# Integer programming approach to statistical learning graphical models

Milan Studený

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

Fields Institute for Research in Mathematical Sciences
Toronto, Canada, April 17, 2012, 14:00

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. Linear programming approach
4. Integer programming approach
   - Characteristic imset
5. Comparison with other approaches
   - Straightforward zero-one encoding of a directed graph
6. LP relaxation of the characteristic imset polytope
7. Conclusions

# Motivation: learning Bayesian network structure

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

# Motivation: learning Bayesian network structure

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

The original motivation for our research was learning BN structure from data (= *model selection*) by a score-and-search method.

# Motivation: learning Bayesian network structure

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

The original motivation for our research was learning BN structure from data (= *model selection*) 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 $D$.

# Motivation: learning Bayesian network structure

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

The original motivation for our research was learning BN structure from data (= *model selection*) 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 $D$.

Examples of such criteria are Schwarz's *BIC criterion* and Bayesian *BDE score*.

# Motivation: learning Bayesian network structure

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

The original motivation for our research was learning BN structure from data (= *model selection*) 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 $D$.

Examples of such criteria are Schwarz's *BIC criterion* and Bayesian *BDE score*.

Here, the general aim is to develop a method for finding *global maximum* of $\mathcal{Q}$ based on tools of *linear programming* (LP).

# Basic concepts: Bayesian network structure

| | |
|---|---|
| $N$ | a non-empty finite set of *variables* |
| $\mathsf{X}_i$, $\lvert\mathsf{X}_i\rvert \geq 2$ | the individual sample spaces (for $i \in N$) |

# Basic concepts: Bayesian network structure

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

# Basic concepts: Bayesian network structure

| | |
|---|---|
| $N$ | a non-empty finite set of *variables* |
| $\mathsf{X}_i$, $\lvert\mathsf{X}_i\rvert \geq 2$ | 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 $G$.

# Basic concepts: Bayesian network structure

| | |
|---|---|
| $N$ | a non-empty finite set of *variables* |
| $\mathsf{X}_i$, $\lvert\mathsf{X}_i\rvert \geq 2$ | 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 $\mathsf{X}_N \equiv \prod_{i \in N} \mathsf{X}_i$ 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 $\mathsf{X}_N$ that factorize according to $G$.

# Basic concepts: Bayesian network structure

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

The (discrete) *Bayesian network* (BN) is a pair $(G, P)$, where $G \in$ 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 $G \in$ DAGS $(N)$, (the statistical model of) a *BN structure* is the class of all distributions $P$ on $\mathsf{X}_N$ that factorize according to $G$.

This statistical model can equivalently be defined in terms *conditional independence* (CI) – thus, it is a special model of a CI structure.

# Basic concepts: Bayesian network structure

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

The (discrete) *Bayesian network* (BN) is a pair $(G, P)$, where $G \in$ 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 $G \in$ DAGS ($N$), (the statistical model of) a *BN structure* is the class of all distributions $P$ on $\mathsf{X}_N$ that factorize according to $G$.

This statistical model can equivalently be defined in terms *conditional independence* (CI) – thus, it is a special model of a CI structure.

Two different acyclic directed graphs over $N$ may describe the same BN structure; a common unique graphical representative of the equivalence class of these graphs is so-called *essential graph*.

# Learning concepts: score-and-search method

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

Provided the individual sample spaces $\mathsf{X}_i$ for $i \in N$ are fixed,

$x^1, \ldots, x^d$ a sequence of elements of $\mathsf{X}_N$ of the length $d \geq 1$ called a *database of the length d* or a *sample of the size d*

# Learning concepts: score-and-search method

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

Provided the individual sample spaces $\mathsf{X}_i$ for $i \in N$ are fixed,

| | |
|---|---|
| $x^1, \ldots, x^d$ | a sequence of elements of $\mathsf{X}_N$ 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 $d$ |

# Learning concepts: score-and-search method

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

Provided the individual sample spaces $\mathsf{X}_i$ for $i \in N$ are fixed,

| | |
|---|---|
| $x^1, \ldots, x^d$ | a sequence of elements of $\mathsf{X}_N$ of the length $d \geq 1$ called a *database of the length d* or a *sample of the size d* |
| DATA $(N, d)$ | the set of all databases over $N$ of the length $d$ |

**Definition (quality criterion)**

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

# Learning concepts: score-and-search method

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

Provided the individual sample spaces $\mathsf{X}_i$ for $i \in N$ are fixed,

| | |
|---|---|
| $x^1, \ldots, x^d$ | a sequence of elements of $\mathsf{X}_N$ 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 $d$ |

## 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 $D$ (formal definition of *statistical consistency* is omitted).

