

# On Bayesian Criteria for Learning Bayesian Network Structure

Milan Studený

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

Durham, UK, July 1, 2008, 9:30-10:20

# Summary of the talk

- 1 Introduction
- 2 Basic concepts
- 3 The idea of an algebraic approach
- 4 Parameterization of a discrete BN model (revision lesson)
- 5 Bayesian terminology (a course for M.S.)
- 6 Bayesian approach to the derivation of quality criteria
- 7 Convex geometry (a course for M.S.)
- 8 Geometric view on standard imsets: an example of GES failure
- 9 Open questions

# Introduction: Bayesian networks

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

By a *quality criterion*, also named a *score metric* or a *score*, is meant a special real function  $Q$  of the BN structure, usually represented by a graph  $G$ , and of the database  $D$ . The value  $Q(G, D)$  should quantify how the BN structure given by  $G$  fits the database  $D$ .

There are two important technical requirements on a quality criterion  $Q$  brought in connection with the maximization problem. One of them is that  $Q$  should be *score equivalent* (Bouckaert 1995), the other is that  $Q$  should be *decomposable* (Chickering 2002).

## Introduction: algebraic approach

The basic idea of an algebraic approach to learning BN structure (Studený 2005) is to represent both the BN structure and the database by a real vector.

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 every score equivalent and decomposable criterion  $\mathcal{Q}$  is an affine function (= linear function plus a constant) of the standard imset. More specifically, one has

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

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

# Introduction: Bayesian criteria

There are different methodological approaches to the derivation of quality criteria (Cowell *et. al.* 1999; chapter 11):

- maximized likelihood → classic information criteria AIC, BIC
- predictive assessment → prequential validation (Dawid 1984)
- marginal likelihood → Bayesian approach

This talk deals with the Bayesian approach. A part of it is an attempt to reformulate (in mathematical terms) the assumption(s) on which the Bayesian approach is based (Heckerman *et. al.* 1995). The resulting criterion is the *logarithm of the marginal likelihood (LML)*, also named *BDe metric* (BDe = Bayesian Dirichlet equivalence).

This allows one to give a mathematical formula for the data vector relative to the LML criterion in terms of the hyper-potential for Dirichlet priors.

The formulas for the data vectors relative to AIC and BIC were derived earlier.

## Introduction: geometric view

Another aim of this talk is to emphasize the geometric interpretation: *the set of all standard imsets over a fixed set of variables  $N$  can be viewed as the set of points in the corresponding Euclidean space.*

A recent result (Studený Vomlel 2008) says that the set of standard imsets is *the set of vertices (= extreme points) of a certain polytope.*

Therefore, once one succeeds to describe the above mentioned polytope in the form of a (bounded) polyhedron, one gets a classic task of linear programming: **to maximize/minimize a linear function over a polyhedron.**

**This seems to be an interesting (and promising ?) research topic ...**

The idea of possible use of the *simplex method* (Schrijver 1986) motivated the concept of the *geometric neighborhood* for standard imsets. Its comparison with common *inclusion neighborhood* led to an example of the failure of the “standard” GES algorithm (Chickering 2002).

## Basic concepts: Bayesian network structure

One of possible definitions of a (discrete) *Bayesian network* is that it is a pair  $(G, P)$ , where  $G$  is an acyclic directed graph over a (non-empty finite) set of nodes (= variables)  $N$  and  $P$  a discrete probability distribution over  $N$  that is Markovian with respect to  $G$ . (Lauritzen 1996)

Having fixed individual (finite) sample spaces  $X_i$  for variables  $i \in N$ , the corresponding (BN) *statistical model* is the class of all positive probability distributions  $P$  on  $X_N \equiv \prod_{i \in N} X_i$  that are Markovian with respect to  $G$ .

To avoid trivial mistakes and silly omissions, throughout this talk we assume  $|X_i| \geq 2$  for every  $i \in N$ .

To name the shared features of distributions in this class one can use the phrase *BN structure*.

