On Geometric Approach to Structural Learning Bayesian Nets

Milan Studený

Department of Decision Making Theory
Institute of Information Theory and Automation (ÚTIA)
Academy of Sciences of the Czech Republic (AV ČR)

based on joint work with Jiří Vomlel and Raymond Hemmecke

Prague, seminar of the research center DAR,
group "Decision Making and Control under Uncertainty"
February 2, 2010, 14:00

Summary of the talk

1. The idea of structural learning Bayesian nets
2. Algebraic approach
3. Geometric view
4. A short trip to convex geometry
5. Recent results and conjectures
6. Conclusions

Motivation: 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.

Motivation: 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 structural component of a BN is an *acyclic directed graph*, whose nodes correspond to variables into consideration (= symptoms, factors). The probabilistic component is a (discrete) probability distribution that *recursively factorizes* according to the graph.

Motivation: 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 structural component of a BN is an *acyclic directed graph*, whose nodes correspond to variables into consideration (= symptoms, factors). The probabilistic component is a (discrete) probability distribution that *recursively factorizes* according to the graph.

If the graph is sparse, then the distribution can effectively be stored in the memory of a computer and the *method of local computation* allows one to compute efficiently conditional probabilities.

Motivation: 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 structural component of a BN is an *acyclic directed graph*, whose nodes correspond to variables into consideration (= symptoms, factors). The probabilistic component is a (discrete) probability distribution that *recursively factorizes* according to the graph.

If the graph is sparse, then the distribution can effectively be stored in the memory of a computer and the *method of local computation* allows one to compute efficiently conditional probabilities.

In probabilistic expert systems, the structure (= the graph) was typically provided by an expert, but later, people became interested in the question of how to learn structure from statistical data.

Motivation: learning structure of a BN

Note that it may happen that two different (acyclic directed) graphs describe the same structure of a BN.

Motivation: learning structure of a BN

Note that it may happen that two different (acyclic directed) graphs describe the same structure of a BN.

From the point of view of statistics, the *BN structure* is modeled as a statistical model (= a parameterized class of probability distributions).

Motivation: learning structure of a BN

Note that it may happen that two different (acyclic directed) graphs describe the same structure of a BN.

From the point of view of statistics, the *BN structure* is modeled as a statistical model (= a parameterized class of probability distributions).

Formally, it is a class of probability distributions (on a fixed discrete joint sample space) that factorize according to the graph. These distributions are called *Markovian distributions*.

Motivation: learning structure of a BN

Note that it may happen that two different (acyclic directed) graphs describe the same structure of a BN.

From the point of view of statistics, the *BN structure* is modeled as a statistical model (= a parameterized class of probability distributions).

Formally, it is a class of probability distributions (on a fixed discrete joint sample space) that factorize according to the graph. These distributions are called *Markovian distributions*.

An equivalent definition of a Markovian distribution is that it is a distribution satisfying some *conditional independence* (CI) constraints, determined by the respective graphical separation criterion.

Motivation: learning structure of a BN

Note that it may happen that two different (acyclic directed) graphs describe the same structure of a BN.

From the point of view of statistics, the *BN structure* is modeled as a statistical model (= a parameterized class of probability distributions).

Formally, it is a class of probability distributions (on a fixed discrete joint sample space) that factorize according to the graph. These distributions are called *Markovian distributions*.

An equivalent definition of a Markovian distribution is that it is a distribution satisfying some *conditional independence* (CI) constraints, determined by the respective graphical separation criterion.

Traditional unique graphical representative of a BN structure is a special chain graph, called the *essential graph*. It (somehow) encodes shared features of acyclic directed graphs defining the structure.

Introduction: methods for structural learning

The methods for structural learning BNs can be divided into two groups:

Introduction: methods for structural learning

The methods for structural learning BNs can be divided into two groups:

- algorithms based on *significance tests* (= statistical CI tests),

Introduction: methods for structural learning

The methods for structural learning BNs can be divided into two groups:

- algorithms based on *significance tests* (= statistical CI tests),
- procedures based on the maximization of a suitable *quality criterion*.

Introduction: methods for structural learning

The methods for structural learning BNs can be divided into two groups:

- algorithms based on *significance tests* (= statistical CI tests),
- procedures based on the maximization of a suitable *quality criterion*.

The motivation for this talk is *learning Bayesian network structure* from data by the method of maximization of a quality criterion.

Introduction: methods for structural learning

The methods for structural learning BNs can be divided into two groups:

- algorithms based on *significance tests* (= statistical CI tests),
- procedures based on the maximization of a suitable *quality criterion*.

The motivation for this talk is *learning Bayesian network structure* from data by the method of maximization of a quality criterion.

