

# An Algebraic Approach to Structural Learning Bayesian Networks

Milan STUDENÝ

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

2nd international workshop Data - Algorithms - Decision Making

December 9-12, 2006, Třešť

# Summary of the talk

- 1 Introduction
- 2 Quality criteria
- 3 Problem of representative choice
- 4 Local search methods
- 5 Inclusion neighborhood
- 6 Algebraic approach: standard imsets
- 7 Discussion (algorithms)

# Introduction

The methods for learning probabilistic conditional independence (CI) structure models can be divided into two groups:

- the methods based on *significance tests*,
- the methods based on the maximization of a suitable *quality criterion*.

Note that some methods can be classified in both groups. Moreover, there is a simulation method, namely MCMC, applicable to learning graphical models which does not belong to either of these two groups.

Most of the learning methods were developed for **learning Bayesian network models**.

# Introduction

The methods for learning probabilistic conditional independence (CI) structure models can be divided into two groups:

- the methods based on *significance tests*,
- the methods based on the maximization of a suitable *quality criterion*.

Note that some methods can be classified in both groups. Moreover, there is a simulation method, namely MCMC, applicable to learning graphical models which does not belong to either of these two groups.

Most of the learning methods were developed for *learning Bayesian network models*.

# Introduction

The methods for learning probabilistic conditional independence (CI) structure models can be divided into two groups:

- the methods based on *significance tests*,
- the methods based on the maximization of a suitable *quality criterion*.

Note that some methods can be classified in both groups. Moreover, there is a simulation method, namely MCMC, applicable to learning graphical models which does not belong to either of these two groups.

Most of the learning methods were developed for **learning Bayesian network models**.

# Introduction

The methods for learning probabilistic conditional independence (CI) structure models can be divided into two groups:

- the methods based on *significance tests*,
- the methods based on the maximization of a suitable *quality criterion*.

Note that some methods can be classified in both groups. Moreover, there is a simulation method, namely MCMC, applicable to learning graphical models which does not belong to either of these two groups.

Most of the learning methods were developed for *learning Bayesian network models*.

# Introduction

The methods for learning probabilistic conditional independence (CI) structure models can be divided into two groups:

- the methods based on *significance tests*,
- the methods based on the maximization of a suitable *quality criterion*.

Note that some methods can be classified in both groups. Moreover, there is a simulation method, namely MCMC, applicable to learning graphical models which does not belong to either of these two groups.

Most of the learning methods were developed for [learning Bayesian network models](#).

# Quality criteria

This talk deals with methods for *learning Bayesian networks* based on the *maximization of a quality criterion*.

- Let  $N$  be a finite non-empty set of variables,
- $\text{DAGS}(N)$  will denote the class of acyclic directed graphs over the set of nodes  $N$ ,
- $\text{DATA}(N, d)$  will denote the set of all databases over  $N$  of the length  $d$ ,  $d \geq 1$  (some finite sample spaces are fixed).

## Definition

*Quality criterion* (for learning Bayesian networks) is a real function  $\mathcal{Q}$  on  $\text{DAGS}(N) \times \text{DATA}(N, d)$ .

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



# Quality criteria

This talk deals with methods for *learning Bayesian networks* based on the *maximization of a quality criterion*.

- Let  $N$  be a finite non-empty set of variables,
- $\text{DAGS}(N)$  will denote the class of acyclic directed graphs over the set of nodes  $N$ ,
- $\text{DATA}(N, d)$  will denote the set of all databases over  $N$  of the length  $d$ ,  $d \geq 1$  (some finite sample spaces are fixed).

## Definition

*Quality criterion* (for learning Bayesian networks) is a real function  $\mathcal{Q}$  on  $\text{DAGS}(N) \times \text{DATA}(N, d)$ .

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

# Quality criteria

This talk deals with methods for *learning Bayesian networks* based on the *maximization of a quality criterion*.

- Let  $N$  be a finite non-empty set of variables,
- $\text{DAGS}(N)$  will denote the class of acyclic directed graphs over the set of nodes  $N$ ,
- $\text{DATA}(N, d)$  will denote the set of all databases over  $N$  of the length  $d$ ,  $d \geq 1$  (some finite sample spaces are fixed).

