Polyhedral approach to statistical learning graphical models

Milan Studený

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Prague

The Second CREST-SBM International Conference
Harmony of Gröbner Bases and the Modern Industrial Society
Osaka, Japan, June 30, 2010, 10:40-11:20

based on joint work with Raymond Hemmecke, Jiří Vomlel and Silvia Lindner

Summary of the talk

1. Motivation: learning Bayesian network structure
2. Basic concepts
3. Towards the outer description of the polytope
4. Edges of the polytope: geometric neighborhood
5. Lattice points in the polytope
6. Characteristic imset and restricted learning
7. Conclusions

Motivation: learning Bayesian networks

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

The motivation for the research reported here is 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 of the BN structure (= of a graph G, usually) and of the database D.

The value $\mathcal{Q}(G, D)$ should say how much the BN structure given by G is good 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 an algebraic approach is 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 criterion $\mathcal{Q}$ for learning BN structure is an affine 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 $\mathcal{Q}$).

Motivation: geometric view

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 the above paper is 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 the task to maximize an affine function over the above-mentioned polytope P.

This problem has been treated thoroughly within the *linear programming* community.

Motivation: some research topics

The intention to apply linear programming methods in the area of learning BN structure motivated a series of (open mathematical) questions concerning the above-mentioned polytope P.

Specifically, we were interested in:

- the outer (= polyhedral) description of the polytope P,
- characterizing the geometric edges of P,
- finding all lattice points within the polytope.

Moreover, an elegant solution to the last question, based on the application of a suitable transformation, opened further research topics:

- an alternative BN structure representative: *characteristic imset*,
- (possible) application to learning decomposable models and forests.

Basic concepts: BN structure

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

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

Given $G \in \mathsf{DAGS}\,(N)$, (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 the *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 $\mathbb{R}^{\mathcal{P}(N)}$]

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 combination of these basic imsets.

Basic concepts: standard imset

Definition (standard imset)

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

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

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.

Nevertheless, it follows from the definition that u_G has at most $2 \cdot |N|$ non-zero values. Thus, 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 emphasized 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 the $s_D^{\mathcal{Q}} \in \mathbb{R}$ and the vector $t_D^{\mathcal{Q}} \in \mathbb{R}^{\mathcal{P}(N)}$ do not depend on G.

Basic concepts: geometric view

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

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

Theorem (Studený, Vomlel, Hemmecke 2010)

S *is the set of vertices of a rational polytope* $\mathsf{P} \subseteq \mathbb{R}^{\mathcal{P}(N)}$.

This polytope P will be called the *standard imset polytope*.
The above results imply that the task to maximize $\mathcal{Q}$ over $G \in \mathsf{DAGS}\,(N)$ is equivalent to the task to minimize the linear function $u \mapsto \langle t_D^{\mathcal{Q}}, u\rangle$ over P.

Attempts at deeper analysis of the polytope were the topic of a paper:

M. Studený and J. Vomlel (2010). On open questions in the geometric approach to structural learning Bayesian nets. Accepted in *International Journal of Approximate Reasoning*, special issue WUPES 09.

Towards the outer description of the polytope

In order to apply the classic version of the simplex method, one needs an explicit *outer description* of the polytope via finitely many linear inequalities (= in the form of a polyhedron).

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

Linear constraints:

On the basis of a detailed analysis case $|N| = 4$, we have established a class of *necessary linear constraints* on the elements of the polytope P:

- trivial *equality constraints*, denoted by (A),
- *non-specific inequality* constraints, denoted by (B),
- *specific* inequality constraints, denoted by (C).

Equality constraints are fully characterized. Their number is $|N| + 1$.

Inequality constraints

The observation that the polytope P is a part of a formerly studied cone led to a set of *non-specific constraints*. These correspond to the extreme (standardized) supermodular functions.

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

$\|N\|$	2	3	4	5
numbers	1	5	37	117978

Specific constraints are in correspondence with certain classes of subsets of N closed under supersets, and, thus, with log-linear models over N.

Table: Numbers of specific inequality constraints:

$\|N\|$	2	3	4	5
before	4	18	166	7576

$\|N\|$	2	3	4
after reduction	1	8	114

Conjecture about the outer description

The constraints (A)-(C) have several consequences, perhaps not evident at first sight, for example:

$$u(S) \geq -1 \qquad \text{for any } S \subseteq N.$$

We have shown that (A)-(C) are necessary constraints on points in P, but we have also some reasons to conjecture this:

Conjecture

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

The conjecture has been verified for $|N| \leq 4$; current task is to confirm or disprove it for $|N| = 5$.

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 the polytope along its edges until an optimal vertex is reached.

Definition (geometric neighbors, differential imset)

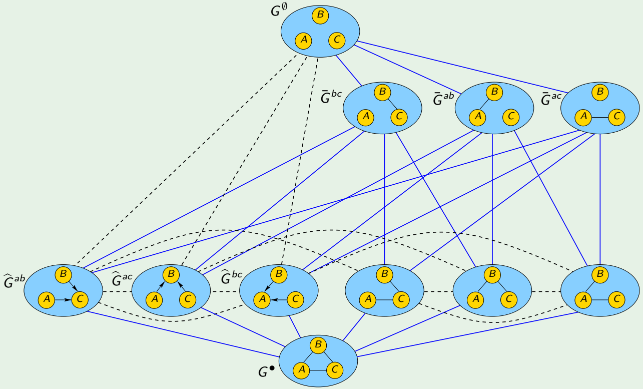
Distinct standard imsets u and v are called *geometric neighbors* if the segment $[u, v]$ is an edge of the polytope P. Given standard imsets u, v, their difference $w = u - v$ is called the *differential imset*.

