International Journal of Approximate Reasoning 51 (2010) 573–586

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

A geometric view on learning Bayesian network structures

Milan Studený [a,*], Jiří Vomlel [a], Raymond Hemmecke [b]

[a] Institute of Information Theory and Automation of the ASCR, Prague, Czech Republic
[b] Zentrum Mathematik, Technische Universität Munich, Germany

ARTICLE INFO

Article history:
Received 27 November 2008
Accepted 20 April 2009
Available online 28 January 2010

Keywords:
Learning Bayesian networks
Standard imset
Inclusion neighborhood
Geometric neighborhood
GES algorithm

ABSTRACT

We recall the basic idea of an algebraic approach to learning Bayesian network (BN) structures, namely to represent every BN structure by a certain (uniquely determined) vector, called a *standard imset*. The main result of the paper is that the set of standard imsets is the set of vertices (=extreme points) of a certain polytope. Motivated by the geometric view, we introduce the concept of the *geometric neighborhood* for standard imsets, and, consequently, for BN structures. Then we show that it always includes the *inclusion neighborhood*, which was introduced earlier in connection with the greedy equivalence search (GES) algorithm. The third result is that the global optimum of an affine function over the polytope coincides with the local optimum relative to the geometric neighborhood.

To illustrate the new concept by an example, we describe the geometric neighborhood in the case of three variables and show it differs from the inclusion neighborhood. This leads to a simple example of the failure of the GES algorithm if data are not "generated" from a perfectly Markovian distribution. The point is that one can avoid this failure if the search technique is based on the geometric neighborhood instead. We also found out what is the geometric neighborhood in the case of four and five variables.

1. Introduction

The motivation for this theoretical paper is learning a Bayesian network (BN) structure from data by the method of maximizing a quality criterion (=the score and search method). By a *quality criterion*, also named a *score metric* or simply a *score* by some other authors, we mean a real function $\mathcal{Q}$ of the BN structure, usually represented by a graph G, and of the database D. The value $\mathcal{Q}(G,D)$ "evaluates" how the BN structure given by G fits the observed database D.

An important related question is how to represent a BN structure in the memory of a computer. Formerly, each BN structure was represented by an arbitrary acyclic directed graph defining it, which led to the non-uniqueness in its description. Later, researchers calling for methodological simplification came up with the idea to represent every BN structure with a unique representative. The most popular graphical representative is the *essential graph*. It is a chain graph describing shared features of acyclic directed graphs defining the BN structure. The adjective "essential" was proposed by Andersson et al. [2], who gave a graphical characterization of (graphs that are) essential graphs.

Since direct maximizing a quality criterion $\mathcal{Q}$ seems, at least at first sight, to be infeasible, various *local search methods* have been proposed. The basic idea is that one introduces a neighborhood relation between BN structure representatives, also named *neighborhood structure* by some authors [3]. Then one is trying to find a local maximum with respect to the chosen neighborhood structure. This is an algorithmically simpler task because one can utilize various greedy search techniques for this purpose. On the other hand, the algorithm can get stuck in a local maximum and fail to find the global maximum. A

* Corresponding author.
E-mail address: studeny@utia.cas.cz (M. Studený).

0888-613X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2010.01.014

typical example of these techniques is the *greedy equivalence search* (GES) algorithm proposed by Meek [12]. The neighborhood structure utilized in this algorithm is the *inclusion neighborhood*, which comes from the conditional independence interpretation of BN structures. Chickering [5] proposed a modification of the GES algorithm, in which he used essential graphs as (unique) BN structure representatives.

There are two important technical requirements on a quality criterion $\mathcal{Q}$ brought in connection with the local search methods, namely to make them computationally feasible. One of them is that $\mathcal{Q}$ should be *score equivalent* [3], which means it ascribes the same value to equivalent graphs (=defining the same BN structure, that is, having ascribed the same essential graph). The other requirement is that $\mathcal{Q}$ should be *decomposable* [5], which means that $\mathcal{Q}(G,D)$ decomposes into contributions which correspond to the factors in the factorization according to the graph G.

The basic idea of an algebraic approach to learning BN structures, presented in Chapter 8 of [18], is to represent both the BN structure and the database with real vectors. More specifically, an algebraic representative of the BN structure defined by an acyclic directed graph G is a certain integer-valued vector u_G, called the *standard imset* (for G). It is also a unique BN structure representative because $u_G = u_H$ for equivalent graphs G and H. Another advantage of standard imsets is that one can read practically immediately from the differential vector $u_G - u_H$ whether the BN structures defined by graphs G and H are neighbors in the sense of inclusion neighborhood. However, the crucial point is that every score equivalent and decomposable criterion $\mathcal{Q}$ is an affine function (=linear function plus a constant) of the standard imset. More specifically, it is shown in Section 8.4.2 of [18] that one has

$$\mathcal{Q}(G,D) = s_D^{\mathcal{Q}} - \langle t_D^{\mathcal{Q}}, u_G \rangle,$$

where $s_D^{\mathcal{Q}}$ is a real number, $t_D^{\mathcal{Q}}$ a vector of the same dimension as the standard imset u_G (these parameters both depend solely on the database D and the criterion $\mathcal{Q}$) and $\langle *, * \rangle$ denotes the scalar product. The vector $t_D^{\mathcal{Q}}$ is named the *data vector* (relative to the criterion $\mathcal{Q}$).

We believe that the above-mentioned result paves the way for future application of efficient linear programming methods in the area of learning BN structures. This paper is a further step in this direction: its aim is to enrich the algebraic approach by a geometric view. One can imagine the set of all standard imsets over a fixed *set N of variables* (=the set of nodes for graphs) as the set of points in the corresponding Euclidean space (of a higher dimension). The main result of the paper is that it is the set of vertices (=extreme points) of a certain polytope P. We derive this geometric fact from former theoretical results on BNs. A consequence of the result is as follows: since every "reasonable" quality criterion $\mathcal{Q}$ can be viewed as (the restriction of) an affine function on the corresponding Euclidean space, the task to maximize $\mathcal{Q}$ over BN structures is equivalent to the task to maximize an affine function over the above-mentioned polytope.

This maximization problem has been treated thoroughly within the linear programming community. Deep algorithmic theory was developed and fast software codes are available that can handle problems with vectors having thousands or even millions of components. A classic tool to solve linear programming problems is the *simplex method* [15]. One of possible interpretations of this method is that it is a kind of an augmentation algorithm (=a search method), in which one moves between vertices of a polytope along its edges (in the geometric sense) until an optimal vertex is reached. Although it has not yet been decided whether the simplex method can be modified to get a polynomial-time algorithm,[1] it performs extremely well in practice.

In order to apply the (classic) simplex method, one needs an explicit description of the polytope via finitely many linear inequalities. Such a description always exists by Weyl–Minkowski theorem [15], which says that any polytope can equivalently be introduced as a (bounded) polyhedron, that is, the intersection of finitely many (affine) half-spaces. Note that the implicit knowledge about these inequalities is often enough to solve a linear program at hand. As concerns the standard imset polytope P, for $|N| = 3$ and $|N| = 4$ a minimal such system has 13 and 154 inequalities, respectively. However, it is already a challenge to existing software packages to find such a minimal inequality description of P for $|N| = 5$ (given by 8782 vertices). Thus, for general $|N|$, one definitely needs to classify these inequalities implicitly in order to apply the classic tools from linear programming.

Because such a "polyhedral" description of the polytope P is not available for arbitrarily high $|N|$ we propose an alternative approach that mimics the walk along the edges in the simplex method. The basic idea is to introduce the concept of *geometric neighborhood* for standard imsets, and, therefore, for BN structures as well. The standard imsets u_G and u_H will be regarded as (geometric) neighbors if the line-segment connecting them is an edge of the polytope P in the geometric sense. An important observation is that the above-mentioned inclusion neighborhood is always contained in the geometric one. Nevertheless, the crucial fact is that, for any affine function, its local maximum relative to the geometric neighborhood is necessarily its global maximum over the polytope. We give the proof of both these observations in the paper.

Thus, once one succeeds in characterizing explicitly or implicitly the geometric neighborhood structure, one can apply the following augmentation algorithm: start at a standard imset (=a vertex of P), for example $u_G = 0$, and keep moving to (geometrically) adjacent standard imsets (=via edges of P) with a higher value of the criterion $\mathcal{Q}$ until one reaches its local maximum relative to the geometric neighborhood. Thus, by the above-mentioned observation, the *global maximum* of $\mathcal{Q}$ over the standard imsets must be found. Note that the resulting standard imset can then be transformed to the corresponding essential graph by a polynomial-time algorithm described in [19].

[1] This is a long-standing open question in linear programming.

We have succeeded to compute the geometric neighborhood structure for $|N| = 3,\ 4,\ 5$. Our computations suggest that, for most standard imsets, there are many more geometric neighbors than the inclusion neighbors. To illustrate the concept of the geometric neighborhood we characterize it for three variables in the paper. The notions of the inclusion neighborhood and of the geometric one already differ in this elementary case. This observation has a simple but notable consequence: the GES algorithm, which is based on the inclusion neighborhood, may fail to find the global maximum of a quality criterion. We give such an example and claim that this is an inevitable defect of the inclusion neighborhood, which may occur whenever a special *data faithfulness assumption* is not guaranteed. In our view, the data faithfulness relative to a perfectly Markovian distribution is a very strong unrealistic assumption except for the case of artificially generated data.

In the Conclusions we discuss further research directions.

2. Basic concepts

In this section we recall basic concepts and some results concerning learning BN structures.

2.1. Bayesian network structures

One of the possible definitions of a (discrete) *Bayesian network* is that it is a pair (G, P), where G is an acyclic directed graph over a (non-empty finite) set of nodes (=variables) N and P a discrete probability distribution over N that (recursively) factorizes according to G [13]. A well-known fact is that P factorizes according to G if and only if it is Markovian with respect to G, which means it satisfies the conditional independence restrictions determined by the graph G through the corresponding (directed) separation criterion [14,11]. Having fixed (non-empty finite) individual sample spaces X_i for variables $i \in N$, the respective (BN) *statistical model* is the class of all probability distributions P on the joint sample space $\mathsf{X}_N \equiv \prod_{i\in N}\mathsf{X}_i$ that factorize according to G. To name the shared features of the distributions in this class one can use the phrase "*BN structure*". Of course, the structure is determined by the graph G, but it may happen that two different graphs over N describe the same structure.

