Learning Bayesian networks viewed as an optimization problem

Milan Studený

Institute of Information Theory and Automation of the AS CR
Prague

COSA Workshop

Combinatorial Optimization, Statistics, and Applications

Munich, Germany, March 15, 2011, 10:45

the presentation is based on joint work with

David Haws, Raymond Hemmecke, Silvia Lindner and Jiří Vomlel

Summary of the talk

1. Motivation: learning Bayesian network structure
2. Basic concepts
3. Original research goals
   - Edges of the polytope
   - Polyhedral characterization of the polytope
   - Lattice points in the polytope
4. New research topics
   - Characteristic imset
   - Plain zero-one encoding of a directed graph
5. Recent findings
6. Conclusions

Motivation: learning Bayesian network structure

Bayesian networks are special graphical models widely used both in artificial intelligence and statistics. They are described by *acyclic directed graphs*, whose nodes correspond to variables.

The motivation for our research has been learning Bayesian network (BN) structure from data by a score-and-search method.

By a *quality criterion*, also called a *score*, is meant a real function $\mathcal{Q}$ of the BN structure (= of a graph G, typically) and of the observed database D.

The value $\mathcal{Q}(G, D)$ should say how much the BN structure given by G is suitable to explain the occurrence of the database D.

The aim is to maximize $G \mapsto \mathcal{Q}(G, D)$ given the observed database D.

An example of such a criterion is Schwarz's *BIC criterion*.

Motivation: algebraic approach to learning

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

The basic idea of the proposed algebraic approach was to represent the BN structure given by an acyclic directed graph G by a certain vector u_G having integers as components, called the *standard imset* (for G).

The point is that then every reasonable quality criterion $\mathcal{Q}$ for learning BN structure appears to be an affine function of the standard imset.

More specifically, one has then

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

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

Motivation: geometric view and optimization task

M. Studený, J. Vomlel and R. Hemmecke (2010). A geometric view on learning Bayesian network structures. *International Journal of Approximate Reasoning* **51**:578-586.

The main result of this paper was that the set of standard imsets over a fixed set of variables N is the *set of vertices (= extreme points) of a certain polytope* P.

In particular, the task to maximize $\mathcal{Q}$ over BN structures (= acyclic directed graphs) is equivalent to a **linear optimization problem**, namely to maximize an affine function over the above-mentioned polytope P.

This problem has been treated thoroughly within the *linear programming* community. Nevertheless, to apply efficient methods of combinatorial optimization in this area one needs to solve some open mathematical problems (of geometric nature concerning the polytope).

Overview of our research goals

M. Studený and J. Vomlel (2011). On open questions in the geometric approach to structural learning Bayesian nets. To appear in *International Journal of Approximate Reasoning*, a special issue devoted to WUPES 09.

Specifically, we are interested in:

- describing the geometric edges of P,
- polyhedral characterization of the polytope P,
- finding all lattice points within the polytope P.

Later, we extended our interests to:
(in cooperation with R. Hemmecke, S. Lindner and D. Haws)

- alternative BN structure representatives,
- complexity tasks and application to learning restricted Bayesian network structures.

Basic concepts: Bayesian network structure

N	a non-empty finite set of *variables*
X_i, $\lvert\mathrm{X}_i\rvert \geq 2$	the individual sample spaces (for $i \in N$)
$\mathrm{DAGS}\,(N)$	collection of all acyclic directed graphs over N

The (discrete) *Bayesian network* (BN) is a pair (G, P), where $G \in \mathrm{DAGS}\,(N)$ and P is a probability distribution on the joint sample space $\mathrm{X}_N \equiv \prod_{i \in N} \mathrm{X}_i$ which (recursively) factorizes according to G.

Given $G \in \mathrm{DAGS}\,(N)$, (the statistical model of) a *BN structure* is the class of all distributions P on X_N that factorize according to G.

Since two different graphs over N may describe the same BN structure, one is interested in describing the BN structure by a unique representative. A classic such graphical representative is so-called *essential graph*.

