Geometric View on Learning Bayesian Network Structures

M. Studený and J. Vomlel
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Prague, CZ 18208, Czech Republic
studeny@utia.cas.cz, vomlel@utia.cas.cz

Abstract

We recall the basic idea of an algebraic approach to learning Bayesian network (BN) structure, namely to represent every BN structure by a certain (uniquely determined) vector, called *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. To illustrate this concept by an example, we describe the geometric neighborhood in the case of three variables and show it differs from the *inclusion neighborhood*, which was introduced earlier in connection with the GES algorithm. Some results in the case of four variables are also available.

1 INTRODUCTION

The motivation for this theoretical paper is learning Bayesian network (BN) structure from data by the method of maximization of a quality criterion. By a *quality criterion*, also named a *score metric* or simply a *score* by 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 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 a non-unique way of its description. Later, researchers calling for methodological simplification came with the idea to represent every BN structure by 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, Madigan and Perlman (1997), who also gave graphical characterization of graphs that are essential.

Since direct maximization of a quality criterion $\mathcal{Q}$ seems, 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 (Bouckaert 1995). The point is that, instead of the global maximum of $\mathcal{Q}$, one is trying to find a local maximum with respect to the chosen neighborhood structure. This is algorithmically simpler task because one can utilize various greedy search techniques for this purpose. A typical example of these techniques is *greedy equivalence search* (GES) algorithm proposed by Meek (1997). The neighborhood structure utilized in this algorithm is the *inclusion neighborhood*, which comes from the conditional independence interpretation of BN structures (Kočka 2001). Chickering (2002) proposed a modification of the GES algorithm, in which he used the 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* (Bouckaert 1995), by which is meant it ascribes the same value to equivalent graphs (= graphs defining the same BN structure). The other requirement is that $\mathcal{Q}$ should be *decomposable* (Chickering 2002), which means $\mathcal{Q}(G, D)$ decomposes into contributions that correspond to factors in the respective factorization according to the graph G.

The basic idea of an algebraic approach to learning BN structures, presented in Chapter 8 of (Studený 2005) is to represent both the BN structure and the database by a real vector. More specifically, an al-

gebraic 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 boon of standard imsets is that one can read practically immediately from the differential imset $u_G - u_H$ whether the BN structures defined by 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 § 8.4.2 of (Studený 2005) 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 (they both depend solely on the database D and the criterion $\mathcal{Q}$) and $\langle *, * \rangle$ denotes the scalar (= inner) product. The vector $t_D^{\mathcal{Q}}$ is named the *data vector* (relative to $\mathcal{Q}$).

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

Now, a well-known classic result on convex sets in the Euclidean space, Weyl-Minkowski theorem, says that a polytope can equivalently be introduced as a bounded polyhedron. Thus, once one succeeds to describe the above mentioned polytope in the form of a (bounded) polyhedron, one gets a classic task of linear programming, namely to maximize/mimimize a linear function over a polyhedron. There are efficient methods, like the *simplex method*, to tackle this problem (Schrijver 1986). To illustrate the idea we describe the above mentioned (standard imset) polytope in the form of a bounded polyhedron in the case $|N| = 3$ in the paper and give a web reference for the case $|N| = 4$.

However, because it is not clear at this moment how to find the "polyhedral" description of the polytope for arbitrary $|N|$,[1] we propose an alternative idea in this paper. 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 a face of the polytope (= the edge of the polytope in the geometric sense). The motivation is as follows: one of possible interpretations of the simplex method is that it is a kind of "greedy search" method in which one moves between vertices (of the polyhedron) along the edges - see § 11.1 of (Schrijver 1986); c.f. § 3.2 in (Ziegler 1995). Thus, provided one succeeds to characterize the geometric neighborhood, one can possibly use greedy search techniques to find the global maximum of $\mathcal{Q}$ over the polytope, and, therefore, over the set of standard imsets.

Again, to illustrate the concept of geometric neighborhood we characterize it for 3 variables in the paper and give a web reference to the characterization in the case of 4 variables. The finding is that the inclusion neighborhood and geometric neighborhood differ already in the case of 3 variables. The last part of the paper is the list of open questions.

2 BASIC CONCEPTS

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

2.1 BN STRUCTURES

One of 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 (Neapolitan 2004). A well-known fact is that P factorizes according to G iff it is Markovian with respect to G, by which is meant it satisfies conditional independence restrictions determined by the respective (directed) separation criterion (Pearl 1988; Lauritzen 1996). Having fixed (non-empty finite) sample spaces X_i for variables $i \in N$, the respective (BN) *statistical model* is the class of all probability distributions P on $\mathsf{X}_N \equiv \prod_{i \in N} \mathsf{X}_i$ that factorize according to G. To name the shared features of 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.

