How matroids occur in the context of learning Bayesian network structure

Milan Studený
the Institute of Information Theory and Automation of the CAS,
Pod Vodárenskou věží 4, Prague, 182 08, Czech Republic

Abstract

It is shown that any connected matroid having a non-trivial cluster of BN variables as its ground set induces a facet-defining inequality for the polytope(s) used in the ILP approach to globally optimal BN structure learning. The result applies to well-known k-cluster inequalities, which play a crucial role in the ILP approach.

1 INTRODUCTION

The motivation for this theoretical paper is learning *Bayesian network* (BN) structure. Some hidden connection of the theory of *matroids* to a recent trend in optimal BN structure learning is revealed; specifically, matroids are related to the application of *integer linear programming* (ILP) methods in this area. To explain the motivation, in this introductory section, the latest developments in the ILP approach to learning BN structure are recalled. Matroid theory is also briefly mentioned and the structure of the rest of the paper is described.

1.1 ILP APPROACH TO LEARNING

The usual score-based approach to BN structure learning consists in maximizing a *scoring criterion* $G \mapsto \mathcal{Q}(G, D)$, where G is an acyclic directed graph and D the observed database (Neapolitan, 2004). The value $\mathcal{Q}(G, D)$ says how much the BN structure defined by the graph G fits the database D. Nevertheless, some researchers are used to identify the BN structure with the respective equivalence class of graphs and prefer to talk about learning the class.

Since the classic heuristic greedy equivalence search (GES) algorithm (Chickering, 2002) is known not to guarantee to find a globally optimal BN structure, researches started to look for alternative methods. One of them was based on the idea of *dynamic programming* (Silander, Myllymäki, 2006), which, however, was limited in the number of BN variables (= nodes of the graph) because of the exponential growth of memory demands.

This limitation has been overcome by the ILP approach based on *family-variable vector* representation of the graphs suggested independently in (Jaakkola, Sontag, Globerson, Meila, 2010) and (Cussens, 2010). An important technical step to overcome the limitation was the idea of the reduction of the search space developed by de Campos and Ji and published in a later journal paper (2011).

Jaakkola *et al.* (2010) introduced an important class of *cluster*-based inequalities approximating the respective family-variable polytope and came with a method of gradual adding these special constraints. Cussens (2010) first applied the family-variable vector representation in the restricted context of pedigree learning. However, his next paper (Cussens, 2011), which was inspired by (Jaakkola *et al.*, 2010), dealt with general BN structure learning and came with a standard *cutting plane approach* offering a more efficient way of adding the (cluster) inequality constraints, based on solving a special simple sub-ILP problem. Moreover, his experiments with general-purpose cutting planes, the so-called Gomory cuts, lead him to the idea to introduce a wider class of *k-cluster inequalities*, where k is a natural number less than the cardinality of the cluster. Bartlett and Cussens (2013) extended later the cutting plane method to a more general *branch-and-cut* approach; they included a lot of fine improvements and achieved much better running times. One of the morals of their paper was that using additional facet-defining inequalities for the family-variable polytope can speed up the computation.

An alternative ILP approach based on *characteristic-imset* representation of BN structures appeared in (Hemmecke, Lindner, Studený, 2012); its motivational sources date back to the method of imsets from (Studený, 2005). Unlike the family-variable vectors, the characteristic imsets uniquely correspond to BN structures. However, although this ILP approach is also feasible, the computational experiments reported in (Studený, Haws, 2014) have not resulted in better running times in comparison with the 2013-year version of GOBNILP software (Cussens and Bartlett, 2015).

Our recent manuscript (Cussens, Haws, Studený, 2015) has been devoted to the comparison of the facet-defining inequalities for the family-variable polytope and for the characteristic-imset polytope. Note that the facet-defining inequalities appear to be the most useful ones in the cutting plane approach to solving ILP problems; see the reasons in § 9.1-9.2 of (Wolsey, 1998). In (Cussens *et al.*, 2015), we established a one-to-one correspondence between *extreme supermodular set functions* and certain facet-defining inequalities for both polytopes. An important special case of such facet-defining inequalities are the above mentioned k-cluster constraints, which can be transformed to the characteristic-imset context.

1.2 MATROID THEORY

The theory of matroids had been formed in the 1930's as an abstract theory of independence and since then became one of important topics in combinatorial optimization. The reader is referred to Oxley's book (1992) for numerous examples of how matroids pervade various branches of discrete mathematics and for how they appear to be useful in computer science.

In (Cussens *et al.*, 2015) we observed an interesting relation of the above mentioned k-cluster inequalities to the so-called *connected uniform matroids*, which gives an elegant interpretation to those inequalities.

In this paper, I extend our former observation and, using an old result by Nguyen (1978) from matroid theory, show that any *connected matroid* over a cluster of BN variables involving at least two variables induces a facet-defining inequality both for the family-variable polytope and for the characteristic-imset polytope.

In my opinion, this theoretical result broadens the class of available facet-defining inequalities which can be used in the cutting plane approach to solving ILP problems arising in the optimal BN structure learning area.

1.3 PAPER STRUCTURE

In § 2 formal definitions of basic concepts are given: from the area of BN structure learning, the theory of polytopes, and matroid theory. The next § 3 then recalls a few observations on the cone of supermodular set functions and the results from (Cussens *et al.*, 2015), (Nguyen, 1978) that are needed. These allows one to formulate and prove the main result in § 4. An illustrating example is given in § 5. The conclusions and the discussion are in § 6.

2 BASIC CONCEPTS

Let N be a finite set of *BN variables*; to avoid a trivial case, assume $n := |N| \geq 2$. In statistical context, the elements of N correspond to random variables; in graphical context, they correspond to nodes of (acyclic directed) graphs. Its subsets $C \subseteq N$ with $|C| \geq 2$, called *clusters* in this paper, will serve as ground sets of the matroids discussed here.

2.1 STRUCTURE LEARNING CONCEPTS

The symbol $\mathsf{DAGs}\,(N)$ will denote the collection of all acyclic directed graphs over N, which means the graphs having N as the set of nodes. Given $G \in \mathsf{DAGs}\,(N)$ and $a \in N$, the symbol $\mathrm{pa}_G(a) := \{\, b \in N : \; b \to a \;\text{ in } G\}$, is 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 $i = 1, \ldots, n$, $\mathrm{pa}_G(a_i) \subseteq \{a_1, \ldots, a_{i-1}\}$; we say then the order and the graph are *consonant*.