Basic concepts: learning by a score-and-search method

Data are assumed to have the form of a complete database:

$x^1, \ldots, x^d$	a sequence of elements of X_N of the length $d \geq 1$ called a *database of the length d* or a *sample of the size d*
$\mathrm{DATA}\,(N, d)$	the set of all databases over N of the length d (provided the individual sample spaces X_i for $i \in N$ are fixed)

Definition (quality criterion)

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

The value $\mathcal{Q}(G, D)$ should somehow evaluate how the statistical model given by G fits the database D.

Thus, the aim is to maximize the function $G \mapsto \mathcal{Q}(G, D)$ given the observed database $D \in \mathrm{DATA}\,(N, d)$.

Basic concepts: imsets

Definition (imset)

An *imset* u over N is an integer-valued function on $\mathcal{P}(N) \equiv \{A;\ A \subseteq N\}$, the power set of N.

It can be viewed as a vector whose components are integers, indexed by subsets of N. [= a lattice point in the Euclidean space $\mathbb{R}^{\mathcal{P}(N)}$]

A trivial example of an imset is the *zero imset*, denoted by 0.
Given $A \subseteq N$, the symbol δ_A will denote this *basic imset*:

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

Since $\{\delta_A;\ A \subseteq N\}$ is a linear basis of $\mathbb{R}^{\mathcal{P}(N)}$, any imset can be expressed as a linear combination of these basic imsets (with integers as coefficients).

Basic concepts: standard imset

Definition (standard imset)

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

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

where $pa_G(i) = \{ j \in N;\ j \to i \ \text{ in } G \}$ denotes the set of *parents* of i in G.

Note that the terms in the above formula can both sum up and cancel each other. Of course, it is a vector of an exponential length in $|N|$.

However, it follows from the definition that u_G has at most $2 \cdot |N|$ non-zero values. In particular, the memory demands for representing standard imsets are polynomial in $|N|$.

Basic concepts: algebraic approach to learning

Lemma (Studený 2005)

Given $G, H \in \mathrm{DAGS}\,(N)$, one has $u_G = u_H$ iff G and H describe the same BN structure.

Thus, the standard imset is a unique representative of the BN structure.

There are two important technical requirements on quality criteria introduced by researchers in computer science: they should be *score equivalent* and *decomposable*.

Theorem (Studený 2005)

Every score equivalent and decomposable criterion $\mathcal{Q}$ has the form

$$\mathcal{Q}(G, D) = s_D^{\mathcal{Q}} - \langle t_D^{\mathcal{Q}}, u_G \rangle \quad \textit{for } G \in \mathrm{DAGS}\,(N),\ D \in \mathrm{DATA}\,(N, d), d \geq 1$$

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

Basic concepts: geometric view

Definition (standard imset polytope)

Having fixed the set of variables N, let us put:

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

The above polytope P will be called the *standard imset polytope*.

Theorem (Studený, Vomlel, Hemmecke 2010)

S *is the set of vertices of the integral polytope* P*.*

Example Distinguished vertices of P are:

- the zero imset 0 (= the standard imset for the full graph),
- the imset $u^{\emptyset} \equiv \delta_N - \sum_{i \in N} \delta_{\{i\}} + (|N| - 1) \cdot \delta_{\emptyset}$
  (= the standard imset for the empty graph).

In case $|N| = 3$, P is the intersection of two cones, with origins in 0 and $u^{\emptyset}$.

Edges of the polytope: geometric neighborhood

One of possible interpretations of the simplex method is that it is a kind of search method, in which one moves between vertices of a polytope along its edges until an optimal vertex is reached.

Analogous idea is at the core of current computer science optimization techniques, like the *GES algorithm*. Specifically, a particular concept of *inclusion neighborhood* has been introduced for (acyclic directed) graphs and one moves between neighboring graphs (in this sense).

[Inclusion neighbors have simple graphical interpretation: edge removal/adding.]