2.1.1. Equivalence of graphs

Two acyclic directed graphs over N will be named *Markov equivalent* if they define the same BN statistical model. To avoid trivial cases and troubles, throughout (the rest of) the paper we assume $|\mathsf{X}_i| \geqslant 2$ for every $i \in N$, that is, every variable has at least two different possible values. In this case, the graphs are Markov equivalent iff they are *independence equivalent*, by which is meant they determine the same collection of conditional independence restrictions – cf. Section 2.2 in [13]. Both Frydenberg [8], and Verma and Pearl [20] gave a classic graphical characterization of independence equivalence: two acyclic directed graphs G and H over N are independence equivalent if and only if they have the same underlying undirected graph and *immoralities*, that is, induced subgraphs of the form $a \to c \leftarrow b$, where $[a, b]$ is not an edge in the graph.

2.1.2. Learning a BN structure

The goal of (structural) learning is to determine the BN structure on the basis of data. These are assumed to have the form of a *complete database* $D : x^1, \ldots, x^d$ of the length $d \geqslant 1$, that is, of a sequence of elements of the joint sample space X_N. Provided the individual sample spaces X_i with $|\mathsf{X}_i| \geqslant 2$ for $i \in N$ are fixed, let $\mathsf{DATA}(N, d)$ denote the collection of all databases over N of the length d. Moreover, let $\mathsf{DAGS}(N)$ denote the collection of all acyclic directed graphs over N. Then we take a real function $\mathcal{Q}$ on $\mathsf{DAGS}(N) \times \mathsf{DATA}(N, d)$ for a *quality criterion*. The value $\mathcal{Q}(G, D)$ should reflect how the statistical model determined by G is suitable for explaining the (occurrence of the) database D. The learning procedure based on $\mathcal{Q}$ then consists in maximizing the function $G \mapsto \mathcal{Q}(G, D)$ over $G \in \mathsf{DAGS}(N)$ if the database $D \in \mathsf{DATA}(N, d),\ d \geqslant 1$ is given.

A classic example of a quality criterion is Jeffreys–Schwarz *Bayesian information criterion* (BIC), defined as the maximum of the likelihood minus a penalty term, which is a multiple of the number of free parameters in the statistical model [16]. To give a direct formula for BIC (in our case) we need a notational convention. Given $i \in N$, let $r(i)$ denote the cardinality $|X_i|$, $pa_G(i) \equiv \{j \in N;\ j \to i\}$ the set of *parents* of i in $G \in \mathsf{DAGS}(N)$, and $q(i, G) \equiv |\prod_{j\in pa_G(i)}\mathsf{X}_j|$ the number of parent configurations for i (in G).[2] Provided $i \in N$ is fixed, the letter k will serve as a generic symbol for (the code of) an element of X_i (=a node configuration) while j as a generic symbol for (the code of) a parent configuration. Given a database D of the length $d \geqslant 1$ let d_{ijk} denote the number of occurrences in D of the (marginal) parent-node configuration encoded by j and k; put $d_{ij} = \sum_{k=1}^{r(i)} d_{ijk}$. Here is the formula – see Corollary 8.2 in [18]:

$$\mathsf{BIC}(G, D) = \sum_{i\in N}\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)} d_{ijk}\cdot \ln\frac{d_{ijk}}{d_{ij}} - \frac{\ln d}{2}\cdot\sum_{i\in N} q(i, G)\cdot[r(i) - 1]. \tag{1}$$

In this brief overview we omit the question of statistical consistency of quality criteria; we refer the reader to the literature on this topic [5,13]. However, we recall two other important concepts. A quality criterion $\mathcal{Q}$ will be named *score equivalent* [3] if, for every $D \in \mathsf{DATA}(N, d),\ d \geqslant 1$,

[2] If $pa_G(i) = \emptyset$ then $q(i, G) = 1$ by a convention, because the only parent configuration for i is the empty configuration then.

$$\mathcal{Q}(G,D) = \mathcal{Q}(H,D) \quad \text{if } G,H \in \mathsf{DAGS}(N) \text{ are independence equivalent.}$$

Moreover, $\mathcal{Q}$ will be called *decomposable* [5] if there exists a collection of functions $q_{i|B} : \mathsf{DATA}(\{i\} \cup B, d) \to \mathbb{R}$, where $i \in N$, $B \subseteq N \setminus \{i\}$, $d \geqslant 1$, such that, for every $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N,d)$, $d \geqslant 1$ one has

$$\mathcal{Q}(G,D) = \sum_{i \in N} q_{i|pa_G(i)}(D_{\{i\} \cup pa_G(i)}),$$

where $D_A : x_A^1, \ldots, x_A^d$ denotes the projection of the database D to the marginal space $\mathsf{X}_A \equiv \prod_{i \in A} \mathsf{X}_i$ for $\emptyset \neq A \subseteq N$.

2.1.3. Inclusion neighborhood

The basic idea of local search methods for maximizing a quality criterion (=score and search methods) has already been explained in the Introduction. Now, we define the inclusion neighborhood formally. Given $G \in \mathsf{DAGS}(N)$, let $\mathscr{I}(G)$ denote the collection of conditional independence restrictions determined by G. Given $G, H \in \mathsf{DAGS}(N)$, if $\mathscr{I}(H) \subset \mathscr{I}(G)$,[3] but there is no $F \in \mathsf{DAGS}(N)$ with $\mathscr{I}(H) \subset \mathscr{I}(F) \subset \mathscr{I}(G)$, then we say H and G are *inclusion neighbors*. Of course, this terminology can be extended to the corresponding BN structures and their representatives.

Note that one can test graphically whether $G, H \in \mathsf{DAGS}(N)$ are inclusion neighbors; this follows from transformational characterization of inclusion $\mathscr{I}(H) \subseteq \mathscr{I}(G)$ provided by Chickering [5].

2.1.4. Essential graphs

Given an (independence) equivalence class $\mathscr{G}$ of acyclic directed graphs over N, the respective *essential graph* G^* is a hybrid graph (=a graph with both directed and undirected edges) defined as follows:

- $a \to b$ in G^* if $a \to b$ in every $G \in \mathscr{G}$,
- $a - b$ in G^* if there are $G, H \in \mathscr{G}$ such that $a \to b$ in H and $a \leftarrow b$ in G.

It is always a chain graph (=an acyclic hybrid graph); this follows from graphical characterization of (graphs that are) essential graphs by Andersson et al. [2]. Chickering [5] used essential graphs as unique graphical BN structure representatives in his version of the GES algorithm.

2.2. Standard imsets

By an *imset* u over N will be meant an integer-valued function on the power set of N, that is, on $\mathscr{P}(N) \equiv \{A;\ A \subseteq N\}$. We will regard it as a vector whose components are integers and are indexed by subsets of N. Actually, any real function $m : \mathscr{P}(N) \to \mathbb{R}$ can be interpreted as a (real) vector in the same way, that is, identified with an element of $\mathbb{R}^{\mathscr{P}(N)}$. The symbol $\langle m, u \rangle$ will denote the scalar product of two vectors of this type:

$$\langle m, u \rangle \equiv \sum_{A \subseteq N} m(A) \cdot u(A).$$

A trivial example of an imset is the *zero imset*, which ascribes the value zero to every subset $A \subseteq N$; to denote it we use the universal zero symbol 0. To write formulas for imsets we introduce a natural notational convention. Given $A \subseteq N$, the symbol δ_A will denote the following basic imset (=vector):

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

The collection of imsets $\{\delta_A;\ A \subseteq N\}$ forms a linear basis of the Euclidean space $\mathbb{R}^{\mathscr{P}(N)}$. This allows one to express any imset over N as a linear combination of these basic terms.

By an *elementary imset* is meant an imset

$$u_{\langle a,b|C \rangle} = \delta_{\{a,b\} \cup C} + \delta_C - \delta_{\{a\} \cup C} - \delta_{\{b\} \cup C},$$

where $C \subseteq N$ and $a, b \in N \setminus C$ are distinct. It is quite simple imset, it has only four non-zero values: it ascribes +1 to C and $\{a,b\} \cup C$ and -1 to $\{a\} \cup C$ and $\{b\} \cup C$. In our algebraic framework this imset encodes an elementary conditional independence statement $a \perp\!\!\!\perp b | C$.

Given $G \in \mathsf{DAGS}(N)$, the *standard imset* for G, denoted by u_G, is given by the formula

$$u_G = \delta_N - \delta_\emptyset + \sum_{i \in N} \left\{ \delta_{pa_G(i)} - \delta_{\{i\} \cup pa_G(i)} \right\}. \tag{2}$$

Note that, in this formula, the terms can both cancel each other and sum up. For example, there is always at least one $i \in N$ with $pa_G(i) = \emptyset$ and, therefore, the term $-\delta_\emptyset$ cancels against one of the terms $\delta_{pa_G(i)}$ with $pa_G(i) = \emptyset$. However, there could be

[3] Here, $\mathscr{I} \subset \mathscr{J}$ denotes strict inclusion, that is, $\mathscr{I} \subseteq \mathscr{J}$ but $\mathscr{I} \neq \mathscr{J}$.

several $i \in N$ with $pa_G(i) = \emptyset$; thus, the value $u_G(\emptyset)$ will be the number of $i \in N$ with $pa_G(i) = \emptyset$ minus one. To illustrate the formula (2) observe that the standard imset for the empty graph over N has the form

$$u = \delta_N - \sum_{i \in N} \delta_i + (|N| - 1) \cdot \delta_\emptyset,$$

and the standard imset for any complete (acyclic directed) graph over N is the zero imset. Another interesting observation is that any elementary imset is the standard imset for a special acyclic directed graph.

It follows from (2) that u_G has at most $2 \cdot |N|$ non-zero values. Hence, one can only keep its non-zero values in the memory of a computer, which means that the memory demands for representing standard imsets are polynomial in the number of variables.

Note that every standard imset u_G can be written (in a non-unique way) as the sum of elementary imsets, possibly an empty sum (for the zero imset). However, the number of summands in such decomposition only depends on u_G. This number will be called the *degree* of u_G and denoted by $deg(u_G)$. The degree corresponds to the number $\mathsf{a}(G)$ of arrows in G as follows:

$$deg(u_G) = \frac{1}{2} \cdot |N| \cdot (|N| - 1) - \mathsf{a}(G), \tag{3}$$

and can be computed from u_G directly by a special formula

$$deg(u_G) = \langle m_*, u_G \rangle, \quad \text{where } m_* : \mathscr{P}(N) \to \mathbb{Z} \tag{4}$$

is given by $m_*(A) = \frac{1}{2} \cdot |A| \cdot (|A| - 1)$ for $A \subseteq N$; for the proofs of these facts see Lemma 7.1 and Section 4.2 in [18]. To illustrate the formulas (3) and (4) recall that the degree of the zero imset is 0 while the maximal degree $\frac{1}{2} \cdot |N| \cdot (|N| - 1)$ among standard imsets over N is achieved by the above-mentioned imset for the empty graph. Of course, elementary imsets have the degree 1.