## Definition

*Quality criterion* (for learning Bayesian networks) is a real function  $\mathcal{Q}$  on  $\text{DAGS}(N) \times \text{DATA}(N, d)$ .

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

# Quality criteria

This talk deals with methods for *learning Bayesian networks* based on the *maximization of a quality criterion*.

- Let  $N$  be a finite non-empty set of variables,
- $\text{DAGS}(N)$  will denote the class of acyclic directed graphs over the set of nodes  $N$ ,
- $\text{DATA}(N, d)$  will denote the set of all databases over  $N$  of the length  $d$ ,  $d \geq 1$  (some finite sample spaces are fixed).

## Definition

*Quality criterion* (for learning Bayesian networks) is a real function  $Q$  on  $\text{DAGS}(N) \times \text{DATA}(N, d)$ .

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

# Quality criteria

This talk deals with methods for *learning Bayesian networks* based on the *maximization of a quality criterion*.

- Let  $N$  be a finite non-empty set of variables,
- $\text{DAGS}(N)$  will denote the class of acyclic directed graphs over the set of nodes  $N$ ,
- $\text{DATA}(N, d)$  will denote the set of all databases over  $N$  of the length  $d$ ,  $d \geq 1$  (some finite sample spaces are fixed).

## Definition

*Quality criterion* (for learning Bayesian networks) is a real function  $\mathcal{Q}$  on  $\text{DAGS}(N) \times \text{DATA}(N, d)$ .

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

# Quality criteria

This talk deals with methods for *learning Bayesian networks* based on the *maximization of a quality criterion*.

- Let  $N$  be a finite non-empty set of variables,
- $\text{DAGS}(N)$  will denote the class of acyclic directed graphs over the set of nodes  $N$ ,
- $\text{DATA}(N, d)$  will denote the set of all databases over  $N$  of the length  $d$ ,  $d \geq 1$  (some finite sample spaces are fixed).

## Definition

*Quality criterion* (for learning Bayesian networks) is a real function  $\mathcal{Q}$  on  $\text{DAGS}(N) \times \text{DATA}(N, d)$ .

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

# Equivalence of graphs

One statistical model can be described by different graphs. Two graphs  $G, H \in \text{DAGS}(N)$  are *Markov equivalent* if they define the same class of Markovian distributions.

## Definition

Two graphs  $G, H \in \text{DAGS}(N)$  are *independence equivalent* if they define the same collection of CI restrictions. Let us write  $G \approx H$  then.

There exists a graphical characterization of independence equivalence (Verma and Pearl 1991).

# Equivalence of graphs

One statistical model can be described by different graphs. Two graphs  $G, H \in \text{DAGS}(N)$  are *Markov equivalent* if they define the same class of Markovian distributions.

## Definition

Two graphs  $G, H \in \text{DAGS}(N)$  are *independence equivalent* if they define the same collection of CI restrictions. Let us write  $G \approx H$  then.

There exists a graphical characterization of independence equivalence (Verma and Pearl 1991).

# Equivalence of graphs

One statistical model can be described by different graphs. Two graphs  $G, H \in \text{DAGS}(N)$  are *Markov equivalent* if they define the same class of Markovian distributions.

## Definition

Two graphs  $G, H \in \text{DAGS}(N)$  are *independence equivalent* if they define the same collection of CI restrictions. Let us write  $G \approx H$  then.

There exists a graphical characterization of independence equivalence (Verma and Pearl 1991).

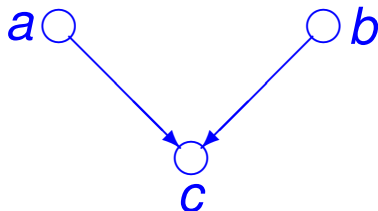


# Graphical characterization of equivalence

## Theorem

$G \approx H$  iff they have the same underlying graph and immoralities.

An *immorality* in a chain graph is its induced subgraph of this form:



# Score-equivalent criteria

## Definition

A quality criterion  $Q$  is *score-equivalent* iff

$\forall G, H \in \text{DAGS}(N), \forall D \in \text{DATA}(N, d)$

whenever  $G \approx H$  then  $Q(G, D) = Q(H, D)$ .