[1] This is an open problem and what worries us is that it may be the case the number of half-spaces defining the polyhedron may occur to be super-exponential in $|N|$.

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. If $|\mathsf{X}_i| \geq 2$ for every $i \in N$ then this is equivalent to the condition they are *independence equivalent*, by which is meant they determine the same collection of conditional independence restrictions - cf. § 2.2 in (Neapolitan 2004). Verma and Pearl (1991) gave classic graphical characterization of independence equivalence: two acyclic directed graphs G and H over N are independence equivalent iff they have the same underlying undirected graph and *immoralities*, by which are meant induced subgraphs of the form $a \to c \leftarrow b$ where $[a, b]$ is not an edge in the graph.

2.1.2 Learning 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 $d \geq 1$, that is, of a sequence of elements of X_N. Provided the sample spaces X_i with $|\mathsf{X}_i| \geq 2$ 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 by a *quality criterion* will be meant a real function $\mathcal{Q}$ on $\mathsf{DAGS}(N) \times \mathsf{DATA}(N, d)$. The value $\mathcal{Q}(G, D)$ should reflect how the statistical model determined by G is suitable to explain the occurrence of the database D. The learning procedure based on $\mathcal{Q}$ then consists in the maximization of the function $G \mapsto \mathcal{Q}(G, D)$ over $G \in \mathsf{DAGS}(N)$ if the database $D \in \mathsf{DATA}(N, d)$, $d \geq 1$ is given. In this brief overview we omit examples of quality criteria and the question of their statistical consistency; we refer the reader to the literature on this topic (Chickering 2002; Castelo 2002; Neapolitan 2004).

However, we recall two important definitions. A quality criterion $\mathcal{Q}$ will be named *score equivalent* if, for every $D \in \mathsf{DATA}(N, d)$, $d \geq 1$, one has

$$\mathcal{Q}(G, D) = \mathcal{Q}(H, D) \quad \text{whenever } G, H \in \mathsf{DAGS}(N)$$

are independence equivalent. Moreover, $\mathcal{Q}$ will be called *decomposable* if there exists a collection of functions $q_{i|B} : \mathsf{DATA}(\{i\} \cup B, d) \to \mathbb{R}$ where $i \in N$, $B \subseteq N \setminus \{i\}$, $d \geq 1$ such that, for every $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N, d)$ 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 D to the marginal space $\mathsf{X}_A \equiv \prod_{i \in A} \mathsf{X}_i$ for $\emptyset \neq A \subseteq N$ and $pa_G(i) \equiv \{j \in N;\, j \to i\}$ the set of *parents* of $i \in N$.

2.1.3 Inclusion neighborhood

The basic idea of local search method for the maximization of a quality criterion has already been explained in the Introduction. Now, we define the inclusion neighborhood formally. Given $G \in \mathsf{DAGS}(N)$, let $\mathcal{I}(G)$ denote the collection of conditional independence restrictions determined by G. Given $G, H \in \mathsf{DAGS}(N)$ if $\mathcal{I}(H) \subset \mathcal{I}(G)$,[2] but there is no $F \in \mathsf{DAGS}(N)$ with $\mathcal{I}(H) \subset \mathcal{I}(F) \subset \mathcal{I}(G)$ 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 $\mathcal{I}(H) \subseteq \mathcal{I}(G)$ provided by (Chickering 2002) together with his modification of the GES algorithm.

2.1.4 Essential graph

Given an (independence) equivalence class $\mathcal{G}$ 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 $G \in \mathcal{G}$,
- $a - b$ in G^* if there are $G, H \in \mathcal{G}$ such that $a \to b$ in H and $a \leftarrow b$ in G.

It is always a chain graph (= acyclic hybrid graph); this follows from graphical characterization of (graphs that are) essential graphs by Andersson, Madigan and Perlman (1997). Chickering (2002) used essential graphs as unique graphical BN structure representatives in his version of the GES algorithm. Note one can characterize all inclusion neighbors of a given BN structure in terms of the respective essential graph (Studený 2005b).

