A Geometric Approach to Learning BN Structures

Milan Studený and Jiří Vomlel

Institute of Information Theory and Automation of the ASCR
Prague, Czech Republic

PGM 2008, Hirtshals,

September 17, 2008, 11:30-12:00

Summary of the talk

1. Introduction
2. Learning concepts
3. Algebraic approach
4. A course from convex geometry
5. The main result
6. Geometric neighborhood
7. Conlusions

Introduction: Bayesian networks

Bayesian networks (BN) are popular (graphical) models in the area of probabilistic reasoning. Most working probabilistic expert systems are based on the mathematical theory related to Bayesian networks.

The motivation for this talk is *learning Bayesian network structure* from data by the method of maximization of a quality criterion (= score and search method).

By a *quality criterion*, also named a *score metric* or a *score*, is meant a special real function $\mathcal{Q}$ of the BN structure, usually represented by a graph G, and of the database D.

There are two important technical requirements on a quality criterion $\mathcal{Q}$ brought in connection with the maximization problem. One of them is that $\mathcal{Q}$ should be *score equivalent* (Bouckaert 1995), the other is that $\mathcal{Q}$ should be *decomposable* (Chickering 2002).

Introduction: algebraic approach

The basic idea of an algebraic approach to learning BN structures (Studený 2005) is to represent both the BN structure and the database by a real vector.

The algebraic representative of the BN structure given by an acyclic directed graph G is a certain integral (= integer-valued) vector u_G, called the *standard imset* (for G).

The crucial point is that every score equivalent and decomposable criterion $\mathcal{Q}$ is an affine function (= sum of a constant and a linear function) of the standard imset. More specifically, one has

$$\mathcal{Q}(G,D) \,=\, s_D^{\mathcal{Q}} - \langle t_D^{\mathcal{Q}}, u_G \rangle \,, \qquad \text{where } s_D^{\mathcal{Q}} \in \mathbb{R},$$

$t_D^{\mathcal{Q}}$ is a real vector of the same dimension as u_G and $\langle *, * \rangle$ denotes the scalar product. The vector $t_D^{\mathcal{Q}}$ is named the *data vector* (relative to $\mathcal{Q}$).

Introduction: geometric view

The aim of this contribution 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 corresponding Euclidean space. The result presented here is that it is *the set of vertices (= extreme points) of a certain polytope*.

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: to maximize/minimize a linear function over a polyhedron.

Motivated by the idea of possible use of the *simplex method* (Schrijver 1986), we have introduced the concept of *geometric neighborhood* for standard imsets and made a drafty analysis in the case 3 and 4 variables.

Learning concepts: Bayesian network structure

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 is Markovian with respect to G. (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 are Markovian with respect to G.

To name the shared features of distributions in this class one can use the phrase *BN structure*.

Learning concepts: quality criterion

Data 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 $\mathrm{DATA}\,(N,d)$ denote the collection of databases over N of the length d. Moreover, let $\mathrm{DAGS}\,(N)$ denote the collection of all acyclic directed graphs over N.

Definition (quality criterion)

Quality criterion or a *score* (for learning BN structures) is a real function $\mathcal{Q}(G,D)$ on $\mathrm{DAGS}\,(N) \times \mathrm{DATA}\,(N,d)$.

In this brief overview we omit examples of quality criteria and the question of their statistical consistency.

Learning concepts: score equivalent criterion

Since the aim of the learning procedure is to get the BN structure it is natural to require that the quality criterion satisfies the following condition:

Definition (score equivalent criterion)

A quality criterion $\mathcal{Q}$ will be named *score equivalent* if, for every $D \in \mathrm{DATA}\,(N, d)$, $d \geq 1$, one has

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

are independence equivalent.

Most quality criteria used in practice are score equivalent.

Learning concepts: decomposable criterion

Definition (decomposable criterion)

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

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

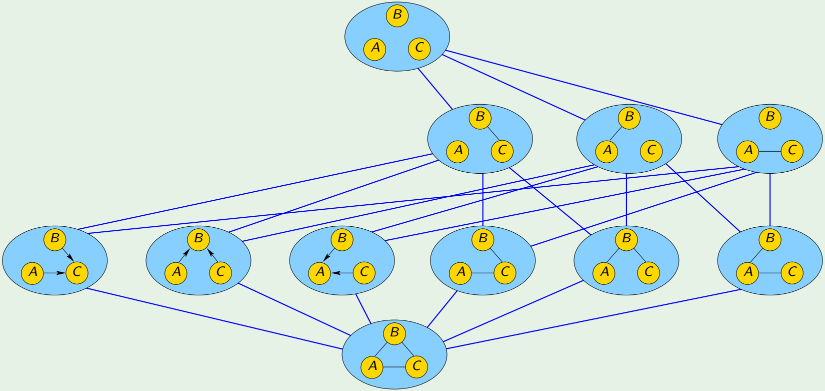
where $D_A : x_A^1, \ldots, x_A^d$ denotes the projection of D to the marginal space $\mathsf{X}_A \equiv \prod_{i \in A} \mathsf{X}_i$ for $\emptyset \neq A \subseteq N$ and $pa_G(i) \equiv \{j \in N;\, j \to i\}$ the set of *parents* of $i \in N$.