It was shown as Corollary 7.1 in [18] that, given $G, H \in \mathsf{DAGS}(N)$, one has $u_G = u_H$ if and only if they are independence equivalent. Moreover, Corollary 8.4 in [18] implies that $G, H \in \mathsf{DAGS}(N)$ are inclusion neighbors if and only if either $u_G - u_H$ or $u_H - u_G$ is an elementary imset. Finally, Lemmas 8.3 and 8.7 in [18] together claim that every score equivalent and decomposable criterion $\mathcal{Q}$ necessarily has the form:

$$\mathcal{Q}(G, D) = s_D^{\mathcal{Q}} - \langle t_D^{\mathcal{Q}}, u_G \rangle \quad \text{for } G \in \mathsf{DAGS}(N),\ D \in \mathsf{DATA}(N, d),\ d \geqslant 1, \tag{5}$$

where the constant $s_D^{\mathcal{Q}} \in \mathbb{R}$ and the *data vector* $t_D^{\mathcal{Q}} : \mathscr{P}(N) \to \mathbb{R}$ do not depend on G. These parameters only depend on the database D and the criterion $\mathcal{Q}$.

The reader may object that the dimension of $t_D^{\mathcal{Q}}$ grows exponentially with $|N|$, making the method unfeasible for many "real-world" problems. However, since $2^{|N|} \leqslant |\mathsf{X}_N|$, the representation of a database D in the form of a data vector may even appear to be more effective than (one of the traditional ways of its representation) in the form of a table of counts (=contingency table)! Another point is that to compute $\langle t_D^{\mathcal{Q}}, u_G \rangle$ one only needs at most $2 \cdot |N|$ components of the data vector (since u_G has at most $2 \cdot |N|$ non-zero values). Thus, one can decide to compute the necessary components of $t_D^{\mathcal{Q}}$ only when one actually needs them during a search procedure. In brief, we believe that whenever one is able to represent the database in the memory of a computer then one should be able to take care of the data vector as well.

3. Some geometric concepts

Here we recall some well-known concepts and facts from the theory of convex polytopes. For a more thorough treatment and proofs see for example [15,22].

3.1. Polytopes and polyhedra

Polytopes and polyhedra are special subsets of the Euclidean vector space $\mathbb{R}^K$ for some non-empty finite set K. The points in this space are vectors $\mathsf{v} = [v_s]_{s \in K}$. Given $\mathsf{x}, \mathsf{v} \in \mathbb{R}^K$ their scalar product is $\langle \mathsf{v}, \mathsf{x} \rangle = \sum_{s \in K} v_s \cdot x_s$.

We call $\mathsf{P} \subseteq \mathbb{R}^K$ a *polytope* if there is a finite set $\mathsf{V} \subseteq \mathbb{R}^K$ such that P equals the convex hull $\mathsf{conv}(\mathsf{V})$ of V, that is, if P is the set of all finite convex combinations $\sum_t \lambda_t \cdot \mathsf{v}_t$ (with $\lambda_t \geqslant 0$ for all t and $\sum_t \lambda_t = 1$) of elements $\mathsf{v}_t \in \mathsf{V}$. If there exists a finite set $\mathsf{V} \subseteq \mathbb{Q}^K$ with $\mathsf{P} = \mathsf{conv}(\mathsf{V})$ then P is called a *rational polytope*.

We call $\mathsf{P} \subseteq \mathbb{R}^K$ a (convex) *polyhedron* if it is the intersection of finitely many affine half-spaces of $\mathbb{R}^K$. Herein, an *affine half-space* in $\mathbb{R}^K$ is a set of the form $\{\mathsf{x} \in \mathbb{R}^K;\ \langle \mathsf{v}, \mathsf{x} \rangle \leqslant \alpha\}$ for some $0 \neq \mathsf{v} \in \mathbb{R}^K$ and $\alpha \in \mathbb{R}$. Consequently, P is the set of solutions to a finite system of linear inequalities over $\mathbb{R}^K$. Note that the parameters v and α defining a half-space are unique up to a positive scaling factor. If all half-spaces defining P can be represented via rational parameters v and α, the polyhedron P is called *rational*. A polyhedron is *bounded* if it does not contain a ray $\{\mathsf{x} + \alpha \cdot \mathsf{w};\ \alpha \geqslant 0\}$ for any $\mathsf{x}, \mathsf{w} \in \mathbb{R}^K,\ \mathsf{w} \neq 0$.

A non-trivial fundamental result in polyhedral geometry relates these two notions:

Theorem 1. *A set $\mathsf{P} \subseteq \mathbb{R}^K$ is a (rational) polytope if and only if it is a bounded (rational) polyhedron.*

It is a challenging algorithmic task to change between an *inner description* $\mathsf{P} = \mathsf{conv}(\mathsf{V})$ and an *outer description* $\mathsf{P} = \{\mathsf{x} \in \mathbb{R}^K;\ \mathsf{A} \cdot \mathsf{x} \leqslant \mathsf{b}\}$, where $\mathsf{A} \in \mathbb{R}^{L \times K},\ \mathsf{b} \in \mathbb{R}^L$ ($L \neq \emptyset$ finite), and back. Standard software packages that allow one to switch between both descriptions are for example *4ti2* [1], *cdd* [7], or *Convex* [6].

The *dimension* $\dim(\mathsf{P})$ of a set $\mathsf{P} \subseteq \mathbb{R}^K$ is the dimension of its affine hull $\mathrm{aff}(\mathsf{P})$, that is, $\mathrm{aff}(\mathsf{P})$ is the set of all finite affine combinations $\sum_t \lambda_t \cdot \mathsf{v}_t$ (with $\lambda_t \in \mathbb{R}$ for all t and $\sum_t \lambda_t = 1$) of elements $\mathsf{v}_t \in \mathsf{V}$. By a convention, the dimension of the empty set is -1. A polytope is *full-dimensional* if $\dim(\mathsf{P}) = |K|$.

3.2. Faces of a polytope

Faces are special sub-polytopes of a polytope; they play an important role both for its inner and outer description. We recapitulate the notion of a face and prove an elementary observation to be used later.

Given $\mathsf{v} \in \mathbb{R}^K$ and $\alpha \in \mathbb{R}$, we call the inequality $\langle \mathsf{v}, \mathsf{x} \rangle \leqslant \alpha$ *valid* for $\mathsf{P} \subseteq \mathbb{R}^K$ if it is satisfied for every $\mathsf{x} \in \mathsf{P}$, that is, if P belongs to the affine half-space $\{\mathsf{x} \in \mathbb{R}^K;\ \langle \mathsf{v}, \mathsf{x} \rangle \leqslant \alpha\}$. If $\langle \mathsf{v}, \mathsf{x} \rangle \leqslant \alpha$ is valid for a polytope P, we call the set $\mathsf{F} = \mathsf{P} \cap \{\mathsf{x} \in \mathbb{R}^K;\ \langle \mathsf{v}, \mathsf{x} \rangle = \alpha\}$ a *face* of P. Then we call the hyperplane $\{\mathsf{x} \in \mathbb{R}^K;\ \langle \mathsf{v}, \mathsf{x} \rangle = \alpha\}$ a *supporting hyperplane* of F.

The sets $\emptyset$ and P are always faces of a polytope P, defined by the valid inequalities $\langle 0, \mathsf{x} \rangle \leqslant 1$ and $\langle 0, \mathsf{x} \rangle \leqslant 0$, respectively. All faces of a polytope P are polytopes and can be classified by their dimension. Faces of the dimension 0 are points in $\mathbb{R}^K$, called *vertices*. Faces of the dimension 1 are line-segments, called *edges*. Faces of the dimension $\dim(\mathsf{P}) - 1$ are called *facets*. Since the vertices of faces are vertices of P, one has only finitely many faces for each polytope. We wish to emphasize for later use that one can conclude from the definition that v is a vertex of P iff it is an *extreme point* of P, that is, if v cannot be written as a convex combination of elements in $\mathsf{P} \setminus \{\mathsf{v}\}$. Another observation is that a line-segment

$$[\mathsf{u}, \mathsf{v}] \equiv \{\alpha \cdot \mathsf{u} + (1 - \alpha) \cdot \mathsf{v};\ \alpha \in [0, 1]\};$$

is an edge of P iff u, v are distinct vertices of P and $\mathsf{P} \setminus [\mathsf{u}, \mathsf{v}]$ is convex.

Theorem 2. *Every polytope* P *is the convex hull of its vertices and the vertices of* P *are the unique inclusion-minimal set* V *with* $\mathsf{P} = \mathrm{conv}(\mathsf{V})$. *Every full-dimensional polytope* P *can be represented as a polyhedron using only the valid inequalities defining the facets of* P. *The facet-defining inequalities form a unique (up to positive scalar factors) inclusion-minimal inequality system defining* P.

The following elementary result will be needed below.

Lemma 3. *Let* $\mathsf{P} \subseteq \mathbb{R}^K$ be a polytope and V the set of its vertices. If $\mathsf{u}, \mathsf{v} \in \mathsf{V}$ are distinct then the following three conditions are equivalent:

(a) *the line-segment* $[\mathsf{u}, \mathsf{v}]$ *is an edge of* P,
(b) *there exists a linear function* U *on* $\mathbb{R}^K$ *such that* $U(\mathsf{u}) > U(\mathsf{v}) > U(\mathsf{w})$ *for any* $\mathsf{w} \in \mathsf{V} \setminus \{\mathsf{u}, \mathsf{v}\}$,
(c) *there exists a linear function* L *on* $\mathbb{R}^K$ *such that* $L(\mathsf{u}) < L(\mathsf{v}) \leqslant L(\mathsf{w})$ *for any* $\mathsf{w} \in \mathsf{V} \setminus \{\mathsf{u}, \mathsf{v}\}$.

Proof. (a) $\Rightarrow$ (b) If $[\mathsf{u}, \mathsf{v}]$ is an edge, there exists a valid inequality $\langle \mathsf{a}, \mathsf{x} \rangle \leqslant \alpha$ for P defining it as a 1-dimensional face of P. As u is a vertex of P there exists a valid inequality $\langle \mathsf{b}, \mathsf{x} \rangle \leqslant \beta$ for P defining it as a 0-dimensional face. We put

$$e = \min_{\mathsf{w} \in \mathsf{V} \setminus \{\mathsf{u}, \mathsf{v}\}} [\langle \mathsf{a}, \mathsf{v} \rangle - \langle \mathsf{a}, \mathsf{w} \rangle] > 0, \quad q = \max_{\mathsf{w} \in \mathsf{V} \setminus \{\mathsf{u}\}} [\langle \mathsf{b}, \mathsf{w} \rangle - \langle \mathsf{b}, \mathsf{v} \rangle] \geqslant 0$$

and choose $\gamma > \frac{q}{e} \geqslant 0$. Then the linear function $U(\mathsf{x}) = \gamma \cdot \langle \mathsf{a}, \mathsf{x} \rangle + \langle \mathsf{b}, \mathsf{x} \rangle$ satisfies the required conditions from (b).

