Mathematical Programming series A manuscript No.
(will be inserted by the editor)

Polyhedral aspects of score equivalence in Bayesian network structure learning

James Cussens · David Haws · Milan Studený

Received: date / Accepted: date

Abstract This paper deals with faces and facets of the *family-variable polytope* and the *characteristic-imset polytope*, which are special polytopes used in integer linear programming approaches to statistically learn Bayesian network structure. A common form of linear objectives to be maximized in this area leads to the concept of *score equivalence* (SE), both for linear objectives and for faces of the family-variable polytope.

We characterize the linear space of SE objectives and establish a one-to-one correspondence between SE faces of the family-variable polytope, the faces of the characteristic-imset polytope, and standardized supermodular functions. The characterization of SE facets in terms of extremality of the corresponding supermodular function gives an elegant method to verify whether an inequality is SE-facet-defining for the family-variable polytope.

We also show that when maximizing an SE objective one can eliminate linear constraints of the family-variable polytope that correspond to non-SE facets. However, we show that solely considering SE facets is not enough as a counter-example shows; one has to consider the linear inequality constraints that correspond to facets of the characteristic-imset polytope despite the fact that they may not define facets in the family-variable mode.

J. Cussens
University of York,
Department of Computer Science & York Centre for Complex Systems Analysis
Deramore Lane, York, YO10 5GH, UK

D. C. Haws
Thomas J. Watson Research Center,
Yorktown Heights, 1101 Kitchwan Road 134, NY, USA

M. Studený
Institute of Information Theory and Automation of the CAS,
Pod Vodárenskou věží 4, Prague, 182 08, Czech Republic
Tel.: +420-26605-2304 Fax: +420-286890378
E-mail: studeny@utia.cas.cz

Keywords family-variable polytope · characteristic-imset polytope · score equivalent face/facet · supermodular set function

Mathematics Subject Classification (2010) 52B12 · 90C27 · 68Q32

1 Introduction

The motivation for our paper is learning *Bayesian network* (BN) structure statistically. Bayesian networks are popular models used in statistics [14] and probabilistic reasoning [18]. Directed acyclic graphs, whose nodes correspond to random variables in consideration, are used to describe the probabilistic conditional independence structures behind the statistical models [19].

Specifically, our motivation comes from the *integer linear programming* (ILP) approach to the statistical learning task to determine the structural model on the basis of observed data. Nowadays, the most popular is the score-based approach consisting in maximizing a *scoring criterion* $G \mapsto \mathcal{Q}(G, D)$, where G is a directed acyclic graph, D the observed database and the value $\mathcal{Q}(G, D)$ says how much the BN structure defined by the graph G explains the occurrence the database D [15].

The point of the ILP approach is that the criteria used in practice can be viewed as (the restrictions of) affine functions of suitable vector representatives of BN structures. The most common is the *family-variable* vector representation of directed acyclic graphs suggested independently in [13] and [7]. Very good running times have recently been achieved using this vector representation and the branch-and-cut approach [1,9]. The corresponding *family-variable polytope*, defined as the convex hull of these vector representatives, is one of the topics of interest in this paper.

Another ILP approach based on the *characteristic-imset* vector representation of BN structures was suggested in [12]; its motivational sources date back to [19]. Unlike the family-variable vectors, the characteristic imsets uniquely correspond to BN structures. This ILP approach is also feasible [23], but has not resulted in better running times than those achieved using GOBNILP software [9]. The other polytope we are interested in this paper is the *characteristic-imset polytope*, defined as the convex hull of all characteristic imsets.

Our paper is devoted to the comparison of facet-defining inequalities for the two above-mentioned polytopes, because such inequalities appear to be the most useful ones in the cutting plane approach to solving ILP problems [26]. There were some former results on this comparison topic in [22], but the present paper brings further and deeper findings.

The structure of the paper is as follows. In Section 2 we introduce our notation and recall basic concepts; elementary facts on polytopes we need later are gathered in Section 3. Some fundamental observations on facets of the family-variable polytope, on which our later considerations are based, are in Section 4; some of these facts are also shown using different arguments in a parallel paper [10].

In Section 5 we pinpoint the concept of *score equivalence* (SE), both for linear objectives to be maximized and for faces of the family-variable polytope. We characterize the linear space of SE objectives in Section 6. Later, in Section 7, we establish a one-to-one correspondence between SE faces of the family-variable polytope and standardized supermodular set functions. The most beneficial seems to be the characterization of SE facets as those that correspond to extreme supermodular functions.

Section 8 deals with well-known (generalized) *cluster inequalities* applied dominantly in contemporary ILP approaches to BN structure learning. We find the corresponding supermodular functions and show they are extreme. This gives a simple proof that the generalized cluster inequalities are facet-defining for the family-variable polytope; note that another proof of this fact, based on different arguments, will appear in [10]. We also interpret the generalized cluster inequalities in terms of connected uniform matroids.

Another one-to-one correspondence between SE faces of the family-variable polytope and faces of the characteristic-imset polytope is then established in Section 9. To illustrate this correspondence we derive the form of cluster inequalities in the characteristic-imset mode. A few simple examples of the polytopes and their facet-defining inequalities are given in Section 10.

Further important observations of ours are in Section 11: when maximizing an SE objective, one actually need not apply the linear facet-defining constraints on the family-variable polytope that are not SE. We also reveal the hidden importance of the linear constraints that correspond to facets of the characteristic-imset polytope. On the other hand, considering only SE facets is not enough as a later counter-example in Section 12 shows. Section 13 contains a few remarks on computational aspects whose aim is to explain why the ILP approach is applicable also in the case of higher number of nodes (in the graphs). We also pinpoint the significance of our theoretical results and observations for (the design of) practical ILP learning procedures.

The appendix contains the proof of an auxiliary combinatorial identity (Section A), a catalogue of SE facets in the case of four BN variables (Section B) and a catalogue of remaining facets of the characteristic-imset polytope in the case of four BN variables (Section C).

2 Notation and basic concepts

Let N be a finite non-empty set of BN variables; $n := |N| < \infty$, consider the non-trivial case $2 \leq n$. Let $\mathsf{DAGS}\,(N)$ denote the collection of directed acyclic graphs over N, that is, directed graphs which have N as the set of nodes and are without directed cycles. Note that we follow common usage in terminology even though the unambiguous term "acyclic directed graph" is perhaps more appropriate from a grammatical point of view. An example of such a graph is the *empty graph*, which is a graph over N without adjacencies. By a *complete graph* we will mean any directed acyclic graph over N in which every pair of distinct nodes is adjacent. Given $G \in \mathsf{DAGS}\,(N)$ and a node $a \in N$, the symbol $pa_G(a) := \{b \in N : \ a \leftarrow b \ \text{ in } G\}$ will denote the *parent set* of the node a. A well-known equivalent definition of acyclicity of a directed graph G over N is the existence of a total order $a_1, \ldots, a_n$ of nodes in N such that, for every $i = 1, \ldots, n$, $pa_G(a_i) \subseteq \{a_1, \ldots, a_{i-1}\}$; we say then that the order and the graph are *consonant*. An *immorality* in G is an induced subgraph of G of the form $a \rightarrow c \leftarrow b$, where the nodes a and b are not adjacent in G.

The symbol $G \sim H$ for $G, H \in \mathsf{DAGS}\,(N)$ will mean that the graphs G and H are *Markov equivalent*, that is, in graphical terms, they have the same adjacencies and immoralities; for references see [14, p. 60] or [19, p. 48-49]. An example of a Markov equivalence class is the set of complete graphs over N.

A node a together with its parent set B will be called a *family*. Note that any directed graph over N is determined by its $n = |N|$ families. Throughout the paper, the index set of family-variable vectors will be

$$\Upsilon := \{\, (a \,|\, B) \ : \ a \in N \ \ \& \ \ \emptyset \neq B \subseteq N \setminus \{a\} \,\} \,.$$

Note that families with empty parent sets are not included.

Given $b \in N$ and $Z \subseteq N \setminus \{b\}$ the symbol $I_{b \leftarrow Z}$ will be used to denote the identifier of this pair, that is, an element of $\mathbb{R}^{\Upsilon}$ given by

$$I_{b \leftarrow Z}(a \,|\, B) = \begin{cases} 1 & \text{if } a = b \text{ and } B = Z, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for any } (a \,|\, B) \in \Upsilon.$$

In case $Z = \emptyset$, for any $b \in N$, $I_{b \leftarrow Z} = I_{b \leftarrow \emptyset}$ is the zero vector. The symbol η_G will be used to denote the family-variable vector encoding $G \in \mathsf{DAGS}\,(N)$, that is, the DAG-code for the (directed acyclic) graph G:

$$\eta_G(a \,|\, B) = \begin{cases} 1 & \text{if } B = pa_G(a), \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for } (a \,|\, B) \in \Upsilon.$$

The *family-variable polytope* can be defined as the convex hull of the set of all possible DAG-codes over N:

$$\mathsf{F} := \mathrm{conv}\,(\{\, \eta_G \in \mathbb{R}^{\Upsilon} \ : \ G \in \mathsf{DAGS}\,(N) \,\}) \,.$$

Note that examples of the family-variable polytope F in cases $n = 3$ and $n = 4$ are given in Section 10. Clearly, the dimension of F, defined as the dimension of its linear hull, is $\dim(\mathsf{F}) = |\Upsilon| = n \cdot (2^{n-1} - 1)$. It is easy to see that none of the DAG-codes is a non-trivial convex combination of the others. In particular, the set of vertices (= extreme points) of F is just the set of DAG-codes.

Given two vectors $v, w \in \mathbb{R}^{\Gamma}$, where Γ is a non-empty finite index set, say $\Gamma = \Upsilon$, their scalar product will be denoted by $\langle v, w \rangle_{\Gamma}$, or just by $\langle v, w \rangle$ if there is no danger of confusion. We also consider alternative index sets.

Specifically, the *characteristic imset* of $G \in \mathsf{DAGS}\,(N)$, introduced in [12] and denoted below by c_G, is an element of $\mathbb{R}^{\Lambda}$ with

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

Recall from [22, § 3.3.2] and [1, § 2] that c_G is a many-to-one linear function of η_G; the transformation is $\eta \mapsto \mathsf{c}_\eta$, where

$$\mathsf{c}_\eta(S) = \sum_{a\in S} \sum_{B\,:\, S\setminus\{a\}\subseteq B\subseteq N\setminus\{a\}} \eta(a\,|\,B) \qquad \text{for any } S\subseteq N,\ |S|\geq 2. \tag{1}$$

A further fundamental observation is that $G \sim H$ for $G, H \in \mathsf{DAGS}\,(N)$ iff $\mathsf{c}_G = \mathsf{c}_H$; see [12, § 3] for more detailed justification. The *characteristic-imset polytope* is defined as follows:

$$\mathsf{C} := \operatorname{conv}\left(\{\, \mathsf{c}_G \in \mathbb{R}^\Lambda \ :\ G \in \mathsf{DAGS}\,(N)\,\}\right).$$

Examples of the characteristic-imset polytope C for $n = 3$ and $n = 4$ are also given in Section 10. One can show that $\dim(\mathsf{C}) = |\Lambda| = 2^n - n - 1$. Of course, C is the image of F by the linear map (1).

Moreover, the power set $\mathcal{P}(N) := \{A \ :\ A \subseteq N\}$ will serve as an index set for vectors, used as auxiliary tools in a later proof in Section 9. Given $A \subseteq N$, let us denote its indicator vector by

$$\delta_A(S) = \begin{cases} 1 & \text{if } \ S = A\,, \\ 0 & \text{if } \ S \subseteq N,\ S \neq A\,, \end{cases}$$

and define the *standard imset* for $G \in \mathsf{DAGS}\,(N)$ as an element of $\mathbb{R}^{\mathcal{P}(N)}$:

$$\mathsf{u}_G := \delta_N - \delta_\emptyset + \sum_{a\in N} \left\{\, \delta_{pa_G(a)} - \delta_{\{a\}\cup pa_G(a)}\,\right\}. \tag{2}$$

Recall from [22, § 3.3] that c_G is a one-to-one affine function of u_G, specifically

$$\mathsf{c}_G(T) = 1 - \sum_{S\colon T\subseteq S\subseteq N} \mathsf{u}_G(S) \qquad \text{for } T\subseteq N,\ |T|\geq 2. \tag{3}$$

In particular, the combination of a former characterization [20, Theorem 4] of the vertices of the standard-imset polytope with (3) implies that the set of vertices (= extreme points) of the characteristic-imset polytope C is just the set of characteristic imsets c_G for $G \in \mathsf{DAGS}\,(N)$. In other words, no characteristic imset is a non-trivial convex combination of the others.

3 Elementary facts on facets and some conventions

Recall the basic concept of a face/facet of a polytope.

Definition 1 (dimension, face, facet)
Let P be a polytope in $\mathbb{R}^\Gamma$, where $\Gamma \neq \emptyset$ is finite. Its *dimension* is defined as the dimension of its affine hull, which is a translate of a linear subspace of $\mathbb{R}^\Gamma$. A set $F \subseteq \mathsf{P}$ is called a *face* of P if there exists a vector $o \in \mathbb{R}^\Gamma$ and a constant $u \in \mathbb{R}$ such that

– $\mathsf{P} \subseteq \{\, v \in \mathbb{R}^\Gamma \ :\ \langle o, v\rangle \leq u\,\}$, and

– $F = \{\, v \in \mathsf{P} \,:\, \langle o, v\rangle = u \,\}$.

We say then that the face F is defined by the inequality $\langle o, v\rangle \le u$. Every face of a polytope is a (possibly empty) polytope, as well; thus, its dimension is defined. A *facet* of P is a face of dimension $\dim(\mathsf{P}) - 1$.

The function $v \in \mathbb{R}^{\Gamma} \mapsto \langle o, v\rangle$, where $o \in \mathbb{R}^{\Gamma}$, is typically a linear objective to be maximized by a linear program; with a small abuse of terminology we will call $o \in \mathbb{R}^{\Gamma}$ an *objective*.

Note that the dimension of a face is one less than the maximum number of affinely independent vectors in the face. An alternative equivalent definition of a facet is that it is a sub-maximal face with respect to inclusion.

Lemma 1 Given a polytope P in $\mathbb{R}^{\Gamma}$, $0 < |\Gamma| < \infty$, a face $F \subset \mathsf{P}$ is a facet of P iff the only face F' of P with $F \subset F'$ is $F' = \mathsf{P}$ itself.

Proof The sufficiency follows from the fact that, for every pair of faces $F_1 \subset F_2$ of P with $\dim(F_1) < d < \dim(F_2)$, a face F_3 of P exists with $F_1 \subset F_3 \subset F_2$ and $\dim(F_3) = d$; see, for example, [4, Corollary 9.7]. For the necessity realize that, if $F_1 \subset F_2$ are faces of P then $\dim(F_1) < \dim(F_2)$; see [4, Corollary 5.5].

The consequence is an auxiliary observation, applied later in the paper.

Corollary 1 Let $\mathsf{P} \subseteq \mathbb{R}^{\Gamma}$, $0 < |\Gamma| < \infty$ be a polytope and let $\langle o_1, v\rangle \le u_1$ and $\langle o_2, v\rangle \le u_2$ be valid inequalities for $v \in \mathsf{P}$ such that

$$\exists\, w_1 \in \mathsf{P} \,:\, \langle o_1, w_1\rangle < u_1 \;\&\; \langle o_2, w_1\rangle = u_2 \;\text{ and }\; \exists\, w_2 \in \mathsf{P} \,:\, \langle o_2, w_2\rangle < u_2. \quad (4)$$

Then no combination of these inequalities $\langle \alpha \cdot o_1 + \beta \cdot o_2, v\rangle \le \alpha \cdot u_1 + \beta \cdot u_2$ with $\alpha, \beta > 0$ is a facet-defining inequality for P.

Proof Let F_1, F_2 and F be the faces of P defined by inequalities $\langle o_1, v\rangle \le u_1$, $\langle o_2, v\rangle \le u_2$ and their combination $\langle \alpha\cdot o_1 + \beta\cdot o_2, v\rangle \le \alpha\cdot u_1 + \beta\cdot u_2$, respectively. Given $v \in F$ one has

$$\alpha \cdot \{\underbrace{\langle o_1, v\rangle - u_1}_{\le 0}\} + \beta \cdot \{\underbrace{\langle o_2, v\rangle - u_2}_{\le 0}\} = 0,$$

which implies that the expressions in braces must vanish. In other words, $F \subseteq F_1 \cap F_2$. Assume for a contradiction that F is a facet. By Lemma 1 observe that either $F_1 = \mathsf{P}$ or $F_1 = F$; the same for F_2. Since (4) implies $w_1 \in \mathsf{P} \setminus F_1$ and $w_2 \in \mathsf{P} \setminus F_2$, one necessarily has $F_1 = F = F_2$. However, this contradicts the existence of $w_1 \in F_2 \setminus F_1$ assumed in (4).

In this paper we mainly deal with the family-variable polytope F. Every face of F can be identified with a set of directed acyclic graphs. Specifically:

$$F \subseteq \mathsf{F} \;\text{ a face of } \mathsf{F} \quad \longleftrightarrow \quad \mathsf{S} = \{G \in \mathsf{DAGS}\,(N) :\; \eta_G \in F\}\,.$$

This correspondence preserves inclusion, that is, $F_1 \subseteq F_2$ for faces of F iff $\mathsf{S}_1 \subseteq \mathsf{S}_2$ for the corresponding sets of graphs $\mathsf{S}_i \subseteq \mathsf{DAGS}\,(N)$. The identification is possible owing to a basic fact from the theory of polytopes that every face F of a polytope P is the convex hull of the set of vertices of P which belong to F, see [2, Lemma VI.1.1] or [27, Proposition 2.3(i)]. Since the vertices of F are just the DAG-codes η_G, where $G \in \mathsf{DAGS}\,(N)$, every face F of F can be identified with a subset of $\mathsf{DAGS}\,(N)$. This leads to the following convention.