## Basic concepts: equivalence of graphs

It may happen that two different graphs over  $N$  describe the same BN structure.

Two acyclic directed graphs over  $N$  will be named *Markov equivalent* if they define the same BN statistical model.

If  $|X_i| \geq 2$  for every  $i \in N$  then this is equivalent to the condition they are *independence equivalent*.

Verma and Pearl (1991) gave classic graphical characterization of independence equivalence: two acyclic directed graphs  $G$  and  $H$  over  $N$  are independence equivalent iff they have the same *underlying undirected graph and immoralities*.



## Basic concepts: quality criterion

Data are assumed to have the form of a *complete database*  $D : x^1, \dots, x^d$  of the length  $d \geq 1$ , that is, of a sequence of elements of  $X_N$ . Statisticians may prefer the term a *sample of the size  $d$*  instead.

Provided the individual sample spaces  $X_i$  for  $i \in N$  are fixed let **DATA**  $(N, d)$  denote the set of all databases over  $N$  of the length  $d$ . Moreover, let **DAGS**  $(N)$  denote the collection of all acyclic directed graphs over  $N$ .

### Definition (quality criterion)

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

There are various methods to derive quality criteria. Most of them come from the idea of the BN statistical model as a parameterized class of distributions.

## Basic concepts: score equivalent criterion

Since the aim of the learning procedure is to get the BN structure it is quite natural to require that the quality criterion satisfies the following condition:

### Definition (score equivalent criterion)

A quality criterion  $Q$  will be named *score equivalent* if, for every  $D \in \text{DATA}(N, d)$ ,  $d \geq 1$ , one has

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

are independence equivalent.

Most quality criteria used in practice are score equivalent.

## Basic concepts: decomposable criterion

### Definition (decomposable criterion)

A criterion  $Q$  will be called *decomposable* if there exists a collection of functions  $q_{i|B} : \text{DATA}(\{i\} \cup B, d) \rightarrow \mathbb{R}$  where  $i \in N$ ,  $B \subseteq N \setminus \{i\}$ ,  $d \geq 1$  such that, for every  $G \in \text{DAGS}(N)$ ,  $D \in \text{DATA}(N, d)$  one has

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

where  $D_A : x_A^1, \dots, x_A^d$  denotes the projection of  $D$  to the marginal space  $X_A \equiv \prod_{i \in A} X_i$  for  $\emptyset \neq A \subseteq N$  and  $pa_G(i) \equiv \{j \in N; j \rightarrow i\}$  the set of *parents* of  $i \in N$ .

All criteria used in practice are decomposable.

## Algebraic approach: imset

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

### Definition (imset)

An *imset*  $u$  (over  $N$ ) is a function  $u : \mathcal{P}(N) \mapsto \mathbb{Z}$ .

We will regard an imset over  $N$  as a vector whose components are integers and are indexed by subsets of  $N$ .

Actually, any **real function**  $m : \mathcal{P}(N) \rightarrow \mathbb{R}$  will 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: elementary imset

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

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

### Definition (elementary imset)

By an *elementary imset* is meant an imset of the form

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

where  $C \subseteq N$  and  $a, b \in N \setminus C$  are distinct.

In our framework, this imset encodes an elementary conditional independence statement  $a \perp\!\!\!\perp b \mid C$ .

## Algebraic approach: standard imset

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

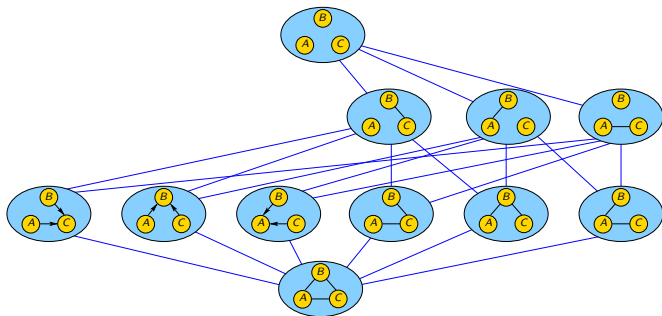
The standard imset is a uniquely determined representative of the Bayesian network structure.