Most of the criteria used in practice are score-equivalent:

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

# Score-equivalent criteria

## Definition

A quality criterion  $Q$  is *score-equivalent* iff

$\forall G, H \in \text{DAGS}(N), \forall D \in \text{DATA}(N, d)$

whenever  $G \approx H$  then  $Q(G, D) = Q(H, D)$ .

Most of the criteria used in practice are score-equivalent:

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

# Decomposable criteria

## Definition

A quality criterion  $Q$  is *decomposable* if there exists a collection of functions  $q_{i|B} : \text{DATA}(B \cup \{i\}, d) \rightarrow \mathbf{R}$ ,  $i \in N$ ,  $B \subseteq N \setminus \{i\}$  such that  $\forall G \in \text{DAGS}(N), \forall D \in \text{DATA}(N, d)$

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

where  $D_A$  denotes the restriction of a database  $D$  for  $A \subseteq N$ .

This technical requirement was brought by researchers in computer science in connection with the method of local search.

All criteria used in practice are (strongly) decomposable.

# Decomposable criteria

## Definition

A quality criterion  $Q$  is *decomposable* if there exists a collection of functions  $q_{i|B} : \text{DATA}(B \cup \{i\}, d) \rightarrow \mathbf{R}$ ,  $i \in N$ ,  $B \subseteq N \setminus \{i\}$  such that  $\forall G \in \text{DAGS}(N), \forall D \in \text{DATA}(N, d)$

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

where  $D_A$  denotes the restriction of a database  $D$  for  $A \subseteq N$ .

This technical requirement was brought by researchers in computer science in connection with the method of local search.

All criteria used in practice are (strongly) decomposable.

# Problem of representative choice

How to represent a Bayesian network model in the memory of a computer?

There are two basic methods:

- represent it by arbitrary  $G \in \text{DAGS}(N)$  in the respective independence equivalence class,
- represent it by a special unique representative.

The most popular graphical representative is the **essential graph**. It is a chain graph obtained from the equivalence class  $\mathcal{G}$  by a special construction.

Later, I will mention an alternative algebraic representative, called the *standard imset*.

# Problem of representative choice

How to represent a Bayesian network model in the memory of a computer?

There are two basic methods:

- represent it by arbitrary  $G \in \text{DAGS}(N)$  in the respective independence equivalence class,
- represent it by a special unique representative.

The most popular graphical representative is the **essential graph**. It is a chain graph obtained from the equivalence class  $\mathcal{G}$  by a special construction.

Later, I will mention an alternative algebraic representative, called the *standard imset*.

# Problem of representative choice

How to represent a Bayesian network model in the memory of a computer?

There are two basic methods:

- represent it by arbitrary  $G \in \text{DAGS}(N)$  in the respective independence equivalence class,
- represent it by a special unique representative.

The most popular graphical representative is the **essential graph**. It is a chain graph obtained from the equivalence class  $\mathcal{G}$  by a special construction.

Later, I will mention an alternative algebraic representative, called the *standard imset*.



# Problem of representative choice

How to represent a Bayesian network model in the memory of a computer?

There are two basic methods:

- represent it by arbitrary  $G \in \text{DAGS}(N)$  in the respective independence equivalence class,
- represent it by a special unique representative.

The most popular graphical representative is the **essential graph**. It is a chain graph obtained from the equivalence class  $\mathcal{G}$  by a special construction.

Later, I will mention an alternative algebraic representative, called the *standard imset*.

# Problem of representative choice

How to represent a Bayesian network model in the memory of a computer?

There are two basic methods:

- represent it by arbitrary  $G \in \text{DAGS}(N)$  in the respective independence equivalence class,
- represent it by a special unique representative.

The most popular graphical representative is the **essential graph**. It is a chain graph obtained from the equivalence class  $\mathcal{G}$  by a special construction.

Later, I will mention an alternative algebraic representative, called the *standard imset*.

# Local search methods

Direct maximization of a quality criterion is typically infeasible. To avoid this problem various heuristic *local search methods* were developed.