Definition 2 (a set of graphs interpreted as a face)
We will call a set $\mathsf{S} \subseteq \mathsf{DAGS}\,(N)$ a *face* (of the family-variable polytope F) if $\mathrm{conv}\,(\{\,\eta_G \in \mathbb{R}^{\Upsilon} :\ G \in \mathsf{S}\})$ is a face of F. Analogously, $\mathsf{S} \subseteq \mathsf{DAGS}\,(N)$ will be called a *facet* (of F) if $\mathrm{conv}\,(\{\,\eta_G \in \mathbb{R}^{\Upsilon} :\ G \in \mathsf{S}\})$ is a facet of F.

A direct method to show that a face $\mathsf{S} \subseteq \mathsf{DAGS}\,(N)$ is a facet is to show that the respective geometric face F has the dimension $\dim(\mathsf{F}) - 1$, which means, to find $\dim(\mathsf{F})$ affinely independent vectors in F. Since the vertices of F are just the family-variable vectors for $G \in \mathsf{S}$, the task, more or less, reduces to the question of finding a subset $\mathsf{S}' \subseteq \mathsf{S}$ of cardinality $|\Upsilon| = n \cdot (2^{n-1} - 1)$ such that the vectors $\{\,\eta_G \in \mathbb{R}^{\Upsilon} :\ G \in \mathsf{S}'\}$ are affinely independent.

We accept a *standardization convention* that valid inequalities for vectors $\eta \in \mathsf{F}$ in the family-variable polytope will be written in the upper-bound form:

$$\langle o, \eta \rangle \leq u \quad \text{where } o \in \mathbb{R}^{\Upsilon} \text{ is an objective and } u \in \mathbb{R} \text{ an upper bound.} \tag{5}$$

Note that any lower-bound inequality $\langle o', \eta \rangle \geq l$ can be replaced by $\langle o, \eta \rangle \leq u$ where $o = -o'$ and $u = -l$. Since F is a rational polytope, its facets are defined by inequalities with rational coefficients, that is, by (5) with $o \in \mathbb{Q}^{\Upsilon}$. By multiplying it by a suitable positive factor one can get (unique) integer vector objective $o \in \mathbb{Z}^{\Upsilon}$ whose components have no common prime divisor. Since the vertices of F are zero-one vectors, the tight upper bound in (5) must be then an integer as well: $u \in \mathbb{Z}$.

Moreover, a couple of special *extension conventions* for vectors in $\mathbb{R}^{\Upsilon}$ and $\mathbb{R}^{\Lambda}$ will be accepted to simplify some later formulas:

- for every *objective* $o \in \mathbb{R}^{\Upsilon}$, assume $o(b\,|\,\emptyset) = 0$ for any $b \in N$,
- for any $m \in \mathbb{R}^{\Lambda}$, put $m(S) = 0$ for $S \subseteq N$, $|S| \leq 1$.

4 Observations on facets of the family-variable polytope

In this section, we present a few general facts concerning faces and facets of F and describe explicitly those facets which contain the empty graph. Note that some of these basic observations are also mentioned and used in a parallel paper [10]. We keep the standardization convention from Section 3. The basic division of facet-defining inequalities is on the basis of the upper bound value u.

Lemma 2 Assume that (5), that is, the inequality $\langle o, \eta \rangle \leq u$ with $o \in \mathbb{R}^{\Upsilon}$ and $u \in \mathbb{R}$, is a valid inequality for all $\eta \in \mathsf{F}$. Then $u \geq 0$.

(i) One has $u = 0$ iff the corresponding face of F contains the empty graph.

(ii) If $u = 0$ then the objective coefficients are non-positive:

$$o(a\,|\,B) \leq 0 \qquad \text{for each} \;\; (a\,|\,B) \in \Upsilon.$$

(iii) The facet-defining inequalities tight at the empty graph are just

$$-\eta(a\,|\,B) \leq 0 \qquad \text{for each} \;\; (a\,|\,B) \in \Upsilon. \tag{6}$$

(iv) If (5) is a facet-defining inequality for F with $u > 0$ then the objective coefficients are non-negative and increasing in the following sense:

$$o(a\,|\,B) \geq o(a\,|\,A) \geq 0 \quad \text{whenever } a \in N \text{ and } \; \emptyset \neq A \subseteq B \subseteq N \setminus \{a\}.$$

Note that an alternative proof of Lemma 2(iv) is in [10].

Proof The zero vector in $\mathbb{R}^{\Upsilon}$ is the code for the empty graph and, therefore, belongs to F. The substitution of $\eta = 0$ into (5) gives $0 \leq u$. It is clear that the inequality is tight for $\eta = 0$ iff $u = 0$, which gives (i).

As concerns (ii), assume for a contradiction that $(a\,|\,B) \in \Upsilon$ such that $o(a\,|\,B) > 0$ exists in (5) with $u = 0$. Consider $G \in \mathsf{DAGS}\,(N)$ with $\eta_G = I_{a \leftarrow B}$. Then $\langle o, \eta_G \rangle = o(a\,|\,B) > 0 = u$ contradicts the validity of (5).

As concerns (iii), an elementary fact is that, for every $(b\,|\,D) \in \Upsilon$, all the inequalities in (6) with $(a\,|\,B) \neq (b\,|\,D)$ are tight for the family-variable vector $\eta = I_{b \leftarrow D} \in \mathsf{F}$ but not the inequality corresponding to $(b\,|\,D)$. This allows us to observe that any inequality in (6) is facet-defining for F. Indeed, any such inequality is valid for F and, having fixed $(a\,|\,B) \in \Upsilon$, the respective inequality $-\eta(a\,|\,B) \leq 0$ is tight for $|\Upsilon|$ affinely independent vectors, namely the zero vector in $\mathbb{R}^{\Upsilon}$ and vectors $I_{b \leftarrow D}$ for $(b\,|\,D) \neq (a\,|\,B)$. The second step is to show that every facet F of F containing the empty graph is defined by (6). Former observations (i) and (ii) imply that the facet-defining inequality for F must have the form $\langle o, \eta \rangle \leq 0$ with $o \in (-\infty, 0]^{\Upsilon}$. Thus, the inequality is a conic combination of those from (6). Since F is assumed to be a facet, Corollary 1 can be used to show that at most one coefficient in the combination is non-zero. Indeed, if two coefficients $o(a\,|\,B)$ and $o(b\,|\,D)$ are non-zero, the above elementary fact implies for the inequalities $o(a\,|\,B) \cdot \eta(a\,|\,B) \leq 0$ and $\sum_{(c\,|\,E) \neq (a\,|\,B)} o(c\,|\,E) \cdot \eta(c\,|\,E) \leq 0$ that the condition (4) from Corollary 1 is fulfilled with $w_1 = I_{a \leftarrow B}$ and $w_2 = I_{b \leftarrow D}$. On the other hand, at least one coefficient must be non-zero, since otherwise $F = \mathsf{F}$. Therefore, the facet F must be defined by one of the inequalities in (6).

As concerns (iv), owing to the extension convention from Section 3, the statement means $o(a\,|\,B) \geq o(a\,|\,A)$ for $a \in N$ and $A \subseteq B \subseteq N \setminus \{a\}$. Assume for a contradiction that $a \in N$ and $A \subset B \subseteq N \setminus \{a\}$ exist such that one has $o(a\,|\,B) < o(a\,|\,A)$ and define $\tilde{o} \in \mathbb{R}^{\Upsilon}$ in the following way:

$$\tilde{o}(b\,|\,D) := \begin{cases} o(b\,|\,D) & \text{for } (b\,|\,D) \in \Upsilon,\; (b\,|\,D) \neq (a\,|\,B), \\ o(a\,|\,A) & \text{for } (b\,|\,D) = (a\,|\,B). \end{cases}$$

The next observation is that $\langle \tilde{o}, \eta \rangle \leq u$ is a valid inequality for F. Specifically, given $G \in \mathsf{DAGS}\,(N)$, construct $\tilde{G} \in \mathsf{DAGS}\,(N)$ such that $\eta_{\tilde{G}}(a\,|\,B) = 0$ and $\langle \tilde{o}, \eta_G \rangle = \langle \tilde{o}, \eta_{\tilde{G}} \rangle$. Indeed, if $pa_G(a) \neq B$ then simply $\tilde{G} := G$, otherwise put $pa_{\tilde{G}}(a) = A$ and $pa_{\tilde{G}}(b) = pa_G(b)$ for $b \in N \setminus \{a\}$, which gives

$$\langle \tilde{o}, \eta_G \rangle - \langle \tilde{o}, \eta_{\tilde{G}} \rangle = \tilde{o}(a\,|\,B) - \tilde{o}(a\,|\,A) = o(a\,|\,A) - o(a\,|\,A) = 0\,.$$

The definition of $\tilde{o}$ implies $\langle \tilde{o}, \eta_{\tilde{G}} \rangle - \langle o, \eta_{\tilde{G}} \rangle = \{\tilde{o}(a\,|\,B) - o(a\,|\,B)\} \cdot \eta_{\tilde{G}}(a\,|\,B) = 0$. Because (5) is valid for $\eta_{\tilde{G}}$ one can observe

$$\langle \tilde{o}, \eta_G \rangle = \langle \tilde{o}, \eta_{\tilde{G}} \rangle = \langle o, \eta_{\tilde{G}} \rangle \leq u\,, \quad \text{which was desired.}$$

Thus, (5) is the sum of the valid inequality $\langle \tilde{o}, \eta \rangle \leq u$ with a positive multiple of the valid inequality $-\eta(a\,|\,B) \leq 0$, namely by $\beta := o(a\,|\,A) - o(a\,|\,B) > 0$. The condition (4) from Corollary 1 is fulfilled with $w_1 = 0$ and $w_2 = I_{a \leftarrow B}$, which implies a contradictory conclusion that (5) is not facet-defining.

This implies the following observation.

Corollary 2 Let S be a facet of F in the sense of Definition 2 which does not contain the empty graph. Then S is closed under super-graphs in the sense:

$$\text{if } G \in \mathsf{S} \text{ is a subgraph of } H \in \mathsf{DAGS}\,(N) \qquad \text{then } H \in \mathsf{S}.$$

Moreover, for every $(a\,|\,B) \in \Upsilon$, there exists $G \in \mathsf{S}$ with $pa_G(a) = B$.

The second statement in Corollary 2 is also derived in [10] using slightly different arguments.

Proof It is enough to verify the first claim when H differs from G in only one parent set, that is, when $a \in N$ exists with $A = pa_G(a) \subset pa_H(a) = B$ and $pa_H(b) = pa_G(b)$ for $b \in N \setminus \{a\}$. By Lemma 2(i), we know that S is given by the inequality (5) with $u > 0$. Thus, by Lemma 2(iv), one can write $\langle o, \eta_H \rangle - \langle o, \eta_G \rangle = o(a\,|\,B) - o(a\,|\,A) \geq 0$. Assuming $G \in \mathsf{S}$, the inequality (5) is tight for η_G and one has

$$u = \langle o, \eta_G \rangle \leq \langle o, \eta_H \rangle \leq u \qquad \text{because (5) is valid for } \eta_H.$$

Hence, $\langle o, \eta_H \rangle = u$, that is, (5) is tight for η_H, saying that $H \in \mathsf{S}$.

As concern the second claim assume for a contradiction that $(a\,|\,B) \in \Upsilon$ exists with $pa_G(a) \neq B$ for any $G \in \mathsf{S}$. That means, S is contained in the face defined by $-\eta(a\,|\,B) \leq 0$. Since $\mathrm{conv}\,(\{\,\eta_G \in \mathbb{R}^{\Upsilon} :\; G \in \mathsf{S}\})$ is a facet of F, by Lemma 1, observe that it coincides with the face defined by $-\eta(a\,|\,B) \leq 0$. This implies a contradictory conclusion that S contains the empty graph.

An obvious modification of natural convexity constraints gives the following valid inequalities for the family-variable polytope:

$$\sum_{B\,:\,\emptyset \neq B \subseteq N \setminus \{a\}} \eta(a\,|\,B) \leq 1 \qquad \text{for any } a \in N. \tag{7}$$

Except for a degenerate case $n = 2$, these inequalities are facet-defining; see also [10].

Lemma 3 If $n \geq 3$ then, for every $a \in N$, (7) defines a facet of F.

Proof We find $|\Upsilon|$ affinely independent vectors on the face. Specifically, for $\emptyset \neq B \subseteq N \setminus \{a\}$ put $\eta_{(a\,|\,B)} = I_{a\leftarrow B}$, while for $b \in N$, $b \neq a$ and $(b\,|\,D) \in \Upsilon$ put $\eta_{(b\,|\,D)} = I_{a\leftarrow N\setminus\{a,b\}} + I_{b\leftarrow D}$. These vectors linearly generate $\mathbb{R}^{\Upsilon}$. Hence, they are linearly independent, and, therefore, affinely independent.

5 Score equivalence concept

The score-based approach to structural learning Bayesian networks consists in maximization of a function $G \in \mathsf{DAGS}\,(N) \mapsto \mathcal{Q}(G,D)$, where D is the database of observed values and $\mathcal{Q}$ a suitable *quality criterion*, also called a *scoring criterion* [15, p. 437], which evaluates how the graph G fits the database D. The criteria used in practice turn out to be affine functions of the family-variable vector, that is, $\mathcal{Q}(G,D) = k + \langle o, \eta_G \rangle_{\Upsilon}$ with $k \in \mathbb{R}$ and $o \in \mathbb{R}^{\Upsilon}$ encoding both D and $\mathcal{Q}$. Thus, theoretically speaking, the learning task turns into an LP problem to maximize a linear function over the vertices of the family-variable polytope F.

Since the goal is typically to learn the structure, described by a Markov equivalence class of graphs, most of criteria used in practice do not distinguish between Markov equivalent graphs, that is, one has

$$\mathcal{Q}(G,D) = \mathcal{Q}(H,D) \quad \text{whenever } G \text{ and } H \text{ are Markov equivalent.}$$

In the machine learning community, quality criteria satisfying the above condition are called *score equivalent* [3,6]. This motivates the following terminology.

Definition 3 (score equivalent objective)
We say that a vector $o \in \mathbb{R}^{\Upsilon}$ is a *score equivalent objective* (abbreviated below as an *SE objective*) if it satisfies

$$\forall\, G, H \in \mathsf{DAGS}\,(N) \quad G \sim H \;\Rightarrow\; \langle o, \eta_G \rangle = \langle o, \eta_H \rangle\,. \tag{8}$$

Clearly, the set of SE objectives is a linear subspace of $\mathbb{R}^{\Upsilon}$.

The faces and facets of F are defined in terms of normal vectors, which leads to the following concept.

Definition 4 (SE face/facet, closed under Markov equivalence)
We will name a face F of F *score equivalent* (SE) if there exists an SE objective $o \in \mathbb{R}^{\Upsilon}$ and a constant $u \in \mathbb{R}$ such that two conditions from Definition 1 hold for $\mathsf{P} = \mathsf{F}$. An *SE facet* is a facet of F which is a score equivalent face.

A related concept is the next one: a set $\mathsf{S} \subseteq \mathsf{DAGS}\,(N)$ of directed acyclic graphs is *closed under Markov equivalence* if

$$\forall\, G, H \in \mathsf{DAGS}\,(N) \quad G \sim H \qquad G \in \mathsf{S} \quad \Rightarrow \quad H \in \mathsf{S}. \tag{9}$$

Remark 1 Note that an objective determining a face is not uniquely determined. Only in the case of a facet (of a full-dimensional polytope), is it unique up to a positive multiple. Therefore, one has to be careful when testing score equivalence of a face F which is not a facet, because one of the face-defining objectives for F could be SE and another objective for F need not be. Our definition requires the *existence* of at least one SE objective defining the face.

The following observation is straightforward.

Lemma 4 A set of graphs on an SE face is closed under Markov equivalence.

Proof Given an SE objective o with $F = \{\, \eta \in \mathsf{F} \,:\, \langle o, \eta \rangle = u \}$ for some $u \in \mathbb{R}$ and $G \in \mathsf{DAGS}\,(N)$ with $\langle o, \eta_G \rangle = u$, (8) implies for $H \sim G$ that $\langle o, \eta_H \rangle = u$.

An open question is whether the converse is true.

Conjecture 1
Every face $\mathsf{S} \subseteq \mathsf{DAGS}\,(N)$ of F closed under Markov equivalence is an SE face.

We managed to confirm the conjecture for facets; see Theorem 1 in Section 7. The arguments there are slightly special and do not apply to general faces. However, we were able to verify Conjecture 1 for $n = |N| = 3$ by an exhaustive analysis. By means of a computer, we verified for $n = 4$ that every inclusion-submaximal face among those closed under Markov equivalence is already an SE face. Our computational attempts to find a counter-example for $n = 5$ have not been successful.

6 Characterization of SE objectives

Recall that to present the characterization of the linear space of SE objectives in an elegant way we use the extension conventions from Section 3.

Lemma 5 A vector $o \in \mathbb{R}^{\Upsilon}$ is an SE objective if and only if either of the following two conditions (a) and (b) holds. The two conditions are equivalent: the first one holds if and only if the second one does.

(a) For any $Z \subseteq N$ and $a, b \in N \setminus Z$, $a \neq b$ one has

$$o(b \,|\, \{a\} \cup Z) + o(a \,|\, Z) = o(a \,|\, \{b\} \cup Z) + o(b \,|\, Z)\,. \tag{10}$$

(b) There exists $m \in \mathbb{R}^{\Lambda}$ such that

$$o(a \,|\, B) = m(\{a\} \cup B) - m(B) \qquad \text{for any } a \in N,\ B \subseteq N \setminus \{a\}. \tag{11}$$

In particular, the dimension of the linear subspace of SE objectives is $2^n - n - 1$.