A *BN model* is a pair (G, P), where $G \in \mathsf{DAGs}\,(N)$ and P a probability distribution on the joint sample space $\mathsf{X}_N := \prod_{a \in N} \mathsf{X}_a$, the Cartesian product of individual non-empty finite sample spaces X_a for variables $a \in N$, which factorizes according to G. An equivalent characterization of the factorization property is that P is *Markovian* with respect to G, which means it satisfies the conditional independence restrictions determined by G (Lauritzen, 1996). The *BN structure* defined by G is formally the class of Markovian probability distributions with respect to G.

Different graphs over N could be *Markov equivalent*, which means they define the same BN structure. The classic graphical characterization of equivalent graphs by Verma and Pearl (1991) states that two graphs are Markov equivalent if and only if they have the same adjacencies and immoralities. Recall that an *immorality* in $G \in \mathsf{DAGs}\,(N)$ is an induced subgraph of G of the form $a \to c \leftarrow b$, where the nodes a and b are not adjacent in G. Markov equivalence of $G, H \in \mathsf{DAGs}\,(N)$ will be denoted by $G \sim H$.

The task of *learning* the BN structure is to determine it on the basis of an observed (complete) *database* D, which is a sequence $x_1, \ldots, x_m$, $m \geq 1$ of elements of the joint sample space X_N. This is often done by maximizing some *quality criterion*, also called a *score* or a *scoring criterion*, which is a bivariate real function $(G, D) \mapsto \mathcal{Q}(G, D)$, where $G \in \mathsf{DAGs}\,(N)$ and D a database. Examples of such criteria are Schwarz's (1978) Bayesian information criterion (BIC) and Bayesian Dirichlet equivalence (BDE) score (Heckerman, Geiger, Chickering, 1995). The reader is referred to (Neapolitan, 2004) for the definition of a relevant concept of *statistical consistency*.

A crucial technical assumption from the computational point of view (Chickering, 2002) is that $\mathcal{Q}$ should be additively *decomposable*, which means, it has the form

$$\mathcal{Q}(G, D) \;=\; \sum_{a \in N} q_D(a \,|\, \mathrm{pa}_G(a))\,, \qquad (1)$$

where the summands $q_D(* \,|\, *)$ are called *local scores*. All criteria used in practice satisfy this requirement.

Given an observed database D, the goal is to maximize $G \mapsto \mathcal{Q}(G, D)$. Since the aim is learn the BN structure, a natural assumption is that the criterion $\mathcal{Q}$ to be maximized is *score equivalent* (Bouckaert, 1995), which means, for every database D and $G, H \in \mathsf{DAGs}\,(N)$,

$$\mathcal{Q}(G, D) = \mathcal{Q}(H, D) \quad \text{whenever } G \sim H.$$

Most of the criteria used in practice satisfy that.

2.2 POLYTOPES FOR LEARNING

We recall a few basic concepts from polyhedral geometry; see (Barvinok, 2002) or (Ziegler, 1995) for more details.

Below we deal with the Euclidean real vector spaces $\mathbb{R}^\Gamma$, where $\Gamma \neq \emptyset$ is a non-empty finite index set. Given two vectors $v, w \in \mathbb{R}^\Gamma$, their scalar product will be denoted by

$$\langle v, w \rangle_\Gamma := \sum_{i \in \Gamma} v_i \cdot w_i \,,$$

or just by $\langle v, w \rangle$ if there is no danger of confusion.

A *polytope* P is the convex hull of finitely many vectors from $\mathbb{R}^\Gamma$; we only consider non-empty P. The *dimension* of P, denoted by $\dim(\mathsf{P})$, is the dimension of its affine hull, which is nothing but a translate of a linear subspace. The maximal number of affinely independent vectors in P is then $\dim(\mathsf{P}) + 1$.

Given $o \in \mathbb{R}^\Gamma$ and $u \in \mathbb{R}$, a linear inequality $\langle o, v \rangle \leq u$ for $v \in \mathbb{R}^\Gamma$ is called *valid* for P if it holds for any $v \in \mathsf{P}$. The inequality is then called *tight* for a vector $w \in \mathsf{P}$ if $\langle o, w \rangle = u$. Given such valid linear inequality for P the corresponding *face* of P is its subset $F \subseteq \mathsf{P}$ of the form

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

One usually deals with valid inequalities that are tight for at least one vector $w \in \mathsf{P}$ in which case $F \neq \emptyset$. Then we will name the respective inequality *face-defining*. The function $v \in \mathbb{R}^\Gamma \mapsto \langle o, v \rangle$ is typically a linear objective to be maximized; with little abuse of terminology we will then call $o \in \mathbb{R}^\Gamma$ an *objective*.

A *facet* of a polytope P is its face of the dimension $\dim(\mathsf{P}) - 1$. The corresponding inequality will be then called *facet-defining*. Given a (non-empty) facet $F \subseteq \mathsf{P}$ of a full-dimensional polytope P in $\mathbb{R}^\Gamma$, its facet-defining inequality is unique up to a positive multiple (of both $o \in \mathbb{R}^\Gamma$ and $u \in \mathbb{R}$). A well-known fundamental result in polyhedral geometry is that every full-dimensional polytope P with non-empty facets is specified as the set of vectors $v \in \mathbb{R}^\Gamma$ satisfying all facet-defining inequalities for P. Thus, P is a *bounded polyhedron* and the facet-defining inequalities provide its minimal description in terms of inequalities.

2.2.1 Family-Variable Polytope

The index set for our family-variable vectors will be

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

Given $G \in \mathsf{DAGs}\,(N)$, the symbol η_G will denote the *family-variable vector* encoding it:

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

The *family-variable polytope* is defined as the convex hull of the collection of all such vectors:

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

Clearly, $\dim(\mathsf{F}) = |\Upsilon| = n \cdot (2^{n-1} - 1)$.

One can re-write (1) in terms of η_G in this way:

$$\mathcal{Q}(G, D) \;=\; \sum_{a \in N} \sum_{B \subseteq N \setminus \{a\}} q_D(a \,|\, B) \cdot \eta_G(a \,|\, B), \quad (2)$$

which allows one to interpret $\mathcal{Q}$ as (the restriction of) a linear function of η_G. In particular, the maximization of $\mathcal{Q}$ over $\mathsf{DAGs}\,(N)$ turns into the task to maximize a linear function with the objective $o(a \,|\, B) = q_D(a \,|\, B)$ for $(a \,|\, B) \in \Upsilon$ over the family-variable polytope F. In other words, the local scores become the components of the respective objective vector $o \in \mathbb{R}^\Upsilon$.