Since every standard imset over  $N$  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  $|N|$* .

A common graphical representative is so-called *essential graph*.

## Example: the case of three variables

In the case of 3 variables one has 11 standard imsets and they break into 5 types.



- The zero imset corresponds to the complete graph.
- Six elementary imsets break into two types, namely  $u_{\langle a,b|\emptyset \rangle}$  and  $u_{\langle a,b|c \rangle}$ ; the respective essential graphs are  $a \rightarrow c \leftarrow b$  and  $a - c - b$ .
- Three “semi-elementary” imsets of the form  $u_{\langle a,bc|\emptyset \rangle} \equiv \delta_{abc} + \delta_\emptyset - \delta_a - \delta_{bc}$  define one type. The corresponding essential graph has just one edge.
- The imset  $\delta_N - \sum_{i \in N} \delta_i + 2 \cdot \delta_\emptyset$  corresponds to the empty essential graph.

## Parameterization of a discrete BN model: convention

For every  $A \subseteq N$ , we fix a total ordering of configurations for  $A$ .

### Definition (Notational convention)

- The symbol  $i$  will be a generic symbol for *variables*/nodes. Moreover,  $r(i)$  will denote  $|X_i|$ , that is, the *number of node configurations for  $i$*  and, given  $G \in \text{DAGS}(N)$ ,  $q(i, G)$  will denote  $|X_{pa_G(i)}|$ , that is, the *number of parent configurations for  $i$* .
- The symbol  $k$  will serve as a generic symbol for a *code of a node configuration*. Specifically, if  $y_i^1, y_i^2, \dots, y_i^{r(i)}$  is the fixed ordering of  $X_i$ , then  $k \in \{1, \dots, r(i)\}$  encodes the  $k$ -th node configuration  $y_i^k$ .
- The symbol  $j$  will serve as a generic symbol for a *code of a parent configuration*. If  $z_i^1, \dots, z_i^{q(i, G)}$  is the fixed ordering of  $X_{pa(i)}$ , then  $j \in \{1, \dots, q(i, G)\}$  encodes the  $j$ -th parent configuration  $z_i^j$ .



## Parameterization of a BN model: parameter space

The elementary parameters in the parameterization correspond to triplets:

$$\left[ \underbrace{\text{variable} = \text{node}}_i, \underbrace{\text{parent configuration}}_j, \underbrace{\text{node configuration}}_k \right].$$

This is concordance with the mentioned notation convention:

$$\theta_{ijk}, \quad i \in N, \quad j = 1, \dots, q(i, G), \quad k = 1, \dots, r(i).$$

### Definition (Parameter space)

Given  $G \in \text{DAGS}(N)$ , the parameter space is

$$\Theta_G = \prod_{i \in N} \prod_{j=1}^{q(i, G)} \Theta_{(ij)} \quad \text{where } \Theta_{(ij)} = \{[\theta_{ijk}]_{k=1}^{r(i)}; \theta_{ijk} > 0, \sum_{k=1}^{r(i)} \theta_{ijk} = 1\}.$$

## Parameterization of a BN model: formula for density

The interpretation of the elementary parameter  $\theta_{ijk}$  is the value of the conditional probability  $p_{i|pa_G(i)}^\theta$  for the node configuration encoded by  $k$  given the parent configuration encoded by  $j$ .

### Definition (Formula for the theoretical distribution)

Given  $G \in \text{DAGS}(N)$ , and a parameter vector  $\theta \in \Theta_G$  the density of the corresponding Markovian distribution is

$$p^\theta(x) = \prod_{i \in N} \theta_{i j(i,x) k(i,x)}$$

where  $j(i, x)$  denotes the code of  $x_{pa_G(i)}$  and  $k(i, x)$  the code of  $x_i$ .

