A Geometric Approach to Learning BN Structures

M. Studený and J. Vomlel
Institute of Information Theory and Automation of the ASCR
Prague, CZ 18208, Czech Republic

Abstract

We recall the basic idea of an algebraic approach to learning a 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. This leads to an 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 greedy search technique is based on the geometric neighborhood instead.

1 Introduction

The motivation for this theoretical paper is learning a Bayesian network (BN) structure from data by the method of maximization of a quality criterion (= the score and search method). By a *quality criterion*, also named a *score metric* 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 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, Madigan and Perlman (1997), who gave a graphical characterization of essential graphs.

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). 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 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. 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), which means it ascribes the same value to equivalent graphs. The other requirement is that $\mathcal{Q}$ should be *decomposable* (Chickering, 2002), 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 (Studený, 2005), 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 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 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 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 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. One 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 of maximizing 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 in describing the above-mentioned polytope in the form of a (bounded) polyhedron, one gets a classic task of linear programming, namely to find an extremal value of 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 $|N| = 4$.

However, because it is not clear at this moment how to find the "polyhedral" description of the polytope for arbitrary $|N|$, we propose an alternative approach 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). Thus, provided one succeeds at characterizing 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. 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 neighbor-

hood and geometric neighborhood differ already in the case of 3 variables. 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 $\mathcal{Q}$. We give such an example and claim that this is an inevitable defect of the inclusion neighborhood, which may occur whenever the data faithfulness 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.

2 Basic Concepts

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

2.1 BN 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 (Neapolitan 2004). A well-known fact is that P factorizes according to G iff 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 (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.

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*, which means they determine the same collection of conditional independence restrictions – cf. § 2.2 in (Neapolitan, 2004). Both Frydenberg (1990), and 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*, i.e. 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 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 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.

A classic example 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 (Schwarz, 1978). To give a direct formula for BIC (in this 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). 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 for (the code of) a parent configuration. Given a database D of the length $d \geq 1$ let d_{ijk} denote the

number of occurences 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 (Studený, 2005):

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

In this brief overview we omit the question of statistical consistency of quality criteria; we refer the reader to the literature on this topic (Chickering, 2002; Neapolitan, 2004). A quality criterion $\mathcal{Q}$ will be named *score equivalent* if, for every $D \in \mathsf{DATA}(N,d)$, $d \geq 1$,

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

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 $\emptyset \neq A \subseteq N$.

2.1.3 Inclusion neighborhood

The basic idea of local search methods for the maximization of 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 $\mathcal{I}(G)$ denote the collection of conditional independence restrictions determined by G. Given $G,H \in \mathsf{DAGS}(N)$, if $\mathcal{I}(H) \subset \mathcal{I}(G)$,[1] 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.

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

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).

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.

2.2 Standard Imset

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:

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

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

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

By an *elementary imset* is meant an imset

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

where $C \subseteq N$ and $a, b \in N \setminus C$ are distinct. 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

$$u_G = \delta_N - \delta_\emptyset + \sum_{i \in N} \{ \delta_{pa_G(i)} - \delta_{\{i\} \cup pa_G(i)} \}. \quad (1)$$

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, which means the memory demands for representing standard imsets are polynomial in the number of variables.

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:

$$\mathcal{Q}(G, D) = s_D^{\mathcal{Q}} - \langle t_D^{\mathcal{Q}}, u_G \rangle \quad (2)$$

for $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N, d)$, $d \geq 1$ where the constant $s_D^{\mathcal{Q}} \in \mathbb{R}$ and the (data) vector $t_D^{\mathcal{Q}} : \mathcal{P}(N) \to \mathbb{R}$ do not depend on G.

The reader can object that the dimension of $t_D^{\mathcal{Q}}$ grows exponentially with $|N|$, making the method unfeasible for many "real-world" problems. However, since $2^{|N|} \leq |\mathsf{X}_N|$, the representation of a database D in the form of a data vector may appear to be even more effective than (one of the traditional ways) in the form of a contingency table! Another point is that to compute $\langle t_D^{\mathcal{Q}}, u_G \rangle$ one only needs at most $2 \cdot |N|$ values of the data vector. 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

In this section we recall well-known concepts and facts from the theory of convex polytopes (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. 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 $\mathbb{R}^K$; if the set consists of points in $\mathbb{Q}^K$, the polytope is *rational*. It is straightforward that the smallest set of points whose convex hull is a polytope P is the set of its *vertices* ($\equiv$ *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 vertices of P is finite. The *dimension* $\dim(\mathsf{P})$ of $\mathsf{P} \subseteq \mathbb{R}^K$ is the dimension of its affine hull $\mathrm{aff}(\mathsf{P})$, which is the collection of affine combinations $\sum_{\boldsymbol{v} \in \mathsf{R}} \lambda_{\boldsymbol{v}} \cdot \boldsymbol{v}$, where $\emptyset \neq \mathsf{R} \subseteq \mathsf{P}$ is finite and $\lambda_{\boldsymbol{v}} \in \mathbb{R}$, $\sum_{\boldsymbol{v} \in \mathsf{R}} \lambda_{\boldsymbol{v}} = 1$.[2] A polytope is *full-dimensional* if $\dim(\mathsf{P}) = |K|$.