All criteria used in practice are decomposable.

Learning concepts: local search methods

The basic idea is that one introduces a *neighborhood relation* between BN structure representatives. The one uses *greedy search* techniques to find a local maximum of $\mathcal{Q}$ with respect to that neighborhood concept.

Example (essential graphs and inclusion neighborhood for 3 variables)



Algebraic approach: imset

- N ... a finite set of variables
- $\mathcal{P}(N) \equiv \{A; A \subseteq N\}$ ... the power set of N

Definition (imset)

An imset u is a function $u : \mathcal{P}(N) \mapsto \mathbb{Z}$.

We will regard is 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. The symbol $\langle m, u \rangle$ will then denote the scalar product of two vectors of this type:

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

Algebraic approach: elementary imset

Given $A \subseteq N$, the symbol δ_A will denote a special imset given by:

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

Definition (elementary imset)

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

In this algebraic framework it encodes an elementary conditional independence statement $a \perp\!\!\!\perp b \,|\, C$.

Algebraic approach: standard imset

Definition (standard imset)

The *standard imset* for an acyclic directed graph G is given by the formula

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

Here $pa_G(a) \equiv \{ b \in N;\ b \rightarrow a \text{ in } G \}$ denotes the set of *parents* of the node a.

The standard imset is uniquely determined representative of a Bayesian network structure.

Since every standard imset over N has at most $2 \cdot |N|$ non-zero values, it can be represented in the memory of a computer *with polynomial complexity* with respect to $|N|$.

Convex geometry: polytopes and polyhedrons

Consider the Euclidean space $\mathbb{R}^K$, where K is a non-empty finite set.

Definition (polytope)

A *polytope* in $\mathbb{R}^K$ is the convex hull of a finite set of points in $\mathbb{R}^K$. Its *dimension* $\dim(\mathrm{P})$ is the dimension of its affine hull.

The least set of points whose convex hull is a polytope P is the set of its *extreme points*.

Definition (polyhedron)

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

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

where $0 \neq \mathbf{v} \in \mathbb{R}^K$ is a non-zero vector 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 $\{\mathbf{x} + \alpha \cdot \mathbf{w};\, \alpha \geq 0\}$ for any $\mathbf{x}, \mathbf{w} \in \mathbb{R}^K$, $\mathbf{w} \neq 0$.

Convex geometry: Weyl-Minkowski theorem

Theorem (Weyl-Minkowski theorem)

A set $\mathrm{P} \subseteq \mathbb{R}^K$ *is a polytope iff it is a bounded polyhedron.*

A further important observation is that if P is a full-dimensional polytope then its *irredundant description* in the form of a polyhedron is unique.

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.

They are computer packages that allow one to get the description in the form of a polyhedron on the basis of the description in the form of a (rational) polytope.

Result: standard imsets are vertices of a polytope

Theorem (main result)

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 $2^{|N|} - |N| - 1$.

Now, recall that every score equivalent and decomposable criterion $\mathcal{Q}$ necessarily has the form:

$$\mathcal{Q}(G, D) = s_D^{\mathcal{Q}} - \langle t_D^{\mathcal{Q}}, u_G \rangle \qquad \text{for any } G \in \mathrm{DAGS}\,(N), D \in \mathrm{DATA}\,(N, d)\,,$$

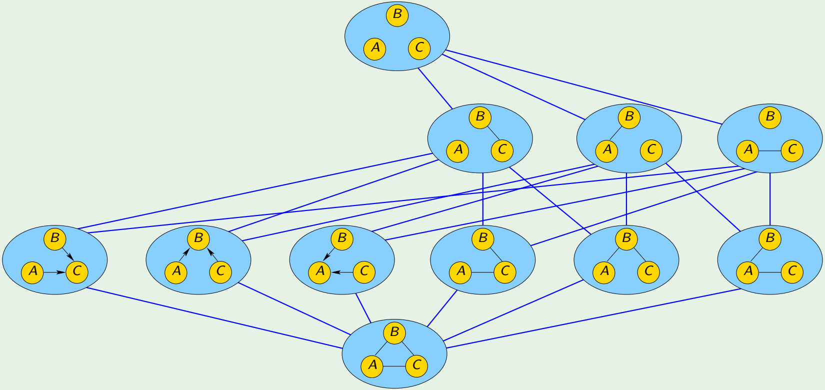
where $s_D^{\mathcal{Q}} \in \mathbb{R}$ and $t_D^{\mathcal{Q}} : \mathcal{P}(N) \to \mathbb{R}$ do not depend on G.