# Parameterization of a BN model: well-known facts

## Theorem (Parameterization)

*The mapping  $\theta \mapsto p^\theta$  is a one-to-one mapping from  $\Theta_G$  onto the class of strictly positive distributions on  $X_N$  that are Markovian with respect to  $G$ .*

Another important fact is that  $\{p^\theta; \theta \in \Theta_G\}$  is an exponential family:

$$p^\theta(x) = c(\theta) \cdot u(x) \cdot \exp\left(\sum_{s=1}^m q_s(\theta) \cdot t_s(x)\right),$$

where

- $m = \sum_{i \in N} \sum_{j=1}^{q(i,G)} \sum_{k=1}^{r(i)} 1 = \sum_{i \in N} r(i) \cdot q(i, G)$ ,
- $c(\theta) = 1$ ,  $u(x) = 1$ ,
- $q_s(\theta) = \ln \theta_{ijk}$  for  $(ijk) \sim s$ ,
- $t_s(x) = \begin{cases} 1 & \text{if } x_{pa_G(i)} = z_i^j \text{ and } x_i = y_i^k, \\ 0 & \text{otherwise,} \end{cases}$  for  $(ijk) \sim s$ .

# Parameterization of a BN model: likelihood function

## Definition (Database convention)

Given  $D \in \text{DATA}(N, d)$ , we introduce notation (for marginal counts):

$$d_{ijk} = |\{\ell; 1 \leq \ell \leq d, (x^\ell)_{pa_G(i)} = z_i^j \ \& \ (x^\ell)_i = y_i^k \}|$$

$$d_{ij} = |\{\ell; 1 \leq \ell \leq d, (x^\ell)_{pa_G(i)} = z_i^j \}| \equiv \sum_{k=1}^{r(i)} d_{ijk}.$$

## Theorem (Formula for the likelihood)

$\forall \theta \in \Theta_G, \forall D \in \text{DATA}(N, d)$

$$L(\theta, D) = \prod_{i \in N} \prod_{j=1}^{q(i,G)} \prod_{k=1}^{r(i)} (\theta_{ijk})^{d_{ijk}}.$$

## Bayesian terminology: basic input components

- $(\mathbb{X}, \mathcal{X}) \dots$  *sample space* here  $\mathbb{X} = \underbrace{X_N \times \dots \times X_N}_{d \text{ times}} = \text{DATA}(N, d)$
- $(\Theta, \mathcal{A}) \dots$  *parameter space* here  $\Theta \equiv \Theta_G$
- $\{P_\theta; \theta \in \Theta\} \dots$  *sampling probabilities*  
typically given by densities  $p(x|\theta)$  w.r.t. a “standard” dominating measure on  $(\mathbb{X}, \mathcal{X})$   
here  $p(x|\theta) = L(\theta, D)$  is the above mentioned likelihood function

What is specific in the Bayesian approach is this:

- $\pi(\theta) \dots$  *prior density*  
w.r.t. a “standard” ( $\sigma$ -finite) dominating measure on  $(\Theta, \mathcal{A})$   
here  $\pi(\theta)$  will be a product of Dirichlet distributions

These components establish a *Bayesian experiment*.

## Bayesian terminology: output components

From mathematical point of view, the Bayesian experiment is characterized by the (joint) distribution  $\Pi$  on  $(\Theta \times \mathbb{X}, \mathcal{A} \times \mathcal{X})$  with the density

$$\Pi(\boldsymbol{\theta}, x) = \pi(\boldsymbol{\theta}) \cdot p(x|\boldsymbol{\theta}) \quad \text{for } \boldsymbol{\theta} \in \Theta, x \in \mathbb{X}.$$

The point is that  $\Pi$  has a dual decomposition:

$$\Pi(\boldsymbol{\theta}, x) = p(x) \cdot \pi(\boldsymbol{\theta}|x) \quad \text{for } \boldsymbol{\theta} \in \Theta, x \in \mathbb{X},$$

where

- $p(x)$  is the marginal density of  $\Pi$  on  $\mathbb{X}$ , sometimes called the *predictive probability*. Here, it will be called the *marginal likelihood* for it is obtained by integrating the likelihood  $L(\boldsymbol{\theta}, D) \equiv p(x|\boldsymbol{\theta})$  after the prior  $\pi(\boldsymbol{\theta})$ .
- $\{\pi(\boldsymbol{\theta}|x); x \in \mathbb{X}\}$  is the system of *posterior densities*.