Definition

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. That is, $\mathcal{Q} : \mathrm{DAGS}\,(N) \times \mathrm{DATA}\,(N,d) \mapsto \mathbb{R}$.

Introduction: methods for structural learning

The methods for structural learning BNs can be divided into two groups:

- algorithms based on *significance tests* (= statistical CI tests),
- procedures based on the maximization of a suitable *quality criterion*.

The motivation for this talk is *learning Bayesian network structure* from data by the method of maximization of a quality criterion.

Definition

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. That is, $\mathcal{Q} : \mathrm{DAGS}\,(N) \times \mathrm{DATA}\,(N,d) \mapsto \mathbb{R}$.

A quality criterion should be *statistically consistent*, but there are other reasonable requirements.

Introduction: examples of quality criteria

Some examples of quality criteria used in practice:

Introduction: examples of quality criteria

Some examples of quality criteria used in practice:

- MLL (maximized log-likelihood criterion),

Introduction: examples of quality criteria

Some examples of quality criteria used in practice:

- MLL (maximized log-likelihood criterion),
- AIC (Akaike's information criterion),
- BIC (Jeffrey-Schwarz criterion),

Introduction: examples of quality criteria

Some examples of quality criteria used in practice:

- MLL (maximized log-likelihood criterion),
- AIC (Akaike's information criterion),
- BIC (Jeffrey-Schwarz criterion),
- some Bayesian criteria (this depends on the choice of 'priors').

Introduction: examples of quality criteria

Some examples of quality criteria used in practice:

- MLL (maximized log-likelihood criterion),
- AIC (Akaike's information criterion),
- BIC (Jeffrey-Schwarz criterion),
- some Bayesian criteria (this depends on the choice of 'priors').

There are two other 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).

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.

Algebraic approach: idea

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.

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

Algebraic approach: idea

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.

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

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).

Algebraic approach: idea

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.

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

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 the standard imset is a unique (algebraic) representative of the BN stucture and that every score equivalent and decomposable criterion $\mathcal{Q}$ is an affine function (= linear function plus a constant) of the standard imset.

Algebraic approach: some details

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

Algebraic approach: some details

- 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 $u : \mathcal{P}(N) \mapsto \mathbb{Z}$.

Algebraic approach: some details

- 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 it as a vector whose components are integers and are indexed by subsets of N.

Algebraic approach: some details

- 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 it as a vector whose components are integers and are indexed by subsets of N.

Actually, any real function $m : \mathcal{P}(N) \to \mathbb{R}$ can be interpreted as a (real) vector in the same way.

Algebraic approach: some details

- 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 it as a vector whose components are integers and are indexed by subsets of N.

Actually, any real function $m : \mathcal{P}(N) \to \mathbb{R}$ can 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: standard imset

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

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

Algebraic approach: standard 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 (standard imset)

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

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

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

Algebraic approach: standard 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 (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 $pa_G(a) \equiv \{ b \in N;\ b \to a \text{ in } G \}$ denotes the set of *parents* of the node a.

Because every standard imset 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 $|N|$.

Standard imsets: main observations

Standard imset is another uniquely determined representative of a BN structure. Actually, there exists a polynomial algorithm which translates the standard imset into the essential graph and conversely:

M. Studený and J. Vomlel (2009). A reconstruction algorithm for the essential graph. *International Journal of Approaximate Reasoning* **50**: 385-413.

Standard imsets: main observations

Standard imset is another uniquely determined representative of a BN structure. Actually, there exists a polynomial algorithm which translates the standard imset into the essential graph and conversely:

M. Studený and J. Vomlel (2009). A reconstruction algorithm for the essential graph. *International Journal of Approaximate Reasoning* **50**: 385-413.

Theorem (Studený 2005)

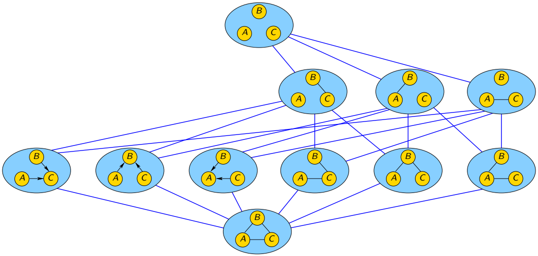
Let $\mathcal{Q}$ be a score equivalent and decomposable criterion, G any acyclic directed graph over N and D a database (over N). Then one has

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

$t_D^{\mathcal{Q}}$ is a real vector in $\mathbb{R}^{\mathcal{P}(N)}$ and $\langle *, * \rangle$ denotes the scalar product. The vector $t_D^{\mathcal{Q}}$ is named the *data vector* (relative to $\mathcal{Q}$).