Therefore, the aim is to maximize the function $G \mapsto \mathcal{Q}(G, D)$ given the observed database $D \in \mathrm{DATA}\,(N, d)$. This was traditionally done by special *search methods*, which however, in general, do not ensure finding a global maximizer.

# Learning concepts: technical requirements on criteria

**Notation**: Given an acyclic directed graph $G$ over $N$ and its node $i \in N$, $pa_G(i) \equiv \{j \in N;\, j \to i \text{ in } G\}$ is (called) the set of *parents* of $i$.

# Learning concepts: technical requirements on criteria

**Notation**: Given an acyclic directed graph $G$ over $N$ and its node $i \in N$, $pa_G(i) \equiv \{j \in N;\, j \to i \text{ in } G\}$ is (called) the set of *parents* of $i$.

## Definition (score equivalent and decomposable criterion)

A quality criterion $\mathcal{Q}$ will be named *score equivalent* if, for any database $D$,

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

# Learning concepts: technical requirements on criteria

**Notation**: Given an acyclic directed graph $G$ over $N$ and its node $i \in N$, $pa_G(i) \equiv \{j \in N;\, j \to i \text{ in } G\}$ is (called) the set of *parents* of $i$.

### Definition (score equivalent and decomposable criterion)

A quality criterion $\mathcal{Q}$ will be named *score equivalent* if, for any database $D$,

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

It will be called (additively) *decomposable* if it has the form

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

where $D_A$ is the projection of $D$ to the marginal space $\mathsf{X}_A$ for $A \subseteq N$. The terms $q_{i|B}(*|*)$ for $i \in N$ and $B \subseteq N \setminus \{i\}$ are called *local scores*.

# Learning concepts: technical requirements on criteria

**Notation**: Given an acyclic directed graph $G$ over $N$ and its node $i \in N$, $pa_G(i) \equiv \{j \in N;\, j \to i \text{ in } G\}$ is (called) the set of *parents* of $i$.

### Definition (score equivalent and decomposable criterion)

A quality criterion $\mathcal{Q}$ will be named *score equivalent* if, for any database $D$,

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

It will be called (additively) *decomposable* if it has the form

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

where $D_A$ is the projection of $D$ to the marginal space $\mathsf{X}_A$ for $A \subseteq N$. The terms $q_{i|B}(*|*)$ for $i \in N$ and $B \subseteq N \setminus \{i\}$ are called *local scores*.

Quality criteria used in practice are score equivalent and decomposable.

# Linear algebraic approach: imset

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

## 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 $N$.

# Linear algebraic approach: imset

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

**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 $\mathbb{R}^{\mathcal{P}(N)}$]

# Linear algebraic approach: imset

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

**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.

# Linear algebraic approach: imset

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

**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 $\delta_A$ will denote this *basic imset*:

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

# Linear algebraic approach: imset

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

## 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 $\delta_A$ will denote this *basic imset*:

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

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).

# Algebraic concepts: standard imset

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 $G$).

# Algebraic concepts: standard imset

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 $G$).

## Definition (standard imset)

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

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

# Algebraic concepts: standard imset

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 $G$).

## Definition (standard imset)

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

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

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 $|N|$.

# Algebraic concepts: standard imset

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 $G$).

### Definition (standard imset)

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

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

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 $|N|$.

# Algebraic approach to learning

The standard imset is a unique representative of the BN structure.

**Lemma (Studený 2005)**

Given $G, H \in \mathrm{DAGS}\,(N)$, $u_G = u_H$ iff $G$ and $H$ are equivalent.

# Algebraic approach to learning

The standard imset is a unique representative of the BN structure.

**Lemma (Studený 2005)**

Given $G, H \in \mathrm{DAGS}\,(N)$, $u_G = u_H$ iff $G$ and $H$ are equivalent.

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

# Algebraic approach to learning

The standard imset is a unique representative of the BN structure.

**Lemma (Studený 2005)**

Given $G, H \in \mathrm{DAGS}(N)$, $u_G = u_H$ iff $G$ and $H$ are equivalent.

