamma$ then $\lambda(S) = 1 - |S \cap \Lambda| = \lambda(S \cap \Lambda)$, where we accept the convention that $\lambda(L) = 0$ whenever $L \subseteq \Lambda$, $|L| = 1$. Verify $\mu(S) = \mu(S \cap \Lambda)$ under an analogous convention $\mu(L) = 0$ for $L \subseteq \Lambda$, $|L| = 1$.

To this end, provided $\Lambda \setminus S \neq \emptyset$, choose $c \in S \cap \Lambda$, $d \in \Lambda \setminus S$ and a tree J over $\Lambda \setminus S$ in which d is a leaf. Then construct:

- the graph G which has cliques S, singletons in $\Gamma \setminus S$ and, optionally in case $\Lambda \setminus S \neq \emptyset$, also $\{c, d\}$ and two-element cliques of J,
- the graph H whose complete sets are determined as subsets of $S \cap \Lambda$, of singletons in Γ and, optionally in case $\Lambda \setminus S \neq \emptyset$, also of $\{c, d\}$ and two-element cliques of J.

Since $S \in \mathcal{K}$, one also has $S \cap \Lambda \in \mathcal{K}$; thus, the cliques of G and H are in $\mathcal{K}$. The formula for m_G in $\mathbb{R}^{\mathcal{K} \cap \Upsilon}$ consists of δ_S plus an optional term

$$\delta_{\{c,d\}} + \sum_{\{u,v\} \in \mathcal{J}} \delta_{\{u,v\}}, \quad \text{where } \mathcal{J} \text{ is the class of cliques of } J.$$

The formula for m_H consists of $\delta_{S \cap \Lambda} \in \mathbb{R}^{\mathcal{K} \cap \Upsilon}$ (meaning that $\delta_{S \cap \Lambda} = 0$ in case $|S \cap \Lambda| = 1$) plus the same optional term. Hence, the RHS in (D.3) for both m_G and m_H is $(1 - |S \cap \Lambda|) - |\Lambda \setminus S| = 1 - |\Lambda| = \lambda_0$ and (D.3) holds with equality for them. Therefore, we can include G and H into $\mathcal{G}$ and subtracting of equations (D.4) for G and H gives

$$0 = \sum_{T \in \mathcal{K} \cap \Upsilon} \mu(T) \cdot [\mathsf{m}_G(T) - \mathsf{m}_H(T)] = \mu(S) - \mu(S \cap \Lambda)\,,$$

where we have the convention $\mu(L) = 0$ for $L \subseteq \Lambda$, $|L| = 1$.

C. If $S \in \mathcal{K} \cap \Upsilon$ with $S \subseteq \Gamma$ and $S \notin \mathcal{L}^{\uparrow}$ then one has $\lambda(S) = 0$. Verify $\mu(S) = 0$.

To this end, choose a tree J over Λ and construct:

- the graph G which has as cliques S, the cliques of J and all the singletons in the set $\Gamma \setminus S$,
- the graph H which has as cliques those of J and singletons in Γ.

Both graphs are chordal and have cliques in $\mathcal{K}$. Observe that, in $\mathbb{R}^{\mathcal{K}\cap\Upsilon}$, one has

$$\mathsf{m}_H = \sum_{\{u,v\}\in\mathcal{J}} \delta_{\{u,v\}}, \qquad \text{where } \mathcal{J} \text{ is the class of cliques of } J,$$

while $\mathsf{m}_G = \mathsf{m}_H + \delta_S$. Since $\sum_{\{u,v\}\in\mathcal{J}} \lambda(\{u,v\}) = (|\Lambda|-1)\cdot(-1) = \lambda_0$ both graphs belong to the face determined by (D.3). Including G and H into $\mathcal{G}$ allows one to subtract the respective equations in (D.4) and obtain $0 = \mu(S)$.

D. If $S \in \mathcal{K}\cap\Upsilon$ with $S \subseteq \Gamma$ and $S \in \mathcal{L}^{\uparrow}$ then $\lambda(S) = 1$. The details of the consideration depend on $|\Lambda|$, but in any case the next step will be needed.

D.1. Given $L \in \mathcal{L}$ and $S \in \mathcal{K}$ with $L \subseteq S \subseteq \Gamma$ verify that $\mu(S) = \mu(L)$.