The basic idea is that one introduces a *neighborhood structure* in the set  $\text{DAGS}(N)$ , respectively in the set  $\text{DAGS}(N)/\approx$ . Instead of the global maximum of  $Q$  one is trying to find a *local maximum* with respect to that neighborhood structure.

Every graph  $G$  (respectively equivalence class  $\mathcal{G}$ ) is assigned a relatively small set of neighboring graphs (respectively equivalence classes)  $nei(G)$ . They typically differ in the presence of one edge.

The point is that, for a decomposable criterion  $Q$ , the difference  $Q(G, D) - Q(H, D)$  for neighboring graphs  $G, H \in \text{DAGS}(N)$  is easy to compute.

# Local search methods

Direct maximization of a quality criterion is typically infeasible. To avoid this problem various heuristic *local search methods* were developed.

The basic idea is that one introduces a *neighborhood structure* in the set DAGS ( $N$ ), respectively in the set DAGS ( $N$ )/ $\approx$ . Instead of the global maximum of  $Q$  one is trying to find a **local maximum** with respect to that neighborhood structure.

Every graph  $G$  (respectively equivalence class  $\mathcal{G}$ ) is assigned a relatively small set of neighboring graphs (respectively equivalence classes)  $nei(G)$ . They typically differ in the presence of one edge.

The point is that, for a decomposable criterion  $Q$ , the difference  $Q(G, D) - Q(H, D)$  for neighboring graphs  $G, H \in \text{DAGS}(N)$  is easy to compute.

## Local search methods

Direct maximization of a quality criterion is typically infeasible. To avoid this problem various heuristic *local search methods* were developed.

The basic idea is that one introduces a *neighborhood structure* in the set  $\text{DAGS}(N)$ , respectively in the set  $\text{DAGS}(N)/\approx$ . Instead of the global maximum of  $Q$  one is trying to find a **local maximum** with respect to that neighborhood structure.

Every graph  $G$  (respectively equivalence class  $\mathcal{G}$ ) is assigned a relatively small set of neighboring graphs (respectively equivalence classes)  $nei(G)$ . They typically differ in the presence of one edge.

The point is that, for a decomposable criterion  $Q$ , the difference  $Q(G, D) - Q(H, D)$  for neighboring graphs  $G, H \in \text{DAGS}(N)$  is easy to compute.

## Local search methods

Direct maximization of a quality criterion is typically infeasible. To avoid this problem various heuristic *local search methods* were developed.

The basic idea is that one introduces a *neighborhood structure* in the set  $\text{DAGS}(N)$ , respectively in the set  $\text{DAGS}(N)/\approx$ . Instead of the global maximum of  $Q$  one is trying to find a **local maximum** with respect to that neighborhood structure.

Every graph  $G$  (respectively equivalence class  $\mathcal{G}$ ) is assigned a relatively small set of neighboring graphs (respectively equivalence classes)  $nei(G)$ . They typically differ in the presence of one edge.

The point is that, for a decomposable criterion  $Q$ , the difference  $Q(G, D) - Q(H, D)$  for neighboring graphs  $G, H \in \text{DAGS}(N)$  is easy to compute.

# Inclusion neighborhood

Is there a natural neighborhood structure for DAGS  $(N)/\approx$ ?

Let  $\mathcal{I}(G)$  denote the collection of CI restrictions given by  $G \in \text{DAGS}(N)$ . Given  $K, L \in \text{DAGS}(N)$ ,  $\mathcal{I}(K) \subset \mathcal{I}(L)$  means  $\mathcal{I}(K) \subseteq \mathcal{I}(L)$  but  $\mathcal{I}(K) \neq \mathcal{I}(L)$ .

If  $\mathcal{I}(K) \subset \mathcal{I}(L)$  and there is no  $G \in \text{DAGS}(N)$  such that  $\mathcal{I}(K) \subset \mathcal{I}(G) \subset \mathcal{I}(L)$  then we will say that  $\mathcal{I}(L)$  is an *upper inclusion neighbor* of  $\mathcal{I}(K)$ , respectively  $\mathcal{I}(K)$  is a *lower inclusion neighbor* of  $\mathcal{I}(L)$ . We will then write  $\mathcal{I}(K) \sqsubset \mathcal{I}(L)$ .