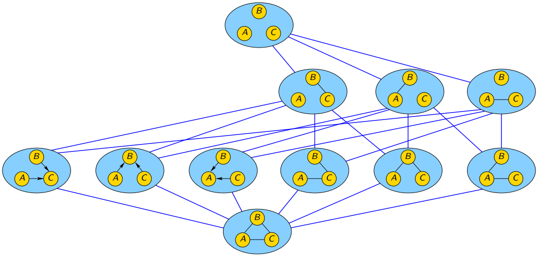
Example: the case of three variables

In the case of 3 variables one has 11 BN structures breaking into 5 types.



Example: the case of three variables

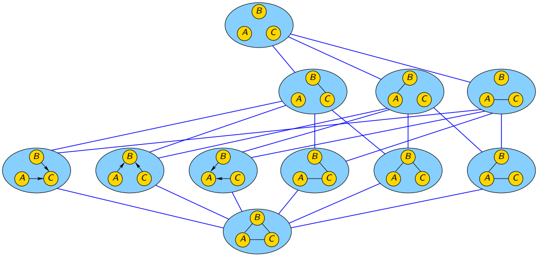
In the case of 3 variables one has 11 BN structures breaking into 5 types.



- The zero imset corresponds to the complete (undirected) essential graph.

Example: the case of three variables

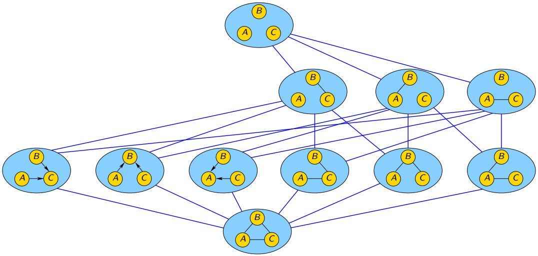
In the case of 3 variables one has 11 BN structures breaking into 5 types.



- The zero imset corresponds to the complete (undirected) essential graph.
- Six "elementary" imsets break into two types, namely $\delta_{ab} + \delta_{\emptyset} - \delta_a - \delta_b$ (encodes $a \perp\!\!\!\perp b | \emptyset$) and $\delta_{abc} + \delta_c - \delta_{ac} - \delta_{bc}$ (encodes $a \perp\!\!\!\perp b | c$); the respective essential graphs are $a \rightarrow c \leftarrow b$ and $a — c — b$.

Example: the case of three variables

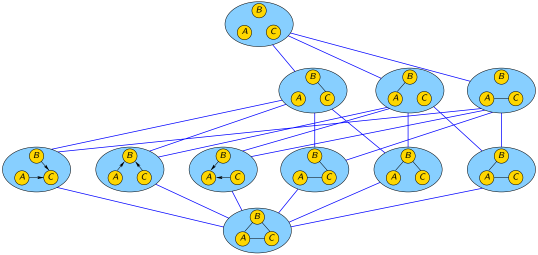
In the case of 3 variables one has 11 BN structures breaking into 5 types.



- The zero imset corresponds to the complete (undirected) essential graph.
- Six "elementary" imsets break into two types, namely $\delta_{ab} + \delta_{\emptyset} - \delta_{a} - \delta_{b}$ (encodes $a \perp\!\!\!\perp b|\emptyset$) and $\delta_{abc} + \delta_{c} - \delta_{ac} - \delta_{bc}$ (encodes $a \perp\!\!\!\perp b|c$); the respective essential graphs are $a \rightarrow c \leftarrow b$ and $a — c — b$.
- Three "semi-elementary" imsets of the form $\delta_{abc} + \delta_{\emptyset} - \delta_{a} - \delta_{bc}$ (encodes $a \perp\!\!\!\perp bc|\emptyset$) define one type; the respective graphs have just one edge.

Example: the case of three variables

In the case of 3 variables one has 11 BN structures breaking into 5 types.



- The zero imset corresponds to the complete (undirected) essential graph.
- Six "elementary" imsets break into two types, namely $\delta_{ab} + \delta_{\emptyset} - \delta_{a} - \delta_{b}$ (encodes $a \perp\!\!\!\perp b|\emptyset$) and $\delta_{abc} + \delta_{c} - \delta_{ac} - \delta_{bc}$ (encodes $a \perp\!\!\!\perp b|c$); the respective essential graphs are $a \to c \leftarrow b$ and $a — c — b$.
- Three "semi-elementary" imsets of the form $\delta_{abc} + \delta_{\emptyset} - \delta_{a} - \delta_{bc}$ (encodes $a \perp\!\!\!\perp bc|\emptyset$) define one type; the respective graphs have just one edge.
- The imset $\delta_N - \sum_{i \in N} \delta_i + 2 \cdot \delta_{\emptyset}$ corresponds to the empty essential graph.