The assumption of score equivalence of $\mathcal{Q}$ then implies the respective objective satisfies, for every $G, H \in \mathsf{DAGs}\,(N)$,

$$G \sim H \;\Rightarrow\; \langle o, \eta_G \rangle_\Upsilon = \langle o, \eta_H \rangle_\Upsilon \,. \quad (3)$$

Thus, if (3) holds for $o \in \mathbb{R}^\Upsilon$ we will say that it is a *score equivalent objective*, abbreviated as *an SE objective*.

Given a cluster $C \subseteq N$, $|C| \geq 2$, of BN variables and a natural number $k = 1, \ldots, |C| - 1$ the inequality

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

is valid for any $G \in \mathsf{DAGs}\,(N)$: as 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. In particular, the inequality is valid for any $\eta \in \mathsf{F}$ in place of η_G and one can transform it into a standardized form:

$$\sum_{a \in C} \sum_{B \subseteq N \setminus \{a\} \,:\, |B \cap C| \geq k} \eta(a \,|\, B) \;\leq\; |C| - k \,. \quad (4)$$

We will call (4) the *k-cluster inequality* for C; its version for $k = 1$ is simply the *cluster inequality* for C. Every k-cluster inequality is facet-defining for F and the objective defining (4) is SE; see (Cussens *et al.*, 2015, Corol 4).

Example 1 Consider $N = \{a, b, c, d\} = C$ and $k = 2$. Then (4) takes the following form:

$$\begin{aligned} &[\,\eta(a\,|\,bc) + \eta(a\,|\,bd) + \eta(a\,|\,cd) + \eta(a\,|\,bcd)\,] \\ +\,&[\,\eta(b\,|\,ac) + \eta(b\,|\,ad) + \eta(b\,|\,cd) + \eta(b\,|\,acd)\,] \\ +\,&[\,\eta(c\,|\,ab) + \eta(c\,|\,ad) + \eta(c\,|\,bd) + \eta(c\,|\,abd)\,] \\ +\,&[\,\eta(d\,|\,ab) + \eta(d\,|\,ac) + \eta(d\,|\,bc) + \eta(d\,|\,abc)\,] \quad \leq \quad 2\,. \end{aligned} \tag{5}$$

2.2.2 Characteristic-Imset Polytope

The *characteristic imset* of $G \in \mathsf{DAGs}\,(N)$, introduced in (Hemmecke *et al.*, 2012) and denoted below by c_G, is an element of the vector space $\mathbb{R}^\Lambda$ where

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

Recall from (Studený, Haws, 2013) that c_G is a many-to-one linear function of η_G, the transformation is $\eta \mapsto \mathsf{c}_\eta$:

$$\mathsf{c}_\eta(S) = \sum_{a \in S} \sum_{B\,:\,S\setminus\{a\} \subseteq B \subseteq N\setminus\{a\}} \eta(a\,|\,B) \tag{6}$$

for $S \in \Lambda$. Thus, given $G \in \mathsf{DAGs}\,(N)$, (6) can serve as the definition of c_G in which one substitutes $\eta = \eta_G$. A fundamental observation is that, for $G, H \in \mathsf{DAGs}\,(N)$, $G \sim H$ if and only if $\mathsf{c}_G = \mathsf{c}_H$ (Hemmecke *et al.*, 2012). In particular, the characteristic imset is a unique representative of the corresponding BN structure.

The *characteristic-imset polytope* is defined as follows:

$$\mathsf{C} := \mathrm{conv}\,(\{\, \mathsf{c}_G \in \mathbb{R}^\Lambda \,:\, G \in \mathsf{DAGs}\,(N) \,\})\,.$$

One can show that $\dim(\mathsf{C}) = |\Lambda| = 2^n - n - 1$. Of course, C is the image of F by the linear mapping (6).

A notable fact is that any valid inequality for C induces a valid inequality for F: if $\langle \mathsf{z}, \mathsf{c}\rangle_\Lambda \leq u$, where $\mathsf{z} \in \mathbb{R}^\Lambda$ and $u \in \mathbb{R}$, is a valid inequality for $\mathsf{c} \in \mathsf{C}$, substitute (6) into $\langle \mathsf{z}, \mathsf{c}_\eta\rangle_\Lambda \leq u$ and re-arrange the terms after the components of η. Indeed, since the image of η_G by (6) is just c_G, one gets an inequality valid for any η_G, $G \in \mathsf{DAGs}\,(N)$. Moreover, the induced inequality for $\eta \in \mathsf{F}$ is given by an SE objective: if $G \sim H$, one has $\mathsf{c}_G = \mathsf{c}_G$ and, therefore, $\langle \mathsf{z}, \mathsf{c}_G\rangle_\Lambda = \langle \mathsf{z}, \mathsf{c}_H\rangle_\Lambda$.

In fact, there is a one-to-one correspondence between the valid inequalities for C and the valid inequalities for F given by SE objectives (Cussens *et al.*, 2015). Thus, these special valid inequalities for F can also be viewed as the valid inequalities for C, that is, interpreted in the context of C. This concerns many facet-defining inequalities for F: the k-cluster inequality (4) takes the following form in the characteristic-imset context, see (Cussens *et al.*, 2015, § 9):

$$\sum_{S \subseteq C,\, |S| \geq k+1} \mathsf{z}(S) \cdot \mathsf{c}(S) \leq |C| - k\,,$$

where $\mathsf{z}(S) = (-1)^{|S|-k-1} \cdot \binom{|S|-2}{|S|-k-1}$ for any such S.

Example 2 Consider $N = \{a, b, c, d\}$. Then the 2-cluster inequality for $C = \{a, b, c, d\}$ takes the form

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

Indeed, the substitution of (6) in it gives just (5).

2.3 CONCEPTS FROM MATROID THEORY

Let us recall some definitions and basic facts from matroid theory; see (Oxley, 1992, chapters 1,2,4) for more details.

A *matroid* is a pair $(C, \mathcal{I})$ where C is a finite set, called its *ground set*, and $\mathcal{I}$ a non-empty class of subsets of C, called the *independent sets* (of the matroid), which is closed under subsets: $I \in \mathcal{I}$, $J \subseteq I$ implies $J \in \mathcal{I}$ and satisfies the *independence augmentation axiom*:

> if $I, J \in \mathcal{I}$ and $|J| < |I|$
> then $a \in I \setminus J$ exists with $J \cup \{a\} \in \mathcal{I}$.

We will also say that the matroid is *on the set* C.