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

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

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

To write formulas for imsets we introduce the following notational convention. Given $A \subseteq N$, the symbol δ_A will denote a special imset given by:

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

By an *elementary imset* is meant an imset of the form

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

where $C \subseteq N$ and $a, b \in N\backslash C$ are distinct. In our algebraic framework it 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

$$u_G \ = \ \delta_N - \delta_\emptyset + \sum_{i\in N} \{\, \delta_{\mathrm{pa}_G(i)} - \delta_{\{i\}\cup \mathrm{pa}_G(i)} \,\}. \qquad (1)$$

The reader may ask whether one can efficiently represent standard imsets in the memory of a computer. Fortunately, it follows from (1) that u_G has at most $2 \cdot |N|$ non-zero values. Thus, one can keep only its non-zero values in the memory of a computer, which means 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. However, the number of summands in such decomposition only depends on u_G. It will be called the *degree* of u_G and denoted by $deg\,(u_G)$. The degree corresponds to the number $\boldsymbol{a}(G)$ of arrows in G as follows:

$$deg\,(u_G) = \frac{1}{2} \cdot |N| \cdot (|N| - 1) - \boldsymbol{a}(G), \qquad (2)$$

and can be computed from u_G directly by the formula

$$deg\,(u_G) = \langle m_*, u_G\rangle \quad \text{where } m_* : \mathcal{P}(N) \to \mathbb{Z} \qquad (3)$$

is given by $m_*(A) = \frac{1}{2} \cdot |A| \cdot (|A| - 1)$ for $A \subseteq N$; for the proof of these facts see Lemma 7.1 and § 4.2 in (Studený 2005).

It was shown as Corollary 7.1 in (Studený 2005) that, given $G, H \in \mathsf{DAGS}(N)$, one has $u_G = u_H$ iff they are independence equivalent. Moreover, Corollary 8.4 in (Studený 2005) says that $G, H \in \mathsf{DAGS}(N)$ are inclusion neighbors iff either $u_G - u_H$ or $u_H - u_G$ is an elementary imset. Finally, Lemmas 8.3 and 8.7 in (Studený 2005) together claim that every score equivalent and decomposable criterion $\mathcal{Q}$ necessarily has the form:

$$\mathcal{Q}(G, D) \ = \ s^{\mathcal{Q}}_D - \langle t^{\mathcal{Q}}_D, u_G\rangle$$

for any $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N, d)$, where $s^{\mathcal{Q}}_D \in \mathbb{R}$ and $t^{\mathcal{Q}}_D : \mathcal{P}(N) \to \mathbb{R}$ do not depend on G.

3 SOME GEOMETRIC CONCEPTS

In this section we recall some well-known concepts and facts from the theory of convex polytopes (Ziegler 1995) because we are not sure whether every reader is familiar with them. The proofs can be found in textbooks on linear programming (Schrijver 1986).

3.1 POLYTOPES AND POLYHEDRONS

These sets are special subsets of the Euclidean space $\mathbb{R}^K$, where K is a non-empty finite set.[3] Thus, the points in this space are vectors $\boldsymbol{v} = [v_i]_{i\in K}$. Given $\boldsymbol{x}, \boldsymbol{v} \in \mathbb{R}^K$ their scalar product is $\langle \boldsymbol{v}, \boldsymbol{x}\rangle = \sum_{i\in K} v_i \cdot x_i$.

A *polytope* in $\mathbb{R}^K$ is the convex hull of a finite set of points in $\mathbb{R}^K$. It is straightforward that the least set of points whose convex hull is a polytope P is the set of its *extreme points*, that is, of those points in P which cannot be written as convex combinations of the other points in P. In particular, the set of extreme points of P is finite. The *dimension* $\dim(\mathsf{P})$ of $\mathsf{P} \subseteq \mathbb{R}^K$ is the dimension of its affine hull

$$\mathrm{aff}(\mathsf{P}) = \{\sum_{\boldsymbol{v}\in\mathsf{R}} \lambda_{\boldsymbol{v}} \cdot \boldsymbol{v};\ \text{finite}\, \mathsf{R} \subseteq \mathsf{P},\ \lambda_{\boldsymbol{v}} \in \mathbb{R}, \sum_{\boldsymbol{v}\in\mathsf{R}} \lambda_{\boldsymbol{v}} = 1\}\,.^4$$

By convention, the dimension of the empty polytope is -1. A polytope is *full-dimensional* if $\dim(\mathsf{P}) = |K|$.

By an *affine half-space* in $\mathbb{R}^K$ is meant the set

$$H^+ = \{\boldsymbol{x} \in \mathbb{R}^K;\ \langle \boldsymbol{v}, \boldsymbol{x}\rangle \leq \alpha\}\,,$$

where $0 \neq \boldsymbol{v} \in \mathbb{R}^K$ is a non-zero vector and $\alpha \in \mathbb{R}$.[5] A *polyhedron* is the intersection of finitely many affine half-spaces. It is *bounded* if it does not contain a ray $\{\boldsymbol{x} + \alpha \cdot \boldsymbol{w};\ \alpha \geq 0\}$ for any $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^K$, $\boldsymbol{w} \neq 0$.