In particular, this defines the concept of *geometric neighborhood* for BN structures.

Comparison with the inclusion neighborhood

We have succeeded to compute the geometric neighborhood for $|N| = 3, 4, 5$. As a by-product we compared it for $|N| = 3$ with the *inclusion neighborhood*, which is at the core of current computer science search techniques.

Example (geometric neigborhood in the case of three variables)



Catalogue of the geometric neighborhood for $|N| = 4$.

Our computations suggest that, for most standard imsets, there are many more geometric neighbors than the inclusion neighbors.

This observation has a simple but notable consequence from the statistical point of view: the *GES algorithm may fail to find the global maximum* of a quality criterion.

Actually, we think that this is an inevitable defect of the inclusion neighborhood, which may occur whenever a special statistical *data faithfulness assumption* is not guaranteed.

The result of our analysis in the case $|N| = 4$ was an electronic catalogue of differential imsets for geometric neighbors:

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

Lattice points in the polytope

Raymond Hemmecke made earlier some computations for $|N| \leq 5$ to find out whether there exists an imset in the interior of the polytope P and the result was negative.

This led him to a hypothesis that every lattice point within the standard imset polytope is already the standard imset.

Theorem

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

The original proof was quite technical, but it has recently been substantially simplified. The idea is to use an elegant linear (affine) transformation.

Transformation to the characteristic imset

- M. Studený, R. Hemmecke and S. Lindner (2010). Characteristic imset: a simple algebraic representative of a Bayesian network structure. Submitted to *Proceedings of PGM 2010*.
- R. Hemmecke, S. Lindner, M. Studený, and J. Vomlel. Characteristic imsets for learning Bayesian network structures. In preparation.

Definition

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

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

Characteristic imset

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 an affine function of the characteristic imset as well.

However, crucial observation is this:

Theorem

Asume $|N| \geq 2$. Given an acyclic directed graph G over N one has $c_G(A) \in \{0, 1\}$ for any $A \subseteq N$, $|A| \geq 2$.
Moreover, one has $c_G(A) = 1$ iff there exists $i \in A$ with $A \setminus \{i\} \subseteq pa_G(i)$.

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.

Characteristic imset and the essential graph

The characteristic imset is much closer to the graphical description than the standard imset:

Corollary

Let G be an acyclic directed graph over N and a, b (and c) are distinct nodes. Then

(i) *a and b are adjacent in G iff $c_G(\{a, b\}) = 1$.*

(ii) *$a \rightarrow c \leftarrow b$ is an induced subgraph of G iff $c_G(\{a, b, c\}) = 1$ and $c_G(\{a, b\}) = 0$.*

There is a direct formula for the characteristic imset on basis of the essential graph and a simple polynomial algorithm for getting the essential graph on basis of the characteristic imset.

The idea of restricted learning BN structures

If one considers a subclass of the class of BN structures, then this corresponds to a subset of the set of standard imsets. Geometrically, this subset specifies a sub-polytope of P.

There are some classes of models which fall withing this frame:

- the class of *decomposable models* (described by chordal graphs),
- (undirected) graphical models that correspond to *trees* or *forests*.

The characteristic imset is quite simple in the case of decomposable models. The situation is particularly transparent for undirected forests:

Corollary

Let H be an undirected forest. Then the corresponding characteristic imset c_H vanishes for sets of cardinality 3 and, more.
For distinct $a, b \in N$ one has $c_H(\{a, b\}) = 1$ iff a and b are adjacent in H.

Restricted learning: forests

The respective sub-polytope (spanned by characteristic imsets for forests) has been formerly studied within matroid theory.

A. Schrijver (2003). *Combinatorial Optimization - Polyhedra and Efficiency, volume B*, Springer Verlag, Berlin.

The point is that both edges and the outer description of that *independent set polytope* are known, as well as highly efficient method for finding maximum weight forests utilizing the *greedy algorithm*.

This is, actually, gives geometric interpretation to an old classic method:

C.K. Chow and C.N. Liu (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory* **14**: 462-467.

Indeed, the above classic method can be interpreted as a method for maximizing *maximized log-likelihood* (MLL) criterion over *trees*.

Conclusions

We would like either to confirm or to disprove for $|N| = 5$ the conjecture about the outer description of P. We plan to do so using a computer program (task for Raymond Hemmecke).

The catalogue of differential imsets for geometric neighbors is meant as a step towards a *deeper analysis of the geometric neighborhood*.

For example, we would like to find out whether there is a graphical interpretation of geometric neighborhood.

The observations made by Silvia Lindner indicate the chance to extend efficient learning procedures (based on the greedy algorithm) to *forests* and other quality criteria, for example *Bayesian criteria*.