A number of equivalent descriptions of the matroid $(C, \mathcal{I})$ exists. Any matroid can be described by the class $\mathcal{B}$ of its *bases*, which are inclusion-maximal independent sets. The above independence augmentation axiom implies that the sets in $\mathcal{B}$ have the same cardinality. The shared cardinality of bases of a matroid is called its *rank*. A well-known fact is that $\mathcal{B} \subseteq \mathcal{P}(C)$ is the class of bases of a matroid on C iff it is a non-empty class of subsets of C satisfying the following *basis exchange axiom*:

> if $I, J \in \mathcal{B}$ and $a \in I \setminus J$
> then $b \in J \setminus I$ exists with $(I \setminus \{a\}) \cup \{b\} \in \mathcal{B}$.

The class $\mathcal{D}$ of *dependent sets* of $(C, \mathcal{I})$ consists of those subsets of C that are not independent sets. The *circuits* of the matroid are the inclusion-minimal dependent sets. A class $\mathcal{C} \subseteq \mathcal{P}(C)$ is the class of circuits of a matroid on C iff it is a class of non-empty inclusion-incomparable subsets of C satisfying the following *circuit elimination axiom*:

> if $K, L \in \mathcal{C}$, $K \neq L$ and $a \in K \cap L$
> then $M \in \mathcal{C}$ exists with $M \subseteq (K \cup L) \setminus \{a\}$.

We will also use the description of the matroid $(C, \mathcal{I})$ in terms of its *rank function*, which is a function r on $\mathcal{P}(C)$ defined as follows:

$$r(J) = \max\{\, |I| \,:\, I \subseteq J \;\&\; I \in \mathcal{I} \,\} \quad \text{for any } J \subseteq C.$$

The rank functions of matroids on C are characterized as integer-valued set functions $r : \mathcal{P}(C) \to \mathbb{Z}$ satisfying the following three conditions:

- if $I \subseteq C$ then $0 \leq r(I) \leq |I|$,

- if $J \subseteq I \subseteq C$ then $r(J) \leq r(I)$,
- if $I, J \subseteq C$ then $r(I \cup J) + r(I \cap J) \leq r(I) + r(J)$.

A set $S \subseteq C$ is called a *separator* of a matroid $(C, \mathcal{I})$ if

$$r(S) + r(C \setminus S) = r(C)\,.$$

The matroid is called *connected* if it has no other separators except for the trivial ones $S = \emptyset$ and $S = C$. Observe that if $(C, \mathcal{I})$ is connected and $|C| \geq 2$ then $\bigcup \mathcal{B} = \bigcup \mathcal{I} = C$, for otherwise $\bigcup \mathcal{I}$ is a non-trivial separator or $r \equiv 0$. An equivalent definition of a connected matroid on C is that, for every pair $a, b \in C$, $a \neq b$, a circuit $D \in \mathcal{C}$ exists with $a, b \in D$, see (Oxley, 1992, Prop 4.1.4).

The *dual matroid* to a matroid over C having $\mathcal{B} \subseteq \mathcal{P}(C)$ as its class of bases is the matroid on C having

$$\mathcal{B}^* := \{C \setminus B \,:\, B \in \mathcal{B}\}$$

as its class of bases. The formula for the rank function r^* of the dual matroid is as follows:

$$r^*(J) = |J| - r(C) + r(C \setminus J) \quad \text{for } J \subseteq C, \qquad (7)$$

see (Oxley, 1992, Prop 2.1.9). Another well-known basic fact is that the dual matroid to a connected matroid is connected as well, see (Oxley, 1992, Corol 4.2.8).

Example 3 Consider $C = \{a, b, c, d\}$ and put

$$\mathcal{B} = \{\ \{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\}\ \}.$$

Clearly, $\mathcal{B}$ is the class of bases of a matroid on C. The independent sets in it are subsets of C of cardinality at most two. The circuits are subsets of C of cardinality 3. The rank function only depends on the cardinality:

$$r(J) = \min\{\,|J|\,, 2\,\} \qquad \text{for } J \subseteq C.$$

The form of r implies that the only separators are $S = \emptyset$ and $S = C$. In particular, the matroid is connected. The dual matroid is itself.

The above example is a special matroid in a certain sense: for any integer $0 \leq k \leq |C|$ the *uniform matroid* on C of the rank k has the collection of subsets of C of the cardinality at most k as its class of independent sets.

In this paper, the attention is limited to matroids which have clusters of BN variables $C \subseteq N$, $|C| \geq 2$ as their ground sets. Any matroid $(C, \mathcal{I})$ on such cluster C can be interpreted as a matroid on N because $\mathcal{I} \subseteq \mathcal{P}(C)$ can be viewed as a class of subsets of N. This kind of *trivial extension* on N leads to the rank function $\overline{r} : \mathcal{P}(N) \to \mathbb{Z}$ given by

$$\overline{r}(S) = r(C \cap S) \qquad \text{for any } S \subseteq N.$$

3 SUPERMODULAR FUNCTIONS

In (Cussens *et al.*, 2015, § 7) a one-to-one correspondence has been established between extreme supermodular set functions and certain important facets of F, respectively of C. The relevant concepts are recalled in this section.

A real function $m : \mathcal{P}(N) \to \mathbb{R}$ on subsets of the set N of BN variables will be called *standardized* if $m(S) = 0$ for $S \subseteq N$, $|S| \leq 1$, and *supermodular* if

$$\forall\, U, V \subseteq N \;\; m(U) + m(V) \leq m(U \cup V) + m(U \cap V)\,.$$

Mirror images of supermodular functions are *submodular* functions, defined by the converse inequalities; recall from § 2.3 that rank functions of matroids are submodular.

3.1 EXTREME SUPERMODULAR FUNCTIONS

The collection of standardized supermodular functions on $\mathcal{P}(N)$, viewed as a set of vectors in $\mathbb{R}^{\mathcal{P}(N)}$, is a pointed polyhedral cone. Therefore, it has finitely many extreme rays, which makes the following definition meaningful.

A standardized supermodular function $m : \mathcal{P}(N) \to \mathbb{R}$ will be called *extreme* if it generates an extreme ray of the standardized supermodular cone. Recall that a non-zero vector v in a cone generates its extreme ray if the only summands in (positive) convex combinations of vectors from the cone yielding v are non-negative multiples of v.

Theorems 1 and 2 from (Cussens *et al.*, 2015) together give the next observation.