Proof The condition (8) for $o \in \mathbb{R}^{\Upsilon}$ means $\langle o, \eta_G - \eta_H \rangle = 0$ if $G, H \in \mathsf{DAGS}\,(N)$ are such that $G \sim H$. A well-known transformational characterization of Markov equivalence [5, Theorem 2] says that $G \sim H$ if and only if there exists a sequence $G = G_1, \ldots, G_m = H$, $m \geq 1$ in $\mathsf{DAGS}\,(N)$ such that, for $i = 1, \ldots, m-1$, the graph G_{i+1} is obtained from G_i by "covered arc reversal". This means that G_i has an arrow $a \to b$ with $pa_{G_i}(b) = \{a\} \cup pa_{G_i}(a)$ and G_{i+1} is obtained from G_i by replacing $a \to b$ in G_i by $b \to a$ in G_{i+1}; the remaining arrows are unchanged. In particular, $G_i \sim G_{i+1}$ and, provided $Z = pa_{G_i}(a)$ one has

$$\eta_{G_i} - \eta_{G_{i+1}} = I_{b \leftarrow \{a\} \cup Z} + I_{a \leftarrow Z} - I_{a \leftarrow \{b\} \cup Z} - I_{b \leftarrow Z}.$$

Hence, we easily derive that (8) holds for $o \in \mathbb{R}^{\Upsilon}$ iff (10) holds.

It remains to show that (10) is equivalent to the existence of $m \in \mathbb{R}^{\Lambda}$ such that o is given by (11). The sufficiency of (11) is easy: then both the LHS and the RHS in (10) have the form $m(\{a,b\} \cup Z) - m(Z)$.

The necessity of (11) can be shown by an inductive construction. Take $Z = \emptyset$ in (10) and get $o(b \,|\, \{a\}) = o(a \,|\, \{b\})$. One can put $m(\{a,b\}) := o(b \,|\, \{a\})$ for any pair of distinct $a, b \in N$. Thus, owing to the above conventions, (11) holds in case $|B| \leq 1$. To confirm (11) for B with $|B| = r \geq 2$ accept the inductive hypothesis that it holds for B' with $|B'| \leq r - 1$. The task is to define $m(D)$ for $D \subseteq N$ with $|D| = r + 1$ so that (11) holds for B with $|B| \leq r$. Having fixed such a set D, for any pair of distinct elements $a, b \in D$ put $Z = D \setminus \{a, b\}$ and observe from (10) by means of the inductive premise:

$$o(b \,|\, \{a\} \cup Z) + m(\{a\} \cup Z) - m(Z) = o(a \,|\, \{b\} \cup Z) + m(\{b\} \cup Z) - m(Z)\,.$$

The cancellation of $m(Z)$ implies the function $b \mapsto o(b \,|\, D \setminus \{b\}) + m(D \setminus \{b\})$ for $b \in D$ is constant on D. Thus, one can put $m(D) := o(b \,|\, D \setminus \{b\}) + m(D \setminus \{b\})$ for any such $b \in D$, which verifies the inductive step.

The correspondence between o and m in (11) is evidently a one-to-one linear mapping, which implies the claim about the dimension.

Corollary 3 Let $o \in \mathbb{R}^{\Upsilon}$ be an SE objective and let $m \in \mathbb{R}^{\Lambda}$ satisfy (11). Then for any $T \in \Lambda$ and arbitrary $b \in T$ with $R := T \setminus \{b\}$ one has

$$\sum_{\emptyset \neq K \subseteq R} (-1)^{|R \setminus K|} \cdot o(b \,|\, K) = \sum_{L \in \Lambda :\, L \subseteq T} (-1)^{|T \setminus L|} \cdot m(L)\,. \tag{12}$$

In particular, the LHS of (12) does not depend on the choice of $b \in T$.

Proof Having in mind the extension conventions from Section 3 write by (11):

$$\begin{aligned}
\sum_{\emptyset\neq K\subseteq R} (-1)^{|R\setminus K|}\cdot o(b\,|\,K) &= (-1)^{|R|}\cdot \sum_{K\subseteq R} (-1)^{|K|}\cdot o(b\,|\,K) \\
&\overset{(11)}{=} (-1)^{|R|}\cdot \sum_{K\subseteq R} (-1)^{|K|}\cdot \{\, m(\{b\}\cup K) - m(K)\,\} \\
&= (-1)^{|R|}\cdot (-1)\cdot \sum_{K\subseteq R} (-1)^{|K|+1}\cdot m(\{b\}\cup K) \\
&\qquad + (-1)^{|R|}\cdot (-1)\cdot \sum_{K\subseteq R} (-1)^{|K|}\cdot m(K) \\
&= (-1)^{|T|}\cdot \sum_{L\subseteq T} (-1)^{|L|}\cdot m(L) = \sum_{L\in\Lambda:\, L\subseteq T} (-1)^{|T\setminus L|}\cdot m(L)\,,
\end{aligned}$$

which concludes the proof.

Another relevant observation is the following.

Lemma 6 Any face of F containing the whole Markov equivalence class of complete graphs is given by an SE objective.

Proof Assume $\langle o,\eta\rangle \leq u$ is an arbitrary defining inequality for such a face F of F, with $o\in\mathbb{R}^{\Upsilon}$, $u\in\mathbb{R}$. By Lemma 5(a), it is enough to show o satisfies (10). Note that, for any $Z\subseteq N$ and distinct $a,b\in N\setminus Z$, complete graphs G and H over N exist with $\eta_G-\eta_H = I_{b\leftarrow\{a\}\cup Z}+I_{a\leftarrow Z}-I_{a\leftarrow\{b\}\cup Z}-I_{b\leftarrow Z}$. Hence, $\langle o,\eta_G\rangle = u = \langle o,\eta_H\rangle$ implies that (10) is true for that particular choice of nodes a,b and the set Z.

It follows from Lemma 6 that every face of F which contains the class of complete graphs is an SE face. Therefore, no counter-example to Conjecture 1 is among the faces containing a complete graph. Indeed, since they must be closed under Markov equivalence, they necessarily contain the whole set of complete graphs.

7 Correspondence to supermodular functions

In this section we characterize those facets of F which contain the set of complete graphs. We show they coincide with SE facets and establish their relation to extreme supermodular functions.

The previous results allow us to confirm Conjecture 1 for facets.

Theorem 1 *The following conditions are equivalent for a facet* $\mathsf{S}\subseteq \mathsf{DAGS}\,(N)$*:*

(a) S *is closed under Markov equivalence,*
(b) S *contains the whole equivalence class of complete graphs,*
(c) S *is SE.*

Proof To show (a)⇒(b) note, by Lemma 2(iii), that S cannot contain the empty graph, since otherwise it is not closed under Markov equivalence. Clearly, S must be non-empty, because otherwise it is not a facet of F (note we assume $n = |N| \geq 2$). Thus, $G \in \mathsf{S}$ exists and one can construct a complete graph $H \in \mathsf{DAGS}\,(N)$ such that G is a subgraph of H. By Corollary 2, $H \in \mathsf{S}$. Since S is closed under Markov equivalence, all complete graphs belong to S. The implication (b)⇒(c) follows from Lemma 6. The implication (c)⇒(a) was mentioned as Lemma 4 in Section 5.

The next step is to recall the definition of a supermodular set function.

Definition 5 (standardized supermodular function)
Any vector $m \in \mathbb{R}^{\mathcal{P}(N)}$ can be viewed as a real set function $m : \mathcal{P}(N) \to \mathbb{R}$. Such a set function will be called *standardized* if $m(S) = 0$ for $S \subseteq N$, $|S| \leq 1$, and *supermodular* if

$$\forall\, U, V \subseteq N \quad m(U) + m(V) \leq m(U \cup V) + m(U \cap V)\,. \tag{13}$$

The following (non-negative) characteristics are ascribed to any supermodular function m: for any $a, b \in N$, $a \neq b$ and $Z \subseteq N \setminus \{a, b\}$, we will denote

$$\varDelta m(a, b \,|\, Z) \,:=\, m(\{a, b\} \cup Z) + m(Z) - m(\{a\} \cup Z) - m(\{b\} \cup Z)\,.$$

It is easy to see that a set function m is supermodular iff $\varDelta m(a, b \,|\, Z) \geq 0$ for any respective triplet $(a, b \,|\, Z)$; see, for example, [25, Theorem 24(iv)]. The point is that standardized supermodular functions correspond to valid inequalities for the family-variable polytope that are tight at all complete graphs.

Lemma 7 An inequality $\langle o, \eta \rangle \leq u$, where $o \in \mathbb{R}^{\Upsilon}$ and $u \in \mathbb{R}$, is valid for all $\eta \in \mathsf{F}$ and tight at any complete graph over N iff it corresponds to a standardized supermodular function m in the sense:

– o is given by (11): $o(a \,|\, B) = m(\{a\} \cup B) - m(B)$ for $a \in N$, $B \subseteq N \setminus \{a\}$,
– u is the shared value $\langle o, \eta_H \rangle$ for complete graphs H over N.

Moreover, the correspondence is one-to-one and preserves a conic combination.

Proof Given such an inequality, Lemma 6 implies that o is an SE objective and Lemma 5(b) says it has the form (11). To show that m is necessarily supermodular observe $\varDelta m(a, b \,|\, Z) \geq 0$ for any $(a, b \,|\, Z)$. To this end, note that, given a triplet $(a, b \,|\, Z)$, a complete graph H over N and $G \in \mathsf{DAGS}\,(N)$ exist such that $\eta_H - \eta_G = I_{b \leftarrow \{a\} \cup Z} - I_{b \leftarrow Z}$. Indeed, consider a total order of elements in N in which Z precedes a after which b and $N \setminus (\{a, b\} \cup Z)$ follow and take H as the complete graph consonant with this order and G is the graph obtained from H by the removal of the arrow $a \to b$. Hence, $\langle o, \eta_G \rangle \leq u = \langle o, \eta_H \rangle$ implies $0 \leq \langle o, \eta_H - \eta_G \rangle = o(b \,|\, \{a\} \cup Z) - o(b \,|\, Z) \stackrel{(11)}{=} \varDelta m(a, b \,|\, Z)$.

Conversely, given a supermodular m, Lemma 5(b) says the objective o given by (11) is SE and the complete graphs H over N share the value $\langle o, \eta_H \rangle$.

Thus, it is enough to show that, for any $G \in \mathsf{DAGS}\,(N)$, a complete graph H exists with $\langle o, \eta_G \rangle \leq \langle o, \eta_H \rangle$. Indeed, consider a total order consonant with G, denote by $pre(a)$ the set of (strict) predecessors of $a \in N$ in that order and by H the complete graph consonant with the order. Then write

$$\langle o, \eta_H - \eta_G \rangle = \sum_{a \in N} \{\, o(a \,|\, pre(a)) - o(a \,|\, pa_G(a)) \,\} =$$

$$= \sum_{a \in N} \{\, \underbrace{m(\{a\} \cup pre(a)) - m(\{a\} \cup pa_G(a)) - m(pre(a)) + m(pa_G(a))}_{\geq 0} \,\} \geq 0\,.$$

Since (11) defines an invertible linear transformation, the last claim is easy.

By the extension convention from Section 3, any vector in $m \in \mathbb{R}^{\Lambda}$ could be identified with a standardized set function. With a small abuse of terminology, we say that $m \in \mathbb{R}^{\Lambda}$ is supermodular if its zero extension $m : \mathcal{P}(N) \to \mathbb{R}$ is a supermodular set function. By its definition, the set of supermodular vectors in $\mathbb{R}^{\Lambda}$ is a polyhedral cone. Since it is pointed, it has finitely many extreme rays. This motivates the next definition.

Definition 6 (extreme supermodular function)
A standardized supermodular set function $m : \mathcal{P}(N) \to \mathbb{R}$ is called *extreme* if it generates an extreme ray of the standardized supermodular cone.

The following fact follows from a specific characterization of extremality of supermodular functions.

Lemma 8 Let $m_1, m_2 \in \mathbb{R}^{\mathcal{P}(N)}$ generate distinct extreme rays of the standardized supermodular cone. Then the faces of F determined by the corresponding inequalities, as described in Lemma 7, are inclusion-incomparable.

Proof The argument is based on the result saying that a supermodular set function m is extreme iff the structural independence model produced by m is sub-maximal; see [19, Lemma 5.6] or [25, Corollary 30]. More specifically, it says m is extreme iff any supermodular function m' with

$$\forall\,(a,b\,|\,Z) \quad \varDelta m(a,b\,|\,Z) = 0 \quad \Rightarrow \quad \varDelta m'(a,b\,|\,Z) = 0$$

either satisfies, for any triplet $(a,b\,|\,Z)$, $\varDelta m'(a,b\,|\,Z) = 0 \;\Leftrightarrow\; \varDelta m(a,b\,|\,Z) = 0$ or even, $\varDelta m'(a',b'\,|\,Z') = 0$ for any $(a',b'\,|\,Z')$, which, for a standardized m', means that m' must be a non-negative multiple of m. Since m_1, m_2 generate distinct rays, a triplet $(a,b\,|\,Z)$ must exist such that $\varDelta m_1(a,b\,|\,Z) > 0$ and $\varDelta m_2(a,b\,|\,Z) = 0$. As in the proof of Lemma 7, construct a complete graph H over N and $G \in \mathsf{DAGS}\,(N)$ with $\eta_H - \eta_G = I_{b \leftarrow \{a\} \cup Z} - I_{b \leftarrow Z}$. Then $\varDelta m_1(a,b\,|\,Z) > 0$ implies that the inequality $\langle o_1, \eta \rangle \leq u_1$ given by m_1 through (11) is not tight for η_G because

$$u_1 - \langle o_1, \eta_G \rangle = \langle o_1, \eta_H - \eta_G \rangle = o_1(b\,|\,\{a\} \cup Z) - o_1(b\,|\,Z) \stackrel{(11)}{=} \varDelta m_1(a,b\,|\,Z) > 0\,,$$

while $\varDelta m_2(a,b\,|\,Z) = 0$ implies that $\langle o_2, \eta \rangle \leq u_2$ is tight for η_G. Hence, the face of F determined by m_2 is not contained is the one determined by m_1. The role of generators m_1 and m_2 is clearly exchangeable.

Now, we are ready to characterize SE facets.

Theorem 2 *An inequality* $\langle o, \eta\rangle \le u$ *for* $\eta \in \mathsf{F}$, *where* $o \in \mathbb{R}^{\Upsilon}$ *and* $u \in \mathbb{R}$, *defines an SE facet for* F *iff there exists an extreme standardized supermodular set function* m *such that* o *is determined by* (11) *and* $u = m(N)$.

Proof By Theorem 1(b), $\langle o, \eta\rangle \le u$ defines an SE facet iff it is facet-defining for F and tight at all complete graphs over N. We show this occurs iff an extreme standardized supermodular set function m exists such that o is determined by (11) and u is the shared value of $\langle o, \eta_H\rangle$ for complete graphs H over N. Of course, by (11), one has $\langle o, \eta_H\rangle = m(N) - m(\emptyset) = m(N)$.

Firstly, using Lemma 1, we show that any extreme standardized supermodular function m_i gives a facet of F. Thus, assume F' is a face containing the face F_i determined by m_i. Lemma 7 applied to m_i says that the face F_i contains the class of complete graphs, and so F' does. Again by Lemma 7 applied to the inequality defining F', the face F' is given by a supermodular function m', which must be a conic combination of finitely many generators of (all) the extreme rays: $m' = \sum_j \alpha_j \cdot m_j$, $\alpha_j \ge 0$.

The assumption $F_i \subseteq F'$ implies that, for any $k \neq i$, the coefficient α_k must vanish. Specifically, one can derive from the last claim in Lemma 7 that $\alpha_k > 0$ forces $F' \subseteq F_k$. Indeed, the inequality defining F' is a conic combination of the inequality corresponding to m_k (= defining F_k) and the inequality corresponding to $\sum_{j\neq k} \alpha_j \cdot m_j$. Any vector $\eta \in F'$ is tight for the (conic) combination of these two inequalities, and, since both these inequalities are valid for F, η must be tight for both of them (this is the same consideration as in the proof of Corollary 1). Thus, η is tight for the inequality corresponding to m_k, that is, $\eta \in F_k$. Thus, when $\alpha_k > 0$ one has $F_i \subseteq F' \subseteq F_k$.

However, for distinct i and k, the respective faces F_i and F_k are inclusion-incomparable, by Lemma 8. Thus, one has $m' = \alpha_i \cdot m_i$, which means either $m' = 0$, in which case $F' = \mathsf{F}$, or m' is a positive multiple of m_i, in which case $F' = F_i$. Hence, by Lemma 1, F_i is a facet of F.

Secondly, we show that any facet F of F involving all complete graphs is given by an extreme standardized supermodular function. Apply Lemma 7 to F and write the respective standardized supermodular function m as a conic combination $m = \sum_j \alpha_j \cdot m_j$, $\alpha_j \ge 0$ of extreme ones. Let us assume for a contradiction that $\alpha_i \neq 0 \neq \alpha_k$ for distinct i and k. By Lemma 8, the faces corresponding to m_i and m_k are incomparable. In particular, provided $\langle o_j, \eta\rangle \le u_j$ denotes the inequality for $\eta \in \mathsf{F}$ corresponding m_j, we know that $w_1 \in \mathsf{F}$ exists satisfying $\langle o_i, w_1\rangle = u_i$ and $\langle o_k, w_1\rangle < u_k$, and $w_2 \in \mathsf{F}$ exists satisfying $\langle o_i, w_2\rangle < u_i$. The inequality corresponding to m is the sum of $\sum_{j\neq i} \alpha_j \cdot \langle o_j, \eta\rangle \le \sum_{j\neq i} \alpha_j \cdot u_j$ and of the α_i-multiple of the inequality $\langle o_i, \eta\rangle \le u_i$. The assumption (4) of Corollary 1 is fulfilled for the vectors w_1 and w_2 above, which gives a contradictory conclusion that F is not a facet. Thus, at most one of the coefficients α_j is non-zero. Since m must be non-zero, it is a positive multiple of some m_j.