A well-known classic, but non-trivial, result is that $\mathsf{P} \subseteq \mathbb{R}^K$ is a polytope iff it is a bounded polyhedron - see Corollary 7.1.c in (Schrijver 1986) or Theorem 1.1 in (Ziegler 1995). A further important observation is that if P is a full-dimensional polytope then its *irredundant description* in the form of a polyhedron[6] is

[3] This is nothing but the "ordinary" finite-dimensional vector space $\mathbb{R}^k$ with $k = |K| \geq 1$. By the chosen notation we want to emphasize that enumerating elements of K (= fixing a total order of components in our vectors) is immaterial in our considerations.

[4] There is a unique linear subspace $\mathsf{L} \subseteq \mathbb{R}^K$ such that $\mathrm{aff}(\mathsf{P}) = \boldsymbol{w} + \mathsf{L}$ for some $\boldsymbol{w} \in \mathbb{R}^K$. The dimenstion of $\mathrm{aff}(\mathsf{P})$ is defined as the dimension of L.

[5] The "parameters" $\boldsymbol{v}$ and α defining the half-space H^+ are not uniquely determined. They are unique up to multiplication by a positive factor. Moreover, the same half-space can alternatively be written in the "dual" form $\{\boldsymbol{x} \in \mathbb{R}^K;\ \langle \boldsymbol{w}, \boldsymbol{x}\rangle \geq \beta\}$ with $0 \neq \boldsymbol{w} \in \mathbb{R}^K$, $\beta \in \mathbb{R}$ given by $\boldsymbol{w} = -\boldsymbol{v}$, $\beta = -\alpha$.

[6] By this we mean the intersection of such a collection of

unique - see claim (17) on page 102 of (Schrijver 1986). Provided that the polytope is *rational*, that is, it is the convex hull of a finite subset of $\mathbb{Q}^K$, the respective (irredundant) half-spaces are given by rational vectors and constants - see a note on p. 99 of (Schrijver 1986).

3.2 FACES OF A POLYTOPE

By an *affine hyperplane* in $\mathbb{R}^K$ is meant any set

$$H = \{\boldsymbol{x} \in \mathbb{R}^K;\, \langle \boldsymbol{v}, \boldsymbol{x}\rangle = \alpha\}\,,$$

where $0 \neq \boldsymbol{v} \in \mathbb{R}^K$ and $\alpha \in \mathbb{R}$. Evidently, the hyperplane deliminates two half-spaces in $\mathbb{R}^K$, namely H^+ mentioned above and $H^- = \{\boldsymbol{x} \in \mathbb{R}^K;\, \langle \boldsymbol{v}, \boldsymbol{x}\rangle \geq \alpha\}$.

We say that a hyperplane H *isolates* a polytope (= a bounded polyhedron) $\mathsf{P} \subseteq \mathbb{R}^K$ if either $\mathsf{P} \subseteq H^+$ or $\mathsf{P} \subseteq H^-$. A *face* F of P is the intersection of P with a hyperplane isolating it:

$$F = \mathsf{P} \cap H \quad \text{where either } \mathsf{P} \subseteq H^+ \text{ or } \mathsf{P} \subseteq H^-\,.^{7}$$

It follows from the definition that every face F of P is again a bounded polyhedron, and, therefore, a polytope. An equivalent definition of a face of a polytope is this: it is a convex (closed) subset $F \subseteq \mathsf{P}$ such that $\forall\, \boldsymbol{v}, \boldsymbol{w} \in \mathsf{P}$ and $\forall\, \alpha \in (0,1)$ whenever $\alpha \cdot \boldsymbol{v} + (1-\alpha)\cdot \boldsymbol{w} \in F$ then $\boldsymbol{v}, \boldsymbol{w} \in F$ - this follows from Theorem 7.5 in (Brøndsted 1983).

A crucial result - see Theorem 2.7 in (Ziegler 1995) - is as follows. The collection $\mathcal{F}(\mathsf{P})$ of all faces of a polytope P is finite. It is a lattice[8] with respect the ordering given by inclusion, called *face lattice*. Moreover, it is a graded lattice with the rank function $h(F) = \dim(F) + 1$, $F \in \mathcal{F}(\mathsf{P})$.[9] In particular, one can classify faces of P by their dimension.

The least element of $\mathcal{F}(\mathsf{P})$ is the empty face $\emptyset$ with dimension -1. The minimal non-empty faces of P have the dimension 0, that means, these are nothing but points in $\mathbb{R}^K$. They are named *vertices* of P. It follows from the equivalent definition of a face above that vertices of P coincide with its extreme points.