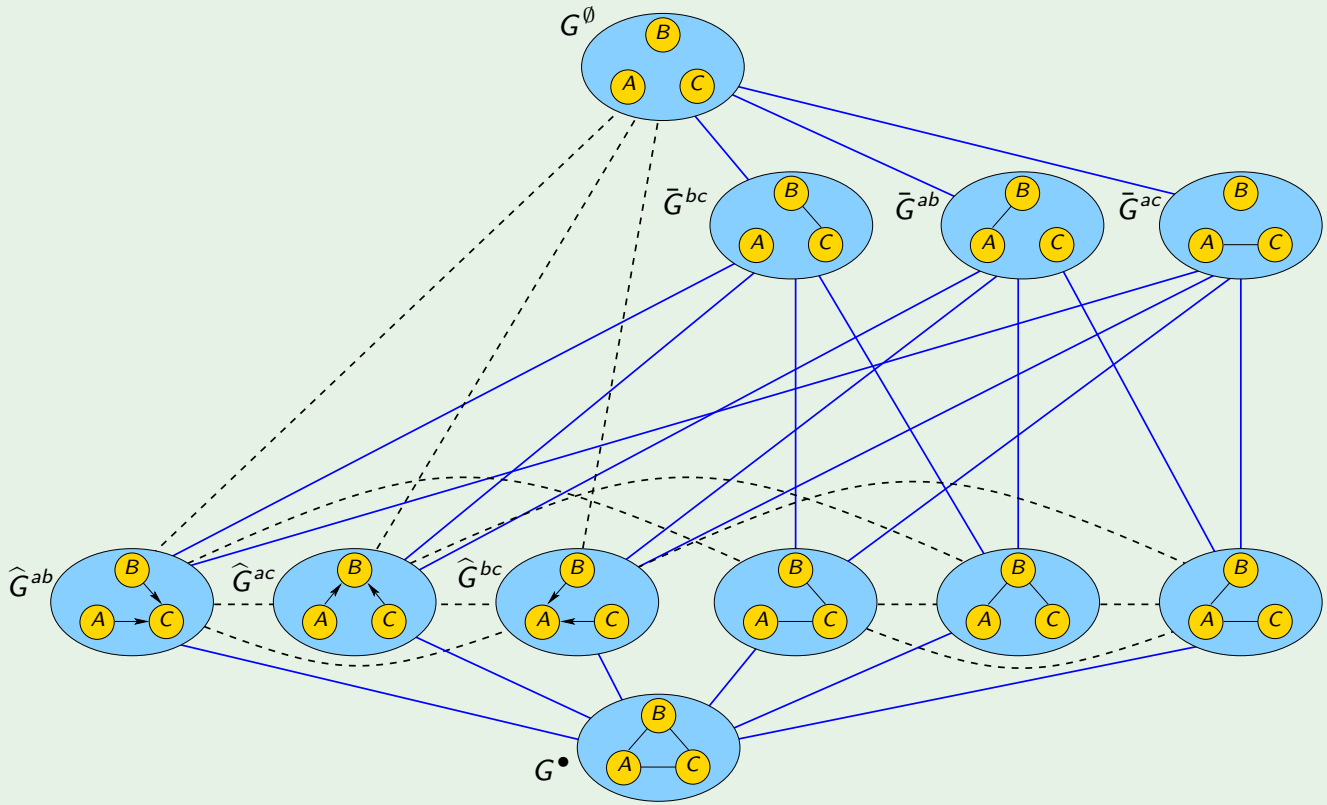
Definition (geometric neighbors)

Distinct standard imsets u and v are called *geometric neighbors* if the segment $[u, v]$ is an edge of the standard imset polytope P.

This defines the concept of *geometric neighborhood* for BN structures.

Comparison with the inclusion neighborhood for $|N| = 3$

Example (geometric neigborhood in the case of three variables)



Observation: for $|N| = 3$, the inclusion neighborhood is strictly contained in the geometric one. This has a simple but notable consequence from the statistical point of view: the *GES algorithm may fail to find the global maximum*.