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

**Theorem (Studený 2005)**

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

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

*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 $G$.*

The vector $t_D^{\mathcal{Q}}$ is called the *data vector* with respect to $\mathcal{Q}$.

# Geometric view on learning

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

# Geometric view on learning

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

## Definition (standard imset polytope)

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

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

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

# Geometric view on learning

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

## Definition (standard imset polytope)

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

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

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

In the set S each BN structure is represented by just one vector! We have shown $\mathsf{S} = \mathrm{ext}(\mathsf{P})$. Thus, maximizing $\mathcal{Q}$ over BN structures is equivalent to finding an optimum of an affine function over P.

# Geometric view on learning

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

### Definition (standard imset polytope)

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

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

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

In the set S each BN structure is represented by just one vector! We have shown $\mathsf{S} = \mathrm{ext}(\mathsf{P})$. Thus, maximizing $\mathcal{Q}$ over BN structures is equivalent to finding an optimum of an affine function over P.

However, to apply classic tools of LP, like the simplex method, one has to have a *polyhedral description* of the domain P. An alternative approach could be based is a characterization of geometric edges of P (= 2-faces).

# Technical problems with the direct LP approach

M. Studený and J. Vomlel (2011). On open questions in the geometric approach to structural learning Bayesian nets. *International Journal of Approximate Reasoning* **52**:627-640.

# Technical problems with the direct LP approach

M. Studený and J. Vomlel (2011). On open questions in the geometric approach to structural learning Bayesian nets. *International Journal of Approximate Reasoning* **52**:627-640.

The are at least two reasons why such a direct LP approach does not seem to lead to a progress in solving practical learning tasks:

# Technical problems with the direct LP approach

M. Studený and J. Vomlel (2011). On open questions in the geometric approach to structural learning Bayesian nets. *International Journal of Approximate Reasoning* **52**:627-640.

The are at least two reasons why such a direct LP approach does not seem to lead to a progress in solving practical learning tasks:

- the number of inequalities in the conjectured outer description seems to be *super-exponential* in $|N|$,

# Technical problems with the direct LP approach

M. Studený and J. Vomlel (2011). On open questions in the geometric approach to structural learning Bayesian nets. *International Journal of Approximate Reasoning* **52**:627-640.

The are at least two reasons why such a direct LP approach does not seem to lead to a progress in solving practical learning tasks:

- the number of inequalities in the conjectured outer description seems to be *super-exponential* in $|N|$,
- the description of most of the inequalities is *implicit*. To apply them in practice one still would need to characterize them explicitly.

# Technical problems with the direct LP approach

M. Studený and J. Vomlel (2011). On open questions in the geometric approach to structural learning Bayesian nets. *International Journal of Approximate Reasoning* **52**:627-640.

The are at least two reasons why such a direct LP approach does not seem to lead to a progress in solving practical learning tasks:

- the number of inequalities in the conjectured outer description seems to be *super-exponential* in $|N|$,
- the description of most of the inequalities is *implicit*. To apply them in practice one still would need to characterize them explicitly.

The reason is that most of the inequalities correspond to *extreme supermodular functions* and one has to characterize these explicitly, which looks like a difficult open theoretical problem.

# Technical problems with the direct LP approach

M. Studený and J. Vomlel (2011). On open questions in the geometric approach to structural learning Bayesian nets. *International Journal of Approximate Reasoning* **52**:627-640.

The are at least two reasons why such a direct LP approach does not seem to lead to a progress in solving practical learning tasks:

- the number of inequalities in the conjectured outer description seems to be *super-exponential* in $|N|$,
- the description of most of the inequalities is *implicit*. To apply them in practice one still would need to characterize them explicitly.

The reason is that most of the inequalities correspond to *extreme supermodular functions* and one has to characterize these explicitly, which looks like a difficult open theoretical problem.

The result of our preliminary analysis of the *geometric edges* was an observation that P has a huge number of edges, and, at this stage, there is no hope for their complete characterization.

# Integer programming approach

The idea is to apply advanced methods of linear optimization. The point is that the considered polytope P is integral, that is, all its vertices are lattice points.

To apply the methods of *integer programming* (IP) one need not necessarily find a completed outer (= facet) description of the polytope.

# Integer programming approach

The idea is to apply advanced methods of linear optimization. The point is that the considered polytope P is integral, that is, all its vertices are lattice points.