Indeed, $U(\mathsf{u}) > U(\mathsf{v})$ since $\langle \mathsf{a}, \mathsf{u} \rangle = \langle \mathsf{a}, \mathsf{v} \rangle = \alpha$ and $\beta = \langle \mathsf{b}, \mathsf{u} \rangle > \langle \mathsf{b}, \mathsf{v} \rangle$. Moreover, for $\mathsf{w} \in \mathsf{V} \setminus \{\mathsf{u}, \mathsf{v}\}$, the inequality

$$\gamma \cdot [\langle \mathsf{a}, \mathsf{v} \rangle - \langle \mathsf{a}, \mathsf{w} \rangle] \geqslant \gamma \cdot e > q \geqslant \langle \mathsf{b}, \mathsf{w} \rangle - \langle \mathsf{b}, \mathsf{v} \rangle$$

gives $U(\mathsf{v}) = \gamma \cdot \langle \mathsf{a}, \mathsf{v} \rangle + \langle \mathsf{b}, \mathsf{v} \rangle > \gamma \cdot \langle \mathsf{a}, \mathsf{w} \rangle + \langle \mathsf{b}, \mathsf{w} \rangle = U(\mathsf{w})$.

(b) $\Rightarrow$ (c) is evident; put $L = -U$.

(c) $\Rightarrow$ (a) Let L be the linear function from (c). Since v is a vertex of P there exists a linear function V on $\mathbb{R}^K$ such that $V(\mathsf{v}) < V(\mathsf{x})$ for any $\mathsf{x} \in \mathsf{P} \setminus \{\mathsf{v}\}$. Observe that $\beta \equiv \frac{L(\mathsf{v}) - L(\mathsf{u})}{V(\mathsf{u}) - V(\mathsf{v})} > 0$ and consider the affine function

$$Q(\mathsf{x}) = L(\mathsf{x}) - L(\mathsf{v}) + \beta \cdot [V(\mathsf{x}) - V(\mathsf{v})] \quad \text{for } \mathsf{x} \in \mathbb{R}^K.$$

Then $Q(\mathsf{v}) = Q(\mathsf{u}) = 0$ and $Q(\mathsf{w}) > 0$ for $\mathsf{w} \in \mathsf{V} \setminus \{\mathsf{u}, \mathsf{v}\}$. Since Q has the property $Q(\sum_{\mathsf{w} \in \mathsf{V}} \alpha_{\mathsf{w}} \cdot \mathsf{w}) = \sum_{\mathsf{w} \in \mathsf{V}} \alpha_{\mathsf{w}} \cdot Q(\mathsf{w})$ whenever $\alpha_{\mathsf{w}} \geqslant 0,\ \sum_{\mathsf{w} \in \mathsf{V}} \alpha_{\mathsf{w}} = 1$, one has $Q(\mathsf{x}) = 0$ for $\mathsf{x} \in \mathrm{conv}(\{\mathsf{u}, \mathsf{v}\}) = [\mathsf{u}, \mathsf{v}]$ and $Q(\mathsf{x}) > 0$ for any $\mathsf{x} \in \mathsf{P} \setminus [\mathsf{u}, \mathsf{v}]$. This already implies that $[\mathsf{u}, \mathsf{v}]$ is an edge of P. □

4. Main result

In this section we give the main result and illustrate it by an example with three variables. Let S denote the set of standard imsets[4] over N:

[4] To avoid misunderstanding recall that distinct $G, H \in \mathrm{DAGS}(N)$ may give the same standard imset $u_G = u_H$; however, the set S contains only one vector for each independence equivalence class of graphs.

$$\mathsf{S} \equiv \{u_G;\ G \in \mathrm{DAGS}(N)\} \subseteq \mathbb{R}^{\mathscr{P}(N)}.$$

Theorem 4. *The set* S *of standard imsets over* N *is the set of vertices of a rational polytope* $\mathsf{P} \subseteq \mathbb{R}^{\mathscr{P}(N)}$. *The dimension of the polytope is* $2^{|N|} - |N| - 1$.

The proof is given in the Appendix A; it is based on some information-theoretical tools from [18] and certain classic results on BNs.

Example 1. Let us describe the situation in the case of three variables. Then one has 11 standard imsets and they break into five types (=permutation equivalence classes). They can also be classified by their degrees, that is, by the numbers of edges in the corresponding essential graph. More specifically:

- The zero imset $u = 0$ has the degree 0 and corresponds to the complete (undirected) essential graph.
- Six elementary imsets have the degree 1. They break into two types, namely $u_{\langle a,b|\emptyset\rangle} \equiv \delta_\emptyset - \delta_a - \delta_b + \delta_{ab}$ and $u_{\langle b,c|a\rangle} \equiv \delta_a - \delta_{ab} - \delta_{ac} + \delta_{abc}$; the respective essential graphs are $a \to c \leftarrow b$ and b — a — c.
- Three "semi-elementary" imsets of the form $u_{\langle b,ac|\emptyset\rangle} \equiv \delta_\emptyset - \delta_b - \delta_{ac} + \delta_{abc}$ define one type of the degree 2. The corresponding essential graphs have just one undirected edge.
- The imset $2 \cdot \delta_\emptyset - \sum_{i\in\{a,b,c\}} \delta_i + \delta_{abc}$ has the degree 3 and corresponds to the empty essential graph.

The situation is illustrated by Fig. 1, where the lines between the graphs indicate the inclusion neighbors.

By Theorem 4, the dimension of the polytope generated by these 11 imsets is 4. To get its outer (=polyhedral) description one should embed it (as a full-dimensional polytope) into a 4-dimensional space. We decided to identify every standard imset over N with its restriction to $\mathscr{K} \equiv \{A \subseteq N;\ |A| \geqslant 2\}$. Then we used the computer package *4ti2* [1] to get all 13 facet-defining inequalities. They break into seven types and can be classified as follows:

- Five inequalities hold with equality for the zero imset. They break into three types: $0 \leqslant 2 \cdot u(abc) + u(ab) + u(ac) + u(bc)$, $0 \leqslant u(abc) + u(ab)$ and $0 \leqslant u(abc)$.
- Eight inequalities achieve equality for the standard imset for the empty graph. They break into four types, namely $u(abc) \leqslant 1$, $u(abc) + u(ab) \leqslant 1$, $u(abc) + u(ab) + u(ac) \leqslant 1$ and $u(abc) + u(ab) + u(ac) + u(bc) \leqslant 1$.

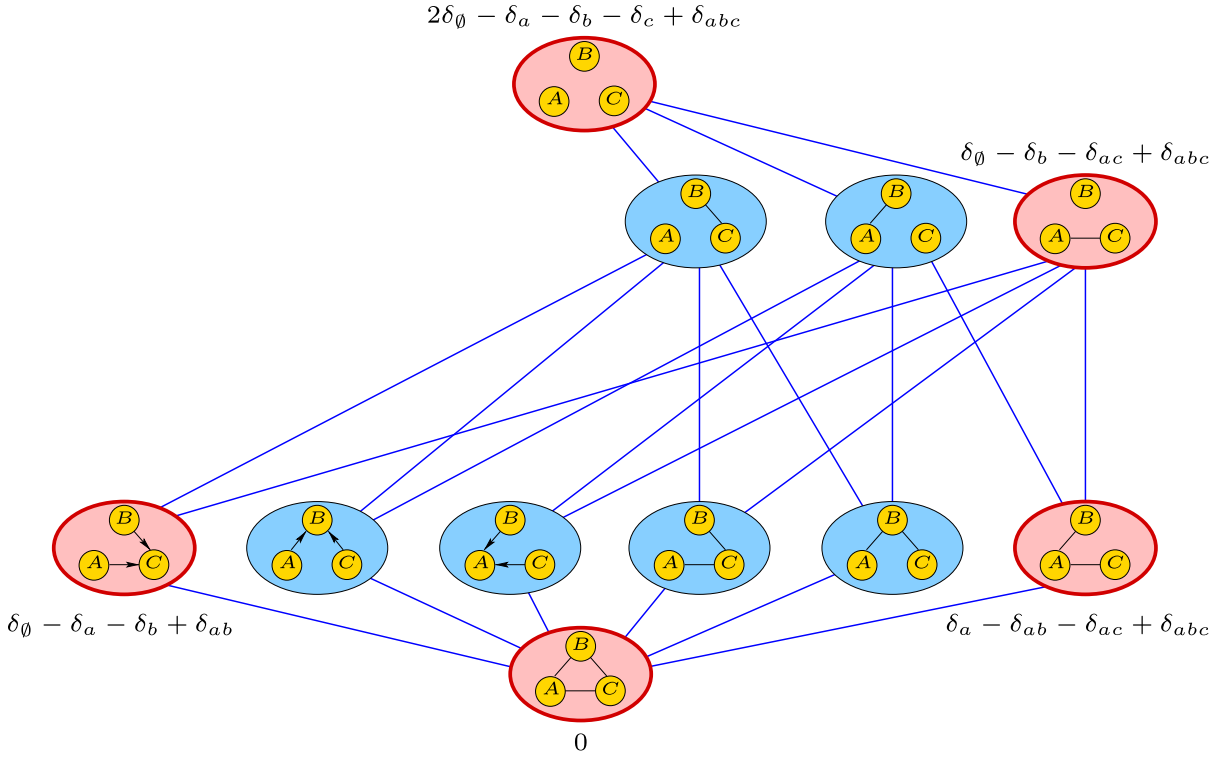

Fig. 1. The essential graphs and (some) standard imsets in the case of three variables.