The inclusion neighborhood was characterized in graphical terms.

There are some arguments why general neighborhood structure in a local search method should include the inclusion neighborhood.

# Inclusion neighborhood

Is there a natural neighborhood structure for  $\text{DAGS}(N)/\approx$ ?

Let  $\mathcal{I}(G)$  denote the collection of CI restrictions given by  $G \in \text{DAGS}(N)$ . Given  $K, L \in \text{DAGS}(N)$ ,  $\mathcal{I}(K) \subset \mathcal{I}(L)$  means  $\mathcal{I}(K) \subseteq \mathcal{I}(L)$  but  $\mathcal{I}(K) \neq \mathcal{I}(L)$ .

If  $\mathcal{I}(K) \subset \mathcal{I}(L)$  and there is no  $G \in \text{DAGS}(N)$  such that  $\mathcal{I}(K) \subset \mathcal{I}(G) \subset \mathcal{I}(L)$  then we will say that  $\mathcal{I}(L)$  is an *upper inclusion neighbor* of  $\mathcal{I}(K)$ , respectively  $\mathcal{I}(K)$  is a *lower inclusion neighbor* of  $\mathcal{I}(L)$ . We will then write  $\mathcal{I}(K) \sqsubset \mathcal{I}(L)$ .

The inclusion neighborhood was characterized in graphical terms.

There are some arguments why general neighborhood structure in a local search method should include the inclusion neighborhood.



# Inclusion neighborhood

Is there a natural neighborhood structure for  $\text{DAGS}(N)/\approx$ ?

Let  $\mathcal{I}(G)$  denote the collection of CI restrictions given by  $G \in \text{DAGS}(N)$ . Given  $K, L \in \text{DAGS}(N)$ ,  $\mathcal{I}(K) \subset \mathcal{I}(L)$  means  $\mathcal{I}(K) \subseteq \mathcal{I}(L)$  but  $\mathcal{I}(K) \neq \mathcal{I}(L)$ .

If  $\mathcal{I}(K) \subset \mathcal{I}(L)$  and there is no  $G \in \text{DAGS}(N)$  such that  $\mathcal{I}(K) \subset \mathcal{I}(G) \subset \mathcal{I}(L)$  then we will say that  $\mathcal{I}(L)$  is an *upper inclusion neighbor* of  $\mathcal{I}(K)$ , respectively  $\mathcal{I}(K)$  is a *lower inclusion neighbor* of  $\mathcal{I}(L)$ . We will then write  $\mathcal{I}(K) \sqsubset \mathcal{I}(L)$ .

The inclusion neighborhood was characterized in graphical terms.

There are some arguments why general neighborhood structure in a local search method should include the inclusion neighborhood.

# Inclusion neighborhood

Is there a natural neighborhood structure for  $\text{DAGS}(N)/\approx$ ?

Let  $\mathcal{I}(G)$  denote the collection of CI restrictions given by  $G \in \text{DAGS}(N)$ . Given  $K, L \in \text{DAGS}(N)$ ,  $\mathcal{I}(K) \subset \mathcal{I}(L)$  means  $\mathcal{I}(K) \subseteq \mathcal{I}(L)$  but  $\mathcal{I}(K) \neq \mathcal{I}(L)$ .

If  $\mathcal{I}(K) \subset \mathcal{I}(L)$  and there is no  $G \in \text{DAGS}(N)$  such that  $\mathcal{I}(K) \subset \mathcal{I}(G) \subset \mathcal{I}(L)$  then we will say that  $\mathcal{I}(L)$  is an *upper inclusion neighbor* of  $\mathcal{I}(K)$ , respectively  $\mathcal{I}(K)$  is a *lower inclusion neighbor* of  $\mathcal{I}(L)$ . We will then write  $\mathcal{I}(K) \sqsubset \mathcal{I}(L)$ .

The inclusion neighborhood was characterized in graphical terms.

There are some arguments why general neighborhood structure in a local search method should include the inclusion neighborhood.

# Algebraic approach

The basic idea is to describe a Bayesian network model by a certain integral vector. This is motivated by a more general algebraic method for describing probabilistic CI structures.