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 Euclidean space $\mathbb{R}^{\mathcal{P}(N)}$.

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 Euclidean space $\mathbb{R}^{\mathcal{P}(N)}$.

This is a crucial observation made in a recent paper:

Theorem (Studený Vomlel Hemmecke 2010)

Having N fixed, none of the standard imsets is the convex combination of others. In particular, the set of standard imsets over N is the set of vertices (= extreme points) of a polytope $\mathsf{P} \subseteq \mathbb{R}^{\mathcal{P}(N)}$.

M. Studený, J. Vomlel and R. Hemmecke (2010). A geometric view on learning Bayesian network structures, to appear in *International Journal of Approximate Reasoning*.

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 Euclidean space $\mathbb{R}^{\mathcal{P}(N)}$.

This is a crucial observation made in a recent paper:

Theorem (Studený Vomlel Hemmecke 2010)

Having N fixed, none of the standard imsets is the convex combination of others. In particular, the set of standard imsets over N is the set of vertices (= extreme points) of a polytope $\mathsf{P} \subseteq \mathbb{R}^{\mathcal{P}(N)}$.

M. Studený, J. Vomlel and R. Hemmecke (2010). A geometric view on learning Bayesian network structures, to appear in *International Journal of Approximate Reasoning*.

We will call the polytope P the *standard imset polytope*.

Convex geometry: polytopes and polyhedrons

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

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* $\dim(P)$ is the dimension of its affine hull.

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* $\mathrm{dim(P)}$ 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*.

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* $\dim(\mathsf{P})$ 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

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

where $0 \neq \mathbf{v} \in \mathbb{R}^K$ is a non-zero vector and $\alpha \in \mathbb{R}$.

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

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

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.

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* $\dim(\mathsf{P})$ 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

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

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

This is a fundamental result in convex geometry (not easy to prove).

Theorem (Weyl-Minkowski theorem)

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

Convex geometry: Weyl-Minkowski theorem

This is a fundamental result in convex geometry (not easy to prove).

Theorem (Weyl-Minkowski theorem)

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

If P is a full-dimensional polytope then its *"minimal" description* in the form of a polyhedron is unique.

Convex geometry: Weyl-Minkowski theorem

This is a fundamental result in convex geometry (not easy to prove).

Theorem (Weyl-Minkowski theorem)

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

If P is a full-dimensional polytope then its *"minimal" description* in the form of a polyhedron is unique.

Thus, every polytope has two types of unique description:

Convex geometry: Weyl-Minkowski theorem

This is a fundamental result in convex geometry (not easy to prove).

Theorem (Weyl-Minkowski theorem)

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

If P is a full-dimensional polytope then its *"minimal" description* in the form of a polyhedron is unique.

Thus, every polytope has two types of unique description:

- *inner description* (= in terms of extreme points),

Convex geometry: Weyl-Minkowski theorem

This is a fundamental result in convex geometry (not easy to prove).

Theorem (Weyl-Minkowski theorem)

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

If P is a full-dimensional polytope then its *"minimal" description* in the form of a polyhedron is unique.

Thus, every polytope has two types of unique description:

- *inner description* (= in terms of extreme points),
- *outer description* (= in terms of half-spaces).

Convex geometry: Weyl-Minkowski theorem

This is a fundamental result in convex geometry (not easy to prove).

Theorem (Weyl-Minkowski theorem)

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

If P is a full-dimensional polytope then its *"minimal" description* in the form of a polyhedron is unique.

Thus, every polytope has two types of unique description:

- *inner description* (= in terms of extreme points),
- *outer description* (= in terms of half-spaces).

They are several computer packages that allow one to change between these two types of description of a polytope.

Convex geometry: faces (of a polytope)

Definition (face)

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

$$\mathsf{H} = \{\mathbf{x} \in \mathbb{R}^K;\ \langle \mathbf{v}, \mathbf{x} \rangle = \alpha\} \qquad \text{where } 0 \neq \mathbf{v} \in \mathbb{R}^K \text{ and } \alpha \in \mathbb{R}\,.$$

Convex geometry: faces (of a polytope)

Definition (face)

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

$$\mathsf{H} = \{\mathbf{x} \in \mathbb{R}^K;\ \langle \mathbf{v}, \mathbf{x} \rangle = \alpha\} \qquad \text{where } 0 \neq \mathbf{v} \in \mathbb{R}^K \text{ and } \alpha \in \mathbb{R}\,.$$