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

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

where $0 \neq \boldsymbol{v} \in \mathbb{R}^K$ and $\alpha \in \mathbb{R}$. 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). A further important observation is that if P is a full-dimensional polytope then its *irredundant description* in the form of a polyhedron[3] is unique – see claim (17) on page 102 of (Schrijver, 1986).

Finally, there are software packages that allow one, on the basis of the list of vertices of a rational polytope P, to compute all inequalities defining an irredundant polyhedral description of P, e.g. (Franz, 2006).

[2] 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 dimension of $\mathrm{aff}(\mathsf{P})$ is defined as the dimension of L.

[3] By this is meant the intersection of such a collection of half-spaces in which no half-space can be dropped without changing the polyhedron.

4 Main Result

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

$$S \equiv \{u_G;\ G \in \mathsf{DAGS}(N)\} \subseteq \mathbb{R}^{\mathcal{P}(N)}.^4$$

Theorem 1. *The set S 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 $2^{|N|} - |N| - 1$.*

Because of a limited scope for this paper we skip the proof, which can be found in (Studený and Vomlel, 2008).

EXAMPLE Let us describe the situation in the case of three variables. Then one has 11 standard imsets and they break into 5 types (= permutation equivalence classes). They can also be classified by the number of edges in the corresponding essential graph. (c.f. Figure 1 below)

- The zero imset corresponds to the complete (undirected) essential graph.
- Six elementary imsets break into two types, namely $u_{\langle a,b|\emptyset\rangle}$ and $u_{\langle a,b|c\rangle}$; the essential graphs are $a \rightarrow c \leftarrow b$ and $a - c - b$.
- Three "semi-elementary" imsets of the form $u_{\langle a,bc|\emptyset\rangle} \equiv \delta_{abc} + \delta_\emptyset - \delta_a - \delta_{bc}$ define one type; the essential graphs have just one undirected edge.
- The imset $\delta_N - \sum_{i\in N} \delta_i + 2\cdot\delta_\emptyset$ 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 $\mathcal{K} \equiv \{A \subseteq N;\ |A| \geq 2\}$. Then we used the computer package Convex (Franz, 2006) to get all 13 polyhedron-defining inequalities. They break into 7 types and can be classified as follows:

- Five inequalities hold with equality for the zero imset. They break into 3 types: $0 \leq 2\cdot\delta_{abc} + \delta_{ab} + \delta_{ac} + \delta_{bc}$, $0 \leq \delta_{abc} + \delta_{ab}$ and $0 \leq \delta_{abc}$.
- Eight inequalities achieve equality for the imset corresponding to the empty graph. They 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 polytope is 11. The number of corresponding polyhedron-defining inequalities is 154 – see vertex-facet table in (Vomlel and Studený, 2008).

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 (2), 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 (Studený, 2005):

$$t_D^{\mathsf{BIC}}(A) = d\cdot H(\hat{P}_A|\prod_{i\in A}\hat{P}_i) - \frac{\ln d}{2}\cdot\{\,|A| - 1 + \prod_{i\in A} r(i) - \sum_{i\in A} r(i)\}$$

for $A \subseteq N$, where $H(*|*)$ is the relative entropy and $\hat{P}_A$ is the marginal empirical distribution given by (the projection of the database) D_A.

5 Geometric Neighborhood

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

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

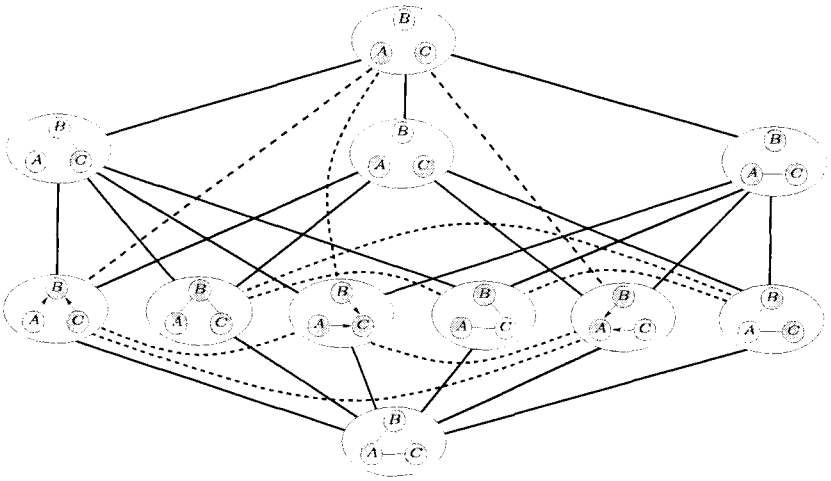

Figure 1: The geometric and inclusion neighborhood (for essential graphs) in the case of 3 variables.

EXAMPLE We characterized the geometric neighborhood in the case of three variables and compared it with the inclusion neighborhood. We found out that the inclusion neighborhood is contained in the geometric one. 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 numbers of edges.