To apply the methods of *integer programming* (IP) one need not necessarily find a completed outer (= facet) description of the polytope.

**Definition (LP relaxation)**

By an *LP relaxation* of a polytope P is meant a polyhedron R containing the polytope ($\mathsf{P} \subseteq \mathsf{R}$), with the property that the lattice points contained in P and R coincide ($\mathsf{P} \cap \mathbb{Z}^* = \mathsf{R} \cap \mathbb{Z}^*$).

# Integer programming approach

The idea is to apply advanced methods of linear optimization. The point is that the considered polytope P is integral, that is, all its vertices are lattice points.

To apply the methods of *integer programming* (IP) one need not necessarily find a completed outer (= facet) description of the polytope.

**Definition (LP relaxation)**

By an *LP relaxation* of a polytope P is meant a polyhedron R containing the polytope ($\mathsf{P} \subseteq \mathsf{R}$), with the property that the lattice points contained in P and R coincide ($\mathsf{P} \cap \mathbb{Z}^* = \mathsf{R} \cap \mathbb{Z}^*$).

Then the maximization task can be re-formulated in the form of *integer programing* (IP) problem:

$$\min \{ \langle t_D^{\mathcal{Q}}, u \rangle ;\ u \in \mathsf{R} ,\ u \in \mathbb{Z}^* \}$$

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

# Integer programming approach

The idea is to apply advanced methods of linear optimization. The point is that the considered polytope P is integral, that is, all its vertices are lattice points.

To apply the methods of *integer programming* (IP) one need not necessarily find a completed outer (= facet) description of the polytope.

**Definition (LP relaxation)**

By an *LP relaxation* of a polytope P is meant a polyhedron R containing the polytope ($\mathsf{P} \subseteq \mathsf{R}$), with the property that the lattice points contained in P and R coincide ($\mathsf{P} \cap \mathbb{Z}^* = \mathsf{R} \cap \mathbb{Z}^*$).

Then the maximization task can be re-formulated in the form of *integer programing* (IP) problem:

$$\min \{ \langle t_D^{\mathcal{Q}}, u \rangle ;\; u \in \mathsf{R} ,\; u \in \mathbb{Z}^* \}$$

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

There are software packages, which efficiently solve IP problems (CPLEX). In IP is often advantageous to have a polytope, whose vertices are zero-one vectors.

# Transformation to the characteristic imset

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

# Transformation to the characteristic imset

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

## 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

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

# Transformation to the characteristic imset

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

### 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

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

Clearly, the characteristic imset is obtained from the standard one by an invertible affine transformation. In particular, every score equivalent and decomposable criterion is an affine function of the characteristic imset!

# Transformation to the characteristic imset

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

**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

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

Clearly, the characteristic imset is obtained from the standard one by an invertible affine transformation. In particular, every score equivalent and decomposable criterion is an affine function of the characteristic imset!

The motivation for the terminology was that, if $G$ is a forest, then $c_G$ is the (zero extension of the) characteristic vector of its edge-set.

# Characteristic imset: basic observation

**Theorem (Studený, Hemmecke, Lindner 2010)**

*Assume $|N| \geq 2$. Given an acyclic directed graph $G$ over $N$ one has*

$$c_G(A) \in \{0,1\} \qquad \textit{for any} \;\; A \subseteq N,\; |A| \geq 2\,.$$

# Characteristic imset: basic observation

**Theorem (Studený, Hemmecke, Lindner 2010)**

*Assume $|N| \geq 2$. Given an acyclic directed graph $G$ over $N$ one has*

$$c_G(A) \in \{0,1\} \qquad \textit{for any} \;\; A \subseteq N,\; |A| \geq 2\,.$$

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!

# Characteristic imset: basic observation

**Theorem (Studený, Hemmecke, Lindner 2010)**

*Assume $|N| \geq 2$. Given an acyclic directed graph $G$ over $N$ one has*

$$c_G(A) \in \{0,1\} \qquad \text{for any} \;\; A \subseteq N, \; |A| \geq 2\,.$$

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.

# Characteristic imset: basic observation

**Theorem (Studený, Hemmecke, Lindner 2010)**

*Assume $|N| \geq 2$. Given an acyclic directed graph $G$ over $N$ one has*

$$c_G(A) \in \{0,1\} \qquad \textit{for any} \;\; A \subseteq N,\; |A| \geq 2\,.$$

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.