THEOREM 1 *An inequality $\langle o, \eta\rangle_\Upsilon \leq u$ for $\eta \in \mathbb{R}^\Upsilon$, where $o \in \mathbb{R}^\Upsilon$, $u \in \mathbb{R}$, is facet-defining for F and defined by an SE objective o iff there exists an extreme standardized supermodular function m on $\mathcal{P}(N)$ such that o is given by*

$$o(a \,|\, B) = m(\{a\} \cup B) - m(B) \quad \textit{for } (a \,|\, B) \in \Upsilon, \quad (8)$$

and u is the shared value of $\langle o, \eta_H\rangle_\Upsilon$ for full graphs H over N, that is, for such $H \in \mathsf{DAGs}(N)$ in which every pair of distinct nodes is adjacent.

Example 4 Consider $N = \{a, b, c, d\}$ and the set function

$$m(S) = \begin{cases} 2 & \text{if } S = N, \\ 1 & \text{if } |S| = 3, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for } S \subseteq N.$$

Clearly, m is a standardized supermodular function; finer arguments why m is extreme are given later (Example 5). The formula (8) leads to the inequality (5). By Theorem 1, (5) is facet-defining for the family-variable polytope F.

Moreover, Corollary 6 in (Cussens *et al.*, 2015) says what is the role of the inequalities from Theorem 1 in the characteristic-imset context; here we have in mind the correspondence of the inequalities mentioned in § 2.2.2.

COROLLARY 1 *Facet-defining inequalities* $\langle o,\eta\rangle_\Upsilon \le u$ *for* $\eta \in \mathsf{F}$ *with SE objectives correspond to facet-defining inequalities* $\langle \mathsf{z},\mathsf{c}\rangle_\Lambda \le u$ *for* $\mathsf{c} \in \mathsf{C}$ *tight for the 1-imset, which is the vector in* $\mathbb{R}^\Lambda$ *whose all components are ones.*

3.2 SUBMODULARITY AND RANK FUNCTIONS

Mirror images of supermodular functions are submodular ones, which play an important role in matroid theory. It follows from the facts mentioned in § 2.3 that every rank function of a matroid belongs to the cone of non-decreasing submodular functions r with $r(\emptyset) = 0$. This is a pointed polyhedral cone and has finitely many extreme rays.

Nguyen (1978) was interested in the question when the rank function of a matroid generates an extreme ray of that cone. The next fact follows from his Theorem 2.1.5.

THEOREM 2 *Let* C *be a non-empty finite set and* $(C,\mathcal{I})$ *a matroid on it such that* $C = \bigcup\mathcal{I}$. *Then its rank function* r *generates an extreme ray of the cone of non-decreasing submodular functions on* $\mathcal{P}(C)$ *satisfying* $r(\emptyset) = 0$ *iff the corresponding matroid* $(C,\mathcal{I})$ *is connected.*

4 MAIN RESULT

LEMMA 1 Given a connected matroid $(C,\mathcal{I})$ on $C \subseteq N$, $|C| \ge 2$ with the rank function $r : \mathcal{P}(C) \to \mathbb{Z}$, the function

$$m(S) := |C \cap S| - r(C \cap S) \quad \text{for } S \subseteq N, \qquad (9)$$

is extreme standardized supermodular function on $\mathcal{P}(N)$.

Proof: Let us denote by $\mathsf{R}[C]$, for $C \subseteq N$, the cone of submodular functions r^* on $\mathcal{P}(C)$ such that $r^*(\emptyset) = 0$ and $r^*(C) - r^*(C \setminus \{a\}) = 0$ for any $a \in C$. Any function r^* in $\mathsf{R}[C]$ is necessarily non-decreasing. The dual matroid to $(C,\mathcal{I})$ is connected; by Theorem 2, its rank function r^* given by (7) generates an extreme ray of the non-decreasing submodular cone. Since $(C,\mathcal{I})$ is connected, $\bigcup\mathcal{I} = C$ says $r(\{a\}) = 1$ for any $a \in C$. Moreover, the dual matroid to the dual matroid is again $(C,\mathcal{I})$, which gives

$$1 = r(\{a\}) = r^{**}(\{a\}) \overset{(7)}{=} 1 - r^*(C) + r^*(C \setminus \{a\})$$

for any $a \in C$; hence, r^* belongs to the smaller cone $\mathsf{R}[C]$. This easily implies that r^* generates an extreme ray of $\mathsf{R}[C]$, which fact allows one to observe by a minor consideration that its trivial extension

$$\overline{r^*}(S) := r^*(C \cap S) \quad \text{for } S \subseteq N,$$

generates an extreme ray of $\mathsf{R}[N]$. Finally, the formula

$$m(S) = \overline{r^*}(N) - \overline{r^*}(N \setminus S) \quad \text{for } S \subseteq N,$$

defines a one-to-one linear transformation of the cone $\mathsf{R}[N]$ onto the cone of standardized supermodular functions m on $\mathcal{P}(N)$ (in fact, the transformation is self-inverse). In particular, $\overline{r^*} \mapsto m$ maps generators of extreme rays to generators of extreme rays, implying that m is extreme in the respective cone. Thus, one can write for any $S \subseteq N$:

$$\begin{aligned} m(S) &= \overline{r^*}(N) - \overline{r^*}(N \setminus S) = r^*(C) - r^*(C \setminus S) \\ &\overset{(7)}{=} \{\,|C| - r(C)\,\} - \{\,|C \setminus S| - r(C) + r(C \cap S)\,\} \\ &= |C \cap S| - r(C \cap S)\,, \end{aligned}$$

which gives (9). □

Example 5 Consider $N = \{a,b,c,d\} = C$ and take the uniform matroid on C of rank 2 from Example 3. It is a connected matroid and, by Lemma 1, it induces through (9) an extreme supermodular function m from Example 4.

THEOREM 3 *Given a connected matroid* $(C,\mathcal{I})$ *on a cluster* $C \subseteq N$, $|C| \ge 2$ *of BN variables, the inequality*

$$\sum_{a\in C} \sum_{B\subseteq N\setminus\{a\}:\, \exists\, D\in\mathcal{C}\ a\in D\subseteq B\cup\{a\}} \eta(a\,|\,B) \le |C| - k, \quad (10)$$

where k *is the rank of* $(C,\mathcal{I})$ *and* $\mathcal{C}$ *the collection of its circuits, is a facet-defining inequality for* F.

Proof: By Lemma 1, (9) gives an extreme standardized supermodular function; one can apply Theorem 1 then. The upper bound u in the respective facet-defining inequality $\langle o,\eta\rangle_\Upsilon \le u$ for $\eta \in \mathsf{F}$ is the shared value $\langle o,\eta_H\rangle_\Upsilon$ for full graphs $H \in \mathsf{DAGs}(N)$. Using (8) one gets $u = m(N)$, that is, $u = m(N) \overset{(9)}{=} |C| - r(C) = |C| - k$.