Thus, Theorem 2 transforms the problem of testing certain facets of F into the task of verifying whether the respective supermodular function is extreme.

Note that a simple linear criterion for testing extremality of a standardized supermodular function m has recently been proposed in [25]. The criterion consists in solving a linear equation system determined by the combinatorial structure of the so-called *core polytope* ascribed to m:

$$\mathcal{C}(m) := \{\, [v_a]_{a\in N} \in \mathbb{R}^N \ : \ \sum_{a\in N} v_a = m(N) \ \ \& \ \forall\, S\subseteq N \ \sum_{a\in S} v_a \geq m(S)\,\}\,.$$

We hope that the criterion from [25] will appear to be useful in our context.

8 Generalized cluster inequalities and uniform matroids

An important class of inequalities for the family-variable polytope is discussed in this section. We apply Theorem 2 from the previous section to show they define SE facets and reveal their hidden connection to uniform matroids.

Jaakkola, Sontag, Globerson and Meila introduced in [13] an interesting class of cluster-based inequalities for F, whose purpose was to express the acyclicity restrictions. To shorten the terminology we call them the *cluster inequalities.* Specifically, if the family-variable vector η_G encoding $G \in \mathsf{DAGS}\,(N)$ is extended by additional components for the empty parent sets $\eta_G(a\,|\,\emptyset)$, $a\in N$, then the inequality ascribed to a cluster $C\subseteq N$, $|C|\geq 2$, has the form

$$1 \leq \sum_{a\in C} \ \sum_{B\subseteq N\,:\,B\cap C=\emptyset} \eta_G(a\,|\,B)\,.$$

The interpretation is clear: since the induced subgraph G_C is acyclic, there is at least one node a in C which has no parent in C. An important fact is that the only integral vectors in the polyhedron specified by the cluster inequalities, and, for any $a\in N$, by the *convexity constraints* $\eta_G(a\,|\,B)\geq 0$, $B\subseteq N\setminus\{a\}$ and $\sum_{B\subseteq N\setminus\{a\}} \eta_G(a\,|\,B) = 1$, are the DAG-codes [22, Lemma 2].

The cluster inequalities have appeared to have a crucial role in the integer linear programming (ILP) approach learning BN structure. This was confirmed computationally in [8] by the first author of this paper, who also introduced *generalized cluster inequalities.* Specifically, to every cluster $C\subseteq N$, $|C|\geq 2$, and $k=1,\ldots,|C|-1$ one can ascribe the inequality

$$k \leq \sum_{a\in C} \ \sum_{B\subseteq N\setminus\{a\}\,:\,|B\cap C|<k} \eta_G(a\,|\,B)\,.$$

Its interpretation is analogous: since the induced subgraph G_C is acyclic, the first k nodes in a total order of nodes in C consonant with G_C have at most $k-1$ parents in C. Note that for $k=|C|$ and $k=0$ the inequalities are tight at any $G\in\mathsf{DAGS}\,(N)$ and are, therefore, omitted. In particular, we only consider the generalized cluster inequalities for $k=1,\ldots,|C|-1$; this also enforces $|C|\geq 2$. To transform them into standardized inequality constraints on a vector η in $\mathsf{F} \subseteq \mathbb{R}^{\Upsilon}$ we use the above convexity equality constraints and get for any $C \subseteq N$, $|C| \geq 2$, and $k = 1, \ldots, |C| - 1$,

$$\sum_{a \in C} \; \sum_{B \subseteq N \setminus \{a\} \,:\, |B \cap C| \geq k} \eta(a \,|\, B) \;\leq\; |C| - k\,. \tag{14}$$

The point is that this k-cluster inequality (14) corresponds to an extreme standardized supermodular set function in sense of Theorem 2.

Lemma 9 For any $C \subseteq N$, $|C| \geq 2$, and $k = 1, \ldots, |C| - 1$, the formula

$$m_{C,k}(S) = \max\{\, 0, |S \cap C| - k \,\} \quad \text{for any} \quad S \subseteq N, \tag{15}$$

gives an extreme standardized supermodular function which determines through the formula (11) the objective coefficients in (14).

Proof Easily, the objective coefficient for $(a \,|\, B) \in \Upsilon$ is

$$o_{C,k}(a \,|\, B) \stackrel{(11)}{=} m_{C,k}(\{a\} \cup B) - m_{C,k}(B) = \begin{cases} 1 & \text{if } a \in C \text{ and } |B \cap C| \geq k, \\ 0 & \text{otherwise,} \end{cases}$$

and the value of $\langle o_{C,k}, \eta_H \rangle$ for any complete graph H over N is $|C| - k$. Hence, (15) determines through (11) the inequality (14).

It remains to show that $m_{C,k}$ generates an extreme ray of the cone K of standardized supermodular functions. Recall m is supermodular iff, for any triplet $A, B, Z \subseteq N$ of pairwise disjoint sets, one has

$$\Delta m(A, B \,|\, Z) \;:=\; m(A \cup B \cup Z) + m(Z) - m(A \cup Z) - m(B \cup Z) \geq 0\,,$$

which is a re-formulation of (13), but it is enough to verify $\Delta m(a, b \,|\, Z) \geq 0$ for any $a, b \in N$, $a \neq b$ and $Z \subseteq N \setminus \{a, b\}$. It is easy to observe $m(S) \geq 0$ for any $m \in \mathsf{K}$ and $S \subseteq N$. Since $m_{C,k}(S) = m_{C,k}(S \cap C)$ for any $S \subseteq N$, one has

$$\Delta m_{C,k}(A, B \,|\, Z) = \Delta m_{C,k}(A \cap C, B \cap C \,|\, Z \cap C) \quad \text{for disjoint } A, B, Z \subseteq N.$$

To show $m_{C,k} \in \mathsf{K}$ observe that, for any triplet $(a, b \,|\, Z)$ with $\{a, b\} \cup Z \subseteq C$,

$$\begin{array}{ll} \Delta m_{C,k}(a, b \,|\, Z) = 1 & \text{if } |\{a, b\} \cup Z| = k + 1, \text{ and} \\ \Delta m_{C,k}(a, b \,|\, Z) = 0 & \text{otherwise.} \end{array}$$

We have to verify that, if $m_{C,k} = \alpha \cdot m_1 + (1 - \alpha) \cdot m_2$, $\alpha \in (0, 1)$ is a non-trivial convex combination of $m_1, m_2 \in \mathsf{K}$ then m_1 and m_2 are non-negative multiples of $m_{C,k}$. To show $m = \gamma \cdot m_{C,k}$ for some $\gamma \geq 0$ it is enough to verify:

(i) $m(S) = 0$ for $S \subseteq N$ with $|S \cap C| \leq k$,
(ii) $m(S) = m(S \cap C)$ for any $S \subseteq N$,
(iii) $m(S) = m(T)$ for $S, T \subseteq C$, $|S| = |T| = k + 1$,
(iv) if γ is the shared value from (iii) then $m(S) = \gamma + m(R)$ for any pair of sets $R \subseteq S \subseteq C$ are such that $|S| = |R| + 1 \geq k + 1$.

To verify (i) for m_1, m_2 with some such $S \subseteq N$ write

$$0 = m_{C,k}(S) = \alpha \cdot m_1(S) + (1-\alpha) \cdot m_2(S)\,.$$

The RHS here is a convex combination of non-negative terms; therefore, they both vanish, which means $0 = m_1(S) = m_2(S)$. To verify (ii) for m_1, m_2 with some $S \subseteq N$ consider $(A, B \,|\, Z) = (S \cap C, S \setminus C \,|\, \emptyset)$ and observe

$$0 = \Delta m_{C,k}(A, B \,|\, Z) = \alpha \cdot \Delta m_1(A, B \,|\, Z) + (1-\alpha) \cdot \Delta m_2(A, B \,|\, Z)\,.$$

Hence, for $i = 1, 2$, $\Delta m_i(A, B \,|\, Z) = 0$, implying together with (i) for m_i that $m_i(S) = m_i(S \cap C)$. To verify (iii) it is enough to observe $m_i(S) = m_i(T)$ in the case $|S| = |T| = k+1$ with $S \setminus T = \{s\}$ and $T \setminus S = \{t\}$. Choose $r \in S \cap T$, put $R = (S \cap T) \setminus \{r\}$ and consider the triplets $(r, t \,|\, R \cup \{s\})$ and $(r, s \,|\, R \cup \{t\})$. Since both $0 = \Delta m_{C,k}(r, t \,|\, R \cup \{s\})$ and $0 = \Delta m_{C,k}(r, s \,|\, R \cup \{t\})$, one has $0 = \Delta m_i(r, t \,|\, R \cup \{s\}) = \Delta m_i(r, s \,|\, R \cup \{t\})$, for $i = 1, 2$. Hence, by (i) for m_i,

$$0 = \Delta m_i(r, t \,|\, R \cup \{s\}) - \Delta m_i(r, s \,|\, R \cup \{t\}) = m_i(T) - m_i(S)\,.$$

The condition (iv) can be verified by induction on $|S|$: (i) and (iii) for m_i say (iv) holds for m_i and $|S| = k+1$. If $|S| > k+1$ and $S \setminus R = \{s\}$ then choose $t \in R$ and put $T = S \setminus \{t\}$. Because $0 = \Delta m_{C,k}(s, t \,|\, R \cap T)$ one gets $0 = \Delta m_i(s, t \,|\, R \cap T)$, that is, $m_i(S) - m_i(R) = m_i(T) - m_i(R \cap T) = \gamma$ for $i = 1, 2$, by the inductive assumption.

Corollary 4 Any generalized cluster inequality (14) defines an SE facet of F.

Proof Combine Lemma 9 with Theorem 2.

The rest of this section is an observation which makes sense for a reader familiar with elementary notions in matroid theory. Thus, we assume the reader knows basic equivalent definitions of a matroid in terms of independent sets, bases and the rank function, as given, for example, in [17, Chapter 1].

The link between generalized cluster inequalities and certain matroids is based on a duality relationship of supermodular functions and their mirror images, submodular functions. Recall that $r \in \mathbb{R}^{\mathcal{P}(N)}$ is *submodular* if

$$r(U \cup V) + r(U \cap V) \leq r(U) + r(V) \qquad \text{for any } U, V \subseteq N\,.$$

In fact, there is a one-to-one linear mapping from the cone K of standardized supermodular functions onto the cone of submodular functions $r : \mathcal{P}(N) \to \mathbb{R}$ satisfying $r(\emptyset) = 0$ and $r(N) = r(N \setminus \{a\})$ for any $a \in N$. The point is that the rank functions of non-degenerate matroids fall within this submodular cone. Specifically, one can consider the *duality transformation* which ascribes to any $m \in \mathsf{K}$ the set function r given by

$$r(T) = m(N) - m(N \setminus T) \qquad \text{for any } T \subseteq N.$$

This self-inverse transformation maps the supermodular function $m_{C,k}$ for $C \subseteq N$, $|C| \geq 2$, and $k = 1, \ldots, |C|-1$, onto the submodular function

$$r_{C,k}(T) = \min\{\, |T \cap C|, |C| - k \,\} \quad \text{for any} \quad T \subseteq N, \tag{16}$$

which is the rank function of a matroid on N. However, it can be viewed as a kind of trivial "loop-adding" extension of a matroid which has C as its ground set. Indeed, the function (16) can be identified with its restriction to $\mathcal{P}(C)$, which is the rank function of the uniform matroid of rank $|C| - k$ on C; see [17, Example 1.2.7]. The bases of this matroid are just the subsets of C of the cardinality $|C| - k$. Two remaining uniform matroids on C, namely those of the ranks 0 and $|C|$, differ in the property they are not connected: that means a set $\emptyset \subset S \subset C$ exists with $r(C) = r(S) + r(C \setminus S)$, where r is their rank function; see [17, § 4.2] for this concept. Therefore, one can summarize our observation by saying that the generalized cluster inequalities for $C \subseteq N$ are in a one-to-one correspondence with *connected uniform matroids* on C.

Remark 2 Note that the duality transformation is not the only one-to-one linear mapping between the considered supermodular and submodular cones; see [25, § 7.2] for the details. However, this fact is not important in our context since the use of the other transformation leads to the same conclusion, the difference is that the uniform matroid on C of the rank k is ascribed to $m_{C,k}$ instead. On the other hand, the duality transformation has the property that the vertices of the core polytope ascribed to $m_{C,k}$, as defined in the end of Section 7, are just the incidence vectors for bases of the uniform matroid of rank $|C| - k$.

Note that the correspondence of generalized cluster inequalities and connected uniform matroids can be extended. It has been recently shown in [24] on the basis of results of the present paper and some classic results from matroid theory that *any connected* matroid which has $C \subseteq N$, $|C| \geq 2$, as its ground set induces an SE facet of F.

9 On the faces of the characteristic-imset polytope

In this section, we introduce a one-to-one correspondence between faces of the characteristic-imset polytope C and SE faces of the family-variable polytope F. This allows us to characterize those faces of C that correspond to SE facets.

Let $\langle z, \mathsf{c} \rangle_\Lambda \leq u$, where $z \in \mathbb{R}^\Lambda$ and $u \in \mathbb{R}$, be a valid inequality for c in the characteristic-imset polytope C. It defines a face of C:

$$\bar{F} = \{\, \mathsf{c} \in \mathsf{C} \;:\; \langle z, \mathsf{c} \rangle_\Lambda = u \,\}.$$

By substituting (1) into the inequality $\langle z, \mathsf{c}_\eta \rangle_\Lambda \leq u$ and re-arranging terms after the components of η one gets an inequality for $\eta \in \mathbb{R}^\Upsilon$ valid for any η_G, $G \in \mathsf{DAGS}\,(N)$. Indeed, this is because the image of η_G by (1) is just c_G. Moreover, the objective on the LHS of the obtained inequality is SE because whenever $G \sim H$, one has $\mathsf{c}_G = \mathsf{c}_G$ and, therefore, $\langle z, \mathsf{c}_G \rangle_\Lambda = \langle z, \mathsf{c}_H \rangle_\Lambda$. Thus, any face of C defines an SE face of F. Nevertheless, the converse is true.

Lemma 10 Given an SE objective $o \in \mathbb{R}^\Upsilon$, there exists unique $z_o \in \mathbb{R}^\Lambda$ such that the following holds:

$$\forall\, \eta \in \mathbb{R}^\Upsilon \qquad \langle o, \eta \rangle_\Upsilon = \langle z_o, \mathsf{c}_\eta \rangle_\Lambda \,. \tag{17}$$

Specifically, one has

$$z_o(T) := \sum_{\emptyset \neq K \subseteq R} (-1)^{|R \setminus K|} \cdot o(b \,|\, K) \tag{18}$$
$$\text{for } T \in \Lambda, \text{ with any } b \in T \text{ and } R := T \setminus \{b\}.$$

In particular, the expression in (18) does not depend on the choice of $b \in T$.

Proof We are going to show that $z_o \in \mathbb{R}^\Lambda$ given by (18) satisfies

$$\forall\, G \in \mathsf{DAGS}\,(N) \qquad \langle o, \eta_G \rangle_\Upsilon = \langle z_o, \mathsf{c}_G \rangle_\Lambda \,. \tag{19}$$

By Corollary 3, we know that z_o takes the form

$$z_o(T) \overset{(12)}{=} \sum_{L \in \Lambda \,:\, L \subseteq T} (-1)^{|T \setminus L|} \cdot m(L) \quad \text{for } T \in \Lambda, \tag{20}$$
$$\text{where } m \in \mathbb{R}^\Lambda \text{ given by (11).}$$

The next step is to note that (20) is equivalent to the relation

$$m(S) = \sum_{T \in \Lambda \,:\, T \subseteq S} z_o(T) \quad \text{for any } S \in \Lambda, \tag{21}$$

which can be verified by substituting (20) into the RHS of (21). To verify (19) substitute (11) into the expression for $\langle o, \eta_G \rangle_\Upsilon$, then use the definitions of η_G and that of the standard imset u_G (see Section 2):

$$\begin{aligned}
\langle o, \eta_G \rangle_\Upsilon &\overset{(11)}{=} \sum_{(a\,|\,B) \in \Upsilon} \{\, m(\{a\} \cup B) - m(B) \,\} \cdot \eta_G(a \,|\, B) \\
&= \sum_{\emptyset \neq S \subseteq N} m(S) \cdot \left\{ \sum_{(a\,|\,B)} \eta_G(a \,|\, B) \cdot \delta_S(\{a\} \cup B) - \sum_{(a\,|\,B)} \eta_G(a \,|\, B) \cdot \delta_S(B) \right\} \\
&= \sum_{\emptyset \neq S \subseteq N} m(S) \cdot \left\{ \sum_{a \in N} \delta_S(\{a\} \cup \mathrm{pa}_G(a)) - \sum_{a \in N} \delta_S(\mathrm{pa}_G(a)) \right\} \\
&= \sum_{\emptyset \neq S \subseteq N} m(S) \cdot \sum_{a \in N} \{\, \delta_{\{a\} \cup \mathrm{pa}_G(a)}(S) - \delta_{\mathrm{pa}_G(a)}(S) \,\} \\
&\overset{(2)}{=} \sum_{\emptyset \neq S \subseteq N} m(S) \cdot \{ \delta_N(S) - \mathsf{u}_G(S) \} \,.
\end{aligned}$$

Further, we substitute the relation (21) into the above expression and get this:

$$
\begin{aligned}
\langle o, \eta_G \rangle_\Upsilon &= \sum_{\emptyset \neq S \subseteq N} m(S) \cdot \{\delta_N(S) - \mathsf{u}_G(S)\} = \sum_{S \in \Lambda} m(S) \cdot \{\delta_N(S) - \mathsf{u}_G(S)\} \\
&\overset{(21)}{=} \sum_{S \in \Lambda} \sum_{T \in \Lambda \,:\, T \subseteq S} z_o(T) \cdot \{\delta_N(S) - \mathsf{u}_G(S)\} \\
&= \sum_{T \in \Lambda} z_o(T) \cdot \sum_{S \colon T \subseteq S \subseteq N} \{\delta_N(S) - \mathsf{u}_G(S)\} \\
&= \sum_{T \in \Lambda} z_o(T) \cdot \{1 - \sum_{S \colon T \subseteq S \subseteq N} \mathsf{u}_G(S)\} \overset{(3)}{=} \sum_{T \in \Lambda} z_o(T) \cdot \mathsf{c}_G(T) = \langle z_o, \mathsf{c}_G \rangle_\Lambda \,.
\end{aligned}
$$

Since the codes η_G for $G \in \mathsf{DAGS}\,(N)$ linearly span $\mathbb{R}^\Upsilon$ the relation (19) implies (17). The uniqueness of the vector z_o in the formula (17) is easy because the codes c_G for $G \in \mathsf{DAGS}\,(N)$ span $\mathbb{R}^\Lambda$.