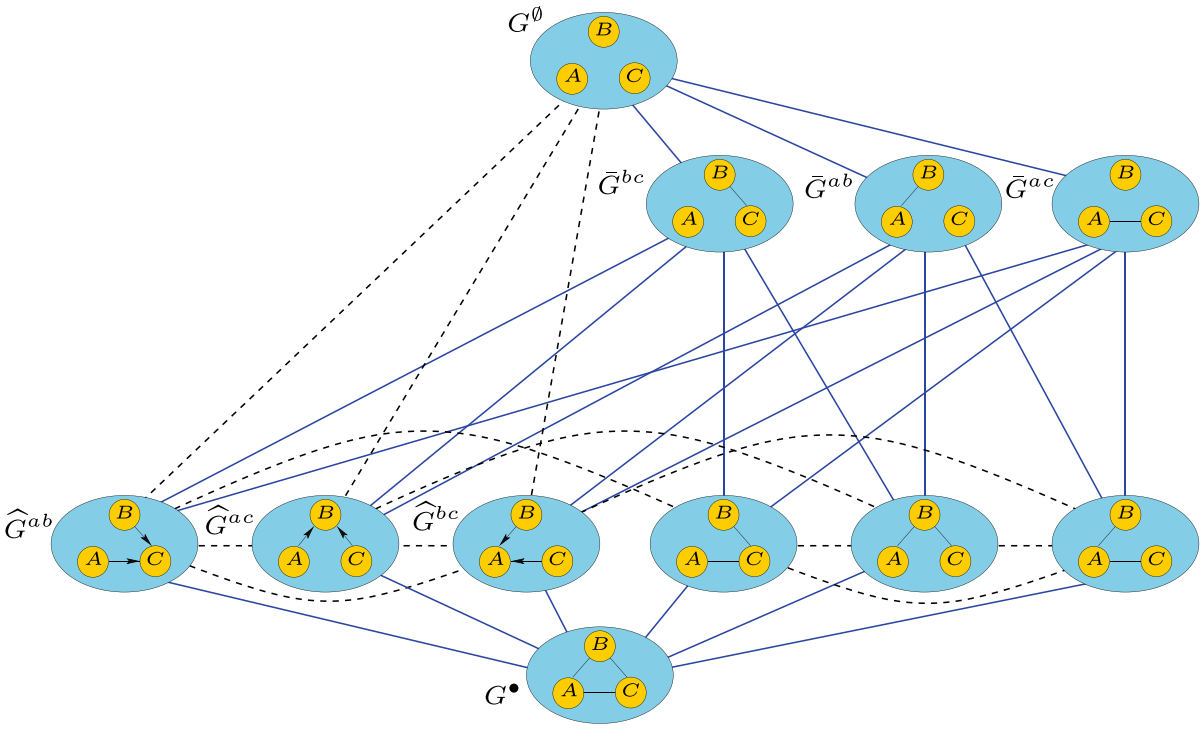

Fig. 2. The geometric and inclusion neighborhood in the case of three variables.

We also made analogous computation in the case $|N| = 4$. In this case one has 185 standard imsets breaking into 20 types. The dimension of the polytope is 11. The number of corresponding facet-defining inequalities is 154 – see vertex-facet table in [21].

Thus, in the case of three and four variables, the polyhedral description of the polytope P was found. In particular, the task to maximize a (score equivalent and decomposable) quality criterion $\mathcal{Q}$ is, by (5), equivalent to a standard linear programming problem, namely to minimize a linear function $u \mapsto \langle t_D^{\mathcal{Q}}, u \rangle$ over the domain specified by those 13, respectively 154, inequalities. Note that the formula for the data vector relative to BIC is also known, see (8.39) in [18]:

$$t_D^{\mathrm{BIC}}(A) = d \cdot H\left(\widehat{P}_A | \prod_{i \in A} \widehat{P}_i\right) - \frac{\ln d}{2} \cdot \left\{ |A| - 1 + \prod_{i \in A} r(i) - \sum_{i \in A} r(i) \right\} \quad \text{for } A \subseteq N,$$

where $H(P|Q)$ is the relative entropy of P with respect to Q and $\widehat{P}_A$ the marginal empirical distribution given by (the projection of the database) D_A.

5. Geometric neighborhood

We say that two standard imsets $u, v \in \mathsf{S}$ are *geometric neighbors* if the line-segment connecting them in $\mathbb{R}^{\mathscr{P}(N)}$ is an edge of the polytope P (generated by S). Given $u \in \mathsf{S}$ the set of its geometric neighbors will be denoted by $geo(u)$. The motivation for this concept has already been explained in the Introduction. The concept of geometric neighborhood can be extended to the corresponding BN structures, and to the essential graphs as well.

An important observation concerning the geometric neighborhood is that it includes the inclusion neighborhood (see Section 2.1.3).

Theorem 5. *If two standard imsets $u, v \in \mathsf{S}$ are inclusion neighbors*[5] *then they are geometric neighbors.*

The proof is given in the Appendix B; it is a kind of extension of the proof of Theorem 4.

However, the substantial statement is as follows.

Theorem 6. *Let* P *denote the convex hull of the set of standard imsets* S *over N and $u \in \mathsf{S}$. If an affine function $\mathcal{Q}$ on $\mathbb{R}^{\mathscr{P}(N)}$ achieves its local maximum relative to the geometric neighborhood in u, that is, if*

[5] That means, the corresponding BN structures are inclusion neighbors.

$$\mathcal{Q}(u) \geqslant \mathcal{Q}(v) \quad \textit{for every } v \in geo(u),$$

then it achieves in u its global maximum over P*, that is,*

$$\mathcal{Q}(u) \geqslant \mathcal{Q}(x) \quad \textit{for every } x \in \mathsf{P}.$$

In particular, the global maximum of $\mathcal{Q}$ *over* S ($\equiv$ *over* P) *coincides with the local maximum of* $\mathcal{Q}$ *relative to the geometric neighborhood.*

In the Appendix C, we give a proof of this fact for any polytope P in $\mathbb{R}^K$ and a linear function L on $\mathbb{R}^K$. Obviously, this implies what is said in Theorem 6. Recall that every score equivalent and decomposable criterion $\mathcal{Q}$ has the required form (see (5) in Section 2.2); thus, Theorem 6 is applicable to $\mathcal{Q}$.

Example 2. We characterized the geometric neighborhood in the case of three variables and compared it with the inclusion neighborhood. The result is depicted in Fig. 2, in which BN structures are represented by essential graphs, solid lines join inclusion neighbors and dashed lines geometric neighbors that are not inclusion neighbors. Different levels correspond to the degrees of the corresponding standard imsets.

We made similar computations also in the case of four variables – see vertex–vertex table in [21]. Our method was as follows: we first computed the outer description of P by standard software packages and then, on basis of that, we computed the edges of P by our own computer program. In that computer program, we have utilized the following two simple facts. First, a line connecting vertices is an edge iff it is the intersection of facets. Second, every face is characterized by the list of vertices contained in it. For more details see the description in [21]. This method, however, was not able to get the results in reasonable time for five variables.

Nevertheless, computing for 2.5 days on a Sun Fire V890 Ultra Sparc IV, 1200 MHz, we have found the geometric neighborhood for $|N| = 5$ by another method that avoids computing the outer description of P. For this, we used the following simple fact. If V denotes the set of vertices of P and if $u, v \in \mathsf{V}$, then the line-segment $[u, v]$ is an edge of P if and only if $v - u$ is not a non-negative linear combination of the vectors $\{w - u : w \in \mathsf{V} \setminus \{u, v\}\}$. But this can be tested for each pair $u, v \in \mathsf{V}$ using linear programming software such as *Cplex* [10]. The results for $|N| = 5$ are also downloadable through [21]; however, they are not in the form of a neat table intended for human users. The case of six variables seems to be a challenge to existing software packages.

6. GES failure

What does it mean that standard imsets $u, v \in \mathsf{S}$ are geometric but not inclusion neighbors? It follows from Lemma 3(b) that then there exists a linear function U on $\mathbb{R}^{\mathscr{P}(N)}$ such that $U(u) > U(v) > U(w)$ for any $w \in \mathsf{S} \setminus \{u, v\}$. Thus, provided u and v are not inclusion neighbors, U achieves in v its local maximum relative to the inclusion neighborhood structure; however, its global maximum over S is achieved in u.

It has already been explained in the Introduction that every "reasonable" quality criterion $\mathcal{Q}$ is (the restriction of) an affine function on $\mathbb{R}^{\mathscr{P}(N)}$. Thus, the reader may ask whether the above phenomenon may occur with $\mathcal{Q}$ in place of U. Indeed, this is the case for three variables, the imset $u = u_{\langle a,b|\emptyset\rangle}$, which corresponds to an "immorality" $\widehat{G}^{ab} : a \to c \leftarrow b$ and the imset v corresponding to the empty graph $G^{\emptyset}$ – see Fig. 2.

Example 3. There exists a database D (of the length $d = 4$) over $N = \{a, b, c\}$ such that the BIC criterion (see Section 2.1.2) achieves its local maximum relative to the inclusion neighborhood in the empty graph $G^{\emptyset}$ and its global maximum in (any of) the graph(s) $\widehat{G}$ of the type $a \to c \leftarrow b$. Put $\mathsf{X}_i = \{0, 1\}$ for $i \in N$ and $D : x^1, x^2, x^3, x^4$, where

$$x^1 = (0, 0, 0), \quad x^2 = (0, 1, 1), \quad x^3 = (1, 0, 1), \quad x^4 = (1, 1, 0).$$

Now, the direct computation of BIC values using the formula (1), which is left to the reader, gives

$$\mathrm{BIC}(\widehat{G}) = -14 \ln 2, \quad \mathrm{BIC}(G^{\emptyset}) = -15 \ln 2 \quad \text{and} \quad \mathrm{BIC}(\overline{G}) = -16 \ln 2$$

for any graph $\overline{G}$ over N having just one edge. Thus, BIC achieves its local maximum relative to the inclusion neighborhood in $G^{\emptyset}$, but it is not the global maximum since $\mathrm{BIC}(\widehat{G}) > \mathrm{BIC}(G^{\emptyset})$ (see Fig. 2 for illustration). Note that BIC exhibits the same behavior if the database D is multiplied,[6] which is a kind of simulation of the situation the data are "generated" from the empirical distribution $\widehat{P}$ given by D, whose density $\hat{p} : \mathsf{X}_N \to [0, 1)$ is given by

$$\hat{p}(0, 0, 0) = \hat{p}(0, 1, 1) = \hat{p}(1, 0, 1) = \hat{p}(1, 1, 0) = 1/4$$

and $\hat{p}(x) = 0$ for remaining $x \in \mathsf{X}_N$.

[6] That means, $D : x^1, \ldots, x^d$, $d = 4 \cdot r$ for $r > 1$, where $x^i = x^{i-4}$ for $5 \leqslant i \leqslant d$.

Now, let us recapitulate a few details concerning the GES algorithm from [5]. It always starts with the empty graph. In the first phase, it searches for the increase in the value of the criterion among the lower inclusion neighbors, in the second phase among the upper inclusion neighbors.[7] Thus, it follows from the description of the algorithm, that, in Example 3, the GES algorithm will immediately end with its starting iteration $G^{\emptyset}$. On the other hand, it is clear that (any of the graphs) $\widehat{G}$ is a more appropriate BN structure approximation of the "actual" conditional independence structure given by $\widehat{P}$. Indeed, this is because none of the conditional independence statements $a \perp\!\!\!\perp b|c,\ a \perp\!\!\!\perp c|b$ and $b \perp\!\!\!\perp c|a$ is valid with respect to $\widehat{P}$.