**Definition (characteristic imset polytope)**

*Characteristic imset polytope* is the convex hull of the set of characteristic imsets: $\mathsf{C} = \mathrm{conv}(\{c_G;\; G \in \mathrm{DAGS}\,(N)\})$

# Characteristic imset: directly from the graph

## Theorem (equivalent definition of a characteristic imset)

*Let $c_G$ be the characteristic imset for an acyclic directed graph $G$ over $N$. For $S \subseteq N$, $|S| \geq 2$ one has*

$$c_G(S) = 1 \text{ iff there exists some } i \in S \text{ with } S \setminus \{i\} \subseteq pa_G(i).$$

# Characteristic imset: directly from the graph

## Theorem (equivalent definition of a characteristic imset)

*Let $c_G$ be the characteristic imset for an acyclic directed graph $G$ over $N$. For $S \subseteq N$, $|S| \geq 2$ one has*

$$c_G(S) = 1 \text{ iff there exists some } i \in S \text{ with } S \setminus \{i\} \subseteq pa_G(i).$$

## Corollary (crucial components of the characteristic imset)

*Let $i, j$ (and $k$) are distinct nodes in $G$. Then:*

- *there is an edge between distinct nodes $i, j \in G$ if and only if $c_G(\{i, j\}) = 1$,*

# Characteristic imset: directly from the graph

**Theorem (equivalent definition of a characteristic imset)**

*Let $c_G$ be the characteristic imset for an acyclic directed graph $G$ over $N$. For $S \subseteq N$, $|S| \geq 2$ one has*

$$c_G(S) = 1 \text{ iff there exists some } i \in S \text{ with } S \setminus \{i\} \subseteq pa_G(i).$$

**Corollary (crucial components of the characteristic imset)**

*Let $i, j$ (and $k$) are distinct nodes in $G$. Then:*

- *there is an edge between distinct nodes $i, j \in G$ if and only if $c_G(\{i, j\}) = 1$,*
- *there is an immorality $i \rightarrow k \leftarrow j$ in $G$ if and only if $c_G(\{i, j, k\}) = 1$ and $c_G(\{i, j\}) = 0$.*

# Characteristic imset: directly from the graph

## Theorem (equivalent definition of a characteristic imset)

*Let $c_G$ be the characteristic imset for an acyclic directed graph $G$ over $N$. For $S \subseteq N$, $|S| \geq 2$ one has*

$$c_G(S) = 1 \text{ iff there exists some } i \in S \text{ with } S \setminus \{i\} \subseteq pa_G(i).$$

## Corollary (crucial components of the characteristic imset)

*Let $i, j$ (and $k$) are distinct nodes in $G$. Then:*

- *there is an edge between distinct nodes $i, j \in G$ if and only if $c_G(\{i, j\}) = 1$,*
- *there is an immorality $i \to k \leftarrow j$ in $G$ if and only if $c_G(\{i, j, k\}) = 1$ and $c_G(\{i, j\}) = 0$.*

*The characteristic imset $c_G$ is determined uniquely by its values for sets of cardinality 2 and 3.*

# Characteristic imset: directly from the graph

## Theorem (equivalent definition of a characteristic imset)

*Let $c_G$ be the characteristic imset for an acyclic directed graph $G$ over $N$. For $S \subseteq N$, $|S| \geq 2$ one has*

$$c_G(S) = 1 \text{ iff there exists some } i \in S \text{ with } S \setminus \{i\} \subseteq pa_G(i).$$

## Corollary (crucial components of the characteristic imset)

*Let $i, j$ (and $k$) are distinct nodes in $G$. Then:*

- *there is an edge between distinct nodes $i, j \in G$ if and only if $c_G(\{i, j\}) = 1$,*
- *there is an immorality $i \rightarrow k \leftarrow j$ in $G$ if and only if $c_G(\{i, j, k\}) = 1$ and $c_G(\{i, j\}) = 0$.*

*The characteristic imset $c_G$ is determined uniquely by its values for sets of cardinality 2 and 3.*

However, the values $c_G(S)$ for $|S| \geq 4$ do not depend linearly on them.

# Overview of methods for getting the global optimimum

Traditional score-and-search methods, like the *greedy equivalence search* (GES) algorithm, do not guarantee to find the global maximum of $\mathcal{Q}$.