Every face of the characteristic-imset polytope C can be identified with a set of directed acyclic graphs closed under Markov equivalence:

$$
\bar{F} \subseteq \mathsf{C} \;\text{ a face of } \mathsf{C} \quad \longleftrightarrow \quad \mathsf{S} = \{G \in \mathsf{DAGS}\,(N) : \; \mathsf{c}_G \in \bar{F}\}\,.
$$

Indeed, the arguments given above Definition 2 are also valid for $\mathsf{P} = \mathsf{C}$ and, since the vertices of C are just the characteristic imsets, its faces can be viewed as sets of characteristic imsets. These, however, correspond to equivalence classes of graphs over N. Thus, every face of C can be identified with a set of such graphs, namely with the union of the respective equivalence classes. These are just the graphs whose characteristic imsets belong to the face. It is easy to see that the correspondence preserves inclusion: $\bar{F}_1 \subseteq \bar{F}_2$ for faces of C iff $\mathsf{S}_1 \subseteq \mathsf{S}_2$ for the corresponding sets of graphs $\mathsf{S}_i \subseteq \mathsf{DAGS}\,(N)$.

Corollary 5 There is a one-to-one correspondence between SE faces of F and faces of C which preserves inclusion: given SE faces F_1, F_2 of F and the corresponding faces $\bar{F}_1, \bar{F}_2$ of C one has $F_1 \subseteq F_2$ if and only if $\bar{F}_1 \subseteq \bar{F}_2$. Specifically, the SE face of F given by an inequality $\langle o, \eta \rangle_\Upsilon \leq u$ corresponds to the face of C given the inequality $\langle z_o, \mathsf{c} \rangle_\Lambda \leq u$. This correspondence has the property that the sets of graphs identified with the faces coincide.

Proof It is easy to see that $\langle o, \eta \rangle_\Upsilon \leq u$ is valid for $\eta \in \mathsf{F}$ iff $\langle z_o, \mathsf{c} \rangle_\Lambda \leq u$ is valid for $\mathsf{c} \in \mathsf{C}$. Moreover, by Lemma 10, the set of $G \in \mathsf{DAGS}\,(N)$ such that $\langle o, \eta \rangle_\Upsilon \leq u$ is tight for η_G coincides with the set of $G \in \mathsf{DAGS}\,(N)$ such that $\langle z_o, \mathsf{c} \rangle_\Lambda \leq u$ is tight for c_G. Thus, an SE face of F and the corresponding face of C have the same sets of "belonging" graphs. This observation easily implies the claim about preserving the inclusion of faces.

There are two distinguished vertices of the characteristic imset polytope C. One of them is the 0-imset, the zero vector in $\mathbb{R}^\Lambda$, which is the characteristic imset of the empty graph over N. The other one is the 1-imset, a vector in $\mathbb{R}^\Lambda$ whose all components are ones, which is the characteristic imset of any of the complete graphs over N. It plays a crucial role in the description of faces of C corresponding to SE facets of F.

Corollary 6 SE facets of the family-variable polytope F correspond to those facets of the characteristic-imset polytope C that contain the 1-imset. None of those facets of C include the 0-imset.

Proof Let $\bar{F}$ be a face of C corresponding to an SE facet F of F. We show, using Lemma 1, that $\bar{F}$ is a facet of C. For a face $\bar{F}'$ of C with $\bar{F} \subset \bar{F}'$ the respective SE face F' of F satisfies, by Corollary 5, $F \subset F'$. Thus, necessarily $F' = \mathsf{F}$, which implies $\bar{F}' = \mathsf{C}$. By Theorem 1, F contains the whole equivalence class of complete graphs and, by Corollary 5, $\bar{F}$ must contain the 1-imset.

Conversely, let F be an SE face of F which corresponds to a facet $\bar{F}$ of C containing the 1-imset. Using Lemma 1 observe that F is a facet of F. Indeed, since F contains the whole equivalence class of complete graphs, the same is the case for any face F' of F with $F \subset F'$. By Lemma 6, F' is SE; hence, it has the corresponding face $\bar{F}'$ of C. By Corollary 5 one has $\bar{F} \subset \bar{F}'$; therefore, $\bar{F}' = \mathsf{C}$, which implies $F' = \mathsf{F}$.

The last claim follows easily by contradiction: otherwise the corresponding SE facet contains the empty graph and, by Lemma 2(iii), it is determined by (6). But none of these facets of F is SE.

By combining Corollary 6 and Theorem 2 one observes that the facets of C containing the 1-imset correspond to extreme (standardized) supermodular functions. On the other hand, it follows from Corollaries 5 and 6 that the SE faces of F corresponding to facets of C not containing the 1-imset are *sub-maximal SE faces* with respect to inclusion, but *not SE facets*. That means, these are SE faces F of F such that there is no other SE face F' of F such that $F \subset F'$ except $F' = \mathsf{F}$ but F is not a facet of F since $\dim(F) < \dim(\mathsf{F}) - 1$. Example 3 in Section 10 shows what such sub-maximal SE faces look like.

To illustrate Corollary 5 we transform the generalized cluster inequalities (14) from Section 8 into the characteristic-imset frame. Specifically, having fixed a cluster $C \subseteq N$, $|C| \geq 2$, and $k \in \{1, \ldots, |C|-1\}$, the coefficients $z(S)$ for $S \in \Lambda$ in the transformed corresponding k-cluster inequality vanish outside subsets of C and only depend on the cardinality of the set S:

$$z(S) = \begin{cases} (-1)^{|S|-k-1} \cdot \binom{|S|-2}{|S|-k-1} & \text{if } S \subseteq C \text{ and } |S| \geq k+1, \\ 0 & \text{otherwise.} \end{cases} \quad \text{for } S \in \Lambda. \tag{22}$$

The proof is based on an auxiliary combinatorial identity (28) from Section A.

Lemma 11 In the context of the characteristic-imset polytope, the k-cluster inequality (14) for $C \subseteq N$, $|C| \geq 2$, and $k \in \{1, \ldots, |C|-1\}$, takes the form

$$\sum_{S \in \Lambda} z(S) \cdot \mathsf{c}(S) \leq |C| - k\,, \quad \text{where } z(S) \text{ are given by (22).} \tag{23}$$

Proof By suitable substitutions we re-write (23) into the desired form (14):

$$\sum_{S\in\Lambda} z(S)\cdot \mathsf{c}_\eta(S) \overset{(22),(1)}{=} \sum_{S\subseteq C\,:\,|S|\geq k+1} z(S)\cdot \sum_{a\in S}\ \sum_{B\,:\,S\setminus\{a\}\subseteq B\subseteq N\setminus\{a\}} \eta(a\,|\,B)$$
$$= \sum_{a\in C}\ \sum_{B\subseteq N\setminus\{a\}\,:\,|B\cap C|\geq k} \eta(a\,|\,B)\cdot \underbrace{\sum_{S\,:\,|S|\geq k+1,\, a\in S,\, S\setminus\{a\}\subseteq B\cap C} z(S)}_{1}\ .$$

It remains to show that, for fixed $a\in C$ and $B\subseteq N\setminus\{a\}$ with $|B\cap C|\geq k$, the indicated expression is indeed 1. We put $\ell := |B\cap C|$, $s := \ell-k$ and write:

$$\sum_{S\,:\,|S|\geq k+1,\, a\in S,\, S\setminus\{a\}\subseteq B\cap C} z(S) \;=\; \sum_{R\subseteq B\cap C\,:\,|R|\geq k} z(\{a\}\cup R)$$
$$\overset{(22)}{=} \sum_{R\subseteq B\cap C,\,|R|\geq k} (-1)^{|R|-k}\cdot \binom{|R|+1-2}{|R|-k} = \sum_{r=k}^{\ell} \binom{\ell}{r}\cdot(-1)^{r-k}\cdot\binom{r-1}{r-k}$$
$$= \sum_{m=0}^{\ell-k} \binom{\ell}{k+m}\cdot(-1)^m\cdot\binom{m+k-1}{m}$$
$$= \sum_{m=0}^{s} (-1)^m\cdot\binom{k+s}{k+m}\cdot\binom{m+k-1}{m} \overset{(28)}{=} 1\,,$$

which concludes the proof.

Thus, it follows from Lemma 11 using Corollaries 4 and 6 that (23) defines a facet of C containing the 1-imset.

10 Simple illustrating examples

To illustrate the achieved results we analyze completely the situation in the case of three BN variables and comment on the case of four BN variables.

We have observed that the following inequalities are facet-defining for the family-variable polytope F in case $|N|=n\geq 3$:

- the *non-negativity* constraints (6) (see Lemma 2(iii)),
- the *modified convexity* constraints (7) (see Lemma 3), and
- the *generalized cluster* inequalities (14) (see Corollary 4).

This is a complete list of facets of F in the case of three BN variables. The following example illustrates the observations from Section 8; we use a shorthand $\eta(a\,|\,bc)$ for $\eta(a\,|\,\{b,c\})$ below.

Example 1 If $N=\{a,b,c\}$ one has $|\Upsilon|=9$. The 9-dimensional polytope F has 25 vertices and 17 facets. Five of its facets are SE and are defined by the generalized cluster inequalities. They decompose into 3 permutation types:

- $\eta(a\,|\,b)+\eta(a\,|\,bc)+\eta(b\,|\,a)+\eta(b\,|\,ac)\leq 1$ (3 inequalities of this type), the (generalized) cluster inequality for $C=\{a,b\}$ (and $k=1$), the extreme supermodular function is $m_{\{a,b\},1}=\delta_{\{a,b,c\}}+\delta_{\{a,b\}}$,
- $\eta(a\,|\,bc)+\eta(b\,|\,ac)+\eta(c\,|\,ab)\leq 1$ (1 inequality of this type), the generalized cluster inequality for $C=\{a,b,c\}$ and $k=2$, it corresponds to the extreme supermodular function $m_{\{a,b,c\},2}=\delta_{\{a,b,c\}}$,
- $\eta(a\,|\,b)+\eta(a\,|\,c)+\eta(a\,|\,bc)+\eta(b\,|\,a)+\eta(b\,|\,c)+\eta(b\,|\,ac)$ $+\,\eta(c\,|\,a)+\eta(c\,|\,b)+\eta(c\,|\,ab)\leq 2$ (1 inequality of this type), the (generalized) cluster inequality for $C=\{a,b,c\}$ (and $k=1$), the supermodular function is $m_{\{a,b,c\},1}=2\cdot\delta_{\{a,b,c\}}+\delta_{\{a,b\}}+\delta_{\{a,c\}}+\delta_{\{b,c\}}$.

If one adds nine non-negativity constraints

- $-\eta(a\,|\,b)\leq 0$ (6 inequalities of this type),
- $-\eta(a\,|\,bc)\leq 0$ (3 inequalities of this type),

to those five generalized cluster inequalities then one obtains a polytope with 28 vertices. Besides the 25 vertices of F it has 3 additional integral vertices of the type $I_{a\leftarrow\{b\}}+I_{a\leftarrow\{c\}}$. By adding the modified convexity constraints

- $\eta(a\,|\,b)+\eta(a\,|\,c)+\eta(a\,|\,bc)\leq 1$ (3 inequalities of this type),

one completes the list of facet-defining inequalities for F.

In the case of four BN variables there are facet-defining inequalities for F other than those given by (6), (7) and (14). In fact,

– there are SE facets other than those given by clusters in (14),
– there are facets besides the SE facets and those given by the non-negativity constraints (6) and modified convexity constraints (7).

Example 2 If $N=\{a,b,c,d\}$ one has $|\Upsilon|=28$ and the 28-dimensional polytope F has 543 vertices and 135 facets. There exist 37 SE facets of F which decompose into 10 permutation types. In Section B we give the list of those types. Six of those types are the generalized cluster inequalities (14), but the remaining four of them are not.

The substantial difference from the case of three BN variables is that the polyhedron F^* specified by 37 SE facet-defining inequalities, 28 non-negativity constraints and 4 modified convexity constraints differs from F. We computed the vertices of F^* and found that, besides all the 543 DAG-codes, it has 786 additional fractional vertices in comparison with F, which decompose into 37 permutation types. Here we give three examples of them:

$$
\begin{aligned}
\eta_1 &= \frac{1}{2}\cdot I_{a\leftarrow\{b\}} + \frac{1}{2}\cdot I_{a\leftarrow\{d\}} + \frac{1}{2}\cdot I_{b\leftarrow\{a,c\}} + \frac{1}{2}\cdot I_{c\leftarrow\{a\}} \\
&\quad + \frac{1}{2}\cdot I_{c\leftarrow\{b,d\}} + \frac{1}{2}\cdot I_{d\leftarrow\{a,b,c\}}, \\
\eta_2 &= \frac{1}{3}\cdot I_{a\leftarrow\{c\}} + \frac{1}{3}\cdot I_{a\leftarrow\{d\}} + \frac{1}{3}\cdot I_{a\leftarrow\{b,c,d\}} + \frac{1}{3}\cdot I_{b\leftarrow\{a\}} + \frac{1}{3}\cdot I_{b\leftarrow\{a,c,d\}} \\
&\quad + \frac{1}{3}\cdot I_{c\leftarrow\{b\}} + \frac{1}{3}\cdot I_{c\leftarrow\{d\}} + \frac{1}{3}\cdot I_{c\leftarrow\{a,b\}} + \frac{1}{3}\cdot I_{d\leftarrow\{a,b,c\}}, \\
\eta_3 &= \frac{1}{6}\cdot I_{a\leftarrow\{b\}} + \frac{1}{3}\cdot I_{a\leftarrow\{d\}} + \frac{1}{3}\cdot I_{b\leftarrow\{c\}} + \frac{1}{3}\cdot I_{b\leftarrow\{a,c,d\}} \\
&\quad + \frac{1}{3}\cdot I_{c\leftarrow\{a\}} + \frac{1}{3}\cdot I_{c\leftarrow\{d\}} + \frac{1}{3}\cdot I_{c\leftarrow\{a,b,d\}} + \frac{1}{3}\cdot I_{d\leftarrow\{b,c\}}.
\end{aligned}
$$

Therefore, the family-variable polytope F necessarily has, besides the above mentioned facets, additional non-SE facets. There are 66 such facet-defining inequalities which decompose into five permutation types; see [10] for details.

The next example is devoted to the characteristic-imset polytope C and illustrates the observations from Section 9. In the case $|N| = 3$, every facet of C either contains the 1-imset or contains the 0-imset.

Example 3 If $N = \{a, b, c\}$ one has $|\Lambda| = 4$. The 4-dimensional polytope C has 11 vertices and 13 facets; they were already discussed in [22, Examples 5,8]. There are five facet-defining inequalities tight for the 1-imset; they correspond to SE facets of F mentioned in Example 1. Here is their overview in both modes; they decompose into 3 permutation types:

- $\mathsf{c}(ab) \le 1$ (3 inequalities of this type),
  in family variables $\eta(a\,|\,b) + \eta(a\,|\,bc) + \eta(b\,|\,a) + \eta(b\,|\,ac) \le 1$,
- $\mathsf{c}(abc) \le 1$ (1 inequality of this type),
  in family variables $\eta(a\,|\,bc) + \eta(b\,|\,ac) + \eta(c\,|\,ab) \le 1$,
- $\mathsf{c}(ab) + \mathsf{c}(ac) + \mathsf{c}(bc) - \mathsf{c}(abc) \le 2$ (1 inequality of this type)
  in family variables
  $\eta(a\,|\,b) + \eta(a\,|\,c) + \eta(a\,|\,bc) + \eta(b\,|\,a) + \eta(b\,|\,c) + \eta(b\,|\,ac)$
  $\quad + \eta(c\,|\,a) + \eta(c\,|\,b) + \eta(c\,|\,ab) \le 2$.