# Bayesian terminology: conjugate family

## Definition (Conjugate family)

A system  $\mathcal{S}$  of probability distributions on the parameter space  $(\Theta, \mathcal{A})$  will be called a *conjugate family* for a system of probability distributions  $\mathcal{T}$  on the sample space  $(\mathbb{X}, \mathcal{X})$  if the following condition holds:

*whenever the prior is in  $\mathcal{S}$  and sampling probabilities are in  $\mathcal{T}$ , then every posterior is in  $\mathcal{S}$ .*

Formally:

$$\pi \in \mathcal{S} \ \& \ [\forall \theta \in \Theta \ p(*|\theta) \in \mathcal{T}] \ \Rightarrow \ [\forall x \in \mathbb{X} \ \pi(*|x) \in \mathcal{S}].$$

## Dirichlet distribution: parameters and hyperparameters

Natural conjugate family for (the class of) discrete (strictly positive) sampling distributions is the class of Dirichlet distributions.

### Definition (Parameter space and hyperparameters)

Consider the parameter space for sampling distributions with  $r \geq 2$  outputs:

$$\Theta[r] = \{[\theta_k]_{k=1}^r; \theta_k > 0, \sum_{k=1}^r \theta_k = 1\}.$$

It will serve as the sample space for the class of Dirichlet distributions. The *(hyper)parameters* for these distributions then belong to the set:

$$(\Xi)[r] = \{[\alpha_k]_{k=1}^r; \alpha_k > 0\}.$$



## Dirichlet distribution: formula for the density

### Definition (Formula for density of Dirichlet distribution)

The formula for Dirichlet density on  $\Theta[r]$ ,  $r \geq 2$  is as follows:

$$\forall \boldsymbol{\alpha} \equiv [\alpha_k]_{k=1}^r, \alpha_k > 0 \quad f([\theta_1, \dots, \theta_r]) = \frac{\Gamma(\sum_{k=1}^r \alpha_k)}{\prod_{k=1}^r \Gamma(\alpha_k)} \cdot \prod_{k=1}^r (\theta_k)^{\alpha_k - 1},$$

where  $\Gamma$  denotes the Gamma function  $\Gamma(\alpha) = \int_0^{+\infty} e^{-t} \cdot t^{\alpha-1} dt$  for  $\alpha > 0$ . Given a vector of hyperparameters  $\boldsymbol{\alpha} = [\alpha_k]_{k=1}^r \in (\Xi)[r]$  the corresponding Dirichlet distribution will be denoted by  $\mathcal{D}([\alpha_k]_{k=1}^r)$ .

**Warning** The density is *not* with respect to  $(r-1)$ -dimensional Lebesgue measure on the affine space  $\{[\theta_k]_{k=1}^r; \sum_{k=1}^r \theta_k = 1\}$ , but with respect to its  $\frac{1}{\sqrt{r}}$ -multiple! This is the image of the Lebesgue measure on  $\mathbb{R}^{r-1}$  by any “lifting”:

$$\forall i \in \{1, \dots, r\} \quad [\theta_k]_{k \in \{1, \dots, r\} \setminus \{i\}} \mapsto ([\theta_k]_{k \in \{1, \dots, r\} \setminus \{i\}}, \theta_i \equiv 1 - \sum_{k=1, k \neq i}^r \theta_k).$$

# Bayesian approach: summarized assumption

## Definition (Compatibility assumption)

There exists a (*hyper*)potential  $\alpha : X_N \rightarrow (0, +\infty)$  such that, for every  $G \in \text{DAGS}(N)$ , the prior on the parameter space  $\Theta_G$  (for the corresponding BN model) is determined as follows:

- Hyperparameters (for local Dirichlet priors) are given by marginalizing the potential  $\alpha$ :

$$\forall i \in N \quad \forall j = 1, \dots, q(i, G) \quad \forall k = 1, \dots, r(i) \\ \alpha_{ijk}^G = \sum \{ \alpha(x); x \in X_N, x_{i \cup pa_G(i)} = [y_i^k, z_i^j] \}$$

- Global prior is the product of these local Dirichlet priors:

$$\pi_\alpha^G = \prod_{i \in N} \prod_{j=1}^{q(i, G)} \mathcal{D}([\alpha_{ijk}^G]_{k=1}^{r(i)}).$$

## Bayesian approach: the definition of the criterion

### Definition (LML criterion)

Under the compatibility assumption the *LML criterion* corresponding to a hyper-potential  $\alpha : X_N \rightarrow (0, +\infty)$  is the *logarithm of the marginal likelihood*:

$$\text{LML}[\alpha](G, D) = \ln \int_{\Theta_G} L(\theta, D) d\pi_\alpha^G(\theta)$$

for every  $G \in \text{DAGS}(N)$ ,  $D \in \text{DATA}(N, d)$ ,  $d \geq 1$ .

There is a direct (closed-form) formula for this criterion in terms of hyperparameters (of local Dirichlet priors) and marginal counts:

$$\text{LML}[\alpha](G, D) = \sum_{i \in N} \sum_{j=1}^{q(i, G)} \left\{ \ln \frac{\Gamma(\alpha_{ij}^G)}{\Gamma(\alpha_{ij}^G + d_{ij})} - \sum_{k=1}^{r(i)} \ln \frac{\Gamma(\alpha_{ijk}^G)}{\Gamma(\alpha_{ijk}^G + d_{ijk})} \right\}$$

for every  $G \in \text{DAGS}(N)$ ,  $D \in \text{DATA}(N, d)$ ,  $d \geq 1$ .

## Bayesian approach: LML data vector definition

### Definition (Data vector for LML)

Let  $\alpha : X_N \rightarrow (0, +\infty)$  be a hyperpotential for priors and  $D : x^1, \dots, x^d$  a database of the length  $d \geq 1$  (= a sample of the size  $d$ ).

Given  $A \subseteq N$ , let  $\alpha_A$  denotes the *marginal potential* of  $\alpha$  for  $A$  and  $d_A$  the *marginal contingency table* corresponding to  $D$ :

$$\alpha_A(y) = \sum \{ \alpha(x) ; x \in X_N, x_A = y \} \quad \text{for } y \in X_A,$$

$$d_A(y) = |\{ \ell ; 1 \leq \ell \leq d, (x^\ell)_A = y \}| \quad \text{for } y \in X_A.$$

Then a (non-standardized) *data vector corresponding to  $\alpha$*  can be introduced as follows:

$$t_D^{\text{LML}[\alpha]}(A) = \sum_{y \in X_A} \ln \frac{\Gamma(\alpha_A(y) + d_A(y))}{\Gamma(\alpha_A(y))} \quad \text{for any } A \subseteq N.$$

## Bayesian approach: the required formula

### Theorem (Formula for the LML criterion)

*Under the compatibility assumption, the LML criterion is score equivalent and decomposable. Moreover, it can be expressed in the form*

$$\text{LML}[\alpha](G, D) = s_D^{\text{LML}[\alpha]} - \langle t_D^{\text{LML}[\alpha]}, u_G \rangle,$$

*where  $u_G$  is the standard imset for  $G$ ,  $t_D^{\text{LML}[\alpha]}$  the LML data vector introduced above and*

$$s_D^{\text{LML}[\alpha]} = t_D^{\text{LML}[\alpha]}(N) - t_D^{\text{LML}[\alpha]}(\emptyset).$$

**Remark** Given a criterion  $\mathcal{Q}$  and  $D \in \text{DATA}(N, d)$ ,  $d \geq 1$ , the corresponding data vector  $t_D^{\mathcal{Q}}$  is not uniquely determined. However, it is unique under additional standardization conditions.