The formula for the objective coefficients $o(a\,|\,B)$, where $a \in N$ and $B \subseteq N \setminus \{a\}$ (possibly empty) is then

$$\begin{aligned} o(a\,|\,B) &\overset{(8)}{=} m(\{a\} \cup B) - m(B) \\ &\overset{(9)}{=} |C \cap \{a\}| - r(C \cap (\{a\} \cup B)) + r(C \cap B)\,, \end{aligned}$$

implying $o(a\,|\,B) = 0$ if $a \in N \setminus C$. In case $a \in C$ one has

$$o(a\,|\,B) = 1 - r(C \cap (\{a\} \cup B)) + r(C \cap B) = o(a\,|\,C \cap B)\,.$$

Therefore, in the rest of the proof, we assume $a \in C$ and $B \subseteq C \setminus \{a\}$; our goal is to verify

$$o(a\,|\,B) = \begin{cases} 1 & \exists\, D \in \mathcal{C} \text{ with } a \in D \,\&\, D \subseteq B \cup \{a\}, \\ 0 & \text{otherwise,} \end{cases}$$

which clearly gives (10). We come from the above formula

$$o(a\,|\,B) = 1 - r(\{a\} \cup B) + r(B)\,. \qquad (11)$$

Having fixed $a \in C$, the coefficient are monotone

$$E \subseteq B \subseteq C \setminus \{a\} \;\Rightarrow\; o(a\,|\,B) \ge o(a\,|\,E) \qquad (12)$$

because of submodularity of r :

$$\begin{aligned} &o(a\,|\,B) - o(a\,|\,E) \\ &\overset{(11)}{=} r(B) + r(\{a\} \cup E) - r(E) - r(\{a\} \cup B) \ge 0\,. \end{aligned}$$

As $(C, \mathcal{I})$ is connected one has $r(\{a\}) = 1$ for any $a \in C$, which gives

$$o(a \,|\, \emptyset) \overset{(11)}{=} 1 - r(\{a\}) + r(\emptyset) = 1 - 1 + 0 = 0.$$

Since the dual matroid is also connected, one has

$$o(a \,|\, C \setminus \{a\}) \overset{(11)}{=} 1 - r(C) + r(C \setminus \{a\}) \overset{(7)}{=} r^*(\{a\}) = 1.$$

In particular, the objective coefficients are either zeros or ones. In case a circuit $D \in \mathcal{C}$ exists with $a \in D \subseteq B \cup \{a\}$, it is enough to show $o(a \,|\, D \setminus \{a\}) = 1$ and apply (12). Indeed, by the definition of a circuit, $D \setminus \{a\} \in \mathcal{I}$ and $r(D \setminus \{a\}) = |D| - 1$. One cannot have $r(D) = |D|$, for otherwise $D \in \mathcal{I}$ contradicts the assumption $D \in \mathcal{C}$. Thus, $r(D) = |D| - 1$ and one has

$$o(a \,|\, D \setminus \{a\}) \overset{(11)}{=} 1 - r(D) + r(D \setminus \{a\}) = 1\,.$$

It remains to show that $o(a \,|\, B) = 0$ in the complementary case that no such $D \in \mathcal{C}$ exists for B. By the definition of the rank function r, a set $J \subseteq B$ exists with $J \in \mathcal{I}$ and $|J| = r(B)$. It is enough to show $\{a\} \cup J \in \mathcal{I}$ because then $r(\{a\} \cup B) = |J| + 1$ (use submodularity of r) and

$$o(a \,|\, B) \overset{(11)}{=} 1 - r(\{a\} \cup B) + r(B) = 1 - (|J| + 1) + |J| = 0\,.$$

Thus, assume for a contradiction that $\{a\} \cup J \in \mathcal{D}$ is a dependent set and, by the definition of circuits, find $D \in \mathcal{C}$ with $D \subseteq \{a\} \cup J$. Necessarily $a \in D$, for otherwise a contradictory conclusion $J \in \mathcal{D}$ is derived. This implies $a \in D \subseteq \{a\} \cup J \subseteq \{a\} \cup B$ contradicting the assumption that no such circuit $D \in \mathcal{C}$ exists for B. □

The observation that the k-cluster inequalities (4) are facet-defining for the family-variable polytope easily follows from Theorem 3. Indeed, any uniform matroid on $C \subseteq N$, $|C| \geq 2$ of the rank k, $1 \leq k \leq |C| - 1$ is connected. This fact is illustrated by the following simple example.

Example 6 Consider $N = \{a, b, c, d\}$, $C = \{a, b, c\}$ and $k = 1$. The uniform matroid on C of rank 1 has the bases $\{a\}$, $\{b\}$ and $\{c\}$. Thus, the class of its circuits is

$$\mathcal{C} = \{\ \{a, b\}, \{a, c\}, \{b, c\}\ \}\,.$$

Since every pair of BN variables in C is contained in a circuit, it is a connected matroid. To get the specific form of the inequality (10) in this case realize that $a \in C$ is contained in two circuits $D \in \mathcal{C}$, namely in $\{a, b\}$ and in $\{a, c\}$. Thus, one has in (10) those terms $\eta(a \,|\, B)$ where $B \subseteq N \setminus \{a\}$ and either $b \in B$ ($\Leftrightarrow$ $\{a, b\} \subseteq B \cup \{a\}$) or $c \in B$. Thus, (10) takes the form

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

which is just the cluster inequality (4) for C with $k = 1$. Theorem 3 claims it is a facet-defining inequality for F.

Another instance of a uniform matroid was mentioned in Example 3; in this case, the inequality (10) turns into (5) from Example 1. The next example goes beyond the scope of k-cluster inequalities and uniform matroids.

Example 7 Consider $C = \{a, b, c, d\} = N$ and put

$$\mathcal{B} = \{\ \{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}\ \}.$$

Clearly, $\mathcal{B}$ is the class of bases of a matroid on C of the rank 2. The rank function has the form

$$r(J) = \begin{cases} 0 & \text{if } J = \emptyset, \\ 1 & \text{if } J = \{c, d\} \text{ or } |J| = 1, \\ 2 & \text{otherwise,} \end{cases} \qquad \text{for } J \subseteq C,$$

while the class $\mathcal{C}$ of its circuits is

$$\mathcal{C} = \{\ \{a, b, c\}, \{a, b, d\}, \{c, d\}\ \}.$$

As every pair of elements in C is contained in a circuit, it is a connected matroid. Theorem 3 says that the inequality

$$\begin{aligned}
&[\,\eta(a \,|\, bc) + \eta(a \,|\, bd) + \eta(a \,|\, bcd)\,] \\
&+ [\,\eta(b \,|\, ac) + \eta(b \,|\, ad) + \eta(b \,|\, acd)\,] \\
&+ [\,\eta(c \,|\, d) + \eta(c \,|\, ab) \\
&\qquad + \eta(c \,|\, ad) + \eta(c \,|\, bd) + \eta(c \,|\, abd)\,] \\
&+ [\,\eta(d \,|\, c) + \eta(d \,|\, ab) \\
&\qquad + \eta(d \,|\, ac) + \eta(d \,|\, bc) + \eta(d \,|\, abc)\,] \quad \leq \quad 2\,.
\end{aligned} \tag{13}$$

is facet-defining for F. An interesting observation is that the inequality (13) defines the so-called 4B-type facet found by Bartlett and Cussens (2013); see (13) in § 6 of their paper where $\{v_1, v_4\} = \{a, b\}$ and $\{v_2, v_3\} = \{c, d\}$.