# Overview of methods for getting the global optimimum

Traditional score-and-search methods, like the *greedy equivalence search* (GES) algorithm, do not guarantee to find the global maximum of $\mathcal{Q}$.

T. Silander, P. Myllymäki (2006). A simple approach for finding the globally optimal Bayesian network structure. In Proceedings of the 22th UAI conference, AUAI Press, pp. 445-452. method of dynamic programming

# Overview of methods for getting the global optimimum

Traditional score-and-search methods, like the *greedy equivalence search* (GES) algorithm, do not guarantee to find the global maximum of $\mathcal{Q}$.


T. Silander, P. Myllymäki (2006). A simple approach for finding the globally optimal Bayesian network structure. In Proceedings of the 22th UAI conference, AUAI Press, pp. 445-452. method of dynamic programming

C.P. de Campos, Z. Zeng, Q. Ji (2009). Structure learning Bayesian networks using constraints. In Proceedings of the 26th ICML conference, pp. 113-120. general branch-and-bound principle

# Overview of methods for getting the global optimimum

Traditional score-and-search methods, like the *greedy equivalence search* (GES) algorithm, do not guarantee to find the global maximum of $\mathcal{Q}$.


T. Silander, P. Myllymäki (2006). A simple approach for finding the globally optimal Bayesian network structure. In Proceedings of the 22th UAI conference, AUAI Press, pp. 445-452.

method of dynamic programming


C.P. de Campos, Z. Zeng, Q. Ji (2009). Structure learning Bayesian networks using constraints. In Proceedings of the 26th ICML conference, pp. 113-120.

general branch-and-bound principle


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.

ILP approach

# Overview of methods for getting the global optimimum

Traditional score-and-search methods, like the *greedy equivalence search* (GES) algorithm, do not guarantee to find the global maximum of $\mathcal{Q}$.


- T. Silander, P. Myllymäki (2006). A simple approach for finding the globally optimal Bayesian network structure. In Proceedings of the 22th UAI conference, AUAI Press, pp. 445-452. method of dynamic programming
- C.P. de Campos, Z. Zeng, Q. Ji (2009). Structure learning Bayesian networks using constraints. In Proceedings of the 26th ICML conference, pp. 113-120. general branch-and-bound principle
- 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. ILP approach
- J. Cussens (2010). Maximum likelihood pedigree reconstruction using integer programming. In Proceedings of CBMB workshop, pp. 9-19. ILP approach
- J. Cussens (2011). Bayesian network learning with cutting planes. In Proceedings of the 27th UAI conference, pp. 153-160. ILP approach

# Straightforward zero-one encoding for a directed graph

Both (Jaakkola *et al.* 2010) and (Cussens 2010, 2011) used a simple zero-one vector $\eta_G$ to encode a directed graph $G$ over $N$.

# Straightforward zero-one encoding for a directed graph

Both (Jaakkola *et al.* 2010) and (Cussens 2010, 2011) used a simple zero-one vector $\eta_G$ 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:

**Definition (straightforward zero-one code of a directed graph)**

Let $G$ be a(n acyclic) directed graph over $N$. Then we put

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

# Straightforward zero-one encoding for a directed graph

Both (Jaakkola *et al.* 2010) and (Cussens 2010, 2011) used a simple zero-one vector $\eta_G$ 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:

**Definition (straightforward zero-one code of a directed graph)**

Let $G$ be a(n acyclic) directed graph over $N$. Then we put

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

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

# Straightforward zero-one encoding for a directed graph

Both (Jaakkola *et al.* 2010) and (Cussens 2010, 2011) used a simple zero-one vector $\eta_G$ 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:

**Definition (straightforward zero-one code of a directed graph)**

Let $G$ be a(n acyclic) directed graph over $N$. Then we put

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

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.

# LP relaxation offered by Jaakkola *et al.*

Their polyhedron J was given by the following constraints:

# 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 $(i|B)$,

# 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 $j \in N$,

# 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

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

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.

# 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.

# 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 $N$.

# 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

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

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.

# Comparison with Jaakkola *et al.*'s approach

M. Studený, D. Haws (2011). On polyhedral approximations of polytopes for learning Bayes nets, research report n. 2303, Institute of Information Theory and Automation of the ASCR, `http://arxiv.org/abs/1107.4708`.