In particular, the problem of maximization of LML is equivalent to the task to minimize a linear function  $u \mapsto \langle t_D^{\text{LML}[\alpha]}, u \rangle$  on the respective domain.

# 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.

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

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

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

## Theorem (Weyl-Minkowski theorem)

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

A further important observation is that if  $P$  is a full-dimensional polytope then its *irredundant description* in the form of a polyhedron is unique.

Provided that the polytope is *rational*, that is, it is the convex hull of a finite subset of  $\mathbb{Q}^K$ , the respective (irredundant) half-spaces are given by rational vectors and constants.

There is a geometric concept of a *face* of a polytope (*whose definition is omitted in this talk*). Faces can be classified by their dimension. Faces of dimension 0 are *vertices*. Important faces are (geometric) *edges*, faces of dimension 1. These are special line-segments connecting vertices.

# Geometric view: standard imsets are vertices of a polytope

## Theorem (Geometric view on standard imsets)

*The set  $S$  of standard imsets over  $N$  is the set of vertices of a rational polytope  $P \subseteq \mathbb{R}^{\mathcal{P}(N)}$ . The dimension of the polytope is  $2^{|N|} - |N| - 1$ .*

Now, recall that every score equivalent and decomposable criterion  $Q$  necessarily has the form:

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

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

The consequence is as follows: the task to maximize  $Q$  over BN structures (= standard imsets) is equivalent to the task to minimize a linear function  $u \mapsto \langle t_D^Q, u \rangle$  over the above mentioned polytope.



## Geometric view: neighborhood concept

Classic task of linear programming is to maximize/minimize a linear function over a polyhedron. There are efficient methods, like the *simplex method*, to tackle this problem (Schrijver 1986). One of possible interpretations of this method is that it is a kind of “greedy search” in which one moves between polyhedron’s vertices along its edges.

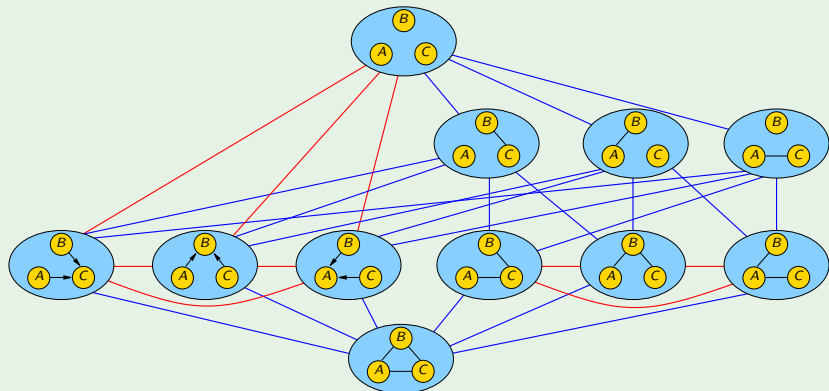
### Definition (geometric neighborhood)

We say that two standard imsets  $u, v \in S$  are *geometric neighbors* if the line-segment connecting them in  $\mathbb{R}^{\mathcal{P}(N)}$  is an edge of the polytope  $P$  (generated by the set  $S$  of standard imsets).

The geometric neighborhood for the case of 3 variables was characterized in (Studený Vomlel 2008). We found out it differs from the *inclusion neighborhood*, which was introduced in connection with common ML methods for maximizing quality criteria (Chickering 2002).

# Geometric view: search space for three variables

## Example



# GES failure: example

## Example

We put  $N = \{a, b, c\}$ ,  $\forall i \in N \quad X_i = \{0, 1\}$  and consider the database

$$D : \quad x^1 \equiv (0, 0, 0), \quad x^2 \equiv (0, 1, 1), \quad x^3 \equiv (1, 1, 0), \quad x^4 \equiv (1, 0, 1).$$

Possible  $n$ -repetition is given by  $x^{i+4t} \equiv x^i$  for  $t = 1, \dots, n-1$  and  $i = 1, 2, 3, 4$ .