The reader may object that this is perhaps a rare casual example because of a very special form of the database. However, we have some arguments why (we think) this is, actually, the asymptotic behavior of any (statistically) consistent score equivalent decomposable criterion $\mathcal{Q}$, provided the database is "generated" from the empirical distribution $\widehat{P}$ given by D. First, we re-formulate informally the definition of a consistent criterion from [13]. It is a criterion $\mathcal{Q}$ that satisfies two conditions:

(i) If we have $G, H \in \mathrm{DAGS}(N)$ such that the "generating" distribution P for the database (of the length d) belongs to the statistical model given by G (see Section 2.1) but not to the statistical model given by H, then with $d \to \infty$, the probability that $\mathcal{Q}(G, D(\omega)) > \mathcal{Q}(H, D(\omega))$ approaches to 1.[8]

(ii) If we have $G, H \in \mathrm{DAGS}(N)$ such that P is contained in the both statistical models but the underlying graph for H strictly includes the underlying graph for G, then, again, with $d \to \infty$, the probability that $\mathcal{Q}(G, D(\omega)) > \mathcal{Q}(H, D(\omega))$ approaches to 1.

Now, let us come back to Example 3 and consider the distribution $\widehat{P}$. This distribution is contained in the statistical models given by any of the graphs $\widehat{G}$ and the full graph $G^{\bullet}$. Indeed, this is because each of conditional independence statements $a \perp\!\!\!\perp b|\emptyset,\ a \perp\!\!\!\perp c|\emptyset$ and $b \perp\!\!\!\perp c|\emptyset$ is valid with respect to $\widehat{P}$. On the other hand, $\widehat{P}$ is not contained in any other statistical model of BN structure over $N = \{a, b, c\}$. This is because, for any $G \in \mathrm{DAGS}(N)$ defining such a BN structure, the conditional independence restrictions determined by G involve one of the statements $a \perp\!\!\!\perp b|c,\ a \perp\!\!\!\perp c|b$ and $b \perp\!\!\!\perp c|a$, none of which is valid with respect to $\widehat{P}$. Hence, given a consistent criterion $\mathcal{Q}$, by the condition (i), the global maximum of $\mathcal{Q}$ should be "asymptotically" among graphs of the type $\widehat{G}$ and $G^{\bullet}$. Moreover, by the condition (ii), one should have "asymptotically" $\mathcal{Q}(\widehat{G}, D(\omega)) > \mathcal{Q}(G^{\bullet}, D(\omega))$. That is, the global maximum of $\mathcal{Q}$ should be within the graphs of the type $\widehat{G}$.

The point of our consideration is that if $\mathcal{Q}$ is score equivalent and decomposable then, by formula (5), $\forall D \in \mathrm{DATA}(N, d),\ d \geqslant 1$, one has

$$\mathcal{Q}(\widehat{G}^{ab}, D) - \mathcal{Q}(G^{\bullet}, D) = -\langle t_D^{\mathcal{Q}}, u_{\langle a,b|\emptyset\rangle}\rangle = \mathcal{Q}(G^{\emptyset}, D) - \mathcal{Q}(\overline{G}^{ab}, D),$$
$$\mathcal{Q}(\widehat{G}^{ac}, D) - \mathcal{Q}(G^{\bullet}, D) = -\langle t_D^{\mathcal{Q}}, u_{\langle a,c|\emptyset\rangle}\rangle = \mathcal{Q}(G^{\emptyset}, D) - \mathcal{Q}(\overline{G}^{ac}, D),$$
$$\mathcal{Q}(\widehat{G}^{bc}, D) - \mathcal{Q}(G^{\bullet}, D) = -\langle t_D^{\mathcal{Q}}, u_{\langle b,c|\emptyset\rangle}\rangle = \mathcal{Q}(G^{\emptyset}, D) - \mathcal{Q}(\overline{G}^{bc}, D).$$

In particular, by the condition (ii), $\mathcal{Q}$ should "asymptotically" have the local maximum in $G^{\emptyset}$ – see Fig. 2 for illustration.

The reader may ask how the arguments above are related to the result from [5] about the asymptotic optimality of the GES algorithm. Recall that it says that if the database is "generated" from a distribution which is perfectly Markovian with respect to some $G \in \mathrm{DAGS}(N)$ then, with $d \to \infty$, the probability that the GES algorithm ends with the essential graph for G approaches to 1. Herein, a distribution is called *perfectly Markovian* with respect to G if it satisfies those and only those conditional independence restrictions that are given by G. The point is that the distribution $\widehat{P}$ in Example 3 is **not** perfectly Markovian with respect to any acyclic directed graph.

In our view, the above-mentioned example of the failure of the GES algorithm may occur whenever a disputable *data faithfulness assumption* is not fulfilled.[9] This assumption is "valid" if data are artificially generated, but, in our view, one can hardly ensure its validity for "real" data.

The point of our example is that the GES algorithm is based on the inclusion neighborhood. It follows from Theorem 6 that this cannot happen if the greedy search technique is based on the geometric neighborhood. Therefore, we think the concept of geometric neighborhood is quite important.

7. Conclusions

We have introduced the standard imset polytope P and showed that the problem of learning a BN structure (by the score and search method) is equivalent to the task to maximize a linear function over this polytope. A better understanding of the underlying geometry of the problem is important in itself and may lead to new solution approaches or improvements in existing algorithms.

Therefore, an important open question is to characterize (all) facets and edges of P for general $|N|$ in order to employ a dual simplex method or a (greedy) augmentation algorithm based on the geometric neighborhood. We have already made

[7] If $G, H \in \mathrm{DAGS}(N)$ are inclusion neighbors with $\mathscr{I}(H) \subset \mathscr{I}(G)$ then G is named the *upper* inclusion neighbor, and H the *lower* inclusion neighbor.

[8] Here, a random sample from the distribution P is substituted for the database. To indicate the dependence on a random event ω we write $D(\omega)$ instead of D.

[9] By this we mean the assumption that data are "generated" from a distribution which is **perfectly Markovian** with respect to an acyclic directed graph.

some basic observations. For example, it is not difficult to show that the geometric neighbors of the zero imset $u = 0$ coincide with its inclusion neighbors, that is, with elementary imsets. Consequently, the facet-defining inequalities of P containing $u = 0$ coincide with the facet-defining inequalities of the polyhedral cone generated by elementary imsets. In other words, P is a "cut" out of that cone.

We showed in this paper that the inclusion neighbors are always geometric ones. Thus, the GES algorithm can be interpreted as an augmentation algorithm that moves only along those edges of P that correspond to the inclusion neighbors. Our computations for $|N| = 3,\ 4,\ 5$ (for results see [21]) show that the number of inclusion neighbors is much smaller than the number of geometric neighbors and this proportion decreases as $|N|$ gets bigger. Therefore, from the geometric point of view, it is quite probable (and not surprising at all) that the GES algorithm will fail to reach an optimal vertex of P, which means, the GES algorithm will fail to come with an optimal BN structure.

Note that what we propose is not the only option for finding the global maximum of a quality criterion. In [17] another method based on dynamic programming was presented, and quite impressively, it is claimed there that it is possible to use that method to learn optimal BN structures with about 30 variables. However, due to its enumerative nature, the problem sizes that can be handled by that approach (also in future) are clearly bounded and new ideas are needed to push these bounds. Moreover, this dynamic programming approach cannot take advantage of a good or even optimal solution obtained via some heuristics.

Obviously, a quick and good starting solution can tremendously speed up any augmentation algorithm such as the one we suggest using the geometric neighborhood. Often one generates heuristically an optimal solution (without knowing it) and is left with the task of proving its optimality. From the theoretical point of view, the complexity of solving the full optimization problem and the complexity of proving optimality of a given candidate solution differ only by a polynomial factor. From a practical point of view, however, even such a polynomial factor can make a huge difference.

Our approach can easily be modified to some cases of restricted learning. For example, one may wish to restrict oneself to learning BN structures that are given by graphs which have a prescribed upper limit for the number of arrows. In our context it leads to an elegant simplification step, namely considering a sub-polytope of P determined by a subset of its vertices.

Let us gather some important open questions concerning P:

(1) Characterize all facets and edges of P. An interesting related conjecture is that the only lattice points (=vectors whose components are integers) within P are its extreme points, that is, standard imsets.
(2) Characterize differential imsets $u_G - u_H$, where $G, H \in \mathrm{DAGS}(N)$, for geometric neighbors. Find out whether they can be interpreted in graphical terms. In other words: What is the relation between the corresponding essential graphs if u_G and u_H are geometric neighbors?
(3) Apply the presented geometric approach to restricted learning BN structures. Again, we face the task to describe (all) facets and edges of some polytopes, but these polytopes can appear to be much simpler than P.

All these questions concern the complexity of a potential future (greedy) search procedure for maximization of a quality criterion Q based on the geometric neighborhood. We hope that the analysis of the results for $|N| \leqslant 5$ will give a clue for the (mathematical) characterization of the geometric neighborhood, which is a great theoretical challenge.

Acknowledgements

This research has been supported by the Grants GAČR No. 201/08/0539, GAAVČR No. IAA100750603, MŠMT No. 1M0572, and 2C06019. We thank anonymous reviewers for their comments.

Appendix A. The proof of Theorem 4

We introduce $\mathsf{P} \subseteq \mathbb{R}^{\mathscr{P}(N)}$ as the convex hull of the set of standard imsets S. Obviously, it is a rational polytope and the set of extreme points of P has to be a subset of S. Throughout the rest of the proof we assume $|N| \geqslant 2$ because if $|N| = 1$ then S only contains the zero imset, $\mathsf{P} = \mathsf{S}$ and the claims in Theorem 4 are trivial.

To show the first statement, claiming that each $u = u_G \in \mathsf{S},\ G \in \mathrm{DAGS}(N)$, is a vertex of P, it suffices to construct a supporting hyperplane for the face $\{u\}$. This is equivalent to constructing a linear function L on $\mathbb{R}^{\mathscr{P}(N)}$ that is uniquely minimized in $u \in \mathsf{P}$, that is, $L(u) < L(v)$ for any $v \in \mathsf{S} \setminus \{u\}$.[10] The function will be of the form $L(x) = \langle m, x\rangle,\ x \in \mathbb{R}^{\mathscr{P}(N)}$, where $m : \mathscr{P}(N) \to \mathbb{R}$ is a suitable real function (=a point in $\mathbb{R}^{\mathscr{P}(N)}$).

In the construction we utilize the properties of the *multiinformation function* m_P for a (discrete) probability distribution P over N – see Section 2.3.4 in [18]. It is a function $m_P : \mathscr{P}(N) \to [0, \infty)$ which ascribes to every $A \subseteq N$ the multiinformation of the corresponding marginal P_A of P for A.[11] The basic property of the multiinformation function m_P is that it is *supermodular*[12] and characterizes conditional independence statements in P by algebraic identities. In particular, the independence structure

[10] Note it implies $L(u) < L(x)$ for any $x \in \mathsf{P} \setminus \{u\}$.
[11] The multiinformation of R (over A) is the relative entropy $H(R|Q)$ of R with respect to the product $Q = \prod_{i \in A} R_i$ of its one-dimensional marginals.
[12] This means $m_P(C \cup D) + m_P(C \cap D) \geqslant m_P(C) + m_P(D)$ for any $C, D \subseteq N$.