A *face* F of a polytope $\mathsf{P} \subseteq \mathbb{R}^K$ is the intersection of P with a hyperplane isolating it:

$$\mathsf{F} = \mathsf{P} \cap \mathsf{H} \quad \text{where} \quad \text{either} \quad \mathsf{P} \subseteq \mathsf{H}^{+} \equiv \{\mathbf{x} \in \mathbb{R}^K;\ \langle \mathbf{v}, \mathbf{x} \rangle \leq \alpha\},$$
$$\text{or} \quad \mathsf{P} \subseteq \mathsf{H}^{-} \equiv \{\mathbf{x} \in \mathbb{R}^K;\ \langle \mathbf{v}, \mathbf{x} \rangle \geq \alpha\}\,.$$

Convex geometry: faces (of a polytope)

Definition (face)

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

$$\mathsf{H} = \{\mathbf{x} \in \mathbb{R}^K;\ \langle \mathbf{v}, \mathbf{x} \rangle = \alpha\} \qquad \text{where } 0 \neq \mathbf{v} \in \mathbb{R}^K \text{ and } \alpha \in \mathbb{R}\,.$$

A *face* F of a polytope $\mathsf{P} \subseteq \mathbb{R}^K$ is the intersection of P with a hyperplane isolating it:

$$\mathsf{F} = \mathsf{P} \cap \mathsf{H} \quad \text{where} \quad \text{either} \quad \mathsf{P} \subseteq \mathsf{H}^+ \equiv \{\mathbf{x} \in \mathbb{R}^K;\ \langle \mathbf{v}, \mathbf{x} \rangle \leq \alpha\},$$
$$\text{or} \quad \mathsf{P} \subseteq \mathsf{H}^- \equiv \{\mathbf{x} \in \mathbb{R}^K;\ \langle \mathbf{v}, \mathbf{x} \rangle \geq \alpha\}\,.$$

Of course, every face F of a polytope P is again a polytope.

Convex geometry: faces (of a polytope)

Definition (face)

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

$$\mathsf{H} = \{\mathbf{x} \in \mathbb{R}^K;\ \langle \mathbf{v}, \mathbf{x} \rangle = \alpha\} \qquad \text{where } 0 \neq \mathbf{v} \in \mathbb{R}^K \text{ and } \alpha \in \mathbb{R}\,.$$

A *face* F of a polytope $\mathsf{P} \subseteq \mathbb{R}^K$ is the intersection of P with a hyperplane isolating it:

$$\mathsf{F} = \mathsf{P} \cap \mathsf{H} \quad \text{where} \quad \text{either} \quad \mathsf{P} \subseteq \mathsf{H}^+ \equiv \{\mathbf{x} \in \mathbb{R}^K;\ \langle \mathbf{v}, \mathbf{x} \rangle \leq \alpha\},$$
$$\text{or} \quad \mathsf{P} \subseteq \mathsf{H}^- \equiv \{\mathbf{x} \in \mathbb{R}^K;\ \langle \mathbf{v}, \mathbf{x} \rangle \geq \alpha\}\,.$$

Of course, every face F of a polytope P is again a polytope.

Faces can be classified by their dimensions. The most important are *vertices* (= extreme points), *edges* and *facets*.

The outer description of a polytope corresponds to its facets.

Linear programming and simplex method

The classic problem of *linear programming* (LP) it to maximize/minimize a linear function over a polyhedron.

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

Linear programming and simplex method

The classic problem of *linear programming* (LP) it to maximize/minimize a linear function over a polyhedron.

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

Since an affine function differs from a linear function by a constant, the task to maximize it over a polytope (= bounded polyhedron) is, in fact, an LP problem.

Linear programming and simplex method

The classic problem of *linear programming* (LP) it to maximize/minimize a linear function over a polyhedron.

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

Since an affine function differs from a linear function by a constant, the task to maximize it over a polytope (= bounded polyhedron) is, in fact, an LP problem.

A classic tool to solve the LP problems is the *simplex method*.

Linear programming and simplex method

The classic problem of *linear programming* (LP) it to maximize/minimize a linear function over a polyhedron.

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

Since an affine function differs from a linear function by a constant, the task to maximize it over a polytope (= bounded polyhedron) is, in fact, an LP problem.

A classic tool to solve the LP problems is the *simplex method*.

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

Linear programming and simplex method

The classic problem of *linear programming* (LP) it to maximize/minimize a linear function over a polyhedron.

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

Since an affine function differs from a linear function by a constant, the task to maximize it over a polytope (= bounded polyhedron) is, in fact, an LP problem.

A classic tool to solve the LP problems is the *simplex method*.

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

Again, there are efficient software packages to solve the LP problems.

Algorithmic issue: plan

Recall two main formerly mentioned results:

Algorithmic issue: plan

Recall two main formerly mentioned results:

- every "reasonable" quality criterion $\mathcal{Q}$ can be viewed as (the restriction of) an affine function on (the set of) standard imsets,

Algorithmic issue: plan

Recall two main formerly mentioned results:

- every "reasonable" quality criterion $\mathcal{Q}$ can be viewed as (the restriction of) an affine function on (the set of) standard imsets,
- the set of standard imsets over a fixed set of variables N is the set of vertices of a certain polytope P, called *standard imset polytope*.