The consequence is as follows: the task to maximize $\mathcal{Q}$ over BN structures (= standard imsets) is equivalent to the task to maximize an affine function over the above mentioned polytope.

Example: the case of three variables (essential graphs)

In this case, one has 11 BN structures and they break into 5 types (= permutation equivalence classes).

Example



Example: the case of three variables (standard imsets)

The standard imsets can also be classified by the number of edges in the corresponding essential graph.

- 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.

The dimension of the polytope generated by these 11 imsets is 4.

Example: the case of three variables (polyhedron)

To get its irredundant description in the form of a polyhedron we embedded it in a 4-dimensional space. Then we used the computer package Convex (Franz 2006) to get all 13 polyhedron-defining inequalities, which break into 7 types. They can be classified as follows:

- Five inequalities hold with equality for the zero imset. They break into 3 types: $0 \leq 2 \cdot u(abc) + u(ab) + u(ac) + u(bc)$, $0 \leq u(abc) + u(ab)$ and $0 \leq u(abc)$.
- Eight inequalities achieve equality for the imset corresponding to the empty graph. They break into 4 types, namely $u(abc) \leq 1$, $u(abc) + u(ab) \leq 1$, $u(abc) + u(ab) + u(ac) \leq 1$ and $u(abc) + u(ab) + u(ac) + u(bc) \leq 1$.

In the case of 4 variables one has 185 BN structures breaking into 20 types. The dimension of the polytope is 11. The number of corresponding polyhedron-defining inequalities is 154.

Geometric neighborhood

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 its (geometric) edges. This motivated the definition:

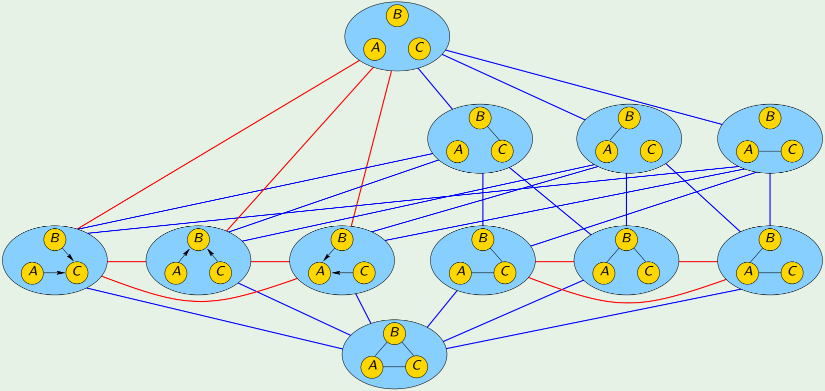
Definition (geometric neighborhood)

We say that two standard imsets u, v 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 the set of standard imsets), which means $\mathsf{P} \setminus E$ is convex.

We characterized the geometric neighborhood in the case of three and four variables and compared it with the inclusion neighborhood.

Geometric neighborhood: search space for three variables

Example



Geometric neighborhood: GES Failure

What are the consequences?

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 BN structure given empty graph G^0 and its global maximum for (any of) the graph(s) $\widehat{G}$ of the type $a \longrightarrow b \longleftarrow c$.

Put $\mathsf{X}_i = \{0, 1\}$ for $i \in N$ and
$D: \quad x^1 = (0, 0, 0), \;\; x^2 = (0, 1, 1), \;\; x^3 = (1, 0, 1), \;\; x^4 = (1, 1, 0).$

Actually, this is the 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. In particular, the GES algorithm (Chickering 2002) should (asymptotically) learn the empty graph G^0, while it is clear that (any of the graphs) $\widehat{G}$ gives a more appropriate BN structure approximation for $\hat{P}$.

Conlusions

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 valid. The point of the preceding example 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.

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.

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.

The conjecture that the inclusion neighborhood is always contained in the geometric one has recently been confirmed by Raymond Hemmecke.

Some relevant literature

- 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/`
- S.L. Lauritzen (1996). *Graphical Models.* Oxford: Clarendon Press.
- A. Schrijver (1986). *Theory of Linear and Integer Programming.* Chichester: John Wiley.
- M. Studený (2005). *Probabilistic Conditional Independence Structures.* London: Springer-Verlag.