$\mathscr{M}(m_P)$ produced by m_P through the respective algebraic test[13] coincides with the collection $\mathscr{I}(P)$ of conditional independence statements represented in P : $\mathscr{M}(m_P) = \mathscr{I}(P)$.[14]

A further preparatory observation concerns standard imsets. Lemma 7.1 in [18] says that every standard imset $u = u_G$ for $G \in \mathsf{DAGS}(N)$ belongs to a wider class of combinatorial imsets, and, therefore, to an even wider class of *structural imsets* – see Section 4.2.3 in [18].[15] Moreover, Lemma 7.1 [18] also says that the independence structure $\mathscr{M}(u_G)$ induced by the imset u_G through the respective algebraic criterion coincides with the collection $\mathscr{I}(G)$ of conditional independence restrictions in G (determined by the respective graphical separation criterion): $\mathscr{M}(u_G) = \mathscr{I}(G)$.

In the sequel, we will use the following notation: given $v \in \mathsf{S}$ the symbol $\mathscr{I}(v)$ will denote the collection of conditional independence restrictions $\mathscr{I}(H)$ determined by (any) graph $H \in \mathsf{DAGS}(N)$ with $v = u_H$.[16] A further observation is that, for every $u, v \in \mathsf{S}$ the (strict) inclusion $\mathscr{I}(v) \subset \mathscr{I}(u)$ implies $deg(v) < deg(u)$. Indeed if $u = u_G$ and $v = u_H$, where $G, H \in \mathsf{DAGS}(N)$, then $\mathscr{I}(H) \subset \mathscr{I}(G)$. Thus, it follows from the characterization of inclusion in [5] that H has a higher number of arrows than G : $\mathsf{a}(H) > \mathsf{a}(G)$. Hence, by (3), one has $deg(u_H) < deg(u_G)$.

Since now we fix $u \in \mathsf{S}$ and one of the graphs $G \in \mathsf{DAGS}(N)$ with $u = u_G$. Let us put

$$\mathsf{S}(u) = \{v \in \mathsf{S};\ \mathscr{I}(v) \subseteq \mathscr{I}(u)\}.$$

The crucial step in our proof is that we utilize a well-known result [9] on the existence of a perfectly Markovian distribution for an acyclic directed graph. If this result is applied to our fixed $G \in \mathsf{DAGS}(N)$ it says there exists a discrete probability distribution P over N such that $\mathscr{I}(P) = \mathscr{I}(G)$. We take such a distribution P, fix it, consider its multiinformation function m_P and interpret it as a point in $\mathbb{R}^{\mathscr{P}(N)}$.

The next step is to realize that $\langle m_P, v\rangle \geqslant 0$ for any $v \in \mathsf{S}$ and one has $\langle m_P, v\rangle = 0$ if and only if $v \in \mathsf{S}(u)$. The first claim follows from Proposition 5.1(i) in [18] saying that $\langle \tilde{m}, v\rangle \geqslant 0$ for any supermodular function $\tilde{m}$ and a structural imset v. As explained above, these assumptions are valid for m_P and any $v \in \mathsf{S}$. As concerns the second claim, Proposition 5.6 in [18] says, under the same assumptions, that $\langle m_P, v\rangle = 0$ if and only if $\mathscr{M}(v) \subseteq \mathscr{M}(m_P)$. However, as explained above, one has $\mathscr{M}(m_P) = \mathscr{I}(P)$ and provided that $v = u_H,\ H \in \mathsf{DAGS}(N)$ one also has $\mathscr{M}(v) = \mathscr{M}(u_H) = \mathscr{I}(H) = \mathscr{I}(v)$. Thus, $\langle m_P, v\rangle = 0$ if and only if $\mathscr{I}(v) \subseteq \mathscr{I}(P)$. However, since $\mathscr{I}(P) = \mathscr{I}(G) = \mathscr{I}(u)$ the inclusion $\mathscr{I}(v) \subseteq \mathscr{I}(P)$ means $\mathscr{I}(v) \subseteq \mathscr{I}(u)$, that is, $v \in \mathsf{S}(u)$.

Thus, because $\langle m_P, v\rangle > 0$ for any $v \in \mathsf{S} \setminus \mathsf{S}(u)$ we know that

$$k \equiv \min_{v \in \mathsf{S}\setminus\mathsf{S}(u)} \langle m_P, v\rangle > 0.$$

Put $R \equiv \max_{v\in\mathsf{S}} deg(v)$. Actually, we know by (3) that $R = \frac{1}{2}\cdot|N|\cdot(|N|-1)$, and, therefore, $R > 0$. Let us choose $\varepsilon > 0$ with $\varepsilon < \frac{k}{R}$ and put

$$m \equiv m_P - \varepsilon \cdot m_*,$$

where $m_* : \mathscr{P}(N) \to \mathbb{Z}$ is the function from (4). Finally, we define a linear function $L : \mathbb{R}^{\mathscr{P}(N)} \to \mathbb{R}$ by the formula

$$L(x) \equiv \langle m, x\rangle \quad \text{for } x \in \mathbb{R}^{\mathscr{P}(N)}. \tag{A.1}$$

It follows from (4) that $\forall v \in \mathsf{S}$ one has $L(v) = \langle m_P, v\rangle - \varepsilon \cdot deg(v)$. In particular, $L(u) = -\varepsilon \cdot deg(u)$. If $v \in \mathsf{S} \setminus \mathsf{S}(u)$ then

$$L(v) = \langle m_P, v\rangle - \varepsilon \cdot deg(v) \geqslant k - \varepsilon \cdot R > 0 \geqslant L(u). \tag{A.2}$$

Moreover, if $v \in \mathsf{S}(u),\ v \neq u$ then $\mathscr{I}(v) \subset \mathscr{I}(u)$[17] and, by the above observation, $deg(u) > deg(v)$. Hence, since $\langle m_P, v\rangle = 0 = \langle m_P, u\rangle$ one has

$$L(v) - L(u) = \varepsilon \cdot (deg(u) - deg(v)) > 0. \tag{A.3}$$

Thus, $L(u) < L(v)$ for any $v \in \mathsf{S} \setminus \{u\}$, which concludes the proof of the first statement of Theorem 4.

As concerns the second statement, realize every standard imset $u : \mathscr{P}(N) \to \mathbb{Z}$ satisfies $\sum_{A\subseteq N} u(A) = 0$ and $\sum_{A\subseteq N, i\in A} u(A) = 0$ for every $i \in N$. In particular, u is uniquely determined by its restriction to $\mathscr{K} \equiv \{A \subseteq N;\ |A| \geqslant 2\}$. This defines a one-to-one linear transformation between $\mathbb{R}^{\mathscr{K}}$ and a linear subspace of $\mathbb{R}^{\mathscr{P}(N)}$ containing S. Thus, to prove what is desired it suffices to show that the linear hull of $\mathscr{K}$-restrictions of standard imsets is just $\mathbb{R}^{\mathscr{K}}$, which has the dimension $|\mathscr{K}| = 2^{|N|} - |N| - 1$. Because every elementary imset is standard,[18] it is enough to show that, for every $A \in \mathscr{K}$, the ($\mathscr{K}$-restriction of the) imset δ_A is a linear combination of $\mathscr{K}$-restrictions of elementary imsets. This can be done easily by induction on $|A|$.

[13] We omit the definition of that algebraic test because these details are not necessary to understand the arguments given in this paper.

[14] See Proposition 5.3 in [18] for the respective arguments.

[15] Again, we omit the definitions of these imsets and of the independence structures defined by them. These definitions are not substantial in our considerations.

[16] The definition is correct since one has $u_G = u_H$ if and only if $G, H \in \mathsf{DAGS}(N)$ are independence equivalent (cf. Section 2.2).

[17] The unique $v \in \mathsf{S}$ with $\mathscr{I}(v) = \mathscr{I}(u)$ is u itself.

[18] Given $a \perp\!\!\!\perp b|C$, consider a total order of N in which C precedes $\{a, b\}$ and $\{a, b\}$ precedes $N \setminus (C \cup \{a, b\})$, direct edges of the complete graph over N according to this order and remove the arrow between a and b.

Appendix B. The proof of Theorem 5

Assume without loss of generality that $u = u_G$ and $v = u_H$ for $G, H \in \mathrm{DAGS}(N)$ such that $\mathscr{I}(H) \subset \mathscr{I}(G)$, but there is no $F \in \mathrm{DAGS}(N)$ with $\mathscr{I}(H) \subset \mathscr{I}(F) \subset \mathscr{I}(G)$. Thus, it follows from the characterization of inclusion by Chickering [5] that H has just one more arrow than $G : \mathsf{a}(H) = \mathsf{a}(G) + 1$. Hence, by (3), one has $deg(u) = deg(v) + 1$.

Now, we can repeat the considerations from the proof of Theorem 4 and, having fixed $u = u_G$, define a linear function L on $\mathbb{R}^{\mathscr{P}(N)}$ by the formula (A.1). Thus, it follows from (A.2) and (A.3) that $L(w) > 0$ for $w \in \mathsf{S} \setminus \mathsf{S}(u)$ while $L(w) = -\varepsilon \cdot deg(w)$ for $w \in \mathsf{S}(u)$. In particular,

$$L(u) = -\varepsilon \cdot deg(u) < -\varepsilon \cdot deg(v) = L(v) \leqslant L(w) \quad \text{for any } w \in \mathsf{S} \setminus \{u, v\}.$$

Indeed, for $w \in \mathsf{S} \setminus \mathsf{S}(u)$ one has $L(v) \leqslant 0 < L(w)$ while, for $w \in \mathsf{S}(u) \setminus \{u\}, deg(w) \leqslant deg(u) - 1 = deg(v)$ and $L(v) = -\varepsilon \cdot deg(v) \leqslant -\varepsilon \cdot deg(w) = L(w)$.

Thus, by Lemma 3(c), the line-segment connecting u and v is an edge of P, that is, u and v are geometric neighbors.