COROLLARY 2 *The inequality* (10) *from Theorem 3 has the following form in the characteristic-imset mode:*

$$\sum_{T \in \Lambda,\, T \subseteq C} \mathsf{z}(T) \cdot \mathsf{c}(T) \leq |C| - k \quad \textit{for } \mathsf{c} \in \mathbb{R}^{\Lambda}, \tag{14}$$

$$\textit{where } \mathsf{z}(T) \quad = \quad -\sum_{L \subseteq T} (-1)^{|T \setminus L|} \cdot r(L)$$

are determined by the rank function r of the matroid. The inequality (14) *defines a facet of* C *containing the 1-imset.*

Proof: This follows from Lemma 10 and the formula (20) in (Cussens *et al.*, 2015) saying that $\langle o, \eta \rangle_{\Upsilon} = \langle \mathsf{z}, \mathsf{c}_{\eta} \rangle_{\Lambda}$ where the coefficients $\mathsf{z}(T)$ for $T \in \Lambda$ are given by the Möbius transform of the corresponding standardized supermodular function m, that is, by

$$\mathsf{z}(T) = \sum_{L \subseteq T} (-1)^{|T \setminus L|} \cdot m(L) \qquad \text{for } T \in \Lambda.$$

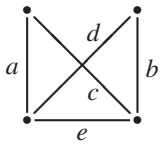

Figure 1: Edges of the graph define a matroid.

In our case, m is given by (9), which implies $\mathsf{z}(T) = 0$ whenever $T \setminus C \neq \emptyset$. Moreover, the Möbius transform of the first term in (9) vanishes for $T \in \Lambda$, $T \subseteq C$, which gives (14). The second claim follows from Corollary 1. □

Example 8 Consider again the matroid from Example 7. The formula (14) applied to the rank function r gives

$$\mathsf{z}(T) = \begin{cases} -1 & \text{if } T = C, \\ 1 & \text{if } T \in \mathcal{C}, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for } T \in \Lambda,\ T \subseteq C.$$

In particular, the inequality (13) has the following form in the characteristic-imset mode:

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

5 FIVE VARIABLES EXAMPLE

Note that in case of four BN variables there is no other matroid-based facet-defining inequality for F except the k-cluster inequalities and (13). However, there are more matroid-based inequalities in case of five BN variables.

An important class of matroids are the so-called *graphic matroids* (Oxley, 1992, § 1.1). In fact, any undirected graph $\mathcal{G}$ defines a matroid on *the set of its edges*. Specifically, a set I of edges in $\mathcal{G}$ is considered to be *independent* (in the graphic matroid) if the edge-subgraph of $\mathcal{G}$ consisting of edges in I is a *forest*, that is, has no undirected cycle.

The *circuits* of this graphic matroid are then the sets D of edges in $\mathcal{G}$ forming edge-minimal cycles in $\mathcal{G}$, which means the removal of any edge from D results in a forest. The idea is illustrated by an example.

Example 9 Consider $C = \{a, b, c, d, e\} = N$ and define a matroid on C by means of the graph in Figure 1, where the edges are identified with the elements of C. It makes no problem to observe that the matroid has three circuits:

$$\mathcal{C} = \{\ \{a, b, c, d\}, \{a, c, e\}, \{b, d, e\}\ \}\,,$$

while the number of bases is eight: these are all 3-element subsets of C except for $\{a, c, e\}$ and $\{b, d, e\}$. Of course, these are just the sets of edges defining spanning trees for the graph from Figure 1. It is easy to see that the matroid is connected and has rank $k = 3$. Like in Example 6 one can determine the terms $\eta(* \,|\, B)$ which occur in (10). For example, $a \in C = N$ is contained in two circuits $D \in \mathcal{C}$, namely in $\{a, c, e\}$ and $\{a, b, c, d\}$. In particular, one has in (10) those terms $\eta(a \,|\, B)$ where $B \subseteq N \setminus \{a\}$ and either $\{c, e\} \subseteq B$ or $\{b, c, d\} \subseteq B$. The same principle applies to b, c, d and e which results in the following abbreviated form of (10):

$$\begin{aligned} &\sum_{ce \subseteq B \,\vee\, bcd \subseteq B} \eta(a \,|\, B) + \sum_{de \subseteq B \,\vee\, acd \subseteq B} \eta(b \,|\, B) \\ + &\sum_{ae \subseteq B \,\vee\, abd \subseteq B} \eta(c \,|\, B) + \sum_{be \subseteq B \,\vee\, abc \subseteq B} \eta(d \,|\, B) \\ &+ \sum_{ac \subseteq B \,\vee\, bd \subseteq B} \eta(e \,|\, B) \quad \leq \quad 2\,. \end{aligned} \tag{15}$$

Thus, by Theorem 3, the inequality (15) if facet-defining for F. We leave to the reader to derive the rank function r of the matroid and observe that its Möbius transform only has 4 non-zero values: -1 for circuits in $\mathcal{C}$ and $+1$ for $C = N$. In particular, by Corollary 2, (15) has the following simple form in the characteristic-imset mode:

$$\mathsf{c}(abcd) + \mathsf{c}(ace) + \mathsf{c}(bde) - \mathsf{c}(abcde) \leq 2\,.$$

Example 9 indicated a way one can search for connected matroids, and, thus, for facet-defining inequalities to be used in the ILP approach to BN structure learning. Graphic matroids are common examples of matroids; but there are many matroids which are not graphic, like the uniform matroid from Example 3.