Algorithmic issue: plan

Recall two main formerly mentioned results:

- every "reasonable" quality criterion $\mathcal{Q}$ can be viewed as (the restriction of) an affine function on (the set of) standard imsets,
- the set of standard imsets over a fixed set of variables N is the set of vertices of a certain polytope P, called *standard imset polytope*.

In particular, the task to maximize $\mathcal{Q}$ over BN structures is equivalent to the task to maximize an affine function over that polytope P.

Algorithmic issue: plan

Recall two main formerly mentioned results:

- every "reasonable" quality criterion $\mathcal{Q}$ can be viewed as (the restriction of) an affine function on (the set of) standard imsets,
- the set of standard imsets over a fixed set of variables N is the set of vertices of a certain polytope P, called *standard imset polytope*.

In particular, the task to maximize $\mathcal{Q}$ over BN structures is equivalent to the task to maximize an affine function over that polytope P.

However, this can is, in fact, an LP problem. Thus, efficient LP methods and software packages can perhaps be applied in future also in the area of structural learning Bayesian nets.

Algorithmic issue: plan

Recall two main formerly mentioned results:

- every "reasonable" quality criterion $\mathcal{Q}$ can be viewed as (the restriction of) an affine function on (the set of) standard imsets,
- the set of standard imsets over a fixed set of variables N is the set of vertices of a certain polytope P, called *standard imset polytope*.

In particular, the task to maximize $\mathcal{Q}$ over BN structures is equivalent to the task to maximize an affine function over that polytope P.

However, this can is, in fact, an LP problem. Thus, efficient LP methods and software packages can perhaps be applied in future also in the area of structural learning Bayesian nets.

This goal led us in the conclusions of (Studený Vomlel Hemmecke 2010) to formulate several research directions (= open mathematical questions). They concern the complexity of a potential future LP procedure for maximization of a quality criterion $\mathcal{Q}$.

Open questions (research directions)

- Get the *outer description* of the polytope P (for any $|N|$).

  The standard version of the simplex method assumes that the region is described in the form of a polyhedron. Therefore, the general characterization of P in this form would be an ideal solution.

Open questions (research directions)

- Get the *outer description* of the polytope P (for any $|N|$).
  The standard version of the simplex method assumes that the region is described in the form of a polyhedron. Therefore, the general characterization of P in this form would be an ideal solution.
- A related question is how "thick" or "thin" the polytope P is. More specifically, is there any imset in its interior?

Open questions (research directions)

- Get the *outer description* of the polytope P (for any $|N|$).
  The standard version of the simplex method assumes that the region is described in the form of a polyhedron. Therefore, the general characterization of P in this form would be an ideal solution.
- A related question is how "thick" or "thin" the polytope P is. More specifically, is there any imset in its interior?
- Describe the *edges* of the polytope P.
  Because of the above-mentioned interpretation of the simplex method, a potential alternative LP approach may be based only on vertices and edges. Vertices are known, it remains to find edges, which is sometimes easier than finding facets (= outer description).

Open questions (research directions)

- Get the *outer description* of the polytope P (for any $|N|$).
  The standard version of the simplex method assumes that the region is described in the form of a polyhedron. Therefore, the general characterization of P in this form would be an ideal solution.
- A related question is how "thick" or "thin" the polytope P is. More specifically, is there any imset in its interior?
- Describe the *edges* of the polytope P.
  Because of the above-mentioned interpretation of the simplex method, a potential alternative LP approach may be based only on vertices and edges. Vertices are known, it remains to find edges, which is sometimes easier than finding facets (= outer description).
- Try to apply the geometric approach to *restricted learning* BN structures (e.g. learning undirected trees).
  Learning within a subclass of BN structures is equivalent to maximization over a sub-polytope, which may appear to be simpler!

Recent results: towards outer description

The above mentioned questions were a topic of a conference contribution:

M. Studený and J. Vomlel (2009). On open questions in the geometric approach to learning BN structures. In *Proceedings of WUPES 09* (T. Kroupa, J. Vejnarová eds.), Liblice CZ, September 19-23, 2009.

Recent results: towards outer description

The above mentioned questions were a topic of a conference contribution:

M. Studený and J. Vomlel (2009). On open questions in the geometric approach to learning BN structures. In *Proceedings of WUPES 09* (T. Kroupa, J. Vejnarová eds.), Liblice CZ, September 19-23, 2009.