## Definition

An *imset over  $N$*  is an integer-valued function on the power set of  $N$ .

Given  $A \subseteq N$ ,  $\delta_A$  is the identifier of the set  $A$ .

## Definition

Given CI statement  $a \perp\!\!\!\perp b \mid C$ ,  $a, b \in N$ ,  $a \neq b$ ,  $C \subseteq N \setminus \{a, b\}$  the respective *elementary imset* has the form

$$U_{\langle a, b \mid C \rangle} = \delta_{\{a, b\} \cup C} + \delta_C - \delta_{\{a\} \cup C} - \delta_{\{b\} \cup C}.$$

# Algebraic approach

The basic idea is to describe a Bayesian network model by a certain integral vector. This is motivated by a more general algebraic method for describing probabilistic CI structures.

## Definition

An *imset over  $N$*  is an integer-valued function on the power set of  $N$ .

Given  $A \subseteq N$ ,  $\delta_A$  is the identifier of the set  $A$ .

## Definition

Given CI statement  $a \perp\!\!\!\perp b \mid C$ ,  $a, b \in N$ ,  $a \neq b$ ,  $C \subseteq N \setminus \{a, b\}$  the respective *elementary imset* has the form

$$U_{\langle a,b \mid C \rangle} = \delta_{\{a,b\} \cup C} + \delta_C - \delta_{\{a\} \cup C} - \delta_{\{b\} \cup C}.$$

# Algebraic approach

The basic idea is to describe a Bayesian network model by a certain integral vector. This is motivated by a more general algebraic method for describing probabilistic CI structures.

## Definition

An *imset over  $N$*  is an integer-valued function on the power set of  $N$ .

Given  $A \subseteq N$ ,  $\delta_A$  is the identifier of the set  $A$ .

## Definition

Given CI statement  $a \perp\!\!\!\perp b \mid C$ ,  $a, b \in N$ ,  $a \neq b$ ,  $C \subseteq N \setminus \{a, b\}$  the respective *elementary imset* has the form

$$u_{\langle a,b|C \rangle} = \delta_{\{a,b\} \cup C} + \delta_C - \delta_{\{a\} \cup C} - \delta_{\{b\} \cup C}.$$

# Standard imsets

## Definition

Let  $G \in \text{DAGS}(N)$ . Then the respective *standard imset* has the form

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

Note that every standard imset is a *structural imset*, that is, a combination of elementary imsets with non-negative rational coefficients.

In particular, it describes a certain special probabilistic CI structure: the one that corresponds to the respective Bayesian network model.

As the standard imset has many 'zeros' it can be easily kept in the memory of a computer.

# Standard imsets

## Definition

Let  $G \in \text{DAGS}(N)$ . Then the respective *standard imset* has the form

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

Note that every standard imset is a *structural imset*, that is, a combination of elementary imsets with non-negative rational coefficients.

In particular, it describes a certain special probabilistic CI structure: the one that corresponds to the respective Bayesian network model.

As the standard imset has many 'zeros' it can be easily kept in the memory of a computer.

# Standard imsets

## Definition

Let  $G \in \text{DAGS}(N)$ . Then the respective *standard imset* has the form

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

Note that every standard imset is a *structural imset*, that is, a combination of elementary imsets with non-negative rational coefficients.

In particular, it describes a certain special probabilistic CI structure: the one that corresponds to the respective Bayesian network model.

As the standard imset has many ‘zeros’ it can be easily kept in the memory of a computer.



## Some basic results on standard imsets

### Lemma

*Given  $G, H \in \text{DAGS}(N)$  one has  $G \approx H$  iff  $u_G = u_H$ .*

Thus, the standard imset can serve as a unique representative of the respective Bayesian network model.

### Lemma

*Given  $K, L \in \text{DAGS}(N)$  one has  $\mathcal{I}(K) \sqsubset \mathcal{I}(L)$  iff  $u_L - u_K$  is an elementary imset.*

## Some basic results on standard imsets

### Lemma

*Given  $G, H \in \text{DAGS}(N)$  one has  $G \approx H$  iff  $u_G = u_H$ .*

Thus, the standard imset can serve as a unique representative of the respective Bayesian network model.