The hyperpotential  $\alpha : X_N \rightarrow (0, +\infty)$  will be  $\alpha \equiv 1$ .

In fact, this is a kind of so-called BDeu-metric (Heckerman *et al.* 1995).

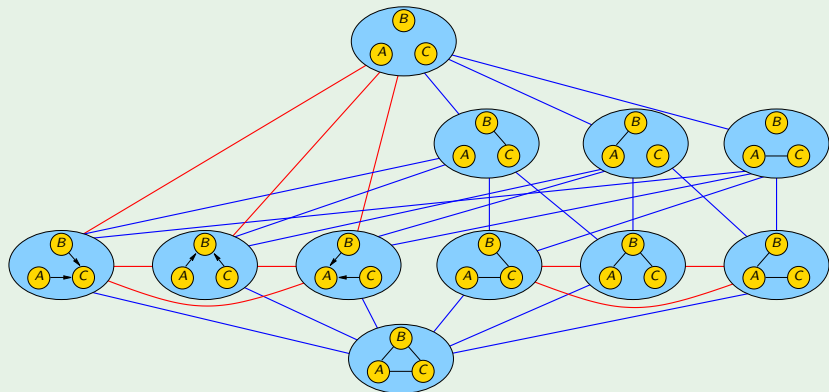
$$\text{LML}[\alpha] (\text{empty}) - \text{LML}[\alpha] (\text{one-edge}) = \ln \frac{125}{99} > 0,$$

$$\text{LML}[\alpha] (\text{immorality}) - \text{LML}[\alpha] (\text{empty}) = \ln \frac{99}{80} > 0.$$

This also happens for any  $n$ -repetition,  $n \geq 1$ .

# GES failure: picture again

## Example



## GES failure: reason

Thus, the well-known *greedy equivalence search* (GES) algorithm (Chickering 2002), which starts in the empty graph and search for the increase in the criterion  $Q$  within inclusion neighbors, gets stuck in the empty (essential) graph.

Nevertheless, the global maximum of  $Q = \text{LML}[\alpha]$  is achieved in any of the immoralities and any of these graphical models gives better explanation of the occurrence of  $D$ .

The reason for this phenomenon is that the database “generated” from a distribution which is **not** perfectly Markovian with respect to any  $G \in \text{DAGS}(N)$ .

This cannot happen if the maximization technique is based on the geometric neighborhood. Then one is guaranteed to find the global maximum of the criterion  $Q$  over BN structures.

## Some open questions

The formula for the LML data vector opens further research topics:











- What is the asymptotic behavior of the data vector?
- The question of statistical consistency of Bayesian quality criteria should be re-examined.

As concerns the geometric neighborhood, the conjecture that it always contains the inclusion neighborhood has recently been confirmed.

- Is it possible to find the “polyhedral” description of the standard imset polytope  $P$  for arbitrary  $|N|$ ?
- How “dense” the geometric neighborhood is, that is, what is the “average” number of geometric neighbors of a given BN structure? What are differential imsets for geometric neighbors?

These questions concern the complexity of a potential greedy search procedure for maximization of a quality criterion  $Q$  based on the geometric neighborhood.

# Some relevant literature

-  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.
-  R.G. Cowell, A.P. Dawid, S.L. Lauritzen and D.J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York.
-  A.P. Dawid (1984). Statistical theory: prequential approach. *Journal of the Royal Statistical Society A* **147**: 277-305.
-  D. Heckerman, D. Geiger, D.J. Chickering (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* **20**: 194-243.
-  S.L. Lauritzen (1996). *Graphical Models*. Clarendon Press, Oxford.
-  A. Schrijver (1986). *Theory of Linear and Integer Programming*. John Wiley, Chichester.
-  M. Studený (2005). *Probabilistic Conditional Independence Structures*. Springer-Verlag, London.
-  M. Studený and J. Vomlel (2008). A geometric approach to learning BN structures. To appear in Proceedings of PGM 2008.
-  M. Studený (2008). Mathematical aspects of learning Bayesian networks: Bayesian quality criteria. A research report in preparation.