# Comparison with Jaakkola *et al.*'s approach

M. Studený, D. Haws (2011). On polyhedral approximations of polytopes for learning Bayes nets, research report n. 2303, Institute of Information Theory and Automation of the ASCR, `http://arxiv.org/abs/1107.4708`.

We have observed that the standard imset $u_G$ is an affine (many-to-one) function of $\eta_G$ and the characteristic imset $c_G$ is even its linear function:

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

# Comparison with Jaakkola *et al.*'s approach

M. Studený, D. Haws (2011). On polyhedral approximations of polytopes for learning Bayes nets, research report n. 2303, Institute of Information Theory and Automation of the ASCR, http://arxiv.org/abs/1107.4708.

We have observed that the standard imset $u_G$ is an affine (many-to-one) function of $\eta_G$ and the characteristic imset $c_G$ is even its linear function:

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

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

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

# Comparison with Jaakkola *et al.*'s approach

M. Studený, D. Haws (2011). On polyhedral approximations of polytopes for learning Bayes nets, research report n. 2303, Institute of Information Theory and Automation of the ASCR, `http://arxiv.org/abs/1107.4708`.

We have observed that the standard imset $u_G$ is an affine (many-to-one) function of $\eta_G$ and the characteristic imset $c_G$ is even its linear function:

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

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 *et al.*'s linear constraints to our framework(s) and to compare them with our constraints.

# Recent findings: inequalities translation

A positive finding was that the *cluster inequalities can easily be transformed* to the framework of standard imsets. They come to inequalities we already knew from our former analysis. Specifically, they correspond to certain extreme supermodular functions:

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

# Recent findings: inequalities translation

A positive finding was that the *cluster inequalities can easily be transformed* to the framework of standard imsets. They come to inequalities we already knew from our former analysis. Specifically, they correspond to certain extreme supermodular functions:

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

It was a slightly tough technical problem to transform *Jaakkola et al.s' non-negativity and equality constraints*. We found out that they are transformed to certain already known inequalities for the standard imset polytope and that the transformation raises the number of inequalities!

# Recent findings: inequalities translation

A positive finding was that the *cluster inequalities can easily be transformed* to the framework of standard imsets. They come to inequalities we already knew from our former analysis. Specifically, they correspond to certain extreme supermodular functions:

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

It was a slightly tough technical problem to transform *Jaakkola et al.s' non-negativity and equality constraints*. We found out that they are transformed to certain already known inequalities for the standard imset polytope and that the transformation raises the number of inequalities!

The consequence of the above observations is that the polyhedron conjectured in (Studený, Vomlel 2011) to be an outer description of the standard imset polytope P is indeed its LP relaxation.

# Towards LP relaxation of the characteristic imset polytope

We have also transformed Jaakkola *et al.*'s inequalities in the framework of characteristic imsets.

# Towards LP relaxation of the characteristic imset polytope

We have also transformed Jaakkola *et al.*'s inequalities in the framework of characteristic imsets.

Our question has been whether the transformed inequalities define an LP relaxation of the characteristic imset polytope.

# Towards LP relaxation of the characteristic imset polytope

We have also transformed Jaakkola *et al.*'s inequalities in the framework of characteristic imsets.

Our question has been whether the transformed inequalities define an LP relaxation of the characteristic imset polytope.

This appeared to be related to the *unimodularity* of the respective transformation-defining matrix. We have succeeded to confirm the conjecture that it is indeed the case.

# Towards LP relaxation of the characteristic imset polytope

We have also transformed Jaakkola *et al.*'s inequalities in the framework of characteristic imsets.

Our question has been whether the transformed inequalities define an LP relaxation of the characteristic imset polytope.

This appeared to be related to the *unimodularity* of the respective transformation-defining matrix. We have succeeded to confirm the conjecture that it is indeed the case.

Nevertheless, although we got an LP relaxation of the characteristic imset polytope, this particular one does not seem to be ideal for practical purposes, for the high number of inequalities.

# LP relaxation for an extended vector-code

S. Lindner (2012). Discrete optimization in machine learning - learning Bayesian network structures and conditional independence implication. PhD thesis, TU Munich.

Both (Cussens 2010) and (Lindner 2012) used another trick: they added some additional components to their vector codes. These additional components correspond to ordered pairs of variables.