### Lemma

*Given  $K, L \in \text{DAGS}(N)$  one has  $\mathcal{I}(K) \sqsubset \mathcal{I}(L)$  iff  $u_L - u_K$  is an elementary imset.*

## Some basic results on standard imsets

### Lemma

*Given  $G, H \in \text{DAGS}(N)$  one has  $G \approx H$  iff  $u_G = u_H$ .*

Thus, the standard imset can serve as a unique representative of the respective Bayesian network model.

### Lemma

*Given  $K, L \in \text{DAGS}(N)$  one has  $\mathcal{I}(K) \sqsubset \mathcal{I}(L)$  iff  $u_L - u_K$  is an elementary imset.*

# Standard imsets and quality criteria

## Theorem

Let  $\mathcal{Q}$  be a *score-equivalent decomposable criterion*. Then it has the following form:  $\forall G \in \text{DAGS}(N), \forall D \in \text{DATA}(N, d)$

$$\begin{aligned} \mathcal{Q}(G, D) &= k_{\mathcal{Q}}(D) + \sum_{S \subseteq N} u_G(S) \cdot t_D^{\mathcal{Q}}(S) \\ &= k_{\mathcal{Q}}(D) + \langle u_G, t_D^{\mathcal{Q}} \rangle. \end{aligned}$$

where  $t_D^{\mathcal{Q}}$  is a real vector representing the database  $D$  (relative to  $\mathcal{Q}$ ) and  $k_{\mathcal{Q}}(D)$  is a constant (depending on data). In particular, for any pair of graphs  $K$  and  $L$ ,  $\mathcal{Q}(L, D) - \mathcal{Q}(K, D) = \langle u_L - u_K, t_D^{\mathcal{Q}} \rangle$ .

Thus, from purely mathematical point of view, the maximization of a quality criterion leads to the *problem of maximization of a (shifted) linear function*.

# Standard imsets and quality criteria

## Theorem

Let  $\mathcal{Q}$  be a *score-equivalent decomposable criterion*. Then it has the following form:  $\forall G \in \text{DAGS}(N), \forall D \in \text{DATA}(N, d)$

$$\begin{aligned} \mathcal{Q}(G, D) &= k_{\mathcal{Q}}(D) + \sum_{S \subseteq N} u_G(S) \cdot t_D^{\mathcal{Q}}(S) \\ &= k_{\mathcal{Q}}(D) + \langle u_G, t_D^{\mathcal{Q}} \rangle. \end{aligned}$$

where  $t_D^{\mathcal{Q}}$  is a real vector representing the database  $D$  (relative to  $\mathcal{Q}$ ) and  $k_{\mathcal{Q}}(D)$  is a constant (depending on data). In particular, for any pair of graphs  $K$  and  $L$ ,  $\mathcal{Q}(L, D) - \mathcal{Q}(K, D) = \langle u_L - u_K, t_D^{\mathcal{Q}} \rangle$ .

Thus, from purely mathematical point of view, the maximization of a quality criterion leads to the *problem of maximization of a (shifted) linear function*.

# Standard imsets and quality criteria

## Theorem

Let  $\mathcal{Q}$  be a *score-equivalent decomposable criterion*. Then it has the following form:  $\forall G \in \text{DAGS}(N), \forall D \in \text{DATA}(N, d)$

$$\begin{aligned} \mathcal{Q}(G, D) &= k_{\mathcal{Q}}(D) + \sum_{S \subseteq N} u_G(S) \cdot t_D^{\mathcal{Q}}(S) \\ &= k_{\mathcal{Q}}(D) + \langle u_G, t_D^{\mathcal{Q}} \rangle. \end{aligned}$$

where  $t_D^{\mathcal{Q}}$  is a real vector representing the database  $D$  (relative to  $\mathcal{Q}$ ) and  $k_{\mathcal{Q}}(D)$  is a constant (depending on data). In particular, for any pair of graphs  $K$  and  $L$ ,  $\mathcal{Q}(L, D) - \mathcal{Q}(K, D) = \langle u_L - u_K, t_D^{\mathcal{Q}} \rangle$ .