The starting analysis of the geometric neighborhood

We have proved in 2010 that the inclusion neighborhood is always contained in the geometric one (for any $|N|$).

We have also succeeded to compute the geometric neighborhood for $|N| = 3, 4, 5$. Our computations suggest that, for most standard imsets, there are many more geometric neighbors than the inclusion neighbors.

An output of our computations was also an electronic catalogue of types of geometric neighbors in the case $|N| = 4$:

`http://staff.utia.cas.cz/vomlel/imset/catalogue-diff-imsets-4v.html`

The intention has been to find out whether it is possible to interpret geometric neighbors in graphical terms.

Recent findings on the geometric neighborhood

Already in 2010, we observed that the geometric neighbors of the full graph ($\sim$ the zero imset 0) coincide with its inclusion neighbors.
[These are graphs with one missing edge.]

Recently, we have developed some methods to show/disprove whether two given (standard) imsets are geometric neighbors.

We have also recognized a further general type of geometric neighbors, which has graphical interpretation.
[Our preliminary naming of it is "immorality rotation".]

Finally, we succeeded to characterize the geometric neighbors of the empty graph ($\sim$ the imset $u^{\emptyset}$). These appear to be graphs with just one non-initial node: $G \in \mathrm{DAGS}(N)$ such that $\exists!\, i \in N$ with $pa_G(i) \neq \emptyset$.

These recent findings have been derived using an alternative algebraic representative of a BN structure, called the *characteristic imset*.

Polyhedral characterization of the polytope

In classic formulation of a linear optimization problem, the domain is specified in the form of a polyhedron, that is, via finitely many linear inequalities.

Therefore, a lot of our effort was devoted to the attempts to find the *outer description* of the (standard imset) polytope P ($\equiv$ characterization of all facets of P).

[Even implicit facet description can appear to be useful.]

In 2010, our state of knowledge about this was as follows:
(a result of computations)

$\lvert N\rvert$	3	4	5
vertices	11	185	8782
facets	13	154	??

Classification of our linear constraints

Nevertheless, we made a detailed analysis in the case $|N| = 4$. The result was a classification of all tight linear inequalities we had been aware of:

- trivial *equality restrictions*,
- so-called *non-specific* inequality constraints,
- so-called *specific* inequality constraints.

We have also shown that these are *necessary linear constraints* on the elements of the polytope P (for any $|N|$).

Equality restrictions have the form

$$\sum_{S \subseteq N} u(S) = 0 \quad \text{and} \quad \sum_{S \subseteq N; i \in S} u(S) = 0 \quad \text{for any } i \in N.$$

Their number $|N| + 1$ is final.

Non-specific linear constraints

Definition (non-specific inequality)

By a *non-specific inequality* (for P) we mean the facet-defining inequality for a facet of P containing the zero imset 0.

The point is that the cone generated by geometric neighbors of 0 was studied earlier and we knew its facets correspond to the extreme supermodular functions.

In particular, each non-specific inequality has the following form:

$$\sum_{T \subseteq N} m(T) \cdot u(T) \geq 0 \qquad \text{where } m : \mathcal{P}(N) \to \mathbb{Z}^+ \text{ satisfies}$$

$m(C \cup D) + m(C \cap D) \geq m(C) + m(D)$ for $C, D \subseteq N$ and $m(S) = 0$ for $|S| \leq 1$.

Table: Numbers of non-specific inequality constraints for $|N| \leq 5$:

$\lvert N \rvert$	2	3	4	5
numbers	1	5	37	117978

Specific linear constraints

Remaining *specific constraints* are in correspondence with non-empty classes of non-empty subsets of N closed under supersets, that is,

$$\text{classes } \emptyset \neq \mathcal{A} \subseteq \{T \subseteq N;\ |T| \geq 1\} \quad \text{with} \quad S \in \mathcal{A}, S \subseteq T \ \Rightarrow \ T \in \mathcal{A}.$$

Given a class of sets $\mathcal{A}$ of this kind the corresponding *non-specific linear inequality* has quite simple form:

$$\sum_{T \in \mathcal{A}} u(T) \leq 1 \ .$$

Nevertheless, not all of these inequalities are facet-defining for P.