The remaining eight facet-defining inequalities of C are tight for the 0-imset and decompose into 4 permutation types:

- $-\mathsf{c}(ab) \le 0$ (3 inequalities of this type),
  in family variables $-\eta(a\,|\,b) - \eta(a\,|\,bc) - \eta(b\,|\,a) - \eta(b\,|\,ac) \le 0$,
- $-\mathsf{c}(abc) \le 0$ (1 inequality of this type),
  in family variables $-\eta(a\,|\,bc) - \eta(b\,|\,ac) - \eta(c\,|\,ab) \le 0$,
- $-\mathsf{c}(ab) - \mathsf{c}(ac) + \mathsf{c}(abc) \le 0$ (3 inequalities of this type),
  in family variables $-\eta(a\,|\,b) - \eta(a\,|\,c) - \eta(a\,|\,bc) - \eta(b\,|\,a) - \eta(c\,|\,a) \le 0$,

- $-\mathsf{c}(ab) - \mathsf{c}(ac) - \mathsf{c}(bc) + 2 \cdot \mathsf{c}(abc) \leq 0$ (1 inequality of this type),
  in family variables $-\eta(a\,|\,b) - \eta(a\,|\,c) - \eta(b\,|\,a) - \eta(b\,|\,c) - \eta(c\,|\,a) - \eta(c\,|\,b) \leq 0$.

These eight inequalities define in family variables sub-maximal SE faces of F that are not facets: they are implied by the non-negativity constraints. The η-polyhedron F' given by all 13 above-mentioned SE inequalities is unbounded, it has a linear subspace of the dimension 5. This polyhedron is, in fact, the pre-image of the polytope C by the characteristic transformation (1).

As concerns the case of four BN variables, unlike the case of three BN variables, there are facets of the characteristic-imset polytope C which neither contain the 0-imset nor the 1-imset.

Example 4 In case $N = \{a, b, c, d\}$ one has $|\Lambda| = 11$ and the 11-dimensional polytope C has 185 vertices and 154 facets. Thus, it has 358 fewer vertices than the family-variable polytope F, but 19 more facets than F. Besides those 37 facets that correspond to SE facets of F and contain the 1-imset (Corollary 6), there exist 117 facets of C that do not contain the 1-imset. They decompose into 20 permutation types, which are listed in Section C.

With the exception of one permutation type all these inequalities are tight for the 0-imset. The exception is

$$\begin{aligned} & -\mathsf{c}(bc) - \mathsf{c}(bd) - \mathsf{c}(cd) \\ & +\mathsf{c}(abc) + \mathsf{c}(abd) + \mathsf{c}(acd) + 2 \cdot \mathsf{c}(bcd) - 2 \cdot \mathsf{c}(abcd) \leq 1, \end{aligned} \tag{24}$$

in family variables
$$\begin{aligned} & +\eta(a\,|\,bc) + \eta(a\,|\,bd) + \eta(a\,|\,cd) + \eta(a\,|\,bcd) \\ & -\eta(b\,|\,c) - \eta(b\,|\,d) - \eta(c\,|\,b) - \eta(c\,|\,d) - \eta(d\,|\,b) - \eta(d\,|\,c) \leq 1, \end{aligned}$$

consisting of 4 inequalities. The η-version of the inequality (24), therefore, defines an inclusion sub-maximal SE face of F which is not a facet. Clearly, (24) follows from the modified convexity constraint

$$\eta(a\,|\,b) + \eta(a\,|\,c) + \eta(a\,|\,d) + \eta(a\,|\,bc) + \eta(a\,|\,bd) + \eta(a\,|\,cd) + \eta(a\,|\,bcd) \leq 1$$

and the non-negativity constraints $-\eta(a\,|\,b) \leq 0, \ldots, -\eta(d\,|\,c) \leq 0$.

11 The sufficiency of SE faces

As explained in Section 5, the statistical task of learning BN structure can be viewed as an LP problem to maximize an SE objective over the family-variable polytope F. When solving such problems by means of the tools of (integer) linear programming an important question is what are the inequalities specifying the feasible set. The computational complexity depends on how many inequalities we actually/potentially use, how complex they are, how closely we are able to approximate the true feasible set, which is the family-variable polytope F in our case.

In general, facet-defining inequalities for a polytope P are suitable when one maximizes a linear objective over P [26]. Since our goal is to maximize quite special linear objectives over F a natural question is whether we really need all facet-defining inequalities for F. Indeed, the correspondence between SE faces of F and faces of the characteristic-imset polytope C explained in Section 9 allows one to transform the LP problem to maximize an SE objective over F into the task to maximize a general linear function over C. This indicates that those facets of F that are not SE are perhaps superfluous. Note that we know from Example 2 that there are many non-SE facets of F besides those given by the non-negativity (6) and modified convexity (7) constraints.

On the other hand, transforming our LP problems completely into the frame of characteristic imsets does not appear to be advantageous from the point of view of computational complexity as observed in the conclusions of [23]. The main theoretical reason is that simple non-negativity and modified convexity constraints are represented in the characteristic-imset frame by a much higher number of more complex specific inequalities [22].

This motivates the idea of combining both polyhedral approaches to benefit from their different strengths. The constraints that are tight at the empty graph are clearly better represented in the family-variable frame while the constraints that are tight at (all) the complete graphs are more naturally expressed in the characteristic-imset frame. Why not stay in the family-variable frame, utilize (6) and (7) there and combine them with SE constraints, which encode the constraints on the characteristic-imset polytope?

In this section we show that this is indeed possible. However, the *original conjecture* we started with, namely that one can limit oneself to the inequalities defining SE facets and the non-negativity and modified convexity constraints is false; a counter-example in given in Section 12.

The basic observation is that one can limit to SE *faces*.

Lemma 12 Let o be an SE objective. Then the LP problem to maximize $\eta \mapsto \langle o, \eta \rangle_\Upsilon$ over $\eta \in \mathbb{R}^\Upsilon$ from the polyhedron F' specified by the inequalities defining SE faces of F has the same optimal value as the LP problem to maximize that function over the family-variable polytope F.

Proof A basic observation is that the image of F by the transformation (1) is C, which can be viewed as the polyhedron specified through its faces. The pre-image of C with (1) is, therefore, the polyhedron F' of η-vectors specified by the respective inequalities in the η-mode, which are, by Corollary 5, just those defining SE faces of F. Since both F and F' have C as its image by (1), it follows from Lemma 10 that the maximization of $\eta \mapsto \langle o, \eta \rangle_\Upsilon = \langle z_o, \mathsf{c}_\eta \rangle_\Lambda$ over any of them has the same optimal value as the maximization of $\mathsf{c} \mapsto \langle z_o, \mathsf{c} \rangle_\Lambda$ over c in the polytope C.

However, most of the inequalities defining SE faces of F are superfluous. That redundant list can be reduced as follows.

Theorem 3 *Let o be an SE objective. Then the LP problem to*

$$\textit{maximize } \eta \mapsto \langle o, \eta \rangle_\Upsilon \ \textit{ over } \eta \in \mathsf{F}$$

has the same optimal value as the LP problem to maximize the same function over the polyhedron specified by

- *the inequalities defining SE faces that correspond to those facets of* C *that do not contain the 0-imset,*
- *the non-negativity and modified convexity constraints* (6) *and* (7).

Proof We extend the arguments given in the proof of Lemma 12. The polytope C can be viewed as the polyhedron specified by its facet-defining inequalities. In particular, the pre-image F' of C by (1) can equivalently be defined as the polyhedron specified by the facet-defining inequalities for C which are re-written into the η-mode.

Of course, the conclusion of Lemma 12 on the same optimal value holds for any polyhedron F'' such that $\mathsf{F} \subseteq \mathsf{F}'' \subseteq \mathsf{F}'$. Thus, in place of F'' one can take the polyhedron specified by the corresponding facet-defining inequalities for C, the non-negativity and modified convexity constraints.

The last observation is that the facet-defining inequalities for C that are tight for the 0-imset are implied by the non-negativity constraints. Indeed, the η-versions of such inequalities are tight at the empty graph and the observation follows from Lemma 2(i)-(ii). Therefore, they can be dropped from the specification of F''.

Thus, our aim, when maximizing an SE objective, to eliminate non-SE facets of F except for (6) and (7) seems to be achieved. The price for it is that one has to include inequalities that are not facet-defining for F, namely some of the facet-defining inequalities for C written in the η-mode.

Remark 3 It follows from the proof of Theorem 3 that the modified convexity constraints (7) are superfluous there. However, Theorem 3 can be strengthened using a stronger result from [22] which says that one can exclude the so-called *specific inequalities* from the list of the inequalities given by SE objectives. These specific inequalities are shown in [22, § 4.1] to be exact translations of the constraints (6) and (7) into the frame of the characteristic imsets. The list of specific inequalities in case $|N| = 4$ is given in Section C; only one of them, namely (24), needs (7) for its derivation. Thus, in that strengthening of Theorem 3, the modified convexity inequality (7) must be included.

It turns out that, in the case of $|N| = 4$ the original conjecture, namely that SE *facets* plus (6) and (7) are sufficient, is true.

Corollary 7 If $|N| = 4$ then the LP problem to maximize $\eta \mapsto \langle o, \eta \rangle_{\Upsilon}$ over $\eta \in \mathsf{F}$ with an SE objective $o \in \mathbb{R}^{\Upsilon}$ has the same optimal value as the LP problem to maximize the same objective over the polyhedron specified by (6) and (7) and the inequalities defining SE facets.

Proof Use Theorem 3; as shown in Example 4, the only facets of C not implied solely by (6) are those in (24), implied by the combination of (6) and (7).

12 A counter-example to the original conjecture

Recall from Section 11 that our *original conjecture about the sufficiency of SE facets* was as follows:

> for any SE objective o, the maximum of o over the family-variable polytope F coincides with the maximum of o over a (larger) polyhedron in $\mathbb{R}^{\Upsilon}$ specified by
> - non-negativity constraints (6),
> - modified convexity constraints (7), and
> - inequalities defining SE facets of F.

Now we show that the above hypothesis is not true. Our counter-example is based on an example by Orlinskaya [16], who disproved Conjecture 1 from [21] about the facet description of the standard-imset polytope P, a polytope which is affinely equivalent to the characteristic-imset polytope C.

Specifically, in our context, we can re-phrase the finding from Orlinskaya's thesis [16, p. 43] as follows: she found a new facet-defining inequality for the characteristic-imset polytope C in the case $N = \{a,b,c,d,e\}$, which is neither tight for the 1-imset nor one of the earlier-mentioned *specific inequalities*, whose η-versions are derivable from (6) and (7). The inequality has this form:

$$\begin{aligned} &-\mathsf{c}(ab)+2\cdot\mathsf{c}(ac)+3\cdot\mathsf{c}(ae)+\mathsf{c}(bc)-\mathsf{c}(bd)+2\cdot\mathsf{c}(cd)\\ &+5\cdot\mathsf{c}(ce)+3\cdot\mathsf{c}(de)+2\cdot\mathsf{c}(abc)+4\cdot\mathsf{c}(abd)+3\cdot\mathsf{c}(abe)\\ &+\mathsf{c}(acd)-2\cdot\mathsf{c}(ace)+2\cdot\mathsf{c}(bcd)-\mathsf{c}(bce)-3\cdot\mathsf{c}(cde)-5\cdot\mathsf{c}(abcd)\\ &-2\cdot\mathsf{c}(abce)-3\cdot\mathsf{c}(abde)-\mathsf{c}(acde)+\mathsf{c}(bcde)+5\cdot\mathsf{c}(abcde)\leq 16\,. \end{aligned} \tag{25}$$

The substitution of (1) gives the family-variable version of the inequality:

$$\begin{aligned} &-\eta(a\,|\,b)+2\cdot\eta(a\,|\,c)+3\cdot\eta(a\,|\,e)+3\cdot\eta(a\,|\,bc)+3\cdot\eta(a\,|\,bd)\\ &+5\cdot\eta(a\,|\,be)+3\cdot\eta(a\,|\,cd)+3\cdot\eta(a\,|\,ce)+3\cdot\eta(a\,|\,de)+3\cdot\eta(a\,|\,bcd)\\ &+5\cdot\eta(a\,|\,bce)+6\cdot\eta(a\,|\,bde)+3\cdot\eta(a\,|\,cde)+6\cdot\eta(a\,|\,bcde)\\ &-\eta(b\,|\,a)+\eta(b\,|\,c)-\eta(b\,|\,d)+2\cdot\eta(b\,|\,ac)+2\cdot\eta(b\,|\,ad)\\ &+2\cdot\eta(b\,|\,ae)+2\cdot\eta(b\,|\,cd)-\eta(b\,|\,de)+2\cdot\eta(b\,|\,acd)+2\cdot\eta(b\,|\,ace)\\ &+2\cdot\eta(b\,|\,ade)+2\cdot\eta(b\,|\,cde)+5\cdot\eta(b\,|\,acde)+2\cdot\eta(c\,|\,a)\\ &+\eta(c\,|\,b)+2\cdot\eta(c\,|\,d)+5\cdot\eta(c\,|\,e)+5\cdot\eta(c\,|\,ab)+5\cdot\eta(c\,|\,ad)\\ &+5\cdot\eta(c\,|\,ae)+5\cdot\eta(c\,|\,bd)+5\cdot\eta(c\,|\,be)+4\cdot\eta(c\,|\,de)+5\cdot\eta(c\,|\,abd)\\ &+5\cdot\eta(c\,|\,abe)+4\cdot\eta(c\,|\,ade)+7\cdot\eta(c\,|\,bde)+7\cdot\eta(c\,|\,abde)-\eta(d\,|\,b)\\ &+2\cdot\eta(d\,|\,c)+3\cdot\eta(d\,|\,e)+3\cdot\eta(d\,|\,ab)+3\cdot\eta(d\,|\,ac)+3\cdot\eta(d\,|\,ae)\\ &+3\cdot\eta(d\,|\,bc)+2\cdot\eta(d\,|\,be)+2\cdot\eta(d\,|\,ce)+3\cdot\eta(d\,|\,abc)+3\cdot\eta(d\,|\,abe)\\ &+2\cdot\eta(d\,|\,ace)+4\cdot\eta(d\,|\,bce)+5\cdot\eta(d\,|\,abce)+3\cdot\eta(e\,|\,a)\\ &+5\cdot\eta(e\,|\,c)+3\cdot\eta(e\,|\,d)+6\cdot\eta(e\,|\,ab)+6\cdot\eta(e\,|\,ac)+6\cdot\eta(e\,|\,ad)\\ &+4\cdot\eta(e\,|\,bc)+3\cdot\eta(e\,|\,bd)+5\cdot\eta(e\,|\,cd)+6\cdot\eta(e\,|\,abc)\\ &+6\cdot\eta(e\,|\,abd)+5\cdot\eta(e\,|\,acd)+5\cdot\eta(e\,|\,bcd)+8\cdot\eta(e\,|\,abcd)\leq 16\,. \end{aligned} \tag{26}$$

Consider the corresponding SE objective $o^* \in \mathbb{R}^{\Upsilon}$, that is, for any $(a\,|\,B) \in \Upsilon$, $o^*(a\,|\,B)$ is the coefficient with $\eta(a\,|\,B)$ in (26). It follows immediately from Lemma 2(iv) that (26) is *not facet-defining for* F because some coefficients are negative. In fact, the respective SE face of F given by $\langle o^*, \eta\rangle_{\Upsilon} = 16$, denoted below by F_*, has the dimension 53, which is far from 74, the dimension of facets of F. We checked this fact by means of a computer: we found all 153 codes of directed acyclic graphs on F_*; at most 54 of them are affinely independent.

On the other hand, the inequality (25) *is facet-defining for* C. We have computed 59 characteristic imsets on this face of C, denoted below by $\bar{F}_*$, and found 26 of them affinely independent. This implies the dimension of $\bar{F}_*$ is 25, which is the dimension of facets of C.