Note that the assumption implies $L \in \Upsilon$ and we also know from (D.3) that $\lambda(S) = \lambda(L) = 1$. We choose $c \in \Lambda$ and a tree J over Λ in which c is a leaf. The corresponding construction is as follows:

- the graph G has cliques S, $L\cup\{c\}$, all two-element cliques of J and the singletons in $\Gamma\setminus S$,
- the graph H has as cliques $L\cup\{c\}$, all the two-element cliques of J and the singletons in $\Gamma\setminus L$.

By the assumption $L\cup R \in \mathcal{K}$ for any $L, R \in \mathcal{L}$ we are sure that $L\cup\{c\} \in \mathcal{K}$, and, by construction, both graphs over N are chordal and have cliques in $\mathcal{K}$. The formulas for superset Möbius inversions in $\mathbb{R}^{\mathcal{K}\cap\Upsilon}$ are

$$\begin{aligned}
\mathsf{m}_H &= \delta_{L\cup\{c\}} + \sum_{\{u,v\}\in\mathcal{J}} \delta_{\{u,v\}}, \quad \text{where } \mathcal{J} \text{ is the class of cliques of } J,\\
\mathsf{m}_G &= \mathsf{m}_H + \delta_S - \delta_L\,.
\end{aligned}$$

Hence, the RHS in (D.3) for both m_G and m_H is

$$\lambda(L\cup\{c\}) + \sum_{\{u,v\}\in\mathcal{J}} \lambda(\{u,v\}) = 0 + (-1)\cdot(|\Lambda|-1) = \lambda_0\,.$$

Since G and H are tight for $\mathcal{L}$ they can be included into $\mathcal{G}$. By subtracting the equations (D.4) for m_G and m_H one gets

$$0 = \sum_{T\in\mathcal{K}\cap\Upsilon} \mu(T)\cdot[\mathsf{m}_G(T) - \mathsf{m}_H(T)] = \mu(S) - \mu(L)\,.$$

D.2. There exists a shared value μ° for $\mu(S)$ for $S \in \mathcal{K} \cap \mathcal{L}^\uparrow$ with $S \subseteq \Gamma$.

For every pair $L, R \in \mathcal{L}$ with $L, R \subseteq \Gamma$ one has $L \cup R \in \mathcal{K}$, which allows one to deduce $\mu(L) = \mu(L \cup R) = \mu(R)$ by the previous step D.1. Thus, there is a shared value μ° for $\mu(L)$ for $L \in \mathcal{L}$ with $L \subseteq \Gamma$. By applying the observation in D.1 again we obtain the desired conclusion.

D.3. (in case $|\Lambda| \geq 2$ and $\mathcal{L} \cap \Upsilon \neq \emptyset$) observe that the shared value μ^* from the step A.2 coincides with $-\mu^\circ$, where μ° is the shared value from D.2.

Because $|\Lambda| \geq 2$, we can choose different $a, b \in \Lambda$ and a tree J over $\Lambda \setminus \{a\}$ in which b is a leaf. Because $\mathcal{L} \cap \Upsilon \neq \emptyset$, one can also choose a set $L \in \mathcal{L}$ such that $L \subseteq \Gamma$. The construction is as follows:

- the graph G has cliques $L \cup \{a\}$, $L \cup \{b\}$, two-element cliques of J, and singletons in $\Gamma \setminus L$,
- the graph H has as cliques $\{a, b\}$, two-element cliques of J and singletons in Γ.

As $L \cup R \in \mathcal{K}$ for any $L, R \in \mathcal{L}$ we know that $L \cup \{a\}, L \cup \{b\} \in \mathcal{K}$; thus, G and H are chordal graphs over N with cliques in $\mathcal{K}$. Moreover, in $\mathbb{R}^{\mathcal{K} \cap \Upsilon}$, one has

$$\begin{aligned} \mathsf{m}_H &= \delta_{\{a,b\}} + \sum_{\{u,v\} \in \mathcal{J}} \delta_{\{u,v\}}, \quad \text{where } \mathcal{J} \text{ is the class of cliques of } J, \\ \mathsf{m}_G &= \mathsf{m}_H - \delta_{\{a,b\}} + \delta_{L \cup \{a\}} + \delta_{L \cup \{b\}} - \delta_L \,. \end{aligned}$$

Hence, the RHS in (D.3) for m_H is

$$\lambda(\{a, b\}) + \sum_{\{u,v\} \in \mathcal{J}} \lambda(\{u, v\}) = 1 - |\Lambda| = \lambda_0 \,,$$

and, because $-\lambda(\{a, b\}) + \lambda(L \cup \{a\}) + \lambda(L \cup \{b\}) - \lambda(L) = +1 + 0 + 0 - 1 = 0$, the same holds for m_G. Hence, G and H can be included into $\mathcal{G}$. By subtracting the equations (D.4) for m_G and m_H one gets

$$\begin{aligned} 0 = & \sum_{T \in \mathcal{K} \cap \Upsilon} \mu(T) \cdot [\mathsf{m}_G(T) - \mathsf{m}_H(T)] \\ = & \; -\mu(\{a, b\}) + \mu(L \cup \{a\}) + \mu(L \cup \{b\}) - \mu(L) = -\mu^* + 0 + 0 - \mu^\circ, \end{aligned}$$

by the cases A.2, B and D.2. This means $\mu^* = -\mu^\circ$, which was the goal.