Appendix C. The proof of Theorem 6

Let $\mathsf{P} \subseteq \mathbb{R}^K$, $K \neq \emptyset$ finite, be a polytope. The set of its vertices will be denoted by $\mathsf{V} \equiv vert(\mathsf{P})$ throughout Section C. Given (distinct) $\mathsf{z}, \mathsf{w} \in \mathbb{R}^K$, we accept a shorthand notation for the interior of the line-segment $[\mathsf{z}, \mathsf{w}]$:

$$(\mathsf{z}, \mathsf{w}) \equiv \{\alpha \cdot \mathsf{z} + (1 - \alpha) \cdot \mathsf{w}; \quad \alpha \in (0, 1)\}.$$

We will also utilize the following equivalent definition of a face of P (use Theorem 7.5 in [4]): it is a (closed) convex subset $\mathsf{F} \subseteq \mathsf{P}$ such that one has

$$\forall \mathsf{z}, \mathsf{w} \in \mathsf{P} \quad \text{if } (\mathsf{z}, \mathsf{w}) \cap \mathsf{F} \neq \emptyset \text{ then } \mathsf{z}, \mathsf{w} \in \mathsf{F}.$$

Given $\mathsf{x} \in \mathsf{V}$, let us denote by $ne_{\mathsf{P}}(\mathsf{x})$ the set of those vertices $\mathsf{y} \in \mathsf{V}$ such that the line-segment $[\mathsf{x}, \mathsf{y}]$ connecting x and y in $\mathbb{R}^K$ is an edge of P.

We base our proof on the next observation, mentioned as Lemma 3.6 in [22].

Lemma 7. *Given a polytope $\mathsf{P} \subseteq \mathbb{R}^K$ and $\mathsf{x} \in vert(\mathsf{P})$ one has*

$$\mathsf{P} \subseteq \mathsf{x} + cone(\{\mathsf{y} - \mathsf{x}; \ \mathsf{y} \in ne_{\mathsf{P}}(\mathsf{x})\}),$$

where $cone(\mathsf{A}) \equiv \{\sum_{\mathsf{x} \in \mathsf{A}} \alpha_{\mathsf{x}} \cdot \mathsf{x}; \ \alpha_{\mathsf{x}} \geqslant 0\}$ denotes the cone generated by a finite set $\mathsf{A} \subseteq \mathbb{R}^K$.

The geometric meaning of Lemma 7 is that the polytope P belongs to the cone with the origin x and with the rays determined by the edges of P coming out of x. Lemma 7 has the following consequence.

Corollary 8. *Let $\mathsf{Q} \subseteq \mathbb{R}^K$ be a polytope, $\mathsf{u} \in vert(\mathsf{Q})$ and L is a linear function on $\mathbb{R}^K$. If $L(\mathsf{u}) = L(\mathsf{v})$ for any $\mathsf{v} \in ne_{\mathsf{Q}}(\mathsf{u})$ then L is constant on Q, that is, $L(\mathsf{x}) = L(\mathsf{u})$ for any $\mathsf{x} \in \mathsf{Q}$.*

Proof. Let us put $\gamma \equiv L(\mathsf{u})$; the assumption says that, for every $\mathsf{v} \in ne_{\mathsf{Q}}(\mathsf{u})$, the function L has the same value γ for two distinct points in the ray $\mathsf{R}_{\mathsf{v}} = \{\mathsf{u} + \alpha \cdot (\mathsf{v} - \mathsf{u}); \ \alpha \geqslant 0\}$. Since L is linear, it has to be constant on R_{v} with the value γ. Because every point in the set $\mathsf{u} + cone(\{\mathsf{v} - \mathsf{u}; \ \mathsf{v} \in ne_{\mathsf{Q}}(\mathsf{u})\})$ is a convex combination of the points in the rays $\mathsf{R}_{\mathsf{v}}, \mathsf{v} \in ne_{\mathsf{Q}}(\mathsf{u})$, it implies L has the same value γ in the whole cone. In particular, by Lemma 7, L has the same value γ in the whole set $\mathsf{Q} \subseteq \mathsf{u} + cone(\{\mathsf{v} - \mathsf{u}; \ \mathsf{v} \in ne_{\mathsf{Q}}(\mathsf{u})\})$. □

Now, we are ready to show the following property which clearly implies what is said in Theorem 6, because an affine function differs from a linear function by a constant.

Lemma 9. *Given a polytope $\mathsf{P} \subseteq \mathbb{R}^K$ with the vertex set $\mathsf{V} \equiv vert(\mathsf{P}), \mathsf{u} \in \mathsf{V}$ and a linear function L on $\mathbb{R}^K$, the condition*

$$L(\mathsf{u}) \geqslant L(\mathsf{v}) \quad \textit{for every } \mathsf{v} \in ne_{\mathsf{P}}(\mathsf{u}), \tag{C.1}$$

implies that L achieves in u its global maximum over P, that is,

$$L(\mathsf{u}) \geqslant L(\mathsf{x}) \quad \textit{for every } \mathsf{x} \in \mathsf{P}. \tag{C.2}$$

Proof. Let us put $\gamma = L(\mathsf{u})$ and $\mathsf{Q} = \mathsf{P} \cap \{\mathsf{x} \in \mathbb{R}^K; \ L(\mathsf{x}) \geqslant \gamma\}$. Obviously, Q is a bounded polyhedron, and, therefore, a non-empty polytope. Moreover, u is a vertex of $\mathsf{Q} : \mathsf{u} \in vert(\mathsf{Q})$.[19] To show (C.1)⇒(C.2) we only need to show that (C.1) implies $L(\mathsf{x}) = \gamma$ for $\mathsf{x} \in \mathsf{Q}$. By Corollary 8, to this end it is enough to show $L(\mathsf{v}) = \gamma$ for every $\mathsf{v} \in ne_{\mathsf{Q}}(\mathsf{u})$. Thus, assume for a contradiction

$$\exists \mathsf{v} \in ne_{\mathsf{Q}}(\mathsf{u}) \quad L(\mathsf{v}) > \gamma, \tag{C.3}$$

[19] Realize that this means the same as an extreme point of $\mathcal{Q}$.

while (C.1) holds.[20] To get a contradictory conclusion it suffices to observe that $[u,v]$ is an edge of P.[21] That means, we have to show (see above)

$$\text{if } z,w \in P \text{ are such that } (z,w) \cap [u,v] \neq \emptyset \text{ then } z,w \in [u,v]. \tag{C.4}$$

To show that choose $x \in (z,w) \cap [u,v]$ and observe $x \neq u$.[22] This implies $L(x) > \gamma$.[23] However, $x \in (z,w)$ and the linearity of L on $[z,w]$ gives $\max\{L(z), L(w)\} > \gamma$.[24] Without loss of generality assume $L(w) > \gamma$, that is, $w \in Q$. Observe by a contradiction that then $L(z) \geqslant \gamma$.

Indeed, if $L(z) < \gamma$ then the linearity of L on $[z,x] \subseteq [z,w]$ implies there exists unique $y \in (z,x)$ such that $L(y) = \gamma$. Then $y,w \in Q$ and $x \in (y,w) \cap [u,v]$. Since $[u,v]$ is an edge of Q, this implies $y,w \in [u,v]$. Hence, $y \in [u,v]$ while $L(y) = \gamma$. But the only point in $[u,v]$ having γ as the value of L is just u. Therefore $y = u$. This, however, means $u \in (z,w)$ for $z,w \in P, z \neq w$ contradicting $u \in \mathit{vert}(P)$.

Thus, $L(z) \geqslant \gamma$ means $z \in Q$. Therefore, $z,w \in Q$ and $(z,w) \cap [u,v] \neq \emptyset$. Because $[u,v]$ is an edge of Q we observe $z,w \in [u,v]$, which was desired to verify (C.4). □

References

[1] 4ti2 team, 4ti2, a software package for algebraic, geometric and combinatorial problems on linear spaces. Available at <http://www.4ti2.de>.
[2] S.A. Andersson, D. Madigan, M.D. Perlman, A characterization of Markov equivalence classes for acyclic digraphs, The Annals of Statistics 25 (1997) 505–541.
[3] R.R. Bouckaert, Bayesian belief networks: from construction to evidence, Ph.D. Thesis, University of Utrecht, 1995.
[4] A. Brøndsted, An Introduction to Convex Polytopes, Springer-Verlag, 1983.
[5] D.M. Chickering, Optimal structure identification with greedy search, Journal of Machine Learning Research 3 (2002) 507–554.
[6] M. Franz, Convex, a Maple package for convex geometry. Available at <http://www.math.uwo.ca/~mfranz/convex>.
[7] K. Fukuda, cdd and cdd+, an implementation of the double description method. Available at <http://www.ifor.math.ethz.ch/~fukuda/cdd_home/cdd.html>.
[8] M. Frydenberg, The chain graph Markov property, Scandinavian Journal of Statistics 17 (1990) 333–353.
[9] D. Geiger, J. Pearl, On the logic of causal models, in: Proceedings of the 4th Conference on Uncertainty in Artificial Intelligence, North-Holland, 1990, pp. 3–14.
[10] Ilog team, Cplex, mathematical programming optimizer, version 10.1. Available at <www.ilog.com/products/cplex>.
[11] S.L. Lauritzen, Graphical Models, Clarendon Press, 1996.
[12] C. Meek, Graphical models, selecting causal and statistical models, Ph.D. Thesis, Carnegie Melon University, 1997.
[13] R.E. Neapolitan, Learning Bayesian Networks, Pearson Prentice Hall, 2004.
[14] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, 1988.
[15] A. Schrijver, Theory of Linear and Integer Programming, John Wiley, 1986.
[16] G. Schwarz, Estimation the dimension of a model, The Annals of Statistics 6 (1978) 461–464.
[17] T. Silander, P. Myllymäki, A simple approach for finding the globally optimal Bayesian network structure, in: Proceedings of the 22th Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2006, pp. 445–452.
[18] M. Studený, Probabilistic Conditional Independence Structures, Springer-Verlag, 2005.
[19] M. Studený, J. Vomlel, A reconstruction algorithm for the essential graph, International Journal of Approximate Reasoning 50 (2009) 385–413.
[20] T. Verma, J. Pearl, Equivalence and synthesis of causal models, in: Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence, Elsevier, 1991, pp. 220–227.
[21] J. Vomlel, M. Studený, Geometric neighborhood for Bayesian network structures over three and four variables, 2008. Web page, see <http://www.utia.cas.cz/vomlel/imset/polytopes-3v-and-4v.html>.
[22] G.M. Ziegler, Lectures on Polytopes, Springer-Verlag, 1995.

[20] Since $L(x) \geqslant \gamma$ for $x \in Q$ there is no $v \in ne_Q(u) \subseteq Q$ with $L(v) < \gamma$.
[21] Because then $v \in ne_P(u)$ and by (C.1) $\gamma \equiv L(u) \geqslant L(v)$ contradicts (C.3).
[22] If $x = u$ then $u \in (z,w)$ contradicts $u \in \mathit{vert}(P)$.
[23] This is because $L(u) = \gamma, L(v) > \gamma$ by (C.3) and the linearity of L implies $L(y) > \gamma$ for $y \in [u,v] \setminus \{u\}$.
[24] A linear function has not a local maximum in an internal point of a line-segment.