Table: Numbers of specific inequality constraints:

$\lvert N\rvert$	2	3	4	5
before	4	18	166	7579

$\lvert N\rvert$	2	3	4
after reduction	1	8	117

Conjecture about the outer description

The result of our analysis in the case $|N| = 4$ was also an observation that, in this case, P is not already the intersection of two cones (with origins in 0 and $u^{\emptyset}$); there are 4 more facets of P that do not contain either 0 or $u^{\emptyset}$.

Conjecture (stronger version)

The above-mentioned linear constraints together form a necessary and sufficient condition for $u \in \mathbb{R}^{\mathcal{P}(N)}$ to belong to P.

Conjecture (weaker version)

The above constraints give together a necessary and sufficient condition for $u \in \mathbb{Z}^{\mathcal{P}(N)}$ to belong to $\mathrm{P} \cap \mathbb{Z}^{\mathcal{P}(N)}$.

Recent success: The weaker conjecture has been verified for $|N| = 5$!

Actually, finding suitable LP relaxation of the integral polytope P would make it possible to apply advanced methods of integer linear programming.

Lattice points in the polytope

Already in 2009, Raymond Hemmecke made some computations for $|N| \leq 5$ to find out whether there exists a lattice point in the interior of P. The answer was negative. This led him to a hypothesis that P is *thin* in the sense $\mathrm{P} \cap \mathbb{Z}^{\mathcal{P}(N)} = \mathrm{ext}\,(\mathrm{P}) = \mathrm{S}$.

Theorem

The only lattice points within the polytope P *are its vertices.*

The original 2009 proof was quite technical, but it has been substantially simplified in 2010. The idea is to use an elegant affine transformation!

M. Studený, R. Hemmecke and S. Lindner (2010). Characteristic imset: a simple algebraic representative of a Bayesian network structure. In *Proceedings of the 5th European Workshop PGM*, pp. 257-264.

Transformation to the characteristic imset

Definition (characteristic imset)

Assume $|N| \geq 2$. Given an acyclic directed graph G over N, let u_G be the corresponding standard imset. The *characteristic imset* for G is given by the formula

$$c_G(T) = 1 - \sum_{S, T \subseteq S \subseteq N} u_G(S) \qquad \text{for } T \subseteq N, |T| \geq 2.$$

Clearly, the characteristic imset is obtained from the standard one by an affine transformation. Moreover, this mapping is invertible.

In particular, every score equivalent and decomposable criterion is also an affine function of the characteristic imset!

Characteristic imset

However, the crucial observation about characteristic imsets is as follows:

Theorem

Asume $|N| \geq 2$. *Given an acyclic directed graph* G *over* N *one has* $c_G(A) \in \{0, 1\}$ *for any* $A \subseteq N$, $|A| \geq 2$.

The above-mentioned affine transformation maps lattice points to lattice points. Since there is no lattice point in the interior of 0-1 hypercube, there is no lattice point in the interior of the standard imset polytope P!

The characteristic imset is also much closer to the graphical description than the standard imset. There is a simple polynomial algorithm for getting the essential graph on basis of the characteristic imset.

Simple zero-one encoding of a directed graph

T. Jaakkola, D. Sontag, A. Globerson, and M. Meila (2010). Learning Bayesian network structure using LP relaxations. In *Proceedings of the 13th International Conference on AI and Statistics*, pp. 358-365.