To utilize fully the matroid-based approach some computer scientists may take the following exhaustive "brute-force" approach: given a (presumably) small cluster C, $|C| \geq 2$ generate by means of a computer all (permutation) types of classes $\mathcal{C}$ of inclusion-incomparable subsets of C such that $\forall\, a, b \in C$, $a \neq b$, a set $D \in \mathcal{C}$ exists with $a, b \in D$. Then one can check (again with the help of a computer) which of them satisfy the circuit elimination axiom. In this way one gets all types of connected matroids on C and can transform them into facet-defining inequalities for the family-variable polytope or for the characteristic imset polytope.

Other people may prefer to search in the literature on matroid theory. Indeed, researcher in this area have generated various catalogues of (types of) matroids on small ground sets; see, for example (Mayhew, Royle, 2008).

6 CONCLUSIONS

Theorem 3 implies that every connected matroid on a non-trivial cluster of BN variables induces a facet-defining inequality for the family-variable polytope; Corollary 2 says what is the form of that inequality in the context of the characteristic-imset polytope.

This is a quite general theoretical result because the well-known k-cluster inequalities, which play the key role in contemporary ILP approaches to BN structure learning, can be derived in this way. Specifically, they correspond to the prominent (connected) uniform matroids.

The significance of the paper is mainly theoretical: the area of statistical learning is related to a seemingly remote field in discrete mathematics, namely to matroid theory. Although matroids were previously known to have many applications in *combinatorial optimization*, this particular intimate link to BN structure learning could be surprising. The advantage of the matroid-based approach to learning is that the inequalities are easy to find and the verification that they are facet-defining is immediate since testing whether a matroid is connected is easy. The result is applicable in both ILP approaches to BN structure learning, that is, both in the context of the family-variable polytope and in the context of the characteristic-imset polytope.

However, the result also has some potential for practical future use because it may lead to bettering certain currently used algorithms. Let me recall in more detail the original motivation, which was the ILP approach to BN structure learning. I have in mind the *cutting plane method* where one solves an ILP optimization problem by the method of iterative reduction of the feasible set. The solution to a *linear relaxation* problem, which is a (non-integer) linear program with a larger feasible set, specified be a small number of inequalities, is typically a fractional vector. The next step is to solve the *separation problem*, that is, to find an inequality from a reservoir of available inequalities (for example from the class of cluster inequalities) which cuts the current fractional solution from the true feasible region, which is the polytope defined as the convex hull of integer vectors in the feasible set, see (Wolsey, 1998, § 8.5)

From the point of view of computational efficiency, it is essential to find such inequality which approximates the polytope as close as possible near the current solution. This leads to the suggestion to look for the most violated inequalities by the current fractional solution; see also the heuristic justification in (Cussens, 2011, § 4.1).

The presented result broadens the reservoir of available facet-defining inequalities in the ILP approach to BN structure learning. In fact, it is claimed by Bartlett and Cussens in (2013, § 6) that the inequality (13) from Example 7 has appeared to be particularly useful in their experiments. Moreover, the other facet-defining inequalities for F, that is, those not based on matroids, have not appeared to be very useful. Their empirical observations are the basis for my hope that the matroid-based inequalities may bring some further progress in this area, perhaps even resulting in better future running times.

Nevertheless, let me emphasize that additional problems have to be solved to reach the practical applicability of general matroid-based inequalities. More specifically, it is necessary to solve the corresponding separation problem, that is, to design a speedy algorithm for finding the (most) violated inequalities by a current (fractional) solution in the class of all general matroid-based inequalities. Thus, the next step towards the practical application of the matroid-based inequalities should be a proposal for such algorithm.

Acknowledgements

The research on this topic has been supported by the grant GAČR n. 13-20012S. I am indebted to my colleague Fero Matúš for giving me some guidance in matroid theory.

References

M. Bartlett, J. Cussens (2013). Advances in Bayesian network learning using integer programming. In Uncertainty in Artificial Intelligence 29, AUAI Press, 182-191.

A. Barvinok (2002). *A Course in Convexity.* Graduate Studies in Mathematics 54, Providence: American Mathematical Society.

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

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

J. Cussens (2010). Maximum likelihood pedigree reconstruction using integer programming. In Proceedings of the Workshop on Constraint Based Methods for Bioinformatics (WCBMB), 9-19.

J. Cussens (2011). Bayesian network learning with cutting planes. In F. Cozman, A. Pfeffer (eds.) Uncertainty in Artificial Intelligence 27, AUAI Press, 153-160.

J. Cussens, M. Bartlett (2015). GOBNILP software. Available at *www.cs.york.ac.uk/aig/sw/gobnilp*.

J. Cussens, D. Haws, M. Studený (2015). Polyhedral aspects of score equivalence in Bayesian network structure learning. Submitted to *Mathematical Programming A*, also available at *arxiv.org/abs/1503.00829.*

C.P. de Campos, Q. Ji (2011). Efficient structure learning Bayesian networks using constraints. *Journal of Machine Learning Research* **12**:663-689.

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

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

T. Jaakkola, D. Sontag, A. Globerson, M. Meila (2010).

Learning Bayesian network structure using LP relaxations. In Y.W. Teh, M. Titterington (eds.) JMLR Workshop and Conference Proceedings 9: AISTATS 2010, 358-365.

S.L. Lauritzen (1996). *Graphical Models.* Oxford: Clarendon Press.

D. Mayhew, G.F. Royle (2008). Matroids with nine elements. *Journal of Combinatorial Theory B* **98**:415-431.

R.E. Neapolitan (2004). *Learning Bayesian Networks.* Upper Saddle River: Pearson Prentice Hall.

H.Q. Nguyen (1978). Semimodular functions and combinatorial geometries. *Transaction of the American Mathematical Society* **238**:355-383.

J.G. Oxley (1992). *Matroid Theory.* Oxford: Oxford University Press.

G.E. Schwarz (1978). Estimation of the dimension of a model. *Annals of Statistics* **6**:461-464.

T. Silander, P. Myllymäki (2006). A simple approach for finding the globally optimal Bayesian network structure. In R. Dechter, T. Richardson (eds.) Uncertainty in Artificial Intelligence 22, AUAI Press, 445-452.

M. Studený (2005). *Probabilistic Conditional Independence Structures.* London: Springer.

M. Studený, D.C. Haws (2013). On polyhedral approximations of polytopes for learning Bayesian networks. *Journal of Algebraic Statistics* **4**:59-92.

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

T. Verma, J. Pearl (1991). Equivalence and synthesis of causal models. In Uncertainty in Artificial Intelligence 6, Elsevier, 220-227.

L.A. Wolsey (1998). *Integer Programming*. New York: John Wiley.

G.M. Ziegler (1995). *Lectures on Polytopes.* New York: Springer.