Thus, from purely mathematical point of view, the maximization of a quality criterion leads to the *problem of maximization of a (shifted) linear function*.

# Discussion

To utilize fully the algebraic approach to the local search methods one has to be able to describe the inclusion neighborhood in terms of the standard imset.

There exists characterization of inclusion neighborhood of a given  $\mathcal{G} \in \text{DAGS}(N)/\approx$  in terms of the respective essential graph.

For this reason, it is desirable to translate graphical representatives into algebraic ones and conversely. There exists a formula which gives the standard imset on the basis of the essential graph and a two-stage inverse algorithm (Vomlel Studený 2004).

The middle stage of that procedure is a certain sequence of variable sets, which can perhaps be visualizes in the form of a *hierarchical junction tree* (Puch, Smith and Bielza 2003).

## Discussion

To utilize fully the algebraic approach to the local search methods one has to be able to describe the inclusion neighborhood in terms of the standard imset.

There exists characterization of inclusion neighborhood of a given  $\mathcal{G} \in \text{DAGS}(N)/\approx$  **in terms of the respective essential graph**.

For this reason, it is desirable to translate graphical representatives into algebraic ones and conversely. There exists a **formula** which gives the standard imset on the basis of the essential graph and a **two-stage inverse algorithm** (Vomlel Studený 2004).

The middle stage of that procedure is a certain sequence of variable sets, which can perhaps be visualizes in the form of a *hierarchical junction tree* (Puch, Smith and Bielza 2003).



## Discussion

To utilize fully the algebraic approach to the local search methods one has to be able to describe the inclusion neighborhood in terms of the standard imset.

There exists characterization of inclusion neighborhood of a given  $\mathcal{G} \in \text{DAGS}(N)/\approx$  in terms of the respective essential graph.

For this reason, it is desirable to translate graphical representatives into algebraic ones and conversely. There exists a formula which gives the standard imset on the basis of the essential graph and a two-stage inverse algorithm (Vomlel Studený 2004).

The middle stage of that procedure is a certain sequence of variable sets, which can perhaps be visualizes in the form of a *hierarchical junction tree* (Puch, Smith and Bielza 2003).

## Discussion

To utilize fully the algebraic approach to the local search methods one has to be able to describe the inclusion neighborhood in terms of the standard imset.

There exists characterization of inclusion neighborhood of a given  $\mathcal{G} \in \text{DAGS}(N)/\approx$  in terms of the respective essential graph.








For this reason, it is desirable to translate graphical representatives into algebraic ones and conversely. There exists a formula which gives the standard imset on the basis of the essential graph and a two-stage inverse algorithm (Vomlel Studený 2004).

The middle stage of that procedure is a certain sequence of variable sets, which can perhaps be visualizes in the form of a *hierarchical junction tree* (Puch, Smith and Bielza 2003).

That's all.

Thank you for your attention!

# Some relevant literature

-  S.A. Andersson, D. Madigan and M.D. Perlman (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics* **25**: 505-541.
-  R.R. Bouckaert and M. Studený (2005). Racing for conditional independence inference. In *ECSQARU 2005* (L. Godo ed.), Lecture Notes in AI 3571, Springer-Verlag, 221-232.
-  D.M. Chickering (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research* **3**: 507-554.
-  R.G. Cowell, A.P. Dawid, S.L. Lauritzen and D.J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. New York: Springer-Verlag
-  R.O. Puch, J.Q. Smith and C. Bielza (2004). Hierarchical junction trees, conditional independence preservation and forecasting in dynamic Bayesian networks with heterogeneous evolution. In *Advances in Bayesian Networks* (J.A. Gámez, S. Moral and A. Salmerón eds.) Springer-Verlag: 57-75.
-  M. Studený and J. Vomlel (2004). Transition between graphical and algebraic representatives of Bayesian network models. In *Proceeding of PGM'04*, Leiden (P. Lucas ed.), 193-200.
-  M. Studený (2005). *Probabilistic Conditional Independence Structures*. London: Springer-Verlag.
-  T. Verma and J. Pearl (1991). Equivalence and synthesis of causal models, in *Uncertainty in AI 6*, Elsevier: 220-227.