We made a similar computation also in the case of four variables – see vertex-vertex table in (Vomlel and Studený, 2008). The description of our method for computing the geometric neighborhood is also available at (Vomlel and Studený, 2008).

5.1 GES Failure

What does it mean that $u, v \in S$ are geometric but not inclusion neighbors? The fact that they are geometric neighbors means there exists a linear function on $\mathbb{R}^{\mathcal{P}(N)}$ achieving its maximum over S just in $\{u, v\}$. Analogously, since u is a vertex of P, there exists (another) linear function achieving its maximum just in u. Therefore, by a suitable convex combination of these functions, one can construct a linear function L on $\mathbb{R}^{\mathcal{P}(N)}$ such that $L(u) > L(v) > L(w)$ for any $w \in S \setminus \{u, v\}$. Provided u and v are not inclusion neighbors, L achieves its local maximum (with respect the inclusion neighborhood) in v and the global maximum over S in u.

Now, it has already been explained that every "reasonable" quality criterion $\mathcal{Q}$ is (the restriction of) an affine function on $\mathbb{R}^{\mathcal{P}(N)}$. Thus, the reader may ask whether this may happen for $\mathcal{Q}$ in place of L. Indeed, this is true in the case of three variables for the imset $u = u_{\langle a,c|\emptyset\rangle}$, which corresponds to an "immorality" $a \to b \leftarrow c$ and the imset v corresponding to the empty graph – see Figure 1.

EXAMPLE There exists a database D (of the length $d = 4$) over $N = \{a, b, c\}$ such that the BIC criterion achieves its local maximum in the empty graph G^0 and its global maximum in (any of) the graph(s) $\widehat{G}$ of the type $a \to b \leftarrow c$. Put $\mathsf{X}_i = \{0, 1\}$ for $i \in N$ and $x^1 = (0, 0, 0)$, $x^2 = (0, 1, 1)$, $x^3 = (1, 0, 1)$, $x^4 = (1, 1, 0)$. Then direct computation of BIC (see § 2.1.2) gives $\mathsf{BIC}(\widehat{G}) = -14 \ln 2$, $\mathsf{BIC}(G^0) = -15 \ln 2$ and $\mathsf{BIC}(G') = -16 \ln 2$ for any graph G' over N having just one edge.

The reader may object that this is perhaps a rare casual example because of a short database. However, BIC exhibits the same behavior if the database D is multiplied! The limited scope of this contribution does not allow us to give the arguments why (we think) this is, actually, asymptotic behavior of any consistent score equivalent decomposable criterion $\mathcal{Q}$, provided the database is "generated" from the empirical

distribution $\hat{P}$ given by D. The point is that $\hat{P}$ is **not** perfectly Markovian with respect to any $G \in \mathsf{DAGS}(N)$.

In particular, the GES algorithm – see (Chickering, 2002) for details about this algorithm – should (asymptotically) learn the empty graph G^0, while 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 $\hat{P}$.

6 Conclusion

In our view, this is an example of the failure of the GES algorithm which may occur whenever a disputable *data faithfulness assumption* is not fulfilled.[5] This assumption is "valid" if data are artificially generated, but, in our view, one can hardly ensure its validity for "real" data.

On the other hand, the point of the example from 5.1 is that the GES algorithm is based on the inclusion neighborhood. This cannot happen if the greedy search technique is based on the geometric neighborhood. Indeed, we are able to show that each local maximum (of an affine function) with respect to the geometric neighborhood is necessarily a global maximum (over P). The proof is at the manuscript stage and will be published later. The conjecture that the inclusion neighborhood is always contained in the geometric one has recently been confirmed by Raymond Hemmecke (personal communication). Therefore, we think the concept of geometric neighborhood is quite important. We plan to direct our future research effort to algorithms for its efficient computation.

Acknowledgements

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

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

References

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

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

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

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

M. Frydenberg. 1990. The chain graph Markov property. *Scandinavian Journal of Statistics*, 17:333-353.

S.L. Lauritzen. 1996. *Graphical Models*. Clarendon Press.

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

R.E. Neapolitan. 2004. *Learning Bayesian Networks*. Pearson Prentice Hall.

J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

A. Schrijver. 1986. *Theory of Linear and Integer Programming*. John Wiley.

G. Schwarz. 1978. Estimation the dimension of a model. *The Annals of Statistics*, 6:461-464.

M. Studený. 2005. *Probabilistic Conditional Independence Structures*. Springer-Verlag.

M. Studený and J. Vomlel. 2008. Geometric view on learning Bayesian network structures. A draft available at `http://staff.utia.cas.cz/studeny/c11.html`

T. Verma and J. Pearl. 1991. Equivalence and synthesis of causal models. In *6th Conferece on Uncertainty in Artificial Intelligence*, pages 220–227.

J. Vomlel and M. Studený. 2008. Geometric neighborhood for Bayesian network structures over three and four variables. Web page, see `http://www.utia.cas.cz/vomlel/imset/polytopes-3v-and-4v.html`