The faces of dimension 1, that is, line-segments, are called *edges* of P. By the equivalent definition of a face, an edge is a line-segment $E \subseteq \mathsf{P}$ connecting vertices of P such that $\mathsf{P} \setminus E$ is convex.

The greatest element in $\mathcal{F}(\mathsf{P})$ is P itself, with top dimension $\dim(\mathsf{P})$. The maximal proper faces of P, that is, maximal faces different from P, have dimension $\dim(\mathsf{P}) - 1$. They are named *facets* of P. Facets closely correspond to the unique irredundant description of P in the form of a polyhedron. The affine hulls of facets are just hyperplanes delimiting the respective half-spaces - see § 8.4 in (Schrijver 1986).

Finally, there are software packages that allow one to compute facet-defining inequalities[10] on the basis of the list of vertices of a rational polytope (Franz 2006).

4 MAIN RESULT

In this section we prove the main result and illustrate it by an example with 3 variables.

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

Proof: Having fixed a (non-empty finite) set of variables N, let us denote by S the set of standard imsets over N:

$$S \equiv \{u_G;\, G \in \mathsf{DAGS}(N)\} \subseteq \mathbb{R}^{\mathcal{P}(N)}\,.^{11}$$

We introduce $\mathsf{P} \subseteq \mathbb{R}^{\mathcal{P}(N)}$ as the convex hull of S. Of course, it is a rational polytope by definition and the set of extreme points of P is a subset of S. Since now we assume $|N| \geq 2$ because if $|N| = 1$ then S contains only the zero imset, $\mathsf{P} = S$ and the statements of the theorem are trivial.

As concerns the first statement, it suffices to prove that none of the standard imsets is a convex combination of the other (elements of S). Indeed, since vertices of P coincide with its extreme points one has to show, for any $u = u_G$, $G \in \mathsf{DAGS}(N)$ that u is not a combination of other points in P. Assume for a contradiction that $u = \sum_{t\in T} \alpha_t \cdot t$, where $\alpha_t > 0$, $\sum_t \alpha_t = 1$, for a finite set $T \subseteq \mathsf{P}$ with $u \notin T$. Each $t \in T$ can be written as a convex combination of points from S: $t = \sum_{v\in S} \beta_v^t \cdot v$, $\beta_v^t \geq 0$, $\sum_v \beta_v^t = 1$. Hence, by a substitution we derive $u = \sum_{v\in S}(\sum_{t\in T} \alpha_t \beta_v^t)\cdot v$ and put $\gamma_v = \sum_{t\in T} \alpha_t \beta_v^t$. We have $\gamma_v \geq 0$ and $\sum_{v\in S} \gamma_v = 1$ and, since $u \notin T$, there exists at least one $v \in S\setminus\{u\}$ with $\gamma_v > 0$. Thus, $\sum_{v\in S\setminus\{u\}} \gamma_v = 1 - \gamma_u > 0$ and $u = \sum_{v\in S} \gamma_v \cdot v$ implies

---

half-spaces in which no half-space can be dropped without changing the polyhedron.

[7] This is the definition of a face from (Ziegler 1995); Schrijver (1986) in § 8.3 considers only non-empty faces.

[8] A *lattice* is a partially ordered set in which every pair of elements has both the least upper bound and the greatest lower bound.

[9] By a *graded lattice* is meant a lattice $\mathcal{F}$ with a (rank) function $h : \mathcal{F} \to \{0, 1, \ldots r\}$, $r \in \mathbb{N}$ such that $h(y) = h(x) + 1$ whenever $x, y \in \mathcal{F}$ are such that $x \prec y$ and there is no $z \in \mathcal{F}$ with $x \prec z \prec y$.

[10] These are inequalities $\langle \boldsymbol{v}, \boldsymbol{x}\rangle \leq \alpha$ for $\boldsymbol{x} \in \mathsf{P}$ that define the respective irredundant half-spaces.

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

$u = \sum_{v \in S\setminus\{u\}} \frac{\gamma_v}{1-\gamma_u} \cdot v$. This, however, contradicts the premise for our contradiction proof.

Now, we are going to verify that no $u = u_G \in S$ is a convex combination of points in $S\setminus\{u\}$ by constructing a linear function L on $\mathbb{R}^{\mathcal{P}(N)}$ such that $L(u) < L(v)$ for any $v \in S\setminus\{u\}$.[12] It will be in the form $L(v) = \langle m, v\rangle$, $v \in \mathbb{R}^{\mathcal{P}(N)}$ where $m : \mathcal{P}(N) \to \mathbb{R}$ is a suitable real function (= a point in $\mathbb{R}^{\mathcal{P}(N)}$).