Necessary linear constraints:

On the basis of a detailed analysis case of four variables ($|N| = 4$), we have established a class of *necessary linear constraints* on the elements of the polytope P. Specifically, we classified them into three groups:

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

Recent results: towards outer description

The above mentioned questions were a topic of a conference contribution:

M. Studený and J. Vomlel (2009). On open questions in the geometric approach to learning BN structures. In *Proceedings of WUPES 09* (T. Kroupa, J. Vejnarová eds.), Liblice CZ, September 19-23, 2009.

Necessary linear constraints:

On the basis of a detailed analysis case of four variables ($|N| = 4$), we have established a class of *necessary linear constraints* on the elements of the polytope P. Specifically, we classified them into three groups:

- the *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 $|N| + 1$.
The dimension of the standard imset polytope in $2^{|N|} - |N| - 1$.

Inequality constraints

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

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

$\|N\|$	2	3	4	5
$\|\mathcal{K}_{\ell}^{\diamond}(N)\|$	1	5	37	117978

Inequality constraints

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

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

$\|N\|$	2	3	4	5
$\|\mathcal{K}_\ell^\diamond(N)\|$	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: Number of specific inequality constraints for $|N| \leq 4$.

$\|N\|$	2	3	4
$\sharp$	1	8	114

Conjecture about the outer description

We proved that all above mentioned inequalities are *necessary* for the points in the polytope P (for any N).

Conjecture about the outer description

We proved that all above mentioned inequalities are *necessary* for the points in the polytope P (for any N).

Moreover, in case $|N| = 4$ they give *complete characterization* of those points. Actually, they lead to the outer description of P in that case.

Conjecture about the outer description

We proved that all above mentioned inequalities are *necessary* for the points in the polytope P (for any N).

Moreover, in case $|N| = 4$ they give *complete characterization* of those points. Actually, they lead to the outer description of P in that case.

Conjecture

The above mentioned necessary linear constraints (A)-(C) together form also a sufficient condition for a vector to belong to P.

Conjecture about the outer description

We proved that all above mentioned inequalities are *necessary* for the points in the polytope P (for any N).

Moreover, in case $|N| = 4$ they give *complete characterization* of those points. Actually, they lead to the outer description of P in that case.

Conjecture

The above mentioned necessary linear constraints (A)-(C) together form also a sufficient condition for a vector to belong to P.

We would like either confirm or disprove the conjecture for $|N| = 5$ using a computer programm (task for R. Hemmecke and S. Lindner).

Recent results: integral vectors within 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.

Recent results: integral vectors within 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 integral (= integer-valued) vector within the standard imset polytope is already the standard imset.

Recent results: integral vectors within 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 integral (= integer-valued) vector within the standard imset polytope is already the standard imset.

Theorem

The only imsets within the polytope P *are its vertices.*

Recent results: integral vectors within 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 integral (= integer-valued) vector within the standard imset polytope is already the standard imset.

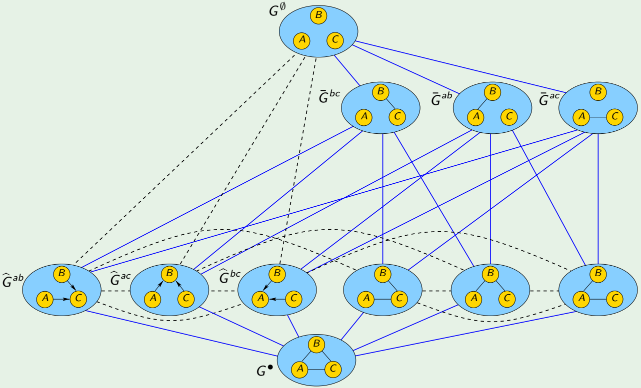
Theorem

The only imsets within the polytope P *are its vertices.*

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

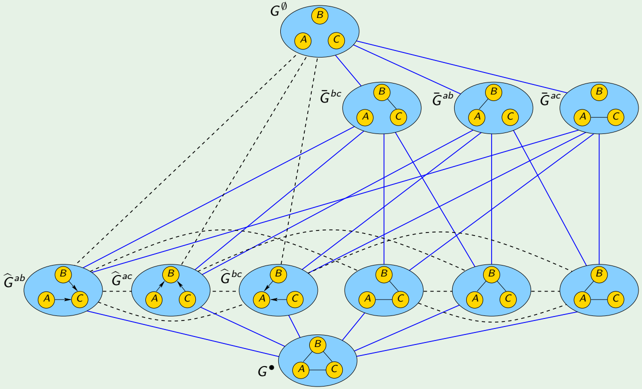
Edges of the standard imset polytope

Example (geometric edges in the case of three variables)