E. Observe that if $|\Lambda| \geq 2$ then $\mu_0 = (|\Lambda| - 1) \cdot \mu^*$ and if $|\Lambda| = 1$ then $\mu_0 = 0$.

To this end we choose a tree J over Λ and construct

- a graph G which has as cliques all the cliques of J and singletons in Γ.

This is a chordal graph over N with cliques in $\mathcal{K}$. Moreover, in $\mathbb{R}^{\mathcal{K}\cap\Upsilon}$, one has

$$\mathsf{m}_G = \sum_{\{u,v\}\in\mathcal{J}} \delta_{\{u,v\}}, \quad \text{where } \mathcal{J} \text{ is the class of cliques of } J.$$

Hence, the RHS in (D.3) for m_G is $1 - |\Lambda| = \lambda_0$ and G can be included into $\mathcal{G}$. The equation (D.4) for m_G says

$$\mu_0 = \sum_{T\in\mathcal{K}\cap\Upsilon} \mu(T) \cdot \mathsf{m}_G(T) = \sum_{\{u,v\}\in\mathcal{J}} \mu(\{u,v\}),$$

which is either zero, in case $|\Lambda| = 1$, or $(|\Lambda| - 1) \cdot \mu^*$, in case $|\Lambda| \geq 2$ by A.2.

Now, putting the observations A-E together implies that the collection of real numbers μ_0 and $\mu(T)$, $T \in \mathcal{K}\cap\Upsilon$, is a multiple of λ_0 and $\lambda(T)$, $T \in \mathcal{K}\cap\Upsilon$, which was desired. Specifically, the multiplicative factor is $-\mu^*$ from A.2 in case $|\Lambda| \geq 2$, respectively μ° from D.2 in case $|\Lambda| = 1$. $\square$

Appendix E. Proof of Lemma 6

This is the result we are going to prove.

Rephrasing Lemma 6: Let M be a non-empty finite set. Recall that a *filter* is a class of sets $\mathcal{F} \subseteq \mathcal{P}(M)$ closed under supersets: $S \in \mathcal{F}$, $S \subseteq T \subseteq M$ implies $T \in \mathcal{F}$. The indicator vector of such class $\mathcal{F}$ will be denoted as follows:

$$\sigma_{\mathcal{F}}(T) \;:=\; \delta(T \in \mathcal{F}) \qquad \text{for } T \subseteq M.$$

Then the filter polytope defined by

$$\mathsf{R} := \mathrm{conv}\,(\,\{\,\sigma_{\mathcal{F}} \in \mathbb{R}^{\mathcal{P}(M)} :\, \mathcal{F} \subseteq \mathcal{P}(M) \text{ is a filter with } \emptyset \notin \mathcal{F}, M \in \mathcal{F}\,\}\,) \tag{17}$$

is characterized the next linear constraints:

$$\sigma(\emptyset) = 0, \quad \sigma(M) = 1, \quad \sigma(B) \leq \sigma(B \cup \{a\}) \quad \text{for } a \in M,\, B \subseteq M \setminus \{a\}\,. \tag{18}$$

Proof. The validity of (18) for $\sigma_{\mathcal{F}} \in \mathsf{R}$ follows immediately from the definition of a filter. We are going to verify that every vector $\sigma \in \mathbb{R}^{\mathcal{P}(M)}$ satisfying (18) is a convex linear combination of vertices of R. This can be shown by induction on $s := |\{T \subseteq M \,:\, \sigma(T) \neq 0\}|$. Note that the inequalities (18) imply that $0 \leq \sigma(T) \leq 1$ for any $T \subseteq M$. The induction premise is immediate: if $s = 1$ then $\sigma = \sigma_{\mathcal{F}^*}$, where $\mathcal{F}^* = \{M\}$ is the filter consisting of the set M only.