In the construction we utilize the properties of the *multiinformation function* m_P for a (discrete) probability distribution P over N - see § 2.3.4 in (Studený 2005). It is a function $m_P : \mathcal{P}(N) \to [0, \infty)$ which ascribes to every $A \subseteq N$ the multiinformation of the respective marginal P^A of P for A.[13] The basic property of the multiinformation function is that it is *supermodular*[14] and characterizes conditional independence statements in P by algebraic identities. In particular, (the proof of) Proposition 5.3 in (Studený 2005) implies that the independence structure $\mathcal{M}(m_P)$ produced by m_P through the respective algebraic test[15] coincides with the collection $\mathcal{I}(P)$ of conditional independence statements represented in P.

A further preparatory observation concerns standard imsets. Lemma 7.1 in (Studený 2005) 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 § 4.2.3 in (Studený 2005).[16] Moreover, Lemma 7.1 also says that the independence structure $\mathcal{M}(u_G)$ induced by the imset u_G through the respective algebraic criterion coincides with the collection $\mathcal{I}(G)$ of conditional independence restrictions in G (by the respective graphical separation criterion).

In the sequel, we will use the following notation: given $v \in S$ the symbol $\mathcal{I}(v)$ will denote the collection of conditional independence restrictions $\mathcal{I}(H)$ determined by (any) graph $H \in \mathsf{DAGS}(N)$ with $v = u_H$.[17] A further observation is that, for every $u, v \in S$ the (strict) inclusion $\mathcal{I}(v) \subset \mathcal{I}(u)$ implies $deg(v) < deg(u)$. Indeed if $u = u_G$ and $v = u_H$, where $G, H \in \mathsf{DAGS}(N)$, then $\mathcal{I}(H) \subset \mathcal{I}(G)$. Thus, it follows from the characterization of inclusion in (Chickering 2002) that H has a higher number of arrows than G: $\boldsymbol{a}(H) > \boldsymbol{a}(G)$. Hence, by (2) one has $deg(u_H) < deg(u_G)$.

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

$$S(u) = \{v \in S;\, \mathcal{I}(v) \subseteq \mathcal{I}(u)\}\,.$$

The crucial step in our proof is that we utilize a well-known result (Geiger and Pearl 1990) on the existence of a perfectly Markovian distribution for an acyclic directed graph. If applied to our fixed $G \in \mathsf{DAGS}(N)$ it says there exists a discrete probability distribution P over N such that $\mathcal{I}(P) = \mathcal{I}(G)$. We take such a distribution P, fix it, consider its multiinformation function m_P and interpret it as a point in $\mathbb{R}^{\mathcal{P}(N)}$.

The next step is to realize that $\langle m_P, v\rangle \geq 0$ for any $v \in S$ and one has $\langle m_P, v\rangle = 0$ iff $v \in S(u)$. The first claim follows from Proposition 5.1(i) in (Studený 2005) saying that $\langle \tilde{m}, v\rangle \geq 0$ for any supermodular function $\tilde{m}$ and a structural imset v. As explained above, these assumptions are valid for m_P and any $v \in S$. As concerns the second claim, Proposition 5.6 in (Studený 2005) says, under the same assumptions, that $\langle m_P, v\rangle = 0$ iff $\mathcal{M}(v) \subseteq \mathcal{M}(m_P)$. However, as explained above, one has $\mathcal{M}(m_P) = \mathcal{I}(P)$ and provided that $v = u_H$, $H \in \mathsf{DAGS}(N)$ one also has $\mathcal{M}(v) = \mathcal{M}(u_H) = \mathcal{I}(H) = \mathcal{I}(v)$. Thus, $\langle m_P, v\rangle = 0$ iff $\mathcal{I}(v) \subseteq \mathcal{I}(P)$. However, since $\mathcal{I}(P) = \mathcal{I}(G) = \mathcal{I}(u)$ the inclusion $\mathcal{I}(v) \subseteq \mathcal{I}(P)$ means $v \in S(u)$.