# LP relaxation for an extended vector-code

S. Lindner (2012). Discrete optimization in machine learning - learning Bayesian network structures and conditional independence implication. PhD thesis, TU Munich.

Both (Cussens 2010) and (Lindner 2012) used another trick: they added some additional components to their vector codes. These additional components correspond to ordered pairs of variables.

The point was that only a polynomial number of additional components (in $|N|$) was added, but this step allowed them overcome technical problems (= to prove they get an LP relaxation of what they want).

# LP relaxation for an extended vector-code

S. Lindner (2012). Discrete optimization in machine learning - learning Bayesian network structures and conditional independence implication. PhD thesis, TU Munich.

Both (Cussens 2010) and (Lindner 2012) used another trick: they added some additional components to their vector codes. These additional components correspond to ordered pairs of variables.

The point was that only a polynomial number of additional components (in $|N|$) was added, but this step allowed them overcome technical problems (= to prove they get an LP relaxation of what they want).

Cussens extended the $\eta_G$-vector and used the new components to encode a total order of variables in $N$ consonant with the arrows in (acyclic directed) graph $G$.

# LP relaxation for an extended vector-code

S. Lindner (2012). Discrete optimization in machine learning - learning Bayesian network structures and conditional independence implication. PhD thesis, TU Munich.

Both (Cussens 2010) and (Lindner 2012) used another trick: they added some additional components to their vector codes. These additional components correspond to ordered pairs of variables.

The point was that only a polynomial number of additional components (in $|N|$) was added, but this step allowed them overcome technical problems (= to prove they get an LP relaxation of what they want).

Cussens extended the $\eta_G$-vector and used the new components to encode a total order of variables in $N$ consonant with the arrows in (acyclic directed) graph $G$.

Lindner considered an extension of the characteristic imset $c_G$. She used the additional components to encode the direction of arrows in an acyclic directed graph $G$ inducing $c_G$.

# Conclusions

Both (Cussens 2010, 2011) and (Lindner 2012) have done some practical computational experiments with this new ILP approach. They, unlike (Jaakkola *et al.* 2010), used some ILP software packages.

# Conclusions

Both (Cussens 2010, 2011) and (Lindner 2012) have done some practical computational experiments with this new ILP approach. They, unlike (Jaakkola *et al.* 2010), used some ILP software packages.

My future research direction in this area is as follows: consider an extension of the characteristic imset $c_G$ with additional components encoding the direction of arrows in the respective *essential graph*!

# Conclusions

Both (Cussens 2010, 2011) and (Lindner 2012) have done some practical computational experiments with this new ILP approach. They, unlike (Jaakkola *et al.* 2010), used some ILP software packages.

My future research direction in this area is as follows: consider an extension of the characteristic imset $c_G$ with additional components encoding the direction of arrows in the respective *essential graph*!

Actually, the idea is to encode the arrows in a graph which falls within a special wider class of graph, involving both all acyclic directed graphs inducing $c_G$ (= equivalent to $G$) and the respective essential graph for $G$.

# Conclusions

Both (Cussens 2010, 2011) and (Lindner 2012) have done some practical computational experiments with this new ILP approach. They, unlike (Jaakkola *et al.* 2010), used some ILP software packages.

My future research direction in this area is as follows: consider an extension of the characteristic imset $c_G$ with additional components encoding the direction of arrows in the respective *essential graph*!

Actually, the idea is to encode the arrows in a graph which falls within a special wider class of graph, involving both all acyclic directed graphs inducing $c_G$ (= equivalent to $G$) and the respective essential graph for $G$.

The essential graph can then be obtained by an additional simple auxiliary ILP problem (of polynomial complexity in $|N|$).

# Conclusions

Both (Cussens 2010, 2011) and (Lindner 2012) have done some practical computational experiments with this new ILP approach. They, unlike (Jaakkola *et al.* 2010), used some ILP software packages.

My future research direction in this area is as follows: consider an extension of the characteristic imset $c_G$ with additional components encoding the direction of arrows in the respective *essential graph*!

Actually, the idea is to encode the arrows in a graph which falls within a special wider class of graph, involving both all acyclic directed graphs inducing $c_G$ (= equivalent to $G$) and the respective essential graph for $G$.

The essential graph can then be obtained by an additional simple auxiliary ILP problem (of polynomial complexity in $|N|$).

Of course, I plan to work on it in cooperation with colleagues (abroad).