To get the desired counter-example we consider a convex combination $\eta_\dagger$ of all 153 codes of directed acyclic graphs on F_* with the coefficients $\frac{1}{153}$:

$$\begin{aligned}
\eta_\dagger := \ & \frac{4}{153}\cdot I_{a\leftarrow\{c\}} + \frac{4}{153}\cdot I_{a\leftarrow\{d\}} + \frac{14}{153}\cdot I_{a\leftarrow\{e\}} + \frac{1}{153}\cdot I_{a\leftarrow\{b,c\}}\\
&+ \frac{10}{153}\cdot I_{a\leftarrow\{b,d\}} + \frac{3}{153}\cdot I_{a\leftarrow\{b,e\}} + \frac{8}{153}\cdot I_{a\leftarrow\{c,d\}} + \frac{7}{153}\cdot I_{a\leftarrow\{c,e\}}\\
&+ \frac{7}{153}\cdot I_{a\leftarrow\{d,e\}} + \frac{1}{153}\cdot I_{a\leftarrow\{b,c,d\}} + \frac{3}{153}\cdot I_{a\leftarrow\{b,c,e\}} + \frac{24}{153}\cdot I_{a\leftarrow\{b,d,e\}}\\
&+ \frac{3}{153}\cdot I_{a\leftarrow\{c,d,e\}} + \frac{18}{153}\cdot I_{a\leftarrow\{b,c,d,e\}} + \frac{8}{153}\cdot I_{b\leftarrow\{c\}} + \frac{6}{153}\cdot I_{b\leftarrow\{e\}}\\
&+ \frac{6}{153}\cdot I_{b\leftarrow\{c,d\}} + \frac{6}{153}\cdot I_{b\leftarrow\{c,d,e\}} + \frac{66}{153}\cdot I_{b\leftarrow\{a,c,d,e\}} + \frac{4}{153}\cdot I_{c\leftarrow\{a\}}\\
&+ \frac{4}{153}\cdot I_{c\leftarrow\{b\}} + \frac{2}{153}\cdot I_{c\leftarrow\{d\}} + \frac{33}{153}\cdot I_{c\leftarrow\{e\}} + \frac{8}{153}\cdot I_{c\leftarrow\{a,b\}}\\
&+ \frac{15}{153}\cdot I_{c\leftarrow\{a,d\}} + \frac{13}{153}\cdot I_{c\leftarrow\{a,e\}} + \frac{11}{153}\cdot I_{c\leftarrow\{b,d\}} + \frac{1}{153}\cdot I_{c\leftarrow\{b,e\}}\\
&+ \frac{2}{153}\cdot I_{c\leftarrow\{a,b,d\}} + \frac{1}{153}\cdot I_{c\leftarrow\{a,b,e\}} + \frac{21}{153}\cdot I_{c\leftarrow\{b,d,e\}}\\
&+ \frac{15}{153}\cdot I_{c\leftarrow\{a,b,d,e\}} + \frac{4}{153}\cdot I_{d\leftarrow\{a\}} + \frac{2}{153}\cdot I_{d\leftarrow\{c\}} + \frac{38}{153}\cdot I_{d\leftarrow\{e\}}\\
&+ \frac{10}{153}\cdot I_{d\leftarrow\{a,b\}} + \frac{12}{153}\cdot I_{d\leftarrow\{a,c\}} + \frac{13}{153}\cdot I_{d\leftarrow\{a,e\}} + \frac{1}{153}\cdot I_{d\leftarrow\{b,c\}}\\
&+ \frac{1}{153}\cdot I_{d\leftarrow\{a,b,c\}} + \frac{2}{153}\cdot I_{d\leftarrow\{a,b,e\}} + \frac{6}{153}\cdot I_{d\leftarrow\{b,c,e\}}\\
&+ \frac{13}{153}\cdot I_{d\leftarrow\{a,b,c,e\}} + \frac{8}{153}\cdot I_{e\leftarrow\{a\}} + \frac{3}{153}\cdot I_{e\leftarrow\{b\}} + \frac{23}{153}\cdot I_{e\leftarrow\{c\}}\\
&+ \frac{19}{153}\cdot I_{e\leftarrow\{d\}} + \frac{15}{153}\cdot I_{e\leftarrow\{a,b\}} + \frac{12}{153}\cdot I_{e\leftarrow\{a,c\}} + \frac{17}{153}\cdot I_{e\leftarrow\{a,d\}}\\
&+ \frac{3}{153}\cdot I_{e\leftarrow\{b,d\}} + \frac{2}{153}\cdot I_{e\leftarrow\{c,d\}} + \frac{1}{153}\cdot I_{e\leftarrow\{a,b,c\}} + \frac{4}{153}\cdot I_{e\leftarrow\{a,b,d\}}\\
&+ \frac{2}{153}\cdot I_{e\leftarrow\{b,c,d\}} + \frac{14}{153}\cdot I_{e\leftarrow\{a,b,c,d\}}\,.
\end{aligned}$$

It is tedious but straightforward to verify $\langle o^*, \eta_\dagger\rangle = 16$. One can also easily check that none of five modified convexity constraints is tight for $\eta_\dagger$. We also verified that the vector $\mathsf{c}_\dagger \in \mathbb{R}^\Lambda$ ascribed to $\eta_\dagger$ by (1) is in the relative interior of $\bar{F}_* \subseteq \mathsf{C}$. For this purpose, we have first used 59 vertices of $\bar{F}_*$ to compute its 55 facets. Then we verified computationally that $\mathsf{c}_\dagger$ does not belong to any of the 55 facets of $\bar{F}_*$.

The above observation implies that none of the SE-facets of F contains $\eta_\dagger$. Indeed, assume for a contradiction that $\eta_\dagger$ belongs to some SE-facet F of F. Then, by Lemma 10 and Corollary 5, $\mathsf{c}_\dagger$ belongs to the corresponding face $\bar{F}$ of C, which is, by Corollary 6, a facet of C. The facet $\bar{F}$ does not fully contain $\bar{F}_*$ since otherwise, by Lemma 1 applied to C and $\bar{F}_*$, one has $\bar{F} = \bar{F}_*$ and, by Corollary 5, $F = F_*$, contradicting the above mentioned fact that (26) is not facet-defining for F. Therefore, $\mathsf{c}_\dagger \in \bar{F}_* \cap \bar{F} \subset \bar{F}_*$, which is a contradiction with $\mathsf{c}_\dagger$ belonging to the relative interior of $\bar{F}_*$.

These observations are enough to derive the existence of a counter-example, which is the vector $\eta_\star := (1+\epsilon)\cdot\eta_\dagger$, where $\epsilon > 0$ is small enough. Indeed, $\eta_\star$ satisfies all non-negativity constraints and all other inequalities, namely five modified convexity constraints and SE-facets of F are valid for $\eta_\dagger$ but not tight for it: $\langle o, \eta_\dagger\rangle < u$ for the respective $o \in \mathbb{R}^\Upsilon$ and $u > 0$. Since the number of these inequalities is finite, a small ϵ-perturbation retains $\langle o, \eta_\star\rangle < u$ for any of them. On the other hand, the value of the considered SE objective $o^* \in \mathbb{R}^\Upsilon$ for $\eta_\star$ is

$$\langle o^*, \eta_\star\rangle = (1+\epsilon)\cdot\langle o^*, \eta_\dagger\rangle = (1+\epsilon)\cdot 16 > 16\,.$$

Thus, the maximum of the linear SE objective $\eta \mapsto \langle o^*, \eta\rangle$, $\eta \in \mathbb{R}^\Upsilon$ on F is 16, while its value in $\eta_\star$, which satisfies (6), (7) and all SE facet-defining inequalities for F, exceeds 16. This gives the desired counter-example.

13 Computational point of view

As explained earlier in this paper, the task of learning BN network structure can be re-formulated in the form of an LP problem to maximize a linear/affine objective over a polytope, either over the family-variable polytope F or over the characteristic-imset polytope C. Nevertheless, solving this optimization task directly as an LP problem is unrealistic for at least two reasons:

- the dimension of both polytopes grows exponentially with the number of nodes $n = |N|$,
- the number of inequalities specifying the polytopes seem to grow even more, as suggested by Theorem 2 and Corollary 6.

In practical computational optimization, the first obstacle is overcome by means of a special dimension reduction procedure, we name *pruning*, while the second obstacle is overcome by the application of advanced methods of *integer linear programming*.

We are not going to explain all details of the *pruning* procedure in this theoretical paper; we only mention the main idea, elaborated in more detail in [11]. The point is that typical databases occurring in practice are limited in the number of items and an advantageous form of common scoring criteria allows one to conclude that the optimal graph does not have nodes with really large parent sets. This is based on a simple observation that in case $o(a\,|\,B) < o(a\,|\,C)$ for $C \subset B \subseteq N \setminus \{a\}$ one has $\eta_G(a\,|\,B) = 0$ in any $G \in \mathsf{DAGS}\,(N)$ maximizing $G \mapsto \langle o, \eta_G \rangle$. Hence, one can exclude from consideration many components of η because they have to vanish in the family-variable vector η_G for an optimal graph G. An analogous conclusion can be reached in the case of characteristic imsets [23]. The pruning procedure is time consuming, but useful: as reported in [11, § 6], in practical cases it usually results in the reduction of the parent set cardinality to at most 5. Thus, in practical situations the actual length of BN vector representatives is "polynomial" in the number of nodes $n = |N|$.

As concerns the second obstacle, both considered polytopes are integral, that is, all their vertices have integers as components. Thus, one can further reformulate the optimization task as an *integer linear programming* problem to maximize a linear objective over integral vertices within a relaxed polyhedron P' for the original polytope P. For example, the basic non-negativity and modified convexity constraints (6), (7) together with the cluster inequalities mentioned in Section 8 define a relaxation F' for the family-variable polytope F, which has the property that the vectors in F' having integers as components coincide with the vertices of F. The problem with the exponential number of cluster inequalities can be solved by an iterative constraint adding method [13]. Advanced combinatorial optimization methods can be applied in this area, like the *cutting plane* approach [8] or its combination with the branch-and-bound approach, known as the *branch-and-cut* method [1]. Good running time were achieved by GOBNILP [9] even in cases where $n = |N|$ approaches 100, which testifies to the feasibility of the ILP approach to BN structure learning.

Computational experiments have also confirmed empirically the importance of facet-defining inequalities as predicted theoretically in [26, § 9.1-9.2]: one can substantially speed up the ILP solving procedure by including them as potential cutting planes. Thus, although Theorem 2 only gives an implicit description of SE facets of F, it provides the way of verifying whether a prospective inequality defines an SE facet and offers a potential method to generate useful cutting planes. The observation from Theorem 3 about the redundancy of non-SE facets of F when maximizing an SE objective has also been confirmed empirically. To summarize: the theoretical results in our paper can help one to design computationally efficient ways of doing the BN structure learning, although one cannot directly "read off" the best practical way of solving this optimization problem from them.

14 Conclusions

Let us summarize the main achievements of the paper. We dealt with two distinguished polytopes used in the ILP approach to BN structure learning, namely the *family-variable polytope* and the *characteristic-imset polytope*. Being motivated by a common form of linear objective to be maximized in BN structure learning we introduced the concept of a *score equivalent (SE) face* of the family-variable polytope. We further characterized the linear space of the corresponding SE objectives (Lemma 5). A correspondence has been established between SE faces of the family-variable polytope F and the faces of the characteristic-imset polytope C, which preserves the inclusion of faces (Corollary 5).

We observed that SE facets of F correspond to those facets of C which contain a distinguished vector, called the 1-imset (Corollary 6) and succeeded in characterizing SE facets in terms of the respective collection of graph codes (Theorem 1). The SE facets of F were also shown to correspond to *extreme supermodular functions*, which gives an elegant method to verify that an inequality is SE-facet-defining for F (Theorem 2). To illustrate the method we showed that the well-known (generalized) cluster inequalities are facet-defining for F (Corollary 4) and derived their form in the context of the characteristic-imset polytope (Lemma 11). The correspondence with extreme supermodular set functions may appear to be useful because of a recent extremality criterion for supermodular functions from [25].

Since a typical linear objective appearing in the ILP approach to learning BN structure is special, namely SE, we raised the question of whether all facets of F are needed to specify the feasible sets for (integer) linear programs when such an objective is maximized. We succeeded in showing that one can eliminate those facets of F that are not SE, that is, defined by a non-SE normal vectors (Lemma 12, Theorem 3). Nevertheless, our starting original conjecture that one can, besides simple non-negativity and modified convexity constraints, limit oneself only to SE facets of F turned out not to be true (a counter-example is given in Section 12). The moral is that one has to consider the inequalities defining facets of the characteristic-imset polytope C although they may not define facets in the context of the family-variable polytope F.

This leads to a suggestion to use a combined coding of BN structures in the ILP approach. One can encode a BN structure by a concatenation of the family-variable vector and the characteristic imset and utilize the linear relation (1). Linear constraints tight at the empty graph are better represented by simple non-negativity and modified convexity inequalities in the family-variable part, while the other SE linear inequality constraints can be more naturally represented in the characteristic-imset part.

We left some of questions open. One of them is whether a simple condition of being closed under Markov equivalence characterizes the sets of graph-codes belonging to SE faces of F (Conjecture 1). However, it looks like the answer to this question is not essential for the practical application of ILP methods in BN structure learning.

Acknowledgements The research of Milan Studený has been supported by the grants GAČR n. 13-20012S and 16-12010S. James Cussens was supported by the UK Medical Research Council, grant G1002312 and senior postdoctoral fellowship SF/14/008 from KU

Leuven. Our special thanks are devoted to Fero Matúš, who helped us to find an easy proof of the combinatorial identity from Lemma 13. We also express our gratitude to the reviewer for valuable comments.

References

1. Bartlett, M., Cussens, J.: Advances in Bayesian network learning using integer programming. In: Uncertainty in Artificial Intelligence 29, pp. 182-191, AUAI Press, Corvallis (2013).
2. Barvinok, A.: A Course in Convexity. Graduate Studies in Mathematics 54. American Mathematical Society, Providence (2002).
3. Bouckaert, R.R.: Bayesian belief networks – from construction to evidence. PhD thesis, University of Utrecht, 1995.
4. Brøndsted, A.: An Introduction to Convex Polytopes. Springer, New York (1983).
5. Chickering, D.M.: A transformational characterization of equivalent Bayesian network structures. In: Uncertainty in Artificial Intelligence 11, pp. 87–98, Morgan Kaufmann, San Francisco (1995).
6. Chickering, D.M.: Optimal structure identification with greedy search. Journal of Machine Learning Research 3, 505–554 (2002).
7. Cussens, J.: Maximum likelihood pedigree reconstruction using integer programming. In: Proceedings of the Workshop on Constraint Based Methods for Bioinformatics (WCBMB), pp. 9–19 (2010).
8. Cussens, J.: Bayesian network learning with cutting planes. In: Uncertainty in Artificial Intelligence 27, pp. 153–160, AUAI Press, Corvallis (2011).
9. Cussens, J., Bartlett, M.: GOBNILP software; a web page (updated 2016) www.cs.york.ac.uk/aig/sw/gobnilp/ .
10. Cussens, J., Järvisalo, M., Korhonen, J.H., Bartlett, M.: Bayesian network structure learning with integer programming: polytopes, facets, and complexity. Submitted to Journal of Artificial Intelligence Research (2016); available at arxiv.org/abs/1605.04071 .
11. de Campos, C.P., Ji, Q.: Efficient structure learning Bayesian networks using constraints. Journal of Machine Learning Research 12, 663–689 (2011).
12. Hemmecke, R., Lindner, S., Studený, M.: Characteristic imsets for learning Bayesian network structure. International Journal of Approximate Reasoning 53, 1336-1349 (2012).
13. Jaakkola, T., Sontag, D., Globerson, A., Meila, M.: Learning Bayesian network structure using LP relaxations. In: Journal of Machine Learning Research Workshop and Conference Proceedings 9: AISTATS 2010, pp. 358–365 (2010).
14. Lauritzen, S.L.: Graphical Models. Clarendon Press, Oxford (1996).
15. Neapolitan, R.E.: Learning Bayesian Networks. Pearson Prentice Hall, Upper Saddle River (2004).
16. Orlinskaya G.: Linear constraints on standard and characteristic imsets for learning Bayesian network structures. Diploma thesis, TU Munich, 2014.
17. Oxley, J.G.: Matroid Theory. Oxford University Press, Oxford (1992).
18. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo (1988).
19. Studený, M.: Probabilistic Conditional Independence Structures. Springer, London (2005).
20. Studený M., Vomlel J., Hemmecke R.: A geometric view on learning Bayesian network structures. International Journal of Approximate Reasoning 51, 573–586 (2010).
21. Studený M., Vomlel J.: On open questions in the geometric approach to structural learning Bayesian nets. International Journal of Approximate Reasoning 52, 627–640 (2011).
22. Studený M., Haws D.C.: On polyhedral approximations of polytopes for learning Bayesian networks. Journal of Algebraic Statistics 4, 59–92 (2013).
23. Studený M., Haws D.: Learning Bayesian network structure: towards the essential graph by integer linear programming tools. International Journal of Approximate Reasoning 55 1043–1071 (2014).
24. Studený, M.: How matroids occur in the context of learning Bayesian network structures. In: Uncertainty in Artificial Intelligence 31, pp. 832–841, AUAI Press, Corvallis (2015).

25. Studený M., Kroupa T.: Core-based criterion for extreme supermodular functions. Discrete Applied Mathematics 206, 122–151 (2016).
26. Wolsey L.A.: Integer Programming. John Wiley, New York (1998).
27. Ziegler G.M.: Lectures on Polytopes. Springer, New York (1995).

A Combinatorial identity

Lemma 13 For every non-negative integer $s \geq 0$, $k \geq K \geq 0$ one has

$$\sum_{m=0}^{s} (-1)^m \cdot \binom{k+s}{k+m} \cdot \binom{m+k-K}{m} = \binom{s+K-1}{K-1}, \tag{27}$$

with conventions $\binom{n}{0} = \binom{n}{n} = 1$ for any $n \in \mathbb{Z}$ and $\binom{n}{-1} = \binom{n}{n+1} = 0$ for any non-negative $n \in \mathbb{Z}$. In particular,

$$\forall\, s \geq 0,\ k \geq 1 \text{ integers} \qquad \sum_{m=0}^{s} (-1)^m \cdot \binom{k+s}{k+m} \cdot \binom{m+k-1}{m} = 1\,. \tag{28}$$

Proof The proof relies on Pascal's triangle identity

$$\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1} \quad \text{valid for integers } n \geq 1,\ n \geq r \geq 0.$$

Let us denote the sum in (27) by $\Sigma(s,k,K)$; the basic idea of the proof is the induction on $s+K$. First, we verify (27) in the case $s = 0$:

$$\Sigma(s=0,k,K) = (-1)^0 \cdot \binom{k+0}{k+0} \cdot \binom{k-K}{0} = 1 = \binom{0+K-1}{K-1}.$$

Further special easy case is $s \geq 1$ and $K = 0$, in which case

$$\begin{aligned}
&\Sigma(s \geq 1, k, K=0) = \sum_{m=0}^{s} (-1)^m \cdot \binom{k+s}{k+m} \cdot \binom{m+k}{m} \\
&= \sum_{m=0}^{s} (-1)^m \cdot \frac{(k+s)!}{(k+m)! \cdot (s-m)!} \cdot \frac{(k+m)!}{m! \cdot k!} \\
&= \frac{(k+s)!}{k! \cdot s!} \cdot \sum_{m=0}^{s} (-1)^m \cdot \frac{s!}{m! \cdot (s-m)!} = \binom{k+s}{k} \cdot \sum_{m=0}^{s} (-1)^m \cdot \binom{s}{m} \\
&= \binom{k+s}{k} \cdot (-1+1)^s = 0 = \binom{s-1}{-1}.
\end{aligned}$$

Thus, (27) holds in cases $s = 0$ and $K = 0$; in particular, if $s + K \leq 1$. In case $s, K \geq 1$ the induction premise means (27) holds for $s', K' \geq 0$ with $s' + K' \leq s + K - 1$. To verify the induction step write by the identity

$$\binom{k+s}{k+m} = \binom{k+s-1}{k+m} + \binom{k+s-1}{k+m-1}, \quad \text{use} \quad \binom{k+s-1}{k+s} = 0\,,$$

apply the induction premise and use Pascal's triangle identity again:

$$\begin{aligned}
\Sigma(s,k,K) &= \sum_{m=0}^{s} (-1)^m \cdot \binom{k+s}{k+m} \cdot \binom{m+k-K}{m} \\
&= \sum_{m=0}^{s-1} (-1)^m \cdot \binom{k+s-1}{k+m} \cdot \binom{m+k-K}{m} \\
&\qquad + \sum_{m=0}^{s} (-1)^m \cdot \binom{k+s-1}{k+m-1} \cdot \binom{m+k-K}{m} \\
&= \Sigma(s-1,k,K) + \Sigma(s,k-1,K-1) \\
&= \binom{s+K-2}{K-1} + \binom{s+K-2}{K-2} = \binom{s+K-1}{K-1},
\end{aligned}$$

which gives the desired result. Putting $K = 1$ gives (28).