Thus, because $\langle m_P, v\rangle > 0$ for any $v \in S \setminus S(u)$ we know that $k \equiv \min_{v \in S\setminus S(u)} \langle m_P, v\rangle > 0$. Put $D \equiv \max_{v \in S} deg(v)$. Actually, we know by (2) that $D = \frac{1}{2} \cdot |N| \cdot (|N| - 1)$, and, therefore, $D > 0$. Let us choose $\varepsilon > 0$ with $\varepsilon < \frac{k}{D}$ and put

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

where $m_* : \mathcal{P}(N) \to \mathbb{Z}$ is the function from (3). Finally, we define a linear function $L : \mathbb{R}^{\mathcal{P}(N)} \to \mathbb{R}$ by the formula

$$L(v) \equiv \langle m, v\rangle \quad \text{for } v \in \mathbb{R}^{\mathcal{P}(N)}\,.$$

It follows from (3) that $\forall\, v \in S$ one has $L(v) = \langle m_P, v\rangle - \varepsilon \cdot deg(v)$. In particular, $L(u) = -\varepsilon \cdot deg(u)$. If $v \in S \setminus S(u)$ then

$$L(v) = \langle m_P, v\rangle - \varepsilon \cdot deg(v) \geq k - \varepsilon \cdot D > 0 \geq L(u)\,.$$

Moreover, if $v \in S(u)$, $v \neq u$ then $\mathcal{I}(v) \subset \mathcal{I}(u)$[18] and, by the above observation, $deg(u) > deg(v)$. Hence, since $\langle m_P, v\rangle = 0 = \langle m_P, u\rangle$ one has

$$L(v) - L(u) = \varepsilon \cdot (deg(u) - deg(v)) > 0\,.$$

[12] This is enough, for if we put $\ell = \min_{v \in S\setminus\{u\}} L(v)$ then $u = \sum_{v \in S\setminus\{u\}} \alpha_v \cdot v$, $\alpha_v \geq 0$, $\sum_v \alpha_v = 1$ gives a contradiction: $L(u) = \sum_v \alpha_v \cdot L(v) \geq \ell > L(u)$.

[13] The multiinformation of R (over A) is the relative entropy of R with respect to the product $Q = \prod_{i \in A} R^i$ of its one-dimensional marginals: $H(R|Q) \equiv \sum_x r(x) \cdot \ln \frac{r(x)}{q(x)}$.

[14] This means $m_P(C \cup D) + m_P(C \cap D) \geq m_P(C) + m_P(D)$ for any $C, D \subseteq N$.

[15] In this paper we omit the definition of that algebraic test because it is not relevant here.

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

[17] The definition is correct since one has $u_G = u_H$ iff $G, H \in \mathsf{DAGS}(N)$ are independence equivalent.

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

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

As concerns the second statement, realize that every standard imset $u : \mathcal{P}(N) \to \mathbb{Z}$ 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 $\mathcal{K} \equiv \{A \subseteq N;\, |A| \geq 2\}$. This defines a one-to-one linear transformation. Thus, to prove what is desired it suffices to show that the linear hull of $\mathcal{K}$-restrictions of standard imsets is just $\mathbb{R}^{\mathcal{K}}$, which has the dimension $|\mathcal{K}|$. Because every elementary imset is standard,[19] it is enough to show that, for every $A \in \mathcal{K}$, the ($\mathcal{K}$-restriction of the) imset δ_A is a linear combination of ($\mathcal{K}$-restrictions of) elementary imsets. This can be done easily by induction on $|A|$. □

Example Let us describe the situation in the case of 3 variables. Then one has 11 standard imsets and they break into 5 types (= permutation equivalence classes). They can also be classified by their degree. More specifically:

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

By the theorem above, the dimension of the polytope generated by these 11 imsets is 4. To get its irredundant description in the form of a polyhedron it is suitable to have it embedded (as a full-dimensional polytope) in a 4-dimensional space. To this end, we decided to identify every standard imset over N with its restriction to $\mathcal{K} \equiv \{A \subseteq N;\, |A| \geq 2\}$. The we used the computer package Convex (Franz 2006) to get 13 facet-defining inequalities, which break into 7 types. They can be classified as follows:

- Five facets include the zero imset. The respective inequalities break into 3 types, namely $0 \leq \delta_{abc}$, $0 \leq \delta_{abc} + \delta_{ab}$ and $0 \leq 2 \cdot \delta_{abc} + \delta_{ab} + \delta_{ac} + \delta_{bc}$.
- Eight facets include the imset of maximal degree 3. The respective inequalities break into 4 types, namely $\delta_{abc} \leq 1$, $\delta_{abc} + \delta_{ab} \leq 1$, $\delta_{abc} + \delta_{ab} + \delta_{ac} \leq 1$ and $\delta_{abc} + \delta_{ab} + \delta_{ac} + \delta_{bc} \leq 1$.

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 respective polytope is 11. The number of its facets is 154.

5 GEOMETRIC NEIGHBORHOOD

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

Example We characterized the geometric neighborhood in the case of 3 variables and compared it with the inclusion neighborhood. In this case, the inclusion neighborhood is contained in the geometric one, which leads us to a conjecture this holds universally. The result is depicted in Figure 1, 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 degree of the respective standard imset.