They use a simple 0-1-vector $\boldsymbol{\eta}_G$ to encode a directed graph G over N. The vector has components indexed by pairs $(i|B)$, where $i \in N$ and $B \subseteq N \setminus \{i\}$. More specifically:

$$\eta_G(i|B) = 1 \text{ iff } B = pa_G(i), \qquad \eta_G(i|B) = 0 \text{ otherwise.}$$

The main difference: different equivalent graphs have different representatives! Their vectors are even longer than ours; have $|N| \cdot 2^{|N|-1}$ components.

They also turned the BN learning task into a linear optimization problem.

They start with a polyhedral upper approximation of their polytope and combine the LP approach with other methods, like *branch-and-bound* principle.

LP relaxation offered by Jaakkola *et.al.*

Their polyhedron J was given by the following constraints:

- simple *non-negativity constraints* $\eta(i|B) \geq 0$ for every $(i|B)$,
- *equality constraints* $\sum_{B \subseteq N \setminus \{j\}} \eta(j|B) = 1$ for any $j \in N$,
- *cluster inequalities*, which correspond to sets

$$C \subseteq N,\ |C| \geq 2 \text{ (called } \textit{clusters}\text{)}: \quad 1 \leq \sum_{i \in C} \sum_{B \subseteq N \setminus C} \eta(i|B)\,.$$

The cluster inequalities encode acyclicity restrictions to G. The inequality for C means that the induced subgraph G_C has at least one initial node.

There could be non-integral vertices of J.

An interesting observation (which is not difficult to show) is that the only lattice points in J are the codes of acyclic directed graphs over N.

Thus, their polyhedron is an LP relaxation of the convex hull of the set of codes.

Recent findings: transformation to our frameworks

We have observed that the standard imset u_G is an affine function of η_G (which is a many-to-one mapping, of course):

$$u^G(T) = \delta_N(T) - \delta_\emptyset(T) + \sum_{(i|B)} \eta_G(i|B) \cdot \{\delta_B(T) - \delta_{\{i\}\cup B}(T)\} \quad \text{for } T \subseteq N,$$

and the characteristic imset c_G is even a linear function of it:

$$c_G(T) = \sum_{(i|B)} \eta(i|B) \cdot \delta[\, i \in T \,\&\, T \setminus \{i\} \subseteq B\,] \qquad \text{for } T \subseteq N.$$

Therefore, we have three ways of algebraic representation of Bayes nets:

$$\eta_G \quad \longrightarrow \quad u_G \quad \longleftrightarrow \quad c_G\,.$$

Our aim was to transform Jaakkola's linear constraints to our framework(s) and to compare them with our constraints.

Recent findings: inequalities translation

First finding (already in November 2010) was that the *cluster inequalities can be easily transformed* to the framework of standard imsets. They appear to correspond to some non-specific inequalities:

$$\sum_{T \subseteq N} m_C(T) \cdot u(T) \geq 0 \quad \text{where } m_C(T) = \max\{0, |C \cap T| - 1\} \text{ for } T \subseteq N.$$

It was quite a big technical problem to transform *Jaakkola's non-negativity and equality constraints* (owing to many-to-one mapping). Nevertheless, very recently we succeeded to confirm our former conjecture that they *lead exactly to the specific inequalities*.

[Paradox: we reduce the dimension but raise the number of inequalities.]

A consequence of this partial result is that the polyhedron given by (solely) specific inequalities is integral (and bounded).

Conclusions

Thus, our polyhedral approximation is tighter than Jaakkola's one.

We think can possibly utilize their observations:

To confirm the weaker version of our conjecture (= the only lattice points in our polyhedral approximation are standard imsets) it remains to show that the corresponding linear mapping has the following property: *If a integral vector has non-negative pre-image, then it even has an integral non-negative pre-image*.

This appears to be related to the *unimodularity* of the respective matrix. We believe our computations confirmed the unimodularity for $|N| = 3, 4, 5, 6$.

If the weaker version of our conjecture is confirmed, then the transformation to the 0-1 framework of characteristic imsets can perhaps allow us to use advanced methods of integer programming, like *cutting plane* methods and *lift-and-project* approach.

Moreover, there is still a chance that the stronger version of the conjecture is true!