To verify the induction step in case $s > 1$ we put

$$\mathcal{F} := \{\, T \subseteq M \,:\, \sigma(T) > 0 \,\} \quad \text{and} \quad \beta := \min\, \{\, \sigma(T) \,:\, T \in \mathcal{F} \,\} > 0$$

and observe that $\mathcal{F} \subseteq \mathcal{P}(M)$ is a filter with $\emptyset \notin \mathcal{F}$ and $M \in \mathcal{F}$. Note that in case $\beta = 1$ necessarily $\sigma = \sigma_{\mathcal{F}}$ and the induction step is verified. Thus, assume $\beta < 1$ in which case put

$$\sigma' := \frac{1}{1-\beta} \cdot [\, \sigma - \beta \cdot \sigma_{\mathcal{F}} \,] \in \mathbb{R}^{\mathcal{P}(M)} \quad \text{and have} \;\; \sigma = (1-\beta) \cdot \sigma' + \beta \cdot \sigma_{\mathcal{F}}.$$

Observe that since σ satisfies the constraints from (18) σ' does so: $\sigma'(\emptyset) = 0$ and $\sigma'(M) = 1$ is easy; for fixed $a \in M$ and $B \subseteq M \setminus \{a\}$, write

$$\begin{aligned}
&(1-\beta) \cdot [\, \sigma'(B \cup \{a\}) - \sigma'(B) \,] \\
&\qquad = \quad \sigma(B \cup \{a\}) - \beta \cdot \sigma_{\mathcal{F}}(B \cup \{a\}) - \sigma(B) + \beta \cdot \sigma_{\mathcal{F}}(B) \\
&\qquad = \quad \begin{cases} \sigma(B \cup \{a\}) - \sigma(B) \geq 0 & \text{if } B \in \mathcal{F} \text{ or } B \cup \{a\} \notin \mathcal{F}, \\ \sigma(B \cup \{a\}) - \beta \geq 0 & \text{if } B \notin \mathcal{F} \text{ and } B \cup \{a\} \in \mathcal{F}, \end{cases}
\end{aligned}$$

because of the definition of β. Now realize that $\sigma'(T) = 0$ for $T \subseteq M$, $T \notin \mathcal{F}$, and there exists at least one $T \in \mathcal{F}$ with $\sigma(T) = \beta$ and, therefore, $\sigma'(T) = 0$. Thus, $s' = |\{S \subseteq M \,:\, \sigma'(S) \neq 0\}| < s$ and the induction hypothesis says that σ' is a convex combination of vertices of R. The formula $\sigma = (1-\beta) \cdot \sigma' + \beta \cdot \sigma_{\mathcal{F}}$ then completes the proof of the induction step. $\square$

References

[1] 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.

[2] A. Barvinok. *A Course in Convexity*. American Mathematical Society, Providence, 2002.

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

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

[5] 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.

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

[7] 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.

[8] J. Cussens, D. Haws, and M. Studený. Polyhedral aspects of score equivalence in Bayesian network structure learning. Mathematical Programming A, published online `doi:10.1007/s10107-016-1087-2`, November 2016.

[9] 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.

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

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

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

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

[14] 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.

[15] 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.

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

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

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

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

[20] 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.

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

[22] 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.

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

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

[25] M. Studený and J. Cussens. The chordal graph polytope for learning decomposable models. In A. Antonucci, G. Corani, and C. P. de Campos, editors, *JMLR Workshop and Conference Proceedings 52: PGM 2016*, pages 499–510, 2016.

[26] 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.

[27] M. Studený and D. C. Haws. On polyhedral approximations of polytopes for learning Bayesian networks. *Journal of Algebraic Statistics*, 4:59–92, 2013.

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

[29] Y. Xiang, S. K. M. Wong, and N. Cercone. A 'microscopic' study of minimum entropy search in learning decomposable Markov networks. *Machine Learning*, 26(1):65–92, 1997.

*Highlights (for review)

Highlights - 2016 IJAR submission

Towards using the chordal graph polytope in learning decomposable models

by M. Studený and J. Cussens

- we propose zero-one vector representatives for decomposable graphical models, called *characteristic imsets*, to be used in integer linear approach to structural learning,
- we re-formulate the learning task in the form of an (integer) linear problem to maximize a linear objective over a special *chordal graph polytope*, defined as the convex hull of all characteristic imset over a fixed set of variables,
- we introduce a class of *clutter inequalities* valid for the polytope,
- we prove that all these inequalities are facet-defining for the polytope and conjecture that they yield a complete facet description of the polytope,
- we propose a linear programming method to solve the *separation problem* with these inequalities for the use in a cutting plane approach to solving the ILP problem.