An interesting observation is that differential imsets for geometric neighbors (that are not inclusion neighbors) are quite simple in the case of 3 variables: they all have only four non-zero values, like elementary imsets (= differential imsets for inclusion neighbors).

We made similar computation also in the case of 4 variables. All the results and also (the description of) our method for computing the geometric neighborhood is available at

`http://www.utia.cas.cz/vomlel/imset/polytopes-3v-and-4v.html`

6 OPEN PROBLEMS

We conclude our paper with the list of open questions. Of course, we would like to know whether one can possibly get the characterization of the standard imset polytope P in the form of a polyhedron for general N. However, we are slightly sceptical about this direction.

As concerns the geometric neighborhood, we would like to verify the conjecture that it always contains the inclusion neighborhood. It would be nice to know

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

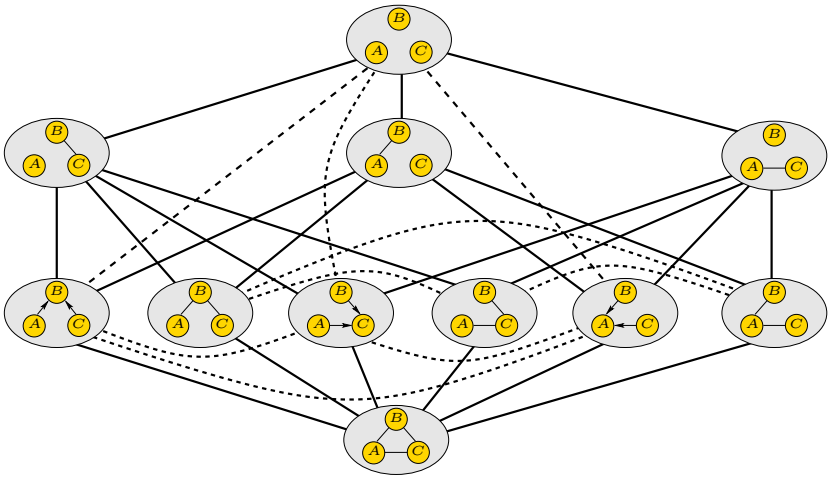

Figure 1: The geometric and inclusion neighborhood (for essential graphs) in the case of 3 variables.

how "dense" the geometric neighborhood is, that is, what is the "average" number of geometric neighbors of a given BN structure. Is the number of neighbors somehow related to the degree of the respective standard imset? We would also like to know what are differential imsets for geometric neighbors.

All these questions concern the complexity of a potential (future) greedy search procedure for maximization of a quality criterion $\mathcal{Q}$ based on the geometric neighborhood. Such a procedure would be particularly valuable, because, as mentioned in the Introduction, it should find the global maximum of $\mathcal{Q}$. Note that the resulting standard imset can be transformed to the respective essential graph by the algorithm described in (Studený and Vomlel 2008).

Acknowledgements

We are grateful to our colleague Tomáš Kroupa for his help with computation. This research has been supported by the grants GAČR n. 201/08/0539 and MŠMT n. 1M0572, and n. 2C06019.

References

S.A. Andersson, D. Madigan and M.D. Perlman (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics* **25**: 505-541.

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

A. Brøndsted (1983). *An Introduction to Convex Polytopes.* New York: Springer-Verlag.

R. Castelo (2002). The discrete acyclic digraph Markov model in data mining. PhD thesis, University of Utrecht.

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

M. Franz (2006). Convex - a Maple package for convex geometry, version 1.1, available at `http://www-fourier.ujf-grenoble.fr/~franz/convex/`

D. Geiger and J. Pearl (1990). On the logic of causal models. in *Uncertainty in AI 4*, North-Holland: 3-14.

T. Kočka (2001). Graphical models: learning and application. PhD thesis, University of Economics Prague.

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

C. Meek (1997). Graphical models, selecting causal and statistical models. PhD thesis, Carnegie Melon University.

R.E. Neapolitan (2004). *Learning Bayesian Networks.* New York: Pearson Prentice Hall.

J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems.* San Mateo: Morgan Kaufmann.

A. Schrijver (1986). *Theory of Linear and Integer Programming.* Chichester: John Wiley.

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

M. Studený (2005b). Characterization of inclusion neighbourhood in terms of the essential graph. *International Journal of Approximate Reasoning* **38**: 283-309.

M. Studený and J. Vomlel (2008). A reconstruction algorithm for the essential graph. Submitted to *International Journal of Approximate Reasoning.*

T. Verma and J. Pearl (1991). Equivalence and synthesis of causal models, in *Uncertainty in AI 6*, Elsevier: 220-227.

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