Edges of the standard imset polytope

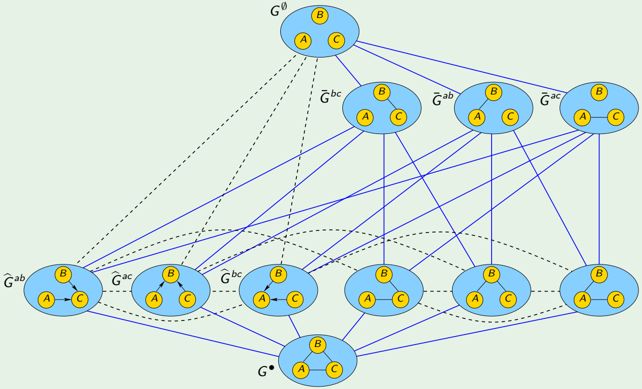
Example (geometric edges in the case of three variables)



The edges of P were found for $|N| \leq 5$ using LP software.

Edges of the standard imset polytope

Example (geometric edges in the case of three variables)



The edges of P were found for $|N| \leq 5$ using LP software.

Nevertheless, what is substantial, are their *directions* (and lengths)!

Recent results: geometric neighborhood

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. Their difference $w = u - v$ is then called the *differential imset*.

Recent results: geometric neighborhood

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. Their difference $w = u - v$ is then called the *differential imset*.

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`

Recent results: geometric neighborhood

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. Their difference $w = u - v$ is then called the *differential imset*.

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`

We have classified their types using three criteria:

- the *squared Euclidean length* of w, i.e. , $\sum_{S \subseteq N} w(S)^2$,
- the number of non-zero imset values, i.e. , $|\{S \subseteq N;\, w(S) \neq 0\}|$, and
- something, which is called the *degree difference* for w.

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.

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 several classes of models which fall withing this frame:

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 several classes of models which fall withing this frame:

- the class of *decomposable models*,

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 several classes of models which fall withing this frame:

- the class of *decomposable models*,
- (undirected) graphical models that correspond to *trees* or *forests*,

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 several classes of models which fall withing this frame:

- the class of *decomposable models*,
- (undirected) graphical models that correspond to *trees* or *forests*,
- the class of BN structures described by acyclic directed graphs with *prescribed order* of nodes (= variables).

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 several classes of models which fall withing this frame:

- the class of *decomposable models*,
- (undirected) graphical models that correspond to *trees* or *forests*,
- the class of BN structures described by acyclic directed graphs with *prescribed order* of nodes (= variables).

Since these subclasses are (sometimes substantially) smaller than the whole class, there is a chance that the corresponding sub-polytope is simpler!

It could be the case the sub-polytope can be characterized in terms of inequalities or in terms of edges, which can result in efficient LP procedures for learning those classes of models.

Restricted learning: forests

Silvia Lindner made some observations concerning the case of *forests*:

R. Hemmecke, S. Lindner, J. Vomlel and M. Studený (2010). Polyhedral studies for learning Bayesian network structures, a manuscript in preparation.

Restricted learning: forests

Silvia Lindner made some observations concerning the case of *forests*:

R. Hemmecke, S. Lindner, J. Vomlel and M. Studený (2010). Polyhedral studies for learning Bayesian network structures, a manuscript in preparation.

The respective sub-polytope can be transformed linearly to a so-called *independent set polytope*, formerly studied within matroid theory.

Restricted learning: forests

Silvia Lindner made some observations concerning the case of *forests*:

R. Hemmecke, S. Lindner, J. Vomlel and M. Studený (2010). Polyhedral studies for learning Bayesian network structures, a manuscript in preparation.

The respective sub-polytope can be transformed linearly to a so-called *independent set polytope*, formerly studied within matroid theory.

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

Restricted learning: forests

Silvia Lindner made some observations concerning the case of *forests*:

R. Hemmecke, S. Lindner, J. Vomlel and M. Studený (2010). Polyhedral studies for learning Bayesian network structures, a manuscript in preparation.

The respective sub-polytope can be transformed linearly to a so-called *independent set polytope*, formerly studied within matroid theory.

The point is that both edges and the outer description of that 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, C.N. Liu (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Infor. Theory* **14**: 462-467.

Restricted learning: forests

Silvia Lindner made some observations concerning the case of *forests*:

R. Hemmecke, S. Lindner, J. Vomlel and M. Studený (2010). Polyhedral studies for learning Bayesian network structures, a manuscript in preparation.

The respective sub-polytope can be transformed linearly to a so-called *independent set polytope*, formerly studied within matroid theory.

The point is that both edges and the outer description of that 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, C.N. Liu (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Infor. 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 R. Hemmecke).

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 R. 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.

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 R. 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*.

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 R. 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*.

We also plan to have a closer look at the promising linear transformation of standard imsets, which may lead to more intuitive way of geometric description of BN structures.