B Catalogue of SE facets in case of four BN variables

There exist 37 SE facets of F in the case $N = \{a,b,c,d\}$ which decompose into 10 permutations types. Below we list all the types of the inequalities, both in the family-variable mode and in the characteristic-imset mode. The generalized cluster inequalities are indicated by $\bullet$, the remaining types by $\circ$.

- the (generalized) cluster inequality for $C = \{a,b\}$ (and $k = 1$),

  $$[\,\eta(a\,|\,b)+\eta(a\,|\,bc)+\eta(a\,|\,bd)+\eta(a\,|\,bcd)\,]+[\,\eta(b\,|\,a)+\eta(b\,|\,ac)+\eta(b\,|\,ad)+\eta(b\,|\,acd)\,] \leq 1\,,$$

  (6 inequalities of this type), in characteristic imsets

  $$\mathsf{c}(ab) \leq 1\,,$$

- the generalized cluster inequality for $C = \{a,b,c\}$ and $k = 2$,

  $$[\,\eta(a\,|\,bc)+\eta(a\,|\,bcd)\,]+[\,\eta(b\,|\,ac)+\eta(b\,|\,acd)\,]+[\,\eta(c\,|\,ab)+\eta(c\,|\,abd)\,] \leq 1\,,$$

  (4 inequalities of this type), in characteristic imsets

  $$\mathsf{c}(abc) \leq 1\,,$$

- the (generalized) cluster inequality for $C = \{a,b,c\}$ (and $k = 1$),

  $$\begin{aligned}
  &[\,\eta(a\,|\,b)+\eta(a\,|\,c)+\eta(a\,|\,bc)+\eta(a\,|\,bd)+\eta(a\,|\,cd)+\eta(a\,|\,bcd)\,] \\
  &+[\,\eta(b\,|\,a)+\eta(b\,|\,c)+\eta(b\,|\,ac)+\eta(b\,|\,ad)+\eta(b\,|\,cd)+\eta(b\,|\,acd)\,] \\
  &+[\,\eta(c\,|\,a)+\eta(c\,|\,b)+\eta(c\,|\,ab)+\eta(c\,|\,ad)+\eta(c\,|\,bd)+\eta(c\,|\,abd)\,] \leq 2\,,
  \end{aligned}$$

  (4 inequalities of this type), in characteristic imsets

  $$\mathsf{c}(ab)+\mathsf{c}(ac)+\mathsf{c}(bc)-\mathsf{c}(abc) \leq 2\,,$$

- the generalized cluster inequality for $C=\{a,b,c,d\}$ and $k=3$,

$$[\,\eta(a\,|\,bcd)+\eta(b\,|\,acd)+\eta(c\,|\,abd)+\eta(d\,|\,abc)\,]\leq 1\,,$$

(1 inequality of this type), in characteristic imsets

$$\mathsf{c}(abcd)\leq 1\,,$$

- the generalized cluster inequality for $C=\{a,b,c,d\}$ and $k=2$,

$$\begin{aligned}&[\,\eta(a\,|\,bc)+\eta(a\,|\,bd)+\eta(a\,|\,cd)+\eta(a\,|\,bcd)\,]\\&+[\,\eta(b\,|\,ac)+\eta(b\,|\,ad)+\eta(b\,|\,cd)+\eta(b\,|\,acd)\,]\\&+[\,\eta(c\,|\,ab)+\eta(c\,|\,ad)+\eta(c\,|\,bd)+\eta(c\,|\,abd)\,]\\&+[\,\eta(d\,|\,ab)+\eta(d\,|\,ac)+\eta(d\,|\,bc)+\eta(d\,|\,abc)\,]\leq 2\,,\end{aligned}$$

(1 inequality of this type), in characteristic imsets

$$\mathsf{c}(abc)+\mathsf{c}(abd)+\mathsf{c}(acd)+\mathsf{c}(bcd)-2\cdot\mathsf{c}(abcd)\leq 2\,,$$

- the (generalized) cluster inequality for $C=\{a,b,c,d\}$ (and $k=1$),

$$\begin{aligned}&[\,\eta(a\,|\,b)+\eta(a\,|\,c)+\eta(a\,|\,d)+\eta(a\,|\,bc)+\eta(a\,|\,bd)+\eta(a\,|\,cd)+\eta(a\,|\,bcd)\,]\\&+[\,\eta(b\,|\,a)+\eta(b\,|\,c)+\eta(b\,|\,d)+\eta(b\,|\,ac)+\eta(b\,|\,ad)+\eta(b\,|\,cd)+\eta(b\,|\,acd)\,]\\&+[\,\eta(c\,|\,a)+\eta(c\,|\,b)+\eta(c\,|\,d)+\eta(c\,|\,ab)+\eta(c\,|\,ad)+\eta(c\,|\,bd)+\eta(c\,|\,abd)\,]\\&+[\,\eta(d\,|\,a)+\eta(d\,|\,b)+\eta(d\,|\,c)+\eta(d\,|\,ab)+\eta(d\,|\,ac)+\eta(d\,|\,bc)+\eta(d\,|\,abc)\,]\leq 3\,,\end{aligned}$$

(1 inequality of this type), in characteristic imsets

$$\begin{aligned}&\mathsf{c}(ab)+\mathsf{c}(ac)+\mathsf{c}(ad)+\mathsf{c}(bc)+\mathsf{c}(bd)+\mathsf{c}(cd)\\&\quad-\mathsf{c}(abc)-\mathsf{c}(abd)-\mathsf{c}(acd)-\mathsf{c}(bcd)+\mathsf{c}(abcd)\leq 3\,,\end{aligned}$$

- ○ non-cluster SE inequality with 13 terms

$$\begin{aligned}&[\,\eta(a\,|\,bc)+\eta(a\,|\,bd)+\eta(a\,|\,cd)+2\cdot\eta(a\,|\,bcd)\,]\\&+[\,\eta(b\,|\,ac)+\eta(b\,|\,ad)+\eta(b\,|\,acd)\,]\\&+[\,\eta(c\,|\,ab)+\eta(c\,|\,ad)+\eta(c\,|\,abd)\,]\\&+[\,\eta(d\,|\,ab)+\eta(d\,|\,ac)+\eta(d\,|\,abc)\,]\leq 2\,,\end{aligned}$$

(4 inequalities of this type), in characteristic imsets

$$\mathsf{c}(abc)+\mathsf{c}(abd)+\mathsf{c}(acd)-\mathsf{c}(abcd)\leq 2\,,$$

- ○ non-cluster SE inequality with 16 terms

$$\begin{aligned}&[\,\eta(a\,|\,b)+\eta(a\,|\,bc)+\eta(a\,|\,bd)+\eta(a\,|\,cd)+\eta(a\,|\,bcd)\,]\\&+[\,\eta(b\,|\,a)+\eta(b\,|\,ac)+\eta(b\,|\,ad)+\eta(b\,|\,cd)+\eta(b\,|\,acd)\,]\\&+[\,\eta(c\,|\,ad)+\eta(c\,|\,bd)+\eta(c\,|\,abd)\,]\\&+[\,\eta(d\,|\,ac)+\eta(d\,|\,bc)+\eta(d\,|\,abc)\,]\leq 2\,,\end{aligned}$$

(6 inequalities of this type), in characteristic imsets

$$\mathsf{c}(ab)+\mathsf{c}(acd)+\mathsf{c}(bcd)-\mathsf{c}(abcd)\leq 2\,,$$

- non-cluster SE inequality with 22 terms

$$\begin{aligned}[\,\eta(a\,|\,b)+\eta(a\,|\,c)+\eta(a\,|\,d)+2\cdot\eta(a\,|\,bc)+2\cdot\eta(a\,|\,bd)+2\cdot\eta(a\,|\,cd)+2\cdot\eta(a\,|\,bcd)\,]\\ +[\,\eta(b\,|\,a)+\eta(b\,|\,ac)+\eta(b\,|\,ad)+\eta(b\,|\,cd)+\eta(b\,|\,acd)\,]\\ +[\,\eta(c\,|\,a)+\eta(c\,|\,ab)+\eta(c\,|\,ad)+\eta(c\,|\,bd)+\eta(c\,|\,abd)\,]\\ +[\,\eta(d\,|\,a)+\eta(d\,|\,ab)+\eta(d\,|\,ac)+\eta(d\,|\,bc)+\eta(d\,|\,abc)\,]\leq 3,\end{aligned}$$

  (4 inequalities of this type), in characteristic imsets

$$\mathsf{c}(ab)+\mathsf{c}(ac)+\mathsf{c}(ad)+\mathsf{c}(bcd)-\mathsf{c}(abcd)\leq 3\,,$$

- non-cluster SE inequality with 26 terms

$$\begin{aligned}[\,\eta(a\,|\,b)+\eta(a\,|\,c)+\eta(a\,|\,d)+\eta(a\,|\,bc)+\eta(a\,|\,bd)+2\cdot\eta(a\,|\,cd)+2\cdot\eta(a\,|\,bcd)\,]\\ +[\,\eta(b\,|\,a)+\eta(b\,|\,c)+\eta(b\,|\,d)+\eta(b\,|\,ac)+\eta(b\,|\,ad)+2\cdot\eta(b\,|\,cd)+2\cdot\eta(b\,|\,acd)\,]\\ +[\,\eta(c\,|\,a)+\eta(c\,|\,b)+\eta(c\,|\,ab)+\eta(c\,|\,ad)+\eta(c\,|\,bd)+2\cdot\eta(c\,|\,abd)\,]\\ +[\,\eta(d\,|\,a)+\eta(d\,|\,b)+\eta(d\,|\,ab)+\eta(d\,|\,ac)+\eta(d\,|\,bc)+2\cdot\eta(d\,|\,abc)\,]\leq 4\,,\end{aligned}$$

  (6 inequalities of this type), in characteristic imsets

$$\mathsf{c}(ab)+\mathsf{c}(ac)+\mathsf{c}(ad)+\mathsf{c}(bc)+\mathsf{c}(bd)-\mathsf{c}(abc)-\mathsf{c}(abd)+\mathsf{c}(abcd)\leq 4\,.$$

C Specific inequalities in case of four BN variables

In the case $|N|=4$, the characteristic-imset polytope C has, besides 37 facets containing the 1-imset and listed in Section B, additionally 117 specific facets that do not contain the 1-imset. They all are defined by means of the so-called *specific inequalities* discussed in [22, § 4.1.2]. Each of these inequalities corresponds to a clutter (= Sperner family = antichain) of non-empty subsets of N, that is, to a class of inclusion-incomparable subsets of N. The 117 specific facets decompose into 20 permutation types listed below. Except four facets belonging to the last type, mentioned earlier in (24), all of them contain the 0-imset.

- $-\mathsf{c}(ab)\leq 0$ (6 inequalities of this type),
  Sperner family is $\mathcal{I}=\{ab\}$,
- $-\mathsf{c}(abc)\leq 0$ (4 inequalities of this type),
  Sperner family is $\mathcal{I}=\{abc\}$,
- $-\mathsf{c}(abcd)\leq 0$ (1 inequality of this type),
  Sperner family is $\mathcal{I}=\{abcd\}$,
- $-\mathsf{c}(ab)-\mathsf{c}(ac)-\mathsf{c}(bc)+2\cdot\mathsf{c}(abc)\leq 0$ (4 inequalities of this type),
  Sperner family is $\mathcal{I}=\{ab,ac,bc\}$,
- $-\mathsf{c}(ab)-\mathsf{c}(acd)-\mathsf{c}(bcd)+2\cdot\mathsf{c}(abcd)\leq 0$ (6 inequalities of this type),
  Sperner family is $\mathcal{I}=\{ab,acd,bcd\}$,

- $-\mathsf{c}(abc) - \mathsf{c}(abd) - \mathsf{c}(acd) + 2\cdot\mathsf{c}(abcd) \leq 0$ (4 inequalities of this type), Sperner family is $\mathcal{I} = \{abc, abd, acd\}$,
- $-\mathsf{c}(abc) - \mathsf{c}(abd) - \mathsf{c}(acd) - \mathsf{c}(bcd) + 3\cdot\mathsf{c}(abcd) \leq 0$ (1 inequality), Sperner family is $\mathcal{I} = \{abc, abd, acd, bcd\}$,
- $-\mathsf{c}(ab) - \mathsf{c}(ac) - \mathsf{c}(ad) - \mathsf{c}(bcd) + \mathsf{c}(abc) + \mathsf{c}(abd) + \mathsf{c}(acd) \leq 0$ (4 inequalities), Sperner family is $\mathcal{I} = \{ab, ac, ad, bcd\}$,
- $-\mathsf{c}(ab) - \mathsf{c}(ac) - \mathsf{c}(ad) - \mathsf{c}(bc) - \mathsf{c}(bd)$ $+2\cdot\mathsf{c}(abc) + 2\cdot\mathsf{c}(abd) + \mathsf{c}(acd) + \mathsf{c}(bcd) - 2\cdot\mathsf{c}(abcd) \leq 0$ (6 inequalities), Sperner family is $\mathcal{I} = \{ab, ac, ad, bc, bd\}$,
- $-\mathsf{c}(ab) - \mathsf{c}(ac) - \mathsf{c}(ad) - \mathsf{c}(bc) - \mathsf{c}(bd) - \mathsf{c}(cd)$ $+2\cdot\mathsf{c}(abc) + 2\cdot\mathsf{c}(abd) + 2\cdot\mathsf{c}(acd) + 2\cdot\mathsf{c}(bcd) - 3\cdot\mathsf{c}(abcd) \leq 0$ (1 inequality), Sperner family is $\mathcal{I} = \{ab, ac, ad, bc, bd, cd\}$,
- $-\mathsf{c}(ab) - \mathsf{c}(ac) + \mathsf{c}(abc) \leq 0$ (12 inequalities of this type), Sperner family is $\mathcal{I} = \{ab, ac\}$,
- $-\mathsf{c}(abc) - \mathsf{c}(abd) + \mathsf{c}(abcd) \leq 0$ (6 inequalities of this type), Sperner family is $\mathcal{I} = \{abc, abd\}$,
- $-\mathsf{c}(ab) - \mathsf{c}(acd) + \mathsf{c}(abcd) \leq 0$ (12 inequalities of this type), Sperner family is $\mathcal{I} = \{ab, acd\}$,
- $-\mathsf{c}(ab) - \mathsf{c}(cd) + \mathsf{c}(abcd) \leq 0$ (3 inequalities of this type), Sperner family is $\mathcal{I} = \{ab, cd\}$,
- $-\mathsf{c}(ab) - \mathsf{c}(ac) - \mathsf{c}(ad) + \mathsf{c}(abc) + \mathsf{c}(abd) + \mathsf{c}(acd) - \mathsf{c}(abcd) \leq 0$ (4 inequalities), Sperner family is $\mathcal{I} = \{ab, ac, ad\}$,
- $-\mathsf{c}(ab) - \mathsf{c}(ac) - \mathsf{c}(bd) + \mathsf{c}(abc) + \mathsf{c}(abd) \leq 0$ (12 inequalities of this type), Sperner family is $\mathcal{I} = \{ab, ac, bd\}$,
- $-\mathsf{c}(ab) - \mathsf{c}(ac) - \mathsf{c}(bcd) + \mathsf{c}(abc) + \mathsf{c}(abcd) \leq 0$ (12 inequalities of this type), Sperner family is $\mathcal{I} = \{ab, ac, bcd\}$,
- $-\mathsf{c}(ab) - \mathsf{c}(ac) - \mathsf{c}(bc) - \mathsf{c}(cd) + 2\cdot\mathsf{c}(abc) + \mathsf{c}(acd) + \mathsf{c}(bcd) - \mathsf{c}(abcd) \leq 0$ (12 inequalities of this type), Sperner family is $\mathcal{I} = \{ab, ac, bc, cd\}$,
- $-\mathsf{c}(ab) - \mathsf{c}(ad) - \mathsf{c}(bc) - \mathsf{c}(cd)$ $+\mathsf{c}(abc) + \mathsf{c}(abd) + \mathsf{c}(acd) + \mathsf{c}(bcd) - \mathsf{c}(abcd) \leq 0$ (3 inequalities), Sperner family is $\mathcal{I} = \{ab, ad, bc, cd\}$,
- $-\mathsf{c}(bc) - \mathsf{c}(bd) - \mathsf{c}(cd) + \mathsf{c}(abc) + \mathsf{c}(abd) + \mathsf{c}(acd) + 2\cdot\mathsf{c}(bcd) - 2\cdot\mathsf{c}(abcd) \leq 1$ (4 inequalities of this type), Sperner family is $\mathcal{I} = \{a, bc, bd, cd\}$.